Application of Next Generation Sequencing and Bioinformatic Approaches to the Study of Influenza Virus Strains and MiRNA Profiling in Liver Diseases

by

Zhen Lin

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Medical Sciences - Shantou in Medicine

University of Alberta

ABSTRACT

Analysis and interpretation of next generation sequencing data has posed a significant challenge to researchers, especially to those who are not equipped with expertise in bioinformatics. In this thesis, I investigated three different biomedical questions using NGS and bioinformatic approaches, and explored the application of these methods to meet different needs.

Influenza epidemics result in up to 5 million severe cases worldwide, causing approximately 500,000 deaths annually. The characterization of emerging new strains can be challenging due to the diversity and high dissimilarity of influenza subtypes. Herein, we used NGS to analyze the RNA from the lungs of ferrets infected with influenza A/California/07/2009. Several bioinformatic approaches were used to catalogue the viral genes and investigate the quasi-species. Using the optimal bioinformatic approach, we were able fully characterize viral isolates following NGS.

For the next project, we used NGS to study microRNA expression in the liver of patients with end stage liver disease. These non-coding RNAs regulate levels of cellular mRNA and we sought to investigate whether miRNA profiles were associated with different disease processes. However, we found heterogeneous miRNA expression within samples derived from patients with the same disorder, which has limited our ability to link specific profiles with one disease process. Further multidimensional analysis and hierarchical clustering revealed the existence of two sub-groups in several diseases. A sizable number of miRNAs were found differentially expressed between those two groups, including miR-29 and miRNAs from miR-17-92 cluster. These miRNAs regulate the PI3K-AKT signaling and focal adhesion pathways that play a critical role in modulating diverse cellular functions including metabolism, growth, proliferation, survival and oncogenesis. However, the exact role that miR-29 and miRNAs from miR-17-92 cluster exert in end stage liver disease requires further clarification.

An analysis comparing miRNA expression in hepatocellular carcinoma and in the surrounding liver tissues was more productive. We therefore expanded the investigation to include more paired

samples; this form of analyses allows statistical modeling to incorporate both the variance between patients and the variance between diseased and non-diseased tissues. Moreover, by profiling miRNA expression and gene expression in paired human HCC tumor and non-tumor tissues, we were able to characterize the interaction between miRNAs and genes in tumor tissues. We found known oncomiRs (e.g. miR-21) and tumor suppressive miRNAs (e.g. miR-200 family) and observed reciprocal changes in gene expression. We also found a set of rarer oncomiRs such as miR-1269a, miR-518b, and miR-512-3p; the latter two miRNAs come from the largest miRNA cluster, C19MC. These studies provide new avenues for investigation for deregulated gene expression in HCC.

In summary, I have used a series of bioinformatic approaches to investigate complex NGS datasets to better understand the biology of diverse processes, including infectious diseases, non-infectious diseases and cancer.

PREFACE

This thesis is part of a research project approved by Health Research Ethic Board, University of Alberta. Project name: Viral discovery in liver disease, inflammatory bowel disease and other idiopathic disorders. Study ID: Pro00005105, Date: February 19, 2013

Chapter 3 was published as Lin Z, Farooqui A, Li G, Wong GK, Mason AL, Banner D, Kelvin AA, Kelvin DJ, León AJ. Next-generation sequencing and bioinformatic approaches to detect and analyze influenza virus in ferrets. J. Infect. Dev. Ctries. 2014 Apr 15;8(4):498-509. doi: 10.3855/jidc.3749. I was responsible for data collection, data analysis and writing.

All experiments and data presented in Chapter 4 and 5 were produced by myself, with data analysis support from Dr. Juan Jovel. Dr. Andy Mason and Dr. Gane Wong provided intellectual support.

Scripts in Appendix A were written by Dr. Juan Jovel and applied by myself.

ACKNOWLEGEMENT

I would like to thank my supervisors Dr. Andy Mason, Dr. Gane Wong and Dr. David Kelvin, for the guidance and support throughout the training. I am also very grateful to Dr. Juan Jovel, for his time and effort on guiding my studies, from study design, lab training to bioinformatic analysis and manuscript preparation. Many thanks to Dr. Alberto Leon and Dr. Alyson Kelvin for the guidance and advice in the influenza studies.

I am particularly thankful to the Sino-Canadian program coordinator, Dr. Paul Melançon. His advice and support has been very helpful in some very challenging moments during the study.

I want to thank my other supervisory committee members, Dr. Edan Foley and Dr. Wan Jiao for their time and precious advice. Also I want to express my gratitude to graduate coordinate Dr. Sean McMurtry and graduate student advisor Ms. Maggie Hill for their support.

Throughout the years in this program between China and Canada, I have been working with many great people in different labs, including Dr. Kelvin's lab in Shantou University Medical College and in UHN Toronto, Dr. Mason's lab and Dr. Wong's lab in University of Alberta. I'm very thankful to all the past and present coworkers for their help, friendship and support.

I would like to thank my husband Xu and my parents. This thesis would not be possible without their encouragement and support.

Lastly, I am very grateful to the Li Ka-Shing foundation, who initiated the joint PhD program and provided with generous scholarship. Also I'm thankful to Canadian liver foundation who has sponsored my study during the past year.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 1 Introduction

1.1 Sequencing technologies

1.1.1 Development of sequencing technologies – the three generations

First generation sequencing

Based on the so-called primer-extension strategy, Fredrick Sanger and colleagues developed in 1970s a DNA sequencing method using chain-terminating inhibitors (Sanger and Coulson, 1975) (Sanger et al., 1977). This method uses selectively incorporated dideoxynucleotides to terminate the chain extension during DNA replication catalyzed by DNA polymerase. Canonically, a DNA polymerase catalyzes the extension of a nascent oligonucleotide by adding deoxynucleotides harboring a 3'-hydroxyl group. If chemical analogs lacking the 3'-hydroxyl, technically called 2',3'-dideoxynucleotides triphosphates, are added to the reaction, polymerization will terminate exactly at the place where the native base should have been incorporated (Sanger et al., 1977). Sanger sequencing became automated in 1980s and became cheaper and faster owing to the development of the dye-terminator technique and the addition of capillary electrophoresis and an automatic DNA sequence analyzer. Since then, sequencing technologies based on Sanger sequencing have been widely applied in medical and research labs. The fast development and wide spreading of this technology laid the foundation for proposing the Human Genome Project in 1980s and enabled its final completion in 2003. Together with shotgun cloning, automated Sanger sequencing is considered as the first-generation sequencing technology.

Massively parallel sequencing (Next generation sequencing)

Although the state-of-art sequencing technology at that time was adopted in the Human Genome Project, it still took the scientists from hundreds of international labs thirteen years and 3.4 billion US dollars on total to complete the project. It was clear that for sequencing large genomes, faster and cheaper sequencing methods with higher throughput needed to be developed. In 2005 Life Sciences commercialized the 454 genome sequencer, which was able to sequence 25 million bases in a single four-hour run, with a read length of 100 bases (Margulies et al., 2005). That was the start of the next generation sequencing (NGS) era. Soon other companies joined the market, including the Illumina/Solexa sequencing platform in 2006 and the SOLiD (Sequencing by Oligo Ligation Detection) by Applied Biosystem in 2007 (Valouev et al., 2008). The Illumina and SOLiD sequencers generated reads of 35 bp long, but in a much larger amount (30 and 100 million reads, respectively) compared to 200 thousand reads generated by the 454 sequencer. In 2010, the

Personal Genome Machine (PGM) was released by Ion Torrent (now Life Technologies). This system resembles the 454 technology but offers higher speed and lower cost, owing to a new technology for detection of nucleotides. It delivered up to 270 Mb of sequence in reads that were up to 100 nt long. For a detailed description of the technology behind each system see reviews (Metzker, 2010) (Liu et al., 2012). In addition to the major platforms mentioned above, other less-popular NGS technologies were developed, such as Heliscope single molecule sequencer (Pushkarev et al., 2009) . For a more complete list of sequencing technologies, see review (Zhang et al., 2011). NGS technologies share some common advantages. First, the sequencing no longer relies on bacterial cloning of DNA fragments. Second, millions of sequencing reactions and base interrogations are happening in parallel without the need for electrophoresis. These new methods enabled the sequencing of big genomes at an ultrafast speed and high throughput.

Single molecule sequencing (Third generation sequencing)

One drawback of NGS is its dependence on PCR amplification, an intrinsic undesirable attribute of the sequencing-by-synthesis technology. A new sequencing method developed by Pacific Biosciences, known as PacBio, overcame this shortcoming by conducting “real-time DNA sequencing from single polymerase molecules” (Eid et al., 2009). PacBio has additional advantages over NGS, such as higher throughput, longer read length, faster runtime, higher accuracy and lower cost; this technology is referred to as “third-generation sequencing” (Schadt et al., 2010). Another technology for single-molecule sequencing is the nanopore sequencing (Clarke et al., 2009), represented by the Oxford Nanopore MinION sequencer released in 2015. In the view of some experts, however, “it will require dramatic decreases in error rates before it lives up to its promise” (Mikheyev and Tin, 2014).

1.1.2 Application of NGS technologies in biomedical research

As mentioned above, one of the first applications of NGS technology is whole genome sequencing (WGS). Concomitant with advances in throughput, costs and turnaround time of sequencing has been significantly reduced, thus enabling the proposal of large-scale human WGS studies, such as the 1000 Genomes Project (Abecasis et al., 2010). Since then, more population-scale genome studies have been launched, which has greatly enhanced the understanding of the association between human genetic variation and phenotype (Kilpinen and Barrett, 2013). In addition, WGS has facilitated a wide range of clinical and research studies, from forensic genetics (Weber-Lehmann et al., 2014) to clinical diagnostics (Gasser et al., 2011; Palomaki et al., 2012; Vilarino-Guell et al., 2011); from single cell genomics (Shapiro et al., 2013) to metagenomics studies

(Morgan and Huttenhower, 2014). Another important application is to identify pathogens strains associated with outbreaks as well as surveillance of strain evolution (Lin et al., 2014; Lipkin, 2013).

In addition to genome sequencing, NGS technologies have also been applied to transcriptome sequencing. The transcriptome is formally defined as “the complete set of transcripts in a cell, and their abundance, in a specific developmental stage or physiological condition” (Wang et al., 2009). It includes both coding and non-coding transcripts (Wang et al., 2009). For gene expression profiling, RNA-Seq has surpassed traditional methods such as microarray or qPCR owing to the following virtues: 1) it has the ability to detect novel transcripts and variants without the need of reference transcriptome or genome; 2) it has very low background signal and much broader dynamic range for detection; 3) it is more accurate in quantifying expression levels of transcripts; 4) it requires less starting material (Wang et al., 2009). Furthermore, with the emergence of robust tools for splicing analysis, RNA-Seq has been able to identify novel splicing isoforms in recent years (Anders et al., 2012; Blekhman et al., 2010).

Small non-coding RNAs, especially microRNAs (miRNAs), have attracted great research interest since they were discovered, given their important roles in post-transcriptional regulation of many biological processes (Bentley et al., 2008). MiRNAs are small RNAs with length around 22 nucleotides, which suppress transcription through binding to the 3’ untranslated region (UTR) of messenger RNAs (Figure 1.1). MicroRNA-seq, accordingly, has been successfully applied in miRNA research after NGS technologies were introduced to the market. Compared to microarray and qPCR, miRNA-seq is more powerful in profiling known miRNAs. More importantly, it offers the opportunity to discover and characterize novel miRNAs and miRNA isoforms (isomiRs) (Zhou et al., 2010).

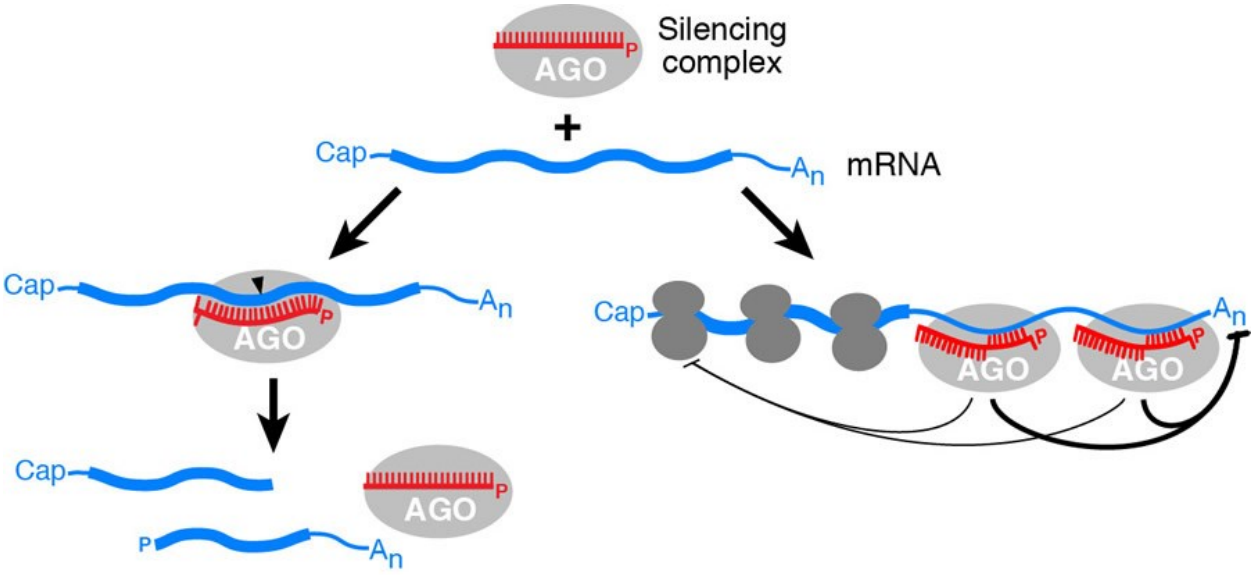


Figure 1.1 Simplified schematic of miRNA regulation of gene expression on post-transcription level. A miRNA (red) was bound within an argonaute protein (AGO) and other proteins (not shown) to form a silencing complex. After pairing to mRNAs (blue), miRNA can direct gene repression through two different pathways. (Left) If base-pairing between the miRNA and the targeted mRNA is perfect, the mRNA can be cleaved by AGO. Alternatively (right), the mRNA can be repressed if pairing is less extensive but perfect in seed region, through inhibiting translational initiation and destabilizing the mRNA by shortening its poly-A tail. This figure is cited from http://www.hhmi.org/research/regulation-mrna-translation-and-decay

Despite the vast advantages of NGS technologies in many research fields, sequencing data interpretation has been a significant challenge to researchers. Since the output from NGS sequencers includes large amounts of relatively shorter reads, analysis and interpretation of NGS data requires extensive expertise in bioinformatics. However, the emergence and fast development of NGS technologies has greatly provoked the advance of bioinformatics, new algorithms and software have been timely developed and are becoming more user-friendly than ever. Furthermore, the improvements in computational techniques have increased the utilization of NGS technologies and the amount of information that can be extracted from sequencing experiments. More about data analysis is discussed in a later section of this chapter.

Among the manufacturers, Illumina sequencers have been the most popular choice for expression studies using RNA-Seq. As the leader of the NGS industry, Illumina system offers highest throughput at lowest per-base cost (Liu et al., 2012). Furthermore, both the read length and the read number generated per run have been increasing steadily throughout the years since its

introduction in 2007. Currently, the read length of Illumina sequencers goes up to 300 bp, which fits into a wider range of applications. Depending on the specific system, each Illumina sequencer differs in output range, maximum read length and turnaround time (Table 1.1).

Table 1.1 Features of Illumina sequencers used in this project

Feature	Genome Analyzer IIx	MiSeq sequencer	HiSeq 2000 Sequencer
Flow cells per run	1	1	1 or 2
Maximum Read Length	2x150bp	2x300bp	2x250bp
Reads per Flow Cell	320 million	25 million	300 million
Output Range	10-95GB	0.3-5GB	10-300GB
Run time	2-14 days	5-55 hours	7-60 hours

1.1.3 Illumina sequencing technology

The typical Illumina sequencing workflow, similar to other NGS systems, is comprised by sample preparation, cluster generation, sequencing and data analysis

(http://www.illumina.com/technology.html).

Sample preparation, or library construction, refers to the processing of DNA or RNA samples into the appropriate format for the sequencer. In general, library construction involves steps of fragmentation, end-repair, 5' and 3' adapter ligation, followed by enrichment and purification of libraries (Figure 1.2A).

Clustering is a process whereby each fragment of the samples is isothermally amplified. Illumina technology uses a flow cell as a solid phase for amplification and imaging. The flow cell is surface-bounded with a lawn of oligoes, which are complimentary to the sequences of the adapters ligated to the fragments in the library construction step. The denatured DNA molecule is captured on the flow cell and the initial copy of the template is generated. Then the original template is washed away and the bounded copy is amplified by bridge amplification. In bridge amplification the strand fold over and the adapter region hybridizes into the other type of oligo on the flow cell. Polymerases generate the complimentary strand, forming a double-stranded bridge. The bridge is then denatured and the reverse strand is washed away. This process repeats and occurs

simultaneously for millions of clusters, resulting in clonal amplification for all the fragments, each of which has a colony of up to 1000 identical copies (Figure 1.2B).

Illumina applies sequencing-by-synthesis (SBS) technology with fluorescent reversible terminator chemistry (Bentley et al., 2008). During sequencing, one single fluorescently labeled deoxynucleoside triphosphate (dNTP) is incorporated at a time to the nucleic acid chain on the flow cell in repetitive cycles. The fluorescent label also serves as a reversible terminator for polymerization, so after incorporation and imaging, the label is cleaved to allow incorporation of the next nucleotide. All the images are recorded during each cycle for the millions of clusters on the flow cell in parallel, and base callings are made directly from signal intensity measurement (Figure 1.2C)

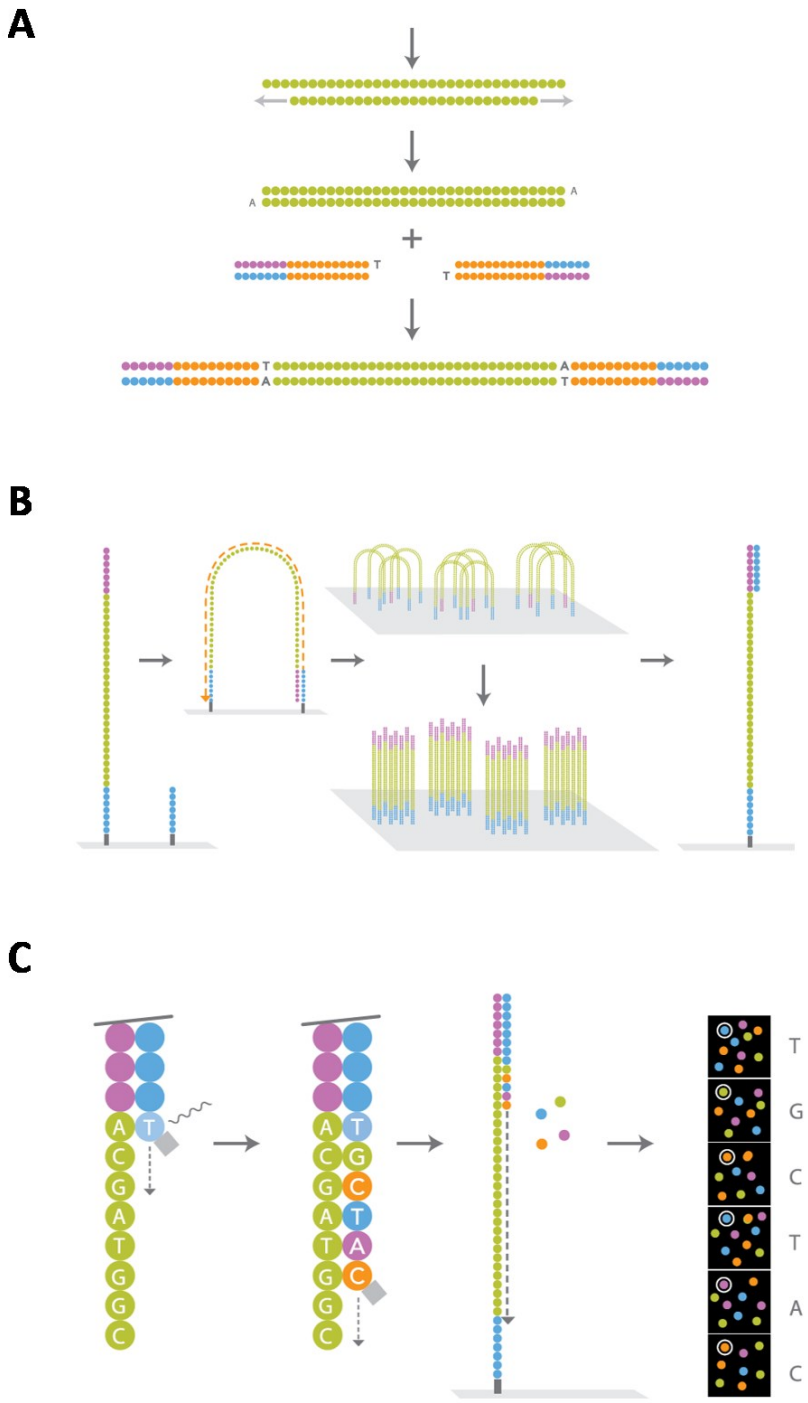


Figure 1.2 Illumina library construction (TruSeq), amplification and sequencing. **A)** The ends of the fragmented DNA are repaired and the 3' end is adenylated. Adapters are then ligated to the DNA, forming the sequencing libraries. **B)** The template is attached to the primers on the flow cell, copied and cloned in to clusters by bridge amplification. **C)** Clusters are sequenced by synthesis with fluorescent reversible terminator technology.

1.2 Bioinformatic approaches for analyzing NGS data

1.2.1 General bioinformatic approaches for analysis of sequencing data

The output data from Illumina sequencers are usually in fastq format, which is a four-line text-based format that stores the nucleotide sequence and the corresponding quality score. Although there is no existing standard pipeline for processing RNA sequencing data, the general scheme of bioinformatics usually includes: (1) pre-processing (i.e. removal of adapters and low-quality reads), (2) mapping reads to certain reference (e.g. reference genome, miRBase) by alignment tools (the scenario that there is none reference is not discussed here), (3) quantification and generating a count table for each library, (4) normalization of the expression files across all samples, (5) statistical analysis to address specific biological questions (e.g. differentially expressed genes) (Figure 1.2). For influenza virus genome sequencing, the data process usually only includes pre-processing, aligning, assembly and quantification for obtaining the consensus genome sequence. The specific approaches applied on sequencing of viral RNA, microRNAs, and transcriptome are introduced individually hereafter.

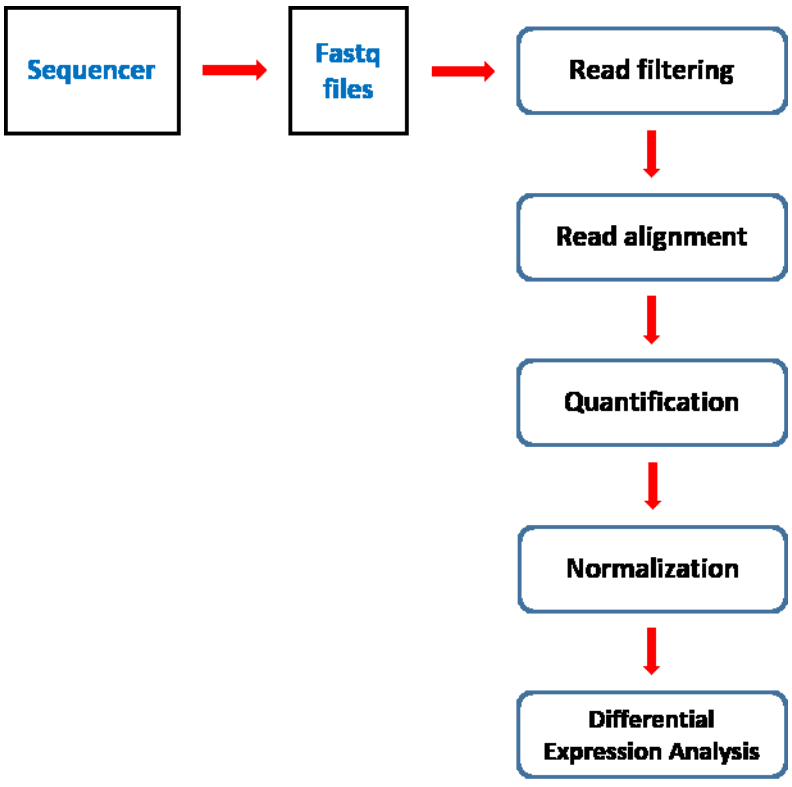


Figure 1.3 Bioinformatics pipeline with typical steps involved in sequencing analysis.

1.2.2 Alignment and assembly tools for influenza virus detection and characterization

In scenarios intended to confirm the presence, quantify or study minor sequence variations of known viruses, the use of a short-reads aligner such as Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2010) or SOAP2 (Li et al., 2009b), provides a well-established and rich framework when used in combination with other downstream analysis tools.

De novo assembly is to assemble short reads without the aid of a reference sequencing. For viral discovery studies, or when a significant dissimilarity is expected between the short-read sequences and the viral reference, de novo assemblers such as ABySS (Birol et al., 2009), Velvet (Zerbino, 2010) or SOAPdenovo (Li et al., 2010) can be used to generate longer sequence contigs. Subsequently, BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) analysis, carefully adapting the parameters to the necessities of each case, allows the detection of viral sequences in a collection of contigs by aligning them against a database of known viral sequences. Alternatively, the short-read libraries can be aligned against viral reference sequences followed by assembly of the consensus sequence. Direct BLAST analysis of short-reads can be very sensitive, when there are good references such as in discovery of novel influenza virus strains, although the short-reads need to be of sufficient length so that the analysis can “absorb” the differences without causing a dramatic decrease in the similarity score.

1.2.3 Tools for transcriptome analysis

Alignment

As a first step during transcriptome analysis, transcripts recovered in RNA-seq libraries are mapped onto the corresponding genome. One of the most popular mapping tools used in transcriptome profiling is TopHat (Trapnell et al., 2009). Using BWA as short read aligner, TopHat implements novel splicing junction discovery alignment in order to reconstruct transcripts from RNA-seq data (Garber et al., 2011). Developed from TopHat, TopHat 2 integrates Bowtie2 as core mapping engine, which is able to efficiently detect short insertions and deletions (indels). For detection of larger deletions, inversions and gene translocations, the TopHat-Fusion algorithm (Kim and Salzberg, 2011) was incorporated into TopHat2, so that the discovery of fusion gene products could be achieved by simply including the relevant parameters into the command line (Kim et al., 2013).

Count normalization

It has been shown that normalization is an essential step in RNA-seq data analysis because of its strong impact to the differential expression (DE) results (Bullard et al., 2010) (Zyprych-Walczak et al., 2015). The purpose of normalization is to transform the data derived from different samples to be comparable by considering some main factors such as difference in library size and experimental conditions. But the ideal method using for normalization has constantly been a matter of debate. One major issue is that most normalization approaches are based on a presumption that every cell produces similar amount of RNA. Under this presupposition, gene expression across samples can be comparable after normalizing for library size through introducing same amount of RNA (Mortazavi et al., 2008; Oshlack and Wakefield, 2009). However, this presumption is not necessarily correct (Lin et al., 2012; Nie et al., 2012). A good solution is to use spike-in standard, which allows normalization to cell number and thus more accurate in detecting genes that differentially expressed (Loven et al., 2012). But according to a recent study, external spike-ins are still not reliable enough to be used as standard normalization procedures and a strategy called remove unwanted variation (RUV) was proposed instead (Risso et al., 2014).

Among well-recognized normalization approaches, TMM normalization (trimmed mean of M-value, implemented in edgeR) and DESeq normalization are considered robust in most simulation studies (Dillies et al., 2013; Guo et al., 2013; Oshlack and Wakefield, 2009). Additionally, both algorithms has encoded to correct the intrinsic distribution issue.

EdgeR and DESeq are Bioconductor packages developed based on the assumption that RNA-seq data fits into a negative binomial distribution and that most genes are not differentially expressed (Robinson and Oshlack, 2010) (Anders and Huber, 2010). In TMM normalization, the scaling factor is computed between the test sample and any reference sample, based on a weighted mean after trimming the log fold-changes (M) and absolute gene expression level (A) (Robinson and Oshlack, 2010). TMM is implemented in edgeR and the default cutoff for M-value is 30% and A-value 5% (Robinson et al., 2010). In DESeq normalization, a size factor for a given library and for each gene is calculated as the median of the ratio of its observed counts to its geometric mean across all samples (Anders and Huber, 2010). In both cases the normalized read count of each gene is obtained by dividing the raw counts by the normalization factor of that library.

Differential expression analysis

When aiming at identifying differentially expressed genes between groups (e.g. mutant vs. wild type, disease vs. non-disease, etc.), an appropriate differential expression method is required.

Along with the fast spread of RNA-seq application, many software/packages have been developed for differential expression analysis, including DESeq, edgeR, BaySeq (Hardcastle and Kelly, 2010), EBSeq (Leng et al., 2013), NOISeq (Tarazona et al., 2011) and SAMSeq (Li and Tibshirani, 2013), among others. Although no agreement exists on which method performs best on a given dataset, DESeq and edgeR remain top performers according to several comparative studies with simulated data (Guo et al., 2013; Seyednasrollah et al., 2015; Soneson and Delorenzi, 2013) and spike-in (Seyednasrollah et al., 2015). In a comparison between edgeR and DESeq, the former was found more relaxed and the latter more conservative (Seyednasrollah et al., 2015; Soneson and Delorenzi, 2013). DESeq2 (Love et al., 2014), as an upgraded version of DESeq, has higher sensitivity according to the authors (Love, 2014).

Cuffdiff 2 (Trapnell et al., 2013) is a differential expression analysis package incorporated in an popular RNA-seq analysis pipeline named Tuxedo suite (Trapnell et al., 2012). Nevertheless, Cuffdiff 2 was found less favorable due to the high false positive rate with no increase in sensitivity (Rapaport et al., 2013).

1.2.4 Tools for miRNA-seq data analysis

MiRNAs are short non-coding RNAs playing important roles in regulating messenger RNA expression (for details see introduction in Chapter 4). The bioinformatic analysis of miRNA-seq data is largely inherited from RNA-seq expression analysis and shares the general scheme of transcriptome data processing (Fig. 1.2). Within the past ten years there have been a number of programs developed for preprocessing and alignment of miRNA-seq data, such as miRDeep (Friedlander et al., 2008), Shortran (Gupta et al., 2012), miRanalyzer (Hackenberg et al., 2011; Hackenberg et al., 2009), CAPmiRSeq (Sun et al., 2014) etc., which differ in the settings for adapter removal, reads filtering and the choice of aligner. While filtering algorithms have some influence on the recovery of miRNAs profile, greater impact is exerted by the aligners (Tam et al., 2015). Among the alignment algorithms, Bowtie and BWA are the most popular choices, and the effectiveness of them was confirmed in a recent study using spike-in dataset to simulate the features observed in real experiments (Tam et al., 2015).

Similar to transcriptome analysis, there is also no agreement on the optimal normalization and differential expression analysis method for miRNA-seq data processing. Although still a common practice, normalizing read counts to count-per-million (cpm) by sequencing depth has been shown insufficient (Dillies et al., 2013; Garmire and Subramaniam, 2012). From the several comparison studies using different simulation datasets and aligners, TMM and upper quartile are best

supported (Dillies et al., 2013; Tam et al., 2015; Zhou et al., 2013). Although few study has explored the pros and cons of current tools for miRNA differential expression analysis, edgeR and DESeq remain the top choices.

Common interests regarding miRNA profiling studies usually include identifying and characterizing isomiRs and novel miRNAs. While aligning tools can usually find isomiRs, packages such as IsomiRex (Sablok et al., 2013) and miRSpring (Humphreys and Suter, 2013) are more convenient options in term of speed and accuracy. For novel miRNA prediction, miRDeep and miRDeep2 (Friedlander et al., 2012) are the most recognized ones.

1.2.5 MiRNA target prediction tools -TargetScan and DIANA-microT

MiRNAs regulate gene expression through binding to mRNAs and repressing the translation and/or inducing degradation of targets (Krol et al., 2010). Characterization of deregulated miRNAs is usually followed by determination of their targets and the biological implications of miRNA/mRNA interactions. Computational prediction of miRNA targets is a fast and cost-effective approach to advance miRNA functional studies; it integrates cell biology, biochemistry and statistical parameters to provide reliable predictions of target genes (Vlachos and Hatzigeorgiou, 2013). For reviews on miRNA target predicting tools, see (Liu et al., 2014; Peterson et al., 2014). Among the many published tools and pipelines, only a few are actively maintained and upgraded throughout years, which include TagetScan (Agarwal et al., 2015; Friedman et al., 2009; Garcia et al., 2011; Grimson et al., 2007; Lewis et al., 2003) and DIANA-microT (Maragkakis et al., 2009; Paraskevopoulou et al., 2013; Reczko et al., 2012).

TargetScan was one of the first algorithms available for miRNA target prediction. The prediction is accomplished by searching conserved 7mer or 8mer sites on 3'UTR that are complementary to the seed region of a miRNA (bases two to eight in its 5' end). Targets are ranked by context+ score, which indicates the probability of the effective tageting. In later versions of TargetScan, more features were added to compute the score, such as local AU-rich composition, 3'-supplementary and position contribution (Grimson et al., 2007) as well as preferentially conserved targeting (*P*ct) score (Friedman et al., 2009), seed-paring stability and target abundance (Garcia et al., 2011). In the latest version TargetScan 7.0, the site type and 14 additional features were combined to develop the context ++ model, by which the most effectively targeted mRNAs are predicted (Agarwal et al., 2015).

The fifth version of DIANA-microT, DIANA-microT-CDS, takes into account the miRNA recognition elements (MREs) located in the coding DNA sequence (CDS) region, besides the 3'-UTR. In brief,

DIANA microT incorporates features such as binding category, conservation across species, target sites accessibility, pair stability, AU content in MRE flanking, etc. into the program. The algorithm is trained on an experimental dataset and provides decent sensitivity and high level of precision (Paraskevopoulou et al., 2013; Reczko et al., 2012).

Although both programs present high sensitivity and precision, studies have found that prediction by different algorithms can vary substantially. Therefore it has been a common practice to combine different algorithms to control the false positive rate. But if only working on the intersection, there will be inevitable increase to false negative rate. Therefore, we applied both TargetScan and DIANA microT-CDS, but not restricted to the intersection.

1.3 Aims and hypotheses

In this thesis we aimed to explore the application of NGS and bioinformatic approaches in investigating the two main biomedical questions. First one was to detect and characterize unknown influenza virus strain in an accurate and timely way. Second one was to identify deregulated miRNAs in end-stage liver diseases and HCC.

Accordingly the hypotheses were 1) application of NGS sequencing technology and carefully designed bioinformatic approaches are competent in characterizing novel influenza strain fast and accurately; 2) Deregulated miRNA signatures in end-stage liver diseases and HCC could be identified by miRNA-seq profiling.

References

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

Agarwal, V., Bell, G.W., Nam, J.W., Bartel, D.P., 2015. Predicting effective microRNA target sites in mammalian mRNAs. Elife 4.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome biology 11, R106.

Anders, S., Reyes, A., Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. Genome research 22, 2008-2017.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J.,

Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E.C.M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-59.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M., Gascoyne, R.D., Marra, M.A., Jones, S.J., 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25, 2872-2877.

Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M., Gilad, Y., 2010. Sex-specific and lineage-specific alternative splicing in primates. Genome research 20, 180-189.

Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11, 94.

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H., 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nature nanotechnology 4, 265-270.

Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., Jaffrezic, F., French StatOmique, C., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14, 671-683.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133-138.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., Rajewsky, N., 2008. Discovering microRNAs from deep sequencing data using miRDeep. Nature biotechnology 26, 407-415.

Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W., Rajewsky, N., 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40, 37-52.

Friedman, R.C., Farh, K.K., Burge, C.B., Bartel, D.P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome research 19, 92-105.

Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8, 469-477.

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., Bartel, D.P., 2011. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 18, 1139-1146.

Garmire, L.X., Subramaniam, S., 2012. Evaluation of normalization methods in mammalian microRNA-Seq data. RNA 18, 1279-1288.

Gasser, T., Hardy, J., Mizuno, Y., 2011. Milestones in PD genetics. Mov Disord 26, 1042-1048.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., Bartel, D.P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27, 91-105.

Guo, Y., Li, C.I., Ye, F., Shyr, Y., 2013. Evaluation of read count based RNAseq analysis methods. BMC Genomics 14 Suppl 8, S2.

Gupta, V., Markmann, K., Pedersen, C.N., Stougaard, J., Andersen, S.U., 2012. shortran: a pipeline for small RNA-seq data analysis. Bioinformatics 28, 2698-2700.

Hackenberg, M., Rodriguez-Ezpeleta, N., Aransay, A.M., 2011. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res 39, W132-138.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M., Aransay, A.M., 2009. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 37, W68-76.

Hardcastle, T.J., Kelly, K.A., 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11, 422.

Humphreys, D.T., Suter, C.M., 2013. miRspring: a compact standalone research tool for analyzing miRNA-seq data. Nucleic Acids Res 41, e147.

Kilpinen, H., Barrett, J.C., 2013. How next-generation sequencing is transforming complex disease genetics. Trends Genet 29, 23-30.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36.

Kim, D., Salzberg, S.L., 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome biology 12, R72.

Krol, J., Loedige, I., Filipowicz, W., 2010. The widespread regulation of microRNA biogenesis, function and decay. Nature reviews. Genetics 11, 597-610.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10, R25.

Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., Haag, J.D., Gould, M.N., Stewart, R.M., Kendziorski, C., 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29, 1035-1043.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., Burge, C.B., 2003. Prediction of mammalian microRNA targets. Cell 115, 787-798.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.
Li, J., Tibshirani, R., 2013. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 22, 519-536.
Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.
Li, Y., Hu, Y., Bolund, L., Wang, J., 2010. State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics 4, 271-277.
Lin, C.Y., Loven, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., Young, R.A., 2012. Transcriptional amplification in tumor cells with elevated c-Myc. Cell 151, 56-67.
Lin, Z., Farooqui, A., Li, G., Wong, G.K., Mason, A.L., Banner, D., Kelvin, A.A., Kelvin, D.J., Leon, A.J., 2014. Next-generation sequencing and bioinformatic approaches to detect and analyze influenza virus in ferrets. Journal of infection in developing countries 8, 498-509.
Lipkin, W.I., 2013. The changing face of pathogen discovery and surveillance. Nature reviews. Microbiology 11, 133-141.
Liu, B., Li, J., Cairns, M.J., 2014. Identifying miRNAs, targets and functions. Brief Bioinform 15, 1-19.
Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. Journal of biomedicine & biotechnology 2012, 251364.
Love, M.I., 2014. Assessment of DESeq2 performance through simulation.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15, 550.
Loven, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., Young, R.A., 2012. Revisiting global gene expression analysis. Cell 151, 476-482.
Maragkakis, M., Reczko, M., Simossis, V.A., Alexiou, P., Papadopoulos, G.L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., Hatzigeorgiou, A.G., 2009. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 37, W273-276.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.
Metzker, M.L., 2010. Sequencing technologies - the next generation. Nature reviews. Genetics 11, 31-46.
Mikheyev, A.S., Tin, M.M., 2014. A first look at the Oxford Nanopore MinION sequencer. Molecular ecology resources 14, 1097-1102.
Morgan, X.C., Huttenhower, C., 2014. Meta'omic analytic techniques for studying the intestinal microbiome. Gastroenterology 146, 1437-1448 e1431.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628.
Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D.R., Tessarollo, L., Casellas, R., Zhao, K., Levens, D., 2012. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell 151, 68-79.
Oshlack, A., Wakefield, M.J., 2009. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4, 14.

Palomaki, G.E., Deciu, C., Kloza, E.M., Lambert-Messerlian, G.M., Haddow, J.E., Neveux, L.M., Ehrich, M., van den Boom, D., Bombard, A.T., Grody, W.W., Nelson, S.F., Canick, J.A., 2012. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet Med 14, 296-305.

Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., Hatzigeorgiou, A.G., 2013. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res 41, W169-173.

Peterson, S.M., Thompson, J.A., Ufkin, M.L., Sathyanarayana, P., Liaw, L., Congdon, C.B., 2014. Common features of microRNA target prediction tools. Front Genet 5, 23.

Pushkarev, D., Neff, N.F., Quake, S.R., 2009. Single-molecule sequencing of an individual human genome. Nature biotechnology 27, 847-850.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome biology 14, R95.

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., Hatzigeorgiou, A.G., 2012. Functional microRNA targets in protein coding sequences. Bioinformatics 28, 771-776.

Risso, D., Ngai, J., Speed, T.P., Dudoit, S., 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology 32, 896-902.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology 11, R25.

Sablok, G., Milev, I., Minkov, G., Minkov, I., Varotto, C., Yahubyan, G., Baev, V., 2013. isomiRex: web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets. FEBS letters 587, 2629-2634.

Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94, 441-448.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74, 5463-5467.

Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Human molecular genetics 19, R227-240.

Seyednasrollah, F., Laiho, A., Elo, L.L., 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 16, 59-70.

Shapiro, E., Biezuner, T., Linnarsson, S., 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature reviews. Genetics 14, 618-630.

Soneson, C., Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14, 91.

Sun, Z., Evans, J., Bhagwate, A., Middha, S., Bockol, M., Yan, H., Kocher, J.P., 2014. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. BMC Genomics 15, 423.

Tam, S., Tsao, M.S., McPherson, J.D., 2015. Optimization of miRNA-seq data preprocessing. Brief Bioinform.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A., 2011. Differential expression in RNA-seq: a matter of depth. Genome research 21, 2213-2223.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology 31, 46-53.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562-578.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., Sidow, A., Fire, A., Johnson, S.M., 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome research 18, 1051-1063.

Vilarino-Guell, C., Wider, C., Ross, O.A., Dachsel, J.C., Kachergus, J.M., Lincoln, S.J., Soto-Ortolaza, A.I., Cobb, S.A., Wilhoite, G.J., Bacon, J.A., Behrouz, B., Melrose, H.L., Hentati, E., Puschmann, A., Evans, D.M., Conibear, E., Wasserman, W.W., Aasly, J.O., Burkhard, P.R., Djaldetti, R., Ghika, J., Hentati, F., Krygowska-Wajs, A., Lynch, T., Melamed, E., Rajput, A., Rajput, A.H., Solida, A., Wu, R.M., Uitti, R.J., Wszolek, Z.K., Vingerhoets, F., Farrer, M.J., 2011. VPS35 mutations in Parkinson disease. Am J Hum Genet 89, 162-167.

Vlachos, I.S., Hatzigeorgiou, A.G., 2013. Online resources for miRNA analysis. Clinical biochemistry 46, 879-900.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics 10, 57-63.

Weber-Lehmann, J., Schilling, E., Gradl, G., Richter, D.C., Wiehler, J., Rolf, B., 2014. Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Sci Int Genet 9, 42-46.

Zerbino, D.R., 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11, Unit 11 15.

Zhang, J., Chiodini, R., Badr, A., Zhang, G., 2011. The impact of next-generation sequencing on genomics. Journal of genetics and genomics = Yi chuan xue bao 38, 95-109.

Zhou, X., Oshlack, A., Robinson, M.D., 2013. miRNA-Seq normalization comparisons need improvement. RNA 19, 733-734.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., Yu, J., 2010. The next-generation sequencing technology and application. Protein Cell 1, 520-536.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Gorczak, K., Klamecka, K., Figlerowicz, M., Siatkowski, I., 2015. The Impact of Normalization Methods on RNA-Seq Data Analysis. Biomed Res Int 2015, 621690.

Chapter 2 Materials and Methods

This chapter includes the materials and methods of three experiments. For the ease of reading, the materials and methods for each experiment were described individually.

2.1 Experiment #1: Detection and analyzation of influenza virus by NGS and bioinformatic approaches

2.1.1 Sample collection from virus infected ferrets

Ethics statement

All experiments with ferrets were performed in accordance with the guidelines of Canadian Council of Animal Care (CCAC). The animal use protocols were approved by the Animal Care Committee (ACC) of the University Health Network (UHN). UHN has certification under the Animals for Research Act (Permit Number: #0045 and #0085 of the Ontario Ministry of Agriculture, Food and Rural Affairs). Infections and sample collections were conducted under 5% isoflurane anaesthesia with all effort to reduce suffering.

Virus

H1N1pandemic virus strain, A/California/07/2009, is provided by Center for Disease Control and Prevention ([CDC], Atlanta, GA, USA) in aliquots of embryonic chicken egg allantoic fluid.

Infection and ferrets monitoring

Ferrets were maintained and monitored as previously described (Rowe et al., 2010). In brief, 4~6 months old male ferrets were purchased from Triple F Farms (Sayre, PA, USA) or bred on-site (University Health Network, Toronto, ON, Canada). Ferrets were screened to ensure seronegative against circulating influenza A and B strains prior to infection, with hemagglutinin inhibition assay as previously described (Rowe et al., 2010). Before infection ferrets were randomly chosen and pair-housed in cages in laminar-flow clean-room enclosures (Lab Products, Seaford, DE) in the BSL-2 area, and a temperature transponder (BioMedic Data Systems, Inc., Seaford, DE) was implanted subcutaneously in each animal. Body temperature and weight were recorded on Day 0 as a baseline. Upon infection, ferrets were anesthetized by breathing in 5% isoflurane with oxygen. 1ml 1×10^6 50% egg infectious doses (EID50) of influenza A/California/07/2009 (H1N1) was inoculated intranasally to each animal as 0.5ml per nostril. Clinical signs including body

temperature, weight, activity level, symptoms of nasal discharge and sneezing were observed and recorded at the same time every day until euthanasia.

Sample collection

On days 1, 3, 5, and 14 post infection (p. i.), three ferrets were euthanized and lung tissues from mid-lobe were collected and stored in RNALater at -80°C for later experiments, including pathological examination, virus titration and microarray analysis. A total of four lung tissue samples from days 1, 3, 5 and 14 post-infection, respectively, were selected for analysis by deep sequencing. The virus was not detectable in the lung sample day 14 post-infection and it was included in this study as negative control. Detailed information about the infection procedures, clinical data and the results of the microarray and NGS analysis were describe in a previous publication from our group (Leon et al., 2013).

2.1.2 Methods

2.1.2.1 RNA isolation

TriPure reagent (Roche, Indianapolis, IN, USA) and TRIzol reagent (Invitrogen) were used for RNA isolation from the ferret lung and the human liver tissues, respectively, according to manufacturer's instructions. Briefly 50 mg of tissue was homogenized in 1 ml of reagent followed by addition of 0.3 ml chloroform. After centrifugation the colorless aqueous supernatant was extracted, mixed with 1 volume of isopropanol and then centrifuged. The RNA pellet was washed with 1 ml of 75% ethanol, twice, air-dried and finally re-suspended in RNase free water. The quality of the RNA was verified in an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, California).

2.1.2.2 cDNA library construction and sequencing from infected ferrets

From the ferret RNA, cDNA libraries were constructed and sequenced at BGI (Shenzhen, Guangdong, China) according to previously published procedures (Wang et al., 2010). Briefly, the mRNA was isolated and fragmented, double-stranded cDNA was synthesized followed by adapter ligation; DNA fragments were selected by excising the 200 ± 25 bp band in an agarose gel electrophoresis followed by PCR for library enrichment.

Paired-end 90 bp sequencing was performed using an Illumina Genome Analyzer IIx sequencer. Adapter sequences were removed and those reads with more than 10% Q<20 bases were filtered out.

2.1.2.3 Detection and characterization of viruses

Flowchart

First the short reads data from sequencing were processed by three different types of programs: short-reads aligners Bowtie, general-purpose aligner BLAST and *de novo* assembler Abyss. Next, Consensus sequences were generated by coupling the results with specific tools for biological interpretation, including quantification of viral genes (Iliad assembler or EdgeR) and SNP calling for detection of viral quasispecies (VarScan) (Figure 2.1).

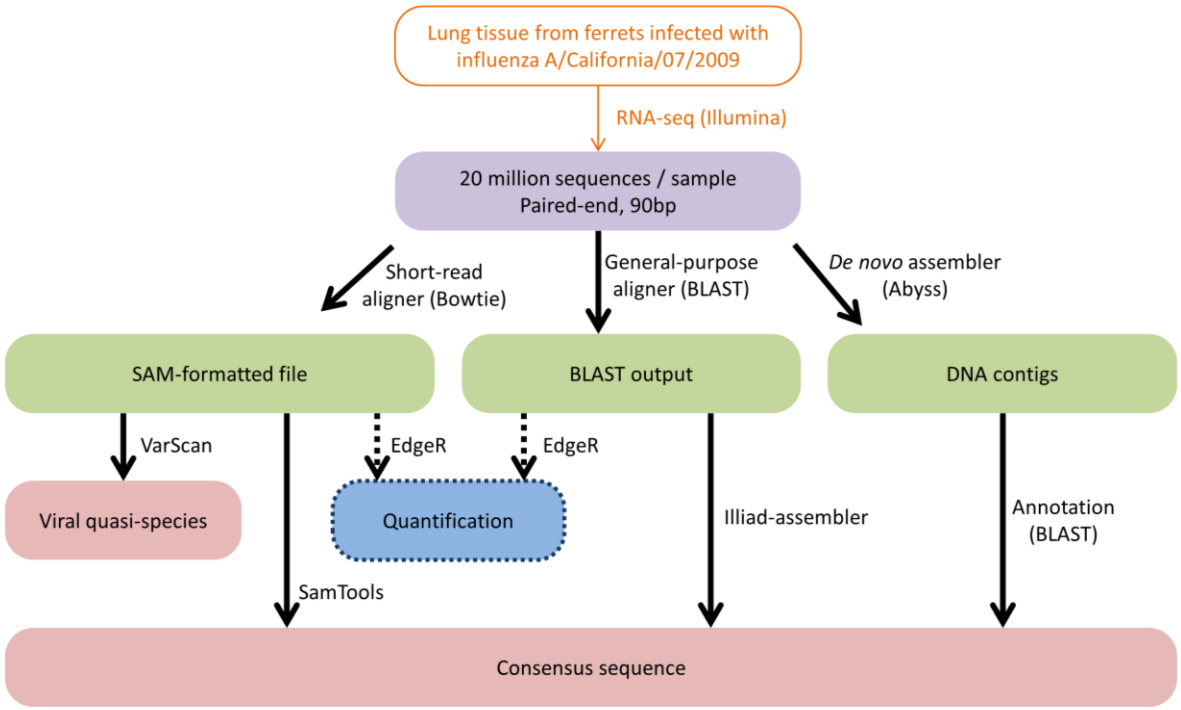


Figure 2.1 Bioinformatic analysis for detection and characterization of viruses. The first step of the bioinformatic process can be performed by three different types of programs: short-reads aligners, general-purpose aligners and *de novo* assembly. Afterwards, biological interpretation requires coupling with specific tools to generate the consensus sequence, quantification of viral genes and SNP calling for detection of viral quasispecies. The chart is comprised by the following elements: biological samples and sequencing (orange colour); data (coloured boxes); bioinformatic analysis (black arrows) which were performed (continuous lines) and not performed in this study but of relevance in the field (dotted lines).

Detection of influenza virus-matching reads using Bowtie and downstream analysis

Bowtie2 (v2.0.2, Linux 64 version) was downloaded (http://bowtie-bio.sourceforge.net/index.shtml) and it was executed under Linux Ubuntu desktop 11.04. Nucleotide sequences of all the viral segments of A/California/07/2009 (Garten et al., 2009) were retrieved from Genbank: FJ966976 (polymerase PB2 subunit), FJ966978 (polymerase PB1 subunit), FJ966977 (polymerase PA subunit), FJ966974 (hemagglutinin, HA), FJ969536 (nucleocapsid protein, NP), FJ984386 (neuraminidase, NA), FJ966975 (matrix proteins, MP) and FJ969528 (non-structural genes, NS). A Bowtie2 index containing the sequences of the viral segments was generated with the bowtie2-build program. The analysis of the short-reads was performed by using Bowtie in paired-end mode with the -S option to obtain the output in SAM (Sequence Alignment Map) format, which is a generic format for storing large nucleotide sequence alignments (http://samtools.sourceforge.net) (Li et al., 2009a). The SAM Tools-0.1.12 (Linux 64 version) package was downloaded (http://sourceforge.net/projects/samtools/files/samtools/0.1.12/) and used to process the sequence alignment files in SAM format sequentially using the SAM Tools commands *view*, *sort* and *pileup*. The resulting output is in pileup format and it describes the base-pair information at each position (http://samtools.sourceforge.net/pileup.shtml). Next, the consensus sequence for each viral segment was generated by running the script "*samtools.pl pileup2fq*" with a minimum coverage per-base of 3. Additionally, SAM files were imported in the assembly visualization tool Tablet v1.11.08.29 (http://bioinf.scri.ac.uk/tablet/) (Milne et al., 2010) and the number of times that each position had been covered by the aligned short-reads was determined. Finally, to study the variations present in the influenza-matching short-read sequences of each sample, the previously generated *pileup* files were analyzed with VarScan-2.2.5 software (http://varscan.sourceforge.net/) (Koboldt et al., 2012) by executing the program with the *pileup2snp* command.

