Navigating the micronome

A systematic study of both the external effects of microRNAs on gene repression networks, and the contribution of microRNA terminal loops to microRNA function.

by

Elize Astghik Shirdel

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Medical Biophysics
University of Toronto

© Copyright by Elize Shirdel 2012
Navigating the micronome

Elize Astghik Shirdel
Doctor of Philosophy
Medical Biophysics
University of Toronto
2012

Abstract

The first aim of this thesis is to examine relationships between microRNAs targeting gene networks, combining knowledge from microRNA prediction databases into our microRNA Data Integration Portal (mirDIP). Modeling the microRNA:transcript interactome – referred to as the micronome – to build microRNA interaction networks of signalling pathways, we find genes within signalling pathways to be co-targeted by common microRNAs suggesting an unexpected level of transcriptional control. We identify two distinct classes of microRNAs; universe microRNAs, which are involved in many signalling pathways; and intra-pathway microRNAs, which target multiple genes within one signalling pathway. We find universe microRNAs to have more targets, to be more studied and more involved in cancer signalling than their intrapathway counterparts.

The second aim was to undertake a more focused view, analyzing the characteristics of microRNAs within the micronome itself beginning with a focus on the under-examined microRNA terminal loop across the micronome to determine if this region of the microRNA structure might contribute to microRNA functioning. We have identified 2 main classes of microRNAs based on loop structure – perfect and occluded, which show biological relevance.
We found regulatory motifs within microRNA terminal loops and found a large number of Frequently Occurring Words (FOWs) significantly overrepresented across the micronome. Set analysis of in vitro secreted microRNAs, microRNA expression across a panel of normal tissues, and microRNAs shown to be secreted in lung cancer shows that specific microRNA loop motifs within these groups are significantly overrepresented – suggesting that microRNA terminal loops harbour sequences bearing microRNA processing and localization signals.
Dedication

This is for my dad. Although both of my parents would have loved to see this thesis – I think that it would have made by dad the proudest. He did warn me about going into academia, but I think that he, himself, understood the mysterious draw. Now that I have a son of my own, with a mind of his own -- I can see how I must have driven my dad crazy when I was younger – always wanting to do things my way, figuring it out on my own and discounting his years of experience and his depth of knowledge. But seeing my independent and stubborn toddler makes me realize that it’s a little bit innate, that irreverence – and I’m sure that he must have understood that, too. I hope.
Acknowledgements

The path to this thesis has been a long and winding road, with many peaks and many valleys. I have to give credit to Dr. Tak Mak for first getting me hooked on the idea of genetic analysis of any kind, luring me into the world of biology with breast cancer CGH arrays – which for me, at the time, were a holy grail of hidden information that took over my life.

I’m quite sure that I owe a lot of my degree to Dr. Igor Jurisica. He helped me to pick up all the pieces when it looked like it everything had fallen apart – then stepped back to let me find my own way. Always there to offer a suggestion, to share in excitement or to help think through a problem, I doubt that I would have finished without his presence.

Special thanks to the guidance and time generously donated by my committee members, Drs. Lea Harrington, Mitsu Ikura and Mark Minden.

I must thank my in-laws Rick and Kate Smith for their encouragement and family support for the last half of my graduate work, which coincided with additions to our family

It has been a pleasure to work with all members of the Jurisica lab, but there are a few that I should thank in particular. Christian Cuumba always solved my cluster-related issues, Kevin Brown for getting me started, Max Kotylar always offered helpful advice and positive encouragement, Dan Strumpf offered encouragement and different perspectives, Rohan Ramnarine for collaborations, Chiara Pastrello for enthusiasm during our out-of-the-box brainstorming sessions, Kristen for quiet support, and Wing for mirDIP brilliance.

During my time in the Mak lab, there are a number of people who kindly took time to help get my feet wet at the bench and who taught me many techniques, including Thorsten Berger, Sophie Vasseur, Jennifer Liepa, Margareta Wilhelm, Enrico Arpaia, Christiane Knobbe, Patrick Reilly, Susan McCracken, Arda Shahinian and Gordon Duncan.

Along with tears in both celebration, and in failure, many informal lessons on biology, and the kind of emotional support that one can only get from fellow graduate students on the same exciting but arduous path, I am grateful to David McIlwain, Stephanie Sue, Valentina Lapin and Megan Nelles.

My parents, Zhorik and Judy Shirdel, died in the year before I began this degree. I think that they suspected that a Ph.D. was on the horizon. Although I can’t thank them for the huge hole that they left in my life, I can thank them for the spectacular family life that we had, the weekends full of puzzles and cryptoquips, the debates and the discussions and the endless laughter that we shared for 24 years.

I must thank my only sister, Hilda Shirdel – for unconditional love and support through what has been a tumultuous few years. Everybody should be so lucky as to have a sister like her. My maternal grandmother, Maud Snowdon, has always been a source of support and pride for me. As has Lucy. Finally, I must admit - I have the greatest husband in the world. Isaac Smith is a
true partner, splitting parental leaves for both Milton and Expected-Baby-Number-2 with me so that my work could continue on as uninterrupted as possible, and always having confidence in me, no matter where my doubts lie. Finally thanks to Milton for being such a ball of sunshine – and for finally sleeping through the night. At 19 months. Better late than never, my baby :}
Table of Contents

Abbreviations ................................................................. xi

Measures for Evaluating Predictions ........................................ xiv

List of Tables ........................................................................ xv

List of Figures ........................................................................ xvi

List of Appendices .................................................................. xviii

1 Introduction ......................................................................... 1

1.1 MicroRNAs ......................................................................... 2

1.1.1 Discovery ................................................................. 2

1.1.2 MicroRNA Processing .................................................. 4

1.1.3 MicroRNA Function ...................................................... 9

1.1.4 MicroRNA Structure .................................................... 15

1.1.5 MicroRNA Regulation .................................................. 20

1.1.6 Secreted or Circulating microRNAs ................................. 22

1.1.7 MicroRNA Associations with Disease ............................... 25

1.1.8 MicroRNA Targets ....................................................... 27

1.1.9 Classification ............................................................. 29

1.2 Biological Interaction Networks .......................................... 30

1.2.1 Protein Interaction Networks ......................................... 30

1.2.2 MicroRNA Interaction Networks .................................... 31

1.2.3 Gene Regulatory Networks ........................................... 32

1.3 Summary of Research Contributions .................................. 32

1.3.1 Chapter 2 – Navigating the micronome .......................... 32
1.3.2 Chapter 3 - A Global Analysis of microRNA Loops Identifying Secretion-related Motifs and Structures .......................................................... 33

2 NAViGating the microNome ........................................................................ 34
   2.1 Abstract ................................................................................................. 34
   2.2 Background .......................................................................................... 36
   2.3 Results and Discussion ......................................................................... 37
      2.3.1 Characteristics of microRNA predicting databases ......................... 37
      2.3.2 MicroRNA prediction database similarities .................................... 43
      2.3.3 Comparing microRNA prediction databases to the truth .................. 47
      2.3.4 Construction of microRNA interaction networks .............................. 55
      2.3.5 Universe and Intrapathway microRNAs ............................................ 65
   2.4 Conclusions .......................................................................................... 74
      2.4.1 MicroRNA Prediction Databases ...................................................... 74
      2.4.2 mirDIP ............................................................................................. 75
      2.4.3 Discovery of Universe and Intrapathway MicroRNAs in Interaction Networks .. 75
   2.5 Materials and Methods ....................................................................... 76

3 A Global Analysis of MicroRNA Loops Identifying Secretion-related Motifs and Structures ................................................................. 82
   3.1 Abstract ............................................................................................... 82
   3.2 Background .......................................................................................... 83
   3.3 Results and Discussion ....................................................................... 88
      3.3.1 MicroRNA Seeds may not be Revealing the Whole Story .............. 88
      3.3.2 An Examination into MicroRNA Loops .......................................... 93
      3.3.3 Conservation across Loops but not Seeds Exists for Some microRNA Families ................................................................. 95
3.3.4 Common loop motifs cannot be explained by similar stem loops containing identical seeds ................................................................. 96
3.3.5 Binding-Based Interactions .......................................................................................................................... 99
3.3.6 Loop Structure-based Classification ........................................................................................................... 109
3.3.7 Clues to MicroRNA Loop Function ........................................................................................................... 117
3.4 Conclusions .................................................................................................................................................. 133
3.5 Materials and Methods .................................................................................................................................. 134
3.6 Acknowledgements ....................................................................................................................................... 137
4 Discussion ..................................................................................................................................................... 138

4.1 Navigating the Micronome ............................................................................................................................ 139
  4.1.1 mirDIP .................................................................................................................................................... 139
  4.1.2 Combinatorial MicroRNA Action ........................................................................................................... 139
  4.1.3 Pathway regulation at the RNA level ...................................................................................................... 141
  4.1.4 Classification of microRNAs by Targeting Characteristics ................................................................ 142
  4.1.5 Contributions and Limitations ............................................................................................................... 143

4.2 MicroRNA Loops .......................................................................................................................................... 144
  4.2.1 A Broader role for microRNAs through the Recognition of Distinct Classes based on Loop Structures ........................................... 144
  4.2.2 Emerging Roles for MicroRNAs as Master Regulators ........................................................................ 145
  4.2.3 Loops as recognition sequences ............................................................................................................. 146

4.3 Future Directions .......................................................................................................................................... 147
  4.3.1 MicroRNA Organization .......................................................................................................................... 147
  4.3.2 MicroRNA involvement in Translation through tRNA Interactions .................................................... 148
  4.3.3 mirDIP Expansion .................................................................................................................................. 150

4.4 Conclusions .................................................................................................................................................. 151
Copyright Acknowledgements.............................................................................................................. 153

Appendices................................................................................................................................................ 154

5 Chapter 2 Supplementary Figures ........................................................................................................ 155

5.1.1 mirDIP – Interface description........................................................................................................ 155

5.1.2 Using mirDIP ...................................................................................................................................... 157

5.2 Chapter 2 Supplemental Materials ....................................................................................................... 161

5.3 Chapter 2 Supporting Methods ............................................................................................................. 162

5.4 Additional Co-authored Publication .................................................................................................... 163

Thesis References ........................................................................................................................................ 203
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGO</td>
<td>Argonaute (Member of the family)</td>
</tr>
<tr>
<td>ARE</td>
<td>AU-rich elements</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine Triphosphate</td>
</tr>
<tr>
<td>BMP</td>
<td>Bone Morphogenic Protein</td>
</tr>
<tr>
<td>CD4</td>
<td>Cluster of Differentiation 4</td>
</tr>
<tr>
<td>CD8</td>
<td>Cluster of Differentiation 8</td>
</tr>
<tr>
<td>ceRNA</td>
<td>Competing Endogenous RNA</td>
</tr>
<tr>
<td>DDX17</td>
<td>Dead-box Polypeptide 17</td>
</tr>
<tr>
<td>DDX20</td>
<td>Dead-box Polypeptide 20</td>
</tr>
<tr>
<td>DDX5</td>
<td>Dead-box Protein 5</td>
</tr>
<tr>
<td>DGCR8</td>
<td>DiGeorge Syndrome Critical Region 8</td>
</tr>
<tr>
<td>EGFR</td>
<td>Epidermal Growth Factor Receptor</td>
</tr>
<tr>
<td>EIF4E</td>
<td>Eukaryotic Translation Initiation Factor 4E</td>
</tr>
<tr>
<td>ERBB2</td>
<td>v-Erb-b2 Erythroblastic Leukemia Viral Oncogene Homologue 2</td>
</tr>
<tr>
<td>FGFR</td>
<td>Fibroblast Growth Factor Receptor</td>
</tr>
<tr>
<td>FOW</td>
<td>Frequently Occurring Word</td>
</tr>
<tr>
<td>FOXO3</td>
<td>Forkhead Box O3</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GPCR</td>
<td>G Protein-Coupled Receptor</td>
</tr>
<tr>
<td>hnRNP A1</td>
<td>heterogeneous nuclear ribonucleoprotein A1</td>
</tr>
<tr>
<td>HSP70</td>
<td>Heat Shock Protein 70</td>
</tr>
<tr>
<td>HSP90</td>
<td>Heat Shock Protein 90</td>
</tr>
<tr>
<td>I2D</td>
<td>Interologous Interaction Database</td>
</tr>
<tr>
<td>IRES</td>
<td>Internal Ribosome Entry Site</td>
</tr>
<tr>
<td>Jak-STAT</td>
<td>Janus Kinase-Signal Transducer and Activator of Transcription</td>
</tr>
<tr>
<td>kDa</td>
<td>kiloDalton</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>KSRP</td>
<td>KH-type Splicing Regulatory Protein</td>
</tr>
<tr>
<td>MAPK</td>
<td>Mitogen-Activated Protein Kinase</td>
</tr>
<tr>
<td>mirDIP</td>
<td>MicroRNA Data Integration Portal</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>MSC</td>
<td>Mesenchymal Stem Cells</td>
</tr>
<tr>
<td>mTOR</td>
<td>Mechanistic Target of Rapamycin</td>
</tr>
<tr>
<td>NAViGaTOR</td>
<td>Network Analysis, Visualization and Graphing Toronto</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>NSCLC</td>
<td>Non-Small Cell Lung Carcinoma</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>PAX</td>
<td>Paired Box</td>
</tr>
<tr>
<td>PDGF</td>
<td>Platelet Derived Growth Factor</td>
</tr>
<tr>
<td>PDK1</td>
<td>Pyruvate Dehydrogenase Kinase isozyme 1</td>
</tr>
<tr>
<td>PI3K</td>
<td>Phosphoinositide 3-Kinase</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>Phosphoinositide 3-Kinase, Catalytic, Alpha Peptide</td>
</tr>
<tr>
<td>PIWI</td>
<td>P-element Induced Wimpy Testes</td>
</tr>
<tr>
<td>pSILAC</td>
<td>Pulsed Stable Isotope Labeling by Amino Acids in Cell Culture</td>
</tr>
<tr>
<td>PTEN</td>
<td>Phosphatase and Tensin Homolog</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-Induced Silencing Complex</td>
</tr>
<tr>
<td>RPS6KB1</td>
<td>Ribosomal Protein S6 Kinase beta-1</td>
</tr>
<tr>
<td>TARBP2/TRBP</td>
<td>TAR RNA-binding protein 2</td>
</tr>
<tr>
<td>TP53</td>
<td>Tumour Protein 53</td>
</tr>
<tr>
<td>TSC1</td>
<td>Tuberous Sclerosis Protein 1</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
<tr>
<td>VEGF</td>
<td>Vascular Endothelial Growth Factor</td>
</tr>
</tbody>
</table>
Measures for Evaluating Predictions

**Precision**

\[
\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}
\]

**Recall**

\[
\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}
\]
List of Tables

Table 2-1 MicroRNA Prediction Databases used in our study...................................................... 38
Table 2-2 Characteristics of MicroRNA Prediction Databases. ................................................. 39
Table 2-3 Characteristics of High-throughput Experiments used in our study......................... 47
Table 3-1 Determination of MicroRNA Loops drawn from mirbase [25, 26]............................ 94
Table 3-2 Complements and Reverse Complements within the set of Frequently Occurring Words ........................................................................................................................................ 106
Table 3-3 Classes of MicroRNA Structures.............................................................................. 111
Table 3-4 Gene Enrichment of Mesenchymal Stem Cell Secreted MicroRNAs..................... 122
Table 3-5 Lung Cancer MicroRNA Studies ............................................................................... 128
Table 3-6 Lung Cancer MicroRNAs by Secretionary Status ................................................... 130
List of Figures

Figure 1-1 MicroRNA Species .......................................................... 4
Figure 1-2 MicroRNA Processing ...................................................... 6
Figure 1-3 Precursor microRNA Structure (pre-microRNA) ............ 7
Figure 1-4 Methods of Protein Expression Repression by microRNAs and the RISC complex ... 11
Figure 1-5 Differential Processing Between Plant and Animal microRNAs .......... 17
Figure 1-6 Mir-181a/c-1 function in T-cells is pre-microRNA stem-loop dependent [95] .... 19
Figure 2-1 MicroRNA prediction database characteristics .................. 45
Figure 2-2 MicroRNA target prediction processing and evaluation .......... 49
Figure 2-3 MicroRNA interaction network for assembly of PI3K subunits ....... 58
Figure 2-4 MicroRNA interaction network elements of the PI3K pathway ...... 60
Figure 2-5 MicroRNA interaction network characteristics ................... 62
Figure 2-6 Micronome Roadmap .................................................... 66
Figure 2-7 Comparison of universe and intra-pathway microRNAs .......... 68
Figure 2-8 Expression of universe and intra-pathway microRNAs ............ 73
Figure 3-1 The importance of microRNA seed sequences across microRNA prediction databases .................................................................................................................... 89
Figure 3-2 Precursor microRNA (pre-microRNA) ............................... 91
Figure 3-3 Properties the MicroRNA Seed Space ............................... 92
Figure 3-4 Alignments of the mir-154 and let-7 families ....................... 95
Figure 3-5 Loop sequence alignment of all microRNAs that share a seed sequence ........ 97
Figure 3-6 Common seeds cannot explain most microRNAs with shared loop sequences .... 98
Figure 3-7 Data Processing Pipeline for FOWs .................................. 99
Figure 3-8 Likelihood of FOWs........................................................................................................... 100
Figure 3-9 Basepair Distribution of microRNA species......................................................................... 101
Figure 3-10 Wordle showing Maximal Loop 6-letter FOWs................................................................. 102
Figure 3-11 The Sequences of 3-letter Apexes of MicroRNA Hairpin loops are Biased Towards
Complement/Reverse Complement Pairs................................................................................................. 107
Figure 3-12 Kissing Complexes involved in HIV-1 replication................................................................. 108
Figure 3-13 Loop-Loop Kissing Complexes in R1 replication................................................................. 109
Figure 3-14 Physical Characteristics of MicroRNA Loops...................................................................... 112
Figure 3-15 Examining MicroRNA Loop Structure in the Human Let-7 Family................................. 114
Figure 3-16 Comparison of Perfect and Occluded Loops.................................................................... 116
Figure 3-17 Network Diagram of Common Loop Motifs from Mesenchymal Secreted MicroRNAs
............................................................................................................................................................ 119
Figure 3-18 Heatmap of Normal Tissues Clustered by Common MicroRNA Loop Motif............ 124
Figure 3-19 KSRP-dependent MicroRNAs............................................................................................ 125
Figure 3-20 MicroRNA Loop Structures in KSRP-dependent MicroRNAs........................................... 126
Figure 3-21 MicroRNA Loop Motifs by Secretory Status.................................................................... 132
Figure 4-1 Integration of multiple lung cancer studies to investigate microRNA secretion..... 150

xvii
List of Appendices

5 Chapter 2 Supplementary Figures ........................................................................................................ 154
  5.1.1 mirDIP – Interface description ..................................................................................................... 154
  5.1.2 Using mirDIP ............................................................................................................................... 156
5.2 Chapter 2 Supplemental Materials .................................................................................................... 160
5.3 Chapter 2 Supporting Methods ........................................................................................................ 161
5.4 Additional Co-authored Publication ............................................................................................... 162
Chapter 1

Introduction

This thesis takes a bioinformatics approach to the study of microRNAs. Using computational approaches, we examine the micronome as a system, focusing on both: the relationships between the elements of the micronome and other entities within the cell, specifically the combinatorial targeting of genes by multiple microRNAs using a network-based approach; and then examining the relationships between the elements of the micronome focusing our study on one structural element of microRNAs – the terminal loop, which has long been down-played as unimportant in microRNA processing and functionality.

Just as genes have been classified based on their structure, expression, motifs – we begin to realize that microRNAs are a diverse group of genetic elements – possibly functioning in extremely distinct manners and we are impeding our understanding of the cellular system by lumping all microRNAs together in current models postulating the common function of the elements of the micronome.
1.1 MicroRNAs

1.1.1 Discovery

MicroRNAs are short, but important non-coding RNA sequences that regulate gene expression [1]. They are thought to primarily target the 3’ Untranslated Regions (UTRs) of mRNA, disrupting their ability to be translated into proteins, sometimes repressing the expression of the mRNA itself [2-9]. MicroRNA prediction algorithms generally pair the seed region of the microRNA (bases 2-7 from the 5’ end of the mature microRNA) to a cognate mRNA sequence. However, this binding is complicated by many factors, not the least of which is that imperfect microRNA:mRNA binding occurs, and thus single base-pair mismatches and G:U wobble base-pairs must be considered.

Discovery of the first microRNA – lin-4 in worm (C. Elegans) [10], its further characterization in 1989 [11], annotation as a non-coding RNA in 1993 with a sequence complementary to the lin-14 3’ UTR [1, 12], and functional characterization as having a translational repression effect later that year [13] opened a rich research field. Many subsequent in vitro experiments and computational predictions aimed at uncovering microRNA:target relationships to fathom microRNA effects on gene expression regulation. With the discovery of a second nematode microRNA – let-7, which targets lin-41 and hbl-1, the concept of microRNAs made the jump from worms to higher species, since let-7 had well-known homologues even in humans [14-16]. Coining the term “microRNA” for this class of non-coding gene regulators in three back-to-back Science papers in 2001 [17-19], the discovery of microRNAs had crossed over to the human domain, and finding microRNA targets became a high priority. After the first
bioinformatics attempt at predicting plant microRNAs [5], many microRNA prediction algorithms, for both fly (*D. melanogaster*) and human (*H. sapiens*), were developed [20-22]. More than 10 public databases for microRNA:mRNA target prediction have been created, all using different algorithms and approaches, but the challenge of identifying and validating targets remains a tough one. [20, 23-40].

Databases used in our study are discussed in Chapter 2.
1.1.2 MicroRNA Processing

Most microRNAs are transcribed by RNA polymerase II in the nucleus into primary transcripts (pri-miRNAs), which can be longer than 1 kb in length [41, 42] (See Figure 1-1 for a schematic of microRNA precursor species).

**Figure 1-1 MicroRNA Species**
Schematic of Primary MicroRNAs, Precursor MicroRNAs and Mature MicroRNAs.
These pri-miRNAs have a stem-loop structure resulting from the complementary base pairing between 2 different regions within the primary transcript that form the stem. On the other end of the stem from the loop structure are the base-pair terminal overhangs which end in either the 5’ cap or the poly(A) tail, making the microRNA structures quite similar those of protein coding genes [43]. Occasionally, microRNAs are clustered together in the genome and are transcribed into one long transcript that forms several stem-loop structures [17] [18].

Still in the nucleus, the first processing step occurs when the pri-miRNA transcript is bound first by DGCR8, and then by Drosha, an RNAse III endonuclease, forming the Microprocessor complex. This leads to cleavage of the pri-miRNA eliminating the terminal overhangs, producing a 60-70 base pair long stem-loop structure called a pre-miRNA or microRNA precursor [41, 44-46] (Figure 1-2). This cleavage leaves a staggered cut pattern at the base of the stem resulting in a 5’ phosphate and a 3’ overhang by ~2 nucleotides [47, 48]. This cutting action defines one end of the mature microRNA that will be later be loaded into the RISC complex [48]. After Drosha cleavage this precursor is exported out of the nucleus by Ran-GTP and Exportin 5. [49, 50] [51]
Figure 1-2 MicroRNA Processing.
MicroRNA processing begins in the nucleus with transcription of the microRNA gene into the pri-mRNA. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer [52], copyright (2006)
Once in the cytoplasm, the single mammalian Dicer – another RNAse III endonuclease - recognizes the double stranded RNA, perhaps by its 5’ phosphate tail opposite a 3’ overhang [48]. Dicer, minimally partnered with an RNA-binding protein TAR RNA-binding protein 2 (TARBP2/TRBP), cuts the precursor microRNA approximately 2 helical turns from the base of the stem loop[53]. This chopping removes the loop of the precursor microRNA and leaves another 5’ phosphate and 3’ ~2 nucleotide overhang opposite the first one. [54-57]

This small (~22 nucleotide) double stranded RNA is now the microRNA:microRNA* duplex. One of the strands of RNA is the mature sequence that will be loaded into the RISC complex and the other will eventually be a substrate of the complex and cleaved. The microRNA:microRNA* duplex is likely highly transient in the cell compared to the mature microRNA as observed by its low cloning frequency relative to mature microRNAs in efforts to clone C. elegans microRNAs [58]. Figure 1-3 shows the location of the mature microRNA sequences and seeds within the pre-microRNA sequence.

**Figure 1-3 Precursor microRNA Structure (pre-microRNA).**
Schematic of microRNA components and nomenclature.
Early assumptions were that the strand designated as the mature strand of the miRNA:miRNA* duplex was the one with the least tight 5’ end pairing – meaning having the end the smallest run of matching base pairs [59, 60]. Research now shows that both strands may participate equally or differentially in target binding relationships (although in any duplex only one strand will be incorporated into the RISC, with the other being degraded), and we are getting away from the mature- / passenger(*)- strand nomenclature and moving towards a less-biased 5p/-3p- naming convention to recognize that both microRNA strands are important depending on the context [61].

1.1.2.1 The RISC Complex

The actual functional work done by mature microRNAs to affect mRNA levels in vivo is done through the association of the mature microRNA strand with the RNA Induced Silencing Complex, abbreviated RISC. The RISC is composed of several proteins that play key roles in incorporating mature microRNAs into the complex and performing cleavage and or repression actions on specific mRNA targets determined by the sequence of the mature microRNA.

There are four human Argonaute (AGO) proteins, with varying expression levels across different cell types. Human AGO proteins are characterized by 3 domains: PAZ and MID domains hold the 3’ and 5’ ends of the microRNA respectively and the PIWI domain – characterized by an Asparagine-Asparagine-Histidine motif, which is the site of RNA cleavage for those Argonautes capable of inducing it [62, 63]. It seems that the sole human cleavage-competent Argonaute family member is Argonaute2 (AGO2, EIF2C2), which has been identified as the endonuclease
component of the complex that performs mRNA cleavage of target messages [63-66] and it is likely that the particular combination of microRNA and AGO in the RISC complex determines the biological functioning of the complex [51]. Other proteins implicated in the RISC include: TRBP, which has been shown to perform the recruitment of mature microRNA to AGO2 after the Dicer activity has occurred [67] and Gemin3 and Gemin4 (DDX20) - 2 RNA helicases. [68] It seems likely that more RISC component proteins exist.

1.1.2.2 RISC Maturation

Outside proteins help in loading the mature duplex onto the AGO protein (possibly Heat Shock Proteins 70 and 90 (HSP70, HSP90)). This is an ATP-dependent activity, due to the changing conformation of the AGO proteins to accept the duplex [69, 70]. The maturation step occurs with the loss of the secondary duplex strand, likely a dissociation through ATP-independent unwinding, since most human AGOs lack the slicer activity that was initially thought to be involved[70, 71].

1.1.3 MicroRNA Function

Whether a particular target mRNA is targeted for degradation or simply translational repression, and where exactly in the process of translation that repression occurs is the subject of much debate. Studies have indicated that the 5’ end of the microRNA is the most important core element of the microRNA for target binding action. Residues 2-7 from the 5’ end of the mature
microRNA are highly conserved and tend to show a higher degree of complementarity with the target mRNA sequence [21, 22, 36, 58, 72], likely because this is the portion of the microRNA sequence that the RISC complex presents to the target mRNA to induce repression/cleavage [36].

1.1.3.1 Methods of mRNA Repression by microRNA

There are several possibilities for the mechanism of action of microRNA gene expression inhibition whether at the protein or mRNA level. With reference to Figure 1-4, they are discussed below.

1.1.3.1.1 Translational Repression

It was initially thought that the complementarity between the microRNA and the target sequence would dictate whether the induction of cleavage or translational repression of the protein coded by the target mRNA could occur. It turns out the situation may not be so clear cut. [73] [4]
Figure 1-4 Methods of Protein Expression Repression by microRNAs and the RISC complex. From [74]. Reprinted from Cell, 132, Eulalio, A., Hunztinger, E., Izaurralde, E., Getting to the root of miRNA-mediated gene silencing, Pages 9-14, Copyright 2006, with permission from Elsevier.
1.1.3.1.2 Post Initiation Translational Repression

Early studies in *C. elegans* and later studies in cell culture suggest that loss of protein expression is through a post-translational initiation step, since the presence of polyribosomes – clusters of ribosomes involved in active translation of the mRNA – that are sensitive to drugs that inhibit translation have been observed [6, 73, 75, 76]. To reconcile the observation that while active translation of the microRNA target seems to be occurring, the protein product is undetectable, Nottrott et al. proposed that concurrent protein degradation must be occurring (Figure 1-4 B), based on a lack of evidence for the protein’s existence rather than evidence for the concurrent degradation – and the proteases responsible for this proposed degradation remain unknown (and the proteasome ruled out) [76]. Petersen et al. proposed the idea of a ribosome drop-off model for protein repression (Figure 1-4 A). Their system involved overexpression of a reporter with multiple target sites partially complementary to a microRNA mimic. Expression of the reporter saw association with polyribosomes, with the reporter expression inhibited by the mimic. The clinching piece of evidence was that under translational-initiation-blocking conditions, dissociation of ribosomes from the reporter targeted by the microRNA mimic was much quicker compared to a control. The implication being that ribosomes stayed associated with the microRNA-bound reporter for less time overall – perhaps falling off during the translational process. Since reporters containing 5’ UTR Internal Ribosome Entry Site (IRES) sequences are able to be repressed by microRNAs, they presented further evidence that microRNA repression is a 5’ cap-independent process, since IRES sequences initiate translation without the presence of the cap [73].
1.1.3.1.3 Inhibition of the Initiation of Translation

Contradicting studies show a likely inhibition of translation initiation model. Pillai et al. showed in 2005 that mRNA targeted by microRNAs and the RISC complex were not associated with polyribosomes and this study and others shows that mRNAs translated through IRES mechanisms were unresponsive to repression by microRNAs – showing support for the 5’ cap-dependent nature of microRNA translational repression [77-80]. There is some evidence in the literature that Argonout proteins have sequence similarity to eIF4E – a cytoplasmic cap-binding protein - might be able to displace it, preventing translation initiation (Figure 1-4 C), although the physical logistics of the interaction have yet to be established in a physiologically relevant setting[81]. In 2007 Chendrimada et al. proposed a model of interference between AGO2 and eIF6 and the large ribosomal subunit [82]. This might prevent the association of the large and small ribosomal subunits, thus preventing translation (Figure 1-4 D).

