Differential microRNA Expression in Barrett's Esophagus correlates with regulation of Posterior Homeotic Genes

Reilly June Clark
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the Biomedical Engineering and Bioengineering Commons

Repository Citation
https://corescholar.libraries.wright.edu/etd_all/2181

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
DIFFERENTIAL MICRORNA EXPRESSION IN BARRETT’S ESOPHAGUS CORRELATES WITH REGULATION OF POSTERIOR HOMEOTIC GENES

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

By

REILLY JUNE CLARK
B.S., Boise State University, 2014

2019
Wright State University

Scott Everet Baird, Ph.D.
Dissertation Director

Mill W. Miller, Ph.D.
Director, Biomedical Sciences
Ph.D. Program

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

Committee on Final Examination

Scott Everet Baird, Ph.D.

Paula A. Bubulya, Ph.D.

Michael L. Raymer, Ph.D.

Sangeeta Agrawal, M.D.

Nancy Kelley-Loughnane, Ph.D.
ABSTRACT

Clark, Reilly June, Ph. D., Biomedical Sciences, Wright State University. 2019. Differential microRNA expression in Barrett’s Esophagus Correlates with Regulation of Posterior HOX Genes.

Barrett’s Esophagus (BE) is characterized by the appearance of an intestinal-like epithelium in the distal esophagus. The molecular mechanisms behind BE development are unknown. BE is often preceded by Gastroesophageal Reflux Disease (GERD) and predisposes patients to esophageal adenocarcinoma (EAC). Due to the high mortality rate associated with EAC, BE patients are continuously monitored through upper endoscopy with biopsy for progression to low grade dysplasia (LGD), high grade dysplasia, and EAC. This monitoring technique poses numerous risks, so alternative surveillance and diagnostic techniques for BE pathogenesis are continually studied. microRNA biomarkers in BE pathogenesis may provide alternative means of diagnosis as well as a greater understanding of BE and its progression to EAC. Here, small RNA-sequencing of serum and tissue from GERD, BE, LGD, and EAC patients revealed three candidate tissue microRNAs differentially expressed in BE compared to GERD patients. Differential expression of the three candidate microRNAs was validated in a second cohort of BE and GERD patient tissues by quantitative PCR. Gene target analysis revealed two candidate microRNAs are homeobox (HOX) microRNAs, which directly target central and posterior HOX genes. HOX genes are transcription factors which regulate gene expression along the anterior/posterior axis. BE resembles a homeotic transformation, which could be due to aberrant expression of posterior HOX genes in the esophagus, an anterior organ. The third candidate microRNA targets a component of Polycomb Repressive Complex 1, a transcriptional repressor of HOX genes. Thus, the three candidate microRNAs may modulate posterior HOX gene expression associated with BE development.
# TABLE OF CONTENTS

Abstract iii  
Table of Contents iv  
List of Figures viii  
List of Tables ix  
Acknowledgements x  

I. Introduction 1  
   Barrett’s Esophagus (BE) 1  
   Models for BE Development 3  
      Transdifferentiation 3  
      Residual Embryonic Cell Populations 5  
      Stem Cell Migrations 6  
   Homeotic Genes in BE Development 8  
      BE Lesions Resemble Homeotic Transformations 8  
      HOX Genes Regulate Positional Expression of Intestine-type Genes 10  
      ParaHox Genes Regulate Posterior Gut Development as well as HOX Gene Expression 11  
Protein Markers of BE 12  
Surveillance of BE Development and Progression to EAC 13  
   Upper GI Endoscopy with Biopsy 13  
   Alternative Surveillance Techniques 15  
   Alternative Diagnostic Techniques 16  
Utilizing microRNA Signatures to Understand BE Pathogenesis 17
II. Materials and Methods

Small RNA-Sequencing of microRNAs in Patient Tissue and Serum Biopsies

Collection of Fresh-Frozen Tissue and Serum from Human Patients

Small RNA Isolation from Patient Tissue Samples

Small RNA Isolation from Patient Serum Samples

cDNA Library Preparation of Tissue small RNAs

cDNA Library Preparation of Serum small RNAs

Small RNA-sequencing of microRNAs in Patient Tissue and Serum Biopsies

Preparing small RNA sequences for Analysis of Differential microRNA Expression

Pre-Alignment Processing Removes Low Quality Bases from Raw Sequences

Aligning the Raw Sequence to miRBase v.21 eliminates non-microRNA Sequences from Analysis

Quantification of aligned reads to miRBase v.21 results in microRNA read counts

Normalization of microRNA Read Counts by Trimmed Mean of M-values

Filtering based on TMM-normalized microRNA read count applied to Serum and Tissue Datasets

Identification of Potential microRNA Signatures from Normalized microRNA Read Counts

Pair-wise Comparisons Between Pathology Groups to Identify Candidate microRNAs

Gene Target Analysis for Selected Candidate microRNA Biomarkers

Quantitative Real-Time PCR Validation of Candidate microRNAs in Human Patient Tissues

III. Results

Small RNA-sequencing of microRNAs from Patient Serum and Fresh-Frozen Tissue

Construction of cDNA libraries for small RNA-sequencing

Selection of cDNA Libraries for small RNA-sequencing
IV. Discussion

Limitations in Identifying Biomarkers for BE Pathogenesis 96

Serum 96

Tissue 100

HOX microRNAs may Regulate Posterior HOX Genes as a part of a Negative Feedback Mechanism 102

