

MicroRNAs in Metastasis

*Investigating miRNAs in colorectal cancer
derived liver metastasis*

Eirik Høye



Thesis submitted for the degree of
Master of Science in Molecular Bioscience
60 credits

Department of Bioscience
Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

December / 2016

MicroRNAs in Metastasis

© Eirik Høyve

2016

MicroRNAs in Metastasis

Eirik Høyve

<http://www.duo.uio.no/>

Trykk: Reprosentralen, Universitetet i Oslo

Abstract

Colorectal Cancer (CRC) is one of the leading causes of cancer related deaths in the western world, and metastatic progression is the dominating cause of mortality. The primary site of CRC metastasis is the liver, followed by the lungs and peritoneal cavity, and prognosis for patients with metastatic CRC (mCRC) is poor, with only 10 % five-year survival. Although much is known about progression and metastasis of CRC; how primary CRC (pCRC) differs from mCRC on the molecular level are incompletely understood. This is important for our understanding of the disease, but also could have significant implications with respect to detection and treatment of CRC derived liver metastases.

On numerous occasions microRNAs have been shown to be key elements in cancer progression and are candidate biomarkers detectable in blood. However, recent reports on mCRC failed to identify microRNA signatures of metastatic progression. To address this, a small RNA sequencing approach was used focusing on primary tumors and a set of liver metastases. Further, the highly-curated microRNA reference MirGeneDB.org was used to ensure that only bona fide microRNAs were studied.

Although global miRNA expression was not distinguishable between primary tumor and colorectal derived liver metastasis, a number of individual miRNAs were significantly different between pCRC and mCRC of the liver. Surprisingly, Mir-339-3p and Mir-1247-5p were validated in a meta-analysis of published data that hadn't reported them. Specific isoforms (isomiRs) were also found to be differentially expressed.

This study underlines the importance of using high quality microRNA reference dataset, and lays the foundation for more in-depth investigations of miRNA role in this deadly disease.

Preface

This thesis was written as part of the Master's Program in Molecular Biology at the University of Oslo. The author has a bachelor's degree in Molecular Biology, also from the university. Thesis opponents are Trine Rounge from the Cancer Registry of Norway and Kamran Shalchian-Tabrizi from the Department of Biosciences, UiO.

Several contributors have helped this thesis become what it is, most notably I would like to thank my main supervisor Bastian Fromm, who's friendly yet firm attitude has helped me exceed my expectations. I would also like to thank Jon Bråte, my internal supervisor, Kjersti Flatmark, our PI, and all members of our research group, who have been more than welcoming and have always been in a positive and enthusiastic spirit.

In writing this thesis, I have acquired an extensive skill set necessary for data intensive biology, including programming and high performance computing. These skills were not taught to me in my background as a molecular biologist, and will surely help me in my future scientific career.

Eirik Høye

December 2016

Table of contents

MicroRNAs in Metastasis	III
Abstract	V
Preface	VII
Table of contents	VIII
List of abbreviations	XII
1 Introduction	14
1.1 MicroRNAs	15
1.1.1 miRNA induced silencing	18
1.1.2 Functional role	20
1.1.3 Evolution	22
1.1.4 Annotation and nomenclature	24
1.1.5 IsomiRs	28
1.1.6 Sequential and structural motifs	29
1.2 Cancer	31
1.2.1 Metastasis	34
1.2.2 Colorectal Cancer	36
1.3 MiRNA as biomarkers	38
1.4 Goals	40
2 Materials and methods	40
2.1 Clinical studies	42
2.2 RNA isolation and quality control	44
2.3 NGS Library preparation and sequencing	46
2.4 Preprocessing and read mapping	49
2.5 Sample distances and hierarchical clustering	52
2.6 Differential expression analysis	53
2.6.1 Target prediction	54
2.7 IsomiRs	55
2.8 Sequential motifs	56
3 Results	57
3.1 RNA extraction	58
3.2 NGS-results	60
3.3 Clustering and sample distance	63
3.3.1 nCR and nLi are distinct from each other	63

3.3.2	pCRC is distinct from nCR	64
3.3.3	CLM is distinct from nLi	65
3.3.4	No distinction of pCRC and CLM.....	66
3.4	Differential expression	67
3.4.1	Venn downregulated signature miRNA.....	67
3.4.2	Venn upregulated miRNA.....	68
3.4.3	Volcano plot nCR versus nLi	70
3.4.4	Volcano plot pCRC versus CLM.....	71
3.4.5	Validation in Neerincx and Röhr.....	73
3.5	Target prediction.....	76
3.6	IsomiRs	78
3.6.1	Table isomiRs downregulated in CLM	78
3.6.2	Table isomiRs upregulated in CLM	79
3.7	Sequential motifs.....	81
4	Discussion	85
	References	88
	Attachments.....	93
4.1.1	Top 10 miRNA per Tissue	93
4.1.2	Volcano plot nCR vs pCRC.....	94
4.1.3	Volcano plot nLi versus CLM.....	95
4.1.4	Volcano plot of isomiRs in nCR and nLi	96
4.1.5	Volcano plot of isomiRs in pCRC and CLM.....	97

Figure 1.1	miRNA Structure.....	15
Figure 1.2	miRNA Biogenesis	17
Figure 1.3	miRNA Induced Silencing	18
Figure 1.4	miRNA Targeting	19
Figure 1.5	Rheostat model of miRNA 'resistance'	20
Figure 1.6	miRNA Regulatory Network.....	21
Figure 1.7	miRNA and Tissue Complexity	23
Figure 1.8	miRNA Annotation	25
Figure 1.9	Rejected miRNA MirGeneDB	26
Figure 1.10	IsomiR Definition.....	28
Figure 1.11	miRNA Structural Motifs.....	29
Figure 1.12	Hallmarks of cancer proposed by Hanahan et al 2011	31
Figure 1.12	Stages of progression of colorectal cancer.....	36
Figure 2.1	TrueSeq small RNA sample preparation.....	46
Figure 2.2	FASTQ format	49
Figure 2.3	Removing 3p adapter sequence	49
Figure 2.4	summarizeOverlaps.....	50
Figure 2.5	DESeq2 shrinkage of dispersion.....	53
Figure 2.6	Defining pri-miRNA Sequential and Structural Motifs.....	56
Table 3.1	Result of RNA extraction and sequencing	59
Figure 3.1	Density Plots.....	60
Figure 3.2	MeanSdPlot	60
Figure 3.3	Counts per Sample	61
Figure 3.4	Clustering nCR vs nLi	63
Figure 3.5	Clustering nCR vs pCRC.....	64
Figure 3.6	Clustering nLi vs CLM.....	65
Figure 3.7	Clustering pCRC vs mCRC	66
Figure 3.8	Venn Diagram of downregulated signature miRNA.....	67
Figure 3.9	Venn diagram of upregulated signature miRNA	68
Figure 3.10	Volcano Plot nCR vs nLi	70
Table 3.2	Signature miRNA in nCR vs nLi.....	70
Figure 3.11	Volcano Plot pCRC vs CLM	71
Table 3.3	Differentially Expressed miRNAs in pCRC vs CLM	71
Table 3.4	miRNA Differentially Expressed for pCRC vs CLM and Controlled for Normal Tissue	72
Figure 3.12	Boxplot of signature miRNA	74
Table 3.5	Signature miRNA Controlled for Normal Tissues	75
Table 3.6	RPM in tissues	75
Table 3.7	Top 10 Hsa-Mir-1247_5p target sites predicted by TargetScan	76
Table 3.8	Top 10 Hsa-Mir-339_3p target sites predicted by TargetScan	77
Table 3.9	Top 15 IsomiRs Downregulated in CLM.....	78
Table 3.10	Top 15 IsomiRs Upregulated in CLM.....	79
Figure 3.13	Number of miRNA in MirGeneDB with pri-miRNA motifs.....	81
Figure 3.14	Bar Plot of Human MirGeneDB annotated miRNA genes with mismatch GHG motif	82
Table 3.11	Motifs in Signature miRNA	83
Table 3.12	Motifs in Signature isomiRs	84

Appendix Figure 1	Top 10 miRNA in nCR, pCRC, CLM and nLi.....	93
Appendix Figure 2	Volcano Plot nCR vs pCRC	94
Appendix Table 1	Top 10 signature miRNA in nCR vs nLi	94
Appendix Figure 3	Volcano Plot nLi vs CLM	95
Appendix Figure 4	IsomiR Volcano Plots nCR vs nLi	96
Appendix Figure 5	IsomiR Volcano Plots pCRC vs CLM.....	97

List of abbreviations

ceRNA	Competing Endogenous RNA
CLM	Colorectal Derived Liver Metastasis
CTC	Circulating Tumor Cells
EMT	Epithelial Mesenchymal Transition
HTS	High Throughput Sequencing
IsomiR	miRNA Isoform
LFC	Log2 Fold Change
lncRNA	Long Non-coding RNA
mCRC	Metastatic Colorectal Cancer
miRISC	MicroRNA Induced Silencing Complex
miRNA	MicroRNA
mRNA	Messenger RNA
nCR	Normal Colorectum
NGS	Next Generation Sequencing
padj	Adjusted p-Value
pCRC	Primary Colorectal Cancer
pre-miRNA	Precursor miRNA Transcript
pri-miRNA	Primary miRNA Transcript
RIN	RNA Integrity Number
RNAi	RNA Interference
RNAseq	RNA Sequencing
RPM	Counts Per Million
UTR	Untranslated Region
3p	Three Prime
5p	Five Prime

1 Introduction

1.1 MicroRNAs

MicroRNAs (miRNAs) are small, non-coding RNAs that regulate gene expression in most plants and animals. MiRNAs are 20 – 26 nucleotides long molecules that can target messenger RNA (mRNA) and inhibit their translation into proteins. Typically, a miRNA binds to a mRNA based on the complementarity of a 7-8 nucleotide long, so called “seed” sequence, to target sites in the 3p untranslated region of mRNAs. As a result miRNAs function as guiding strands for the so called miRNA Induced Silencing Complex (miRISC), allowing it to locate and degrade targeted mRNAs before they get translated into proteins [1]. MiRNAs are therefore part of the cells post transcriptional gene regulatory network.

Biogenesis and function

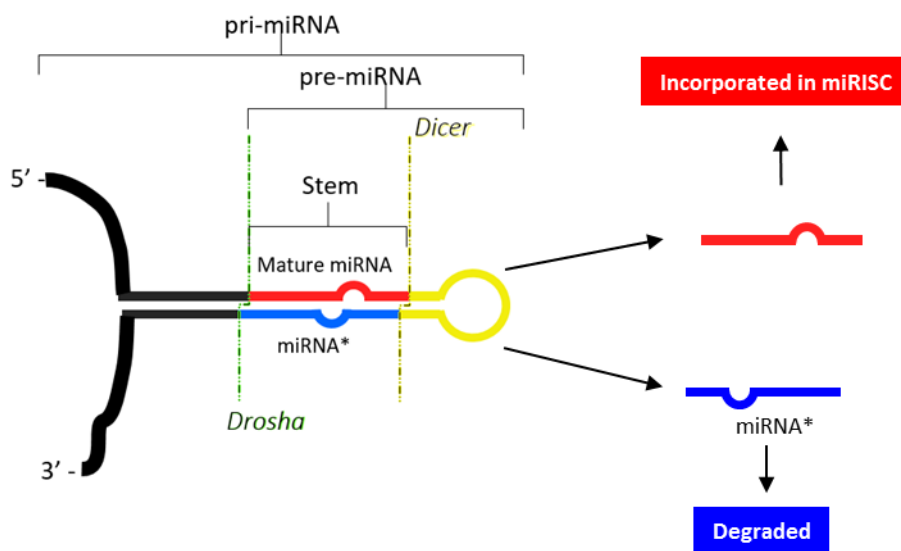


Figure 1.1 miRNA Structure Canonical miRNA biogenesis pathway, defining different stages of processing. Pri-miRNA include hairpin stem-loop, as well as 5p- and 3p- primary transcript arms. After drosha cleavage in nucleus, pre-miRNA includes the hairpin stem-loop. Pre-miRNA is transported into the cytosol, where dicer cleaves off the loop sequence, leaving the hairpin stem. The mature miRNA is incorporated into miRISC complex, while miRNA* is degraded. Flatmark et al, 2016 [2]

In the canonical miRNA biogenesis pathway, a miRNA gene is transcribed by RNA polymerase II [3] into a primary miRNA transcript (pri-miRNA). This pri-miRNA contains one or more sequential units that can form hairpin structures, the stem of which is made of complementary nucleotides which comprise the ~22 nucleotide mature and star miRNA sequences. These hairpin structures act as substrates for the RNase III enzyme Drosha [4], which, with the help of its binding partner, DGCR8 [5], cleaves off the stem loop with a 2

nucleotide offset, leaving the precursors miRNA (pre-miRNA) stem loop. This pre-miRNA is then transported out of the nucleus by Exportin 5 [6].

In the cytosol, the enzyme Dicer cleaves the pre-miRNA by removing the loop sequence, leaving a double stranded RNA molecule called the miRNA/miRNA* duplex, with a 2-nucleotide offset at the 3p-ends, [7]. In the canonical miRNA biogenesis pathways, the miRNA*, or passenger strand, is degraded, leaving what is now the ~22 nucleotide long mature miRNA which exerts biological function. Deep sequencing of miRNAs shows that the vast majority of miRNA genes follow this mature/star pattern either expressing the 5p or the 3p-strand, but some miRNA genes to show similar read counts for both strands (co-mature pattern). The mechanism by which one strand is determined over the other is not determined [8].

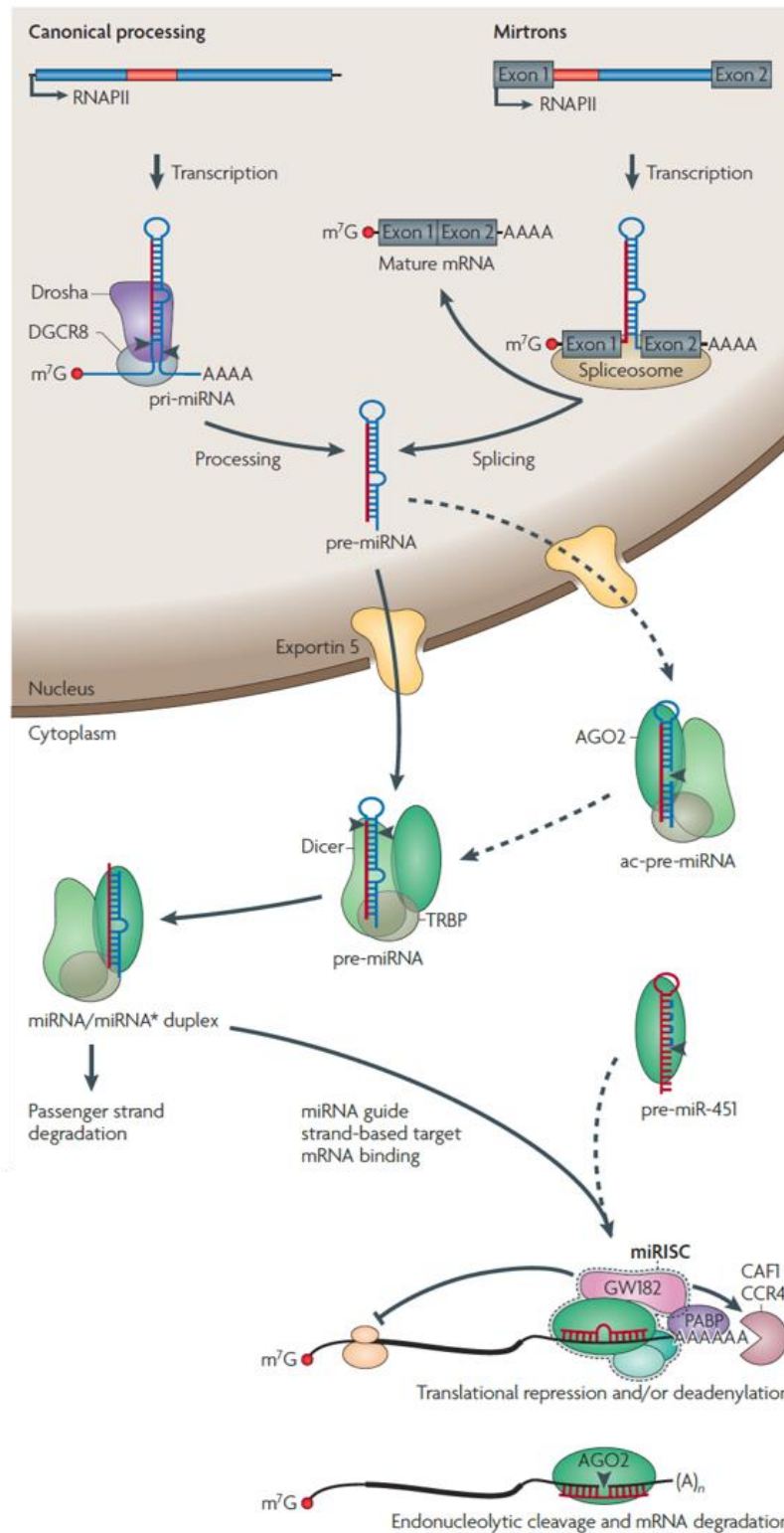


Figure 1.2 miRNA Biogenesis Two biogenesis pathways are described, canonical and a non-canonical. Canonical pathway has primary transcript, pri-miRNA, cleaved by Drosha, leaving pre-miRNA stem-loop structure. Pre-miRNA is exported by Exportin 5 into cytosol, where Dicer cleaves off the loop sequence, leaving miRNA/miRNA* duplex. miRNA* is degraded, and miRNA is incorporated into miRISC complex. The described non-canonical pathway starts off with a miRNA gene located inside the intron of a coding gene. In this case, spliceosome cleavage results in the finished pre-miRNA, no Drosha cleavage required. The remaining steps are identical to canonical pathway. Krol et al 2010 [9].

1.1.1 miRNA induced silencing

The mature miRNA is then incorporated into miRISC. The mature miRNA guides the miRISC complex to mRNA molecules that have complementary sequences on their 3p-UTR. Key protein of miRISC is Argonaute, along with various other protein factors. The miRISC complex silence gene expression either by destabilizing and degrading mRNA, or by repressing ribosomal translation (Figure 1.3) [10].

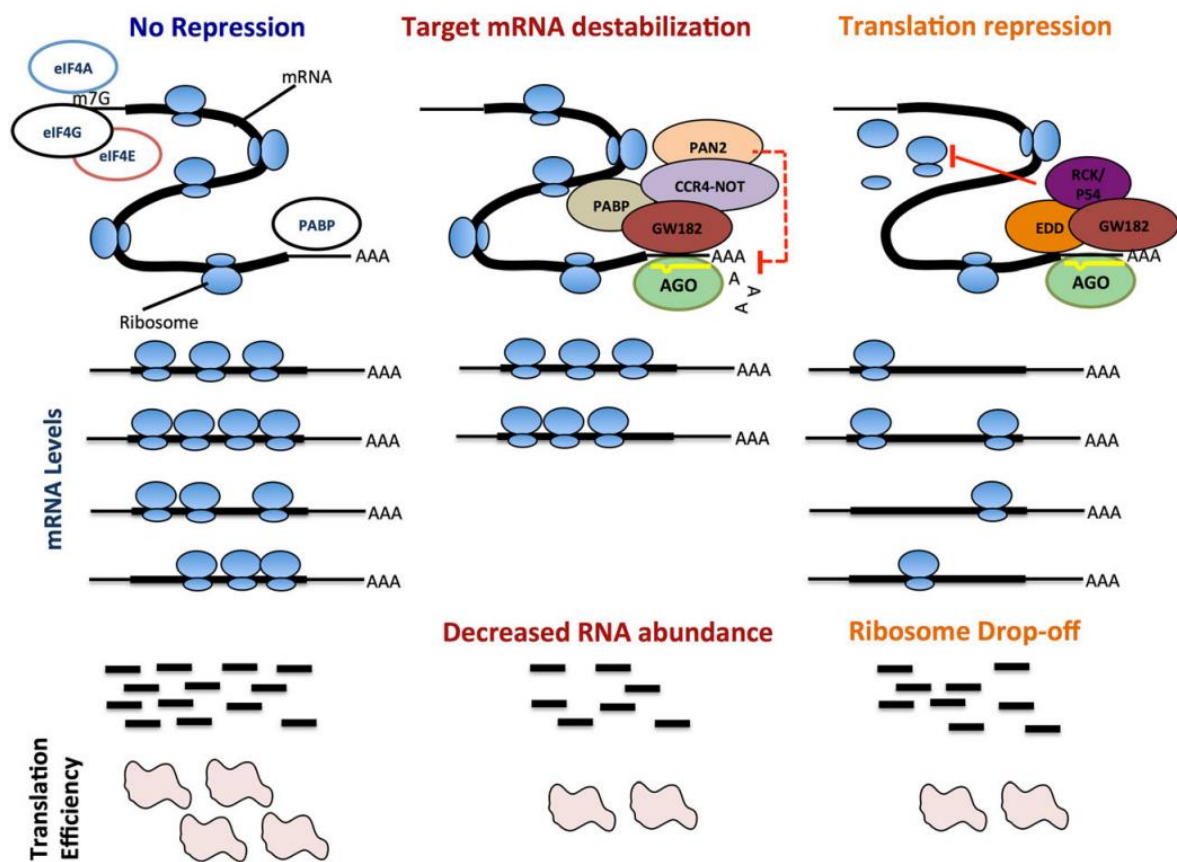


Figure 1.3 miRNA Induced Silencing With no miRNA silencing, mRNA is abundant and ribosomes free to bind and translate. With miRNA silencing, mRNAs are destabilized, and their abundance drops, while ribosomes are repressed from translation. Ramalho-Carvalho et al, 2016 [10].

Animal miRNA induced targeting of mRNA typically requires perfect Watson Crick pairing in the 5p-end nucleotides 2-7, called the ‘seed’ region [11]. There are three canonical types of miRNA target sites. First, the 7mer-A1 (**Figure 1.4a**), where the miRNA ‘seed’ form Watson crick pairing with the corresponding mRNA target site and the target site also having an adenine at position 1. Second, the 7mer-m8, the mRNA target site forms (**Figure 1.4b**) Watson crick pairing with the miRNA 2-7 “seed” plus an eight nucleotide. The 8mer site has Watson crick pairing for the “seed” and nucleotide 8, and an adenine at mRNA position 1. (**Figure 1.4c**) [11]

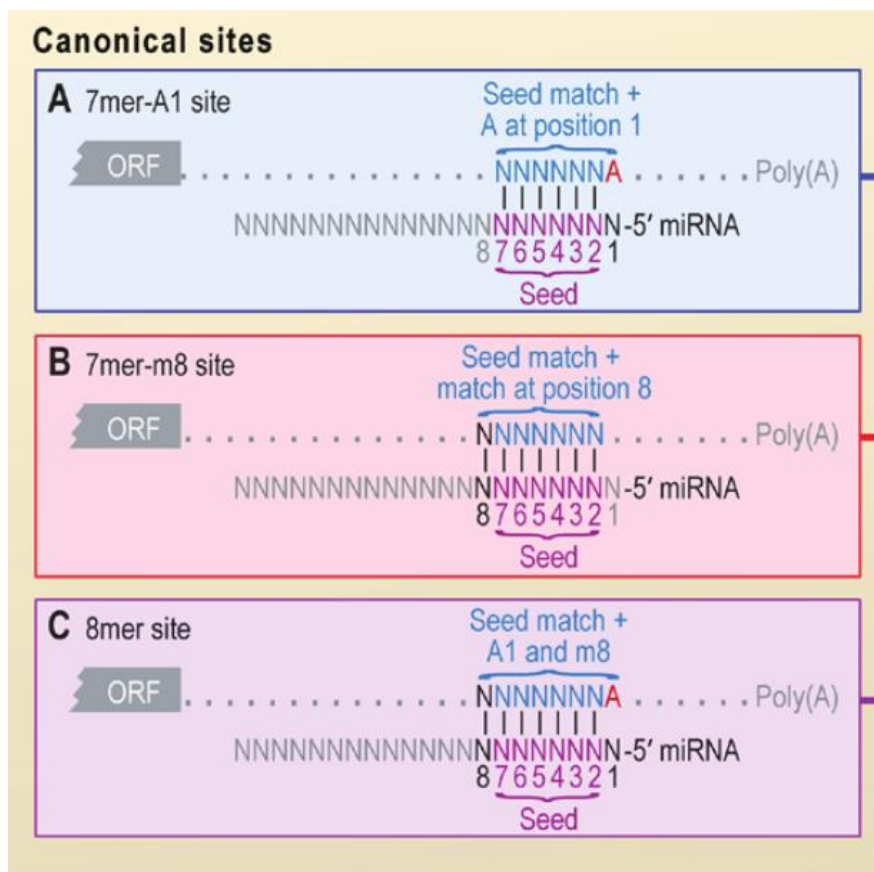


Figure 1.4 miRNA Targeting (a) 7mer-A1, with adenine in position 1, and Watson Crick pairing for nucleotides 2-7. (b) 7mer-m8, with Watson Crick pairing for both 2-7 seed region as well as nucleotide 8. (c) 8mer site with adenine at position 1 and Watson Crick pairing in both the 2-7 seed, as well as nucleotide 8. Bartel 2009 Figure 1 [11].

1.1.2 Functional role

MiRNA function as posttranscriptional gene regulators by guiding silencing protein complexes to their mRNA targets before those targets get translated into protein. Thus, miRNAs form an additional layer in the elaborated gene regulatory repertoire of cells. In contrast to other much stronger regulatory mechanisms, they only modulate gene expression. Evidence suggests miRNA targeted genes only display about a 2 to 4 fold change in the corresponding protein level [12]. Although significant, it's not enough to switch the gene off entirely. Rather, they are suggested to function as dampeners of gene expression, allowing cells in different tissues to fine tune their gene expression to suit their specific needs [11, 13]. Bartel et al, 2004, made the analogy to rheostats [13], where miRNA infer a resistance to a gene in two ways, by the number of miRNAs expressed in the cell, and the number of complementary and occupancy sites present on the mRNA.

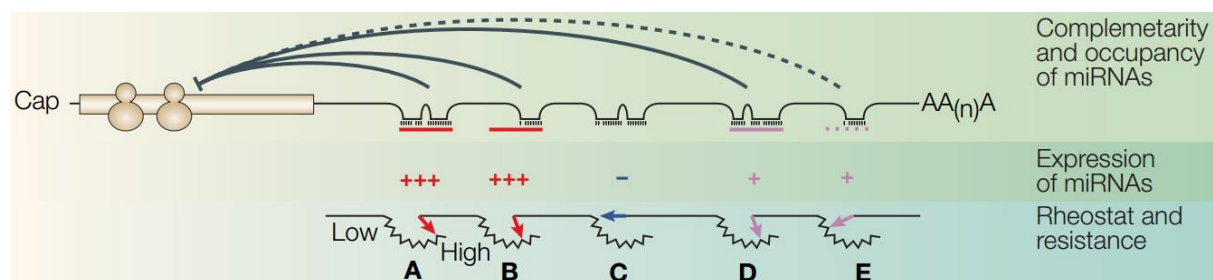


Figure 1.5 Rheostat model of miRNA 'resistance' Expression of miRNA and number of target sites both determine the amount of 'resistance' imparted by miRNA. **A** has both high miRNA expression and high complementarity and occupancy, imparting high resistance. **B** and **D** both also impart high resistance, even though **B** has fewer occupancy sites and **D** has less miRNA expression. **E** has low imparted resistance due to low miRNA expression and few occupancy sites, while **C** has no imparted resistance even with abundant complementary sites, since no miRNAs are expressed. Bartel et al, 2004 [13].

