Characterization of Altered MicroRNA Expression in Cervical Cancer

by

Christine How

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Medical Biophysics
University of Toronto

© Copyright by Christine How 2013
Characterization of Altered MicroRNA Expression in Cervical Cancer

Christine How

Doctor of Philosophy

Graduate Department of Medical Biophysics
University of Toronto

2013

Abstract

Cervical cancer is the third most common cancer among women worldwide, and the fourth leading cause of cancer mortality. Despite significant declines in the incidence and mortality rates of cervical cancer in Canada, it remains the 4th most common cancer in women aged 20-29 years. In order to gain novel insights into cervical cancer tumourigenesis and clinical outcome, we investigated and characterized the alterations in microRNA (miRNA) expression in this disease. Firstly, we performed global miRNA expression profiling of cervical cancer cell lines (n=3), and patient specimens (n=79). From this analysis, we identified miR-196b to be significantly down-regulated in cervical cancer, and characterized its role in regulating the HOXB7~VEGF axis. The global miRNA expression data also led to the development of a candidate 9-miRNA signature that was prognostic for disease-free survival in patients with cervical cancer, although we were unable to validate this signature in an independent cohort. This report describes important considerations concerning the development and validation of microRNA signatures for cervical cancer.

Our investigations also led us to a comparison of three methods for measuring miRNA abundance: the TaqMan Low Density Array, the NanoString nCounter assay, and single-well quantitative real-time PCR. Our findings demonstrated limited concordance between the TLDA and NanoString platforms, although each platform correlated well with PCR, which is considered the gold standard for nucleic acid quantification. Furthermore, we examined biases created by amplification protocols for microarray studies. Our analysis demonstrated that performing a
correction using the LTR-method (linear transformation of replicates) could help mitigate, but not completely eliminate such biases.

Overall, this report presents insights into the role of miRNAs in cervical cancer, as well as an evaluation of technical considerations concerning miRNA and mRNA expression profiling studies.
Acknowledgments

First and foremost, I would like to acknowledge my PhD supervisor, Dr. Fei-Fei Liu, for giving me the opportunity to pursue my graduate studies in her lab and collaborate with world-class scientists and clinicians. You have been an invaluable teacher and role model, and I am truly thankful for your mentorship and support during my PhD studies. My achievements and accolades could not have been possible without your continuous encouragement and guidance. Special thanks to my supervisory committee, Dr. Anthony Fyles and Dr. Mark Minden, for providing me with keen advice and valuable feedback. You have always encouraged and challenged me to think more broadly, which I will continue to strive to do in the future.

It has been a pleasure to work with all members of the Liu Lab, and there are a few current and past members whom I would like to acknowledge in particular: Angela Hui, for providing guidance and sharing your wealth of knowledge, you have played an instrumental role in conceptualizing and developing my projects. Willa Shi, for providing encouragement and advice, and assisting with specimen preparation and immunohistochemistry. Nehad Alajez, for your advice and assistance with experimental design and data analysis. Jeff Bruce, for always being willing and enthusiastic in assisting with my projects. Michelle Lenarduzzi, for being a great lab bench neighbor, conference roommate, and friend. Winnie Yue, for assisting with animal work and always offering encouragement. Other colleagues I would like to thank for their contributions to this thesis include: Melania Pintilie, Trudey Nicklee, Blaise Clarke, Rui Yan, and Paul Boutros.

Finally, I owe my sincere gratitude to my family and friends for always supporting me throughout the highs and lows of graduate school. To my mother (Mary) and father (Gary), thank you for your unconditional love, support and patience. I am humbled and grateful for everything you have done for me, and I would not be the person I am today without you. I hope I have made you proud. To my sister (Maxine) and brother (Andrew), thank you for keeping me motivated and reminding me to have fun. To my dearest friends, thank you for always trying to put a smile on my face, and for believing in me even when I didn’t believe in myself.
# Table of Contents

Abstract ........................................................................................................................................... ii  
Acknowledgments .......................................................................................................................... iv  
Table of Contents ............................................................................................................................. v  
List of Tables ................................................................................................................................... x  
List of Figures ................................................................................................................................ xi  
List of Abbreviations ................................................................................................................... xiii

CHAPTER 1: INTRODUCTION ....................................................................................................1  

1.1 MicroRNAs ..................................................................................................................................2  
1.1.1 Overview ..................................................................................................................................2  
1.1.2 History ..................................................................................................................................2  
1.1.3 Biogenesis and Processing ....................................................................................................2  
1.1.4 RNA-induced Silencing Complex (RISC) Assembly and Function ...............................6  
1.1.5 Target Prediction Databases ...............................................................................................7  
1.1.6 MicroRNAs in Cancer ........................................................................................................7  
1.1.7 MicroRNA Expression Profiling ..........................................................................................9  
1.1.8 Platforms for MicroRNA Expression Profiling ..............................................................10  
1.1.9 microRNAs as Biomarkers .............................................................................................12  

1.2 Cervical Cancer .......................................................................................................................13  
1.2.1 Background .......................................................................................................................13  
1.2.2 Human Papillomavirus .......................................................................................................14  
1.2.3 Treatment ............................................................................................................................16  
1.2.4 Gene Expression Signatures for Cervical Cancer ..........................................................16  
1.2.5 MicroRNA Expression in Cervical Cancer .................................................................17  

1.3 Homeobox Genes ....................................................................................................................18
## 1.3.1 Overview ..................................................................................................................18
## 1.3.2 Role in Development .............................................................................................20
## 1.3.3 Role in Cancer ........................................................................................................20

### 1.4 Vascular Endothelial Growth Factor .................................................................................21

#### 1.4.1 Overview ................................................................................................................21
#### 1.4.2 Receptors................................................................................................................21
#### 1.4.3 Isoforms .................................................................................................................22
#### 1.4.4 VEGF in Cervical Cancer ......................................................................................22

### 1.5 Research Objectives ...........................................................................................................24

## CHAPTER 2: MICRORN-196B REGULATES THE HOMEBOX-B7/VASCULAR ENDOTHELIAL GROWTH FACTOR AXIS IN CERVICAL CANCER ........................................................25

### 2.1 Chapter Abstract ................................................................................................................26
### 2.2 Introduction ........................................................................................................................26
### 2.3 Materials and Methods ......................................................................................................27

#### 2.3.1 Cell Lines and Transfections ................................................................................27
#### 2.3.2 miRNA Expression Profiling of Cell Lines ..........................................................27
#### 2.3.3 miRNA Expression Profiling of Patient Tissues ....................................................28
#### 2.3.4 Quantitative Real-time PCR Analysis of miRNAs and mRNAs ......................29
#### 2.3.5 Cell Viability, Proliferation and Colony-Formation Assays ................................31
#### 2.3.6 Cell Migration and Invasion Assays ....................................................................31
#### 2.3.7 Cell Cycle Analysis .................................................................................................31
#### 2.3.8 In vivo Experiments ...............................................................................................32
#### 2.3.9 Tri-modal Approach for miRNA Target Identification ...................................32
#### 2.3.10 Luciferase Reporter Assay .................................................................................33
#### 2.3.11 Immunoblotting ....................................................................................................33
#### 2.3.12 Enzyme-linked Immunosorbent Assay (ELISA) ..............................................33
### 2.3.13 5-aza-2’-deoxycytidine Treatment

### 2.3.14 Statistical Analysis

### 2.4 Results

2.4.1 miR-196b was significantly down-regulated in primary cervical cancer tissues and cell lines, and was strongly associated with DFS

2.4.2 miR-196b over-expression significantly reduced cell viability, clonogenicity, proliferation and invasion

2.4.3 miR-196b over-expression suppressed tumor angiogenesis and tumor cell proliferation *in vivo*

2.4.4 miR-196b directly targets HOXB7

2.4.5 VEGF is a relevant downstream target of HOXB7

2.4.6 Knockdown of HOXB7 or VEGF recapitulated the biological effects observed following mir-196b over-expression

### 2.5 Discussion

### CHAPTER 3: CHALLENGES ASSOCIATED WITH DEVELOPING A PROGNOSTIC MICRORNA SIGNATURE FOR HUMAN CERVICAL CARCINOMA

3.1 Chapter Abstract

3.2 Introduction

3.3 Materials and Methods

3.3.1 Clinical specimens

3.3.2 Sample processing

3.3.3 Data normalization

3.3.4 Survival analysis

3.4 Results

3.4.1 The majority of significantly differentially-expressed miRNAs were downregulated

3.4.2 Patient miRNA expression was associated with disease-free survival

3.4.3 Validation of miRNA signature

3.4.4 Independent corroboration of the Hu *et al.* 2-miRNA signature
3.5 Discussion .......................................................................................................................... 68

CHAPTER 4: IDENTIFICATION AND CORRECTION OF AMPLIFICATION PROTOCOL BIAS IN MICROARRAY STUDIES ................................................................. 73

4.1 Chapter Abstract ................................................................................................................ 74

4.2 Introduction ....................................................................................................................... 75

4.3 Materials and Methods ................................................................................................... 76

4.3.1 Samples .................................................................................................................. 76

4.3.2 Sample processing ................................................................................................. 77

4.3.3 Microarray hybridization ....................................................................................... 77

4.3.4 Pre-processing ........................................................................................................ 78

4.3.5 Unsupervised machine-learning ............................................................................. 78

4.3.6 Correlation analyses ............................................................................................... 79

4.3.7 Statistical analysis of microarray data ................................................................... 79

4.3.8 Gene subset division .............................................................................................. 80

4.3.9 Functional analysis ................................................................................................. 80

4.3.10 Sequence analysis .................................................................................................. 80

4.3.11 Visualization .......................................................................................................... 81

4.4 Results ................................................................................................................................ 81

4.4.1 Experimental approach .......................................................................................... 81

4.4.2 Nugen and express data are not directly comparable ............................................. 82

4.4.3 Protocol introduces systematic bias .......................................................................... 85

4.4.4 Protocol noise is larger than biological signal ....................................................... 85

4.4.5 Protocol difference biases specific biological functions ......................................... 88

4.4.6 Sequence characteristics do not explain protocol bias ........................................... 88

4.4.7 Batch effects adjustment partially mitigates protocol bias .................................... 91

4.4.8 Replication analysis ............................................................................................... 91
4.5 Discussion........................................................................................................................................93

CHAPTER 5: SYSTEMATIC COMPARISON OF PLATFORMS FOR MICRORNA PROFILING.................................................................103

5.1 Chapter Abstract ................................................................................................................................104

5.2 Introduction.......................................................................................................................................104

5.3 Materials and Methods....................................................................................................................106
  5.3.1 Samples and sample processing..................................................................................................106
  5.3.2 Data description ........................................................................................................................107
  5.3.3 Data pre-processing ...................................................................................................................109
  5.3.4 Unsupervised machine-learning.................................................................................................109
  5.3.5 Statistical analysis .....................................................................................................................109
  5.3.6 Visualization ............................................................................................................................110

5.4 Results.............................................................................................................................................110
  5.4.1 Experimental design..................................................................................................................110
  5.4.2 Biological differences are detectable by platforms....................................................................111
  5.4.3 Data from NanoString and TLDA are significantly different..................................................111
  5.4.4 PCR is more similar to TLDA than NanoString.........................................................................112
  5.4.5 Fold-changes are poorly correlated between NanoString vs. TLDA, but differentially-expressed miRNAs are highly correlated across platforms.........113

5.5 Discussion.......................................................................................................................................120

CHAPTER 6: DISCUSSION.........................................................................................................................................123

6.1 Research Summary ............................................................................................................................124

6.2 Future Directions ..............................................................................................................................124

6.3 Conclusions......................................................................................................................................126

References...............................................................................................................................................127


### List of Tables

<table>
<thead>
<tr>
<th>Table 2.1</th>
<th>Clinical parameters of 79 cervical cancer patients. .............................................. 29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.2</td>
<td>Primer Sequences used for qRT-PCR..................................................................................... 30</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Clinical parameters of patients in the training and validation cohorts............................ 62</td>
</tr>
<tr>
<td>Table S3.1</td>
<td>Significantly differentially-expressed miRNAs in cervical cancer ..................................... 71</td>
</tr>
<tr>
<td>Table S4.1</td>
<td>Sample descriptions (Data: cervix data) ............................................................................... 966</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Data descriptions for breast cancer datasets ........................................................................ 108</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Data descriptions for cervical cancer datasets .................................................................... 108</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Data descriptions for NPC datasets.................................................................................... 108</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure Reference</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>MicroRNA biogenesis and processing</td>
<td>4</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Map of HPV genome</td>
<td>15</td>
</tr>
<tr>
<td>Figure 1.3</td>
<td>Human HOX gene clusters</td>
<td>19</td>
</tr>
<tr>
<td>Figure 1.4</td>
<td>VEGF family members and receptors</td>
<td>23</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>miR-196b down-regulation in cervical cancer</td>
<td>35</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>\textit{In vitro} effects of miR-196b over-expression</td>
<td>37</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Identification of HOXB7 as an mRNA target of miR-196b</td>
<td>40</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Identification of VEGF as a target of the HOXB7 transcription factor</td>
<td>42</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Downstream effects of siHOXB7 and siVEGF</td>
<td>43</td>
</tr>
<tr>
<td>Figure S2.1</td>
<td>\textit{In vitro} effects of 5-aza-DCT treatment and miR-196b over-expression</td>
<td>48</td>
</tr>
<tr>
<td>Figure S2.2</td>
<td>Effect of miR-196b on \textit{in vivo} tumor growth</td>
<td>50</td>
</tr>
<tr>
<td>Figure S2.3</td>
<td>Effect of miR-196b on \textit{in vivo} angiogenesis and apoptosis</td>
<td>51</td>
</tr>
<tr>
<td>Figure S2.4</td>
<td>Tri-modal strategy for target identification</td>
<td>52</td>
</tr>
<tr>
<td>Figure S2.5</td>
<td>Transcript levels of putative miR-196b targets</td>
<td>53</td>
</tr>
<tr>
<td>Figure S2.6</td>
<td>\textit{In vitro} effects of treatment with siHOXB7, siVEGF, or pre-miR-196b</td>
<td>54</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Kaplan-Meier analysis of DFS according to 9-miRNA signature</td>
<td>64</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Application of 9-miRNA signature to validation cohort</td>
<td>65</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Kaplan-Meier analysis of DFS according to Hu \textit{et al.} 2-miRNA signature</td>
<td>67</td>
</tr>
<tr>
<td>Figure S3.1</td>
<td>Application of TCGA Small RNA-Seq data to cervical cancer signatures</td>
<td>72</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.1</td>
<td>Experimental design</td>
<td>833</td>
</tr>
<tr>
<td>4.2</td>
<td>Correlations between technical replicates</td>
<td>844</td>
</tr>
<tr>
<td>4.3</td>
<td>Specific genes are affected by protocol bias</td>
<td>866</td>
</tr>
<tr>
<td>4.4</td>
<td>Differentially expressed genes between biological and technical replicates</td>
<td>87</td>
</tr>
<tr>
<td>4.5</td>
<td>Differing characteristics of significantly differentially expressed genes</td>
<td>899</td>
</tr>
<tr>
<td>4.6</td>
<td>Application of LTR algorithm to remove platform specific biases</td>
<td>92</td>
</tr>
<tr>
<td>S4.1</td>
<td>Distributions of Selected Genes (Dataset: Cervix data)</td>
<td>97</td>
</tr>
<tr>
<td>S4.2</td>
<td>Overlap of significant GO terms by Protocol or sample type</td>
<td>98</td>
</tr>
<tr>
<td>S4.3</td>
<td>Correlations between technical replicates (Data: kidney data)</td>
<td>999</td>
</tr>
<tr>
<td>S4.4</td>
<td>Distributions of mRNA abundances (Data: kidney data)</td>
<td>100</td>
</tr>
<tr>
<td>S4.5</td>
<td>Correlation comparisons between cervix and kidney data.</td>
<td>101</td>
</tr>
<tr>
<td>S4.6</td>
<td>Distribution of Y-chromosome filter process for cervix data</td>
<td>102</td>
</tr>
<tr>
<td>5.1</td>
<td>Experimental design</td>
<td>114</td>
</tr>
<tr>
<td>5.2</td>
<td>Correlations between technical replicates (TLDA vs. NanoString)</td>
<td>1155</td>
</tr>
<tr>
<td>5.3</td>
<td>Distributions and correlations of miRNA abundances across platforms</td>
<td>1166</td>
</tr>
<tr>
<td>5.4</td>
<td>Correlations between technical replicates with PCR</td>
<td>117</td>
</tr>
<tr>
<td>5.5</td>
<td>Sample-wise and gene-wise correlations between PCR vs. NanoString and PCR vs. TLDA</td>
<td>118</td>
</tr>
<tr>
<td>5.6</td>
<td>Correlations of tumour/normal fold-changes between NanoString and TLDA</td>
<td>119</td>
</tr>
</tbody>
</table>
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>aCGH</td>
<td>Array comparative genomic hybridization</td>
</tr>
<tr>
<td>AGO</td>
<td>Argonaute</td>
</tr>
<tr>
<td>AGO2</td>
<td>Argonaute 2</td>
</tr>
<tr>
<td>ALL</td>
<td>Acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>AML</td>
<td>Acute myeloid leukemia</td>
</tr>
<tr>
<td>ANKHD1</td>
<td>Ankyrin repeat and KH domain containing 1</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>ARI</td>
<td>Adjusted rand index</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>AU</td>
<td>Adenine-uracil</td>
</tr>
<tr>
<td>BCL2</td>
<td>B-cell Lymphoma 2</td>
</tr>
<tr>
<td>BCL6</td>
<td>B-cell Lymphoma 6</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CD31</td>
<td>Cluster of differentiation 31</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CLL</td>
<td>Chronic lymphocytic leukemia</td>
</tr>
<tr>
<td>COX-2</td>
<td>Cyclooxygenase-2</td>
</tr>
<tr>
<td>Ct</td>
<td>Threshold cycle</td>
</tr>
<tr>
<td>CT</td>
<td>Computed tomography</td>
</tr>
<tr>
<td>CTDSP2</td>
<td>CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase 2</td>
</tr>
<tr>
<td>CRT</td>
<td>Chemo-radiotherapy</td>
</tr>
<tr>
<td>DCE-MRI</td>
<td>Dynamic contrast-enhanced magnetic resonance imaging</td>
</tr>
<tr>
<td>DFS</td>
<td>Disease-free survival</td>
</tr>
<tr>
<td>DGCR8</td>
<td>DiGeorge Syndrome Critical Region 8</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>eIF4E</td>
<td>Eukaryotic translation initiation factor 4E</td>
</tr>
<tr>
<td>eIF4G</td>
<td>Eukaryotic translation initiation factor 4 gamma</td>
</tr>
<tr>
<td>eIF6</td>
<td>Eukaryotic translation initiation factor 6</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>ER</td>
<td>Estrogen receptor</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal bovine serum</td>
</tr>
<tr>
<td>FDR</td>
<td>False-discovery rate</td>
</tr>
<tr>
<td>FFPE</td>
<td>Formalin-fixed paraffin-embedded</td>
</tr>
<tr>
<td>FGFR1</td>
<td>Fibroblast growth factor receptor 1</td>
</tr>
<tr>
<td>FIGO</td>
<td>International Federation of Gynecologists and Obstetricians</td>
</tr>
<tr>
<td>FLT1</td>
<td>Fms-related tyrosine kinase 1</td>
</tr>
<tr>
<td>FLT4</td>
<td>Fms-related tyrosine kinase 4</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>FOXO1</td>
<td>Forkhead box O1</td>
</tr>
<tr>
<td>GAPDH</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>GO</td>
<td>Gene ontology</td>
</tr>
<tr>
<td>GTP</td>
<td>Guanosine triphosphate</td>
</tr>
<tr>
<td>HDAC</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>HE</td>
<td>Hematoxylin and eosin</td>
</tr>
<tr>
<td>HER2</td>
<td>Human epidermal growth factor receptor 2</td>
</tr>
<tr>
<td>HOXA</td>
<td>Homeobox-A</td>
</tr>
<tr>
<td>HOXA7</td>
<td>Homeobox-A7</td>
</tr>
<tr>
<td>HOXA9</td>
<td>Homeobox-A9</td>
</tr>
<tr>
<td>HOXB</td>
<td>Homeobox-B</td>
</tr>
<tr>
<td>HOXB7</td>
<td>Homeobox-B7</td>
</tr>
<tr>
<td>HPV</td>
<td>Human Papillomavirus</td>
</tr>
<tr>
<td>Ig</td>
<td>Immunoglobulin</td>
</tr>
<tr>
<td>IGFR1</td>
<td>Insulin-like growth factor receptor 1</td>
</tr>
<tr>
<td>IMRT</td>
<td>Intensity modulated radiotherapy</td>
</tr>
<tr>
<td>KDR</td>
<td>Kinase insert domain receptor</td>
</tr>
<tr>
<td>KRT8</td>
<td>Keratin 8</td>
</tr>
<tr>
<td>LTR</td>
<td>Linear transformation of replicates</td>
</tr>
<tr>
<td>MAQC</td>
<td>MicroArray Quality Control</td>
</tr>
<tr>
<td>MEIS1</td>
<td>Myeloid ecotropic viral integration site 1 homolog</td>
</tr>
<tr>
<td>MEM</td>
<td>Minimal essential medium</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>MLL</td>
<td>Mixed lineage leukemia</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MTS</td>
<td>3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium</td>
</tr>
<tr>
<td>NC</td>
<td>Negative Control</td>
</tr>
<tr>
<td>NPC</td>
<td>Nasopharyngeal carcinoma</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>OCT</td>
<td>Optimal cutting temperature</td>
</tr>
<tr>
<td>OS</td>
<td>Overall survival</td>
</tr>
<tr>
<td>Pap</td>
<td>Papanicolaou</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PlGF</td>
<td>Placenta growth factor</td>
</tr>
<tr>
<td>pRb</td>
<td>Retinoblastoma protein</td>
</tr>
<tr>
<td>Pre-miRNA</td>
<td>Precursor microRNA</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Name</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Pri-miRNA</td>
<td>Primary microRNA</td>
</tr>
<tr>
<td>PTEN</td>
<td>Phosphatase and tensin homolog</td>
</tr>
<tr>
<td>PUM2</td>
<td>Pumilio homolog 2</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Quantitative real-time PCR</td>
</tr>
<tr>
<td>RIN</td>
<td>RNA integrity number</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>RMA</td>
<td>Robust multi-array average</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNAP II</td>
<td>RNA Polymerase II</td>
</tr>
<tr>
<td>RNAP III</td>
<td>RNA Polymerase III</td>
</tr>
<tr>
<td>RT</td>
<td>Radiotherapy</td>
</tr>
<tr>
<td>SCC</td>
<td>Squamous cell carcinoma</td>
</tr>
<tr>
<td>SCJ</td>
<td>Squamocolumnar junction</td>
</tr>
<tr>
<td>SCID</td>
<td>Severe combined immunodeficiency</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard error of the mean</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small interfering RNA</td>
</tr>
<tr>
<td>SLC9A6</td>
<td>solute carrier family 9, subfamily A (NHE6, cation proton antiporter 6), member 6</td>
</tr>
<tr>
<td>SMC3</td>
<td>Structural maintenance of chromosomes 3</td>
</tr>
<tr>
<td>SMG7</td>
<td>smg-7 homolog, nonsense mediated mRNA decay factor (C. elegans)</td>
</tr>
<tr>
<td>snoRNA</td>
<td>Small nucleolar RNA</td>
</tr>
<tr>
<td>SR140</td>
<td>U2 snRNP-associated SURP domain containing</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TLDA</td>
<td>TaqMan Low Density Array</td>
</tr>
<tr>
<td>TRBP</td>
<td>Transactivating response RNA-binding protein</td>
</tr>
<tr>
<td>TUNEL</td>
<td>Terminal deoxynucleotidyl transferase-mediated dUTP nick end labeling</td>
</tr>
<tr>
<td>TZ</td>
<td>Transformation zone</td>
</tr>
<tr>
<td>URR</td>
<td>Upstream regulatory region</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>VEGF</td>
<td>Vascular endothelial growth factor</td>
</tr>
<tr>
<td>VEGFR</td>
<td>Vascular endothelial growth factor receptor</td>
</tr>
<tr>
<td>XRN1</td>
<td>5’-3’ exoribonuclease 1</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION
1.1 MicroRNAs

1.1.1 Overview

MicroRNAs (miRNAs) are a class of short, non-coding RNAs that regulate gene expression in a sequence-specific manner (1, 2). These single-stranded RNAs of ~20-22 nucleotides in length are endogenous in plants and animals, and show cross-species evolutionary conservation (3). As of February 2013, there were 21,262 miRNAs catalogued in the miRBase database, including 2042 human miRNAs (miRBase Release 19, www.mirbase.org). According to bioinformatic analyses, up to 1/3 of all human genes are predicted to be regulated by miRNAs (4), rendering them as one of the largest classes of gene regulators in the mammalian system (5).

1.1.2 History

miRNAs were first described in Caenorhabditis elegans (C elegans) in 1993 (6). Ambros had previously observed that worms with mutations in the lin-4 and lin-14 genes did not develop normally (7). Upon further characterization of these two genes, Lee et al. discovered that the lin-4 gene was transcribed into a 22-nucleotide RNA (lin-4) that did not encode a protein (6). Furthermore, lin-4 was found to be complementary to a repeated sequence element in the 3’ untranslated region (UTR) of lin-14 messenger RNA (mRNA), and could regulate protein translation of lin-14 by binding to its 3’UTR.