1.1.3.2 mRNA Cleavage

Although initial reports showed that animal microRNAs affected protein expression levels with no effect on mRNA levels, recent reports have negated this result [83, 84]. Further, Lim et al. in 2005 showed that overexpression of a microRNA results in lower expression of mRNAs containing binding sites for that particular microRNA[85].

In the case where complete complementarity results in mRNA cleavage, the target sequence is cut by AGO2 in the RISC complex between the 10th and 11th base pairs of the sequence complementary to the mature microRNA, counting from the 5’ end of the mature microRNA.
Once cleavage occurs, the microRNA and the RISC remain intact and can proceed to cleave other mRNA [3, 56, 86].

1.1.3.3 mRNA Degradation through Deadenylation and Decapping

In the case of incomplete complementarity, mRNAs are thought to be degraded by the general mRNA degradation machinery – where microRNAs have been shown to accelerate deadenylation and decapping of their cognate sequences (Figure 1-4 F) [84, 87]. This decay process has been noted to require the presence of many proteins including: Argonautes, GW182 (a component of p-bodies), a deadenylase complex, decapping enzymes and decapping activators [84, 88]. Eulalio et al. have shown that the process of either mRNA decay or translational repression is not guided solely by the microRNA itself, due to the observation that one microRNA can show degradation in one circumstance while decay is guided by another. Suggestions that this decision is based on the nature of the microRNA-mRNA duplexes, the number of binding sites in the 3’ UTR along with their proximity to one another have all been proposed [27, 89]. Figure 1-4 E outlines a model proposed by Wakiyama et al. who advocate a deadenylation situation through their observations that both a 5’ cap and poly(A) tail were required for silencing. They suggest that once deadenylation is occurring due to microRNA targeting, the cap and the poly(A) tail are no longer able to circularize the mRNA through their interactions with PABPC1 and eIF4G, greatly decreasing translational efficiency [80].
1.1.3.4 Sequestration in P-bodies

Many components involved in mRNA degradation by microRNAs have been found in p-bodies including the microRNAs themselves, their targets and Argonaute proteins, along with many of the elements of the mRNA decaying process mentioned in the previous section. Several groups have proposed a model that this presence is a sequestration of microRNAs and their targets from the translational machinery preventing their expression within the cell – although it’s not clear if this is the cause of the decreased protein expression, or a result of it, due to changes in translation. Studies in cells lacking p-bodies do not show any impairment of microRNA gene silencing – suggesting the latter [77, 79].

1.1.4 MicroRNA Structure

1.1.4.1 MicroRNA Terminal Loops

Although microRNA terminal loops were originally deemed to be unimportant by Han et al. in 2006 – recent literature provides mounting evidence to the contrary [90].

Our interest in microRNA terminal loops and their structures parallels ideas on research into protein secondary structures which define the three-dimensional shapes formed by segments of a protein based on their amino acid composition and the hydrogen bonds between them. Dictated by sequence, protein secondary structures include turns, loops, alpha helices and beta sheets, which can be analyzed by software programs to model secondary and tertiary structures and has proven useful for predicting the functions of these structures [91]. For example, DNA binding motifs are often composed of alpha helices and membrane-spanning G protein-coupled receptors.
consist of 7 alpha helices in their membrane-crossing regions [92]. This sequenced-based structural knowledge provides clues to the classification of proteins of unknown function.

Transfer RNA also harbours a distinct clover-leaf structure which is recognizable by sequence information. Complementary bases within stretches of tRNA sequences form bonds to create stems, leaving loop structures which define crucial components to tRNA functioning; the TψC loop, the D-loop and the anticodon loop which recognizes codon sequences within mRNA for translation. In a similar manner, we are interested in applying this logic to microRNA terminal loop structures to determine if there are characteristics of microRNA loops that are functionally relevant and which might give us an idea of possible specialization within this large class of tiny genes.

1.1.4.1. Evidence for the Importance of microRNA Terminal Loops

There has been limited assessment of the functions of microRNA precursors – pri- and pre-microRNAs, which retain the loop structure - due to the lack of assays able to isolate and examine them in mature-microRNA-free settings, since the lack of either Drosha or Dicer has huge impacts within the cell, while also affecting downstream microRNA species, including mature microRNAs [93].

Terminal loops were first discounted as important in 2006, when Han et al. examined the role of the terminal loop in Drosha processing and deemed it unessential[90]. It wasn’t until several years later that the topic of microRNA terminal loops was reopened for study.
One simple but suggestive piece of evidence for terminal loop importance lies in their differential processing in plant and animal cells. While both plants and animals lose their terminal loops by the time the mature microRNA is loaded into the RISC complex – higher order organisms retain their loops much longer and into the cytoplasm (Figure 1-5) [94], suggesting a possible regulatory role in more complex species.

Figure 1-5 Differential Processing Between Plant and Animal microRNAs. Reprinted from Cell, 136, Carthew, R.W., Sontheimer, E.J., Origins and Mechanisms of miRNAs and siRNAs, Pages 642-55, Copyright 2009, with permission from Elsevier. [94]
An extremely compelling piece of evidence for the importance of terminal loops was provided by Liu et al. in 2008 [95]. Comparing 2 near-identical mir-181 family members – differing by only one nucleotide in their mature sequence (Figure 1-6 A) – they found that mir-181a and mir-181c have distinct activities in early T-cell development. They were able to measure this activity, and then set out to mix and match different components of the microRNA stem-loop structures(Figure 1-6 B) to examine which part of the microRNA was responsible for the differential function (Figure 1-6). Interestingly, throughout the combinations of stem-loop components, they concluded that the function of the microRNA went along with the loop – meaning that the terminal loop was somehow playing a role in determining microRNA function.

Figure 1-6 Panel C shows the Relative Activity of different stem-loop combinations. The grey panel of the graph shows the relative activity of the baseline set of microRNAs Wild Type mir-181a-1 and mir-181c. The yellow panel shows mutants with differing stems – a direct swap of microRNA mir-181a-1 and mir-181c stem regions, but not the loop regions, does not change their activity levels. The orange and green panels showed reversed activity levels – and the constant reversed region of the microRNA in these cases is the loop sequence.
Further, the terminal loop plays a role in the recognition of the pri-microRNA by the Drosha-DGCR8 complex, and acts as a starting point for measuring the site of cleavage [96]. Loop mutations have been shown to affect the expression of mir-16, affecting Drosha processing by changing the structure of the terminal loop [97]. Zang and Zheng have shown that large flexible loops are more conducive to efficient Drosha and Dicer processing [97].
1.1.5 MicroRNA Regulation

The 3 different forms of microRNAs: pri-miRNAs, pre-miRNAs and mature microRNAs have varying levels of expression within cell types, suggesting levels of regulation after the initial transcription step which can be regulated by epigenetic or transcription factors [98-100]

1.1.5.1 Drosha and Dicer Binding Proteins

Although the Drosha-DGCR8 complex is sufficient to process pri-miRNAs into pre-miRNAs, Drosha has been shown to be part of larger complexes containing other proteins that influence microRNA processing in specific situations. DDX5 (protein p68) associates with SMAD when induced by TGF-beta, and binds to Drosha and DGCR8, to enhance the processing of pri-mir-21, and with p53 to enhance pri-mir-61-1 and pre-mir-143 processing[101-103]. When complexed with DDX17 (protein p72), p68 results in inhibition of pri-mir-16, pri-mir-125a, pri-mir-143, pri-mir-145 and pri-mir-195 processing upon ER-alpha stimulation by estradiol[104].

Dicer’s main interaction partners in microRNA processing are TRBP and one of the 4 Argonaut proteins – but preferentially AGO2. TRBP phosphorylation mediated by the MAPK signalling pathway has been shown to stabilize the Dicer complex resulting in the increased levels of mir-17, mir-20a and mir-92 – all growth promoting microRNAs[105]

1.1.5.2 Terminal Loop Binding Proteins

The processing of primary transcripts and microRNA precursors has been show to be regulated by terminal loop binding proteins. The first instance of this loop-binding was shown during
differentiation in mouse embryonic stem cells, and subsequently in embryocarcinoma cells where the processing of members of the let-7 family were inhibited by RNA-binding protein Lin-28 [106-112].

The RNA binding protein heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1) has been shown to enhance processing of hsa-mir-18a, through binding of similar sequences in both the terminal loop and the stem of the primary-transcript, resulting in relaxation of the stem-structure which makes the microprocessing step more efficient – although its effects on Dicer processing were not examined [113, 114]. Since hsa-mir-18a is a member of the mir-17-92 cluster – this is an interesting example of differential regulation of microRNAs that are transcribed together, since other members of this cluster were not affected by hnRNP A1. hnRNP A1 has also been shown to bind pri-let-7a-1 and pri-mir-101, suggesting a more general role in microRNA terminal loop binding and regulation [115-117].

KH-type splicing regulatory protein (KSRP) is another RNA-binding protein shown to be involved in microRNA regulation through binding the terminal loop sequence. Known for its propensity to bind AU-rich elements (AREs) to direct mRNA decay, it has been shown to participate in both the Drosha and Dicer complexes[118, 119]. KSRP has been shown to enhance the processing of a several microRNAs including hsa-mir-21, hsa-mir-16 and let-7a, by binding their terminal loops and promoting processing by Drosha or Dicer through protein interactions with the processing complexes[119].
1.1.6 Secreted or Circulating microRNAs

Valadi et al., in 2007 revealed the presence of microRNAs in exosomes from cell lines [120]. The first evidence of the presence of detectable levels of microRNAs exterior to cells and circulating in the serum came from a British group in 2008, who showed that circulating levels of tumour-associated microRNAs (hsa-mir-155, hsa-mir-210, hsa-mir-21) were higher in serum from diffuse large B-cell lymphoma patients than in control serum[121]. Hunter et al., then showed later that year that microRNAs are present in exosomes in the blood of healthy patients and that there wasn’t a large degree of variation in microRNA expression based on age or gender [122].

Exosomes are 30-100 nm membraned vesicles that are secreted by various cell types to transmit information and signals to other cells – often over large distances. Exosomes were first suggested to be merely garbage bins extracting and disposing cellular waste [123]. But they have since been found to travel in the extracellular environment and deliver their contents to receptor cells, often distantly located, through the fusion of the vesicular body to the plasma membrane of the receptor cell. Many types of cells have been shown to release exosomes including T-cells, B-cells, epithelial cells and tumour cells [124]. All exosomes are similar in nature in that they share some proteins and lipids in their structures, although exosomes do contain some proteins and lipids which are distinct to their cell of origin [124].

A comprehensive study by Chen et al. presented many important characteristics of circulating microRNAs [125]. This group showed that microRNAs are stable under many harsh conditions including freezing, boiling and varying pH levels. MicroRNAs are present in the serum of many
species of mammals including humans, mice, rats and horses, and they showed that members of the same species have similar microRNA expression profiles[125].

1.1.6.1 Structure
There are conflicting reports on the microRNA species that is present in microvesicles in serum. Zhang et al. has suggested that plant microRNAs found to be circulating in human serum, uptaken as food, are likely to be taken up in their mature form [126]. Keller et al. studied serum-microRNAs across a spectrum of diseases using probes designed for the mature microRNAs [127]; however, this does not rule out the presence of precursor microRNA species. Working in cell lines, Chen et al. observed that both let-7b and let-7g were present in conditioned media and mesenchymal stem cells in both mature and pre-microRNA forms. However they commented that they were likely secreted predominantly as pre-microRNA[128].

1.1.6.2 Possible Mechanisms of Secretion

It remains unclear how, why and in what form microRNAs are secreted from their cells of origin. We enumerate several possible mechanisms that lead to their identification in the blood, in the presence of a primary tumour. The following concepts are derived from personal communications and brain storming sessions [129].
1.1.6.2.1 Shedding
Shedding of microRNAs from primary cancer cells through vesicular mediated transport pathways, resulting in rapid change of gene expression in the primary cell as a means to gain metastatic potential. This would present as an inverse relationship between microRNAs downregulated in tumours -- particularly those of late-stage primaries -- and microRNAs upregulated in the blood.

1.1.6.2.2 Secretion from Primary Tumours
Secretion of microRNAs from the primary tumours into the blood stream to home to tissues of possible metastases, initiating new programs of gene expression within cells uptaking the microRNA to begin the transformation process. This might be seen as an overexpression in the blood, and either normal expression levels or an accompanying overexpression within the cell.

1.1.6.2.3 Dead and Dying Cells
Accruement of microRNAs in the serum resulting from dead and dying cancer cells as they are attacked by the immune system, which is unrelated to any ongoing cancer process in the blood, but simply reflects the overall tumour microRNA expression. This would result in a similar pattern of expression between microRNAs in the tumour cells and in the blood.

1.1.6.2.4 Active Transport
Active transport of microRNAs from cells via energy-dependent transmembrane protein export. This process is dependent on a transporter protein and is a more efficient way of exporting
microRNAs from cells since vesicular formation is not necessary. This process would result in exosome- or vesicle-free microRNAs in the blood.

1.1.6.2.5 Travelling Tumour Cells

There is a possibility that the observation of microRNAs in the blood is the result of the identification of microRNAs lysed from within viable tumour cells in the blood. These cells are potentially travelling along a migration path from the initial tumour cells to the metastasis, to resettle in a secondary site.

1.1.6.2.6 Alternate Cells of Origin

The final case addresses the possibility that microRNAs in the blood don’t originate from the cancer cells at all. They may be produced by another cell type all together, and may enter the blood to perform a task contrary to the progression of the cancer itself --ie. microRNAs deriving from immune-related cells.

1.1.7 MicroRNA Associations with Disease

1.1.7.1 Non-circulating MicroRNAs

Evidence is mounting to support the idea that each cancer has a specific microRNA profile – and such a profile links the cancer with its primary cell of origin -- even for poorly differentiated cancers -- surpassing the ability of gene microarray to do so [130]. MicroRNA expression patterns can also be used to distinguish cancerous cells from their adjacent normals, and some
microRNAs – hsa-mir-21 in the adenocarcinoma setting - were shown to carry prognostic value showing an association with poor survival and therapeutic outcome[131].

The first data to show that microRNAs could be implicated in cancer came in 2002 with the identification of deleted and downregulated microRNAs mir-15 and mir-16 in chronic lymphocytic leukemia [132]. Since then, microRNAs have been identified as both tumour suppressors and oncogenes across a large panel of cancer tissues. This has led to a re-evaluation of suspicious regions of the genome where the lack of genes has previously eluded study, with an added focus on identifying microRNAs. MicroRNA microarray studies were soon performed demonstrating differential expression of many microRNAs in various cancer types – showing that microRNAs could serve as signatures of particular cancers [130, 133] [132, 134-137]. Since microRNA expression is tissue-specific and initial array coverage was poor, microRNA profiles of solid tumour types have also lacked a panel of common deregulated microRNAs – although initial reports stated that trend was for microRNA down regulation in solid primary tumours[130], later studies reported the converse to be true for several tumour types [133].

1.1.7.2 Circulating microRNAs
The associations of microRNAs in serum with their primary tumours are strengthened by studies showing that the removal of primary tumours greatly reduces the serum levels of certain previously elevated microRNAs in patients [124]. This was shown in colon cancer with mir-17-3p and mir-92 [138]. In ovarian cancer, Taylor et al. showed that the levels of circulating tumour-derived exosomes – exosomes containing microRNAs shown to be overexpressed in the
primary tumour -- were much higher in ovarian cancer patients than in both normal controls and in women with benign ovarian cysts. They also showed that the levels of circulating exosomes increased with the invasiveness and stage of the cancer [139]. In breast cancer, only smaller subsets of microRNAs have been studied and to date, serum microRNAs have been less able to distinguish healthy from benign tumours than in other solid tumours. However, mir-155 has been shown to correlate with progesterone positive breast cancers [140]. Mir-195 levels have shown a correlation between primary tumours and serum, and a decrease in mir-195 and let-7a levels was seen in post-operative breast cancer patients, suggesting the tumour as the source of the serum overexpression and the potential use of mir-195 as a measure of post-operative success. This same study also showed correlations between decreased let-7a serum levels and lymph node involvement, and a correlation of circulating mir-21 and mir-10b levels with Estrogen Negative disease [141]. Serum microRNAs in pancreatic cancer – one of the most difficult cancers to detect, and also the most lethal – have been studied and mir-210 has been identified by 2 studies as being increased in the serum of pancreatic cancer patients, detectable at levels four-fold higher than normal controls [142, 143].

1.1.8 MicroRNA Targets

1.1.8.1 MicroRNA Prediction Databases

It is difficult to validate predicted microRNA targets. Since microRNAs may not alter the transcript level of the genes that they target, high-throughput microarrays and real time PCR experiments serve as only a snap-shot of the potential situation, awaiting the development of
reliable high-throughput protein assays. This has lead to particular interest in the accurate prediction of microRNA targets computationally by a sequence based approach. More than 10 public databases for microRNA:mRNA target prediction have been created, all using different algorithms and approaches. Considering varying degrees of sequence similarity, conservation, site accessibility and different targeted regions of the mRNA – all databases add a novel level of complexity to the microRNA question [20, 23-40].

1.1.8.2 MicroRNA Prediction Integration
At the time of publication, there were several online sources that integrated microRNA target prediction data, either across sources or with other datasets. Three are discussed below.

1.1.8.2.1 miRGator v1.0
MirGator v1.0 [144] integrates three microRNA prediction algorithms: miRanda [34], PicTar [31] and TargetScanS[24]. It enables the comparison of microRNA target predictions from these three databases to expression data from public datasets to compare the correlations of gene and microRNA expression. It also enables searches by subsets of genes by gene ontology groups, pathways and disease terms. It does not provide the user with any flexibility in terms of how they would like to combine microRNA prediction databases, nor does it allow combinatorial searches.
1.1.8.2.2 miRecords

miRecords [145] integrates 11 microRNA target prediction databases. It is searchable by gene ID or a subset of microRNAs. It returns microRNA target predictions across all the databases, allowing users to see how frequently an interaction is predicted – but only in a low-throughput way with visual icons indicating commonalities. It lacks the flexibility of combining prediction algorithms as well as lacking the ability to return targets of multiple microRNAs.

1.1.8.2.3 miRNAMap

miRNAMap [146] houses both pertinent microRNA sequences and expression information, but also microRNA target predictions, albeit only those from TargetScanS. It is very useful for assembling information about a given microRNA, but provides minimal support for microRNA target predictions since it only uses one source.

1.1.9 Classification

Much work has been done to segment the protein coding genome into identifiable groups of genes, as a first step to make this large amount of data more understandable. Genetic classification by locus, by protein domain (which is sequence driven) and by motif (for example upstream transcription factor start sites or protein binding sites) are obvious early candidates. Deeper understanding of sets of genes has been gained using The Gene Ontology and Reactome’s work to curate robust and accurate classification sets, including genes related by
pathway membership, disease-related genes and the classifications of genes according to
function or as components of a complex\cite{147-149}. We borrow ideas from this approach and
propose that it can be similarly applied to the set of microRNAs – that they are distinct enough in
structure, tissue expression patterns, loop motifs and function that they can be further subdivided
to gain greater insight into their roles in the cell.

1.2 Biological Interaction Networks

1.2.1 Protein Interaction Networks

Modeling the set of protein interactions within the cell is one of the great challenges of both
cellular and systems biology. Since proteins have many binding partners which are expressed
across different cell types it is not a one-dimensional problem and there are as many different
sets of protein-protein interactions as there are cell types. Protein interaction networks borrow
concepts from computer science, including graph and network theory to provide a way to display
the set of interactions in a particular cellular setting. Nodes in the network diagrams are
representative of proteins and edges join any 2 proteins that have been shown to interact. One
protein can have many binding partners, since protein interaction networks are generally not
temporally sensitive and hence they display all of the partners of one protein across many
different time points and tissues and environmental conditions.
Protein interaction networks are useful in several ways. They allow us to manage and display the results of high throughput experiments, which was previously difficult to visualize. They allow us to make predictions as to how the perturbation of one protein in the system might affect others. They allow us to see the effects of drugs on proteins downstream of their initial targets. They are often able to identify neighbourhoods or clusters of genes that interact with each other. They allow us to see beyond a protein of interest to examine its interacting partners, perhaps enabling us to see a commonality between them.

1.2.2 MicroRNA Interaction Networks

To visualize and analyze complex relationships between different predictions of microRNA:mRNA target mappings, we borrowed ideas from protein-protein interactions and gene regulatory networks. We used our lab’s software -- NAViGaTOR (Network Analysis, Visualization and Graphing Toronto) [150] to analyze and visualize the resulting network of microRNA:mRNA target mappings – the microRNA interaction network (*micronome*) – a novel application of biological interaction network theory. NAViGaTOR was particularly appropriate for use in this setting due the flexible nature of the loading of the interaction set – which enabled us to use microRNA-mRNA interactions as the backbone of our networks, rather than protein-protein interactions.
1.2.3 Gene Regulatory Networks

In addition to protein-protein interaction networks and microRNA interactions networks, cellular functionality is also frequently analyzed though other molecular networks – such as genetic networks representing regulatory or functional relationships among genes.

1.3 Summary of Research Contributions

1.3.1 Chapter 2 – Navigating the micronome

There are many microRNA prediction databases all using different approaches to perform the same task – making accurate predictions of targets of microRNAs. We review microRNA database characteristics, evaluate their prediction quality by comparing known microRNA interactions to their predicted targets, and integrate their predictions into mirDIP - our data integration portal - to allow users to manipulate and combine the results of different targeting algorithms to maximize their usefulness and quality scores for their particular applications. Using mirDIP we identified of 2 classes of microRNAs based on their targeting behavior, and introduced the concepts of microRNA interaction networks, showing that genes within a pathway are more tightly connected by microRNA interactions than one would expect by chance, suggesting some degree of signalling pathway regulation at the transcriptional level.
1.3.2 Chapter 3 - A Global Analysis of microRNA Loops Identifying Secretion-related Motifs and Structures

Our earlier work led us to the observation that microRNA prediction databases are extremely seed centric. We undertook an investigation of microRNA loops – an often ignored element of the microRNA structure to see if we might be able to unearth a functional role for microRNA terminal loops. Not only did we find that microRNA terminal loops differed drastically in size and structure, we found that they carried many common RNA motifs. Further, these common motifs had complements and reverse complements appearing frequently in other terminal loops, suggesting the possibility of microRNA:microRNA precursor binding. Using publicly available data, we found evidence that microRNA loop motifs may play roles in the secretion process of microRNAs from cells, possibly with their terminal loops acting as recognition sequences. These findings enabled further classifications of microRNAs in addition to the intrapathway and universe classes introduced in Chapter 2.
Chapter 2

NAViGating the microNome

Modified From:


2.1 Abstract

MicroRNAs are a class of small RNAs thought to regulate gene expression at the transcript level, the protein level, or both. Since microRNA binding is sequence-based but possibly structure-specific, work in this area has resulted in multiple databases storing predicted microRNA:target relationships computed using diverse algorithms. We integrate prediction databases, compare predictions to in vitro data, and use cross-database predictions to model the microRNA:transcript interactome – referred to as the micronome – to study microRNA involvement in well-known
signalling pathways as well as associations with disease. We make this data freely available with a flexible user interface as our microRNA Data Integration Portal — mirDIP (http://ophid.utoronto.ca/mirDIP).

mirDIP integrates prediction databases to elucidate accurate microRNA:target relationships. Using NAViGaTOR to produce interaction networks implicating microRNAs in literature-based, KEGG-based and Reactome-based pathways, we find these signalling pathway networks have significantly more microRNA involvement compared to chance (p < 0.05), suggesting microRNAs co-target many genes in a given pathway. Further examination of the micronome shows two distinct classes of microRNAs; universe microRNAs, which are involved in many signalling pathways; and intra-pathway microRNAs, which target multiple genes within one signalling pathway. We find universe microRNAs to have more targets (p < 0.0001), to be more studied (p < 0.0002), and to have higher degree in the KEGG cancer pathway (p < 0.0001), compared to intra-pathway microRNAs.

Our pathway-based analysis of mirDIP data suggests microRNAs are involved in intra-pathway signalling. We identify two distinct classes of microRNAs, suggesting a hierarchical organization of microRNAs co-targeting genes both within and between pathways, and implying differential involvement of universe and intra-pathway microRNAs at the disease level.
2.2 Background

MicroRNAs are short, but important non-coding RNA sequences that regulate gene expression [1]. They are thought to target the 3’ Untranslated Regions (UTRs) of mRNA, disrupting their ability to be translated into proteins, sometimes repressing the expression of the mRNA itself [2-9]. MicroRNA prediction algorithms generally pair the seed region of the microRNA (bases 2-7 from the 5’ end of the microRNA) to a cognate mRNA sequence. However, this binding is complicated by many factors, not the least of which is that imperfect microRNA:mRNA binding occurs, and thus single base-pair mismatches and G:U wobble base-pairs must be considered.

Discovery of the first microRNA – lin-4 in worm (C. Elegans) [10], its further characterization in 1989 [11], annotation as a non-coding RNA in 1993 with a sequence complementary to the lin-14 3’ UTR [1, 12], and functional characterization as having a translational repression effect later that year [13] opened a rich research field. Many subsequent in vitro experiments and computational predictions aimed at uncovering microRNA:target relationships to fathom microRNA effects on gene expression regulation. With the discovery of a second nematode microRNA – let-7, which targets lin-41 and hbl-1, the concept of microRNAs made the jump from worms to higher species, since let-7 had well-known homologues even in humans [14-16]. Coining the term “microRNA” for this class of non-coding gene regulators in three back-to-back Science papers in 2001 [17-19], the discovery of microRNAs had crossed over to the human domain, and finding microRNA targets became a high priority. After the first bioinformatics attempt at predicting plant microRNAs [5], many microRNA prediction algorithms, for both fly (D. melanogaster) and human (H. sapiens), were developed [20-22].
More than 10 public databases for microRNA:mRNA target prediction have been created, all using different algorithms and approaches. Considering varying degrees of sequence similarity, conservation, site accessibility and different targeted regions of the mRNA – all databases add a novel level of complexity to the microRNA question [20, 23-40].

To visualize and analyze these complex relationships between different predictions of microRNA:mRNA target mappings, we borrow ideas from protein-protein interactions and gene regulatory networks. We first integrate all databases into a freely available data portal – mirDIP (microRNA Data Integration Portal) – and use NAViGaTOR (Network Analysis, Visualization and Graphing Toronto) [150] to analyze and visualize the resulting network of microRNA:mRNA target mappings – the microRNA interaction network (micronome).

2.3 Results and Discussion

2.3.1 Characteristics of microRNA predicting databases

There are many characteristics of microRNA:mRNA target binding that are taken into account - in different combinations - for each microRNA prediction database. We begin with a review of these criteria. Table 2-1 shows all databases considered in this research.
Table 2-1 MicroRNA Prediction Databases used in our study.

* Not used in all comparisons, nor the construction of microRNA interaction networks since it is a superset of the top database predictions.

$ Not used in all comparisons, nor the construction of microRNA interaction networks since it was not available for bulk download at the time of data curation

To enable more informed integration of these predictions, we consider characteristics of individual microRNA prediction algorithms in detail, and summarize them in Table 2-2.
<table>
<thead>
<tr>
<th></th>
<th>Targetscan Conserved</th>
<th>Targetscan Non-Conserved</th>
<th>RNA22 3' UTR</th>
<th>RNA22 5' UTR</th>
<th>RNA22 CDS</th>
<th>microRNA.org</th>
<th>microCosm (formerly mirBase)</th>
<th>PITA Top Hits</th>
<th>PITA All Hits</th>
<th>picTar 4-way</th>
<th>picTar 5-way</th>
<th>DIANA microT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation</td>
<td>X*</td>
<td></td>
<td>X</td>
<td>X</td>
<td>§</td>
<td>§</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Site Accessibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Local AU content</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Multiple Binding Sites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>(1 microRNA)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Multiple Binding Sites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X§</td>
<td></td>
<td>X§</td>
<td>X§</td>
<td>X§</td>
<td>X§</td>
</tr>
<tr>
<td>(&gt;1 microRNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Uses miRanda</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Free Energy of Duplex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Examines surrounding Sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Weighted 5' end or</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>considers seed type</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2-2 Characteristics of MicroRNA Prediction Databases.**

*Targetscan Conserved uses conservation, but it is not integrated into the context score.