De novo assembly with Abyss and annotation of the generated contigs

To reduce the level of complexity and the computational requirements, short-reads matching the ferret genome were subtracted from the short-read libraries; the 1,871 genomic scaffolds that comprise the ferret genome assembly MusPutFur1.0 were downloaded from GenBank (accessions GL896898-GL898768). A Bowtie index containing these genomic scaffolds was built. The short-read libraries were analyzed with Bowtie (v0.12.7) and the unaligned reads were stored on separate files. The *de novo* assembler ABySS-1.2.7 (http://www.bcgsc.ca/platform/bioinfo/software/abyss) was run on a Linux environment (Ubuntu desktop 11.04) with 12Gb of RAM, using the parameter k=32 and paired-end mode. Next, the

resulting contigs were subjected to BLAST analysis to select those contigs showing high degree of similarity with the influenza sequences.

Pre-selection of influenza-matching short-reads with BLAST and assembly of the consensus sequences with Iliad Assembler

The software BLAST 2.2.25+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) was installed and run locally under Windows 7. BLAST databases containing the short-read sequences of the different samples were constructed with the makeblastdb program included in the BLAST package. The nucleotide sequences of all the viral segments of A/California/07/2009 (accession numbers shown above) were used as "query" for the BLAST analysis. Different combinations of settings were tested to optimize the BLAST analysis. Iliad Assembler is a software tool developed by our group which falls in the category of guided assemblers and it relies on BLAST to perform the alignments (http://www.ferretscience.org/2012/02/iliad-assembler.html). The program generates the consensus sequence by using a reference sequence together with a set of pre-selected short-reads; additionally, it finds the correct position for the contigs even when unresolved areas are present. Given the flexibility of BLAST alignments, the program offers a good performance in situations when high dissimilarity between the reference and the reads are present (manuscript under preparation).

2.2 Experiment #2: profiling miRNA expression in human end-stage livers and HCC-affected livers using NGS and bioinformatic approaches

2.2.1 Sample collection

Sample collections were approved by Health Research Ethic Board, University of Alberta. Explanted livers were collected immediately after removal from recipient patients, directly in the operating rooms (OR) of the University of Alberta Hospital and then immediately transferred to the nearby pathology laboratory for collection of samples with the assistant of a pathologist. In general, samples were taken within a lapse of 30 minutes after each liver was explanted. Tissues were snap frozen in liquid nitrogen. For miRNA profiling in multiple end-stage liver diseases, 63 samples from eight different diseases were collected (Table 2.1). For the HCC study, four livers were sampled including paired tumor and non-tumor tissues from each liver (Table 2.2). Non-tumor tissues were collected from a non-adjacent region.

Table 2.1 Clinical information of liver tissues from end-stage livers

Liver Diseases	Abbreviation	Number of samples
Autoimmune hepatitis	AIH	9
Cryptogenic cirrhosis	CRP	6
Alcoholic cirrhosis	ETH	9
Hepatocellular carcinoma (non-tumor tissue)	HCC	10
Cirrhosis with hepatitis C virus infection	HCV	3
Nonalcoholic steatohepatitis	NSH	5
Primary biliary cirrhosis	PBC	10
Primary sclerosing cholangitis	PSC	11
Total number of libraries		63

Table 2.2 Clinical information for the HCC samples collected from explanted livers

Sample	Age (y)	Gender	Diagnosis	
			Cirrhosis	Virus infection
311	61	f	alcoholic cirrhosis	HCV
314	56	f	cryptogenic cirrhosis	no
315	52	m	alcoholic cirrhosis	HCV
338	56	f	alcoholic cirrhosis	no

For HCC samples from resected livers, paired samples from tumors and non-tumor tissue were acquired from the Cross Cancer Institute (CCI) tumor bank. To ensure that degradation of RNA was minimized and transcriptional status was preserved, golden quality samples were requested, which means that tissues were harvested directly from the patients, either from intraoperative biopsy or following surgical devascularisation, and snap frozen in less than 30 minutes post devitalisation. Additional information of the patients is listed in Table 2.3.

Table 2.3 Description of HCC and control samples from resected livers

No.	Age	Gender	Primary diagnosis and comorbidities		
			Cirrhosis	**Non-hepatic diagnosis and therapy**	**Virus infection**
1	70	m	alcoholic cirrhosis	Marfan syndrome	no
2	73	m	N/A[a]	N/A	N/A
3	63	f	cirrhosis	portal vein embolization	HBV
4	45	f	no cirrhosis	N/A	HCV
5	77	m	no cirrhosis	hypercholesterolemia	no
6	63	m	no cirrhosis	Lamivudine	HBV
7	53	m	alcoholic cirrhosis	N/A	N/A[b]
8	64	m	no information	diabetes	N/A
9	65	m	hemochromatosis cirrhosis	N/A	N/A
10	68	m	alcoholic cirrhosis	N/A	no
11	71	m	non-alcoholic steatohepatitis	diabetes, hyperlipidemia	no
12	61	m	cirrhosis	Tenofovir	HBV

a Information not available
b Chronic viral infection mentioned, not sure which type

2.2.2 Methods

2.2.2.1 RNA isolation

TRIzol reagent (Invitrogen) was used for RNA isolation according to manufacturer's instructions. Other details are the same as in section 2.1.2.1

2.2.2.2 Small RNA library construction and sequencing

Small RNA libraries were constructed using the TruSeq small RNA kit (Illumina) following the sample prep guide from the manufacturer. The theory behind such method is to ligate the 5' end of a pre-adenylated linker to the free hydroxyl (-OH) group at the 3' end of miRNAs in an ATP-free reaction. Such ligation reaction was conducted by a mutant T4 DNA-RNA ligase that exhibits preference for such substrates, thereby minimizing non-specific ligations. This is followed by ligation of a 5' adapter, reverse transcription, indexing PCR, size selection, titration and quantification (Figure 2.2). Details are described in the following paragraphs.

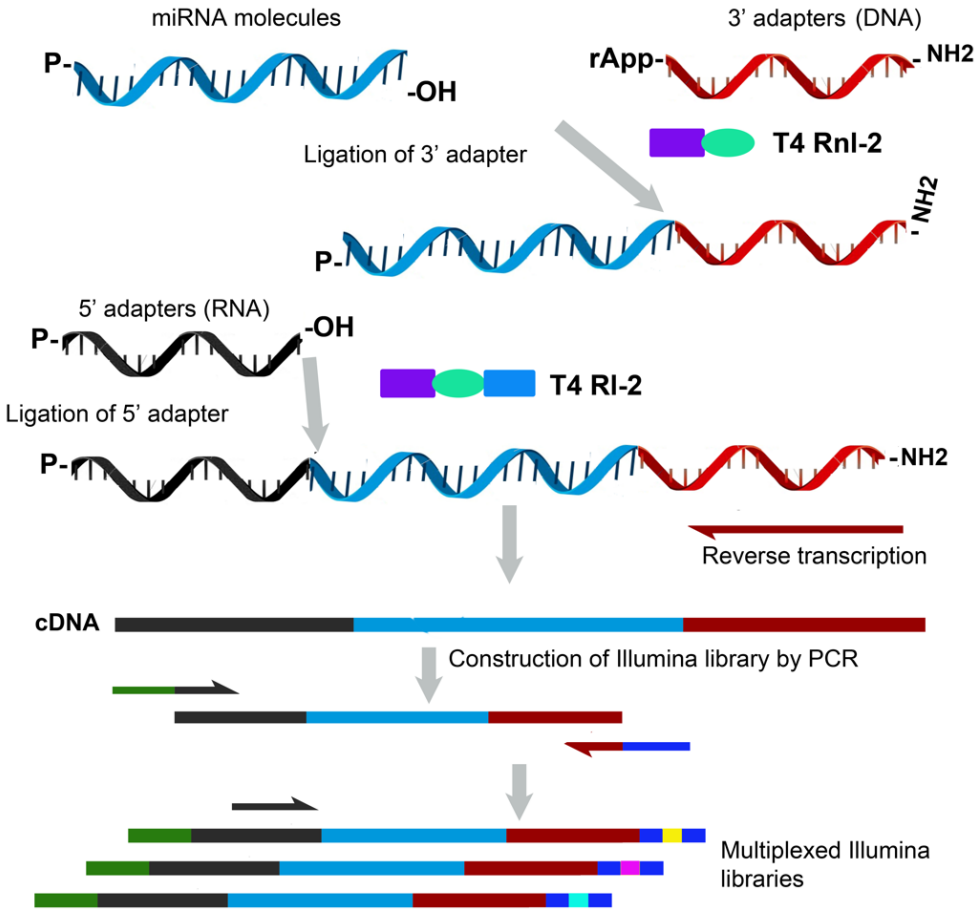


Figure 2.2 Small RNA library construction. A pre-adenylated linker was ligated to the free hydroxyl (-OH) group at the 3' end of miRNAs by a mutant T4 DNA-RNA ligase. Then a 5' adapter was ligated to the RNA, followed by reverse transcription, PCR amplification and indexing.

After quantification, 1 μg total RNA was heated at 70°C for two minutes. RNAs was then incubated with 3' adapters and T4 RNA ligase 2 (truncated) at 28°C for one hour, to selectively ligate the 3' adapters to small RNAs. Then RNAs were ligated with 5' adapters, followed by reverse transcription and PCR amplification. During PCR amplification, a different index was added to each sample. To check for the amplification of mature miRNAs, the libraries were loaded into a Bioanalyzer DNA1000 chip to check for the existence of peaks at ~145bp (See Figure 2.3 B, C). Libraries were then purified with the PCR purification kit (MinElute, Qiagen) and re-suspended in 20ul elution buffer.

Next the cDNAs were run on a 12% polyacrylamide gels for size selection. To completely separate the ~145 bp band from the neighbouring ~155 bp prominent band (which represents other small

regulatory RNAs), the electrophoresis was run for 4-5hours under 160v. The bands were then stained with 1:10000 SYBR Safe for 15 min and visualized under a UV transilluminator. The ~145bp band was carefully excised (Figure 2.3A) and transferred into gel-breaking tubes. After centrifugation, 0.4 M sodium chloride was added to the gel shreds and the tubes were rotated overnight at 4°C.

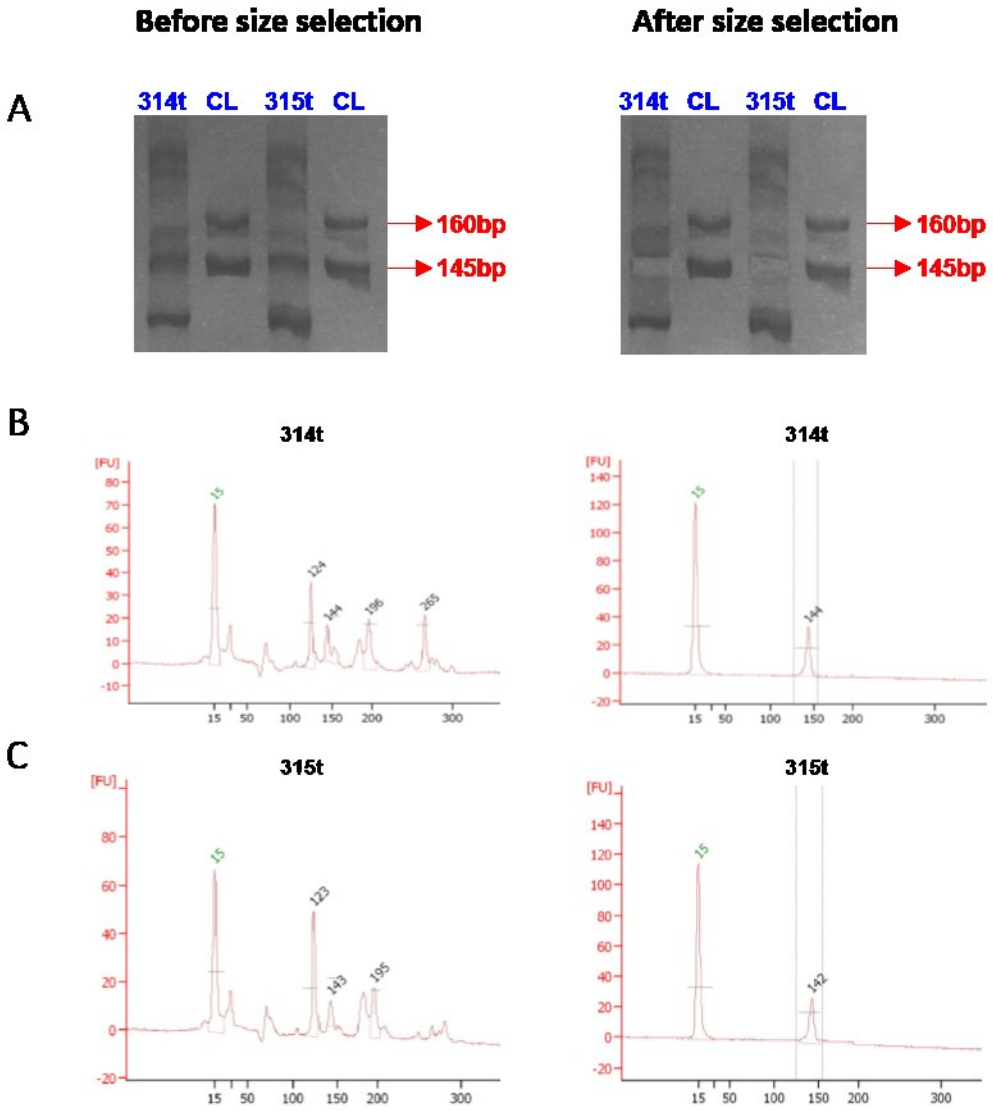


Figure 2.3 Gel purification of small RNA libraries. A) Picture of PAGE gel shows the bands of small RNA libraries (left) and excision of ~145bp bands (right) which represent miRNA libraries. **B) C)** Bioanalyzer validation of construction of small RNA libraries (left) and purification of small RNA libraries (right). CL, customer ladder

Next day, the eluted cDNAs were transferred into a new tube, precipitated in the presence of glycogen and isopropanol, and re-suspended in EB buffer. After 5 PCR cycles, the libraries were cleaned up using PCR purification columns and quantified by Qubit. For validation, samples were loaded into a Bioanalyzer DNA1000 chip to check the size and purity of the libraries (See Figure 2.3 B, C). Finally the libraries were diluted to 2nM and pooled together prior to sequencing.

The small RNA libraries were sequenced using the Illumina MiSeq platform with a single-end 50 cycles protocol, in the presence of 1% PhiX control libraries.

2.2.2.3 Analysis pipeline of miRNA expression

An in-house bioinformatics pipeline was used for analysis of miRNA sequences (Figure 2.4).

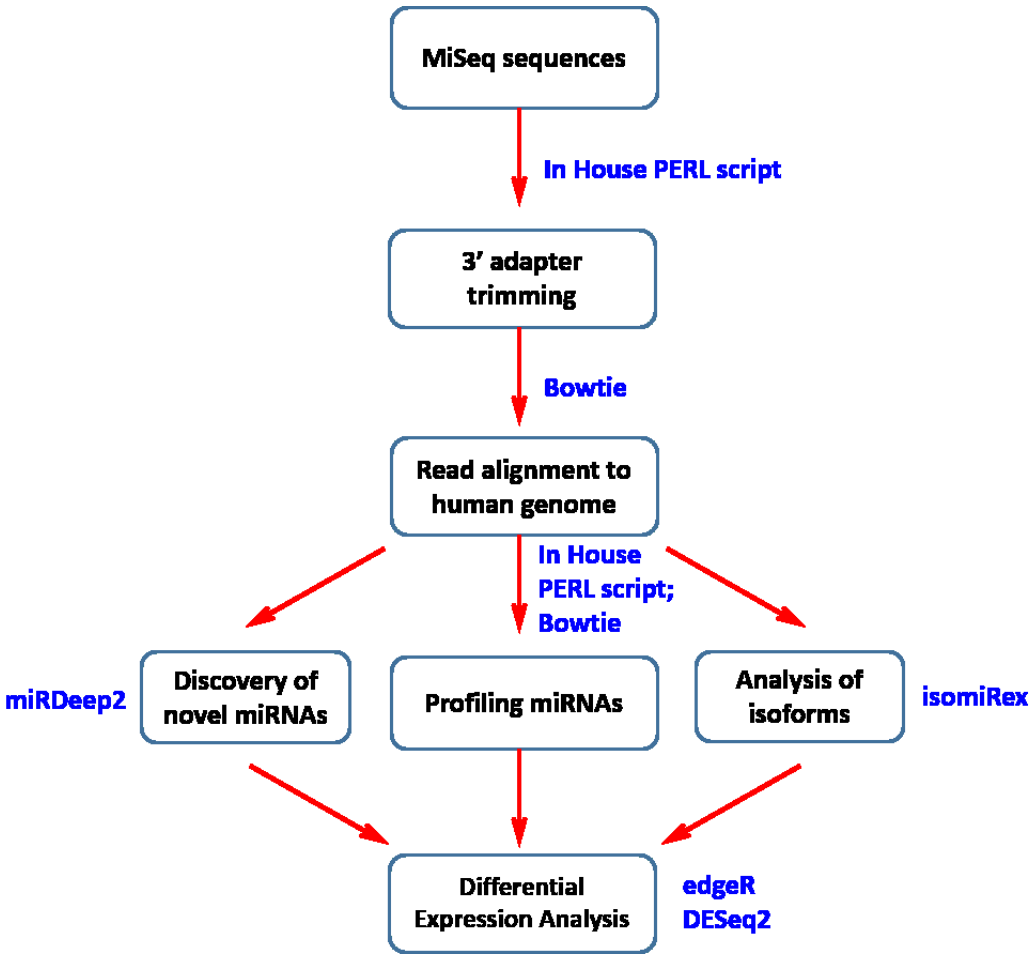


Figure 2.4 In-house bioinformatics pipeline for analysis of miRNA sequences. Libraries are sequenced in a MiSeq instrument and reads were trimmed and aligned to human genome and miRBase to profile miRNA expression. Other human sequences were used for discovering novel miRNAs. Isoforms of known miRNAs were also profiled. Finally, differential expression analysis was conducted.

2.2.2.4 Profiling miRNA expression

Analyses were run on a Gentoo Base System release 2.2 unless otherwise specified. Several analyses steps were performed with in-house Perl scripts; such scripts are included as appendix and its usage is described in the body of each of them.

Preprocessing

The MiSeq instrument outputs sequencing results in FASTQ format, which includes a sequence for each read with a corresponding quality score for each base, recorded as a letter according to American Standard Code for Information Interchange (ASCII) character-encoding scheme. The raw data is filtered of reads with ambiguous base calls according to quality score by the sequencer in real-time. Similar to the Phred scoring scheme in Sanger's sequencing, the Illumina Q score is computed as $Q = -10 \log_{10}(e)$, where e is the estimated probability of the base calling to be incorrect. Illumina started to use Sanger's transformation Phred +33 since the release of CASAVAv1.8 (a software distributed by Illumina). (http://www.illumina.com/science/education/sequencing-quality-scores.html). Illumina Q scores fit into a 0-41 scale, where Q10 indicates a potential error every 10 bases, Q20 a potential error every 100 bases, and so on.

Trimming adapters

Since the MiSeq workflow removes the 5' adapter of the small RNAs libraries, but not the 3' one, the 3' adapter was trimmed by an in-house Perl script. The sequence of the 3' adapter is the following 'TGGAATTCTCGGGTGCCAAGG'. All sequences generated by the MiSeq were 50 base pairs. The script screens each sequence, starting from its 3' end, for the presence of the first six bases of the adapter and removes the 3' moiety once such hexamer is found; if a perfect match to such hexamer is not found, then the script searches for the first nine nucleotides of the adapter in the same fashion, but in this case allows one mismatch (no indel is allowed however). Only sequences with length between 15 and 30 nt are written in a FASTQ output file (see Script 1: *trim_3_adapter.pl* in Appendix A). As quality control, FASTQ files were inspected before and after adapter trimming, with the software *fastqc*. In Figure 2.5, it is shown that the script *trim_3_adapter.pl* is able to remove all adapters from the original FASTQ file. As exemplified in Figure 2.5, the quality of our sequences was higher than 25 in all libraries. Thus, no quality trimming was required.

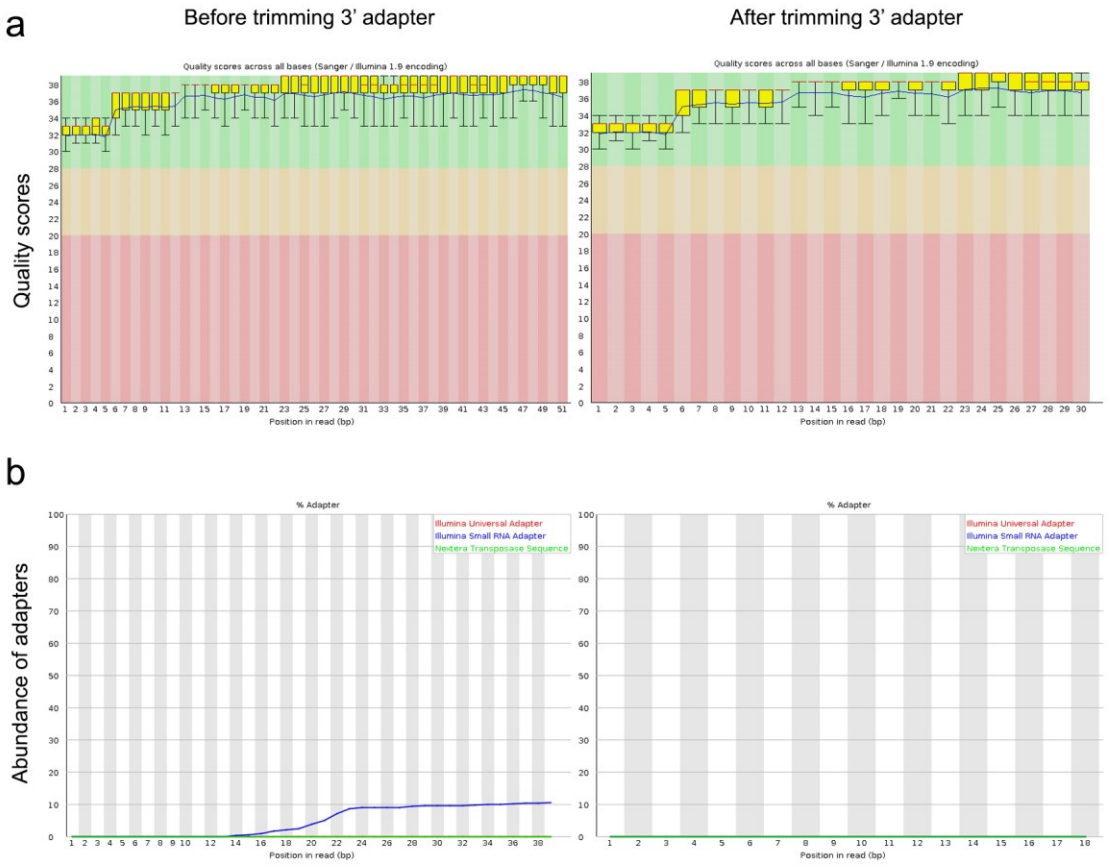


Figure 2.5 Preprocessing MiSeq sequences. a) Average quality scores of bases in all reads before and after adapter trimming; **b)** Frequency plot of 3' adapters showing that they were successfully removed by script *trim_3_adapter.pl.*

Aligning short reads against the miRNA database

The miRBase (Griffiths-Jones, 2004; Griffiths-Jones et al., 2006; Griffiths-Jones et al., 2008; Kozomara and Griffiths-Jones, 2011, 2014) (www.mirbase.org) is a public microRNAs database that annotates and curates microRNA sequences from all species for which sequencing data is available. During the course of this study, our reference miRNAs database was regularly updated whenever new versions were made available at the miRBase site. All HCC samples were aligned against the release 21 of the miRBase, while all other samples were aligned against the release 20. The miRBase reports miRNAs from all available species in a single FASTA file containing RNA sequences. For example, two typical entries in release 21 are as follows:

```
>cel-miR-90-3p MIMAT0000061 Caenorhabditis elegans miR-90-3p
```

```
UGAUAUGUUGUUUGAAUGCCCCU
>hsa-let-7a-5p MIMAT0000062 Homo sapiens let-7a-5p
UGAGGUAGUAGGUUGUAUAGUU
```

The first step was to extract only the human miRNAs from miRBase, which use the identifier 'hsa' and also contains the "Homo sapiens" name in the header of each human miRNA. The 'U' RNA base was transliterated into 'T' and a tail of 20 't' characters was appended to the 3' end of each miRNAs, to make them artificially longer than the query sequences and therefore suitable for alignments with the short reads aligner Bowtie. Spaces in the headers of the miRNAs were also removed, to prevent that any downstream tool from truncating that name. Those preparative steps were done with the script *prepare_miRNA_database.pl* (Script 2 in Appendix A).

Here, we selected Bowtie, and not Bowtie2, because it has been reported to have higher sensitivity for alignment of small RNA libraries. The database of human miRNAs prepared as above was indexed for Bowtie with the following command:

```
bowtie-build hs_miRBase21.fa hs_miRBase21
```

In order to align all files in a set of libraries in batch mode, the script *bowtie_miRNAs.pl* was used (Script 3 in Appendix A). Essentially, this script guides the user to perform bowtie alignments with the following bowtie command line:

```
bowtie -q -v 0 -p 5 <index_file> <input_file> <output_file>
-q: indicates that input file is in FASTQ format
-v: zero mismatches are allowed
-p: number of processors to use
```

Creating count table for miRNA expression

Bowtie results were parsed to produce two count tables, one with the raw data and a second one with counts normalized by million of sequences that mapped to the miRNA database used for alignment. This is called total sum normalization. Parsing of the bowtie results is accomplished with the script *count_miRNA_hits.pl,* which can be found as Script 4 in Appendix A. Such script will only count hits that mapped uniquely to the miRNA database, i.e. if a sequence aligned with zero mismatches to more than one miRNA in the database, such hit will be excluded from the count tables. Total sum normalized counts are useful for creation of descriptive plots like bar graphs and box plots, for example, but were not used for differential expression analyses here.

2.2.2.5 MiRNA count normalization and differential expression analyses

In general, the goal of normalization is to ameliorate the influence of sample-specific technical effects on the difference observed between samples, so that the remaining difference can be assumed to be biological in nature. However, the distribution of gene expression data for the whole transcriptome (RNA composition) may differentially affect the relative expression of specific sets of genes. For example, if a few genes express to exceptionally high levels in some but not all samples, and as a result occupy 50% of the corresponding libraries, the rest of genes will occupy only the other 50%. When compared to samples where the same genes do not express to such aberrantly high levels, many low-expressing genes will appear artefactually down-regulated. This effect is independent of library size. Thus, although normalizing gene expression according to sequencing depth seems correct, it may be insufficient in several biological contexts.

As mentioned in the introduction chapter, normalization of count data from NGS libraries is a matter of intense debate. Although there is no consensus on which is the most appropriate normalization procedure, the internal normalization methods implemented by the R packages edgeR and DESeq2 have received good evaluations in several comparative studies. We therefore used both approaches and combined the results obtained.

edgeR

edgeR implements the so-called trimmed mean of M-values (TMM) normalization (Robinson and Oshlack, 2010), where M is the log fold-change between two samples. Formally, to moderate the degree of overdispersion across genes, edgeR uses an empirical Bayes procedure, in which the dispersion towards a consensus value is shrunk by borrowing information between genes (Robinson and Smyth, 2007). The idea here is to minimize the log fold-changes between samples for most genes, which is accomplished by excluding the counts of those genes that express to very high or to very low levels. Once the scale factor has been calculated using TMM, the effective library size is obtained multiplying the actual library size by the scaling factor. The effective library size is then used for subsequent analyses (Chen Y, 2015). Our own implementation of edgeR (version 2.4.0) is presented in the script *edgeR_glm.R*, see Script 5 in Appendix A.

In general, the analysis conducted by our script includes the following steps: 1) importing a counts table, 2) converting counts table into an edgeR object, filtering out low-count reads and calculating the normalization factors, 3) generating a multi-dimensional scaling plot, 4) estimating dispersion, and 5) doing exact test and outputting the list of differential expressed miRNAs.

Determination of normalization factors can be accomplished by simply applying the calcNormFactors of edgeR:

```
miRcountTable <- calcNormFactors(miRcountTable)
```

However it is not necessary, since the differential expression analysis routine will implement normalization automatically.

In our case, to compare the miRNA expression in paired tumor and non-tumor tissues, two sources of variation were included into a generalized linear model (GLM). First a new factor was created to group paired samples inside each patient (exemplified here for the four pairs of explanted tumors):

```
Patient <- factor(c(1,1,2,2,3,3,4,4))
```

A second factor to model the difference attributed to hepatocellular carcinoma was created:

```
Tissue <- factor(c("N","T","N","T","N","T","N","T"))
```

And the general linear model was as follows:

```
design <- model.matrix(~Patient+Tissue)
```

edgeR uses the Benjamini-Hochberg procedure (Benjamini Y, 1995) for correction of p values and determination of the false discovery rate (FDR). Differential expression results from edgeR were parsed manually, using a p value < 0.05 and a FDR < 0.05.

DESeq2

As edgeR, DESeq2 also uses a negative binomial distribution to model the counts assigned to each gene. The mean of such distribution is the abundance of a gene multiplied by a normalization factor. The size factors for each sample are calculated as follows: i) the geometric mean of counts across samples is calculated for each gene; ii) the original counts are divided by the geometric mean; iii) the median of such ratios in each sample is the size factor for each sample. This is called the median-of-ratios method, originally developed in a previous version of this package called DESeq (Anders and Huber, 2010).

Similar to edgeR, steps for differential expression analysis of DESeq2 are: 1) estimation of size factors (which control for differences in the library size of the sequencing experiments), 2) estimation of dispersion for each gene, fitting a generalized linear model, and finally 3) a Wald test. Our own implementation of DESeq2 (version 1.8.1) is presented in the script *DESeq2.R*, see Script 6 in Appendix A.

Count data was imported into a DESeqDataSet object, which also included a metadata file that contains relevant information about the sources of variability to be considered during differential expression analysis. For example, for our four explanted liver samples, the metadata file was as follows:

SampleName	condition	patient	label
nt311	nonTumor	pt311	nt311
t311	tumor	pt311	t311
nt314	nonTumor	pt314	nt314
t314	tumor	pt314	t314
nt315	nonTumor	pt315	nt315
t315	tumor	pt315	t315
nt338	nonTumor	pt338	nt338
t338	tumor	pt338	t338

To perform differential expression with DESeq2, we simply used the function DESeq, which automatically implements the normalization procedure. Namely:

```
ddsMat <- DESeq(ddsMat, test="LRT", reduced = ~ patient)
```

Extraction of differentially expressed miRNAs, using a p value < 0.05 and a FDR < 0.05 was conducted with the following commands:

```
res <- results(ddsMat)
res.05 <- results(ddsMat, alpha=0.05)
table(res.05$padj < 0.05)
```

edgeR and DESeq2 analyses were performed under the R statistical computing environment (R core team, 2015) version 3.2.0 (2015-04-16) downloaded from https://www.r-project.org/ using the R studio on platform: x86_64-w64-mingw32/x64 (64-bit).

2.2.2.6 Prediction of miRNA targets and gene pathways

In this study we applied the most widely used algorithms, TargetScan and DIANA-microT-CDS, to predict miRNA targets. DIANA mirPath was used to predict the pathway deregulated miRNA may affect. All the programs are free-accessed online resources.

TargetScan

The TargetScan Human release 7.0 was accessed at http://www.targetscan.org/. To predict the targets of a certain miRNA, one simply inputs the full name of the miRNA and click 'submit'. The output page gives a table of top target genes ranked by a cumulated weighed context++ score, and prediction information including number of 3p-seq tags supporting UTR +5, link to site on UTR, target site counts, representative miRNA and aggregate P_{CT}.

DIANA-microT-CDS

The 5th version of the microT-CDS web server can be accessed on http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index. The output is a list of targets ranked by miTG score from highest to lowest, whereas the default threshold of miTG score is 0.7. The prediction can be tailored more sensitively or more stringently by adjusting this score in advanced filter options. Additional details of prediction can be shown, such as target region, binding type and position, if a binding site is conserved or not, by clicking the arrow in the right end of the lane. The table also shows whether the predicted mRNA is also the target given by the packages miRanda and TargetScan.

DIANA mirPath

The miRNA pathway analysis web-server version 2.0 can be accessed at http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpath/index. The program utilizes predicted miRNA targets provided by microT-CDS or DIANA-TarBase, and combines with sophisticated merging and a meta-analysis algorithm to locate the most likely pathways that the miRNAs of interest may affect. By clicking the pathway name one can also visualize the pathway picture in which all targets are marked in different colors.

2.3 Experiment #3: profiling gene expression in human HCC-affected livers by NGS and bioinformatics approaches

2.3.1 Sample information

The same sample for gene expression profiling are the same four pairs of paired tumor and non-tumor tissues taken from explanted livers (Table 2.2). Please see details in section 2.2.1.

2.3.2 Methods

2.3.2.1 Transcriptome library construction and sequencing from explanted livers

RNA-seq libraries were constructed using the TruSeq RNA sample prep kit (Illumina), following the recommendations of the manufacturer.

In brief the poly-A-containing mRNAs molecules were purified from total RNAs using poly-T oligo-attached magnetic beads and then chemically fragmented to desired size. Fragmented mRNAs were reverse transcribed and the second strand cDNA was synthesized. Then the reactions were cleaned up using AmPure XP beads, the ends of cDNA were repaired and poly-A tails were added. In next steps, the adapters, which contain indices, were ligated to the cDNA, followed by two rounds of beads cleaning. The libraries were enriched by PCR for 15 cycles, followed by beads clean up. The resulting libraries were then titrated to 2 nM and quantified by an Agilent Bioanalyzer.

RNA-seq libraries were sequenced on an Illumina HiSeq 2500 instrument at an approximate depth of 50 million reads per transcriptome using a paired-end 150 cycles protocol. Bases with a quality score below 30 were trimmed off and sequences shorter than 75 base pairs were discarded.

2.3.2.2 Bioinformatic analysis of RNA-seq data

A central step during analysis of RNA-seq data is the detection of splicing junction, the point where two exons are joined after a separating intron is excised during splicing. This allows determining the relative abundance of several splicing isoforms. Several tools for mapping reads to a reference genome in a splicing-aware manner have been developed, including STAR (Dobin et al., 2013) and TopHat2 (Kim et al., 2013).

A pipeline that includes TopHat2 for mapping of reads, and edgeR and DESeq2 for differential expression analysis was recently described by the developers of the edgeR and DESeq2 packages (Anders et al., 2013). Such pipeline was implemented here for the analysis of RNA-seq data. Such protocol has a modular structure, and therefore can be entered through multiple points. Thus, the implementation presented here does not faithfully adhere to Anders et al., 2013, but follows their general workflow (Figure 2.6).

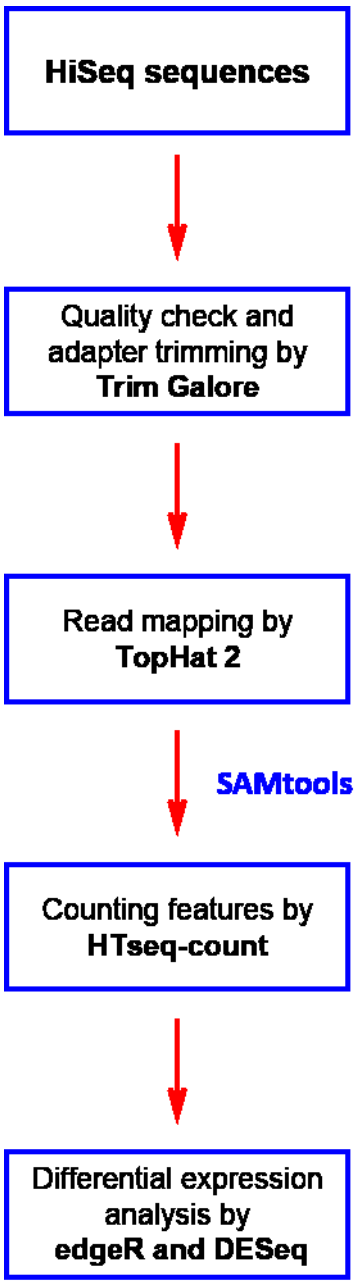


Figure 2.6 Bioinformatics pipeline for analysis of RNA-seq sequences. HiSeq sequences were trimmed and aligned to the reference. The resulted BAM files were then sorted and count table was generated. Finally the normalization and differential expression analysis was performed.

A Gene Transfer Format (GTF) file is needed for the alignment of reads in a splicing-aware manner. A GTF file is a tab-delimited format that provides information about the structure and location of different elements in the genome. To avoid inconsistencies in coordinates, we downloaded the human reference genome (hg19) and the corresponding GTF file from the same source at https://ccb.jhu.edu/software/tophat/igenomes.shtml.

Due to budget limitations, we conducted RNA-seq experiments for only the four explanted livers. Initially, adapter sequences and low-quality bases (Q < 20) were removed from libraries using the software 'Trim Galore!', which was obtained from the following site http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, and run as follows:

```
trim_galore --paired -phred33 -q 20 -a CTGTCTCTTATACACATCT -a2 AGATGTGTATAAGAGACAG sample1_R1.fastq.gz sample1_R2.fastq.gz
```

Subsequently, the high quality reads were aligned to the reference genome using the wrapper TopHat2. Because alignment of RNA-seq data is much more compute-intensive than the analogous task with miRNAs, TopHat2 alignments were run on a compute node, using the following an in-house batch script (see Script 7: *batch_script.pl*)

After alignment, the binary version of the Sequence Alignment Map (SAM) files generated by TopHat2 were sorted using the SAM tools (Li et al., 2009a):

```
for DIR in 3*thout; do samtools sort -n ${DIR}/accepted_hits.bam ${DIR}/accepted_hits_sn; done
```

Sorted BAM files were converted into SAM ones, with the SAMtools:

```
for DIR in 3*thout; do samtools view -o ${DIR}/accepted_hits_sn.sam ${DIR}/accepted_hits_sn.bam; done
```

The newly generated format is used for the counting with the HTseq-count tool (Anders et al., 2015). To accomplish such task, we first exported the path indicating the location of the GTF file:

```
export gtf='/data/…/Homo_sapiens/…/Genes/genes.gtf'
```

Then, aligned reads were counted:

```
for DIR in 3*thout; do htseq-count -s no -a 10 ${DIR}/accepted_hits_sn.sam $gtf >${DIR}/${DIR}.counts; done
```

Finally, results from all files were merged in a single counts table, as follows:

```
paste <(cut -f 1-2 311n.counts) <(cut -f 2 314n.counts) <(cut -f 2
315n.counts) <(cut -f 2 338n.counts) <(cut -f 2 311t.counts) <(cut -f 2
314t.counts) <(cut -f 2 315t.counts) <(cut -f 2
338t.counts) >4HCC.RNAseq.counts.xls
```

Table 4HCC.RNAseq.counts.xls contains a series of columns: the first one is the name of the genes and then a column for each sample, containing the counts (abundance) for that gene. Such count table was then imported either into edgeR or DESeq2 for differential expression analysis using the script 5 and 6 in Appendix A as described in section 2.2.2.5.

References

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome biology 11, R106.

Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., Robinson, M.D., 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8, 1765-1786.

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Benjamini Y, H.Y., 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. . J R Stat Soc B Methodol 57, 289-300.

Chen Y, M.D., Robinson M, Smyth GK, 2015. edgeR: differential expression analysis of digital gene expression data User's Guide.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Garten, R.J., Davis, C.T., Russell, C.A., Shu, B., Lindstrom, S., Balish, A., Sessions, W.M., Xu, X., Skepner, E., Deyde, V., Okomo-Adhiambo, M., Gubareva, L., Barnes, J., Smith, C.B., Emery, S.L., Hillman, M.J., Rivailler, P., Smagala, J., de Graaf, M., Burke, D.F., Fouchier, R.A., Pappas, C., Alpuche-Aranda, C.M., Lopez-Gatell, H., Olivera, H., Lopez, I., Myers, C.A., Faix, D., Blair, P.J., Yu, C., Keene, K.M., Dotson, P.D., Jr., Boxrud, D., Sambol, A.R., Abid, S.H., St George, K., Bannerman, T., Moore, A.L., Stringer, D.J., Blevins, P., Demmler-Harrison, G.J., Ginsberg, M., Kriner, P., Waterman, S., Smole, S., Guevara, H.F., Belongia, E.A., Clark, P.A., Beatrice, S.T., Donis, R., Katz, J., Finelli, L., Bridges, C.B., Shaw, M., Jernigan, D.B., Uyeki, T.M., Smith, D.J., Klimov, A.I., Cox, N.J., 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325, 197-201.

Griffiths-Jones, S., 2004. The microRNA Registry. Nucleic Acids Res 32, D109-111.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J., 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34, D140-144.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., Enright, A.J., 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res 36, D154-158.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 22, 568-576.

Kozomara, A., Griffiths-Jones, S., 2011. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39, D152-157.

Kozomara, A., Griffiths-Jones, S., 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42, D68-73.

Leon, A.J., Banner, D., Xu, L., Ran, L., Peng, Z., Yi, K., Chen, C., Xu, F., Huang, J., Zhao, Z., Lin, Z., Huang, S.H., Fang, Y., Kelvin, A.A., Ross, T.M., Farooqui, A., Kelvin, D.J., 2013. Sequencing, annotation, and characterization of the influenza ferret infectome. Journal of virology 87, 1957-1966.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., Marshall, D., 2010. Tablet--next generation sequence assembly visualization. Bioinformatics 26, 401-402.

R core team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology 11, R25.

Robinson, M.D., Smyth, G.K., 2007. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23, 2881-2887.

Rowe, T., Leon, A.J., Crevar, C.J., Carter, D.M., Xu, L., Ran, L., Fang, Y., Cameron, C.M., Cameron, M.J., Banner, D., Ng, D.C., Ran, R., Weirback, H.K., Wiley, C.A., Kelvin, D.J., Ross, T.M., 2010. Modeling host responses in ferrets during A/California/07/2009 influenza infection. Virology 401, 257-265.

Chapter 3

Next-generation sequencing and bioinformatic approaches to detect and analyze influenza virus in ferrets

A version of this chapter was published in:

Lin Z, Farooqui A, Li G, Wong GK, Mason AL, Banner D, Kelvin AA, Kelvin DJ, León AJ. Next-generation sequencing and bioinformatic approaches to detect and analyze influenza virus in ferrets. J. Infect. Dev. Ctries. 2014 Apr 15;8(4):498-509. doi: 10.3855/jidc.3749.

3.1 Introduction

Human influenza remains a major burden of disease and concern in public health (Suk and Semenza, 2011). According to World Health Organization (WHO), seasonal epidemics affect 5-10% adults and 20-30% children globally in a single year. Symptoms are usually sudden onset of high fever, dry cough, sore throat and running nose, in company with fatigue, muscle ache and headache. While most patients recover within a week, certain populations – such as children younger than two years, seniors elder than 65 years, pregnant women and individuals with underlying chronic diseases – are more likely to suffer clinical complications and even death. It is estimated that annual seasonal influenza epidemics cause three to five million severe infections, and a quarter to half million deaths worldwide (WHO, 2014a). When an influenza epidemic spreads all over the world, it is called influenza pandemic. When a pandemic with a new influenza virus occurs, a larger proportion of the population is usually affected with more severe clinical signs due to the lack of prior immunity against the emerging virus. The most notorious pandemic in human history is the 1918 'Spanish Flu', which affected 500 million people and was responsible for more than 40 million deaths (Taubenberger and Morens, 2006). The most recent influenza pandemic emerged in 2009 and killed 280,000 people world widely, according to a modelling study (Dawood et al., 2012).

Influenza viruses, including type A, B and C, constitute three genera in the Orthomyxoviridae family. The genome of influenza A virus comprises eight segmented, negative-sense, single-stranded RNAs (Figure 3.1). The major proteins these RNA segments encode are PB2, PB1, PA, HA (hemagglutinin), NP, NA (neuraminidase), M1, M2, NS1 and NS2 (or NEP) (Bouvier and Palese, 2008). HA and NA proteins are envelope glycoproteins (Figure 3.1), which are important for viral invasion and replication (Bouvier and Palese, 2008). They are also antigenic proteins that the host immune response targets (Iwasaki and Pillai, 2014).

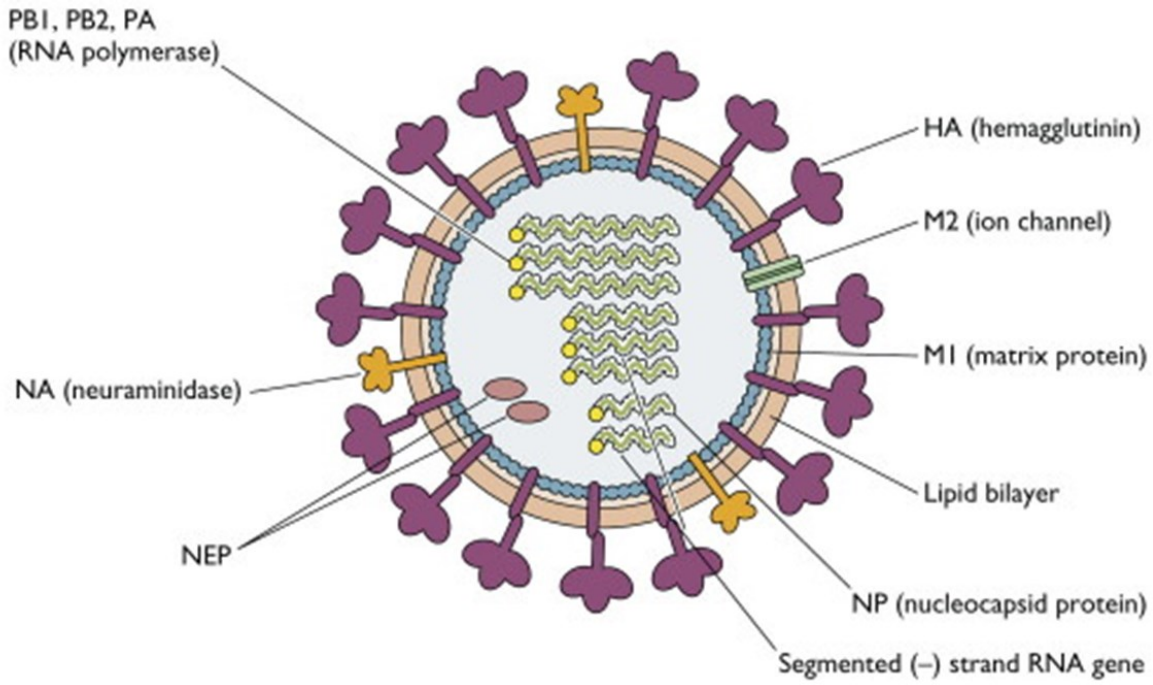


Figure 3.1 Molecular structure of influenza A virus. The genome is composed of eight segmented RNAs which encode core proteins PB2, PB1, PA, HA, NP, NA, M1, M2, NS1 and NEP. Among them HA and NA are surface proteins. This figure is cited from http://www.virology.ws

Changes in influenza virus are due to antigenic drifting (i.e. differences in seasonal H1N1 and H3N2), antigenic shifting (i.e. a new emerging pandemic strain) or a combination of both (i.e. changes in avian influenza H7N9) (Farooqui et al., 2015). Antigenic drift is the accumulation of point mutations in the HA and NA genes which results in the change of their antigenicity and is responsible for the continuous evolution of circulating strains (Bouvier and Palese, 2008; Smith

et al., 2004); these antigenic changes cause a partial antigenic mismatch with the pre-existing host immunity that was established during previous infections with a parent virus or vaccination (Carrat and Flahault, 2007) (Medina and Garcia-Sastre, 2011). The major change in virus antigenicity, however, is caused by antigenic shift. Antigenic shift of influenza virus occurs by the combination and reassortment of antigenically distant versions of the gene segments encoding HA or NA after coinfection of two or more viruses in a single host (Taubenberger and Kash, 2010). This process can happen not only in humans but also in domestic mammals and in avian species, which are the natural host of influenza virus. Since the majority of the human population is immunologically naïve to novel subtypes, influenza pandemics arise as a consequence antigenic shift, exampled by the 1918 "Spanish Flu" (Bouvier and Palese, 2008) and the 2009 H1N1 pandemic (Garten et al., 2009).

Due to the antigenic diversity of the influenza strains and their elevated mutation rates, as compared to other respiratory viruses, molecular characterization of the circulating strains is necessary for updating the vaccines against seasonal strains (Carrat and Flahault, 2007), rapid development of vaccines against emerging strains (Carrat and Flahault, 2007), detection of resistance to neuraminidase inhibitors and consequent treatment updates (Stephenson et al., 2009), as well as veterinary management of influenza-susceptible species (Medina and Garcia-Sastre, 2011), etc.

Traditionally, PCR-based methods have been adopted widely in the diagnostics of influenza viruses. However due to the high mutation rates of influenza genes it is not unusual for surveillance studies to show a fraction of samples which are influenza A positive but unsubtypable (Farooqui et al., 2011; McHardy and Adams, 2009). Microarray-based approaches to viral detection, such as ViroChip (Wang et al., 2002), constitute a good alternative for viral screening; however, the rapidly evolving nature of many viruses makes it difficult for microarrays to deliver the same level of detail as sequencing-based approaches (Greninger et al., 2010). Next-generation sequencing (NGS) allows the detection of pathogens when only little prior knowledge of their genomes is available and without the need for target-specific PCR primers. Additionally, NGS technologies deliver a large amount of genomic information that allows the study of additional aspects such as development of resistance to antivirals, variety of quasi-species and determinants of adaptation to different host species (Ghedin et al., 2011; Kuroda et al., 2010).

The common process behind most NGS approaches begins with random fragmentation of the template DNA chains and binding to a solid substrate, followed with parallel PCR amplification that results in spatially separated clonal populations of DNA which can be sequenced

independently (Metzker, 2010). Interpretation of NGS data presents bioinformatic challenges due to the large size and complexity of the sequencing data (Nowrousian, 2010). The initial approach to the analysis of NGS data can be done using three different types of tools: short-read aligners, de novo assemblers and general-purpose aligners. In those scenarios intended to confirm the presence, quantify or study minor sequence variations of known viruses, short-read aligners such as Bowtie(Langmead et al., 2009), Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010) or Short Oligonucleotide Analysis Package 2 (SOAP2) (Li et al., 2009b), provide a well-established and rich framework when used in combination with other downstream analysis tools. For viral discovery studies or when a significant dissimilarity is expected between the short-read sequences and the viral reference, de novo assemblers such as ABySS (Birol et al., 2009), Velvet (Zerbino, 2010) or SOAPdenovo (Li et al., 2010) can be used to generate longer sequence contigs; subsequent BLAST (Altschul et al., 1990) analysis by carefully adapting the parameters to the necessities of each scenario allows the detection of viral sequences in a collection of contigs by aligning them with a database of known viral sequences. Another approach to scenarios with high sequence dissimilarity is to use the general purpose aligner BLAST to interrogate directly the short-read libraries against the viral reference sequences followed by assembly of the consensus sequence. Direct BLAST analysis of short-reads can be very sensitive, provided that the short-reads are of sufficient length so that the analysis can “absorb” a number of indels and mismatches without causing a dramatic decrease in the similarity score.

In this study, we explore different existing paths intended to analyze the data generated by NGS in the context of viral research. Using short-read sequences generated from lung tissue of ferrets experimentally infected with influenza A/California/07/2009 (H1N1), we illustrate in detail the bioinformatic process to classify those short-reads matching influenza sequences and the subsequent generation of the consensus sequences. We simulated the characterization of an “unknown” influenza virus and explored the viral variants or quasispecies within a sample. Finally, we evaluate different options that must be considered during the design of any NGS-based strategy for viral detection, such as NGS platform and sequencing length and depths, which can cause a dramatic impact in the study results.