Mir-215-3p and miR-196-5p are involved in Indirect Regulation of HOX Gene Expression 108

miR-215-3p targets BMI1, a Component of PRC1 108

Multiple Lineage Leukemia (MLL) Complex Promotes miR-196-5p Expression 109

A Model for Homeotic Gene Regulation in BE Development 110

V. Literature Cited 114

VI. Appendices 133

Appendix I: small RNA sequencing of Serum and Tissue cDNA Libraries 133

Appendix II: cDNA Libraries from Serum and Tissue small RNAs 134

Appendix III: Variance of TMM-normalized microRNA reads of Selected microRNAs 138
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessment of small RNA and microRNA Yields following small RNA Extraction</td>
</tr>
<tr>
<td>2</td>
<td>Construction of cDNA libraries from Serum and Tissue small RNAs</td>
</tr>
<tr>
<td>3</td>
<td>cDNA Library Selection is Based on the size of the 94-114 bp peak</td>
</tr>
<tr>
<td>4</td>
<td>Distribution of Raw microRNA Read Counts in Serum and Tissue Samples</td>
</tr>
<tr>
<td>5</td>
<td>Hierarchical Clustering of Serum and Tissue Samples</td>
</tr>
<tr>
<td>6</td>
<td>Heat Maps of Serum and Tissue Samples</td>
</tr>
<tr>
<td>7</td>
<td>Serum microRNAs cannot distinguish among the Four Stages of BE Pathogenesis</td>
</tr>
<tr>
<td>8</td>
<td>Tissue microRNAs cannot distinguish among the Four Stages of BE Pathogenesis</td>
</tr>
<tr>
<td>9</td>
<td>Decision Trees Can Separate Tissues into Four Pathology Groups</td>
</tr>
<tr>
<td>10</td>
<td>randomForest classification on Serum and Tissue Datasets</td>
</tr>
<tr>
<td>11</td>
<td>randomForest regression on Serum and Tissue Datasets</td>
</tr>
<tr>
<td>12</td>
<td>Three microRNAs show significant increased expression in BE Tissues vs GERD Tissues</td>
</tr>
<tr>
<td>13</td>
<td>miR-423-5p is not significantly differentially expressed between GERD and BE tissues</td>
</tr>
<tr>
<td>14</td>
<td>miR-423-5p does not display constant expression between GERD and BE FFPE tissues</td>
</tr>
<tr>
<td>15</td>
<td>miR-196a-5p, miR-196b-5p, and miR-215-3p all display significantly increased expression in BE tissues in comparison to GERD tissues</td>
</tr>
<tr>
<td>16</td>
<td>Collinear Expression of HOX genes on the Anterior/Posterior Axis</td>
</tr>
<tr>
<td>17</td>
<td>A Model for Homeotic Gene Regulation in BE Development</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human ANTP-class Homeotic Genes Differentially Regulated or Mutated in BE Pathogenesis</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Gene Target Analysis Utilizes Four Distinct Methods</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Small RNA Extracted from Fresh-Frozen Tissue Biopsis</td>
<td>39</td>
</tr>
<tr>
<td>4</td>
<td>Composition of Serum and Tissue Sequencing Data Sets</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>Selected Serum and Tissue Samples Processed and Aligned in Partek® Flow®</td>
<td>49</td>
</tr>
<tr>
<td>6</td>
<td>Seven microRNAs display Differential Expression in Stages of BE Pathogenesis</td>
<td>75</td>
</tr>
<tr>
<td>7</td>
<td>Power Analysis for qPCR validation</td>
<td>81</td>
</tr>
<tr>
<td>8</td>
<td>Total RNA Extracted from Formalin Fixed Paraffin Embedded Tissues</td>
<td>82</td>
</tr>
<tr>
<td>9</td>
<td>Change in Expression of selected microRNAs in BE vs GERD FFPE tissues</td>
<td>91</td>
</tr>
<tr>
<td>10</td>
<td>Selected Gene Targets for the Three Candidate Tissue microRNA Biomarkers</td>
<td>93</td>
</tr>
<tr>
<td>11</td>
<td>Serum and Tissue microRNAs Differentially Expressed in BE Literature</td>
<td>101</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

This work would not have been possible without joint funding from Dayton VA Hospital (to Dr. Sangeeta Agrawal) and Wright State University Boonshoft School of Medicine (to Dr. Madhavi Kadakia). I was financially supported by the Science, Mathematics, and Research for Transformation (SMART) Scholarship-for-Service program for the entirety of my PhD. I am especially grateful to Dr. Scott E. Baird, Department of Biological Sciences, who worked actively to provide me with the mentorship I needed to complete my dissertation.

I wish to acknowledge my friends and family whose love and support was and continues to be essential to my professional success and personal well-being. My brother, Tyler Clark, cheered me on with his unconditional love, approval, and acceptance. Nathaniel Wentland was a true friend who kept up a continuous stream of motivating messages over the years. I am particularly thankful for the love and guidance provided by my exceptional MOM, Teri Clark, who was always there to listen and advise.
INTRODUCTION

Barrett’s Esophagus (BE)

The esophagus serves as an intermediary organ in the human gastrointestinal system between the oral cavity and the stomach. Unlike either of these organs, the esophagus does not play a role in digestion or absorption of nutrients. In a normal human adult, the entire esophagus is lined with a stratified squamous epithelium [1]. Several pathological conditions of the esophagus exist including esophagitis, esophageal squamous cell carcinoma, gastroesophageal reflux disease (GERD), Barrett’s Esophagus (BE), and Esophageal Adenocarcinoma (EAC) [2, 3]. The three latter conditions are thought to be linked.