§PITA does not explicitly use conservation in scoring targets. However, accessible microRNA binding sites tend to show high conservation.

¥picTar does have predictions for multiple microRNAs binding to a single 3’ UTR; however, that data was not used in this study.
All main groups of features used for prediction are described below:

(1) Seed Sequence match. All prediction algorithms depend on this criterion. Allowing for base-pair mismatches and G:U wobbles, which have been shown to be important in microRNA binding [151], prediction algorithms look for high degree of complementarity between the 5’ end of the microRNA and the 3’ end of the mRNA target sequence. Particular attention is paid to the seed region (bases 2-7 from the microRNA 5’ end).

(2) Conservation. Many prediction algorithms take into account the conservation of the microRNA binding sequence in the mRNA target across species. Generally used as a filtering step, a highly conserved target site is thought to produce a more reliable prediction. Conservation is not directly used in some databases (Probability of Interaction by Target Accessibility (PITA)) [30, 34], is not directly incorporated into the score in others (Targetscan) [24, 27, 33], and is not used at all in others (RNA22) [34]. Interestingly, PITA results suggest that considering site accessibility is analogous to considering conservation, since accessible 3’ UTR microRNA binding sites tend to fall in conserved regions [30]. To reduce bias in our analyses, we use both predictions with and without conservation.

(3) Free Energy of microRNA:mRNA duplex. The Free Energy of the microRNA:mRNA duplex (ΔG), is often calculated with the Vienna Folding package [152-154] or RNA hybrid [155]. It evaluates the energy required for the formation of the microRNA:mRNA duplex from a completely dissociated state – a more negative value indicates a larger inclination for the two RNAs to bind.
(4) **Site accessibility.** Site accessibility is not considered in many prediction algorithms. Measured as $\Delta\Delta G$ for use in PITA, it compares the energy requirement for the already folded 3’ UTR to unfold to allow the microRNA accessibility to the target site, and to refold into the microRNA:mRNA duplex [30]. A more negative $\Delta\Delta G$ indicates a favourable folding energy for the microRNA:mRNA configuration.

(5) **Contribution of multiple binding sites.** Many algorithms reward microRNAs that have multiple binding sites within the 3’ UTR of a particular gene, reasoning that the microRNAs will be able to exert a dose-dependent effect on target expression. Binding sites can be for a single microRNA or for multiple different microRNAs that show co-operativity resulting in synergistic gene repression [156]. Several studies have shown that the ideal inter-binding site distance falls between 8-40 base-pairs [27, 157].

(6) **Local AU content.** Considered in Targetscan's context score, Grimson et al. have shown that an enrichment of A or U base-pairs in the 30 nucleotides up or downstream of the microRNA binding site in the 3' UTR tends to favorably associate with repression in target expression [27, 33].

(7) **Local mRNA sequence.** The consideration of sequence surrounding the microRNA binding site on the 3’ UTR is sometimes taken into account. Algorithms may examine local sequence effect on site accessibility, or examine sequence content for particular nucleotides [27, 30, 33].

(8) **Ribosomal shadow.** Considered in Targetscan, the 15 nucleotides after the stop codon in a 3’ UTR form poor microRNA target binding sites that show little ability to repress expression. It has been postulated that this is due to a ribosomal shadow effect [27].
(9) *Uses miRanda.* miRanda [20, 29] is the first microRNA alignment algorithm, is similar to the Smith-Waterman algorithm for sequence alignment and uses rules of thumb previously established in sequence alignment [159-161]. It forms the basis of several microRNA prediction algorithms. miRanda considers several features described below:

- **Sequence match** – a reward of +5 for a G=C or A=U match, +2 for G:U wobble. A penalty of -2 for a Gap Extension and -8 for a Gap Opening. The cutoff for S, the result of these sequence matches is generally $S > 80$ (flies), $S > 50$ (humans).
- **Scaling** – Matches in positions 1-11 of the microRNA (from the 5’ end) are given twice the weight of matches elsewhere to reflect the asymmetry of microRNA binding [29].
- **Four empirical rules:**
  - No mismatch in bases 2-4;
  - <5 mismatches in bases 3-12;
  - At least 1 mismatch in bases 9 to (Length-5);
  - < 2 mismatches in the final 5 base-pairs.
- **Vienna Package Folding** assumes the microRNA is linked to the 3’ UTR by 8 –x– base-pairs that cannot bind anything. This single structure is then folded. The $\Delta G$ cutoff is usually set as $\Delta G < -14$ kcal/mol for flies and $\Delta G < -17$ kcal/mol for humans.

The final score is the total energy and total score of all hits between those of a microRNA and a 3’ UTR.
• Conservation – a filtering step requiring 90% conservation or more between human and rat or mouse and 80% conservation between *D. melanogaster* and *D. pseudoobscura* or *A. gambiae*.

(10) **Position effects.** Positional effects reward microRNA target sites that fall within the first quartile of the 3’ UTR after the stop codon (+15 base-pairs) or within the final quartile of the 3’ UTR, near the poly(AAAA) tail. This effect is more pronounced in long UTRs [27].

(11) **3’ Pairing.** Aside from strong seed region pairing, many algorithms that aren’t based on miRanda also require nucleotide binding between the microRNA and the target mRNA between bases 12-17 of the 3’ end of the microRNA [27].

### 2.3.2 MicroRNA prediction database similarities

Since microRNA:mRNA target prediction algorithms use different combinations of features to perform the same task, it is useful to analyze the distribution of these predictions across databases. There is an expected trend – with far fewer predictions being made that transcend six or more databases than those that are present in just one database. We count over 2 million predictions present in only one database, falling off to a surprisingly small 18 predictions identified in 8 of the 9 databases considered (As indicated in Table 2-1, we do not consider PITA All Targets nor picTar 4-way in this part of our analysis to avoid double-counting. Nor do we consider microT, since bulk download was not available at the time of data curation) (Figure 2-1A). Figure 2-1B compares all database predictions to microRNA.org, the largest catalog of conserved interactions (excluding PITA ALL Targets – which is a superset of PITA TOP
Targets, and microT which was not available in its entirely at the time of publication) – indicating that although we see low total overlap among all databases, in reference to the largest conservation-considering database there is considerable similarity between at least five database prediction schemes. Although DIANA microT v3.0 [39, 40] was not included in our extended database analysis and comparison, since it was not available for bulk download when our study began, we have included it in this figure for the sake of comparison.
Figure 2-1 MicroRNA prediction database characteristics.
Panel C: Overlap of microRNA prediction databases Targetscan and picTar, since both consider degree of conservation as part of their scoring scheme.
Panel D: Percentage of overlapping microRNA predictions across two or more databases.

Although most databases consider conservation, they each handle it differently. Bartel’s Targetscan publishes dual lists of targets based on either conserved or non-conserved sites. Thirty-one percent of these microRNA:mRNA target predictions are shared by both lists (Figure 2-1C, left panel), demonstrating that there is a strong tendency for genes to contain both conserved and non-conserved microRNA binding sites along the length of their 3’ UTR. On the other hand, picTar considers grades of conservation in their prediction algorithm. Publishing both a 4-way and 5-way conservation scheme (human, mouse, rat, dog vs. human, mouse, rat, dog, chicken) picTar suggests degree of conservation correlates with robustness of prediction. In this case we can see that one list is clearly a subset of the other, and moving from a less conserved setting to a more conserved setting reduced the number of predicted targets to 30% (Figure 2-1C, right panel). When combining datasets, Figure 2-1D shows the percentage of predictions preserved per prediction scheme when requiring a microRNA:mRNA target prediction to occur in at least three databases. Targetscan and PITA Top Hits have the most remaining interactions after applying this filter.
2.3.3 Comparing microRNA prediction databases to the truth

2.3.3.1 MicroRNA target filtering is vital

To examine whether a combination of microRNA prediction databases would outperform any one source, data from 15 publicly available microRNA over-expression/knockdown experiments followed by microarray [85, 162-169] was assembled (Table 2-3).

<table>
<thead>
<tr>
<th>Study</th>
<th>microRNA</th>
<th>Platform</th>
<th>Cell Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lim et al., 2005</td>
<td>hsa-mir-1</td>
<td>Rosetta 25k v2.2.1</td>
<td>HeLa</td>
</tr>
<tr>
<td>Baek et al., 2008</td>
<td>hsa-mir-1</td>
<td>Agilent Whole Genome 4x44k</td>
<td>HeLa</td>
</tr>
<tr>
<td>Linsley et al., 2007</td>
<td>hsa-mir-106b</td>
<td>Rosetta/Merck 44k 1.1</td>
<td>HeLa</td>
</tr>
<tr>
<td>Lim et al., 2005</td>
<td>hsa-mir-124</td>
<td>Rosetta 25k v2.2.1</td>
<td>HeLa</td>
</tr>
<tr>
<td>Wang et al., 2006</td>
<td>hsa-mir-124</td>
<td>Affymetrix U133 plus2</td>
<td>HepG2</td>
</tr>
<tr>
<td>Baek et al., 2008</td>
<td>hsa-mir-124</td>
<td>Agilent Whole Genome 4x44k</td>
<td>HeLa</td>
</tr>
<tr>
<td>Ceppi et al., 2009</td>
<td>hsa-mir-155</td>
<td>Affymetrix U133 plus2</td>
<td>MDDS</td>
</tr>
<tr>
<td>Linsley et al., 2007</td>
<td>hsa-mir-16</td>
<td>Rosetta/Merck 44k 1.1</td>
<td>HeLa</td>
</tr>
<tr>
<td>Baek et al., 2008</td>
<td>hsa-mir-181</td>
<td>Agilent Whole Genome 4x44k</td>
<td>HeLa</td>
</tr>
<tr>
<td>Gennarino et al., 2009</td>
<td>hsa-mir-26b</td>
<td>Affymetrix U133 2</td>
<td>HeLa</td>
</tr>
<tr>
<td>Tavazoie et al., 2008</td>
<td>hsa-mir-335</td>
<td>Affymetrix U133 plus2</td>
<td>LM2</td>
</tr>
<tr>
<td>Huang et al., 2008</td>
<td>hsa-mir-373</td>
<td>Wistar Illumina V6</td>
<td>MCF-7</td>
</tr>
<tr>
<td>Huang et al., 2008</td>
<td>hsa-mir-520c</td>
<td>Wistar Illumina V6</td>
<td>MCF-7</td>
</tr>
<tr>
<td>Webster et al., 2009</td>
<td>hsa-mir-7</td>
<td>Affymetrix U133 plus2</td>
<td>A549</td>
</tr>
<tr>
<td>Gennarino et al., 2009</td>
<td>hsa-mir-98</td>
<td>Affymetrix U133 2</td>
<td>HeLa</td>
</tr>
</tbody>
</table>

Table 2-3 Characteristics of High-throughput Experiments used in our study.
As discussed in the Methods section, when comparing microRNA target predictions to actual microRNA targets (as determined by microarray experiments) two filtering steps were performed to increase the suitability of the target predictions for the data – filtering by both microarray and by cell type. Filtering by microarray (Table 2-3 column 3) eliminates targets not present on the particular chip in the experiment, and thus having no chance of appearing in the final target set. Filtering by cell type (Table 2-3 column 4) eliminates genes expressed at only low levels in the cell line (which would reduce their chances of showing a knock-down effect). This two-step filtering drastically changes the predictions. As illustrated in Figure 2-2A, beginning with an identical set of mir-1 predicted targets across all databases and filtering by cell type and chip type to make the target predictions suitable for comparison to 2 different experiments results in significantly different final prediction sets – with overlapping targets numbering only 60% of the sets – clearly demonstrating the need to tailor predictions to the setting in which the experiment was done before any comparisons are undertaken. This filtering exercise shows how critical it is to consider tissue specificity when examining microRNAs of interest. Clearly, with the availability of more in vitro and in vivo data, it will become crucial to ensure that data is organized in a tissue-specific manner to enable more accurate modelling of the interactions present in particular settings. This might also include environmental conditions: normoxic vs hypoxic conditions, presence or absence of growth factors and other stressors.
Figure 2-2 MicroRNA target prediction processing and evaluation.
Panel A: MicroRNA targets are highly dependent on an experimental setup. Filtering by cell type and microarray platform on an identical initial prediction set can cause a divergence of up to 40% in the final target lists.

Panel B: MicroRNA over-expression in different experimental settings results in poor overlap. Venn diagram of discrepancy between in vitro microRNA over-expression experiments of mir-124.

Panel C: Comparison of precision and recall across microRNA prediction databases, measured by computing the average values for all microRNA predictions by a particular database compared to their matched low stringency “ground truths”.

Panel D: The percentage of remaining predictions by considering overlap of interactions in all databases occurring in 2, 3, 4, and 5 prediction databases.

Panel E: Precision measurements for microRNA:target predictions by number of prediction databases indicating the percentage of predicted targets that were shown to be true across in vitro experiments. Stringency levels refer to confidence in the microarray data and were determined by either p-value or percentage knockdown as discussed in the methods.

Panel F: Recall measurements for microRNA:target predictions by number of prediction databases indicating the percentage of in vitro targets covered by predictions. Stringency levels refer to confidence in the microarray data and were determined by either p-value or percentage knockdown as discussed in the methods.

2.3.3.2 High-throughput target validation experiments are not always in agreement

Ideally, high-throughput experiments would provide clear and concise answers in simple over-expression experiments. Unfortunately, we have not found that to be entirely the case.

Examining the filtered results for the 2 microRNAs with high-throughput experimental results by multiple groups – there is remarkably little overlap between reported targets. Using mir-124 over-expression as an example, comparing the Baek et al. [162], Lim et al. [85] and Wang et al. [168] data sets – at the least stringent confidence level for targets, allowing the overlap between experiments to be maximal – we see only 10 common targets between all 3 lists, 3.7% of the smallest target list (Figure 2-2B). Expanding the overlap to include “true” targets predicted on 2
in vitro lists improves the situation, yet covers less than 50% of the smallest dataset. Similar results are seen in duplicate mir-1 experiments putting the overlap at 8%. One possible explanation for such observations is over-dosing with transfections, resulting in deregulation of gene expression due to a massive influx of microRNA molecules[170].

2.3.3.3 Comparing predictions to ground truth
PITA Top Targets, picTar 5-way Conservation and TargetScan Conserved Targets are all suitable candidates for top microRNA prediction database. Not only do they retain many predictions passing through a filter requiring predictions to be present in 3 or more databases (79% (165,060 predictions), 64% (11,024 predictions), 57% (107,771 predictions) respectively) (Figure 2-1D), they also perform well when evaluating database performance on both precision and recall when compared to publicly available high-throughput microarray data (Figure 2-2C). While all three databases have many retained cross-database predicted targets, PITA and Targetscan Conserved do tend to outperform picTar 5-way when both precision and recall are considered – that is, when we require database prediction sets to not only contain many true positives, but also to predict many of the actual true targets. As discussed in the methods section, all in vitro experiments were analyzed to produce high- med- and low- stringency sets of targets based on the degree of gene expression knockdown. Examining the least stringent in vitro “ground truth” data: PITA Top Targets, picTar 5-way Conservation and Targetscan Conserved have precision and recall values of: 30%, 9%; 38%, 2%; 32%, 12% respectively. This
demonstrates that although many of picTar 5-way’s predictions are true, it performs exceptionally poorly when measuring the proportion of real targets that picTar actually predicts.

In the balance between precision and recall one might suggest using these databases as follows: 1) when looking for confirmatory evidence of a particular interaction between a microRNA and a specific target – it is better to use a database with superior recall such as Targetscan Conserved, Targetscan Non-Conserved or microCosm (formerly mirBase), which are more likely to include a target prediction if one exists; 2) when identifying any possible target for a particular microRNA to form the basis for in vitro or in vivo experiments, it would be best to consult picTar 5-way; 3) when finding in silico evidence for an interaction of a microRNA and a gene of a certain family or function, it is best to use a database with a more even balance between precision and recall such as PITA Top Targets.

2.3.3.4 Comparing predictions to Tarbase
Tarbase [171], curated by the DIANA Lab, provides a running list of microRNA interactions that have been shown to be true or false by either microarray experiments, pSILAC experiments or some other manner of specific probing for a particular microRNA:target interaction. Although Tarbase does not represent a non-biased list of microRNA targets, it is interesting to compare our list of 2+DB microRNA interactions (those present in 2 or more prediction databases) to those present in their database. Thirty-nine percent of Tarbase-reported True mRNA repression targets, 48% of Tarbase-reported True mRNA cleavage targets, 67% of Tarbase-reported targets of
unknown effect, 32% of Tarbase-reported pSILAC tested interactions and 62% of Tarbase-reported microarray tested interactions were present in our 2+DB set of interactions.

Since microRNAs act through translational inhibition more frequently than they do through mRNA degradation, it is obvious that examining microarray data is not the perfect setting in which to evaluate microRNA targets. The subset of targets that have been transcribed but not translated will still be expressed in the data and as such they will be missed. However, it has been shown that proteins repressed by more than 30% also tend to destabilize at the transcript level [162] – meaning that examination of expression levels is a reasonable surrogate for large translational repressions. Another possible source for incorrect predictions includes off-target effects. MicroRNA overexpression is thought to produce some false positives, perhaps due to dosage issues [170]. However, these off-target effects will occur less frequently than in synthetic siRNA overexpression systems.

High-throughput proteomics approaches such as pSILAC experiments are exciting new techniques that are emerging at the forefront of microRNA target research, and which allow the quantitative comparison of the proteomes of two different samples. Although an improvement on expression analyses for microRNA target research, examinations at the protein level will still suffer from the inability to distinguish primary from secondary effects. Furthermore, they are neither as high-throughput as expression analyses nor as time-efficient to run, and the set-up costs to run mass spectrometry experiments are far higher than microarrays at the present time. Optimal microRNA target analysis would require experiments where it can be shown that actual microRNA:mRNA binding is occurring with an associated reduction in mRNA or protein
expression. Only then could we be certain an interaction is occurring – and such high-throughput experiment series remain a future challenge.

2.3.3.5 Integrating prediction databases in mirDIP

Due to the massive amount of genomic information being deciphered on a daily basis, there is an inevitable bottleneck between computational prediction and identification of binding sites, and the in vitro or in vivo validation of such interactions. Clearly, it would be useful to be able to prioritize microRNA:mRNA target predictions to reduce excessive false leads and unnecessary experiments. It has been previously shown, and confirmed here that none of the microRNA prediction databases does a perfect job of target identification [162, 172], although they are all suitable to provide an initial prediction. Integrating multiple databases can provide a means to balance out the precision and recall values. Comparing microRNA predictions made by a minimum of either two or three databases to all truth files, enables us to retain 24% and 8% of filtered target predictions (Figure 2-2D), and obtain precision and recall values of 25%, 19%; 29%, 11% respectively (Figures 2-2E, 2-2F), providing a more balanced precision:recall ratio.

To enable this analysis, we introduce mirDIP – the microRNA Data Integration Portal – a free and publicly available data portal integrating up-to-date microRNA target predictions from eleven individual source prediction databases [20, 23-35]. Similar to our Interologous Interaction Database (I2D) maintenance program, we will update it at minimum twice a year to ensure that the latest microRNA:target prediction data from all sources is available to users. Importantly, to ensure consistency and enable accurate re-analysis in the future using new and older data, we
keep track of versions of individual resources, and all mirDIP releases will be able to search the most current, or older versions.

Similar to mirGator, which amalgamates three microRNA databases (miRanda, picTar and TargetScan) with expression data while also providing enrichment analysis [144], mirDIP allows the user to take more control over the prediction data that they consider. Not only does our resource conveniently integrate twelve different prediction databases in one place, it allows users to choose which combinations of databases they would like to consider – refining options by database or by database characteristics – when selecting prediction data. This empowers users to capitalize on their knowledge of the workings of different databases, compensating for strengths and weaknesses of individual databases – choosing to focus on schemes considering different variables to create a customized prediction set based on the user’s preferences and tailored to application-specific tasks, taking into account the need for either high precision or high recall as discussed above. Section 5.1 in the Appendices introduces the mirDIP interface (Figure 5-1) and describes several search scenarios. Supporting Figures 5-2 to 5-5 display screenshots of the mirDIP search parameters. In the sections that follow, we describe how mirDIP can be used along with NAViGaTOR [150] – a scalable, network analysis and visualization system – to perform novel microRNA:target prediction visualization.

2.3.4 Construction of microRNA interaction networks

For the construction of microRNA interaction networks based on gene signalling pathways, we have refrained from using only targets from \textit{in vitro} or \textit{in vivo} experiments due to the obvious
bias present in such data. Rather, we have chosen to use interactions appearing at two different confidence levels – those present in at least 2, or at least 3 microRNA prediction databases (2+DB, 3+DB) as a threshold for robust microRNA:target predictions. Further, drawing from nine of the twelve databases indicated in Table 2-1 to determine the 2+DB and 3+DB datasets (eliminating the risk of double-counting by omitting the PITA All Targets and picTar 4-way databases and not including microT), we draw from 4/9 databases using conservation as a target site algorithm criterion and 5/9 databases not considering it. As such, we ensure that the requirement of sequence conservation does not influence the construction of microRNA networks in either direction. As discussed in the methods, permutation analysis was done by constructing microRNA interaction networks from the same microRNA interaction set (2+DB or 3+DB), based on the same number of starting nodes as the pathway or set of genes being examined. Random starting nodes were selected from the same interaction file being used to map the genes being examined.

Beginning with the well-known Phosphoinositide 3-Kinase (PI3K) pathway, we examined two aspects of this pathway with respect to microRNA involvement, garnering our pathway information from reviews discussing member-genes [173-176].
2.3.4.1 PI3K subunit regulation.

To examine the relevance of mapping microRNAs into signalling pathways, we chose to examine 2 separate coordinate signalling scenarios in the PI3K pathway. Well known for its control of a broad range of down-stream effector genes, the PI3K pathway is involved in cell growth, proliferation, differentiation, cell death, motility and survival. Implicated in many cancers, not only does it count as members many oncogenes, at the top of the pathway lies the most potent breast cancer oncogene known to date – receptor tyrosine kinase HER2 (also known as ERBB2) – a key receptor at the top of the signal transduction chain.

The PI3K family is divided into 3 classes. Members of each class of PI3K molecules comprise 2 subunits – a regulatory subunit and a catalytic subunit. These subunits are distinct proteins coded in different regions of the genome as either distinct genes or splice variants transcribed out of a similar locus producing translated proteins of varying sizes. The particular assembled combination of the 2 subunits of PI3K determine the molecule’s structure and function, and varying combinations of subunits are active in entirely different cellular settings [176]. Using interactions at the 3+DB robustness level, we map the microRNAs targeting genes involved in the assembly of Class 1 PI3K (Figure 2-3).

Immediately, it becomes evident that the possibility for PI3K subunit regulation at a post-transcriptional level is real. The network resulting from the input of all Class 1 PI3K subunit genes (PIK3CA/B/C/D, PIK3R1/2/3/4/5/6) contains five primary nodes (the other subunit genes are missing due to the lack of microRNAs targeting them in a sufficient number of databases), 181 secondary nodes and 206 interactions. Permutation analysis of randomly selected 5-node
networks confirmed that this provides a significant enrichment ($p < 0.05$) for number of nodes and interactions in the network. The most striking feature of the network is the participation of primary nodes in interactions with at minimum two other nodes – indicating that this network is significantly more connected through microRNAs than one would expect by chance alone ($p < 0.01$).

**Figure 2-3 MicroRNA interaction network for assembly of PI3K subunits.** Mapping PI3K subunits to microRNA interactions produces a network that is significantly more connected than at random ($p < 0.05$). Green nodes are regulatory subunits and yellow nodes are catalytic subunits.
2.3.4.2 Regulation of PI3K signalling.

To further examine microRNA involvement in this pathway, we use a model of the downstream signalling components of this pathway as indicated in a recent review [173]. Here we unveil a second highly-connected microRNA network (Figure 2-4) (based on 2000 permutations: \( p < 0.05 \) for number of nodes in the network, \( p < 0.05 \) for number of interactions in the network, \( p < 0.05 \) for number of nodes with degree \( \geq 4 \)).

It is quite surprising to see the number of microRNAs that can co-target potent tumour suppressors and oncogenes. We find a microRNA – hsa-mir-19b – that concurrently targets PTEN-TSC1-PI3KCA-TP53, and others that co-target RPS6KB1-PDK1-TSC1-PTEN and PTEN-RPS6KB1-FOXO3-TSC1. In addition, there are many microRNAs that target pairs of elements of this pathway: 15 microRNAs target RPS6KB1 and PTEN, 8 microRNAs target both RPS6KB1 and TSC1, and 4 microRNAs target both EIF4E and RPS6KB1. Clearly, we are only beginning to understand the level of regulation possible with microRNAs co-targeting many different genes, but it is becoming increasingly evident that this level of network complexity governs some interesting and previously hidden relationships between potent oncogenes and tumour suppressors in the cell.
Figure 2-4 MicroRNA interaction network elements of the PI3K pathway. Mapping the elements of the PI3K pathway based on a literature review [173], produces a network where many genes are targeted by common microRNAs, suggesting a novel microRNA role of pathway regulation.
2.3.4.3 Examination of KEGG and Reactome pathway-based MicroRNA Networks.

2.3.4.3.1 Basic Network Parameters

After initially testing our hypothesis on the PI3 Kinase pathway, we decided to undertake a more extensive and rigorous examination of signalling pathways within the cell. Since pathway definitions have not been unanimously settled and there is still much debate as to which resource defines a signalling pathway most accurately and comprehensively, we decided to use pathways delineated by the Kyoto Encyclopedia of Genes and Genomes database (KEGG) [177, 178] and pathways defined by the Reactome [148, 149, 179] database to further support the microRNA networks built based on expert-curated pathway reviews in the previous section. Examining interactions predicted at 2 threshold levels: 2+DB and 3+DB, we created microRNA networks for both the canonical signalling pathways and for 2000 permutations of pathways created with the same number of primary node genes. Our findings showed a similar trend for most interaction sets and signalling pathways that we examined. We found that true signalling pathways tend to involve more microRNAs and contain more interactions, as well as having more high degree nodes (degree ≥ 4) than pathways created out of a random set of starting nodes. We examined 9 KEGG pathways and 12 Reactome pathways at the 2+DB and 3+DB interaction thresholds. The pathways with the lowest average p-values (that is the average of p-values across the 4 measured parameters - number of network interactions, number of network microRNAs, number of network nodes with degree ≥ 4 and network density) were KEGG pathways: ERBB signalling pathway (hsa04012) (2+DB), mTOR signalling pathway (hsa04150) (2+DB), Wnt signalling pathway (hsa04310) (2+DB), MAPK signalling pathway (hsa04010) (3+DB) and...
Pathways in cancer (hsa05200) (3+DB) with average p-values of $p < 0.0006$, $p < 0.0009$, $p < 0.002$, $p < 0.002$, $p < 0.007$, respectively (Figure 2-5).

**Figure 2-5 MicroRNA interaction network characteristics.**
Examination of four microRNA interaction network characteristics across well-known signalling pathways using KEGG (panels A and B) and Reactome pathway databases (panels C and D). Signalling pathways tend to be enriched for the number of microRNAs, the number interactions and the number of high degree nodes mapped.

Of the pathways described in both the KEGG and Reactome databases (NOTCH, VEGF and WNT), WNT results were the least conserved across both databases – showing significance in
KEGG (average p-values of $p < 0.002$ and $p < 0.036$ for 2+DB and 3+DB respectively), but not in Reactome (average p-values of $p < 0.64$ and $p < 0.68$ for 2+DB and 3+DB respectively), while NOTCH measured parameters were the most likely to be consistent across the two databases (average p-values of $p < 0.102$ and $p < 0.105$ for 2+DB and 3+DB respectively in KEGG and average p-values of $p < 0.256$ and $p < 0.139$ for 2+DB and 3+DB respectively in Reactome). We found that some pathways had greater tendencies than others to show significance – for example the FGFR and Cell Cycle Genes pathways (which, it could be argued, is not a signalling pathway and hence does not fit within this study and hence acts as our negative control) described only by the Reactome database had a tendency towards higher p-values than other pathways examined (Reactome FGFR pathway average p-values of $p < 0.35$ and $p < 0.4$ for 2+DB and 3+DB respectively and Reactome Cell Cycle Genes average p-values of $p < 0.78$ and $p < 0.45$ for 2+DB and 3+DB respectively). The measured parameters found to be most frequently significant across all studied scenarios were the number of microRNA nodes in the network with degree $\geq 4$ (significant at $p < 0.05$ in 30/42 tested scenarios), and the number of total microRNA:target interactions in the network (significant at $p < 0.05$ in 27/42 tested scenarios). As highlighted in Figure 2-5 – one can find enrichments that are supported by both pathway databases, while other enrichments are highlighted in the analysis using one or the other pathway database. Examining expert-curated pathways, KEGG pathways and Reactome pathways with similar findings gives us confidence that this phenomenon is in fact real.
2.3.4.3.2 Centrality Measures.