3.2 Results

3.2.1 Overview of the Illumina sequencing data output

RNA was purified from the lung tissues collected on days 1, 3, 5 and 14 post infection; for each of those time-points, one sample was subjected to NGS analysis at BGI, Shenzhen, China. The

sequencing analysis produced 20 million paired-end reads per sample (totally 40 million reads per sample), 90 base-pairs long. The vast majority of the sequences correspond to the ferret mRNA and only a small fraction to influenza virus (determined by BLAST analysis, details are shown below).

3.2.2 Detection of influenza virus by using the fast aligner Bowtie and SAM Tools.

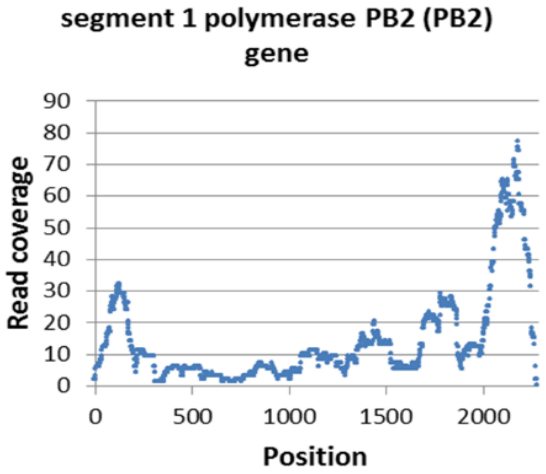
A Bowtie index was generated using the eight viral segments of A/California/07/2009. Later, the paired-end reads from each sample, which were contained in their respective fastq format files, were aligned with Bowtie and the resulting output files were generated in SAM format. Next, the consensus sequences for all the viral genes were generated with the pileup command of SAM Tools; as expected, the resulting sequences were well formed and they showed a high degree of similarity with respect to the reference sequences. The alignments were loaded in the visualization tool Tablet obtaining the number of short-read alignments for each viral gene (Table 3.1) and the sequencing coverage at every nucleotide position (Figure 3.2).

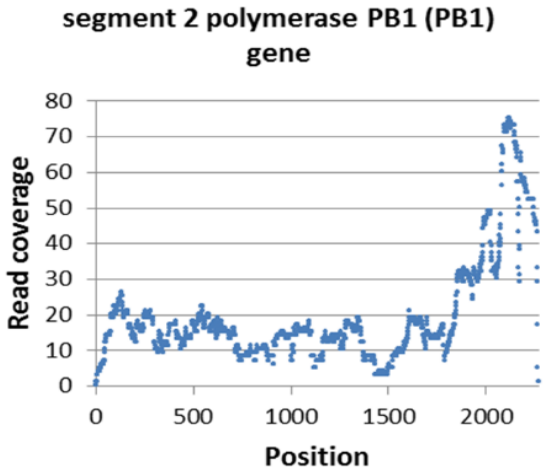
Table 3.1 Analysis of next-generation sequencing data with Bowtie2 and BLASTn to characterize the genomic segments of influenza virus. The table shows the number of reads that match the influenza segments and the % length of the consensus sequence with respect to each reference sequence at different times post-infection (PI).

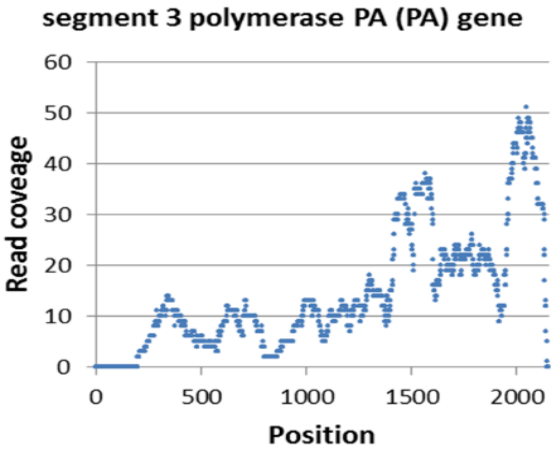
Segment[a]	Ferret day1 PI	Ferret day3 PI	Ferret day5 PI	Ferret day14 PI
Bowtie2 and SAM Tools[b]				
PB2	50 (24.0)	227 (72.9)	350 (91.3)	0 (0)
PB1	156 (79.0)	345 (93.0)	482 (99.5)	0 (0)
PA	27 (11.1)	125 (60.9)	340 (87.2)	0 (0)
HA	257 (95.2)	1,461 (99.4)	3,154 (99.9)	0 (0)
NP	494 (96.8)	2,678 (99.7)	6,172 (100)	0 (0)
NA	119 (85.4)	742 (97.4)	1,553 (99.5)	0 (0)
MP	321 (92.3)	1,730 (97.3)	4,421 (99.8)	0 (0)
NS	334 (85.1)	1,969 (96.1)	5,092 (98.6)	0 (0)
Total influenza	1,758	9,277	21,564	0
% library	0.0043	0.0231	0.0539	0
BLASTn and Iliad Assembler				
PB2	50 (24.9)	232 (91.4)	381 (96.4)	0 (0)
PB1	157 (91.3)	386 (95.2)	505 (99.3)	0 (0)
PA	29 (17.8)	132 (76.3)	358 (90.2)	0 (0)
HA	274 (98.1)	1,508 (99.5)	3,363 (99.6)	0 (0)
NP	329 (99.6)	1,752 (99.6)	4,489 (99.1)	0 (0)
NA	121 (91.2)	759 (99.4)	1,578 (99.9)	0 (0)
MP	511 (98.2)	2756 (99.6)	6,461 (99.9)	0 (0)
NS	340 (94.4)	2055 (97.7)	5,362 (98.4)	0 (0)
Total influenza	1,811	9,580	22,497	0
% library	0.0045	0.0239	0.0562	0

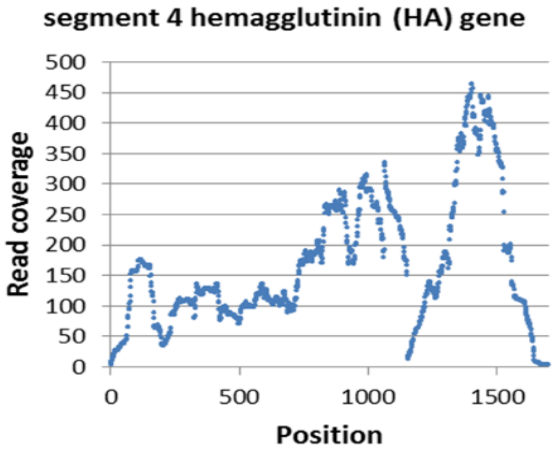
a The nucleotide sequences of influenza A/California/07/2009 (H1N1) were used as reference. The GenBank accession numbers were as follows: FJ966976 (PB2), FJ966978 (PB1), FJ966977 (PA), FJ966974 (HA), FJ969536 (NP), FJ984386 (NA), FJ966975 (MP) and FJ969528 (NS).

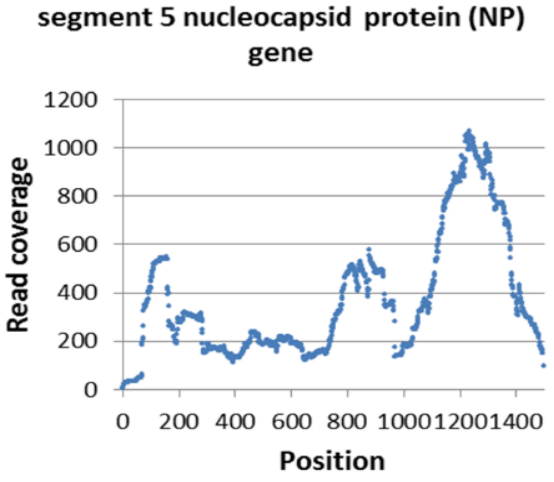
b Only positions with a minimum coverage of 3 where considered.

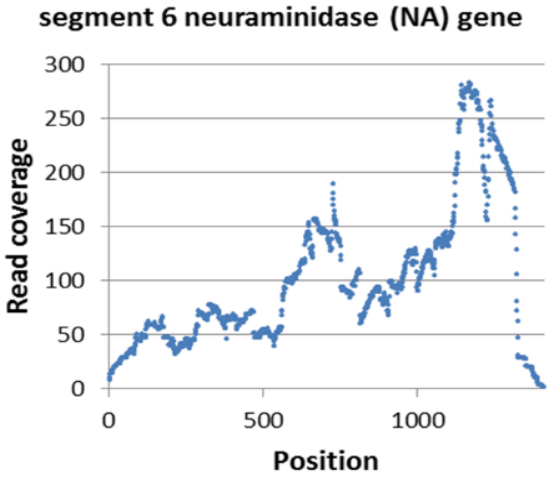

segment 1 polymerase PB2 (PB2) gene
Read coverage
Position
0
10
20
30
40
50
60
70
80
90
0
500
1000
1500
2000


segment 2 polymerase PB1 (PB1) gene
Read coverage
Position
0
10
20
30
40
50
60
70
80
0
500
1000
1500
2000


segment 3 polymerase PA (PA) gene
Read coveage
Position
0
10
20
30
40
50
60
0
500
1000
1500
2000


segment 4 hemagglutinin (HA) gene
Read coverage
Position
0
50
100
150
200
250
300
350
400
450
500
0
500
1000
1500


segment 5 nucleocapsid protein (NP) gene
Read coverage
Position
0
200
400
600
800
1000
1200
0
200
400
600
800
1000
1200
1400


segment 6 neuraminidase (NA) gene
Read coverage
Position
0
50
100
150
200
250
300
0
500
1000

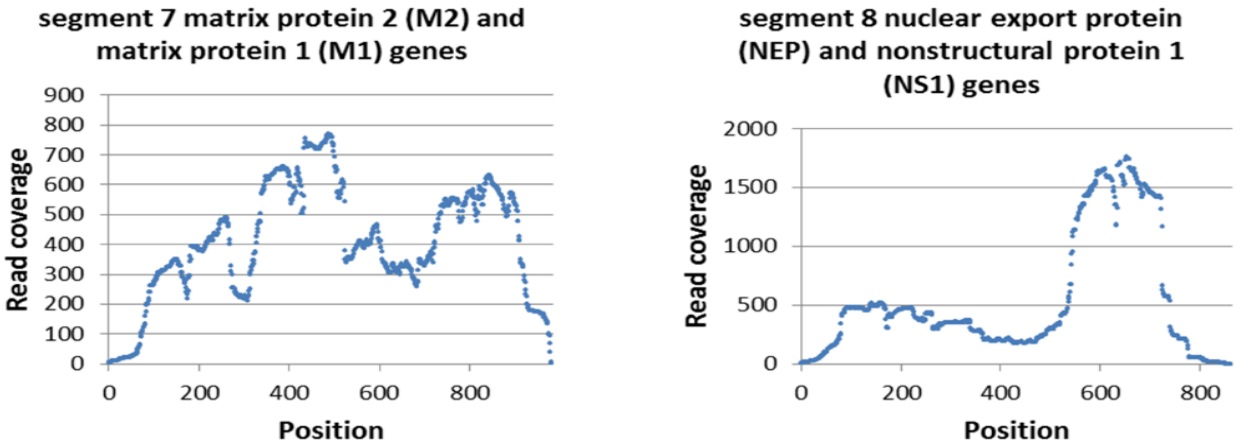


Figure 3.2 Sequencing coverage at every nucleotide position for the genomic segments of influenza virus. Short-reads from day 5 post-infection were aligned to the sequences from A/California/07/2009 using Bowtie2; the resulting SAM file was loaded in the visualization tool Tablet to generate coverage summaries, and later, these were plotted with Microsoft Excel.

3.2.3 Detection of influenza virus by BLAST analysis and Iliad Assembler

BLAST was used to pre-select the short-reads that match the viral genes and the generation of the consensus sequence was performed by guided assembly with Iliad Assembler. First, we generated BLAST databases containing the reads from each sample; the short-read sequences were subsequently used as the BLAST "subject" during the analysis. Since this method does not allow the processing of paired-end reads, all the reads were treated as single-end. Next, each viral segment of A/California/07/2009 was independently used as the BLAST "query" using the following BLASTn parameters: word_size 12, evalue 1E-12, reward 2 and penalty -3; the results are shown in Table 1. The number of short-reads mapped to influenza genes and the percentage of flu-matching reads per sample were as follows: day 1: 1,811 reads (0.005%), day 3: 9,580 reads (0.024%), day 5: 22,497 reads (0.056%) and on day 14 no influenza sequences were detected. The differences in the number of reads among time-points are in accordance with the evolution of the viral loads previously described for this infection model (Rowe et al., 2010). Finally, the pre-selected short-reads were processed with Iliad Assembler to generate the consensus sequence and to calculate the coverage percentage (Table 3.1). BLAST analysis of the final

assembly of the hemagglutinin gene (day 5 post-infection data) showed 1,694 out of 1,695 identities and zero gaps with respect to the reference sequence.

3.2.4 Detection of influenza virus by de novo assembly

We achieved a reduction in the size of the short-read libraries of around 50% (Table 3.2) by subtracting those reads matching the ferret genomic DNA; this led to a significant reduction of the computing workload during the de novo assembly process. After performing several preliminary runs to optimize the program settings, de novo assembly was performed using the subtracted libraries and the ABySS option k=32 and paired-end mode. Contigs that significantly matched influenza sequences were identified with BLAST and Iliad Assembler was used to calculate % length of the assembly with respect to each reference sequence (Table 3.2). Even when the number of available reads was low, the results were comparable to those obtained with Bowtie2 or direct BLAST analysis (Table 3.1).

Table 3.2 Summary of de novo assembly with ABySS and subsequent identification of influenza-matching contigs by BLAST analysis at different times post-infection (PI).

	Ferret Day 1 PI		**Ferret Day 3 PI**		**Ferret Day 5 PI**	
Short-read sequences library size [a]						
Total short-reads	40 million		40 million		40 million	
After subtraction	19.0 million		18.2 million		20.3 million	
Number of contigs assembled by ABySS [b]						
Total (≥50bp)	442,693		433,748		398,298	
≥200bp	40,713		36,089		37,711	
≥500bp	13,344		11,846		13,755	
≥1000bp	4,412		3,797		4,802	
Influenza-matching contigs [c]	Contigs	% coverage	Contigs	% coverage	Contigs	% coverage
Segment 1 (PB2)	3	24.8	9	82.6	2	96.3
Segment 2 (PB1)	6	77.3	4	97.0	1	100.0
Segment 3 (PA)	5	23.7	5	68.3	3	90.7
Segment 4 (HA)	3	99.2	1	99.9	11	100.0
Segment 5 (NP)	1	100.0	3	99.5	21	99.6
Segment 6 (NA)	3	89.0	1	99.4	2	99.8
Segment 7 (MP)	1	92.5	5	99.0	24	99.9
Segment 8 (NS)	4	94.4	7	92.4	20	96.8

[a] Reads matching ferret sequences (MusPutFur1.0) were subtracted with Bowtie (v0.12.7).
[b] ABySS was run in paired-end mode and k=32
[c] Sequences from A/California/07/2009 (H1N1) were used as reference. % coverage was calculated with Iliad Assembler.

3.2.5 Simulation of virus discovery using BLAST

We aimed to use a realistic scenario to explore the challenges involved in the characterization of new viruses when using other previously known viruses with high degrees of dissimilarity as reference for the sequence alignments. We focused this simulation on the hemagglutinin gene because this gene presents the highest degree of sequence variability among strains, and therefore, it is the most challenging gene to resolve in newly isolated viruses. It was assumed that our 90bp short-reads from day 5 post-infection would contain an "unknown" influenza A virus from 2009, and only hemagglutinin sequences from 2008 isolates deposited in GenBank were used as reference (Figure 3.3). In a preliminary stage, the sequence from hemagglutinin of A/Brisbane/59/2007 was used as reference and several BLAST analyses with different levels of stringency were performed. We found that word_size was the most relevant parameter when trying to detect influenza sequences with high degrees of dissimilarity; the use of an evalue of 10E-6 was stringent enough to discriminate influenza sequences from those belonging to the host species (Figure 3.4). Using the optimal BLAST parameters, word_size 7 and evalue 10E-6, 1,419 reads were pre-selected and a consensus sequence was generated with 71.7% length coverage (Figure 3.3). This sequence was queried against a BLAST database containing all the influenza sequences published in 2008; the closest match was A/swine/Ohio/02026/2008(H1N1)-HA (GenBank accession CY09915), showing 90.3% homology between them. Finally, BLAST analysis was performed using the sequence from A/swine/Ohio/02026/2008(H1N1)-HA as reference; after assembling the HA-matching reads, the resulting consensus sequence showed 1,698/1,701 identities with the A/California/07-HA sequence.

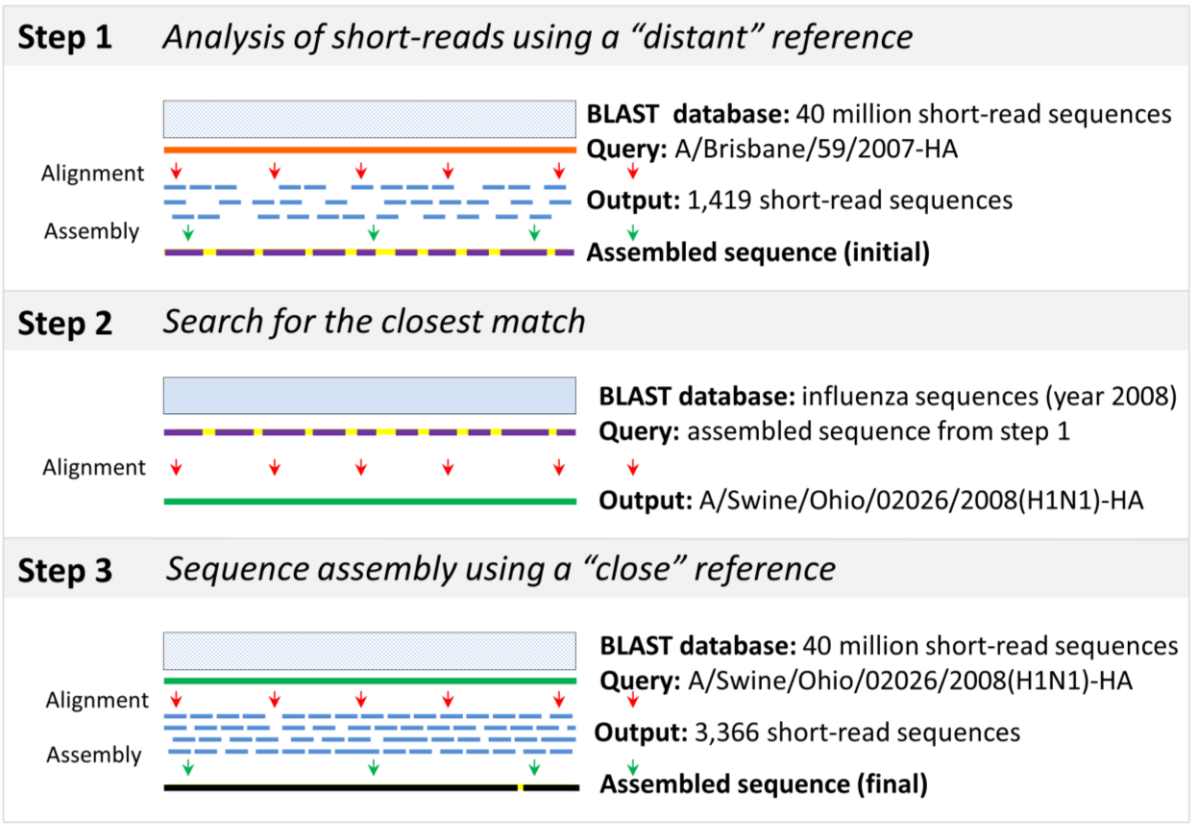


Figure 3.3 Simulation of virus discovery using direct BLAST analysis of NGS data. Step 1: a BLAST database containing the sequencing data from day 5 post-infection was screened by BLAST using a "distant" reference (orange line) from a virus of a lineage that was circulating at that moment (A/Brisbane/59/2007-HA); a number of short-reads were selected (short blue lines) and they were used to generate the consensus sequence (initial) by Iliad Assembler (purple-yellow dashed line). Step2: BLAST analysis of the assembled sequence using as reference influenza isolates from 2008 revealed that the closest match was a strain of swine origin, A/Swine/Ohio/02026/2008 (H1N1)-HA. Step 3) the database was searched using A/Swine/Ohio/02026/2008-HA as the "close" reference (green line) and the consensus sequence (final) was assembled (black line with a yellow dash). BLASTn settings were *word_size* 12, *reward* 2 and *penalty* -3. Yellow dashes indicate uncharacterized areas of the assembled sequences. Red arrows: BLAST alignment. Green arrows: sequence assembly.

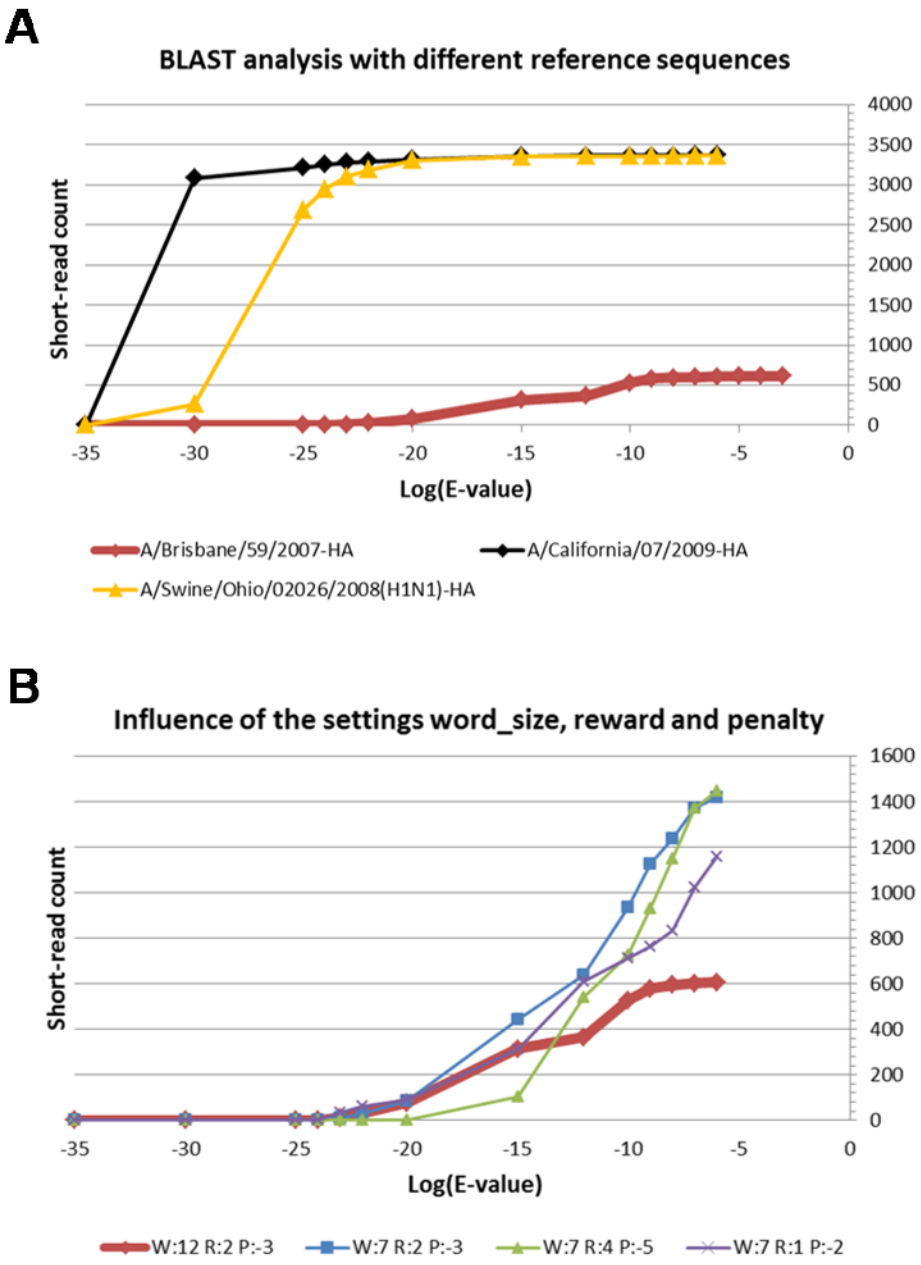


Figure 3.4 Factors influencing the output of BLASTn analysis when detecting sequences from influenza A/California/07/2009 in the short-read library from 5 days post-infection. (A) Counts of aligned short-reads obtained when using the sequences of several influenza strains as reference and different BLAST Expect value (E-value) thresholds. BLASTn settings were *word_size* 12, *reward* 2 and *penalty* -3. (B) Effect of different BLASTn settings in the short-read counts when using A/Brisbane/59/2007-hemagglutinin (HA) as reference.

3.2.6 Detection of virus subpopulations with Varscan

To investigate the capacity of deep sequencing to detect virus subpopulation or quasispecies within a biological sample, the alignment files in SAM format previously generated by Bowtie2 were analyzed using VarScan software as described in Chapter 2. The quality thresholds to discriminate allele variants from sequencing errors were empirically adjusted by using ferret mRNA beta actin as reference (data not shown). VarScan analysis was run with a minimum base quality of 50, and only those allele variants with more than two supporting reads in both plus and minus strands were considered. Our sequencing strategy was based on direct RNA sequencing without prior PCR amplification. Consequently, the vast majority of the viral genome did not have sufficient coverage to allow the detection of variants with low frequency (Figure 3.5). Three coding variants with 100% frequency were shared between the samples from days 3 and 5 post-infection (Table 3.3), which indicates that these changes were introduced before the inoculation of ferrets, possibly during the viral expansion in eggs. Additionally, three more variants with low frequency were found in the RNA sample from day 5 post-infection, one of which was a coding mutation in the hemagglutinin gene, suggesting the presence of viral quasispecies.

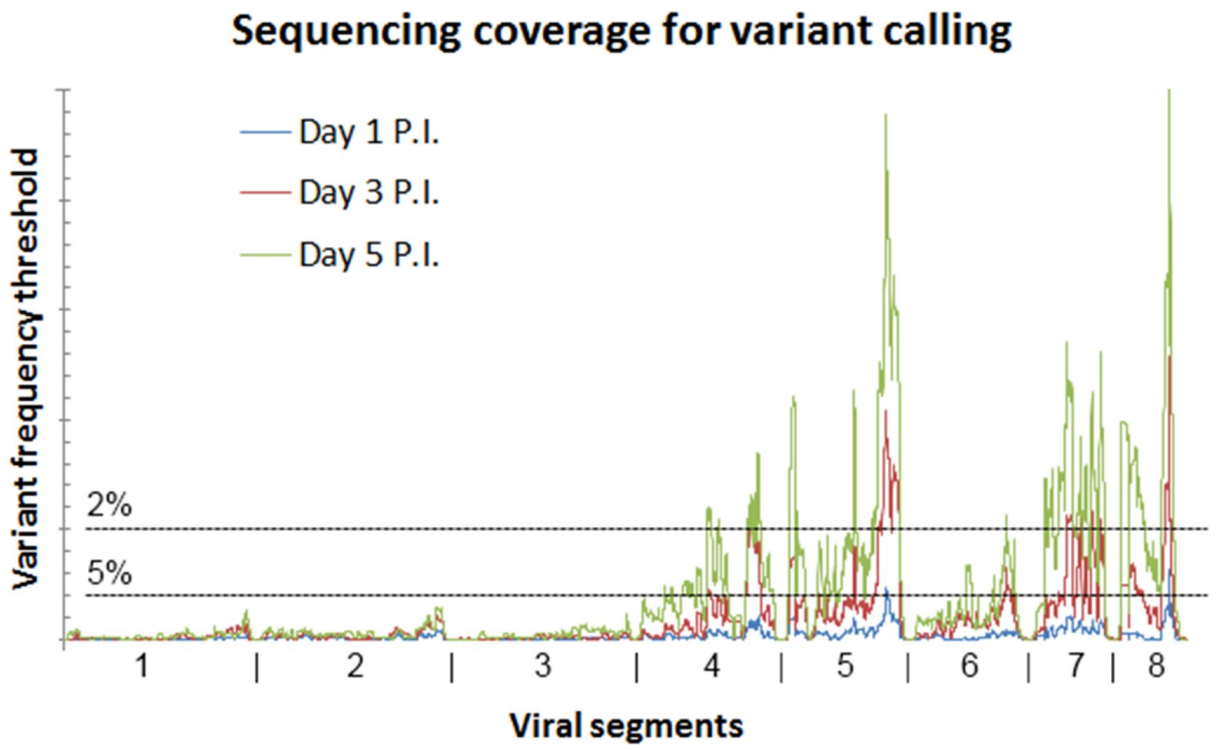


Figure 3.5 Overview of regions of the influenza genome in which nucleotide variants can be called using the NGS data from our study at different times post-infection (PI). A sufficient number of reads from both forward and reverse complementary strands need to support the presence of a nucleotide variant; for each position, the read count from the strand with the lowest coverage was plotted. Frequency thresholds were set with a minimum requirement of more than 2 supporting reads in both the plus and minus strands.

Table 3.3 Variants detected in the influenza sequences by VarScan analysis at different times post-infection (PI)

Viral segment	Nucleotide change	Amino acid change	Variant Frequency	Supporting Reads Reference (Plus/Minus)	Supporting Reads Variant (Plus/Minus)	Average Quality of Reads (Reference/ Variant)
Ferret day 1 PI						
NP(5)	159 T/G	D53E	100%	0 / 0	15 / 20	- / 68
NP(5)	365 T/A	L122Q	100%	0 / 0	4 / 7	- / 66
Ferret day 3 PI						
HA(4)	598 T/C	S200P	100%	0 / 0	26 / 13	- / 67
NP(5)	159 T/G	D53E	100%	0 / 0	76 / 77	- / 67
NP(5)	365 T/A	L122Q	100%	0 / 0	27 / 37	- / 65
Ferret day 5 PI						
HA(4)	598 T/C	S200P	100%	0 / 0	69 / 41	- / 68
HA(4)	1546 G/A	E516K	18.72%	59 / 93	16 / 19	66 / 67
NP(5)	159 T/G	D53E	100%	0 / 0	220 / 319	- / 67
NP(5)	365 T/A	L122Q	100%	0 / 0	92 / 61	- / 64
NS(8)	291 A/G	-	3.78%	134 / 196	6 / 7	65 / 67
NS(8)	588 A/G	-	2.97%	1260 / 212	37 / 8	66 / 67

3.2.7 Overview of the Roche 454 GS FLX sequencing data output

The sequencing run produced 265,484 reads of an average length of 386bp, which is in the range of a successful analysis according to the manufacturer's standards. BLAST analysis revealed that the number of sequences matching each viral segment was PB2: 0, PB1: 1, PA: 0, HA: 12, NP: 6, NA: 6, MP: 9 and NS: 16.

3.3 Discussion

After having performed different studies focused on the host immune responses during respiratory viral infections involving microarray analysis (Cameron et al., 2008; Danesh et al., 2011; Rowe et al., 2010), our group decided to use RNA-seq to better characterize transcriptional variations during experimental influenza infections in ferrets. As part of this work, we also evaluated the presence of influenza virus in our samples. Although the detection of sequences matching the influenza genome can be regarded as a technically simple task, we found that a robust data analysis requires careful consideration of different aspects. Hence this paper is intended to explore the complexities of sequencing data analysis in the context of viral detection and discovery.

Some NGS studies reported the use of prior PCR amplification of the viral segments (Hoper et al., 2011; Nakamura et al., 2009) or enrichment of viral sequences by using probe-capture methods (Ramakrishnan et al., 2009). Our results (Table 3.1) and also previously published studies (Ghedin et al., 2011; Kuroda et al., 2010) show that direct RNA sequencing can provide very high coverage of the viral genome; however, there can be clinical samples where the number of virus-matching reads is low (Yongfeng et al., 2011) and pre-amplification is still an option that may need consideration. The selection of the sequencing platform will determine the number of reads that can be obtained. Roche 454 FLX GS was the first NGS platform commercially available; however, its technical capacities were surpassed shortly after by the Illumina sequencers (Cheval et al., 2011). We sequenced one RNA sample using both platforms, and although the experimental design was not intended to make direct comparisons between technologies, we were able to conclude that direct RNA sequencing in the 454 platform can resolve only the most highly expressed viral genes (as shown in results), a scenario that highly resembles the results from a previously published paper (Ghedin et al., 2011); therefore, viral analysis by 454 sequencing requires prior enrichment of the viral segments by PCR amplification or sequence-specific capture. Meanwhile, Illumina sequencing obtained much higher coverage and succeeded

in delivering information from all the viral segments (Table 3.1). Unless otherwise indicated, the results and the discussion refer to the data obtained by Illumina GAIIx sequencing".

Bowtie2 is a fast aligner widely used in different NGS applications, such as re-sequencing of mammal genomes and study of gene splicing variants (Langmead and Salzberg, 2012). Unlike other aligners such as BWA (Li and Durbin, 2010) and SOAP2 (Li et al., 2009b) that only search for ungapped alignments, Bowtie2 is capable gapped-read alignment, which results in an increase in the number of correct alignments. To perform successful alignments, these tools require a high degree of similarity between the reference and the experimentally obtained short-reads. Here, Bowtie2 was able to align the short-reads to the sequences from A/California/07/2009 used as reference (Table 3.1) and the subsequent consensus sequence was obtained through SAM Tools. The quality thresholds must be set in accordance with the degree of confidence demanded by each application. For example, when obtaining the consensus sequence from the PB2 segment in day 1 post-infection, we found that when using different minimum depths of 3, 5 or 8, the resulting percentage coverage was 23.9, 17.6 and 7.9, respectively. It should be noted that the way in which quality thresholds are implemented varies among methods of analysis; therefore, the percentage coverage alone is not suitable to establish direct performance comparisons. As expected, our simulation shows that the length of the reads and the number of reads mapped to a certain gene have a dramatic impact in the coverage (Table 3.4).

Table 3.4 The resulting % coverage varies with the sequencing platform, length and depth

NGS platform[a]	Length of reads (bp)	Reads matching California/07-HA	% coverage
Illumina Genome Analyzer IIx[b]	90	1,000	95.8
		500	95.1
		250	92.7
		125	78.7
	60	1,000	94.6
		500	93.7
		250	83.6
		125	65.1
	30	1,000	94.6
		500	77.0
		250	60.1
		125	36.2
Roche 454 GS FLX [c]	287 (average)	13	67.3

a Total RNA from the lung tissue of one ferret infected with A/California/07/2009, 5 days post-infection, was analyzed using two different sequencing platforms.

b The sequencing run generated $20x10^6$ paired-end reads, 90bp. In order to study variations in the % length coverage, a number influenza-matching sequences were randomly selected and trimmed to the desired length. Alignments were performed with Bowtie and the % coverage was calculated with Samtools.

c The sequencing run using the Roche GS FLX 454 platform generated 265,454 reads with an average length of 386bp. Out of these, 13 reads matched the sequence from California/07-HA, showing and average length of 287bp. The % coverage was calculated with Iliad Assembler.

Direct BLAST analysis of short-reads is one of the key approaches that should be considered when little or no sequence information is available, or when a significant degree of dissimilarity is expected between a new virus and the previously known strains. Because of the sustained increase in the read length that upcoming NGS platforms can deliver, direct BLAST analysis of short-reads will probably be embraced more widely. Influenza genes have a low degree of homology with respect to the genes of mammal hosts, making their identification easy within libraries where the majority of the sequences correspond to the host species. On the other hand, the great flexibility that BLAST offers through careful selection of the alignment parameters makes it a tool of great value in a variety of studies. After pre-selecting the reads that match the reference genome, they must be assembled to generate the consensus sequence; we used Iliad Assembler to perform this task, a flexible tool written by our group that is well suited for the assembly of complex transcripts (manuscript under preparation). Nonetheless, other tools can be used to perform the assembly of the pre-selected reads such as SSAKE (Short Sequence Assembly by

K-mer search and 3' read Extension) (Warren et al., 2007); alternatively, this task can be performed by de novo assemblers.

De novo assembler programs are designed to find overlaps in the short-reads to generate longer contig sequences; these need to be later identified by using a general-purpose aligner. This approach has been previously used to resolve genomic sequences of influenza virus; for example, Greninger et al. reported that de novo assembled contigs had 90.3% coverage using 60bp short-reads (Greninger et al., 2010), which is in accordance with our results (Table 3.2). Given the short length of the influenza genome and the structural simplicity of their genes, as compared with most mammal genes, the number of reads covering the target sequence is the most important factor for the success of this approach rather than the election of the assembler. Next, we tried to simulate the conditions of the analysis that occur during outbreaks of new influenza strains in which only “distant” viral sequences are available. We found that direct BLAST analysis of the short-reads is a viable option (Figure 3.4); when using the sequence from A/Brisbane/59/2007-HA as reference, we were able to obtain 72% coverage of the “new” virus, and 99% coverage was obtained when using A/Swine/Ohio/02026/2008 as the intermediate reference. On the other hand, when BLAST is used for either direct short-read analysis or identification of de novo assembled contigs, the selection of the database of reference sequences is of critical importance; the analysis needs to biologically rich while keeping the computing requirements at reasonable levels.

The sequencing data allowed us to obtain the consensus sequences of all the viral genes, and they were almost identical to the previously published sequences of A/California/07/2009. For example, the consensus sequence of hemagglutinin showed only one mismatch with respect to the reference sequence, and the fact that this variation was present in all three virus-containing samples suggests that the introduction of this mutation possibly occurred during viral expansion in embryonated eggs prior to ferret infection. Also, we found that the number of reads matching influenza sequences increased gradually from day 1 to day 5 after infection, and none was found on day 14 (Table 3.1). This trend correlates well with the viral titers that were previously observed in the lung tissue from which those samples were retrieved (Rowe et al., 2010). However, given the lack of biological replicates, we were unable to obtain any statistically significant conclusions regarding differences in the quantity of virus among the different experimental groups. The methods for estimating differential expression levels using NGS data are still an active area of research. RPKM-based methods (Mortazavi et al., 2008) are widely used to determine differential gene expression; however, they rely on information from both the transcripts and the genomic

DNA to make certain statistical assumptions and therefore they may not be well suited to study levels of virus expression. Other methods such as edgeR (Robinson et al., 2010) and DEGSeq (Wang et al., 2010) rely on only the number of short-reads per transcript; therefore, they can be used to study the relative expression of viral genes.

SNP calling is a valuable tool that can help to track the changes that viral segments undergo during different adaptation processes (McHardy and Adams, 2009). To characterize the variants or quasispecies of influenza virus that are present in the lungs of infected ferrets the sequencing data was analyzed with VarScan (Koboldt et al., 2012) (Table 3.3). Unfortunately, we found that direct sequencing of RNA samples does not provide sufficient coverage of the viral genome to study these sub-populations (Figure 3.5); therefore, the analysis of viral quasispecies requires preliminary amplification of the viral segments by PCR and subsequent deep sequencing. An approach that has been gaining traction in recent years to enrich the content of influenza segments consists on performing whole-genome multi-segment RT-PCR amplification using primers against conserved flanking sequences of the viral segments (Zhou et al., 2009) and subsequent Illumina sequencing (Farooqui et al., 2015).

In conclusion, the combination of NGS technology with an adequate strategy of data analysis (Table 3.5) constitutes a major leap forward in surveillance and diagnostics of influenza virus. The increased capacity in the acquisition of sequencing data means that nearly full characterization of the viral genomes can now be performed routinely. Also, increased coverage allows influenza to be approached as populations rather than just isolates, which will boost the characterization of determinants of pathogenicity and drug resistance.

Table 3.5 Overview of the analysis techniques used in this study and their performance

Type of analysis	Performance
Illumina GAIIx sequecing	
Sequencing of RNA without prior amplification	Good for viral detection Good for characterization of viral segments
Fast-aligner Bowtie2 + Samtools	Near complete characterization of viral segments in samples with high viral loads (ferret day 5 post-infection).
General purpose aligner BLAST + Iliad Assembler	
De novo assembler Abyss	
SNP calling with VarScan	Insufficient coverage, prior enrichment of viral sequences is required
Roche 454 GS FLX sequencing	
Sequencing of RNA without prior amplification	Sufficient coverage for viral detection Insufficient coverage for characterization of viral segments

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M., Gascoyne, R.D., Marra, M.A., Jones, S.J., 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25, 2872-2877.

Bouvier, N.M., Palese, P., 2008. The biology of influenza viruses. Vaccine 26 Suppl 4, D49-53.

Cameron, C.M., Cameron, M.J., Bermejo-Martin, J.F., Ran, L., Xu, L., Turner, P.V., Ran, R., Danesh, A., Fang, Y., Chan, P.K., Mytle, N., Sullivan, T.J., Collins, T.L., Johnson, M.G., Medina, J.C., Rowe, T., Kelvin, D.J., 2008. Gene expression analysis of host innate immune responses during Lethal H5N1 infection in ferrets. Journal of virology 82, 11308-11317.

Carrat, F., Flahault, A., 2007. Influenza vaccine: the challenge of antigenic drift. Vaccine 25, 6852-6862.

Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F., Berthet, N., Brisse, S., Moszer, I., Bourhy, H., Manuguerra, C.J., Lecuit, M., Burguiere, A., Caro, V., Eloit, M., 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. Journal of clinical microbiology 49, 3268-3275.

Danesh, A., Cameron, C.M., Leon, A.J., Ran, L., Xu, L., Fang, Y., Kelvin, A.A., Rowe, T., Chen, H., Guan, Y., Jonsson, C.B., Cameron, M.J., Kelvin, D.J., 2011. Early gene expression events in ferrets in response to SARS coronavirus infection versus direct interferon-alpha2b stimulation. Virology 409, 102-112.

Dawood, F.S., Iuliano, A.D., Reed, C., Meltzer, M.I., Shay, D.K., Cheng, P.Y., Bandaranayake, D., Breiman, R.F., Brooks, W.A., Buchy, P., Feikin, D.R., Fowler, K.B., Gordon, A., Hien, N.T., Horby, P., Huang, Q.S., Katz, M.A., Krishnan, A., Lal, R., Montgomery, J.M., Molbak, K., Pebody, R., Presanis, A.M., Razuri, H., Steens, A., Tinoco, Y.O., Wallinga, J., Yu, H., Vong, S., Bresee, J., Widdowson, M.A., 2012. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. The Lancet. Infectious diseases 12, 687-695.

Farooqui, A., Lei, Y., Wang, P., Huang, J., Lin, J., Li, G., Leon, A.J., Zhao, Z., Kelvin, D.J., 2011. Genetic and clinical assessment of 2009 pandemic influenza in southern China. Journal of infection in developing countries 5, 700-710.

Farooqui, A., Leon, A.J., Huang, L., Wu, S., Cai, Y., Lin, P., Chen, W., Fang, X., Zeng, T., Liu, Y., Zhang, L., Su, T., Chen, W., Ghedin, E., Zhu, H., Guan, Y., Kelvin, D.J., 2015. Genetic diversity of the 2013-14 human isolates of influenza H7N9 in China. BMC infectious diseases 15, 109.

Garten, R.J., Davis, C.T., Russell, C.A., Shu, B., Lindstrom, S., Balish, A., Sessions, W.M., Xu, X., Skepner, E., Deyde, V., Okomo-Adhiambo, M., Gubareva, L., Barnes, J., Smith, C.B., Emery, S.L., Hillman, M.J., Rivailler, P., Smagala, J., de Graaf, M., Burke, D.F., Fouchier, R.A., Pappas, C., Alpuche-Aranda, C.M., Lopez-Gatell, H., Olivera, H., Lopez, I., Myers, C.A., Faix, D., Blair, P.J., Yu, C., Keene, K.M., Dotson, P.D., Jr., Boxrud, D., Sambol, A.R., Abid, S.H., St George, K., Bannerman, T., Moore, A.L., Stringer, D.J., Blevins, P., Demmler-Harrison, G.J., Ginsberg, M., Kriner, P., Waterman, S., Smole, S., Guevara, H.F., Belongia, E.A., Clark, P.A., Beatrice, S.T., Donis, R., Katz, J., Finelli, L., Bridges, C.B., Shaw, M., Jernigan, D.B., Uyeki, T.M., Smith, D.J., Klimov, A.I., Cox, N.J., 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325, 197-201.

Ghedin, E., Laplante, J., DePasse, J., Wentworth, D.E., Santos, R.P., Lepow, M.L., Porter, J., Stellrecht, K., Lin, X., Operario, D., Griesemer, S., Fitch, A., Halpin, R.A., Stockwell, T.B., Spiro, D.J., Holmes, E.C., St George, K., 2011. Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. The Journal of infectious diseases 203, 168-174.

Greninger, A.L., Chen, E.C., Sittler, T., Scheinerman, A., Roubinian, N., Yu, G., Kim, E., Pillai, D.R., Guyard, C., Mazzulli, T., Isa, P., Arias, C.F., Hackett, J., Schochetman, G., Miller, S., Tang, P., Chiu, C.Y., 2010. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PloS one 5, e13381.

Hoper, D., Hoffmann, B., Beer, M., 2011. A comprehensive deep sequencing strategy for full-length genomes of influenza A. PloS one 6, e19075.

Iwasaki, A., Pillai, P.S., 2014. Innate immunity to influenza virus infection. Nature reviews. Immunology 14, 315-328.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 22, 568-576.

Kuroda, M., Katano, H., Nakajima, N., Tobiume, M., Ainai, A., Sekizuka, T., Hasegawa, H., Tashiro, M., Sasaki, Y., Arakawa, Y., Hata, S., Watanabe, M., Sata, T., 2010. Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. PloS one 5, e10256.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10, R25.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.

Li, Y., Hu, Y., Bolund, L., Wang, J., 2010. State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics 4, 271-277.

McHardy, A.C., Adams, B., 2009. The role of genomics in tracking the evolution of influenza A virus. PLoS pathogens 5, e1000566.

Medina, R.A., Garcia-Sastre, A., 2011. Influenza A viruses: new research developments. Nature reviews. Microbiology 9, 590-603.

Metzker, M.L., 2010. Sequencing technologies - the next generation. Nature reviews. Genetics 11, 31-46.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628.

Nakamura, S., Yang, C.S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., Iida, T., Nakaya, T., 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PloS one 4, e4219.

Nowrousian, M., 2010. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryotic cell 9, 1300-1310.

Ramakrishnan, M.A., Tu, Z.J., Singh, S., Chockalingam, A.K., Gramer, M.R., Wang, P., Goyal, S.M., Yang, M., Halvorson, D.A., Sreevatsan, S., 2009. The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies. PloS one 4, e7105.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Rowe, T., Leon, A.J., Crevar, C.J., Carter, D.M., Xu, L., Ran, L., Fang, Y., Cameron, C.M., Cameron, M.J., Banner, D., Ng, D.C., Ran, R., Weirback, H.K., Wiley, C.A., Kelvin, D.J., Ross, T.M., 2010. Modeling host responses in ferrets during A/California/07/2009 influenza infection. Virology 401, 257-265.

Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D., Fouchier, R.A., 2004. Mapping the antigenic and genetic evolution of influenza virus. Science 305, 371-376.

Stephenson, I., Democratis, J., Lackenby, A., McNally, T., Smith, J., Pareek, M., Ellis, J., Bermingham, A., Nicholson, K., Zambon, M., 2009. Neuraminidase inhibitor resistance after oseltamivir treatment of acute influenza A and B in children. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 48, 389-396.

Suk, J.E., Semenza, J.C., 2011. Future infectious disease threats to Europe. American journal of public health 101, 2068-2079.

Taubenberger, J.K., Kash, J.C., 2010. Influenza virus evolution, host adaptation, and pandemic formation. Cell host & microbe 7, 440-451.

Taubenberger, J.K., Morens, D.M., 2006. 1918 Influenza: the mother of all pandemics. Emerging infectious diseases 12, 15-22.

Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., DeRisi, J.L., 2002. Microarray-based detection and genotyping of viral pathogens. Proceedings of the National Academy of Sciences of the United States of America 99, 15687-15692.

Wang, L., Feng, Z., Wang, X., Zhang, X., 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-138.

Warren, R.L., Sutton, G.G., Jones, S.J., Holt, R.A., 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23, 500-501.

WHO, 2014. Influenza (Seasonal). Fact sheet N°211 http://www.who.int/mediacentre/factsheets/fs211/en/.

Yongfeng, H., Fan, Y., Jie, D., Jian, Y., Ting, Z., Lilian, S., Jin, Q., 2011. Direct pathogen detection from swab samples using a new high-throughput sequencing technology. Clin Microbiol Infect 17, 241-244.

Zerbino, D.R., 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11, Unit 11 15.

Chapter 4

Profiling miRNA in end-stage liver diseases using next generation sequencing and bioinformatic approaches

4.1 Introduction

4.1.1 MiRNA biogenesis pathway

MiRNAs are a class of conserved endogenous small non-coding RNA molecules, about 22 nucleotides (nt) long, that regulate gene expression by binding to target mRNAs (Bartel, 2004). After transcribed by the RNA polymerase II (Lee et al., 2004) or RNA polymerase III (Borchert et al., 2006) in less cases, primary miRNAs (pri-miRNA) are processed by the endoribonuclease Drosha (Lee et al., 2003) and its partner DGCR8 (Denli et al., 2004; Gregory et al., 2004), to form ~70nt long hairpin-structured precursor miRNAs (pre-miRNA). Exportin 5 then transports pre-miRNAs from the nucleus to the cytoplasm (Yi et al., 2003), where these latter are cleaved by the Dicer nuclease, generating a miRNA duplex (Bernstein et al., 2001). Duplexes are subsequently loaded onto an Argonaute protein (AGO) to form an effector complex called RNA-induced silencing complex (RISC) (Chendrimada et al., 2005; Gregory et al., 2005). One of the miRNAs strand, the passenger strand, is removed from the duplex to generate a mature RISC, which is guided by nts 2-7 from the 5' end of the mature miRNA (seed region) to target mRNAs through base pairing (Khvorova et al., 2003; Schwarz et al., 2003). Usually mRNA binding sites are located in the 3' untranslated region (UTR), but recent reports showed that the binding sites can also be found in the coding sequence (CDS) or 5' UTR (Forman et al., 2008; Lytle et al., 2007). Cellular factors are then recruited by RISC to induce translational repression, mRNA deandenylation and mRNA decay, reviewed in (Ameres and Zamore, 2013; Ha and Kim, 2014; Krol et al., 2010).

Conserved clusters of miRNAs genes have been found in human and other mammalian chromosomes (Altuvia et al., 2005; Bartel, 2004). In humans, up to 42% of miRNA pri-miRNAs were found organized in clusters, when a cutoff of 3Kb for inter-miRNA distance was used (Altuvia et al., 2005). Studies have shown that miRNAs from the same cluster are transcribed as polycistrons and their expression patterns are highly correlated (Baskerville and Bartel, 2005; Cullen, 2004). The miR-17-92 cluster, composed by miR-17, miR-18a, miR-19a, miR-19b and miR-92, is the first discovered and best studied human miRNA cluster (Mogilyansky and

Rigoutsos, 2013; Ota et al., 2004). miRNAs in this cluster have been reported to be associated with several human diseases (Mendell, 2008; Mogilyansky and Rigoutsos, 2013). Notably, accumulating evidence indicates that clustered miRNAs collectively modulate molecular pathways through co-targeting of the same genes and/or targeting different individual genes in the same pathway (Hausser and Zavolan, 2014).

4.1.2 MiRNA profiling using NGS

Global miRNA profiling refers to the parallel determination of expression of all miRNAs in a specific tissue. MiRNA profiles comparison between different statuses can provide valuable insights into the molecular determinants causing a disease. Currently various platforms are available for miRNA profiling, such as qPCR, microarray and NGS. qPCR is fast, easy and sensitive, but only available in low-density formats (Takada and Asahara, 2012). Besides, primer design for qPCR and probes for microarrays depends on the known miRNA sequences and therefore they cannot be used to identify unknown miRNAs.

NGS provides not only the expression of every known miRNA, but it is also able to identify miRNA isoforms and potential novel miRNAs, owing to its high accuracy and high throughput properties. Furthermore, mRNA sequencing can be conducted in parallel to correlate miRNAs and transcripts abundance and finally to infer possible interactions.

4.1.3 Roles of miRNAs in the liver

MiRNAs have been reported to be involved in virtually all biological processes. In liver, miRNAs are key regulators in diverse physiological and developmental processes, including cell proliferation and differentiation, metabolism, inflammation and fibrosis (Hsu and Ghoshal, 2013; Szabo and Bala, 2013) among others.