GERD describes a chronic disorder in which gastric juice is refluxed past the lower esophageal sphincter into the distal esophagus. GERD patients experience continual pain and inflammation due to the chronic exposure of the esophageal epithelium to the acidic refluxate [4, 5]. The incidence of GERD has been on the rise in the United States [6]. Obesity can be a precursor condition to GERD. Rising rates of adult obesity in the United States are thought to be a contributing factor to GERD’s increased incidence [7]. However, the causes of GERD are varied and cannot always be linked to higher BMI [4]. Hiatal hernias and structural abnormalities in the esophageal sphincter are also associated with this chronic disorder [8].
In BE, regions of normal squamous epithelium are replaced by a simple columnar epithelium interspersed with intestinal-type cells, termed intestinal metaplasia [9]. BE is most associated with the presence of goblet cells in the distal esophagus, a typical cell type in the small intestinal epithelium [10, 11]. A BE diagnosis is frequently preceded by GERD [6, 12-14]. BE lesions do not present any additional symptoms, and there are no treatments to reverse or prevent BE. The only recourse following a BE diagnosis are antacids and proton pump inhibitors to increase the pH of the refluxate, mediating the pain and inflammation characteristic of GERD [15]. Several studies have shown the acidic environment created by gastric reflux can alter epithelial cell signaling [16, 17]. Introduction of bile acids from the small intestine through duodenal esophageal reflux can promote intestinal-type gene expression in the esophageal epithelium, most notably caudal-related homeobox 2 (CDX2) and acid mucin 2 (MUC2) [18]. It is still unclear whether the intestinal-type signaling promoted by acid reflux is the initiating event of BE development or a later event following replacement of the normal squamous epithelia.

A BE diagnosis does significantly increase patient risk for Esophageal Adenocarcinoma (EAC). A risk evaluation by Cameron et al (1985) has shown that a BE patient’s risk for EAC can reach 100 fold that of a patient without BE [19]. EAC diagnoses comprise less than 1% of all cancer diagnoses every year in the United States [11, 20][11, 20][11, 20][11, 20]. Yet, EAC mortality rate has been reported as high as 80% [21]. EAC tumors are difficult to diagnose in their early stages, as the epithelium of the distal esophagus is impossible to self-monitor. Unless EAC is caught early enough to enable ablation or surgical resection of the tumor, the five year survival rate is less than 5% [21]. Surgical resection is the most successful treatment available for this cancer. Smoking and alcohol
consumption are minor risk factors for EAC, but the only known major risk factor for EAC is BE [22]. EAC tumors have been shown to originate from BE lesions [23]. BE pathogenesis follows a sequence: GERD, BE, BE with low grade dysplasia, BE with high grade dysplasia, and finally EAC. Both the American Gastroenterological Association (AGA) and the American College of Gastroenterology (ACG) recommend that GERD and BE patients are monitored every 3-5 years for signs of BE lesions, dysplasia, and EAC [10, 20]. Upon detection of dysplastic or cancerous tissue, the affected tissues are either removed or subjected to more frequent monitoring. BE lesions without dysplastic characteristics can remain benign for decades. It is estimated that only 0.2-2.1% of BE patients progress to dysplasia or EAC every year [24]. Since it is not yet possible to predict which BE patients will develop EAC, a BE diagnosis in a GERD patient does not affect the AGA/ACG-recommended monitoring regimen.

**Models for BE Development**

Hypotheses abound concerning the origin of BE lesions [25]. Three major hypotheses have been explored in the literature pertaining to BE development: transdifferentiation of resident squamous epithelial stem cells, emergence of residual embryonic stem cells, and migration of stem cells from a distinct location.

*Transdifferentiation*

In terms of the number of peer-reviewed studies, the “transdifferentiation” hypothesis for BE development has been the most popular. This hypothesis posits that the squamous epithelial stem cells of the distal esophagus transdifferentiate upon some stimulus into
columnar epithelia cells to form BE lesions [26, 27]. In other tissues, it has been shown that adult stem cells can transdifferentiate into other cell types [28, 29]. One study offers an apparent intermediate stage in BE transdifferentiation, in which an epithelial cell had markers of both squamous and columnar epithelia [30]. Alternately, Jiang et al (2017) identified a transitional epithelium in the distal esophagus which contained a squamous–columnar junction basal cell population [31]. Upon exposure to CDX2, these cells can differentiate into intestinal-type cells. In the small intestine, columnar epithelia transdifferentiate into goblet cells and other intestinal-type cells through the actions of CDX1 and CDX2 [32]. CDX gene expression in the adult esophageal epithelium is a well-validated marker of intestinal metaplasia [33-35]. In BE, CDX1 is known to upregulate markers of epithelial differentiation, including cytokeratin 20 and villin [36]. Intestinal metaplasia in BE is positive for CDX2, MUC2, and villin expression [37]. CDX1 also targets effectors of Wnt, Retinoic Acid, and Fibroblast Growth Factor (FGF) signaling [38-40]. CDX1 expression can induce transdifferentiation of esophageal keratinocyte cell lines to a cell type which stains positive for acid mucins by Alcian Blue staining [41]. It is therefore possible that CDX1 and CDX2 expression in BE results in transdifferentiation of cells within an existing simple columnar epithelium to goblet cells and other intestinal-type cells. However, the molecular signaling events that would lead to squamous epithelial transdifferentiation have not been revealed.