We further examined network betweenness centrality (using Brandes’ algorithm [180] in R [181] using the RBGL package [182, 183]) as well as the average betweenness centrality of the top 10 genes and microRNAs by degree, and the average shortest path length between the top 10 genes. In general, these measures were not found to be significantly different in true signalling pathways from the random networks across most pathways. For the KEGG 3+DB signalling pathways network betweenness centrality – a measure of the difference between the node with the highest betweenness centrality (the node on the most shortest paths) to all other nodes in the network – we did see a small trend towards pathway networks having lower betweenness centrality (p < 0.0001 (WNT pathway) to p < 0.837 (VEGF pathway). This trend suggests that true signalling pathways have a more balanced centrality structure with fewer “hub” nodes than random networks do. However, we did not see any difference in the betweenness centrality of the top ten microRNAs by degree or the top ten genes by degree in the signalling pathways (p < 0.089 to p < 0.687 for microRNAs and p < 0.37 to p < 0.987 for genes). Further, due to the distributions of the network values for average and maximum shortest paths (measured with Dijkstra’s algorithm [184]) between the top 10 genes we were unable to conclusively evaluate these parameters (95% of average shortest path values were 3 and almost 75% of maximum shortest path values were infinite). This lack of conclusive significance in centrality measures can be explained by the fact that we did not model interactions between proteins in our networks, choosing to examine only interactions between genes and microRNAs. Thus, our networks tended to have a particular structure requiring all pathways to alternate between gene and
microRNA due to the lack of protein-protein connections. Integration of protein-protein interactions with microRNA-target interactions in a network could be re-examined at a later date.

2.3.4.3.3 Network Hubs
We also examined the possibility that hubs in these microRNA networks might be more likely to be date or party hubs as defined in Han et al.'s paper [185]. Using our I2D database [186, 187] we examined known human protein-protein interactions for a binomial distribution to define such hubs, and failed to find such a distribution, hence we are unable to further study any such relationship.

2.3.5 Universe and Intra-pathway microRNAs.
Upon realizing that microRNAs play a large role within signalling pathways – we produced a road map to delineate the inter-pathway connections (Figure 2-6).
It quickly became clear that there are distinct classes of microRNAs. Examining microRNAs with degree greater than two in any signalling pathway, we were able to identify 77 microRNAs that act only in an intra-pathway manner, affecting multiple targets but only within one single pathway. These microRNAs tend to target the ERBB, mTOR, MAPK, WNT and Jak-STAT pathways and no intra-pathway microRNAs appear to target the VEGF, NOTCH and PI3K pathways. We further identified 61 microRNAs targeting all 8 KEGG pathways that we examined at the 3+DB level, which we refer to as universe microRNAs. In attempts to validate
this classification of microRNAs into intra-pathway and universe classes, we went to the
literature. Searching for total PubMed articles, we see a significant difference between universe
and intra-pathway microRNAs (p < 0.0002) – with universe microRNAs discussed more
frequently (Figure 2-7A). Further, the most discussed microRNAs, hsa-mir-15a, hsa-mir-16 and
hsa-mir-34a have high degree in the many pathways in which they are involved (hsa-mir-15a has
intra-pathway ranking of 2(ERBB), 1(Jak-STAT), 2(MAPK), 3(VEGF), 4(mTOR), 1(WNT),
27(NOTCH)). This observation makes sense when one considers that many decisions regarding
the selection of microRNAs to study are based on high-throughput experiments, through over-
expression of a library of microRNAs and examination of several simple read-out conditions.
Figure 2-7 Comparison of universe and intra-pathway microRNAs.
Panel A: Universe microRNAs have a significantly larger number of PubMed papers compared to intra-pathway microRNAs ($p = 0.0002$).
Panel B: Universe microRNAs have significantly higher degree in the KEGG Pathways in Cancer 3+DB network ($p < 0.0001$).
Panel C: Universe microRNAs have significantly more predicted target interactions than intra-pathway microRNA across several different microRNA prediction databases ($p < 0.0001$),
Panel D: Top – Universe microRNA targets (red) tend to have more OMIM hits than intra-pathway microRNAs (blue). Bottom – Universe microRNAs themselves have more “cancer” PhenomiR hits than intra-pathway microRNAs ($p < 0.0001$), supporting the result in panel A.

It follows that microRNAs with involvement in many pathways – universe microRNAs – might be able to produce large changes within the cell, resulting in measurable outcomes compared to controls. As such, these microRNAs might be selected for further study, resulting in more PubMed articles. When constructing the microRNA road map from known signalling pathways in KEGG, we did not include the Pathways in Cancer gene network, since it is not a signalling pathway in its own right. Overlaying universe and intra-pathway microRNAs with the Pathways in Cancer Network built for Figure 2-5, we see that universe microRNAs have much higher degree than intra-pathway microRNAs in the Pathways in Cancer network ($p < 0.0001$) (Figure 2-7B). Considering that this type of effect could have been induced by our filtering methods, we examined our 3+DB interaction set for the number of targets predicted for both universe and intra-pathway microRNAs. We did see significantly more predicted targets for universe microRNAs than for intra-pathway microRNAs ($p < 0.0001$). However, this distribution was replicated in TargetScan predicted targets ($p < 0.0001$), PITA predicted targets ($p < 0.0001$) and picTar predicted targets ($p < 0.0001$) (Figure 2-7C). Since this distribution transcends any filters that we have applied and it holds for these individual database prediction sets we suggest that
universe microRNAs simply tend to have more targets, and are therefore able to exert a broader program of control over the cell than are intra-pathway microRNAs.

At this point, we would like to address the issue of bias in the data and distinguish microRNA interaction sets from protein interaction datasets. There is one large and obvious difference between protein and microRNA interactions. Protein-protein interactions are often curated through highly-biased information gathering methods; literature searches, which are biased towards highly-studied proteins, and high-throughput experiments focusing on finding all partners for one protein of interest, while considering a library of potential partners. Although useful interaction generating techniques, they cannot be relied upon to uncover protein-protein interactions evenly across the human proteome. MicroRNA:target relationships are different. The information upon which our study is built is entirely sequence-based. The databases considered do use different algorithms to make their predictions; however, the predictions are free from bias due to the ground truth that everything studied is sequence-based. Conservation of a binding site, binding site accessibility and presence or absence of a seed-region depend entirely on the coded gene, its transcribed RNA and the sequence of the microRNA that might bind to it, freeing us from the requirement to compensate for any bias in microRNA:target predictions. That being said, one possible bias that we cannot decouple from our current analysis is the relationship of the length of a given gene’s 3’ UTR and the number of microRNAs that target it. It remains unclear if the fact that genes with long 3’ UTRs tend to have more predicted targeting microRNAs is due to the fact that this is the way that the biology works or if it is simply related to the odds of having more binding sites in longer UTRs.
Finally, we examined the differences between universe and intra-pathway microRNAs in a disease setting. First, we examined the cumulative number of Online Mendelian Inheritance in Man (OMIM) database [188] hits for all targets of each microRNA (Figure 2-7D Top Panel). Arranging them in decreasing order we show that universe microRNAs have many more OMIM hits than do intra-pathway microRNAs. It should be noted here that we did not normalize for number of targets per microRNA. The lower panel shows the universe microRNAs having significantly more cancer hits for each microRNA in the PhenomiR database [189] compared to intra-pathway microRNAs (p < 0.0001). We see a strong distinction between universe microRNAs and intra-pathway microRNAs for disease association, again supporting our hypothesis that universe microRNAs are a subset of microRNAs that can target many genes within the cell, acting as master controls.

As further explanation for why universe microRNAs have been more studied than intra-pathway microRNAs, a comparison of microRNA “number” - their unique identifying IDs, which were assigned in approximate order of discovery - shows that universe microRNAs have a lower average identifying number than intra-pathway microRNAs (mean ID for universe microRNAs = 51 vs intra-pathway microRNAs = 84, p < 0.0001).

This may be either because universe microRNAs have been discovered earlier purely by chance and thus were more studied, or they may truly be more universal and thus were easier to discover under many conditions. To provide additional evidence to answer this question, we considered expression of microRNAs across a panel of tissues from Landgraf et al. [85]. Figure 2-8 shows a heatmap comparing universe and intra-pathway microRNA expression across tissues, confirming
that universe microRNAs are more widely expressed than intra-pathway microRNAs. Thus, it is more likely that universe microRNAs are more broadly affecting varying cell types, and through their misexpression, universe microRNAs have the opportunity to create a more global change quickly by affecting genes in many different pathways. To further understand their role in human disease thus warrants further research.

This work in microRNA interaction networks provides more evidence for the possibility that microRNAs are in fact working in a coordinated fashion with each other and within signalling pathways. It has been previously noted that many microRNAs might co-bind to a UTR [2, 31, 157, 190], and perhaps our results support that view, since genes in a common pathway share many more common microRNAs than one would expect by chance (p < 0.0035 to p < 0.365 for KEGG 2+DB). This thinking opens the door for many exciting in vitro experiments to examine this co-regulation and co-binding, and raises the questions, how many microRNAs might actually be occupying a 3’ UTR at once? Is it a sequential or a parallel microRNA process?

Future work to determine the layout of such microRNA binding sites in the untranslated regions might provide further insight here – and a within pathway study of the degree of overlap and layout of microRNA binding sites on interacting genes would provide insight into the microRNA regulatory network. Another interesting expansion of this work would be to determine predicted binding sites in 5’ UTRs and coding regions of target genes, and integrate them with RNA22 predictions already included in mirDIP to allow cross-database comparisons. While the majority of confirmed microRNA binding sites fall into 3’ UTRs, functional binding sites have been shown in other regions [191-193] and attempts to include them in mirDIP would result in a more complete representation of true microRNA target genes within the cell.
Figure 2-8 *Expression of universe and intra-pathway microRNAs.* Universe microRNAs are expressed in a broader panel of tissues than intra-pathway microRNAs [194].
The identification of two distinct classes of microRNAs – universe and intra-pathway microRNAs – lays the framework for possibly hierarchical organization of pathway- and gene-level control and execution of gene regulation. Using PhenomiR, we provide the first disease-associated evidence that universe microRNAs may be more likely to be involved in cancer specifically – showing significantly more involvement in breast (p < 0.07), ovarian (p < 0.005) and lung cancers (p < 0.05) and in carcinogenesis overall (p < 0.0001) while also showing involvement in human disease in general (p < 0.0001), and this information will allow us to focus our disease-driven microRNA-associated research towards a smaller subset of these potent cellular regulators.

2.4 Conclusions

2.4.1 MicroRNA Prediction Databases

Similar to work done by other groups, we have examined microRNA prediction databases to determine that PITA Top, picTar 5-way and Targetscan Conserved provide the most accurate microRNA:target predictions. Using different prediction algorithms, individual predictions overlap only partially and they differ in their precision and recall when compared to in vitro truth data. However, each has a particular application where it might be best suited for use. We have further examined the importance of filtering target predictions before making microRNA database comparisons, and have determined that filtering by both experiment cell type and microarray chip type are crucial steps that alter gene prediction sets by up to 40%. We suggest that when searching for true microRNA targets, it is useful to consider such steps.
2.4.2 mirDIP

We have presented a unique database to aid researchers in determining the optimal microRNA prediction databases to use for application-specific microRNA:target searches. mirDIP allows users to focus their searches on any subset of microRNA prediction databases, in either “high precision” or “high recall” databases depending on their path of study.

2.4.3 Discovery of Universe and Intrapathway MicroRNAs in Interaction Networks

Using data from mirDIP, we found that microRNAs are significantly more involved in known signalling pathways compared to random chance, producing networks with more interactions (p < 0.1 in 76% of tested pathways). Signalling pathways contain many microRNAs that target multiple elements of the pathway, perhaps suggesting a level of transcriptional regulation not previously described. Our data suggest a possible co-regulation of signalling proteins at the post-transcriptional level – whether concurrent or sequential – which opens a new line of research to study hierarchical organization of microRNAs. Further, we have identified two novel classes of microRNAs: universe and intra-pathway microRNAs, which are significantly differentiated by the degree of their involvement in signalling pathways within the cell and their association with cancer (p < 0.0001) and human disease (p < 0.0001). Universe microRNAs are involved in regulation of many known signalling pathways, while intra-pathway microRNAs are pathway-specific and do not appear to play a global role in cellular regulation.
2.5 Materials and Methods

MicroRNA predictions were downloaded from the individual microRNA prediction sites:

- [http://microrna.org](http://microrna.org) (Sept. 2008),
- [http://microrna.sanger.ac.uk/targets/v5/](http://microrna.sanger.ac.uk/targets/v5/) (ver. 5) (now [http://ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/](http://ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/)),
- [http://genie.weizmann.ac.il](http://genie.weizmann.ac.il) (ver. 6 Aug. 2008),
- [http://targetsan.org](http://targetsan.org) (Release 5.0 Dec. 2008),

Target Prediction Files

All target prediction files were processed to contain the same information in the same format. The UCSC Genome Browser ([http://genome.ucsc.edu/](http://genome.ucsc.edu/)) [195-197] and Galaxy [198, 199] were used to convert all files to include HUGO gene names for all interactions according to human genome version hg18. RNA22, picTar and DIANA microT required intermediate mapping steps using Ens54 [200] and RefSeq May 2006 [201] assemblies. All files were then combined to produce one file of all predictions. A filtering step produced the interaction files for
NAViGaTOR – eliminating all interactions present in less than 2 or 3 microRNA prediction databases. To avoid double-counting interactions present in 2 databases from the same source compiled with different stringency requirements, only the most stringent PITA and picTar microRNA prediction files were used as inputs into the integration and filtering steps.

MicroRNA microarray Truth Files

Files used to compare microRNA prediction files to truths were obtained from the following GEO Datasets: GSE2075[85], GSM306946[162], GSE6838/GSM155064[166], GSE6207[168], GSM302945[162], GSE13296[163], GSE6838[166], GSM302995[162], GSE12091[164], GSE9586[167], GSE9742[165], GSE14507, GSE12092[164]. Thresholds for low-, med- and high-confidence truths were established using p-values of p<0.1, p<0.05, p<0.01, where replicates were present, and otherwise at three step-wise incremental knockdown or over-expression thresholds dependent on the distribution of target knockdown – 50%-25%-10% for mir-335 (GSE9586) and mir-7 (GSE14507) (in this case – since there were only 2 replicates, we also required the replicates to be within 15% of each other), 75%-25%-10% for mir-155 (GSE13296) and 25%-20%-10% for mir-124 (GSE6207).

Target Filtering

To filter target predictions prior to prediction database comparison, we used genes present in the bottom quartile of the control cell line microarray experiment. In most cases, one or more
negative control sample values were present and those values were averaged and then ranked by intensity value. When filtering by experimental cell type, only genes not present in this bottom quartile passed through our filter. In the few cases where it was not possible to extract the control cell line values from the experiment (mir-1 Lim et al. [5] and mir-124 Lim et al. [5]), filtering genes from the negative controls from a microarray experiment in the same cell line were used (mir-98 negative controls Gennarino et al. [53]). We further filtered by the presence of the predicted target gene on the microarray chip used in the experiment, information available at GEO datasets.

NAViGaTOR Networks

NAViGaTOR networks [186, 187] were built based on the microRNA:target interaction files discussed above, with two levels of robustness: interactions present in two or more databases (2+DB) or interactions present in three or more databases (3+DB). Note that out of the twelve databases examined in the first section, only nine are used for the microRNA interaction networks due to the fact that PITA Top Targets (used) is a subset of PITA All Targets (not used) and picTar 5-way (used) is a subset of picTar 4-way (not used). Using groups of Associated Genes of interest (as determined by well known sub-units [176], pathways extracted from the literature [173] and KEGG [177, 178] and Reactome [148, 149, 179] databases) as primary nodes – networks were built to examine the interactions between the given associated gene set at the microRNA level. Associated gene network significance was evaluated based on four characteristics: 1) the number of nodes in the network, 2) the number of interactions in the
network, 3) the number of nodes with degree greater than three, and 4) the measured network density, and compared to values obtained from 2000 random networks constructed from the same number of primary nodes (genes randomly selected from the interaction file, hence genes that have been identified as participating in a microRNA interaction by at least two or three prediction databases). KEGG pathway HUGO IDs were used to create networks, while Reactome Swiss Protein IDs were mapped in the UCSC Genome Browser to HUGO IDs before networks were built. Networks were built using the graph (ver. 1.24.1) [183] and RBGL (ver. 1.22.0) packages [182] of the R Statistical Package software (ver. 2.8.1) [181]. When comparing pathways represented in both KEGG and Reactome databases, comparisons were made between the differences of the sums of the p-values of the four network parameters. All analysis was done using NAViGaTOR ver. 2.1.13 [1] (http://ophid.utoronto.ca/navigator).

Examination of Date and Party Hub Nodes

In our examination of human protein-protein interactions to determine whether a bimodal date and party hub distribution was present, I2D human source interactions were used [186, 187].

Universe and Intra-pathway microRNAs

Using NAViGaTOR ver. 2.1.13 to display the microRNA:pathway interactions from the KEGG 3+DB study, we laid out the micronome roadmap to identify universe and intra-pathway
microRNAs. Comparisons between the two classes of microRNAs and number of associated PubMed articles were done using biopython (v1.50) (http://biopython.org). OMIM [188] hits and PhenomiR (v1.0) [189] hits were drawn from their respective sources (http://www.ncbi.nlm.nih.gov/omim/ (accessed Feb. 2010) and http://mips.helmholtz-muenchen.de/phenomir/).

Tarbase Comparison

We used Tarbase V5.0 [171] to compare our 2+DB interaction set to the best curated set of microRNA interactions existing. We used only human interactions, eliminated the support_type=FALSE interactions and mapped by the HGNC column.

Details about mirDIP can be found in Supporting Methods S1 in the Appendices.

Supplemental Materials and Supporting Methods can be found in the Appendices Section.

Acknowledgements

The authors would like to thank Isaac R. Smith for his assistance with our network figures and Milton R. S. Smith for being a baby who slept – for the first month anyway.
Chapter 3

A Global Analysis of MicroRNA Loops Identifying Secretion-related Motifs and Structures

3.1 Abstract

Pursuant to previous work examining relationships external to the micronome through microRNAs targeting gene networks, we have now undertaken a study more inwardly focused on analyzing the characteristics of microRNAs within the micronome itself. Beginning with a study on microRNA seeds, we found that many microRNAs share common seeds. We compared microRNA targets of pairs of microRNAs with identical seeds, to pairs with a 1 basepair seed-difference across several microRNA prediction databases. We found that a one basepair difference in microRNA seed results in a >80% loss of common microRNA targets. Work in our previous paper and by others, has outlined the poor performance of individual microRNA prediction databases. Given their seed-centricity, we focused on systematically characterizing the under-examined microRNA terminal loop across the micronome to determine if this region
of the microRNA structure might contribute to microRNA functioning, in a manner not captured by microRNA prediction databases. We have identified 5 broad classes of microRNAs based on loop-structure: perfect loops, occluded loops, closed loops, small loops and runover loops – with 90% of microRNA loop structures falling into the occluded and perfect loop classes. We have been able to show class enrichment in several biologically relevant groups; analysis of a study of in vitro mesenchymal stem cell secreted microRNAs shows an enrichment in the occluded class (p<0.05), KH-type splicing regulatory protein (KSRP)-dependent microRNAs also show an occluded-type enrichment (p<0.05). To determine if there might be regulatory motifs within microRNA terminal loops, we looked for commonly occurring words of varying length and found a large number of Frequently Occurring Words (FOWs) significantly overrepresented across the micronome. Set analysis of in vitro secreted microRNAs, microRNA expression across a panel of normal tissues, and microRNAs shown to be secreted in lung cancer shows that specific microRNA loop motifs within these groups are significantly overrepresented, suggesting that microRNA terminal loops harbour sequences bearing microRNA processing and localization signals.

3.2 Background

MicroRNAs are short, non-coding RNA sequences that regulate gene expression [1] through their targeting of mRNA Untranslated Regions (UTRs), disrupting their target’s ability to be translated into proteins or by repressing the expression of the target mRNA itself [2-9].
Since their discovery and characterization in worm (C. Elegans) in 1989 [10-13], the flood gates have opened to new and exciting research to determine just how these tiny but powerful regulators operate. Once it was determined that microRNAs likely have human homologues, after the discovery of let-7 in worm – the term microRNA was coined and the search for microRNA targets became a top priority. [14-19]

Working at the RNA level, microRNAs and their targets are related through nucleotide binding between the seed region of the microRNA (bases 2-7 from the 5’ end of the mature microRNA sequence) and the 3’ UTR region of the gene transcript. This problem lent itself beautifully to a bioinformatics approach and work began in earnest in plants [5], fly and humans [20-22] and many public databases exist to tackle the problem – each taking a slightly different approach [20, 23-35, 39, 40].

As bioinformatics work focused on seed region pairing, increasing evidence pointed out that microRNA loop sequences aren’t limited to the silent ride-along structural sequences that they have been assumed to be for so long. The first clue comes in the difference in microRNA processing between plants and animals. In more complex organisms microRNAs retain their loops until they are in the cytoplasm, at which point they are acted upon by Dicer to produce the mature microRNA sequence to be loaded into the RISC complex. Plant microRNAs lose their loops much earlier – before they are exported from the nucleus. This may suggest that in higher organisms, the microRNA loop structure plays some recognition or localization role in the cytoplasm [94, 202].
Pri-microRNAs are the longest microRNA precursor transcripts, which can be hundreds of nucleotides in length and may contain the sequences of one or more microRNA stem-loops coded proximally in the genome and which exist only in the nucleus. Pre-microRNAs – the Drosha-processed pri-microRNA now consisting of a single stem-loop – and pri-microRNAs have always been considered to be transitory elements in the microRNA processing pathway. Once the loop has been lost and the microRNA loaded into the RISC – only then does the critical microRNA function occur. However, early studies into shRNA design show that the loop sequence is critical for efficient messenger RNA silencing, likely because Dicer processing occurs near the loop sequence [203]. Further, Liu et al. undertook an interesting series of experiments on mir-181a-1 and mir-181c – two microRNAs from the same microRNA family with similar mature sequences (a one nucleotide difference), but distinct loops. The two microRNAs show differing activities in T-cells, where ectopic mir-181a-1 expression promotes CD4/CD8 Double Positive development, but mir-181c does not. By mixing and matching different stem and loop regions between the 2 microRNAs, they narrowed down the loop region as being critical to maintaining the ability of hsa-mir-181a-1 to promote CD4 and CD8 double-positive T-cell development when ectopically expressed in thymic progenitor cells. They argue that although the mature microRNA sequences are similar, divergent loop regions between family members – which are conserved across species -- specialize similar microRNAs to particular tasks [95]. The importance of loop regions is further supported by the fact that loss-of-function phenotypes for lin-4 and let-7 cannot be rescued by the mature microRNA sequences alone – insertion of the pri-microRNA sequence is necessary [1, 16].
Several recent papers suggest mechanisms by which microRNA loops may play important roles as binding sites for regulatory proteins. A first example of a protein binding to the terminal loop nucleotides for the Drosha-mediated processing of a microRNA was identified in 2008, when several groups established a role for the pre-microRNA loop in the regulated production of let-7 [109-112]. They found that mouse embryonic stem cells contain a Drosha inhibitor, an embryonic stem cell specific protein Lin-28, that specifically binds to the conserved nucleotides in the loop region of pre-let-7, resulting in its regulated production.

Further study of let-7 shows that KH-type splicing regulatory protein (KSRP) is a promoter of microRNA biogenesis at both the Drosha and Dicer processing steps, through its ability to bind terminal loop sequences[119]. This is particularly relevant since KSRP has been shown to be a member of both the Drosha and Dicer processing complexes [119]. KSRP-loop binding seems to be a microRNA-specific binding action and relevant to a subset of microRNAs containing a GGG motif – different from the AU-rich motif, which KSRP has generally been known to bind. Studies have suggested that KSRP and Lin28 act in an antagonistic fashion, since they bind to the let-7 terminal loop in a mutually exclusive manner, even though they bind unique regions of the loop – perhaps sterically hindering each other from their optimal binding position[204].

Michlewski et al., found that Heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1), a protein implicated in different aspects of RNA processing, plays a role in the posttranscriptional regulation of pri-microRNAs harbouring conserved terminal loops [114]. They show that the binding of hnRNP A1 to its consensus binding site UAGGGA/U in the loop of pri-microRNA-18a is essential for its processing and this processing can be blocked by oligonucleotide
sequences complementary to the terminal loop. The authors believe that the highly conserved loops act as landing pads for trans-acting factors, which influence microRNA processing. Loop targeting oligos designed against a number of pri-microRNAs with conserved loop sequences were able to specifically block the processing of the corresponding pri-microRNAs.

Yet another possibility for the function of microRNA loops involves a role in microRNA secretion. The suggestion of microRNAs in serum as biomarkers for pregnancy by Gilad et al. in 2008, and other recent studies have shown that microRNAs are not only cell bound [121, 139, 205, 206]. The classification of cancers based on serum microRNA expression [127], the discovery of microRNAs in exosomes and vesicles in cell media [120], in exosomes in human breast milk [207] and even the discovery of plant microRNAs in human serum – after food processing and after having passed through the entire digestive tract [208] - are all leading us to the conclusion that microRNAs are much more stable, mobile and ubiquitous that was previously thought. Since some studies suggest that it is the pre-microRNA forms that are present in an extra-cellular setting[128], perhaps in some form of cell-cell communication, there will be many levels of regulation involved in this secretionary path and we also investigate whether microRNA loop structure could play a role here.

The whole picture of microRNA processing, mechanism, export and gene modulation functions are not yet clearly mapped and understood. The work that follows undertakes a broader examination of human microRNA loop sequences, examining characteristics of loops, their sequences, relationships between loops of different microRNAs and discusses several possible
unexplored functions of the loop, including implications in recognition and export from the primary cell of origin.

3.3 Results and Discussion

3.3.1 MicroRNA Seeds may not be Revealing the Whole Story

3.3.1.1 MicroRNA Targeting Database Predictions are highly seed-centric

Initial clues to the potential importance of another non-seed region of the microRNA emerged with examination of microRNA target-prediction sources. Our previous work has indicated that individual microRNA prediction databases do not perform particularly well when compared to validated microRNA targets. We were interested to understand why. We found that existing microRNA prediction databases are highly seed-centric. Figures 3-1 A and B indicate the effect of a one base pair change in seed on microRNA target predictions.
Figure 3-1 *The importance of microRNA seed sequences across microRNA prediction databases*

For all microRNA prediction databases, all pairs of microRNAs with identical seed sequences (Panel A) are compared to all pairs of microRNAs with seeds differing by one base pair (Panel B) in terms of the number of shared targets.

Figure 3-1 A examines all pairs of microRNAs with the identical seed. It is evident that microRNA seed is the driving force behind target prediction with all databases showing a percentage match peak above 60% -- meaning that over 60% of predicted targets are shared between microRNAs with common seeds – with most databases reporting shared targets at the 80-100% level. To see the effect of a small change in seed, Panel B shows the percentage of common predicted targets between microRNAs with a 1-bp difference in seed. In this case, the common targets are abolished and the sets are largely disjoint – showing just how heavily microRNA target prediction databases rely on microRNA seed region, and how biased thinking
is towards the seed region as the only important region for microRNA functioning. We have discussed that that mir-181a and mir-181c have different functions in T-cells, which implies different target genes or processing steps despite identical seed sequences [95]. The human let-7 family also shares identical seed sequences, but shows differential temporal expression during development [209]. Let-7a-3 has been identified as being oncogenic, while most other human let-7 family members have been described as tumour suppressors [52, 210]. Although not concrete evidence that predictions would benefit from examining regions of the microRNA outside the seed sequence, the fact that microRNAs with identical seeds do not behave in identical manners in identical settings, suggests that there is more to the story. Surprisingly, the microRNA literature lacks a study directly comparing two microRNAs with identical seeds and their ability and effectiveness at binding identical target sites to evaluate the actual contributions of other regions of the microRNA. Since microRNA prediction databases are not performing particularly well in the translation to the bench – we and others have reported that the very best precision scores between predicted and actual targets from any individual microRNA prediction database is roughly 40% from picTar, but this results in an extremely low recall value of below 5% – we suspect that more information has to be considered than just simple seed targeting within the microRNA cell of origin [162, 211].

The most obvious, and generally overlooked, non-seed region in the microRNA structure is the terminal loop (Figure 3-2).
Figure 3-2 Precursor microRNA (pre-microRNA).
Schematic of microRNA components and nomenclature.

Considering the importance of the loop sequences in other stem-loop structures – particularly tRNA, where the cloverleaf structure of the molecule is integral to its function and the sequence of the anticodon loop is of utmost importance [212] – we began to examine microRNA terminal loops in more detail.