MiR-122, the liver-specific miRNA, has been reported to regulate the balance of hepatocyte proliferation and differentiation during liver development (Xu et al., 2010). While miR-21 and miR-221 have been shown to promote hepatocyte proliferation by regulating different genes (Ng et al., 2012; Yuan et al., 2013), miR-34a and miR-127 were identified to have an anti-proliferative role to hepatocytes in animal models (Chen et al., 2011; Pan et al., 2012). Besides hepatocytes, miRNAs were also found regulating cholangiocyte development (Rogler et al., 2009) and hepatic stellate cell activation and proliferation (Roderburg et al., 2011; Sekiya et al., 2011; Venugopal et al., 2010).

Emerging evidence suggests that miRNAs play a crucial role in liver metabolism. miR-122 was the first miRNA reported as key regulator of lipid metabolism (Esau et al., 2006), followed by identification of more miRNAs as modulators of lipid and cholesterol regulatory genes, such as miR-370 (Iliopoulos et al., 2010), miR-33 (Rayner et al., 2010), miR-185, miR-96, miR-223 (Wang et al., 2013), and miR-27a (Vickers et al., 2013), among others. Additionally, glucose metabolism as well as insulin signalling pathways in hepatocytes are also under regulation of miRNAs, as reviewed in (Chen and Verfaillie, 2014).

MiRNAs are also proposed to fine-tune liver inflammation by regulating various signalling molecules (O'Neill et al., 2011). MiR-155 has been reported as promoter of liver inflammation by targeting tumor necrotic factor after alcohol feeding (Bala et al., 2011), and miR-146a is a negative regulator of TLR signaling (Zhao et al., 2011). Other miRNAs that have been implicated in inflammatory responses include miR-132, miR-125b, let-7 and miR-21 (O'Neill et al., 2011).

MiRNAs have been reported as central players in both hepatic anti-fibrotic and profibrotic processes as well, such as miR-29, miR-19, miR-199a, miR-200, miR-150 and miR-194 (Noetel et al., 2012; Szabo and Bala, 2013). It has been reported that the miR-17/92 cluster targets connective tissue growth factor, which is a central regulator of fibrosis (Kodama et al., 2011).

4.1.4 Roles of miRNAs in liver diseases

In general, end-stage liver diseases refer to the terminal stage of any chronic liver disease, in which the liver is no longer functional. Liver transplantation is usually the main treatment for patients with end-stage liver diseases. Causes of cirrhosis, in addition to virus infection, include alcoholic liver disease (ALD), non-alcoholic steatohepatitis (NASH) related cirrhosis, autoimmune hepatitis (AIH), and cryptogenic cirrhosis (CRP) as well as diseases affecting the bile ducts such as primary biliary cirrhosis/cholangitis (PBC) (Beuers et al., 2015), and primary sclerosing cholangitis (PSC). Accumulating evidence suggest an important role of miRNA in those liver diseases (Munoz-Garrido et al., 2012; Wang et al., 2012b), but no study has focused on the miRNA deregulation in end-stage livers.

Over consumption of alcohol is the second leading cause of cirrhosis in North America, after hepatitis C infection. Alcoholic cirrhosis develops from simple steatosis of alcoholic fatty liver (AFL), alcoholic steatohepatitis (ASH), and fibrosis. All of them are collectively called alcoholic liver disease (ALD). MiRNAs have been found dysregulated in chronic alcohol-fed rats with ALD and associated with tumour necrosis factor-alpha (TNF-α), a key inflammatory agent involved in

liver fibrosis. Common deregulated miRNAs in ALD include decreased miR-125b, miR-181, miR-199a, miR-200a and increased miR-155, miR-21, miR-34a, and miR-37, among others (Dippold et al., 2013; McDaniel et al., 2014). However, so far few studies have explored the miRNA profiles of ALD in human samples.

Non-alcoholic fatty liver disease (NAFLD) is the most common liver disease in developed countries and its worldwide prevalence is rising (Satapathy and Sanyal, 2015; Starley et al., 2010). NAFLD is associated mostly with obesity, type 2 diabetes or hyperlipidemia (Farrell and Larter, 2006). NASH is an aggressive form of NAFLD, featured by liver inflammation and accumulation of fat and fibrous tissue. It has been estimated that 10-20% of NASH cases develop into NASH-related cirrhosis, which has become the third leading cause for liver transplantation in developed countries (Zezos and Renner, 2014). Insulin resistance, oxidative stress, ER impairment and apoptosis have been recognized as the main pathogenic factors for development and progression of NAFLD (Wang et al., 2006). miRNAs have been found dysregulated in NAFLD and now it is known that they play important roles in regulating its pathogenesis (Ferreira et al., 2014). Additionally, miRNAs have also been proposed as promising biomarkers for predicting disease progression and HCC development (Gori et al., 2014).

AIH, PBC, and PSC are the three major categories of autoimmune liver diseases (Washington, 2007). While in AIH the hepatocytes are the main targets of the immune system, in PBC and PSC the bile ducts are attacked. Elevated levels of circulating transaminases, autoantibodies and γ-globulin, and histologic evidence of interface hepatitis characterize AIH. Most AIH patients respond well to immune-suppressors while some refractory cases develop into cirrhosis and end-stage liver disease (Liberal et al., 2015). The pathogenesis of AIH is not completely understood and the role of miRNA needs further investigation. PBC and PSC are both autoimmune diseases affecting bile ducts, whereas PBC is characterized by immune destruction of intrahepatic cholangiocytes, PSC causes inflammation and fibrosis in larger bile ducts. A miRNA expression profiling study in PBC using microarray found 35 deregulated miRNAs, among which decreased expression of microRNA-122a and miR-26a and up-regulation of miR-328 and miR-299-5p were validated by PCR (Padgett et al., 2009). In another study using Illumina deep sequencing to assess serum miRNA expression, miR-505 and miR-197-3p were found decreased in PBC patients (Ninomiya et al., 2013). Although little is known on the association of miRNA with PSC, preliminary studies of miRNA expression in bile of cholangiocarcinoma (CCA) patients, a cancer developed in 10% of PSC patients, have already highlighted the potential utility of miRNAs as biomarker for diagnosing CCA (Li et al., 2014; Shigehara et al., 2011).

Cryptogenic cirrhosis (CC), by definition, refers to a form of cirrhosis whose cause is unknown; therefore, it is clinically diagnosed by exclusion of other better-defined pathologies. The prevalence of CC ranges from 5% to 30% in liver cirrhosis patients and CC accounts for about 10% of liver transplants (Caldwell, 2010). In recently years, CC has been associated with NAFLD (Maheshwari and Thuluvath, 2006). Other common conditions involved in CC include silent autoimmune hepatitis, occult ethanol consumption and non-B, non-C hepatitis (Caldwell, 2010). Since CC is a collection of diseases rather than a single disease, few research has focused on the pathogenesis of CC.

4.1.5 Chapter overview

MiRNAs are key players in human health and disease because of their role in many biological processes. End-stage liver diseases are by definition terminal and for this reason the only treatment is liver-transplantation. So far, little is known about the key driving factors for development of cirrhosis and their progression into end-stage liver diseases. Furthermore, no study has focused on the simultaneous examination of end-stage livers affected by different diseases, in search for common or specific pathological mechanisms associated with different diseases.

In this chapter, NGS and bioinformatic approaches were applied to characterize the miRNA expression profiles in 63 end-stage livers from eight different liver diseases. Explanted livers affected by autoimmune hepatitis (AIH, n = 9), cryptogenic cirrhosis (CRP, n = 3), alcoholic cirrhosis (ETH, n = 10), hepatocellular carcinoma (HCC, n = 10; this included three with HCV infection and alcoholic cirrhosis, three with HBV infection, two with cryptogenic cirrhosis, one with only alcoholic cirrhosis, and one for which no additional diagnostic was recorded), hepatitis C infection (HCV, n = 3), non-alcoholic steatohepatitis (NSH, n = 5), primary biliary cirrhosis (PBC, n = 10) and primary sclerosing cholangitis (PSC, n = 11) were included in this study. After characterizing the distinct patterns of miRNA expression between diseases, bulk samples could be stratified into two major subgroups and eventually this allowed the identification of a set of miRNAs that may play vital roles in liver fibrogenesis.

4.2 Results

4.2.1 Baseline characteristics of sequencing

Small RNA libraries that contained less than 50,000 aligned reads to the miRNAs database were discarded. Baseline characteristics of the remaining 63 liver libraries from are listed in Table 4.1.

On average, 0.94 million reads were sequenced per library (range 0.21 – 1.7 million reads). Only sequences that perfectly and uniquely aligned to sequences in the miRBase 21 were extracted for further analyses. The percentage of sequenced reads aligned ranged from 15 to 81%. The average number of unique miRNAs detected per library was 461 (range 299-658) with no significant difference between diseases (Kruskal Wallis test, p=0.72).

Table 4.1 Baseline characteristics of hepatic miRNA libraries from patients with end-stage liver disease

Diagnosis	Number of samples	Range of reads per library	Range of aligned reads per library	Range of aligned reads % per library	Range of known miRNA
AIH	9	277587 ~ 1631027	133344 ~ 732670	35.47% ~ 64.44%	319 ~ 658
CRP	6	925732 ~ 1280246	368023 ~ 873706	39.75% ~ 69.56%	389 ~ 522
ETH	9	428903 ~ 1448476	190407 ~ 807354	28.02% ~ 68.34%	410 ~ 584
HCC	10	914901 ~ 1683780	188784 ~ 1015268	17.52% ~ 73.65%	325 ~ 540
HCV	3	928516 ~ 1183224	374091 ~ 591725	40.29% ~ 50.01%	481 ~ 506
NSH	5	373593 ~ 1545516	161701 ~ 1073164	14.95% ~ 69.44%	350 ~ 606
PBC	10	219114 ~ 1191572	88494 ~ 530467	20.20% ~ 67.64%	299 ~ 516
PSC	11	467584 ~ 1737511	308600 ~ 1410348	36.29% ~ 81.17%	353 ~ 545

4.2.1 Pairwise comparison between diseases

To search for disease-specific miRNA expression features, pairwise comparisons were performed on each disease. Surprisingly, more than half of the comparisons resulted in no significant difference (Table 4.2). It is likely that there exist differences in the microRNA profiles between AIH and ETH, for example that were not captured in the prior analyses. There are several potential explanations including the heterogeneity of expression or functional redundancy of microRNAs.

Nevertheless, a relatively large number of miRNAs were found differentially expressed between specific diseases. For example, expression of 123 miRNAs in AIH samples was significantly different when compared to CRP samples. When HCC was compared to PSC, 85 miRNAs were deregulated. Interestingly, miRNAs from liver samples from either AIH or PSC demonstrated increased variance as compared with other disorders. Additional analyses were performed on the diseases demonstrating the greatest number of differentially expressed miRNAs (Table 4.2: AIH vs. HCC and HCC vs. PSC).

Table 4.2 Numbers of deregulated miRNAs in pairwise comparisons for all end-stage liver diseases (DESeq2)

	AIH	CRP	ETH	HCC	HCV	NSH	PBC	PSC
AIH		54	0	123	0	55	0	2
CRP			0	0	0	0	2	40
ETH				0	9	0	0	1
HCC					0	0	7	85
HCV						0	0	0
NSH							0	38
PBC								0
PSC								

4.2.1.1 Two subgroups were defined in AIH samples by MDS and unsupervised clustering for miRNA expression

Different analyses that compute the complexity or relationship of multiple microRNAs can provide additional evidence towards disease association. For example, multidimensional scaling (MDS) analysis is an ordination technique that can be used to find samples with similar characteristics and place them close to each other usually in a two-dimensional space, while more dissimilar samples will be place far apart from each other (Quinn GP, 2002). In Figure 4.1, two groups were observed when comparing AIH with HCC using MDS analysis. One group was confined to four AIH samples and the larger assembly was composed of five AIH and eight HCC samples, whereas two additional HCC samples were located elsewhere. The relevance of these findings remains to be resolved, however.

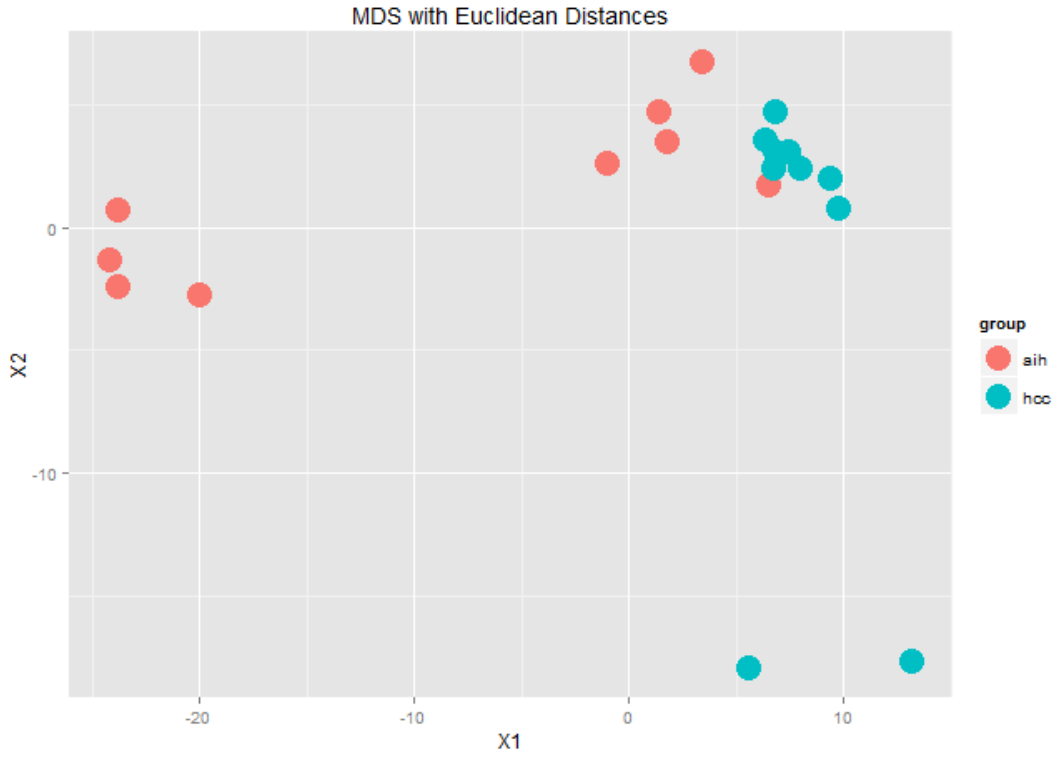


Figure 4.1 Multidimensional scaling analysis of samples diagnosed with AIH (red dots) and HCC (blue dots). MDS separated AIH samples into two subgroups on the first dimension. Samples in one of such subgroups (n=4) clustered relatively tightly and located on one end of the first dimension axis, whereas the rest of AIH samples (n=5) were in close proximity to most of the HCC samples.

One way of defining specific miRNAs associated with a biological process may be derived using unsupervised clustering. For example, the heatmap below was constructed using the top 50 differentially expressed microRNA from AIH and HCC samples (Figure 4.2). Similar to the MDS analysis, the heatmap shows two distinct groups. Samples aih4, aih7, aih8, aih9 formed a discernable cluster, whereas the other AIH samples grouped together with the HCC samples. A notable feature was the greater than four-fold increase of expression in a set of miRNAs [miR-424-5p, miR-101-3p, miR-29c-3p, miR-142-3p, miR-542-3p, miR-19a-3p and miR-19b-3p].

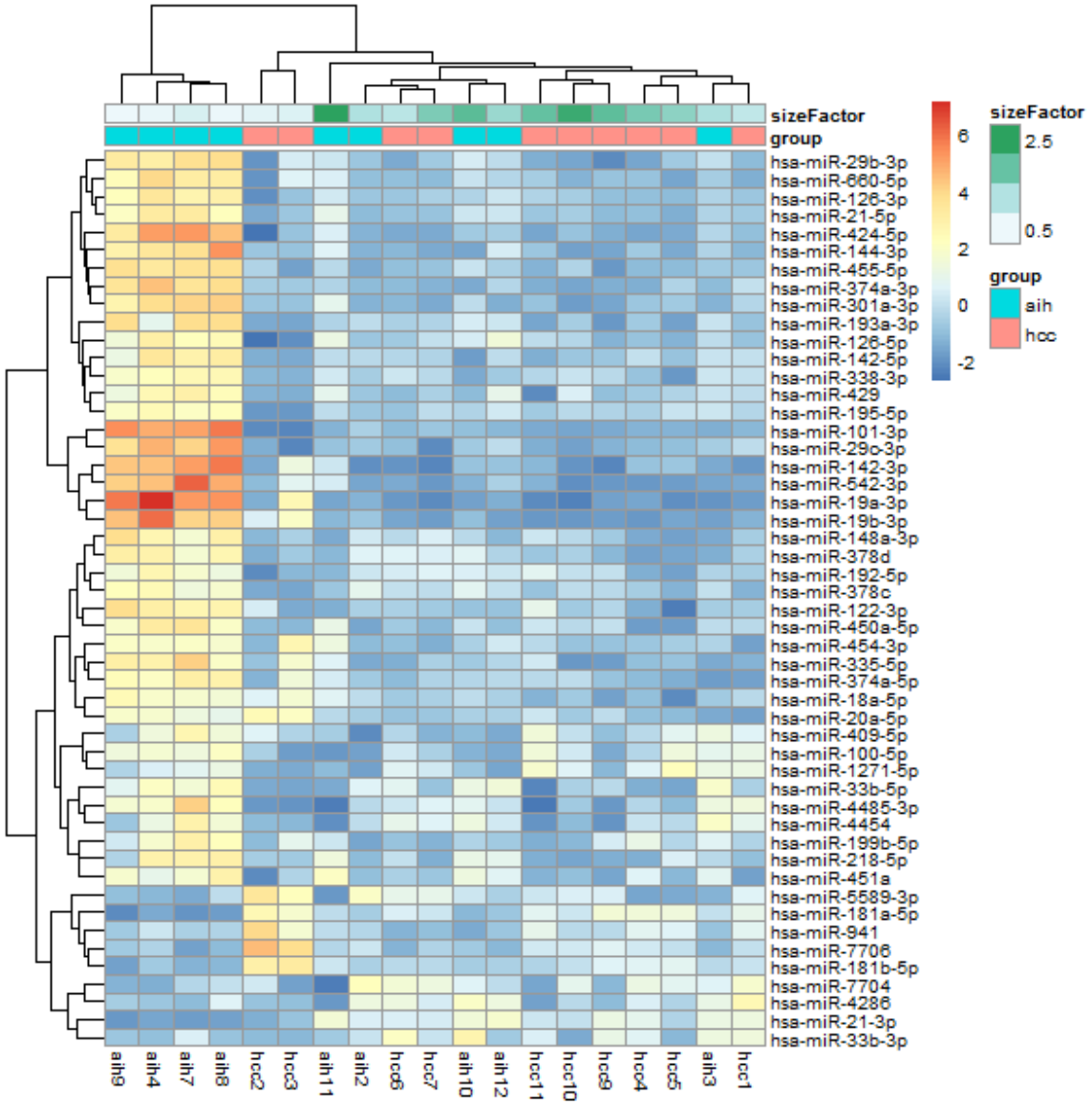


Figure 4.2 Unsupervised hierarchical clustering of samples diagnosed with AIH and HCC by 50 miRNAs expressed with largest Euclidian distance. A discrete cluster on the left is formed by four AIH samples, featured by up-regulation of a relatively large number of miRNAs, with miR-424-5p, miR-101-3p, miR-29c-3p, miR-142-3p, miR-542-3p, miR-19a-3p and miR-19b-3p being the top upregulated ones in such cluster.

4.2.1.2 Two subgroups were defined in PSC samples by MDS and unsupervised clustering for miRNA expression

We had previously observed a large differential expression of miRNAs in HCC and PSC samples. Therefore, we conducted similar analyses to further explore the difference between their miRNAs expression profiles. Interestingly, bimodal subgrouping was also found in the MDS analysis (Figure 4.3), where five PSC samples clustered as an individual group and the other six PSC samples gathered together with most HCC samples.

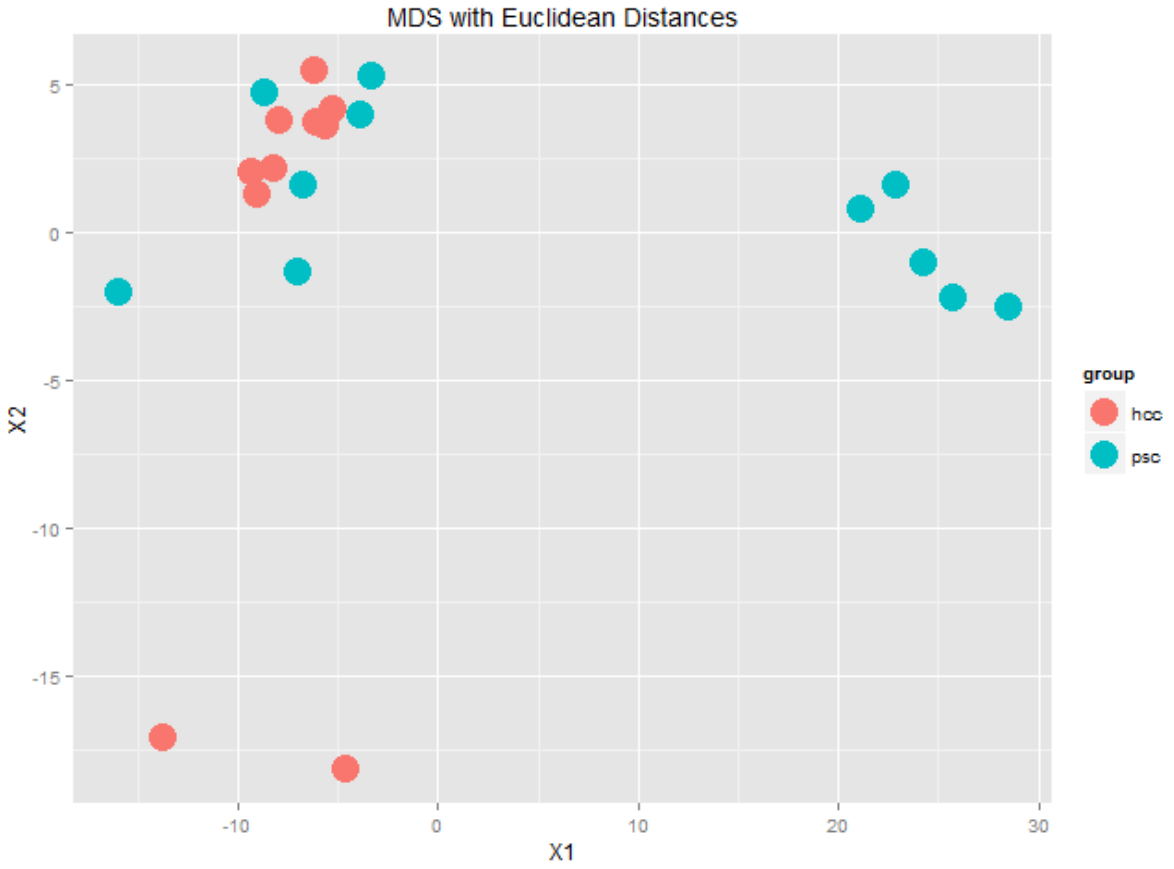


Figure 4.3 Multidimensional scaling analysis of samples diagnosed with HCC (red dots) and PSC (blue dots). PSC samples were separated into two subgroups, five on the right and six on the left. The six-PSC subgroup is mixed with most HCC samples, forming a big group in the upper left corner. Two HCC sample are clustered in the bottom, far from all other samples.

We then performed an unsupervised clustering heatmap using expression of the top 50 differed miRNA among HCC and PSC samples (Figure 4.4). Samples psc7, psc8, psc9, psc10 and psc12

were clustered together, showing a very different miRNA expression pattern from other samples (Figure 4.4). Interestingly, the same set of miRNAs found upregulated in samples aih4, aih7, aih8 and aih9 in Figure 4.2 were also increased in expression in samples psc7, psc8, psc9, psc10 and psc12 in this heatmap.

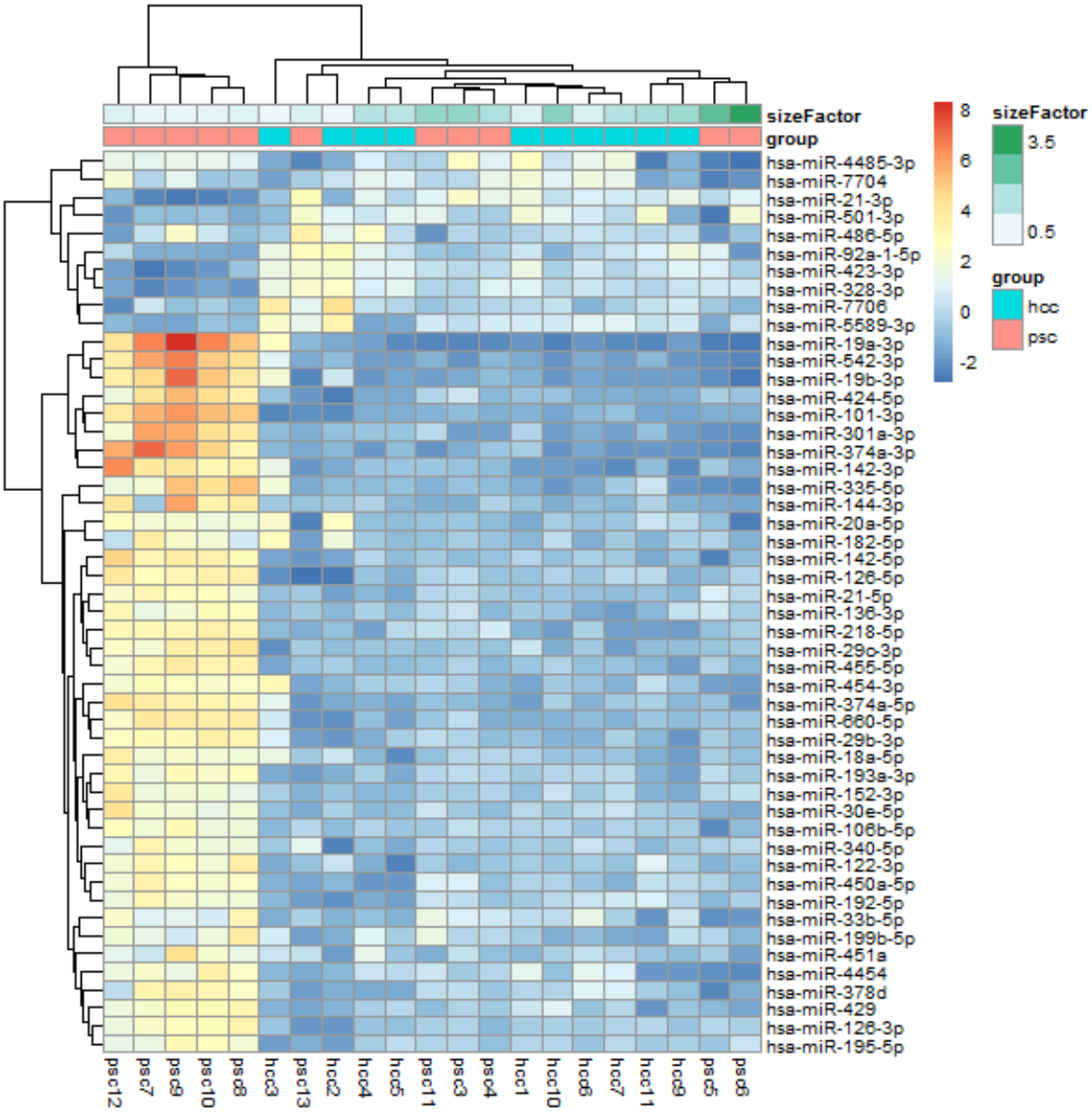


Figure 4.4 Unsupervised hierarchical clustering of samples diagnosed with PSC and HCC by 50 miRNAs expressed with largest Euclidian distance. A unique cluster on the left is formed by five PSC samples, featured by up-regulation of a big panel of miRNAs led by miR-19a-3p, miR-542-3p, miR-19b-3p, miR-424-5p, miR-101-3p, miR-301a-3p and miR-374a-3p.

4.2.1.3 Merging of datasets shows that AIH and PSC samples group together by MDS and unsupervised clustering for miRNA expression

As both AIH and PSC samples shared similar miRNA expression, we addressed the hypothesis that the two groups may similarly merge when analysed together. Therefore, another MDS analysis was conducted using all samples with the diagnosis of AIH, PSC and HCC (Figure 4.5). Interestingly, in this plot the small subgroups of four AIH and five PSC did indeed form a single cluster while the remaining AIH and PSC samples grouped with eight HCC as previously observed.

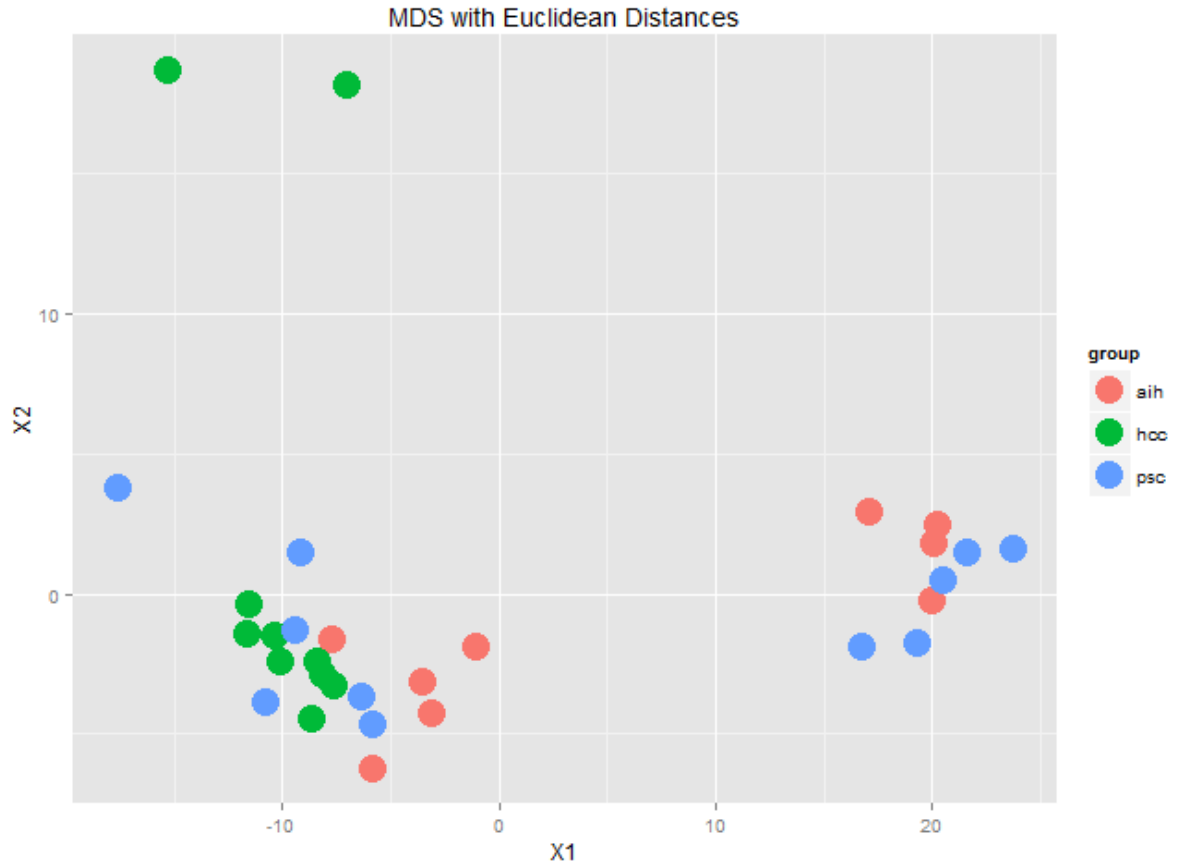


Figure 4.5 Multidimensional scaling analysis of samples diagnosed with AIH (red dots), HCC (green dots) and PSC (blue dots). Both AIH PSC samples were separated into two subgroups, one on the right and the other on the left. The subgroups on the right form a single cluster, while the subgroups on the left constitute a bigger cluster together with eight HCC samples.

The same analyses was then performed using hierarchical clustering of the 50 miRNA expressed with largest variance across all AIH, PSC and HCC samples (Figure 4.6). Once again, we observed aih4, aih7, aih8 and aih9 clustered with psc7, psc8, psc9, psc10 and psc12. A differential miRNA expression pattern can be easily observed with increased expression of a set of more than 10 miRNAs as shown [miR-19a-3p and miR-19b-3p, miR-542-3p, miR-424-5p, miR-660-5p, miR-29b-3p, miR-142-3p, miR-301a-3p, miR-101-3p, miR-374a-3p].

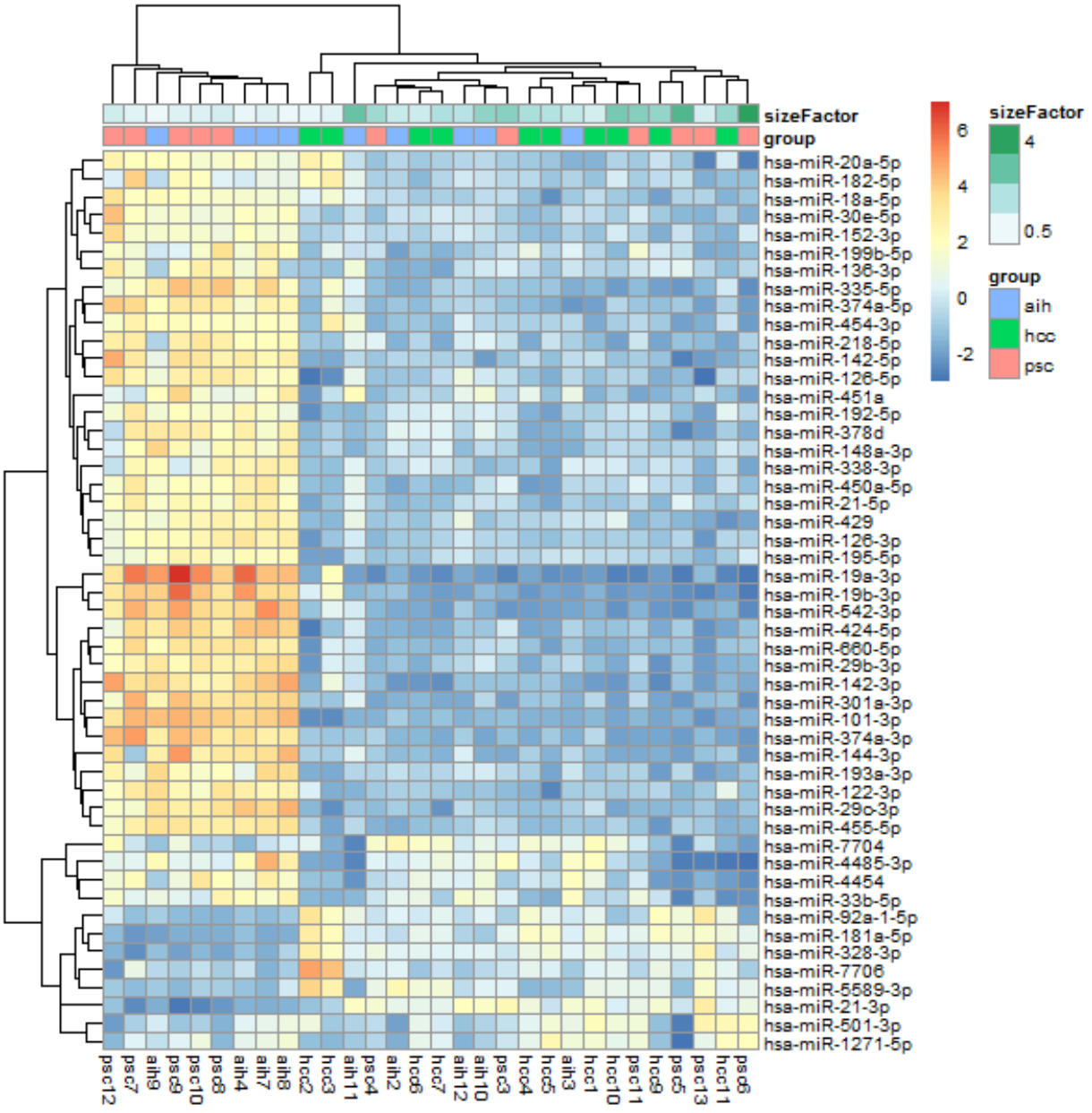


Figure 4.6 Unsupervised hierarchical clustering of AIH, PSC and HCC by 50 top variably expressed miRNAs. Four AIH and five PSC form a distinctive cluster on the left, featured by increased expression of a sizable panel of miRNAs.

This observation then led to a further exploration of miRNA expression profiles in all end-stage liver diseases. We wondered whether a similar dichotomy of miRNA expression can also be found in other end stage liver diseases.

4.2.2 Two distinct groups were defined by ordination and hierarchical clustering

In a two-dimensional MDS plot, all 63 samples from the eight liver diseases were analysed (Figure 4.7). The pattern showed a clear bimodal distribution of samples, where 13 samples clustered together and separated well from the other 50 samples (Figure 4.8A). After closer inspection, it was found that the group of 13 samples was mostly composed by AIH, PSC and, to a lesser extent, PBC, all of which are autoimmune liver diseases. Two alcoholic cirrhosis (ETH) samples were also in this small group. For the sake of simplicity, the smaller group will be referred to as group A and the larger group B. The latter consisted of 50 libraries in all eight diseases, including the five AIH, seven ETH, eight PBC, five PSC and all samples from CRP, NSH, HCV and HCC (Figure 4.7, Table 4.3). Unsupervised hierarchical clustering reflected the same two-group pattern (Figure 4.8).

Table 4.3 Number of samples in Group A and Group B

Disease	Number of samples		
	All	Group A	Group B
AIH	9	4	5
CRP	6	0	6
ETH	9	2	7
HCC	10	0	10
HCV	3	0	3
NSH	5	0	5
PBC	10	2	8
PSC	11	5	6
Total number	63	13	50

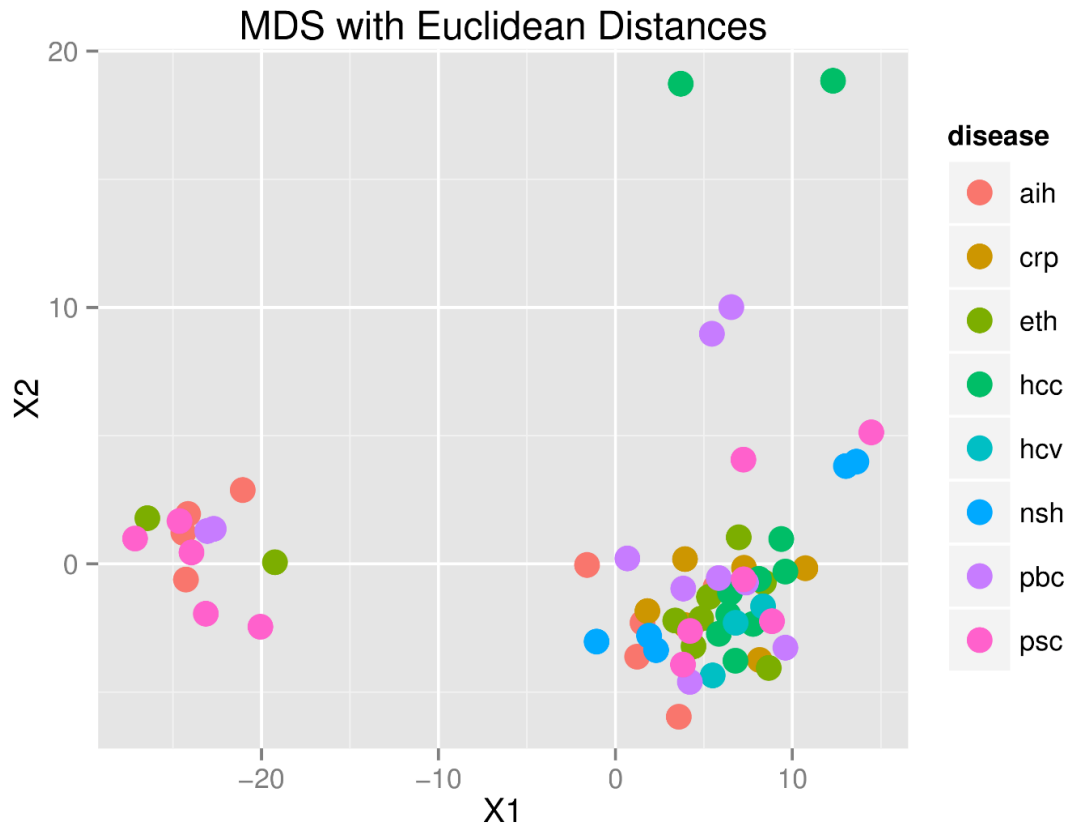


Figure 4.7 MDS plot of miRNA expression profiles in all 63 end-stage livers. Two distinct groups were defined, each of which was composed by different diseases. The smaller group on the left is referred to as Group A (N=13), which contains 13 samples diagnosed with AIH, PSC, PBC and ETH while the bigger group on the right is referred as Group B (N=50) which consists other 50 samples with all eight diseases.

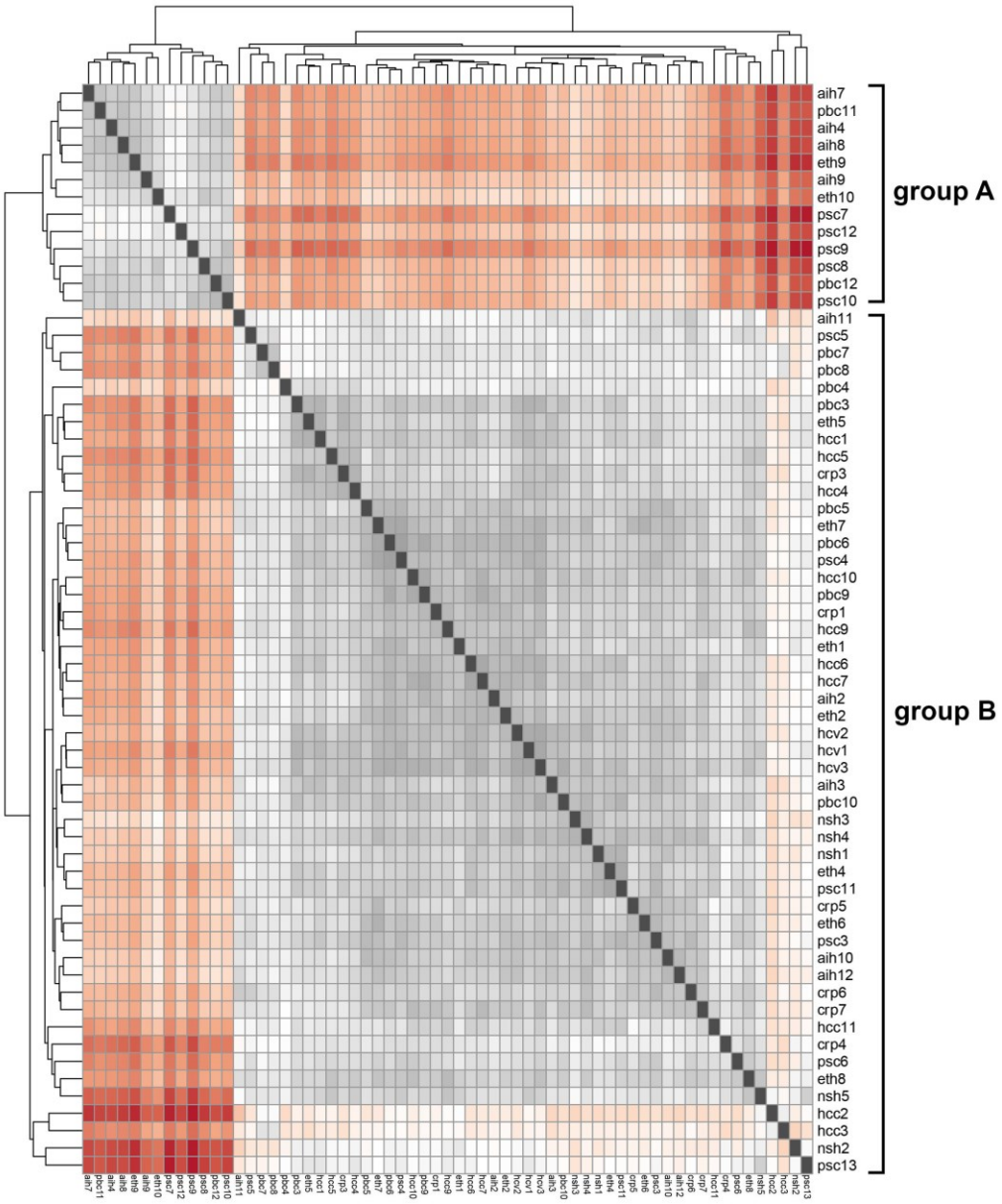


Figure 4.8 Hierarchical clustering by Euclidean distance of all sample in group A and group B also shows a two-cluster pattern of all samples, group components of each are the same as shown in Figure 4.7.

When plotting the 50 miRNAs expressed with largest variance on a heatmap, the grouping of A and B was clearly preserved (Figure 4.9). Several discrete blocks of miRNAs that were either upregulated or downregulated in group A, with regard to group B, are clearly observed in the heatmap. The upper part shows a group of 16 miRNAs expressed in higher levels in group A,

compared to group B, followed by nine miRNAs expressed to lower levels in group A than in group B. In the lower part of the plot, a large number of upregulated miRNAs can be seen in group A, followed by some miRNAs that show variable patterns of expression in the two groups.

Figure 4.9 Unsupervised hierarchical clustering of 63 end-stage livers by the top 50 miRNAs with largest Euclidian distance in expression levels across all samples. Group A samples (labelled in pink) and group B samples (labelled in blue) formed two distinct clusters in the plot.

Once the differential expression analysis between group A and B was completed, a total of 203 miRNAs were found to be differentially expressed between group A and group B. A total of 106 miRNAs were upregulated and 97 miRNAs were downregulated in group A. The top differentially expressed miRNAs according to adjusted P value are shown in Table 4.4.

Table 4.4 Top deregulated miRNAs between group A and group B samples

microRNA	BM[1]	Log2FC[2]	P value	P adj[3]
hsa-miR-455-5p	123.04	-5.65	9.65E-136	2.70E-133
hsa-miR-101-3p	1624.85	-7.42	2.48E-86	3.47E-84
hsa-miR-660-5p	120.59	-5.47	1.90E-83	1.77E-81
hsa-miR-542-3p	29.82	-7.93	4.27E-75	2.99E-73
hsa-miR-374a-3p	20.44	-8.40	6.20E-75	3.47E-73
hsa-miR-374a-5p	26.14	-5.67	9.77E-67	4.56E-65
hsa-miR-122-3p	134.04	-5.00	1.32E-66	5.27E-65
hsa-miR-142-5p	43.13	-6.15	2.36E-66	8.25E-65
hsa-miR-126-3p	1151.92	-4.43	1.36E-65	4.22E-64
hsa-miR-29c-3p	236.15	-6.33	3.22E-63	9.01E-62
hsa-miR-424-5p	648.67	-6.13	6.44E-60	1.64E-58
hsa-miR-142-3p	76.74	-7.05	1.10E-58	2.57E-57
hsa-miR-21-5p	8676.94	-4.39	5.80E-55	1.25E-53
hsa-miR-301a-3p	14.78	-7.06	1.21E-51	2.42E-50
hsa-miR-19b-3p	467.73	-7.02	1.22E-50	2.29E-49
hsa-miR-29b-3p	109.64	-5.20	1.63E-49	2.86E-48
hsa-miR-148a-3p	29163.09	-4.35	1.28E-46	2.10E-45
hsa-miR-195-5p	221.59	-3.54	3.66E-45	5.69E-44
hsa-miR-335-5p	15.03	-6.49	1.86E-40	2.74E-39
hsa-let-7c-5p	3886.90	2.78	8.48E-39	1.19E-37
hsa-miR-502-3p	22.38	-2.77	4.04E-38	5.39E-37
hsa-miR-152-3p	150.21	-3.57	4.86E-37	6.19E-36
hsa-miR-181a-5p	13883.44	3.90	6.67E-37	8.12E-36
hsa-miR-193a-3p	30.84	-4.63	1.10E-36	1.29E-35
hsa-miR-144-3p	14.43	-7.28	4.43E-36	4.96E-35
hsa-miR-30e-5p	78.10	-3.69	2.19E-35	2.36E-34
hsa-let-7a-5p	26930.06	2.26	9.74E-35	1.01E-33
hsa-let-7d-3p	253.62	3.37	2.99E-34	2.99E-33

1. BM: Base mean, the mean expression of this miRNA 2. Log2FC: log2 fold change
3. P adj: adjusted P value, equivalent to false discovery rate

4.2.3 MiRNA expression in group A and B compared with controls

To better understand the differences between both groups of samples with end stage liver disease (A and B), we used patient liver samples described in Chapter 5 (controls). These "control" samples were used for investigating miRNA in paired liver samples and tumor samples derived from the same patient. The controls included patients with chronic hepatitis C virus infection, chronic hepatitis B virus infection, ETH, non-alcoholic fatty liver disease and hemochromatosis. All samples were derived from surgical resection and therefore had less advanced liver disease, as compared to patients who had undergone liver transplantation. To further investigate the differential miRNA expression in groups A and B, we then included this control group data into the analyses (see below). The miRNA expression profiles derived from the 12 non-tumor control samples clustered close to each other on the MDS plot. MDS analysis of the three groups resulted in three well-defined clusters of samples (Figure 4.10).

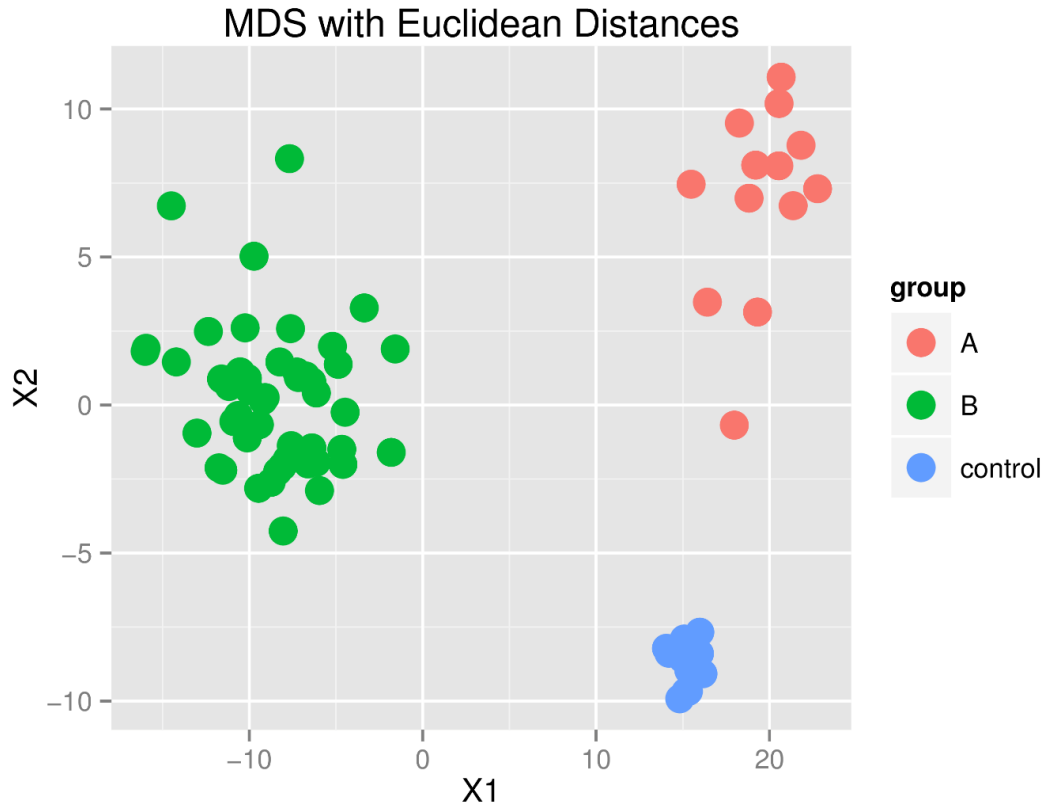


Figure 4.10 MDS analysis for samples in Group A (red dots), Group B (green dots) and controls (blue dots) showed three well-defined groups. The control samples were grouped into a tight cluster while the other two groups were scattered. Overall the control group was positioned closer to group A than to group B.

