2012-08-25

Investigating circulating microRNAs with next generation sequencing

Author: Mikael Lindberg

Onsjögatan 1A Lgh 0901
222 41 Lund

Supervisor: Carlos Rovira

Examiner: Torbjörn Säll

External reviewer: Marita Cohn

Master's degree project in Molecular Biology, 45 hp, 2011-2012

Department of Oncology, Kamprad, Lund University

Abstract

In recent years focused research aimed at finding new specific and sensitive biomarkers for detection, classification and, monitoring of human cancers has been ongoing. MicroRNAs (miRNAs) are small, 19-24 nucleotides, noncoding RNAs that are stably expressed in body fluids such as blood. These circulating RNAs have shown great promise as biomarkers, because the unique expression patterns strongly correlate with certain human diseases, including various types of cancer. These miRNAs act mainly through down regulation of specific target mRNAs in the cytosol and have been shown to be key regulatory players in several important cellular processes. In this project sequence data of circulating small RNAs, 18-30 nucleotides, from blood serum were analyzed and compared to sequence data from blood cells and two blood cancer cell lines. The results showed that a surprisingly large part of the small RNAs in human serum did not match the human genome. Moreover, the results indicate the existence of several novel miRNAs in the human genome. When comparing the most frequent RNA molecules, differences between the compared samples were striking and several well described miRNAs showed up amongst these differently expressed RNAs. Despite huge progress in recent years, there are still more research needed in this field; most likely several novel miRNAs are still to be found in the human genome and, the involvement of certain miRNAs as key regulators in important cellular processes are to be expected. In some years to come, probably some miRNAs will prove to be reliable biomarkers for some specific trait.

Introduction

The field of small regulatory RNA molecules is a relatively young discipline. Twenty years ago, in 1993, Victor Ambros and Gary Ruvkun published two papers describing the regulatory roles of small RNA molecules involved in the developmental control of the nematode *Caenorhabditis elegans* (Lee *et al.* 1993, and Wightman *et al.* 1993). They called these molecules microRNAs (miRNAs) and suggested that they are able to control and regulate the expression of target genes by interacting with the 3'UTR (untranslated region) of their target mRNAs. In 2006, Andrew Fire and Craig Mello received the Nobel Prize in physiology and medicine for their discovery of the process named RNA interference (RNAi) (Fire *et al.* 1998, and Martin-Muller *et al.* 2010). In short, it describes how double-stranded RNA triggers suppression of gene activity in a homology-dependent manner. Biochemical dissection of the mechanisms behind miRNA and RNAi revealed that these were two closely connected cellular processes. In 2001, a breakthrough came revealing a large number of microRNAs (Lau *et al.* 2001; Lagos-Quintana *et al.* 2001; Lee *et al.* 2001). Since then, a lot of new miRNAs have been reported, and at present there are 2,237 unique mature human miRNAs in the Sanger Centre miRBase Database (release 19), the central repository of miRNA sequences (miRBase 2012, and Griffiths-Jones *et al.* 2006).

The biogenesis and nomenclature of microRNAs

The miRNAs can be positioned in introns or in exons within genes and they can also be positioned independently of any protein coding gene with its own promoter. miRNA-like small RNAs have even been found within tRNA genes (Pederson *et al.* 2010, and Haussecker *et al.* 2010), snoRNAs (Ender *et al.* 2008) and vault RNAs (Persson *et al.* 2009). Some miRNA are organized into clusters and transcribed as a long polycistronic precursor. Most known miRNA to date are transcribed by RNA polymerase II (Lee *et al* 2004). The transcript is called pri-miRNA, can vary in size and contains the characteristics of RNA polymerase II transcripts e.g. a poly-A tale and a 5'cap. The pri-miRNA is folded back to create a hairpin stem-loop structure. These precursors are spliced in the nucleus by an RNase III enzyme called Drosha to create a ~70 nt pre-miRNA. Drosha forms a small "microprocessor" complex that recognizes the double stranded miRNA with the help of several proteins, among others a protein called DGCR8 (also known as Pasha) (Davis *et al.* 2010). The stem-loop structure is kept intact whereas the 5' and 3' ends are trimmed by Drosha (Davis *et al.* 2010). The pre-miRNA is transported to the cytoplasm by exportin-5, a process that is guanosine triphosphate (GTP)-dependent (Martin-Muller *et al.* 2010).

In the cytoplasm, a second RNase III enzyme named Dicer cuts the pre-miRNA to create the ~22 nucleotide (nt) mature miRNA and its complementary strand, the star sequence (*), disrupting the stem-loop. The mature miRNAs are single-stranded RNA molecules that act as posttranscriptional regulators of gene expression. The nomenclature of miRNAs states that a mature miRNA is written miR-… and the complementary star sequence is written miR-…*, where the dots are the specific number of the miRNA. Sometimes only one molecule is active and sometimes both have targets which they act on. The two molecules can also be called 3p and 5p, distinguishing them by the means of from which end of the pre-miRNA stem-loop they originate from, the 3' or 5' end (Ambros *et al.* 2003). There are some microRNAs, like the let-family, that do not follow the common nomenclature because the names of these microRNAs were already established when the nomenclature was introduced.

The miRNAs are identified and incorporated in the RNA interference silencing complex (RISC). The RISC consists of several different proteins which differ in different species. The most important RISC

component is the argonaute protein which binds the miRNA double strand. In humans there are eight different argonaute proteins (Mallory *et al.* 2010). They seem to have various roles in the silencing machinery. One of the miRNAs (the miR or miR*) is chosen and a complementary target mRNA is identified. Which mRNA to target is decided by the seed sequence of the miRNA. The seed sequence consists of the 2-8 nt from the 5' end (Martin-Muller *et al.* 2010). Depending on the seed sequence miRNAs are divided into families which are predicted to have the same targets.

Following identification of the target mRNA one of several things can happen. If the complementarity of the seed sequence and the target mRNA is perfect, the mRNA is cut by the argonaute protein (Ago2) and the RISC can bind another mRNA. If the seed sequence is not of perfect match the mRNA is not cut but instead the RISC complex hinders ribosomes from binding and translation to begin (Lim *et al.* 2005). These two actions are the most common actions and are both examples of negative regulation of mRNAs. Which action a specific miRNA will have on its target mRNA is not known. The RISC with the miRNA/mRNA complex is believed to be located into processing bodies (P-bodies). P-bodies are cellular structures in the cytosol enriched in mRNA-degradation enzymes (Martin-Muller *et al.* 2010). As mentioned above most miRNAs act as negative regulators of mRNAs but there have also been reports of positive regulating miRNAs. For example, miR-10a has been shown to bind to the 5'UTR of mRNAs encoding ribosomal proteins and enhancing their translation (Anderson *et al.* 2008).

One miRNA can have several different target mRNAs and a mRNA can be targeted by several different miRNAs (Lim *et al.* 2005). Predictions on the human mRNA estimate that more than 60% are potential targets for miRNA regulation (Freidman *et al.* 2009). These estimates are made on target sites in the 3'UTR of the mRNAs because most miRNAs has been reported to target this region. Some miRNAs can also target inside the coding sequence of their target mRNAs and as mentioned earlier, miR-10a binds its targets in the 5'UTR (Anderson *et al.* 2008). With this in mind the amount of mRNAs under the regulatory control of miRNA might even be higher than 60%.

MicroRNAs circulating in the blood

MiRNA was first described circulating in the human blood serum and plasma in 2008 (Chen *et al.* 2008, Chim *et al.* 2008, Lawrie *et al.* 2008, and Michell *et al.* 2008). In these studies it was shown that miRNAs were remarkably stable and in some way protected from endogenous RNase activity in the blood. Since then, several papers have been published in the attempt to describe these circulating miRNAs. The miRNA profiles have been shown to differ extensively between healthy individuals (with gender, age, pregnancy etc.) and between different medical conditions, e.g. different types of cancers and, diabetes (Zen and Zhang 2010). Human microRNA-genes have also been shown to frequently being located at fragile sites and genomic regions involved in various cancers (Calin *et al.* 2004).

The fact that circulating miRNAs have shown to be highly stable in plasma and are relatively easy to measure with great sensitivity, combined with the strong association with disease states, have made circulating miRNAs strong candidates as a new class of biomarkers. The underlying reason for the high stability of miRNAs in plasma is not known, but it has been suggested that they are protected by encapsulations into lipid vesicles such as exosomes. There are several different types of lipid vesicles circulating in the blood and some are called microvesicles (MVs). Recently, MVs have been shown to shuttle specific miRNAs promoting proliferation and mobility between macrophages and breast cancer tumors (Yang *et al.* 2011). However, in a recent paper evidence is presented showing that a majority of circulating miRNAs are associated with the Ago2 protein in the blood and only a minority are associated to MVs and other lipid vesicles (Arroyo *et al.* 2011).

The origin of the circulating miRNAs in the blood is not known; with some exceptions of miRNAs with a known tissue specific origin, e.g. miR-122, which is expressed by liver cells (Jopling 2012). In a recent study it was shown that out of 79 previously reported solid tumor circulating miRNAs, 58 % were highly expressed in one or more blood cell type (Pritchard *et al.* 2012). Consequently they argue that the origin of these miRNA could be from blood cells and not the tumors. To this can be added that the miRNA population in blood cells has been shown to shift and reflect the presence of diseases that are not necessarily blood-borne diseases (Keller *et al.* 2011). The underlining cause of this change is not known but it could be as a response to changes in the environment caused by the disease, the authors of these findings argue.

Experimental techniques, next generation sequencing

There are three methods commonly used to detect miRNAs: next generation sequencing, microarrays, and reverse transcription quantitative polymerase chain reaction (RT qPCR). Next generation sequencing refers to the DNA sequencing technologies developed since 2005 that have reduced the cost and increased the data throughput (Nelson *et al.* 2011). Next generation sequencing is also referred to as high-throughput sequencing or deep sequencing.

With the introduction of next generation sequencing the science community realized that most miRNAs exist in various lengths. A specific miRNA can vary in length due to some variation in processing by Drosha and Dicer and/or can exist with several different additions both in the 3' and 5' ends. Nucleotide substitutions have also been reported. First variation in the 3' end was coined isomiRs (Morin *et al.* 2008). Later this term has been broadened to also include variations in the 5' ends of miRNAs. The variations in the 5' end changes the seed sequence of the miRNA and has thereby been controversial because the predicted targets are also changed. Often one form is the most common one and then this form is the one that is described in miRBase. In a recent paper, Lee and coworkers presented a comprehensive study over isomiRs from human and mouse samples (Lee *et al.* 2010). They argue that the heterogeneity amongst miRNAs are nonrandom and does not originate as a defect of the sequencing technique which was previously raised as a concern.

The project

In this project the initial plan was to investigate the miRNA profile in blood plasma from breast cancer patients that had established metastases. More specifically, the aim was to identify potential miRNA biomarkers that could indicate in what tissue an aggressive breast cancer would metastasize. The technique to be used to map the miRNA was next generation sequencing and the patients were to be matched to healthy controls. However, due to complications in the laboratory, no final sequence data was obtained even if all the experimental steps (RNA isolation, quality control, sample preparation for next generation sequencing and the actual sequencing) were performed by me in the lab. The final data analysis of sequence data were performed on fourteen datasets publically available in the Gene Expression Omnibus (GEO) database. The datasets originate from two papers published in 2010 and 2012 (Vaz *et al.* 2010 and Zhang *et al.* 2012). The RNA extraction and library creation protocols used in both these experiments were highly similar to the procedure performed and described here. In these studies, ten sequencing datasets were generated from pooled serum samples from healthy Chinese individuals, while two datasets originate from blood cells from two healthy Indian individuals, and two datasets originate from two different cancer cell lines originating from blood cancer cells.

Zhang *et al.* 2012 focused on foreign/non-human miRNAs, e.g. miR-168 from plant, circulating in the blood stream. They presented, through several experiments, the first evidence that exogenous plant miRNA, through food intake, can enter the blood stream, be taken up by various tissues and cells, and there regulate the expression of target genes in mammals. We found that a surprisingly large portion of

the small RNA sequences in the ten serum libraries were non-human, e.g. not aligned to the human genome. This is in line with the findings of Zhang *et al.* 2012. Whereas Zhang *et al.* 2012 run their libraries against all known miRNAs in the miRBase we run their libraries against the entire human genome. Other small RNA molecules than the known miRNAs will be identified when running against the entire human genome. In this way the sequence data that was obtained will differ and giving some scientific relevance to our data analysis.