The existence of miRNAs in higher organisms was first suggested in 2000 when the second miRNA, let-7, was identified in C elegans (8). Let-7 had several known homologues in higher organisms, including humans, and was reported to regulate lin-41 and lin-57/hbl-1 in C elegans (6, 9, 10). Interestingly, the term “microRNA” was first used to describe this novel class of small, non-coding RNAs in 2001, in three papers published simultaneously in volume 294(5543) of Science (11-13).

1.1.3 Biogenesis and Processing

The biogenesis and processing of miRNAs occurs via a complex pathway outlined in Figure 1.1. They are initially transcribed inside the nucleus, usually by RNA polymerase II (RNAP II) (14), to form large RNA precursors called primary miRNAs (pri-miRNAs) that can be longer than 1 kb (14, 15). Some miRNAs are transcribed by RNA polymerase III (RNAP III),
including those surrounded by Alu elements (16). Pri-miRNAs contain at least one hairpin structure, as well as a 3’ polyadenylated tail, and a 5’ 7-methylguanylate cap, similar to mRNA transcripts (17).

While still inside the nucleus, the hairpin structure of the pri-miRNA is recognized and bound by DiGeorge Syndrome Critical Region 8 (DGCR8), which recruits Drosha, an RNase III endonuclease, to form the microprocessor complex (18, 19). Guided by DGCR8, Drosha cleaves the pri-miRNA precisely 11 base pairs (bp) away from the base of the hairpin stem (20). This cleavage yields a ~60-70 nt stem-loop precursor miRNA (pre-miRNA) with a 2-nt overhang at the 3’ end, which is characteristic of products that have been cleaved by RNase III endonucleases, and important for downstream steps in the processing pathway (21). Exportin-5 recognizes the pre-miRNA based on its stem-loop structure and 3’ overhang (22, 23), and binds to the pre-miRNA with high affinity immediately after cleavage by Drosha (24). After binding, Exportin-5 transports the pre-miRNA to the cytoplasm using Ran-GTP (25, 26).

In the cytoplasm, the pre-miRNA is further cleaved by Dicer, another RNase III endonuclease (27). Dicer, aided by transactivating response RNA-binding protein (TRBP), cleaves off the pre-miRNA terminal loop, yielding a ~22-nt double-stranded miRNA duplex (28). Although Dicer alone is sufficient for this step, TRBP stabilizes Dicer and improves its processing efficiency (29). Only one strand of the miRNA duplex, referred to as the guide strand, is loaded onto the RNA-induced silencing complex (RISC), whereas the other strand, referred to as the passenger strand, is subsequently degraded. The guide strand was traditionally believed to be the strand with the less stable base pair at the 5’ end (30); however, it has been recently reported that either strand could be incorporated into the RISC in some instances (31).
miRNAs are initially transcribed inside the nucleus as large RNA precursors (pri-miRs), and then cleaved by the microprocessor complex into ~70 nt hairpin structures (pre-miRs). The pre-miRs
are exported to the cytoplasm where they are further cleaved by the RNase III enzyme Dicer into ~22 nt miRNA duplexes. The less stable of the two strands in the duplex, referred to as the “guide” strand, is incorporated into the RNA-induced silencing complex (RISC), while the opposite strand, referred to as the “passenger” strand, is destroyed. The miRNA-RISC complex can bind to target mRNAs with perfect complementarity, leading to mRNA degradation; or with imperfect complementarity, blocking protein translation. miRNA, microRNA; RNAP II, RNA polymerase II; pri-miRNA, primary microRNA; pre-miRNA, precursor microRNA; DGCR8, DiGeorge Syndrome Critical Region 8; TRBP, transactivating response RNA-binding protein; RISC, RNA-induced silencing complex.
1.1.4 RNA-induced Silencing Complex (RISC) Assembly and Function

The RISC is composed of the guide strand, Dicer, TRBP and Argonaute2 (AGO2) (32). TRBP remains bound to the mature miRNA duplex after cleavage by Dicer and recruits it to the AGO proteins (33). In humans, there are four Argonaute (AGO) proteins but only one member of this family, AGO2, harbours endonuclease activity and is the catalytic unit of the RISC (34-36). The RISC machinery binds via the guide strand to complementary target site(s) in the 3’UTR of the target mRNA (37), or in some cases, located in the 5’UTR (38), or even coding regions (39). The seed region, a 6 to 8 nt sequence at the 5’ end of the guide strand, determines the specificity of this binding (2, 40). Seed regions show a high degree of complementarity with their corresponding binding sites in target mRNAs, and are highly conserved across species (41, 42).

miRNAs regulate their target mRNAs by one of two methods: destabilization of the target mRNA, or translational repression. Destabilization of the target mRNA occurs when the seed region of the guide strand has perfect complementarity with binding site(s) in the target mRNA. AGO2 in the RISC initiates the mRNA degradation process by cleaving the target mRNA via its endonuclease domain (43). The cleaved target mRNA is subsequently degraded by 5’-3’ exoribonuclease 1 (XRN1) and the exosome (44).

When the seed region of the guide strand has partial complementarity with the binding site(s) in the target mRNA, translational repression of the target mRNA occurs, either by inhibition of translational initiation or repression of translational elongation. Inhibition of translational initiation is a 5’ cap-dependent process (45). Normally, translational initiation begins when the 5’ cap of the mRNA is recognized by eukaryotic translation initiation factor 4E (eIF4E), a component of the translational initiation complex (46). It has been reported that AGO2 from the RISC and eIF4E are similar in sequence, which allows AGO2 to displace eIF4E, thereby inhibiting translational initiation (47). In another study, Chendrimada et al. showed that miRNAs can also inhibit translational initiation by blocking ribosome assembly (48). The authors demonstrated that RISC associates with the 60S large ribosomal subunit and eIF6, which is a ribosome inhibitory protein known to prevent the 40S and 60S ribosomal subunits from assembling together to form the 80S ribosome. The evidence that translational repression can occur during elongation came from polyribosome experiments with miRNAs, which showed that
active translation of the target was occurring even though the protein product could not be detected (49-51). Unlike repression of translational initiation, the repression of translational elongation is a 5’ cap-independent process (51).

1.1.5 Target Prediction Databases

With their ability to bind to target mRNAs with either perfect or partial complementarity, each miRNA can potentially bind to numerous target mRNAs; conversely, each mRNA can potentially be regulated by numerous miRNAs (52). Hence, one of the major challenges in miRNA biology is to accurately identify mRNA targets that are truly regulated by any single miRNA. Multiple publically-available in silico databases have been generated to assist in identifying targets of miRNAs, using algorithms that incorporate criteria such as seed sequence alignment (40), thermodynamic stability (53), conservation in different species (54), or binding site accessibility (55). Some algorithms also consider the contribution of multiple binding sites and their proximity to each other within the 3’UTR of the target mRNA (56, 57), or enrichment of AU base-pairs in the 30 nt region flanking the binding site (58).

Currently there are over ten miRNA target prediction databases, each utilizing a unique algorithm for determining putative target mRNAs, including: Targetscan (58), RNA22 (59), PicTar (60), Diana-microT (61), miRDB (62), GenMir++ (63), miRanda (64, 65), TarBase (66), PITA (55), and miRBase (4). These target prediction databases often provide hundreds of candidate target mRNAs for any single miRNA; however, the algorithms used have been shown to have high false discovery and high false negative rates (40, 67). Therefore, functional experiments are necessary to validate putative target mRNAs that were determined using in silico prediction databases. We used a tri-modal approach for miRNA target identification (68), which combines: i) in silico target prediction; ii) clinical gene expression data from patients; and iii) in vitro functional experiments. This tri-modal approach is described in further detail in Chapter 2.

1.1.6 MicroRNAs in Cancer

It is not surprising that alterations in miRNA expression have been linked to most human cancers, given the ability of miRNAs to regulate the expression of genes involved in important cellular activities, including cell proliferation and apoptosis (69). Over-expression of oncogenic miRNAs can contribute to cancer progression by down-regulating target genes that function as
tumour suppressors. For example, miR-21, which is frequently over-expressed in many types of solid tumours, targets the tumour suppressor phosphatase and tensin homolog (PTEN) (70), and pro-apoptotic genes (71, 72). Conversely, under-expression of tumour suppressor-like miRNAs can result in up-regulation of oncogenes. In chronic lymphocytic leukemia (CLL), miR-15a and miR-16-1 are often down-regulated, and have been reported to target B Cell Lymphoma 2 (BCL2) (73), a well-known oncogene (74).

There are multiple mechanisms that can lead to dysregulation of miRNA expression, including: epigenetic alterations (75-77), chromosomal aberrations (73, 78), and deregulated expression of miRNA processing enzymes (79-81). Altered miRNA expression has been attributed to epigenetic events for a number of human malignancies, which is unsurprising considering bioinformatic analyses have demonstrated that 13 and 27% of human miRNA genes are located within 3 and 10 kb of a CpG island, respectively (76). Aberrant CpG methylation can lead to silencing of tumour suppressor-like miRNAs, such as miR-376a and miR-376c in melanoma (77). The authors of this study showed that both of these miRNAs were downregulated in melanoma and directly targeted insulin-like growth factor receptor 1 (IGFR1), which may contribute to IGFR1 overexpression in melanoma and thereby promote tumourigenesis and metastasis. Another study reported that miR-127 was epigenetically silenced in bladder cancer, and expression could be reactivated in vitro after treatment with 5-aza-2'-deoxycytidine and 4-phenylbutyric acid, which inhibit DNA methylation and histone deacetylase, respectively (82). Reactivation of miR-127 expression was accompanied by a corresponding downregulation of its target B-cell CLL/lymphoma 6 (BCL6), which is a known proto-oncogene.

Approximately half of miRNA genes are located in fragile chromosomal sites, which are prone to chromosomal aberrations such as amplifications, deletions or translocations (83). For example, the oncogenic miR-17~92 cluster is located in a region (13q31.3) that is frequently amplified in several human cancers such as Burkitt’s lymphomas, follicular lymphomas, diffuse large B-cell lymphomas, and lung cancer (84). As expected, overexpression of the miR-17~92 cluster, which encodes six miRNAs (miR-17, miR-18a, miR-19a, miR-20a, miR-19b-1, and miR-92-1), is frequently observed in a number of human cancers including lymphomas and lung cancer (85, 86). Conversely, the aforementioned tumour suppressor-like miR-15a and miR-16-1
are located in a region (13q14.3) that has been reported to be deleted in several hematopoietic and solid malignancies (73, 74).

A number of studies have reported dysregulation of miRNA processing enzymes, either up- or down-regulation, in various cancers, including colorectal (87, 88), skin (89, 90), renal (91), and ovarian (92). In colorectal cancer, down-regulation of Dicer was significantly associated with higher FIGO stage, tumour grade and nodal metastasis, and low Dicer expression was correlated with shortened overall survival, independent of other prognostic factors (87). Merritt et al. demonstrated that Dicer and Drosha were down-regulated in over 50% of the 111 ovarian cancer patients in their study (92). Furthermore, patients with low expression of both Dicer and Drosha had worse overall survival, again independent of other prognostic factors.

1.1.7 MicroRNA Expression Profiling

miRNA expression profiling of human cancers has been used to distinguish between different types of cancers (86), as well as categorize cancer sub-types (93, 94), including poorly-differentiated tumours (95). Due to their small sizes, miRNAs are extremely stable and can be readily extracted from cell lines and various types of clinical specimens, including frozen and formalin-fixed paraffin embedded (FFPE) tissues, blood, serum, plasma, urine, and saliva (96-102). The ability to utilize FFPE specimens for miRNA expression profiling is a powerful tool for biomarker discovery, since FFPE processing is a universally standard procedure, and FFPE specimens are often linked to clinical databases with extensive patient follow-up. In contrast, there are many technical limitations associated with mRNA expression profiling using FFPE specimens, since mRNAs are significantly longer than miRNAs; thus are easily degraded during formalin fixation (103, 104).

Several reports have demonstrated highly concordant data when the same platform is used to measure the expression of a panel of miRNAs in tissue-matched frozen and FFPE tissue specimens; this has been documented for kidney, breast, prostate, lung, skin, glioblastoma and melanoma specimens (96, 105-112). However, some of these studies have also shown poorly correlated patterns of under- or over-expression for select miRNAs in the frozen vs. FFPE specimens (105, 108, 109, 111, 112). For example, Leite et al. analyzed a panel of 14 miRNAs in five prostate cancer specimens (5 frozen, 5 FFPE) and reported that four of the 14 miRNAs demonstrated opposing patterns of tumour-normal fold changes in tissue-matched sample pairs.
Furthermore, Weng et al. reported that miRNAs with significant tumour-normal fold changes were poorly correlated between tissue-matched frozen vs. FFPE renal carcinoma samples; when the lists of dysregulated miRNAs were compared between matched samples, the correlation peaked at 55% for the top 20 miRNAs (108).

Most methods of miRNA expression profiling utilize total RNA that contains small RNA (<200 nt), and do not require separation or enrichment of small RNAs in the sample. Methods such as TRIzol and the Ambion mirVana miRNA Isolation Kit use phenol:chloroform extraction followed by alcohol precipitation. Other methods use column-based RNA extraction, including the Norgen Total RNA Purification Kit and the Ambion RecoverAll Total Nucleic Acid Isolation Kit. Furthermore, methods have been developed and optimized for extracting high-quality RNA from FFPE specimens, such as the Ambion RecoverAll Total Nucleic Acid Isolation Kit, which was shown to recover the most amplifiable RNA in a study that compared ten RNA extraction methods for FFPE samples (113).

1.1.8 Platforms for MicroRNA Expression Profiling

Various platforms for global miRNA expression profiling have been developed, including the Applied Biosystems TaqMan Low Density Array Human MicroRNA Array (TLDA), NanoString nCounter Human miRNA Expression Assay (NanoString), and deep sequencing. Individual quantitative real-time PCR (qRT-PCR), which many consider to be the gold standard for nucleic acid quantification, is most often used to examine the expression of a panel of selected miRNAs, since it is difficult to analyze a large number of miRNAs simultaneously using this method. Due to their small sizes, miRNAs are not compatible with standard PCR primers that require a minimum template size of 40 nt. miRNA-specific stem-loop primers elongate the template RNA during the reverse transcription step, and are specific enough to discriminate between mature miRNAs that differ by a single nucleotide (114). During the PCR stage, the reverse transcription product is subjected to 40 cycles where TaqMan probes hybridize with their target miRNA sequences. The raw data consists of the threshold cycle (Ct), which is defined as the cycle number at which the fluorescence signal emitted passes the threshold. Hence, there is an inverse relationship between Ct values and the amount of template RNA in the
sample; higher Ct values indicate lower levels of the template miRNA, whereas lower Ct values indicate higher levels of the template miRNA.

The TLDA method simultaneously measures the expression of 377 miRNAs and four endogenous controls using a microfluidic card. This method utilizes the same stem-loop primers as individual qRT-PCR, in a multiplex reverse transcription step. The reverse transcription product is then loaded onto the microfluidic card with 384 wells, which are each pre-loaded with PCR primers specific for one target sequence (there is one well per miRNA, and four wells for the U6 control). The same TaqMan chemistry as individual qRT-PCR is used to measure the level of each miRNA, and the raw data is also outputted as Ct.

NanoString is a method that provides a direct count of miRNAs using optical quantification of fluorescently-tagged miRNA molecules. First, the sample is prepared by ligating the miRNA to a miRtag sequence. The miRtagged miRNA is then hybridized to a probe pair consisting of a reporter probe, which carries the fluorescent signal/barcode, and a capture probe, which allows the complex to be immobilized in the next step. Excess probes are removed during the purification step and then probe-target miRNA complexes are immobilized. In the final step, the colour barcode for each target miRNA is counted by the nCounter Digital Analyzer. The raw data consists of a direct count of the number of target miRNA molecules, and there is no bias due to amplification steps or enzymatic reactions.

Deep sequencing has been noted for its increased specificity and sensitivity, as well as its ability to detect both known and novel miRNAs (115, 116). The first step of this method is the reverse transcription of small RNAs to a cDNA library. Adapters are then ligated to the cDNAs, which allows the library either to be affixed to beads for emulsion PCR (e.g. Applied Biosystems SOLiD, Roche GS FLX), or a surface for solid-phase PCR (e.g. Illumina RNA-Seq). Next, millions of individual cDNAs from the library are sequenced in parallel, with the number of sequence reads for each small RNA species determined digitally. For each miRNA, the relative quantification is provided, which is the number of sequence reads for a given miRNA relative to the total reads in the sample.
1.1.9 **microRNAs as Biomarkers**

Biomarkers are used in the clinic to guide treatment decisions. A prognostic biomarker indicates the likely clinical outcome of a patient with a particular disease, regardless of treatment, whereas a predictive biomarker indicates the likely response of a patient to a particular treatment. Established biomarkers include estrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2), which are both used for breast cancer patients. The absence or presence of ER is used to determine whether or not a patient will respond to treatment with tamoxifen, which targets the ER in breast tissue. HER2 is used to predict sensitivity to trastuzumab, a monoclonal antibody against HER2. Patients with HER2 positive tumours, about 15% of all primary breast cancers, benefit significantly from trastuzumab therapy (117).

Despite the large quantity of reports describing prognostic factors for cancer, very few of these prognostic factors have been developed as biomarkers for clinical use. Simon et al. described the key steps required for the development and validation of therapeutically relevant biomarkers for the clinic (118). Firstly, a biomarker should be developed using a training cohort in which patients are adequately homogenous and are receiving the same treatment. Next, the biomarker should be internally validated, and its prognostic value should be compared to standard prognostic factors for the given disease. If the biomarker performs better than existing standard prognostic factors, it should be translated to a platform that can be broadly used in the clinic. The biomarker assay should be standardized and highly reproducible. Finally, the biomarker should be independently validated with an external cohort of patients. The steps required to develop and validate clinical biomarkers require robust experimental design and can be very costly, which explains why so few prognostic factors have been developed into clinical biomarkers.

Given their notable stability in clinical specimens, miRNAs could potentially be used as biomarkers for cancer. Multiple reports to date have demonstrated a correlation between miRNA expression and clinical outcome in cancer patients. The first reported miRNA signature was a 13-miRNA set associated with prognosis and progression in patients with CLL (119). Since then, a number of prognostic miRNA signatures have been published for various cancers, including cervical (120), colorectal (121), esophageal (122), gastric (123), hepatocellular (124), lung (125-127), nasopharyngeal (128), prostate (129), and acute myeloid leukemia (AML) (130).
These studies highlight the prognostic value and potential translational impact of miRNA expression profiling in cancer.

1.2 Cervical Cancer

1.2.1 Background

The cervix is covered by both mucin-secreting columnar epithelium, which line the endocervix, and non-keratinising stratified squamous epithelium, which line the ectocervix. These two types of epithelia meet at the squamocolumnar junction (SCJ). The original SCJ is the location at which the endocervix and ectocervix meet at birth, while the new SCJ consists of SCJ that is newly formed by the continuous remodeling process that occurs throughout the life of a female. The transformation zone (TZ), which is the border between the old and new SCJ, is the most common area where cervical cancer develops. The TZ undergoes metaplasia under normal physiological conditions, which makes cells in the TZ more susceptible to malignant transformation. Squamous cell carcinoma (SCC) and adenocarcinoma are the two most common histological types, comprising 80-90% and 10-15% of all cervical cancers, respectively. SCC’s arise from the squamous epithelial cells of the ectocervix, whereas adenocarcinomas arise from the columnar epithelial cells of the endocervix. The International Federation of Gynecology and Obstetrics (FIGO) system is used for staging cervical cancer, which is primarily based on clinical examination.

Globally, cervical cancer is the third most common cancer amongst women, with 530,000 new cases and 275,000 deaths each year, making it the fourth leading cause of cancer mortality (131). The vast majority of these cases and deaths (greater than 85%) occur in low income countries in South America, Africa and South-Central Asia. In high income countries, such as Canada, the United States and Australia, cervical cancer incidence and mortality rates have fortunately declined significantly over the past few decades, primarily due to widespread Papanicolaou (Pap) test screening programs that allow for early detection of pre-cancerous lesions (132, 133). Although the incidence and mortality rates of cervical cancer in Canada have been declining by 1.4% and 2.9% per year, respectively; it still remains as the 4th most common cancer in women aged 20-29 years in this country (134).
1.2.2 Human Papillomavirus

Infection with human papillomavirus (HPV) has been identified as a necessary cause of cervical cancer (135), with 99.7% of all cervical cancers being HPV-positive (136). The connection between HPV and cervical cancer was first postulated in 1974 due to the morphologic changes in the cancer lesions which were suggestive of a viral infection (137). About a decade later, the causal relationship was established when HPV16 DNA was identified and cloned in cervical cancer lesions by Professor Harold zur Hausen (138). Due to the rigour of his science, and the impact of this discovery, which subsequently led to the development of the HPV vaccine (139, 140), Professor zur Hausen was awarded the Nobel Prize in Physiology and Medicine in 2008 (http://www.nobelprize.org/nobel_prizes/medicine/laureates/2008/). To date, there are 15 high-risk HPV subtypes (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73 and 82) that are considered carcinogenic (141), with HPV16 and 18 accounting for 70% of all HPV-positive cases of cervical cancer (135).

HPV is a small, non-enveloped DNA virus with a double-stranded, circular genome of approximately 8 kb (Figure 1.2). The HPV genome encodes six non-structural regulatory proteins (E1, E2, E4, E5, E6 and E7) from the early region, and two structural viral capsid proteins (L1 and L2) from the late region (142). The upstream regulatory region (URR) contains the viral origin of replication, and encodes cis-acting elements that regulate viral gene transcription and synthesis of viral DNA. E1 and E2 are involved in viral DNA replication; E1 encodes an ATP-dependent helicase, and E2 encodes a transcription factor. E4 facilitates viral assembly and release, while E5 enhances cell immortalization.

The E6 and E7 viral oncoproteins mediate cervical cancer development by inactivating the p53 and retinoblastoma protein (pRb) tumour suppressor proteins, respectively. P53 induces cell cycle arrest upon recognizing DNA damage, and can initiate apoptosis if the damage is irreparable. E6 ubiquitinates p53, which results in its degradation through the proteasomal pathway. pRb binds to and inhibits E2F transcription factor, which prevents the cell from replicating damaged DNA. However, when E7 competes for pRb, E2F is free to transactivate its target genes, which facilitates the transition from G1 to S phase, thereby promoting cell cycle progression.
The HPV genome, indicated by the circle, is a 7,857 bp double-stranded DNA genome. The coloured boxes outside the circle represent the open reading frames that encode the viral proteins. The upstream regulatory region (URR) contains the viral origin of replication, and encodes cis-acting elements that regulate viral gene transcription and synthesis of viral DNA.
1.2.3 Treatment

Cervical cancer patients with non-bulky, early-stage disease can be effectively treated with surgery alone, achieving 5-year overall survival (OS) rates of ~80% (143). Unfortunately, patients with high-risk, early-stage cervical cancer, such as those with lymph node metastases, parametrial invasion, or deeply invasive lesions, fare much worse with surgery, with 5-year OS rates ranging from 50-70% (144-148).

Radiotherapy (RT) has been used to treat cervical cancer patients with advanced-stage disease since the 1930’s (149). Since then, many improvements have been made to RT, and chemotherapy has been added to the standard treatment regimen, based on a recommendation made by the National Cancer Institute in 1999. This decision was made after five major randomized clinical trials had provided consistent evidence that RT combined with concurrent chemotherapy was superior to RT alone, in terms of patient survival (150-154). Hence, current standard therapy for advanced stages of cervical cancer consists of concurrent chemo-RT (CRT) using cisplatin. Typically, external-beam RT is applied to the primary cervical tumour and pelvic lymph nodes (45 to 50 (Gy) total, in 1.8-to-2.0 Gy daily fractions with 18-to-25-MV photons), combined with weekly doses of cisplatin (40 mg/m² total, 5 doses).

1.2.4 Gene Expression Signatures for Cervical Cancer

Prognostic gene signature development for cervical cancer has been slow compared to other types of cancers, likely due to the low incidence of this disease in high income countries with research capabilities. To date, there have been four gene signatures reported for cervical cancer, but none of them have been validated in an independent patient cohort. The earliest reported gene signature was a 58-gene (mRNA) set associated with recurrence in cervical cancer patients treated with CRT (155). The second was a 7-gene (mRNA) set to predict pelvic relapse, derived from a group of patients participating in a Phase II clinical trial evaluating the benefit of a COX-2 inhibitor Celecoxib in addition to CRT using cisplatin (156). A third signature consisted of a 7-gene set used to identify patients who were likely to respond to RT alone; hence could be spared from the toxicity induced by the addition of chemotherapy (157). Another unvalidated signature was a 63-gene set associated with metastasis of advanced tumours following treatment with RT (158). The clinical relevance of these gene signatures remains unknown, since none have been validated in an independent cohort. Furthermore, all of these
studies used either frozen tissues (156, 158) or fresh tissues that were immediately stored in RNAlater stabilization solution upon tissue collection (155, 157); neither of which are routine practice or cost-efficient compared to FFPE tissues.