Several studies have investigated the initiating factors which may prompt transdifferentiation in BE, including bile acids and retinoic acid [18, 42, 43]. In a 2015 study by Sun et al, bile acid treatment in a rat model can induce expression of markers of BE, including CDX2 and MUC2 in BE-like metaplastic lesions [42]. However, in that
study, the formation of the metaplasia was believed to result from the heightened inflammatory response in the esophagus upon bile acid treatment. While it has been shown that bile acids can be potent modulators of cell signaling, especially in a neutral pH refluxate, no evidence has been put forth that shows its role in BE transdifferentiation. In the embryonic esophagus, a simple columnar epithelium is differentiated into a stratified squamous epithelium [44]. In embryogenesis, such differentiation of epithelial stem cells is triggered by retinoic acid signaling [45]. Retinoic acid activity is increased in BE lesions, which may elucidate the increased expression of CDX genes in those same tissues [45]. CDX1 and CDX2 are downstream effectors of retinoic acid signaling [38, 46, 47]. Together, this indicates some role of retinoic acid signaling in BE development.

Residual embryonic cell populations

Tumor protein p63 is responsible for maintenance of stratified epithelia [48]. It has been documented that BE tissues do not express p63, which distinguishes them from the normal esophageal squamous epithelium [49]. In p63-null mice, stratified squamous epithelial tissues are replaced by a simple columnar epithelium which resembles BE lesions [48, 49]. Since p63 is not necessary for differentiation of the squamous epithelium, a 2011 study by Wang et al contended that BE may not form from transdifferentiation, but instead a physical replacement of the squamous cells by a residual embryonic population of columnar epithelia [50]. In mouse embryonic development of the esophagus and proximal stomach, a simple columnar epithelia precedes the formation of the mature stratified squamous epithelium [51]. Also in the Wang 2011 study, both human BE and the esophagus of a p63-null mouse were observed
to exhibit similar expression of known BE biomarkers, including Keratins 8, 18, and 20, as well as Anterior gradient 2 (Agr2), Trefoil factors (TFF) 1-3, and Villin1 [50, 52, 53]. CDX2, a marker of the intestinal-type cells which occur in BE lesions, is not expressed in p63-null mouse esophagus [50]. Perhaps, expression of CDX2 expression in the distal esophagus, and thus the appearance of intestinal-type cells, is a separate event from the development of the simple columnar epithelium.

**Stem Cell Migrations**

The hypothesis that BE is the result of stem cell migration is also popular. The initial location of the migrating stem cells is a matter of much debate. Since BE was first described in the 1950s, there has been a number of studies investigating the idea that the stem cells maintaining these lesions originate in the gastric cardia, the region of the stomach immediately adjacent to the gastroesophageal junction[54-56]. The epithelium of the gastric cardia is columnar, resembling the columnar epithelia characteristic of BE [55]. Intestinal metaplasia does occur in the gastric cardia [57]. These lesions differ from BE, as they do not possess the same drastically increased risk of cancer [58].

An alternate source for migrating BE stem cells are the esophageal submucosal glands. Intestinal metaplasia has been discovered in these glands and is referred to colloquially as buried metaplasia [59]. Some studies indicate that columnar epithelial cells are present in these submucosal glands and even share mutations with adjacent BE lesions [60]. Surveillance through upper endoscopy with biopsy can miss the lesions [61]. Therefore, it is currently unknown whether buried metaplasia is the source of BE lesions or simply an
extension of adjacent lesions. Alternatively, Garman and McCall argue that buried metaplasia may not be intestinal metaplasia but instead the result of acinar ductal metaplasia [25].

Sarosi et al (2008) hypothesized and showed in a rat reflux model that bone marrow derived cells found in the esophagus could be linked to formation of the intestinal metaplasia [62]. This study was performed in a rat esophageal reflux model, and formation of the intestinal metaplasia was observed after a bone marrow transplant. This hypothesis remains controversial, because its applicability to human BE patients with chronic acid reflux was not clearly established [17, 63].

While hypotheses concerning the cellular origin of BE abound in the literature, it has become obvious that the inability to observe BE development in vivo, either in an animal model or a 3-D human cell culture model has been a large roadblock in the field’s progress [64]. Thus, the drive for insight into the origin of BE has turned towards the differential signaling which occurs in each stage of BE pathogenesis [65]. Although each hypothesis discussed regarding the origin of BE has merit, there is much to be learned about these lesions. Our incomplete understanding of how BE develops and what induces its progression to dysplasia and EAC hinders our ability to appropriately diagnosis, treat, or prevent BE and EAC. Identifying pertinent signaling pathways and biomarkers for each stage of BE pathogenesis may aid in these processes [66].
Homeotic Genes in BE Development

BE lesions Resemble Homeotic Transformations

Homeotic transformation refers to the development of structures in an inappropriate body region. The classic example occurs in Drosophila where a leg develops in the place of an antennae, a phenomenon known as antennapedia [67]. Several human gene disorders are attributed to homeotic transformations [68]. BE occurs in the esophagus, but it is characterized by the appearance of an epithelium resembling a more posterior portion of the gastrointestinal tract, the small intestine. BE developments in adulthood, rather than during embryonic development, but its resemblance to homeotic transformations is striking [9, 69]. Homeotic transformations are the result of dysregulated homeotic gene expression [70].