The other paper, Vaz *et al.* 2010, is a comparison of the miRNA profile of normal white blood cells (leukocytes) and two different cancer cell lines both originating from blood cells. In this study the authors show that there is a difference in miRNA expression patterns, and they also present 370 novel miRNA. They ran their libraries against the human genome in a similar fashion as we did, but at the time of their analysis there were not as many known human miRNAs in the miRBase as of today. Our analysis of the Vaz *et al.* 2010 data confirmed their findings. The potential blood cell origin of circulating miRNAs is being debated in the scientific community, and therefore a comparison between the small RNA profiles of the normal white blood cells and the small RNA profiles from the serum libraries, is of interest. This was the reason for choosing the Vaz *et al.* 2010 datasets.

The two papers and the fourteen datasets are presented and described in more depth in the materials and methods. There, questions are presented on comparison between the libraries and the answers to the questions are later presented in the results. The questions and results that came out of the data analysis are not directly related to the questions on breast cancer metastases, as of the questions in the original plan. Instead, the results highlight some interesting things about circulating miRNA in a more broad sense, like gender and age differences, along with the interesting comparison to the original papers.

Materials and Methods

This section contains a description of techniques, procedures and protocols used according to the initial project plan. Even though, no final sequence data were obtained in this study, the preparations are nevertheless, as mentioned earlier, highly similar to the preparations made in the two papers that generated the datasets used in this report.

When working with RNA, it is of outmost importance to take precautions to avoid exposure to RNase enzymes. Compared to DNases, RNases are much more stable. There are RNases on the skin of humans so gloves should be worn at all times when handling samples, tubes and equipment used for RNA preparation and analysis (Bustin 2004). Special RNase free water was used at all times and the laboratory work was done on lab benches and in rooms dedicated for working only with RNA.

Blood sample collection and preparations

All blood material that was used was blood from the author himself collected at two different dates, 2011-09-30 and 2011-10-28 (4 tubes with 6 ml, in total 24 ml, at both occasions). The blood was first collected in EDTA tubes (K2E 10.8mg Plus Blood Collection Tubes manufactured by BD Vacutainer) and then incubated at room temperature for 30 minutes. EDTA is an anticoagulant that prevents blood from clotting. Following the incubation the tubes were centrifuged at 2,000 x g for 10 minutes at room temperature. After centrifugation the blood is separated in three different layers. The lowest layer consists of red blood cells and on top of this layer are the white blood cells. The upper layer is the

plasma which consists of the cell free blood and is approximately 55% of the volume. The plasma was taken from the tube and put in 1.5 ml microcentrifuge tubes in aliquots of 300 µl. These tubes were stored in a -70°C freezer. Blood plasma is easily confused with blood serum which is the cell free fluid that is left when blood has been allowed to clot. When collecting blood with the aim of recovering serum, other tubes not containing any anticoagulant are used. One could roughly say that serum is plasma without fibrinogen and other clotting factors because these are pelleted along with the cells after clotting and centrifugation. It has been shown that miRNA profiles in plasma and serum, taken from the same individual, are very similar and in this field of research there has not been any decision whether to use serum or plasma as a standard.

Instead of using EDTA tubes sometime Heparin coated tubes are used to collect plasma. Heparin is a naturally occurring polysaccharide found in the artery walls and is an anticoagulant just as EDTA (Kim *et al.* 2012). It has been shown that Heparin interferes with PCR reactions and other downstream applications, and therefore is not recommended when investigating miRNAs and similar molecules. Additional heparin is commonly given to patients who undergo transplantations heart surgery or have had a stroke or similar problems connected to blood clotting (Kim *et al.* 2012). Because of this the kind of medicine given to the patients and controls are of outermost importance when investigating miRNA. It has been shown that heparinized samples treated with heparinase, an enzyme degrading heparin, prior to RT-PCR increase detection of otherwise undetected miRNAs (Kim *et al.* 2012).

Synthetic spiked-in RNA as controls

Synthetic RNA spikes similar in size to miRNAs can be used as controls when investigating and comparing miRNA level in parallel samples. The spikes are added to all samples in an early step of the preparations. By knowing that the concentrations of spikes were the same in the early stage one can in the end analysis use this to normalize the sample data to each other. In this way the data can be compared in a more true sense. The normalization takes away differences in the data that is due to handling differences, i.e. unequal loss of material due to small pipetting errors, incomplete precipitations, pellet recoveries etc.

To each plasma sample (of 300 µl) a spike pool was added containing synthetic RNAs spikes in the size of miRNAs. In total twelve spikes were used, see table 1. Ten of the spikes were 24 nt long and these were diluted in a series by a fifth (starting from 0.01 nM going down to $5.12*10^{-9}$ nM in the spike pool). Two of the spikes were in the same concentration, 0.005 nM, in the spike pool. These two were 19 nt in length and differed in sequence only in the eleventh nt, which was a G in one and an A in the other. In table 1 the synthetic RNAs are presented in name, sequence, length, concentration in the spike pool, and approximate amount of copies added to the plasma sample prior to RNA extraction. The way to estimate the amount of copies was to multiply the spike pool concentrations with the Avogadro constant ($6.02*10^{23}$), the volume added (10 µl = $10*10^{-6}$ liter), and the abbreviation nano (10^{-9}) in front of the molar sign.

Table 1. The sequence and length of 12 synthetic RNAs. The table also presents the concentrations of the spike pool and the approximate amount of copies added to the plasma samples prior to the RNA extraction.

Name	Sequence 5'-> 3'	Length nt	Spike Pool Conc. nM	Approximate amount of copies added
RNA spike 1	GCGGUCCAACGAAUAUCGA	19	0.005	30000000
RNA spike 2	GCGGUCCAACAAAUAUCGA	19	0.005	30000000
Control-11	GAUUAGCGGUUCCGGUAUGGGCAC	24	0.01	60000000
Control-12	AUUCCUACGACGCGGCGUUCGUUG	24	0.002	12000000
Control-13	CGGACGCAACGGACGAUAACAACC	24	0.0004	2400000
Control-14	AUUCUGCGACGUGUUGCGCUCGGA	24	0.00008	480000
Control-15	CAUACGGACGUUUACGCACGUUCA	24	0.000016	96000
Control-16	CCGUGACGUGUACCGCCGCCUAUA	24	0.0000032	19200
Control-17	CCGACGGUACUACAUAGACGACGU	24	0.00000064	3840
Control-18	CUACGCACGACUUAGUCGUUCGAG	24	0.000000128	768
Control-19	UCGGAUCGUUAUAGGUAUUGCGCA	24	2.56E-08	154
Control-20	CGCAUCACGUCCGUUGCGUUAGAU	24	5.12E-09	31

The spikes were phosphorylated in the 5'-end in a reaction with T4 Polynucleotide Kinase gene PNK (NEB, New England BioLabs). This was done to allow subsequent ligation of adapters on the spikes later on in the creation of libraries. The reaction setup were a phosphorylation step at 37°C for 30 minutes, followed by an enzyme heat inactivation step at 60°C for 20 minutes according to the manufacturer's instructions.

Test runs were made on plasma containing different dilutions of the spike pool. The plasma samples used in these tests were from the collection described above and the RNA extraction and library creation followed the procedure described in the forthcoming text. The reason for running these tests were to practice and validate the program and to figure out what concentration of the spike pool to add. The spikes should be in concentrations that are detectable in all samples, but should not be a too big part of the overall reads in the sequence data.

The spikes were ordered from a company called Integrated DNA Technology (IDT) (web site: http://eu.idtdna.com/site). The spikes were tested and validated prior to use with quantitative PCR (qPCR) to verify their detectability and stability in plasma. The qPCR method that was used is called miR-specific quantitative RT-PCR with DNA primers (Balclls *et al.* 2011). In short the method uses a poly(A) polymerase to extend the 3' end of the targets, followed by a cDNA synthesis with a reverse transcriptase and an anchored poly(T) primer. Then the qPCR reaction is run with SYBR-Green, a target specific forward primer and a reverse primer. In qPCR, the amplified DNA product is measured after each round of amplification with the use of the SYBR-Green molecule that fluoresces when binding to double stranded DNA (dsDNA). Since the amplification in theory is equal for different starting materials, the amount of end product is proportional to the starting material. Consequently, in theory one could calculate the starting amount of a specific amplified product and compare different amounts of different samples. None of the spike validation data has been included in the results of this report.

Total RNA extraction and isolation

In order to create libraries for sequencing, one must start by extracting and isolating the total RNA in the plasma samples. There are several techniques used to extract RNA from biological samples. There is not one specific method that has been proven to be much superior to the other. The technique used her is based on phase separation whit phenol and chloroform, followed by precipitation of the RNA with ethanol. TRIzol® LS reagent provided by Invitrogen, which is designed for liquid solutions, was

the phenol reagent that was used. The following text describes the extraction and isolation protocol of total RNA.

Each plasma sample tube contained 300 µl. To each tube 900 µl TRIzol LS reagent (3:1 volumes) was added and the samples were homogenized with a pipette. Then 10 µl of the spike pool were added to each tube and the tubes were incubated 10 minutes in room temperature for total disassociation of RNA-protein complexes. Thereafter 240 µl of absolute chloroform (99%) were added and the tubes were shaken vigorously 15 seconds and incubated 2-3 minutes at room temperature. The tubes were centrifuged at 12,000 x g for 15 min at 4°C. The samples separate into three phases: a transparent upper aqueous phase containing RNA, a white thin interphase containing DNA, and a red-pink lower phenol phase. The aqueous phase was carefully transferred to a fresh tube. Then 1 µl glycogen (20µg/µl) and 600 µl absolute ethanol (99%) was added to each tube. The tubes were inverted several times to mix, incubated 15 minutes at room temperature, and then centrifuged at 12,000 x g for 30 minutes at 4°C. The supernatants were carefully removed and the pellets were washed with 1.2 ml 75% ethanol by centrifugation at 12,000 x g for 10 minutes at 4°C. The supernatants were removed carefully and the pellets were air-dried 5-10 minutes in room temperature. Then the pellets were dissolved in 10 µl nuclease-free H_2O and the samples were stored in -70°C for later use.

A NanoDrop was used to measure the concentration and quality of the samples. NanoDrop is a photometrical technique that is widely used to measure concentrations of nucleic acids and proteins. A small volume (1 µl) of the sample is put on the NanoDrop and the absorbance at different wavelengths in the ultraviolet light spectrum is measured. The A260/A280 ratio should be 1.8-2 and the A260/A230 ratio should optimally be 2 or above for RNA. A lower A260/A280 ratio may be caused by contaminating DNA or protein. A lower A260/A230 ratio than 2 may be caused by traces of contaminating phenol, which could interfere with downstream experiments (Bustin 2004).

Pre-sequencing preparation and cDNA library construction

In order to run the sequencing on the samples, libraries needed to be constructed. The protocol used was a modification of the "Small RNA v1.5 Sample Preparation Guide" from Illumina. The procedure is described in detail in the following text and a schematic overview is presented in figure 1.

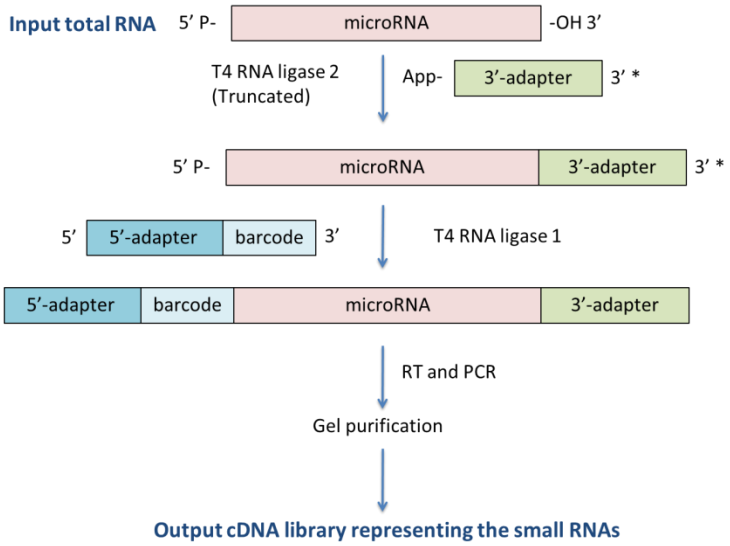

Figure 1. Schematic illustrative overview of the preparations and construction of small RNA libraries for sequencing. The star on the 3'-adapter indicates a chemical modification that eliminates ligation on this end of the adapter.

When constructing libraries for sequencing, one can choose amongst several different approaches. Since the modern sequencing techniques read huge amounts of sequences, one can pool several libraries and still get good coverage of all sequences in each library. To be able to pool libraries and still be able to distinguish them, a unique barcode is incorporated which is sequenced along with the target sequences. This technique is referred to as multiplexing. We chose an approach where we used one 3'-adapter, the same for all, and different 5'-adapters with unique barcodes for each library.