1.2.5 MicroRNA Expression in Cervical Cancer

The first report to examine miRNA expression in cervical cancer used direct sequencing to compare global miRNA expression in six cervical cancer cell lines and five normal cervix tissue samples, and identified six differentially expressed miRNAs (one upregulated: miR-21; five downregulated: miR-143, miR-196b, let-7b, let-7c, miR-23b) (159). When the expression of two of these dysregulated miRNAs, miR-21 and miR-143, were examined in 29 matched pairs of cervical cancer and normal tissues, it was observed that the expression variation in the cell lines was reproduced in the tissue samples. The second published study identified 70 miRNAs that were differentially expressed in ten early-stage invasive squamous cell carcinomas (SCC’s) compared to ten normal cervical epithelial samples (160). Furthermore, up regulation of miR-127 was found to be significantly associated with lymph node metastases and was validated in an independent group of 31 patient samples.

Using age-matched cervical cancer and normal cervix tissue samples, Wang et al. reported significant down regulation of miR-126, miR-143 and miR-145, along with up regulation of miR-15b, miR-16, miR-146a and miR-155 (161). Functional experiments in cervical cancer cell lines demonstrated that miR-146a promoted cell proliferation, whereas miR-143 and miR-145 inhibited cell growth. A study by Pereira et al. examined miRNA expression in four SCC’s, five high-grade squamous intraepithelial lesions, nine low-grade squamous epithelial lesions, and 19 normal cervix tissues, but could not derive a miRNA signature for cervical cancer due to high variability in miRNA expression, particularly among the normal cervix samples (162). Hu et al. reported a 2-miRNA signature (miR-200a and miR-9) that was associated with overall survival (OS) (120). This was the only study that had an adequate number of patient samples to develop a prognostic miRNA signature (training set: n=60; validation set: n=40).

A recent report described five dysregulated miRNAs that were directly linked with frequent chromosomal alterations in cervical cancer, namely miR-9 (1q23.2), miR-15b (3q25.32), miR-28-5p (3q27.3), miR-100 and miR-125b (both 11q24.1) (163). Functional
experiments performed in vitro demonstrated that overexpression of miR-9, which is linked to a chromosomal gain of 1q, promoted cell viability, anchorage-independent growth, and cell migration, supporting the oncogenic role of miR-9. Another recent study identified nine down regulated (miR-211, miR-145, miR-223, miR-150, miR-142-5p, miR-328, miR-195, miR-199b, and miR-142-3p), and two up regulated (miR-182 and miR-183) miRNAs in cervical cancer cell lines (164). The most up regulated miRNA, miR-182, was characterized further and shown to be significantly upregulated in cervical cancer tissues, correlating with advanced stage of disease. In vitro experiments demonstrated that miR-182 was involved in apoptosis and associated with forkhead box O1 (FOXO1) regulation, and in vivo experiments confirmed that inhibiting miR-182 could reduce tumour growth.

1.3 Homeobox Genes

1.3.1 Overview

The Homeobox genes, which were first reported in 1984, are a family of regulatory genes that encode transcription factors that play important roles in directing the formation and patterning of body structures during embryonic development (165, 166). Homeobox genes contain a 180 bp homeobox sequence encoding a highly conserved 60-residue helix-turn-helix protein domain, called the homeodomain (166). The homeodomain is a DNA-binding motif that recognizes and binds to DNA elements in a sequence-specific manner (167). Homeobox genes are contained in genomic clusters, and are expressed in a spatial and temporal manner (168).

Human class I human homeobox (HOX) genes consist of 39 genes organized in four clusters (HOXA, HOXB, HOXC and HOXD) located on different chromosomes, 7p, 17q, 12q, and 2q, respectively (Figure 1.3). Within each cluster, the HOX genes are assigned to 13 paralog groups, with each cluster containing 9 to 11 HOX genes.
Figure 1.3  Human HOX gene clusters
Schematic representation of the 39 human HOX genes assigned to 13 paralog groups in four HOX clusters (A-D) located on different chromosomes (7p, 17q, 12q, 2q). During embryonic development, the HOX genes are expressed sequentially from the 3’ to 5’, with the early genes (shown in orange/yellow) being expressed in anterior regions, and late genes (shown in blue) being expressed in posterior regions.
1.3.2 **Role in Development**

During embryogenesis, HOX genes are fundamentally important for proper development along the anterior-posterior axis (169). The HOX genes are expressed sequentially from the 3’ to the 5’ end of each cluster, with 3’ end genes, such as HOXB1, expressed early in development in anterior regions, and 5’ end genes, such as HOXB13, expressed later in development in the posterior regions (170). HOX genes have also been shown to play a role in directing patterns development in limbs (171, 172), lungs (173-175), and genitals (176, 177).

1.3.3 **Role in Cancer**

The expression of HOX genes in normal adult tissues indicates a role for these genes beyond embryonic development (178). Takahasi *et al.* reported that HOX genes were expressed in adult organs in a highly specific manner, and deregulated HOX gene expression was associated with tumour development (178). Given that HOX genes have been shown to play important roles in normal hematopoiesis (179-181), it is thus not surprising that aberrant HOX gene expression has been linked to leukemias (182, 183). Ferrando *et al.* reported consistent up regulation of HOXA9, HOXA10 and HOXC6 in acute lymphoblastic leukemia (ALL) patients with chromosomal translocations involving the mixed lineage leukemia (MLL) gene (184). In a microarray study of 6817 genes tested in patients with acute myeloid leukemia (AML), HOXA9 was identified as a diagnostic marker, and was the only gene that was strongly correlated with clinical outcome (185). Furthermore, inhibition of HOXA9 was shown to reduce proliferation and induce apoptosis in AML, especially in samples with MLL translocations (186).

Aberrant HOX gene expression has also been reported in solid tumours. Three mechanisms of HOX gene deregulation in cancer have been described by Abate-Shen (187). In the first mechanism, which is responsible for the majority of cases of HOX gene deregulation, HOX genes that are normally only expressed during embryogenesis are re-expressed in tumour cells. Secondly, HOX genes can be aberrantly expressed in tumour cells derived from tissues in which the gene is not normally expressed. The third and final mechanism involves down regulation of HOX genes in tumour cells derived from tissues in which the gene is normally expressed.
In prostate cancer, HOXC4, HOXC5, HOXC6 and HOXC8 were found to be overexpressed in tumours and nodal metastases (188). Interestingly, HOXC8 overexpression has been linked with loss of differentiation in prostate cancer (189), as well as increased cell invasion in vitro (190). HOXA5 under expression in breast cancer was associated with loss of p53 transcript and protein levels (191). In renal carcinoma, HOXB5 and HOXB9 have been reported to be down regulated, whereas HOXB2 was up regulated compared to normal kidney tissues (192).

1.4 Vascular Endothelial Growth Factor

1.4.1 Overview

The Vascular endothelial growth factor (VEGF) family proteins are secreted platelet-derived growth factors (193). There are five members of the VEGF family: VEGF-A, VEGF-B, VEGF-C, VEGF-D and placenta growth factor (PlGF) (Figure 1.4) (194). VEGF-A was the first to be discovered over 20 years ago, and is the most well-characterized family member (195). It was identified as a key regulator of signaling pathways involved in vasculogenesis and angiogenesis of endothelial cells (196). VEGF-A also induces chemotaxis in various types of cells (197-199), and stimulates vasodilation mediated via nitric oxide release from endothelial cells (200). VEGF-B shares high sequence homology with VEGF-A, and was initially described to be an angiogenic factor (201). However, more recent studies have shown that VEGF-B can both positively and negatively regulate angiogenesis, depending on the conditions (202). The role of VEGF-B remains unclear at this time; VEGF-C and VEGF-D appear both to be involved in angiogenesis and lymphangiogenesis (203, 204).

1.4.2 Receptors

The VEGF family members bind to and activate three VEGF receptors (VEGFR’s): VEGFR-1 (FLT1), VEGFR-2 (KDR), and VEGFR-3 (FLT4) (205, 206). Each VEGF ligand has distinct binding specificities for each of the VEGFRs, which contributes to their diverse functions (Figure 1.4). The VEGFRs are receptor tyrosine kinases with an extracellular portion composed of seven immunoglobulin (Ig)-like domains, a single membrane-spanning region, and an intracellular portion with a conserved intracellular tyrosine kinase domain containing a kinase insert (207, 208). FLT1 has the highest affinity for VEGF and can also bind VEGF-B and PlGF. This receptor is expressed on monocytes, macrophages and hematopoietic stem cells. VEGFR-2
is the most widely expressed receptor and also has a high affinity for VEGF. It is mainly responsible for mediating mitogenesis and angiogenesis. VEGFR-3 is mainly expressed on the lymphatic endothelium, and is involved in lymphangiogenesis.

1.4.3 Isoforms

VEGF-A, herein referred to as VEGF, undergoes alternative splicing to yield different isoforms which are denoted as VEGF_{xxx}, where xxx represents the number of residues present in the protein, excluding the signaling peptide. To date, eight VEGF isoforms have been identified in humans, ranging in length from 121 to 206 residues (209). The most common and well-studied isoforms are VEGF_{121}, VEGF_{165} and VEGF_{189}, which are expressed in most VEGF-producing cells (210). The five minor isoforms (VEGF_{111}, VEGF_{145}, VEGF_{148}, VEGF_{183} and VEGF_{206}) are less common, and their functions are not clearly understood (211-216).

1.4.4 VEGF in Cervical Cancer

VEGF levels in patients undergoing radiotherapy for cervical cancer, as measured in serum by ELISA (217, 218), have been shown to be a prognostic marker of disease-free survival (DFS), whereby high pre-treatment VEGF levels are associated with worse DFS. Bevacizumab is a humanized anti-VEGF monoclonal antibody (A.4.6.1) that recognizes all biologically active isoforms of VEGF and prevents them from binding to their receptors. This drug has been shown to be effective in treating several malignancies by blocking angiogenesis (219). Thus far, bevacizumab has been approved by the U.S. Food and Drug Administration (FDA) for treatment of metastatic colorectal cancer, non-small cell lung cancer, glioblastoma, and metastatic kidney cancer. Recently, the National Institutes of Health announced the results of a multi-centred randomized Phase III clinical trial (GOG240), where Bevacizumab was evaluated in patients with advanced, recurrent or persistent cervical cancer who could not be cured with standard treatment (220). The preliminary data from this trial reported that the addition of Bevacizumab to chemotherapy with cisplatin plus paclitaxel significantly improved OS by a median of 3.7 months.
The five VEGF family members (VEGF-A, VEGF-B, VEGF-C, VEGF-D, and PlGF) and the three principal VEGF receptors (VEGFR-1, VEGFR-2, VEGFR-3). VEGF-A binds to both VEGFR-1 and VEGFR-2. VEGF-B and PlGF bind to only VEGFR-1. VEGF-C and VEGF-D bind to both VEGFR-2 and VEGFR-3.
1.5 **Research Objectives**

Novel insights are required, to develop new therapeutic approaches that could improve clinical outcome for patients with advanced stages of cervical cancer. Altered miRNA expression has been shown to play a significant role in many human malignancies (69), including cervical cancer (120, 159, 161, 162), and target genes that contribute to tumor progression have been described for several miRNAs (68, 221, 222). The overarching objective of this thesis is to investigate and characterize the role of altered miRNA expression in cervical cancer, and to provide analyses of the currently available methods for miRNA expression profiling.

Firstly, global miRNA expression profiling will be performed using cervical cancer cell lines and patient specimens, to identify differentially-expressed miRNAs in cervical cancer. From the list of differentially-expressed miRNAs, we will select miRNA(s) of interest and identify target genes that contribute to tumour progression; we will characterize the selected miRNA(s) and their target genes(s) *in vitro* and *in vivo*, using various assays that assess colony formation, cell migration and invasion, cell viability, and tumour formation (Chapter 2). The global miRNA expression data from the patient specimens (n = 79) will also be used to develop a prognostic miRNA signature, which we will attempt to validate with an independent cohort of patients (n = 87) (Chapter 3).

Analyses of methods for genome-wide miRNA and mRNA expression studies will be conducted. We will examine the effect of amplification protocols used for microarray studies, and determine if biases can be corrected computationally (Chapter 4). Furthermore, two platforms for measuring global miRNA expression will be evaluated: the TaqMan Low Density Array and the NanoString nCounter assay; we will also compare these two platforms with qRT-PCR, which is considered to be the gold-standard for nucleic acid quantification (Chapter 5).

Overall, this thesis will provide insights into the important role of miRNAs in cervical cancer, as well as an evaluation of technical considerations associated with genome-wide expression profiling studies. The findings from this thesis will potentially guide future studies and help improve therapeutic approaches for cervical cancer patients.
CHAPTER 2: MICRORNA-196B REGULATES THE HOMEBOX-B7/VASCULAR ENDOTHELIAL GROWTH FACTOR AXIS IN CERVICAL CANCER

The data in this chapter have been published in *PLoS ONE*:

2.1 Chapter Abstract

The down-regulation of microRNA-196b (miR-196b) has been reported, but its contribution to cervical cancer progression remains to be investigated. In this study, we first demonstrated that miR-196b down-regulation was significantly associated with worse disease-free survival (DFS) for cervical cancer patients treated with combined chemo-radiation. Secondly, using a tri-modal approach for target identification, we determined that homeobox-B7 (HOXB7) was a bona fide target for miR-196b, and in turn, vascular endothelial growth factor (VEGF) was a downstream transcript regulated by HOXB7. Reconstitution of miR-196b expression by transient transfection resulted in reduced cell growth, clonogenicity, migration and invasion \textit{in vitro}, as well as reduced tumor angiogenesis and tumor cell proliferation \textit{in vivo}. Concordantly, siRNA knockdown of HOXB7 or VEGF phenocopied the biological effects of miR-196b over-expression. Our findings have demonstrated that the miR-196b/HOXB7/VEGF pathway plays an important role in cervical cancer progression; hence targeting this pathway could be a promising therapeutic strategy for the future management of this disease.

2.2 Introduction

Worldwide, cervical cancer is the third most frequently diagnosed malignancy and the fourth leading cause of cancer mortality in women, with an estimated 530,000 new cases and 275,000 deaths each year (131). Although cervical cancer incidence and mortality rates have declined over the past thirty years in the United States (223), the 5-year survival rate has remained below 40% for patients diagnosed with Stage III or IV disease (224). Novel insights are required to better understand the mechanisms that contribute to disease progression, in order to design improved therapies for patients with locally advanced cervical cancer.

Micro-RNAs (miRNAs) are short, non-coding RNAs that regulate gene expression post-transcriptionally (1, 2), and aberrant miRNA expression has been shown to be important in many human malignancies (69). Gene targets that contribute to tumor progression have been described for several miRNAs (68, 221, 222); however, the biological function of the majority of miRNAs still remains unknown. One of the major challenges to miRNA target identification is the ability of miRNAs to bind mRNA targets with imperfect complementarity; hence a single miRNA can potentially regulate several hundreds or thousands of genes (52). Unfortunately, the currently
available in silico miRNA target prediction algorithms have high false-discovery and false-negative rates (40, 67); thereby mandating experimental validation of miRNA targets.

Down-regulation of miR-196b in cervical cancer has been previously reported (159), but its role in tumor progression in this disease has not been previously investigated. Herein, we report down-regulation of miR-196b in primary human cervical cancer tissues and cell lines. Furthermore, we identified the HOXB7 transcription factor as a novel, direct and specific target of miR-196b, which in turn, regulates VEGF in cervical cancer. Most importantly, miR-196b down-regulation was associated with worse DFS in patients treated with chemo-radiation, highlighting the biological importance of miR-196b in cervical cancer progression.

2.3 Materials and Methods

2.3.1 Cell Lines and Transfections

Human cervical cancer cell lines (ME-180, SiHa, and HT-3) were obtained from American Type Culture Collection (ATTC), and grown in α-MEM supplemented with 10% FBS at 37°C, 5% CO₂. All cells were authenticated every six months at the Centre for Applied Genomics (Hospital for Sick Children, Toronto, Canada) using the AmpF/STR Identifier PCR Amplification Kit (Applied Biosystems), and determined to be free from Mycoplasma contamination using the MycoAlert Mycoplasma Detection Kit (Lonza). ME-180 and SiHa cells were transfected using the LipofectAMINE 2000 (Invitrogen) forward transfection protocol, according to the manufacturer’s instructions. Pre-miR Negative Control #1 (NC), Pre-miR-196b (Ambion), All Stars Negative Control (siNEG), siHOXB7 and siVEGF (Qiagen) were all transfected at a final concentration of 30 nmol/L.

2.3.2 miRNA Expression Profiling of Cell Lines

Total RNA was isolated from SiHa, ME-180 and HT3 cervical cancer cell lines using the mirVana miRNA Isolation Kit (Ambion) according to the manufacturer’s instructions. FirstChoice® Total RNA: Human Normal Cervix Tissue (Ambion) from 3 different tissue donors was utilized as normal comparators. Expression levels of 377 miRNAs and 3 snoRNAs (controls) were assayed in the cervical cancer cell lines and normal cervix tissues using the TaqMan® Low Density Array (TLDA) Human MicroRNA Panel (Applied Biosystems), with the Applied Biosystems 7900HT Real-Time PCR System, as we have previously described (96).
2.3.3 miRNA Expression Profiling of Patient Tissues

Flash-frozen punch biopsies were obtained from patients with locally advanced cervical cancer who were planned to receive primary treatment with standard chemo-radiation, consisting of external-beam radiotherapy to the primary cervical tumor and pelvic lymph nodes (45 to 50 Gy total, in 1.8-to-2-Gy daily fractions with 18-to-25-MV photons), combined with weekly doses of cisplatin (40 mg/m2 total, 5 doses). FIGO (International Federation of Gynecologists and Obstetricians) staging was determined using a combination of: pretreatment evaluation under anesthesia, computed tomography (CT) scans of the abdomen and pelvis, chest x-ray, and magnetic resonance imaging (MRI) of the pelvis. MRI was also used to determine lymph node status; pelvic and para-aortic lymph nodes were classified as positive for metastatic disease if the MRI short-axis dimension was >1 cm and equivocal if it was 8 to 10 mm. After biopsy, the specimens were placed in optimal cutting temperature (OCT) storage medium for histopathologic examination, then flash-frozen in liquid nitrogen. H&E-stained tissue sections were cut from the OCT-embedded material and evaluated by a gynecologic pathologist (B Clarke). Total cell content (stroma and tumor cells) was estimated for all tissue samples using a light microscope, and only samples containing >70% tumor cells were considered for further analysis (n=79). The clinical characteristics of these 79 patients are provided in Table 2.1. The median follow-up time for this cohort was 3 years. Flash-frozen normal cervix tissues obtained from 11 patients who underwent hysterectomy for benign causes served as normal comparators.

Two sections of 50-micron thickness were cut from the OCT-embedded flash-frozen tissues and placed in a nuclease-free microtube. Total RNA was isolated using the Norgen Total RNA Purification Kit (Norgen Biotek), according to the manufacturer’s instructions. Global miRNA expression was measured in the cervical cancer and normal cervix tissues with the TaqMan® Low Density Array (TLDA) Human MicroRNA A Array v2.0 (Applied Biosystems) using the Applied Biosystems 7900HT Real-Time PCR System, as already described (96).
Table 2.1  Clinical parameters of 79 cervical cancer patients.

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>N = 79</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years</td>
<td>Median</td>
<td>48</td>
<td>61%</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>26-84</td>
<td></td>
</tr>
<tr>
<td>Tumour size</td>
<td>≤5 cm</td>
<td>48</td>
<td>61%</td>
</tr>
<tr>
<td></td>
<td>&gt;5 cm</td>
<td>31</td>
<td>39%</td>
</tr>
<tr>
<td>FIGO stage</td>
<td>IB</td>
<td>24</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>IIA</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>IIB</td>
<td>35</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td>IIIA</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>IIIB</td>
<td>18</td>
<td>23%</td>
</tr>
<tr>
<td>Histology</td>
<td>Squamous</td>
<td>74</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>Adenocarcinoma</td>
<td>4</td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>Adenosquamous</td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>Pelvic node involvement</td>
<td>Positive</td>
<td>25</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td>Equivocal</td>
<td>15</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>39</td>
<td>49%</td>
</tr>
</tbody>
</table>

2.3.4  Quantitative Real-time PCR Analysis of miRNAs and mRNAs

Total RNA was isolated from the cell lines using the Total RNA Purification Kit (Norgen Biotek), according to the manufacturer’s instructions. The expression of miR-196b was measured by quantitative real-time polymerase chain reaction (qRT-PCR) using the standard TaqMan MicroRNA Assay (Applied Biosystems). Briefly, RNA was first reverse transcribed using the TaqMan MicroRNA Reverse Transcription Kit and a stem-loop primer specific for miR-196b (Applied Biosystems) (114). The 2-ΔΔCt method was used to calculate relative levels of miR-196b expression, using RNU44 as a reference gene (225).

The RT products were amplified with miR-196b-specific primers, as we previously described (96). The expression levels of previously-described (c-myc, BCL2, HOXA9, MEIS1), and candidate mRNA targets for miR-196b (ANKHD1, CTDSP2, FGFR1, HDAC, HOXA7, HOXB7, KRT8, PUM2, SLC9A6, SMC3, SMG7, SR140), and HOXB7 (VEGF, Ku70, Ku80, DNA-PK, FGF2, MMP2, WNT5a, PDGFA, THBS2) were also measured by qRT-PCR. One
microgram of total RNA was reverse-transcribed using SuperScript II Reverse Transcriptase (Invitrogen) according to the manufacturer’s instructions. Quantitative RT-PCR was performed using SYBR Green PCR Master Mix (Applied Biosystems) and PCR primers (Table 2.2) designed using Primer 3 Input. The 2-ΔΔCt method was used to calculate relative levels of gene expression, with GAPDH as a reference gene (225).

Table 2.2 Primer Sequences used for qRT-PCR

<table>
<thead>
<tr>
<th>Gene</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOXB7</td>
<td>5'GTGGTTACTAGTCCAGCTCT 5'GGAACGTTGAATC3'</td>
<td>5'GGGAACGTTGAACGCTCTC 3'</td>
</tr>
<tr>
<td>KRT8</td>
<td>5'CAAAGGCCAGAGGTTC 3'</td>
<td>5'GGGAACGTTGAACGCTCTC 3'</td>
</tr>
<tr>
<td>HDAC</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>PUM2</td>
<td>5'GGGAGCTTCTCACCATTCA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>SMG7</td>
<td>5'GATTATACCCTCCAGTTGCACACAC 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>SR140</td>
<td>5'CAGGACGGGAGCTTTACGTTAT 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>ANKHDI</td>
<td>5'CGAAGTGCTACGTTGAATAC 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>SMC3</td>
<td>5'AAAGAAACAGAGGCAAACG 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>CTDSP2</td>
<td>5'AAGCAAGGCCCTGCTCC 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>FGFR1</td>
<td>5'CGATGCAGAGCATCAACT 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>HOXA7</td>
<td>5'CAGTGACCTGCCAAGGAGG 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>SLC9A6</td>
<td>5'CATTGGGCTTGGGCACT 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>Ku70</td>
<td>5'CGACAGTGGTGGCTGAGAA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>Ku80</td>
<td>5'AATCCAGGACAGAACATATG 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>DNA-PK</td>
<td>5'TGCAAGGTTTAAACGCAAAA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>VEGF</td>
<td>5'GCTACTGCTGCAACATCAGGAAGAATGTCGCTTACTATGGTA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>HOXA9</td>
<td>5'CCACGAGGCTTGGGAAAC 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>BCL-2</td>
<td>5'GAGGATTGTGGCCCTTGGTTCG 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>MEIS1</td>
<td>5'TGTAGGGCTGGGACAACAGTG 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>MYC</td>
<td>5'TAGTGGAAACGACAGCCTACG 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>FGF2</td>
<td>5'CCTGGGTCTTAAATGTACGGTGGTGAAGAATAGAAGTAGATGTCGCTTACTATGGTA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>MMP2</td>
<td>5'TGGGCAACAAAAATGAAAGAAGAATAGAAGTAGATGTCGCTTACTATGGTA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>WNT5a</td>
<td>5'ATTCTTGGTTGCTCAGTGTTGTAAGTGGACAGAGGAAGAATAGAAGTAGATGTCGCTTACTATGGTA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>PDGFA</td>
<td>5'CATGTTTCTGCGGCGAGGAAAGAATAGAAGTAGATGTCGCTTACTATGGTA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
<tr>
<td>THBS2</td>
<td>5'ACTCGCAGGAAAGAAGAATAGAAGTAGATGTCGCTTACTATGGTA 3'</td>
<td>5'CCTGACCGACCCAGTTGCACACAC 3'</td>
</tr>
</tbody>
</table>
2.3.5 Cell Viability, Proliferation and Colony-Formation Assays

The viability of transfected cells was assessed by the Trypan blue exclusion assay. ME-180 and SiHa cells were transfected in triplicate with 30 nmol/L of pre-miR-196b, NC, siHOXB7, siVEGF, or siNEG and incubated at 37°C, 5% CO₂. At 48 and 72 hours post-transfection, cells were trypsinated, stained with Trypan blue and counted using a hemocytometer. Cell proliferation was examined using the CellTiter 96 Non-Radioactive Cell Proliferation Assay (MTS Assay) (Promega BioSciences), according to the manufacturer’s instructions. For colony formation assays, cells were transfected with 30 nmol/L of pre-miR-196b, Pre-miR Negative Control #1, siHOXB7, siVEGF, or siNEG and incubated at 37°C, 5% CO₂. At 48 hours post-transfection, cells were re-seeded at low density in 6-well plates in triplicate. Cells were incubated at 37°C, 5% CO₂ for 10-12 days, then fixed and stained with 0.1% crystal violet in 50% methanol. The number of colonies containing at least 50 cells was counted, and the surviving fraction was calculated by comparison with cells transfected with Negative Control.