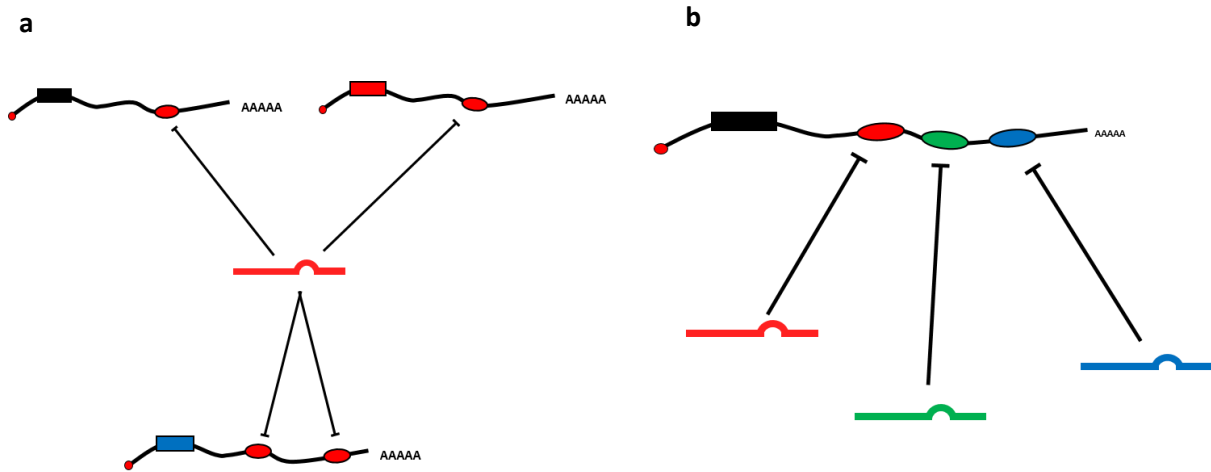


Figure 1.6 miRNA Regulatory Network (a) a single miRNA may have target sites for multiple mRNA, while (b) multiple miRNA may target the same mRNA. Flatmark et al, 2016 [2].

Since a single mRNA UTR may have target sites for multiple miRNAs, and conversely, a single miRNA may target multiple mRNA, one may suppose there are elaborate miRNA mediated regulatory networks which allow precise fine tuning of expression levels not possible by the standard, more crude, transcription factors. Furthermore, some RNAs can function as miRNA sponges, these are RNA molecules that have abundant miRNA target sites and can therefore bind much of the available miRNA in the cell, leaving very few available to repress the remainder RNA with those target sites. Such interactions of miRNAs with competing endogenous RNAs (ceRNAs) have been proposed as an important new mechanism by Salmena et al 2011 [14].

Briefly, only a finite number of miRNAs is present and available to repress mRNAs at any one time. As such, how many RNAs containing target sites for a given miRNA are present in the cell will affect their ability to repress a specific mRNA. In this way, all RNA molecules with binding sites for the same miRNA may compete with each other for repression. This would allow for a mechanism whereby separate genes may interact with each other, forming elaborate competitive endogenous RNA regulatory networks.

CeRNAs have been shown to be functionally important in muscle cells, where long-noncoding RNA linc-MD1 acts as ceRNA for miRNA regulating muscle differentiation [15], and in prostate cancer [16], glioblastoma [17] and melanoma [18] where ceRNA influence the miRNA complement available to regulate PTEN, a tumor suppressor gene. However,

recent studies have started to cast doubt about the overall physiological relevance of such a ceRNA, with Denzler et al 2014 [19] and Denzler et al 2016 [20], suggesting the likelihood of observing such a ceRNA effect is much lower than originally thought.

1.1.3 Evolution

RNA interference (RNAi) originated early in eukaryotic evolution. Cerutti et al, 2006 [21], analyzed key members of the RNAi machinery in five eukaryotic ‘supergroups’, and found that all had at least one of Argonaute-, Piwi- or Dicer-like proteins, and one RNA-dependent RNA polymerase. On this basis, they suggest that RNA interference was already present in the last common ancestor of eukaryotes, likely originating as a defense mechanism against transposable elements. This early machinery would already have the capability of transcript degradation. Interestingly however, current evidence suggests several separate emergences of miRNAs and the miRNA processing machinery in plants and animals [22]. Explanation for this seeming conundrum - given the extraordinary level of conservation within higher plants and animals - may be that miRNAs became integral for complex organisms in regulating multi-cellularity and increased cell- and tissue-complexity of an organism.

The evolution of complex organisms with multiple cell and tissue types cannot be explained by an expansion of the organism’s protein coding gene repertoire. Analyzing the protein coding genes of organisms with widely differing number of cell types show that there are about 20,000 genes required for animals to form their morphology, regardless of complexity [23], and that this genetic “toolbox” fully developed early in the Metazoa [24]. As such, increasing complexity of animal morphology and tissue types must be explained through an expansion in gene regulatory network. It is not the total number of genes but the spatial and temporal activation of those genes which allows for organismal complexity. MiRNAs, subsequently, show dramatic increase in gene number and gene families as species evolve more diverse cell and tissue types [24], and conversely, *devolution*, defined as an organisms loss of complexity over lineages, show a decrease in miRNA gene number and gene families [24].

In this view, miRNAs would be one of the foundations upon which larger, more complex organisms could emerge. The model upon which this works would be as follows. A miRNA exerting an evolutionary beneficial regulatory effect on a gene transcript would be preserved through evolution. If said miRNA, or the target site(s) on the corresponding gene transcript,

where to undergo a mutation that removed mentioned beneficial regulatory effect, it would have a negative effect and selected against. Interestingly however, other genes might also undergo a mutation in their 3'-UTR, allowing the same miRNA to exert a regulatory effect upon them as well. If the regulatory effect was beneficial, the mutation would be conserved through evolution. Gradually, increasingly complex miRNA gene regulatory networks would emerge as both more miRNA genes arise, and more genes come under their control. Conversely, in organisms where tissue complexity is lost, highly elaborate gene regulatory networks may no longer be selectively advantageous, and loss of miRNA genes may improve fitness.

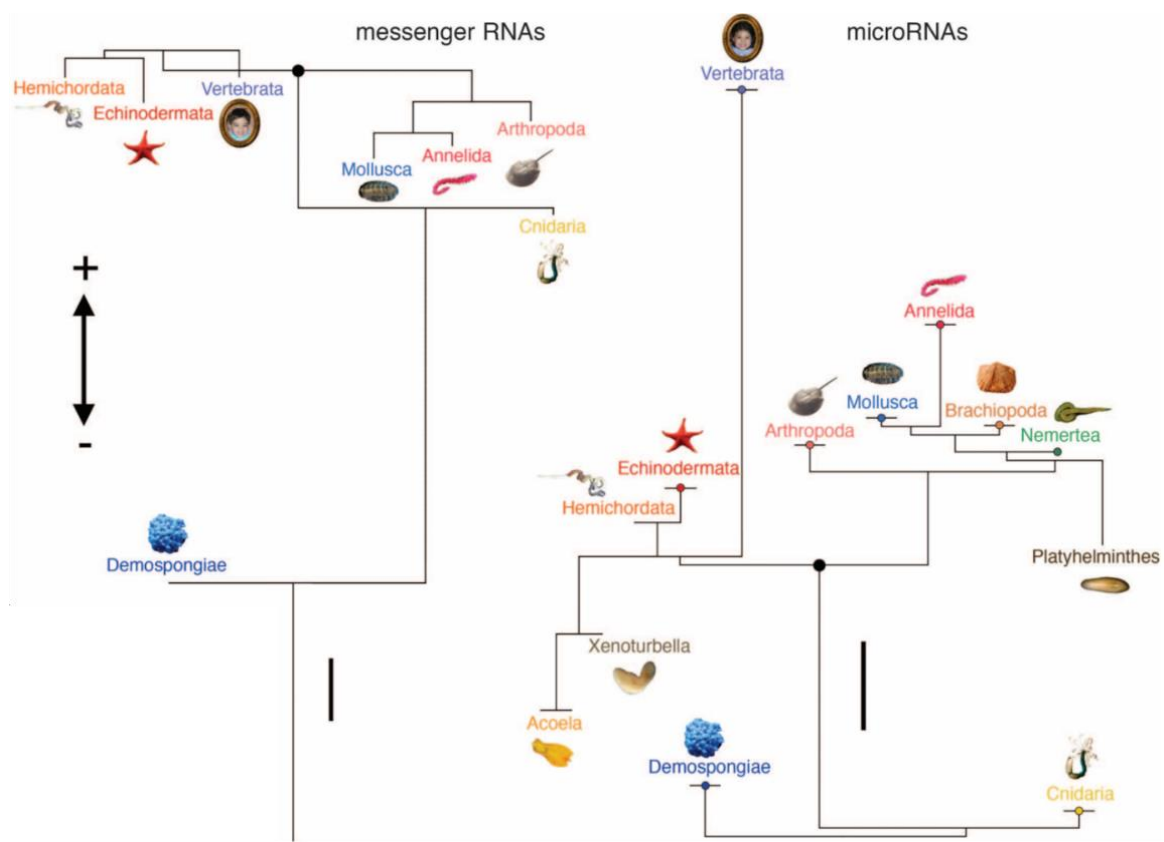
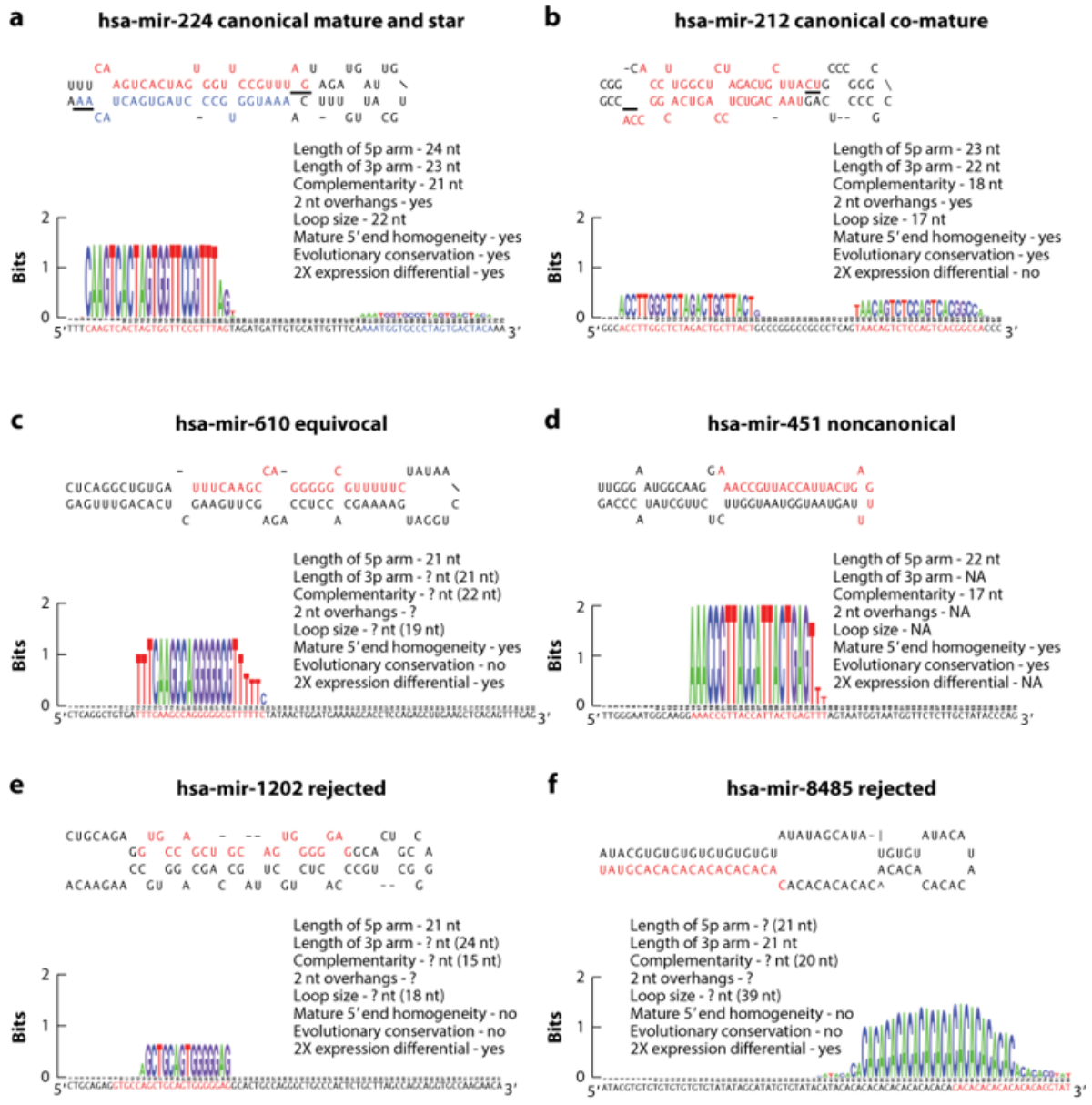


Figure 1.7 miRNA and Tissue Complexity Plot of acquisition of mRNA genes (left) and miRNA genes (right) through increasingly complex animal lineages. Branches indicate gene acquisition; lineages going up have more genes, lineages going downward have fewer. Scale bars correspond to 10 genes. The miRNA complements rapidly increase from demosponges to cnidarians and bilaterians, after which the complement remains more or less flat. MiRNA complement, meanwhile, see extensive gains in the bilaterian lineage compared to the cnidarian. Increasingly complex species see increasing miRNA complement. *Xenoturbella* and *Acoela* groups have undergone simplifications in their morphology and complexity, and see a drop in their miRNA complement. Erwin et al, 2011 [24].

1.1.4 Annotation and nomenclature

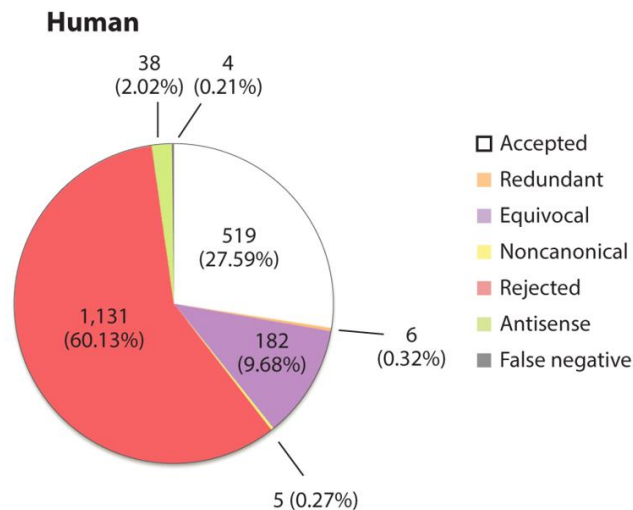
The advent of next generation sequencing has resulted in a dramatic increase in the reported miRNA. MiRBase, the current repository for annotated miRNA[25], contains 1881 human miRNA entries. MiRBase is not a curated database, and it has long been suspected that a large proportion of entries are false annotations [26-28]. Fromm et al 2015 [29] set out to ascertain the validity of the human miRNA complement in miRBase by establishing a set of criteria to define miRNAs, then compare all human entries in miRBase against those criteria. They established that miRNAs have 20-26 nt long reads expressed from both stem loop arms, these must have 2 nt offset, 5p homogeneity, the two arms must have 16 nt complementarity or more, and the loop sequence must be 8 nt in length or more. Figure 1.8 illustrates these criteria when applied to six putative miRBase “miRNAs” testing their validity.



Fromm B, et al. 2015.
 Annu. Rev. Genet. 49:213–42

Figure 1.8 miRNA Annotation Applying a consistent set of miRNA annotation criteria to six putative miRBase miRNAs. Both (a) and (b) fulfill annotation criteria, with (a) having a clear 5p expression preference, while (b) appear to be a co-mature. (c) does not have any expression of the 3p arm, while (d) is noncanonical with the mature miRNA making up the loop sequence. (e) and (f) are rejected due to not fulfilling the annotation criteria. Fromm et al, 2015 [29]

Applying these criteria to all entries in miRBase, Fromm et al, 2015 [29] showed that less than two thirds off all human entries fulfilled criteria (**Figure 1.9**). As a result, a new curated open access miRNA gene database, MirGeneDB (<http://mirgenedb.org>) was established to provide the research community with a repository of high quality, correctly annotated miRNAs.




 Fromm B, et al. 2015.
Annu. Rev. Genet. 49:213–42

Figure 1.9 Rejected miRNA MirGeneDB Pie chart of miRNA that fulfill criteria set out by Fromm et al, 2016 [29]. Less than one third of all 1881 miRNAs in miRBase fulfilled annotation criteria. Fromm et al, 2015 [29].

Additionally, a revised nomenclature system was implemented. The conventional miRNA naming was introduced by Ambros et al 2003 [30] by naming each miRNA with the prefix “miR”, followed by a number based upon the sequential order in which the miRNA was discovered. Identical miRNA has identical names, with very similar sequences given a number or letter suffix, to distinguish between them. Their coding genes are named similarly, except using italics and capital letters in the conventional manner.

This naming system does not take into account what evolutionary relationship between miRNAs. This is of huge importance for the vast numbers of miRNAs described not only for humans. Therefore, to arrive at a system where orthologous and paralogous miRNAs can be

identified based on their name; a revised nomenclature system was implemented. To avoid confusion, existing gene names were used where possible, and where miRNA genes were shown to be homologous, their names were merged. To distinguish the new nomenclature system from the old, gene names start with uppercase Mir-, and miRNA families use whole capital MIR-. MiRNA families contain only genes that are not paralogous to miRNAs outside that family. Paralogous genes within a family are annotated with a P followed by a number, starting with the first member of the family. In cases where a second duplication event has occurred, the P letter and number designation is followed by a letter indicating sequence of duplication. Lastly, all orthologues in all species are given the same name, to avoid confusion when comparing miRNA between species.

Needless to say, using a database as reference where more than two thirds of annotated miRNAs are false annotations would at best be a waste of time. At worst false conclusions might be drawn. Therefore, this study used the curated MirGeneDB as reference.

1.1.5 IsomiRs

IsomiRs, defined as variants of the canonical miRNA, have been shown to be real, physiologically active participants in the cells gene regulatory machinery [31]. IsomiRs may be polymorphic, their sequence containing mismatches compared to canonical miRNA. They may also be elongated or truncated at their 5p or 3p. Elongated isomiRs can have both canonical additions, identical to the pre-miRNA sequence, or non-canonical, where the additions differ from the pre-miRNA. By far the most common are 3p isomiRs [32-35] Some studies suggest isomiRs are dynamically and actively regulated by cells [32, 34]. Koppers-Lalic et al 2014 [36] showed that 3p uridylated isomiRs are more abundant in exosomes, while 3p adenylated isomiRs are more abundant in cytosol of cells. IsomiR modification have been reported to influence the stability of the miRNA as well as Argonaute loading [31]

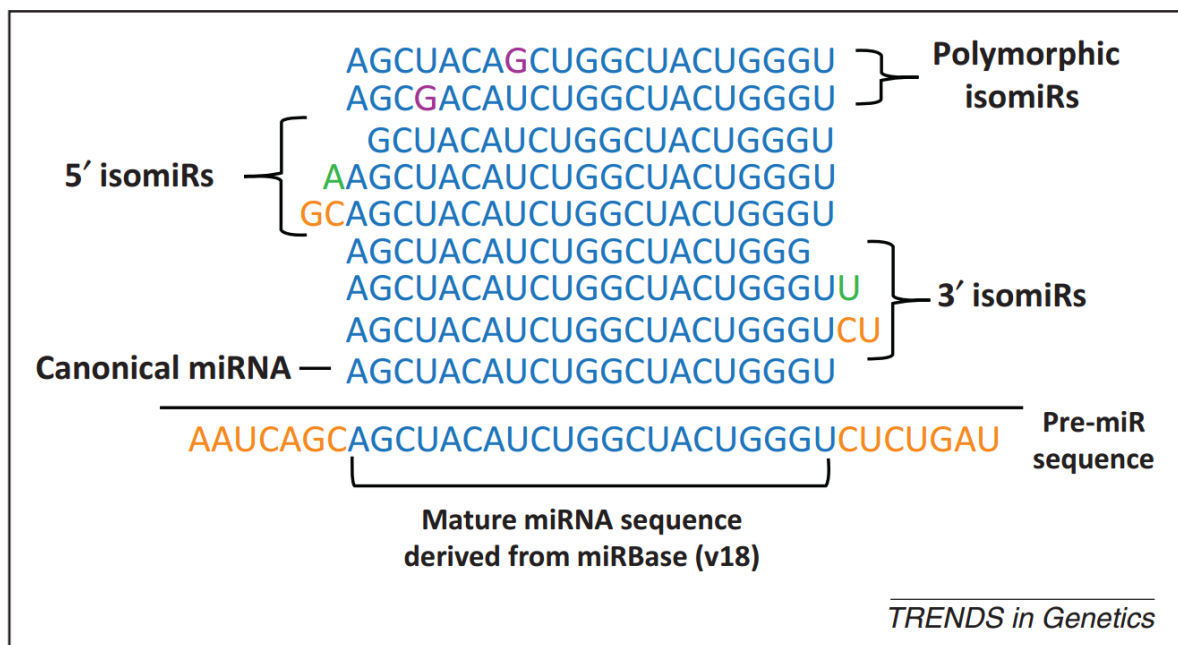


Figure 1.10 IsomiR Definition IsomiRs may be polymorphic, with distinct nucleotides (purple) compared to template strand (blue). IsomiRs may be 5p and 3p truncated, or 5p and 3p elongated, with either non-templated (green) or templated (orange) additions. Neilsen et al, 2012 [31].

1.1.6 Sequential and structural motifs

An unresolved question regarding miRNA biogenesis is how Drosha selects and cleaves hairpin sequences from transcribed miRNA genes, but avoids the remainder of transcripts containing hairpins. Estimates suggest upwards of 11 million regions of the genome may form hairpin structures if transcribed [37]. Drosha cleavage of all such transcribed, non-miRNA sequences would not only be a waste of energy for the cell, but could also lead to transcriptional abnormalities. Auyeung et al 2013 [38, 39] suggested sequential motifs in pri-miRNA stem loop as one mechanism by which such a selection mechanism might work. The sequential motifs UG 14 nucleotides upstream of Drosha cut in the 5p-prime lower stem, UGUG in the loop sequence, and CNNC motifs at position 16, 17, or 18 in the 3p-prime lower stem (**Figure 1.11**), where shown to enhance processing of mutated variants of the miRNAs miR-16, miR-30 and miR-125.

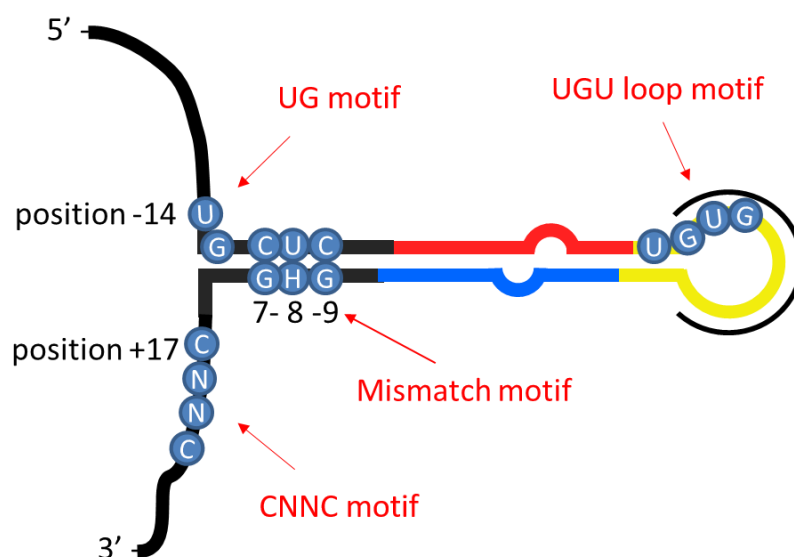


Figure 1.11 miRNA Structural Motifs Illustrates the location of the different motifs in the pri-miRNA hairpin structure. Changed after Fromm 2016 [40]

Presumably, these sequential motifs enhance binding affinity of Drosha and its interacting partner DGCR8 to the pri-miRNA stem, although Kwon et al 2016 where unable to find any

residues that would closely associate with these nucleotides after unraveling the 3D structure of Drosha [5]. An additional structural motif has been suggested by Fang et al, 2015 [41]. They propose a key component of miRNAs is a stem of double stranded RNA at 35 +/- 1 bases, stretching from the pri-miRNA basal stem region where upstream and downstream single stranded RNA fuse to form double stranded RNA, to the loop region (**Figure 1.11**). According to this model, Drosha would cut at position 13, counting from the upstream basal region, and position 11, from the downstream basal region. This implies that the upstream UG motif at position 14 upstream of Drosha cut site would lie right at the spot where single stranded RNA is fused to form double stranded RNA. They then propose a mismatch motif at position 7-8-9 for both upstream and downstream strands, where position 7 and 9 form Watson crick pairs, while position 8 are mismatching. This was again shown to enhance Drosha processing in mutated variants of miR-16, miR-30 and miR-125. (**Figure 1.11**) Previous studies (Kwon et al, 2016 Auyeung et al 2013, Fang et al, 2015, Nguyen et al 2015, [5, 38, 39, 41]) have used experimental approach for motif discovery and verification, but a comprehensive verification in a large dataset of curated miRNA genes has yet to be accomplished.

1.2 Cancer

Cancer is a multitude of diseases related by the fact they all involve uncontrolled cell growth and proliferation. As multicellular organisms are made up of tissues with trillions of cells, strict control of cell growth and division is essential. Cells have a plethora of checks and balances to ensure none of them escapes this controlled environment; however, mutations and chromosomal alterations mean eventually there is still a small probability some cells manage to circumvent them and reach tumorous growth and proliferation. The danger this poses to the patient depends on where in the body the tumor originates, at what time the tumor is discovered and the specific genetic alterations unique to that individual tumor.

There are a plethora of obstacles preventing a cell from reaching the cancerous stage. Hanahan and Weinberg therefore suggested a series of hallmarks common to cancerous tumors, first in 2000 [42], then refined later in 2011 [43].

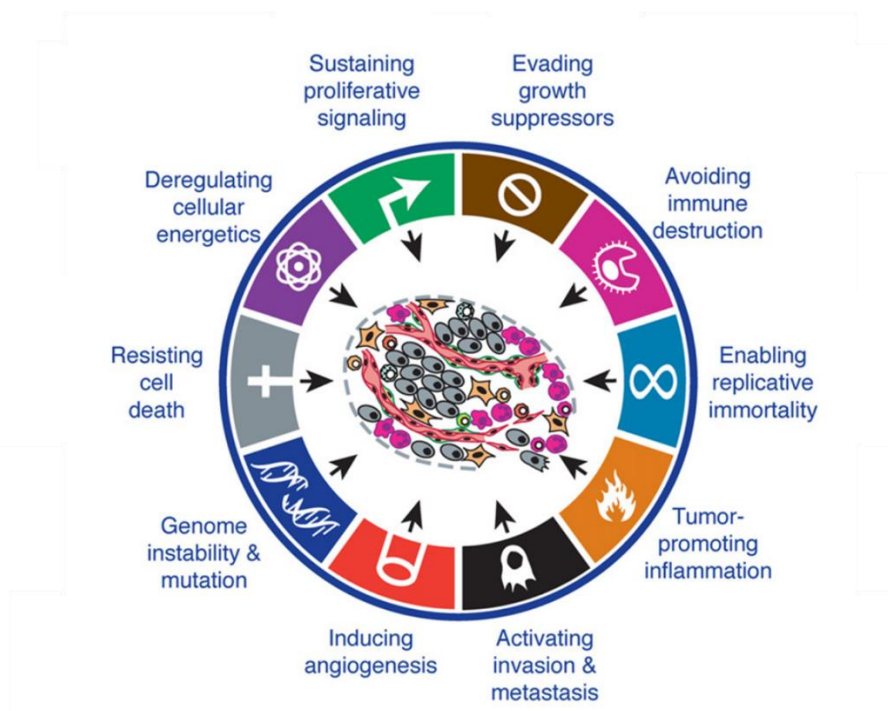


Figure 1.12 Hallmarks of cancer proposed by Hanahan et al 2011. Hanahan et al 2011 [43]

One is ensuring sustained proliferative signaling. Cells require a steady input of growth factor ligands before they can divide and proliferate. In healthy tissue, these growth signals are strictly controlled. Cancers circumvent this in a number of ways. One of them is secretion of their own growth ligands, leading to autocrine growth stimulus and tissue independence. Alternatively, secretion of signaling molecules to nearby healthy cells may trigger them to

secrete stimulatory molecules [44, 45]. Additional strategies involve increasing the number or altering the structure of growth signaling cell surface receptors, increasing response to the same stimulus. Alterations in signaling molecules downstream of the cell surface receptor may also provide growth signaling independence, leaving the growth signal permanently switched on.