Homeotic genes are transcription factors which contain a homeobox domain and determine segment identity along the anterior/posterior axis. The ANTP class, the largest class of homeotic genes, is only found in metazoans [71]. The name of this class is derived from Antennapedia gene in Drosophila. 255 ANTP-class homeotic genes are present in the human genome. Many of these are in the HOX, ParaHOX, and Nyx gene clusters. Eight of the known human homeotic genes have been shown to be differentially regulated in BE tissues (Table 1). The HOX and ParaHOX gene clusters are heavily involved in the normal development of the human GI tract. These gene clusters are hypothesized to have originated from the same ancestral ProtoHOX gene clusters, based on the similarities in their homeobox domains [72].
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Regulation/Mutation in BE</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDX1</td>
<td>caudal type homeobox 1</td>
<td>Upregulated</td>
<td>[1]</td>
</tr>
<tr>
<td>CDX2</td>
<td>caudal type homeobox 2</td>
<td>Upregulated</td>
<td>[2-5]</td>
</tr>
<tr>
<td>HOXB5</td>
<td>homeobox B5</td>
<td>Upregulated</td>
<td>[6]</td>
</tr>
<tr>
<td>HOXB6</td>
<td>homeobox B6</td>
<td>Upregulated</td>
<td>[6]</td>
</tr>
<tr>
<td>HOXB7</td>
<td>homeobox B7</td>
<td>Upregulated</td>
<td>[6, 7]</td>
</tr>
<tr>
<td>PDX1</td>
<td>pancreatic and duodenal</td>
<td>Upregulated</td>
<td>[8-10]</td>
</tr>
<tr>
<td></td>
<td>homeobox 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARX1</td>
<td>BARX homeobox 1</td>
<td>SNP¹</td>
<td>[10-14]</td>
</tr>
<tr>
<td>MSX1</td>
<td>msh homeobox 1</td>
<td>SNP¹</td>
<td>[15]</td>
</tr>
</tbody>
</table>

¹ SNP: single nucleotide polymorphism
Members of the HOX gene clusters are expressed throughout the body [72]. In embryogenesis, HOX genes are expressed in the mesoderm in a certain pattern which specifies each body segment, termed the HOX code [73]. As embryonic development progresses, HOX gene expression is initiated in the posterior regions of the body and spreads anteriorly [74]. This promotes “posterior prevalence” which endures past embryonic development and into adulthood. HOX genes are essential in the formation and specification of the three GI segments, the foregut, midgut, and hindgut, in embryonic development [75-77].

In bilateral metazoans, HOX gene spatial and temporal expression along the AP axis is essential in maintaining segmental identities [78]. As a result of genome duplications in human evolution, HOX genes occur in four clusters: HOXA, HOXB, HOXC, and HOXD [79, 80]. The expression of HOX genes in a cluster is collinear: the order of a HOX gene on the chromosome determines its positional expression in the body. Posterior HOX genes from the HOXB cluster have been shown to be upregulated in BE tissues (Table 1) [69].

HOX genes are regulated by protein complexes which modify chromatin structure. Polycomb Repressive Complexes (PRC) promote closed chromatin structure and repress HOX gene expression, while Multiple Lineage Leukemia (MLL) complexes promote open chromatin structure and activate HOX gene expression [81-84]. It is unknown how these complexes are recruited to the appropriate HOX genes to regulate gene patterning along the AP axis.
ParaHOX genes regulate posterior gut development as well as HOX gene expression

The ParaHOX gene cluster includes PDX1, CDX1, CDX2 and CDX4 [72]. Expression of the ParaHOX genes are confined to the endoderm and regulate the expression of other homeotic genes in the gut along the AP axis [72]. CDX2 is the most posterior ParaHOX gene. All three CDX genes are expressed in the gut endoderm of the human embryo and have roles in gene expression patterning in posterior segments of the gut [85]. Together, these genes act as transcriptional activators for more than 3,000 genes [86]. They are significant players in the maturation, differentiation, and maintenance of the intestinal epithelium [86]. Both CDX1 and CDX2 are not expressed in the normal squamous epithelium of the distal esophagus, yet aberrant expression of both CDX genes has been widely documented in BE [17, 33-36, 87-89].

In a CDX2-null mouse embryo, an apparent homeotic transformation occurs, wherein the intestine develops with a mucosa resembling the esophagus and proximal stomach instead of the mature differentiated epithelium [87]. In BE, CDX1 is known to upregulate markers of epithelial differentiation, including cytokeratin 20 and villin [90]. CDX1 also targets effectors of Wnt, Retinoic Acid, and Fibroblast Growth Factor (FGF) signaling [38-40]. CDX gene activity can be regulated by Wnt/B-catenin signaling and BMP4 signaling, which are dysregulated in BE [91].

In the small intestine, CDX1 and CDX2 regulate expression of the HOX genes [39]. Central genes from all four HOX gene clusters can be targeted directly by CDX2 [92]. HOXA genes have been found to be activated by CDX1. CDX1 binds to a CDX-binding site in the HOXA gene cluster and acetylates Histone H3 at Lysine 27, an activating
epigenetic mark [93]. CDX gene inactivation can affect central HOX gene expression, which can manifest as homeotic transformations in vertebral column [47].

**Protein Markers of BE**

BE tissue can be differentiated from normal esophageal tissue due to the presence of intestinal-type cells and associated molecular signaling. Numerous signaling pathways involved in inflammation, epithelial differentiation, and tumorigenesis are dysregulated in BE lesions [94]. Sonic Hedgehog, canonical Notch, and NF-κB signaling all appear to be dysregulated in BE tissues [95-99]. Intestinal-type proteins including various columnar cytokeratins, AGR2, TFF1-3, villin, the acid mucins, and CDX1 and CDX2 are aberrantly expressed in BE [37].