3.3.1.2 The microRNA seed-space is relatively empty – yet many microRNAs share seeds
Historically, microRNA research has been narrowly focused on the binding relationships between microRNAs and their cognate mRNAs. This focus has driven microRNA databases to be highly seed-centric, regardless of the fact that there is a remarkable redundancy in microRNA seed composition. Our analysis reports that there are many more microRNAs with common
seeds than one would expect given the nomenclature. In fact, considering the size of the microRNA seed space (all the possible combinations of microRNA seeds – which is $4^6=4096$ sequences), there are relatively few seed sequences. Based on all 1734 annotated mature microRNAs (mirbase version 17) examining bases 2-7 of the mature microRNA (the seed region), there are 1157 distinct seeds (Figure 3-3).

![Mature MicroRNA Space Properties](image)

**Figure 3-3 Properties the MicroRNA Seed Space**
Comparison of properties of mature microRNA sequences based on mirbase version 17, and the available number of distinct microRNA seeds.

This represents only 28% of possible seeds. The most common seed AGCAGC occurs 8 times (mir-15a/b, mir-16, mir-195, mir-424, mir-497, mir-503, mir-646) and there are 819 seeds occurring in only a single microRNA. 915/1734 of microRNAs share a seed with at least one other microRNA. Since the seed space is so large and empty, it is intriguing that so many microRNAs would share common seed regions given that seed region is thought to be the driving force regulating microRNA binding. Aside from redundancy, we investigate whether
there are other regions of the microRNA that might be implicated in function and which might suggest a way to diversify this common-seed scenario – perhaps suggesting a further specialized function for two microRNAs with identical seed sequences. We hypothesize that perhaps another region of the microRNA sequence could harbour code to control or regulate a mechanism to guide microRNA co-localization to specific tissues, between entities within the cell, along the lengths of 3’ UTRs or within compartments of the cell, perhaps even marking microRNAs as those eligible for cellular export through vesicular pathways.

3.3.2 An Examination into MicroRNA Loops

3.3.2.1 Definition of a microRNA Loop

To further investigate the significance of the terminal loop across human microRNAs, we first had to establish exactly what constituted a microRNA loop. The determination of the starting and ending points of a microRNA loop were not trivial. Due to varying degrees of pairings between opposing bases in the loop region, we decided to impose two different loop definitions. Basing our interpretations on information parsed from mirbase structure diagrams displayed on individual microRNA loop pages [25, 26] (see example here: http://mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0001150), Minimal Loops were defined from the loop region of the microRNA, advancing towards the stem. Proceeding from the tip of the loop region, and considering the base pairs on each side of the stem -- all bases not involved in any pairing were included in the loop definition, until the first bonded base pair was encountered – signalling the end of the Minimal Loop. See Table 3-1 for several microRNA examples. Maximal Loops were
determined from the opposing end of the stem loop. Walking along the stem from the feet of the hairpin, Maximal Loops begin at the first unbound base after the final base of the mature microRNA (Table 2-1). Mature sequences were stripped from html code on the latter portion of the individual mirbase microRNA pages. Definitions based on the star (*) strand, soon to be called -3p sequences, were used in place of the mature -5p sequences should the -5p end of the hairpin not contain a mature sequence [213].

<table>
<thead>
<tr>
<th>MicroRNA</th>
<th>Minimal Loop</th>
<th>Maximal Loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>-- uc uu uc</td>
<td>uc</td>
<td>ucaaccuu</td>
</tr>
<tr>
<td>gugagagucag</td>
<td>-- cc uguu</td>
<td>ucaaccuuuc</td>
</tr>
<tr>
<td>acugggcg</td>
<td>gu</td>
<td>ucaaccuu</td>
</tr>
<tr>
<td>gugagagucag</td>
<td>gu</td>
<td>ucaaccuuuc</td>
</tr>
<tr>
<td>gu</td>
<td>-- cc uguu</td>
<td>ucaaccuuuc</td>
</tr>
<tr>
<td>hsa-mir-196b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MicroRNA</th>
<th>Minimal Loop</th>
<th>Maximal Loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>-- uu uugccua</td>
<td>a cugacgcac</td>
<td>accuga</td>
</tr>
<tr>
<td>gugagagucag</td>
<td>gu</td>
<td>cugaccacuga</td>
</tr>
<tr>
<td>gu</td>
<td>-- cc uguu</td>
<td>ggaau</td>
</tr>
<tr>
<td>hsa-mir-1269</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MicroRNA</th>
<th>Minimal Loop</th>
<th>Maximal Loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>-- aa uguu</td>
<td>gguu</td>
<td>gcua</td>
</tr>
<tr>
<td>gugagagucag</td>
<td>gu</td>
<td>None</td>
</tr>
<tr>
<td>gu</td>
<td>-- cc uguu</td>
<td></td>
</tr>
<tr>
<td>hsa-mir-924</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3-1 Determination of MicroRNA Loops drawn from mirbase [25, 26].
3.3.3 Conservation across Loops but not Seeds Exists for Some MicroRNA Families

To highlight the importance of microRNA loops within some microRNA families, we performed an alignment of microRNAs within two families, the well-studied let-7 family and the mir-154 family (Figure 3-4)[214, 215]. As might be expected, the let-7 family of microRNAs aligned predominantly around their seed sequence (shown in blue), while showing similar but non-identical loop structures (loop bases shown in red).

Figure 3-4 Alignments of the mir-154 and let-7 families
MicroRNA stem-loop sequences of 2 families of microRNAs showing distinct patterns of sequence alignment. The let-7 family (Top Alignment) aligns by seed region, whereas the mir-154 family (Bottom Alignment) aligns by loop region.

However, the mir-154 family of microRNAs aligned in the converse manner. They predominantly share a loop sequence, with variations occurring in the seed sequence. The determination of functionally important elements through conservation analysis is often used to
justify the importance of the microRNA seed region. In this case – we show that similar reasoning can be used to make an argument for the importance of the terminal loop regions, which are conserved mainly in primates. miRBase groups microRNAs into families based on sequence conservation. The microRNAs in the alignment in Figure 3-4 belong to the same family – meaning it is likely that they have arisen from a common ancestor, followed by drift - thereby acquiring mutations and resulting in variations across the length of the hairpin structure. However, the terminal loop region remains conserved across these 8 microRNAs in the mir-154 family likely because it has a function – perhaps acting as nucleotide motifs that interact with trans-acting factors influencing microRNA processing. Furthermore, in contrast to the let-7 family, which is conserved all the way to worms, the mir-154 family exists only in placental mammals. Although highly conserved across most of the hairpin, there is some variation in mir-154 across the second half of the loop sequence, perhaps making a case for speciation through the specialization of microRNA loops across species.

3.3.4 Common loop motifs cannot be explained by similar stem loops containing identical seeds

Out of 1734 mature microRNAs across the micronome, we considered those 915 which share seeds. We clustered microRNAs hierarchically by their hairpin loop sequence alignments using Clustal W [214, 215] and laid out the results in a polar dendrogram using Dendroscope [216] (Figure 3-5). We identified the most commonly occurring microRNA seed sequences by colouring leaves representing those microRNAs similarly. This enables us to see that although
there are distinct clusters based on microRNA loop region sequences, they are not associated with microRNAs sharing seed sequences.

**Figure 3-5 Loop sequence alignment of all microRNAs that share a seed sequence.**
Polar Dendrogram showing the alignments of all microRNAs in mirbase version 17 which share a loop sequence with at least one other microRNA. Sets of microRNAs sharing the most common seeds are coloured by seed sequence.
Figure 3-6 *Common seeds cannot explain most microRNAs with shared loop sequences.*
Plots showing the percentage match in seed sequences between sets of microRNAs sharing loop motifs for Minimal Loops (Panel A) and Maximal Loops (Panel B) by motif length.

Looking across sets of microRNAs sharing common motifs of varying lengths in their terminal loops, very few microRNAs that share a motif in their terminal loop also have an identical seed, in fact less than 20% of microRNAs in any group shared a seed sequence with any other microRNA in the group, confirming that loop motifs are not solely picked up due to full length stem-loop conservation across microRNA families, and might indeed identify novel subgroups of microRNAs.
3.3.5 Binding-Based Interactions

3.3.5.1 Commonly occurring RNA loop-motifs suggest functional significance

To systematically analyze the sequences of microRNA loops, we created 1000 simulated loop micronomes by permuting the nucleotides of the actual microRNA loop sets. We compared the number of occurrences of motifs ranging in size from 4-8 bases for Maximal Loops and 3-7 for Minimal Loops in the actual loop-space to the number of the identical motifs in the simulated loop-spaces. We selected motifs that were present in the generated micronomes at levels greater than or equal to the actual micronome less than 10 times, giving us a p < 0.01 significance level. This set of loop motifs we refer to as “Frequently occurring words” (FOWs). Figure 3-7 shows the data processing pipeline for the analysis for FOWs.

Figure 3-7 Data Processing Pipeline for FOWs
Figures 3-8 panels A and B show a further display of the likelihood of these FOWs in the micronome.
Figure 3-8 Likelihood of FOWs. Comparisons of the size of the motif space for x-length motifs, and the number of x-length motifs that fit into the microRNA loop space for Maximal Loops (Panel A) and Minimal Loops (Panel B). Frequency of the most common x-length motif is in red, and the expected frequency of all x-length motifs are in black.

In panel A, the grey bars represent the total number of word combinations for a particular length word. In black is the total number of X-length words that will fit into the microRNA loop-space. The black number represents this ratio – the expected number of any given X-length word in the micronome. The red numbers indicate the frequency of the most commonly occurring word of X-length in the actual micronome. In all cases, the frequency of the most commonly occurring word is significantly higher than the expected frequency. The probabilities that a motif occurs by chance at least as frequently as Figure 3-8 A and B suggest for Maximal Loops results in highly significant p-values with $p<6.2\times10^{-13}$ as the upper bound (Binomial Test).
3.3.5.2 Frequently Occurring Words are AU-rich

This set of FOWs has two interesting properties. First, this subset of microRNA is AU-rich. AU-rich elements (AREs) have long been known to be involved in mRNA turnover, transport and translation through their interactions with RNA-binding proteins [217-219]. In Maximal Loops and Minimal Loops respectively, the basepair distributions are: 27.75% U, 26.04% A, 21.63% G, 24.57% C, and 30.54% U, 30.15% A, 20.06% C, 19.24% G. Figure 3-9 demonstrates that as we traverse from microRNA hairpins, to mature sequences to terminal loops, microRNA basepair content becomes more and more AU-rich.

![Basepair Content of Human microRNA entities](image)

**Figure 3-9 Basepair Distribution of microRNA species**
Comparison of the distributions of base pairs within the micronome across varying lengths and of species.
Figure 3-10 uses Wordle software to display the most common 6-letter Maximal Loop motifs [220]. The three most common 6-letter motifs in microRNA Maximal terminal loops are UUAUUU (occurring 14 times), it’s reflection, UUUAUU (occurring 11 times and possibly facilitating recognition by the same protein) and the very similar UUACUU (occurring 11 times with Cytosine being very similar to Uracil, in fact deamination of Cytosine – a spontaneously occurring phenomenon in nature – produces a Uracil base).

Figure 3-10 Wordle showing Maximal Loop 6-letter FOWs
Motif size is related to the frequency of the motif in terminal loops across the micronome.
Beyond their propensity to form loops, these frequent and similarly-structured AU-rich words suggest the presence of a binding site for a cis-acting protein among microRNA terminal loops.

An interesting corollary to this AU-rich characteristic comes from the observation that microRNA binding sites preferentially reside within locally AU-rich regions[27], perhaps hinting at the possibility that microRNA loops are present late enough in the processing steps -- perhaps with Dicer excision and RISC-loading happening at or very near to the target site -- to have some sort of interaction with the 3’ UTRs. This remains a possibility with Maniataki noting that in animals, Dicer and AGO loading are possibly closely linked processes [221], meaning that the microRNA loop might be kept until just before RISC loading. Evidence of microRNAs that skip the dicer-processing step also lend credence to this loop-UTR binding possibility. Hsa-mir-451 and hsa-mir-144 are transcribed into one pri-microRNA. Mir-451 is extremely short -- its stem is only 17-nt long, and it has been shown to be a Dicer-independent microRNA -- meaning that, perhaps due to its length, it cannot be processed by Dicer and takes another processing pathway to RISC loading, while its pri-microRNA transcript mate, mir-144 follows the canonical path. Instead, mir-451 is directly loaded into the AGO RISC and it appears to be cleaved there by the AGO2 [222-224]. This is particularly relevant to our studies because this would mean that mir-451’s terminal loop structure would be maintained well into the RISC, and it shows the possibility of microRNA loops playing a role at later stages in the biogenesis pathway, nearer to the UTR sequence.

Another possibility is that the proteins that recognize the AU-rich sequences in the microRNA terminal loops will also recognize AU-rich regions within mRNA 3’ UTRs and may play a role
in guiding the mature strand and the RISC complex to its cognate binding site once it has been excised from the loop fragment by the Dicer complex.

Another obvious possibility for the recurrence of AU-rich themes in terminal loops is the existence of binding sites for transacting factors that interact with microRNA terminal loops. KH-type splicing regulatory protein (KRSP) which generally binds AU-rich sequences in mRNAs mediating decay, has been shown to be an element of both the Drosha and Dicer complexes binding to the terminal loops of microRNAs thereby regulating biogenesis (in this case binding a GG\_G tetramer)[119]. HuR proteins are also known to bind AU-rich mRNA sequences and have been implicated in microRNA processing, likely through their ability to bind AU-rich sites in 3’ UTR regions - which are proximal to microRNA binding sites - in antagonistic fashion promoting their stability from microRNA processing[225, 226].

There are many candidate proteins with preferences for binding AU-rich regions in RNA and this lends some support for the importance of the AU-rich regions that we have identified as frequently occurring and potentially important in microRNA terminal loops.

### 3.3.5.3 FOWs have Frequently Occurring Complements and Reverse Complements

The second characteristic of the frequently occurring terminal loop motifs that we have identified is that these frequently occurring microRNA words have complements and reverse complements which are also frequently occurring. This implies that pairs of microRNA terminal loops have complementary regions that may be inclined to form bonds. This was an unexpected and
somewhat confusing discovery. Table 3-2 shows the frequency of complement and reverse-complement pairs within the set of frequently occurring words.

Statistically, the likelihood of these complementary-pairs co-occurring in this set – for all word lengths – is extremely small, less than $p < 0.0001$. Figure 3-11 displays the data on the base pairs at the very tips of microRNA loops. Triplets are plotted along the x-axis in descending order of frequency across the micronome loop apexes. Bars are coloured by complementary pairs (Top Graph) and by reverse-complementary pairs (Bottom Graph). This figure drives home the relationships between frequently occurring motifs within loops, since triplets with their frequently occurring complements or reverse complements are clearly clustered among the most common triplets.
### Minimal Loops

<table>
<thead>
<tr>
<th>Word Length</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOWs</td>
<td>23</td>
<td>52</td>
<td>110</td>
<td>208</td>
<td>318</td>
</tr>
<tr>
<td>Complement/Reverse Complement Pairs</td>
<td>28</td>
<td>53</td>
<td>48</td>
<td>47</td>
<td>62</td>
</tr>
<tr>
<td>Total Possible Words</td>
<td>64</td>
<td>256</td>
<td>1024</td>
<td>4096</td>
<td>16384</td>
</tr>
<tr>
<td>Significance</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

### Maximal Loops

<table>
<thead>
<tr>
<th>Word Length</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOWs</td>
<td>78</td>
<td>201</td>
<td>420</td>
<td>821</td>
<td>1331</td>
</tr>
<tr>
<td>Complement/Reverse Complement Pairs</td>
<td>95</td>
<td>184</td>
<td>241</td>
<td>180</td>
<td>134</td>
</tr>
<tr>
<td>Total Possible Words</td>
<td>256</td>
<td>1024</td>
<td>4096</td>
<td>16384</td>
<td>65536</td>
</tr>
<tr>
<td>Significance</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

Table 3-2 Complements and Reverse Complements within the set of Frequently Occurring Words.
The Sequences of 3-letter Apexes of MicroRNA Hairpin loops are Biased Towards Complement/Reverse Complement Pairs.

Complement and Reverse Complement Triples in order of descending frequency in microRNA loops. Complementary (Top graph) and reverse-complementary (Bottom graph) pairs of triplets are coloured identically.

Several pieces of evidence in the literature support the relevance of this finding, showing that loop-to-loop binding between stem-loop structures occurs naturally and is an important regulatory step in several processes. In Hiv-1, tat RNA makes a loop-loop kissing structure that
is involved in viral replication[227]. This dimerization increases the degree of recombination between the 2 copies of the viral genome during the replication process.

Figure 3-12 Kissing Complexes involved in HIV-1 replication.
Proceedings of the National Academy of Science, 93, Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C., Marquet, R., A loop-loop “kissing” complex is the essential part of the dimer linkage of genomic HIV-1 RNA., Pages 5572-7, [227], Copyright 1996 National Academy of Sciences, U.S.A.

Kissing complexes have also been described in R1 plasmids as the rate-limiting step in controlling copy number[228, 229]. This is done by a loop-loop interaction between an anti-sense RNA (CopA) and the RepA mRNA (Figure 3-13).
3.3.6 Loop Structure-based Classification

3.3.6.1 MicroRNA Loops can be subdivided into 5 classes based on terminal-loop structure

We were first able to divide all microRNAs into 5 broad classes based on their loop structure (Table 3-3). Starting with the 1424 microRNA hairpin structures in mirbase version 17 we find that 35% of microRNA stem-loops are perfect loops. These microRNAs have the classical microRNA structure. They have a double-stranded RNA stem and a loop without any bases inclined to form bonds. Fifty-seven percent of microRNA hairpins form what we are calling occluded loops, with structures similar to the first group, except that they have loop sequences that have one or more opposite base-pair complements, resulting in bond formation within the loop. Zhang and Zeng postulate that since restraining the terminal loop through the addition of

---

**Figure 3-13 Loop-Loop Kissing Complexes in R1 replication.**
Reprinted from The Journal of Biological Chemistry[229], copyright 1997.
bonds result in fewer processed microRNA transcripts, it’s likely that this increase in stability affects microRNA processing at both the Drosha and Dicer steps[97].

<table>
<thead>
<tr>
<th>MicroRNA</th>
<th>Loop Class</th>
<th>Number Of Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-mir-10a</td>
<td>Perfect</td>
<td>532</td>
</tr>
<tr>
<td>hsa-mir-21</td>
<td>Occluded</td>
<td>871</td>
</tr>
<tr>
<td>hsa-mir-762</td>
<td>Closed</td>
<td>18</td>
</tr>
<tr>
<td>hsa-mir-34c</td>
<td>Small Loop</td>
<td>112</td>
</tr>
<tr>
<td>hsa-mir-331</td>
<td>Runover</td>
<td>19</td>
</tr>
</tbody>
</table>
Table 3-3 Classes of MicroRNA Structures

We speculate that perhaps this increased stability of occluded loops could play a role in the identification and maintenance of microRNA stem-loop integrity of microRNAs that are destined for export from the cell. The small loop class is a subset of the perfect loop class, but with loop size smaller than 5 base pairs. Examples of notable members of all loop classes and their frequencies are indicated in Table 3-3. The 19-member runover loop class is defined by the location of the mature microRNA sequence with respect the beginning of the Minimal Loop. In runover loops at least one basepair of the mature sequence is within the loop structure itself – perhaps having implications for the processing steps of these microRNAs – particularly for runover loops that have no bonded bases, making their Maximal and Minimal Loops identical and leaving no doubt as to the commencement of the loop structure. Finally there are closed loops – which are barely identifiable as a stem-loop structure since they lack the terminal loop entirely.

Figure 3-14 expands on characteristics of microRNA stem loops across the micronome. Panel A shows that Minimal Loops are less than 7 bps long on average, while Maximal Loops are 14 bps long (Panel B) and that the average microRNA has 30 bonds in the entire stem-loop RNA structure (Panel C). Panel D displays the percentage of bonded bases in microRNA loops – showing 2 distinct peaks comprising both perfect loops centered at 0% bondage and occluded loops centered at 40% (which is expanded in Panel E). Panel F shows the distribution of the length of the entire stem-loop structures, with a peak at 40 basepair – indicating that while microRNAs can be characterized by loop type – there is considerable variance surrounding the
length of the stem loop structure of pre-microRNAs ranging from side lengths of 20 bases to over 60 bases - making for an extremely diversified micronome.

**Figure 3-14 Physical Characteristics of MicroRNA Loops**
Distributions of properties of microRNA loops. Panel A: Number of bases within the Minimal Loop, Panel B: Number of bases within the Maximal Loop, Panel C: Number of bonds in the entire stem-loop structure, Panel D: Degree of binding in the loop structure, Panel E: Degree of bonding (occluded class only), Panel F: Length of non-loop microRNA stem
To further explore these classes, we focus on the let-7 family of microRNAs, since it is well conserved and well studied. Figure 3-15 A shows the mapping of the let-7 family into classes of terminal loop structures. Despite their similarities by seed sequences, this well-conserved family is extremely heterogeneous when it comes to the structures of their loops. Figure 3-15 B shows the results of clustering by loop structure using ClustalW [214, 215]. Unexpectedly, the loop sequences of members of the family with identical seed sequences but which are mapped to different regions of the genome (let-7a-1, let-7a-2, let-7a-3, let-7f-1, let-7f-2), do not cluster together, and in some cases are more closely related to other members of the family than they are to their own.
Figure 3-15 Examining MicroRNA Loop Structure in the Human Let-7 Family
Although the let-7 family members have identical seed sequences and similar mature microRNA sequences, they have distinct loop characteristics.
A global analysis into the associations of microRNA loop classes with various microRNA characteristics reveals several class-based trends, which might lend some clues as to their functions Figure 3-16. Although there are 1.64 times as many occluded microRNAs in the micronome as perfect microRNAs (Panel A) – there are 2.7 times as many experimentally validated interactions (Panel B) between microRNAs with occluded loops as there are with microRNAs with perfect loops[171], perhaps suggesting a greater effect or involvement of microRNAs with occluded loops in cellular processes, although we recognize the potential for simple experimental bias. Since they do not have a statistically significantly different number of
targets (Panel C), this effect is more likely due to the magnitude of importance of the genes that they target making them more likely to be picked out of a screen for further study. There is a small but significant difference ($p < 0.01$) in the number of transcripts targeted by these two groups of microRNAs (Panel D) lending support to the idea that microRNAs are somehow influential in the realm of splicing, an idea already hypothesized by others[230, 231]. We also noted that occluded microRNAs target 3’ UTRs in a more complex manner, with 116 targets of occluded genes having 4 or more binding sites within one UTR, compared to perfect loops in an analysis of PITA-predicted binding sites (Panel E) [30]. These distinct characteristics of different classes of microRNAs point to the possibility that the micronome is indeed very heterogeneous. And just how our understanding of the non-microRNA transcriptome has evolved to the point where we see specialization in tasks for different families of genes, perhaps our understanding of the micronome will undergo such a shift, resulting in recognition of specialization of similar microRNAs from the under the broad umbrella of expression-repressing function that we understand today.
Figure 3-16 Comparison of Perfect and Occluded Loops
3.3.7 Clues to MicroRNA Loop Function

3.3.7.1 Secreted microRNAs share similar loop motifs

Recent studies have established that not only do microRNAs make their way out of cells by way of exosomes and vesicles, differential plasma microRNA-expression levels between diseased and healthy patients point to the possibility of blood-level testing for the presence or absence of cancer and other autoimmune diseases [120, 127, 128]. Since - contrary to prokaryotic microRNAs - eukaryotic microRNA loops are preserved into the cytoplasm[202], and some studies indicate the presence of pre-microRNAs (still containing loops) in the vesicles of mesenchymal stem cell media [128], we sought to determine whether microRNA loops might play a role in this secretion process, perhaps by providing a structure- or sequence-based binding site to increase stability and facilitate recognition for transport out of the cell.

We turned our study to microRNAs contained within microparticles – cholesterol-rich phospholipid vesicles, present in the conditioned media of human embryonic-derived mesenchymal stem cells (MSCs). Chen et al. showed that some pre-microRNAs are present, at higher concentrations outside the cells within lipid vesicles than within the cells, at significantly higher concentrations than their mature microRNA counterparts [128]. Importantly, components of the RISC complex were lacking in the identified exosome cargo – supporting the possibility of a transport theory for the pre-microRNAs.
Using data from this paper – we took 27 mapped microRNA hairpins with at least a 2-fold expression increase in the media compared to the cell cytoplasm – to examine the possibility of loop sequence involvement as a signal of export from the cell. Comparing the loop sequences of this set – we were able to identify several 4-letter motifs that recur in the loop regions of members of this microRNA set. Figure 3-17 displays the relationships between these MSC-secreted microRNAs in a network diagram. The top three motifs: GGGC (involving 12/27 nodes), UGUC (9/27) and UGGG (7/27) are significantly enriched in this set of secreted microRNAs (p < 2.3x10^{-10}, p < 7.6x10^{-6}, 7.79x10^{-5} respectively, Hypergeometric test). Unlike the AU-rich loop sequences characterizing the overall micronome loop sequences, this set of microRNAs secreted from MSCs is GC rich, suggesting a mesenchymal stem cell secretion specific motif, making them a unique and independent set. The GGGC and UGGG motifs are particularly interesting due to their similarity to the GGG sequence that KSRP has been shown to bind in microRNA terminal loops. Perhaps this GGG motif, which has been shown to promote Dicer-mediated processing of the pre-microRNA transcripts, is necessary for microRNAs which leave their cell of origin to be processed at a secondary site – to kick start the processing pathway in a new environment.

Interestingly, a large number of these secreted microRNAs are identified as occluded. Of these 27 microRNAs, 23 have occluded loops (p < 0.02), supporting the hypothesis that bonded terminal loops result in increased hairpin stability and decreased microRNA processing and expression [97]. We further hypothesize that this possibly facilitates their retained structure during the transport from the cell, and may even enable us to predict (using structure alone) which microRNAs are most likely secreted into the blood stream.
Figure 3-17 *Network Diagram of Common Loop Motifs from Mesenchymal Secreted MicroRNAs*

Nodes represent microRNAs in the study, edges represent common loop motifs. MicroRNAs with perfect loops are red circles, those with occluded loops are blue triangles.
A gene enrichment analysis of Gene Ontology terms for the target genes of these Mesenchymal Stem Cell secreted microRNAs based on our previously published 3+DB set of interactions has identified 4 broad categories of significant overexpression (Table 3-4) [211, 232, 233].

Intriguingly, 13 GO terms related to activities exterior to the cell turned up in the GO analysis at significance levels $p < 1 \times 10^{-9}$, including cell adhesion, cell-cell signalling, blood circulation and response to external stimulus. This suggests that microRNAs secreted from cells are targeting mRNA involved in regulating the external environment of the cell, although whether they are meant to target the donor cell or the receptor cell, we can’t be certain. This may imply that the secreted microRNAs travel to distant locations to help regulate the external environment of the recipient cells preparing the microenvironment, priming it for future secretory events and signals. Six different developmental processes are overrepresented by targets of secreted microRNAs at $p < 1 \times 10^{-16}$ and five different immune response terms at $p < 1 \times 10^{-6}$. In terms of GO Molecular Functions, receptor binding and activity, as well as protein binding and extracellular matrix structural constituents were all enriched at levels less than $p < 1 \times 10^{-15}$. A final confirmation of the importance of the activities of targets of secreted microRNAs to elements exterior to the cell, enrichment was found in GO cellular components extracellular matrix and region ($p < 1 \times 10^{-27}$), and an underrepresentation of elements classified as being intracellular ($p < 1 \times 10^{-3}$).
## GO BIOLOGICAL PROCESSES

<table>
<thead>
<tr>
<th>Activities Exterior to the Cell</th>
<th>Actual</th>
<th>Expected</th>
<th>Direction</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell adhesion</td>
<td>135</td>
<td>43.39</td>
<td>+</td>
<td>6.18E-33</td>
</tr>
<tr>
<td>cell-cell adhesion</td>
<td>98</td>
<td>26.81</td>
<td>+</td>
<td>1.33E-28</td>
</tr>
<tr>
<td>signal transduction</td>
<td>247</td>
<td>132.69</td>
<td>+</td>
<td>2.20E-27</td>
</tr>
<tr>
<td>cell communication</td>
<td>252</td>
<td>138.23</td>
<td>+</td>
<td>9.78E-27</td>
</tr>
<tr>
<td>cell-cell signalling</td>
<td>117</td>
<td>42.32</td>
<td>+</td>
<td>6.42E-24</td>
</tr>
<tr>
<td>response to stimulus</td>
<td>126</td>
<td>53.77</td>
<td>+</td>
<td>8.97E-20</td>
</tr>
<tr>
<td>cell surface receptor linked signal transduction</td>
<td>146</td>
<td>68.43</td>
<td>+</td>
<td>1.99E-19</td>
</tr>
<tr>
<td>cell-matrix adhesion</td>
<td>39</td>
<td>6.5</td>
<td>+</td>
<td>1.56E-18</td>
</tr>
<tr>
<td>response to external stimulus</td>
<td>35</td>
<td>9.75</td>
<td>+</td>
<td>1.69E-10</td>
</tr>
<tr>
<td>blood coagulation</td>
<td>35</td>
<td>9.75</td>
<td>+</td>
<td>1.69E-10</td>
</tr>
<tr>
<td>blood circulation</td>
<td>27</td>
<td>6.76</td>
<td>+</td>
<td>2.40E-09</td>
</tr>
</tbody>
</table>

| Developmental Processes                         |        |          |           |            |
| system development                              | 160    | 66.51    | +         | 5.63E-27   |
| ectoderm development                            | 122    | 48.19    | +         | 8.47E-22   |
| developmental process                           | 184    | 95.39    | +         | 2.21E-20   |
| mesoderm development                            | 119    | 49.74    | +         | 2.67E-19   |
| skeletal system development                     | 57     | 16.25    | +         | 5.71E-16   |
| nervous system development                      | 106    | 42.69    | +         | 3.17E-18   |

| Immune Response                                 |        |          |           |            |
| immune system process                           | 181    | 78.4     | +         | 5.90E-29   |
| complement activation                           | 26     | 4.87     | +         | 1.16E-11   |
| response to external stimulus                   | 35     | 9.75     | +         | 1.69E-10   |
| macrophage activation                           | 34     | 9.42     | +         | 2.69E-10   |
3.3.7.2 MicroRNA Loops contain tissue-specific motifs

Since we now know that microRNAs are not cell-bound, we were interested to see if the loop sequences might carry signals relating to the particular tissues in which they are expressed.