The flip side is that cancer cells must also avoid growth suppressors, another hallmark. Cells have multiple tumor suppressor genes whose function is to detect and prevent uncontrolled tumor growth. Two prominent examples are retinoblastoma-associated gene, RB, and TP53. RB is involved in cell cycle regulation, where it controls passing through the R point in the G₁ cell cycle phase. In its hypophosphorylated state, RB binds to transcription factor E2F, inactivating it. Hyperphosphorylated, RB is unable to bind to E2F, allowing E2F to induce transcription of genes driving the cell cycle through G₁ to S phase. [46, 47] Loss of function mutation in both RB alleles is thus advantageous for cancerous cells. TP53 meanwhile, play key role in cellular response to various cellular stresses, including DNA damage. TP53 integrates input from various stress sensors, deciding if cell cycle is allowed to continue or must come to a halt. Alternately if damage is too high, TP53 may commit the cell to apoptosis. Other hallmarks include activation of the cells telomerase genes to initiate replicative immortality. Due to the genome replication process inability to replicate chromosomal ends, each successive cell replication shortens the chromosome slightly. After a finite number of replications, the cell is no longer able to replicate. In adult individuals, only stem cells have active telomerases which extends the chromosomal ends back to their original ends. Cancerous cells must therefore reactivate their telomerases if they are going to continue to proliferate.

First reported involvement of miRNAs in cancer progression was reported by the Croce laboratory in 2002 [48]. Since then, miRNAs have been shown to be involved in all cancer hallmarks [49]. An interesting concept, as alluded to in the discussion on miRNA evolution, is the importance of miRNA in maintaining organismal complexity. As was shown, organisms increase their miRNA complement along with an increase in their complexity, while devolution, species evolving into less complex organisms, such as parasites, lose miRNA genes [50, 51]. In a sense, cancer-cells could be seen as cellular attempts at escaping the confines of complex, multicellular organisms, and becoming more akin to their ancient, less complex and single celled predecessors. One might therefore suppose cancers would also

see a drop in their miRNA complement. Some earlier studies have indeed suggested this to be the case, where Lu et al, 2005 [52], reported a general downregulation of miRNA in cancers.

1.2.1 Metastasis

Metastasis is the process by which cells of a primary tumor disseminate from its site of origin and spread through the body to colonize other organs. Accomplishing this task requires that a tumor cell acquires the ability to disseminate from tissue of origin, survive in circulation, escape the blood vessels at a distant site, then survival and growth to form colonies at distant sites [53]. Many cancer types exhibit specificity in the locations they metastasize, where, according to Stephen Paget's 1889 'seed and soil' hypothesis [54], factors in the cancer cell 'seed' and the distant organ environment 'soil' determine likelihood of metastasis. Anatomical and physiological parameters also factor into where cancers metastasize [53]. The direction of blood flow from the primary tumor determines the first organ reached by circulating tumor cells, CTCs, where their circulation may be arrested by the smaller capillaries of the organ. The local structure of capillary walls in distant organs also plays an important role in CTCs ability to extravasate, or leave the blood stream. Liver capillary walls, for instance, consists of thin, fenestrated endothelium [55], where gaps allow CTCs to pass through. Lung capillary walls, by contrast, consist of tight endothelium. Several genes however, have been identified that allow extravasation of cancer cells even through lung capillaries [55, 56].

The molecular biology underlying the complex morphological and phenotypic developments driving these processes has only recently starting to be understood [43, 57, 58]. A key component is Epithelial Mesenchymal Transition, EMT, whereby immobile and polar epithelial cells alter their morphology to motile and nonpolar mesenchymal cells, allowing them to escape the epithelial layer to the underlying mesenchymal layer [59, 60].

Characteristic of EMT is the loss of E-cadherin and γ -cadherin, while acquiring expression of N-cadherin [61]. EMT plays a key role during embryonic development and is an elaborate process requiring change in expression levels of a myriad of genes. Orchestrating this process are a plethora of transcription factors, including SNAIL, ZEB1, ZEB2 and E47, which directly suppress E-cadherin by repressing its promoter.

MiRNAs have been shown to play a role in EMT by targeting EMT regulating transcription factors. MiR-200, a family of miRNAs, has been shown to target ZEB1 and ZEB2 [62-64]. The miR-200 regulatory network therefore act as repressors of EMT and metastasis, with both clinical and cell line samples showing a correlation between miR-200 levels and the expression of E-cadherin [63, 64], as well as the level of primary tumor dissemination in the

presence of miR-200 overexpression [65]. Other miRNA regulating EMT include miR-9, which promote metastasis by directly target E-cadherin coding mRNA, as well as Leukemia Inhibitory Factor Receptor, LIFR, which suppress metastasis by again targeting YAP, a metastasis promoting gene. Further miR-148a suppresses EMT by targeting Met and Snail, two proteins involved in E-cadherin expression [66], while miR-29c has been shown to stimulate EMT by targeting of PTP4A and GNA13, respective members of the ERK/GSK3 β / β -catenin and AKT/GSK3 β / β -catenin pathways [67]. The p53 induced miR-34a has been shown to target transcription factor SNAIL, with suppression of miR-34a upregulating SNAIL and stimulating invasion and migration. Conversely, increased miR-34a expression downregulates SNAIL and represses invasion and migration [68]. MiR-363 and miR-335 have also been shown to repress EMT by targeting of Sox4, a gene involved in embryonic development [69-71].

MiRNA have also been shown to play roles in other parts of metastasis biology. Mir-21 drives invasion and metastasis by targeting programmed cell death 4, PDCD4, a tumor and invasion suppressor gene, as well as tumor suppressor gene tropomyosin, TPM1, and Maspin, also involved in invasion and metastasis [72, 73]. MiR-182 has been shown to be involved in metastasis by targeting transcription factor FOXO3 and microphthalmia associated transcription factor MTF. MiR-30b and miR-30d are drivers of metastasis by repressing GALNT1 and GALNT2, both suppressors of migration and invasion [74]. An interesting case are the miRNAs miR-551a and miR-483, which prevent invading tumor cell survival at distant site by targeting the gene creatine kinase brain-type (CKB) [75]. CKB is exploited by tumor cells to help survive in the hypoxic tumor environment by phosphorylating creatine to phosphocreatine, used by the cell to replenish its ATP supply.

1.2.2 Colorectal Cancer

Colorectal Cancer (CRC) is the second most prevalent cancer in the western world, with a reported 447,000 new cases in Europe in 2012, and a reported 215,000 deaths [76]. The main cause of death is metastasis in the liver, as well as metastatic spread to lungs and peritoneal cavity [77, 78]. Early detection is a key factor in reducing patient mortality. Colorectal cancer progression is divided into four stages. If the cancer is detected during stage II or earlier, an operation has a 90 % chance of curing the patient of the disease [79]. If the cancer is first detected at stage IV, at which point the tumor has started to progress through the colon wall and disseminate through blood and lymphatic system, the five-year survival rate is 10 % [79]. As with other cancers, heterogeneous underlying genetic disorders cause colorectal cancer, with common risk factors including lack of physical activity, old age, diet, obesity and smoking [80].

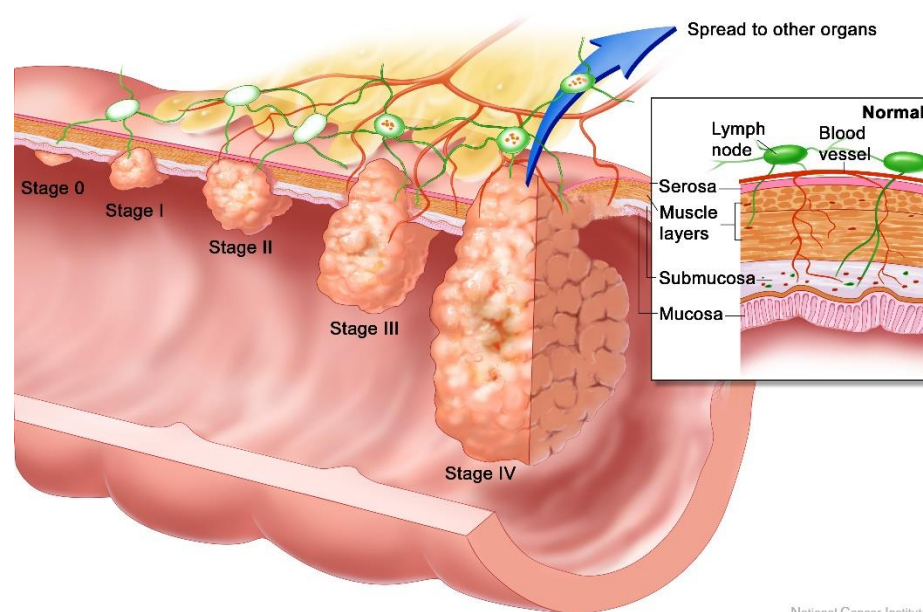


Figure 1.12 Stages of progression of colorectal cancer. If detected by stage II, patients have 90 % likelihood of survival [79]. By stage IV, the tumor has progressed through the colon wall, tumor cells disseminate into blood and lymph vessels, and metastasis to distant sites. Source: National Cancer Institute

Typical molecular pathways of CRC development include chromosomal instability, CIN, microsatellite instability, MSI, and CpG island methylator phenotype, CIMP. Of these, about 65-70 % of CRC patients have CIN [81], which leads to an increase or decrease in both the number of chromosomes as well as alterations in the chromosome structure. By deleting a region of the chromosome, the cell can disable tumor suppressor genes, such as APC, a key component of the Wnt signaling pathway, SMAD4, which is part of the TGF β pathway, and

p53, a key tumor suppressor gene checking for DNA damage and numerous other stress signals.

MSI occurs due to loss of function of mismatch repair genes. The cell is therefore no longer able to repair mismatching base pairs, leading to likely mutations in oncogenes and tumor suppressor genes. Mismatch repair deficiency may be identified by observing abnormalities in repetitive elements of the genome. In the absence of mismatch repair genes, any erroneous insertions by DNA polymerase will not be corrected, leading to frame shift mutations. If this occur in coding region of genes, the protein may will seize functioning.

Lastly, CIMP is caused by methylation of so called CpG sites in promoter regions. At CpG dinucleotides, which is shorthand for *5p-C-phosphate-G-3p*, the cytosine can be methylated, forming 5-methylcytosine. If the methylation occurs inside a gene promoter region, that gene is silenced since the transcription factors are no longer able to locate the promoter. Thus, CpG island methylation may promote cancer in one of two ways, hypomethylation, abnormally low methylation levels which may increase expression of oncogenes, and hypermethylation, abnormally high methylation levels, which may silence tumor suppressor genes.

In a recent study, Guinney et al, 2015 [82], attempted to obtain a consensus view of molecular subtypes of colorectal cancer. They observed preexisting classification systems, and after computational analysis derived at four consensus molecular subtypes, or CMSs, for colorectal cancer. CMS1, or MSI immune, make up 14 % of CRCs, and are characterized by MSI, CIMP, *BRAF* mutations and immune infiltration, and patients experience worse survival rate after relapse. CMS2, or Canonical, make up 37 % of CRCs, and are characterized by high Somatic Copy Number Alterations, or SCNA, as well as WNT and MYC activation. CMS3, or Metabolic, make up 13 % of CRCs, and have mixed MSI status, low SCMA and CIMP status, as well as *KRAS* mutations and metabolic deregulation. CMS4, or Mesenchymal, make up 23 %, and has high SCNA status, and characterized by stromal infiltration, TGF- β activation, angiogenesis and worse patient relapse-free survival.

1.3 MiRNA as biomarkers

MiRNAs have been proposed as biomarkers in cancer. Key miRNA properties make them potentially well suited as biomarkers. MiRNA can be released from the cells that produced them, and are stable in blood and tissue, allowing detection in samples that have been stored over longer periods of time [83]. Methods for detection and quantification are readily available. Furthermore, miRNAs are tissue specific [84, 85], and have been shown to play a role in all hallmarks of cancer [49, 86].

Clinical biomarkers can be classified in several ways [87]. One is based on how they are measured, for instance, if they are measured intracellularly or extracellularly. Extracellular biomarkers can further be divided into their level of invasiveness, from invasive, minimally invasive, and non-invasive. Biomarkers extracted from blood and urine samples would, for instance, be examples of non-invasive extracellular biomarkers. Prostate Specific Antigen, PSA is a non-invasive extracellular biomarker used as a predictor of prostate cancer [88]. Meanwhile, estrogen and hormone receptor levels [89], and mutations in BRCA1/2[90], are biomarkers for breast cancer, but require invasive tissue sampling. Ease of sampling and non-invasiveness are key characteristics to good biomarkers. The stability in of miRNAs in circulation, and the fact that cells can secrete miRNA into the blood stream, make them potentially well suited as biomarkers.

However, as is outlined in Flatmark et al 2016 [2], challenges still abound. A clinical study, after identifying six promising miRNAs involved in colorectal cancer and investigating their expression level in a cohort of 200 stage I-III patients, found few associations between miRNA expression level and the pathological state of patients [91, 92]. Another study found that colorectal cancers did display a clear miRNA expression profile distinct from healthy colon tissue [92], however as of now there are no miRNA colorectal cancer biomarker is in clinical use [87].

Flatmark et al, 2016 [2], describes several possible explanations. As miRNAs involvement in cancer was only discovered in 2002 [48], and their biology is still being unraveled, it is hardly surprising progress has been less than originally hoped. However, several compounding issues complicate the matter. For instance, although miRNA expression is able to distinguish tissue types, not all miRNAs are necessarily tissue specific. MiRNA expression

often has considerable overlap between tissues, with only moderate fold changes for some miRNA.

Secondly, as described earlier, a miRNA may have target sites on multiple genes, and what role a miRNA play in one tissue may entirely different in another tissue type, depending on the circumstances; this is essentially the presence of how many putative targets in the light of miRNA molecules. The biology of miRNAs is therefore highly complex. To be an effective biomarker however, moderate fold changes are not enough. There must be a large fold change between the disease state and the normal state, and this difference must be consistent.

Thirdly, much of the miRNA literature has contradictory results: the same miRNAs might be reported as upregulated and downregulated in the same tissue by different studies. This is likely due to different experimental methods used for detection. Microarray and qRT-PCR are reliant upon correct annotation in the reference database used. As described earlier, miRNA research currently suffers from a plethora of misannotated miRNAs, a problem that is only recently being addressed [29]. Additionally there are limitations when profiling known probes rather than discovery of new molecule. Some of these problems should be rectified by moving to next generation sequencing for profiling, however, studies have found these platforms have problems with reproducibility [93-95]. Efforts are however being made to standardize small RNA workflows to provide more reliable biomarker signatures [96].

1.4 Goals

Metastatic progression from colorectal cancer to primary organs is a dominating cause of cancer related deaths in the western world. And although our knowledge of this process has greatly expanded in the last decades, molecular differences between primary and metastatic colorectal cancer are poorly understood. MiRNAs are candidate biomarkers and key players in cancer progression with demonstrated changes of expression in different cancers. Only few studies have explored miRNA involvement in mCRC, with a recent a study failing to find differentially expressed miRNAs in pCRC and CLM [98]. Furthermore, studies by miRNA biology have been plagued by reliance on miRBase, a non-curated database where as much as two thirds of human entries have been shown to be false annotations [30].

This study therefore set out to investigate miRNA expression in pCRC and CLM, using a small RNA sequencing approach focused on liver metastases and adjacent tissue. Using the highly-curated miRNA database MirGeneDB.org as reference ensured only bona fide miRNA were observed. The identification of signature miRNAs could lay the foundation for future, more in-depth investigations of the role miRNA play in metastatic progression and possibly lead to development of biomarkers

2 Materials and methods

2.1 Clinical studies

This study obtained samples from both the Oslo Colorectal liver Metastasis, COMET, study and the Locally Advanced Rectal Cancer and exfoliated peritoneal tumor cells, LARC-EX, study. The COMET study was a randomized controlled study of laparoscopic versus open liver resection for patients undergoing surgery for CLM. Its overall outcome was to observe patient perioperative morbidity, 5 – year survival, recurrence pattern, inflammatory response, pain level and overall patient health. Secondary objective includes generating a biobank of CLM, versus normal liver, nLi, for each patient, with signed consensus form, to be used in downstream molecular analysis [97]. Tissue samples were snap frozen in liquid nitrogen, and stored at -80°C. Molecular analysis on DNA, RNA and protein was performed. In addition, tumor tissue was made available for the purpose of tissue microarrays. The study included a comprehensive list of clinical patient information, allowing correlation to molecular biological analysis with disease outcome. More than 200 paired CLM and nLi samples were gathered and stored in a biobank. Tissue samples from CLM and nLi were extracted and used in this study.

The LARC-EX study is an ongoing study, with the goal of observing exfoliated peritoneal tumor cells derived from LARC. The study set out to explore the possibility of CRC cells escaping into the peritoneal cavity during tumor growth or during surgical intervention. Kristensen et al, 2008 [98], observed that 19 out of 237 patients with LARC had tumor cells in the peritoneal cavity. This finding correlated with poor patient survival. The LARC-EX study follows up on this finding, whereby patients with LARC patients undergo lavage, or washing of the peritoneal cavity with fluid, before and after surgery, and analysis is performed to ascertain tumor cell exfoliation effect on tumor recurrence and patient outcome. Tissue samples from primary Colorectal Cancer, pCRC, normal Colorectum, nCR, and Peritoneal Cavity, PC, were gathered and stored at -80°C. pCRC and nCR samples were extracted and used in this study.

In addition, next generation sequencing data from previous studies were also obtained and used for validation and comparison purposes, or to supplement our data when sample size was small. Schee et al, 2013 [92] characterized a miRNA expression profile between nCR and pCRC, by deep sequencing a large cohort of 88 pCRC samples. The study found a consistent miRNA expression profile in the pCRC distinct from nCR. In addition, Neerincx et al, 2015 [99], made a differential expression analysis of pCRC versus CLM. Samples

included paired nCR and CLM, with samples from multiple metastatic sites, including liver, lung, ovarian and peritoneal tissues. Their study did not attempt to distinguish different metastatic sites, and where not able to make a distinction, neither globally nor on the individual gene level, between miRNA expression in pCRC and metastatic CRC. Röhr et al, 2013 [100], sequenced paired pCRC, mCRC and nCR from 8 patients, 6 of which were liver and 2 were lymph node metastasis.

2.2 RNA isolation and quality control

To isolate total RNA from patient samples, Qiagen Allprep DNA/RNA/miRNA universal kit was used, which permits simultaneous isolation of genomic DNA and total RNA from one sample. This maximizes yields since one doesn't have to split the samples for separate isolation procedures. In this case, genomic DNA and total RNA was purified from tissues stored at -80°C without stabilizing agent. No tissue sample was larger than 30 mg. Tissue samples were stabilized in 600 µl Buffer RLT Plus, added one 5 mm diameter stainless, RNase free steel bead and subsequently homogenized using TissueLyser LT for 2 x 4 min at 40 Hz. The homogenized lysate was transferred into AllPrep DNA Mini spin columns and centrifuged for 30 s at full speed. The spin column containing genomic DNA was stored at 4°C. Flow through containing total RNA was 80 µl Proteinase K and 35 µl 100% ethanol, and incubated for 10 min. Another 750 µl 100% ethanol was added and 700 µl of this mix was transferred to an RNeasy Mini spin column, centrifuged at full speed until all residual ethanol had passed through. The flow-through was discarded. 500 µl Buffer RPE was added to the column, centrifuged 15 s, then 80 µl DNase I incubation mix was added to the spin column, and incubated for 15 min. In the next step 500 µl Buffer FRN was added and centrifuged for 15 s. As the flow-through contained small RNAs, it was reapplied to the spin column, centrifuged for 15 s, then discarded. 500 µl Buffer RPE was added to the RNeasy Mini spin column, and centrifuged for 15 s, then 500 µl 100 % ethanol was spun through the column, to wash the spin column membrane. Purified total RNA was then eluted in 30 µl RNase free water.

To measure total RNA concentration and check for contaminants, ThermoFisher NanoDrop Spectrophotometer was used. Estimating presence of proteins or phenols is accomplished by looking at ratio of 260 nm to 280 nm absorbance. Nucleic acids absorb at 260 nm, while proteins absorb at 280 nm. For RNA, a ratio above 2.0 is generally said to be pure, if the ratio is considerably lower, the sample may contain protein or phenols. Additionally, a 260 nm to 230 nm absorbance ratio below 2.0 indicates the presence of organic compounds that absorb at 230 nm.

When analyzing RNA, an important consideration is the degree of degradation. Most RNA molecules are unstable, and will rapidly degrade when stored at room temperature. To determine the level of RNA degradation in the samples, Agilent Technologies Bioanalyzer RNA 6000 Nano kit for microcapillary electrophoresis was used. This kit allows analysis of

12 samples per chip, requires a volume of 1 μl and has a quantitative range of 25 – 500 ng / μl . Bioanalyzer function by the same principle as gel electrophoresis, whereby a current is applied over a porous gel, allowing charged molecules to pass through it at a rate based on their molecular weight. Bioanalyzer offers higher sensitivity and specificity than traditional gel electrophoresis. Ribosomal RNA makes up > 80 % of total RNA and typically have a 28S:18S ratio of 2:1. Therefore, one can make the assumption that 28S:18S ratio of around 2:1 represent samples with low degree of degradation, whereas if the ratio is considerably lower, the sample has significant degree of degradation. The degree of degradation may be quantified by a RNA Integrity Number, RIN, where 1 is worst and 10 is best. Although the importance of total RNA degradation for miRNA sequencing is controversial [101], high quality RNA is still important so as to avoid sequencing of degradation fragments.

2.3 NGS Library preparation and sequencing

NGS library was prepared using the TruSeq Small RNA Library Prep protocol. Optimal input for this protocol is 1 µg of total RNA in 5 µl of nuclease free water. The procedure is as follows. The first step is ligating adapters to the 3p and 5p ends of all RNA molecules in the sample. The adaptors are necessary for two reasons, firstly to hybridize complementary primers for the reverse transcription step. Secondly, after transformation to cDNA, they hybridize with flowcell oligos before the bridge amplification step during sequencing.

First 1 µl 3p RNA adapter is mixed with 1 µg total RNA in 5 µl nuclease free water, to a total volume 6 µl, and incubated at 70oC for 2 min. To this 2 µl Ligation buffer 1 µl RNase inhibitor and 1 µl T4 RNA Ligase 2, for a total volume of 10 µl are added respectively. The mix is incubated at 28oC for 1 hour. After incubation, 1 µl Stop Solution is added, and incubation continued at 28oC for 15 min. Secondly, 5p-adapter ligation mix was prepared by adding 1.1 µl per sample of 5pprime adapter, incubating at 70oC for 2 min, adding 1.1 µl per sample 10 mM ATP and 1.1 µl per sample T4 RNA Ligase. Of this mix 3 µl was added to the 3p-adapter mixture, for a total volume of 14 µl, and incubated at 28 µl for 1 hour.

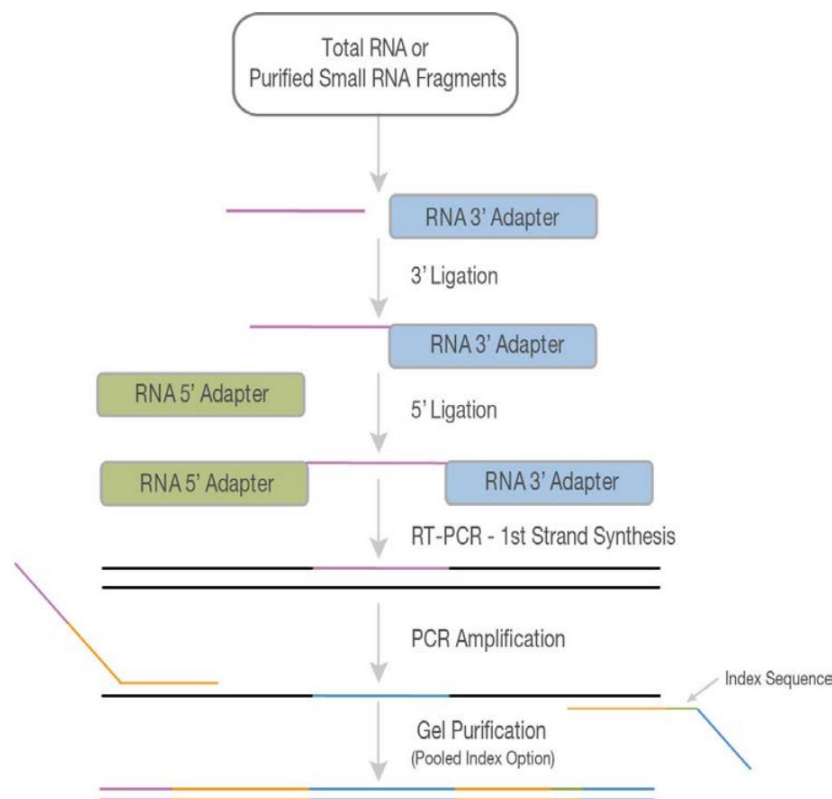


Figure 2.1 TrueSeq small RNA sample preparation TruSeq® Small RNA Sample Preparation Guide
Illumina

The second step is generating cDNA constructs from the RNA fragments ligated with 3p and 5p adapters. By using primers that anneal to the adapter ends, this process selectively amplifies fragments containing 3p- and 5p-adapters. For each sample, 6 µl of the prepared adapter ligated RNA library and 1 µl RNA RT Primer is added to a new 200 µl PCR tube, and incubated at 70°C for 2 min. To this mix is added 2 µl 5X First Strand Buffer, 0.5 µl 12.5 mM dNTP mix, 1 µl 100 mM DTT, 1 µl RNase Inhibitor and SuperScript II Reverse Transcriptase, for a total volume of 12.5 µl. The reverse transcription mix is incubated at 50°C for 1 hour.

After conversion to cDNA comes the library amplification step using PCR. To each library is added 8.5 µl Ultrapure Water, 25 µl PCR Mix, 2 µl RNA PCR Primer and 2 µl RNA PCR Primer Index, for a total volume of 50 µl. This mix is placed on a thermal cycler with the following program:

- Preheated at 100°C
- 98°C for 30 s
- 11 cycles with:
 - o 98°C for 10 s
 - o 60°C for 30 s
 - o 72°C for 15 s
- 72°C for 10 min
- Hold at 4°C

To observe if library preparation was successful, each library was run through an Agilent Bioanalyzer High Sensitivity DNA Assay Chip. A successful library preparation will show up as a distinct peak at length ~22 bp plus adapter sequences.