Figure 4.11 shows the unsupervised clustering of all samples and controls based on the 50 most differentially expressed miRNAs across samples. The separation of the two sample groups and the control group is also evident here. Many of the miRNAs that were found to be upregulated in group A vs. group B were also upregulated in the control group. However, differential expression was also evident between group A and the control group.

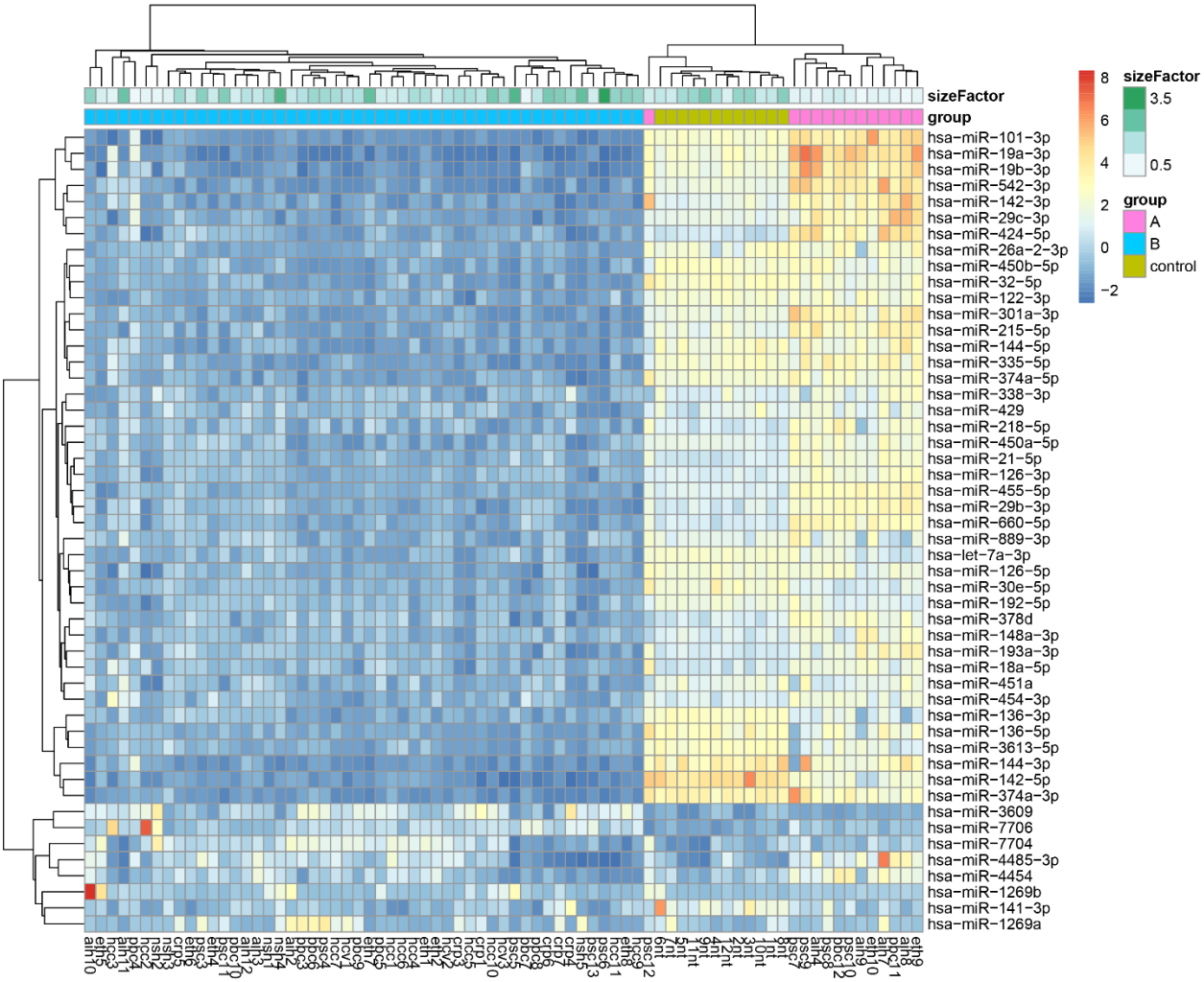


Figure 4.11 Unsupervised hierarchical clustering of miRNA expression of samples in group A, group B and controls by 60 miRNAs with largest variance in expression. In the plot group B samples (labelled in blue), control samples (labelled in green) and group A samples (labelled in pink) formed three different clusters. Control samples expression profile clustered closer to group A than to group B.

When comparing group A vs. controls, 90 miRNAs were found upregulated and 93 miRNAs were found downregulated by differential expression analysis. For group B, 125 miRNAs were significantly decreased and 98 miRNAs were significantly increased in expression vs. the control group. Interestingly, 34 miRNAs that were detected as upregulated in group A with respect to the control group were found downregulated in group B, when compared to the control group (Table 4.5, only the most abundant ones are shown). Conversely, 21 miRNAs that were found downregulated in group A were found upregulated in group B, both regarding the control group (Table 4.6, only most abundant ones are shown).

Table 4.5 Common miRNAs upregulated in Group A and downregulated in Group B compared with control samples

Differential expressed miRNAs	Group B vs. Control			Group A vs. Control		
	BM	Log2FC	Padj	BM	Log2FC	Padj
hsa-miR-101-3p	720.42	5.98	1.67E-56	3540.40	-1.47	3.97E-12
hsa-miR-455-5p	40.61	3.45	1.32E-50	229.33	-2.18	7.68E-29
hsa-miR-29c-3p	72.85	4.22	3.94E-33	446.01	-2.10	1.44E-10
hsa-miR-126-3p	625.06	2.88	3.04E-27	2127.97	-1.50	1.77E-10
hsa-miR-19b-3p	143.91	4.94	2.63E-24	920.86	-2.06	7.30E-15
hsa-miR-148a-3p	16205.25	2.84	3.63E-24	55121.43	-1.48	1.45E-04
hsa-miR-106b-5p	19.02	2.61	1.15E-23	39.17	-0.65	1.31E-02
hsa-miR-660-5p	33.49	2.80	8.27E-21	211.91	-2.65	6.10E-38
hsa-miR-21-5p	3805.97	2.32	7.56E-15	14982.96	-2.06	1.54E-24
hsa-miR-29b-3p	35.87	2.81	1.03E-13	195.10	-2.36	3.37E-23
hsa-miR-193a-3p	12.10	2.42	2.53E-12	53.27	-2.15	8.77E-10
hsa-miR-424-5p	102.02	2.52	4.59E-11	1091.70	-3.53	1.28E-31
hsa-miR-195-5p	121.84	1.61	2.28E-10	343.52	-1.88	1.09E-13
hsa-miR-378d	9.11	2.38	2.73E-10	32.07	-1.73	1.72E-07
hsa-miR-19a-3p	25.63	5.41	1.08E-09	182.50	-2.20	1.34E-16
hsa-miR-18a-5p	25.16	2.14	2.94E-09	78.00	-1.74	4.44E-08
hsa-miR-146b-5p	781.54	1.65	5.34E-09	1465.54	-1.10	1.70E-04
hsa-miR-29a-3p	605.64	1.24	9.09E-07	805.81	-0.72	7.95E-03
hsa-miR-20a-5p	48.86	2.13	2.20E-06	85.74	-0.62	1.11E-02
hsa-miR-151a-3p	568.25	0.71	2.15E-05	685.94	-1.05	4.69E-03
hsa-miR-502-3p	13.87	0.64	1.57E-04	28.83	-2.10	1.05E-10
hsa-miR-152-3p	67.27	0.83	2.80E-03	219.63	-2.69	1.85E-15

Table 4.6 Common miRNAs downregulated in Group A and upregulated in Group B compared with control samples

Differential expressed miRNAs	Group B vs. Control			Group A vs. Control		
	BM	Log2FC	Padj	BM	Log2FC	Padj
hsa-let-7d-3p	319.74	-2.29	5.43E-23	30.12	1.09	1.12E-06
hsa-let-7b-5p	12774.83	-2.18	3.51E-20	1373.25	0.71	6.36E-03
hsa-let-7a-5p	35892.84	-1.13	1.35E-11	6912.07	1.08	4.84E-06
hsa-let-7d-5p	650.97	-0.98	5.64E-11	143.38	0.89	2.00E-04
hsa-miR-320a	805.03	-1.04	5.31E-05	177.90	0.75	6.01E-06
hsa-miR-193b-3p	614.94	-1.22	5.76E-05	109.32	1.21	9.36E-06
hsa-miR-423-5p	558.17	-1.27	6.05E-05	92.86	1.39	2.54E-19
hsa-let-7c-5p	5400.45	-0.72	1.29E-04	1131.51	1.97	5.86E-15
hsa-miR-221-3p	352.67	-0.50	9.64E-04	116.22	0.53	1.76E-02
hsa-miR-125a-5p	6402.73	-0.90	6.26E-03	1235.07	1.82	3.75E-11
hsa-miR-204-5p	318.37	-0.67	8.53E-03	86.27	0.85	3.87E-04
hsa-miR-191-5p	13880.19	-0.63	1.27E-02	2983.57	2.13	1.17E-12
hsa-miR-181a-5p	19302.62	-0.77	1.35E-02	3460.97	3.04	2.35E-53
hsa-miR-378a-5p	123.22	-0.62	3.88E-02	28.53	1.67	1.77E-10
hsa-miR-423-3p	337.16	-0.60	4.56E-02	85.65	1.29	2.72E-03

4.2.4 Pathways deregulated miRNAs may affect

Inspection of the function of individual miRNAs found differentially expressed in group A or B, regarding the control group is problematic, given the large number of miRNAs found deregulated and the opposite roles of some of them. It was, therefore, more plausible to try to understand a putative role of such groups of miRNAs of liver pathology at the pathways level.

Subsequently, 14 miRNAs expressed with largest variance across all groups were input into the pathway prediction program. Those miRNAs include miR-19a-3p, miR-19a-3p, miR-101-3p, miR-542-3p, miR-215, miR-142-3p, miR-29c-3p, miR-424-5p, miR-455-5p, miR-660-5p, miR-29b-3p, miR-148a-3p, miR-301-3p, and miR-378d. The top predicted pathways are focal adhesion and PI3K-AKT signaling pathways (Figure 4.12 and 4.13). The gene numbers those miRNAs target were summarized in Table 4.7.

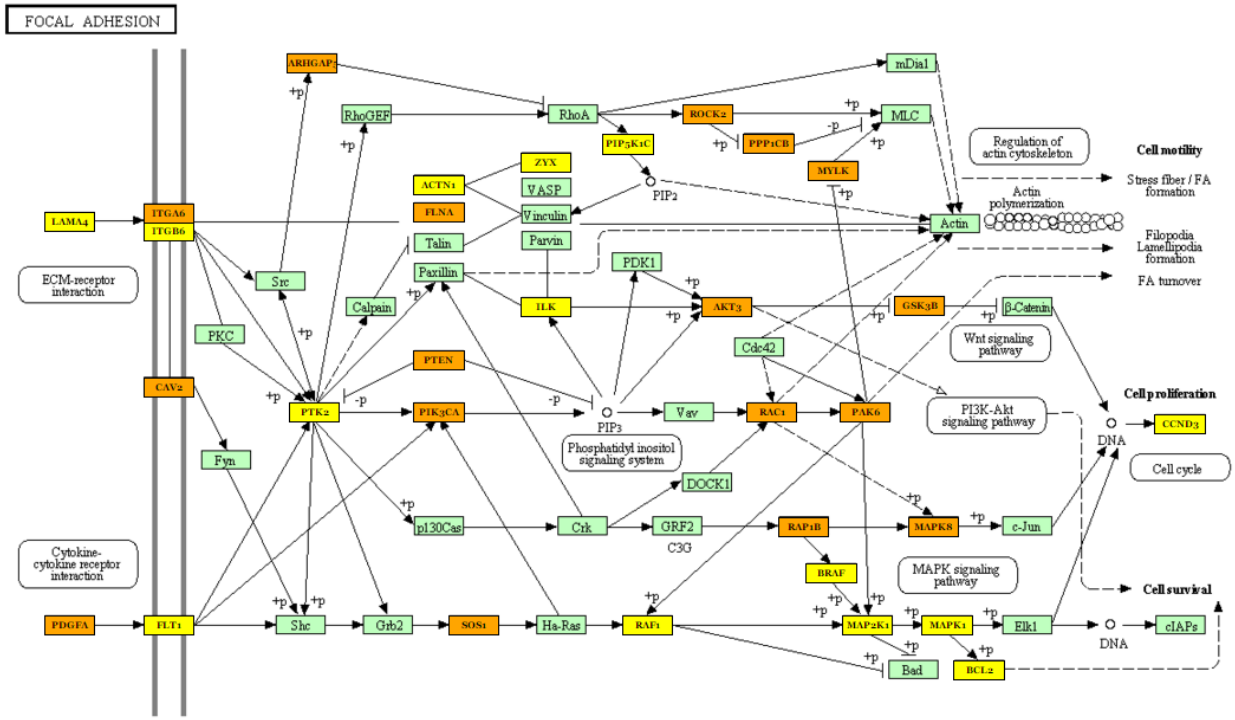


Figure 4.12 Focal adhesion pathway with highlighted targets of the fourteen miRNAs that were upregulated in group A and downregulated group B with largest fold changes. Green box: not a target; yellow box: genes targeted by one miRNA; brown box: genes targeted by more than one miRNAs

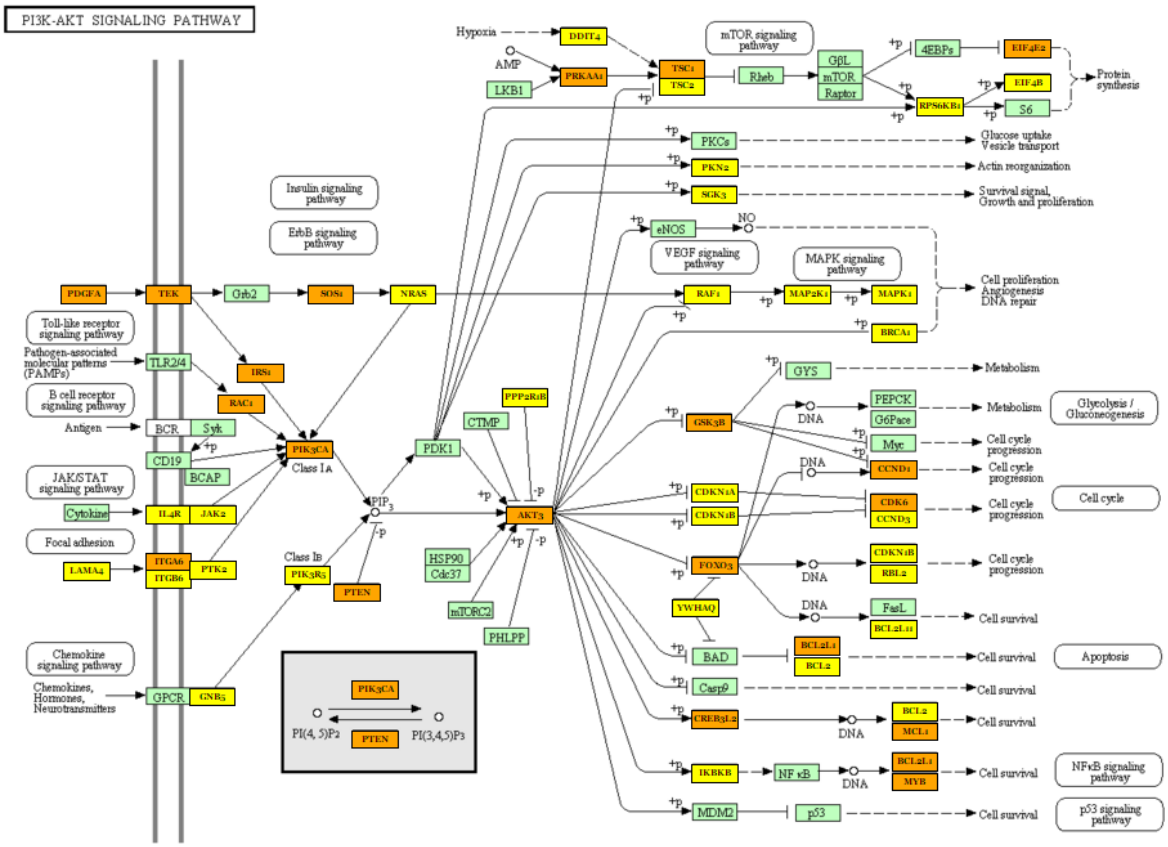


Figure 4.13 PI3K-AKT signaling pathway with highlighted targets of the fourteen miRNAs that were upregulated in group A and downregulated group B with largest fold changes. Green box: not a target; yellow box: genes targeted by one miRNA; brown box: genes targeted by more than one miRNAs

Table 4.7 Number of predicted genes in the top two pathways of 14 miRNAs expressed with largest variance among groups

miRNA	# Targets in pathways	
	Focal Adhesion	PI3K-AKT signaling
hsa-miR-29c-3p	26	31
hsa-miR-29b-3p	26	31
hsa-miR-424-5p	24	36
hsa-miR-19a-3p	16	15
hsa-miR-19b-3p	16	15
hsa-miR-101-3p	14	24
hsa-miR-148a-3p	12	17
hsa-miR-542-3p	9	9
has-miR-301-3p	8	12
hsa-miR-142-3p	4	10
hsa-miR-378d	4	4
hsa-miR-660-5p	3	6
hsa-miR-455-5p	1	1
hsa-miR-215	1	0

4.3 Discussion

The initial goal for profiling miRNA expression in end-stage livers was to search for miRNA expression signatures associated to some, but not all, diseases. In the absence of appropriate non-diseased controls, pairwise comparisons between all diseases were conducted. Some interesting observations were derived from such analyses. For example, a relatively large number of differentially expressed miRNAs were found in comparisons that included AIH or PSC, when compared with CRP, HCC or NASH, especially HCC (Table 4.2).

Further analyses on the AIH versus HCC comparison and the HCC versus PSC comparison unravelled the presence of subgroups in AIH and PSC samples regarding miRNA expression (Figure 4.1 and 4.3). Interestingly, the top deregulated miRNAs in the two comparisons were almost the same (Figure 4.2 and 4.4), and in both differential expression analyses these miRNAs were dramatically increased in the smaller subgroups (data not shown). Additional analysis further

illustrated the similarity of the two subgroups in AIH and PSC (Figure 4.5 and 4.6), which encouraged us to investigate the miRNA expression pattern in all end-stage liver samples.

Remarkably, closer inspection of the global miRNA expression profiles by MDS analysis revealed two well-separated groups of samples along dimension X1 (Figure 4.7). For simplicity, such groups were dubbed A and B. Group A comprised five PSC samples, four AIH samples, two PBC and two ETH, while group B contained the other 50 samples encompassing all diseases under study. Thus, group A was integrated mostly by cholestatic and autoimmune diseases (the ETH samples being the exception). To some extent, this explained the results obtained in pairwise comparisons. For example, all HCC were located in group B, and a large number of miRNAs were found differentially expressed when this disease was compared to AIH or PSC. It is possible that the subset of AIH and PSC samples included in group A largely influenced the results of differential expression analyses. Given the close relationship of miRNAs expression with cellular physiology, we sought to investigate the possible biological differences between livers in groups A and B.

To rule out that the composition of clusters A and B could be the result of a sampling artefact, or be associated to particular clinical observations, we analyzed the metadata that was available. This included gender, age, comorbidities in patients, liver tissue collection time, methods for processing of samples and possible batch effects. None of the above was able to explain the dichotomy.

Then we conducted some more statistical analysis to the two groups. When 60 miRNAs that exhibited the largest variance between samples were analysed by hierarchical clustering, samples in group A and B formed two discrete clusters (as in the MDS analyses) occupying two separate branches on the dendrogram (Figure 4.9). More importantly, we observed a set of 16 miRNAs that were extremely abundant in group A, as compared to samples in group B. In addition to those 16 miRNAs, many other miRNAs deployed the same pattern of higher abundance in group A, but to a less extent. The converse situation was also observed: some miRNAs accumulated to a higher level in group B than in group A. But in such cases the contrast between the two groups was less dramatic. However, from this analysis alone it was not possible to discriminate whether the conspicuously abundant miRNAs were upregulated in group A or downregulated in group B.

Fortunately, in a later stage of this project, we had access to twelve non-tumor tissues from surgically resected HCC livers (see Chapter 5). Theoretically, if a HCC patient is subjected to

tumor resection, the degree of damage of the liver is expected to be less severe than that of patients selected as candidates for liver transplantation. In other words, likely these resected non-tumor tissues are closer to a physiologically normal liver than tissues from most explanted livers. Following this rational, the resected non-tumor samples were used as controls to gain additional insights regarding the condition of samples in group A or group B.

In the MDS plot the three groups resulted in three well-defined clusters (Figure 4.10). The twelve controls samples clustered in a very compact group. The vast similarity of miRNA expression in these tissues may be due to the fact that they were all taken from relatively more functional livers. The group A samples were at approximately the same position on dimension X1 of the plot.-Group B samples, on the contrary, were positioned on the other extreme of dimension X1.

From above we concluded that the resected liver controls, group A and group B constituted three distinct subpopulations. But in some respect, samples in group A were more similar to those in the control group than samples in group B. One possible explanation was that samples in group A had undergone less severe physiological degradation than samples in group B; they were perhaps less diseased samples than those in group B.

In unsupervised hierarchical clustering analyses, we noticed that most miRNAs that were highly abundant in group A as compared to group B were also abundant in the control group. Indeed differential expression analysis revealed that many miRNAs in group A were statistically significantly more abundant than in control group. Given the multiplicity of miRNAs that exhibited this behavior, we decided to centre our attention on a sub-population of miRNAs that show the largest variance in abundance across samples and that exhibited the highest abundance in samples in group A. This resulted in a selection of fourteen miRNAs including miR-29c-3p, miR-29b-3p, miR-424-5p, miR-19a-3p, miR-19b-3p, miR-101-3p, miR-148a-3p, miR-542-3p, miR-301-3p, miR-142-3p, miR-378d, miR-660-5p, miR-455-5p, and miR-215.

Those fourteen miRNAs were subjected to pathway analysis and as a result the focal adhesion and PI3K-AKT signaling pathways were identified as the top pathways possibly regulated by this set of miRNAs. Importantly, many genes in these two pathways, such as ITGA, ITGB, FAK, PDGFA, Actinin, PTK2 have been reported as essential for perpetuating activation of myofibroblasts, which is central to hepatic fibrogenesis (Marra et al., 1999; Melton et al., 2007; Pellicoro et al., 2014; Reif et al., 2003). Moreover, miR-29b, miR-29c, miR-424-5p and miR-17-92 cluster have been reported playing important roles in tumorigenesis. Nevertheless, the

biological significance of these miRNAs and the subgrouping among end-stage livers needs further investigation in terms of wet-lab experiments, which is beyond the scope of this thesis.

In summary, this study profiled miRNA expression in end-stage livers using NGS. We found that despite being affected by different diseases, these livers formed two distinct groups based on their miRNA expression profiles. A set of miRNAs samples accumulated at low abundance in group B and, according to pathway analyses, they have the potential to negatively regulate focal adhesion and PI3K-AKT signaling pathways. Some of those miRNAs have been found playing important roles in suppressing fibrosis and tumorigenesis. Some other miRNAs, such as miR-378d, miR-660-5p, miR-455-5p, and miR-215, have been only partially studied, especially regarding their roles in liver pathogenesis. Although our findings were derived *in silico* from miRNAs expression profiles, they deserve further experimental investigation because our findings are consistent with previous reports in the literature.

One of the main limitations of this study was the large variability in miRNAs profiles found within diseases. This hindered the identification of miRNA expression signatures associated to specific diseases. We reasoned that such heterogeneity should derive from genetic and environmental factor affecting each patient. To circumvent this problem, it was decided to apply the methodology optimized in this study to interrogate the miRNAs profiles in paired tissues from patients affected by hepatocellular carcinoma. Namely, cancer tissues from HCC tumors were compared against non-cancer distal tissues of the same liver. The description of such study is presented in Chapter 5.

References

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M.J., Tuschl, T., Margalit, H., 2005. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 33, 2697-2706.

Ameres, S.L., Zamore, P.D., 2013. Diversifying microRNA sequence and function. Nature reviews. Molecular cell biology 14, 475-488.

Bala, S., Marcos, M., Kodys, K., Csak, T., Catalano, D., Mandrekar, P., Szabo, G., 2011. Up-regulation of microRNA-155 in macrophages contributes to increased tumor necrosis factor {alpha} (TNF{alpha}) production via increased mRNA half-life in alcoholic liver disease. J Biol Chem 286, 1436-1444.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281-297.

Baskerville, S., Bartel, D.P., 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA 11, 241-247.

Bernstein, E., Caudy, A.A., Hammond, S.M., Hannon, G.J., 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409, 363-366.

Beuers, U., Gershwin, M.E., Gish, R.G., Invernizzi, P., Jones, D.E., Lindor, K., Ma, X., Mackay, I.R., Pares, A., Tanaka, A., Vierling, J.M., Poupon, R., 2015. Changing nomenclature for PBC: From 'cirrhosis' to 'cholangitis'. J Hepatol.

Borchert, G.M., Lanier, W., Davidson, B.L., 2006. RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol 13, 1097-1101.

Caldwell, S., 2010. Cryptogenic cirrhosis: what are we missing? Curr Gastroenterol Rep 12, 40-48.

Chen, H., Sun, Y., Dong, R., Yang, S., Pan, C., Xiang, D., Miao, M., Jiao, B., 2011. Mir-34a is upregulated during liver regeneration in rats and is associated with the suppression of hepatocyte proliferation. PloS one 6, e20238.

Chen, Y., Verfaillie, C.M., 2014. MicroRNAs: the fine modulators of liver development and function. Liver Int 34, 976-990.

Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., Shiekhattar, R., 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature 436, 740-744.

Cullen, B.R., 2004. Transcription and processing of human microRNA precursors. Mol Cell 16, 861-865.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., Hannon, G.J., 2004. Processing of primary microRNAs by the Microprocessor complex. Nature 432, 231-235.

Dippold, R.P., Vadigepalli, R., Gonye, G.E., Patra, B., Hoek, J.B., 2013. Chronic ethanol feeding alters miRNA expression dynamics during liver regeneration. Alcohol Clin Exp Res 37 Suppl 1, E59-69.

Esau, C., Davis, S., Murray, S.F., Yu, X.X., Pandey, S.K., Pear, M., Watts, L., Booten, S.L., Graham, M., McKay, R., Subramaniam, A., Propp, S., Lollo, B.A., Freier, S., Bennett, C.F., Bhanot, S., Monia, B.P., 2006. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. Cell Metab 3, 87-98.

Ferreira, D.M., Simao, A.L., Rodrigues, C.M., Castro, R.E., 2014. Revisiting the metabolic syndrome and paving the way for microRNAs in non-alcoholic fatty liver disease. FEBS J 281, 2503-2524.

Forman, J.J., Legesse-Miller, A., Coller, H.A., 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proceedings of the National Academy of Sciences of the United States of America 105, 14879-14884.

Gori, M., Arciello, M., Balsano, C., 2014. MicroRNAs in nonalcoholic fatty liver disease: novel biomarkers and prognostic tools during the transition from steatosis to hepatocarcinoma. Biomed Res Int 2014, 741465.

Gregory, R.I., Chendrimada, T.P., Cooch, N., Shiekhattar, R., 2005. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell 123, 631-640.

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., Shiekhattar, R., 2004. The Microprocessor complex mediates the genesis of microRNAs. Nature 432, 235-240.

Ha, M., Kim, V.N., 2014. Regulation of microRNA biogenesis. Nature reviews. Molecular cell biology 15, 509-524.

Hausser, J., Zavolan, M., 2014. Identification and consequences of miRNA-target interactions--beyond repression of gene expression. Nature reviews. Genetics 15, 599-612.

Hsu, S.H., Ghoshal, K., 2013. MicroRNAs in Liver Health and Disease. Curr Pathobiol Rep 1, 53-62.

Iliopoulos, D., Drosatos, K., Hiyama, Y., Goldberg, I.J., Zannis, V.I., 2010. MicroRNA-370 controls the expression of microRNA-122 and Cpt1alpha and affects lipid metabolism. J Lipid Res 51, 1513-1523.

Khvorova, A., Reynolds, A., Jayasena, S.D., 2003. Functional siRNAs and miRNAs exhibit strand bias. Cell 115, 209-216.

Kodama, T., Takehara, T., Hikita, H., Shimizu, S., Shigekawa, M., Tsunematsu, H., Li, W., Miyagi, T., Hosui, A., Tatsumi, T., Ishida, H., Kanto, T., Hiramatsu, N., Kubota, S., Takigawa, M., Tomimaru, Y., Tomokuni, A., Nagano, H., Doki, Y., Mori, M., Hayashi, N., 2011. Increases in p53 expression induce

CTGF synthesis by mouse and human hepatocytes and result in liver fibrosis in mice. The Journal of clinical investigation 121, 3343-3356.
Krol, J., Loedige, I., Filipowicz, W., 2010. The widespread regulation of microRNA biogenesis, function and decay. Nature reviews. Genetics 11, 597-610.
Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., Kim, V.N., 2003. The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 415-419.
Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., Kim, V.N., 2004. MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23, 4051-4060.
Li, L., Masica, D., Ishida, M., Tomuleasa, C., Umegaki, S., Kalloo, A.N., Georgiades, C., Singh, V.K., Khashab, M., Amateau, S., Li, Z., Okolo, P., Lennon, A.M., Saxena, P., Geschwind, J.F., Schlachter, T., Hong, K., Pawlik, T.M., Canto, M., Law, J., Sharaiha, R., Weiss, C.R., Thuluvath, P., Goggins, M., Shin, E.J., Peng, H., Kumbhari, V., Hutfless, S., Zhou, L., Mezey, E., Meltzer, S.J., Karchin, R., Selaru, F.M., 2014. Human bile contains microRNA-laden extracellular vesicles that can be used for cholangiocarcinoma diagnosis. Hepatology 60, 896-907.
Liberal, R., Vergani, D., Mieli-Vergani, G., 2015. Update on Autoimmune Hepatitis. J Clin Transl Hepatol 3, 42-52.
Lytle, J.R., Yario, T.A., Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proceedings of the National Academy of Sciences of the United States of America 104, 9667-9672.
Maheshwari, A., Thuluvath, P.J., 2006. Cryptogenic cirrhosis and NAFLD: are they related? Am J Gastroenterol 101, 664-668.
Marra, F., Arrighi, M.C., Fazi, M., Caligiuri, A., Pinzani, M., Romanelli, R.G., Efsen, E., Laffi, G., Gentilini, P., 1999. Extracellular signal-regulated kinase activation differentially regulates platelet-derived growth factor's actions in hepatic stellate cells, and is induced by in vivo liver injury in the rat. Hepatology 30, 951-958.
McDaniel, K., Herrera, L., Zhou, T., Francis, H., Han, Y., Levine, P., Lin, E., Glaser, S., Alpini, G., Meng, F., 2014. The functional role of microRNAs in alcoholic liver injury. J Cell Mol Med 18, 197-207.
Melton, A.C., Soon, R.K., Jr., Park, J.G., Martinez, L., Dehart, G.W., Yee, H.F., Jr., 2007. Focal adhesion disassembly is an essential early event in hepatic stellate cell chemotaxis. Am J Physiol Gastrointest Liver Physiol 293, G1272-1280.
Mendell, J.T., 2008. miRiad roles for the miR-17-92 cluster in development and disease. Cell 133, 217-222.
Mogilyansky, E., Rigoutsos, I., 2013. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. Cell death and differentiation 20, 1603-1614.
Munoz-Garrido, P., Garcia-Fernandez de Barrena, M., Hijona, E., Carracedo, M., Marin, J.J., Bujanda, L., Banales, J.M., 2012. MicroRNAs in biliary diseases. World J Gastroenterol 18, 6189-6196.
Ng, R., Song, G., Roll, G.R., Frandsen, N.M., Willenbring, H., 2012. A microRNA-21 surge facilitates rapid cyclin D1 translation and cell cycle progression in mouse liver regeneration. The Journal of clinical investigation 122, 1097-1108.
Ninomiya, M., Kondo, Y., Funayama, R., Nagashima, T., Kogure, T., Kakazu, E., Kimura, O., Ueno, Y., Nakayama, K., Shimosegawa, T., 2013. Distinct microRNAs expression profile in primary biliary cirrhosis and evaluation of miR 505-3p and miR197-3p as novel biomarkers. PloS one 8, e66086.
Noetel, A., Kwiecinski, M., Elfimova, N., Huang, J., Odenthal, M., 2012. microRNA are Central Players in Anti- and Profibrotic Gene Regulation during Liver Fibrosis. Front Physiol 3, 49.
O'Neill, L.A., Sheedy, F.J., McCoy, C.E., 2011. MicroRNAs: the fine-tuners of Toll-like receptor signalling. Nature reviews. Immunology 11, 163-175.

Ota, A., Tagawa, H., Karnan, S., Tsuzuki, S., Karpas, A., Kira, S., Yoshida, Y., Seto, M., 2004. Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma. Cancer Res 64, 3087-3095.

Padgett, K.A., Lan, R.Y., Leung, P.C., Lleo, A., Dawson, K., Pfeiff, J., Mao, T.K., Coppel, R.L., Ansari, A.A., Gershwin, M.E., 2009. Primary biliary cirrhosis is associated with altered hepatic microRNA expression. J Autoimmun 32, 246-253.

Pan, C., Chen, H., Wang, L., Yang, S., Fu, H., Zheng, Y., Miao, M., Jiao, B., 2012. Down-regulation of MiR-127 facilitates hepatocyte proliferation during rat liver regeneration. PloS one 7, e39151.

Pellicoro, A., Ramachandran, P., Iredale, J.P., Fallowfield, J.A., 2014. Liver fibrosis and repair: immune regulation of wound healing in a solid organ. Nature reviews. Immunology 14, 181-194.

Quinn GP, K.M., 2002. Experimental design and data analysis for biologists. Cambridge University Press.

Rayner, K.J., Suarez, Y., Davalos, A., Parathath, S., Fitzgerald, M.L., Tamehiro, N., Fisher, E.A., Moore, K.J., Fernandez-Hernando, C., 2010. MiR-33 contributes to the regulation of cholesterol homeostasis. Science 328, 1570-1573.

Reif, S., Lang, A., Lindquist, J.N., Yata, Y., Gabele, E., Scanga, A., Brenner, D.A., Rippe, R.A., 2003. The role of focal adhesion kinase-phosphatidylinositol 3-kinase-akt signaling in hepatic stellate cell proliferation and type I collagen expression. J Biol Chem 278, 8083-8090.

Roderburg, C., Urban, G.W., Bettermann, K., Vucur, M., Zimmermann, H., Schmidt, S., Janssen, J., Koppe, C., Knolle, P., Castoldi, M., Tacke, F., Trautwein, C., Luedde, T., 2011. Micro-RNA profiling reveals a role for miR-29 in human and murine liver fibrosis. Hepatology 53, 209-218.

Rogler, C.E., Levoci, L., Ader, T., Massimi, A., Tchaikovskaya, T., Norel, R., Rogler, L.E., 2009. MicroRNA-23b cluster microRNAs regulate transforming growth factor-beta/bone morphogenetic protein signaling and liver stem cell differentiation by targeting Smads. Hepatology 50, 575-584.

Satapathy, S.K., Sanyal, A.J., 2015. Epidemiology and Natural History of Nonalcoholic Fatty Liver Disease. Semin Liver Dis 35, 221-235.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., Zamore, P.D., 2003. Asymmetry in the assembly of the RNAi enzyme complex. Cell 115, 199-208.

Sekiya, Y., Ogawa, T., Iizuka, M., Yoshizato, K., Ikeda, K., Kawada, N., 2011. Down-regulation of cyclin E1 expression by microRNA-195 accounts for interferon-beta-induced inhibition of hepatic stellate cell proliferation. J Cell Physiol 226, 2535-2542.

Shigehara, K., Yokomuro, S., Ishibashi, O., Mizuguchi, Y., Arima, Y., Kawahigashi, Y., Kanda, T., Akagi, I., Tajiri, T., Yoshida, H., Takizawa, T., Uchida, E., 2011. Real-time PCR-based analysis of the human bile microRNAome identifies miR-9 as a potential diagnostic biomarker for biliary tract cancer. PloS one 6, e23584.

Starley, B.Q., Calcagno, C.J., Harrison, S.A., 2010. Nonalcoholic fatty liver disease and hepatocellular carcinoma: a weighty connection. Hepatology 51, 1820-1832.

Szabo, G., Bala, S., 2013. MicroRNAs in liver disease. Nature reviews. Gastroenterology & hepatology 10, 542-552.

Takada, S., Asahara, H., 2012. Current strategies for microRNA research. Modern rheumatology / the Japan Rheumatism Association 22, 645-653.

Venugopal, S.K., Jiang, J., Kim, T.H., Li, Y., Wang, S.S., Torok, N.J., Wu, J., Zern, M.A., 2010. Liver fibrosis causes downregulation of miRNA-150 and miRNA-194 in hepatic stellate cells, and their overexpression causes decreased stellate cell activation. Am J Physiol Gastrointest Liver Physiol 298, G101-106.

Vickers, K.C., Shoucri, B.M., Levin, M.G., Wu, H., Pearson, D.S., Osei-Hwedieh, D., Collins, F.S., Remaley, A.T., Sethupathy, P., 2013. MicroRNA-27b is a regulatory hub in lipid metabolism and is altered in dyslipidemia. Hepatology 57, 533-542.

Wang, D., Wei, Y., Pagliassotti, M.J., 2006. Saturated fatty acids promote endoplasmic reticulum stress and liver injury in rats with hepatic steatosis. Endocrinology 147, 943-951.

Wang, L., Jia, X.J., Jiang, H.J., Du, Y., Yang, F., Si, S.Y., Hong, B., 2013. MicroRNAs 185, 96, and 223 repress selective high-density lipoprotein cholesterol uptake through posttranscriptional inhibition. Mol Cell Biol 33, 1956-1964.

Wang, X.W., Heegaard, N.H., Orum, H., 2012. MicroRNAs in liver disease. Gastroenterology 142, 1431-1443.

Washington, M.K., 2007. Autoimmune liver disease: overlap and outliers. Mod Pathol 20 Suppl 1, S15-30.

Xu, H., He, J.H., Xiao, Z.D., Zhang, Q.Q., Chen, Y.Q., Zhou, H., Qu, L.H., 2010. Liver-enriched transcription factors regulate microRNA-122 that targets CUTL1 during liver development. Hepatology 52, 1431-1442.

Yi, R., Qin, Y., Macara, I.G., Cullen, B.R., 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & development 17, 3011-3016.

Yuan, Q., Loya, K., Rani, B., Mobus, S., Balakrishnan, A., Lamle, J., Cathomen, T., Vogel, A., Manns, M.P., Ott, M., Cantz, T., Sharma, A.D., 2013. MicroRNA-221 overexpression accelerates hepatocyte proliferation during liver regeneration. Hepatology 57, 299-310.

Zezos, P., Renner, E.L., 2014. Liver transplantation and non-alcoholic fatty liver disease. World J Gastroenterol 20, 15532-15538.

Zhao, J.L., Rao, D.S., Boldin, M.P., Taganov, K.D., O'Connell, R.M., Baltimore, D., 2011. NF-kappaB dysregulation in microRNA-146a-deficient mice drives the development of myeloid malignancies. Proceedings of the National Academy of Sciences of the United States of America 108, 9184-9189.

Chapter 5

Identification of miRNAs associated with hepatocellular carcinoma through next generation sequencing and bioinformatic approaches

5.1 Introduction

5.1.1 Importance and risk factors of HCC

Primary liver cancer is the sixth most common cancer and the second leading cause of cancer-related deaths in the world. In 2012 there were 782,000 new cases diagnosed and primary liver cancer was responsible for 745,000 deaths (WHO, 2014b). Hepatocellular carcinoma (HCC) accounts for most primary liver cancers. The prevalence varies by geographical region, where Asian and some parts of Africa suffer the highest incidence rate (Venook et al., 2010). Although relatively uncommon in North America, the incidence of HCC has steadily increased in the past 15 years in both the United States (Berry et al., 2012) and Canada (De et al., 2013). The incidence in men tripled from 2.2 to 6.8 per 100,000 inhabitants in Canada, according to the Canadian Cancer Statistics in 2012, and was projected to increase to 18.5 per 100,000 inhabitants in 2015 (Pocobelli et al., 2008).

HCC is a malignant tumour derived from hepatocytes, and the most frequent hepatic neoplasm in the liver. HCC commonly results from underlying chronic liver diseases including hepatitis B virus (HBV) and hepatitis C virus (HCV) infection, alcoholic cirrhosis, genetic diseases, chemicals (e.g. aflatoxin B1), autoimmune hepatitis, and non-alcoholic steatohepatitis (El-Serag and Rudolph, 2007). The chronic inflammation of the liver leads to fibrosis and gradually develops into cirrhosis and 80-90% of HCC cases arise from cirrhosis (El-Serag, 2011).

To date, therapeutic options for HCC include surgical resection, trans arterial chemo-embolization, radiofrequency ablation and liver transplantation. However, since in most cases HCC is diagnosed at a late stage when patients present an unresectable tumour and/or develop distant metastases, these treatments are no longer viable options. As a result, the overall 5-year survival rate of HCC is as poor as 5-9%. In patients who had a resection, the 5-year survival rate is only 30-40% (Aravalli et al., 2008) mainly due to chemotherapy resistance and recurrence or

previously unnoticed metastases. Currently, early detection of HCC relies on surveillance by ultrasound in patients with cirrhosis or serological test for increased alfa-fetoprotein. However, neither of the methods exceeds 60% sensitivity (Singal et al., 2009; Trevisani et al., 2001). Advances in imaging technologies such as magnetic resonance imaging and computed tomography have largely improved the diagnosis, but these procedures are pricy and not easily accessible in developing countries. Therefore, the study of mechanisms involved in HCC tumorigenesis remains an essential task in order to discover new biomarkers at early stages of the disease as well as to develop therapeutics that is more effective.

5.1.2 Gene expression deregulation during HCC

The pathogenesis of HCC is known as a stepwise process which begins from pre-neoplasia to dysplasia to a final neoplasia stage (Thorgeirsson and Grisham, 2002). Genetic alteration is one of the most important mechanisms associated with HCC initiation and progression. Genetic changes can be observed as early as pre-neoplastic lesions of cirrhotic livers, and they are thought to be the initiating events in hepatocarcinogenesis. The irreversible genetic abnormalities accumulate in hepatocytes, which further lead to disrupted gene expression, and finally to malignant transformation (Aravalli et al., 2008).

Over the last decade, many researches have focused on identifying key genes and signaling pathways involved in the development of HCC. Different mechanisms associated with genetic alteration of cancer cells have been characterized, including sustained proliferation (e.g. over activation of IGF, BRAF and mutation of PTEN) (Colombino et al., 2012; Kim et al., 1998; Yao et al., 1999), evasion of growth suppressors (e.g.TP53 mutation and MDM2 overexpression) (Jablkowski et al., 2005), invasion and metastasis (e.g. up-regulation of E-cadherin and VEGF) (Mise et al., 1996; Wei et al., 2002) and energy metabolism changes (e.g. activation of c-myc, mTORC2 and AKT, or inactivation of TP53) (Beyoglu et al., 2013; Hagiwara et al., 2012; Levine, 1997; Yuneva et al., 2012). The signaling pathways such as p53, b-Catenin, VEGF, EGFR, as well as IGF signaling and Hippo pathway have also been related to the development of HCC. Several reviews have summarized the molecular mechanisms of HCC (Aravalli et al., 2008; Farazi and DePinho, 2006) and the related signaling pathways (Marquardt et al., 2012; Villanueva et al., 2007).

Besides genetic alterations of tumour suppressors and oncogenes, deregulation of miRNAs has been shown to modulate these genes through post-transcriptional mechanisms. Accumulating data highlights the importance of deregulation of miRNA expression in liver carcinogenesis.

Since its discovery, miRNAs have been found to play a crucial role in development of cancer. They can behave as oncomiRs or tumour suppressor miRNAs by directly or indirectly regulating the expression of key proteins involved in cancer-related pathways (Lujambio and Lowe, 2012). Deregulation of miRNAs has been repeatedly described in a variety of cancers such as breast (Zhang et al., 2014), gastrointestinal (Schetter et al., 2012; Song and Ajani, 2013), urinary tract (Catto et al., 2011), lung (Kang and Lee, 2014) and liver (Morishita and Masaki, 2014), among others. Accordingly, miRNAs have been proposed as biomarkers for early diagnosis, disease staging and prognosis (Di Leva et al., 2014), as well as promising targets/tools for cancer treatment (Giordano and Columbano, 2013; Hayes et al., 2014).

In HCC, multiple profiling studies based on microarray and qPCR have identified the deregulation of miRNA in tumors. However, results from those studies are inconsistent for some specific miRNAs, which could be due to multiple reasons such as adoption of different techniques, diversity of samples type and size and selection of different approaches for data analyses (Shi et al., 2015). There have been very few studies using NGS to profile miRNA expression in HCC samples and the results from them have been inconsistent as well (Bao et al., 2014; Hou et al., 2011; Wojcicka et al., 2014). Nevertheless, the deregulation of a number of miRNAs has been consistently reported (Shi et al., 2015; Yang et al., 2014a), as summarized in Table 5.1.

Table 5.1 Most reported HCC associated miRNAs based on 26 miRNA profiling studies using microarray or qPCR[1]

miRNA	Expression in tumors	Chromosome	No. of Studies	microRNA Cluster
miR-223-3p	down	Xq12	9	N/A
miR-214-3p	down	1q24.3	13	miR-199/miR-214
miR-199a-5p	down	19p13.2/1q24.3	14	N/A
miR-199a-3p	down	19p13.2/1q24.3	15	N/A
miR-195-5p	down	17p13.1	14	miR-195/miR-497
miR-150-5p	down	19q13.33	10	N/A
miR-145-5p	down	5q32	13	miR-143/miR-145
miR-130a-3p	down	11q12.1	9	miR-130/miR-301/miR-454
miR-375	down	2q35	3	N/A
miR-93-5p	up	7q22.1	12	miR-25/miR-93/miR-106
miR-224-5p	up	Xq28	12	miR-224/miR-452
miR-222-3p	up	Xp11.3	12	miR-221/miR-222
miR-221-3p	up	Xp11.3	16	miR-221/miR-222
miR-21-5p	up	17q23.1	15	N/A

1. Modified from Shi et al., 2015 and Yang et al., 2014.

Moreover, efforts have been made to associate the deregulation of specific miRNAs to HCC tumourigenesis and identify their role in cancer biology, including cell proliferation, cell-cycle control, apoptosis, cell growth, angiogenesis and tumour metastasis (Hung et al., 2014; Morishita and Masaki, 2014). For example, studies have found that up-regulation of miR-21 was involved in tumour metastasis through targeting of PTEN, RHPB and PDCD4 (Connolly et al., 2010; Meng et al., 2007; Zhu et al., 2012). Other common upregulated miRNAs, such as miR-224, miR-216 and miR-182, were also associated with tumour proliferation, anti-apoptosis activity and early carcinogenesis or metastasis, through regulating different genes involved in these cellular processes (Chen et al., 2012; Wang et al., 2012a; Zhang et al., 2013b). Among downregulated miRNAs in HCC, deregulation of miR-199a has been related to cancer cell invasion and metastasis (Lee et al., 2015; Shen et al., 2010). Previous studies have also found miRNAs in the miR-200 family, which includes miR-200a, miR-200b, miR-200c, miR-141 and miR-429, play key tumour-suppressive roles in epithelial-to-mesenchymal transition (EMT) (Gregory et al., 2008),

and enhancing HCC cell proliferation, migration and metastasis through different mechanisms (Hung et al., 2013; Wong et al., 2015; Yuan et al., 2011; Zhang et al., 2013a).

Given the pre-eminent role of miRNAs on HCC phenotypical traits, researchers have tried to link the alteration of miRNAs expression with clinical application, such as molecular classification and prognosis of the disease, in order to improve the outcome. In a microarray profiling study, Murakami et al. found that the expression level of miR-92, miR-20, miR-18 were reversely correlated with the degree of tumour differentiation (Murakami et al., 2006). More recently, based on miRNA profiles of 89 HCC samples, Toffanin and coworkers proposed a miRNA-based classification of HCC with three main clusters, namely the wingless-type MMTV integration site, interferon-related, and proliferation subclasses (Toffanin et al., 2011). More studies have investigated the prognostic role of miRNA expression in HCC. A panel of 19 miRNA were associated with patient survival (Jiang et al., 2008) and in another study, 20 miRNAs were identified as metastasis-related miRNAs in HCC (Budhu et al., 2008). Researches have also focused on the role of individual miRNAs in predicting prognosis, where over-expression of miR-10b (Li et al., 2012), miR-21, miR-221 (Karakatsanis et al., 2013) and down-regulation of miR-122 (Coulouarn et al., 2009), miR-139 (Wong et al., 2011) and miR-200 family (Dhayat et al., 2014) was correlated with poor prognosis.

Deregulation of miRNAs from the same gene cluster, whose coding genes are located in a proximal distance in a chromosome, was also reportedly associated with the disease prognosis. Augello et al. proposed that the chromosome 19 miRNA cluster (C19MC) was a molecular alteration characteristic of HCC and the high level of C19MC miRNAs indicated poor prognosis with increased risk of tumour recurrence and shorter overall survival time (Augello et al., 2012). Although C19MC is the largest miRNA cluster in human, its importance in HCC is still under-studied.

In addition to classification and prognosis prediction, detection of circulating miRNAs have been proposed to be used in early diagnosis of HCC (Qu et al., 2011; Tomimaru et al., 2012; Zhou et al., 2011), since miRNAs are found to be highly stable in serum. In studies using miRNAs as therapeutic targets or drugs for treating HCC, both targeting of oncomiRs and administration of onco-suppressive miRNAs have shown effectiveness in mice studies (Callegari et al., 2012; Hatziapostolou et al., 2011; Park et al., 2011). Thus, not only are miRNAs of fundamental importance to understand HCC biology, but also they are promissory targets for disease management.

5.1.3 Chapter overview

In this chapter, we describe experiments using NGS and bioinformatic approaches to profile miRNAs and gene expression in paired tumour and non-tumour tissues from human livers affected by HCC. The main goal was to identify deregulated miRNAs and to explore their possible interactions with deregulated genes and/or putative targets. We initially investigated levels of miRNAs and transcripts in four explanted livers (Table 2.2). This preliminary study led to the identification of a handful of deregulated miRNAs and hundreds of deregulated genes in the tumour tissues. In order to improve detection of deregulated miRNAs, we increased our sample size by acquiring twelve additional pairs of samples derived from resected livers (Table 2.3). This allowed the detection of more than one hundred deregulated miRNAs, including several cases not reported in the literature. By using a variety of analytical techniques, several putative interactions of the deregulated miRNAs with deregulated genes, or with their *in silico* predicted targets, were identified. Moreover, pathways analysis suggested that a set of miRNAs found downregulated in HCC tumours might be associated with the activation of the MAPK pathway, and the consequent tumorigenesis and metastasis.

5.2 Results

5.2.1 Description of sequenced libraries

5.2.1.1 Small RNA sequencing libraries

Small RNA libraries were constructed for four pairs of samples, tumour and non-tumour tissue, from explanted livers and twelve pairs of samples from resected livers. The clinical information of these samples is shown in Table 2.2 and Table 2.3 (Table 2). Libraries were sequenced on a MiSeq sequencer, which delivered high quality sequences, at an average depth of ~1.7 million sequences per library (Table 5.2). After alignment of sequences against the version 21 of the miRNAs database (miRBase21), only sequences that perfectly and uniquely aligned to the reference database were counted and used for further analyses. In order to evaluate whether expression of specific miRNAs could be linked with the development of HCC, the number of miRNAs detected in the tumour and non-tumour libraries were counted (out of 2588 contained in the miRBase21) (Table 5.2).