2.3.6 Cell Migration and Invasion Assays

BD BioCoat Matrigel Invasion Chambers and Control Inserts (BD Biosciences) were used to assay migration and invasion of transfected cells. Chambers contained a polyethylene terephthalate membrane with 8 µm pores. ME-180 and SiHa cells were transfected with 30 nmol/L of pre-miR-196b, NC, siHOXB7, siVEGF, or siNEG and incubated at 37°C, 5% CO₂. At 24 hours after transfection, 1.5 x 10^5 cells were re-seeded inside each chamber with medium containing low serum (1% FBS). The chambers were placed in a 24-well plate, with high serum (20% FBS) medium in each lower chamber to serve as a chemo-attractant. Cells were incubated at 37°C, 5% CO₂ for 48 hours, then the membranes were washed, stained, and mounted onto slides. A light microscope was used to count the number of migrating or invading cells. Relative migration was calculated by comparison with cells transfected with the negative control. Percent invasion was calculated as the number of cells that invaded through the Matrigel insert, divided by the number of cells that migrated through the control insert.

2.3.7 Cell Cycle Analysis

Cell cycle analysis was performed on ME-180 and SiHa cells after transfection with 30 nmol/L of pre-miR-196b or NC, to measure the fraction of cells in the sub-G1 phase of the cell
cycle. Cells were harvested and washed twice in FACS buffer (PBS/0.5% BSA), re-suspended in 1 mL of FACS buffer, then fixed in 1 mL of ice-cold 70% ethanol. After 1h of incubation on ice, cells were washed again and re-suspended in 500 µL of FACS buffer containing 40 µg/mL RNase A (Sigma) and 50 µg/mL propidium iodide, then incubated in the dark at room temperature for 30 minutes. Cells were analyzed in the BD FACScalibur (Becton Dickinson) using the FL-2A and FL-2W channels. The flow cytometry data were analyzed using FlowJo 7.5 software (Tree Star).

2.3.8  In vivo Experiments

Six- to 8-week-old severe combined immunodeficient (SCID) female mice were utilized for xenograft experiments, according to guidelines of the Animal Care Committee, Ontario Cancer Institute, University Health Network. Cells were transfected with pre-miR-196b or NC and incubated at 37°C, 5% CO₂. At 48 hours post-transfection, cells were harvested and 5 x 10⁵ viable cells were diluted in 100 µL of growth medium. Cells were injected intramuscularly into the left gastrocnemius muscle of female SCID mice. Tumor plus leg diameter was measured twice a week and mice were euthanized when 15 mm was attained. Tumors were removed at 25 days after implantation and immediately fixed in 10% buffered formalin for 24 h, placed in 70% ethanol for 24 h, embedded in paraffin, and sectioned (5 µM) for immunostaining. In addition to hematoxylin and eosin (H&E) staining, cluster of differentiation 31 (CD31) was used for assessing tumor angiogenesis, Ki-67 for tumor cell proliferation, and terminal deoxynucleotidyl transferase-mediated dUTP nick end labeling (TUNEL) for apoptosis.

2.3.9  Tri-modal Approach for miRNA Target Identification

Candidate mRNA targets of miR-196b were determined using a previously-described tri-modal approach (68) combining: i) all predicted targets of miR-196b from five in silico miRNA target prediction databases (TargetScan, PicTar, GenMir++, miRBase, miRDB); ii) mRNAs up-regulated at least 2-fold in cervical cancer tissues compared to normal cervix tissues, using two independent publicly-available microarray datasets (226, 227); and iii) mRNAs down-regulated at least 0.5-fold at both 24 and 72 hours after transfection with 30 nmol/L of pre-miR-196b, where transcript levels were measured using the Whole Human Genome 4 x 44K One-Color Array (Agilent).
2.3.10  Luciferase Reporter Assay

Wildtype or mutant fragments of the 3’-untranslated region (UTR) of HOXB7 containing
the predicted binding site (position 220-226) for miR-196b were individually amplified by
AmpliTaq Gold DNA Polymerase (Applied Biosystems) using the primers listed in Table S1.
The PCR products were purified, then cloned downstream of the firefly luciferase gene in the
pMIR-REPORT vector (Ambion) at the SpeI and HindIII restriction sites, to produce the pMIR-
HOXB7 or pMIR-HOXB7-mut vector. ME-180 and SiHa cells were co-transfected with 100
nmol/L of pre-miR-196b or NC, and 100 ng of the reporter vector of interest. As a reference
control, 50 ng of pRL-SV vector (Promega) containing the Renilla luciferase gene was also
transfected with each condition. Firefly and renilla luciferase activities were measured at 24
hours post-transfection using the Dual-Glo Luciferase Assay System (Promega) according to the
manufacturer’s instructions.

2.3.11  Immunoblotting

Cells were transfected with either 100 nmol/L of pre-miR-196b or NC, and total protein
extracts were harvested on ice after 48 and 72 hours. Proteins of interest were probed with rabbit
anti-HOXB7 (1:500 dilution; Invitrogen) or mouse anti-GAPDH (1:10,000 dilution; Sigma), and
detected with IRDye fluorescent secondary antibodies (1:20,000 dilution, LI-COR). GAPDH
was used as a loading control. Immunoblots were scanned and quantified using the Odyssey
Infrared Imaging System (LI-COR).

2.3.12  Enzyme-linked Immunosorbent Assay (ELISA)

Cells were transfected with 30 nmol/L of siHOXB7 or siNEG, and the level of secreted
VEGF was measured at 48 and 72 hours post-transfection, using the Human VEGF DuoSet
ELISA (R&D Systems) according to manufacturer’s instructions.

2.3.13  5-aza-2’-deoxycytidine Treatment

ME-180 and SiHa and were seeded in 24-well plates and incubated overnight at 37°C,
5% CO₂. Media containing 2 µM 5-aza-2’-deoxycytidine (5-aza-DCT) (Sigma-Aldrich) was
added to cells, 1 and 3 days after seeding, respectively. Cells were harvested 4 days after seeding
and total RNA was isolated for qRT-PCR analysis of miR-196b expression.
2.3.14 Statistical Analysis

All experiments have been performed at least three independent times, and the data are presented as the mean ± standard error of the mean (SEM). The Student’s t-test function (unpaired, two-tailed) in Microsoft Excel (Microsoft, Redmond, WA) was used to compare two treatment groups. Graphs were plotted using GraphPad Prism software (GraphPad software, San Diego, CA). The Kaplan-Meier method was used for univariate analysis, and the log-rank test was used to examine associations between miR-19b expression and DFS, where the expression was dichotomized at the median value. The relationship between tumor size and miR-196b expression was investigated using Pearson’s correlation. Correlations between miR-196b expression with either FIGO stage or nodal status were analyzed using one-way ANOVA.

2.4 Results

2.4.1 miR-196b was significantly down-regulated in primary cervical cancer tissues and cell lines, and was strongly associated with DFS

Expression of miR-196b was significantly reduced by almost 4-fold in 79 primary cervical cancer tissues compared to 11 normal cervix tissues (P < 0.001; Fig. 2.1A). Importantly, patients with lower than median miR-196b expression level at the time of diagnosis experienced worse DFS compared to those with higher miR-196b expression (P = 0.02; hazard ratio = 0.39; Fig. 2.1B). miR-196b expression was not significantly correlated with tumor size (P = 0.12), FIGO stage (P = 0.14), or nodal status (P = 0.60). Global miRNA expression profiling conducted on three cervical cancer cell lines (ME-180, SiHa and HT-3) also confirmed the down-regulation of miR-196b in cervical cancer. From the 55 miRNAs that were deregulated at least 2-fold in all three cell lines compared to 3 normal cervix tissues, miR-196b was amongst the most significantly down-regulated miRNAs (Fig. 2.1C), consistent with a previously published miRNA expression profiling study (159).

To explore whether miR-196b down-regulation was epigenetically determined, in addition to chromosomal loss [201], ME-180 and SiHa cells were treated with the demethylating agent 5-aza-2’-deoxycytidine (5-aza-DCT). This treatment resulted in only a minimal increase in miR-196b expression (Fig. S2.1A left), indicating that promoter methylation was unlikely to be a major mechanism for miR-196b under-expression, whereas the levels of miR-375, which is known to be epigenetically regulated, increased significantly (Fig. S2.1A right). Furthermore,
examination of publically-available microarray datasets of gene expression in primary cervical cancer tissues did not reveal any significant alterations in the expression of DICER, drosha, DGCR8, Exportin-5, or any subunits of RNA Polymerase II, which are all involved in miRNA biogenesis and processing (data not shown).

Figure 2.1  miR-196b down-regulation in cervical cancer

A) miR-196b expression levels were measured using qRT-PCR in 79 primary cervical cancer samples, compared to 11 normal cervix epithelial tissue controls. B) Kaplan-Meier analysis of DFS in patients with cervical cancer. Red, patients with higher than median miR-196b expression level (n = 39); blue, patients with lower than median miR-196b expression (n = 39); one patient was removed from survival analysis due to missing survival information. C) Basal levels of miR-196b in three cervical cancer cell lines (SiHa, ME-180, and HT-3), as compared to normal cervix epithelial tissues, assayed by qRT-PCR. ** P < 0.01.
2.4.2 miR-196b over-expression significantly reduced cell viability, clonogenicity, proliferation and invasion

To assess the biological significance of miR-196b down-regulation, cells were transfected with 30 nmol/L NC or pre-miR-196b. Up-regulation of miR-196b expression was sustained for up to 72 hours after transfection (Fig. S2.1B). Transfection with pre-miR-196b led to significantly decreased cell viability compared to controls at 48 and 72 hours post-transfection (ME-180: 25% at 48h; 41% at 72h, SiHa: 29% at 48h; 54% at 72h) (Fig. 2.2A). In addition, miR-196b over expression resulted in significant reductions in clonogenicity (ME-180: 57% compared to NC, SiHa: 64% compared to NC) (Fig. 2.2B), proliferation (ME-180: 36% at 48h, 35% at 72h, SiHa: 22% at 48h; 21% at 72h) (Fig. 2.2C), and migration (32% compared to NC), plus invasion (32% vs. 63% for NC) (Fig. 2.2D). Cells treated with pre-miR-196b demonstrated a small, but not statistically significant, increase in the percentage of cells in the sub G1 population, accompanied by a small decrease in the G0-G1 population (Fig. S2.1C).

2.4.3 miR-196b over-expression suppressed tumor angiogenesis and tumor cell proliferation in vivo

Cells transfected with pre-miR-196b did not show a statistically significant difference in tumor growth in mice, compared to NC-treated cells (Fig. S2.2). At 25 days post-implantation however, CD31 immunostaining was nonetheless observed to be reduced to 69% in pre-miR-196b-treated tumors compared to control tumors (Fig. S2.3, top). Furthermore, this was associated with a modest yet significant reduction in Ki-67 expression (46% vs. 53% for cells treated with NC) (Fig. S2.3, middle), and a minor, but not statistically significant increase in TUNEL staining (Fig. S2.3, bottom). Hence, our data suggested that pre-miR-196b contributed to reduced angiogenesis and tumor cell proliferation.
Figure 2.2  *In vitro* effects of miR-196b over-expression

A) Relative viability of ME-180 and SiHa cells were assessed at 48 and 72 hours post-transfection with pre-miR-196b (30 nmol/L), compared to Negative Control pre-miR (30 nmol/L), using the Trypan Blue assay.  B) Clonogenicity of ME-180 and SiHa cells were assessed by transfection with 30 nmol/L of pre-miR 196b or Negative Control pre-miR. At 48
hours post-transfection, cells were harvested then counted and re-seeded at low density in 6-well plates. After 10 days of incubation, cells were fixed and stained and the number of colonies (>50 cells) were counted. C) Relative proliferation of ME-180 and SiHa cells were examined at 24, 48, and 72 hours post-transfection with pre-miR-196b (30 nmol/L), compared to Negative Control pre-miR (30 nmol/L), using the MTS assay. D) Representative images (left) and histograms (right) depicting migratory ability (top) and invasiveness (bottom) of ME-180 cells that were transfected with 30 nmol/L of pre-miR 196b or Negative Control pre-miR, harvested at 24 hours post-transfection, then counted and re-seeded in the invasion chambers. Cell migration (top), and invasion (bottom), assessed at 48 hours after seeding in transwell chambers. All data represent the mean ± SEM from 3 independent experiments. NC, pre-miR Negative Control; * P < 0.05; ** P < 0.01.
2.4.4 miR-196b directly targets HOXB7

The down-regulation of miR-196b in cervical cancer tissues and cell lines, and the significant phenotypic effects of miR-196b over-expression both in vitro and in vivo, indicated that miR-196b appears to be an important mediator of cervical cancer progression. Thus, a tri-modal strategy (68) was utilized to identify potential mRNA targets that could account for these phenotypic changes (Fig. S2.4). This method identified 15 overlapping candidate targets (ANKHD1, CALM3, CLK2, CTDSP2, FGFR1, HDAC, HOXA7, HOXB7, KRT8, PCCB, PUM2, SLC9A6, SMC3, SMG7, SR140). For target validation, ME-180 cells were transfected with pre-miR-196b or NC, and transcript levels at 24 hours post-transfection were measured with qRT-PCR for 12 candidate targets (suitable primers could not be designed for the other 3 transcripts), plus 4 previously-described targets of miR-196b (c-myc, BCL2, HOXA9, MEIS1). None of the previously-described targets were significantly altered after miR-196b over-expression (Fig S2.5A). Only 5 of the 12 tested candidate targets were significantly down-regulated after miR-196b over-expression (Fig. S2.5B); HOXB7 was selected for further functional evaluation since it was the candidate target that demonstrated the greatest level of down-regulation (40% compared to controls). Furthermore, HOXB7 is a member of the Hox gene cluster, a family of genes which have been reported to be dysregulated in various malignancies (228), and the miR-196 family is known to target the mammalian Hox genes (229).

A luciferase binding assay confirmed that miR-196b directly and specifically interacted with the 3’-UTR of HOXB7. In comparison to control cells transfected with pMIR-REPORT, cells transfected with pMIR-HOXB7 showed reduced luciferase activity (ME-180: 68%, SiHa: 71%) when co-transfected with pre-miR-196b (Fig. 2.3A). This inhibitory effect was completely abrogated with pMIR-HOXB7-mut, which contained a mutation in the miR-196b binding site. In addition, transfection with pre-miR-196b resulted in significantly reduced HOXB7 mRNA transcript (ME-180: 49% at 48h, 62% at 72h; SiHa: 55% at 48h, 69% at 72h) (Fig. 2.3B), and protein (74% at 48h; 80% at 72h) (Fig. 2.3C) levels at 48 and 72 hours post-transfection. There appeared to be a greater fold change in HOXB7 at the mRNA level compared to protein, which is not surprising since miRNAs interact directly with mRNA transcripts and not proteins. In addition, protein translation can be affected by a number of mechanisms that occur upstream, such as nuclear export of RNA transcripts and recruitment of ribosomal subunits.
Figure 2.3  Identification of HOXB7 as an mRNA target of miR-196b

A) Relative luciferase activity of ME-180 or SiHa cells at 24 hours after co-transfection with pMIR-REPORT, pMIR-HOXB7 or pMIR-HOXB7-mut vectors and pre-miR-196b or NC (30 nmol/L).  B) Relative HOXB7 mRNA expression levels in ME-180 or SiHa cells after transfection (48, and 72 hrs) with pre-miR-196b or NC (30 nmol/L), as measured by qRT-PCR. Expression levels were normalized to GAPDH expression. C) Representative Western blot image (top), and relative quantification of HOXB7 protein levels (bottom) after transfection (48, and 72 hrs) with pre-miR-196b or NC (30nmol/L). All data represented the mean ± SEM from 3 independent experiments. OD, optical density; NC, pre-miR Negative Control; *P < 0.05; **P < 0.01.
2.4.5 VEGF is a relevant downstream target of HOXB7

Since HOXB7 is a transcription factor, it was necessary to determine which gene(s) regulated by HOXB7 could be relevant in this context for cervical cancer. Hence, cells were transfected with 30 nmol/L of siHOXB7 or siNEG, and mRNA levels of known HOXB7 targets such as Ku70, Ku80, DNA-PK, FGF2, MMP2, WNT5a, PDGFA, thrombospondin 2 (THBS2) and VEGF (230-233) were measured at 48 hours post-transfection. VEGF demonstrated the greatest level of down-regulation (43%) following HOXB7 knockdown (Fig. S2.6A). Concordantly, significant reduction in VEGF mRNA (ME-180: 63% at 48h, 33% at 72h; SiHa: 69% at 48h, 41% at 72h) (Fig. 2.4A) and secreted protein (ME-180: 82% at 48h, 77% at 72h; SiHa: 84% at 48h, 75% at 72h) (Fig. 2.4B) levels were observed at both 48 and 72 hours post-transfection with siHOXB7, confirming that VEGF was indeed a relevant downstream HOXB7 target in this disease. Furthermore, other pro-angiogenic known targets of HOXB7 (FGF2, MMP2, WNT5a, PDGFA, THBS2) were not significantly altered following HOXB7 knockdown (Fig. S2.6A) or miR-196b overexpression (Fig. S2.6E), suggesting that HOXB7 mediates angiogenesis via VEGF.

2.4.6 Knockdown of HOXB7 or VEGF recapitulated the biological effects observed following mir-196b over-expression

To further corroborate this newly-described pathway of miR-196b targeting HOXB7 which in turn regulated VEGF, ME-180 and SiHa cells were treated with siHOXB7 or siVEGF to determine whether these interventions could recapitulate the effects of miR-196b over-expression. Knockdown of HOXB7 and VEGF transcript and protein levels were sustained for up to 72 hours after siRNA transfection (Fig. S2.6B, S2.6C, and S2.6D). We observed that cell viability was reduced significantly after transfection with either siHOXB7 (ME-180: 77% at 48h, 66% at 72h; SiHa: 80% at 48h, 70% at 72h) (Fig. 2.5A, left) or siVEGF (ME-180: 78% at 48h, 60% at 72h; SiHa: 77% at 48h, 68% at 72h) (Fig. 2.5A, right), similar to the effects observed after transfection with pre-miR-196b (ME-180: 25% at 48h, 41% at 72h; SiHa: 29% at 48h, 54% at 72h) (Fig. 2.2A). Clonogenicity decreased significantly following either siHOXB7 (ME-180: 69%; SiHa: 71%) (Fig. 2.5B, left) or siVEGF (ME-180: 78%; SiHa: 75%) (Fig. 2.5B, right) transfection, as previously observed for pre-miR-196b transfection (ME-180: 57% compared to NC; SiHa: 64% compared to NC) (Fig. 2.2B). Cell migration (Fig. 2.5C, top) was also reduced
significantly following transfection with either siHOXB7 (60%) or siVEGF (62%), similar to the results observed following pre-miR-196b transfection (32% vs. NC) (Fig. 2.2D, top). Finally, cell invasion also significantly decreased after transfection with either siHOXB7 (52% vs. 74% for control cells) or siVEGF (42% vs. 74% for control cells) (Fig. 2.5C, bottom), all recapitulating the effects observed after transfection with pre-miR-196b (32% vs. 63% for control cells) (Fig. 2.2D, bottom).

**Figure 2.4** Identification of VEGF as a target of the HOXB7 transcription factor

A) Relative VEGF expression at the mRNA level after transfection of ME-180 or SiHa cells with siHOXB7 or siNEG (30 nmol/L), as determined by qRT-PCR. Expression levels were normalized to GAPDH expression. B) Relative levels of secreted VEGF protein after transfection of ME-180 or SiHa cells with siHOXB7 or siNEG (30 nmol/L), as determined by ELISA. All data represented the mean ± SEM from 3 independent experiments. siNEG, All Stars Negative Control; *$P < 0.05$; **$P < 0.01$. 
Figure 2.5  Downstream effects of siHOXB7 and siVEGF
ME-180 and SiHa cells were transfected with siHOXB7, siVEGF or siNEG (30 nmol/L):  
A) Relative cell viability assessed using the Trypan blue assay at 48 and 72 hours post-transfection 
with siHOXB7 (left) or siVEGF (right);  
B) Cells were harvested at 48 hours post-transfection, 
then counted and re-seeded at low density in 6-well plates; after 10 days of incubation, cells were
fixed and stained and colonies (>50 cells) were counted. Histograms depict the relative number of surviving colonies post-transfection with siHOXB7 (left), and siVEGF (right). C) Representative images (left) and histograms (right) depict migratory ability (top), and invasiveness (bottom) after transfection. All data represented the mean ± SEM from 3 independent experiments. siNEG, All Stars Negative Control; *$P < 0.05$. 

2.5 Discussion

In this study, we identified that miR-196b was significantly down-regulated in both cervical cancer cell lines and primary tissues, which promoted tumour cell proliferation, migration, invasion, and angiogenesis, mediated through VEGF regulation by HOXB7. MicroRNA-196b was one of the six deregulated miRNAs first reported by Lui et al. for human cervical cancer (159). Its biological role was first described for endometriosis (234); since then, miR-196b has been reported to be deregulated in various human malignancies aside from cervical cancer, including over-expression in acute lymphoblastic leukemia (ALL) (235) and colon cancer (236); as well as under-expression in glioblastoma (237) and B-cell lineage ALL (238). These discordant observations highlight the fact that miR-196b can function as either an oncogene or a tumor suppressor, as described for many miRNAs. In glioblastoma, miR-196b levels have been positively correlated with overall survival (237). In contrast, we report for the first time that lower levels of miR-196b were associated with worse DFS for cervical cancer, by promoting cellular proliferation, clonogenicity, migration and invasion in vitro, as well as tumor cell proliferation and angiogenesis in vivo.

The mechanisms for miR-196b down-regulation are complex. Using array comparative genomic hybridization (aCGH) profiling of cervical cancer samples, the chromosomal location of miR-196b (7p15.2) was noted to be a region with a high level of homozygous loss in squamous cell cervical carcinoma (239); hence this might be one explanation for down-regulation of miR-196b in this disease. miR-196b has also been reported to be epigenetically regulated in gastric cancer (240), which was not corroborated based on our data (Figure S2.2).

Few studies have identified mRNA targets of miR-196b, aside from c-myc (238), and the Hox gene cluster (229). The Hox gene family consists of a set of 39 genes which encode transcription factors that direct the basic structure and orientation of an organism during embryonic development (241), regulating many crucial processes such as differentiation, apoptosis, motility, angiogenesis and receptor signaling (228). Aberrant Hox gene expression has been reported to mediate oncogenesis in many human cancers, including hepatocellular (242), ovarian (243), as well as acute myeloid leukemia (AML) (244). There are at least three mechanisms that have been described which can lead to Hox gene deregulation: a) over-expression of Hox genes in a specific tissue type; b) epigenetic deregulation, whereby Hox genes
are silenced in a tissue when they should normally be expressed; and c) temporo-spatial deregulation, whereby Hox gene expression in a tumor arising in a specific tissue temporo-spatially differs from that in normal tissue (187). In this current study, we provide evidence for an alternate mechanism for HOXB7 deregulation, via miR-196b, in cervical cancer.

Our current study also demonstrated that VEGF was a downstream transcriptional target of HOXB7 in cervical cancer, which has also been reported for breast cancer (231), as well as multiple myeloma (233). VEGF is a known key mediator of angiogenesis (245); a major hallmark of human cancers (246). In contrast, other pro-angiogenic factors that are known targets of HOXB7 (FGF2, MMP2, WNT5a and PDGFA) (231, 233) were not significantly altered following HOXB7 knockdown or miR-196b overexpression. Although VEGF was initially regarded to be an endothelial-specific ligand, reports have shown that VEGF can promote cancer cell proliferation (247, 248), migration and invasion (249, 250). Interestingly, serum VEGF levels have been shown to be a prognostic marker for DFS in cervical cancer patients, whereby high pre-treatment VEGF levels were associated with worse survival (217, 218). In addition, alterations in the serum concentration of VEGF have been used to measure treatment response in cervical cancer patients (251). Serum expression of VEGF might indeed be a superior read-out of angiogenic activity in cervix tumors, compared to tissue expression of VEGF. Unfortunately, sera samples were not available from the patients in this current study.