Successfully prepared libraries were then sequenced using Illumina HiSeq 2500 High Throughput Sequencer. These sequencers function on the principle of sequencing by synthesis. The prepared library template sequences are applied to an Illumina HiSeq Flow Cell, on the surface of which are oligonucleotides that hybridize with the templates adapter sequences, causing them to attach to the flow cell surface. Modified nucleotides containing fluorescent terminator caps and DNA polymerase is then added to the flow cell. DNA polymerase will add these capped terminators to the template sequences, but because of the cap, only one nucleotide is added each round. Each nucleotide has its own distinct fluorescent

color, and a camera identifies the added nucleotide. At the end of the round, the cap is removed, and another nucleotide is added. Computer software, in a process called base calling, is able to identify and keep track of all nucleotides in the library. All template strands are stored as 'reads'.

One problem with this approach is that the fluorescent light from a single nucleotide is too weak to be detected by a camera. Therefore, before sequencing by synthesis, each DNA template must go through a process called bridge amplification, where each individual template is amplified into clusters of thousands of identical sequences in close proximity to each other. Therefore, the surface of the flow cell is littered with clusters of DNA templates which light up with a distinct color detectable by a camera. This, however, leads to its own problem, since it depends on all template strands in a cluster being in phase, where they all add the same nucleotide at the same time. If a cluster gets out of sync, the signal deteriorates, and reliability of base calling decline. Every read output by the Illumina sequencer contains information of both the read sequence and the quality of each base in the read.

2.4 Preprocessing and read mapping

For each library, the Illumina sequencer outputs a FASTQ file, containing four lines for every read. The first line is the read ID, second line is the read sequence, third line is redundant while the fourth line contains the Phred quality score for its respective base in line two. A Phred quality score goes from 1 to 40, where a score of 30 indicates 1 in 1,000 probability that the base is called wrong, or 99.9 % accuracy of the base call. A score of 40 indicates 1 in 10,000 probability that the base is called wrong, or 99.99 % accuracy of the base call.

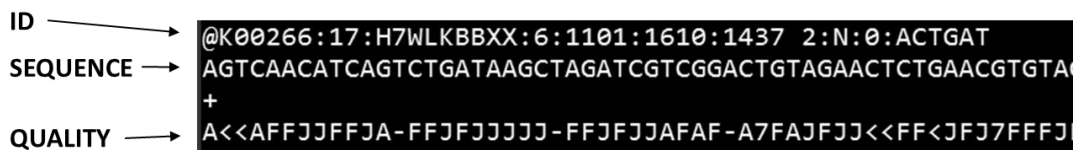


Figure 2.2 FASTQ format First line is read ID, second line is read sequence, third is redundant while fourth line contain quality information for each base call.

The first step in preprocessing is to remove 3p-adapter sequences from all reads. The sequencing by synthesis step starts sequencing in the 5p-end of the actual template strand, and continues into the 3p-adapter sequence. Therefore, one must remove these from all the sequences in the FASTQ file. In this case, `fastx_clipper` from `fastx-toolkit` was used to remove the adapter sequence. The second step is quality trimming, where `fastq_quality_trimmer` was used to remove reads with an average Phred score less than 33. When this is done, the quality information in the FASTQ file is no longer needed, so the FASTQ file was converted to FASTA format, where only read ID and sequence is annotated. After this, all reads shorter than 19 or larger than 26 bp was removed.



Figure 2.3 Removing 3p adapter sequence

Once the preprocessing step is complete, one has a FASTA file for each sample, which contain only those reads where the adapter has been removed, whose base calling accuracy was sufficiently high, and length after adapter trimming between 19 and 26 base pairs. The

next step is mapping the reads against a reference sequence. Since in this case only miRNAs where of interest, the reads where mapped against pri-miRNA sequences with 30 nucleotides 5p- and 3p- derived from MirGeneDB [29], the curated database of miRNA genes. NGS read mapping software goes through each read in the FASTA file and checks if there is a match in the genome. Parameters can be set to determine the maximum numbers of mismatches in the first X bases, and what happens in case a single read matches multiple times. In this case, the read aligner *bowtie* (version 1.0.0) (1) was used, and parameters set to allow 0 mismatches in the first 18 nucleotides of the read, and in cases where a read maps to multiple loci, the read is mapped to each loci.

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Figure 2.4 summarizeOverlaps Union parameter was chosen. <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>

Bowtie outputs a SAM file, or Sequence Alignment/Map Format, containing coordinates for each read, indicating where they mapped against the reference sequence. The SAM file also contains information about location and number of mismatches. The information stored in a SAM file can then be used to generate a count matrix. The SAM file coordinates is compared against annotated coordinates of genes of interest stored in a GFF, or General Feature Format, file. If the coordinates of a read overlap the annotated coordinates of a gene, the count number of said gene is increased by one. In this case, the summarizeOverlaps method from Bioconductor was used to count miRNA genes, using a GFF file derived from annotated

mature miRNA genes in MirGeneDB. SummarizeOverlaps offer several counting modes, in this case, the Union mode was selected whereby a gene is counted assuming any part of a read overlaps the gene. The derived count matrix contains columns for each sample and rows for each miRNA gene, with integers for the number of reads corresponding to a mature miRNA gene in each sample. Notably, miRNA with identical mature sequences will receive the same number of counts. The count matrix is the starting point for all remaining downstream analysis.

2.5 Sample distances and hierarchical clustering

To assess how samples compare to one another and if miRNA expression is able to distinguish between tissue types, hierarchical clustering analysis was performed. Clustering algorithms estimate distance between samples based on multidimensional gene counts data. In this case, Euclidean distance and *complete-linkage* parameters were chosen, grouping samples based on similarity of the furthest sample pair. Agglomerative clustering then produce a dendrogram, successively grouping samples together, starting with the most similar samples, then segmenting the data with gradually larger groups of increasingly distant similarity.

An important source of bias in RNAseq data is the increase in variance of highly expressed genes compared to lower expressed genes. Since only those genes with significant variance across the mean will actually influence clustering outcome, highly expressed genes would infer an outsized influence in untransformed data. To account for this, data was normalized by DESeq2 estimated sizeFactors and log2 transformed.

Four clustering diagrams where made. First comparing nCR and nLi samples, secondly comparing pCRC and nCR, next CLM an nLi were compared and finally pCRC and CLM.

2.6 Differential expression analysis

DESeq2 was used to analyze differentially expressed miRNA between tissue types [102]. In brief, DESeq2 models gene-wise differential expression between sample groups as follows. As input, DESeq2 takes raw, non-transformed counts. First, within-group variation in gene expression is estimated. DESeq2 makes the assumption that genes with similar mean expression also have similar variance. Dispersion is first independently estimated for each individual gene, represented by black dots in **Figure 2.5**. Then, a fitted curve is made from this individual gene data, providing an expected dispersion value per mean expression rate. The gene-wise estimated dispersion is then shrunk to more closely resemble the fitted curve, to give the final dispersion (arrows). Using this method, genes far below the fitted curve are given a substantial increase in dispersion, lowering the statistical power of the potential differential expression of said gene, thus lowering the probability of false positives. On the flip side, genes with dispersion far above the fitted curve are not shrunk, as these may be outliers. Shrinking the dispersion estimate would therefore increase the risk of a false positive, and the original gene-wise dispersion estimate is used instead. Such genes are in **Figure 2.5** shown as a black dot surrounded by a circle.

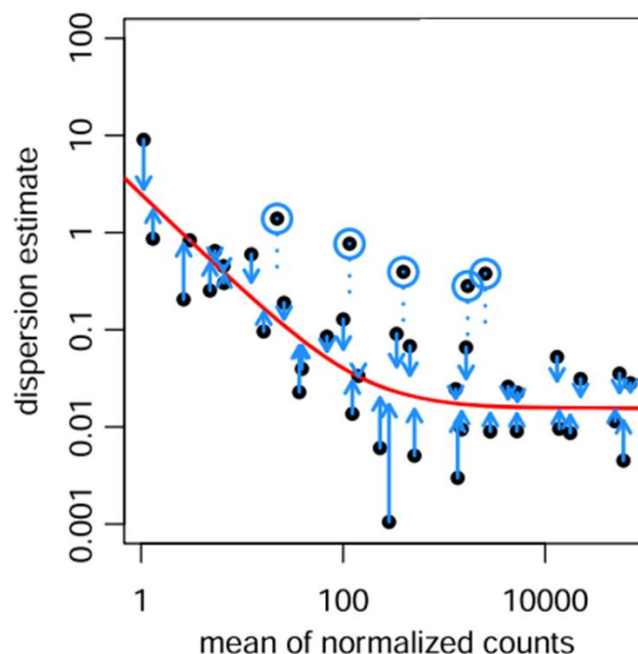


Figure 2.5 DESeq2 shrinkage of dispersion. Black dots are gene-wise dispersions, line shows the fitted curve while arrows show shrunk estimates. Circles around dots represent gene-wise dispersions that did not undergo shrinkage. [102]

Another issue with HTS data is that genes with low count means see stronger variance in LFC as compared to genes with higher mean counts. As explained in the DESeq2 paper Love et al, 2016 [102], this is caused by using counts of discrete values. Low count numbers are inherently noisier than high count numbers. Therefore, DESeq2 reduces the LFC towards zero in cases where there is little information regarding the gene, either due to high dispersion, low counts, or few degrees of freedom. This means it's possible for genes with similar mean expression but different dispersions receiving different degrees of reduction in fold change, preserving LFC in genes with little dispersion while reducing LFC for genes with high dispersion. Net result is decreasing risk of both false positives and false negatives. It is these shrunken LFC values which are used in further downstream tests.

Each gene then undergoes a Wald test, and then independent filtering to address the multiple testing problem. This is important as with a sufficiently large samples size, negligible LFCs will eventually be detected a significant. This results in an adjusted p-value as an estimation of significance. In the case of large sample sizes, a threshold is also set so that only genes showing sufficient difference to be *biologically significant* are considered.

2.6.1 Target prediction

TargetScan was used to estimate targets for miRNA of interest. TargetScan scans through known RNA molecules checking for sites matching the canonical 7mer-A1, 7mer-m8 or 8mer sites targeted by miRNA, as shown in **Figure 1.4**. Target sites containing mismatches are also considered if they contain 3p pairing. Sites are ranked according to their estimated targeting efficacy, and conservation of those targets. [103]

2.7 IsomiRs

Generating a count matrix as described above, by counting a miRNA gene if any part of a mapped read overlaps the genes annotated loci, will inevitably lose some of the information the sequencer provides. NGS data contain information on both mismatches and 5p- or 3p- elongations and truncations. To account for this lost information, a new count matrix of isomiRs was created. IsomiRs were defined as all fragments mapping to MirGeneDB annotated human pri-miRNA sequences, excluding reads identical to canonical miRNA. Bowtie parameters were set to allow 3 mismatches anywhere along the read. A count matrix was made from all samples prepared in this study, along with 3 nCR and 6 pCRC samples from Schee et al 2013 [92], for a total of 33 samples. Only fragments found in at least 50 % of samples were added to the count matrix. Finally, differential expression analysis was performed as described above, using DESeq2.

2.8 Sequential motifs

Sequential and structural motifs were defined as follows. Basal UG motif consists of uracil followed by cytosine positioned 14 nucleotides upstream of Drosha 5p-arm cut site. Apical UGU/GUG motif consists of either a UGU or a GUG sequence at position 1, 2 or 3 after Dicer 5p cut site. Flanking CNNC stem motif consist of a cytosine followed by two random nucleotides, then followed by cytosine, at either position 16, 17, or 18 downstream of Drosha 3p-arm cut site. For the Mismatched GHG motif, the miRNA stem was defined as 35 (± 1) nucleotides long, counting from bottom of hairpin stem, ending at Dicer 5p cut site, see **Figure 2.6**. Mismatched GHG motif resides at position 7-8-9, defined as Watson Crick pairing at position 7 and 9, and a wobble, or non-pairing, at position 8. To verify these motifs, pri-miRNA sequences for every gene annotated in MirGeneDB was used. Each pri-miRNA sequence had 30 nucleotides from their genomic loci added to 5p and 3p ends. For basal UG motif, flanking CNNC motif and apical UGU/GUG motif, the number of miRNA with the respective motifs were counted at each position. For mismatched GHG motif, pri-miRNA sequence was folded using RNAfold (Lorenz et al, 2011 [104]), and the number of miRNA genes with mismatched GHG motifs at each location in the lower hairpin stem was counted.

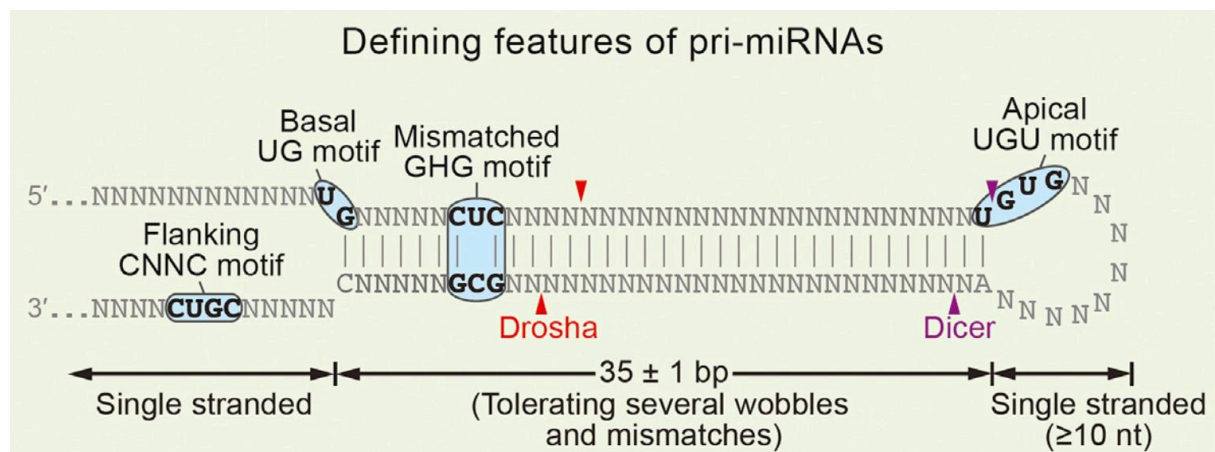


Figure 2.6 Defining pri-miRNA Sequential and Structural Motifs Basal UG motif starting at -14 nucleotides from Drosha 5p cut site. Flanking CNNC motif starting at either 16, 17 or 18 downstream of Drosha 3p cut site. Apical UGU/GUG motif at position 1 after Dicer 5p cut site. Mismatched GHG motif at position 7-8-9 in lower stem, consisting of Watson Crick pairing at position 7 and 9, and mismatch, or wobble, at position 8. Fang et al, 2015 [41].

3 Results

3.1 RNA extraction

RNA extraction and quality assessment using Bioanalyzer and Nanodrop showed a general trend of poor RIN values in nLi and CLM samples. Typically, only samples with RIN above 8.0 are used for sequencing, but in order to obtain sufficient material this threshold was lowered to 5.0. Previous studies have shown high quality NGS libraries can be generated even with lower quality RIN samples [101]. By The cause of disparaging RIN values may be due to differences in the tissue samples themselves. During the sample gathering stage, no sample was left at room temperature for more than 30 minutes. Studies have shown dramatic drops in RIN quality does not occur until 60 minutes [105]. In remaining preparatory steps, samples were at stored at -80°C and kept at dry ice using best practices procedures for handling RNA.

In the end, 9 paired nLi and CLM COMET samples and 3 paired nCR and pCRC, for a total of 24 samples, were found to be of sufficient quality for sequencing.

Returned sequencing data showed significant differences in total number of reads between samples. Total reads ranged from more than 25 million to less than 3 million reads, with an average of 9.5 million reads. This may present a problem in downstream analysis, since some steps, including clustering analysis, requires normalized counts. Widely disparaging total read numbers may induce a bias in the results after normalization. After adapter trimming and quality control, an average of 89 % of total reads remained, while after removing reads longer than 26 bp, an average of 53 % of total reads remained. Of the processed reads, an average of 86 % mapped to MirGeneDB.

Sample Name	Tissue	RIN	260/280	260/230	Conc. (µg/µl)	Initial Reads	After Clipping	Quality Filtering	Reads < 26 bp	Unique	Unique MirGeneDB	Mapped MirGeneDB	Reads > 26 bp mapped
COMET 0003M	liver metastasis	6,3	1,9	2,1	194	19518920	99 %	91 %	64 %	172527	16467	60 %	93 %
COMET 0003N	liver normal	7,2	1,9	1,6	211	3661520	99 %	87 %	60 %	59717	8161	52 %	88 %
COMET 0011M	liver metastasis	8,2	2,0	2,1	197	4120967	99 %	92 %	75 %	56375	8420	70 %	94 %
COMET 0011N	liver normal	6,4	1,9	2,2	198	5901934	99 %	89 %	67 %	73657	9321	62 %	92 %
COMET 0014M	liver metastasis	5,0	2,0	2,2	198	4881422	99 %	90 %	57 %	77506	8507	50 %	87 %
COMET 0014N	liver normal	8,6	1,9	1,4	211	2280732	98 %	84 %	47 %	50305	5589	37 %	80 %
COMET 0016M	liver metastasis	8,3	2,0	2,0	198	3607069	99 %	85 %	49 %	84350	7718	37 %	76 %
COMET 0016N	liver normal	8,5	1,9	1,4	203	3002514	99 %	88 %	57 %	62931	6522	49 %	86 %
COMET 0026M	liver metastasis	6,7	1,9	2,3	199	12822692	99 %	88 %	65 %	130104	11819	57 %	88 %
COMET 0026N	liver normal	6,7	2,0	1,6	199	2292168	98 %	90 %	50 %	39897	5595	44 %	87 %
COMET 0027M	liver metastasis	7,2	1,9	2,1	203	16528813	99 %	90 %	44 %	146162	13515	39 %	89 %
COMET 0027N	liver normal	7,2	2,0	1,7	197	5514267	99 %	93 %	31 %	56216	6599	27 %	87 %
COMET 0028M	liver metastasis	6,1	1,9	2,0	208	10599617	99 %	87 %	56 %	146909	11299	47 %	84 %
COMET 0028N	liver normal	6,4	2,0	1,9	204	4145103	99 %	90 %	53 %	58600	7284	47 %	89 %
COMET 0035M	liver metastasis	6,1	1,9	2,2	209	12440004	99 %	87 %	67 %	154124	13069	58 %	87 %
COMET 0035N	liver normal	6,0	1,8	2,1	200	23619240	99 %	95 %	11 %	96972	7712	9 %	79 %
COMET 0059M	liver metastasis	7,0	2,0	2,0	215	15186554	99 %	85 %	57 %	216221	12069	46 %	81 %
COMET 0059N	liver normal	6,9	2,0	2,2	194	4272569	99 %	89 %	43 %	55844	6969	38 %	88 %
LARC EX 115T	rectum tumor	6,5	2,0	1,89	198	17554448	99 %	83 %	38 %	439527	10383	23 %	61 %
LARC EX 115N	rectum normal	5,0	2,0	2,1	200	8021418	99 %	90 %	47 %	132400	9701	39 %	81 %
LARC EX 138T	rectum tumor	9,2	2,0	1,81	197	8588650	99 %	90 %	52 %	109505	10411	46 %	88 %
LARC EX 138N	rectum normal	5,0	1,7	2,1	198	9282185	99 %	93 %	71 %	80670	10764	66 %	93 %
LARC EX 154T	rectum tumor	5,0	2,0	2,0	201	12482692	99 %	92 %	50 %	130036	10949	44 %	89 %
LARC EX 154N	rectum normal	5,0	2,0	2,0	214	18760603	99 %	93 %	58 %	235570	12232	50 %	86 %
AVERAGE		7	2	2	202	9545254	99 %	89 %	53 %	117113	9331	46 %	86 %

Table 3.1 Result of RNA extraction and sequencing

3.2 NGS-results

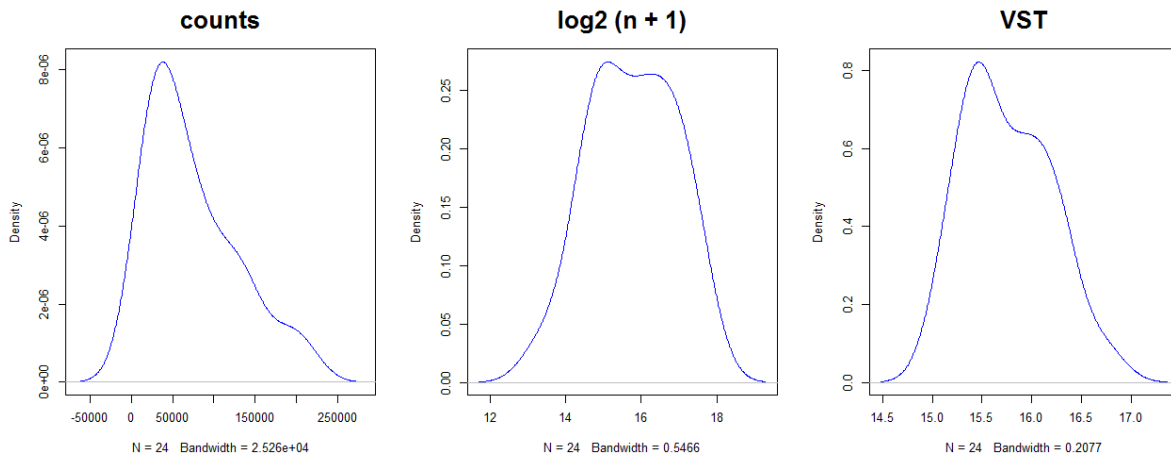


Figure 3.1 Density Plots Density plots of miRNA expression of a randomly chosen gene across 24 samples. Plot of raw counts, $\log_2(n + 1)$ transformed counts and Variance Stabilizing Transformation counts.

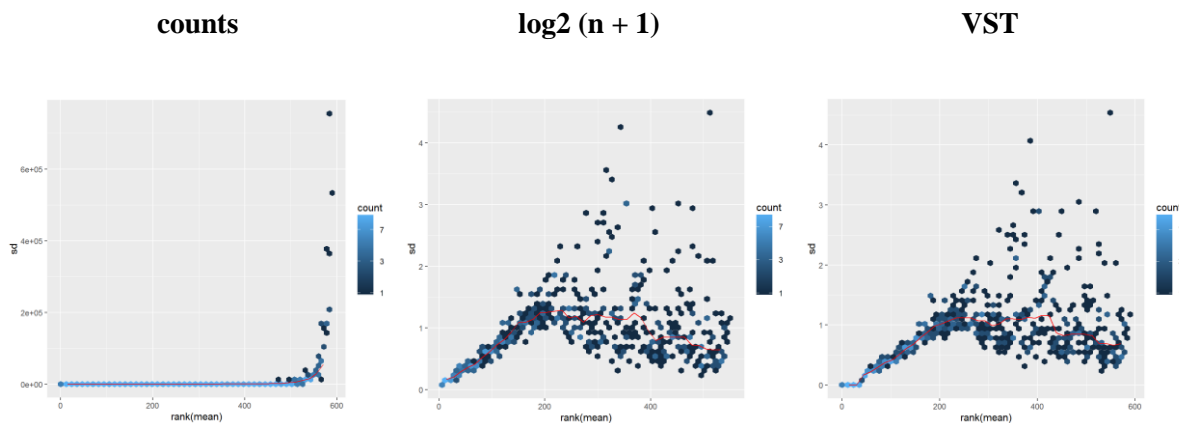


Figure 3.2 MeanSdPlot Plot of per-gene standard deviation versus rank of mean expression across all 24 samples. Plot of raw counts, $\log_2(n + 1)$ transformed counts and Variance Stabilizing Transformation counts.

Density plots of a randomly chosen miRNA gene in the 24 samples indicate that raw counts have a leftward skewed distribution with a significant tail of more extreme values. Both \log_2 transformation and Variance Stabilizing Transformed data appear to remove these more extreme values, as well as the leftward skewed distribution.

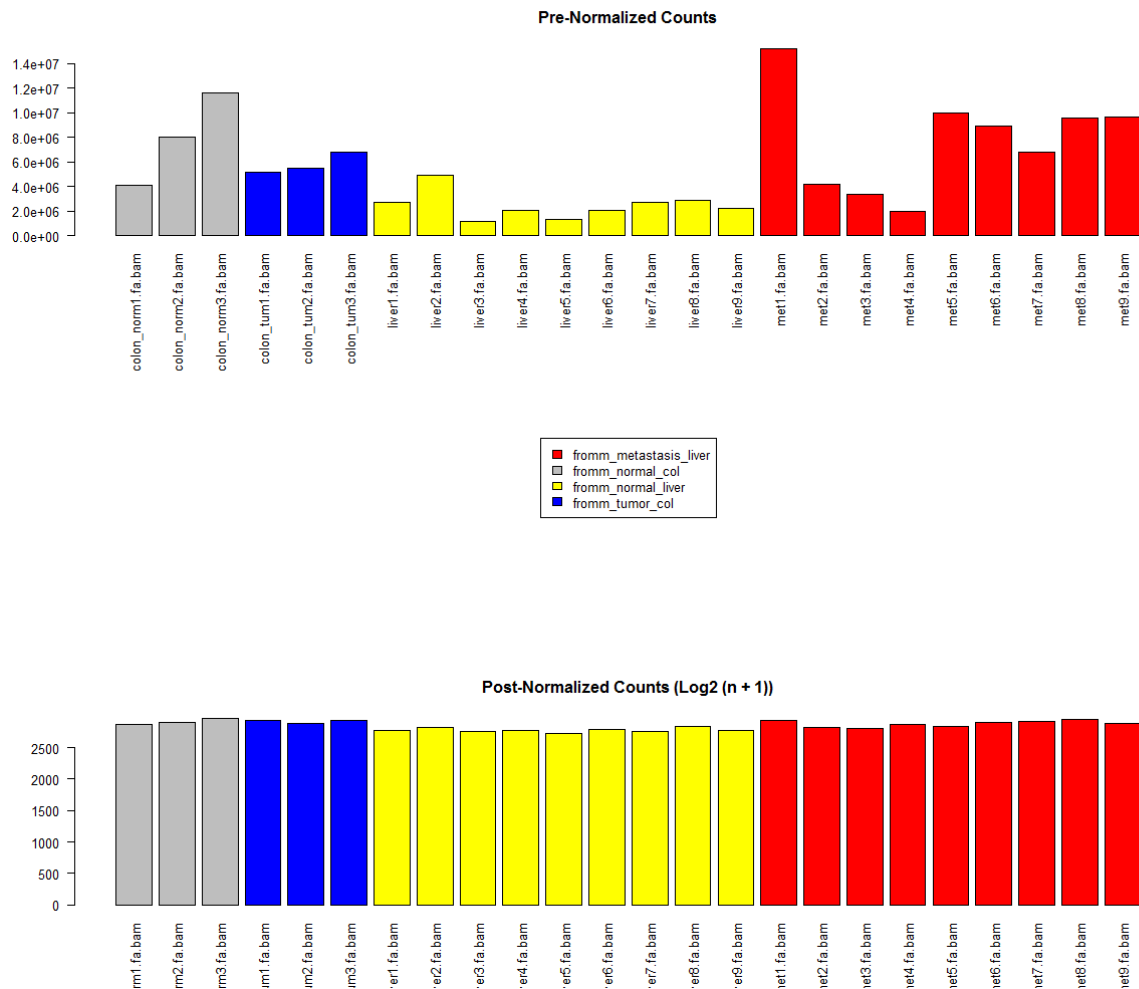


Figure 3.3 Counts per Sample Barplot of pre- and post-normalized counts for all 24 samples. Pre-normalized counts are summation of raw counts for all genes in each sample, while post-normalized counts is $\text{Log}_2(n + 1)$, where counts have been normalized on DESeq2 sample estimated Size Factors, before transformation.