Using data from the Mammalian microRNA expression atlas [194], we took expression data for all microRNAs which were sequenced and present in their panel of normal tissues more than twice. We then counted the occurrences of all the common motifs of microRNAs that were shown to be expressed in each tissue and scaled by the number of expressed microRNAs in the
tissue to avoid giving microRNA-rich tissues an advantage. Figure 3-18 is a heat map showing the distribution of microRNA loop motifs represented in each tissue.

It is particularly interesting to note that body tissues with similar functions clustered together. The motif profiles of normal heart, pancreas, placenta, liver and spleen formed one cluster, ovary, thyroid, testes, uterus, pituitary and cerebellum clustered together and midbrain and frontal cortex formed a subcluster of the latter set. The hippocampus had the most distinct microRNA loop motif profile and did not cluster with any other tissue. With so much hormone related signalling going on in the body, the prospect of microRNA involvement by secretion and homing guided by microRNA loop motifs as secondary site recognition signals is an exciting prospect and further research needs to be done to investigate this preliminary finding.
Figure 3-18 *Heatmap of Normal Tissues Clustered by Common MicroRNA Loop Motif*
Based on characteristics of the expression of microRNAs by loop motif, hormonally-driven tissues cluster together.
3.3.7.3 KSRP-Dependent microRNAs share Loop motifs

Trabucci et al. have described how KSRP binds to terminal loops of microRNA precursors to promote their maturation through its involvement with both the Drosha and Dicer complexes (Figure 3-19). The 12 non-uniquely identified microRNAs within their paper map to 25 microRNA hairpins, twenty of which have an occluded loop structure (p < 0.05, Hypergeometric test) (Figure 3-20) [119].

Figure 3-19 KSRP-dependent MicroRNAs.
Figure 3-20 MicroRNA Loop Structures in KSRP-dependent MicroRNAs
KSRP-dependent microRNAs tend to have occluded loop structures.

This argues microRNA loop structures playing a role in a subset of microRNAs specialized into a family-like group, which are all acted upon by KSRP during microRNA biogenesis, perhaps to go on to perform a similar function within the cell.
3.3.7.4 Lung cancer signalling may involve microRNA loop sequence and structure to facilitate vesicular export of microRNAs from tumour cells

Finally, to investigate whether microRNA loop structure might play a role in the selected export of microRNAs within the cell in a lung cancer setting, we began working on a model to determine how the presence of microRNAs in the blood of tumour patients facilitates primary tumor migration to different organs and helps forming metastases.

Through the integration of fourteen microRNA microarray studies in lung cancer (Table 3-5), we have been able to identify patterns in microRNA expression over an approximated timeline from primary tumour expression, lymph node expression, to primary tumours associated with brain metastases and as well as in studies examining serum microRNA levels at similar time points.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Tissue Source</th>
<th>Timing</th>
<th>Predictor</th>
<th>Paper</th>
<th>Subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>Primary</td>
<td>Cancer Onset</td>
<td>Primary Tumour</td>
<td>Volinia et al., 2006</td>
<td>All Lung</td>
</tr>
<tr>
<td>Set 2</td>
<td>Primary</td>
<td>Cancer Onset</td>
<td>Primary Tumour</td>
<td>Tan et al., 2011</td>
<td>SCC</td>
</tr>
<tr>
<td>Set 3</td>
<td>Primary</td>
<td>Cancer Onset</td>
<td>Primary Tumour</td>
<td>Shen et al., 2011</td>
<td>NSCLC</td>
</tr>
<tr>
<td>Set 4</td>
<td>Primary</td>
<td>Pre-cancer Onset</td>
<td>Primary Tumour</td>
<td>Boeri et al., 2011</td>
<td>All Lung</td>
</tr>
</tbody>
</table>
Table 3-5 Lung Cancer MicroRNA Studies

The literature has presented several clues as to the function of blood-borne microRNAs. Unfortunately there are initial conflicting reports over the forms of microRNAs in serum, whether it is the mature form or pre-microRNAs which are contained within microvesicles, which lack other RISC components. This lack of a functional unit -- as well as an anticipated
lack of their target mRNA within serum, although not necessarily within microvesicles -- implies that microRNAs are being secreted to travel to other cells to influence gene expression at distant sites.

3.3.7.4.1 Cell-bound microRNAs in Lung Cancers Tend to harbour Perfect Loops

Analyses of the temporal relationships derived from Table 3-5 are ongoing and will be the subject of a future paper. However, a key element of this data which is pertinent to involvement of microRNA terminal loops, is the structure of microRNAs shown to be secreted by the studies displayed above. Table 3-6 divides microRNAs from these studies into groups based on evidence of their secretion.

Again, this trend supports the notion that increased stability in microRNA loops as provided by

<table>
<thead>
<tr>
<th>Table 3-6 Lung Cancer MicroRNAs by Secretary Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only sixteen of the microRNAs mapped in the first 2 groups of Table 3-6 have a perfect loop structure, the rest harbour occluded loops. Eleven of these sixteen perfect loops map to the non-secretory group, an unlikely, but not significant occurrence (p&lt;0.25 Hypergeometric test).</td>
</tr>
</tbody>
</table>


additional bonded bases in occluded loop structures may facilitate the recognition and continued
stability of certain microRNAs during the export process.

3.3.7.4.2  Loop motifs are associated with secreted microRNAs in Lung Cancers
A second manner in which microRNA terminal loops may be implicated in cellular export comes
from the examination of microRNA terminal loop sequences of groups of microRNAs known to
be secreted. Comparing the set of microRNAs that show no evidence of leaving the cell, and
contrasting them with the 2 groups that do, we searched for common microRNA loop motifs that
may be implicated in signalling across both Maximal and Minimal loop structures. Figure 3-21
plots the proportional number of occurrences of motifs between the 2 groups of microRNAs by
intra-group rank.
Motifs falling furthest from the diagonal in the upper half of the figure, as well as along the y-axis are overrepresented in cell-bound microRNAs. MicroRNAs overrepresented below the midline, as well as along the x-axis are overrepresented in microRNAs secreted from the cell.

Within the group of motifs overrepresented in secreted microRNA loops, there are many adenine and uracil rich motifs. Importantly, out of 256 possible 4-letter motifs, there are 16 which contain adenine and uracil exclusively. Twenty-seven motifs along the x-axis are seen exclusively in the secreted loop set, and of these 7 are composed of adenine and uracil.
exclusively ($p<0.0003$ Hypergeometric test). This AU enrichment in the loops of secreted microRNAs argues against the idea that terminal loops are AU-rich due only to their tendency to forms loops, and argues that AU-rich loops may indeed have a function at the molecular level, perhaps in a setting-specific manner, as MSC-secreted microRNAs have been shown earlier to be a GC-rich set.

3.4 Conclusions

This first global look at microRNA loop sequences as more than simply a structural component to the pre-microRNA stem loop has unearthed some interesting findings. When considering that the group of microRNAs as a whole may be further subdivided into functional classes based on microRNA structure, we show that it is a distinct possibility that microRNAs are a group of genes with heterogeneous functions, rather than the catch-all expression-repressors as they have been known to date. With the recent finding that microRNAs are no longer cell-bound and can be secreted into the bloodstream, we find characteristics of the microRNA structure, namely the degree of binding in the terminal loop, to be associated with those microRNAs known the be secreted. We speculate that increased terminal loop binding – a phenomenon that we have dubbed an occluded loop – might increase stability for microRNAs fated to be secreted. Not only do microRNAs have frequently occurring motif sequences – occurring at rates much higher than expected by chance – these motifs are AU-rich and also have complements and reverse complements commonly occurring in microRNA loops. These AU-rich motifs could serve as binding sites for RNA binding proteins, or they may be related to secretion signals for
microRNAs that are destined to leave the cell. We found that microRNA loop motifs are also related to the expression patterns of the tissues in which microRNAs are expressed, suggesting the possibility that microRNAs that are secreted need a recognition sequence to be taken up by their secondary cellular site. Our research points to the need for further investigation into the function of microRNA terminal loops to perhaps tell us more about the functions of microRNAs on a systemic level, since the current model has increased significantly in complexity now that they are known to be secreted.

3.5 Materials and Methods

All human precursor microRNA (hairpins) and mature microRNA sequences are taken from mirbase version 17 (hairpin_17.fa and mature_17.fa files) [25, 26, 234]. It should be noted that the mature microRNA file and the hairpin file contain different sets of entities, since some hairpins may contain 2 microRNAs. Structures and their matching mature sequences were pulled using wget from individual microRNA pages since download files containing structure information were not available.

Analysis of microRNA loop classes and characteristics was done using the R statistical software package[235].

In comparing perfect and occluded loops, confirmed interactions were drawn from Tarbase [171], comparisons of number of binding sites for a particular microRNA in a particular UTR used data from PITA [30], the average number of targeted genes used data from mirDIP,
comparisons on number of transcripts per targeted gene used UCSC data from tables hg19.knownGene and crossed ref with the HUGO gene name from kgXref. Galaxy was used for some basic data table manipulations. Secreted microRNAs from MSCs were drawn from Table 1 from [128], and mapped KSRP-dependent microRNAs were drawn from Figure 4 from [119]. and compared by hypergeometric test.

We created 1000 simulated loop micronomes – with the same number of same-sized loops as the actual microRNA-loop space, but randomly permuting the base pairs. We then compared the number of occurrences of motifs ranging in size from 4-8 bases for Maximal Loops and 3-7 for Minimal Loops in the actual loop-space to the number of the identical motifs in the simulated loop spaces. We selected motifs that were present in the generated micronomes at levels greater than the actual micronome less than 10 times, giving us a p<0.01 significance test. This set of loop motifs we refer to as “Frequently occurring words” (FOWs).

Test for the significance of microRNA Frequently Occurring Words were based on 2 sets of data – 1000 artificially constructed micronomes -- and the motifs found therein, as well as from the expected number of words of any given size in the micronome (Figures 3-8 A and B).

Figures comparing triplets at the apex of microRNAs loops were drawn from mirbase files and were made with Prism.

All data surrounding complements and reverse complements of FOWs and their FOW counterparts were generated by programs written in Python using the Bio.Seq module.
The infographic displaying the frequencies of 6-letter microRNA words within the micronome (Figure 3-10) was created using python and Wordle [220].

All clustering of alignment sequences was done using publicly available alignment software ClustalW2 (version 2) using default settings [214, 215].

Comparisons of microRNA prediction targets for microRNAs with various seed compositions was done using mirDIP data [211].

The clustered tree in Figure 3-5 was made using Clustal W2 with default settings, R statistical package for file format conversion and Dendroscope (version 2) to edit the final circular cladogram [214-216].

To determine the percentage of common seeds between sets of microRNAs with a matching minimal/maximal loop, we used the R statistical package and considered all sets of microRNAs sharing an FOW and examined their seed sequences.

The microRNA network diagram in Figure 3-17 was made in NAViGaTOR (version 2.2.1) [150].

Figure 3-18 was done in R using the heatmap function from the stats package in R with hierarchical cluster analysis, using eucliddean distance and complete linkage as parameters [235]. Using data from the Mammalian microRNA expression atlas (2007) [194], we took expression data for all microRNAs which were sequenced and present in their panel of normal tissues more than twice. We then counted the occurrences of all the common motifs of
microRNAs that were shown to be expressed in each tissue and scaled by the number of expressed microRNAs in the tissue to avoid giving microRNA-rich tissues an advantage.

Gene Ontology analyses were done using Panther [232, 233].

Figure 3-21 was plotted by taking all microRNAs from Table 3-6 and classifying them into 2 sets -- secreted and non-secreted -- based on their study of origin. Within these two groups, all pairs of microRNAs were considered for common loop motifs. Counts of all the unique common maximal and minimal loop motifs between intra-group pairs were ranked and assigned a rank ratio to produce Figure 3-21.

3.6 Acknowledgements

Ontario Research Fund (GL2-01-030)

Canada Foundation for Innovation (CFI #12301 and CFI #203383)

Canada Foundation for Innovation (CFI #29272)

Canada Research Chair Program 203373 and 225404

IJ is supported in part by the Canada Research Chair Program.

This research was funded in part by the Ontario Ministry of Health and Long Term Care. The views expressed do not necessarily reflect those of the OMOHLTC.
Chapter 4

Discussion

Greater understanding of microRNA biogenesis, structures and targets could enable a more in-depth understanding of the vast network of control that microRNAs exert over the cell. Since their discovery, microRNA research has suffered from several key weaknesses; the inability to study isolated microRNA species in the processing pathway, the early lack of protocols to solidly prove microRNA binding *in vitro* and *in vivo* – producing a small number of well-confirmed targets from which bioinformatics algorithms can work, and seed-centric bioinformatics algorithms which achieve high precision with low recall or vice versa. mirDIP was initially conceived to combine existing bioinformatics resources to see if a combination of sources might enable integrative analyses and to increase the flexibility of, and confidence in target predictions. We have gone from studying seeds and target binding to studying microRNA loops, a structural component, to attempt to address the question about whether there remain outstanding microRNA species involved in function.
4.1 Navigating the Micronome

4.1.1 mirDIP

mirDIP is useful not only in the context of combining microRNA prediction databases to provide a higher confidence in predicted targets in a task-specific manner, but also as a way to interrogate the combinatorial action of microRNAs on a single target, as well as the common mRNA targets of one or more microRNAs. At the time of publication, mirDIP was the only live resource allowing users to examine co-targeting-microRNAs of multiple genes of interest, which allows users to address the true network nature of microRNA interactions.

4.1.2 Combinatorial MicroRNA Action

An interesting and likely theory of microRNA regulation of mRNA lies in the effects of combinatorial microRNA functions. Since 3’ UTRs are long and contain many binding sites, it is not unreasonable to expect that many microRNAs might co-bind to produce a particular effect. The actual logistics of multi-RISC 3’ UTR binding must investigate the size ratios of RISC complexes compared to the length of the UTR to make sure that it is physically possible. The length of a microRNA binding site is ~7 seeds long, with the mature microRNA spanning ~22 basepairs. The average weight of a nucleotide is 330 Daltons. A 10 kDa protein is roughly the equivalent in size to 270 bp of RNA. The components of the RISC complex are much larger – Argonauit proteins – AGO2 in particular is ~100 kDa in weight, this is equivalent in size to 2700 RNA base pairs. So the argonaute element of the RISC alone is about 122 times larger than the...
microRNA binding site. This would mean that the RISC complex would overshadow hundreds of bases past the edge of the microRNA seed sequence (it would be thousands, but we take into account some protein folding) and would impede subsequent RISC binding to sites closer than this. This follows along with Grimson’s observations that inter-binding site distances up to 500 base pairs are more likely to be provide functional synergy. It may also help to explain the ribosomal shadow effect observed in the same paper – with RISC complexes not able to fit close to the stop codon where ribosomes fall off [27]. Since there will be obvious patterns of which microRNA binding sites are available at the same time due to physical size restrictions of the RISC along the 3’ UTR – we may even find that there are sets of microRNA biding sites -- all just shifted a given RISC-sized spacing – that define several discrete states for the mRNA transcript. Work by Rigoutsos et al. documenting the existence of pyknons, blocks of repetitive motifs in non coding regions of the human genome, contributes interesting ideas to the organization of microRNA binding within UTRs, finding that pyknons tend to occur 22 nucleotides apart and occur in varying combinations in UTR regions of almost all human genes[236].

Although this combinatorial setting makes for a more complicated regulatory situation to be deciphered and understood, it does make sense in a dose dependency context – with more microRNAs bound to a UTR, the effects of their repression are likely to be larger. It also helps to explain the large variance in 3’ UTR length across the genome.
4.1.3 Pathway regulation at the RNA level

We have shown that the case for pathway regulation at the mRNA level is plausible and worthy of further work in vitro and in vivo. Since genes involved in the same signalling pathway are more likely to be targeted by a similar set of microRNAs, they lend themselves to control at the sub-protein level, possibly following the competing endogenous RNA theory.

Our work [211] relates to recent theories examining competing endogenous RNA (ceRNA), and supports the idea of combinatorial action [237]. Since many genes share microRNAs, the idea that a set number of microRNAs exist within the cell, and that they flow between multiple targets in a concerted way, leads us to the analogy of a balance – where RNA-RNA communication is occurring via the shared microRNA profile of sets of genes. This theory includes a role for pseudogenes – transcribed RNA elements, which are extremely similar in nature to their protein coding counterparts, but often lacking introns, yet retaining their 3’ UTR sequences. This scenario implies self-regulating RNAs through the sharing of microRNAs with other protein coding genes - often in the same regulatory pathways as we have shown, as well as with pseudogenes [238]. Based on their assumption of a constant number of microRNAs, any perturbation in the expression of a gene within the network upsets the balance of ceRNAs and hence directly alters the expression of genes with shared binding sites.

Studies in the last decade have uncovered an abundance of transcribed but untranslated loci within the genome, termed long non-coding RNA (ncRNA). Long ncRNA tend not to be conserved – although they do have conserved regions - are thought to be able to evolve more quickly than protein coding transcripts and are present in the order of tens of thousands of
transcripts in the genome, with some studies remarking that only 20% of transcription across the genome is of coding mRNA [239, 240]. Long ncRNAs have been linked to both transcriptional and post-transcriptional regulation, splicing, chromosomal inactivation and imprinting [241]. Since long ncRNAs exert some of their effects via complementary binding to other transcripts, we cannot rule out some sort of interplay between long ncRNAs and microRNAs in this new world of RNA-RNA interactions, in addition to the ceRNA scenario proposed above.

4.1.4 Classification of microRNAs by Targeting Characteristics

This work has also drawn the lines to divide microRNAs into an early classification scheme related to their targeting behaviours. We noted that some microRNAs have a broad reach into pathways within the cell, hitting targets within all canonical signalling pathways identified. These Universe microRNAs have the potential to be master regulators within the cell, with the ability to influence quick and effective changes in patterns of gene expression across the much of the entire scope of cellular signalling activities. They are well studied, have many targets and have been implicated in many cancers. Their counterparts, intrapathway microRNAs – are much more specialized, as seen through their involvement in fewer canonical signalling pathways, limited tissue expression and can perhaps be described as fine-tuning agents to ensure coordinated intra-pathway signalling is maintained.
Contributions and Limitations

While mirDIP integrates microRNA prediction databases to provide tailored microRNA target prediction to improve accuracy on a task-based basis, our study does suffer from some limitations. While integration of multiple sources improves the balance between precision and recall based on the *in vitro* data available for comparison, the overlap between targets predicted by the databases were smaller than we had anticipated. This led us to the task-based mirDIP selection criteria on our portal – to maximize the types of returned results most relevant to the user’s query.

Another important aspect of task-based microRNA prediction that we have discussed earlier, but have not implemented in mirDIP due to the large scale of the undertaking, is a filtering of microRNA results based on tissue-specificity criteria. Since the successful search for a microRNA target is largely dependent on its expression within a tissue, it would be convenient to provide some level of tissue-specific filtering to suggest ideal tissues for wet lab scientists to work in, which express both the target and the microRNA. This would be even more useful at the combinatorial level when an interaction between 2 or more microRNAs is suspected.
4.2 MicroRNA Loops

4.2.1 A Broader role for microRNAs through the Recognition of Distinct Classes based on Loop Structures

Initially, microRNAs were considered one class of small non-coding genes. With the current existence of more than 1700 microRNAs, we propose a change in this thinking. Perhaps microRNAs should no longer be lumped in together and considered to perform similar tasks in similar settings. Although they are distinctly different from the canonical definition of a gene in terms of size, processing, mechanism of action and functional end product, there are nevertheless very diverse subtypes of microRNAs, with different mechanisms, functions and specialized effects across diverse cell and tissue types. In this work we have alluded to two divisions of microRNAs: by function and by structure. The latter subdivides microRNAs by loop structure – identifying a group of microRNAs with loops harbouring intra-loop bonds – occluded loops. We have implicated occluded loops in the microRNA secretion process coinciding with others groups’ suggestions that bonding within the microRNA loop provides additional stability, which we suggest may aid in the secretory process. This largest group of microRNA loop sub-type, numbering 871 in the mirbase 17 set of microRNA hairpins, occluded loops have several other characteristics, including being the most studied set and being implicated in KSRP-dependent microRNA processing.

Classification schemes aside, there are many microRNAs in the micronome. They target distinct sets of genes, they have distinct expression profiles, some have been shown to be secreted, some
are able to be processed in the absence of Dicer, some likely implement their translational repression effect differently from others, some come from long stem loops, some from short, some are implicated in cancer development and progression, some are not. While distinct from protein coding genes, microRNAs are also distinct from one another – and they should be treated with such individuality, rather than lumped under the large umbrella of “microRNAs”. Once we have stopped treating the set of microRNAs as a homogenous population, we shall begin to have a better understanding of the true reach of their functions within the cell. Perhaps we are missing the big picture of what microRNAs do if we limit them to all performing the same tasks.

4.2.2 Emerging Roles for MicroRNAs as Master Regulators

With the discovery that microRNAs can be selectively secreted from cells, within exosomes or vesicles to travel to distant sites, there are many reasons to be excited, and a million new questions to be posed. Most obviously, the use of microRNA as serum biomarkers is a huge opportunity, since it presents many advantages over current diagnostic technologies.

MicroRNAs in serum – which have been shown to be indicative of disease, in several studies – present huge advantages over current biomarkers and lend themselves quite easily to routine screening for various cancers, via non-invasive blood test. Not only are they easier to manage in the laboratory than proteins – microRNA hairpins are more stable and can be measured more easily by PCR-based techniques rather than proteomics based tests - these tests are lower cost, and may be more accurate due to a lack of known post-processing modification. MicroRNAs are easy to extract, and if they are as indicative of disease as early studies suggest – the world of
diagnostics may be about to turn a corner. There is also potential for improvements in treatment-based technology. If serum microRNAs do exert strong systemic signalling effects between primary and metastatic tumours, the potential for serum filtering in addition to surgery and chemotherapy presents another avenue to interfere with tumour progression through signalling inhibition.

The idea that microRNAs do exit the cell, introduces a whole new level of microRNA regulation and complexity – just when we thought that we were getting our heads around the idea of multiple microRNAs co-targeting one 3’ UTR. Why would microRNAs be synthesized in one cell to travel to a receptor cell if their purpose is solely translational repression?

Some of our data show that the set of secreted microRNAs in mesenchymal stem cells target genes enriched in the external environment of the cell. Since the vesicles in which microRNA travel lack components of the RISC, this leads us to conclude that the secreted microRNAs are internalized by the donor cell, incorporated into the RISC at this distant site and then affect the external environment through their target binding and repression of genes that would normally have been secreted or playing an extracellular role, by way of repression.

4.2.3 Loops as recognition sequences

Our research has turned up the possibility that microRNA loops contain motifs that are not only targets for RNA binding proteins related to their biogenesis, but also that relate to their ability to be secreted from the cell. While this is obviously preliminary data that must be tested in vitro,
it’s an interesting issue – if microRNAs are being secreted from cells, in vesicles – how are they identified for secretion? And how are they identifying their final uptake site? Is there some sort of vesicular recognition sequence? Might it be possible that microRNAs can be trans-vesicular – meaning that their stem-loops span the lipid membrane encapsulating the vesicles, leaving the loop portion as an external recognition component? This would explain the tissue-specificity seen in microRNA loop motif sequences and also implies that RNA could work as a recognition and perhaps as a signalling molecule, which is not the standard manner in which RNA molecules have been known to operate, but which cannot be ruled out simply due to lack of consideration, nonetheless.

4.3 Future Directions

4.3.1 MicroRNA Organization

In our view – the microRNA model of biogenesis and repression is lacking a crucial step – one allowing microRNAs to self-organize along the UTR where they will soon bind. Since microRNA action is swift and effective, it seems unlikely that a model of stochastic chaos can explain the mechanism in which microRNAs find their way to the UTR on which they bind in a particular circumstance. Particularly since microRNAs are known to be keen to bind to many different mRNA UTRs. MicroRNAs have been implicated in developmental processes – causing rapid changes in gene expression to change the state of a particular cell as it passes from one stage of development to the next, and it is nearly impossible to accept that the timing for
microRNA binding could be so lax as to depend on such a random model. We think that a model borrowing ideas from transcription factors is a possibility. Transcription factors are often “parked” upstream of their active binding sites until they are required to turn on the transcriptional process. A similar model could be explained in the microRNA setting, with pre-microRNAs lining up along the 3’ UTR to be accessed in short order, processed by dicer and incorporated into the RISC. Such a model would depend on dicer processing proximal to the 3’ UTR – and such an event has not been explored nor ruled out, but might explain the time-efficient manner in which microRNAs seem to act. Such a model would also involve a role for microRNA loop sequences in transient 3’ UTR binding to maintain microRNA proximity to their binding site and could be related to pyknon sequences described be Rigoutsos et al.[236]. The sequence similarity between microRNA loops and 3’ UTR sequences upstream of microRNA binding sites remains on our list of phenomena to investigate. The plausibility of 5’ UTR binding of microRNAs would make this region an interesting point of comparison in this analysis.

4.3.2 MicroRNA involvement in Translation through tRNA Interactions
An interesting possibility for the function of microRNA loops that we would also like to investigate is their similarity to tRNA loop sequences. The loop-loop binding of “kissing complexes” in the literature leaves the possibility that tRNA and microRNA stem-loops might interact to some degree. Since tRNA are actively involved in the translational process, and
microRNA repression occurs through the translational machinery, though no consensus exists as to how, it is surprising that no investigations have been done into a tRNA-microRNA interaction.

4.3.2.1 \textit{In vitro} Confirmation of Loop Motif Function

To test our hypotheses in vitro, we would like to begin with some simple experiments to determine if it is possible to modulate microRNA secretion from within the cell. Starting with a set of cell-bound microRNAs not harbouring any of the identified secretory motifs, we would like to modulate loop sequences to determine if microRNA secretion status could be altered by changing a loop recognition sequences. Such an experiment would be very telling of the relevance of the secretory microRNA motif finding.

Future work in our lab involves the integration of multiple microRNA expression studies across a panel of lung cancers (Figure 4-1). Studies are arranged as follows: On the left are studies in primary tumours, on the right (with red borders) are studies in serum. Oval shaped studies are in primary tumours and rectangular studies are metastatic cases. Orange lines indicate that a microRNA is overexpressed and grey lines indicate a microRNA being downregulated. Dashed lines indicate that the microRNA expression data is from a primary tumour. We intend to use the information within this figure to map out a progression of microRNA-based lung tumour metastasis using a secretionary microRNA model. We would like to understand the reasons and mechanisms for microRNA secretion from cells and to investigate microRNA loop motifs that might facilitate this export event.
Figure 4-1 Integration of multiple lung cancer studies to investigate microRNA secretion

4.3.3 mirDIP Expansion

There is always room for expansion of our mirDIP portal and we have new ideas which we would like to incorporate to make the mirDIP workflow more seamless and applicable to more microRNA research problems. Firstly, we plan to incorporate detailed information about microRNA loop structures and sequences into mirDIP so that fellow researchers are able to expand on our hypothesis of the importance of loop-structure classes. We would also like to
include a large set of microRNA loop motifs in mirDIP. Often interesting sequences come up in scientific study and there is a lack of a single resource to explain all regulatory and recognition motifs shown to be functional in the genome. We would like to make a small contribution to this set of motifs so that in the future, wet lab scientists will be able formulate microRNA-related hypotheses if information within our set of microRNA loop motifs lends itself to their path of study.

Any integration of microRNA cancer-related data in mirDIP would also be of value to users. Once our study of secretion-based lung cancer associated microRNAs is complete, this data would be a valuable addition. We have also done some preliminary work on SNPs falling into microRNA target sites that would benefit from some expansion and would add further functionality to our mirDIP portal.