3' and 5' adapter ligations

The first step in the creation of libraries is to ligate adapters on to the ends of all the isolated RNA molecules. The 3' and 5'-adapters are ligated in two separate reactions using two different T4 RNA Ligases. In the 3'-adapter ligation step the ligase used was T4 RNA Ligase 2 truncated (NEB, New England BioLabs). This enzyme specifically ligates the pre-adenylated 5' end of DNA or RNA to the 3' hydroxyl-terminated end of a RNA molecule. The enzyme does not use ATP for ligation but requires pre-adenylated linkers. The 5'-adapter was ligated with T4 RNA Ligase 1 (NEB, New England BioLabs) which catalyzes the ligation of a 5' phosphoryl-terminated DNA or RNA donor to a 3' hydroxyl-terminated DNA or RNA acceptor with the hydrolysis of ATP to AMP and PP_i.

The adapters used are presented in table 2. The first adapter in the table is the 3'-adapter provided by Illumina. This primer was used in all preparations and is a RNA molecule. The "APP" in the 5' end indicates that the adapter is pre-adenylated which is necessary in order for the ligation to work in the absence of ATP, improving the ligation efficiency avoiding self-ligation or circularization of the small RNAs when using truncated T4 RNA ligase. The "idT" in the 3' end indicates that the last nucleotide is modified, preventing ligations in this end and adapter-dimer formations. The next eight adapters are 5'-adapters with barcodes in the 3' end. These adapters are built up by both RNA and DNA. The last 5 nucleotides are RNA including the barcode, the barcode is highlighted in red and the two other RNA nucleotides are highlighted in green in figure 2. The rest of the adapter is DNA and is highlighted in blue. These eight adapters were designed by the group and ordered from a company called Eurofins MWG Operon (web site: www.eurofinsdna.com). The first four 5'-adapters (number 1-4) have two identical RNA nucleotides then the three letter barcode. The next four 5'-adapters (number 5-8) are two nucleotides longer and have the three letter barcode first then two identical RNA nucleotides. The reason for this difference was that 5'-adapters 1-4 was designed and ordered in an earlier stage. Concerns that the ligations efficiency could differ due to the different nucleotides in the 3' end lead to the design of 5'-adapters 5-8 where the last two nucleotides are the same for all adapters (A and U).

Table 2. The 3' and 5'-adapters used in the preparation of libraries for sequencing. The barcodes in the 5'-adapters are highlighted with red color, the non-barcode RNA nucleotides are in green and, the DNA nucleotides are in blue.

Name	Type	Sequence 5'-> 3'	Length nt
v1.5 sRNA 3' Adapter	3'	APP-UCGUAUGCCGUCUUCUGCUUGUidT	22
GA_ACA_5adapter_1	5'	GTTCAGAGTTCTACAGTCCGACGAUCACA	29
GA_CAU_5adapter_2	5'	GTTCAGAGTTCTACAGTCCGACGAUCCAU	29
GA_GUA_5adapter_3	5'	GTTCAGAGTTCTACAGTCCGACGAUCGUA	29
GA_UGU_5adapter_4	5'	GTTCAGAGTTCTACAGTCCGACGAUCUGU	29
GA_AGC_5adapter_5	5'	GTTCAGAGTTCTACAGTCCGACGATCAGCAU	31
GA_AGC_5adapter_6	5'	GTTCAGAGTTCTACAGTCCGACGATCCUGAU	31
GA_AGC_5adapter_7	5'	GTTCAGAGTTCTACAGTCCGACGATCGACAU	31
GA_AGC_5adapter_8	5'	GTTCAGAGTTCTACAGTCCGACGATCUCGAU	31

The procedure of ligation was as follows: 5.0 µl of total RNA in nuclease-free water was mixed with 1.0 µl of the 3'-adapter and preheated to 70°C for 2 minutes to disrupt any secondary structures and

dimerization. Next a master mix was prepared, for one sample: 1.0 µl 10X T4 RNA Reaction Buffer, 0.8 µl 100 mM $MgCl_2$, 1.5 µl T4 RNA Ligase 2 Truncated (New England BioLabs), 0.5 µl RNase Inhibitor. The volumes of each reagent were multiplied by the numbers of samples being prepared and 10% extra reagent was included to cover for pipetting errors. The master mix was properly mixed by pipette and 3.8 µl of the mix was added to the preheated RNA and adapter mix (total volume 9.8 µl). The ligation mix was incubated on a preset thermal cycler (Bio-Rad C1000) at 22°C for 1 hour. The 3'-adapter (v1.5 sRNA 3' Adapter see table 2) that was used has a modification in the 3' end which stops any ligation to this end and eliminates adapter dimer ligations in this step. The next step was the 5' ligation. First the 5'-adapters were preheated to 70°C for 2 minutes. Then 1.0 µl of the 5'-adapter was added to the ligation mixture along with 1.0 µl 10mM ATP and 1.0 µl T4 RNA Ligase 1 (New England BioLabs). The ligation mixtures were properly mixed with a pipette and incubated at 20°C for 1 hour followed by hold at 4°C. The samples were directly preceded to cDNA synthesis and PCR amplification. Leftovers and backup were stored at -70°C.

cDNA synthesis and PCR amplification

The ligation products were converted to cDNA in a reverse transcription (RT) reaction. A volume of 4.875 µl of the 3' and 5'-adapter ligated RNA was mixed with 1.0 µl of the RT primer (see table 3) and heated to 70°C for 2 minutes then put on ice. A RT master mix was prepared, for one sample: 1.0 µl 10X MulV RT reaction buffer, 0.625 µl 10 mM dNTP Mix, 1.0 µl 100 mM DTT, 0.5 µl RNase inhibitor, 1.0 µl MuLV Reverse Transcriptase (Applied Biosystems). The volumes of each reagent were multiplied by the numbers of samples being prepared and 10% extra reagent was included to cover for pipetting errors. The master mix was properly mixed by pipette and 4.125 µl of the mix was added to the preheated adapter ligated RNA and RT primer (total volume 10 µl). The mixture was put in a thermal cycler (Bio-Rad C1000) with the reaction conditions: 42°C for 1 hour, then 70°C for 5 minutes for enzyme inactivation.

The cDNA were amplified in a polymerase chain reaction (PCR). A PCR master mix was prepared, for one sample: 30.75 µl Nuclease-free water, 5.0 µl 10X Expand High Fidelity Buffer, 1.0 µl Primer GX1, 1.0 µl Primer GX2, 1.5 µl 10 mM dNTP Mix, 0.75 µl DNA Polymerase Expand High Fidelity PCR System (Roche). The volumes of each reagent were multiplied by the numbers of samples being prepared and 10% extra reagent was included to cover for pipetting errors. The master mix was properly mixed by pipette and 40 µl of the mix was transferred to a nuclease-free 200 µl PCR tube. Then 10 µl of the single stranded reverse-transcribed cDNA were added, resulting in reaction volumes of 50 µl. The PCR reaction were performed on a thermal cycler (Bio-Rad C1000) under the following cycle conditions: 98°C for 30 seconds followed by 15 cycles of; 98°C for 10 seconds, 30°C for 30 seconds, 72°C for 15 seconds; then 72°C for 10 minutes, and hold at 4°C. The RT primer and the PCR reverse and forward primers, presented in table 3, were all manufactured by Illumina. The Forward primer has an overhang in relation to the 5'-adapter which it compliments to. This results in an addition in length in the PCR reaction.

Table 3. The three primers by name, type, sequence, length in nt, and concentration used in the RT-PCR and PCR reactions. The reverse transcription (RT) primer and the reverse PCR primer have the same sequence i.e. they are identical.

Name	Type	Sequence 5'-> 3'	Length nt	Concentration
SAR RT	RT	CAAGCAGAAGACGGCATACGA	21	Not known
GX1	Reverse	CAAGCAGAAGACGGCATACGA	21	25 µl
GX2	Forward	AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA	44	25 µl

Gel purification of cDNA constructs

After the PCR the DNA libraries were purified using a polyacrylamide gel electrophoresis (PAGE) gel. The PAGE gel used was a 6% TBE gel. The runtime was 26 min and the voltage was 180 V.

In the process of ligation a byproduct of adapter dimers are formed (one 5'-adapter is ligated with one 3'-adapter). The gel separation is to exclude the adapter dimers and reduce their appearance in the sequencing data later on. The gel purification step also allows for size separation upwards, i.e. enabling exclusion of longer fragments.

In the gel purification step only those bands that matched the desired lengths are to be cut out. In the PCR reaction the forward primer adds 23 nt to the cDNA because of the overhang. The adapter dimer is approximately 74 or 76 nt (23 added from the Forward Primer GX2 + 22 from 3'-adapter + 29 or 31 from 5'-adapter). The bands containing the 18-60 nt RNA fragments with both these adapters are a total of 92-94 to 134-136 nt in length. The band containing 22 nt RNA fragment with both adapters are 96-98 nt in length. The band at 125 nt is containing 48 or 50 nt RNA fragments, i.e. fragments longer than miRNAs. The sequencer used, and described further down in the text, are limited to sequences with lengths up to approximately 2 x 100 bp. So, to use the full capacity of the sequencer one can choose the bands that represent sequences that are approximately 150 bp long because the adaptors add on some nucleotides also.

Gel-pieces containing the constructs of the desired lengths were cut out with a sterile scalpel and put in a sterile nuclease-free 0.5 ml microcentrifuge tube sitting inside a 2.0 ml microcentrifuge collecting tube. The 0.5 ml tubes had been punctured in the bottom from the tube opening, 3-4 times with a sterile 21-gauge needle. The stacked tubes were centrifuged in a benchtop microcentrifuge at 14,000 rpm (approximately 20,000 x g) for 2 minutes at room temperature. During the centrifugation the gel is forced through the holes and shred into smaller pieces. Shredding the gel in this fashion makes the gel pieces similar in size for parallel samples and the smaller pieces enhances the elution of the DNA constructs. The next step was to discard the upper 0.5 ml tube and add 250 µl 1X Gel Elution Buffer, provided by Illumina, onto the gel debris in the 2 ml tube. The DNA constructs were eluted under gentle rotation in a fridge (+2 to +8°C) overnight.

The eluate and the gel debris were transferred to the top of the filter of a Spin-X cellulose acetate filter tube and centrifuged in a benchtop microcentrifuge at 14,000 rpm (approximately 20,000 x g) for 2 minutes at room temperature. The filter was discarded and 1 µl of glycogen (20µg/µl) was added along with 25 µl of 3M NaOAc, and 975 µl of pre-chilled, -15° to -25°C absolute ethanol (99%) to each Spin-X tube. The glycogen and NaOAc cause the DNA to precipitate in absolute ethanol and cold. The tubes were put in -20°C for 1 hour then centrifuged at 14,000 rpm (approximately 20,000 x g) for 30 minutes at 4°C. The supernatants were discarded and the pellets were washed in 500 µl of room temperate 70% ethanol. Then the tubes were centrifuged at 14,000 rpm (approximately 20,000 x g) for 15 minutes at 4°C. The supernatants were discarded and the pellets were left intact. The samples were set to air-dry for 5 minutes then the pellets were resuspended in 10 µl Nuclease-free water.

Library concentration measurement with Qubit fluorometer

A Qubit® fluorometer from Invitrogen was used to measure the concentrations of the libraries. The Qubit is much more accurate than using NanoDrop on low concentrations. When measuring on higher concentrations the techniques are equally good. This is because in the Qubit quantitation platform use molecule-specific fluorescent dyes to measure the concentration of the specific molecule of interest. The NanoDrop technique relies on UV absorbance, which cannot completely distinguish between DNA, RNA, degraded nucleic acids, free nucleotides, and other contaminants. With the Qubit

different assays can be used, with dyes targeting specific types of molecules. The dsDNA High sensitivity assay (assay range: 0.2-100 ng) was used to measure the concentrations of the libraries. The concentrations of the libraries ranged between 10-25 ng/µl.

Sequencing

The sequencing was done in several steps. The following text describes the denaturation, dilutions, clustering and, sequencing steps.

Dilutions and sample preparations prior to cluster generation

First the samples need to be diluted several times prior to cluster generation since only tiny amounts, in the range of pM, are needed. The desired final concentration to use depends on the density of clusters that is desired. 7 pM usually generates 425 000 clusters/mm^2 and was the concentration that was chosen. The concentrations of the DNA are in molar. Therefore the concentration form the Qubit fluorometer, an estimate of an approximate length of the sequences, and an average weight of the nucleotides was used to calculate the molarity. Because the libraries have different barcodes they can be pooled and run together and still be distinguishable from each other.