Epithelial cytokeratins are the intermediate filaments types I and II composing the cytoskeleton of epithelial cells [100, 101]. Different subsets of cytokeratins (CK) are expressed by normal squamous epithelium and BE lesions [102]. The diversity of cytokeratin expression in epithelial tissues is thought to be linked not only to the epithelial cell type but also to the extent of differentiation in the tissue [100, 103]. Squamous epithelia express a larger subset including CK10, CK13m and CK14 [36, 51]. Since BE lesions are a heterogeneous collection of different cells, the subset of cytokeratins expressed can vary. CK7, CK8, CK18, CK19, and CK20 are all associated with BE tissues [36, 51, 102]. Cytokeratins are known to be differentially expressed in tumorigenic tissues and have been considered for serum biomarkers for various cancers [100].
Anterior gradient 2 (AGR2), Villin, and Trefoil Factors 1-3 (TFF1-3) are all proteins expressed in mucosal epithelium, such as the lining of the small intestine [65]. Like the columnar cytokeratins, villin is a cytoskeletal protein essential to form microvilli in the small intestine [104]. The trefoil factors are associated with mucins secreted by columnar epithelial and goblet cells in the small intestine [105]. All three proteins are highly upregulated in BE and EAC [104-107].

The acid mucin, MUC2, is an excellent protein marker of BE lesions. Goblet cells, whose presence is required to diagnose BE, and other intestinal-type cells secrete mucins, including MUC2 [108]. Mucins are not normally expressed in the esophageal epithelium; however, their expression in BE lesions results in an adherent mucous barrier [109]. This barrier contains neutral mucins, acid mucins, sialomucins, and sulfomucins [110, 111]. The extent of expression and the types of mucins expressed vary among BE patients. The expression of mucins is promoted by the abnormal expression of CDX2 in BE [89].

**Surveillance of BE Development and Progression to EAC**

*Upper GI Endoscopy with Biopsy*

The gold standard for surveillance of BE pathogenesis is an upper gastrointestinal endoscopy with histological confirmation from biopsies [112, 113]. This invasive method to diagnose and monitor BE first requires visual examination of the distal esophagus. BE lesions possess a distinct color and texture compared to the normal esophageal epithelium. Upon locating BE lesions, puncture biopsies are obtained along the length of the lesion. In the standard Seattle protocol, biopsies are to be taken every 2 centimeters
[64]. The biopsies are examined for characteristics of BE as well as signs of dysplasia and cancer. A hematoxylin-eosin stain shows the presence of the simple columnar epithelium characteristic of BE, as well as signs of cancerous cells. Confirmation of BE is done through Period-Acid Schiff (PAS) staining combined with an Alcian Blue stain [58]. PAS stains neutral mucins, while Alcian Blue stains acid mucins. Acid mucins are secreted by goblet cells, the defining characteristic of BE lesions [114, 115]. Neutral mucins tend to be associated with the columnar epithelia [114]. While upper endoscopy with biopsy is the most effective tool available to monitor patients for BE pathogenesis, it poses risks to patients. Insertion of the endoscope into the esophagus can result in coughing and gagging [116]. The need for sedation to prevent gagging to safely complete the procedure is also a disadvantage to this method. Inflammation and esophageal tearing can occur, as well as infection at the biopsy sites [117, 118]. Patient anxiety is extremely common with these procedures and can prevent patient compliance to the suggested surveillance regimen.

Dysplasia in BE lesions can be difficult to confirm histologically, as they rely on a pathologist’s interpretation of a small number of biopsies. Montgomery et al (2001) showed inflammation could obscure accurate diagnosis of a BE patient, even when the tissue was examined by multiple experts [119]. This is a critical problem for patients and clinicians. Early detection of dysplasia in BE patients is vital to diagnosing EAC in its earliest stages, when it can be ablated or surgically resected from the esophagus. False results from endoscopy with biopsy can lead to two undesirable scenarios. With a false negative diagnosis, a patient progressing to metastatic EAC could have necessary treatment delayed by months to years. In the case of a false positive diagnosis, a BE
patient without dysplasia would face unnecessary additional surveillance with its associated risks. The prevalence of false diagnoses in BE patients has been documented [120, 121]. The invasiveness and health risks posed by upper endoscopy with biopsy appear to far outweigh its sensitivity and specificity. The cost effectiveness of this method has been controversial.

*Alternative Surveillance Techniques*

Producing an alternative surveillance technique for BE and its pathogenesis with superior sensitivity, accuracy, and safety is a major clinical challenge. Several alternative surveillance techniques to monitor BE pathogenesis exist but are not currently included in AGA/ACG recommendations. These surveillance methods utilize visual examination of the esophagus, epithelial cell collection, and biomarker analysis. A 2017 review by Offman et al showed that many proposed alternative techniques explored rely on less invasive visual examination of the esophagus but still require a puncture biopsy for a confirmed BE diagnosis [116]. Cell collection methods such as the Cytosponge procedure provide less invasive means to collect cells from the distal esophagus for diagnosis confirmation [122]. Contamination by epithelia from the proximal esophagus and the oral cavity is a concern for these alternate methods. A balloon-based cell collection method described by Moinova et al (2018) circumvents this issue by protecting the sample inside a capsule before withdrawal from the distal esophagus [123]. The invasiveness of the original surveillance method is at least partially addressed by these alternatives. The sensitivity and specificity of these techniques for BE pathogenesis still relies upon histological analysis of the esophageal tissues or cells.
Alternative Diagnostic Techniques

An alternative confirmation method is to probe for tissue biomarkers associated with each stage of BE pathogenesis. BE tissues can be differentiated from normal squamous epithelial tissues by the previously introduced protein markers of BE. However, it can be difficult to distinguish non-dysplastic BE from BE with dysplasia, since many of the same proteins are expressed yet the hallmarks of cancer observed in EAC are not present [124, 125].