4.4 Conclusions

This thesis has made several contributions to the scope of knowledge surrounding the micromome. We have made several class-based observations to delve into this large class of genetic entities that has long been considered a homogeneous class. MicroRNA network complexity is growing at an exponential rate. Early work presented in this thesis expanded on that complexity, but attempted to map it by creating microRNA interaction networks for well-known signalling pathways allowing the visualization of microRNA co-targeting, showing that
transcript regulation within pathways thought to be regulated at only the protein level is a real possibility. Our later work has touched on under examined microRNA loop sequences and examined their relevance in several settings of microRNA secretion. Just the idea that microRNAs can be secreted from cells introduces yet another vast level of microRNA complexity, and we aim to contribute just a piece of the genome-wide puzzle by trying to understand how microRNA loop sequences might be related to such secretory events.
Copyright Acknowledgements

**Figure 1-2 MicroRNA Processing.** MicroRNA processing begins in the nucleus with transcription of the microRNA gene into the pri-mRNA. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer[52], copyright (2006)

**Figure 1-4 Methods of Protein Expression Repression by microRNAs and the RISC complex.** From [74]. Reprinted from Cell, 132, Eulalio, A., Hunztinger, E., Izaurrealde, E., Getting to the root of miRNA-mediated gene silencing, Pages 9-14, Copyright 2006, with permission from Elsevier.

**Figure 1-5 Differential Processing Between Plant and Animal microRNAs.** Reprinted from Cell, 136, Carthew, R.W., Sontheimer, E.J., Origins and Mechanisms of miRNAs and siRNAs, Pages 642-55, Copyright 2009, with permission from Elsevier. [94]

**Figure 1-6 Mir-181a/c-1 function in T-cells is pre-microRNA stem-loop dependent** [95]. Reprinted from PLoS ONE, 3, Liu, G, Min, H., Yue, S., Chen, C.Z., Pre-miRNA loop nucleotides control the distinct activities of mir-181a-1 and mir-181c in early T cell development, 3(10). doi:10.1371/2Fjournal.pone.0003592.

**Figure 3-12 Kissing Complexes involved in HIV-1 replication.** Proceedings of the National Academy of Science, 93, Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C., Marquet, R., A loop-loop “kissing” complex is the essential part of the dimer linkage of genomic HIV-1 RNA., Pages 5572-7, [227], Copyright 1996 National Academy of Sciences, U.S.A

**Figure 3-13 Loop-Loop Kissing Complexes in R1 replication.** Reprinted from The Journal of Biological Chemistry[229], copyright 1997.

**Figure 3-19 KSRP-dependent MicroRNAs.** Reprinted by permission from Macmillan Publishers Ltd: Nature, [119], copyright 2009.
Appendices
Chapter 2 Supplementary Figures

5.1.1 mirDIP – Interface description

Figure 5-1 The key component fields of the mirDIP interface.
1. **Gene Symbol Search Box** – used to search for a gene.

2. **MicroRNA Search Box** – used to search for a microRNA.

3. **Normalized Score Range Search Boxes** – used to search for genes or microRNAs with a particular score range.

4. **Source Filter Box** – used to specify which microRNA prediction database(s) should be used in a search.

5. **Prediction Features Box** – used to select specific microRNA prediction database(s) based on characteristics of the prediction algorithms behind them.

6. **Quality Filter Box** – used for supporting multiple scenarios during searching: a) identifying only highly likely predictions, b) identifying all possible predictions, c) optimized balance of both strategies. See the mirDIP Instructions tab for a more detailed discussion.

7. **Database Occurrences Box** – used to specify how many databases must support all returned predictions.

8. **Intersecting IDs Box** – used when searching for multiple genes or microRNAs at once, to ensure that only an intersection of predictions is returned. Can be used with the Source Filter or Quality Filter Boxes.

9. **Search Button** – used to begin a search.
5.1.2 Using mirDIP

5.1.2.1 Example 1

**Figure 5-2 Searching for PTEN Targets.**
The user is interested in the gene PTEN. They would like to do some in vitro experiments and would like to identify high precision (high certainty) microRNAs that target PTEN.

**Steps:**

Enter PTEN in the Gene Symbol box

Choose the High Precision check box under Quality Filter

Click Search

To sort results by normalized score, click on the black score(norm) heading. Click again to sort in the reverse direction
5.1.2.2 Example 2

Figure 5-3 Searching for Common Targets of PTEN, TP53 and TSC1.
The user is interested in the genes PTEN, TP53 and TSC1. They would like to find microRNAs predicted by microRNA.org that might co-target all three genes.

Steps:
Enter PTEN TP53 and TSC1 in the Gene Symbol Box
Click microRNA.org in the Source Filter Box
Click the Intersecting IDs box
Click Search
To sort results by normalized score, click on the black score(norm) heading. Click again to sort in the reverse direction
5.1.2.3 Example 3

**Figure 5-4 Searching for hsa-mir-22 Targets.**

The user is interested in the miRNA hsa-mir-22. They would like to examine some predicted targets from sources that consider conservation as well as site accessibility.

**Steps:**

Enter hsa-mir-22 in the microRNA

Click Conservation and Site Accessibility in the Prediction Features Box

Click Search

To sort results by normalized score, click on the black score(norm) heading. Click again to sort in the reverse direction
5.1.2.4  Example 4

**Figure 5-5 Searching for TP53 Targets predicted by at least 4 sources.**
The user is interested in the gene TP53. They would like to examine microRNAs predicted to target TP53 that occur in at least 4 microRNA prediction databases.

**Steps:**

Enter TP53 in the Gene Symbol

In the Database Occurrences box, select 4 from the drop down list

Click Search

To sort results by normalized score, click on the black score(norm) heading. Click again to sort in the reverse direction
5.2 Chapter 2 Supplemental Materials

Descriptions of the mirDIP interface and sample mirDIP searches.


Key component fields of the microRNA data integration portal.


Sample mirDIP search for microRNAs targeting one gene of interest, requesting high precision target data.


Sample mirDIP search for microRNAs co-targeting three genes of interest using data from one individual microRNA database.

Sample mirDIP search for targets of one particular microRNA, selecting a microRNA prediction algorithm based on specific algorithm criteria.


Sample mirDIP search for microRNAs targeting one gene of interest using targets predicted by 4 or more microRNA prediction databases.


5.3 Chapter 2 Supporting Methods

The mirDIP portal was developed in Java. We used IBM DB2 ver. 9.5 database back-end and deployed it using IBM WebSphere 6.1 web server (http://ophid.utoronto.ca/mirDIP). Our application server has 32 IBM Power5 processors at 2.3 GHz each with 263 GB of RAM. Our database server has 6 IBM Power6 processors at 4.7GHz each with 62 GB of RAM and 1,500 GB of disk storage on a RAID5 configuration. To ensure high availability, our servers are mirrored at data centers in different geographical locations. In addition to hosting mirDIP, we also host other portals such as I2D and GeneCards.
5.4 Additional Co-authored Publication

The Murine Caecal MicroRNA Signature Depends on the Presence of the Endogenous Microbiota

Natasha Singh¹, Elize A. Shirdel²,³, Levi Waldron², Regan-Heng Zhang¹⁺, Igor Jurisica²,³,⁴, and Elena M. Comelli¹†

¹Department of Nutritional Sciences, Faculty of Medicine, University of Toronto, Ontario, Canada

²Ontario Cancer Institute, Princess Margaret Hospital, University Health Network and the Campbell Family Institute for Cancer Research, Toronto, Ontario, Canada

³Department of Medical Biophysics, University of Toronto, Ontario, Canada

⁴Department of Computer Science, University of Toronto, Ontario, Canada

*Present address: The Biomedical Research Centre, Department of Medical Genetics, Faculty of Medicine, The University of British Columbia, Vancouver, Canada

†Corresponding author

Email addresses:

NS: natasha.singh@utoronto.ca
EAS: elizeshirdel@gmail.com
LW: lwaldron.research@gmail.com
RHZ: rh-zhang@brc.ubc.ca
IJ: juris@ai.utoronto.ca
EMC: elena.comelli@utoronto.ca
Abstract

The intestinal messenger RNA expression signature is affected by the presence and composition of the endogenous microbiota, with effects on host physiology. The intestine is also characterized by a distinctive micronome. However, it is not known if microbes also impact intestinal gene expression epigenetically. We investigated if the murine caecal microRNA expression signature depends on the presence of the microbiota, and the potential implications of this interaction on intestinal barrier function. Three hundred and thirty four microRNAs were detectable in the caecum of germ-free and conventional male mice and 16 were differentially expressed, with samples from the two groups clustering separately based on their expression patterns. Through a combination of computational and gene expression analyses, including the use of our curated list of 527 genes involved in intestinal barrier regulation, 2,755 putative targets of modulated microRNAs were identified, including 34 intestinal barrier-related genes encoding for junctional and mucus layer proteins and involved in immune regulation. This study shows that the endogenous microbiota influences the caecal microRNA expression signature, suggesting that microRNA modulation is another mechanism through which commensal bacteria impact the regulation of the barrier function and intestinal homeostasis. Through microRNAs, the gut microbiota may impinge a much larger number of genes than expected, particularly in diseases where its composition is altered. In this perspective, abnormally expressed microRNAs could be considered as novel therapeutic targets.
Key Words

gut microbiota, microRNA, caecum, intestinal barrier, glycosylation, mouse

Introduction

Humans are complex supra-organisms composed of various endosymbionts that stem from all three domains of life including bacteria, archaea, and eukarya in addition to their own cells. Body habitats that are considered “hot spots” for microbial colonization include the skin, oral cavity, gut and urogenital tract (1,2]. However, in utero, the fetus is completely sterile with colonization beginning postnatally. It is this developmental property that enables for the exploitation of germ-free animals, which are devoid of microbes on or within their body, in order to determine the functional properties of host endogenous microbiota.

In fact, the gastrointestinal (GI) tracts of both humans and conventionally raised mice harbor upwards of $10^{14}$ micro-organisms with levels increasing along the cephalocaudal axis. Temporospatially the organization of bacterial cohorts differs with the largest densities residing in the large intestine, with the caecum acting as a fermentation chamber where upwards of $10^{11}$-$10^{12}$ bacteria/gram luminal contents ferment otherwise indigestible polysaccharides leading to the production of short chain fatty acids. The predominant bacteria groups found in human caecal fluid stems from the *E. coli* and *Lactobacillus-Enterococcus* groups that represent 50% of the caecal bacterial ribosomal Ribonucleic Acid (rRNA) whereas, *Bacteroides* (*Bacteroides*, *Porphyromonas* and *Prevotella* spp.) and *Clostridium* groups (*Clostridium*, *Eubacterium* and *Ruminococcus* spp.) represent 13% of caecal bacterial rRNA (3]. Similarly, the murine caecal microbiota establishes gradually during early postnatal life, and its complexity increases with age
until a mature community is reached by 4-6 weeks of age predominately comprising the *Bacteroides* and *Lactobacillus* genera and the *Clostridium coccoidei* group (4). The microbiota residing along the alimentary canal takes advantage of a continuous supply of nutrients and optimal temperature while playing a pivotal role in host physiology, including nutrient processing and generation, affecting energy homeostasis, education of the immune system, and fortifying the intestinal barrier both directly and indirectly [242]. One of the mechanisms underlying this host-microbe mutualistic relationship is the reciprocal impact of host and microbial cells on each other’s gene expression programs [6,7]. In particular, the endogenous microbiota acts as an environmental factor impacting the expression of thousands genes in the host epithelium [8,9], and this is a function of its composition [9,10]. However, the impact of the microbiota on the intestinal gene expression signature at the messenger RiboNucleic Acid (mRNA) level may have thus far been under-evaluated due to a lack of studies linking gut microbiota to epigenetic changes in gene expression particularly, via micro-RiboNucleic Acids (miRNA).

MiRNAs are 20-22 nucleotide, single-stranded, non-coding RNA molecules involved in post-transcriptional gene regulation. Nascent miRNA exist as large hairpin-loop precursor structures that undergo several stages of enzymatic processing. Precursor miRNA molecules are first generated in the nucleus and then exported into the cytosol where they are processed by the enzyme Dicer to form shorter duplexes, with one of the two single-stranded molecules being incorporated as part of the molecular machinery involved in post-transcriptional gene regulation while the other, passenger strand (usually indicated with *), is short-lived and rapidly degraded. The association between the single-stranded miRNA molecule and the enzymatic complex RNA
Induced Silencing Complex (RISC) lends to translational repression, or cleavage of the targeted mRNA via complementary base pairing to the three prime untranslated region (3’UTR) of their target mRNAs, with the degree of complementarity dictating the fate of the target [11]. What has come to light in recent years is that miRNAs can also induce the up-regulation of gene expression through interactions with genes that contain complementary binding sequences in their promoter region [12]. To date, 1,048 miRNAs have been annotated in humans and 672 in mice (miRBase release 16, 2010) [13] with the true number suggested to be well over 1,000 miRNAs that are encoded in the mammalian genome [14]. Indeed, it is estimated that these short non-coding RNA molecules regulate up to 50% of the transcriptome (protein encoding mRNAs) [15], however, the true breadth of their potential lies in the fact that each miRNA can have hundreds of targets [16] and in retrospect, multiple miRNAs can have the same mRNA targets. These properties of miRNAs suggest that a single miRNA can potentially influence multiple biological pathways [17]. In fact, miRNAs whose expression is tissue and developmentally regulated [18], have been shown to affect a broad range of biological processes in plants and animals including; development, differentiation, cell proliferation, apoptosis [19], regulation of innate immunity [20] and defense from viruses and pathogens [21,22].

Whilst few studies have investigated the mammalian intestinal miRNA signature, a recent analysis of the global porcine micronome demonstrated the expression of 332 miRNAs along the intestinal tract with region-specific expression along the longitudinal gut axis [23]. In line with these findings, upwards of 200 known mature miRNAs and 122 miRNA* species were identified in colorectal cell lines [24] with some found in following clinical studies to have a greater affinity for expression in specific regions and most expressed globally in the human GI tract.
Intestinal miRNAs have experimentally proven biological roles ranging from the regulation of neonatal nutrient metabolism to the control of intestinal fluid and electrolyte transport and permeability, besides affecting intestinal epithelial cell differentiation and maturation.

The intestinal miRNA signature has been found to be deregulated in various disease states. MiRNAs can display both oncogenic or tumor suppressive effects in several types of cancers, and recently 11 miRNAs were found to be differentially expressed in the sigmoid colon of patients with active ulcerative colitis (UC) versus healthy controls, with effects on secretion of pro-inflammatory chemokines.

In addition, both plants and animals differentially express miRNAs following sensing of pathogen-associated molecular patterns (PAMPs). For instance, bacterial flagellin-induced upregulation of miR-393 in Arabidopsis thaliana participates in the regulation of the host defense system. In animals, specific miRNAs are induced in response to various bacterial components, such as lipopolysaccharide (LPS) in monocytes and to viral infection such as in Hepatitis B and C. Moreover, miR-155 is upregulated in gastric epithelial cells following Helicobacter pylori infection. All of these changes resulted in downstream regulation of the immune response.

It has been recently suggested that the onset of several intestinal diseases including Inflammatory Bowel Disease (IBD) and Irritable Bowel Syndrome (IBS), are caused by both deregulation of the intestinal barrier function and by microbial factors, but how the two intertwine to affect such conditions is not well understood. The intestinal barrier is a multi-tiered line of
defense localized at the interface between the external environment and internal milieu and comprises physical, chemical and receptor-mediated pathogen sensing components [38]. The endogenous gut microbiota is an important constituent of the barrier in that it not only participates in the formation of the physical and chemical barrier via pathogen exclusion, antimicrobial peptide secretion, and immuno-modulation, but also acts as a vector of change by modulating the mRNA expression of a number of genes involved in intestinal barrier function [9,7]. However, the epigenetic basis of these interactions is yet to be elucidated as there is a lack of studies evaluating modulation of host miRNAs in response to symbiotic microorganisms. Intriguingly, legumes miRNAs are modulated during the establishment and maintenance of the rhizobia symbiosis in root nodules [39]. Though, it is unknown if this is also true for animals who live in a symbiotic relationship with complex microbial communities at various body sites such as the intestine.

We used germ-free and conventionally raised mice to investigate the impact of the endogenous microbiota on the global expression of caecal miRNAs in vivo. We show that the murine miRNA signature in the caecum is comprised of several variously expressed species and that it is indeed affected by the presence of the microbiota. Moreover, we show that several of the putative mRNA targets of the modulated miRNAs encode for genes known to be involved in the regulation of the intestinal barrier function, including glycosylation enzymes, junctional proteins, proteins found in the mucus layers and genes involved in immune regulation.
Materials and Methods

Animals

Swiss Webster male mice were used according to the Regulations of the Animals for Research Act in Ontario and the Guidelines of the Canadian Council on Animal Care. Animal study design and procedures were approved by the animal ethics committee at the University of Toronto (Animal Use Protocol Number: 20008318). Five germ-free and five conventionally raised mice, 6 weeks of age, were obtained from Taconic Farms (Germantown, NY), sacrificed via cervical dislocation and then dissected in sterile conditions. Upon sacrifice, the entire caecum was immediately excised and caecal contents were collected. Caecal tissues were further cleaned with sterile 0.9 % NaCl, divided into two halves longitudinally, snap-frozen in liquid nitrogen and stored at -80°C until further processing. Caecal contents were immediately fixed in 4% paraformaldehyde and used to confirm the germ-free status of the animals by Fluorescence in situ Hybridization with the EUB338 5’-Cy3 labeled 16S rRNA probe specific for all bacteria (5’/5Cy3/GCT GCC TCC CGT AGG AGT-3’) (Integrated DNA Technologies), as previously described [40].

RNA extraction

Small RNA-containing total RNA was extracted from one-half segment of the caecum from both germ-free (n = 5) and conventionally-raised (n = 5) mice, using miRVANA™ miRNA Isolation Kit (Ambion, Austin, TX, USA), as per the manufacturer’s instructions, eluted in 100 μl of RNase-free water and stored at -80°C. Recovered total RNA concentration and purity were spectrophotometrically assessed using Thermoscientific’s Nanodrop 1000 Spectrophotometer.
(Nanodrop Technologies, Wilmington, DE, USA) and ranged between 2.13-2.16 and 2.01-2.12, respectively. RNA integrity was confirmed by denaturing agarose gel electrophoresis.

**Global microRNA expression profiling**

cDNA was synthesized from 1 µg of total RNA (n = 3 per group) using the Taqman® MicroRNA Reverse Transcription Kit in conjunction with Rodent Megaplex™ Primer Pools according to the manufacturer’s protocol (Applied Biosystems, CA, USA). The Rodent Megaplex™ Primer Pools contains two sets of microRNA-specific RT primers, pools A and B, that enable for the RT of 375 microRNAs/6 species-specific controls and 210 microRNAs/6 species-specific controls, respectively. Separate reactions were run for pools A and B in Applied Biosystems’ GeneAmp® PCR System 2700 Thermocycler. Global microRNA expression profiling was conducted by TaqMan quantitative PCR using Applied Biosystems’ Taqman® Rodent MicroRNA Array Set v2.0 (Taqman Low Density Arrays, TLDAs) that comprise two microfluidic cards (plates A and B) containing a total of 384 Taqman® Assays per card (some of which are duplicate probes). cDNA products from the Megaplex RT pools set were independently assessed on both microfluidic cards with plate A enabling for the simultaneous quantification of 375 microRNA targets/6 controls while plate B, 210 microRNA targets/6 controls. The protocol suggested by the manufacturer was followed. Real-Time PCR was performed using an Applied Biosystems 7900 HT Real-Time PCR system and default thermal-cycling conditions for 384-wells Taqman Low Density Arrays.
Real-Time PCR validation of individual microRNAs expression

Ten ng of total RNA (n=5 per group) was reverse transcribed with the Taqman® MicroRNA Reverse Transcription Kit and primers specific for miR-455 (Assay ID: 002455) and the endogenous control snoRNA135 genes (Assay ID: 001230) (Applied Biosystems) in Applied Biosystems’GeneAmp® PCR System 2700 Thermocycler according to the manufacturer’s protocol. Real time PCR was then conducted using undiluted cDNA, TaqMan MicroRNA Assays (miR-455 assay ID: 002455 and snoRNA135 assay ID: 001230) and the TaqMan 2X Universal PCR Master Mix, No AmpEraseUNG² (Applied Biosystems) in a 10 µl PCR reaction. Each reaction was run in triplicates in a 384-well optical plate in Applied Biosystems’ 7900 HT Real-Time PCR machine using the 9600 emulation mode with an initial hold at 95°C for 10 minutes followed by 40 cycles of 95°C for 15 seconds, and 60°C for 60 seconds. Results were expressed as fold change between germ-free and conventional mice as calculated by ΔΔCt method [41] after normalization to sno-135 gene, which was shown to be equally expressed in the caeca of the two groups of mice by the TLDA experiments. Significance of differential gene expression was assessed with the Mann-Whitney test using GraphPad Prism 5 Software (La Jolla, CA, USA).

Analysis of TLDA data

Raw data were pre-processed in SDS 2.3 for individual plates and then concurrently for all plates in SDS RQ Manager 1.2 (Applied Biosystems) for the generation of Ct (Cycle threshold) values. A pre-selection filter was applied to all miRNA TLDA data to reduce noise in the dataset and to reduce the severity of the multiple testing adjustment. To this end, any miRNA not meeting both
of the following criteria were removed from further analysis: a) Presence in all three of either conventional samples or the germ-free group; b) Presence in at least one of the conventional samples to enable the ΔΔCt method to be used for normalization. Following pre-filtering, all duplicate probes for the same miRNA species on the same plate were averaged and the mean Ct value was utilized for further analysis. Data were then normalized by “columnwise mean” normalization, such that the target miRNA is normalized to the mean Ct of all miRNA for each sample, a method that has been suggested as an improvement for high-throughput miRNA Quantitative PCR (qPCR) [243] where the mean abundance of hundreds of targets may be more stable than any endogenous control across samples and experimental groups. To identify differentially expressed genes, the empirical Bayes-moderate t-test was used as implemented in the LIMMA R package [43]. For this approach, missing Ct values were assumed to be unknown rather than imputed to 40, to avoid creating a bimodal distribution of Ct values, which would violate the assumption of the t-test. False discovery rate was calculated by the method of Benjamini and Hochberg, as implemented in the R package multtest (Pollard et al., v.1.22.0).

Supervised heatmaps were created using the R package gplots. Clustering in the heatmaps is based on complete linkage and Euclidean as the distance metric, using default setting for the hclust (hierarchical cluster) function in R [44].

Analysis was performed in the R language and environment for statistical computing (R Development Core Team, 2008, v2.8.1) [45].
**MiRNAs target prediction**

To investigate the relationship between selected miRNAs of interest and the genes that they potentially target, we mapped them into a miRNA network using NAViGaTOR ver. 2.2[46]. We first used high precision miRNA:target relationships in mouse - taken from the Targetscan Conserved Targets (Conserved_Sites_Context_Scores.txt Release 5.1) [47,48,49] or PITA TOP database predictions (PITA_targets_mm9_0_0_TOP.tab.gz, May 2010 download) [50]. Previous work examining miRNA:target relationships suggests that both PITA and Targetscan provide high quality interactions suitable for the construction of an interaction network [51]. Genes identified by this first analysis were then filtered based on their inclusion in an intestinal barrier gene set to assess the potential impact of differentially expressed miRNAs on the intestinal barrier function. A subset of 527 genes important in maintenance of the intestinal barrier function were identified and classified according to function - mainly physical, chemical and pathogen sensing components as per Cummings J. H., et al. [38] (Table S5). Identified miRNA target genes were filtered by the intestinal barrier set prior to being mapped into the miRNA network. This reduced the number of initial miRNA target genes of the 11 miRNAs with predicted targets from 2,755 in the general setting to 34 present in the intestinal barrier setting.

**Analysis of miRNA potential targets biological function**

To further understand the functions of gene targets of miRNAs with altered expression in this study, we examined all 2,755 gene targets of the 11 miRNAs prior to the filtering step. Using the Panther Classification System Version 7.0 [52,53], we examined the over- or under-representation of our miRNA target genes compared to a universe consisting of all genes listed
as miRNA targets in the PITA Top Targets or Targetscan Conserved Targets (as discussed above). Categories examined include Gene Ontology Classes: Biological Processes, Molecular Function, Cellular Component as well as Pathway Analysis and Protein Class Analysis.

**Results**

**Differential expression of miRNAs in the caecum of germ-free and conventionally raised mice**

To assess if the caecal miRNA expression signature is associated with the presence of the endogenous microbiota, we examined small RNA-containing total RNA extracted from the caecum of germ-free (n=3) and conventionally raised mice (n=3). Each sample was independently run on two different Taqman Low Density Arrays (Plates A and B), which combined allow for the analysis of 585 mature miRNAs. Fifty-seven percent of the targeted miRNAs were detectable (Ct<35) in the caeca of both germ-free and conventionally raised mice (Table S1: [http://www.biolsci.org/v08/p0171/ijbsv08p0171s1.xls](http://www.biolsci.org/v08/p0171/ijbsv08p0171s1.xls)). Ranked mean abundance of miRNAs was similar for both groups of mice (Spearman R=0.74, P< 0.0001, 95% CI= 0.68 to 0.78) (Figure 1).
Figure 1. Caecal global microRNA expression is correlated in germ-free and conventional mice.

Scatter plot depicting the relationship between global miRNA expression levels of 336 miRNAs in germ-free (GF) and conventional (Conv) caecal samples (n=3/group) as assessed by qRT-PCR and TLDA plates A and B. Data are presented as mean delta Ct values for each miRNA (mean delta Cts represent the average delta Ct for all three samples/group, with each individual Ct normalized by mean expression value normalization procedure). Of the 585 miRNA species analyzed, 336 remained after filtering and normalization, as explained in Materials and Methods and were used in subsequent analyses.