A humanized anti-VEGF monoclonal antibody (A.4.6.1) that recognizes all biologically active isoforms of VEGF and prevents their binding to VEGF receptors (VEGFR-1 and VEGFR-2), Bevacizumab, has been shown to be effective in treating several human malignancies by blocking angiogenesis (252). Thus far, Bevacizumab has been approved by the FDA for treating metastatic colorectal cancer, non-small cell lung cancer, glioblastoma, and metastatic kidney cancer. A multi-centre randomized Phase III clinical trial (GOG240) has recently been completed, in which Bevacizumab in combination with standard treatment was evaluated in cervical cancer patients with advanced (stage IVB), persistent, or recurrent cervical cancer. A National Cancer Institute (NCI) press release in February 2013 announced that the Phase III trial demonstrated that the addition of Bevacizumab significantly improved median survival by 3.7 months.
In conclusion, we report for the first time that miR-196b is a novel tumor suppressor in cervical cancer, by regulating the transcription factor HOXB7, which in turn, induced VEGF expression. The resulting phenotype of miR-196b down-regulation included increased cell growth, clonogenicity, migration and invasion, as well as increased tumor cell proliferation and vascularity. Furthermore, patients with lower miR-196b expression experienced a worse 5-year DFS. Hence, this novel axis of miR-196b~HOXB7~VEGF might well provide the biological rationale for the potential efficacy of an anti-angiogenic therapeutic strategy for cervical cancer.
Figure S2.1  *In vitro* effects of 5-aza-DCT treatment and miR-196b over-expression

A) qRT-PCR analysis of miR-196b (left) and miR-375 (right) levels in ME-180 and SiHa cells after treatment with DMSO or 2 µM 5-aza-2’-deoxycytidine (5-aza-DCT). Expression levels were normalized to RNU44 expression, relative to cells treated with DMSO. B) qRT-PCR analysis of miR-196b levels in ME-180 and SiHa cells after treatment with NC or pre-miR-196b (30 nmol/L). Expression levels were normalized to RNU44 expression, relative to cells treated with NC. C) Cell cycle analysis performed on ME-180 (left) and SiHa cells (right) using flow cytometry after treatment with pre-miR-196b or NC (30 nmol/L). The data represented the mean
± SEM from 3 independent experiments. NC, pre-miR Negative Control; **$P < 0.01$; $P = \text{ns}$ (not significant).
Figure S2.2  Effect of miR-196b on *in vivo* tumor growth

Tumor-plus-leg diameter measurements of ME-180 tumors in SCID mice after intramuscular injection of cells transfected with pre-miR-196b or Negative Control pre-miR (60nmol/L). The plotted data represent the mean ± SEM from 9 mice in each group.
Figure S2.3  Effect of miR-196b on *in vivo* angiogenesis and apoptosis

Tumours removed at 25 days post-implantation were immunostained for CD31, Ki-67, and TUNEL expression; representative photomicrographs are shown for NC vs. miR-196b for CD31 (top), Ki-67 (middle), and TUNEL (bottom) immuno-expression. The corresponding histograms represented the mean ± SEM scoring obtained from 6 representative regions, from 2 independent tumors. NC, pre-miR Negative Control; *$P < 0.05$; $P =$ ns (not significant).
A tri-modal strategy to elucidate targets of miR-196b in cervical cancer used a combination of: i) All predicted targets of miR-196b from five *in silico* miRNA target prediction databases (*in silico*); ii) mRNA transcripts up-regulated at least 2-fold in primary cervical cancer samples compared to normal cervix tissues [Cervical cancer (Up)]; and iii) mRNA transcripts down-regulated at least 0.5-fold at both 24 and 72 hours after transfection with 30 nmol/L of pre-miR-196b (Exp. Det.).

**Figure S2.4** Tri-modal strategy for target identification
Figure S2.5 Transcript levels of putative miR-196b targets

qRT-PCR analysis of: A) previously described; and B) candidate targets of miR-196b. ME-180 cells were transfected with NC or pre-miR-196b (30 nmol/L) and transcript levels of candidate targets were measured at 24 hours post-transfection. Expression levels were normalized to GAPDH expression, relative to cells transfected with pre-miR Negative Control. The data represent the mean ± SEM from 3 independent experiments. *P < 0.05.
Figure S2.6  *In vitro* effects of treatment with siHOXB7, siVEGF, or pre-miR-196b
A) qRT-PCR analysis of candidate targets of HOXB7. Cells were transfected with siNEG or siHOXB7 (30 nmol/L) and transcript levels of candidate targets were measured at 24 hours post-transfection. B) qRT-PCR analysis of HOXB7 (left) and VEGF (right) transcript levels after treatment with siHOXB7, siVEGF, or siNEG (30 nmol/L). C) Western blot analysis of HOXB7 protein levels after treatment with siNEG or siHOXB7 (30 nmol/L). D) VEGF protein levels as measured by ELISA, after treatment with siNEG or siVEGF (30 nmol/L). E) qRT-PCR analysis of candidate targets of HOXB7. Cells were transfected with NC or pre-miR-196b (30 nmol/L) and transcript levels of candidate targets were measured at 24 hours post-transfection. The data represent the mean ± SEM from 3 independent experiments. NC, pre-miR Negative Control; siNEG, All Stars Negative Control; *P < 0.05; **P < 0.01.
CHAPTER 3: CHALLENGES ASSOCIATED WITH DEVELOPING A PROGNOSTIC MICRORNA SIGNATURE FOR HUMAN CERVICAL CARCINOMA

The data in this chapter have been submitted for publication and is currently under review:

3.1 Chapter Abstract

**Purpose:** Cervical cancer remains the third most frequently diagnosed and fourth leading cause of cancer death in women worldwide. We sought to develop a micro-RNA signature for cervical cancer that was prognostic for disease-free survival, which could potentially allow tailoring of treatment for patients.

**Experimental Design:** The TaqMan Low Density Array (Applied Biosystems) was utilized to measure the expression of 377 unique human micro-RNAs in 79 frozen cervical cancer specimens with five years median clinical follow-up (range, 0.64-10.6 years). LASSO regression was applied to the normalized micro-RNA expression data, to select a subset of micro-RNAs associated with disease-free survival. For our validation cohort (n=87), which were formalin-fixed paraffin-embedded samples, three different methods were used to measure micro-RNA expression: TaqMan Low Density Array, NanoString nCounter Human miRNA Expression Assay, and individual single-well quantitative real-time PCR using TaqMan Micro-RNA Assays.

**Results:** A candidate prognostic 9-micro-RNA signature set for disease-free survival was identified in the frozen samples. Three different approaches however, to validate this signature in an independent cohort of 87 patients with FFPE specimens, were unsuccessful.

**Conclusions:** There are several challenges and considerations associated with developing a prognostic micro-RNA signature for cervical cancer, namely: tumour heterogeneity, lack of concordance between frozen and FFPE specimens, and platform selection for global micro-RNA expression profiling in this disease.

3.2 Introduction

Micro-RNAs (miRNAs) are a class of small, non-coding RNAs that play important roles in regulating target genes by binding to complementary sequences in mRNA transcripts (4). Deregulation of miRNA expression has been reported for numerous solid and hematological malignancies, which is not surprising given their involvement in multiple critical biological processes, including development, proliferation, and apoptosis (2, 253). miRNAs have been described to be extremely stable, and can be readily extracted from cell lines and various types of clinical specimens, including frozen and formalin-fixed paraffin embedded (FFPE) tissues,
blood, serum, plasma, urine, and saliva (96, 98-102). miRNA expression profiling of solid and hematological human malignancies has identified disease-specific miRNA signatures associated with diagnosis, progression, staging, prognosis and response to therapy (254). Numerous prognostic miRNA signatures have since been described for various malignancies, including: lung (125-127), colorectal (121), gastric (123), esophageal (122), hepatocellular (124), prostate (129), nasopharyngeal (128), and cervical cancers (120).

Cervical cancer is the third most frequently diagnosed cancer, and the fourth leading cause of cancer mortality in women worldwide, with an estimated 530,000 new cases and 175,000 deaths each year (131). Although cervical cancer incidence and mortality have declined over the past 30 years in the United States (223), the 5-year survival rate remains less than 40% for patients diagnosed with Stage III disease and above (224). A two-miRNA signature that could predict overall survival in cervical cancer patients was reported by Hu et al. (120), and remains the only reported miRNA signature for cervical cancer to date. However, we were unable to independently corroborate this signature with our own cohort of patients. Given these limitations and the poor survival for patients with advanced stages of cervical cancer, we sought to develop an independent miRNA signature that could predict disease-free survival (DFS) for cervical cancer patients; and potentially allow tailoring of treatment according to risk. Herein, we describe the challenges and considerations associated with developing such a prognostic miRNA signature for cervical cancer.

3.3 Materials and Methods

3.3.1 Clinical specimens

Pre-treatment cancer samples were collected from patients with cervical cancer who were undergoing curative chemo-radiation, consisting of external-beam radiotherapy to the primary cervical tumour and pelvic lymph nodes (45 to 50 Gy total, in 1.8-to-2-Gy daily fractions using 18 or 25MV photons), combined with weekly cisplatin (40 mg/m² total, 5 doses). Patients were staged using the FIGO (International Federation of Gynecologists and Obstetricians) system, with additional clinical information gathered using computed tomography (CT) scans of the abdomen and pelvis and magnetic resonance imaging (MRI) of the pelvis to assess local and
lymphatic disease. Pelvic and para-aortic lymph nodes were classified as positive for metastatic disease if the MRI short-axis dimension was >1 cm, and equivocal if it was 8 to 10 mm.

The training cohort comprised of flash-frozen punch biopsies obtained from 79 patients treated at the Princess Margaret Cancer Centre between 2000 to 2007, inclusively. The biopsy specimens were placed in a storage medium (optimal cutting temperature (OCT) compound) for histopathologic examination, then flash-frozen in liquid nitrogen. H&E-stained tissue sections were cut from the OCT-embedded material, and evaluated by a gynecological pathologist (B. Clarke). The total cell content (stroma and tumour cells) was estimated for all tissue samples using a light microscope, and only samples containing at least 70% tumour cells were considered for further analysis. Flash-frozen normal cervix tissues obtained from 11 patients who underwent total hysterectomy for benign causes served as the normal comparators.

The validation cohort comprised of diagnostic FFPE blocks collected from 87 cervical patients treated between 1999 and 2007, inclusively. All samples contained at least 70% malignant epithelial cells, as determined by a gynecologic pathologist (B. Clarke), or were macro-dissected prior to RNA purification. FFPE normal cervix tissues obtained from 9 patients who underwent hysterectomy for benign causes served as the normal comparators.

### 3.3.2 Sample processing

For the training cohort specimens, two sections of 50-µm thickness were cut from the OCT-embedded flash-frozen tissues and placed in a nuclease-free microtube. Total RNA was isolated using the Norgen Total RNA Purification Kit (Norgen Biotek), according to the manufacturer’s instructions. Global miRNA expression was measured in both the cervical cancer and normal cervix tissues with the TaqMan® Low Density Array (TLDA) Human MicroRNA A Array v2.0 (Applied Biosystems) using the Applied Biosystems 7900HT Real-Time PCR System, as previously described (96).

For the validation cohort specimens, ten sections of 5-µm thickness were cut from the FFPE tissues and placed in a nuclease-free microtube. Total RNA was isolated using the Norgen Total RNA Purification Kit (Norgen Biotek), according to the manufacturer’s instructions. We measured the expression of the 9 miRNAs in our prognostic signature using three methods: 1) Applied Biosystems TLDA Human MicroRNA A Array v2.0; 2) NanoString nCounter Human
miRNA Expression Assay v1.6.0; and 3) individual single-well qRT-PCR using Applied Biosystems TaqMan MicroRNA Assays.

3.3.3 Data normalization

To normalize the miRNA expression data from TLDA, the raw miRNA abundances were loaded into the R statistical environment (v2.15.2). Three control genes were utilized for normalization: RNU44, RNU48, and U6. Normalized miRNA abundances were calculated as

$$-\log_2\left(2^{-(C_T - C_C)}\right),$$

where $C_T$ represents the abundance of a given miRNA, and $C_C$ represents the mean of the control gene abundances.

To normalize the miRNA expression data from NanoString, the R package ‘NanoStringNorm’ (255) was utilized with the following settings:

1) Probe level correction.
2) Code Count Correction = "geo.mean" (geometric mean)
3) Background Correction = "mean.2sd", i.e. mean +/- 2 standard deviations (Background is calculated based on negative controls, the calculated background is subtracted from each sample)
4) Sample Content Correction = "top.geo.mean" (The option 'top.geo.mean' is a method which ranks miRNAs based on the sum of all samples and then takes the geometric mean of the top 75)
5) $\log_2$ transformation

3.3.4 Survival analysis

The normalized data were filtered to remove miRNAs with low expression ($C_T > 35$) in over 20% of samples. LASSO regression was applied to the normalized TLDA miRNA expression data for the training cohort (256), to select a subset of miRNAs associated with DFS. The parameter was obtained through cross-validation (0.087942). The DFS estimates were calculated based on the Kaplan-Meier method. The hazard ratios were obtained from the unadjusted Cox regression method; miRNAs significant for DFS were selected, leading to a 9-miRNA signature prognostic for DFS. A risk score was calculated using the coefficients obtained in LASSO regression and the normalized miRNA expression levels. The risk scores were dichotomized at the median, and the cohort was divided into low and high risk groups. Kaplan-Meier survival analysis was used to illustrate the difference in DFS between the high and low risk groups. The p-value associated with the two curves was determined using the log-rank test.
All statistical analyses were performed with R statistical environment (v2.15.2) using the “Survival” package for survival analysis, and “glmnet” for LASSO analysis. Significance was defined as p-values below 0.05.

Publically-available cervical cancer Illumina Small RNA-Seq data from The Cancer Genome Atlas (TCGA) Data Portal was utilized as another independent validation cohort (n = 48). Level three Small RNA-Seq (isoform_expression) data and clinical information for the 48 cervical cancer patients were downloaded from the Broad Firehose (stddata run 2013_06_06). Reads-per-million (RPM) data were log2 transformed and z-score standardized before the 9-miRNA signature equation was applied to calculate risk scores. Patients were dichotomized by the previously established cut-point (median risk score in training cohort) and compared using a log-rank test.

3.4 Results

3.4.1 The majority of significantly differentially-expressed miRNAs were downregulated

The clinical characteristics of the patients in the training (n = 79) and validation (n = 87) cohorts are provided in Table 3.1. Analysis of tumour-normal (T/N) fold changes revealed that the majority of significantly differentially-expressed miRNAs ($P < 0.01$) were downregulated in our cervical cancer samples (Table S3.1). Of the 29 significantly differentially-expressed miRNAs, only 2 were upregulated (miR-21 and miR-187).
### Table 3.1  Clinical parameters of patients in the training and validation cohorts

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Training cohort (n = 79)</th>
<th>Validation cohort (n = 87)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>Range</td>
<td>26-84</td>
<td>19-83</td>
</tr>
<tr>
<td><strong>Tumour size</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 5 cm</td>
<td>48 (61%)</td>
<td>43 (52%)</td>
</tr>
<tr>
<td>&gt; 5 cm</td>
<td>31 (39%)</td>
<td>39 (48%)</td>
</tr>
<tr>
<td><strong>FIGO stage</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IB</td>
<td>24 (30%)</td>
<td>22 (25%)</td>
</tr>
<tr>
<td>IIA</td>
<td>2 (3%)</td>
<td>5 (6%)</td>
</tr>
<tr>
<td>IIB</td>
<td>35 (44%)</td>
<td>31 (36%)</td>
</tr>
<tr>
<td>IIIA</td>
<td>0</td>
<td>2 (2%)</td>
</tr>
<tr>
<td>IIIB</td>
<td>18 (23%)</td>
<td>27 (31%)</td>
</tr>
<tr>
<td><strong>Pelvic or para-aortic node involvement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>25 (32%)</td>
<td>29 (33%)</td>
</tr>
<tr>
<td>Equivocal</td>
<td>15 (19%)</td>
<td>18 (21%)</td>
</tr>
<tr>
<td>Negative</td>
<td>39 (49%)</td>
<td>40 (46%)</td>
</tr>
<tr>
<td><strong>Overall survival</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deaths</td>
<td>24 (31%)</td>
<td>26 (30%)</td>
</tr>
<tr>
<td><strong>Disease-free survival</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relapses or deaths</td>
<td>28 (35%)</td>
<td>33 (38%)</td>
</tr>
<tr>
<td><strong>Follow-up (years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>6.0</td>
<td>5.3</td>
</tr>
<tr>
<td>Range</td>
<td>0.7-10.6</td>
<td>1.0-10.5</td>
</tr>
</tbody>
</table>
3.4.2 Patient miRNA expression was associated with disease-free survival

Using LASSO regression on the normalized TLDA data from the training cohort, we derived the following model to calculate the risk score for each patient using the expression values of nine miRNAs:

$$\text{Risk Score} = (0.197109 \times E_{\text{let}-7c}) - (0.07048 \times E_{\text{miR}-21}) + (0.045797 \times E_{\text{miR}-222}) - (0.56469 \times E_{\text{miR}-451}) + (0.171838 \times E_{\text{miR}-455-5p}) - (0.01725 \times E_{\text{miR}-134}) + (0.506956 \times E_{\text{miR}-148a}) + (0.203466 \times E_{\text{miR}-218}) + (0.22355 \times E_{\text{miR}-500})$$

Where $E_X$ represents the normalized expression level of the given miRNA.

In this prediction model, a higher risk score would predict for worse DFS. The risk score was calculated for each patient in the training set, and the median risk score (-0.05373) was used to dichotomize the low vs. high risk groups. Our 9-miRNA signature was significantly predictive of DFS for the 79 patients in the training cohort, with a hazard ratio of 9.26 and log-rank p-value of $6.9 \times 10^{-7}$ (Fig 3.1).

3.4.3 Validation of miRNA signature

In our attempts to validate the 9-miRNA signature, we used three different methods to measure miRNA expression in the validation cohort: 1) TLDA ($n = 87$); 2) NanoString ($n = 87$); and 3) qRT-PCR ($n = 68$; 19 samples omitted due to insufficient RNA). A risk-score was calculated for each of the patients by applying the miRNA TLDA expression values to our prediction model. The patients were divided into the high vs. low risk groups based on their risk score, using the same cut-off point as defined in the training cohort (Fig 3.2A). This analysis was repeated for the NanoString and the qRT-PCR datasets (Fig 3.2B-C). Regardless of the method used for measuring miRNA expression, the 9-miRNA signature was not significant when applied to the patients in the validation cohort. We also utilized publically-available cervical cancer Illumina Small RNA-Seq data derived from The Cancer Genome Atlas (TCGA) Data Portal as another independent validation cohort ($n = 48$). Interestingly, this validation attempt approached statistical significance ($p = 0.05251$) (Fig S3.1A).
Figure 3.1  Kaplan-Meier analysis of DFS according to 9-miRNA signature

A risk score was calculated for each patient in the training cohort (n = 79) using our 9-miRNA signature for DFS in cervical cancer. The median risk score was used to divide patients into the high vs. low risk groups. HR; hazard ratio, DFS; disease-free survival, CI; 95% confidence interval.
Figure 3.2  Application of 9-miRNA signature to validation cohort

Kaplan-Meier analysis of DFS. A risk score was calculated for each patient in the validation cohort, by applying our 9-miRNA signature for DFS to the miRNA expression data generated using A) TLDA, B) NanoString, and C) individual qRT-PCR. The same cut-off point from the training set was used. HR; hazard ratio, DFS; disease-free survival, CI; 95% confidence interval.
### 3.4.4 Independent corroboration of the Hu et al. 2-miRNA signature

We attempted to perform an independent corroboration of the only published prognostic miRNA signature to date for cervical cancer by Hu, et al. \((S = 17.9 - 0.284 \times \text{EmiR-9} - 0.376 \times \text{EmiR-200a})\), where \(S\) represents the risk score for each patient, and \(\text{EmiR-9}\) and \(\text{EmiR-200a}\) represent the normalized expression levels of miR-9 and miR-200a in each patient, respectively) (120). This 2-miRNA signature was first applied to the miRNA TLDA expression values from our training frozen samples \((n = 79)\) to calculate a risk-score for each patient. Although the authors used 0 as the cut-off point with their training set, we were not able to use this value for corroboration because the risk scores were all above 0. We thus divided our patients into high vs. low risk groups, with 1/3 in the former and 2/3 in the latter categories, which reflected the Hu et al. population when 0 was used as their cut-off. Based on the Kaplan-Meier analysis, the 2-miRNA signature was not significant when applied to the patients in our frozen cohort utilizing the TLDA platform (Fig 3.3A). This analysis was repeated for the FFPE TLDA, and FFPE NanoString datasets; neither of which could corroborate the 2-miRNA signature (Fig 3.3B-C). In a final attempt to corroborate the Hu et al. signature, we utilized Small RNA-Seq data derived from the TCGA Data Portal as yet another independent cohort \((n = 48)\). Using the same analytical methods, again, the 2-miRNA signature could not be corroborated (Fig S3.1B), although at least the trend was in the correct direction, in contrast to the other 3 datasets.

An attempt was also made to cross-validate our own 9-miRNA signature with the miRNA expression data from the Hu, et al. study; unfortunately, access to their raw data was not provided; hence no further analysis was possible.
Using the Hu et al. 2-miRNA signature, a risk score was calculated for each patient from: A) TLDA frozen cohort (n = 79), B) TLDA FFPE cohort (n = 87), and C) NanoString FFPE cohort (n = 87). HR; hazard ratio, DFS; disease-free survival, CI; 95% confidence interval.
3.5 Discussion

In the clinical management of cancer patients, prognostic evaluation is essential to guide appropriate treatment decisions. Unfortunately, in cervical cancer, there remains significant differences in patient survival despite being assigned to the same clinical stage; underscoring the gaps in the current system, as well as the need to develop more useful prognostic biomarkers. We developed a candidate 9-miRNA signature that was prognostic for DFS in patients with cervical cancer. According to our analysis, our validation cohort (n=87) was large enough to detect a signature with a HR of 2.5 with 83% power. However, we could not validate our signature (HR = 9.26) in our validation cohort, despite trying three separate techniques. Furthermore, attempts to corroborate the only published miRNA signature to date for cervical cancer (120) in three independent patient cohorts were also unsuccessful.

We believe that there are three key reasons as to why we failed to validate our candidate 9-miRNA prognostic signature for cervical cancer: i) intra-tumour heterogeneity; ii) miRNA expression data from frozen and FFPE samples could not be directly compared in this disease; and iii) the current platforms for global miRNA expression profiling are not sufficiently robust.

Firstly, it is well-established that human tumours are intrinsically heterogeneous. Two types of tumour heterogeneity exist: a) inter-tumour heterogeneity, with differences between tumours arising from different patients; and b) intra-tumour heterogeneity, with differences between distinct sub-populations of cancer cells within a single individual’s tumour. In cervical cancer, intra-tumour heterogeneity has been described on various levels. Many studies have demonstrated significant variations in interstitial fluid pressure (257), blood perfusion (258), and oxygen tension (259) in different regions within an individual tumour. At the chromosomal level, intra-tumour heterogeneity has been described with respect to specific genetic mutations and chromosomal abnormalities such as gains and deletions, reflecting the polyclonal derivation of cervical cancer (260-264). At the mRNA transcript level, Bachtiary et al. evaluated intra-tumoural heterogeneity across 11 cervical cancer patients, and demonstrated that multiple biopsies from distinct areas of an individual tumour were necessary to reduce sampling bias, with genes displaying low intra-tumour heterogeneity requiring two to three biopsies, and genes with high intra-tumour heterogeneity requiring more than six biopsies per tumour (265).
To date, intra-tumoural heterogeneity of miRNA expression has only been reported in breast cancer thus far (266). However, natural inter-patient variability has been shown to exist among normal cervix samples, which complicates miRNA expression profiling studies (162). Given that intra-tumour heterogeneity exists in cervical cancer at the macroscopic, chromosomal and transcript levels, it would be a logical extension to assume that this would also apply to miRNAs. In our study, we only utilized one biopsy from each patient in the training and validation cohorts, which is therefore probably insufficient to obtain a representative measure of miRNA expression for the patient’s entire tumour.

A second reason for the lack of signature validation is the lack of concordance with respect to the tissue preservation method used for the two patient cohorts; specifically, the training and validation sets consisted of frozen and FFPE samples, respectively. This experience would appear to contradict several previous studies that have reported high concordance in the miRNA expression data from tissue-matched frozen and FFPE samples derived from various types of human tissues, including: breast (96), lung (107), kidney (108), skin (106), glioblastoma (109), melanoma (111, 267), prostate (105, 112), and lymph nodes (268). However, with the exception of our own single study that used the TLDA platform (96), the remaining reports utilized other miRNA profiling technologies, such as microarray (106, 107, 111, 267, 268), deep sequencing (107, 108), custom PCR array (105), and qRT-PCR using stem-loop (112) or locked-nucleic-acid primers (109). There has only been one report to date that evaluated miRNA expression data from tissue-matched frozen and FFPE cervix samples, which only analyzed 3 cervix specimens, in addition to 3 breast and 2 gall bladder samples (269). This report by Doleshal et al. analyzed 3 miRNAs (miR-24, miR-103, miR-191) using qRT-PCR, and only calculated the ΔCt values between frozen and matched FFPE samples without performing any correlation tests. Furthermore, several of these published reports have demonstrated that although there were high correlations (> 0.5) between tissue-matched frozen and FFPE samples for the overall panel of miRNAs tested, there were specific individual miRNAs that were poorly correlated between the matched samples (108, 109, 111), and sometimes even miRNAs that demonstrated opposing patterns of under- or over-expression in tissue-matched sample pairs (105, 112). The reasons behind these discrepancies, as to why some miRNAs correlate between frozen and FFPE samples; yet others do not, remain unclear.
Lastly, we believe that the current methods used for global miRNA expression profiling are not sufficiently robust. We used the TLDA and NanoString platforms for global miRNA expression profiling, which are both widely-used for miRNA expression profiling. However, in our own recent analyses, the correlation in miRNA expression levels between these two platforms, even using the same FFPE RNA samples, had only a $\rho$ of 0.65 (manuscript in preparation), underscoring the challenges in the technologies to be able to reliably identify clinically useful biomarkers in this disease. This might also provide one technical explanation for the recent review describing the difficulties in validating miRNAs for human malignancies (270). More recent reports have demonstrated the advantages of deep sequencing, including increased sensitivity and specificity with few false-positive calls (115, 116). Given that the majority of differentially-expressed miRNAs in cervical cancer are down-regulated, deep sequencing would likely be a more suitable and robust platform to identify and validate a prognostic miRNA signature in this disease.