The five libraries were diluted first to 10 nM, then to 2 nM in volumes of 10 µl each. Thereafter 5 µl of each library were pooled to one sample of 25 µl containing 2 nM pooled cDNA library construct (the DNA template sample). The solution that was used for the dilutions was called Buffer EB (Elution Buffer) from Qiagen (Lot No.: 130173084). Instead of using this buffer one could use 10 mM Tris-Cl, pH 8.5 + 1 % Tween 20.

The next step in the preparations was to denature the DNA template, i.e. separate double stranded DNA. This was done with 10 µl 0.1 M NaOH added to 10 µl of the 2 nM DNA template in a 1.5 ml microcentrifuge tube. The tube was mixt by vortex and spun down followed by 5 minute incubation in room temperature. Next all the denatured DNA (20 µl) was transferred to a 1.5 microcentrifuge tube containing 980 µl pre-chilled Hybridization buffer (HT1) from Illumina and was put on ice. This dilution created a 20 pM denatured DNA template stock.

To obtain the final concentration of 7 pM, 350 µl of the 20 pM denatured DNA template stock was mixed with 650 µl pre-chilled HT1 solution.

As a control in the cluster generation a synthetic DNA called PhiX provided by Illumina was used. PhiX is fragments of a bacteriophage (viral) genome and have no similar sequence in the human genome. The PhiX comes in a 10 nM stock and was prepared and diluted in similar manner as described above to a concentration of 7 pM. From this sample 10 µl 7 pM PhiX was added to the 1000 µl denatured DNA template with the concentration 7 pM.

Loading the DNA template on the flow cell and cluster generation

Next from the 7 pM denatured DNA template sample 120 µl was used to load in one lane on a flow cell. The flow cell was placed on a cluster station from Illumina, called a cBot and clusters were generated. A flow cell, see figure 2, contains eight lanes/channels and each lane is built up by three columns. Each of the three columns, in turn, consists of 100 squares tiles. The surfaces of these tiles are coated with single stranded primers that are complementary to the end part of the ligated adapters used in the library creation. The 3' and 5' ends are different from each other but the same for all the sequences in the samples so the attached primers are of two types. The first step in the cluster creation is that the samples are loaded on the flow cell and the DNA sequences attach to the surface. As mentioned above, the samples were previous to this diluted, so when the sequences attach to the surface they are spread out over the tile areas. The surface primers are placed with fixed distances

from each other to allow for the attached DNA sequence to bend over and bind to an adjacent primer. This is what happens next, i.e. the DNA sequences bends over and bind to the primers. Unlabeled nucleotides are added and amplification synthesis is initiated, starting from the primer. This phase is referred to as solid-phase bridge amplification. When the synthesis is completed the strands are denatured. Now there are two sequences complementary to each other and both are attached to the tile area. Next several cycles of solid-phase bridge amplification are performed so that clusters containing huge amount of identical and complementary sequences are generated. In this fashion, all sequences in the samples create individual clusters separated with some distance depending upon what concentrations are used. As mentioned in the previous text using a concentration of 7 pM usually generates 425 000 clusters/mm^2. Each cluster contains one unique sequence, and half of the sequences in one cluster are attached in its 3' and half are attached in its 5' end. The clustering took approximately 4 hours.

Figure 2. Illustration of a flow call from Illumina. The flow cell in the picture is in approximately the original size. The flow cell has eight separate lanes.

The sequence reaction on the HiSeq 2000

After the clustering was completed the next step was the actual sequencing which was done on a next generation sequencer HiSeq 2000 from Illumina. The flow cell was placed in the sequencer and reagents were added. The sequencing is performed in several cycles. One cycle starts with the addition of reagents, including all four labeled nucleotides, a DNA polymerase and a primer. The primer that was used is presented in table 4. The fluorescent molecule is attached to the base and each base (A, C, G and T) has its unique color. All the nucleotides contain a block in the 3' end hindering more than one nucleotide to be incorporated at the time. The first base is incorporated and unincorporated bases are removed. The fluorescent signal is detected. Next the fluorescent molecule is removed from the base and the 3' block is removed. Then the cycle is repeated by the addition of new reagents. The run took approximately 48 hours.

Table 4. The Small RNA Sequencing Primer from Illumina used in the sequencing reaction.

Name	Sequence 5'-> 3'	Length nt
Small RNA Sequencing Primer	CGACAGGTTCAGAGTTCTACAGTCCGACGATC	32

Data analysis

Since the sequencing of our samples failed to produce data, all datasets analyzed were downloaded from the Gene Expression Omnibus (GEO) (website: http://www.ncbi.nlm.nih.gov/geo/). The GEO is administrated by the National Center for Biotechnology Information (NCBI) and is a database that was first established in year 2000. From the start the GEO database was meant to be a database containing all published raw data from expression analysis with microarray technology. Later on when sequencing techniques emerged the GEO database adapted to also include these datasets (Barrett *et al.* 2011). The datasets used were fourteen altogether originating from two different publications. One dataset represents one library and all the libraries were prepared in similar ways as described in the text above. Table 5 and 6 below, summarizes the data surrounding the datasets. All fourteen libraries consist of sequences that are between 18-30 nt in length.

Table 5. The table shows a descriptive data summary of ten sequence libraries taken from Zhang *et al.* 2012.

Article: Zhang *et al.* 2012				
GEO reference number: GSE31299				
Individual dataset number	**Description**	**Age (years)**	**individuals in the pool**	**lengths (nt)**
SRR332232	human serum pooled 1 (male)	26.3±1.9	11	~18-30
SRR332233	human serum pooled 2 (female)	24.2±1.3	10	~18-30
SRR332234	human serum pooled 3	50.9±7.9	10	~18-30
SRR332235	human serum pooled 4	50.9±7.9	10	~18-30
SRR332236	human serum pooled 5	50.9±7.9	10	~18-30
SRR332237	human serum pooled 6	50.9±7.9	10	~18-30
SRR332238	human serum pooled 7	50.9±7.9	10	~18-30
SRR332239	human serum pooled 8	50.9±7.9	10	~18-30
SRR332240	human serum pooled 9	50.9±7.9	10	~18-30
SRR332241	human serum pooled 10	50.9±7.9	10	~18-30

Table 6. The table shows a descriptive data summary of four sequence libraries taken from Vaz *et al.* 2010.

Article: Vaz *et al.* 2010				
GEO reference number: GSE19812				
Individual dataset number	**Description**	**Age (years)**	**individuals in the pool**	**lengths (nt)**
SRR039190	Normal Blood cells library 1	-	Not a pool	~18-30
SRR039191	Normal Blood cells library 2	-	Not a pool	~18-30
SRR039192	K562 cancer cell line	-	Not a pool	~18-30
SRR039193	HL60 cancer cell line	-	Not a pool	~18-30

The datasets from the first publication, Zhang *et al.* 2012, consists of ten datasets with the GEO reference number GSE31299. These ten libraries all originate from pools of blood serum samples taken from healthy Chinese individuals. The first library consists of serum from 11 males with mean age 26.3±1.9 years. The second library consists of serum from 10 females with a mean age 24.2±1.3 years. The gender distribution in the other eight libraries were not revealed in the paper and therefore considered as not known. These libraries each consist of serum from pools of 10 individuals. The mean age for all these 80 individuals was 50.9±7.9 years.

The datasets from the second publication, Vaz *et al.* 2010, consists of four datasets with GEO reference number GSE19812. Two of the datasets are libraries created from normal white blood cells form two healthy Indians. The other two datasets are libraries created from two different blood cancer cell lines, K562 and HL60. The K562 cell line was the first human immortalized chronic myeloid leukemia (CML) line to be established and bear some resemblance to undifferentiated granulocytes, i.e. a certain category of white blood cells (Koeffler and Golde 1980). HL60 on the other hand, is a leukemic cell line derived from a human with acute promyelocytic leukemia. This cell line can be chemically stimulated to differentiate to resemble mature granulocytes (Koeffler and Golde 1980). Both cell lines are commonly used in blood related research experiments.

These fourteen libraries were aligned against the human genome (version 64) provided from the website ENSEMBL (www.ensembl.org). The aligning programing was done by a member of our group and is presented in a schematic illustration in figure 3. The figure looks quite complicated but the programing pipeline can be described in short in a few steps. First, all the sequences were stripped of their adapters, filtered and identical reads were clustered together. Then these sequences were aligned against the haploid human genome as well as the human mitochondrial genome. Mismatch in the ends of the sequences were tolerated to constrained limits. This is because additions of nucleotides to the 3' and 5' ends of miRNAs are frequently occurring. Some sequences did not align to the human

genome. Each uniquely sequenced sequence was called a read. Next the aligned sequences were annotated, i.e. information on where in the human genome the read was mapping and what that sequence was known as was collected. In this annotation all the human miRNAs in miRBase were used. The sequences were divided on whether they did or did not annotate. Then information on differential expression of each sequence was registered, i.e. different additions in both the 3' and 5' ends of the sequences.

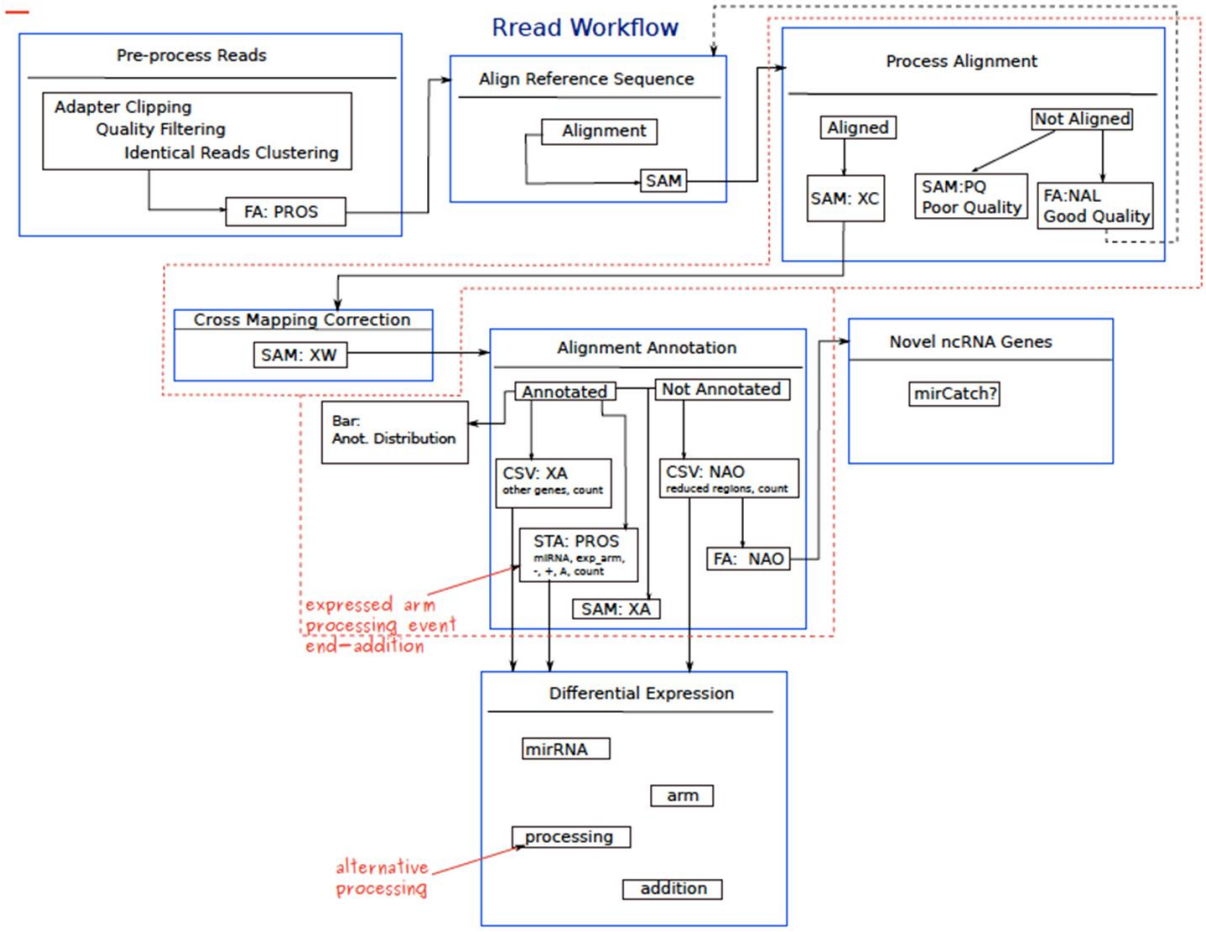

Figure 3. The programing pipeline that was used to align the sequence data to the human genome.

In the result an overview of the distribution of sequences in the libraries are presented in bar charts. The distribution of sequences in each library is also presented in pie charts. To further analyze the differences in the libraries some questions were formulated. These questions were the following.