We found 18 transcripts differentially expressed between the two groups, including both up- and down-regulated miRNAs with a fold change (germ-free vs. conventional) ranging between 0.2 and 4.6 (Table 1) (False Discovery Rate = 0.2). These correspond to 16 unique miRNAs, including mmu-miR-351 and rno-miR-351 - two sequences conserved in mouse and rat - while
Y1 is a rat endogenous small RNA. A second rat miRNA species, rno-miR-664, was found to be significantly up-regulated by 2.85-fold in germ-free samples. The murine homolog, mmu-miR-664, is not represented on the TLDA plates that we used. Sequence analysis using miRviewer [54] demonstrates that miR-664-1 shows sequence similarity in rat, mouse and horse with a greater conservation amongst rat and mouse. Therefore, it is likely that the measured signal is biologically reliable and derives from cross-reaction of the rno-miR-664 TaqMan assay with the homologous murine miRNA species. Five of the sixteen transcripts correspond to passenger miRNA (miRNA*) sequences. Up-regulation of miR-455 in germ-free versus conventional mice was confirmed in a separate experiment using gene-specific single-well TaqMan PCR and RNA from the caecum of five mice per group (fold change germ-free versus conventional=1.7, Mann-Whitney test p=0.0079).
<table>
<thead>
<tr>
<th>miRNA</th>
<th>Fold GF vs Conv</th>
<th>Adjusted P-values</th>
<th>Genomic Location</th>
<th>Mature miRNA Sequence (5' - 3')</th>
<th>Homology</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmu-miR-21*</td>
<td>0.20</td>
<td>0.06</td>
<td>chr1: 85397569-85397660 [1]</td>
<td>55-CAACACGCUAGUUGCCUUG - 77</td>
<td>NA</td>
</tr>
<tr>
<td>mmu-miR-351</td>
<td>1.31</td>
<td>0.06</td>
<td>chr1: 133959260-133959210 [1]</td>
<td>15-UCUCUGGAGGAGCCGUG - 20</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-351</td>
<td>1.33</td>
<td>0.20</td>
<td>chr1: 50404542-50404590 [1]</td>
<td>15-UCCUGAGGACGUCCGUG - 39</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a</td>
<td>0.46</td>
<td>0.12</td>
<td>chr1: 81056641-81056652 [1]</td>
<td>45-AGUGUACAGCGAAUGCCUUG - 71</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-1</td>
<td>0.47</td>
<td>0.06</td>
<td>chr1: 10397975-10399005 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 31</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-2</td>
<td>1.00</td>
<td>0.12</td>
<td>chr1: 10404595-10405107 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 36</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-3</td>
<td>1.50</td>
<td>0.17</td>
<td>chr1: 10405305-10405387 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-4</td>
<td>1.50</td>
<td>0.17</td>
<td>chr1: 10407752-10407844 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-5</td>
<td>1.00</td>
<td>0.12</td>
<td>chr1: 10410258-10410700 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-6</td>
<td>1.00</td>
<td>0.12</td>
<td>chr1: 10412367-10412767 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-7</td>
<td>1.00</td>
<td>0.12</td>
<td>chr1: 10415107-10415219 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-8</td>
<td>1.00</td>
<td>0.12</td>
<td>chr1: 10417007-10417100 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-9</td>
<td>1.00</td>
<td>0.12</td>
<td>chr1: 10420020-10421002 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-487a-10</td>
<td>1.00</td>
<td>0.12</td>
<td>chr1: 10424500-10425102 [1]</td>
<td>15-UAGGCGGAGCUAGAUAGCCG - 35</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-210</td>
<td>0.54</td>
<td>0.19</td>
<td>chr1: 82054020-82054092 [1]</td>
<td>7-AGUGCUGCUAGUUGUACCCU - 39</td>
<td>NA</td>
</tr>
<tr>
<td>mmu-miR-145a</td>
<td>0.58</td>
<td>0.12</td>
<td>chr1: 51219651-51219905 [1]</td>
<td>61-UCAAGUGACUACGAAUCUUG - 82</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-145</td>
<td>1.52</td>
<td>0.17</td>
<td>chr1: 82054020-82054092 [1]</td>
<td>7-UCCUGCUUCGGCUAAUGAU - 29</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-183</td>
<td>1.56</td>
<td>0.17</td>
<td>chr1: 30119656-30119737 [1]</td>
<td>6-UAGGCACUGGUAAGAUUCACU - 27</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-133a-1</td>
<td>1.81</td>
<td>0.12</td>
<td>chr1: 10762901-10762974 [1]</td>
<td>43-UUCGGGCUUGCCCGUUGCACACCUUG - 64</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-133a-2</td>
<td>1.81</td>
<td>0.12</td>
<td>chr1: 10762901-10762974 [1]</td>
<td>43-UUCGGGCUUGCCCGUUGCACACCUUG - 64</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-150</td>
<td>1.65</td>
<td>0.15</td>
<td>chr1: 52277127-52371791 [1]</td>
<td>6-UCUCACCGCUUGUACUGACUG - 27</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>Yt-436735</td>
<td>2.02</td>
<td>0.15</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>mmu-miR-972</td>
<td>2.74</td>
<td>0.06</td>
<td>chr1: 10115151-10115161 [1]</td>
<td>25-GUUAGGCUAGUUGUAGCUUGUAG - 47</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-151a-1</td>
<td>2.79</td>
<td>0.17</td>
<td>chr1: 13985332-13985335 [1]</td>
<td>51-ACCAUCUACUGCUAGAUAGCCG - 76</td>
<td>NA</td>
</tr>
<tr>
<td>Yt-436784-1</td>
<td>2.85</td>
<td>0.15</td>
<td>chr1: 47002544-47002642 [1]</td>
<td>35-UUCGUACCUUAGCUAGAUAGCUUG - 59</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-455</td>
<td>3.00</td>
<td>0.15</td>
<td>chr1: 56278155-56278196 [1]</td>
<td>54-GGAGGUAAGAGGCUGUACACUGA - 74</td>
<td>Mmu Rno</td>
</tr>
<tr>
<td>mmu-miR-139*</td>
<td>4.43</td>
<td>0.12</td>
<td>chr1: 12391094-12392092 [1]</td>
<td>61-GGUGGCUUAGCUACUACAGCA - 82</td>
<td>NA</td>
</tr>
<tr>
<td>mmu-miR-7*</td>
<td>4.60</td>
<td>0.10</td>
<td>chr1: 10591711-10591728 [1]</td>
<td>63-AUGUACAGCUGUACUACAGCUG - 84</td>
<td>NA</td>
</tr>
</tbody>
</table>
Table 1. Differentially expressed microRNA in germ-free versus conventional mice.

Real-time PCR analysis of the global expression of microRNAs in the caecum of germ-free (GF) (n=3) versus conventional (CONV) mice (n=3). Fold-change and statistical significance were calculated after mean expression value normalization. Statistical significance is based on Bayes-moderated t-test with a FDR of 20%. In total 18 transcripts were found to be differentially expressed between the two groups with 16 unique mature microRNAs (Y1 is a rat endogenous small RNA). Fold change corresponds to the ratio of mean expression of the microRNA in GF mice to mean expression in CONV mice. Values < 1 indicate lower expression in GF (down-regulated in GF mice compared to CONV mice), whereas values > 1 indicate higher expression in GF (up-regulated in GF mice compared to CONV mice). Genomic locations and corresponding microRNA sequences (5’–3’) are based on miRBase version 16. Homology of microRNA sequences between diverse species is derived from miRviewer (last updated November 9, 2008).

Analysis of miRNA conservation and their genomic contexts revealed that all of the significantly differently expressed miRNAs belong to various families and cluster separately in terms of their genomic locations with the exception of miR-351 in which both rat miRNA (rno-miR-351), and murine miRNA (mmu-miR-351) were found to belong to the same mir-351 family, based on sequence conservation [13].

Furthermore, supervised hierarchical clustering analysis using the 18 differentially expressed transcripts demonstrated intra-group similarities in miRNA expression with inter-group variation in miRNA expression (Figure 2), showing that the caecal miRNA signatures cluster according to the presence or absence of the endogenous microbiota.
Figure 2. Clustering of caecal microRNAs expression profiles in germ-free and conventionally raised mice.

The profiles of 18 transcripts including 16 microRNAs significantly differently expressed (p<0.05, FDR<20%) between germ-free and conventional caecal samples were visualized using a supervised heatmap (complete linkage and Euclidean distance metric). Expression values range from +2.5 \log_2 to -2.5 \log_2 of ΔCt values normalized using mean expression value normalization with positive values (red) indicating higher expression and negative values (green) indicating lower expression in germ-free versus conventional mice. Dendrograms indicate the correlation between groups of samples and genes. Samples are in columns and transcripts in rows.
Experimentally verified and predicted mRNA targets of the differentially expressed miRNAs

Based on Tarbase V5.0 [55] and miRecords V2. [56], two freely available databases that provide a repository of information pertaining to experimentally validated miRNA targets in several animal species, plants and viruses, six (miR-133a, miR-672, miR-183, miR-148a, miR-145, miR-150) of the sixteen differentially expressed miRNAs have experimentally verified mRNA targets (Table S2: http://www.biolsci.org/v08/p0171/ijbsv08p0171s2.docx). Of these mRNA targets, seven (Serum response factor (Srf), Ras homolog gene family, member A (Rhoa), Cell division cycle 42 homolog (S. cerevisiae) (Cdc42), Peroxiredoxin 6 (PRDX6), Homeo box A9 (Hoxa9), Vascular endothelial growth factor A (Vegfa), and Myeloblastosis oncogene (Myb) were detected with a signal intensity higher than 150 in microarray experiments analyzing gene expression in C57BL/6 mice caeca (n=2) (Gene Expression Omnibus [57] dataset GSE1133[58]). Therefore, based on the miRNA/mRNA co-expression criterion for target validation [59], these genes are good candidates for microbiota-dependent expression modulation via miRNA.

However, each miRNA species is likely to have multiple physiologically relevant targets, most of which are unknown, and several algorithms can be employed for their in silico identification. To predict targets for the 16 miRNAs found to be differentially expressed between germ-free and conventional mice, we adopted a conservative approach. We extracted target predictions from 2 sources: Probability of Interaction by Target Accessibility (PITA) [50] and TargetScan.
thus, our target prediction sets are based on several criteria including conservation of target binding sites and the degree of accessibility of the three prime untranslated regions of the mRNA target. Two thousand seven hundred and fifty-five unique genes were found to map as targets of these miRNAs as predicted by both algorithms (Table S3: http://www.biolsci.org/v08/p0171/ijbsv08p0171s3.xls). These genes were mapped to PANTHER database [52] to assess their group descriptors. We first considered the Gene Ontology classifications and found our gene list to be significantly enriched in several categories: (1) biological processes including development, cell communication, signal transduction (all at p<0.0001), among others; (2) molecular functions including DNA, transcription factor, protein binding (all at p<0.0001), among others; and (3) cellular components such as actin cytoskeleton (p<0.0001) (Table S4). Next, we found our targets enriched in the Wnt signalling pathway, angiogenesis, transforming growth factor-beta (TGF-β) and cadherin signalling pathways and in the transcription factors protein class followed by enzyme modulator and ribosomal proteins (all at p<0.0001), among others (Table S4). Several of the 2,755 global putative targets of the differentially expressed miRNAs were found to map to diverse components of the intestinal barrier function when assessing gene ontologies. These include; (1) Biological Processes such as cell-cell adhesion (p<0.0001), immune system processes (p<0.001) encompassing antigen processing and presentation (p<0.01) and defense response to bacterium (p<0.2) amongst others; (2) Molecular Functions including structural constituents of cytoskeleton (p<0.0001) and (3) Cellular Components such as actin cytoskeleton (p<0.0001), MHC protein complex (p<0.01), cell junction (p<0.01) as well as other factors (Table S4: http://www.biolsci.org/v08/p0171/ijbsv08p0171s4.xls). Similarly, identification of the inflammation mediated by chemokine and cytokine signalling pathway (p<0.001) during
functional analysis of in silico data demonstrate potential implications of microbial-dependent miRNA regulation on the intestinal barrier function.

**Modeling of the microbiota-dependent intestinal barrier micronome.**

Considering the results described above, and the fact that endogenous microbes play an important role in the creation and maintenance of the intestinal barrier we decided to investigate the potential impact of the microbiota-responsive miRNAs on the intestinal barrier function via analysis of intestinal barrier specific gene targets. We first compiled a gene set containing 527 genes involved in the regulation of the intestinal barrier function, as explained in Materials and Methods, which we called the “Intestinal Barrier Gene Set”. The list of genes included in this gene set is provided as (Table S5: http://www.biolsci.org/v08/p0171/ijbsv08p0171s5.xls ). We then filtered the 2,755 target genes by this pre-defined gene set and established all miRNA:target interactions for the remaining genes. Using NAViGaTOR (Network Analysis, Visualization and Graphing Toronto) ver. 2.2 [46], a scalable, network analysis and visualization system, we mapped the miRNA network linking our identified miRNA of interest in order to examine the micronome, as described before [51] (Figure 3). MiR-487b did not have any intestinal barrier targets as per the algorithms employed and therefore it does not appear in the figure. Thirty-four intestinal barrier-related genes were found to be among the potential targets of the intestinal miRNAs the expression of which depends on the endogenous microbiota, and fifteen of these were identified by both algorithms. These include genes involved in glycosylation, cell-cell junction formation, the mucus layer and genes involved in immune regulation particularly MHC I and II proteins amongst others. Closer inspection of this miRNA interaction network reveals several genes co-targeted by the miRNAs identified as differentially expressed between the caecal miRNA signatures in germ-free and conventional mice. Formin 1 (FMN1) is co-targeted
by 2 miRNAs down-regulated in the germ-free mice (miR-351 and miR-467a) as well as one up-regulated miRNA (miR-145). Other genes appearing to be co-targeted by multiple differentially expressed miRNAs are: Cadherin 5 (Cdh5), UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminytransferase 5 (Galnt5), poliovirus receptor-related 1 (Pvrl1), fascin homolog 1, actin bundling protein (Fscn1), Cingulin (Cgn), glucosaminyl (N-acetyl) transferase 1, core 2 (Gcnt1) and UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminytransferase 7 (Galnt7). Of the thirty-four predicted intestinal barrier genes targets, twenty had been previously found to be expressed in the caeca of C57BL/6 mice (n=2), in microarray experiments with a hybridization signal higher than 150 (ATP-binding cassette, sub-family B (MDR/TAP) member 9 (Abcb9), Nicastrin (Ncstn), Spermidine/spermine N1-acetyltransferase 1 (Sat1), Desmoglein 3 (Dsg3), UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 1 (B4galt1), Leucine aminopeptidase 3 (Lap3), beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminytransferase (Gcnt1), CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltransferase (St3gal3), Junction plakoglobin (Jup), Aminopeptidase puromycin sensitive (Npeps), UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminytransferase 7 (Galnt7), Plakophilin 1 (Pkp1), (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 6 (St6galnac6), Carcinoembryonic antigen-related cell adhesion molecule 1 (Ceacam1), Formin-1 (Fmn1), Prostasin (Prss8), glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1 (C1gal1), UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 5 (B4galt5), UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 2 (B4galt2), Myosin, light polypeptide kinase (Mylk) (Gene Expression Omnibus [57] dataset
GSE1133 [58]). This list is conservative, since not all of the genes shown in Figure 3 had microarray probes.

Figure 3. Differentially expressed microRNAs impact on the intestinal barrier.

MicroRNAs significantly differentially expressed between germ-free and conventional mice are represented in this diagram by circles, with the colour corresponding to degree of differential microRNA expression in germ-free and conventional samples. Expression values range from \(+2.5 \log_2\) to \(-2.5 \log_2\) of ∆Ct values with positive values (red) indicating higher expression, and negative values (green), indicating lower expression in germ-free versus conventional mice. Putative intestinal barrier gene targets as identified by the algorithms TargetScan and PITA are represented by triangles. MicroRNAs with a greater number of intestinal barrier targets are symbolized with a larger circle size. Intestinal barrier gene targets that are predicted by both
algorithms are indicated by thicker lines. Differentially expressed microRNAs* were not included in the diagram as they are not present in both prediction databases. MicroRNA-487b did not have any intestinal barrier targets as per the algorithms and therefore was excluded from the figure.

Seventy one percent of the genes included in our gene set were not considered by PITA and Targetscan due to the low conservation of the 3’UTR in homologous genes. Moreover, this analysis did not incorporate passenger miRNAs because they are not considered in PITA and Targetscan databases. Alternatively, when using MicroCosm targets [13] to map these, only miR-let7g* had targets remaining after filtering with the intestinal barrier gene set. These are: C1GALT1-specific chaperone 1 (C1galt1c1), Claudin-7 (Cldn7), Histocompatibility 2, class II antigen A, beta 1 (H2-Ab1), Pancreatitis-associated protein (Pap), Phospholipase A2, group XIIA (Pla2g12a), Phospholipase A2, group IB (Pla2g1b), Spermidine synthase (Srm), Thimet oligopeptidase 1 (Thop1), Toll-like receptor-11 (Tlr11) and Toll-like receptor-13 (Tlr13).

Finally, in order to substantiate the hypothesis that gut commensals impact the intestinal barrier via miRNA expression modulation, we crossed-matched our global list of intestinal barrier genes with genes previously identified to be differentially expressed in the jejunal mucosa of intestinal-specific Dicer knock-out mice [60]. The result of this analysis provides experimental evidence that miRNAs indeed impact on barrier-related gene expression, with potential repercussions on its function. Of particular interest are intestinal barrier genes from our list that were found to be experimentally perturbed (up- or down-regulated) by the conditional knock-out of Dicer [244]. They include: glycosylation enzymes, immuno-inflammatory response genes, components of MHC I and II, junctional proteins, mucus layer associated proteins and defense response proteins, including antimicrobial peptides and Pathogen Associated Molecular Pattern (PAMP)
responsive elements. Although, this analysis shows that miRNAs affect genes that comprise the intestinal barrier, in order to further establish a nexus between microbial induced modulation of miRNAs, which in turn affects barrier function, we combined putative barrier related gene targets of the microbial dependent miRNAs (Figure 3) with the Dicer knock-out mice data [244]. Among the genes differentially expressed in the absence of miRNAs in the jejunal mucosa, we found that seven intestinal barrier related genes were either up-regulated (glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1 (C1galt1), myosin, light polypeptide kinase (Mylk), Aminopeptidase puromycin sensitive (Npepps), UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminytransferase 7 (Galnt7) and Prostasin (Prss8)), or down-regulated (protein kinase C zeta isoform a (Prkcz), beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglicosaminyltransferase (Gcnt1)). More specifically, since these seven genes are in silico targets of the miRNAs modulated by the absence or presence of the microbiota (germ-free versus conventional mice) and experimental evidence points to a role of epigenetic regulation of these genes via miRNAs (Dicer knock-out study) it reiterates a potential novel mechanism of host-microbial cross-talk via microbial dependent regulation of miRNAs that may translate into effects on the host with respect to regulation of the intestinal barrier function.

Discussion

Gene expression modulation is one of the mechanisms underlying the cross-talk between gut endogenous microbiota and host epithelium, and therefore plays a critical role in intestinal homeostasis. Here we show that the presence of the microbiota in the murine intestinal tract, particularly in the caecum, also associates with a distinctive miRNA signature, supporting a role for gut endosymbionts in post-transcriptional regulation of gene expression. Few studies have looked at the relative expression of miRNAs along the cephalocaudal axis of the healthy gut,
particularly with respect to the passenger miRNA* forms, and to the best of our knowledge, this is the first study examining miRNA expression in the murine caecum. We found a characteristic micromome in the caecum, with 334 miRNA species expressed in this region in both germ-free and conventional mice; of these, 74 are miRNA* forms. While miRNAs* are thought to correspond to the rapidly degraded strand of the miRNA duplex, there is evidence suggesting that they may play a so-far unrecognized role within cells [245], and in fact may act in a similar fashion to guide strand miRNAs in terms of abundance and gene regulation [61].

Comparison of the global murine miRNA signature along several intestinal loci, namely the small and large intestinal mucosa, as well as our caecal analysis allows for certain parallels to be drawn. Some of the miRNAs belonging to the 15 miRNAs/miRNA families most expressed in the jejunal and colonic mucosa [60] are also expressed in the caeca of both germ-free and conventional mice (Table S6: http://www.biolsci.org/v08/p0171/ijbsv08p0171s6.xls) and 3 (miR-192, miR-378, miR-29a) of the 15 miRNAs most highly expressed (based on Ct values) in the caecum are also expressed in both the jejunum and colon [60]. Although diverse genes cannot be compared by Ct values, sorting allows for a qualitative measure of the relative level of gene expression and to identify genes that do or do not display a regional expression pattern within the intestine. miR-143 and miR-145 were part of the top expressed miRNAs in common between the jejunum and caecum, and are found in the same genomic cluster (<10kb distance from one another on chromosome 18). miR-200b was found in common as a highly expressed miRNA within both the large intestine and caecum. Interestingly, other miRNAs with sequence similarities to miR-200b were also found to be highly expressed, including miR-200a in the colon, and miR-200c in the caecum, suggesting that members of the miR-8 family play an important physiological role in distal intestinal regions. On the other hand, 19 miRNAs were
found to be expressed in the caecum of conventional mice (Ct<35) but not in the jejunal or colonic mucosa based on the absence of sequence read data [60] in either of the two regions (Table S6), suggesting they may be restricted to the caecum. Intergroup comparisons between germ-free and conventional mice illustrate a relatively high degree of similarity between the top miRNAs expressed in germ-free and conventional caeca with all 15 of the miRNAs with the lowest Ct values overlapping between the two groups (Table S6). Moreover, there is a general concordance between the murine caecal micronome and the human intestinal micronome, which incorporates the caecum. Juxtaposing data on the 13 most highly constitutively expressed miRNAs in both the terminal ileum and colon (caecum, transverse colon, sigmoid colon and rectum) from pinch biopsy samples of healthy adults [25] with the murine caecal miRNA signature of conventional and germ-free mice, revealed that, five miRNAs (miR-143, miR-192, miR-200b, miR-200c and miR-24) found in the intestines of humans were amongst the top 15 mostly highly expressed miRNAs in the murine caecum based on Ct values (Table S6). Moreover, miR-19b which was found to have a 3.2-fold higher expression in the caecum versus the terminal ileum from biopsied samples in humans, was also found, based on our aforementioned analysis, to be part of the 15 most highly expressed miRNAs in the murine caecum (germ-free and conventional mice) but not within the jejunal or colonic mucosa, suggesting this miRNA may exert a more profound effect within the caecum. Although our analysis of miRNA expression between intestinal regions in the mouse is qualitative and cannot be used to determine fold differences in expression, it gives merit into using the mouse as a model organism to investigate intestinal miRNAs as certain parallels can be found in humans.

The expression of these miRNAs may be under genetic and environmental control. The latter is particularly important in the case of the intestine where the epithelium engages in a continuous
cross-talk with the luminal microbes. Here we show that indeed the endogenous microbiota contributes to the physiological miRNA signature in the caecum, which results in 16 miRNAs being differentially expressed between germ-free and conventionally raised mice. Moreover, of these, miR-133a and miR-467a were found to be caecal specific miRNAs when compared with the jejunum and colonic mucosa and miR-145 was a non-selectively expressed miRNA with high levels of expression along the intestine, insinuating a role for microbial control of both regional specific and globally expressed miRNAs that may transcend the boundaries of the caecum. Host miRNA modulation has been so far observed in response to pathogenic insults including prions [62], viruses such as Hepatitis B and C [34] and influenza virus[63], bacteria such as Helicobacter pylori [64], Francisella novicida [65] and Gram negative bacteria LPS [66], the yeast Candida albicans [67] or parasites such as Cryptosporidium parvum [68] and Toxoplasma gondii [69]. To our knowledge, the only host-microbe symbiotic relationship associated to miRNA modulation in the host is the legume-rhizobium symbiosis [39]. In this study, we used whole thickness caeca. Therefore, it is possible that differential miRNA expression was the result of different cellular composition of the caecum of germfree and conventional mice. As well, we could not establish the cellular origin of our measured signals; though, a previous study revealed that at least nine human miRNAs (hsa-miR-145, hsa-miR-150, hsa-miR-133a, hsa-miR-148a, hsa-miR-183, hsa-let-7g*, hsa-miR-181a*, hsa-miR-21*, hsa-miR-27b*) that have sequence homology with our differentially expressed murine miRNAs are indeed expressed in colorectal cell lines [245], suggesting that the intestinal epithelial monolayer is susceptible of responding to the endogenous symbionts or their products, by miRNA modulation.

To date limited information is available on the biological role of these miRNAs; however, several of the miRNAs found to be differentially expressed in this study are known to be altered
in cancer states. MiR-148a, which we found to be expressed more highly in conventional mice, was found to be more highly expressed in tumor samples versus normal colonic epithelium [24], while miR133a and miR-145, which we found to be more highly expressed in germ-free mice, were shown to exhibit significantly higher levels of expression in normal versus tumor tissues [70,71,72]. There is a general consensus in the literature that endogenous gut microbes can alter colon cancer susceptibility and germ-free rats were found to develop less and smaller tumors than their conventional counterparts when using a protocol that induces colorectal cancer [73]. This was attributed to enhanced anticancer immune response. In our study, a novel pathway may be proposed that incorporates microbe signalling to the host and can alter the expression of tumor-suppressors or oncogenes post-transcriptionally via miRNA regulation. Indeed, miR-145 and miR-133a were both predicted by multiple algorithms to target Fascin-1 (FSCN1) (Figure 3), a gene involved in actin cytoskeleton assembly, the down-regulation of which was experimentally found to explain the tumor suppressive effects of miR-145 and miR-133a in bladder, esophageal squamous cell and breast carcinomas [71,74,75].

However, inferring a microbiota-dependent physiological role for differentially modulated miRNA species depends on the identification of their mRNA targets in the caecum. Six of the endogenous microbiota-dependent miRNAs were experimentally proven in previous studies to target various genes, some of which are expressed in the caecum. These genes are categorized in various Gene Ontology classes including development, DNA binding, protein binding, transcription as well as signalling pathways including Wnt receptor signalling suggesting that the microbiota may be an additional factor controlling these functions. These findings are also in line with our PANTHER analysis where experimentally validated targets that are co-expressed in the caecum also map to some of the same functions of the targets predicted in silico.
Moreover, PANTHER, TargetScan and PITA findings collectively reinforce the role that gut bacteria play in organization of the actin cytoskeleton and gut angiogenesis, both previously shown to be affected by gut bacteria at the transcriptional (mRNA) level [9,7,76] suggesting that the impact of gut bacteria on specific pathways is many-sided. Particularly, in terms of angiogenesis, global pathway analysis of targets affected by the microbiota-dependent miRNAs illustrate effects on angiogenesis including the process of angiogenesis (p= 4.56E-13) itself and the Vascular Endothelial Growth Factor (VEGF) signalling pathway (p=7.67E-3). Although both the microbiota [76] and miRNAs [77] have been independently shown to affect vascularization, in silico findings in this study establish a possible link between them, and demonstrate a potential mechanism in which the molecular dialogue between gut bacteria and the host is carried out to affect these functions. Gut bacteria are important in the formation of the intestinal vascular network during postnatal development [76] and miRNAs in general are known to be developmentally regulated. Since the gut microbiota gradually establishes during postnatal life, it is possible for the two processes to intertwine. Indeed, a recent study showed that exposure to LPS from endogenous E. coli in the developing gut of the murine neonate, results in toll-like receptor-4 mediated expression of miR-146a and subsequent down-regulation of interleukin-1 receptor-associated kinase 1 (IRAK1) and the creation of an immunologically tolerant environment [78]. Future studies could examine the postnatal expression pattern of the differentially expressed miRNAs and of their target genes.

Several studies have shown microbial dysbiosis and miRNA deregulation to be important culprits in a number of digestive diseases, including irritable bowel syndrome [36,28] and ulcerative colitis [79,32]. Though, it is not understood if and how the two associate to impact these conditions. Based on our findings, we suggest that deregulation of the microbial
composition in digestive diseases may at least partially affect the miRNA expression signature, and in turn influence the associated pathologies. One line of evidence involves miR-455, which in our study was found to be up-regulated in the caecum of germ-free mice, and found to target heat-shock factor 1 (hsf1) based on bioinformatics analysis. Hsf1 attenuates the effects of experimentally induced colitis in mice models via indirectly inhibiting the production of pro-inflammatory cytokines, cellular apoptosis and cell adhesion molecule induction [80]. Although the authors did not take into account both the microbiota and miRNAs in these mice models we speculate that altered microbial composition in these disease states may affect miRNAs that in turn impact on hsf1 with potential repercussions on gastrointestinal disease states.

In both a healthy situation and disease state one of the primary lines of defense in the gastrointestinal tract is the intestinal barrier, of which the gut microbiota is a critical component. Though, at the same time gut microbes act as a regulator of the barrier function at the mRNA level, by impacting the expression of several genes. Recently, genes regulating the intestinal barrier were found to be differentially expressed in the jejunum of intestinal-specific Dicer knock-out mice, highlighting a role for intestinal miRNAs in the regulated expression of intestinal barrier genes [81]. In line with this finding, we found that a number of genes included in our intestinal barrier gene set are indeed regulated post-transcriptionally in Dicer knock-out mice and therefore depend on miRNAs. Interestingly, these genes are also the potential targets of gut microbiota-dependent miRNAs (Figure 3). These were identified despite a stringent approach excluding 376 of our intestinal barrier genes which are not reported in the PITA and TargetScan databases. Further supporting the existence of an intestinal barrier regulatory network involving miRNAs and the gut microbiota, some of the intestinal barrier genes targeted by our selected miRNAs were found to be up- or down-regulated in Dicer 1-deficient mice.
versus controls [60], suggesting that the microbiota can indirectly impact on the intestinal barrier post-transcriptionally via miRNA regulation. Though, it is important to note that in this study we used whole thickness tissues in order to obtain a comprehensive evaluation of the intestinal miRNA signature response to the commensals.

Physiologically, the basis of this dialogue has yet to be established; nonetheless, an emerging concept incorporates the utilization of toll-like receptors (TLRs) as potential mediators. For example, miR-147 was found to respond to LPS stimulation of TLR4 in murine peritoneal macrophages, resulting in an attenuated release of Interleukin-6 (IL-6) and Tumour Necrosis Factor alpha (TNF-α) [246]. Moreover, miR-146a was also reported to dampen the inflammatory response upon up-regulation through PAMP activated TLRs [83,33]. In turn, these studies show applicability of microbial alterations in miRNA which can impact the barrier function. Therefore, it seems plausible that TLRs which are localized at the interface between the microbiota and the molecular machinery of host cells may be a potential facilitator of this communication.

In summary, this study shows that the murine caecum expresses a large variety of miRNAs, sixteen of which exhibit differential expression in the presence or absence of the endogenous microbiota. Therefore, gut bacteria may impact on intestinal gene regulation not only at the transcriptional level but also post-transcriptionally; thus, contributing to intestinal homeostasis through fine-tuning gene expression. By modulating miRNAs, the gut microbiota may affect a much larger number of genes than so far expected, particularly in a disease situation where the microbiota composition is altered towards less desirable species. In this perspective, abnormally expressed miRNAs could be considered novel therapeutic targets.
Acknowledgments

The authors would like to thank Wen Su and Stefania Arioli for help with dissections and fluorescence *in situ* hybridization. This study was funded by the J. P. Bickell Fdn Medical Research Program. Computational analyses were in part supported by Ontario Research Fund (GL2-01-030), Canada Foundation for Innovation (CFI #12301 and CFI #203383), and Canada Research Chair Program. NS was in part supported by an Ontario Graduate Scholarship (OGS).

Conflict of Interest

No conflicts of interest declared.

References:


10. Sonnenburg JL, Chen CT, Gordon JI. Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host. Plos Biology 2006; 4: e413.


43. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 2009; 25: 765-771.


Thesis References


Carey, V., Long, Li and Gentleman, R. (). *RBGL: An interface to the BOOST graph library.*

Gentleman, R., Whalen, E., Huber, W., Falcon, S. (). *graph: A package to handle graph data structures.*


188. *Online Mendelian Inheritance in Man, OMIM (TM),* McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD),