Our validation attempt using Small RNA-Seq data from the TCGA data portal was promising, with our 9-miR signature predicting worse outcome for high-risk patients than low-risk patients ($p = 0.053$). Interestingly, the datasets from the TCGA cohort and our training cohort were both generated from frozen tissues. With the addition of more cervical cancer samples to the TCGA data portal with survival information and miRNA expression data, we could potentially obtain a statistically significant result in a future validation attempt.

In conclusion, a prognostic miRNA signature for cervical cancer could not be validated, due to intra-tumoural heterogeneity, incompatibilities between miRNA expression data from frozen and FFPE samples, and insufficiently robust technical platforms for global miRNA expression profiling. We propose that this could potentially be resolved in the near future, with the advancement of technologies for miRNA expression profiling such as deep-sequencing. At present, the TCGA Data Portal contains deep sequencing miRNA expression data with clinical annotation for only 48 cervical cancer patients; when this dataset is expanded to include more samples, it could potentially be utilized as an important resource to identify and corroborate potential miRNA signature sets. Our observations provide an important cautionary tale for future miRNA signature studies for cervical cancer, which can also be potentially applicable to miRNA profiling studies involving other types of human malignancies.
**Table S3.1**  Significantly differentially-expressed miRNAs in cervical cancer

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Fold Change (log$_2$)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-149</td>
<td>-4.30</td>
<td>$2.09 \times 10^{-9}$</td>
</tr>
<tr>
<td>let-7c</td>
<td>-3.42</td>
<td>$5.69 \times 10^{-7}$</td>
</tr>
<tr>
<td>miR-218</td>
<td>-3.79</td>
<td>$5.14 \times 10^{-6}$</td>
</tr>
<tr>
<td>miR-139-5p</td>
<td>-2.75</td>
<td>$1.44 \times 10^{-5}$</td>
</tr>
<tr>
<td>miR-203</td>
<td>-3.43</td>
<td>$1.44 \times 10^{-5}$</td>
</tr>
<tr>
<td>miR-125b</td>
<td>-2.80</td>
<td>$1.44 \times 10^{-5}$</td>
</tr>
<tr>
<td>miR-376c</td>
<td>-3.08</td>
<td>$4.77 \times 10^{-5}$</td>
</tr>
<tr>
<td>miR-361-5p</td>
<td>-3.76</td>
<td>$1.15 \times 10^{-4}$</td>
</tr>
<tr>
<td>miR-320a</td>
<td>-1.07</td>
<td>$1.33 \times 10^{-4}$</td>
</tr>
<tr>
<td>miR-324-3p</td>
<td>-1.85</td>
<td>$1.67 \times 10^{-4}$</td>
</tr>
<tr>
<td>miR-196b</td>
<td>-1.59</td>
<td>$2.07 \times 10^{-4}$</td>
</tr>
<tr>
<td>let-7b</td>
<td>-1.48</td>
<td>$4.24 \times 10^{-4}$</td>
</tr>
<tr>
<td>let-7a</td>
<td>-1.84</td>
<td>$4.24 \times 10^{-4}$</td>
</tr>
<tr>
<td>miR-100</td>
<td>-2.61</td>
<td>$4.24 \times 10^{-4}$</td>
</tr>
<tr>
<td>let-7e</td>
<td>-1.29</td>
<td>$1.16 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-193b</td>
<td>-1.14</td>
<td>$1.23 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-99a</td>
<td>-2.86</td>
<td>$1.32 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-328</td>
<td>-1.84</td>
<td>$1.35 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-195</td>
<td>-1.82</td>
<td>$1.35 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-21</td>
<td>1.89</td>
<td>$1.35 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-24</td>
<td>-1.10</td>
<td>$1.43 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-574-3p</td>
<td>-1.57</td>
<td>$2.40 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-375</td>
<td>-5.72</td>
<td>$2.43 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-148a</td>
<td>-2.22</td>
<td>$2.43 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-29a</td>
<td>-1.51</td>
<td>$2.70 \times 10^{-3}$</td>
</tr>
<tr>
<td>let-7g</td>
<td>-0.91</td>
<td>$6.46 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-210</td>
<td>-1.70</td>
<td>$6.80 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-455-5p</td>
<td>-2.26</td>
<td>$8.50 \times 10^{-3}$</td>
</tr>
<tr>
<td>miR-187</td>
<td>2.85</td>
<td>$8.70 \times 10^{-3}$</td>
</tr>
</tbody>
</table>
Figure S3.1  Application of TCGA Small RNA-Seq data to cervical cancer signatures

miRNA expression data from the TCGA miRNASeq cohort (n = 48) was used to test: A) our 9-miRNA signature, and B) the Hu et al. 2-miRNA signature. HR, hazard ratio; CI, 95% confidence interval.
CHAPTER 4: IDENTIFICATION AND CORRECTION OF AMPLIFICATION PROTOCOL BIAS IN MICROARRAY STUDIES

The data in this chapter have been submitted for publication and is currently under review:

Yan R*, How C*, Saha S, Haider S, Hui ABY, Waggott D, Chong LC, Harding NJ, Kasprzyk A, Clarke BA, Fyles A, Reich HN, Liu FF, and Boutros PC. *These authors contributed equally to this work.
4.1 Chapter Abstract

Background: Microarray platforms allow the simultaneous measurement of thousands of transcript abundances, but their clinical use has been limited because surgical specimens often yield low quantities of highly-degraded RNA. To resolve this issue, several novel amplification protocols have been developed. We evaluated the bias introduced by alternative amplification protocols by systematically comparing mRNA abundance profiles generated using standard and low-input/low-quality amplification protocols from matched samples.

Methods: RNA extracted from 12 cervix cancers and 2 unmatched normal cervix samples was divided into two aliquots. One was amplified using a low-input protocol and the other using standard techniques; both samples were analyzed on Affymetrix microarrays. Raw array data were preprocessed using the Robust Multi-array Average (RMA) algorithm, followed by linear-modeling and correlation analyses to identify genes showing different patterns between the two protocols. Agglomerative hierarchical clustering was used to visualize the distribution of the mRNA abundances across the protocols. Significant genes were then extracted for pathway and sequence analyses to determine whether differentially expressed genes are biased in gene structure or function. Lastly, to generalize these conclusions we repeated this experimental approach in two normal kidney samples and one Michigan reference RNA sample (RNA control).

Results: Data generated using differed dramatically between the two amplification protocols, with systematic biases affecting thousands of genes. Altered genes were biased towards specific biological pathways, but did not exhibit unique sequence characteristics. Batch correction protocols reduced, but did not eliminate, protocol bias.

Conclusions: Each amplification protocol creates a unique, systematic bias in microarray data. As a result, experiments performed using different protocols cannot be directly merged. Batch-correction methods partially mitigate these effects, but it remains essential to ensure consistency in the amplification protocol when comparing microarray data.
4.2 Introduction

mRNA microarrays are the most widely used high-throughput technology for functional genomics. Modern microarrays measure the abundances of tens of thousands of transcripts in parallel, based on fluorescent nucleotide labeling and hybridization between complementary sequences (271). Despite the emergence of sequencing-based methods, they remain useful in both basic (272, 273) and translational (274, 275) research, in part because of their cost-effectiveness and favourable signal-to-noise ratios for low-abundance transcripts.

Importantly for clinical applications, the error-characteristics of microarrays have been extensively characterized. Comprehensive empirical studies have led to the development and assessment of sophisticated analysis approaches. For example, the statistical power of a study can be increased by direct integration of intensity-level data from samples measured using two different microarray technologies (276). Despite differing surface chemistry, signal-detection, and hybridization methods, cross-platform comparisons such as the MAQC studies have shown that most platforms are inherently reliable (277, 278). Indeed, biomarkers developed by different teams and using different microarray platforms show remarkable concordance (279). Hence, expression microarrays currently remain the best-studied and most established “omic” technology.

Surprisingly, despite this extensive characterization and consequent wide adoption of microarrays in discovery research, the uptake of microarrays as clinical diagnostic tools remains relatively limited. Outside of a small number of high-profile projects (280), most successful transcriptomic biomarkers are implemented using medium-throughput techniques such as real-time PCR (281-283), even when the initial discovery used microarrays. This occurs for several reasons. First, clinical samples are usually archived after undergoing formalin-fixation followed by embedding in paraffin (FFPE), a procedure used because it maintains morphological features of the tissue for histological analysis. However it is very challenging to extract high-quality RNA from FFPE tissue samples because paraffin embedding does not preserve nucleic acids well (103) and formalin fixation cross-links nucleic acids and proteins (284, 285). Second, clinical specimens are often small in size, thus yield a small quantity of RNA. For example, two 10 µm sections of a breast cancer biopsy yielded only 150 ng of total RNA, an amount insufficient for standard microarray labeling procedures (286). Finally, independent of preservation method and
specimen size, nucleic acids isolated from clinical specimens are often degraded relative to those from model systems. After tumour tissue is surgically excised, it is usually sent for pathological assessment prior to freezing or fixation. This lag between surgery and tissue preservation significantly influences RNA quality, and affects mRNA abundance patterns (287, 288). Similarly, thawing of frozen tissues prior to and during RNA isolation contributes to RNA degradation and changes isoform abundances (289).

Several approaches have been taken to mitigate these challenges and to enable routine microarray profiling of clinical specimens. Unfortunately, to date there have been no systematic comparisons of the accuracy and bias of these protocols, nor a template for how data created using optimized protocols can be merged with data generated from traditional approaches. Here, we take the first step towards resolving these challenges by evaluating the bias introduced by the NuGEN WT-Ovation Pico protocol for the robust linear amplification of limited amounts of significantly degraded RNA. This protocol works with transcripts that have lost their poly-A tail because it combines random- and poly-A priming (290). We considered three issues. First, we determined if the use of a low-input/low-quality RNA amplification protocol altered transcriptional profiles either globally, or for specific genes. Second, we investigated whether these changes altered biological conclusions drawn from the dataset. Third, we examined if there was a bias towards specific classes of genes, either in terms of their sequence characteristics (e.g. length, GC-content) or biological function.

4.3 Materials and Methods

4.3.1 Samples

Twelve flash-frozen punch biopsies were obtained from patients with locally advanced cervix cancer. After biopsy, the specimens were placed in optimal cutting temperature (OCT) storage medium for histopathologic examination, then flash-frozen in liquid nitrogen. H&E-stained tissue sections were cut from the OCT-embedded material and evaluated by a gynecologic oncology pathologist (B Clarke). Total cell content (stroma and tumour cells) were estimated for all tissue samples using a light microscope, to ensure that all samples contained >70% malignant epithelial cells. Flash-frozen normal cervix tissues obtained from two patients who underwent hysterectomy for benign causes served as normal comparators.
Kidney biopsies were performed at University of Health Network Toronto using an ultrasound-guided automated 18-gauge needle system. A section of the third clinical core obtained for kidney biopsy is routinely stored in RNAlater (Life Technologies, Carlsbad CA) at 4C. Once determined that this tissue is not required for clinical diagnosis, this core segment is available for research purposes as per institutional informed consent process. The RNA mixture control sample is a Stratagene Universal human reference RNA sample (Catalog Number: 740000-41). All tissue acquisition protocols (for both cervix and kidney samples) were approved by the Research Ethics Board of the University Health Network.

4.3.2 Sample processing

Two sections of 50-µm thickness were cut from the OCT-embedded flash-frozen cervix tissues and placed in a nuclease-free microtube. Total RNA was isolated using the Norgen Total RNA Purification Kit (Norgen Biotek), according to the manufacturer’s instructions, and aliquoted for further analysis. RNA quantity and quality were determined using the NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies) and the Agilent 2100 Bioanalyser RNA 6000 Pico LabChip kit (Agilent Technologies), respectively.

The kidney biopsy core was micro-dissected using a stereomicroscope to separate the glomerular and tubulo-interstitial compartments as previously described [33]. Following tissue disruption and homogenization total RNA was isolated using the RNeasy kit (Qiagen, Valencia CA), and eluted in 30 µL of RNase-free water. The sample obtained from the tubulo-interstitial compartment was split into two aliquots. RNA quantity and quality were evaluated as described above for cervix samples.

4.3.3 Microarray hybridization

Each aliquot of total RNA from the 12 cervix cancer and 2 normal cervix tissue samples was reverse-transcribed and linearly amplified using one of two methods: 1) 50 ng of total RNA was reverse transcribed into cDNA and linearly amplified using the NuGEN WT Ovation Pico kit, according to the manufacturer’s protocol. The cDNA was fragmented and biotin labelled using the NuGEN FL-Ovation cDNA Biotin Module v2.; 2) 200 ng of total RNA was reverse transcribed into cDNA with the Affymetrix GeneChip One-Cycle cDNA Synthesis Kit and linearly amplified using the Affymetrix 3’ IVT Express kit, according to the manufacturer’s protocol. For both reverse transcription/amplification methods, the quality and quantity of the
amplified, labelled cDNA product was confirmed using the Agilent 2100 Bioanalyzer, and then hybridized to the Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays, according to the manufacturer’s protocol.

4.3.4 Pre-processing

Quantitative microarray data (CEL files) were first loaded into the R statistical environment (v2.15.2) using the affy package (v1.36.0) of the BioConductor library. Samples were measured in three different batches. All RNA samples were subjected to quality assessment: RNA Integrity Numbers (RINs) ranged from 2.2 – 8.5 (Table S4.1); no arrays were excluded. RINs were rounded to the nearest integer for clustering analyses. These data were then pre-processed with Robust Multi-array Average (RMA) algorithm (291). To remove unexpressed genes, we applied a Y chromosome filter (i.e., signal intensity < 5, Figure S4.6), as described previously (265).

For some analyses we considered data produced by the two protocols separately and for others we merged them as a complete dataset. We therefore used the Robust Multi-array Average (RMA) pre-processing algorithm and the Y chromosome filtering method three times, once on the entire dataset and once for each of the two protocol-specific datasets. ProbeSet mappings were updated to modern Entrez Gene ID annotations as implemented in R package hgu133plus2hsentrezgcdf (v15.0.0) (292), and these were annotated with Entrez Gene information from NCBI as of 2012-12-05. Raw and pre-processed data are available in the GEO repository (GSE42764).

4.3.5 Unsupervised machine-learning

Unsupervised hierarchical clustering was used to identify the strongest trends within the dataset. The agglomerative hierarchical clustering using complete linkage and Pearson's correlation were used as a similarity metric chosen for clustering, as implemented in the cluster package (v1.14.3) for the R statistical environment (v2.15.2). The randomness of clustering patterns was assessed using the Adjusted Rand Index (ARI), implemented in the mclust package (v3.5).
4.3.6 Correlation analyses

To assess if Nugen- and Express-processing yielded identical ordering of samples of genes, Spearman's rank-order correlation coefficient was applied between cervix samples processed with the Nugen protocol and those processed with the Express protocol. In the replication analysis (i.e., 2 normal kidney samples and 1 RNA reference), we performed Pearson's correlation instead of Spearman's rank-order correlation because Pearson’s correlation is able to capture the degree of concordances or discordances, while Spearman’s rank-order correlation is likely to overlook the differences between small and large concordances or discordances for small sample size.

4.3.7 Statistical analysis of microarray data

To address the effects of the Nugen and Express protocols, we performed three separate linear-modeling analyses, all using the limma (v3.14.2) package for the R statistical environment (v2.15.2).

First, we pre-processed all data together and fitted a model to explicitly test the effect of protocol on each gene. The model used was: \( Y_i = \alpha_{0,i} + \alpha_{1,i}X_i \), where \( Y \) is a binary indicator value representing protocol (Nugen vs. Express) and \( X \) is the expression of the 14 replicate Nugen and Express arrays for the \( i \)-th gene. A Linear Transformation of Replicates (LTR) was then used in an attempt to remove platform-and batch-specific biases (293). Normalized data from a subset of arrays were fitted into the LTR model, which was then used to adjust the remainder of the data. The LTR analysis was performed with the LTR (v1.0.0) package for R (v2.15.2).

Second, we pre-processed and modeled the Nugen and Express datasets separately to determine if they yielded similar conclusions about biological effects. The model used was again: \( Y_i = \alpha_{0,i} + \alpha_{1,i}X_i \), but here \( Y \) is a binary indicator value representing the biological effect (tumour vs. normal) and \( X \) is the expression of the fourteen replicate Nugen and Express arrays for the \( i \)-th gene. This model was fitted twice, once to the Nugen data and once to the Express data.

After model-fitting, an empirical Bayes moderation of the standard error was applied (294), followed by model-based t-tests to determine if each \( \alpha_{1,i} \) was significantly different from zero. A false-discovery rate (FDR) adjustment was used to control for multiple testing (295).
4.3.8 Gene subset division

For downstream analyses we selected three subsets of genes. First, we identified genes whose mRNA abundance measurements were affected by protocol as those with $p_{\text{adjusted}} < 10^{-4}$ in both protocols. This threshold ensured that the selected genes were statistically significant and the number was reasonable for downstream analyses. Second, we identified genes that showed tumour/normal differential expression in data generated using the Express protocol as those with $p_{\text{adjusted}} < 0.05$ in the Express-only model. Third, we identified genes that showed tumour/normal differential expression in data generated using the Nugen protocol as those with $p_{\text{adjusted}} < 0.05$ in the Nugen-only model.

4.3.9 Functional analysis

Pathway analysis was performed using GOMiner (v2009-09) (296). GOMiner compares a list of differentially expressed genes to the complete set of genes tested and assesses enrichment of Gene Ontology (GO) annotations. We used a false discovery rate threshold of 0.1, generated by 1000 randomizations, and tested all human databases, look-up options, GO evidence codes and ontologies. This analysis was performed on all three of the gene sets described above.

4.3.10 Sequence analysis

Sequence annotations (gene length, GC content, untranslated sequence length and number of exons) for the genes affected in each of the Express and Nugen samples, differentially expressed genes in the protocol and the total genes on the array were retrieved from Ensembl BioMart v61 (Human genome assembly GRCh37). The GC content was calculated using the formula:

$$\frac{G+C}{A+T+G+C} \times 100$$  \hspace{1cm} (1)

The multiple transcripts for each gene were further processed to compute mean untranslated (UTR) sequence length using the formula:

$$UTR_{\text{gene}} = \frac{1}{n} \left( \sum_{i=1}^{n} 5' \text{UTR}_i \right) + \frac{1}{n} \left( \sum_{i=1}^{n} 3' \text{UTR}_i \right)$$  \hspace{1cm} (2)
Where \( n \) is the total number of transcripts for a given gene. 5'UTR and 3'UTR represents five-prime and three-prime untranslated sequence length respectively.

The number of exons across multiple transcripts was averaged to compute the number of exons for the corresponding gene. The GC content of all genes on the array were compared to the GC content of genes that were differentially expressed in the protocols and to the genes that were affected in each of the protocols using two sample unpaired t-test with Welch's correction for heteroscedasticity. Similarly, the gene length, UTR length and the number of exons between the above-mentioned groups of gene sets were compared using the unpaired Wilcoxon rank-sum test. All analyses were performed in the R statistical environment (v2.15.2).

4.3.11 Visualization

Visualization employed the lattice and latticeExtra packages (v0.20-11 and v0.6-24 respectively) for the R statistical environment (v2.15.2). P-value sensitivity plots were produced by plotting all genes altered at 100 \( p_{\text{adjusted}} \) cut-offs spanning from \( p_{\text{adjusted}} = 1 \) to \( p_{\text{adjusted}} = 10^{-12} \). Venn diagrams were created using the VennDiagram package (v1.5.2) (297). All clustering analyses employed agglomerative hierarchical clustering using complete linkage and Pearson’s correlation as the similarity metric as implemented in the cluster package (v1.14.3).

4.4 Results

4.4.1 Experimental approach

Our aim was to evaluate the effects of low-input/low-quality RNA amplification methodologies and to determine if their use introduced any biases (systematic or stochastic) into microarray data. To achieve this, we selected 12 cervix cancers and two unmatched normal cervix samples, which were flash-frozen in liquid nitrogen after biopsy. Total RNA was extracted from each sample and divided into two aliquots. One was reverse transcribed and amplified using the NuGEN WT-Ovation Pico protocol (henceforth simply “Nugen”), while the other was processed using the standard Affymetrix 3’ IVT Express protocol (henceforth “Express”). Each aliquot was hybridized to an Affymetrix HG-U133 Plus 2.0 microarray. We selected this platform because of its widespread use (2,854 GEO experiments encompassing 77,923 samples at writing). Lastly, to generalize these conclusions we repeated this analysis with
two normal kidney samples and one RNA mixture control sample. Figure 4.1 summarizes the experimental design and analysis workflow.

4.4.2 Nugen and express data are not directly comparable

We selected cervix samples with a broad range of RNA quality, with RNA Integrity Numbers (RINs) ranging from 2.2 to 8.5 (Table S4.1). Following pre-processing, we first used unsupervised machine-learning to identify the strongest trends within the dataset (Figure 4.2A). Samples processed with the Nugen protocol (purple annotation bars) clustered separately from those processed with the Express protocol (yellow annotation bars). To test the significance of this observation we calculated the Adjusted Rand Index (ARI), which ranges from -1 to +1, with higher values indicating non-random associations. Despite using identical samples, the ARI for protocol bias was 1.0, compared to ARI of -0.02 for the biological differences between tumour and normal (red/white annotation bars), 0.26 for different microarray hybridization batches (grey-scale annotation bars) and -0.07 for RNA quality (white-to-green annotation bars). This technical bias was not a function of low-expression (Figure S4.1A) or low-variability genes (Figure S4.1B).

These results showed that the two protocols produced distinctly different data. However, these differences might not be sufficient to change biological conclusions. For example, if different protocols created a global scaling artifact that survived data pre-processing, it would lead to rank-order comparable data that could be evaluated using non-parametric statistical methods. We therefore assessed if the two protocols yielded identical ordering of samples for each gene, using Spearman’s correlation coefficient. In total 96% of genes (14,289/14,859) were positively correlated between the two protocols ($\rho > 0$), suggesting a broad similarity between the two methods (Figure 4.2B). However, only 17.6% of genes (2,620) were strongly correlated ($\rho > 0.90$). In fact, 4% of genes (570) showed opposite trends between the two protocols ($\rho < 0$; Figure 4.2C), and these genes were only weakly biased towards lower mRNA abundances (Figure 4.2D). Taken together, these data demonstrated that Nugen & Express data were not directly comparable, even in a rank-order manner.
Figure 4.1  Experimental design

Raw array data were preprocessed by Robust Multi-array Average (RMA) algorithm. Three normalized datasets were generated: Nugen, Express and Protocol (all samples). Agglomerative hierarchical clustering was used to visualize the distribution of the mRNA abundances across the protocols. Linear models were then applied to the three datasets to examine the technical/biological effects. Downstream analyses were performed on significant genes to reveal the effects of technical and biological variables. Significant genes were then extracted for pathway and sequence analyses to determine whether differentially expressed genes are biased in gene structure. The LTR algorithm was used to remove technical biases. Lastly, to generalize these conclusions we repeated this experimental approach in two normal kidney samples and one Michigan reference RNA sample (RNA control). Experiments performed in cervix data and kidney data were indicated as 1 and 2 respectively.
Following pre-processing, we studied the relationship between Nugen-Express effects and other aspects of the microarray. A) On this sample vs. sample heatmap, the white-to-blue colour bar reflects the range of Pearson's correlations. The four-column annotation bars indicate array protocol, biological effects (tumour vs. normal), hybridization batch and RIN. The strong separation by protocol highlights the magnitude of technical artifacts. B) A histogram of gene-wise Spearman’s $\rho$ between Nugen and Express samples was skewed to the right, suggesting most ProbeSets show similar trends. C) However a cumulative plot shows that only a small fraction of genes were highly correlated. D) Correlation between Nugen and Express protocols was not strongly associated with signal intensity (a proxy for mRNA abundance).
4.4.3 Protocol introduces systematic bias

We next sought to determine if the RNA amplification protocol creates a systematic bias that preferentially affects specific genes, or a stochastic one that altered all genes (or a random subset thereof). This is critical because if a reproducible and limited set of genes is affected by protocol, these can be excluded from comparative analyses and data from the two protocols can be safely integrated.

We performed gene-wise linear-modeling to assess protocol effects. Stochastic effects would be expected to yield few reproducible differences between matched samples. Instead, we identified large numbers of genes altered in a reproducible manner by protocol (Figure 4.3A). The dashed line indicates a 1% false-discovery rate; at this stringent threshold, 72% of the array (10,703 genes) was altered by protocol. These effects were very large in magnitude: 86 genes exhibit at least a 10-fold difference between protocols. Both their magnitude/direction (Figure 4.3B) and statistical significance (Figure 4.3C) were independent of expression level. This systematic bias was only weakly related to the poor correlations outlined in Figure 4.2B: ~24% (620/2,620) of the highly correlated genes (\( \rho > 0.9 \)) differed by at least two-fold between protocols. Interestingly, negatively correlated genes showed generally smaller magnitude differences (Figure 4.3D). Thus amplification protocols systematically bias the majority of the transcriptome in a gene-specific manner.