Table 5.2 Description of small RNA libraries on explanted livers from patients undergoing liver transplantation and resected livers from patients undergoing liver surgery

Sample source	Sample ID	sequenced reads[1]	Aligned percentage[2]	Number of detected miRNA[3]
Explanted livers	**311nt**	1981877	80.68%	542
	311t	2129426	67.82%	598
	314nt	1922029	75.31%	581
	314t	2116661	67.45%	605
	315nt	1397401	75.90%	557
	315t	2202415	68.03%	647
	338nt	2022127	69.91%	607
	338t	2083078	69.78%	585
Resected livers	**1NT**	1160792	51.00%	497
	1T	1809148	50.17%	546
	2NT	2304643	39.98%	614
	2T	2091984	69.84%	619
	3NT	1648585	55.87%	602
	3T	1616147	62.43%	557
	4NT	1355445	61.06%	579
	4T	2082864	58.52%	668
	5NT	1752840	57.38%	577
	5T	1686541	62.03%	536
	6NT	1252621	51.06%	523
	6T	822741	50.28%	498
	7NT	1571471	32.22%	496
	7T	1456741	50.22%	650
	8NT	1628282	60.83%	642
	8T	1742431	63.77%	702
	9NT	1928065	69.50%	605
	9T	1677829	68.09%	656
	10NT	1008314	55.52%	491
	10T	998957	64.07%	561
	11NT	1613057	59.44%	582
	11T	1540929	68.18%	562
	12NT	1279906	30.04%	461
	12T	1310130	53.11%	553

1. Sequenced reads refer to the total reads number generated by the sequencer
2. Aligned percentage refers to the percentage that sequenced reads aligned to the miRNA database, miRBase21, which includes 2588 human miRNAs in total
3. Number of miRNAs detected refers to the number of unique miRNAs aligned to miRBase21

5.2.1.2 Description of RNA-seq libraries for transcripts sequencing on tissues from explanted livers

To profile the mRNA expression, RNA-seq libraries were constructed and sequenced including tumour and non-tumour tissues from the 4 explanted livers, but not the 12 resected livers (due to financial constraints). Libraries were sequenced on a HiSeq 2500 instrument at an average depth of ~30 million paired end reads per library, which allowed the detection of ~ 19,000 transcripts per sample. Detailed baseline characteristics of the eight RNA-seq libraries constructed on tumour and non-tumour tissues from explanted livers are included in Table 5.3.

Table 5.3 Description of RNA-seq libraries on samples from explanted livers from patients undergoing liver transplantation

Sample	Sequenced		% aligned		number of detected genes
	End1	End2	End1	End2	
311n	28188225	28188225	94.48%	94.87%	18248
311t	27395487	27395487	93.18%	93.62%	20980
314n	27310008	27310008	94.64%	94.98%	17652
314t	27865353	27865353	95.22%	95.59%	16227
315n	26054838	26054838	94.44%	94.75%	16190
315t	31466586	31466586	94.64%	95.00%	21149
338n	40860228	40860228	95.22%	95.57%	21181
338t	30178941	30178941	94.64%	95.01%	19458

5.2.2 miRNA and gene expression in HCC in tissues from explanted livers

5.2.2.1 Data exploration and differential expression analysis of miRNA and gene expression profiles in tumours and non-tumour tissues from explanted livers

After quality filtering and normalization, the data were subjected to exploratory techniques in search for outliers and for inspection of inter-sample relationships. We conducted multidimensional scaling (MDS) analysis on miRNA and gene expression data from paired tumour and non-tumour tissues taken from the four explanted livers (Figure 5.1A and 5.2A). In general, non-tumour samples gathered into a small group whereas tumour tissues were randomly scattered. Tumour tissue from liver 314, however, was located in proximity to non-tumour samples.

Unsupervised hierarchical clustering was the other data exploratory analysis performed. The miRNA and gene expression profile on the eight tissues were clustered based on the 50 miRNAs

or genes exhibiting the largest variance across samples (Figure 5.1B and 5.2B). Similarly, in both plots the sample 314t also clustered with the non-tumour samples. This suggested that the tumour and paired non-tumour sample from patient 314 was more similar than the other pairs.

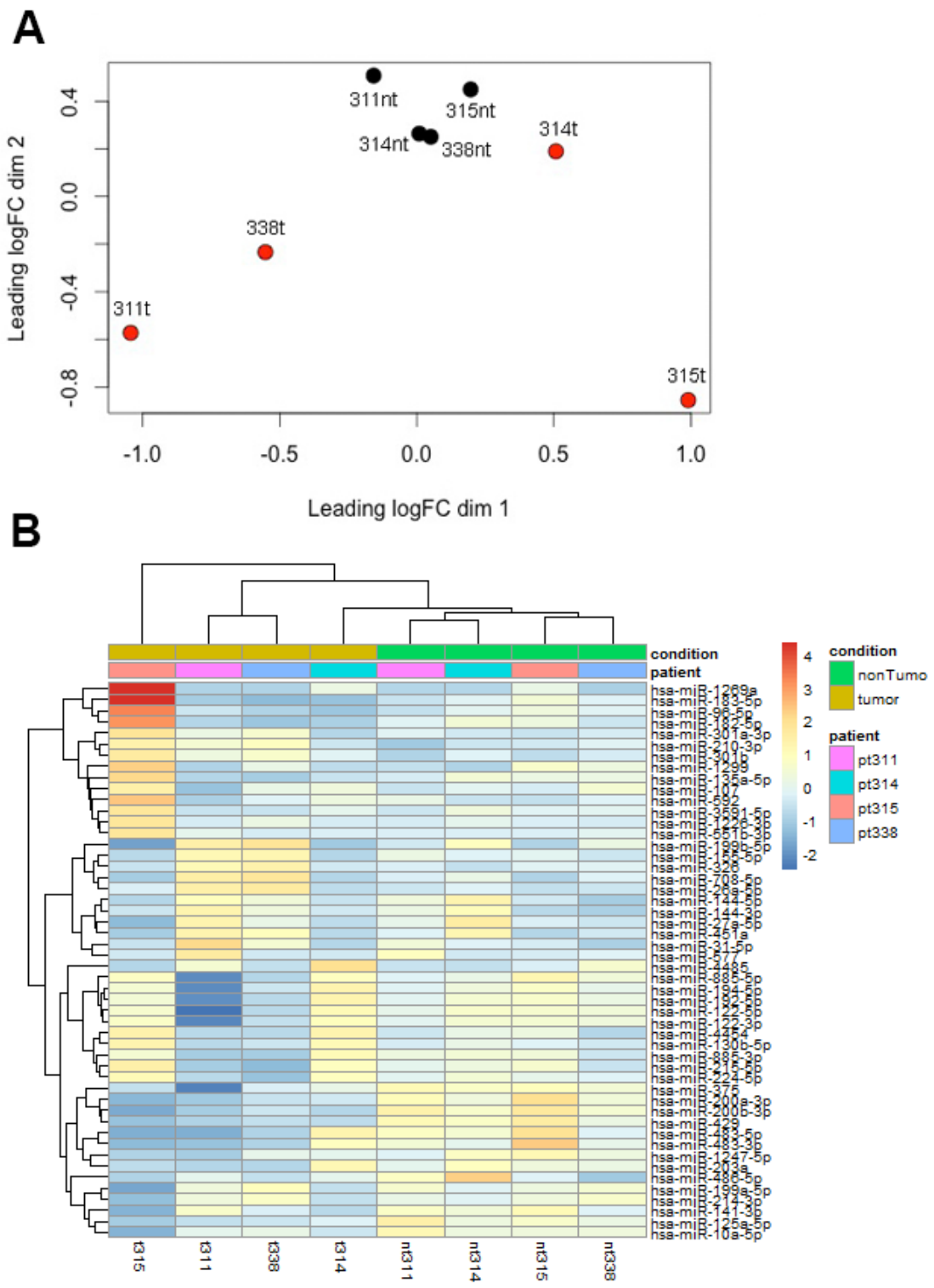


Figure 5.1 miRNA expression profiles in eight samples from four explanted livers reveals close proximity for 314t with the non-tumour samples. A) MDS analysis plot. Red dots: tumour tissues; black dots: non-tumour tissues B) Unsupervised hierarchical clustering of samples based on 50 miRNAs expressed with largest variance across samples

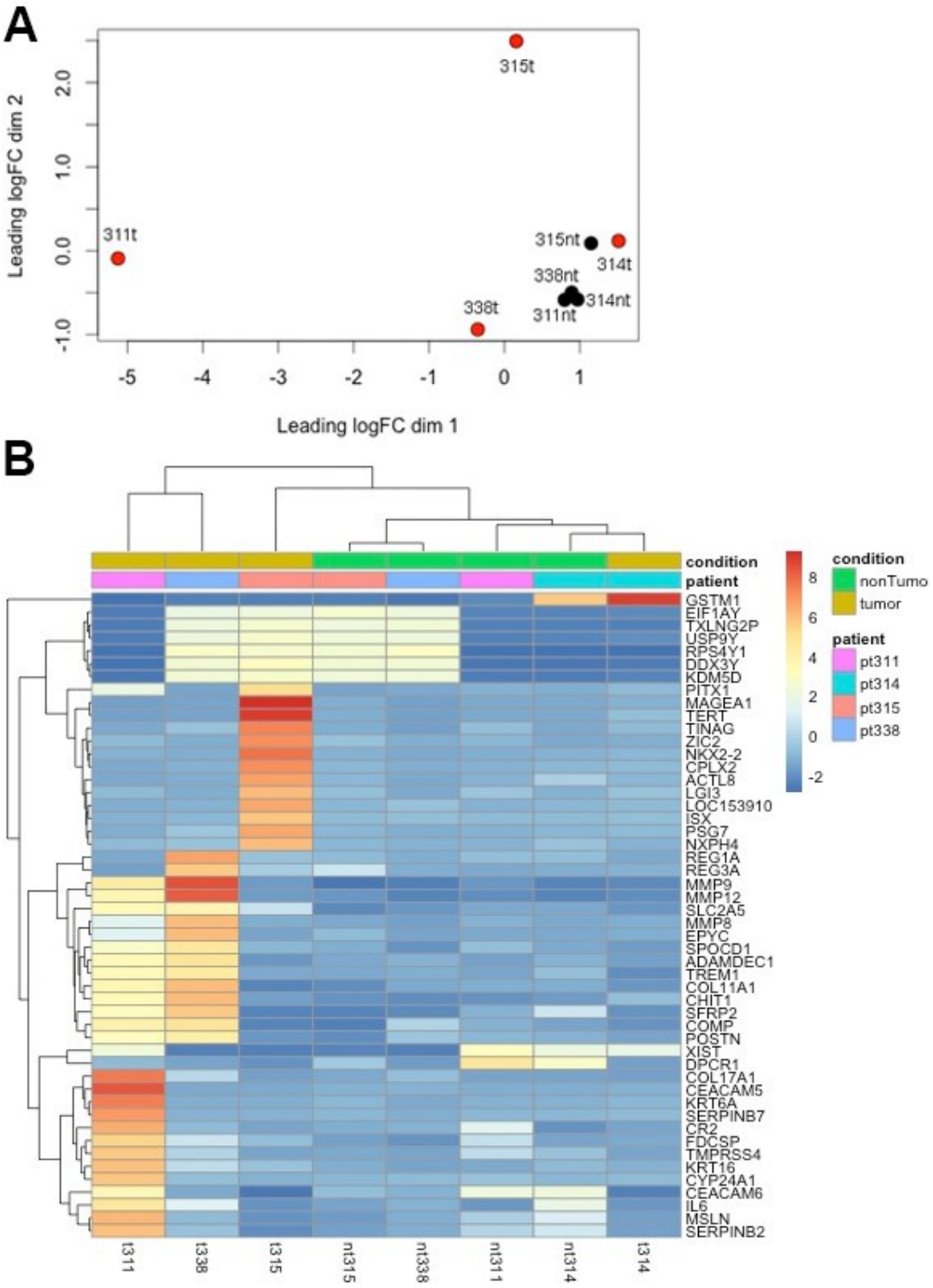


Figure 5.2 Gene expression profiles in eight samples from four explanted livers reflecting similarity of 314t with non-tumour tissues. A) MDS analysis plot. Red dots: tumour tissues; black dots: non-tumour tissues B) Unsupervised hierarchical clustering of samples based on 50 genes expressed with largest variance across samples

5.2.2.2 Differential expression analysis identified deregulated miRNAs in tumour tissues

After differential expression analysis of samples from explanted livers, a small number of miRNAs and genes (7 miRNAs and 120 genes) were found deregulated in the tumour group. Since in exploratory analyses the tumour sample from patient 314 behaved like non-tumour in both miRNA and gene expression data sets, we concerned that a dysplastic or benign tumour was collected as cancerous tumour from liver 314. Therefore, this liver was taken as a biological outlier and the differential analyses of miRNA and gene expression were repeated using three pairs of tissues from explanted livers. However, without pathological validation, there is insufficient evidence to remove 314t from analysis, as statistically it can still be a tail end of the distribution of a group of samples (tumor tissues) with larger variance.

In order to optimize the analysis, two different statistical packages (edgeR and DESeq2) were used to assess the differential expression. Changes in miRNA expression in tumours are described in $\log_2$ fold change. Accordingly, a log2FC = 1 indicates that there is a two-fold increase in expression in the tumour group. A log2FC negative value means that the expression of a miRNA or gene was found decreased in the tumour group. The false discovery rate (FDR) values are also shown in the table and the cut-off of FDR<0.05 was used.

After excluding liver/tumour pair 314, a total of 16 differential expressed miRNAs were detected (Table 5.4). Interestingly, among down-regulated miRNAs, miR-200b-3p, miR-429 and miR-200a-3p belong to the same miRNA family, named miR-200 family. The down-regulation of miRNAs in this family in HCC has been reported before, but mostly individually. MiR-483-3p and miR-485-5p, however, have not been well characterized in the context of HCC.

Among the upregulated miRNAs, miR-1269a was the top upregulated miRNA ($\log_2$FC = 4.56). Inspection of the hierarchical clustering results suggested that this difference was likely derived from the high expression level of this miRNA in sample 315t (Figure 5.1B). This miRNA has been related to HCC tumour invasion and metastasis, but the function of the miR-1269a is poorly understood. Additional upregulated miRNAs included some well-studied oncogenic miRNAs like miR-21-3p, miR-210, miR-26a and miR-301a-3p, but also some miRNAs that have not been previously reported in the context of HCC, like miR-708 and miR-951.

Table 5.4 Differentially expressed miRNAs in tumour tissues from explanted livers

miRNA	Expression in tumours	log2 Fold Change	FDR[1]	Detected by
hsa-miR-483-3p	down	-4.96	1.27E-04	edgeR and DESeq2
hsa-miR-483-5p	down	-4.40	1.58E-04	edgeR and DESeq2
hsa-miR-200b-3p	down	-3.63	1.27E-04	edgeR and DESeq2
hsa-miR-200a-3p	down	-3.20	1.27E-04	edgeR and DESeq2
hsa-miR-429	down	-2.73	1.27E-04	edgeR and DESeq2
hsa-miR-125a-5p	down	-2.09	1.90E-03	edgeR and DESeq2
hsa-miR-375	down	-2.09	4.45E-02	DESeq2
hsa-miR-203a	down	-1.53	4.24E-02	DESeq2
hsa-miR-1269a	up	4.56	1.27E-02	edgeR
hsa-miR-301b	up	3.09	8.35E-03	edgeR and DESeq2
hsa-miR-708-5p	up	2.66	1.00E-02	edgeR
hsa-miR-26a-5p	up	2.39	5.92E-03	edgeR and DESeq2
hsa-miR-21-3p	up	2.07	2.22E-02	edgeR
hsa-miR-210-3p	up	2.01	5.92E-03	edgeR and DESeq2
hsa-miR-301a-3p	up	1.90	3.07E-02	edgeR
hsa-miR-941	up	1.77	3.86E-02	edgeR and DESeq2

1. FDR: false discovery rate

5.2.2.3 Differential expression analysis identified hundreds of deregulated genes in HCC tumours

We then conducted differential expression analysis of genes between tumour and non-tumour tissues from the three explanted livers. Non-abundant transcripts were excluded from the analysis because low-level genes are likely less influential from a biological point of view. In total, 846 genes were detected as differentially expressed in tumours (Figure 5.3). Expression of 155 genes was increased with more than four-fold change (log2FC > 2), and 61 were found downregulated more than 4 times (log2FC < -2) in tumour.

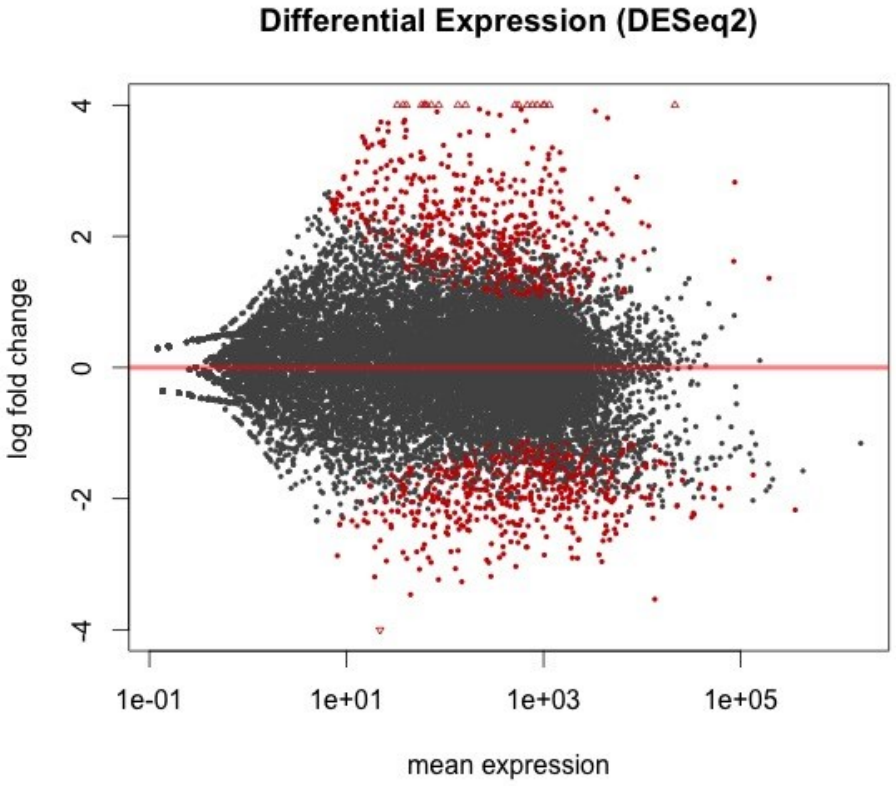


Figure 5.3 Log_2 fold change against mean expression of genes by DESeq2 differential expression analysis between tumour and non-tumour groups from explanted livers. Red dots: genes with FDR <0.05.

5.2.2.4 Putative interaction between deregulated miRNAs and deregulated genes

We found 16 differentially expressed miRNAs and 846 differentially expressed genes in explanted livers. To identify potential interactions between the differential expressed miRNAs and genes, we used TargetScan 7.0 to predict the targets of some top miRNAs and search them in the gene list we found deregulated in our RNA-seq data. Since miRNAs are generally recognized as negative regulators of gene expression, here we only show the miRNA and genes that showed opposite trends of expression (i.e. upregulated miRNAs/downregulated transcripts or vice versa).

To increase specificity in the target prediction analysis, gene targets that have a TargetScan score (context ++) larger than -0.2 were filtered out. According to this targeting prediction program, the lower the target score, the higher is the possibility that the target is a true target of a specific miRNA. In addition, only differentially expressed genes with a log2 fold change > 1.5 or < -1.5

were included. This resulted in discovery of putative interactions between miRNAs and genes (Figure 5.4). Interactions that have a potential role in tumourigenesis are shown in Table 5.5.

Interestingly, some gene targets detected in our analysis may have potential role in cancer, but they have never been studied in the context of HCC. For example, DOCK3 is an activator of the oncogene RAC1 (Sanz-Moreno et al., 2008), but neither the gene nor the miR-125a/DOCK3 regulation have been linked with HCC.

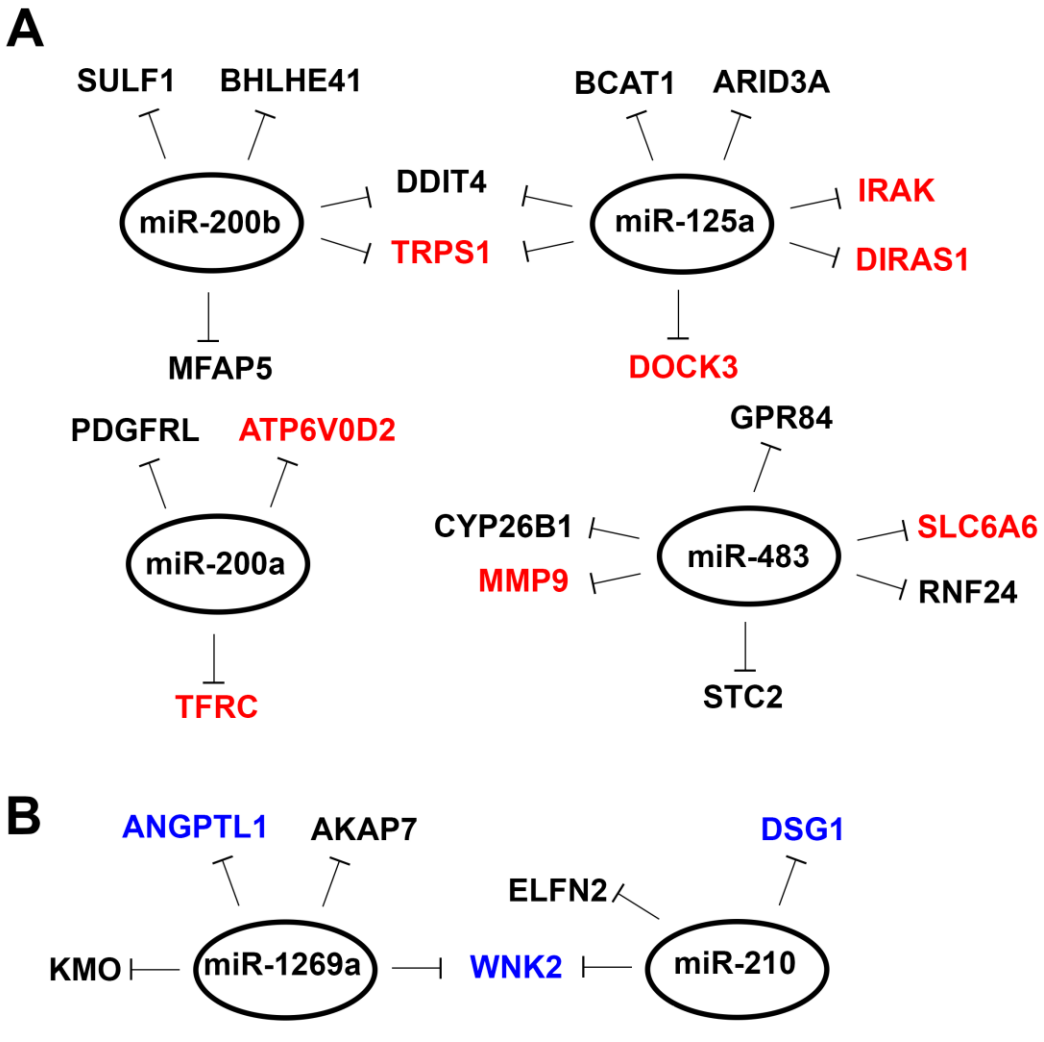


Figure 5.4 Possible interactions between deregulated miRNAs and genes identified in explanted livers. A) Down-regulated miRNAs and their predicted targets among up-regulated genes B) Up-regulated miRNAs and their predicted targets among down-regulated genes. Prediction were performed using TargetScan 7.0. Only interactions having opposite direction between miRNA and gene expression are shown. Putative oncogenes are marked in red and putative tumour-suppressive genes are marked in blue.

Table 5.5 Genes differentially expressed in tumour that are predicted targets of deregulated miRNAs[1]

Gene	Full name	Expression in tumours	log2 FC[2]	Targeted by
TRPS1	trichorhinophalangeal syndrome I	up	2.26	miR-200b, miR-125a
DOCK3	dedicator of cytokinesis 3	up	2.49	miR-125a
IRAK1	interleukin-1 receptor-associated kinase 1	up	1.87	miR-125a
DIRAS1	DIRAS family, GTP-binding RAS-like 1	up	4.68	miR-125a
ATP6V0D2	ATPase, H+ transporting, lysosomal 38kDa, V0 subunit d2	up	3.04	miR-200a
TFRC	transferrin receptor	up	1.59	miR-200a
MMP9	matrix metallopeptidase 9	up	3.05	miR-483
SLC6A6	solute carrier family 6 (neurotransmitter transporter), member 6	up	2.39	miR-483
ANGPTL1	angiopoietin-like 1	down	-1.67	miR-1269a
WNK2	WNK lysine deficient protein kinase 2	down	-1.94	miR-1269a, miR-210-3p
DSG1	desmoglein 1	down	-2.35	miR-210-3p

1. Only genes that have been reported having a role in cancer are listed
2. Log2 FC: log2 fold change

To improve purity of tumour and non-tumour samples as well as to enlarge the study cohort, we included twelve pairs of cancerous tissues and paired distal “normal” tissues from liver resection surgeries of HCC patients. These tissues were snap frozen within 30 min after surgery, which ensured high quality of samples. It is important to mention that livers subjected to resection present less advanced disease compared to patients who undergo liver transplantation. Therefore, the deregulated miRNAs in these livers may better reflect the mechanisms promoting HCC initiation and development.

MiRNA profiling data generated from resected liver tissues were subjected to the same analytical procedures described above. However, due to budgetary reasons gene expression profiling was

not performed in this study, instead pathway prediction were done to find out the potential function of the differential expressed miRNAs in tumours.

5.2.3 MiRNA expression in tissues from HCC resected livers

5.2.3.1 MiRNA expression signature differentiates HCC tumours from non-tumour tissues

Results from MDS analysis of the miRNA expression of the 12 pairs of samples are presented in Figure 5.5. As in the case of explanted livers, the 12 non-tumour tissues formed a tight cluster (black dots). The tumours were distributed in a scattered pattern, however. Although this type of pattern was observed for explanted livers, separation of tumour and non-tumour samples is much clearer in this case.

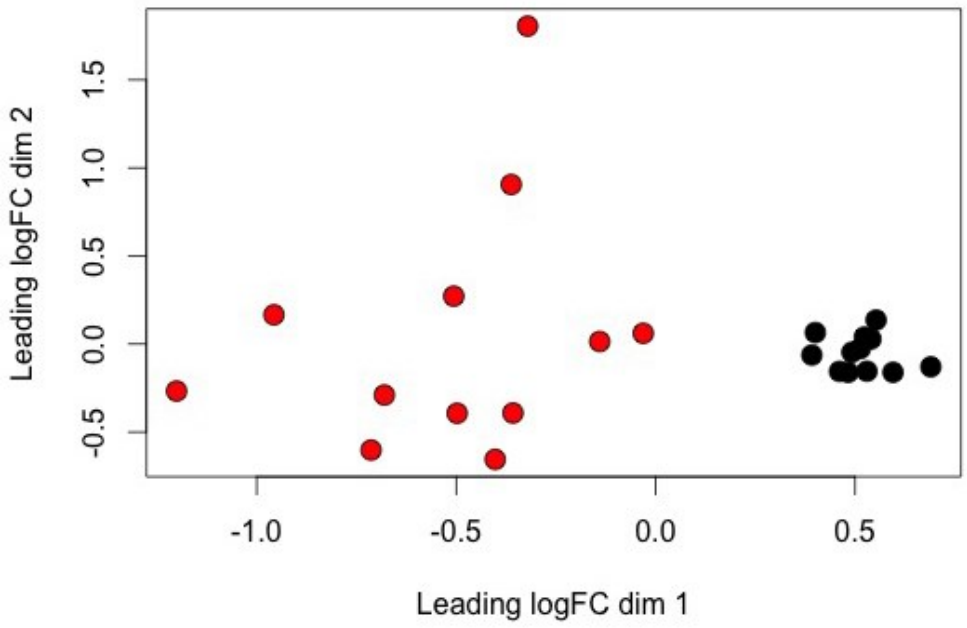


Figure 5.5 MDS analysis of miRNA expression profiles in 24 samples from 12 resected livers. Black dots: non-tumour tissue; red dots: tumour tissue.

Hierarchical clustering of the 50 miRNAs expressed with the largest variance across all samples (Figure 5.6) recapitulated the MDS pattern. Namely, all non-tumour tissues clustered together, while tumour tissues formed two discrete clusters. Interestingly, two tumours (7t and 8t) grouped together on the left side of the plot, with a miRNA expression profile clearly distinct from all other tumour and non-tumour samples. A conspicuous signature in this cluster is the up-regulation of many miRNAs in the C19MC cluster, such as miR-518, miR-512, miR-516, miR-515, miR-526,

miR-520, miR-517, and miR-519, among others. C19MC is the largest miRNA cluster found in the human genome (Bentwich et al., 2005). Interestingly, in one recent miRNA study on HCC liver samples infected with HCV, the up-regulation of miRNAs in this cluster was regarded as feature of one sub-cluster (Toffanin et al., 2011).

In addition, this figure also reveals some features of miRNA expression between the two groups. On the top row of the plot, miR-375 appears as the most downregulated miRNA in the tumours, followed by miRNAs in the miR-200s family, which was also observed in explanted livers. MiR-1269a, at the bottom of the figure, exhibited increased expression in about half of the tumours.

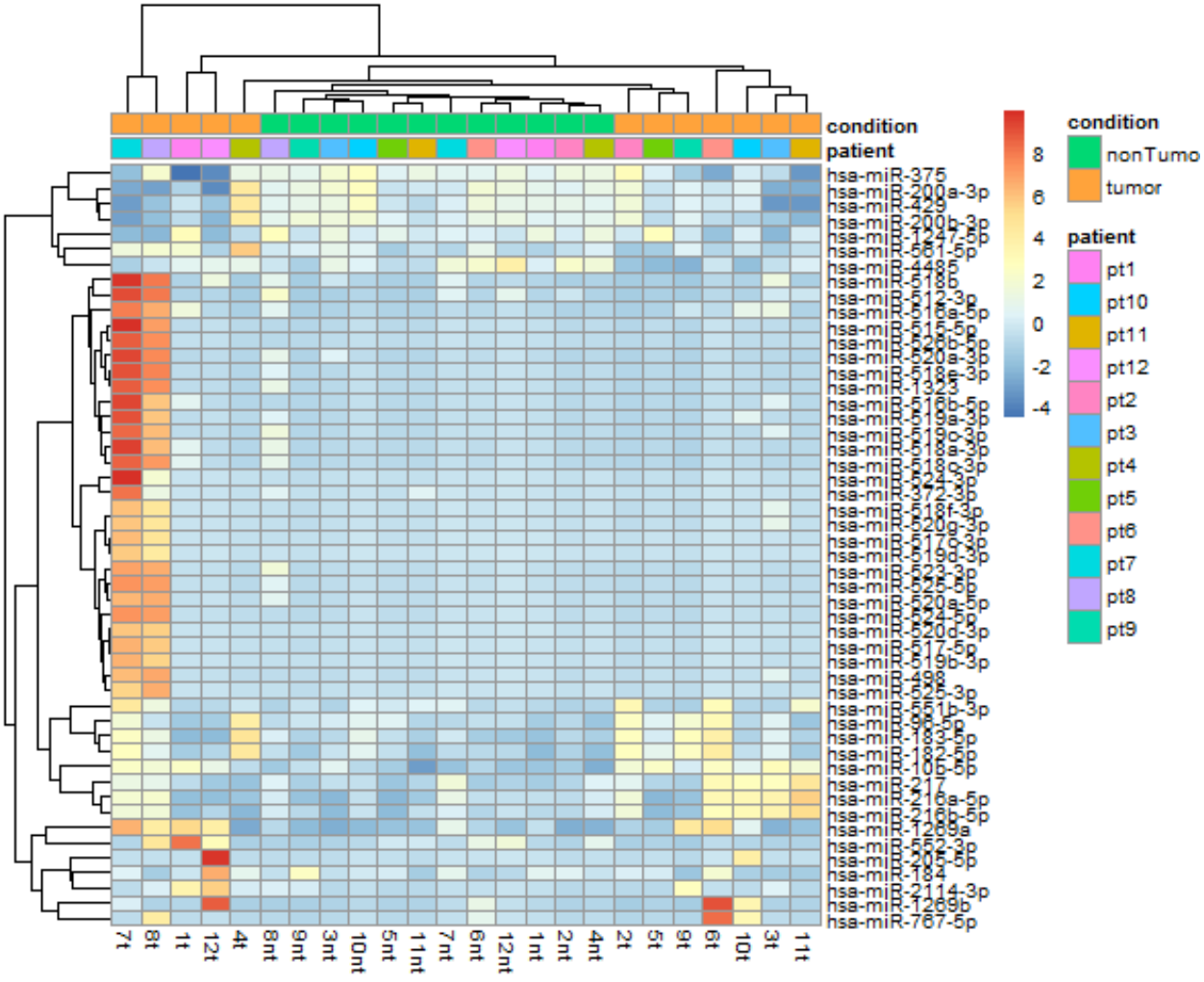


Figure 5.6 Unsupervised hierarchical clustering of miRNA expression in resected livers by 50 miRNAs with largest Euclidian distance in expression levels across all samples. Data was normalized applying the regularized logarithmic transformation of DESeq2.

5.2.3.2 More than one hundred miRNAs were differentially expressed in HCC tumours

Differential expression analysis was conducted as above and detected 152 differentially expressed miRNAs between groups (Figure 5.7).

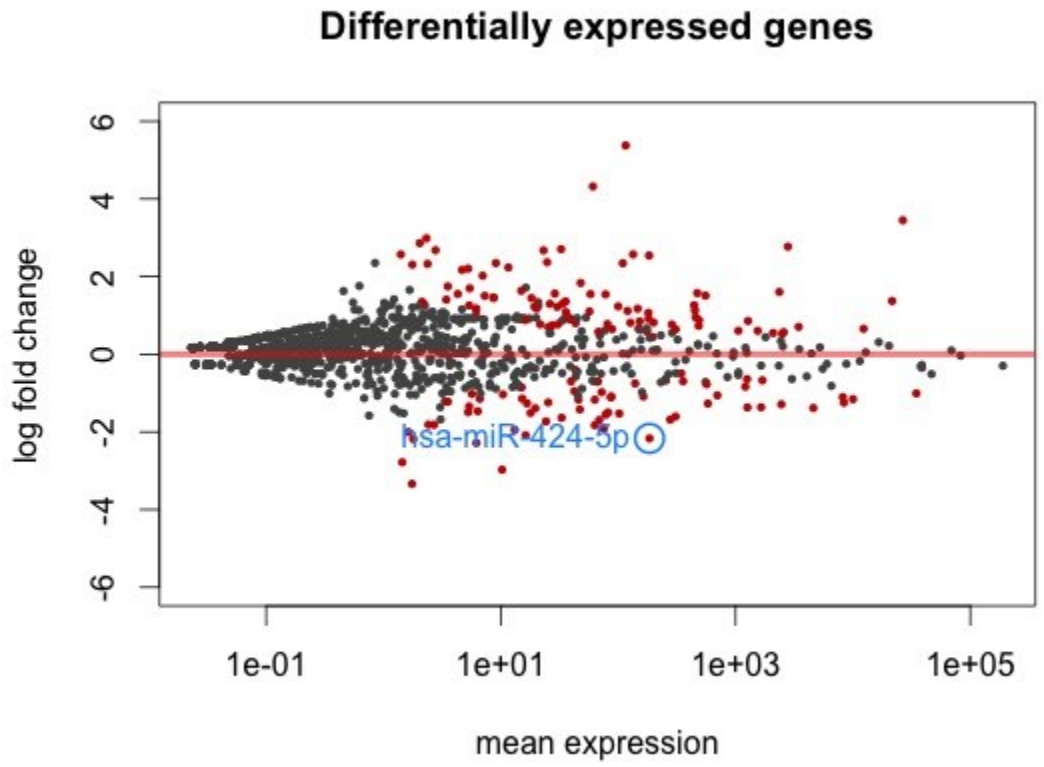


Figure 5.7 Log2 fold change and mean expression of miRNAs by edgeR differential expression analysis between tumour and non-tumour groups from resected livers. Red dots: miRNAs with p values <0.05

The leading down and upregulated miRNAs in tumours detected by either edgeR or DESeq2 program are shown in Table 5.6. Among them, miR-375 and the miR-200 family were the most downregulated ones, followed by miR-214, miR-199, miR-424, miR-139 and miR-144. Among the upregulated miRNAs, miR-1269a stood out as the significantly upregulated miRNA with the largest $\log_2$ fold change (6.24 = ~76 fold change) and lowest FDR. Again, some miRNAs in cluster C19MC, including miR-518b and miR-512, were also found amid the most upregulated miRNAs. This result is likely influenced by their high expression in 7t and 8t (Figure 5.6), but is supported by previous reports in the literature (Toffanin et al., 2011). Other miRNAs reported as differentially expressed in tumours from the explanted livers, including miR-125a, miR-483, miR-310b, miR-

21-3p, miR-210, miR-301a and miR-941, were also identified as differentially expressed in tumours from resected livers (data not shown).

Table 5.6 Deregulated miRNAs expressed in tumour tissues from resected livers

miRNA	Expression in tumour group	Log2 FC	FDR	Detected by
hsa-miR-375	down	-2.82	4.13E-04	edgeR and DESeq2
hsa-miR-200a-3p	down	-2.45	2.55E-03	edgeR
hsa-miR-214-3p	down	-2.29	2.93E-05	edgeR and DESeq2
hsa-miR-199a-5p	down	-2.21	5.83E-04	edgeR and DESeq2
hsa-miR-200b-3p	down	-2.20	1.78E-03	edgeR
hsa-miR-424-5p	down	-2.19	8.46E-12	edgeR and DESeq2
hsa-miR-139-5p	down	-2.17	4.67E-06	edgeR and DESeq2
hsa-miR-429	down	-2.16	5.55E-03	edgeR
hsa-miR-144-3p	down	-2.08	1.32E-03	edgeR and DESeq2
hsa-miR-1269a	up	6.24	5.77E-23	edgeR and DESeq2
hsa-miR-518b	up	4.32	2.47E-07	DESeq2
hsa-miR-10b-5p	up	3.69	5.48E-16	edgeR and DESeq2
hsa-miR-182-5p	up	2.42	9.33E-04	edgeR and DESeq2
hsa-miR-216a-5p	up	2.29	3.82E-02	edgeR and DESeq2
hsa-miR-512-3p	up	2.57	9.24E-03	DESeq2
hsa-miR-183-5p	up	2.03	1.18E-02	edgeR and DESeq2
hsa-miR-452-5p	up	2.46	6.11E-06	edgeR
hsa-miR-224-5p	up	2.39	1.05E-05	edgeR and DESeq2
hsa-miR-216b-5p	up	2.32	2.58E-02	edgeR

5.2.3.3 Pathway analysis of down-regulated miRNAs

Since miRNAs are generally recognized as negative regulators of gene expression, it is more reasonable to input down-regulated miRNAs into pathway prediction programs, to explore the possible effect of the downregulation of these miRNAs on the activation of cancer-related pathways. Thus, we used DIANA mirPath to determine the pathways most likely to be activated in response to the downregulation of miRNAs: miR-375, miR-200a-3p, miR-214-3p, miR-199a-5p, miR-200b-3p, miR-424-5p, miR-139-5p, miR-429 and miR-144-3p.

Interestingly, four out of eight top pathways detected are directly cancer-related pathways, including MAPK signaling pathway, PI3K-Akt signaling pathway, ErbB signaling pathway and

Pathways in cancer (Table 5.7). Amongst them, MAPK signaling pathway ranks the first. In this pathway, 80 genes are predicted targets of the nine miRNAs we found downregulated and many genes are targets of more than one miRNAs (Figure 5.8).

Table 5.7 Predicted pathways that the nine down-regulated miRNAs affected

KEGG pathway	p-value	#genes	#miRNAs
MAPK signaling pathway	9.07E-19	90	9
Focal adhesion	2.65E-17	71	9
PI3K-Akt signaling pathway	7.78E-16	104	9
Gap junction	1.74E-15	36	8
Regulation of actin cytoskeleton	1.36E-14	74	9
Axon guidance	2.68E-12	48	9
ErbB signaling pathway	3.86E-12	33	9
Pathways in cancer	3.86E-12	105	9

KEGG pathway description: http://www.genome.jp/kegg/pathway.html

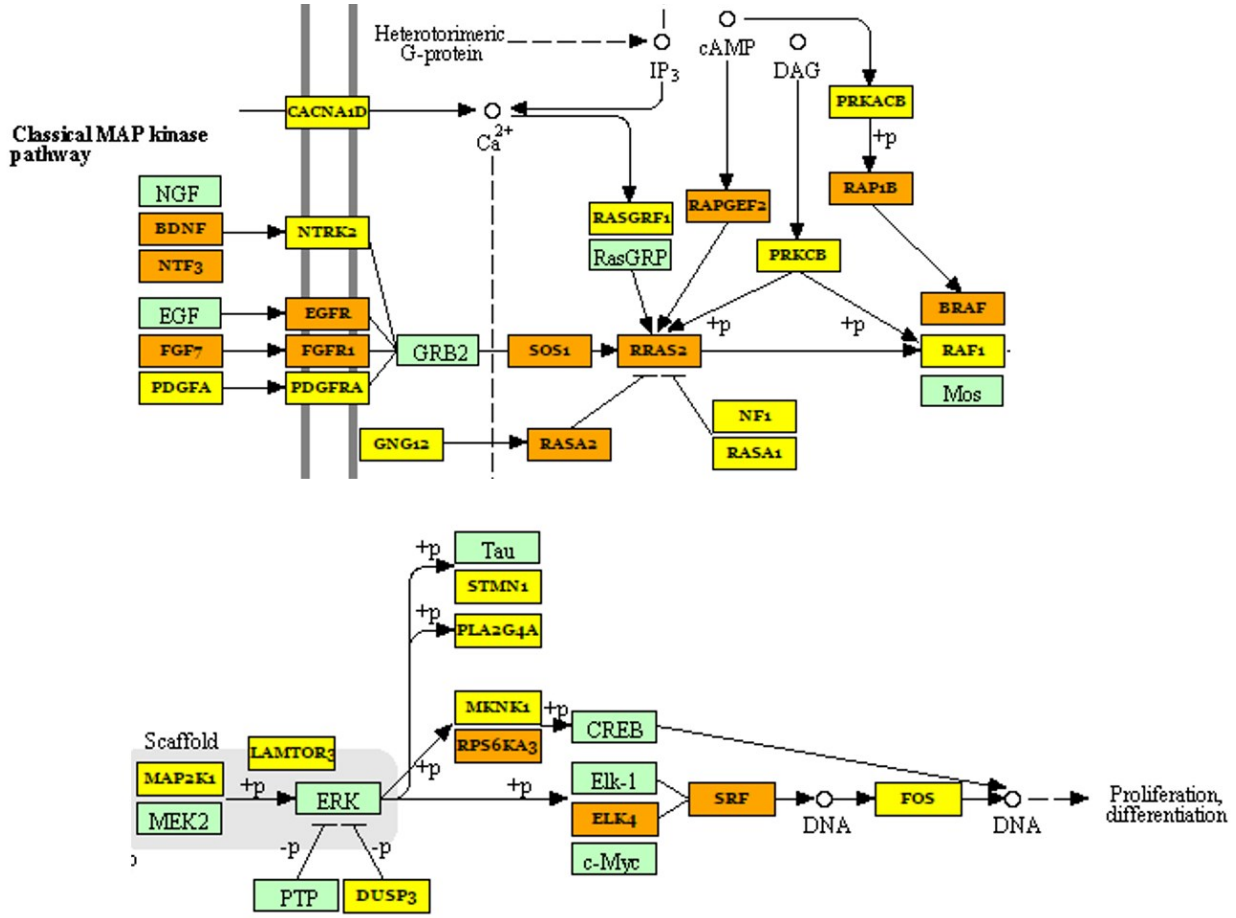


Figure 5.8 MAPK Pathway targeted by nine down-regulated miRNAs. All down regulated miRNA with a log2 fold change more than 2 were assessed using mirPath. Green box: not a target; yellow box: genes targeted by one miRNA; brown box: genes targeted by more than one miRNAs

5.3 Discussion

In this study, miRNAs and genes from paired tumour and non-tumour tissues from livers affected by HCC were profiled using next generation sequencing, with stringent quality control standards for both experimental and bioinformatic procedures. The main findings include **i**) identification of many deregulated miRNAs, including well-known oncogenic microRNAs and tumour suppressors as well as less well studied miRNAs, **ii**) hundreds of deregulated genes in tumour tissues were found, many of which have been reported associated with diverse cancers, or encode proteins with homology to those derived from oncogenes or tumour suppressors, **iii**) A number of putative miRNA-gene interactions between inversely deregulated miRNAs and genes were portrayed, **iv**) pathways related to cancer were found enriched in targets of miRNAs downregulated in tumour tissues.

5.3.1 Deregulated miRNAs identified by miRNA profiling and differential expression analysis

5.3.1.1 Down-regulated miRNAs in tumours

MiRNAs in the miR-200 family were among the most decreased miRNAs in tumours from both explanted and resected livers. This family has been reported downregulated in HCC tumours, in profiling studies using qPCR and microarrays (Jiang et al., 2008; Ladeiro et al., 2008; Murakami et al., 2006). However, results from NGS profiling studies are inconsistent regarding levels of the miR-200 family in HCC tumours. For instance, when profiling miRNA on nine pairs of tumour and non-tumour tissues (Bao et al., 2014) reported that miR-200a was found upregulated. While some other studies reported decreased level of miR-200a in HCC (Hou et al., 2011; Wojcicka et al., 2014). Our results are in line with the reported tumour-suppressive role attributed to these miRNAs in HCC (Gregory et al., 2008; Wong et al., 2015) and further confirm the importance of aberrant expression of this family in the development of HCC.

Among other top downregulated miRNA, miR-424, miR-375, miR-214-3p, miR-199a-5p and miR-139-5p have been reported repeatedly as tumour suppressors in HCC (Jiang et al., 2008; Shi et al., 2015; Wojcicka et al., 2014; Yang et al., 2014a). MiR-144-3p has been reported downregulated in HCC tumours in a single study conducted by qPCR, and was proposed to suppress cell proliferation and metastasis by inhibiting E2F3 (Cao et al., 2014). There has been few report in some less abundant miRNAs, such as miR-483-3p, miR-4485 and hsa-miR-654-3p.

5.3.1.2 up-regulated miRNAs in tumours

Among upregulated miRNAs in tumours, miR-1269a was also detected in both cohorts with high fold change and low false discovery rate value. The expression of miR1269a was increased 58.3% tumour samples (7 out of 12), which confirmed our previous finding in explanted livers and suggested that up-regulation of miR-1269a might play an important role in HCC tumourigenesis, at least in a subgroup of cancers. Consistent with that, miR-1269 was found upregulated in HCC in several studies, including NGS profiling (Wojcicka et al., 2014), microarray-based high-throughput assessment (Yang et al., 2014b) and qPCR measurements (Gan et al., 2015). In the last study, the level of miR-1269 was found positively correlated to tumour invasion and metastasis. In colorectal cancer (CRC), miR-1269 overexpression was found associated with CRC relapse and metastasis due to a positive feedback with TGF-β (Bu et al., 2015).

Other leading upregulated miRNAs in tumours included miR-10b, miR-182, miR-216a/b, miR-183, miR-452 and miR-224. Some of them have been reported upregulated in previous studies such

as miR-10b, miR-182, miR-183 and miR-224 (Bao et al., 2014; Wojcicka et al., 2014; Yang et al., 2014a). The rest has either been proposed as promoter of tumourigenesis and progression (Liu et al., 2015; Zheng et al., 2014) or involved in drug resistance in HCC (Xia et al., 2013) through different mechanisms. Our findings were consistent with those results thus reinforcing the roles of these miRNAs in HCC.

5.3.2 Potential interaction between deregulated miRNAs and targets

In gene expression profiling experiments by RNA-seq, we detected more than 800 deregulated genes, many of which have not been studied. Target predictions correlated some of deregulated genes with the deregulated miRNAs. From the predictions, the most possible interactions were identified. Noticeably, the miR-200a/ATP6V0D2 interaction was predicted. ATP6V0D2 is an ATPase (Smith et al., 2002) that has not been studied in the context of cancer or HCC. The miR-125a/DIRAS1 interaction is also interesting because DIRAS1 is a RAS-like protein (Kontani et al., 2002). RAS-family proteins belong to small GTPase, which are involved in cellular signal transduction (Goodsell, 1999). Although DIRAS1 has been proposed as a novel tumour suppressor in esophageal cancer in one study (Zhu et al., 2013), the role for DIRAS1 in HCC has never be addressed before. Our findings suggest a possible involvement of the genes ATP6V0D2 and DIRAS1 in HCC.

Many other genes found deregulated are interesting candidates for further studies on HCC. For instance, upregulation of DOCK3, an activator of the oncogene RAC1 (Sanz-Moreno et al., 2008); the interleukin-1 receptor-associated kinase (IRAK1), which has been reported to promote tumour growth, metastasis and chemo-resistance (Jain et al., 2014); the transferrin receptor TFRC, which is a marker of several cancers (Daniels et al., 2012) and the matrix metallopeptidase MMT9, with the potential to mediate invasion, metastasis and angiogenesis (Farina and Mackay, 2014), among others (Table 5.5) warrant further investigation in the context of their interaction with downregulated miRNAs. Conversely, tumour suppressor genes like ANGPTL1, an inhibitor of the epithelial-to-mesenchymal transition promoting transcription factor Slug (Kuo et al., 2013); the kinase-independent growth suppressor WNK2 (Hong et al., 2007); and the cell adhesion factor desmoglein1, DSG1, which is a cadherin found downregulated in several cancers (Myklebust et al., 2012; Xin et al., 2014), might form part of the consortium of proteins likely downregulated during HCC progression. The fact that they are putative targets of the identified deregulated miRNA in our analyses, suggests a possible interaction between them (Figure 5.4).

In addition to studying the individual impact of downregulated miRNAs in HCC, it is also useful to see how the combination of these miRNAs affects tumourigenesis. Pathway prediction analysis for the top downregulated miRNAs identified five cancer-related pathways in which MAPD ranked top. The MAPK (mitogen-activated protein kinases) signalling pathway, also called Ras-Raf-MEK-ERK pathway, has been identified previously as a key molecular pathway in tumour development and progression of HCC (Llovet and Bruix, 2008). The downregulation of the inhibitors of the genes in this pathway has been reported implicated in the aberrantly activation of this pathway (Aravalli et al., 2013). However, few studies have focused on the role of down-related tumour-suppressive miRNAs in the activation of MAPK pathway in HCC.

5.3.3 Heterogeneity of HCC tumours

HCC is known as a complex and histologically variable disease and several reports suggest that HCC are highly heterogeneous tumours from a molecular perspective (Hoshida et al., 2010; Zucman-Rossi and Laurent-Puig, 2007). In our study, exploratory multi-dimensional scaling analyses based on expression profiles of miRNAs and genes in all samples positioned non-tumour samples in close proximity to each other, and clearly apart from tumour samples. These latter exhibited a more scattered distribution, for both explanted and resected liver tissues.

One of the noticeable subpopulations included two resected livers with up-regulation of miRNAs in the C19MC cluster, the largest human miRNA cluster located on chr19q13.42 (Bentwich et al., 2005). Recently, a miRNA-based classification in HCV-related HCC tumours proposed three main clusters, with one of them composed by three sub-clusters. Interestingly, deregulation of miRNAs in the C19CM cluster was the hallmark in one of such sub-clusters (Toffanin et al., 2011). Thus, our study provides additional evidence to the existence of a subtype of HCC mainly characterized by upregulation of the C19CM. The rest of tumour samples were rather scattered along the two dimensions of the MDS plot, suggesting high tumour heterogeneity.

5.3.4 Conclusions and significance

Reports on the deregulation of specific miRNAs and their function in HCC have been inconsistent, for studies conducted using NGS or platforms that are more traditional. Our study is the first one in HCC that integrates miRNA and gene expression profiling using NGS and bioinformatic approaches. We were able to identify a set of deregulated miRNAs, including the miR-200 family, miR-1269a and miRNAs in the CM19C cluster, among many others. We also identified deregulated genes such as ATP6V0D2 and DIRAS1 and their putative interaction with deregulated miRNAs. One other finding was the contrasting miRNA and gene expression patterns

in non-tumour (tightly clustered) and tumour (divergent among them) groups. Thus, our study enabled the identification of putative subgroups of HCC tumours, in agreement with previous studies. For instance, upregulation of miRNAs in the cluster C19MC and the miR-1269a seem to be hallmarks of different subtypes of HCC.

In general, our results suggest that HCC is a collection of diseases with distinct gene expression signatures, but may share common mechanisms of tumourigenesis influenced by deregulated miRNAs and their target genes. Our findings enrich the current knowledge of HCC tumourigenesis and provide key clues for future research.