- Which are the 20 most frequent sequences that are common for all the serum libraries, common for the two blood cell libraries and are the most frequent sequences in the cancer libraries?
- Amongst the serum libraries, is there a difference between the 20 most common sequences in the male library and the female library?
- The average age of the males and females in the two libraries are young, ~25 years, compared to the eight other serum libraries which have an older average of ~50 years (see table 4). Is there a difference between the 20 most common sequences regarding age?

In order to extract the information to answer the questions above the data that came out of the alignment were further processed with the statistical program R®. The 100 most common reads for each library were extracted and combined. This presented a list that contained almost 600 unique reads

because the 100 most common reads in the libraries were not all the same between the libraries. The count of each read was divided by the total amount of unique reads in each library, revealing frequencies instead of counts. This was done as a normalization of the data. The sums of the frequencies were calculated for the grouped libraries according to questions above. Then the frequencies were ranked by the sums. As mentioned in the introduction one specific miRNA can vary in length due to some variation in processing and can exist with several different additions both in the 5' and in 3' ends. As a consequence of this, one miRNA will appear several times in the data sets after it has been run through the pipeline and in the list described above, each variant will be a unique read. Most of the non-annotated reads also occur with varying length. To take this into account the reads that mapped to the same chromosome and to the same small region plus minus some nucleotide, were collapsed and the frequencies were added together. In this way the 20 most common sequences were detected. In the results the comparisons are presented with Venn diagrams.

Result

Pre sequencing preparation and cDNA library construction

As mentioned in the materials and methods, a test run were performed where four different concentrations of a spike pool were added to plasma samples. The total RNA content in the plasma samples was isolated and libraries were constructed following the pathway previously described. After the amplification the constructed libraries were run on a 6% PAGE gel. Figure 4 shows three pictures of this gel, taken with a UV trans-illuminator. The four samples with libraries containing different concentrations of the RNA spike pool were loaded in lane 2-5. Lane 6 contained one extra library not of concern for this project. A 25 bp ladder from illumina were loaded in lane 1 and 7. In figure 4, the gel to the left is the whole gel and the gel in the middle is after the short pieces, below 75 nt, were cut out. The gel-picture at the right is after the final cuts, after extracting gel pieces containing library pieces approximately between 75 and 125 nt.

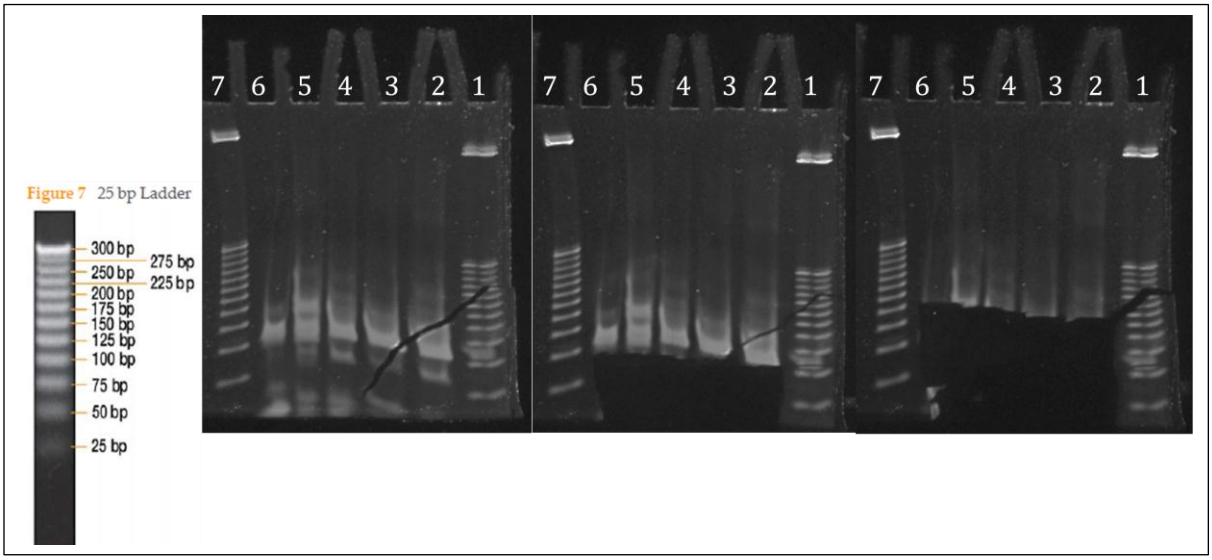

Figure 4. Three pictures of the same 6% PAGE gel and a ladder. The ladder is the 25 bp ladder provided by illumina. The ladder is the same as in the gel. The three gel-pictures illustrate where the cuts were made. The numbers represents the lane numbers.

Investigating the libraries that failed clusters generation

In the beginning of the sequencing run on the HiSeq 2000 it shows if the clustering was successful or not. Unfortunately as mentioned earlier there were hardly any clusters on the flow cell in the experiment using four different libraries with varying concentrations of the spike pool. The few clusters that were there probably were the PhiX control. To investigate why no clusters was created the libraries were amplified in a PCR and run on a 6% PAGE gel. The library samples that were used were the leftovers from the 10 nM samples (se materials and methods). The primers used in the PCR were GX1 Reverse and GX2 Forward, from table 3, which would only give an amplification product if the library construct were successfully created. Figure 5 shows a 6% PAGE gel of the amplified libraries.

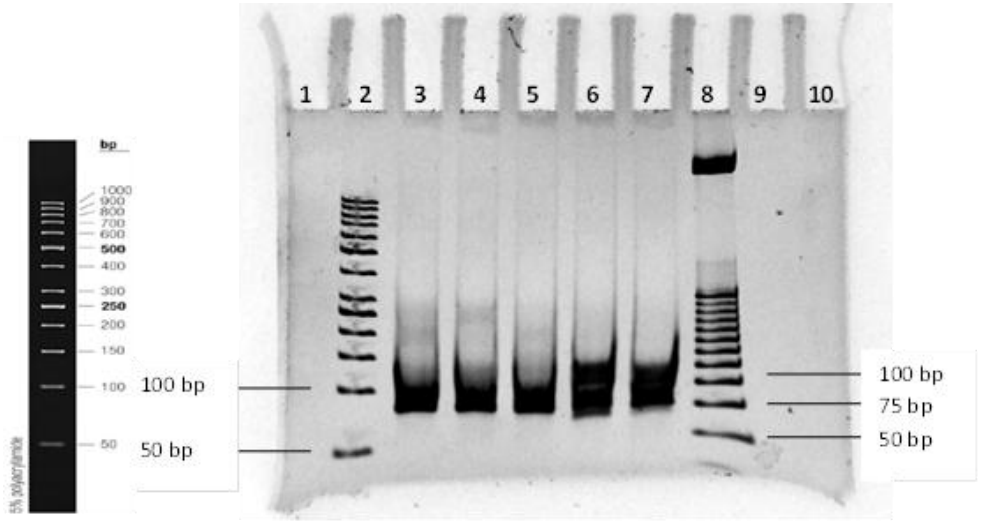

Figure 5. 6% PAGE gel of small RNA cDNA libraries. The lanes are numbered 1-10. Lane 2 contains a 50 bp ladder and lane 8 a 25 bp ladder. Lane 3-7 contains 5 sample libraries which failed in a sequencing attempt.

The amplification was successful in the sense that the bands on the gel in figure 5 are clearly visible. The fragments are between 75 and 100 bp.

Data analysis

As a consequence of the failed sequencing i.e. that the libraries described above did not generate any sequence data; fourteen datasets of libraries publically available were downloaded. In the materials and methods these fourteen datasets are described in table 5 and 6. These datasets were run through our programing pipeline, which compares the sequences against the human genome. The amount of data that came out of the run, contained detailed information on annotations and were of enormous sizes. Each unique sequence constitutes a read. Figure 6 presents two bar charts showing the distribution of the reads, for the average of the serum libraries and the average of the blood cell libraries. The categories in chart A (Average Serum) are: Not annotated, miRNA, Overlap annotated, Rest annotated and, Not aligned. The purple "Rest annotated", represents the rest of the annotated, i.e. all those sequences that map and annotate to some known sequence such as snoRNAs, tRNAs, rRNAs, etc. The "overlap annotated" are sequences that annotate to two or more known sequences, e.g. a miRNA inside a coding sequence or a pseudogene. In chart B (Average Blood cells) snoRNAs are displayed in its own column because the high representation in these libraries.

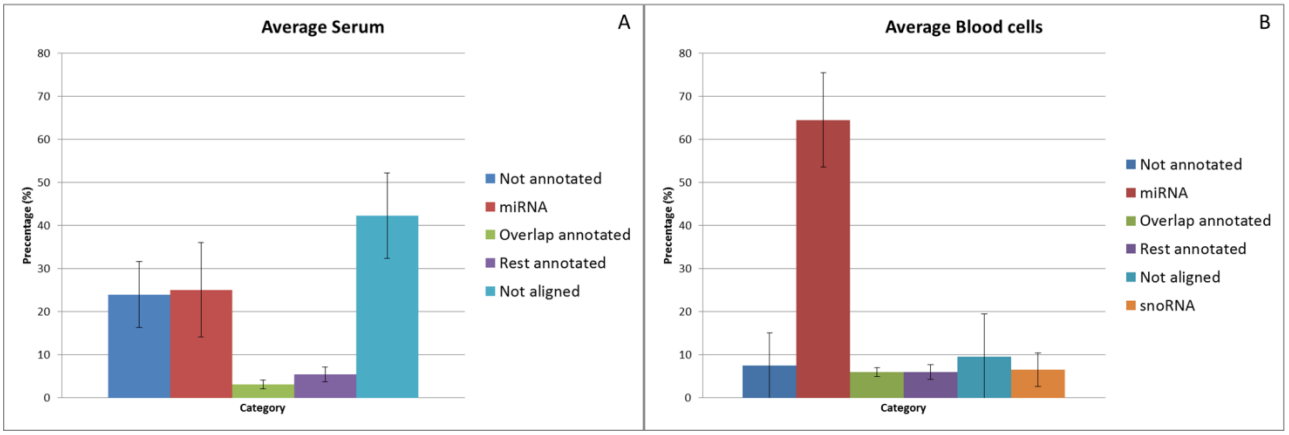

Figure 6. Two bar charts describing the distribution of reads for the average of the ten serum libraries (A) and the average of the two blood cell libraries (B). Al the reads are 18-30 bp in length. The reads are categorized in five categories in the charts: dark blue "Not annotated" are reads that mapped to the genome but did not annotate; read "miRNA" is recognized known miRNA; green "Overlap annotated" are reads that annotate to several known sequences; purple "Rest annotated" is the rest of the reads that aligned and did annotate; light blue "Not aligned" is reads that did not align to the human genome. In the blood cells chart a sixth category, snoRNA, is included. This is because the snoRNAs constituted such a big part of the reads. In A the snoRNAs are included in the purple column (Rest annotated). The error braes represent one standard deviation.

In figure 6 it clearly shows that the overall read distribution profile of serum and blood cells differ a lot. The percentage of reads constituting miRNAs (red bar in figure 6 A and B) is much higher in the blood cells than in the serum. The percentages of sequences that do not align to the human genome (light blue bars) also differ a lot between blood cells and serum.

Another way to visualize the same data as in figure 6 is to show it in pie charts. In figure 7 the read distributions are displayed in pie charts for each serum library individually and the average. The average is the same as presented in bar chart A in figure 6. The reason to display all libraries is to show the diversity. The diversity is huge, for example the miRNA percentage range from 3 to 74% of the overall reads in the different libraries and on average 25%. Notably, the not aligned percentages of the reads are very high, ranging from 17 to 80% and 42% on average. The mapped but not annotated differs a lot also, between 6 and 62% and on average 24%.