Plots of per-gene standard deviation versus rank of mean expression (**Figure 3.2**) show a huge degree of heteroscedasticity in raw counts, where standard deviation rise dramatically among highly expressed genes. This would introduce a bias for data exploratory analysis, where highly expressed genes would infer an undue influence on the result. Interestingly, both the Log_2 transformation and the variance stabilizing transformed data appear equally effective at removing this heteroscedasticity. This is at odds with previously published literature suggesting a Log_2 transformation itself will induce a standard deviation peak at lower mean counts, see Love et al, 2014 [102]. Since no such difference was apparent in our data, the simple Log_2 transformed data was chosen for downstream data exploratory analysis.

The heterogeneity observed in total read numbers between samples was still present in total count numbers shown in **Figure 3.3**. Post-Normalized $\text{Log}_2(n + 1)$ counts are more homogenous.

3.3 Clustering and sample distance

3.3.1 nCR and nLi are distinct from each other

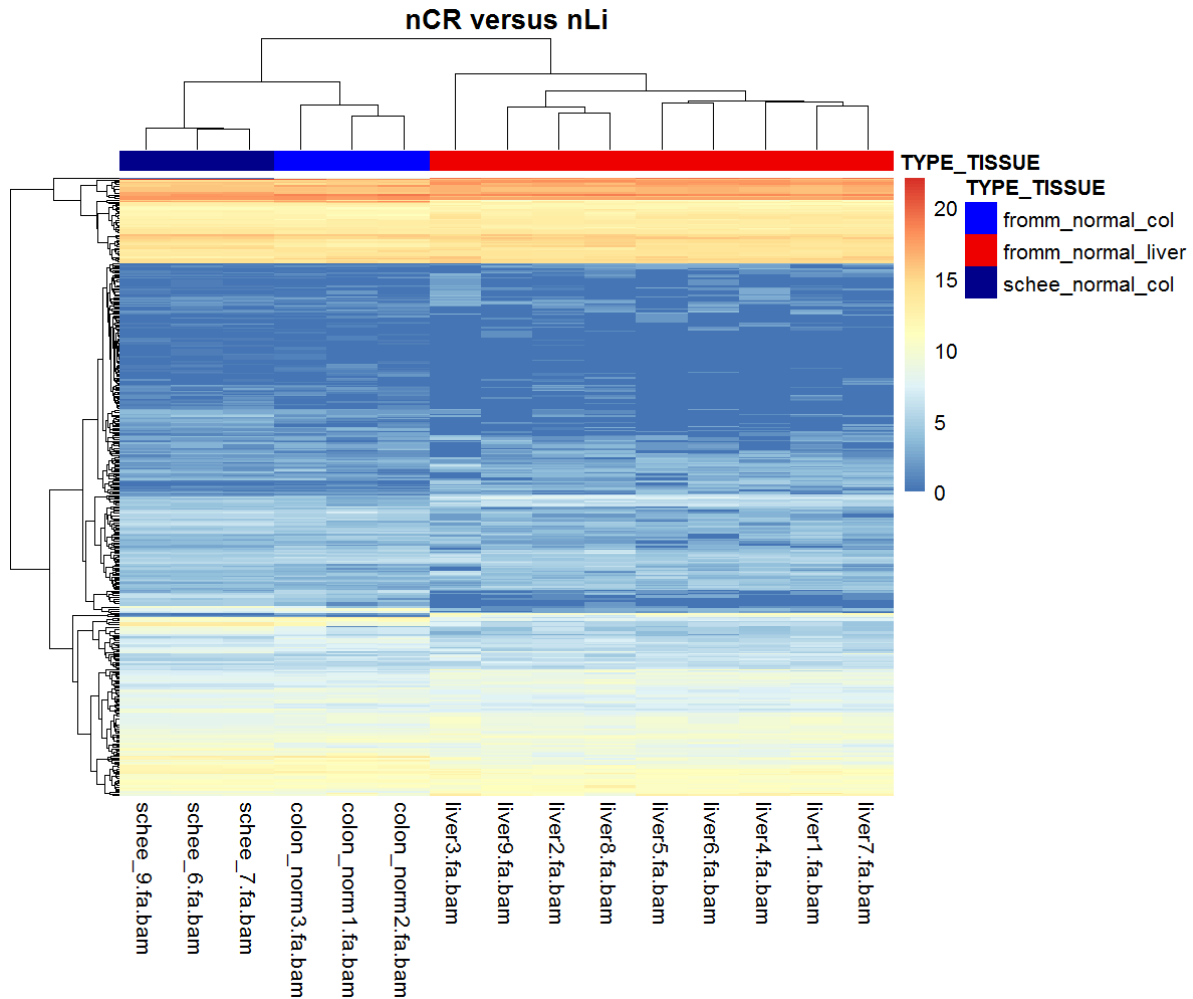


Figure 3.4 Clustering nCR vs nLi Clustering of log₂-transformed miRNA expression levels between normal colorectum and normal liver. MiRNA expression is size-factor normalized and log₂ transformed, clustering based on Euclidean distance and complete-linkage. Sample distance represented by top dendrogram, longer horizontal line between two samples mean longer distance. Samples include 9 nLi and 3 nCR from this study, and 3 nCR from Schee et al, 2013 [92].

Clustering of normal colorectum and normal liver reveals a clear separation, indicating a distinct global miRNA expression profile in each tissue.

3.3.2 pCRC is distinct from nCR

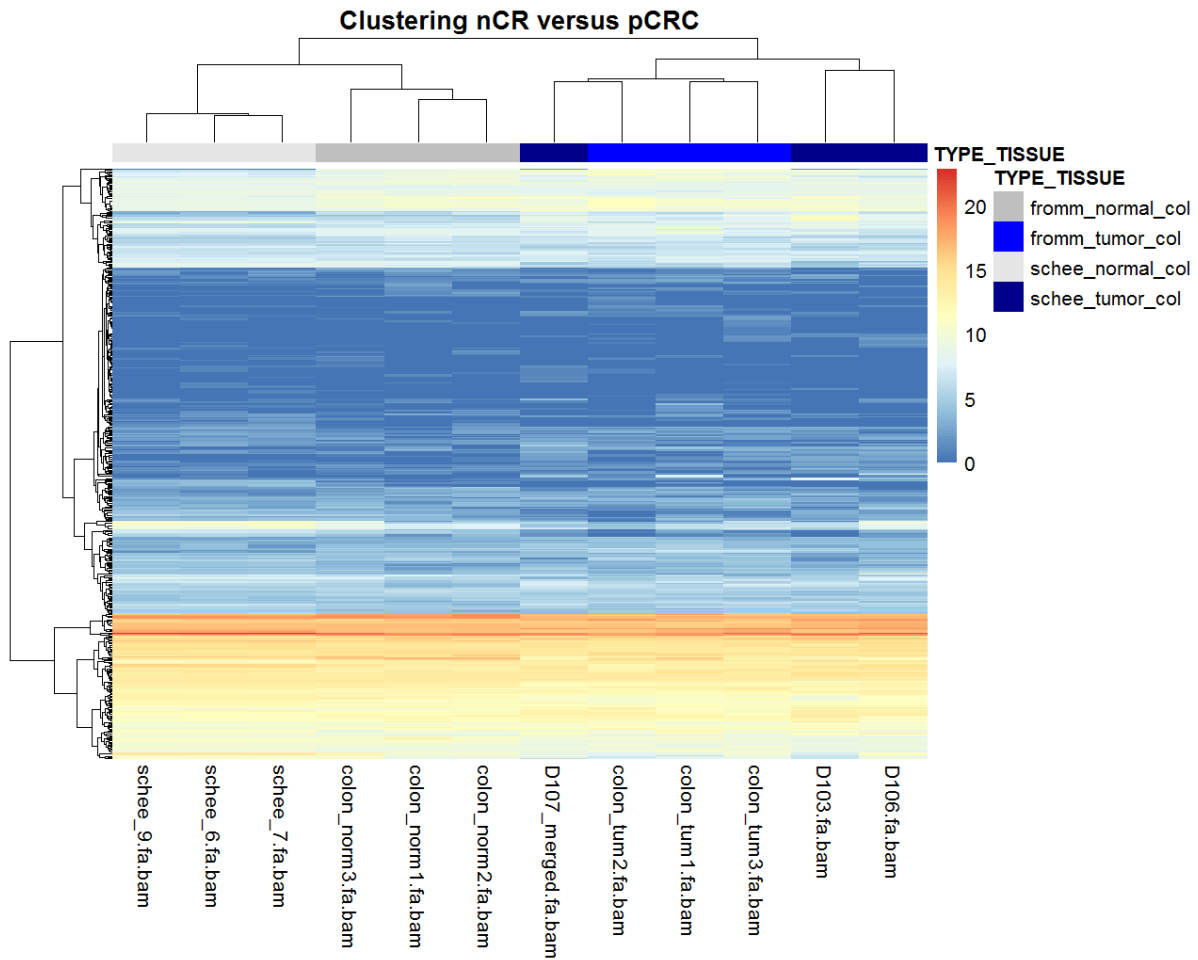


Figure 3.5 Clustering nCR vs pCRC Clustering of log₂-transformed miRNA expression levels between normal colorectum and primary tumor. MiRNA expression is size-factor normalized and log₂ transformed, clustering based on Euclidean distance and complete-linkage. Sample distance represented by top dendrogram, longer horizontal line between two samples mean longer distance. Samples include 3 nCR and 3 pCRC from this study, as well as 3 nCR and 3 randomly chosen pCRC from Schee et al, 2013

Clustering of normal colorectum and primary tumor showed complete separation, indicating a distinct global miRNA expression profile in each tissue. pCRC samples from both studies do not separate, while nCR samples separate into two groups from their respective studies.

3.3.3 CLM is distinct from nLi

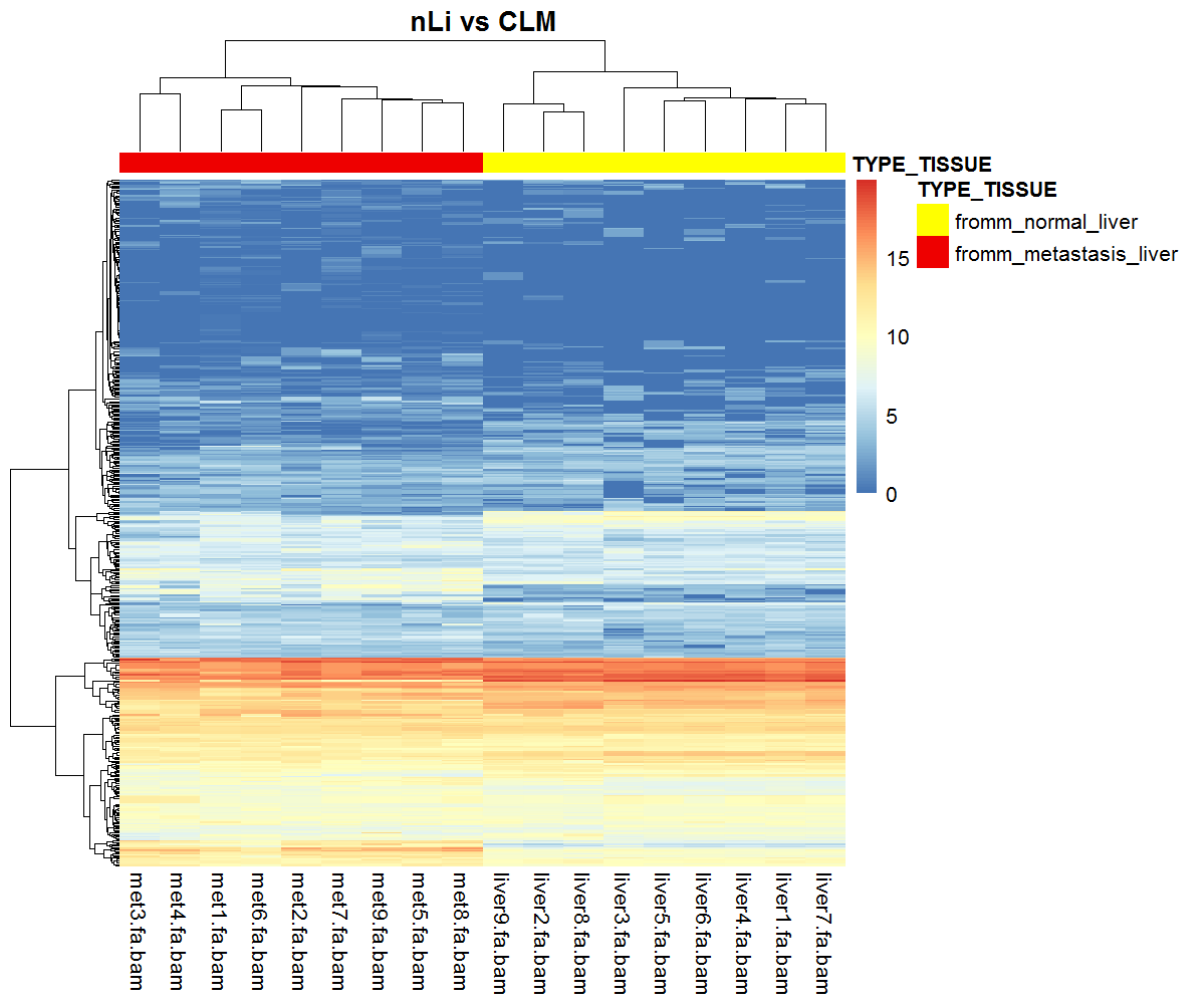


Figure 3.6 Clustering nLi vs CLM Clustering of log₂-transformed miRNA expression levels between nLi and pCRC. MiRNA expression is size-factor normalized and log₂ transformed, clustering based on Euclidean distance and complete-linkage. Sample distance represented by top dendrogram, longer horizontal line between two samples mean longer distance. All 18 samples derived from this study.

Complete separation is observed between normal liver and liver metastasis, indicating a distinct global miRNA expression profile in each tissue.

3.3.4 No distinction of pCRC and CLM

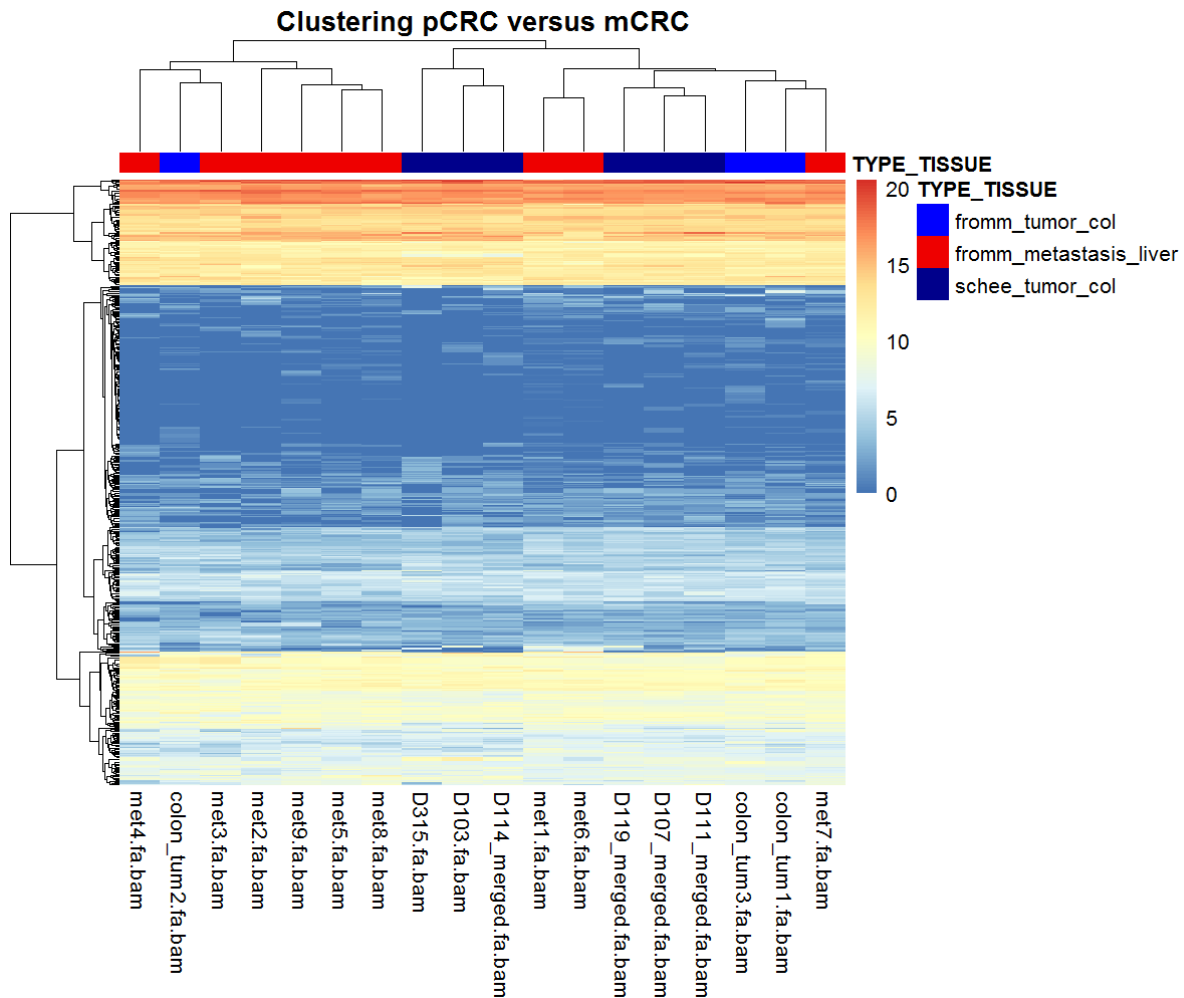
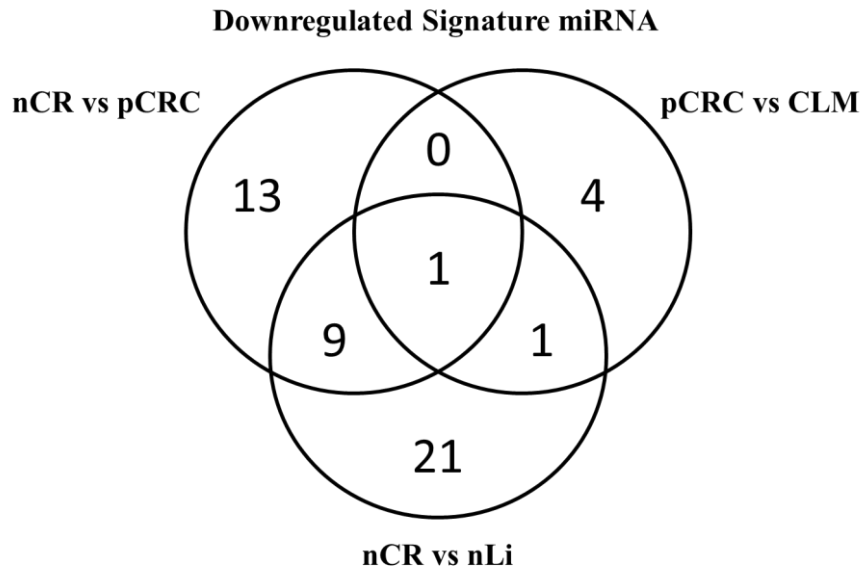


Figure 3.7 Clustering pCRC vs mCRC Clustering of log₂-transformed miRNA expression levels between pCRC and nCRC. MiRNA expression is size-factor normalized and log₂ transformed, clustering based on Euclidean distance and complete-linkage. Sample distance represented by top dendrogram, longer horizontal line between two samples mean longer distance. 12 samples are from this study, as well as six randomly chosen primary tumor samples from Schee et al, 2013

Clustering of primary colorectal cancer tissue and colorectal derived liver metastasis tissue is not able to separate them into two distinct groups, while clustering of colorectal derived liver metastasis and normal liver tissue separate into two distinct groups. This would indicate global miRNA expression profile in liver metastasis predominantly resemble the primary tumors. Global miRNA expression may therefore not dramatically alter as tumor cells metastasize.

3.4 Differential expression

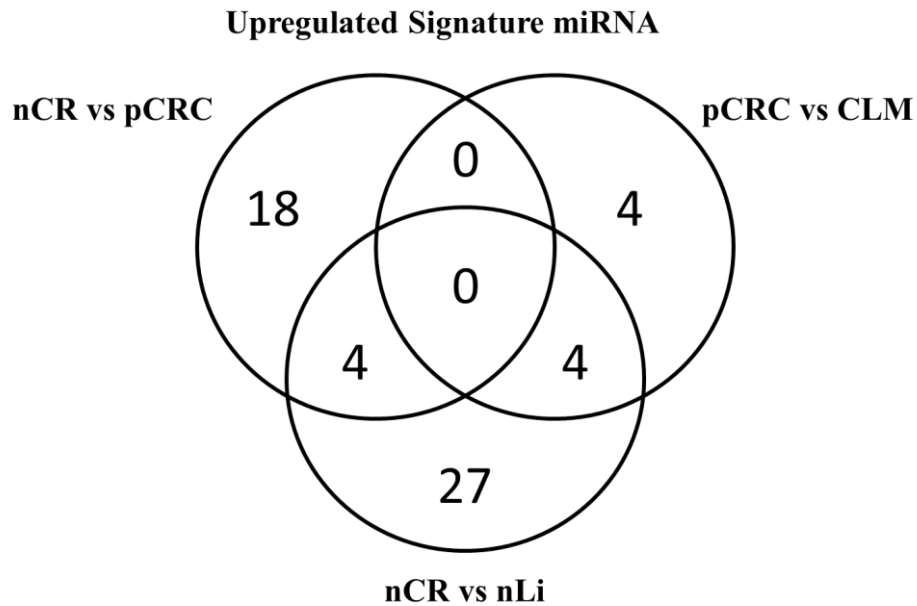
3.4.1 Venn downregulated signature miRNA



nCR vs pCRC	nCR vs pCRC nCR vs nLi	pCRC vs CLM	pCRC vs CLM nCR vs nLi	nCR vs pCRC pCRC vs CLM	All	nCR vs nLi
Mir-10-P3b	Mir-338-P1	Mir-7-P2	Mir-146-P1	None	Mir-10-P1b	Mir-92-P3
Mir-148-P3	Mir-133-P2	Mir-7-P1				Mir-96-P2
Mir-378	Mir-92-P4	Mir-146-P2				Mir-196-P2
Mir-15-P1c	Mir-133-P3	Mir-7-P3				Mir-221-P2
Mir-26-P2	Mir-490					Mir-10-P1a
Mir-26-P1	Mir-145					Mir-196-P1
Mir-10-P3c	Mir-15-P2c					Mir-221-P1
Mir-574	Mir-133-P1					Mir-8-P2b
Mir-10-P3a	Mir-143					Mir-155
Mir-28-P1						Mir-203

Figure 3.8 Venn Diagram of downregulated signature miRNA. Table show top 10 significance level of downregulated signature miRNA

3.4.2 Venn upregulated miRNA



nCR vs pCRC	nCR vs pCRC nCR vs nLi	pCRC vs CLM	pCRC vs CLM nCR vs nLi	nCR vs pCRC pCRC vs CLM	All	nCR vs nLi
Mir-17-P2a	Mir-15-P1d	Mir-150	Mir-10-P3b	None	None	Mir-30-P2b
Mir-224	Mir-92-P1b	Mir-10-P3a	Mir-423			Mir-154-P23
Mir-17-P2b	Mir-148-P1	Mir-1247	Mir-335			Mir-345
Mir-96-P2	Mir-92-P1a	Mir-339	Mir-10-P3c			Mir-885
Mir-135-P3						Mir-122
Mir-29-P2a						Mir-455
Mir-19-P2b						Mir-130-P1b
Mir-21						Mir-197
Mir-17-P4						Mir-941-P2
Mir-17-P1a						Mir-455

Figure 3.9 Venn diagram of upregulated signature miRNA. Table show top 10 significance level of downregulated signature miRNA

Figure 3.8 and **figure 3.9** show the differentially expressed miRNAs represented in a Venn diagram. Signature miRNAs were split into downregulated and upregulated, so one can distinguish cases where a miRNA was upregulated in one tissue type and downregulated in another. Due to the method by which samples were obtained, both pCRC samples and CLM samples contain contaminating cells from their respective surrounding normal tissues. As such, differential expression of pCRC and CLM will contain a lot of *noise* actually caused by the difference in miRNA expression between normal colorectum tissue and normal liver tissue. In the volcano plot of pCRC versus CLM (**Figure 3.11**), signature miRNA found both in nCR versus CLM where colored blue, while those only found in pCRC versus CLM where

labeled red. Additionally, due to the method used to map sequencing reads to MirGeneDB, miRNA with identical mature sequence will all receive the same number of counts. Thus, in some cases, multiple miRNA will have identical counts and identical differential expression. Some of the signature miRNA must therefore be regarded as the same miRNA.

3.4.3 Volcano plot nCR versus nLi

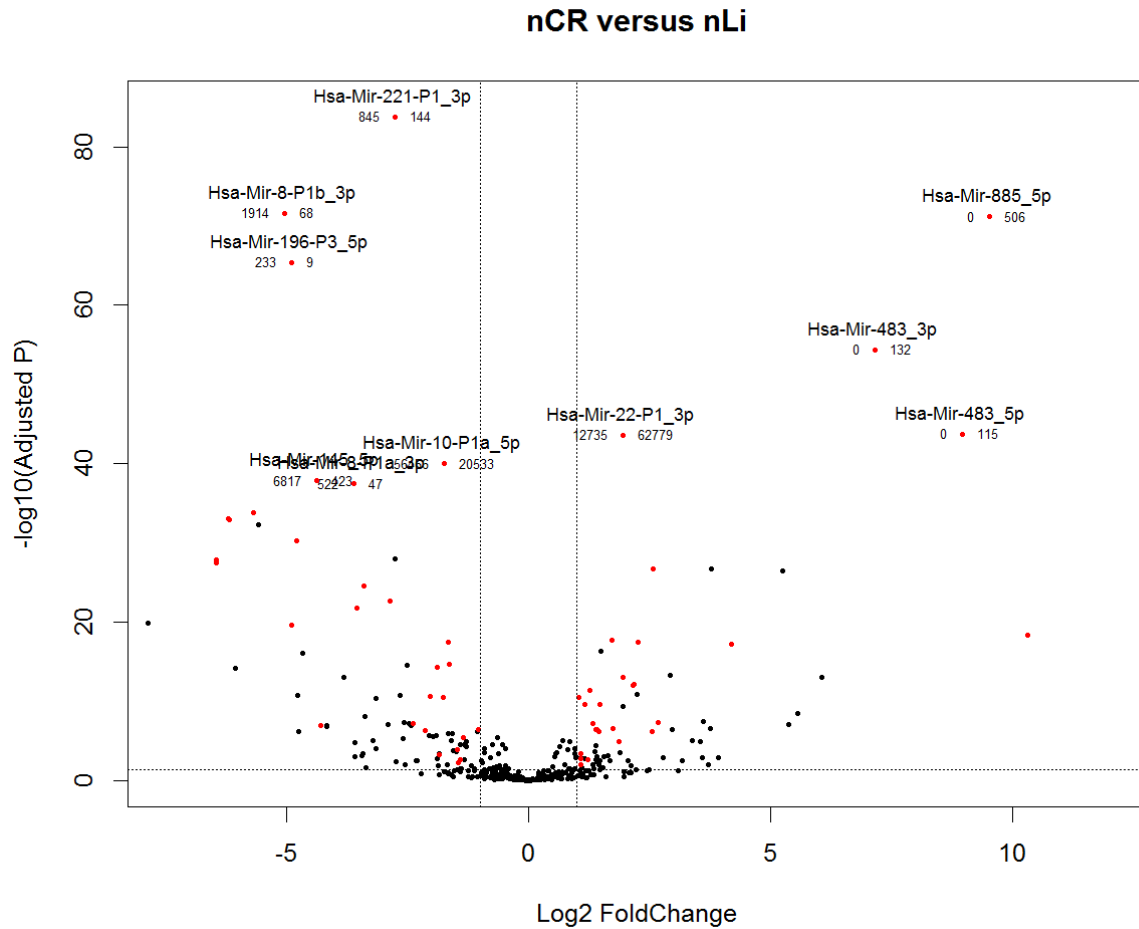


Figure 3.10 Volcano Plot nCR vs nLi Plot of $-\log_{10}$ padj against LFC of nCR vs nLi. Signature miRNA (red) have more than one LFC, padj < 0.05 and one group > 100 RPM. Only top 10 significance level miRNA are highlighted.