Reference

Aravalli, R.N., Cressman, E.N., Steer, C.J., 2013. Cellular and molecular mechanisms of hepatocellular carcinoma: an update. Arch Toxicol 87, 227-247.

Aravalli, R.N., Steer, C.J., Cressman, E.N., 2008. Molecular mechanisms of hepatocellular carcinoma. Hepatology 48, 2047-2063.

Augello, C., Vaira, V., Caruso, L., Destro, A., Maggioni, M., Park, Y.N., Montorsi, M., Santambrogio, R., Roncalli, M., Bosari, S., 2012. MicroRNA profiling of hepatocarcinogenesis identifies C19MC cluster as a novel prognostic biomarker in hepatocellular carcinoma. Liver Int 32, 772-782.

Bao, L., Zhao, J., Dai, X., Wang, Y., Ma, R., Su, Y., Cui, H., Niu, J., Bai, S., Xiao, Z., Yuan, H., Yang, Z., Li, C., Cheng, R., Ren, X., 2014. Correlation between miR-23a and onset of hepatocellular carcinoma. Clin Res Hepatol Gastroenterol 38, 318-330.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., Bentwich, Z., 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nature genetics 37, 766-770.

Berry, D.A., Herbst, R.S., Rubin, E.H., 2012. Reports from the 2010 Clinical and Translational Cancer Research Think Tank meeting: design strategies for personalized therapy trials. Clin Cancer Res 18, 638-644.

Beyoglu, D., Imbeaud, S., Maurhofer, O., Bioulac-Sage, P., Zucman-Rossi, J., Dufour, J.F., Idle, J.R., 2013. Tissue metabolomics of hepatocellular carcinoma: tumor energy metabolism and the role of transcriptomic classification. Hepatology 58, 229-238.

Bu, P., Wang, L., Chen, K.Y., Rakhilin, N., Sun, J., Closa, A., Tung, K.L., King, S., Kristine Varanko, A., Xu, Y., Huan Chen, J., Zessin, A.S., Shealy, J., Cummings, B., Hsu, D., Lipkin, S.M., Moreno, V., Gumus, Z.H., Shen, X., 2015. miR-1269 promotes metastasis and forms a positive feedback loop with TGF-beta. Nat Commun 6, 6879.

Budhu, A., Jia, H.L., Forgues, M., Liu, C.G., Goldstein, D., Lam, A., Zanetti, K.A., Ye, Q.H., Qin, L.X., Croce, C.M., Tang, Z.Y., Wang, X.W., 2008. Identification of metastasis-related microRNAs in hepatocellular carcinoma. Hepatology 47, 897-907.

Callegari, E., Elamin, B.K., Giannone, F., Milazzo, M., Altavilla, G., Fornari, F., Giacomelli, L., D'Abundo, L., Ferracin, M., Bassi, C., Zagatti, B., Corra, F., Miotto, E., Lupini, L., Bolondi, L., Gramantieri, L., Croce, C.M., Sabbioni, S., Negrini, M., 2012. Liver tumorigenicity promoted by microRNA-221 in a mouse transgenic model. Hepatology 56, 1025-1033.

Cao, T., Li, H., Hu, Y., Ma, D., Cai, X., 2014. miR-144 suppresses the proliferation and metastasis of hepatocellular carcinoma by targeting E2F3. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 35, 10759-10764.

Catto, J.W., Alcaraz, A., Bjartell, A.S., De Vere White, R., Evans, C.P., Fussel, S., Hamdy, F.C., Kallioniemi, O., Mengual, L., Schlomm, T., Visakorpi, T., 2011. MicroRNA in prostate, bladder, and kidney cancer: a systematic review. European urology 59, 671-681.

Chen, P.J., Yeh, S.H., Liu, W.H., Lin, C.C., Huang, H.C., Chen, C.L., Chen, D.S., Chen, P.J., 2012. Androgen pathway stimulates microRNA-216a transcription to suppress the tumor suppressor in lung cancer-1 gene in early hepatocarcinogenesis. Hepatology 56, 632-643.

Colombino, M., Sperlongano, P., Izzo, F., Tatangelo, F., Botti, G., Lombardi, A., Accardo, M., Tarantino, L., Sordelli, I., Agresti, M., Abbruzzese, A., Caraglia, M., Palmieri, G., 2012. BRAF and PIK3CA genes are somatically mutated in hepatocellular carcinoma among patients from South Italy. Cell death & disease 3, e259.

Connolly, E.C., Van Doorslaer, K., Rogler, L.E., Rogler, C.E., 2010. Overexpression of miR-21 promotes an in vitro metastatic phenotype by targeting the tumor suppressor RHOB. Molecular cancer research : MCR 8, 691-700.

Coulouarn, C., Factor, V.M., Andersen, J.B., Durkin, M.E., Thorgeirsson, S.S., 2009. Loss of miR-122 expression in liver cancer correlates with suppression of the hepatic phenotype and gain of metastatic properties. Oncogene 28, 3526-3536.

Daniels, T.R., Bernabeu, E., Rodriguez, J.A., Patel, S., Kozman, M., Chiappetta, D.A., Holler, E., Ljubimova, J.Y., Helguera, G., Penichet, M.L., 2012. The transferrin receptor and the targeted delivery of therapeutic agents against cancer. Biochim Biophys Acta 1820, 291-317.

De, P., Dryer, D., Otterstatter, M.C., Semenciw, R., 2013. Canadian trends in liver cancer: a brief clinical and epidemiologic overview. Curr Oncol 20, e40-43.

Dhayat, S.A., Mardin, W.A., Kohler, G., Bahde, R., Vowinkel, T., Wolters, H., Senninger, N., Haier, J., Mees, S.T., 2014. The microRNA-200 family-A potential diagnostic marker in hepatocellular carcinoma? Journal of surgical oncology 110, 430-438.

Di Leva, G., Garofalo, M., Croce, C.M., 2014. MicroRNAs in cancer. Annual review of pathology 9, 287-314.

El-Serag, H.B., 2011. Hepatocellular carcinoma. N Engl J Med 365, 1118-1127.

El-Serag, H.B., Rudolph, K.L., 2007. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 132, 2557-2576.

Farazi, P.A., DePinho, R.A., 2006. Hepatocellular carcinoma pathogenesis: from genes to environment. Nature reviews. Cancer 6, 674-687.

Farina, A.R., Mackay, A.R., 2014. Gelatinase B/MMP-9 in Tumour Pathogenesis and Progression. Cancers (Basel) 6, 240-296.

Gan, T.Q., Tang, R.X., He, R.Q., Dang, Y.W., Xie, Y., Chen, G., 2015. Upregulated MiR-1269 in hepatocellular carcinoma and its clinical significance. Int J Clin Exp Med 8, 714-721.

Giordano, S., Columbano, A., 2013. MicroRNAs: new tools for diagnosis, prognosis, and therapy in hepatocellular carcinoma? Hepatology 57, 840-847.

Goodsell, D.S., 1999. The molecular perspective: the ras oncogene. Oncologist 4, 263-264.

Gregory, P.A., Bert, A.G., Paterson, E.L., Barry, S.C., Tsykin, A., Farshid, G., Vadas, M.A., Khew-Goodall, Y., Goodall, G.J., 2008. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nature cell biology 10, 593-601.

Hagiwara, A., Cornu, M., Cybulski, N., Polak, P., Betz, C., Trapani, F., Terracciano, L., Heim, M.H., Ruegg, M.A., Hall, M.N., 2012. Hepatic mTORC2 activates glycolysis and lipogenesis through Akt, glucokinase, and SREBP1c. Cell Metab 15, 725-738.

Hatziapostolou, M., Polytarchou, C., Aggelidou, E., Drakaki, A., Poultsides, G.A., Jaeger, S.A., Ogata, H., Karin, M., Struhl, K., Hadzopoulou-Cladaras, M., Iliopoulos, D., 2011. An HNF4alpha-miRNA inflammatory feedback circuit regulates hepatocellular oncogenesis. Cell 147, 1233-1247.

Hayes, J., Peruzzi, P.P., Lawler, S., 2014. MicroRNAs in cancer: biomarkers, functions and therapy. Trends in molecular medicine 20, 460-469.

Hong, C., Moorefield, K.S., Jun, P., Aldape, K.D., Kharbanda, S., Phillips, H.S., Costello, J.F., 2007. Epigenome scans and cancer genome sequencing converge on WNK2, a kinase-independent suppressor of cell growth. Proceedings of the National Academy of Sciences of the United States of America 104, 10974-10979.

Hoshida, Y., Toffanin, S., Lachenmayer, A., Villanueva, A., Minguez, B., Llovet, J.M., 2010. Molecular classification and novel targets in hepatocellular carcinoma: recent advancements. Semin Liver Dis 30, 35-51.

Hou, J., Lin, L., Zhou, W., Wang, Z., Ding, G., Dong, Q., Qin, L., Wu, X., Zheng, Y., Yang, Y., Tian, W., Zhang, Q., Wang, C., Zhang, Q., Zhuang, S.M., Zheng, L., Liang, A., Tao, W., Cao, X., 2011. Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. Cancer cell 19, 232-243.

Hung, C.H., Chiu, Y.C., Chen, C.H., Hu, T.H., 2014. MicroRNAs in hepatocellular carcinoma: carcinogenesis, progression, and therapeutic target. Biomed Res Int 2014, 486407.

Hung, C.S., Liu, H.H., Liu, J.J., Yeh, C.T., Chang, T.C., Wu, C.H., Ho, Y.S., Wei, P.L., Chang, Y.J., 2013. MicroRNA-200a and -200b mediated hepatocellular carcinoma cell migration through the epithelial to mesenchymal transition markers. Annals of surgical oncology 20 Suppl 3, S360-368.

Jablkowski, M., Bocian, A., Bialkowska, J., Bartkowiak, J., 2005. A comparative study of P53/MDM2 genes alterations and P53/MDM2 proteins immunoreactivity in liver cirrhosis and hepatocellular carcinoma. J Exp Clin Cancer Res 24, 117-125.

Jain, A., Kaczanowska, S., Davila, E., 2014. IL-1 Receptor-Associated Kinase Signaling and Its Role in Inflammation, Cancer Progression, and Therapy Resistance. Front Immunol 5, 553.

Jiang, J., Gusev, Y., Aderca, I., Mettler, T.A., Nagorney, D.M., Brackett, D.J., Roberts, L.R., Schmittgen, T.D., 2008. Association of MicroRNA expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival. Clin Cancer Res 14, 419-427.

Kang, S.M., Lee, H.J., 2014. MicroRNAs in human lung cancer. Experimental biology and medicine.

Karakatsanis, A., Papaconstantinou, I., Gazouli, M., Lyberopoulou, A., Polymeneas, G., Voros, D., 2013. Expression of microRNAs, miR-21, miR-31, miR-122, miR-145, miR-146a, miR-200c, miR-221, miR-222, and miR-223 in patients with hepatocellular carcinoma or intrahepatic cholangiocarcinoma and its prognostic significance. Mol Carcinog 52, 297-303.

Kim, K.W., Bae, S.K., Lee, O.H., Bae, M.H., Lee, M.J., Park, B.C., 1998. Insulin-like growth factor II induced by hypoxia may contribute to angiogenesis of human hepatocellular carcinoma. Cancer Res 58, 348-351.

Kontani, K., Tada, M., Ogawa, T., Okai, T., Saito, K., Araki, Y., Katada, T., 2002. Di-Ras, a distinct subgroup of ras family GTPases with unique biochemical properties. J Biol Chem 277, 41070-41078.

Kuo, T.C., Tan, C.T., Chang, Y.W., Hong, C.C., Lee, W.J., Chen, M.W., Jeng, Y.M., Chiou, J., Yu, P., Chen, P.S., Wang, M.Y., Hsiao, M., Su, J.L., Kuo, M.L., 2013. Angiopoietin-like protein 1 suppresses SLUG to inhibit cancer cell motility. The Journal of clinical investigation 123, 1082-1095.

Ladeiro, Y., Couchy, G., Balabaud, C., Bioulac-Sage, P., Pelletier, L., Rebouissou, S., Zucman-Rossi, J., 2008. MicroRNA profiling in hepatocellular tumors is associated with clinical features and oncogene/tumor suppressor gene mutations. Hepatology 47, 1955-1963.

Lee, J.M., Heo, M.J., Lee, C.G., Yang, Y.M., Kim, S.G., 2015. Increase of miR-199a-5p by protoporphyrin IX, a photocatalyzer, directly inhibits E2F3, sensitizing mesenchymal tumor cells to anti-cancer agents. Oncotarget 6, 3918-3931.

Levine, A.J., 1997. p53, the cellular gatekeeper for growth and division. Cell 88, 323-331.
Li, Q.J., Zhou, L., Yang, F., Wang, G.X., Zheng, H., Wang, D.S., He, Y., Dou, K.F., 2012. MicroRNA-10b promotes migration and invasion through CADM1 in human hepatocellular carcinoma cells. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 33, 1455-1465.
Llovet, J.M., Bruix, J., 2008. Molecular targeted therapies in hepatocellular carcinoma. Hepatology 48, 1312-1327.
Lujambio, A., Lowe, S.W., 2012. The microcosmos of cancer. Nature 482, 347-355.
Marquardt, J.U., Galle, P.R., Teufel, A., 2012. Molecular diagnosis and therapy of hepatocellular carcinoma (HCC): an emerging field for advanced technologies. J Hepatol 56, 267-275.
Meng, F., Henson, R., Wehbe-Janek, H., Ghoshal, K., Jacob, S.T., Patel, T., 2007. MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. Gastroenterology 133, 647-658.
Mise, M., Arii, S., Higashituji, H., Furutani, M., Niwano, M., Harada, T., Ishigami, S., Toda, Y., Nakayama, H., Fukumoto, M., Fujita, J., Imamura, M., 1996. Clinical significance of vascular endothelial growth factor and basic fibroblast growth factor gene expression in liver tumor. Hepatology 23, 455-464.
Morishita, A., Masaki, T., 2014. miRNA in hepatocellular carcinoma. Hepatology research : the official journal of the Japan Society of Hepatology.
Murakami, Y., Yasuda, T., Saigo, K., Urashima, T., Toyoda, H., Okanoue, T., Shimotohno, K., 2006. Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. Oncogene 25, 2537-2545.
Myklebust, M.P., Fluge, O., Immervoll, H., Skarstein, A., Balteskard, L., Bruland, O., Dahl, O., 2012. Expression of DSG1 and DSC1 are prognostic markers in anal carcinoma patients. Br J Cancer 106, 756-762.
Park, J.K., Kogure, T., Nuovo, G.J., Jiang, J., He, L., Kim, J.H., Phelps, M.A., Papenfuss, T.L., Croce, C.M., Patel, T., Schmittgen, T.D., 2011. miR-221 silencing blocks hepatocellular carcinoma and promotes survival. Cancer Res 71, 7608-7616.
Pocobelli, G., Cook, L.S., Brant, R., Lee, S.S., 2008. Hepatocellular carcinoma incidence trends in Canada: analysis by birth cohort and period of diagnosis. Liver Int 28, 1272-1279.
Qu, K.Z., Zhang, K., Li, H., Afdhal, N.H., Albitar, M., 2011. Circulating microRNAs as biomarkers for hepatocellular carcinoma. J Clin Gastroenterol 45, 355-360.
Sanz-Moreno, V., Gadea, G., Ahn, J., Paterson, H., Marra, P., Pinner, S., Sahai, E., Marshall, C.J., 2008. Rac activation and inactivation control plasticity of tumor cell movement. Cell 135, 510-523.
Schetter, A.J., Okayama, H., Harris, C.C., 2012. The role of microRNAs in colorectal cancer. Cancer journal 18, 244-252.
Shen, Q., Cicinnati, V.R., Zhang, X., Iacob, S., Weber, F., Sotiropoulos, G.C., Radtke, A., Lu, M., Paul, A., Gerken, G., Beckebaum, S., 2010. Role of microRNA-199a-5p and discoidin domain receptor 1 in human hepatocellular carcinoma invasion. Mol Cancer 9, 227.
Shi, K.Q., Lin, Z., Chen, X.J., Song, M., Wang, Y.Q., Cai, Y.J., Yang, N.B., Zheng, M.H., Dong, J.Z., Zhang, L., Chen, Y.P., 2015. Hepatocellular carcinoma associated microRNA expression signature: integrated bioinformatics analysis, experimental validation and clinical significance. Oncotarget 6, 25093-25108.
Singal, A., Volk, M.L., Waljee, A., Salgia, R., Higgins, P., Rogers, M.A., Marrero, J.A., 2009. Meta-analysis: surveillance with ultrasound for early-stage hepatocellular carcinoma in patients with cirrhosis. Aliment Pharmacol Ther 30, 37-47.

Smith, A.N., Borthwick, K.J., Karet, F.E., 2002. Molecular cloning and characterization of novel tissue-specific isoforms of the human vacuolar H(+)-ATPase C, G and d subunits, and their evaluation in autosomal recessive distal renal tubular acidosis. Gene 297, 169-177.

Song, S., Ajani, J.A., 2013. The role of microRNAs in cancers of the upper gastrointestinal tract. Nature reviews. Gastroenterology & hepatology 10, 109-118.

Thorgeirsson, S.S., Grisham, J.W., 2002. Molecular pathogenesis of human hepatocellular carcinoma. Nature genetics 31, 339-346.

Toffanin, S., Hoshida, Y., Lachenmayer, A., Villanueva, A., Cabellos, L., Minguez, B., Savic, R., Ward, S.C., Thung, S., Chiang, D.Y., Alsinet, C., Tovar, V., Roayaie, S., Schwartz, M., Bruix, J., Waxman, S., Friedman, S.L., Golub, T., Mazzaferro, V., Llovet, J.M., 2011. MicroRNA-based classification of hepatocellular carcinoma and oncogenic role of miR-517a. Gastroenterology 140, 1618-1628 e1616.

Tomimaru, Y., Eguchi, H., Nagano, H., Wada, H., Kobayashi, S., Marubashi, S., Tanemura, M., Tomokuni, A., Takemasa, I., Umeshita, K., Kanto, T., Doki, Y., Mori, M., 2012. Circulating microRNA-21 as a novel biomarker for hepatocellular carcinoma. J Hepatol 56, 167-175.

Trevisani, F., D'Intino, P.E., Morselli-Labate, A.M., Mazzella, G., Accogli, E., Caraceni, P., Domenicali, M., De Notariis, S., Roda, E., Bernardi, M., 2001. Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of HBsAg and anti-HCV status. J Hepatol 34, 570-575.

Venook, A.P., Papandreou, C., Furuse, J., de Guevara, L.L., 2010. The incidence and epidemiology of hepatocellular carcinoma: a global and regional perspective. Oncologist 15 Suppl 4, 5-13.

Villanueva, A., Newell, P., Chiang, D.Y., Friedman, S.L., Llovet, J.M., 2007. Genomics and signaling pathways in hepatocellular carcinoma. Semin Liver Dis 27, 55-76.

Wang, J., Li, J., Shen, J., Wang, C., Yang, L., Zhang, X., 2012. MicroRNA-182 downregulates metastasis suppressor 1 and contributes to metastasis of hepatocellular carcinoma. BMC Cancer 12, 227.

Wei, Y., Van Nhieu, J.T., Prigent, S., Srivatanakul, P., Tiollais, P., Buendia, M.A., 2002. Altered expression of E-cadherin in hepatocellular carcinoma: correlations with genetic alterations, beta-catenin expression, and clinical features. Hepatology 36, 692-701.

WHO, 2014. world health organization cancer fact sheet.

Wojcicka, A., Swierniak, M., Kornasiewicz, O., Gierlikowski, W., Maciag, M., Kolanowska, M., Kotlarek, M., Gornicka, B., Koperski, L., Niewinski, G., Krawczyk, M., Jazdzewski, K., 2014. Next generation sequencing reveals microRNA isoforms in liver cirrhosis and hepatocellular carcinoma. Int J Biochem Cell Biol 53, 208-217.

Wong, C.C., Wong, C.M., Tung, E.K., Au, S.L., Lee, J.M., Poon, R.T., Man, K., Ng, I.O., 2011. The microRNA miR-139 suppresses metastasis and progression of hepatocellular carcinoma by down-regulating Rho-kinase 2. Gastroenterology 140, 322-331.

Wong, C.M., Wei, L., Au, S.L., Fan, D.N., Zhou, Y., Tsang, F.H., Law, C.T., Lee, J.M., He, X., Shi, J., Wong, C.C., Ng, I.O., 2015. MiR-200b/200c/429 subfamily negatively regulates Rho/ROCK signaling pathway to suppress hepatocellular carcinoma metastasis. Oncotarget 6, 13658-13670.

Xin, Z., Yamaguchi, A., Sakamoto, K., 2014. Aberrant expression and altered cellular localization of desmosomal and hemidesmosomal proteins are associated with aggressive clinicopathological features of oral squamous cell carcinoma. Virchows Arch 465, 35-47.

Yang, J., Han, S., Huang, W., Chen, T., Liu, Y., Pan, S., Li, S., 2014a. A meta-analysis of microRNA expression in liver cancer. PloS one 9, e114533.

Yang, X.W., Shen, G.Z., Cao, L.Q., Jiang, X.F., Peng, H.P., Shen, G., Chen, D., Xue, P., 2014b. MicroRNA-1269 promotes proliferation in human hepatocellular carcinoma via downregulation of FOXO1. BMC Cancer 14, 909.

Yao, Y.J., Ping, X.L., Zhang, H., Chen, F.F., Lee, P.K., Ahsan, H., Chen, C.J., Lee, P.H., Peacocke, M., Santella, R.M., Tsou, H.C., 1999. PTEN/MMAC1 mutations in hepatocellular carcinomas. Oncogene 18, 3181-3185.

Yuan, J.H., Yang, F., Chen, B.F., Lu, Z., Huo, X.S., Zhou, W.P., Wang, F., Sun, S.H., 2011. The histone deacetylase 4/SP1/microrna-200a regulatory network contributes to aberrant histone acetylation in hepatocellular carcinoma. Hepatology 54, 2025-2035.

Yuneva, M.O., Fan, T.W., Allen, T.D., Higashi, R.M., Ferraris, D.V., Tsukamoto, T., Mates, J.M., Alonso, F.J., Wang, C., Seo, Y., Chen, X., Bishop, J.M., 2012. The metabolic profile of tumors depends on both the responsible genetic lesion and tissue type. Cell Metab 15, 157-170.

Zhang, L., Yang, F., Yuan, J.H., Yuan, S.X., Zhou, W.P., Huo, X.S., Xu, D., Bi, H.S., Wang, F., Sun, S.H., 2013a. Epigenetic activation of the MiR-200 family contributes to H19-mediated metastasis suppression in hepatocellular carcinoma. Carcinogenesis 34, 577-586.

Zhang, W., Liu, J., Wang, G., 2014. The role of microRNAs in human breast cancer progression. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 35, 6235-6244.

Zhang, Y., Takahashi, S., Tasaka, A., Yoshima, T., Ochi, H., Chayama, K., 2013b. Involvement of microRNA-224 in cell proliferation, migration, invasion, and anti-apoptosis in hepatocellular carcinoma. J Gastroenterol Hepatol 28, 565-575.

Zhou, J., Yu, L., Gao, X., Hu, J., Wang, J., Dai, Z., Wang, J.F., Zhang, Z., Lu, S., Huang, X., Wang, Z., Qiu, S., Wang, X., Yang, G., Sun, H., Tang, Z., Wu, Y., Zhu, H., Fan, J., 2011. Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma. J Clin Oncol 29, 4781-4788.

Zhu, Q., Wang, Z., Hu, Y., Li, J., Li, X., Zhou, L., Huang, Y., 2012. miR-21 promotes migration and invasion by the miR-21-PDCD4-AP-1 feedback loop in human hepatocellular carcinoma. Oncol Rep 27, 1660-1668.

Zhu, Y.H., Fu, L., Chen, L., Qin, Y.R., Liu, H., Xie, F., Zeng, T., Dong, S.S., Li, J., Li, Y., Dai, Y., Xie, D., Guan, X.Y., 2013. Downregulation of the novel tumor suppressor DIRAS1 predicts poor prognosis in esophageal squamous cell carcinoma. Cancer Res 73, 2298-2309.

Zucman-Rossi, J., Laurent-Puig, P., 2007. Genetic diversity of hepatocellular carcinomas and its potential impact on targeted therapies. Pharmacogenomics 8, 997-1003.

Bibliography

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

Agarwal, V., Bell, G.W., Nam, J.W., Bartel, D.P., 2015. Predicting effective microRNA target sites in mammalian mRNAs. Elife 4.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M.J., Tuschl, T., Margalit, H., 2005. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 33, 2697-2706.

Ameres, S.L., Zamore, P.D., 2013. Diversifying microRNA sequence and function. Nature reviews. Molecular cell biology 14, 475-488.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome biology 11, R106.

Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., Robinson, M.D., 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8, 1765-1786.

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Anders, S., Reyes, A., Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. Genome research 22, 2008-2017.

Aravalli, R.N., Cressman, E.N., Steer, C.J., 2013. Cellular and molecular mechanisms of hepatocellular carcinoma: an update. Arch Toxicol 87, 227-247.

Aravalli, R.N., Steer, C.J., Cressman, E.N., 2008. Molecular mechanisms of hepatocellular carcinoma. Hepatology 48, 2047-2063.

Augello, C., Vaira, V., Caruso, L., Destro, A., Maggioni, M., Park, Y.N., Montorsi, M., Santambrogio, R., Roncalli, M., Bosari, S., 2012. MicroRNA profiling of hepatocarcinogenesis identifies C19MC cluster as a novel prognostic biomarker in hepatocellular carcinoma. Liver Int 32, 772-782.

Bala, S., Marcos, M., Kodys, K., Csak, T., Catalano, D., Mandrekar, P., Szabo, G., 2011. Up-regulation of microRNA-155 in macrophages contributes to increased tumor necrosis factor {alpha} (TNF{alpha}) production via increased mRNA half-life in alcoholic liver disease. J Biol Chem 286, 1436-1444.

Bao, L., Zhao, J., Dai, X., Wang, Y., Ma, R., Su, Y., Cui, H., Niu, J., Bai, S., Xiao, Z., Yuan, H., Yang, Z., Li, C., Cheng, R., Ren, X., 2014. Correlation between miR-23a and onset of hepatocellular carcinoma. Clin Res Hepatol Gastroenterol 38, 318-330.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281-297.

Baskerville, S., Bartel, D.P., 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA 11, 241-247.

Benjamini Y, H.Y., 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. . J R Stat Soc B Methodol 57, 289-300.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H.,

Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E.C.M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-59.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., Bentwich, Z., 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nature genetics 37, 766-770.

Bernstein, E., Caudy, A.A., Hammond, S.M., Hannon, G.J., 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409, 363-366.

Berry, D.A., Herbst, R.S., Rubin, E.H., 2012. Reports from the 2010 Clinical and Translational Cancer Research Think Tank meeting: design strategies for personalized therapy trials. Clin Cancer Res 18, 638-644.

Beuers, U., Gershwin, M.E., Gish, R.G., Invernizzi, P., Jones, D.E., Lindor, K., Ma, X., Mackay, I.R., Pares, A., Tanaka, A., Vierling, J.M., Poupon, R., 2015. Changing nomenclature for PBC: From 'cirrhosis' to 'cholangitis'. J Hepatol.

Beyoglu, D., Imbeaud, S., Maurhofer, O., Bioulac-Sage, P., Zucman-Rossi, J., Dufour, J.F., Idle, J.R., 2013. Tissue metabolomics of hepatocellular carcinoma: tumor energy metabolism and the role of transcriptomic classification. Hepatology 58, 229-238.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M., Gascoyne, R.D., Marra, M.A., Jones, S.J., 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25, 2872-2877.

Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M., Gilad, Y., 2010. Sex-specific and lineage-specific alternative splicing in primates. Genome research 20, 180-189.

Borchert, G.M., Lanier, W., Davidson, B.L., 2006. RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol 13, 1097-1101.

Bouvier, N.M., Palese, P., 2008. The biology of influenza viruses. Vaccine 26 Suppl 4, D49-53.

Bu, P., Wang, L., Chen, K.Y., Rakhilin, N., Sun, J., Closa, A., Tung, K.L., King, S., Kristine Varanko, A., Xu, Y., Huan Chen, J., Zessin, A.S., Shealy, J., Cummings, B., Hsu, D., Lipkin, S.M., Moreno, V., Gumus, Z.H., Shen, X., 2015. miR-1269 promotes metastasis and forms a positive feedback loop with TGF-beta. Nat Commun 6, 6879.

Budhu, A., Jia, H.L., Forgues, M., Liu, C.G., Goldstein, D., Lam, A., Zanetti, K.A., Ye, Q.H., Qin, L.X., Croce, C.M., Tang, Z.Y., Wang, X.W., 2008. Identification of metastasis-related microRNAs in hepatocellular carcinoma. Hepatology 47, 897-907.

Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11, 94.

Caldwell, S., 2010. Cryptogenic cirrhosis: what are we missing? Curr Gastroenterol Rep 12, 40-48.

Callegari, E., Elamin, B.K., Giannone, F., Milazzo, M., Altavilla, G., Fornari, F., Giacomelli, L., D'Abundo, L., Ferracin, M., Bassi, C., Zagatti, B., Corra, F., Miotto, E., Lupini, L., Bolondi, L., Gramantieri, L., Croce, C.M., Sabbioni, S., Negrini, M., 2012. Liver tumorigenicity promoted by microRNA-221 in a mouse transgenic model. Hepatology 56, 1025-1033.

Cameron, C.M., Cameron, M.J., Bermejo-Martin, J.F., Ran, L., Xu, L., Turner, P.V., Ran, R., Danesh, A., Fang, Y., Chan, P.K., Mytle, N., Sullivan, T.J., Collins, T.L., Johnson, M.G., Medina, J.C., Rowe, T., Kelvin, D.J., 2008. Gene expression analysis of host innate immune responses during Lethal H5N1 infection in ferrets. Journal of virology 82, 11308-11317.

Cao, T., Li, H., Hu, Y., Ma, D., Cai, X., 2014. miR-144 suppresses the proliferation and metastasis of hepatocellular carcinoma by targeting E2F3. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 35, 10759-10764.

Carrat, F., Flahault, A., 2007. Influenza vaccine: the challenge of antigenic drift. Vaccine 25, 6852-6862.

Catto, J.W., Alcaraz, A., Bjartell, A.S., De Vere White, R., Evans, C.P., Fussel, S., Hamdy, F.C., Kallioniemi, O., Mengual, L., Schlomm, T., Visakorpi, T., 2011. MicroRNA in prostate, bladder, and kidney cancer: a systematic review. European urology 59, 671-681.

Chen, H., Sun, Y., Dong, R., Yang, S., Pan, C., Xiang, D., Miao, M., Jiao, B., 2011. Mir-34a is upregulated during liver regeneration in rats and is associated with the suppression of hepatocyte proliferation. PloS one 6, e20238.

Chen, P.J., Yeh, S.H., Liu, W.H., Lin, C.C., Huang, H.C., Chen, C.L., Chen, D.S., Chen, P.J., 2012. Androgen pathway stimulates microRNA-216a transcription to suppress the tumor suppressor in lung cancer-1 gene in early hepatocarcinogenesis. Hepatology 56, 632-643.

Chen Y, M.D., Robinson M, Smyth GK, 2015. edgeR: differential expression analysis of digital gene expression data User's Guide.

Chen, Y., Verfaillie, C.M., 2014. MicroRNAs: the fine modulators of liver development and function. Liver Int 34, 976-990.

Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., Shiekhattar, R., 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature 436, 740-744.

Cheval, J., Sauvage, V., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F., Berthet, N., Brisse, S., Moszer, I., Bourhy, H., Manuguerra, C.J., Lecuit, M., Burguiere, A., Caro, V., Eloit, M., 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. Journal of clinical microbiology 49, 3268-3275.

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H., 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nature nanotechnology 4, 265-270.

Colombino, M., Sperlongano, P., Izzo, F., Tatangelo, F., Botti, G., Lombardi, A., Accardo, M., Tarantino, L., Sordelli, I., Agresti, M., Abbruzzese, A., Caraglia, M., Palmieri, G., 2012. BRAF and PIK3CA genes are somatically mutated in hepatocellular carcinoma among patients from South Italy. Cell death & disease 3, e259.

Connolly, E.C., Van Doorslaer, K., Rogler, L.E., Rogler, C.E., 2010. Overexpression of miR-21 promotes an in vitro metastatic phenotype by targeting the tumor suppressor RHOB. Molecular cancer research : MCR 8, 691-700.

Coulouarn, C., Factor, V.M., Andersen, J.B., Durkin, M.E., Thorgeirsson, S.S., 2009. Loss of miR-122 expression in liver cancer correlates with suppression of the hepatic phenotype and gain of metastatic properties. Oncogene 28, 3526-3536.

Cullen, B.R., 2004. Transcription and processing of human microRNA precursors. Mol Cell 16, 861-865.

Danesh, A., Cameron, C.M., Leon, A.J., Ran, L., Xu, L., Fang, Y., Kelvin, A.A., Rowe, T., Chen, H., Guan, Y., Jonsson, C.B., Cameron, M.J., Kelvin, D.J., 2011. Early gene expression events in ferrets in response to SARS coronavirus infection versus direct interferon-alpha2b stimulation. Virology 409, 102-112.

Daniels, T.R., Bernabeu, E., Rodriguez, J.A., Patel, S., Kozman, M., Chiappetta, D.A., Holler, E., Ljubimova, J.Y., Helguera, G., Penichet, M.L., 2012. The transferrin receptor and the targeted delivery of therapeutic agents against cancer. Biochim Biophys Acta 1820, 291-317.

Dawood, F.S., Iuliano, A.D., Reed, C., Meltzer, M.I., Shay, D.K., Cheng, P.Y., Bandaranayake, D., Breiman, R.F., Brooks, W.A., Buchy, P., Feikin, D.R., Fowler, K.B., Gordon, A., Hien, N.T., Horby, P., Huang, Q.S., Katz, M.A., Krishnan, A., Lal, R., Montgomery, J.M., Molbak, K., Pebody, R., Presanis, A.M., Razuri, H., Steens, A., Tinoco, Y.O., Wallinga, J., Yu, H., Vong, S., Bresee, J., Widdowson, M.A., 2012. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. The Lancet. Infectious diseases 12, 687-695.

De, P., Dryer, D., Otterstatter, M.C., Semenciw, R., 2013. Canadian trends in liver cancer: a brief clinical and epidemiologic overview. Curr Oncol 20, e40-43.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., Hannon, G.J., 2004. Processing of primary microRNAs by the Microprocessor complex. Nature 432, 231-235.

Dhayat, S.A., Mardin, W.A., Kohler, G., Bahde, R., Vowinkel, T., Wolters, H., Senninger, N., Haier, J., Mees, S.T., 2014. The microRNA-200 family-A potential diagnostic marker in hepatocellular carcinoma? Journal of surgical oncology 110, 430-438.

Di Leva, G., Garofalo, M., Croce, C.M., 2014. MicroRNAs in cancer. Annual review of pathology 9, 287-314.

Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., Jaffrezic, F., French StatOmique, C., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14, 671-683.

Dippold, R.P., Vadigepalli, R., Gonye, G.E., Patra, B., Hoek, J.B., 2013. Chronic ethanol feeding alters miRNA expression dynamics during liver regeneration. Alcohol Clin Exp Res 37 Suppl 1, E59-69.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133-138.

El-Serag, H.B., 2011. Hepatocellular carcinoma. N Engl J Med 365, 1118-1127.

El-Serag, H.B., Rudolph, K.L., 2007. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 132, 2557-2576.

Esau, C., Davis, S., Murray, S.F., Yu, X.X., Pandey, S.K., Pear, M., Watts, L., Booten, S.L., Graham, M., McKay, R., Subramaniam, A., Propp, S., Lollo, B.A., Freier, S., Bennett, C.F., Bhanot, S., Monia, B.P., 2006. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. Cell Metab 3, 87-98.

Farazi, P.A., DePinho, R.A., 2006. Hepatocellular carcinoma pathogenesis: from genes to environment. Nature reviews. Cancer 6, 674-687.

Farina, A.R., Mackay, A.R., 2014. Gelatinase B/MMP-9 in Tumour Pathogenesis and Progression. Cancers (Basel) 6, 240-296.

Farooqui, A., Lei, Y., Wang, P., Huang, J., Lin, J., Li, G., Leon, A.J., Zhao, Z., Kelvin, D.J., 2011. Genetic and clinical assessment of 2009 pandemic influenza in southern China. Journal of infection in developing countries 5, 700-710.

Farooqui, A., Leon, A.J., Huang, L., Wu, S., Cai, Y., Lin, P., Chen, W., Fang, X., Zeng, T., Liu, Y., Zhang, L., Su, T., Chen, W., Ghedin, E., Zhu, H., Guan, Y., Kelvin, D.J., 2015. Genetic diversity of the 2013-14 human isolates of influenza H7N9 in China. BMC infectious diseases 15, 109.

Farrell, G.C., Larter, C.Z., 2006. Nonalcoholic fatty liver disease: from steatosis to cirrhosis. Hepatology 43, S99-S112.

Ferreira, D.M., Simao, A.L., Rodrigues, C.M., Castro, R.E., 2014. Revisiting the metabolic syndrome and paving the way for microRNAs in non-alcoholic fatty liver disease. FEBS J 281, 2503-2524.

Forman, J.J., Legesse-Miller, A., Coller, H.A., 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proceedings of the National Academy of Sciences of the United States of America 105, 14879-14884.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., Rajewsky, N., 2008. Discovering microRNAs from deep sequencing data using miRDeep. Nature biotechnology 26, 407-415.

Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W., Rajewsky, N., 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40, 37-52.

Friedman, R.C., Farh, K.K., Burge, C.B., Bartel, D.P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome research 19, 92-105.

Gan, T.Q., Tang, R.X., He, R.Q., Dang, Y.W., Xie, Y., Chen, G., 2015. Upregulated MiR-1269 in hepatocellular carcinoma and its clinical significance. Int J Clin Exp Med 8, 714-721.

Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8, 469-477.

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., Bartel, D.P., 2011. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 18, 1139-1146.

Garmire, L.X., Subramaniam, S., 2012. Evaluation of normalization methods in mammalian microRNA-Seq data. RNA 18, 1279-1288.

Garten, R.J., Davis, C.T., Russell, C.A., Shu, B., Lindstrom, S., Balish, A., Sessions, W.M., Xu, X., Skepner, E., Deyde, V., Okomo-Adhiambo, M., Gubareva, L., Barnes, J., Smith, C.B., Emery, S.L., Hillman, M.J., Rivailler, P., Smagala, J., de Graaf, M., Burke, D.F., Fouchier, R.A., Pappas, C., Alpuche-Aranda, C.M., Lopez-Gatell, H., Olivera, H., Lopez, I., Myers, C.A., Faix, D., Blair, P.J., Yu, C., Keene, K.M., Dotson, P.D., Jr., Boxrud, D., Sambol, A.R., Abid, S.H., St George, K., Bannerman, T., Moore, A.L., Stringer, D.J., Blevins, P., Demmler-Harrison, G.J., Ginsberg, M., Kriner, P., Waterman, S., Smole, S., Guevara, H.F., Belongia, E.A., Clark, P.A., Beatrice, S.T., Donis, R., Katz, J., Finelli, L., Bridges, C.B., Shaw, M., Jernigan, D.B., Uyeki, T.M., Smith, D.J., Klimov, A.I., Cox, N.J., 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325, 197-201.

Gasser, T., Hardy, J., Mizuno, Y., 2011. Milestones in PD genetics. Mov Disord 26, 1042-1048.

Ghedin, E., Laplante, J., DePasse, J., Wentworth, D.E., Santos, R.P., Lepow, M.L., Porter, J., Stellrecht, K., Lin, X., Operario, D., Griesemer, S., Fitch, A., Halpin, R.A., Stockwell, T.B., Spiro, D.J., Holmes, E.C., St George, K., 2011. Deep sequencing reveals mixed infection with 2009 pandemic influenza A

(H1N1) virus strains and the emergence of oseltamivir resistance. The Journal of infectious diseases 203, 168-174.
Giordano, S., Columbano, A., 2013. MicroRNAs: new tools for diagnosis, prognosis, and therapy in hepatocellular carcinoma? Hepatology 57, 840-847.
Goodsell, D.S., 1999. The molecular perspective: the ras oncogene. Oncologist 4, 263-264.
Gori, M., Arciello, M., Balsano, C., 2014. MicroRNAs in nonalcoholic fatty liver disease: novel biomarkers and prognostic tools during the transition from steatosis to hepatocarcinoma. Biomed Res Int 2014, 741465.
Gregory, P.A., Bert, A.G., Paterson, E.L., Barry, S.C., Tsykin, A., Farshid, G., Vadas, M.A., Khew-Goodall, Y., Goodall, G.J., 2008. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nature cell biology 10, 593-601.
Gregory, R.I., Chendrimada, T.P., Cooch, N., Shiekhattar, R., 2005. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell 123, 631-640.
Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., Shiekhattar, R., 2004. The Microprocessor complex mediates the genesis of microRNAs. Nature 432, 235-240.
Greninger, A.L., Chen, E.C., Sittler, T., Scheinerman, A., Roubinian, N., Yu, G., Kim, E., Pillai, D.R., Guyard, C., Mazzulli, T., Isa, P., Arias, C.F., Hackett, J., Schochetman, G., Miller, S., Tang, P., Chiu, C.Y., 2010. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PloS one 5, e13381.
Griffiths-Jones, S., 2004. The microRNA Registry. Nucleic Acids Res 32, D109-111.
Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J., 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34, D140-144.
Griffiths-Jones, S., Saini, H.K., van Dongen, S., Enright, A.J., 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res 36, D154-158.
Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., Bartel, D.P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27, 91-105.
Guo, Y., Li, C.I., Ye, F., Shyr, Y., 2013. Evaluation of read count based RNAseq analysis methods. BMC Genomics 14 Suppl 8, S2.
Gupta, V., Markmann, K., Pedersen, C.N., Stougaard, J., Andersen, S.U., 2012. shortran: a pipeline for small RNA-seq data analysis. Bioinformatics 28, 2698-2700.
Ha, M., Kim, V.N., 2014. Regulation of microRNA biogenesis. Nature reviews. Molecular cell biology 15, 509-524.
Hackenberg, M., Rodriguez-Ezpeleta, N., Aransay, A.M., 2011. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res 39, W132-138.
Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M., Aransay, A.M., 2009. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 37, W68-76.
Hagiwara, A., Cornu, M., Cybulski, N., Polak, P., Betz, C., Trapani, F., Terracciano, L., Heim, M.H., Ruegg, M.A., Hall, M.N., 2012. Hepatic mTORC2 activates glycolysis and lipogenesis through Akt, glucokinase, and SREBP1c. Cell Metab 15, 725-738.
Hardcastle, T.J., Kelly, K.A., 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11, 422.
Hatziapostolou, M., Polytarchou, C., Aggelidou, E., Drakaki, A., Poultsides, G.A., Jaeger, S.A., Ogata, H., Karin, M., Struhl, K., Hadzopoulou-Cladaras, M., Iliopoulos, D., 2011. An HNF4alpha-miRNA inflammatory feedback circuit regulates hepatocellular oncogenesis. Cell 147, 1233-1247.
Hausser, J., Zavolan, M., 2014. Identification and consequences of miRNA-target interactions--beyond repression of gene expression. Nature reviews. Genetics 15, 599-612.

Hayes, J., Peruzzi, P.P., Lawler, S., 2014. MicroRNAs in cancer: biomarkers, functions and therapy. Trends in molecular medicine 20, 460-469.

Hong, C., Moorefield, K.S., Jun, P., Aldape, K.D., Kharbanda, S., Phillips, H.S., Costello, J.F., 2007. Epigenome scans and cancer genome sequencing converge on WNK2, a kinase-independent suppressor of cell growth. Proceedings of the National Academy of Sciences of the United States of America 104, 10974-10979.

Hoper, D., Hoffmann, B., Beer, M., 2011. A comprehensive deep sequencing strategy for full-length genomes of influenza A. PloS one 6, e19075.

Hoshida, Y., Toffanin, S., Lachenmayer, A., Villanueva, A., Minguez, B., Llovet, J.M., 2010. Molecular classification and novel targets in hepatocellular carcinoma: recent advancements. Semin Liver Dis 30, 35-51.

Hou, J., Lin, L., Zhou, W., Wang, Z., Ding, G., Dong, Q., Qin, L., Wu, X., Zheng, Y., Yang, Y., Tian, W., Zhang, Q., Wang, C., Zhang, Q., Zhuang, S.M., Zheng, L., Liang, A., Tao, W., Cao, X., 2011. Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. Cancer cell 19, 232-243.

Hsu, S.H., Ghoshal, K., 2013. MicroRNAs in Liver Health and Disease. Curr Pathobiol Rep 1, 53-62.

Humphreys, D.T., Suter, C.M., 2013. miRspring: a compact standalone research tool for analyzing miRNA-seq data. Nucleic Acids Res 41, e147.

Hung, C.H., Chiu, Y.C., Chen, C.H., Hu, T.H., 2014. MicroRNAs in hepatocellular carcinoma: carcinogenesis, progression, and therapeutic target. Biomed Res Int 2014, 486407.

Hung, C.S., Liu, H.H., Liu, J.J., Yeh, C.T., Chang, T.C., Wu, C.H., Ho, Y.S., Wei, P.L., Chang, Y.J., 2013. MicroRNA-200a and -200b mediated hepatocellular carcinoma cell migration through the epithelial to mesenchymal transition markers. Annals of surgical oncology 20 Suppl 3, S360-368.

Iliopoulos, D., Drosatos, K., Hiyama, Y., Goldberg, I.J., Zannis, V.I., 2010. MicroRNA-370 controls the expression of microRNA-122 and Cpt1alpha and affects lipid metabolism. J Lipid Res 51, 1513-1523.

Iwasaki, A., Pillai, P.S., 2014. Innate immunity to influenza virus infection. Nature reviews. Immunology 14, 315-328.

Jablkowski, M., Bocian, A., Bialkowska, J., Bartkowiak, J., 2005. A comparative study of P53/MDM2 genes alterations and P53/MDM2 proteins immunoreactivity in liver cirrhosis and hepatocellular carcinoma. J Exp Clin Cancer Res 24, 117-125.

Jain, A., Kaczanowska, S., Davila, E., 2014. IL-1 Receptor-Associated Kinase Signaling and Its Role in Inflammation, Cancer Progression, and Therapy Resistance. Front Immunol 5, 553.

Jiang, J., Gusev, Y., Aderca, I., Mettler, T.A., Nagorney, D.M., Brackett, D.J., Roberts, L.R., Schmittgen, T.D., 2008. Association of MicroRNA expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival. Clin Cancer Res 14, 419-427.

Kang, S.M., Lee, H.J., 2014. MicroRNAs in human lung cancer. Experimental biology and medicine.

Karakatsanis, A., Papaconstantinou, I., Gazouli, M., Lyberopoulou, A., Polymeneas, G., Voros, D., 2013. Expression of microRNAs, miR-21, miR-31, miR-122, miR-145, miR-146a, miR-200c, miR-221, miR-222, and miR-223 in patients with hepatocellular carcinoma or intrahepatic cholangiocarcinoma and its prognostic significance. Mol Carcinog 52, 297-303.

Khvorova, A., Reynolds, A., Jayasena, S.D., 2003. Functional siRNAs and miRNAs exhibit strand bias. Cell 115, 209-216.

Kilpinen, H., Barrett, J.C., 2013. How next-generation sequencing is transforming complex disease genetics. Trends Genet 29, 23-30.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36.

Kim, D., Salzberg, S.L., 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome biology 12, R72.

Kim, K.W., Bae, S.K., Lee, O.H., Bae, M.H., Lee, M.J., Park, B.C., 1998. Insulin-like growth factor II induced by hypoxia may contribute to angiogenesis of human hepatocellular carcinoma. Cancer Res 58, 348-351.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 22, 568-576.

Kodama, T., Takehara, T., Hikita, H., Shimizu, S., Shigekawa, M., Tsunematsu, H., Li, W., Miyagi, T., Hosui, A., Tatsumi, T., Ishida, H., Kanto, T., Hiramatsu, N., Kubota, S., Takigawa, M., Tomimaru, Y., Tomokuni, A., Nagano, H., Doki, Y., Mori, M., Hayashi, N., 2011. Increases in p53 expression induce CTGF synthesis by mouse and human hepatocytes and result in liver fibrosis in mice. The Journal of clinical investigation 121, 3343-3356.

Kontani, K., Tada, M., Ogawa, T., Okai, T., Saito, K., Araki, Y., Katada, T., 2002. Di-Ras, a distinct subgroup of ras family GTPases with unique biochemical properties. J Biol Chem 277, 41070-41078.

Kozomara, A., Griffiths-Jones, S., 2011. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39, D152-157.

Kozomara, A., Griffiths-Jones, S., 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42, D68-73.

Krol, J., Loedige, I., Filipowicz, W., 2010. The widespread regulation of microRNA biogenesis, function and decay. Nature reviews. Genetics 11, 597-610.

Kuo, T.C., Tan, C.T., Chang, Y.W., Hong, C.C., Lee, W.J., Chen, M.W., Jeng, Y.M., Chiou, J., Yu, P., Chen, P.S., Wang, M.Y., Hsiao, M., Su, J.L., Kuo, M.L., 2013. Angiopoietin-like protein 1 suppresses SLUG to inhibit cancer cell motility. The Journal of clinical investigation 123, 1082-1095.

Kuroda, M., Katano, H., Nakajima, N., Tobiume, M., Ainai, A., Sekizuka, T., Hasegawa, H., Tashiro, M., Sasaki, Y., Arakawa, Y., Hata, S., Watanabe, M., Sata, T., 2010. Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. PloS one 5, e10256.

Ladeiro, Y., Couchy, G., Balabaud, C., Bioulac-Sage, P., Pelletier, L., Rebouissou, S., Zucman-Rossi, J., 2008. MicroRNA profiling in hepatocellular tumors is associated with clinical features and oncogene/tumor suppressor gene mutations. Hepatology 47, 1955-1963.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10, R25.

Lee, J.M., Heo, M.J., Lee, C.G., Yang, Y.M., Kim, S.G., 2015. Increase of miR-199a-5p by protoporphyrin IX, a photocatalyzer, directly inhibits E2F3, sensitizing mesenchymal tumor cells to anti-cancer agents. Oncotarget 6, 3918-3931.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., Kim, V.N., 2003. The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 415-419.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., Kim, V.N., 2004. MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23, 4051-4060.

Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., Haag, J.D., Gould, M.N., Stewart, R.M., Kendziorski, C., 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29, 1035-1043.

Leon, A.J., Banner, D., Xu, L., Ran, L., Peng, Z., Yi, K., Chen, C., Xu, F., Huang, J., Zhao, Z., Lin, Z., Huang, S.H., Fang, Y., Kelvin, A.A., Ross, T.M., Farooqui, A., Kelvin, D.J., 2013. Sequencing, annotation, and characterization of the influenza ferret infectome. Journal of virology 87, 1957-1966.

Levine, A.J., 1997. p53, the cellular gatekeeper for growth and division. Cell 88, 323-331.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., Burge, C.B., 2003. Prediction of mammalian microRNA targets. Cell 115, 787-798.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009a. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Li, J., Tibshirani, R., 2013. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 22, 519-536.

Li, L., Masica, D., Ishida, M., Tomuleasa, C., Umegaki, S., Kalloo, A.N., Georgiades, C., Singh, V.K., Khashab, M., Amateau, S., Li, Z., Okolo, P., Lennon, A.M., Saxena, P., Geschwind, J.F., Schlachter, T., Hong, K., Pawlik, T.M., Canto, M., Law, J., Sharaiha, R., Weiss, C.R., Thuluvath, P., Goggins, M., Shin, E.J., Peng, H., Kumbhari, V., Hutfless, S., Zhou, L., Mezey, E., Meltzer, S.J., Karchin, R., Selaru, F.M., 2014. Human bile contains microRNA-laden extracellular vesicles that can be used for cholangiocarcinoma diagnosis. Hepatology 60, 896-907.

Li, Q.J., Zhou, L., Yang, F., Wang, G.X., Zheng, H., Wang, D.S., He, Y., Dou, K.F., 2012. MicroRNA-10b promotes migration and invasion through CADM1 in human hepatocellular carcinoma cells. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 33, 1455-1465.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009b. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.

Li, Y., Hu, Y., Bolund, L., Wang, J., 2010. State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics 4, 271-277.

Liberal, R., Vergani, D., Mieli-Vergani, G., 2015. Update on Autoimmune Hepatitis. J Clin Transl Hepatol 3, 42-52.