4.4.4 Protocol noise is larger than biological signal

To put the magnitude of protocol-specific changes in context, we compared this effect to the tumour/normal differential expression between normal and malignant cervix tissues. We compared linear models that were fitted separately on the Nugen and Express data to the technical effects described above. Independent of the statistical threshold, an order of magnitude more genes were altered by protocol than by cancer, with the green curve for the protocol effect shifted far to the right of the red and blue curves in Figure 4.4A. Interestingly, protocol changes led to artifactual up- and down-regulation in equal proportions, whereas an excess of down-regulated genes were observed in cancer (Figure 4.4B), consistent with the reduction of transcriptomic complexity in malignancies (298).
Figure 4.3  Specific genes are affected by protocol bias

This figure reveals how statistically significantly differentially-expressed genes were altered by protocols. Fold change between Nugen and Express is in log2 scale, describing the changes of mRNA abundances between Nugen and Express. P-values were calculated with FDR-adjusted model-based t-tests. A) Large numbers of genes were altered in a reproducible manner by protocol changes. These effects were independent of expression level, as high magnitude (B) and highly significant protocol biases (C) were evident in genes at all signal intensities. The degree of this systematic bias was independent of the previously observed correlation differences (D).
Following a general-linear model, we identified a number of genes at various p-value thresholds. This figure shows how the genes were affected by protocol. A) The $x$-axis is p-value threshold ranging from 1 to $10^{-12}$. The $y$-axis is all genes that are differentially expressed. More genes were altered by technical differences (green curve) than by biological differences (blue and red curves) between samples. B) This figure shows the proportion of up-regulated genes at different p-value thresholds.

Figure 4.4    Differentially expressed genes between biological and technical replicates
4.4.5 Protocol difference biases specific biological functions

Our data demonstrated that changing the amplification protocols introduce systematic artifacts that affect specific genes. To determine if these changes will bias pathway analyses we subjected the set of genes showing differential signal intensity to gene ontology (GO) enrichment analysis. We again put the protocol effects in context of true biology driven by tumour-normal differences. A significant fraction of pathways affected by protocol bias were also altered between normal and tumour cells, and again protocol bias dramatically overwhelms biological signal (Figure S4.2).

4.4.6 Sequence characteristics do not explain protocol bias

Because only some genes were affected by different amplification protocols, and because these genes were affected in different ways, we hypothesized that sequence-specific characteristics were involved. We considered four gene features: GC-content, number of exons, gene length and untranslated region (UTR) length. We compared genes showing protocol differences and those showing tumour-normal differences to the overall gene cohort. Genes affected by protocol were not unusual in their GC content ($P_{P \text{ vs } A} = 1.83 \times 10^{-1}$; Figure 4.5A), number of exons ($P_{P \text{ vs } A} = 3.81 \times 10^{-1}$; Figure 4.5B), length ($P_{P \text{ vs } A} = 7.80 \times 10^{-1}$; Figure 4.5C) or UTR length ($P_{P \text{ vs } A} = 3.83 \times 10^{-1}$; Figure 4.5D), indicating that the protocol bias was not be explained by obvious sequence characteristics. However, we observed a significantly increased number of exons in tumour-associated genes ($P_{TE \text{ vs } A} = 4.88 \times 10^{-2}$, $P_{TN \text{ vs } A} = 2.00 \times 10^{-2}$; Figure 4.5B), which to our knowledge has not previously been described. Interestingly, genes identified as tumour-normal differentially expressed by the Nugen protocol had a higher GC-content than those identified using the Express protocol (red curve, Figure 4.5A), suggesting that the Nugen protocol may be more sensitive at amplifying genes with high GC-content. Traditionally, there have been limitations in the amplification of sequences with high GC-content due to higher affinity bonds between guanine and cytosine nucleotides, compared to those found between adenine and thymine (299). There are clearly many other sequence-based factors that might influence protocol bias, but the most common ones do not appear to play a role.
Figure 4.5  Differing characteristics of significantly differentially expressed genes

We were interested to determine whether differentially expressed genes between protocols differed on any structural characteristics of a gene. A) This figure shows the distribution of GC content in significantly differentially expressed genes between the different protocols, and Nugen and Express tumour and normal samples. Unpaired t-tests calculated p-values of the contrasts between distributions and found a significant difference between all genes (black curve) and Express genes (blue curve). We then studied the distribution of the number of exons (B), gene-length (C) and UTR lengths (D) of significantly differentially expressed genes in protocol, Nugen, and Express. The number of exons in all genes (black curve) versus the Express genes (blue curve) and Nugen genes (red curve) differed significantly. Comparisons for these features
(number of exons, gene-length and UTR length) were conducted using unpaired Wilcoxon rank sum test.
4.4.7 Batch effects adjustment partially mitigates protocol bias

Lastly, we wished to determine whether the technical artifacts created by protocol difference could be mitigated. LTR (Linear Transformation of Replicates) is a linear-modeling-based method that learns the relationship between different microarray batches from replicate hybridizations (293). Normalized data from half of the arrays were used to train an LTR model, which was used to adjust the remainder of the dataset. LTR removal of noise diminished the effect of protocol bias (Figure 4.6), reducing the ARI for protocol effects from 1.0 to -0.04. However a significant number of genes remained poorly correlated between the two protocols, with 11% (1,643) having negative correlations.

4.4.8 Replication analysis

To generalize our findings, we applied the same approach to normal kidney samples and an RNA reference sample. Clustering confirms a strong separation by protocol (Figure S3A, purple/yellow annotation, ARI = 1). The RNA mixture control sample was very highly-correlated between Nugen and Express (\( \rho = 0.9, p < 2.2 \times 10^{-16} \), Figure S4.4A), while the two kidney samples were strongly correlated (\( \rho = 0.71 \) and \( 0.78, p < 2.2 \times 10^{-16} \), Figure S4.4B and C).

We then examined the correlation of each ProbeSet on the microarray. We performed Pearson's correlations between samples processed with the Nugen protocol and those processed with the Express protocol. Most ProbeSets were positively correlated between the two protocols (77% of ProbeSets with \( \rho > 0 \), Figure S4.3B), and 36% (6,840/18,988) were strongly correlated (\( \rho > 0.9 \), Figure S4.3C). Moreover, 16% (1,098 /6,840) of the most strongly correlated genes in the kidney dataset (\( \rho > 0.9 \)) were also strongly positively correlated (\( \rho > 0.9 \)) in the cervix dataset (Figure 4.5). This degree of concordance was independent of expression level (Figure S4.3D). These results thus suggest that amplification bias is a fundamental property of the preparation procedure with only modest sample-specific effects.
Following an LTR-adjustment, we visualized the Pearson's correlations between testing samples (A) before and (B) after LTR-adjustment for the validation cohort. On each of sample vs. sample heatmaps, the white-to-blue colour bar reflects the range of Pearson's correlations. The four-column annotation bars indicate protocol, biological effects, batch and RIN, respectively. The strong separation by protocol in Figure 4.6A highlights the magnitude of technical artifacts, and its disappearance in Figure 4.6B suggested that the LTR algorithm removed protocol bias.

**Figure 4.6** Application of LTR algorithm to remove platform specific biases
4.5 Discussion

mRNA microarrays have become an important high-throughput technology in mRNA expression analysis. Despite their extensive characterization, their use as clinical diagnostic tools remains limited due to the challenges of amplifying low amounts of significantly degraded RNA in clinical samples. Recent studies have focused on protocols that address these challenges to enable routine microarray profiling of clinical samples (300, 301). Furthermore, the emergence of large consortia such as the TCGA and ICGC make it increasingly common to integrate data generated at multiple centres using divergent protocols. It is therefore critical to assess the sources of error so that they can be reported, controlled for, and adjusted statistically. However, to date, very few efforts have been made towards evaluating RNA amplification protocols for expression profiling.

Previous evaluations examined the reproducibility of different protocols by investigating correlations of RNA amplification methods using 8 colon cancer biopsy samples isolated using laser capture microdissection (LCM) (302), or picogram amounts of input RNA from two cell lines (303). Some researchers evaluated RNA integrity and purity (303, 304), and suggested a minimal amount of input RNA (250 pg) for amplification (303). However, none of these studies were systematically comparing accuracy and bias of these protocols, but rather focused on the agreement of different approaches rather than the integration of data generated in diverse labs and from multiple protocols. Our work fills this gap by evaluating the performances of a low-input/low-quality protocol, the NuGEN WT Ovation Pico kit, with a standard protocol, the Affymetrix 3’ IVT Express kit. We investigated 17 unique samples: 2 normal cervix samples, 12 cervix tumour samples, 2 normal kidney samples and 1 RNA mixture control sample. The three key findings from this work are: 1) Nugen and Express data are not directly comparable; 2) technical biases can be partially removed; 3) a template for the evaluation and analysis of different sample-handling protocols is proposed. We discuss these in turn.

The most prominent finding of this evaluation was the inconsistency of the two amplification protocols, given that the two amplification protocols both used linear amplification. For example, only 17.6% of genes were highly correlated ($\rho > 0.9$) between the two protocols, and over 24% of these highly correlated genes showed a difference between the two protocols of over 2-fold. The bias was not a simple global scaling artifact and was
independent of expression level. These results implied that comparing studies generated with
different protocols or mixing protocols within a study would affect differentially expressed
genes, creating artificial disparities and elevating false-positive rates. These findings were
consistent with previous reports that showed microarray data being affected by the RNA
amplification protocol used, and that data generated with the same microarray platform using
different protocols were not directly comparable (305, 306).

Moreover, we found that the technical differences (whether arrays were run with Nugen
or Express reverse transcription/amplification protocols) had a greater effect on differential
expression of genes than biological differences between samples (whether they were tumour or
normal). These discrepancies in the two protocols artificially affected specific biological
pathways and could affect the apparent biological significance of analyses. For these reasons, it
is important to utilize only one amplification methodology within a study. However, if it is
unavoidable to use multiple methodologies, there are steps that can be taken to mitigate the
technical bias. Firstly, it is imperative to use the same type of amplification, either linear or
exponential. Technical bias between two linear-amplification protocols (Nugen and Express) was
greater in magnitude than true biological differences. Hence, it is expected that the technical
artifacts would be significantly greater with mixed types of amplification protocols (linear and
exponential). Secondly, sophisticated algorithms are needed to remove technical artifacts before
interpreting the biological significance of the dataset. In this study, we applied a linear-
modeling-based method, LTR, to remove technical artifacts. The ARI was decreased from 1.00
(highly non-random clustering) to -0.04 (random association) after LTR-adjustment. However,
even after LTR-adjustment, about 11% of genes had negative correlations, suggesting technical
biases were only partially removed. There is a clear need for more sophisticated batch-effect
removal algorithms.

Lastly, this work provides a template for the evaluation and analysis of different sample-
handling protocols and can be extended to other areas. Before conducting any analysis on the
integrated multi-protocol dataset, two generalized steps are suggested: First, data normalization
remains essential, with robust multi-array average (RMA) remaining a typical method for
normalizing Affymetrix data. Second, data cleaning is needed to remove technical bias from the
normalized data. Explicit batch-effect reduction approaches can be used, with some such as LTR
(293) specializing in taking advantage of technical replicates, while others such as ComBat (307)
employing empirical Bayes models to handle unknown error structures. Unsupervised machine-learning methods are valuable for validating whether and how much technical bias is removed, particularly when combined with metrics of clustering-randomness such as ARI, which is infrequently used in bioinformatics. Lastly, linear model analysis can be utilized to fit the integrated data with batch-effect terms to explicitly model technical artifacts.

Many techniques and tools are available to measure mRNA, miRNA and DNA abundances, and this template demonstrates how to evaluate the protocol performance. Protocol-specific changes are expected to increase in magnitude as next-generation sequencing techniques, which require larger amounts of input material, become more widely employed, making careful consideration of this issue particularly important.
Table S4.1  Sample descriptions (Data: cervix data)

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Amplification Protocol</th>
<th>RIN</th>
<th>BatchNumber</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Express</td>
<td>5.4</td>
<td>Batch2</td>
<td>Tumour</td>
</tr>
<tr>
<td>1</td>
<td>Nugen</td>
<td>5.4</td>
<td>Batch3</td>
<td>Tumour</td>
</tr>
<tr>
<td>14</td>
<td>Express</td>
<td>6.5</td>
<td>Batch2</td>
<td>Tumour</td>
</tr>
<tr>
<td>14</td>
<td>Nugen</td>
<td>6.5</td>
<td>Batch3</td>
<td>Tumour</td>
</tr>
<tr>
<td>18</td>
<td>Express</td>
<td>6.0</td>
<td>Batch2</td>
<td>Tumour</td>
</tr>
<tr>
<td>18</td>
<td>Nugen</td>
<td>6.0</td>
<td>Batch3</td>
<td>Tumour</td>
</tr>
<tr>
<td>27</td>
<td>Express</td>
<td>7.1</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>27</td>
<td>Nugen</td>
<td>7.1</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>38</td>
<td>Express</td>
<td>7.9</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>38</td>
<td>Nugen</td>
<td>7.9</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>44</td>
<td>Express</td>
<td>8.5</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>44</td>
<td>Nugen</td>
<td>8.5</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>45</td>
<td>Express</td>
<td>6.7</td>
<td>Batch2</td>
<td>Tumour</td>
</tr>
<tr>
<td>45</td>
<td>Nugen</td>
<td>6.7</td>
<td>Batch3</td>
<td>Tumour</td>
</tr>
<tr>
<td>50</td>
<td>Express</td>
<td>7.0</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>50</td>
<td>Nugen</td>
<td>7.0</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>57</td>
<td>Express</td>
<td>2.2</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>57</td>
<td>Nugen</td>
<td>2.2</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>58</td>
<td>Express</td>
<td>4.9</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>58</td>
<td>Nugen</td>
<td>4.9</td>
<td>Batch1</td>
<td>Tumour</td>
</tr>
<tr>
<td>67</td>
<td>Express</td>
<td>7.4</td>
<td>Batch2</td>
<td>Tumour</td>
</tr>
<tr>
<td>67</td>
<td>Nugen</td>
<td>7.4</td>
<td>Batch3</td>
<td>Tumour</td>
</tr>
<tr>
<td>86</td>
<td>Express</td>
<td>7.4</td>
<td>Batch2</td>
<td>Tumour</td>
</tr>
<tr>
<td>86</td>
<td>Nugen</td>
<td>7.4</td>
<td>Batch3</td>
<td>Tumour</td>
</tr>
<tr>
<td>3856</td>
<td>Express</td>
<td>7.1</td>
<td>Batch2</td>
<td>Normal</td>
</tr>
<tr>
<td>3856</td>
<td>Nugen</td>
<td>7.1</td>
<td>Batch3</td>
<td>Normal</td>
</tr>
<tr>
<td>3926</td>
<td>Express</td>
<td>6.0</td>
<td>Batch2</td>
<td>Normal</td>
</tr>
<tr>
<td>3926</td>
<td>Nugen</td>
<td>6.0</td>
<td>Batch3</td>
<td>Normal</td>
</tr>
</tbody>
</table>

*Legends:*
SampleID: Unique ID for sample used in Nugen or Express
Amplification Protocol: Nugen or Express
RIN: RNA Integrity Number
BatchNumber: Three different batches, Batch 1, Batch 2, and Batch 3
Type: Sample type: Tumour vs. Normal
Figure S4.1  Distributions of Selected Genes (Dataset: Cervix data)

Following pre-processing, we visualized the genes that passed the expression-filtering selection (mRNA abundances > 6, Figure S4.1A) and variance-filtering selection (variance > 1, Figure S4.1B). On each heatmap, rows represent samples and columns represent genes. The white-to-blue colour bar at the bottom indicates scaled signal intensity of a gene. The four-column row-annotation displays four aspects: protocols, biological effects, batch numbers, and RNA Integrity numbers (RIN). The first column indicates the protocol (Nugen in purple; Express in yellow), the second indicates whether a sample is derived from a tumour (red) or a normal (white), the third indicates batch numbers, and the fourth is the discretized RINs. In this case, we see that the samples appear to cluster primarily by protocol. Most Nugen (purple) samples were clustered together, as well as Express (yellow) samples.
Figure S4.2  Overlap of significant GO terms by Protocol or sample type

Overlap of significant GO terms in a three-way interaction between biological and technical cohorts. We estimated the number of GO terms affected by the three groups at $p_{\text{adjusted}} < 0.05$. More GO terms were significant in Nugen than Express, and half of the GO terms that were significant in Express were also differentially expressed in Nugen (approximately 50%).
Following pre-processing, we revealed intra-sample relationships to study how close the samples are. We calculated pair-wise Pearson's correlations of the two kidney samples and then applied agglomerative hierarchical clustering on the correlations. A) On this heatmap, rows and columns represent samples. The white-to-blue colour bar at the bottom reflects the range of correlations [0, 1]. The one-column row-annotation displays the protocol (Nugen in purple and Express in yellow). B) Pearson’s correlation and adjusted p-values were calculated for the expression level of each gene between Nugen and Express for 2 kidney samples and one RNA mixture control sample. C) Cumulative rho tells us what percentage of ProbeSets has a specific correlation. D) Genes with a positive rho between two protocols and with high signal intensity are in the top right corner of the plot.
We examined the distributions of mRNA abundances for: (A) The RNA mixture control sample and two kidney samples (B and C) between Nugen (x-axis) and Express (y-axis). Pairwise Spearman’s rank-order correlation was calculated on each ProbeSet between Nugen and Express.
Figure S4.5  Correlation comparisons between cervix and kidney data.

To examine whether highly correlated genes in one dataset are also consistent in another, we compared the correlations of each ProbeSet on the microarray between cervix data and kidney data. X-axis is the Pearson’s correlations from kidney data and Y-axis is the Spearman’s correlations from cervix data. Highly correlated genes in both datasets can be found on the right top corner.
After RMA normalization, we performed Y-chromosome filter process to remove the noise genes. As data is from Cervix samples, we examined the distribution of genes on chromosome Y (red curve) and genes not on chromosome Y (blue curve) for differentially expressed genes in A) protocol, B) Nugen, and C) Express. Genes on chromosome Y showed low mRNA expression (intensity < 5). This suggested that mRNA with low mRNA expression (<5) was mostly noise, as the mRNA abundance was same as the genes from Y chromosome. Hence, we performed Y-chromosome filter process to remove those genes with mRNA abundance for all cervix samples lower than 5.
CHAPTER 5: SYSTEMATIC COMPARISON OF PLATFORMS FOR MICRO-RNA PROFILING

The data in this chapter have been prepared as a manuscript for submission:

How C*, Yan R*, Waggott D, Bruce J, Yin S, Hui ABY, Chong LC, Harding NJ, Liu FF, and Boutros PC. *These authors contributed equally to this work.
5.1 Chapter Abstract

**Background:** MicroRNAs (miRNAs) are small, non-coding RNAs that are involved in crucial biological processes, including development, proliferation and apoptosis. Altered miRNA expression has been described for numerous diseases, and various methods for miRNA expression profiling have been used for biomarker studies. Here, we evaluated two platforms for miRNA expression profiling, the Applied Biosystems TaqMan Low Density Array Human MicroRNA Array (TLDA) and the NanoString nCounter miRNA Expression Assay (NanoString), and both methods were also compared to a gold standard method, single-well quantitative real-time PCR (PCR).

**Methods:** Total RNA was extracted from 79 cervical cancer flash-frozen punch biopsies, 87 cervical cancer formalin-fixed paraffin-embedded (FFPE) biopsies, 144 nasopharyngeal carcinoma FFPE biopsies, and 40 breast cancer FFPE biopsies. The miRNA abundances were then measured by three different platforms, TLDA, NanoString and PCR. These data were normalized and a series of correlation analyses were then conducted on the common sample- and gene-pairs, to reveal whether genes were biased by the technical differences.

**Results:** Data generated using TLDA and NanoString were weakly, but positively correlated, suggesting that TLDA and NanoString resulted in distinct expression profiles. PCR was found to be comparable to TLDA and NanoString, but gene-wise comparisons were more similar between PCR vs. TLDA than PCR vs. Nanostring.

**Conclusions:** Experiments performed using different platforms for measuring miRNA expression cannot be directly merged, suggesting that it is essential to keep platform consistency when comparing miRNA data.

5.2 Introduction

Micro-RNAs (miRNAs) are small (~22 nucleotides), non-coding RNAs that regulate gene expression by binding to mRNA transcripts in a sequence-specific manner (2). Since miRNAs were initially reported in *C. elegans* in 1993 (6), 2042 human miRNAs have been identified thus far (miRBase Release 19) and the number continues to steadily increase. Altered
miRNA expression has been shown to be important in many human cancers (69), which is unsurprising given that miRNAs are involved in crucial biological processes, including development, proliferation and apoptosis (2, 253).

Due to their small sizes and considerable stability, miRNAs can be easily extracted from both frozen and formalin-fixed paraffin-embedded (FFPE) samples (96, 97). MicroRNA expression profiling of clinical samples has been shown to be particularly useful for distinguishing between various types of cancers (86), and classifying cancer subtypes (93, 94), including poorly-differentiated tumours (95). Furthermore, miRNA expression profiling has led to the identification of disease-specific miRNA signatures associated with diagnosis (126, 308, 309) and prognosis (119, 127, 310) for various solid and haematological malignancies.

The Applied Biosystems TaqMan Low Density Array Human MicroRNA Array (TLDA) and NanoString nCounter miRNA Expression Assay (NanoString) are two commonly used methods for miRNA expression profiling. We sought to compare these two miRNA platforms in cervical, breast and nasopharyngeal (NPC) cancers. Both platforms were also compared to single-well quantitative real-time PCR (PCR), which has long been considered the gold standard for nucleic acid quantification due to its specificity and sensitivity (311). For PCR quantification, we used the Applied Biosystems TaqMan MicroRNA Assays, which specifically measure mature miRNAs using stem-loop reverse transcription followed by TaqMan PCR analysis. The high specificity and sensitivity of this method allows it to distinguish between miRNAs that differ by a single nucleotide (114).

The TLDA method uses multiplex target-specific reverse transcription, where a panel of up to 378 miRNAs and controls are reverse transcribed in a single reaction, using a mixture of miRNA-specific stem-loop primers. After the RT step, the cDNA is loaded into a 384-well microfluidic card containing TaqMan MicroRNA Assays specific for the panel of miRNAs and controls, and the quantity of each miRNA is measured using TaqMan chemistry. A more recently developed method, NanoString, provides a direct count of miRNAs, by fluorescently tagging individual miRNA molecules and then optically scanning and counting the tagged miRNA molecules. This method is a direct measurement of miRNA levels, without bias from amplification or enzymatic reactions (312).
In our analyses of platforms for miRNA expression profiling, we conducted a cross-platform comparison between TLDA, NanoString and PCR, using FFPE and frozen samples from three cancer types (Breast cancer, Cervical cancer and NPC). We aimed to determine if the choice of platform had a significant impact on the detection of differentially-expressed miRNAs. Drawing comparisons between protocol differences and biological differences, we sought to investigate both the biological and technical effects, and the potential influence these could have on miRNA expression profiling studies.

5.3 Materials and Methods

5.3.1 Samples and sample processing

Flash-frozen punch biopsies were obtained from 79 patients with cervical cancer prior to treatment with standard chemo-radiotherapy between 2000 and 2007, inclusive. After removal from the patient, the specimens were placed in a storage medium (optimal cutting temperature compound) for histopathologic examination, then flash-frozen in liquid nitrogen. H&E-stained tissue sections were cut from the optimal cutting temperature-embedded material to ensure samples contained at least 70% tumor cells. Normal cervix tissues obtained from eleven patients who underwent hysterectomy for benign causes were also flash frozen in OCT and served as normal comparators. Two sections of 50-micron thickness were cut from the frozen cervix tissues and placed in a nuclease-free microtube. Total RNA was isolated using the Norgen Total RNA Purification Kit (Norgen Biotek), according to the manufacturer’s instructions for frozen tissue samples.

Diagnostic formalin-fixed paraffin-embedded (FFPE) blocks were collected from 87 human cervical cancer tumours between 1997 and 2000, inclusive. All samples contained at least 70% malignant epithelial cells, or were macrodissected prior to RNA purification. Nine normal cervix FFPE tissues obtained from patients who underwent hysterectomy for benign causes served as normal comparators. Ten sections of 5-micron thickness were cut from the FFPE cervix tissues and places in a nuclease-free microtube. Total RNA was isolated using the Norgen Total RNA Purification Kit, according to the manufacturer’s instructions for FFPE tissue samples.
FFPE primary biopsies were collected from 144 NPC patients diagnosed from 1993 to 2000. A representative section from each biopsy was stained with hematoxylin and eosin stain, and reviewed to determine regions with malignant epithelium for macrodissection. Total RNA was isolated from the macrodissected samples, using the RecoverAll Total Nucleic Acid Isolation Kit for FFPE (Ambion Inc.) according to the manufacturer’s instructions for enrichment for small RNA species.

Thirty-four FFPE lumpectomy blocks and six reduction mammoplasty specimens were macrodissected and RNA was isolated as described previously described (96).