↓ miRNA in nCR vs nLi	↑ miRNA in nCR vs nLi
Hsa-Mir-221-P1_3p	Hsa-Mir-885_5p
Hsa-Mir-8-P1b_3p	Hsa-Mir-483_3p
Hsa-Mir-196-P3_5p	Hsa-Mir-483_5p
Hsa-Mir-10-P1a_5p	Hsa-Mir-22-P1_3p
Hsa-Mir-145_5p	Hsa-Mir-148-P1_3p
Hsa-Mir-8-P1a_3p	Hsa-Mir-122_5p
Hsa-Mir-8-P2b_3p	Hsa-Mir-193-P1b_3p
Hsa-Mir-196-P2_5p	Hsa-Mir-10-P2b_5p
Hsa-Mir-196-P1_5p	Hsa-Mir-455_3p
Hsa-Mir-10-P1b_5p	Hsa-Mir-455_5p

Table 3.2 Signature miRNA in nCR vs nLi. Only top 10 significance level signature miRNA are shown.

3.4.4 Volcano plot pCRC versus CLM

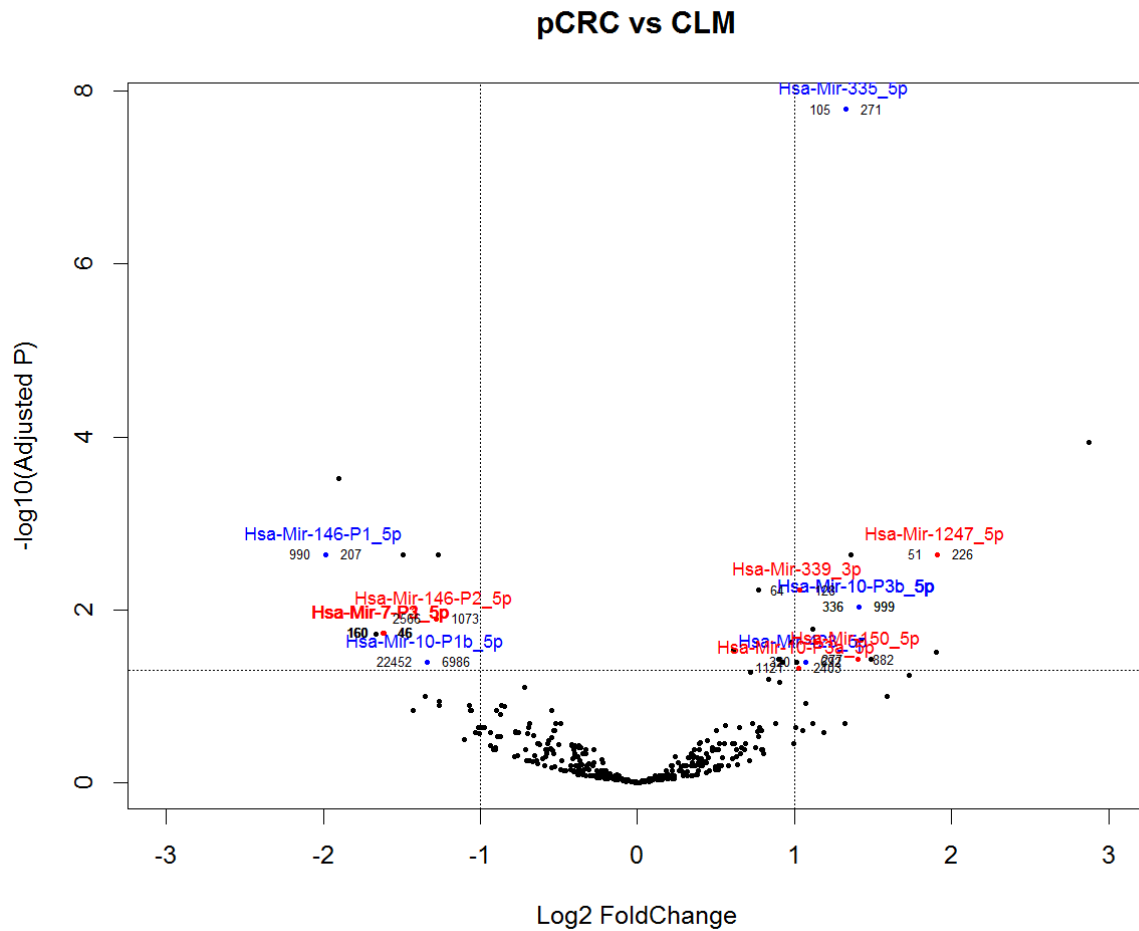


Figure 3.11 Volcano Plot pCRC vs CLM Plot of negative log 10 adjusted p-value against Log2 fold change. Signature miRNA are highlighted, representing miRNA with Log2 fold change above 1 or below -1, adjusted p-value > 0.05 (dotted lines), and where at least one group is above 100 reads per million. Numbers below miRNA ID represents reads per million for each group. (a) Normal Colorectum versus Colorectal Cancer, higher foldchange represents higher expression in CRC. (b) Normal Colorectum versus Normal Liver, higher fold change represents expression in Normal Liver. (c) Primary Colorectal Cancer versus Colorectal derived Liver Metastasis, higher fold change represents higher expression in metastasis. (d) Normal Liver versus Metastasis, higher fold change represents higher expression in Normal Liver versus Metastasis.

↓ miRNA in pCRC vs CLM	↑ miRNA in pCRC vs CML
Hsa-Mir-146-P1_5p	Hsa-Mir-335_5p
Hsa-Mir-146-P2_5p	Hsa-Mir-1247_5p
Hsa-Mir-7-P1/2/3_5p	Hsa-Mir-339_3p
	Hsa-Mir-10-P3b_5p
	Hsa-Mir-10-P3c_5p
	Hsa-Mir-150_5p
	Hsa-Mir-423_5p
	Hsa-Mir-10-P3a_5p

Table 3.3 Differentially Expressed miRNAs in pCRC vs CLM

Figure 3.10 and **3.11** show the result of differential expression analysis. Represented as a volcano plot, the $-\text{Log}_{10}(\text{padj})$ is plotted against LFC. The higher up the y-axis, the lower the adjusted p-value and the higher the fold change, the further to the left or right along the y-axis. Since genes with low fold change also tend to have higher p-values (less significant), these will reside in the bottom center of the plot. Genes with higher fold change will typically have lower p-values, and therefore reside in the upper right or left. Net result is a plot shaped like a volcano, with lowly differentiated genes in the bottom center, highly differentiated genes in upper left or right. Signature miRNA are therefore defined as having a LFC > 1 or < -1 , and an $\text{padj} < 0.05$. Additionally, a requirement was set that at least one of the groups must have more than 100 reads per million. This cutoff was set to prevent lowly expressed genes drowning out the signal of differentially expressed miRNAs. If a miRNA gene has a mean count of 1 in one group and a mean count of 20 in another group, a substantial net 20-fold change. In biological turns, however, the numbers are so small as to be likely insignificant. By implementing a 100 RPM cutoff, only genes with significant expression is considered. A total of 11 microRNA genes was found to be differentially expressed between pCRC and CLM, including Hsa-Mir-335_5p, Hsa-Mir-1247, Hsa-Mir-146-P1_5p, Hsa-Mir-339_3p, Hsa-Mir-10-P3b/c_3p, Hsa-Mir-146-P2_5p, Hsa-Mir-7-P1/2/3_5p, Hsa-Mir-150_5p, Hsa-Mir-10-P1b_5p, Hsa-Mir-423_5p and Hsa-Mir-10-P3a_5p.

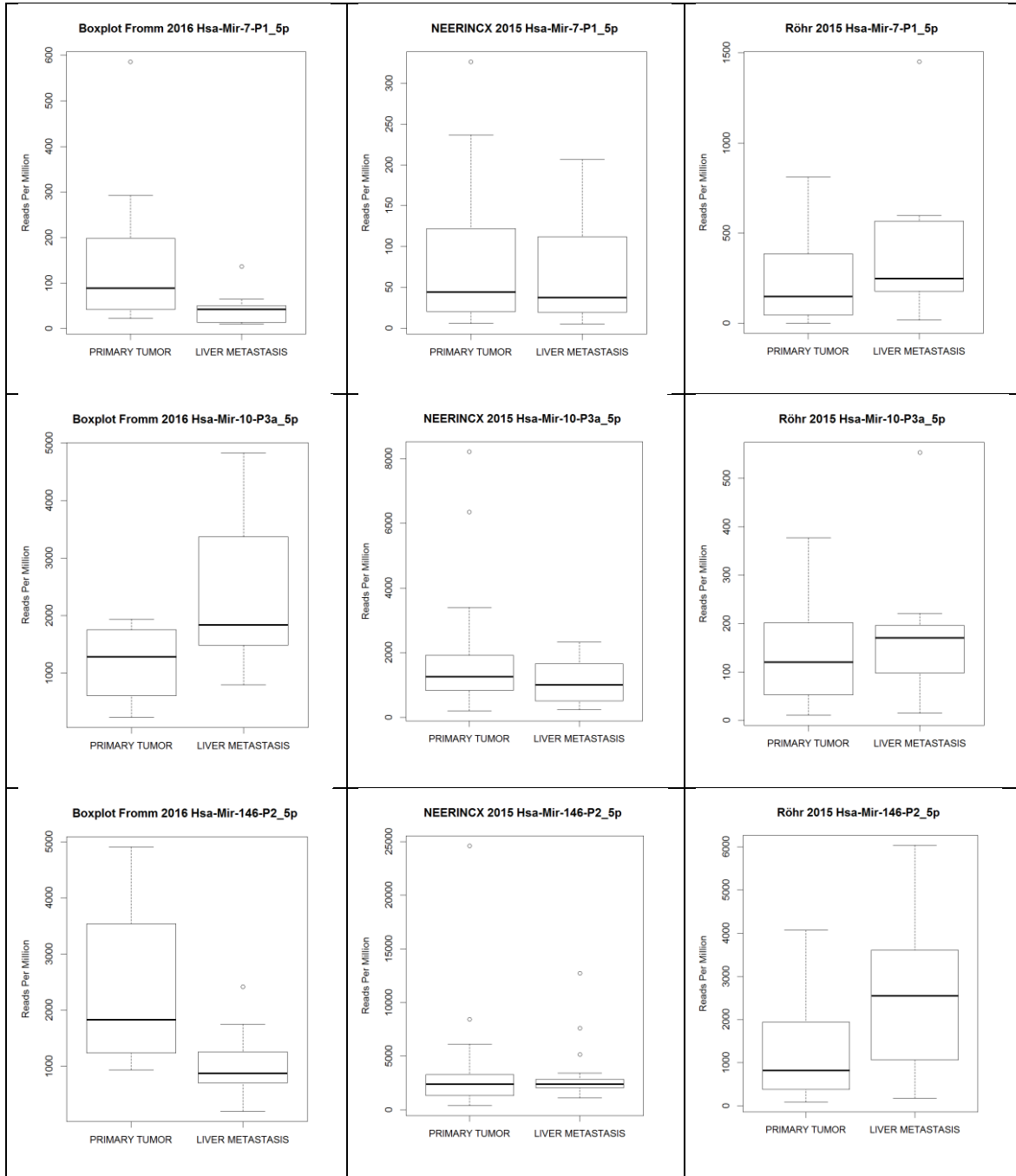
To control for how many of these were caused by different miRNA expression profiles in normal colorectum and normal liver, signature miRNA of the two differential expression analyses were compared. In **Figure 3.11**, signature miRNA found to be differentially expressed in both normal tissues and malignant tissues were colored blue, while signature miRNA exclusively found in malignant tissues were colored red. This left four upregulated miRNAs, Hsa-Mir-1247_5p, Hsa-Mir-339_3p, Hsa-Mir-150_5p, Has-Mir-10-P3a_5p, and two downregulated miRNAs, Has-Mir-146-P2_5p, Hsa-Mir-7-P1/2/3_5p.

↓ miRNA in pCRC vs CLM	↑ miRNA in pCRC vs CML
Hsa-Mir-146-P1_5p	Hsa-Mir-335_5p
Hsa-Mir-146-P2_5p	Hsa-Mir-1247_5p
Hsa-Mir-7-P1/2/3_5p	Hsa-Mir-339_3p
	Hsa-Mir-150_5p
	Hsa-Mir-10-P3a_5p

Table 3.4 miRNA Differentially Expressed for pCRC vs CLM and Controlled for Normal Tissue

3.4.5 Validation in Neerincx and Röhr

Expression levels for the six signature miRNAs in pCRC vs CLM were compared against NGS data from Neerincx et al, 2015 [99] and Röhr et al, 2013 [100]. Previous studies [93-95] have shown that NGS data itself may suffer from replication issues. It is therefore necessary to validate findings, either experimentally, for instance using qRT-PCR, or by comparing data from similar studies. Boxplots of RPM in pCRC and CLM for each of the six signature miRNAs were made for current study, Neerincx and Röhr.



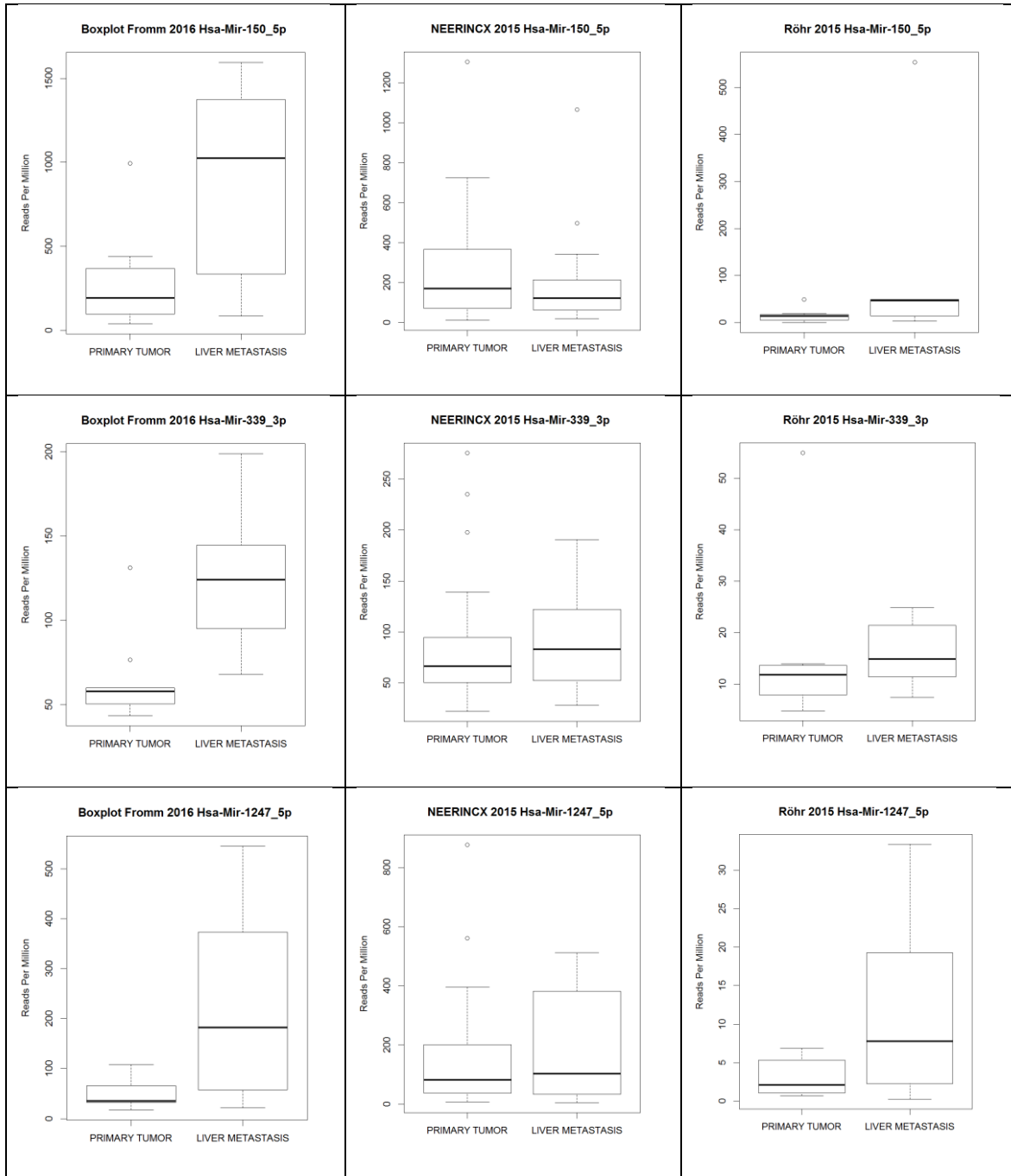


Figure 3.12 Boxplot of signature miRNA Comparison of RPM values for 6 signature miRNA in pCRC and CLM in samples from this study, Neerincx 2015 and Röhr 2013.

Comparing expression levels in Neerincx and Röhr, only 2 miRNA, Hsa-Mir-339_3p and Hsa-Mir-1247_5p, appear to show the same pattern of upregulation in CLM compared to the pCRC. Thus, 2 signature miRNA out of the original 11 were both differentially expressed in pCRC vs CLM and also showing same expression pattern in Neerincx et al 2015 [99] and Röhr et al, 2013 [100].

↓ miRNA in pCRC vs CLM	↑ miRNA in pCRC vs CML
	Hsa-Mir-1247_5p
	Hsa-Mir-339_3p

Table 3.5 Signature miRNA Controlled for Normal Tissues and Showing Same Expression Pattern in Neerinx et al 2015 and Röhr et al 2013.

Hsa-Mir-1247_5p	RPM	Hsa-Mir-339_3p	RPM
CLM:	226.52	CLM:	128.66
pCRC:	51.67	pCRC:	64.83
Brain:	3.35	Heart:	46.99
Lung:	106.65	Spleen:	0.38
Kidney:	29.65	Brain:	84
Liver:	3.28	Lung:	7.1
		Blood:	12.23
		Kidney:	2.31
		Liver:	14.76

Table 3.6 RPM in tissues. Tissues other than CLM and pCRC derived from previously published literature listed at MirGeneDB.org

Hsa-Mir-1247 has 226.52 RPM while Hsa-Mir-339_3p has 128.65 RPM in CLM. Hsa-Mir-339_3p has 12.23 RPM in in serum, while Hsa-Mir-1247 has not yet been detected in serum.

3.5 Target prediction

Tables 3.7 and 3.9 show the top 10 TargetScan predicted target sites for Hsa-Mir-1247_5p and Hsa-Mir-339_3p.

Ortholog of target gene	Gene name	Total sites	8mer sites	7mer-m8 sites	7mer-A1 sites	6mer sites	Representative miRNA
HIST2H2AA3	histone cluster 2, H2aa3	3	2	1	0	0	hsa-miR-1247-5p
CDC14B	cell division cycle 14B	5	4	0	1	0	hsa-miR-1247-5p
AL162389.1	Uncharacterized protein	5	0	5	0	0	hsa-miR-1247-5p
FAM20C	family with sequence similarity 20, member C	2	2	0	0	0	hsa-miR-1247-5p
DVL1	dishevelled segment polarity protein 1	2	2	0	0	0	hsa-miR-1247-5p
KIF26A	kinesin family member 26A	2	2	0	0	0	hsa-miR-1247-5p
HPR	haptoglobin-related protein	1	1	0	0	0	hsa-miR-1247-5p
MBD3	methyl-CpG binding domain protein 3	2	1	1	0	2	hsa-miR-1247-5p
TNFRSF18	tumor necrosis factor receptor superfamily, member 18	1	1	0	0	0	hsa-miR-1247-5p
THEM6	thioesterase superfamily member 6	2	1	1	0	0	hsa-miR-1247-5p

Table 3.7 Top 10 Hsa-Mir-1247_5p target sites predicted by TargetScan

Ortholog of target gene	Gene name	Total sites	8mer sites	7mer-m8 sites	7mer-A1 sites	6mer sites	Representative miRNA
CCDC77	coiled-coil domain containing 77	2	2	0	0	0	hsa-miR-339-3p
TUBB	tubulin, beta class I	1	1	0	0	0	hsa-miR-339-3p
AP001631.10		1	1	0	0	0	hsa-miR-339-3p
FAM19A2	family with sequence similarity 19 (chemokine (C-C motif)-like), member A2	1	1	0	0	0	hsa-miR-339-3p
C15orf37	chromosome 15 open reading frame 37	1	1	0	0	0	hsa-miR-339-3p
GPRC5C	G protein-coupled receptor, family C, group 5, member C	1	1	0	0	0	hsa-miR-339-3p
FAM222A	family with sequence similarity 222, member A	2	1	1	0	0	hsa-miR-339-3p
ASCL5	achaete-scute complex homolog 5 (Drosophila)	2	1	1	0	0	hsa-miR-339-3p
NDUFS7	NADH dehydrogenase (ubiquinone) Fe-S protein 7, 20kDa (NADH-coenzyme Q reductase)	1	1	0	0	6	hsa-miR-339-3p
ATP6V0A4	ATPase, H ⁺ transporting, lysosomal V0 subunit a4	1	1	0	0	0	hsa-miR-339-3p

Table 3.8 Top 10 Hsa-Mir-339_3p target sites predicted by TargetScan

Target sites were detected using TargetScan web interface [103] were top 10 cumulative weighted context++ score genes for both miRNA is listed in **table 3.7** and **table 3.8**. Such target predictions may be used as a starting point for further downstream analysis of the biological role these two miRNAs play in metastasis.

3.6 IsomiRs

Table 3.9 and 3.10 list signature isomiRs in pCRC vs CLM.

3.6.1 Table isomiRs downregulated in CLM

MirGeneDB BLAST	↓ IsomiRs in pCRC vs CLM	IsomiR-TYPE
Hsa-Let-7-P5 or P10_5p	TGAGG C AGTAGATTGTATAGTT	seed-mismatch
Hsa-Let-7-P7_5p	TGAGG C AGTAGGTTGTATAGTT	seed-mismatch
Hsa-Mir-192-P2_5p	ATGACCTATGAATTGACAGACA A	elongated
Hsa-Mir-103-P3_3p	GC CAGCATTGTACAGGGCTA TA	5p and 3p truncated
Hsa-Mir-143_3p	TGAGATGAAGCACTGTAGC C	mismatch
Hsa-Mir-30-P1b_5p	TGTAAACATCCTTGACTGGAAGC G	Non-canonical elongation
Hsa-Mir-192-P2_5p	ATGACCTATGAATTGACAGAC T	mismatch
Hsa-Mir-92-P1b_3p	TATTGCACTTGTCCCGGCTG CA	mismatch
Hsa-Mir-30-P1c_5p	TGTAAACATCCCCGACTGGAAGC G	mismatch
Hsa-Mir-192-P1_5p	CT C ACCTATGAATTGACAGCC	seed-mismatch
Hsa-Mir-30-P1c_5p	TGTAAACATCCCCGACTGGAAGC A	Non-canonical elongation
Hsa-Mir-146-P1_5p	TGAGAAGTGAATTCCATGGGTTG T	Non-canonical elongation
Hsa-Mir-143_3p	GG GAGATGAAGCACTGTAGCT TA	5p and 3p truncated
Hsa-Mir-192-P1_5p	C TGACCTATGAATTGACAGCC	mismatch
Hsa-Mir-17-P3a_5p	TAAAGTGCTTATAGTGCAGGTAG A	Non-canonical elongation

Table 3.9 Top 15 IsomiRs Downregulated in CLM IsomiR sequences were BLASTed against MirGeneDB, selecting miRNA with highest score.

3.6.2 Table isomiRs upregulated in CLM

MirGeneDB BLAST	↑ IsomiRs in pCRC vs CLM	IsomiR-TYPE
Hsa-Mir-1247_5p	ACCCGTCCCGTTCGTCCCCGGA ^T	Non-canonical elongation
Hsa-Mir-26-P1 or P2_5p	TTCAAGTAATCCAGGATAGGC ^{AT}	mismatch, Non-canonical elongation
Hsa-Mir-26-P1 or P2_5p	TTCAAGTAATCCAGGATAGGC ^{AA}	mismatch, Non-canonical elongation
Hsa-Mir-181-P1a or P1b_5p	AACATTCAACGCTGTCGGT	truncated
Hsa-Mir-26-P1 or P2_5p	TTCAAGTAATCCAGGATAGGC	truncated
Hsa-Mir-181-P1a or P1b_5p	AACATTCAACGCTGTCGGTG	truncated
Hsa-Mir-199-P2_5p*	CCCAGTGTTCCAGACTACCTGTTC ^T	Non-canonical elongation
Hsa-Let-7-P2_5p	TGAGGTAGGAGGTTGTATAGTT ^A	Non-canonical elongation
Hsa-Mir-8-P1b_3p	TAACTGTCTGGTAAAGAT	truncated
Hsa-Mir-150_5p	TCTCCCAACCCTTGTACCAGTG ^T	Non-canonical elongation
Hsa-Mir-127_3p	TCGGATCCGTCTGAGCTTGGCT ^{TT}	Non-canonical elongation
Hsa-Mir-361_3p	TCCCCAGGTGTGATTCTGATT	truncated
Hsa-Mir-8-P1b_3p	TAACTGTCTGGTAAAGA ^{AA}	mismatch, truncated
Hsa-Mir-10-P3a_5p	TCCCTGAGACCCTTTAACCTGTG ^G	mismatch
Hsa-Let-7-P2_5p	TGAGGTAGGAGGTTGTATAGTT ^T	Non-canonical elongation

Table 3.10 Top 15 IsomiRs Upregulated in CLM IsomiR sequences were BLASTed against MirGeneDB, selecting miRNA with highest score.

IsomiRs were defined as fragments mapped to MirGeneDB annotated pri-miRNA, and not identical to canonical miRNA. Only fragments present in at least 50 % of samples were counted. Signature isomiRs were defined as isomiRs LFC > 1 or < -1, p-adjusted value > 0.05. At least one of the groups must have > 100 RPM of said isomiR, and neither group may have a mean expression of 0 counts.

Interestingly, one of the isomiRs upregulated in CLM is a non-canonical 3p elongated version of Hsa-Mir-1247_5p, one of the signature miRNA in pCRC vs CLM, with a uracil addition at the 3p. IsomiRs of Hsa-Let-7-P5/10_5p, Hsa-Let-7-P7_5p and Hsa-Mir-192-P1_5p contain a mismatch in the seed region, while isomiRs of Hsa-Mir-103-P3_3p, Hsa-Mir-143_3p show 5p-truncation in the seed. All isomiRs with alterations in the seed region are downregulated in CLM. No isomiR among top 15 upregulated and downregulated show 5p elongation, while a majority of isomiRs have 3p elongations or truncations. Of the 3p elongations, additional nucleotides can both be canonical, containing the same nucleotide as the pri-miRNA at that position, or they can be noncanonical, with a mismatch on the pri-miRNA for that position.

3.7 Sequential motifs

The number of human miRNA in MirGeneDB with Basal UG motif, Apical UGU/GUG motif and Flanking CNNC motif are shown in **Figure 3.13**. Out of 523 human miRNA, only 23 miRNA genes contained all three motifs, 322 contained Flanking CNNC motif, 119 contained Basal UG motif, 140 contained UGU/GUG motifs, and 122 miRNA genes contained no motif.