Epigenetic biomarkers can differentiate BE tissues from normal squamous epithelia. Moinova et al found that DNA methylation at the vimentin and cyclin A1 loci could differentiate BE patients from non-BE patients [123]. Prior to that study, Kaz et al (2016) showed unique DNA methylation signatures between BE and EAC patients [126]. These biomarkers are associated with the simple columnar epithelia in BE, rather than the intestinal-type cells. The microRNAs are another set of promising epigenetic biomarkers for BE pathogenesis. Over 90 microRNAs have been found to be differentially expressed in BE and EAC when compared to normal squamous epithelia [94].

An ideal technique to diagnose BE pathogenesis would be serum biomarker testing. Collection of patient blood is minimally invasive. Such testing requires a panel of serum biomarkers which can differentiate among the stages of BE pathogenesis. Serum metabolites can be used to differentiate BE from GERD and from BE with dysplasia [127]. EAC can be identified by hypermethylation at the adenomatous polyposis coli promoter regions [128]. In their exploratory study, Bus et al (2016) demonstrated that seven serum microRNAs are differentially expressed in BE and EAC [129]. This was
further supported by Chiam et al (2015) which identified serum microRNA signatures which could distinguish BE patients from EAC patients [130].

Despite the potential protein, DNA methylation, and microRNA biomarkers for BE pathogenesis, only the acid mucins are used as diagnostic biomarkers for BE in the clinic. Upper endoscopy with biopsy remains the gold standard for diagnosis of BE and its progression to EAC. One significant impediment to furthering these techniques is the limited molecular understanding of BE development and its progression to EAC.

**Utilizing microRNA Signatures to understand BE Pathogenesis**

This study pursued biomarker signatures which could differentiate each stage of BE pathogenesis. The biomarkers of interest here are the microRNAs. These noncoding RNAs are between 11-22 nucleotides in length and possess the ability to repress translation [131]. Each microRNA contains a seed sequence that is complementary to a region on its target genes’ transcripts. Usually, the mRNA target region is within the 3’ untranslated region (UTR) [132]. microRNAs bind as part of a RNA-induced silencing complex (RISC) [133]. microRNAs repress translation through two mechanisms, either by a physical block of ribosome binding or by recruitment of endonucleases [131] [134]. The RISC prevents ribosome binding to the target mRNA, leading to a reservoir population of microRNA-associated ribonucleoprotein complexes [135]. Some mRNA targets are then degraded, by removal of the poly A tail by GW182 and CCR4-NOT deadenylase complexes [133]. Repression of translation by microRNA does not require perfect complementarity between the seed sequence of the target. One microRNA can target hundreds of gene transcripts across multiple signaling pathways, making them
excellent indicators of disease [136-138]. MicroRNAs can also be introduced into the bloodstream as part of exosomes. MicroRNAs have been identified as biomarkers for colorectal cancer, breast cancer, diabetes mellitus, and traumatic brain injuries as well as many other diseases and conditions [139-143].

Many serum and tissue microRNAs have been found to be differentially expressed in BE and EAC [94]. None of these microRNAs are currently being used as a biomarker for BE, and no study to date has systematically identified any microRNA serum or tissue biomarkers for the development of BE. Despite the diverse roles microRNAs play in molecular signaling, microRNA biomarkers have not been utilized to identify pertinent signaling mechanisms in BE pathogenesis.

This study seeks to examine the relationship between microRNA expression and the development and progression of BE. The first aim of this study was to identify a microRNA signature which could differentiate among all stages of BE pathogenesis. No such signature was identified. The second aim was to reveal microRNAs that were differentially expressed in pair-wise comparisons between stages of BE pathogenesis. Identified microRNAs appeared to be involved in EAC tumorigenesis or BE development. The third and last aim of this study was to determine the biological mechanisms and processes regulated by microRNAs differentially expressed in BE pathogenesis. Certain microRNAs identified in this study have roles in the regulation of posterior homeotic genes, which is consistent with the resemblance of BE to a homeotic transformation.
MATERIALS AND METHODS

Small RNA-sequencing of microRNAs in Patient Tissue and Serum Biopsies

Collection of Fresh-Frozen Tissue and Serum from Human Patients

Tissue and serum sample collection was performed at Dayton Veteran Administration Hospital by Dr. Sangeeta Agrawal. Both tissue and serum samples were collected from each patient during routine upper endoscopy with biopsy appointments. For this study, 76 tissue and serum samples were processed from 38 patients. Upon collection, fresh tissue biopsies were kept in 100mL RNA-later at -80°C. Upon collection, serum biopsies were kept in at -80°C until small RNA could be isolated.

The patients included were classified into five groups: GERD, BE, BE with indefinite dysplasia, BE with low grade dysplasia and EAC. Patients were selected for this study based on their initial diagnosis prior to collection of tissue and serum samples. For this study, BE patients with indefinite dysplasia and BE patients with low grade dysplasia were combined into one group termed LGD.

Tissue and serum samples from patients labelled Normal were also collected, yet this group is not included in the final analysis. The Normal group was intended as a negative control. However, after a preliminary analysis of the small RNA-seq data, it was observed that microRNA expression in the Normal group was more variable than the
other pathology groups in both serum and tissue. The Normal moniker only indicated the absence of GERD or BE in these patients. It did not account for the variety of other conditions which required the patient to undergo an upper GI endoscopy, thus they were not healthy controls.

Prior to sequencing, small RNAs must be isolated from each sample and cDNA libraries prepared from the isolated small RNAs. Under advisement from Life Technologies, the manufacturer of the Ion Proton instrument, the patient tissue and serum samples were prepared for sequencing in slightly differing manners. However, sequencing of tissue and serum followed the same procedure.