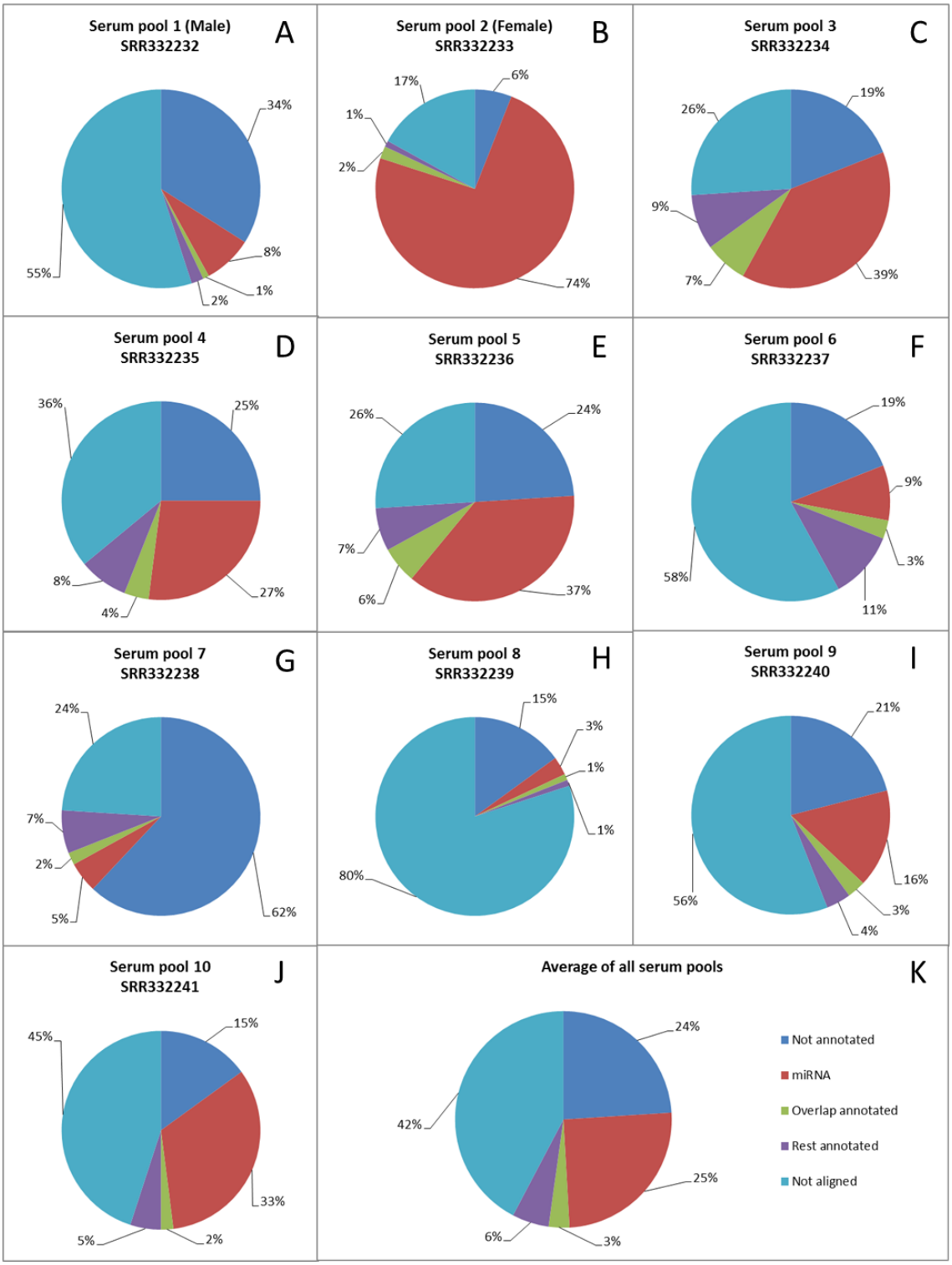

Figure 7. Pie charts over the distribution of reads in ten sequence libraries of small RNAs extracted from serum. The light blue "Not aligned" represents the reads that did not match the human genome and varies between 17 and 80%. The percentage of miRNAs varies between 3 to 75%. Down to the right is the average presented, which is the same as bar chart A in figure 6.

In figure 8 the read distributions are displayed in pie charts for the two blood cell library along with the average of these two and the two cancer cell lines, K562 and HL60. The average is the same as presented in bar chart B in figure 6. The different distribution patterns in the two blood cell libraries are striking.

When comparing the read distribution between the libraries in both figure 7 and 8, there is a striking difference in what percentages the miRNAs constitutes. The read distribution in the two blood cell libraries differ allot from each other, pie chart A and B in figure 8. The K562 cancer library is more similar in the overall read distribution to the first blood cell library (compare pie chart C to A in figure 8) whereas the HL60 cancer library more resembles the other blood cell library (compare pie chart E to B in figure 8).

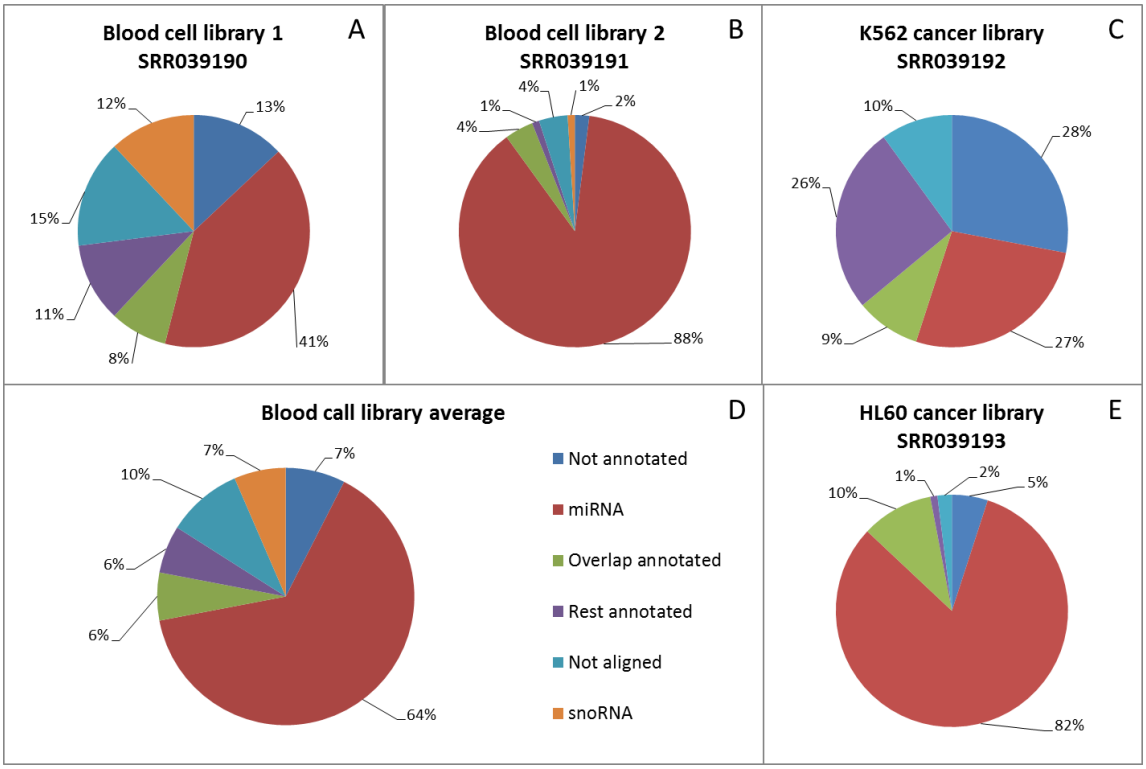

Figure 8. Pie charts over the distribution of reads in different libraries. In the two cancer cell lines the snoRNA are included in the purple category "Rest annotated". In the K562 cancer library the purple category is relatively huge 26%. This part consists of some relatively large subgroups: ~6% rRNA pseudogene, ~5% rRNA, ~4% tRNA, and ~4% snoRNA.

To further analyze what kind of miRNAs and other sequences these fourteen libraries contain, the 20 most frequent sequences in all libraries were extracted. The libraries were filtered so that only sequences that did map to the human genome are amongst the 20 most common sequences. This is important to point out, because such high percentages in some libraries are constituted by sequences that did not map the human genome. In figure 9 the 20 most frequent reads for the two cancer cell lines are compared to the two blood cell libraries together in a Venn diagram. Eleven sequences are in common for all; they are encircled by all three circles. Several of the sequences are miRNAs belonging to the let-family. All the miRNAs have a prefix "has-" in front of the name indicating the human origin. All of the RNAs in the Venn diagram in figure 8 are not miRNAs. There are one mitochondrial tRNA pseudogene (Mt tRNA pseudogene AC073308.1-201), one rRNA pseudogene (rRNA pseudogene AP003035.1-201), and one snoRNA (SNORD78-201), and several NAO (not annotated). The NAOs are reads that did not annotate but did align to the genome (these constitute the dark blue bars and pie pieces in figure 6, 7 and 8). The information within the parenthesis after the NAO describes on which chromosome, which strand (+ or –) and between which nucleotides the sequence mapped. Most of the NAOs existed in several isoforms with varying length in both 3' and 5'. The one presented in the parenthesis are the most common variant, but represents all the isoforms.

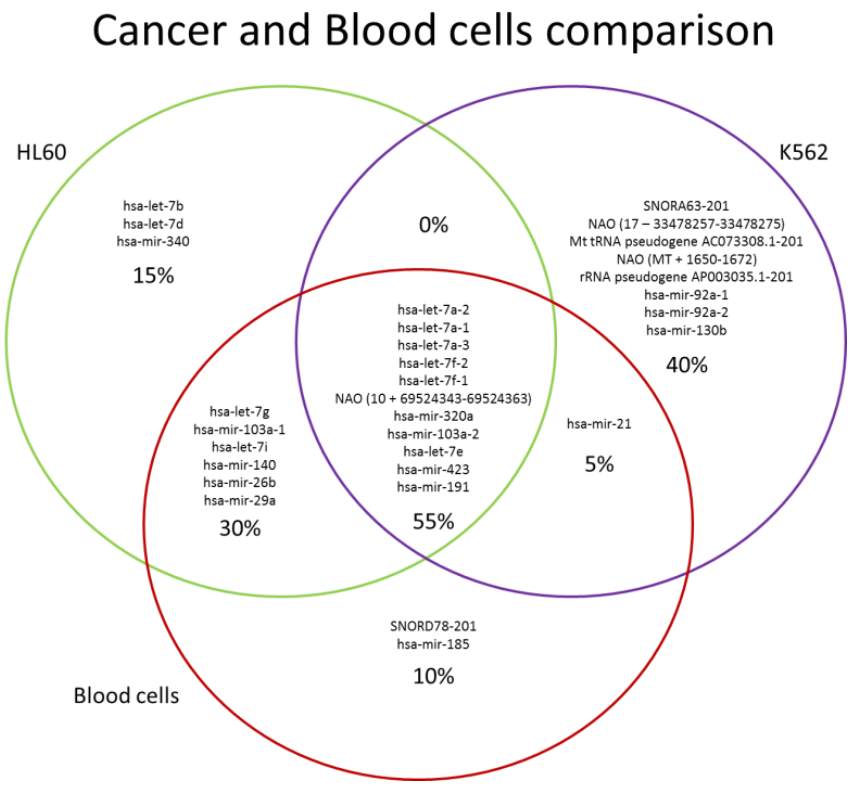

Figure 9. Venn diagram comparing the 20 most common RNAs in the libraries from two different cancer cell lines, HL60 (green circle), K562 (purple circle) and the two human blood cell libraries together (red circle). The names of the RNAs are included as well as the percentage.

Figure 10 shows a Venn diagram comparing the 20 most common sequences for all of the serum libraries together and the two blood libraries together. Only four out of 20 are in common (20%). The blood cells contain several miRNAs in the let-family, whereas the serum total does not contain as many of its 20 most common. The serum total 20 most common, contains among others miR-122 which is specifically expressed in liver cells.

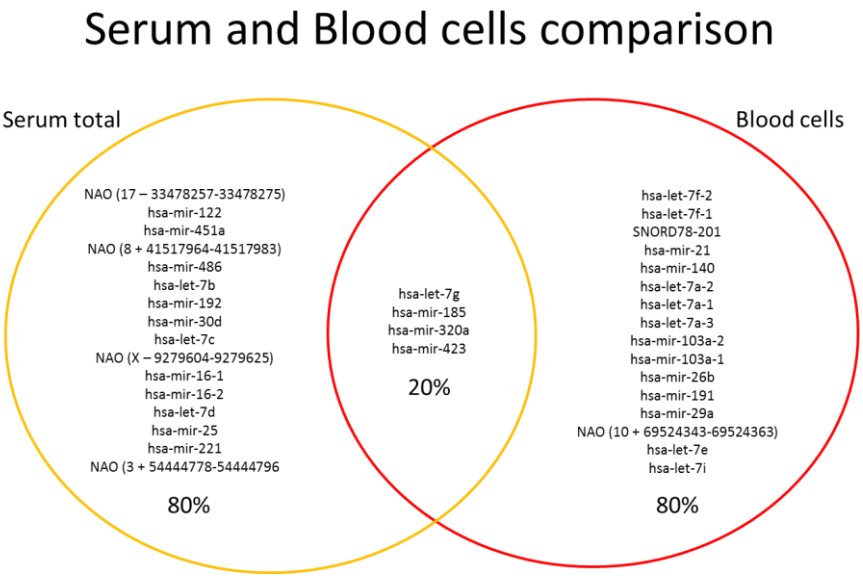

Figure 10. Venn diagram comparing the 20 most common sequences in the ten serum libraries together (yellow circle) and the two human blood cell libraries together (read circle). The names of the RNAs are included as well as the percentage.