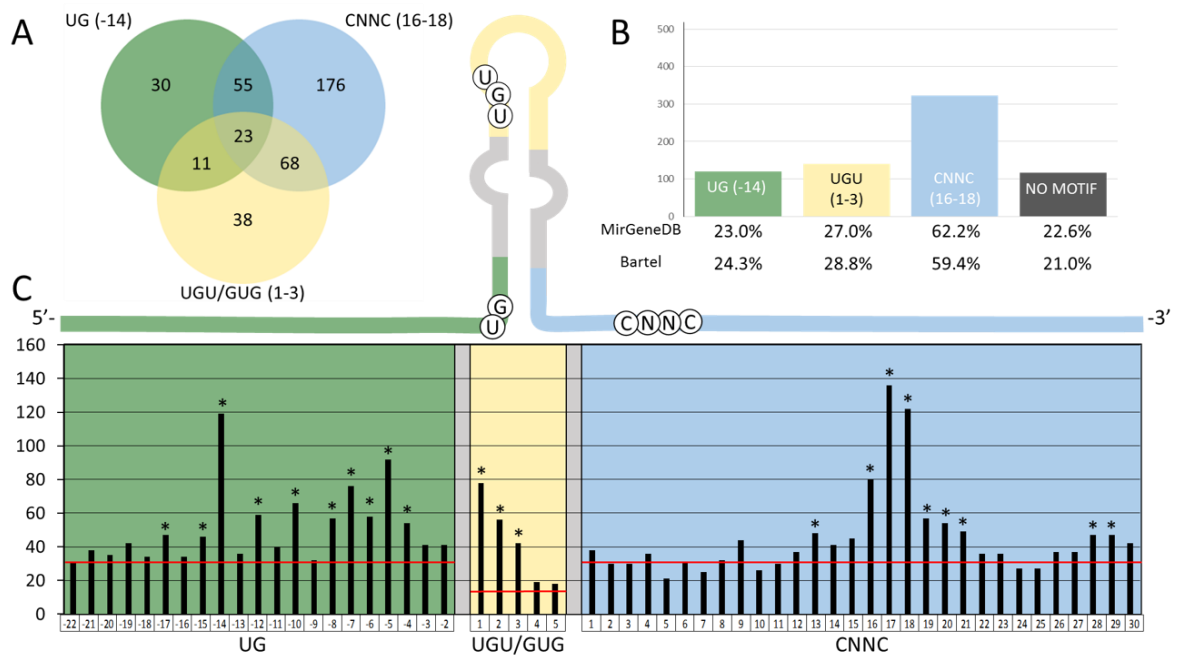


Figure 3.13 Number of miRNA in MirGeneDB with pri-miRNA motifs (a) Venn diagram of miRNAs in MirGeneDB with the three motifs. (b) Barplot of total number of miRNAs in MirGeneDB with motifs, legend show percentage found in MirGeneDB and Auyeung et al, 2013 [38]. (c) Barplot of miRNAs with UG-, UGU- and CNNC- motifs at their respective positions in the 5p-stem, loop sequence, or 3p-stem, respectively. Positions are counted from Drosha cut site for UG- and CNNC-motifs, and DICER cut site for UGU-loop motif.

Figure 3.13 (b) show percentage comparison of miRNA with motif in MirGeneDB, and the number of miRNA reported by Auyeung et al 2013 [38], with a much smaller dataset. Both datasets give comparable results.

For the mismatch GHG motif, folded pri-miRNA sequences were used to observe the secondary structure motif. Results are shown as a bar plot in **figure 3.14**. For all human MirGeneDB annotated genes, there is a clear signal for mismatch GHG motif at lower stem position 7. A bar plot of human miRNA genes used by Fang et al 2015 is also shown, and the signal at position 7 is also present. Therefore, the findings of Fang et al, 2015 is validated with the much larger MirGeneDB data set. Interestingly, a sharp drop in number of motifs is observed upstream of position 0 in MirGeneDB. Upstream of position 0 are single stranded RNA, so mismatching GHG motif is not possible. The ones still present likely noise caused by RNAfold algorithm and random chance.

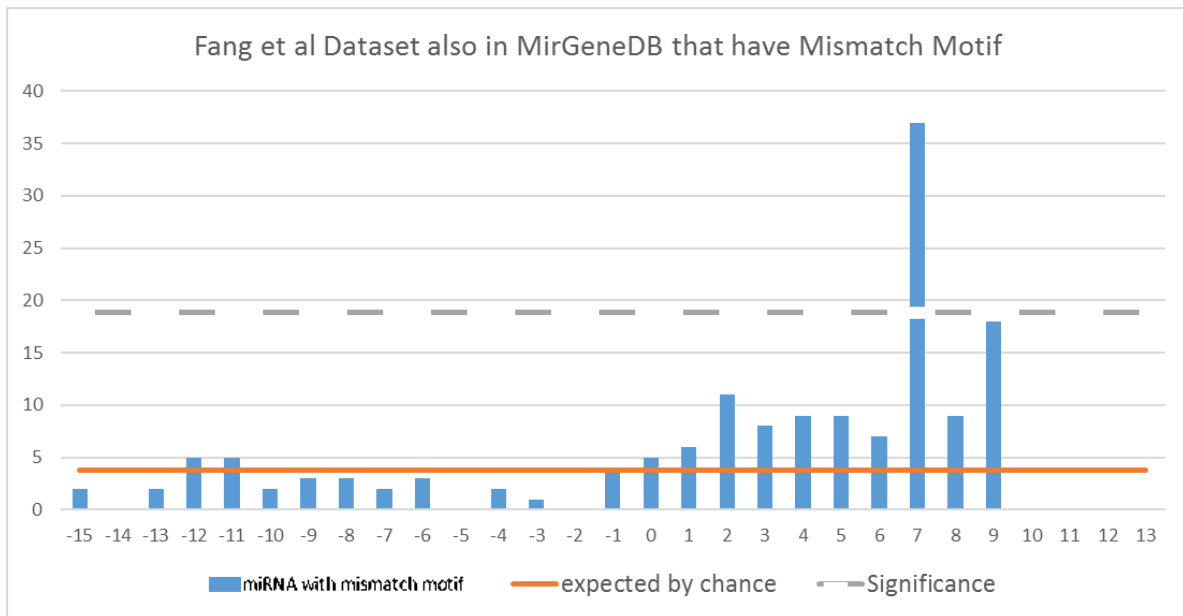
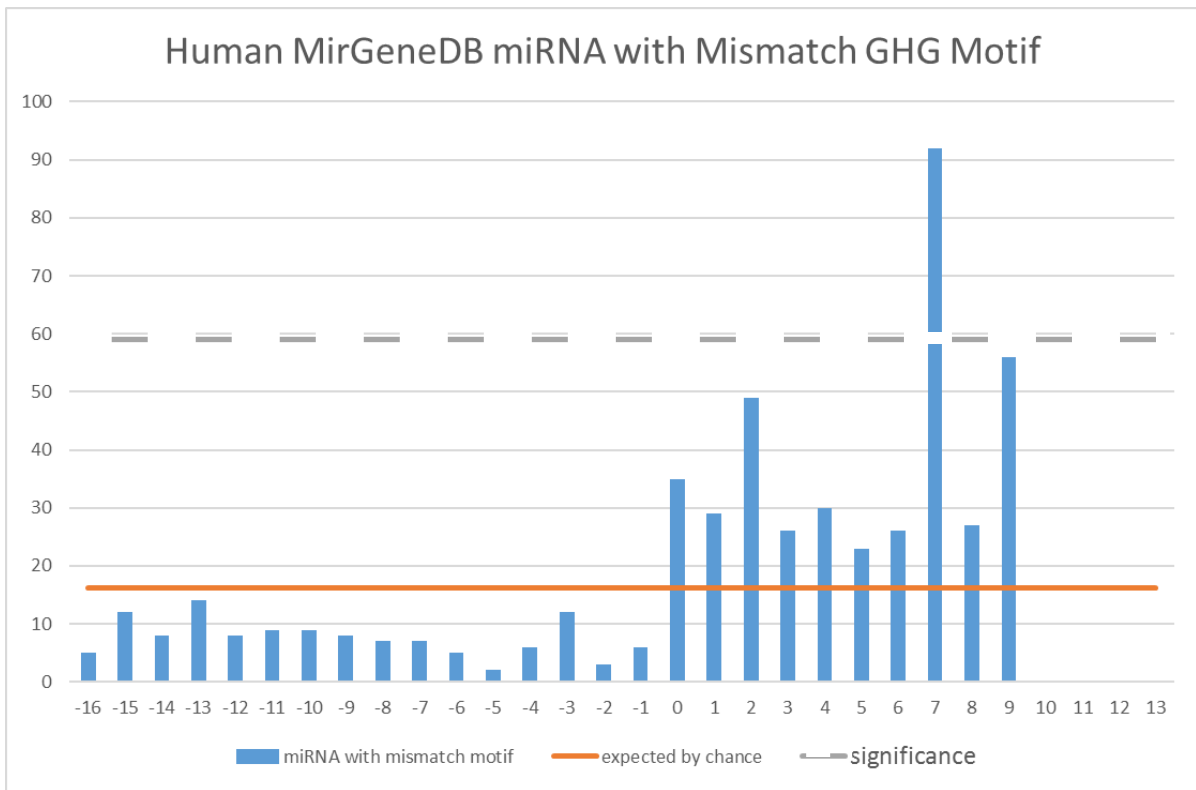


Figure 3.14 Bar Plot of Human MirGeneDB annotated miRNA genes with mismatch GHG motif. Top bar plot represents mismatch GHG motif in all MirGeneDB annotated genes, while bottom bar plot are mismatch GHG motifs of genes used in Fang et al 2013 [41]. Positions are labeled so position 1 is where single stranded RNA forms the double stranded miRNA hairpin stem. A position is counted if said position, plus the two downstream nucleotides, fulfill criteria. For example, position 7 would have mismatch GHG motif at position 7-8-9. Significance represent expectation value (number of motifs expected by random chance) multiplied by 2 standard deviations.

Signature miRNA with Structure	Motifs
<p>Hsa-Mir-1247</p> <pre> 10 20 30 40 50 UCGCCCAGCGCAGCCCCGGCCGC-- A C C UU C ACGUUGC UGGGCGC CC GUC CG CGU CCCGG \ GCCCGCG GG CAG GC GCA GGGCC U GCCCCAGGCGCUUCACCCGAGUCAAA^ 110 100 90 80 70 60 Mismatch GHG Motif </pre>	<p>CNNC at 3p (+17)</p> <p>Mismatch GHG (7-8-9)</p>
<p>Hsa-Mir-339</p> <pre> 10 20 30 40 50 CUGUGCUC CGCAGGGGC-- C CUCC C C A UGU G GGGGCGG CGCU CUGUC UC AGG GCUCACG C CCCCGUC GCGG GACAG AG UCC CGAGUGU C ACCAGGGCCGCGUCUGUGA^ 110 100 90 80 70 60 Mismatch GHG Motif </pre>	<p>UGU in Loop</p> <p>CNNC at 3p (+17)</p> <p>Mismatch GHG (7-8-9)</p>

Table 3.11 Motifs in Signature miRNA Shown is secondary structure of pri-miRNA and any respective sequence and structural motifs.

Checking for motifs in identified signature miRNA showed that Hsa-Mir-1247 contained Flanking CNNC motif at position 17 downstream Drosha 3p cut site, and a mismatch GHG motif at position 7-8-9 in lower stem. Hsa-Mir-339 contained apical UGU/GUG motif in the loop sequence, flanking CNNC motif at position 17 downstream Drosha 3p cut site, and a mismatch GHG at position 7-8-9 in lower stem.

Signature isomiRs with Structure

Motifs

Hsa-Mir-1247 Non-canonical 3p elongation

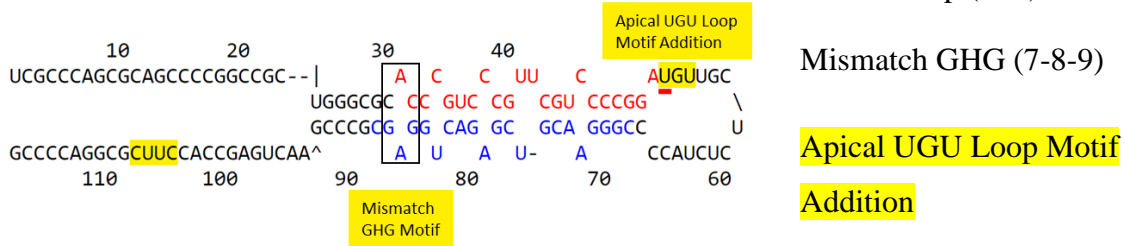


Table 3.12 Motifs in Signature isomiRs Shown is secondary structure of pri-miRNA and any respective sequence and structural motifs they have. Sequence substitution by isomiR marked with red underscore.

One of the signature isomiRs upregulated in CLM was a 3p uridylated isomiR of Hsa-Mir-1247_5p. Interestingly, the added nucleotide is uracil. If this substitution is present in the pri-miRNA, the miRNA would gain an Apical UGU Loop motif.

4 Discussion

Metastatic spread to liver, along with lung and peritoneum, is the main cause of death in colorectal cancer patients. As such, a thorough investigation of the underlying biology of this deadly disease is warranted. MiRNAs have been shown to play key role in all hallmarks of cancer, including colorectal cancer and metastasis, hence substantial effort should be put in place to elucidate their biological involvement in the disease.

Previous studies [99] failed to differentiate miRNA expression levels between pCRC and metastatic tissue, neither at the global expression level, nor individual miRNA gene level. Here, we also failed to distinguish miRNA expression at the global level, indicating CLM cells have a broadly similar miRNA expression profile as their progenitor pCRC cells. However, 6 individual miRNA genes were found to be differentially expressed in pCRC vs CLM. Two of these, Hsa-Mir-1247 and Hsa-Mir-339 also showed the same pattern of differential expression in Neerincx [99] and Röhr [100]. Explanation of the disparaging results may reside in different study designs. Neerincx and Röhr both looked at miRNA expression from multiple metastatic sites, while this study only looked at metastasis to liver. It is possible different metastatic sites have differing miRNA expression profiles. Assigning them to one single group may therefore obscure a site-specific signal. Indeed, when looking exclusively at Neerincx and Röhr Metastatic samples, the two miRNA do show the same pattern of upregulation in Metastasis compared to primary tumor.

For biomarker potential, an important question is whether miRNA is detectable in serum of healthy individuals. MirGeneDB.org shows that previous publications report Hsa-Mir-1247_5p is not at all expressed in serum samples, while Hsa-Mir-339_3p has an expression of 12.23 RPM in serum samples. As such, even if CLM tissue secretes Hsa-Mir-339_3p to serum, it may not be detectable due to high noise to signal ratio. Meanwhile, Hsa-Mir-1247_5p is, based on current literature, not present in serum of normal individuals. Therefore, noise to signal ratio should be less of an issue.

IsomiRs were also found to be differentially expressed between the pCRC and CLM. The vast majority of top 15 up- and downregulated isomiRs showed 3p elongations and truncations, while 6 isomiRs had mismatches or truncations in their seed region. Interestingly, a 3p uridylylated isomiR of Hsa-Mir-1247_5p was detected and found to be the most significantly upregulated signature isomiR in CLM. As shown in Koppers-Lalic et al

2014 [36], 3p uridylated isoforms are enriched in in exosomes, while adenylated isoforms are overrepresented in the cytosol of cells. Previously reported Hsa-Mir-1247_5p isomiRs listed in MirGeneDB.org show 3p uridylation and adenylation to be by far the most common substitution. We propose a hypothesis where CML tissue 3p uridylate Hsa-Mir-1247 for excretion in exosomes.

Alternatively, if isomiR uridylation was a consequence of genomic substitution, Hsa-Mir-1247 pri-miRNA would gain an Apical UGU/GUG Loop Motif, which according to Auyeung et al, 2013 [38] and Nguyen et al, 2015 [39] enhance processing efficiency. We hypothesize a possible mechanism of upregulation by gain of function mutation in the genomic loci of Hsa-Mir-1247, where a gain of Apical UGU/GUG loop motif enhance processing and expression levels. Regardless, a consistently CLM 3p uridylated Hsa-Mir-1247_5p may enhance detection properties in serum, as non-canonical isomiRs may be less likely to be hidden by the cells normally expressed miRNA.

Sequential motifs were successfully validated in MirGeneDB, were percentage of miRNA with motifs stayed consistent in our much larger dataset when compared to dataset used by Auyeung et al [38]. Lower stem GHG structural motifs also showed a clear signal in both datasets at the 7-8-9 site proposed by Fang et al 2015 [41]. Notably however, although a strong signal, the majority of miRNA did not have lower stem structural motif. Any one motif was found only in about a fifth of all miRNA, except the Flanking CNNC motif found in two thirds of miRNA. 122 miRNAs did not have any sequential motif at all. As such, although clearly a feature of miRNA processing, either these motifs are not essential, rather acting as enhancer of miRNA processing than a requirement. Alternatively, these motifs are only part of the picture, with more miRNA structural features yet to be discovered.

Sequencing performed in this study forms the foundation for future investigations into the role miRNA play in colorectal derived liver metastasis. Firstly, although outside the scope of this study, an experimental approach using qRT-PCR to validate expression of the two miRNAs, including isomiRs, is required. More patient samples should also be sequenced, to improve statistical power.

Another notable weakness is the method by which normal adjacent tissue was controlled for. Here, any signature miRNA in nCR vs nLi were removed among signature miRNA in pCRC vs CLM, regardless of differences in LFC or significance levels. This could, for example,

obscure miRNAs with large LFC in pCRC vs CLM, but only modest LFC in nCR vs nLi. Also of note is the question of whether normal tissue really is normal, as a cancer patients 'normal' tissue have also undergone alterations.

Future outlook should therefore be investigating new approaches to control for normal adjacent tissues, by also taking into account the actual size difference in both LFC and significance of signature miRNA showing the same expression pattern in both malignant and normal tissues. Furthermore, small RNA sequencing of colorectal derived lung and PC metastasis samples should be accomplished. If lung and PC metastasis samples display the same signature miRNA, yet these are not found in their respective normal adjacent tissues, it would indicate these are in fact differences between pCRC and the metastasis, not normal adjacent tissues. Our research group is already in the process of sequencing and analysis of colorectal derived lung and PC metastasis.

Additionally, sequencing and validation with qRT-PCR of patient serum samples with known clinicopathological parameters would be very interesting to validate the found signature miRNAs. The Janus Serum Bank at the Norwegian cancer registry, Oslo, contains serum samples of more than three hundred thousand individuals, of those several developed colorectal cancer and metastases. This serum bank therefore represents an important local opportunity to follow up on our findings.

Furthermore, an effort should be made to study variability in genomic loci of signature miRNAs and especially the isomiRs identified in this study. We saw some significant variability that could only partially be explained by post transcriptionally modifications (truncations and additions) and in fact could be based on mutations. A mechanism of miRNA dysregulation of miRNAs in cancer induced by gain and loss of sequential motifs is an interesting hypothesis for future research.

References

1. Pasquinelli, A.E., *MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship*. Nat Rev Genet, 2012. **13**(4): p. 271-82.
2. Flatmark, K., E. Hoye, and B. Fromm, *microRNAs as cancer biomarkers*. Scand J Clin Lab Invest Suppl, 2016. **245**: p. S80-3.
3. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II*. EMBO J, 2004. **23**(20): p. 4051-60.
4. Lee, Y., et al., *The nuclear RNase III Drosha initiates microRNA processing*. Nature, 2003. **425**(6956): p. 415-9.
5. Kwon, S.C., et al., *Structure of Human DROSHA*. Cell, 2016. **164**(1-2): p. 81-90.
6. Bohnsack, M.T., *Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs*. Rna, 2004. **10**(2): p. 185-191.
7. Hutvagner, G.M., J.;Pasquinelli, A.E.;Balint, E.;Tuschl, T.;Zamore,P.D., *A Cellular Function for the RNA-Interface Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA*. Science, 2001.
8. Ameres, S.L. and P.D. Zamore, *Diversifying microRNA sequence and function*. Nat Rev Mol Cell Biol, 2013. **14**(8): p. 475-88.
9. Krol, J., I. Loedige, and W. Filipowicz, *The widespread regulation of microRNA biogenesis, function and decay*. Nat Rev Genet, 2010. **11**(9): p. 597-610.
10. Ramalho-Carvalho, J., et al., *Deciphering the function of non-coding RNAs in prostate cancer*. Cancer Metastasis Rev, 2016. **35**(2): p. 235-62.
11. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions*. Cell, 2009. **136**(2): p. 215-33.
12. Baek, D., et al., *The impact of microRNAs on protein output*. Nature, 2008. **455**(7209): p. 64-71.
13. Bartel, D.P. and C.Z. Chen, *Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs*. Nat Rev Genet, 2004. **5**(5): p. 396-400.
14. Salmena, L., et al., *A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?* Cell, 2011. **146**(3): p. 353-8.
15. Cesana, M., et al., *A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA*. Cell, 2011. **147**(2): p. 358-69.
16. Tay, Y., et al., *Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs*. Cell, 2011. **147**(2): p. 344-57.
17. Sumazin, P., et al., *An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma*. Cell, 2011. **147**(2): p. 370-81.
18. Karreth, F.A., et al., *In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma*. Cell, 2011. **147**(2): p. 382-95.
19. Denzler, R., et al., *Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance*. Mol Cell, 2014. **54**(5): p. 766-76.
20. Denzler, R., et al., *Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression*. Mol Cell, 2016. **64**(3): p. 565-579.
21. Cerutti, H. and J.A. Casas-Mollano, *On the origin and functions of RNA-mediated silencing: from protists to man*. Curr Genet, 2006. **50**(2): p. 81-99.
22. Tarver, J.E., P.C. Donoghue, and K.J. Peterson, *Do miRNAs have a deep evolutionary history?* Bioessays, 2012. **34**(10): p. 857-66.

23. Erwin, D.H., *Early origin of the bilaterian developmental toolkit*. Philos Trans R Soc Lond B Biol Sci, 2009. **364**(1527): p. 2253-61.
24. Erwin, D.H., et al., *The Cambrian conundrum: early divergence and later ecological success in the early history of animals*. Science, 2011. **334**(6059): p. 1091-7.
25. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data*. Nucleic Acids Res, 2014. **42**(Database issue): p. D68-73.
26. Castellano, L. and J. Stebbing, *Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues*. Nucleic Acids Res, 2013. **41**(5): p. 3339-51.
27. Chiang, H.R., et al., *Mammalian microRNAs: experimental evaluation of novel and previously annotated genes*. Genes Dev, 2010. **24**(10): p. 992-1009.
28. Jones-Rhoades, M.W., *Conservation and divergence in plant microRNAs*. Plant Mol Biol, 2012. **80**(1): p. 3-16.
29. Fromm, B., et al., *A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome*. Annu Rev Genet, 2015. **49**: p. 213-42.
30. Ambros, V., *A uniform system for microRNA annotation*. Rna, 2003. **9**(3): p. 277-279.
31. Neilsen, C.T., G.J. Goodall, and C.P. Bracken, *IsomiRs--the overlooked repertoire in the dynamic microRNAome*. Trends Genet, 2012. **28**(11): p. 544-9.
32. Newman, M.A., V. Mani, and S.M. Hammond, *Deep sequencing of microRNA precursors reveals extensive 3' end modification*. RNA, 2011. **17**(10): p. 1795-803.
33. Lee, L.W., et al., *Complexity of the microRNA repertoire revealed by next-generation sequencing*. RNA, 2010. **16**(11): p. 2170-80.
34. Burroughs, A.M., et al., *A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness*. Genome Res, 2010. **20**(10): p. 1398-410.
35. Wyman, S.K., et al., *Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity*. Genome Res, 2011. **21**(9): p. 1450-61.
36. Koppers-Lalic, D., et al., *Nontemplated nucleotide additions distinguish the small RNA composition in cells from exosomes*. Cell Rep, 2014. **8**(6): p. 1649-58.
37. Bentwich, I., et al., *Identification of hundreds of conserved and nonconserved human microRNAs*. Nat Genet, 2005. **37**(7): p. 766-70.
38. Auyeung, V.C., et al., *Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing*. Cell, 2013. **152**(4): p. 844-58.
39. Nguyen, T.A., et al., *Functional Anatomy of the Human Microprocessor*. Cell, 2015. **161**(6): p. 1374-87.
40. Fromm, B., *microRNA Discovery and Expression Analysis in Animals*. 2016: p. 121-142.
41. Fang, W. and D.P. Bartel, *The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes*. Mol Cell, 2015. **60**(1): p. 131-45.
42. Weinberg, R.A. and D. Hanahan, *The Hallmarks Of Cancer*. 2000.
43. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation*. Cell, 2011. **144**(5): p. 646-74.
44. Cheng, N., et al., *Transforming growth factor-beta signaling-deficient fibroblasts enhance hepatocyte growth factor signaling in mammary carcinoma cells to promote scattering and invasion*. Mol Cancer Res, 2008. **6**(10): p. 1521-33.
45. Bhowmick, N.A., E.G. Neilson, and H.L. Moses, *Stromal fibroblasts in cancer initiation and progression*. Nature, 2004. **432**(7015): p. 332-7.
46. Deshpande, A., P. Sicinski, and P.W. Hinds, *Cyclins and cdks in development and cancer: a perspective*. Oncogene, 2005. **24**(17): p. 2909-15.
47. Burkhardt, D.L. and J. Sage, *Cellular mechanisms of tumour suppression by the retinoblastoma gene*. Nat Rev Cancer, 2008. **8**(9): p. 671-82.