Lin, C.Y., Loven, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., Young, R.A., 2012. Transcriptional amplification in tumor cells with elevated c-Myc. Cell 151, 56-67.

Lin, Z., Farooqui, A., Li, G., Wong, G.K., Mason, A.L., Banner, D., Kelvin, A.A., Kelvin, D.J., Leon, A.J., 2014. Next-generation sequencing and bioinformatic approaches to detect and analyze influenza virus in ferrets. Journal of infection in developing countries 8, 498-509.

Lipkin, W.I., 2013. The changing face of pathogen discovery and surveillance. Nature reviews. Microbiology 11, 133-141.

Liu, B., Li, J., Cairns, M.J., 2014. Identifying miRNAs, targets and functions. Brief Bioinform 15, 1-19.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. Journal of biomedicine & biotechnology 2012, 251364.

Llovet, J.M., Bruix, J., 2008. Molecular targeted therapies in hepatocellular carcinoma. Hepatology 48, 1312-1327.

Love, M.I., 2014. Assessment of DESeq2 performance through simulation.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15, 550.

Loven, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., Young, R.A., 2012. Revisiting global gene expression analysis. Cell 151, 476-482.

Lujambio, A., Lowe, S.W., 2012. The microcosmos of cancer. Nature 482, 347-355.

Lytle, J.R., Yario, T.A., Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proceedings of the National Academy of Sciences of the United States of America 104, 9667-9672.

Maheshwari, A., Thuluvath, P.J., 2006. Cryptogenic cirrhosis and NAFLD: are they related? Am J Gastroenterol 101, 664-668.

Maragkakis, M., Reczko, M., Simossis, V.A., Alexiou, P., Papadopoulos, G.L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., Hatzigeorgiou, A.G., 2009. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 37, W273-276.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.

Marquardt, J.U., Galle, P.R., Teufel, A., 2012. Molecular diagnosis and therapy of hepatocellular carcinoma (HCC): an emerging field for advanced technologies. J Hepatol 56, 267-275.

Marra, F., Arrighi, M.C., Fazi, M., Caligiuri, A., Pinzani, M., Romanelli, R.G., Efsen, E., Laffi, G., Gentilini, P., 1999. Extracellular signal-regulated kinase activation differentially regulates platelet-derived growth factor's actions in hepatic stellate cells, and is induced by in vivo liver injury in the rat. Hepatology 30, 951-958.

McDaniel, K., Herrera, L., Zhou, T., Francis, H., Han, Y., Levine, P., Lin, E., Glaser, S., Alpini, G., Meng, F., 2014. The functional role of microRNAs in alcoholic liver injury. J Cell Mol Med 18, 197-207.

McHardy, A.C., Adams, B., 2009. The role of genomics in tracking the evolution of influenza A virus. PLoS pathogens 5, e1000566.

Medina, R.A., Garcia-Sastre, A., 2011. Influenza A viruses: new research developments. Nature reviews. Microbiology 9, 590-603.

Melton, A.C., Soon, R.K., Jr., Park, J.G., Martinez, L., Dehart, G.W., Yee, H.F., Jr., 2007. Focal adhesion disassembly is an essential early event in hepatic stellate cell chemotaxis. Am J Physiol Gastrointest Liver Physiol 293, G1272-1280.

Mendell, J.T., 2008. miRiad roles for the miR-17-92 cluster in development and disease. Cell 133, 217-222.

Meng, F., Henson, R., Wehbe-Janek, H., Ghoshal, K., Jacob, S.T., Patel, T., 2007. MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. Gastroenterology 133, 647-658.

Metzker, M.L., 2010. Sequencing technologies - the next generation. Nature reviews. Genetics 11, 31-46.

Mikheyev, A.S., Tin, M.M., 2014. A first look at the Oxford Nanopore MinION sequencer. Molecular ecology resources 14, 1097-1102.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., Marshall, D., 2010. Tablet--next generation sequence assembly visualization. Bioinformatics 26, 401-402.

Mise, M., Arii, S., Higashituji, H., Furutani, M., Niwano, M., Harada, T., Ishigami, S., Toda, Y., Nakayama, H., Fukumoto, M., Fujita, J., Imamura, M., 1996. Clinical significance of vascular endothelial growth factor and basic fibroblast growth factor gene expression in liver tumor. Hepatology 23, 455-464.

Mogilyansky, E., Rigoutsos, I., 2013. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. Cell death and differentiation 20, 1603-1614.

Morgan, X.C., Huttenhower, C., 2014. Meta'omic analytic techniques for studying the intestinal microbiome. Gastroenterology 146, 1437-1448 e1431.

Morishita, A., Masaki, T., 2014. miRNA in hepatocellular carcinoma. Hepatology research : the official journal of the Japan Society of Hepatology.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628.

Munoz-Garrido, P., Garcia-Fernandez de Barrena, M., Hijona, E., Carracedo, M., Marin, J.J., Bujanda, L., Banales, J.M., 2012. MicroRNAs in biliary diseases. World J Gastroenterol 18, 6189-6196.

Murakami, Y., Yasuda, T., Saigo, K., Urashima, T., Toyoda, H., Okanoue, T., Shimotohno, K., 2006. Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. Oncogene 25, 2537-2545.

Myklebust, M.P., Fluge, O., Immervoll, H., Skarstein, A., Balteskard, L., Bruland, O., Dahl, O., 2012. Expression of DSG1 and DSC1 are prognostic markers in anal carcinoma patients. Br J Cancer 106, 756-762.

Nakamura, S., Yang, C.S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., Iida, T., Nakaya, T., 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PloS one 4, e4219.

Ng, R., Song, G., Roll, G.R., Frandsen, N.M., Willenbring, H., 2012. A microRNA-21 surge facilitates rapid cyclin D1 translation and cell cycle progression in mouse liver regeneration. The Journal of clinical investigation 122, 1097-1108.

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D.R., Tessarollo, L., Casellas, R., Zhao, K., Levens, D., 2012. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell 151, 68-79.

Ninomiya, M., Kondo, Y., Funayama, R., Nagashima, T., Kogure, T., Kakazu, E., Kimura, O., Ueno, Y., Nakayama, K., Shimosegawa, T., 2013. Distinct microRNAs expression profile in primary biliary cirrhosis and evaluation of miR 505-3p and miR197-3p as novel biomarkers. PloS one 8, e66086.

Noetel, A., Kwiecinski, M., Elfimova, N., Huang, J., Odenthal, M., 2012. microRNA are Central Players in Anti- and Profibrotic Gene Regulation during Liver Fibrosis. Front Physiol 3, 49.

Nowrousian, M., 2010. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryotic cell 9, 1300-1310.

O'Neill, L.A., Sheedy, F.J., McCoy, C.E., 2011. MicroRNAs: the fine-tuners of Toll-like receptor signalling. Nature reviews. Immunology 11, 163-175.

Oshlack, A., Wakefield, M.J., 2009. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4, 14.

Ota, A., Tagawa, H., Karnan, S., Tsuzuki, S., Karpas, A., Kira, S., Yoshida, Y., Seto, M., 2004. Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma. Cancer Res 64, 3087-3095.

Padgett, K.A., Lan, R.Y., Leung, P.C., Lleo, A., Dawson, K., Pfeiff, J., Mao, T.K., Coppel, R.L., Ansari, A.A., Gershwin, M.E., 2009. Primary biliary cirrhosis is associated with altered hepatic microRNA expression. J Autoimmun 32, 246-253.

Palomaki, G.E., Deciu, C., Kloza, E.M., Lambert-Messerlian, G.M., Haddow, J.E., Neveux, L.M., Ehrich, M., van den Boom, D., Bombard, A.T., Grody, W.W., Nelson, S.F., Canick, J.A., 2012. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet Med 14, 296-305.

Pan, C., Chen, H., Wang, L., Yang, S., Fu, H., Zheng, Y., Miao, M., Jiao, B., 2012. Down-regulation of MiR-127 facilitates hepatocyte proliferation during rat liver regeneration. PloS one 7, e39151.

Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., Hatzigeorgiou, A.G., 2013. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res 41, W169-173.

Park, J.K., Kogure, T., Nuovo, G.J., Jiang, J., He, L., Kim, J.H., Phelps, M.A., Papenfuss, T.L., Croce, C.M., Patel, T., Schmittgen, T.D., 2011. miR-221 silencing blocks hepatocellular carcinoma and promotes survival. Cancer Res 71, 7608-7616.

Pellicoro, A., Ramachandran, P., Iredale, J.P., Fallowfield, J.A., 2014. Liver fibrosis and repair: immune regulation of wound healing in a solid organ. Nature reviews. Immunology 14, 181-194.

Peterson, S.M., Thompson, J.A., Ufkin, M.L., Sathyanarayana, P., Liaw, L., Congdon, C.B., 2014. Common features of microRNA target prediction tools. Front Genet 5, 23.

Pocobelli, G., Cook, L.S., Brant, R., Lee, S.S., 2008. Hepatocellular carcinoma incidence trends in Canada: analysis by birth cohort and period of diagnosis. Liver Int 28, 1272-1279.

Pushkarev, D., Neff, N.F., Quake, S.R., 2009. Single-molecule sequencing of an individual human genome. Nature biotechnology 27, 847-850.

Qu, K.Z., Zhang, K., Li, H., Afdhal, N.H., Albitar, M., 2011. Circulating microRNAs as biomarkers for hepatocellular carcinoma. J Clin Gastroenterol 45, 355-360.

Quinn GP, K.M., 2002. Experimental design and data analysis for biologists. Cambridge University Press.

R core team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Ramakrishnan, M.A., Tu, Z.J., Singh, S., Chockalingam, A.K., Gramer, M.R., Wang, P., Goyal, S.M., Yang, M., Halvorson, D.A., Sreevatsan, S., 2009. The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies. PloS one 4, e7105.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome biology 14, R95.

Rayner, K.J., Suarez, Y., Davalos, A., Parathath, S., Fitzgerald, M.L., Tamehiro, N., Fisher, E.A., Moore, K.J., Fernandez-Hernando, C., 2010. MiR-33 contributes to the regulation of cholesterol homeostasis. Science 328, 1570-1573.

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., Hatzigeorgiou, A.G., 2012. Functional microRNA targets in protein coding sequences. Bioinformatics 28, 771-776.

Reif, S., Lang, A., Lindquist, J.N., Yata, Y., Gabele, E., Scanga, A., Brenner, D.A., Rippe, R.A., 2003. The role of focal adhesion kinase-phosphatidylinositol 3-kinase-akt signaling in hepatic stellate cell proliferation and type I collagen expression. J Biol Chem 278, 8083-8090.

Risso, D., Ngai, J., Speed, T.P., Dudoit, S., 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology 32, 896-902.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology 11, R25.

Robinson, M.D., Smyth, G.K., 2007. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23, 2881-2887.

Roderburg, C., Urban, G.W., Bettermann, K., Vucur, M., Zimmermann, H., Schmidt, S., Janssen, J., Koppe, C., Knolle, P., Castoldi, M., Tacke, F., Trautwein, C., Luedde, T., 2011. Micro-RNA profiling reveals a role for miR-29 in human and murine liver fibrosis. Hepatology 53, 209-218.

Rogler, C.E., Levoci, L., Ader, T., Massimi, A., Tchaikovskaya, T., Norel, R., Rogler, L.E., 2009. MicroRNA-23b cluster microRNAs regulate transforming growth factor-beta/bone morphogenetic protein signaling and liver stem cell differentiation by targeting Smads. Hepatology 50, 575-584.

Rowe, T., Leon, A.J., Crevar, C.J., Carter, D.M., Xu, L., Ran, L., Fang, Y., Cameron, C.M., Cameron, M.J., Banner, D., Ng, D.C., Ran, R., Weirback, H.K., Wiley, C.A., Kelvin, D.J., Ross, T.M., 2010. Modeling host responses in ferrets during A/California/07/2009 influenza infection. Virology 401, 257-265.

Sablok, G., Milev, I., Minkov, G., Minkov, I., Varotto, C., Yahubyan, G., Baev, V., 2013. isomiRex: web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets. FEBS letters 587, 2629-2634.
Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94, 441-448.
Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74, 5463-5467.
Sanz-Moreno, V., Gadea, G., Ahn, J., Paterson, H., Marra, P., Pinner, S., Sahai, E., Marshall, C.J., 2008. Rac activation and inactivation control plasticity of tumor cell movement. Cell 135, 510-523.
Satapathy, S.K., Sanyal, A.J., 2015. Epidemiology and Natural History of Nonalcoholic Fatty Liver Disease. Semin Liver Dis 35, 221-235.
Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Human molecular genetics 19, R227-240.
Schetter, A.J., Okayama, H., Harris, C.C., 2012. The role of microRNAs in colorectal cancer. Cancer journal 18, 244-252.
Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., Zamore, P.D., 2003. Asymmetry in the assembly of the RNAi enzyme complex. Cell 115, 199-208.
Sekiya, Y., Ogawa, T., Iizuka, M., Yoshizato, K., Ikeda, K., Kawada, N., 2011. Down-regulation of cyclin E1 expression by microRNA-195 accounts for interferon-beta-induced inhibition of hepatic stellate cell proliferation. J Cell Physiol 226, 2535-2542.
Seyednasrollah, F., Laiho, A., Elo, L.L., 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 16, 59-70.
Shapiro, E., Biezuner, T., Linnarsson, S., 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature reviews. Genetics 14, 618-630.
Shen, Q., Cicinnati, V.R., Zhang, X., Iacob, S., Weber, F., Sotiropoulos, G.C., Radtke, A., Lu, M., Paul, A., Gerken, G., Beckebaum, S., 2010. Role of microRNA-199a-5p and discoidin domain receptor 1 in human hepatocellular carcinoma invasion. Mol Cancer 9, 227.
Shi, K.Q., Lin, Z., Chen, X.J., Song, M., Wang, Y.Q., Cai, Y.J., Yang, N.B., Zheng, M.H., Dong, J.Z., Zhang, L., Chen, Y.P., 2015. Hepatocellular carcinoma associated microRNA expression signature: integrated bioinformatics analysis, experimental validation and clinical significance. Oncotarget 6, 25093-25108.
Shigehara, K., Yokomuro, S., Ishibashi, O., Mizuguchi, Y., Arima, Y., Kawahigashi, Y., Kanda, T., Akagi, I., Tajiri, T., Yoshida, H., Takizawa, T., Uchida, E., 2011. Real-time PCR-based analysis of the human bile microRNAome identifies miR-9 as a potential diagnostic biomarker for biliary tract cancer. PloS one 6, e23584.
Singal, A., Volk, M.L., Waljee, A., Salgia, R., Higgins, P., Rogers, M.A., Marrero, J.A., 2009. Meta-analysis: surveillance with ultrasound for early-stage hepatocellular carcinoma in patients with cirrhosis. Aliment Pharmacol Ther 30, 37-47.
Smith, A.N., Borthwick, K.J., Karet, F.E., 2002. Molecular cloning and characterization of novel tissue-specific isoforms of the human vacuolar H(+)-ATPase C, G and d subunits, and their evaluation in autosomal recessive distal renal tubular acidosis. Gene 297, 169-177.
Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D., Fouchier, R.A., 2004. Mapping the antigenic and genetic evolution of influenza virus. Science 305, 371-376.
Soneson, C., Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14, 91.
Song, S., Ajani, J.A., 2013. The role of microRNAs in cancers of the upper gastrointestinal tract. Nature reviews. Gastroenterology & hepatology 10, 109-118.

Starley, B.Q., Calcagno, C.J., Harrison, S.A., 2010. Nonalcoholic fatty liver disease and hepatocellular carcinoma: a weighty connection. Hepatology 51, 1820-1832.

Stephenson, I., Democratis, J., Lackenby, A., McNally, T., Smith, J., Pareek, M., Ellis, J., Bermingham, A., Nicholson, K., Zambon, M., 2009. Neuraminidase inhibitor resistance after oseltamivir treatment of acute influenza A and B in children. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 48, 389-396.

Suk, J.E., Semenza, J.C., 2011. Future infectious disease threats to Europe. American journal of public health 101, 2068-2079.

Sun, Z., Evans, J., Bhagwate, A., Middha, S., Bockol, M., Yan, H., Kocher, J.P., 2014. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. BMC Genomics 15, 423.

Szabo, G., Bala, S., 2013. MicroRNAs in liver disease. Nature reviews. Gastroenterology & hepatology 10, 542-552.

Takada, S., Asahara, H., 2012. Current strategies for microRNA research. Modern rheumatology / the Japan Rheumatism Association 22, 645-653.

Tam, S., Tsao, M.S., McPherson, J.D., 2015. Optimization of miRNA-seq data preprocessing. Brief Bioinform.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A., 2011. Differential expression in RNA-seq: a matter of depth. Genome research 21, 2213-2223.

Taubenberger, J.K., Kash, J.C., 2010. Influenza virus evolution, host adaptation, and pandemic formation. Cell host & microbe 7, 440-451.

Taubenberger, J.K., Morens, D.M., 2006. 1918 Influenza: the mother of all pandemics. Emerging infectious diseases 12, 15-22.

Thorgeirsson, S.S., Grisham, J.W., 2002. Molecular pathogenesis of human hepatocellular carcinoma. Nature genetics 31, 339-346.

Toffanin, S., Hoshida, Y., Lachenmayer, A., Villanueva, A., Cabellos, L., Minguez, B., Savic, R., Ward, S.C., Thung, S., Chiang, D.Y., Alsinet, C., Tovar, V., Roayaie, S., Schwartz, M., Bruix, J., Waxman, S., Friedman, S.L., Golub, T., Mazzaferro, V., Llovet, J.M., 2011. MicroRNA-based classification of hepatocellular carcinoma and oncogenic role of miR-517a. Gastroenterology 140, 1618-1628 e1616.

Tomimaru, Y., Eguchi, H., Nagano, H., Wada, H., Kobayashi, S., Marubashi, S., Tanemura, M., Tomokuni, A., Takemasa, I., Umeshita, K., Kanto, T., Doki, Y., Mori, M., 2012. Circulating microRNA-21 as a novel biomarker for hepatocellular carcinoma. J Hepatol 56, 167-175.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology 31, 46-53.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562-578.

Trevisani, F., D'Intino, P.E., Morselli-Labate, A.M., Mazzella, G., Accogli, E., Caraceni, P., Domenicali, M., De Notariis, S., Roda, E., Bernardi, M., 2001. Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of HBsAg and anti-HCV status. J Hepatol 34, 570-575.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., Sidow, A., Fire, A., Johnson, S.M., 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome research 18, 1051-1063.

Venook, A.P., Papandreou, C., Furuse, J., de Guevara, L.L., 2010. The incidence and epidemiology of hepatocellular carcinoma: a global and regional perspective. Oncologist 15 Suppl 4, 5-13.

Venugopal, S.K., Jiang, J., Kim, T.H., Li, Y., Wang, S.S., Torok, N.J., Wu, J., Zern, M.A., 2010. Liver fibrosis causes downregulation of miRNA-150 and miRNA-194 in hepatic stellate cells, and their overexpression causes decreased stellate cell activation. Am J Physiol Gastrointest Liver Physiol 298, G101-106.

Vickers, K.C., Shoucri, B.M., Levin, M.G., Wu, H., Pearson, D.S., Osei-Hwedieh, D., Collins, F.S., Remaley, A.T., Sethupathy, P., 2013. MicroRNA-27b is a regulatory hub in lipid metabolism and is altered in dyslipidemia. Hepatology 57, 533-542.

Vilarino-Guell, C., Wider, C., Ross, O.A., Dachsel, J.C., Kachergus, J.M., Lincoln, S.J., Soto-Ortolaza, A.I., Cobb, S.A., Wilhoite, G.J., Bacon, J.A., Behrouz, B., Melrose, H.L., Hentati, E., Puschmann, A., Evans, D.M., Conibear, E., Wasserman, W.W., Aasly, J.O., Burkhard, P.R., Djaldetti, R., Ghika, J., Hentati, F., Krygowska-Wajs, A., Lynch, T., Melamed, E., Rajput, A., Rajput, A.H., Solida, A., Wu, R.M., Uitti, R.J., Wszolek, Z.K., Vingerhoets, F., Farrer, M.J., 2011. VPS35 mutations in Parkinson disease. Am J Hum Genet 89, 162-167.

Villanueva, A., Newell, P., Chiang, D.Y., Friedman, S.L., Llovet, J.M., 2007. Genomics and signaling pathways in hepatocellular carcinoma. Semin Liver Dis 27, 55-76.

Vlachos, I.S., Hatzigeorgiou, A.G., 2013. Online resources for miRNA analysis. Clinical biochemistry 46, 879-900.

Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., DeRisi, J.L., 2002. Microarray-based detection and genotyping of viral pathogens. Proceedings of the National Academy of Sciences of the United States of America 99, 15687-15692.

Wang, D., Wei, Y., Pagliassotti, M.J., 2006. Saturated fatty acids promote endoplasmic reticulum stress and liver injury in rats with hepatic steatosis. Endocrinology 147, 943-951.

Wang, J., Li, J., Shen, J., Wang, C., Yang, L., Zhang, X., 2012a. MicroRNA-182 downregulates metastasis suppressor 1 and contributes to metastasis of hepatocellular carcinoma. BMC Cancer 12, 227.

Wang, L., Feng, Z., Wang, X., Zhang, X., 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-138.

Wang, L., Jia, X.J., Jiang, H.J., Du, Y., Yang, F., Si, S.Y., Hong, B., 2013. MicroRNAs 185, 96, and 223 repress selective high-density lipoprotein cholesterol uptake through posttranscriptional inhibition. Mol Cell Biol 33, 1956-1964.

Wang, X.W., Heegaard, N.H., Orum, H., 2012b. MicroRNAs in liver disease. Gastroenterology 142, 1431-1443.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics 10, 57-63.

Warren, R.L., Sutton, G.G., Jones, S.J., Holt, R.A., 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23, 500-501.

Washington, M.K., 2007. Autoimmune liver disease: overlap and outliers. Mod Pathol 20 Suppl 1, S15-30.

Weber-Lehmann, J., Schilling, E., Gradl, G., Richter, D.C., Wiehler, J., Rolf, B., 2014. Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Sci Int Genet 9, 42-46.

Wei, Y., Van Nhieu, J.T., Prigent, S., Srivatanakul, P., Tiollais, P., Buendia, M.A., 2002. Altered expression of E-cadherin in hepatocellular carcinoma: correlations with genetic alterations, beta-catenin expression, and clinical features. Hepatology 36, 692-701.

WHO, 2014a. Influenza (Seasonal). Fact sheet N°211 http://www.who.int/mediacentre/factsheets/fs211/en/.

WHO, 2014b. world health organization cancer fact sheet.

Wojcicka, A., Swierniak, M., Kornasiewicz, O., Gierlikowski, W., Maciag, M., Kolanowska, M., Kotlarek, M., Gornicka, B., Koperski, L., Niewinski, G., Krawczyk, M., Jazdzewski, K., 2014. Next generation sequencing reveals microRNA isoforms in liver cirrhosis and hepatocellular carcinoma. Int J Biochem Cell Biol 53, 208-217.

Wong, C.C., Wong, C.M., Tung, E.K., Au, S.L., Lee, J.M., Poon, R.T., Man, K., Ng, I.O., 2011. The microRNA miR-139 suppresses metastasis and progression of hepatocellular carcinoma by down-regulating Rho-kinase 2. Gastroenterology 140, 322-331.

Wong, C.M., Wei, L., Au, S.L., Fan, D.N., Zhou, Y., Tsang, F.H., Law, C.T., Lee, J.M., He, X., Shi, J., Wong, C.C., Ng, I.O., 2015. MiR-200b/200c/429 subfamily negatively regulates Rho/ROCK signaling pathway to suppress hepatocellular carcinoma metastasis. Oncotarget 6, 13658-13670.

Xin, Z., Yamaguchi, A., Sakamoto, K., 2014. Aberrant expression and altered cellular localization of desmosomal and hemidesmosomal proteins are associated with aggressive clinicopathological features of oral squamous cell carcinoma. Virchows Arch 465, 35-47.

Xu, H., He, J.H., Xiao, Z.D., Zhang, Q.Q., Chen, Y.Q., Zhou, H., Qu, L.H., 2010. Liver-enriched transcription factors regulate microRNA-122 that targets CUTL1 during liver development. Hepatology 52, 1431-1442.

Yang, J., Han, S., Huang, W., Chen, T., Liu, Y., Pan, S., Li, S., 2014a. A meta-analysis of microRNA expression in liver cancer. PloS one 9, e114533.

Yang, X.W., Shen, G.Z., Cao, L.Q., Jiang, X.F., Peng, H.P., Shen, G., Chen, D., Xue, P., 2014b. MicroRNA-1269 promotes proliferation in human hepatocellular carcinoma via downregulation of FOXO1. BMC Cancer 14, 909.

Yao, Y.J., Ping, X.L., Zhang, H., Chen, F.F., Lee, P.K., Ahsan, H., Chen, C.J., Lee, P.H., Peacocke, M., Santella, R.M., Tsou, H.C., 1999. PTEN/MMAC1 mutations in hepatocellular carcinomas. Oncogene 18, 3181-3185.

Yi, R., Qin, Y., Macara, I.G., Cullen, B.R., 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & development 17, 3011-3016.

Yongfeng, H., Fan, Y., Jie, D., Jian, Y., Ting, Z., Lilian, S., Jin, Q., 2011. Direct pathogen detection from swab samples using a new high-throughput sequencing technology. Clin Microbiol Infect 17, 241-244.

Yuan, J.H., Yang, F., Chen, B.F., Lu, Z., Huo, X.S., Zhou, W.P., Wang, F., Sun, S.H., 2011. The histone deacetylase 4/SP1/microrna-200a regulatory network contributes to aberrant histone acetylation in hepatocellular carcinoma. Hepatology 54, 2025-2035.

Yuan, Q., Loya, K., Rani, B., Mobus, S., Balakrishnan, A., Lamle, J., Cathomen, T., Vogel, A., Manns, M.P., Ott, M., Cantz, T., Sharma, A.D., 2013. MicroRNA-221 overexpression accelerates hepatocyte proliferation during liver regeneration. Hepatology 57, 299-310.

Yuneva, M.O., Fan, T.W., Allen, T.D., Higashi, R.M., Ferraris, D.V., Tsukamoto, T., Mates, J.M., Alonso, F.J., Wang, C., Seo, Y., Chen, X., Bishop, J.M., 2012. The metabolic profile of tumors depends on both the responsible genetic lesion and tissue type. Cell Metab 15, 157-170.

Zerbino, D.R., 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11, Unit 11 15.

Zezos, P., Renner, E.L., 2014. Liver transplantation and non-alcoholic fatty liver disease. World J Gastroenterol 20, 15532-15538.

Zhang, J., Chiodini, R., Badr, A., Zhang, G., 2011. The impact of next-generation sequencing on genomics. Journal of genetics and genomics = Yi chuan xue bao 38, 95-109.

Zhang, L., Yang, F., Yuan, J.H., Yuan, S.X., Zhou, W.P., Huo, X.S., Xu, D., Bi, H.S., Wang, F., Sun, S.H., 2013a. Epigenetic activation of the MiR-200 family contributes to H19-mediated metastasis suppression in hepatocellular carcinoma. Carcinogenesis 34, 577-586.

Zhang, W., Liu, J., Wang, G., 2014. The role of microRNAs in human breast cancer progression. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 35, 6235-6244.

Zhang, Y., Takahashi, S., Tasaka, A., Yoshima, T., Ochi, H., Chayama, K., 2013b. Involvement of microRNA-224 in cell proliferation, migration, invasion, and anti-apoptosis in hepatocellular carcinoma. J Gastroenterol Hepatol 28, 565-575.

Zhao, J.L., Rao, D.S., Boldin, M.P., Taganov, K.D., O'Connell, R.M., Baltimore, D., 2011. NF-kappaB dysregulation in microRNA-146a-deficient mice drives the development of myeloid malignancies. Proceedings of the National Academy of Sciences of the United States of America 108, 9184-9189.

Zhou, B., Donnelly, M.E., Scholes, D.T., St George, K., Hatta, M., Kawaoka, Y., Wentworth, D.E., 2009. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. Journal of virology 83, 10309-10313.

Zhou, J., Yu, L., Gao, X., Hu, J., Wang, J., Dai, Z., Wang, J.F., Zhang, Z., Lu, S., Huang, X., Wang, Z., Qiu, S., Wang, X., Yang, G., Sun, H., Tang, Z., Wu, Y., Zhu, H., Fan, J., 2011. Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma. J Clin Oncol 29, 4781-4788.

Zhou, X., Oshlack, A., Robinson, M.D., 2013. miRNA-Seq normalization comparisons need improvement. RNA 19, 733-734.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., Yu, J., 2010. The next-generation sequencing technology and application. Protein Cell 1, 520-536.

Zhu, Q., Wang, Z., Hu, Y., Li, J., Li, X., Zhou, L., Huang, Y., 2012. miR-21 promotes migration and invasion by the miR-21-PDCD4-AP-1 feedback loop in human hepatocellular carcinoma. Oncol Rep 27, 1660-1668.

Zhu, Y.H., Fu, L., Chen, L., Qin, Y.R., Liu, H., Xie, F., Zeng, T., Dong, S.S., Li, J., Li, Y., Dai, Y., Xie, D., Guan, X.Y., 2013. Downregulation of the novel tumor suppressor DIRAS1 predicts poor prognosis in esophageal squamous cell carcinoma. Cancer Res 73, 2298-2309.

Zucman-Rossi, J., Laurent-Puig, P., 2007. Genetic diversity of hepatocellular carcinomas and its potential impact on targeted therapies. Pharmacogenomics 8, 997-1003.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Gorczak, K., Klamecka, K., Figlerowicz, M., Siatkowski, I., 2015. The Impact of Normalization Methods on RNA-Seq Data Analysis. Biomed Res Int 2015, 621690.

Appendix

Script 1 trim_3_adapter.pl

```
#!/usr/bin/perl -w

=head1
This script will remove sequences at the 3' end of reads, starting from a tag
provided by the user. It will iterate throughout all files of a directory
with the extension specified by the user

USAGE: perl trim_3_tail.pl <directory> <extension> <prefix>

prefix = the prefix that will be used to create a log file

=cut

use strict;

chomp (my $dir    = $ARGV[0]);
chomp (my $ext    = $ARGV[1]);
chomp (my $prefix = $ARGV[2]);
my $tag_length = 6;

open (LOGFILE, ">$dir/$prefix.log");

my @files = ReadDir::get_dir_files ($dir, $ext);

opendir (DIR, "PWD") # PWD = present working directory
my @files = grep {/\.fastq|\.fq/} readdir DIR;
closedir DIR;
@files = sort (@files);

foreach my $file (@files) {
     (my $outfile = $file) =~ s/\.$ext/_trim.$ext/;
     open (IN, "$dir/$file");
     open (OUT, ">$dir/$outfile");

     print "Processing File: $file\n";

     my $record_counter = 0;
     my $trimming_counter = 0;

     my $id = '';
     my $seq = '';
     my $sign = '';
     my $qual = '';

     do {
       if ($ext =~ /fa|fasta|fq|fastq/){
            chomp ($id = <IN>);
            chomp ($seq = <IN>);
       }if ($ext =~ /fq|fastq/){
            chomp ($sign = <IN>);
            chomp ($qual = <IN>);
```

```
	}else{
		print "This script processes ONLY files with the following\n";
		print "extensions: fa, fasta, fq, fastq\n";
		exit;
	}

	$record_counter++;
	my $offset = $tag_length;
	my $read_length = length ($seq);

		for (my $i = 0; $i < 29; ++$i){

			my $window = substr($seq, ($read_length - $offset),
$offset);

			# full adapter: TGGAATTCTCGGGTGCCAAGG

		if ($window =~
/TGGAAT|.GGAATTCT|T.GAATTCT|TG.AATTCT|TGG.ATTCT|TGGA.TTCT|TGGAA.TCT/){



				my $trimmed_seq   = substr ($seq, 0,
($read_length - $offset));
				my $trimmed_qual  = substr ($qual, 0,
($read_length - $offset));
			my $trimmed_length = length ($trimmed_seq);

			if ($trimmed_length < 31){

					print OUT "$id\n";
					print OUT "$trimmed_seq\n";
					print OUT "$sign\n";
					print OUT "$trimmed_qual\n";
					$i = 100;
			$trimming_counter++;
			}
			}
			else {
				++$offset;
			}

		}
    } until eof;

    close IN;
    close OUT;
    my $perc_trimmed = sprintf ('%.2f', ($trimming_counter *
100)/$record_counter);
    print LOGFILE "\n$record_counter sequences processed\n";
    print LOGFILE "$trimming_counter sequences trimmed\n";
    print LOGFILE "Percentage of sequences trimmed: $perc_trimmed\n";
    print LOGFILE "\n_____________________________ \n";
}

close LOGFILE;
```

```
exit;
```

Script 2 prepare_miRNA_database.pl

```
#!/usr/bin/perl -w

=head
This script takes an original miRBase FASTA file and extract only miRNAs with
the tags 'hsa' at the beginning of the header line and the corresponding
sequence. In  addition, the RNA sequence gets converted into a DNA sequence
and a string of 't's get appended at the 3' of the miRNA sequence, to enable
alignment

USAGE: perl format_human_miRBase.pl <miRBase_file.fasta>

=cut

use strict;

chomp (my $infile = $ARGV[0]);
open (IN, $infile);

(my $outfile = $infile) =~ s/.f.*/_hsa.fa/;
open (OUT, ">$outfile");

while (my $line = <IN>){
     if ($line =~ /sapiens/){
       my $header = $line;
       $header =~ s/ /_/g;
       chomp (my $seq = <IN>);
       $seq =~ tr/U/T/;
       my $diff = 45 - (length ($seq));
       for (my $i = 0; $i < $diff; $i++){
             $seq .= 'n';
       }
       print OUT "$header$seq\n";
     }
}
close IN;
close OUT;

print "Your results are in $outfile\n";

exit;
```

Script 3 bowtie_miRNAs.pl

```
#!/usr/local/bin/perl -w

=head1

This program will perform bowtie alignments on all files included in a user-
specified directory...

If unaligned reads want to be streamed to separated files, include option --
```

```
un <file_name.un> in the bowtie command (for filterig of ribosome and
mitocondria hits)


USAGE: perl bowtie_miRNA.pl

=cut

print "Directory? -> ";
chomp($dir = <STDIN>);

do{
    print "File format ? 'fastq', 'fq', 'fasta' or 'fa' ONLY -> ";
    chomp ($format = <STDIN>);
}until (($format eq 'fastq') || ($format eq 'fq') || ($format eq 'fasta') ||
($format eq 'fa'));

opendir(DIR, "$dir") or die "$!";
@files = grep {/\.$format$/} readdir DIR;
closedir DIR;

@files = sort(@files);
print "@files \n";
$file_counter = 0;

print "Which database you want to align against:\n";
print "[1] hg19\n";
print "[2] Human miRNAs (miRBase19)\n";
print "[3] Human miRNAs (miRBase20)\n";
print "[4] Human miRNAs (miRBase21)\n";
print "[5] Human Ribo & Mito\n";
print "[6] All viruses in NCBI\n";
chomp ($genome = <STDIN>);

if ($genome eq '1') {$binary_file  ='/data/jane/db/hg19/bwt/hg19.binary';
$prefix = 'hg19'}
if ($genome eq '2') {$binary_file
='/data/jane/projects/miRNA/db/miRNA_current/miRBase19/bwt/mature/miR19.binar
y'; $prefix = 'miR19'}
if ($genome eq '3') {$binary_file
='/data/jane/projects/miRNA/db/miRNA_current/miRBase20/bwt/hs_miR20'; $prefix
= 'miR20'}
if ($genome eq '4') {$binary_file
='/data/jane/projects/miRNA/db/miRNA_current/miRBase21/bwt/mature.hsa';
$prefix = 'miR21'}
if ($genome eq '5') {$binary_file  ='/data/jane/db/ribo/bwt $ ls
hum_rm.binary'; $prefix = 'riboMito'}
if ($genome eq '6') {$binary_file
='/data/jane/db/viruses/dna/all/140505/virus'; $prefix = 'virus'}

if    (($format eq 'fastq') || ($format eq 'fq')) {$input_format = 'q'}
elsif (($format eq 'fa') || ($format eq 'fasta')) {$input_format = 'f'}

foreach $file (@files){
        ($out = $file) =~ s/\.$format$/-$prefix.bwt/;
        $file_counter++;
        print "----------File: $file --------\n";
```

```
        system("bowtie -$input_format -v 0 -p 5 $binary_file $dir/$file
$dir/$out");
}

print "$file_counter files were processed\n";

exit;
```

Script 4 count_miRNA_hits.pl

```
#!/usr/bin/perl -w

=head1

This program counts the number of times a miRNA is reported in a bowtie
alignment summary

The program will work on all *.bwt files of a directory

USAGE: perl count_miRNAs_hits.pl <dir>

=cut

# Example of record in bowtie (input) file
#M02251:29:000000000-A8081:1:1101:15098:1540 1:N:0:4     +       hsa-let-7b-
5p   0       TGAGGTAGTAGGTTGTGTGGTTT >33>>CFFFFFFGGGGGGGGGGH 0

use warnings;
use strict;


chomp (my $dir = $ARGV[0]);
chomp (my $outPrefix = $ARGV[1]);

opendir(DIR, $dir) || die "$!";
my @files = grep {/\.bwt$/} readdir DIR;
close DIR;
@files = sort(@files);
my $number_of_files = @files;

my $miR19_counter   = 0;
my $miR20_counter   = 0;
my $miR21_counter   = 0;
my $noMatch_counter = 0;
foreach my $file (@files) {
    if    ($file =~ /miR19/) {$miR19_counter++}
    elsif ($file =~ /miR20/) {$miR20_counter++}
    elsif ($file =~ /miR21/) {$miR21_counter++}
    else  {$noMatch_counter++}
}

unless (($miR19_counter == $number_of_files) || ($miR20_counter ==
$number_of_files) || ($miR21_counter == $number_of_files)){
    print "All your *.bwt files should come from alignments against the same
miRBase, but does not\n";
    print "seem to be the case: $miR19_counter were aligned against
```

```
miRBase19\n";
    print "                    $miR20_counter were aligned against
miRBase20\n";
    print "                    $miR21_counter were aligned against
miRBase21\n";
    print "                    $noMatch_counter do not specify the database
they were aligned against\n";
    exit;
}
# Assign file containing the miRNAs IDs according to the version of miRBase
used for alignment
my $ids_file = '';
if    ($files[0] =~ /miR19/) { $ids_file
='/data/jane/projects/miRNA/db/miRNA_current/miRBase19/miR19_ids.fa' }
elsif ($files[0] =~ /miR20/) { $ids_file
='/data/jane/projects/miRNA/db/miRNA_current/miRBase20/miR20_ids.fa' }
elsif ($files[0] =~ /miR21/) { $ids_file
='/data/jane/projects/miRNA/db/miRNA_current/miRBase21/bwt/miR21_ids.fa' }
else  { print "Your input files name must contain at least one of the
following sitrings 'miR19, miR20 or miR21'\n";
    exit}
open (IDs, $ids_file);



# create a template hash indexed by all miRNAs in the database used for
alignment. Fill the hash with an array containing as many columns as files
@files and fill each cell of the array with 0.00
my %results  = ();

while (defined (my $miRNA = <IDs>)) {
    chomp $miRNA;

    unless (exists $results{$miRNA}) {
      for (my $i = 0; $i < $number_of_files; $i++) {
            $results{$miRNA}[$i] = 0.00;
      }
    }
}


# Open two files for printing raw counts or normalized (ppm) counts

my $raw_outfile  = $outPrefix . '_raw.xls';
my $norm_outfile = $outPrefix . '_norm.xls';
open (OUTR, ">$dir/$raw_outfile") || die "$!";
open (OUTN, ">$dir/$norm_outfile") || die "$!";


# File counter
my $j = 0;
my @normFactors;
print OUTR "miRNA";
print OUTN "miRNA";

foreach my $file (@files){
    print "Processing file: $file\n";
```

```
    open (IN, "$dir/$file") || die "$!";

    (my $sample_name = $file) =~ s/.bwt//;
    print OUTR "\t$sample_name";
    print OUTN "\t$sample_name";

    my $normalizer = `wc -l $dir/$file`;
    $normalizer =~ s/\s.*//;
    $normalizer /= 1000000;
    push (@normFactors, $normalizer);


    do{
      chomp(my $line = <IN>);
      my @temp = split (/\t/, $line);
      my $strand = $temp[1];
      my $target = $temp[2];
      if(($strand eq '+') and ($temp[6] == 0)) {
            $results{$target}[$j] += 1;
      }
    }until eof;

    $j++;
    close IN;
}

print OUTR "\n";
print OUTN "\n";

# print results

foreach my $key (sort keys %results){
    print OUTR $key;
    print OUTN $key;
    for (my $i = 0; $i < $number_of_files; $i++){
      print OUTR "\t$results{$key}[$i]";
      print OUTN "\t", $results{$key}[$i]/$normFactors[$i];
    }
    print OUTR "\n";
    print OUTN "\n";
}

close OUTR;
close OUTN;

exit;
```

Script 5 edgeR_glm.R

```
##### This script performs differential expression analysis using the edgeR
package
##### Last modified: September 07, 2015

library ("edgeR")

rawdata <- read.delim("/Users/jane/analyses/miRNAs/12_pairs_HCC.txt ",
```

```
                        check.names=FALSE, stringsAsFactors=FALSE,
                        header=TRUE, sep="\t")
head (rawdata)

# Import count data into a DGEList class. Exclude the first column (miRNAs
names). From the pure counts and instead assign it to a new slot in the
DGEList class, called 'genes'
miRcountTable <- DGEList(counts=rawdata[,2:25], genes=rawdata[,1])

# Calculate actual library size
miRcountTable$samples$lib.size <- colSums(miRcountTable$counts)

# Remove rows with less than 50 read/million in at least 3 samples
miRcountTable <- miRcountTable[rowSums(1e+06 *
miRcountTable$counts/expandAsMatrix(miRcountTable$samples$lib.size,
dim(miRcountTable)) > 50) >= 3, ]

# Write to the pwd the table containing the counts of miRNAs after excluding
the non-abundant ones
tempMiRNAs <- miRcountTable$counts
write.table(tempMiRNAs, "afterExcludingNonAbundant.txt", sep="\t", eol="\n")

# Check how many miRNA survived
dim (miRcountTable)

# Calculate normalization factors
miRcountTable <- calcNormFactors(miRcountTable)

# Visualize normalization factors
miRcountTable$samples

# Make multi-dimensional scalling plot
plotMDS( miRcountTable , main = "MDS Plot for Count Data", labels = colnames(
miRcountTable$counts ) )

# Effective library sizes will be the product of normalization factors and
the actual size of the libraries
miRcountTable$samples$lib.size * miRcountTable$samples$norm.factors

# Create a new factor "patient" to be integrated into the model
Patient <- factor(c(1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12))

# Create a new factor "tissue" to be integrated into the model
Tissue <-
factor(c("N","T","N","T","N","T","N","T","N","T","N","T","N","T","N","T","N",
"T","N","T","N","T","N","T"))

# Create a dataframe that includes the new factors paired to the name of each
sample
data.frame(Sample=colnames(miRcountTable),Patient,Tissue)

# The model will discriminate the patient to patient variability and the
cancer to non-cancer variability
design <- model.matrix(~Patient+Tissue)

rownames(design) <- colnames(miRcountTable)
```

```
miRcountTable <- estimateGLMCommonDisp(miRcountTable, design, verbose=TRUE)
names (miRcountTable)
summary( miRcountTable$common.dispersion )
miRcountTable <- estimateGLMTrendedDisp(miRcountTable, design)
names (miRcountTable)
summary( miRcountTable$trended.dispersion )
miRcountTable <- estimateGLMTagwiseDisp(miRcountTable, design)
names (miRcountTable)
summary( miRcountTable$tagwise.dispersion )

meanVarPlot <- plotMeanVar( miRcountTable ,show.raw.vars=TRUE ,
show.tagwise.vars=TRUE ,
                            show.binned.common.disp.vars=FALSE ,
show.ave.raw.vars=FALSE ,
                            NBline = TRUE , nbins = 100 ,
                            #these are arguments about what is plotted
                            pch = 16 , xlab ="Mean Expression (Log10 Scale)"
,
                            ylab = "Variance (Log10 Scale)" , main = "Mean-
Variance Plot" )
                            #these arguments are to make it look prettier

# Plotted will be:
# the raw variances of the counts (grey dots),
# the variances using the tagwise dispersions (light blue dots),
# the variances using the common dispersion (solid blue line),
# variance = mean a.k.a. poisson variance (solid black line).

fit <- glmFit(miRcountTable, design)
lrt <- glmLRT(fit)
de.glmLRT <- topTags(n=134, lrt)

names (lrt)
lrt$comparison

colnames(design)
o <- order(lrt$table$PValue)
cpm(miRcountTable)[o[1:10],]
summary(de <- decideTestsDGE(lrt))
detags <- rownames(miRcountTable)[as.logical(de)]
plotSmear(lrt, de.tags=detags)
abline(h=c(-1, 1), col="blue")

write.table(de.glmLRT, file="glmLRT_results.tsv", sep = "\t" , row.names =
FALSE)
```

Script 6 DESeq2.R

```
# This script implements differential expression using the DESeq2 package and
also perform some exploratory tests
# Last modification: September 07, 2015

library ("DESeq2")
library("pheatmap")
library ("ggplot2")
library ("gplots")
library("RColorBrewer")
```

```
suppressWarnings(warning("testit"))

setwd ('/Users/jane/analyses/miRNAs/DESeq_150611')
countdata <- read.table('4HCC_miRNAs.counts', header=TRUE, sep="\t",
                        row.names = 1)
#sampleInfo <- read.table('metadata_4HCC_miRNA.txt', header = TRUE, sep="\t",
row.names = 1)
sampleInfo <- read.table('metadata', header = TRUE, sep="\t", row.names = 1)


# convert data.frame containing the count data into a matrix
countdata <- data.matrix(countdata)

# Generate the DESeqDataSet object from the count matrix the ddsMat object
will contain all the input data, the information about the procedures
implemented on the data and will also store the results
(ddsMat <- DESeqDataSetFromMatrix(countData = countdata, colData =
sampleInfo,
                                  design = ~ patient + condition))

# Estimate size factors required for differential expression analyses
ddsMat <- estimateSizeFactors(ddsMat)

# Apply a regularized-logarithmic transformation (rlog)
rld <- rlog(ddsMat)

# Calculate Euclidian distances between samples
sampleDists <- dist( t( assay(rld) ) )
sampleDists

# Visualize Euclidian distances in a heatmap
sampleDistMatrix <- as.matrix( sampleDists )
rownames(sampleDistMatrix) <- paste( rld$patient, rld$condition, sep="_" )
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "Reds")) )(255)
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors)


# Visualize Euclidian distances in a heatmap
poisd <- PoissonDistance(t(counts(ddsMat)))

samplePoisDistMatrix <- as.matrix( poisd$dd )
rownames(samplePoisDistMatrix) <- paste( rld$patient, rld$condition, sep="_"
)
colnames(samplePoisDistMatrix) <- NULL
pheatmap(samplePoisDistMatrix,
         clustering_distance_rows=poisd$dd,
         clustering_distance_cols=poisd$dd,
         col=colors)

# Generate an MDS plot
mds <- data.frame(cmdscale(sampleDistMatrix))
mds <- cbind(mds, as.data.frame(colData(rld)))
```

```
mdsPlotEuc <- qplot(X1,X2,color=condition,data=mds)
mdsPlotEuc + geom_point(shape = 20,  size = 6) +
  geom_text(aes(label=label),hjust=0, vjust=1.5, color="black")

# Generate the MDS for the Poisson Distance
mds <- data.frame(cmdscale(sampleDistMatrix))
mds <- cbind(mds, as.data.frame(colData(rld)))
mdsPlotEuc <- qplot(X1,X2,color=group,data=mds)
mdsPlotEuc + geom_point(shape = 20,  size = 7) +
  labs(title="MDS with Euclidean Distances")
  geom_text(aes(label=name),hjust=0, vjust=1.5, color="black")

# Extract "n" genes with the largest variance
topVarGenes <- head(order(-rowVars(assay(rld))),50)

# Plot topVarGenes in a heatmap, but using the distance of each gene counts
to the mean of all samples for the same gene
matTopVarGenes <- assay(rld)[ topVarGenes, ]
matTopVarGenes <- matTopVarGenes - rowMeans(matTopVarGenes)
df <- as.data.frame(colData(rld))
pheatmap(matTopVarGenes, annotation_col=df)

### DIFFERENTIAL EXPRESSION
# This single command performs differential expression analysis
ddsMat <- DESeq (ddsMat)
(res <- results(ddsMat))

# Extract all gene records in the results that have a p-value smaller than
0.05
res.05 <- results(ddsMat, alpha=0.05)

# From those ones, extract only the ones that had an FDR smaller than 0.05
table(res.05$padj < 0.05)

# Visualize how many results have a p-value smaller than 0.05
sum(res$pvalue < 0.05, na.rm=TRUE)

# Visualize how many results have a FDR smaller than 0.05
sum(res$padj < 0.05, na.rm=TRUE)

resSig <- subset(res, padj < 0.1)
head(resSig[ order( resSig$log2FoldChange ), ])

# Export signidicantly differentially expressed results to a text file
write.table(resSig, file = "diffExpAnal_FDR_0.05.txt", sep = "\t", eol =
"\n")


# Create an MA plot (change the ylimits if desired)
plotMA(res, main="Differential Expression (DESeq2)", ylim=c(-4,4))

# To extract the unschrunken data, export to an external file if wanted
resMLE <- results(ddsMat, addMLE=TRUE)
plotMA(resMLE, main="Differential Expression (unschrunken)", ylim=c(-4,4))

# Mark the top gene in the plot
plotMA(res, ylim=c(-6,6))
```

```
with(res[topGene, ], {
  points(baseMean, log2FoldChange, col="dodgerblue", cex=2, lwd=2)
  text(baseMean, log2FoldChange, topGene, pos=2, col="dodgerblue")
})
```

Script 7 uniq.pl

```
#!/usr/bin/perl -w

=head1
This script will extract unique reads from all fastq files in a directory it
will also count their frequency

USAGE perl uniq.pl <dir> <ext>

=cut

use lib "/home/jane/scripts/perl/modules";
use ReadDir;
use strict;

chomp (my $dir = $ARGV[0]);
chomp (my $file_ext  = $ARGV[1]);
my @files = ReadDir::get_dir_files ($dir, $file_ext);


foreach my $file(@files){

    my %all_reads = ();
    my %counts = ();
    my @counts = ();
    my %unique = ();

    (my $outfile = $file) =~ s/\.$file_ext/_unique\.fa/;
        open (IN, "$dir/$file");
        open (OUT, ">$dir/$outfile");

    print "\nDumping file $file into a hash...\n";
    do{
            chomp (my $id   = <IN>);
            chomp (my $seq  = <IN>);
            $all_reads{$id} = $seq;
      if ($file_ext =~ /fastq|fq/){
            chomp (my $sign = <IN>);
            chomp (my $qual = <IN>);
      }

    }until eof;
    close IN;

    print "Pushing uniq sequences into an array\n";
    print "and counting their frequency...\n";

# Each sequence in the %all_reads hash will be extracted into $string and
used as a key for the %counts hash (i.e. if $counts{$string}does not exist,
it will be pushed into another hash %unique. The frequency of each unique
sequence will be stored in %counts
```

```
    my $total_reads = 0;
    foreach my $key (sort keys %all_reads){
      $total_reads++;
            my $string = $all_reads{$key};
            if(not exists $counts{$string}){
                  $unique{$key} = $string;
            }
      $counts{$string}++;
    }

    my @unique_keys = sort keys %unique;

    foreach my $key (@unique_keys){
            push @counts, $counts{$unique{$key}}
    }

    my $unique_counter = 0;
    my $sum_freq_unique = 0;

    print "Printing results into output file $outfile\n";
    foreach my $key (sort keys %unique){
      my $length = length($unique{$key});
      print OUT
">n=$counts[$unique_counter]:Length=$length:Unique_read_$unique_counter\n";
      $sum_freq_unique += $counts[$unique_counter];
      $unique_counter++;
      print OUT "$unique{$key}\n";
    }
    close OUT;
    print "$total_reads read ends were analyzed\n";
    print "$unique_counter unique read ends found\n";
    # Notice that the sum of frequencies of unique reads should equal the
number of read ends analyzed
    print "The sum of frequencies of unique reads = $sum_freq_unique\n";
    print "_____\n";
}


exit;
```