In figure 11 the serum libraries consisting of pools of males and females are compared. Eleven out of the 20 most common reads are the same (55%).

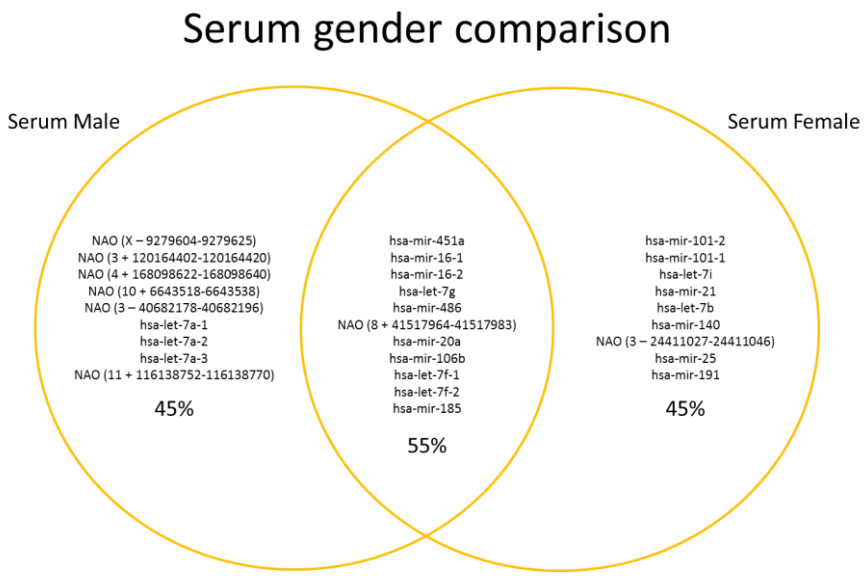

Figure 11. Venn diagram comparing the 20 most common sequences of the serum male and female pool libraries.

In figure 12 the 20 most common sequences are compared between serum from young and old. The young serum is the most common sequences when taking the male and female libraries together, with an average age approximately of 25 years, see table 5 in the materials and methods. The serum old is from the eight other serum pool libraries, with an average age approximately of 50 years, see table 5 in the materials and methods. The gender distribution within the eight serum pools is not known.

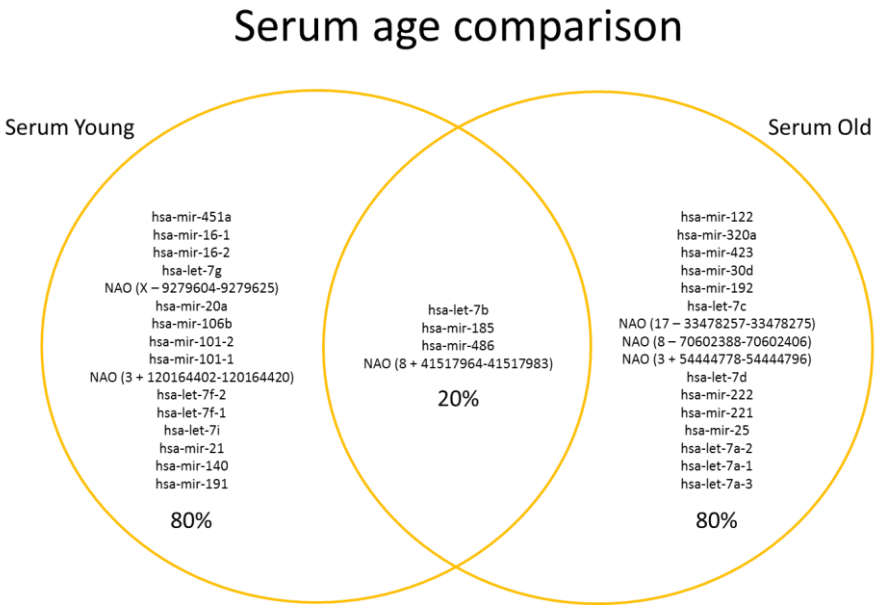

Figure 12. Venn diagram comparing the 20 most common sequences from the serum young and old libraries. The serum young was the most frequent of the male and the female pool libraries together (age ~25 years). The old is the most frequent of the eight serum pool 1 to 8 together (age ~50).

Only four sequences are the same (20%) for both young and old. Notably one miRNAs is present amongst the 20 most frequent sequences in all serum and blood libraries but not in the cancer cell

lines, namely has-mir-185 (miR-185). Two sequences NAO (8 + 41517964-41517983) which is a not annotated sequence and miR-486 is amongst the 20 most common in both serum sample comparisons (figure 11 and 12) but does not show up amongst the 20 most common sequences in the blood cells or the two cancer cell line libraries (figure 9).

Described above are results where the 20 most common RNAs in the samples are compared. Table 7 shows the total percentage of the 20 most common RNAs for the serum libraries, the blood cells libraries and, cancer cell libraries. The table also presents the percentage of the 20 most common sequences in the serum total, serum young (male and female together), serum old (the serum pool libraries 3 to 10 together) and, the blood cell libraries together. In table 7, it clearly shows that there is a huge variation between what percentages these, 20 most common sequences, constitute in the individual libraries. In the HL60 library the 20 most common sequences constitute 90% of that library compared to the K562 library within which only 31% of the library is represented by the 20 most common sequences.

Table 7. The total percentage of the 20 most common reads in the serum libraries, the blood cell libraries and, the blood cancer cell line libraries described in table 5 and 6. The bar represents the percentage; a fully filled field is 100%. The percentage of the 20 most common reads in the serum total, serum young, serum old and, blood cells together are also displayed. These are the 20 most common reads that are compared in the Venn diagrams in figure 9, 10, 11 and, 12.

Library	**Total percent of the 20 most common reads**
Serum pool library 1 (Male)	43%
Serum pool library 2 (Female)	82%
Serum pool library 3	60%
Serum pool library 4	58%
Serum pool library 5	57%
Serum pool library 6	64%
Serum pool library 7	66%
Serum pool library 8	39%
Serum pool library 9	68%
Serum pool library 10	73%
Serum total	58%
Serum young (male and female)	61%
Serum older (pool 3 to 10)	60%
Blood cells library 1	46%
Blood cells library 2	77%
Blood cell libraries together	66%
K562	31%
HL60	90%

Notably, in the HL60 library let-7f-1 and let-7f-2 constitute approximately 50% of all reads (data not shown). Further, let-7g, let-7f1-1 and let-7f-2 constitute approximately 53% of the reads in the serum female library but only 4% of the reads in the serum male library (data not shown).

When comparing the 20 most common for each serum library to the total serum 20 most common, it shows that some libraries contribute more than others (data not shown). Only three out of the 20 most common miRNAs are the same for all the 10 serum libraries. This even further highlights the diversity between the libraries.

Discussion

When investigating components in body fluids like serum and plasma one could expect a huge diversity because of the heterogeneity of these fluids. It is well-known that the blood composition varies a lot due to all imaginable conditions such as diets, drugs, fitness, stress, age, and time of collection during the day. So when investigating miRNAs circulating in the blood one should expect these molecules to occur in similar fashion as other molecules. Despite of this, there are several biomarkers used as indicators of different conditions and diseases.

As described in the materials and methods, RNA spikes were added early on in the creation of libraries as controls. These spikes were intended to be used to normalize the different libraries to each other and thereby get more reliable comparisons between the libraries. The clustering failed and no sequence data were obtained from our libraries with spikes. The fourteen libraries that make up the results did not contain any spikes. In the process of extracting the 20 most common sequences, the count of each read were divided by the total number of reads in that library to display frequencies. If one would have used spikes instead, then one would have subtracted or added the value of the some of the spikes to each reads count before calculating the frequencies. This would have presented frequencies that would have been much better to compare. Since the spikes were added in the same concentrations to all samples and all the samples were handled equivalent, the spikes would tell if there had been any unexpected material losses and the spikes could be used to compensate for this. However, the 20 most common sequences in the fourteen different libraries can still be compared in the way displayed in the result.

When constructing cDNA libraries aimed for sequencing one could do this with several different approaches. We chose an approach where we used one 3'-adapter, the same for all, and different 5'-adapters with unique barcodes for each library. Another approach is to have the barcode in the 3'-adapter. Yet another design could be to incorporate the barcode in one of the PCR-primers, and then the 3'-adapter and 5'-adapter used is the same for all libraries. The last approach is commercially available in a kit from Illumina. The reason for choosing the design with 5'-barcodes was that our group did already have eight 5'-adapters with unique barcodes and to order more of these would be the most economical option. The choice of approach is more of a technical and economic nature and would not affect the quality of the cDNA libraries.

In the materials and methods it is described how libraries were constructed and how, in the end, the clustering failed and no sequence data were created. In an attempted to investigate why it failed the libraries from that experiment was amplified and run on a gel, see figure 5. This amplification was designed so that it would only amplify sequences that contained both ligated adapters, because of the primers used. As shown on the gel in figure 5, there are clearly visible amplification products in all libraries. So why did the clustering fail? One could speculate in that perhaps were the concentrations measured on the Qubit lower in reality, which would give a lower concentration in all next coming steps and perhaps leading to a too low concentration on the flow cell and the clustering station. The amplification control performed would not reveal if this was the case. To control for this one could re do the concentration measurement on the Qubit fluorometer. To further investigate one could run a qPCR on the samples with primers designed so that one primer overlaps one of the ligation sites, while the other primer is specific for a target sequence (some of the expected miRNAs). This would give more specific information on if a target miRNA exist in the sample and if it is correctly constructed. In a qPCR like this one could also estimate the amount of adapter dimers by using primers designed so that one primer overlap the ligation of two adapters (one 3' and one 5' adapter) and the other primer

binding further away from this site. When examining the gel in figure 5 more closely it seems like much of the amplified product are of the size 75 bp which is the size of the predicted adapter dimer.

All fourteen datasets presented in table 5 and 6 are approximately 18-30 nt long. This is much shorter than intended for our libraries and these short reads cover mostly miRNA. The reason for including longer, i.e. up to 120 nt, was to reveal and map other small RNAs, such as tRNAs and snoRNAs. Moreover, this would reveal if molecules such as pre-miRNAs (~70 nt) do exist in the plasma samples. These molecules would not be expected because they have been shown to be less stable in circulation outside the cells compared to mature miRNAs. On the other hand, if huge parts of the circulating RNAs originate from erupted and lysed cells in material from cancer patients, one could expect to find these molecules circulating anyhow.

The high percentage of snoRNAs in the blood cell libraries, the orange category in figure 6 and 8, could be explained with the fact that the snoRNAs are involved in the RNA splicing machinery and blood cells have a high mRNA turnover.

The mapped but not annotated sequences, the dark blue in figure 6, 7 and 8, represent sequences that might just be novel miRNA or other small RNAs that have not yet been described. The overlap annotations in figure 6 represent sequences that annotate to several known sequences in the genome. Many of these are probably miRNAs that sits in coding regions of genes or pseudo genes. With these two observations the actual percentage that represents miRNAs should be somewhat higher.

It is interesting that the not aligned reads, the light blue in figure 6 and 7, are such a high percentage of the overall reads in the serum samples. One reason could be that the parameters chosen for our alignment. For instance reads mapping to several locations in the genome would be discarded since their origin is uncertain. If the parameters chosen for the alignment in the pipeline cause some sequences to be classified as not aligned even though they do align this could be expected to be equally common between libraries and could not explain the huge difference between the different libraries. Also, unexpected biological events, such as RNA editing, would change the sequence on the RNA molecule after transcription, producing mismatches that would affect the alignment. The same is true for short pieces from mRNAs expanding over exon boundaries. When investigating tRNAs, which contains a lot of nucleotide substitutions, with sequencing techniques lots of different variants of the sequences are produced due to these modified nucleotides.

When looking into biological samples, such as blood plasma and serum, one would expect the majority of the molecules to originate from the host itself. The samples originate from the paper were the authors find miRNAs from the plant world in these samples. One could speculate that the origin of these small RNAs could also be from bacteria in the gastrointestinal tract, from viruses, or from other food sources, besides from the plant world. The obvious next step to take would be to run a BLAST search on these sequences against all known genomes to answer what these sequences originate from.

Yet another explanation, to the high levels of not aligned reads in the serum libraries compared to the levels of human miRNAs could be that the libraries are pools. If the individuals within the pools have eaten different diets and thereby have different sets of foreign miRNAs these miRNAs would take up a larger part of the libraries compared to the human miRNAs. With this argument one assumes that the human miRNA profiles are relatively similar and limited to the approximately 1,500 human miRNAs known to date. One could expect some variation in which of these human miRNAs that are present among the individuals in a pool, but the total amount of unique human miRNAs would be expected to be fewer in total than the foreign miRNAs that have a higher potential to vary because they have all kinds of food intake as potential origin.