5.3.2 Data description

This paper studied three types of cancer patients: breast cancer, cervical cancer and NPC. Detailed data descriptions are listed in Tables 5.1, 5.2, and 5.3. Herein, we refer to the miRNAs measured by the platforms as “genes”.

**Breast cancer** FFPE samples were collected and gene expression was measured using three platforms: NanoString nCounter™ Human miRNA Expression Assay (NanoString) (v1.6.0), Applied Biosystems TaqMan Low Density Array Human MicroRNA Array (TLDA) (v1.0), and single-well quantitative real-time PCR (PCR) using Applied Biosystems TaqMan MicroRNA Assays.

**Cervical cancer** data included FFPE and Frozen samples. Gene expression was measured by two platforms: NanoString nCounter™ Human miRNA Expression Assay (NanoString) (v1.7.0) and TLDA Human MicroRNA A Array (v2.0).

**NPC** data included FFPE samples and gene expression was measured by two platforms: NanoString nCounter™ Human miRNA Expression Assay (NanoString) (v1.7.0) and TLDA Human MicroRNA Array (v1.0).
### Table 5.1 Data descriptions for breast cancer datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>FFPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NanoString</td>
<td>742 genes * 16 samples (4 normals, 12 tumours)</td>
</tr>
<tr>
<td>TLDA</td>
<td>365 genes * 40 samples (6 normals, 34 tumours)</td>
</tr>
<tr>
<td>PCR</td>
<td>20 genes * 40 samples (7 normals, 33 tumours)</td>
</tr>
</tbody>
</table>

### Table 5.2 Data descriptions for cervical cancer datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>FFPE</th>
<th>Frozen</th>
</tr>
</thead>
<tbody>
<tr>
<td>NanoString</td>
<td>753 genes * 96 samples (9 normals, 87 tumours)</td>
<td>753 genes * 12 samples (0 normals, 12 tumours)</td>
</tr>
<tr>
<td>TLDA</td>
<td>378 genes * 96 samples (9 normals, 87 tumours)</td>
<td>378 genes * 90 samples (11 normals, 79 tumours)</td>
</tr>
</tbody>
</table>

### Table 5.3 Data descriptions for NPC datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>FFPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NanoString</td>
<td>753 genes * 134 samples (10 normals, 124 tumours)</td>
</tr>
<tr>
<td>TLDA</td>
<td>365 genes * 12 samples (0 normals, 12 tumours)</td>
</tr>
</tbody>
</table>
5.3.3 Data pre-processing

Prior to any further analyses, data pre-processing was performed to remove the technical assay bias (*i.e.*, data normalization) and extract the common genes and samples between platforms (*i.e.*, gene and sample mapping).

We used standard \( \Delta C_T \) methods to normalize gene abundances measured from TLDA and PCR. We subtracted the gene abundance \( C_T \) from the mean of the control gene abundance \( C_C \) and then calculated the normalized gene abundance as: \(-log_2\left(2^{-\left(C_T - C_C\right)}\right)\). The reason for using the log2-space was to be consistent with the amplification processes in PCR and TLDA.

To normalize gene abundances from NanoString, we applied the R package 'NanoStringNorm' (v1.1.11) (255). The gene abundances were adjusted based on the sum of all the samples, which was able to normalize the majority of the technical variations (*i.e.*, NS1). We further normalized gene abundances using the mean of all samples (*i.e.*, NS2). To evaluate the platform performance, we extracted the common samples and genes from every pair of platforms, since not all the genes and samples were measured in all platforms.

5.3.4 Unsupervised machine-learning

An unsupervised machine-learning hierarchical clustering analysis was applied to the normalized mapped gene abundances, to understand how gene abundances were distributed across the platforms. The divisive hierarchical clustering algorithm DIANA was used to cluster all the genes, as implemented in the cluster package (v1.14.3) for the R statistical environment (v2.15.1), and heatmaps were visualized using the lattice (v0.20.10) and latticeExtra (v0.6-24) packages. Pearson's correlation was used as the distance metric.

5.3.5 Statistical analysis

Correlation analyses were first performed to measure the impact of platform on gene expression. Pair-wise comparisons between platforms were calculated using Spearman's rank correlation (rho, \( \rho \)). We measured three different correlations to understand different aspects of the platform effects:

1. Overall correlations: The correlations of gene expression per sample and gene between two platforms were calculated. This gives an overall evaluation of two
platforms regardless of the number of genes and samples.

2. Sample-wise correlations: The correlations of gene expression per sample between two platforms were examined. The correlations depend on the number of genes per sample involved.

3. Gene-wise correlations: Correlations of gene expression per gene between two platforms were examined. The correlations depend on the number of samples per gene involved.

Next, to study the protocol effects on differentially-expressed genes, we calculated the FDR-adjusted p-value and fold change between tumour and normal for each pair of platforms. Fold change was calculated as: \( \frac{\text{mean(tumour)}}{\text{mean(normal)}} \) in \( \log_2 \)-space. A Spearman’s correlation was also calculated based on the fold change and p-value, to measure the correlations between differentially-expressed genes.

5.3.6 Visualization

For visualization, we employed the lattice and latticeExtra packages (v0.19-33 and v0.6-33 respectively) for the R statistical environment (v2.13.2). Venn diagrams were created using the VennDiagram package (v1.1.1) (297). All clustering analyses employed agglomerative hierarchical clustering using complete linkage, and Pearson’s correlation as the distance metric, as implemented in the cluster package (v1.14.1). Heatmaps were created using the lattice (v0.19-33) and latticeExtra (v0.6-19) packages.

5.4 Results

5.4.1 Experimental design

The experimental design for this study is outlined in Figure 5.1. Briefly, tissue samples were collected from patients with: breast cancer (FFPE), cervical cancer (FFPE and frozen), or NPC (FFPE). The gene abundances were measured using: NanoString nCounter™ Human miRNA Expression Assay (NS) (v1.6.0 for breast cancer; v1.7.0 for cervical cancer and NPC), Applied Biosystems TaqMan Low Density Array Human MicroRNA Array (TLDA) (TLDA microRNA array v1.0 for breast cancer and NPC; TLDA microRNA A Array v2.0 for cervical
cancer), and qRT-PCR using Applied Biosystems TaqMan MicroRNA Assays (breast cancer only). The data were pre-processed to find the common samples and genes across the three platforms. The filtered genes and samples were subjected to a series of statistical analyses to reveal the platform relationships. Survival analysis was performed on the cervical cancer data, to understand the impact of platform on the biologically-significant genes. Lastly, these biologically-significant genes were subjected to sequence and pattern analysis, to determine the effect of sequence characteristics.

5.4.2 Biological differences are detectable by platforms

It is important to assess data quality before any analysis. Thus, after normalization, miRNA abundances for every pair of samples across protocols were subjected to unsupervised machine-learning. As expected, all platforms were able to distinguish biological differences. For example, almost all tumour and normal samples were tightly clustered to each other, regardless of the protocol (Figure 5.2A and B), suggesting that the platforms were sensitive to biological differences between cancer and normal samples. We noticed that about 5% to 10% of normalized miRNA abundances from NanoString in breast cancer samples were exceptionally high; for better visualization, these miRNAs were removed from these heatmaps.

5.4.3 Data from NanoString and TLDA are significantly different

To identify the strongest trends in the miRNA abundances between TLDA and NanoString, we first employed an unbiased machine-learning approach (313). The correlations of the miRNA abundance profiles of breast and cervical samples were calculated between each pair of arrays, generating a correlation-matrix. The strongest trend within these two matrices was the separation of samples processed with the TLDA platform (yellow annotation bars) vs. those processed with the NanoString platform (blue annotation bars) (Figure 5.2A for breast samples; Figure 5.2B for cervical FFPE samples). The effects of platform were stronger than the biological differences between cancer and normal samples (red/white annotation bars). The strong technical effects were also observed in cervical frozen samples (Figure 5.2C) and NPC samples (Figure 5.2D). In these two latter cases, cancer samples clustered perfectly by protocol with the absence of normal samples. To test the significance of this observation, we calculated the Adjusted Rand Index (ARI), which ranges from -1 to +1, with higher values indicating non-random associations. The higher the ARI, the more similar two clustering patterns are. An ARI
of 1.0 indicates perfect partitioning and a non-positive ARI (ARI $\leq 0$) indicates independent or random partitioning. The ARI for platform bias was 1.0, indicating that the TLDA and NanoString datasets contained significantly different expression profiles.

To quantify these similarities or dissimilarities, we calculated pair-wise Spearman's correlations on the miRNA abundances between protocols. The correlations between NanoString and TLDA were: 0.32 for breast FFPE samples (Fig 5.3A), 0.32 for cervical FFPE samples (Fig 5.3B), -0.02 for NPC FFPE samples (Fig 5.3C), and 0.42 for cervical frozen samples (Fig 5.3F), again suggesting the weak correlation between miRNA abundances from these two platforms. The exceptionally low correlation ($\rho = -0.02$) for the NPC samples (Fig 5.3C) could be due to the smaller number of samples and miRNAs that were common between the two platforms, with only 194 miRNAs and 12 samples (= 2328 sample pairs), which is 56% the sample pairs compared to the cervical frozen samples (345 miRNAs and 12 samples in common), and 7% of sample pairs compared to the cervical FFPE samples (345 miRNAs and 96 samples in common).

5.4.4 PCR is more similar to TLDA than NanoString

We next compared the miRNA abundances between PCR and NanoString (Figure 5.4A and B), and between PCR and TLDA (Figure 5.4C and D) for the breast samples. Platform effects between PCR (purple annotation bars) and NanoString (blue annotation bars) were not as strong as the biological differences between cancer and normal samples (red/white annotation bars) (Figure 5.4B). Furthermore, the differences between PCR (purple annotation bars) and TLDA (yellow annotation bars) were weaker compared to the biological differences (red/white annotation bars) (Figure 5.4D). These results suggest that the datasets produced from either NanoString or TLDA, were more similar to the dataset produced from PCR than they were to each other.

In contrast to the low correlations observed between NanoString and TLDA for all sample types, the correlations on the miRNA abundances between PCR and NanoString ($\rho = 0.6$; Figure 5.3D), and between PCR and TLDA ($\rho = 0.7$; Figure 5.3E) in the breast FFPE samples were significantly higher. Sample-wise correlations were high for both PCR vs. NanoString ($\rho = 0.75$, $p = 4.32 \times 10^{-44}$; Figure 5.5A), and PCR vs. TLDA ($\rho = 0.70$, $p = 6.69 \times 10^{-58}$; Figure 5.5B). Despite this, gene-wise correlations were not comparable, with almost 50% of genes being
negatively correlated between PCR and NanoString ($\rho < 0$; Figure 5.5C), whereas over 90% of genes were positively correlated between PCR and TLDA ($\rho > 0$; Figure 5.5D).

5.4.5 Fold-changes are poorly correlated between NanoString vs. TLDA, but differentially-expressed miRNAs are highly correlated across platforms

We investigated the relationship of tumour/normal fold-changes between NanoString and TLDA data. This analysis was performed using the breast FFPE and cervical FFPE datasets; the cervical frozen and NPC FFPE datasets were excluded due to their small numbers of samples. Fold-changes were positively- but lowly-correlated between NanoString and TLDA in the breast FFPE samples ($\rho = 0.33$; Fig 5.6A) and cervical FFPE samples ($\rho = 0.37$; Fig 5.6B). However, miRNAs that were significantly differentially-expressed between tumour and normal (fold-change $\geq 2.0$ in both platforms; upper right section of plots in Fig 5.6C and D), were generally significantly highly-correlated between the two platforms ($\rho > 0.5$ and $P_{\text{Rho}} \leq 0.05$). Altogether, these results show that although tumour/normal fold-changes were lowly-correlated between the NanoString and TLDA datasets, significantly differentially-expressed miRNAs were significantly highly-correlated between the two platforms.
Figure 5.1  Experimental design

FFPE tissue samples were collected from three types of cancer: breast (1), cervical (2) and NPC (4). Frozen tissue samples were collected from cervical cancer (3). Total RNA was extracted from each sample and gene abundances were measured using three different platforms: a) NanoString nCounter™ Human miRNA Expression Assay (NanoString, or NS), b) Applied Biosystems TaqMan Low Density Array Human MicroRNA Array (TLDA), and c) qRT-PCR using Applied Biosystems TaqMan MicroRNA Assays (PCR). These data were normalized and common samples and genes were extracted across the three platforms. The common genes and samples were subjected to unsupervised machine-learning to visualize the distribution of the gene abundances across the platforms. A series of correlation analyses were then conducted to investigate the technical and biological effects. To determine how the technical biases affected the biologically-significant genes, statistical analyses were performed and Spearman's correlations were calculated based on the tumour/normal fold changes. Lastly, survival analysis was conducted to understand the effect of platform on the prognostic genes. Datasets involved in the analysis are indicated with 1-4.
Figure 5.2  Correlations between technical replicates (TLDA vs. NanoString)

Following pre-processing, the relationships between platform effects (TLDA vs. Nanostring) and biological effects (tumour vs. normal) were examined for the: A) breast FFPE dataset, B) cervical FFPE dataset, C) cervical frozen dataset, and D) NPC FFPE dataset. On each sample vs. sample heatmap, the white-to-blue colour bar reflects the range of Pearson’s correlations. The two-column annotation bars indicate platform (TLDA in yellow, NanoString in blue) and biological effects (tumour in red, normal in white). The strong separation by platform highlights the magnitude of technical artifacts.
Figure 5.3  Distributions and correlations of miRNA abundances across platforms

Following pre-processing, we compared the distributions of miRNA abundances and calculated the correlations between the NanoString and TLDA datasets for A) breast FFPE samples, B) cervical FFPE samples, C) NPC FFPE samples, and D) cervical frozen samples. We also examined the distributions and correlations of miRNA abundances in the PCR dataset in comparison with the D) NanoString and E) TLDA datasets for breast FFPE samples. P-values were calculated using FDR-adjusted model-based t-tests.
Following pre-processing, we studied the relationships between NanoString-PCR (A and B), and TLDA-PCR effects (C and D), and biological effects (tumour vs. normal) for the breast FFPE dataset. On each sample vs. sample heatmap, the white-to-blue colour bar reflects the range of Pearson's correlations. The two-column annotation bars indicate platform (TLDA in yellow, NanoString in blue, and PCR in purple), and biological effects (tumour in red, normal in white). The strong separation by platform highlights the magnitude of technical artifacts.
Figure 5.5  Sample-wise and gene-wise correlations between PCR vs. NanoString and PCR vs. TLDA

Spearman’s rank correlations were calculated for miRNA abundances in the breast cancer dataset, for every pair of replicate samples between A) PCR vs. NanoString, and B) PCR vs. TLDA, and every pair of genes between C) PCR vs. NanoString and D) PCR vs. TLDA. The dashed line indicates the overall correlation between platforms.
Figure 5.6 Correlations of tumour/normal fold-changes between NanoString and TLDA

Tumour/normal fold-changes (shown in log2-scale) were plotted for the NanoString (NS) and TLDA datasets, for all miRNAs that were common to both platforms, for A) breast FFPE, and B) cervical FFPE samples. P-values were calculated using FDR-adjusted model-based t-tests. Only significantly differentially-expressed miRNAs, those with statistically significant fold-changes ($P_{\text{both}} \leq 0.05$, $P_{\text{TLDA}} \leq 0.05$, or $P_{\text{NS}} \leq 0.05$), were re-plotted for C) breast FFPE, and D) cervical FFPE samples, to show the correlations between NS and TLDA datasets. The yellow-to-purple colour bar at the bottom represents the range of correlations (rho).
5.5 Discussion

The use of global miRNA profiling platforms for biomarker development continues to increase in popularity, largely due to the fact that miRNAs are extremely stable in many types of clinical specimens, and have been implicated in nearly all human diseases. With various platforms available for the high-throughput measurement of miRNA abundance, it is important to consider the platform-specific effects associated with the detection of differentially-expressed miRNAs. The performance of technologies used for miRNA detection are not as well characterized as those used for DNA, mRNA or protein, with respect to their use in clinical research. Many methods for mRNA data normalization and analysis have been adapted for miRNA datasets, however, the accuracy of any methods that require a normal distribution of data may not be as robust, because most global miRNA platforms can only detect several hundred different miRNAs, compared to global mRNA platforms that can detect tens of thousands of different gene transcripts (314).

Previous studies comparing global miRNA profiling technologies have mainly focused on various microarray and next-generation sequencing technologies (315-318), however, there are no published reports to date that directly compare NanoString with TLDA for miRNA expression profiling. Our study fills this gap by evaluating the performance of the NanoString nCounter miRNA Expression Assay and the TaqMan Low Density Array Human MicroRNA Array across three types cancers (breast, cervical and NPC) and two types of specimens (FFPE and frozen tissues). We also compared both of these platforms with PCR, which is widely regarded as the gold standard and the method of choice for validation. The four key findings of our study were: i) miRNA abundances were poorly correlated between NanoString and TLDA, ii) PCR was more similar to TLDA than NanoString, iii) Tumour/normal fold-changes were poorly correlated between NanoString and TLDA, and iv) Significantly differentially-expressed miRNAs were highly-correlated between Nanostring and TLDA.

Consistent with other studies that have reported low cross-platform concordance between various global miRNA profiling platforms (315, 316, 318), we demonstrated poor correlations between the NanoString and TLDA datasets, with respect to miRNA abundances ($\rho = 0.32$ for breast FFPE samples; $\rho = 0.32$ for cervical FFPE samples; $\rho = -0.02$ for NPC FFPE samples, and $\rho = 0.42$ for cervical frozen samples) and tumour/normal fold changes ($\rho = 0.33$ for breast...
FFPE samples; \( \rho = 0.37 \) for cervix FFPE samples). The poor correlation between NanoString and TLDA could potentially be explained by the fundamental differences between the chemistries of these two methods. The TLDA method relies on target-specific reverse transcription and the hybridization of stem-loop primers, while NanoString provides a direct count of miRNAs using optical quantification of fluorescently-tagged miRNA molecules. The possibility of bias from amplification steps or enzymatic reactions is present for data generated using TLDA, but not NanoString.

Despite low cross-platform correlations between NanoString and TLDA, we demonstrated that PCR was highly comparable to both NanoString (\( \rho = 0.6 \)) and TLDA (\( \rho = 0.7 \)) in terms of miRNA abundances, which was consistent with other published reports that found high concordance between PCR and various global miRNA profiling platforms (315, 319). However, gene-wise correlations of replicate samples revealed that almost 50% of genes were negatively correlated between PCR and NanoString, whereas over 90% of genes were positively correlated between PCR and TLDA. The similarity between PCR and TLDA are not surprising, since TLDA is a PCR-based method and utilizes the same stem-loop reverse transcription primers and TaqMan chemistry.

Although tumour/normal fold-changes were poorly-correlated between the NanoString and TLDA datasets (\( \rho = 0.33 \) for breast FFPE samples; \( \rho = 0.37 \) for cervix FFPE samples), miRNAs that were significantly differentially-expressed between tumour and normal were highly correlated between the two platforms. This indicates that biologically-significant miRNAs are generally concordant between the two platforms, and are likely to be identified with either NanoString or TLDA.

In conclusion, our study demonstrated that despite poor cross-platform correlation between NanoString and TLDA for global miRNA quantification, both platforms performed well in comparison with PCR, although TLDA was more similar to PCR. To our knowledge, this is the first reported study describing a systematic comparison of the NanoString and TLDA platforms for global miRNA expression profiling. When selecting a platform for global miRNA profiling, many consideration must be taken into account, including: cost, available facilities, amount of RNA required, and familiarity with platform-specific data normalization and analysis.
methods. For future studies, it would be useful to develop methods for integrating data across platforms.
CHAPTER 6: DISCUSSION
6.1 Research Summary

The focus of this thesis was to examine the role of miRNAs in cervical cancer, and develop new insights into their contributions to tumourigenesis and clinical outcome. During our investigations, several methods were utilized for measuring miRNA expression, and our focus was expanded to include an analysis of various technologies for quantifying miRNA abundance.

MicroRNA-196b was demonstrated to be significantly down regulated in cervical cancer cell lines and patient samples, and low miR-196b expression in patients was shown to be associated with poorer DFS. Furthermore, we observed that miR-196b directly targeted the HOXB7 transcription factor, which in turn, regulated the expression of VEGF. Our miRNA profiling studies in patients also led to the development of a candidate 9-miRNA signature that was prognostic for DFS in cervical cancer patients, independent of tumour size and nodal status. However, although three different methods were utilized to measure these nine miRNAs in an independent cohort of patients, we were unable to validate this signature. Furthermore, our attempts to independently corroborate the only published miRNA signature to date for cervical cancer (120) were similarly unsuccessful. These results highlight the considerations that must be taken into account with regards to the evaluation of miRNAs as prognostic biomarkers.

Our analyses of the TLDA and Nanostring methods for global miRNA expression profiling demonstrated an overall lack of concordance between these two platforms, although they individually correlated well with PCR, the gold-standard for nucleic acid quantification. On a positive note, biologically-relevant miRNAs were found to be highly correlated between the two platforms. In addition, our investigations identified a protocol bias in microarray studies that was created by the choice of the amplification protocol; the LTR-method was able to help remove, but was unable to completely eliminate this bias.

6.2 Future Directions

Our characterization of the role of miR-196b in the HOXB7~VEGF axis was very promising, given the recent announcement of the preliminary data derived from this Phase III clinical trial for Bevacizumab (220). This drug is currently approved for use in metastatic colorectal cancer, non-small cell lung cancer, glioblastoma, and metastatic kidney cancer, and has been shown to impact tumour progression by blocking angiogenesis (219). If Bevacizumab
were to be added to the treatment regimen for cervical cancer patients, further investigations would be warranted to determine if miR-196b levels could possibly be used as a biomarker to predict response to Bevacizumab.

Additional attempts to validate our 9-miRNA signature should be made. One option is to utilize an alternate RNA extraction method for the FFPE samples that make up the validation cohort. After our experiments were completed, a study that compared ten RNA extraction methods for FFPE samples demonstrated that the Ambion RecoverAll Total Nucleic Acid Isolation Kit recovered the most amplifiable RNA (113), which would be well-suited for PCR-based RNA expression profiling methods including the TLDA. Another option is to use a validation cohort consisting of frozen samples, given the lack of consistency in the tissue preservation methods (frozen vs. FFPE) between our training and validation cohorts. One strategy is to collect frozen cervical cancer specimens to create a new validation cohort, and measure the expression of the nine miRNAs in our signature using TLDA or individual real-time PCR. Another strategy is to utilize the frozen cervical cancer samples from the TCGA data portal as another independent validation set, after the number of frozen cervical cancer samples with miRNASeq data and clinical information is expanded; as of June 2013, only 48 samples had survival information.

Alternately, we could try to generate a new prognostic miRNA signature for cervical cancer, using a more robust method for global miRNA expression profiling. Reports have demonstrated the increased sensitivity and specificity of deep sequencing for measuring miRNA expression (115, 116). Since we have shown that the majority of significantly differentially-expressed miRNAs in cervical cancer are down-regulated, deep sequencing should be able to more accurately detect more lowly-expressed miRNAs in our samples that were beyond the limit of detection for TLDA.

Another potential avenue would be to perform a global miRNA expression profiling experiment using multiple biopsies from each individual patient’s tumour, similar to the Bachtiary et al. study (265), to explore intra-tumour heterogeneity of miRNA expression in cervical cancer. It would also be interesting to develop a novel signature using these data, since this should provide a more representative miRNA profile of the tumour, in its entirety.
In addition, we will perform co-analyses of global miRNA and mRNA expression data that we have generated for our 79 cervical cancer and 11 normal cervix frozen samples, similar to studies that have been reported for glioma (320), ovarian cancer (321), primary human lymphoblastoid cell lines (322), and the NCI-60 panel of human cancer cell lines (323). Global miRNA and mRNA levels have been measured using the TLDA Human MicroRNA A Array (v2.0) and the Affymetrix GeneChip Human Genome U133 Plus 2.0, respectively. We will examine the correlation of genome-wide expression profiles between miRNAs and their predicted target mRNAs, and determine if predicted miRNA-mRNA pairings are negatively correlated in expression.

6.3 Conclusions

This thesis has provided insights into the role of microRNAs in cervical cancer, and provided a thorough analysis of the issues and considerations pertaining to prognostic miRNA signature development and validation. miR-196b was identified and characterized to be important in cervical cancer development and progression. These findings are significant contributions to the body of knowledge that will help us to realize the promise of personalized cancer medicine. In addition, various platforms for miRNA expression profiling were investigated and compared to each other, and amplification methods for Affymetrix microarrays were evaluated. In both cases, the low concordance in data produced by different protocols was assessed and methods to mitigate biases were described.
References


17. Cai X, Hagedorn CH, Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 2004;10: 1957-66.


38. Lytle JR, Yario TA, Steitz JA. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. Proc Natl Acad Sci U S A 2007;104: 9667-72.


44. Orban TI, Izaurralde E. Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome. RNA 2005;11: 459-69.


204. Achen MG, Jeltsch M, Kukk E, et al. Vascular endothelial growth factor D (VEGF-D) is a ligand for the tyrosine kinases VEGF receptor 2 (Flk1) and VEGF receptor 3 (Flt4). Proc Natl Acad Sci U S A 1998;95: 548-53.