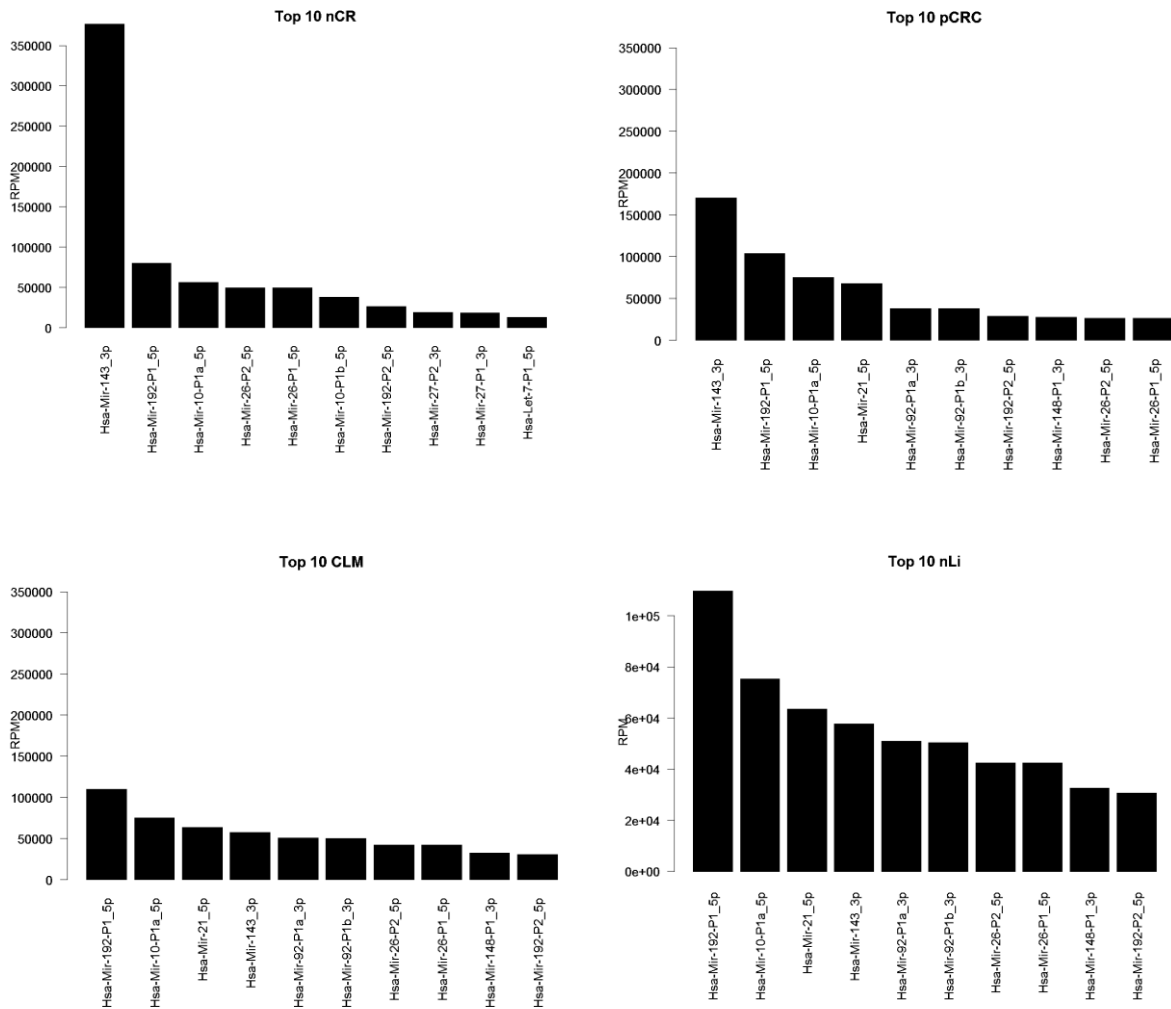
48. Calin, G.A., et al., *Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia*. Proc Natl Acad Sci U S A, 2002. **99**(24): p. 15524-9.
49. Berindan-Neagoe, I., et al., *MicroRNAome genome: a treasure for cancer diagnosis and therapy*. CA Cancer J Clin, 2014. **64**(5): p. 311-36.
50. Fromm, B., et al., *Substantial loss of conserved and gain of novel MicroRNA families in flatworms*. Mol Biol Evol, 2013. **30**(12): p. 2619-28.
51. Philippe, H., et al., *Acoelomorph flatworms are deuterostomes related to Xenoturbella*. Nature, 2011. **470**(7333): p. 255-8.
52. Lu, J., et al., *MicroRNA expression profiles classify human cancers*. Nature, 2005. **435**(7043): p. 834-8.
53. Chambers, A.F., A.C. Groom, and I.C. MacDonald, *Dissemination and growth of cancer cells in metastatic sites*. Nat Rev Cancer, 2002. **2**(8): p. 563-72.
54. Fidler, I.J., *The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited*. Nature, 2002.
55. Massague, J. and A.C. Obenauf, *Metastatic colonization by circulating tumour cells*. Nature, 2016. **529**(7586): p. 298-306.
56. Minn, A.J., et al., *Genes that mediate breast cancer metastasis to lung*. Nature, 2005. **436**(7050): p. 518-24.
57. Talmadge, J.E. and I.J. Fidler, *AACR centennial series: the biology of cancer metastasis: historical perspective*. Cancer Res, 2010. **70**(14): p. 5649-69.
58. Gupta, G.P. and J. Massague, *Cancer metastasis: building a framework*. Cell, 2006. **127**(4): p. 679-95.
59. Yang, J. and R.A. Weinberg, *Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis*. Dev Cell, 2008. **14**(6): p. 818-29.
60. Kalluri, R. and R.A. Weinberg, *The basics of epithelial-mesenchymal transition*. J Clin Invest, 2009. **119**(6): p. 1420-8.
61. Hur, K., et al., *MicroRNA-200c modulates epithelial-to-mesenchymal transition (EMT) in human colorectal cancer metastasis*. Gut, 2013. **62**(9): p. 1315-26.
62. Burk, U., et al., *A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells*. EMBO Rep, 2008. **9**(6): p. 582-9.
63. Park, S.M., et al., *The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2*. Genes Dev, 2008. **22**(7): p. 894-907.
64. Gregory, P.A., et al., *The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1*. Nat Cell Biol, 2008. **10**(5): p. 593-601.
65. Gibbons, D.L., et al., *Contextual extracellular cues promote tumor cell EMT and metastasis by regulating miR-200 family expression*. Genes Dev, 2009. **23**(18): p. 2140-51.
66. Zhang, J.P., et al., *MicroRNA-148a suppresses the epithelial-mesenchymal transition and metastasis of hepatoma cells by targeting Met/Snail signaling*. Oncogene, 2014. **33**(31): p. 4069-76.
67. Zhang, J.X., et al., *MiR-29c mediates epithelial-to-mesenchymal transition in human colorectal carcinoma metastasis via PTP4A and GNA13 regulation of beta-catenin signaling*. Ann Oncol, 2014. **25**(11): p. 2196-204.
68. Siemens, H., et al., *miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions*. Cell Cycle, 2011. **10**(24): p. 4256-71.
69. Hu, F., et al., *MiR-363-3p inhibits the epithelial-to-mesenchymal transition and suppresses metastasis in colorectal cancer by targeting Sox4*. Biochem Biophys Res Commun, 2016. **474**(1): p. 35-42.
70. Png, K.J., et al., *MicroRNA-335 inhibits tumor reinitiation and is silenced through genetic and epigenetic mechanisms in human breast cancer*. Genes Dev, 2011. **25**(3): p. 226-31.

71. Heyn, H., et al., *MicroRNA miR-335 is crucial for the BRCA1 regulatory cascade in breast cancer development*. *Int J Cancer*, 2011. **129**(12): p. 2797-806.
72. Zhu, S., et al., *MicroRNA-21 targets tumor suppressor genes in invasion and metastasis*. *Cell Res*, 2008. **18**(3): p. 350-9.
73. Zhu, S., et al., *MicroRNA-21 targets the tumor suppressor gene tropomyosin 1 (TPM1)*. *J Biol Chem*, 2007. **282**(19): p. 14328-36.
74. Gazieli-Sovran, A., et al., *miR-30b/30d regulation of GalNAc transferases enhances invasion and immunosuppression during metastasis*. *Cancer Cell*, 2011. **20**(1): p. 104-18.
75. Loo, J.M., et al., *Extracellular metabolic energetics can promote cancer progression*. *Cell*, 2015. **160**(3): p. 393-406.
76. Ferlay, J., et al., *Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012*. *Eur J Cancer*, 2013. **49**(6): p. 1374-403.
77. Nguyen, D.X., P.D. Bos, and J. Massague, *Metastasis: from dissemination to organ-specific colonization*. *Nat Rev Cancer*, 2009. **9**(4): p. 274-84.
78. Riihimaki, M., et al., *Patterns of metastasis in colon and rectal cancer*. *Sci Rep*, 2016. **6**: p. 29765.
79. O'Connell, J.B., M.A. Maggard, and C.Y. Ko, *Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging*. *J Natl Cancer Inst*, 2004. **96**(19): p. 1420-5.
80. Johnson, C.M., et al., *Meta-analyses of colorectal cancer risk factors*. *Cancer Causes Control*, 2013. **24**(6): p. 1207-22.
81. Pino, M.S. and D.C. Chung, *The chromosomal instability pathway in colon cancer*. *Gastroenterology*, 2010. **138**(6): p. 2059-72.
82. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer*. *Nat Med*, 2015. **21**(11): p. 1350-6.
83. Meng, W., et al., *Comparison of microRNA deep sequencing of matched formalin-fixed paraffin-embedded and fresh frozen cancer tissues*. *PLoS One*, 2013. **8**(5): p. e64393.
84. Landgraf, P., et al., *A mammalian microRNA expression atlas based on small RNA library sequencing*. *Cell*, 2007. **129**(7): p. 1401-14.
85. Shenoy, A. and R.H. Blelloch, *Regulation of microRNA function in somatic stem cell proliferation and differentiation*. *Nat Rev Mol Cell Biol*, 2014. **15**(9): p. 565-76.
86. Pichler, M. and G.A. Calin, *MicroRNAs in cancer: from developmental genes in worms to their clinical application in patients*. *Br J Cancer*, 2015. **113**(4): p. 569-73.
87. Angelini, T.G. and C. Emanuelli, *MicroRNAs as clinical biomarkers?* *Front Genet*, 2015. **6**: p. 240.
88. Ravery, V., *The significance of recurrent PSA after radical prostatectomy: benign versus malignant sources*. *Semin Urol Oncol*, 1999. **17**(3): p. 127-9.
89. Yang, Y.F., et al., *Discordances in ER, PR and HER2 receptors between primary and recurrent/metastatic lesions and their impact on survival in breast cancer patients*. *Med Oncol*, 2014. **31**(10): p. 214.
90. Drooger, J.C., et al., *Diagnostic and therapeutic ionizing radiation and the risk of a first and second primary breast cancer, with special attention for BRCA1 and BRCA2 mutation carriers: a critical review of the literature*. *Cancer Treat Rev*, 2015. **41**(2): p. 187-96.
91. Schee, K., O. Fodstad, and K. Flatmark, *MicroRNAs as biomarkers in colorectal cancer*. *Am J Pathol*, 2010. **177**(4): p. 1592-9.
92. Schee, K., et al., *Deep Sequencing the MicroRNA Transcriptome in Colorectal Cancer*. *PLoS One*, 2013. **8**(6): p. e66165.
93. Baran-Gale, J., et al., *Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods*. *Front Genet*, 2015. **6**: p. 352.

94. Toedling, J., et al., *Deep-sequencing protocols influence the results obtained in small-RNA sequencing*. PLoS One, 2012. **7**(2): p. e32724.
95. Leshkowitz, D., et al., *Differences in microRNA detection levels are technology and sequence dependent*. RNA, 2013. **19**(4): p. 527-38.
96. Buschmann, D., et al., *Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow*. Nucleic Acids Res, 2016. **44**(13): p. 5995-6018.
97. Fretland, A.A., et al., *Open versus laparoscopic liver resection for colorectal liver metastases (the Oslo-CoMet Study): study protocol for a randomized controlled trial*. Trials, 2015. **16**: p. 73.
98. Kristensen, A.T., et al., *Molecular detection (k-ras) of exfoliated tumour cells in the pelvis is a prognostic factor after resection of rectal cancer?* BMC Cancer, 2008. **8**: p. 213.
99. Neerincx, M., et al., *MiR expression profiles of paired primary colorectal cancer and metastases by next-generation sequencing*. Oncogenesis, 2015. **4**: p. e170.
100. Rohr, C., et al., *High-throughput miRNA and mRNA sequencing of paired colorectal normal, tumor and metastasis tissues and bioinformatic modeling of miRNA-1 therapeutic applications*. PLoS One, 2013. **8**(7): p. e67461.
101. Lopez, J.P., et al., *Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing*. BMC Med Genomics, 2015. **8**: p. 35.
102. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. Genome Biol, 2014. **15**(12): p. 550.
103. Agarwal, V., et al., *Predicting effective microRNA target sites in mammalian mRNAs*. Elife, 2015. **4**.
104. Lorenz, R., et al., *ViennaRNA Package 2.0*. Algorithms Mol Biol, 2011. **6**: p. 26.
105. Hong, S.H., et al., *Effects of delay in the snap freezing of colorectal cancer tissues on the quality of DNA and RNA*. J Korean Soc Coloproctol, 2010. **26**(5): p. 316-23.

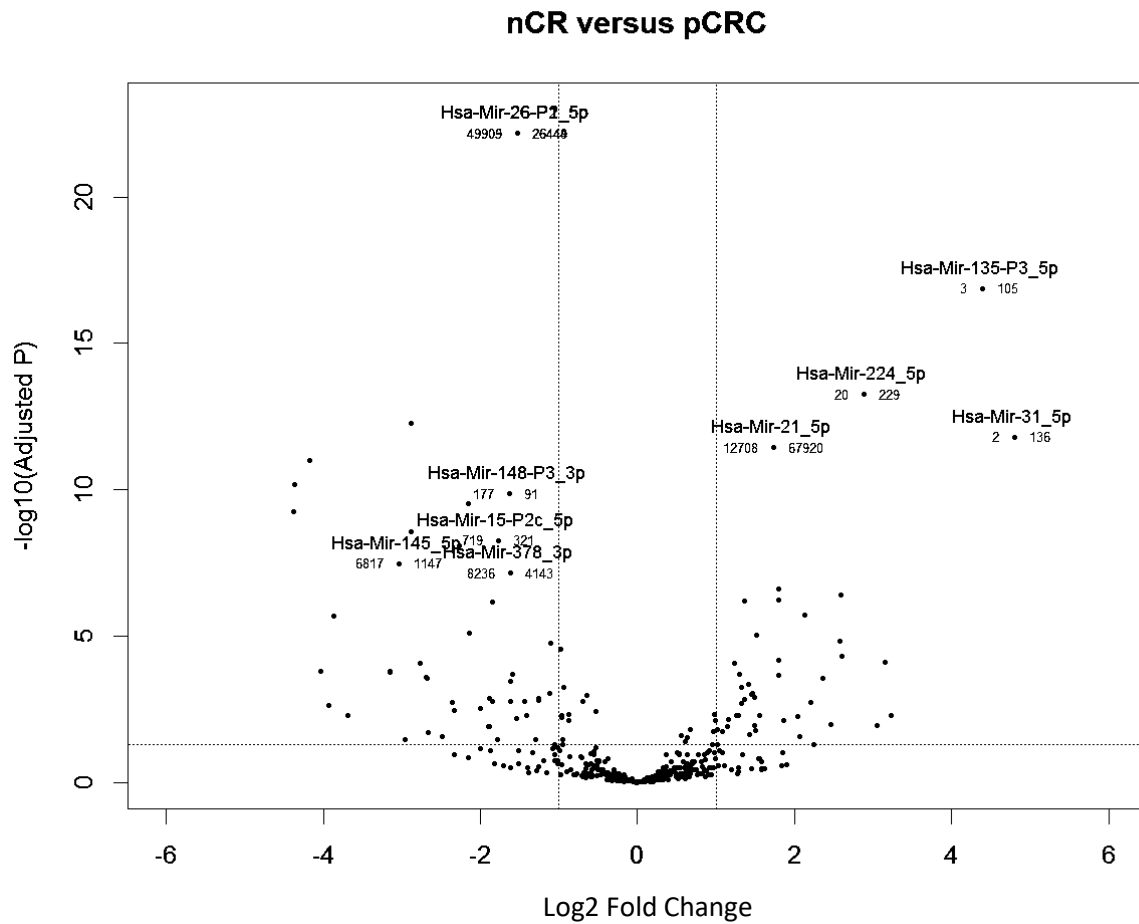
Attachments

4.1.1 Top 10 miRNA per Tissue



Appendix Figure 1 Top 10 miRNA in nCR, pCRC, CLM and nLi.

4.1.2 Volcano plot nCR vs pCRC

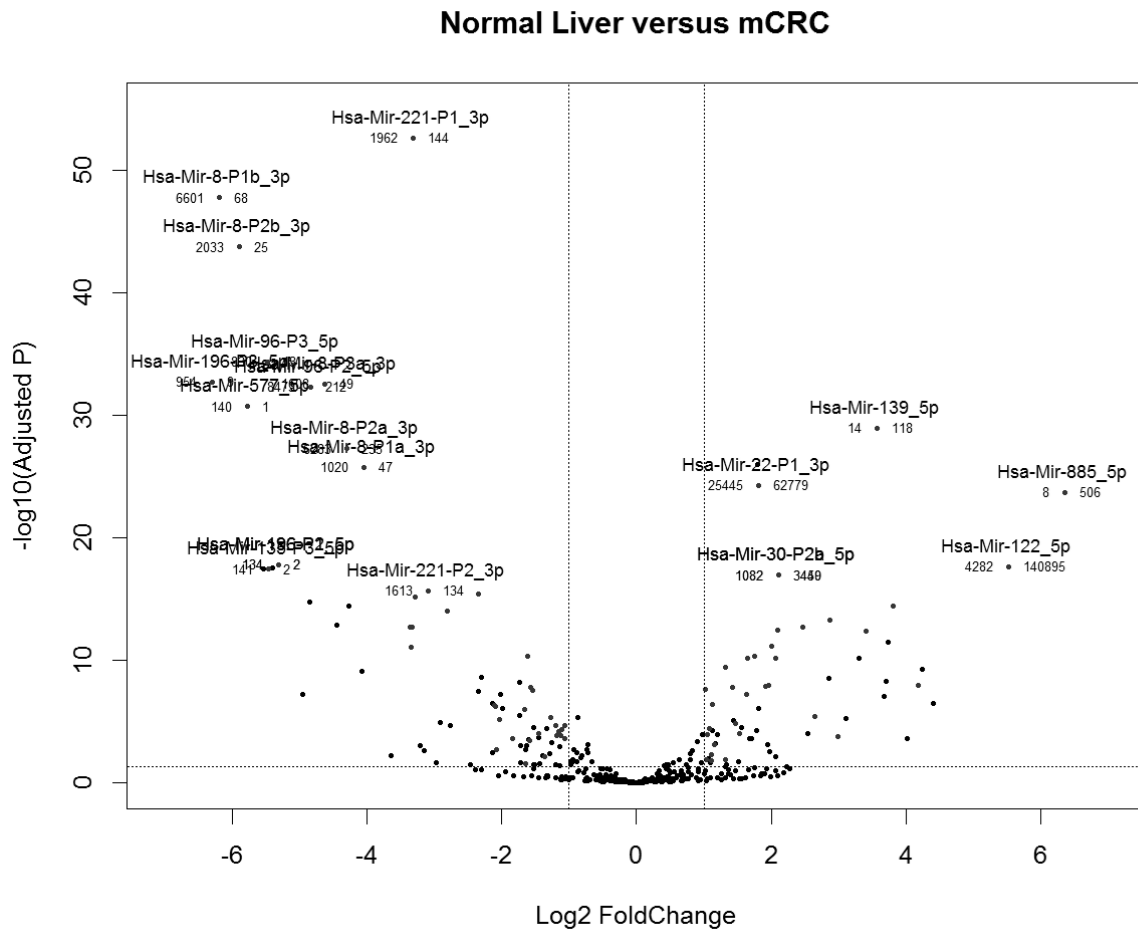


Appendix Figure 2 Volcano Plot nCR vs pCRC Plot of $-\log_{10}$ padj against LFC of nCRC vs pCRC. Signature miRNA (red) have more than one LFC, padj < 0.05 and one group > 100 RPM. Only top 10 significant miRNA highlighted.

↓ miRNA in nCR vs pCRC	↑ miRNA in nCR vs pCRC
Hsa-Mir-26-P1_5p	Hsa-Mir-135-P3_5p
Hsa-Mir-26-P2_5p	Hsa-Mir-224_5p
Hsa-Mir-148-P3_3p	Hsa-Mir-31_5p
Hsa-Mir-15-P2c_5p	Hsa-Mir-21_5p
Hsa-Mir-145_5p	Hsa-Mir-19-P1_3p
Hsa-Mir-378_3p	Hsa-Mir-17-P3a_5p
Hsa-Mir-10-P1b_5p	Hsa-Mir-17-P1a_5p
Hsa-Mir-338-P1_3p	Hsa-Mir-15-P1d_5p
Hsa-Mir-133-P1_3p	Hsa-Mir-17-P4_5p
Hsa-Mir-133-P2_3p	Hsa-Mir-96-P2_5p

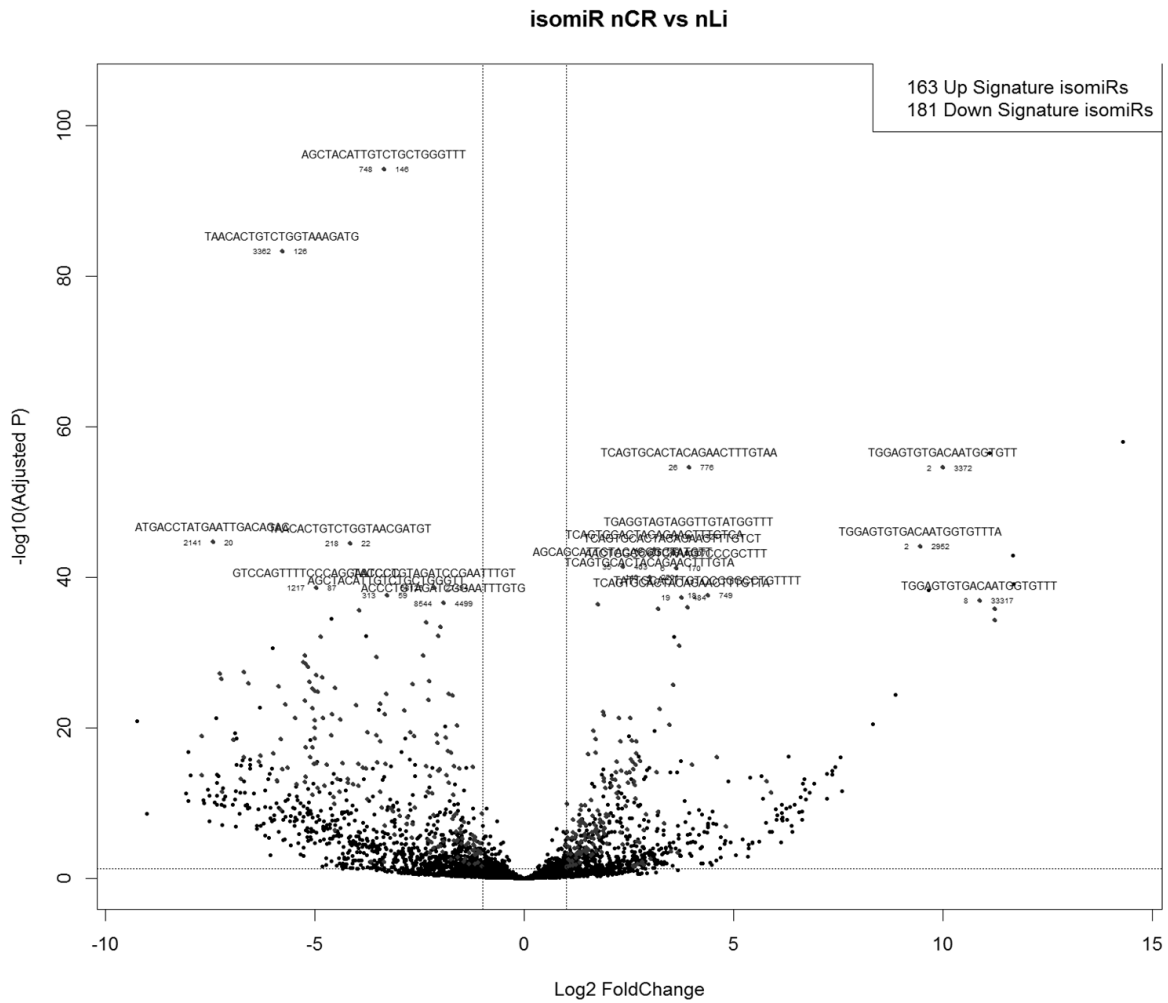
Appendix Table 1 Top 10 signature miRNA in nCR vs nLi. Only top 10 by significance level up- and down regulated are shown

4.1.3 Volcano plot nLi versus CLM



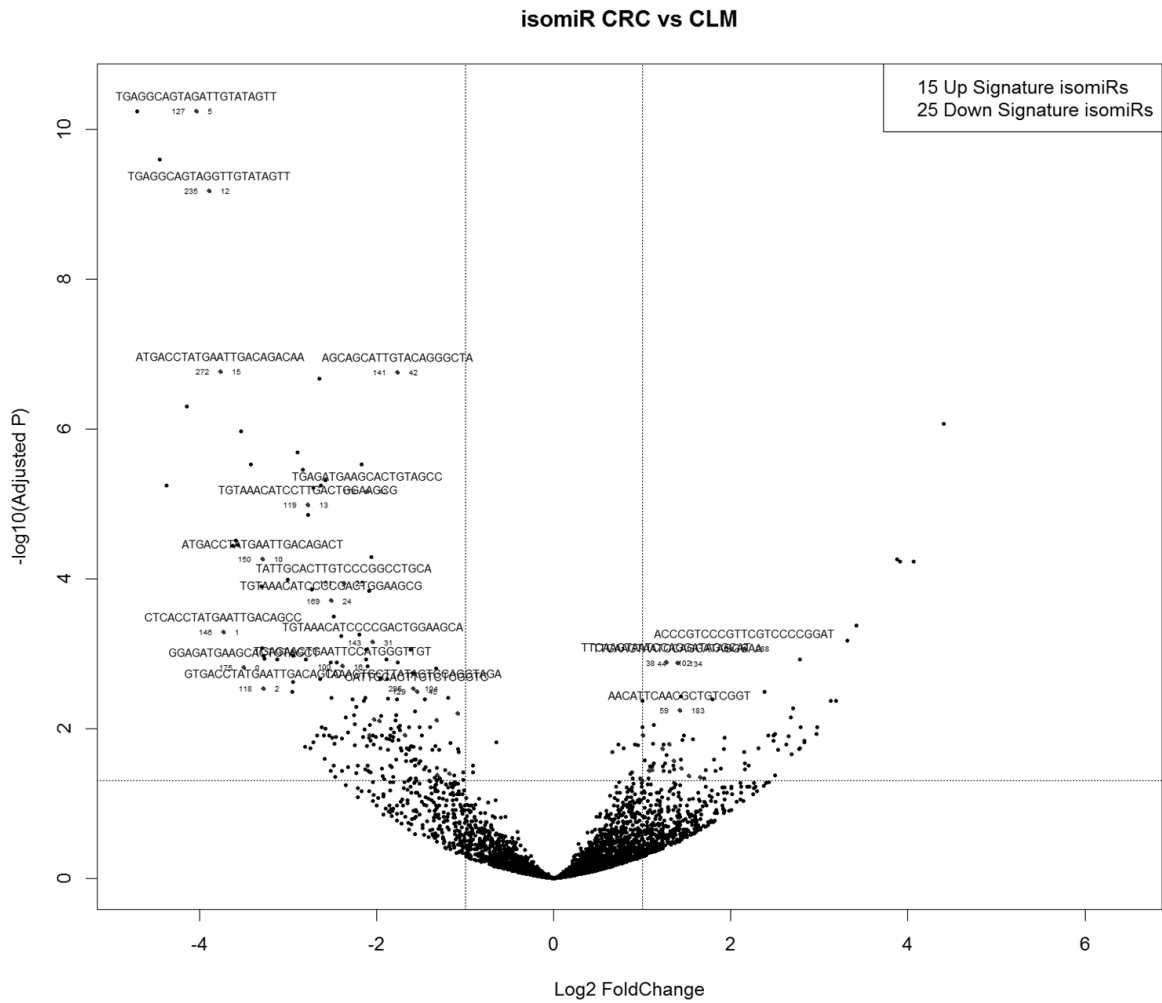
Appendix Figure 3 Volcano Plot nLi vs CLM Plot of $-\log_{10}$ adjusted p-value against Log2 Fold Change of nLi vs CLM Signature miRNA highlighted.

4.1.4 Volcano plot of isomiRs in nCR and nLi



Appendix Figure 4 IsomiR Volcano Plots nCR vs nLi Plots of $-\log_{10}(\text{padj})$ against LFC of nCR vs nLi. IsomiRs defined as all fragments mapped to MirGeneDB annotated pri-miRNA sequences, and fragment not identical to canonical miRNA. IsomiRs must be expressed in at least 50 % of samples. Signature isomiRs must have > 100 RPM in one of the groups, and no signature isomiR may have one group with 0 counts.

4.1.5 Volcano plot of isomiRs in pCRC and CLM



Appendix Figure 5 IsomiR Volcano Plots pCRC vs CLM Plots of $-\log_{10}(\text{padj})$ against LFC of pCRC and CLM (bottom). IsomiRs defined as all fragments mapped to MirGeneDB annotated pri-miRNA sequences, and fragment not identical to canonical miRNA. IsomiRs must be expressed in at least 50 % of samples. Signature isomiRs must have > 100 RPM in one of the groups, no signature isomiR may have one group with 0 counts and no signature isomiR may be signature in normal tissue.