Only one miRNA, miR-185, is amongst the 20 most common for all serum sample comparisons and the blood cells (figure 10, 11 and 12) but not in the cancer cell lines (figure 9). This miR-185 has during the past year been shown to down regulate a well-known oncogene called *c-Myc* which codes for a transcription factor that is known to be constitutively expressed in various cancers (Liao and Lu 2011). In addition to these findings, miR-185 has been shown to target and down regulate DNA methyltransferases 1 which regulate global gene expression through methylations at promoter regions at several important genes that, in turn, codes for regulatory proteins (Zhang *et al.* 2011). With these findings, it is not surprising that miR-185 does not show up amongst the 20 most common sequences in the two cancer cell libraries.

As pointed out in the results, both the not annotated sequence NAO (8 + 41517964-41517983) and miR-486 are amongst the 20 most common in both serum samples comparisons (figure 11 and 12) but is not amongst the 20 most common sequences in the blood cells or the two cancer cell line libraries (figure 9 and 10). The NAO (8 + 41517964-41517983) sequence is complementary to miR-486 but is situated on the other strand of the genome and is consequently not the star (*) sequence but the antisense to miR-486. There are 4 different variants of this sequence in the 100 most common sequences and these vary in length between 19-22 nt. It is highly probable that this sequence will be identified as a miRNA in the future. In a recent paper miR-486 was proposed as a tumor suppressor in gastric cancer (Oh *et al.* 2011). In the paper it was shown that miR-486 was significantly down regulated in gastric cancer and that 25 to 30% out of 106 studied gastric cancers had genomic loss of miR-486. Further, they showed that restoring miR-486 caused suppression of several pro-oncogenic traits in gastric cancer cells, while inhibiting miR-486 in other cells enhanced proliferation. Moreover, the mRNA of an antiapoptotic endogenous glycoprotein OLFM4 was proposed as the target of miR-486 with bioinformatic and experimental support (Oh *et al.* 2011). If miR-486 is indeed a tumor suppressor miRNA, this can explain tis abstains in the cancer cell line libraries but not it's abstains in the blood cell libraries. Perhaps the miR-486 is not normally expressed in high levels in blood cells, but then one could speculate that the two cancer cell lines, which are derived from blood cells, would be expected to also express miR-486 in low levels.

As indicated in the result, on commenting upon the results presented in table 7, looking at the 20 most common sequences has its advantages and disadvantages. In some libraries the 20 most common sequences constitute as much as 90 % of the reads and in some libraries only a few miRNAs constitute as much as 50 % of the sequences. In general the 20 most common sequences represent approximately 50% or more of the total reads in each library. Moreover, as mentioned in the last part of the results, when comparing the 20 most common for each serum library to the total serum 20 most common, only three out of the 20 most common miRNAs are the same for all the 10 serum libraries. This even further highlights the diversity between the libraries.

There are definitely differences amongst the most common sequences in the serum libraries compared to the blood cell libraries as presented in figure 10 in the results. Only 4 out of 20 sequences are the same (20%). Moreover, when locking at what percentage these, the 20 most common sequences, constitutes of the overall reads in there libraries (see table 7) the serum total 20 most common sequences constitutes 58% and the blood cells together 20 most common sequences constitutes 66%. This shows that more than half of the reads are represented by the 20 most common sequences. These results are in line with the notion that most circulating miRNAs are not reflecting the miRNAs in blood cells but can have other origins. As previously mentioned in the introduction, recently the majority of circulating miRNAs have been shown to be associated with the Ago2 protein in the blood (Arroyo *et al.* 2011). Still others have shown a much higher correlation between the miRNA profiles in blood cells and in circulation (Pritchard *et al.* 2012), as also mentioned in the introduction.

Furthermore, also as mentioned in the introduction, a research group recently reported the findings that the miRNA population in blood cells has been shown to change in relation to various diseases that are not necessarily blood-borne diseases (Keller *et al.* 2011). The underlining cause of this change is not known but it could be as a response to changes in the environment caused by the disease, the authors of these findings argue. These findings indicate that several circulating miRNAs might originate from blood cells and or other cells and not from the tumors. This makes the finding of new cancer biomarkers more complex. But one could argue that if the miRNA profile of blood cells changes along with the profile of circulating miRNAs in a cancer patient the miRNAs could still be considered as valid biomarkers for the cancer even though the miRNAs do not originate from the actual cancer cells.

Finally, I believe interesting times are ahead, when it comes to miRNA research. Surely new findings regarding miRNAs acting as key regulators in cellular processes involved in various diseases especially different types of cancers are to be expected in a near future. All the technologies needed like next generation sequencing capable of sequencing entire genomes in a few days combined with powerful computers capable of processing massive amount of data, are now in place. What are needed now are skilled and clever scientists that ask and find the answers to the right questions. Technological breakthroughs like next generation sequencing and bioinformatics will push the knowledge of molecular genetics and cellular biology into new depths.

References

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. E., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G., and Tuschil, T. (2003). A uniformed system for microRNA annotation. *RNA*, 9:277-279.

Anderson, Ö. U., Nielsen, F. C., and Lund, A. H. (2008). MicroRNA-10a Binds the 5'UTR of Ribosomal Protein mRNAs and Enhances Their Translation. *Molecular Cell*, 30:460-471.

Arroyo, J. D., Chevillet, J. R., Kroh, E. M., Ruf, I. K., Pritchard, C. C., Gibson, D. F., Mitchell, P. S., Bennett, C. F., Pogosova-Agadjanyan, E. L., Stirewalt, D. L., Tait, J. F., and Tewari, M. (2011). Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *PANS*, 108(12):5003-5008.

Balcells, I., Cirera, S., and Busk, P. K. (2011). Specific and sensitive RT-PCR of miRNAs with DNA primers. *BMC Biotechnology*, 11(70):1-11.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Philippy, K. H., Sherman, P. M., Muertter, R. N., and Edgar, R. (2011). NCBI GEO: archive for functional genomics data sets – 10 years on. *Nucleic Acids Research*, 37:885-890.

Bustin, S. A. (2004). A-Z of Quantitative PCR. *International University Line*.

Calin, G. A., Seviganani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M., and Croce C. M. (2004).Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *PNAS*, 101(9):2999-3004.

Chen, X., Ba, Y., Ma, L., Cai, X., Yin, Y., Wang, K., Guo, J., Zhang, Y., Chen, J., Guo, X., Li, Q., Li, X., Wang, W., Zhang, Y., Wang, J., Jiang, X., Xiang, Y., Xu, C., Zhang, P., Zhang, J., Li, R., Zhang, H., Shang, X., Gong, T., Ning, G., Wang, J., Zen, K., Zhang, J., and Zhang, C-Y. (2008). Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell research*, 18(10):997-1006.

Chim, S. S. C., Shing, T. K. F., Hung, E. C. W., Leung, T.-Y., Lau, T.-K., Chiu, R. W. K., and Lo, Y. M. D. (2008). Detection and characterization of placental microRNAs in maternal plasma. *Clinical chemistry*, 54(3):482-490.

Davis, B. N., Hilyard, A. C., Nguyen, P. H., Lagna, G., Hata, A. (2010). Smad proteins bind a conserved RNA sequence to promote microRNA maturation by Drosha. *Molecular Cell*, 39:373-84.

Ender, C., Krek, A., Friedländer, M. R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. (2008). A Human snoRNA with MicroRNA-Like Functions. *Molecular Cell*, 32:519-528.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391:806-811.

Friedman, R. C., Farh, K. K-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19:92-105.

Griffiths-Jones, S., Grocock, R. J., Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acid Research*, 34:140-144.

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., and Kay, M. A. (2010). Human tRNA-derived small RNAs in global regulation of RNA silencing. *RNA*, 16:673-695.

Jopling, C. L. (2012). Liver-specific microRNA-122 Biogenesis and function. *RNA Biology*, 9(2):1-6.

Keller, A., Leidinger, P., Bauer, A., ElSharawy, A., Haas, J., Backes, C., Wendschlag, A., Giese, N., Tjaden, C., Ott, K., et al. (2011). Toward the blood-borne miRNome of human diseases. *Nature Methods*, 8(10):841-845.

Kim, D-J., Linnstaedt, S., Palma, J., Park, J. C., Ntrivalas, E., Kwak-Kim J. Y. H., Gilman-Sachs, A., Beaman, K., Hastings, M. L., Martin, J. N., and Duelli, D. M. (2012). Plasma Components Affect Accuracy of Circulating Cancer-Related MicroRNA Quantitation. *The Journal of Molecular Diagnostics*, 14:71-80.

Koeffler, H. P., and Golde, D. W. (1980). Human Myleoid Leukemia Cell Lines: A Review. *Blood*, 56(3):344-350.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, 294:853-858.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294:858-862.

Lawrie, C. H., Gal, S., Dunlop, H. M., Pushkaran, B., Liggins, A. P., Pulford, K., Banham, A. H., et al. (2008). Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *British journal of haematology*, 141(5):672-675.

Lee, L. W., Zhang, S., Etheridge, A., Ma, L., Martin, D., Galas, D., and Wang, K. (2010). Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, 16:2170-2180.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294:862-864.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843-854.

Lee, Y., Kim, M., Han, J., Yeom, K., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *European Molecular Biology Organization Journal*, 23:4051-4060.

Liao, J-M., and Lu, H. (2011). Autoregulatory Suppression of c-Myc by miR-185-3p. *Journal Of Biological Chemistry*, 286(39):33901-33909.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433:769-773.

Mallory, A., and Vaucheret, H. (2010). Form, Function, and Regulation of ARGONAUTE Proteins. *The Plant Cell*, 22:3879-3889.

Martin-Muller, C., Yao, Q., and Chen, C. (2010). miRNA and Human Disease. *Encyclopedia of life Science*.1-9.

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., et al. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(30):10513-10518.

miRBase (2012). The Sanger Centre miRBase Database (release 19), the central repository of miRNA sequences. Web page: [http://www.mirbase.org] 24/08/2012.

Morin, R. D., O´Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., and Marra, M. A. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*, 18:610-621.

Nelson, F. K., Snyder, M., and Gardner, A. (2011). Introduction and Historical Overview of DNA Sequencing. *Current Protocols in Molecular Biology*, 96:7.0.1-7.0.18. doi:10.1002/0471142727.mb0700s96.

Oh, H-K., Tan, A. L-K., and Das, K. (2011). Genomic loss of miR-486 Regulates Tumor Progression and the OLFM4 Antiapoptotic Factor in Gastric Cancer. *Clinical Cancer Research*, 17(9):2657-2667.

Pederson, T. (2010). Regulatory RNAs derived from transfer RNA?. *RNA*, 16(10):1-5.

Persson, H., Kvist, A., Vallon-Christersson, J., Medstrand, P., Borg, Å., and Rovira, C. (2009). The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nature Cell Biology*. 1(10):1268-1272.

Pritchard, C. C., Kroh, E., Wood, B., Arroyo, J. D., Dougherty, K. J., Miyaji, M. M., Tait, J. F., and Tewati, M. (2012). Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. *Cancer Prevention Research*, 5:492-497.

Yang, M., Chen, J., Su, F., Yu, B., Su, F., Lin, L., Liu, Y., Huang, J-D., and Song, E. (2011). Microvesicles secreted by macrophages shuttle invasion-potentiating microRNAs into breast cancer cells. Molecular Cancer, 10(117):1-13.

Vaz, C., Ahmad, H. M., Sharma, P., Gupta, R., Kumar, L., Kulshreshtha, R., and Bhattacharya, A. (2010). Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. *BMC Genomics*, 11(288):1-18.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional Regulation of the Heterochronic Gene *lin-14* by *lin-4* Mediates Temporal Pattern Formation in C. elegans. *Cell*, 75:855-862.

Zen, K., and Zhang, C-Y. (2010). Circulating MicroRNAs: A Novel Class of Biomarkers to Diagnose and Monitor Human Cancers. *Medicinal Research Reviews*, 2:326-348.

Zhang, L., Hou, D., Chen, X., Li. D., Zhu, L., Zhang, Y., Li, J., Bian, Z., et al. (2012). Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Research*, 22:107-126.

Zhang, Z., Tang, H., Wang, Z., Zhang, B., Liu, W., Lu, H., Xiao, L., Liu, X., Wang, R., Li, X., Wu, M., and Li, G. (2011). MiR-185 Targets the DNA Methyltransferases 1and Regulates Global DNA Methylation in human glinoma. *Molecular Cancer*, 10(124):1-16.