Small RNA Isolation from Patient Tissue Samples

Small RNAs were isolated from homogenized tissues following the manufacturer’s protocol provided for the mirVana™ Paris™ RNA and Native Protein Purification kit (#AM1556, ThermoFisher Scientific). Fresh-frozen tissues were removed from RNA-later and homogenized in 500mL Cell Disruption buffer. The tissue fragments were either chopped into small pieces with a razor blade and then homogenized with a small hand-held homogenizer or homogenized with a rotor-stator homogenizer. All tissue homogenization was done on ice. Following small RNA isolation, the small RNAs were treated with the TURBO DNA-free™ kit (#AM1907, ThermoFisher Scientific) to eliminate any possible DNA contamination. The lack of DNA contamination was confirmed by monitoring the isolated product by quantitative polymerase chain reaction (qPCR) for GAPDH amplification (#4333764T, Applied Biosystems).
Yield and quality of the isolated small RNAs were assessed through the Agilent Bioanalyzer small RNA kit (#5067-1548, Agilent Technologies). The Agilent Bioanalyzer (#G2939BA, Agilent Technologies) acts as a digital agarose gel. The virtual gel provides the percentage and concentration of microRNAs based on the size of oligonucleotides in the isolation product. Isolation products which contained over 30% microRNA were selected for sequencing and progressed to cDNA library preparation.

Small RNA Isolation from Patient Serum Samples

Although 1000μL serum is the recommended starting volume for small RNA isolation from serum, 1000μL serum was not available for every patient in this study. In order to limit introduction of variability into cDNA libraries made from the serum samples, the starting volume was reduced to 600μL serum for every patient. Small RNAs were isolated as directed by mirVana™ Paris™ RNA and Native Protein Purification kit. Yield and quality by Agilent Bioanalyzer small RNA chip was not assessed for serum samples. Instead, the 20-30μL isolated product was concentrated to 5μl with a speed vacuum concentrator.

cDNA Library Preparation of Tissue small RNAs

Tissue cDNA libraries were prepared using the Ion Total RNA-seq v2 kit by Ion Torrent™ (#4479789, ThermoFisher Scientific). Per the kit protocol, the small RNA yield is diluted to 10 μg/μl with nuclease-free water, based on the microRNA concentration provided by the Agilent 2100 Bioanalyzer instrument. The small RNAs were hybridized with a proprietary mix of adaptors which were ligated to the 5’ end of
the small RNAs. The adaptors allow for the next steps which are reverse transcription to
cDNAs and subsequent amplification of the cDNAs. The proprietary PCR primers
require the presence of the hybridized adaptors in order to amplify the entire cDNA
population. The 3’ primer also includes a barcode sequence. These barcode primers allow
for patient samples to be differentiated from one another. This enables pooling of
multiple samples during sequencing and is essential later for data analysis. With the
inclusion of the adaptors and PCR primers, cDNAs reverse transcribed from microRNAs
tend to be between 94-114 nucleotides. To filter cDNA libraries for this size range, a
purification and size selection step was performed before and after cDNA amplification.
Since tissue samples contain a vast number of varied nucleic acids, the rigorous protocol
for the total RNA-seq v2 kit was utilized for purification and size selection, as it entails
two ethanol washes.

Following library preparation, the Agilent Bioanalyzer DNA 1000 chip (#5067-1504,
Agilent Technologies) was used to determine the quality of each cDNA library. The
DNA 1000 chip provides size ranges corresponding to small RNAs (50-300 base pairs)
and to microRNAs (94-114 base pairs). These size ranges account for the addition of
proprietary library primers and sequence adaptors to the cDNA libraries. Libraries with a
high percentage of microRNAs (> 30%) and a high concentration of small RNAs (< 100
nmol/l) underwent small RNA-sequencing.

*cdDNA Library Preparation of Serum small RNAs*

Serum cDNA libraries were also prepared according to the Ion Total RNA-seq v2 kit
protocol, except for the purification and size-selection steps. The initial concentration of
small RNA used for each serum cDNA library is unknown. However, the total volume from small RNA isolation was concentrated to 5μl, and 4μl of that volume initiated cDNA library preparation.

Since the concentration of total RNAs in serum is assumed to be lower than fresh-frozen tissue, a less rigorous protocol for purification and size-selection is needed. Therefore, the Total Exosome RNA and Protein Isolation kit (#4478545, ThermoFisher Scientific) protocol was utilized for the purification and size-selection steps only.

As with the tissue samples, quality and quantity of cDNAs derived from small RNA and microRNAs was assessed by the Agilent Bioanalyzer DNA 1000 chip.

**Small RNA-sequencing of microRNAs in Patient Tissue and Serum Biopsies**

In this study, the Ion Proton™ system for Next-Generation Sequencing (Ion Torrent™, Life Technologies) was utilized to sequence small RNAs in each patient tissue and serum sample.

Small RNA-sequencing requires pooling of samples and another amplification of the cDNA libraries by emulsion PCR. First, each cDNA library was diluted to 5nM with nuclease-free water, and then up to 14 samples were pooled together and diluted to a final total cDNA library concentration. The Ion One Touch™ 2 system (#4474779, Life Technologies) was utilized for emulsion PCR. In emulsion PCR, every cDNA library fragment was ligated onto an Ion Sphere™ Particle (ISP) beads and then amplified (Ion PI™ Hi-Q™ OT2 200 kit, #A26434, Life Technologies). The 3’ primer for emulsion PCR contains a B primer sequence which can hybridize with the ISPs as well as the P1
adaptor sequences. A quality control step was included at this stage using the Ion Sphere™ assay for the Qubit™ 2.0 Fluorometer (#4468656, #Q32866, Invitrogen). The quality control metric was the percentage of ISP beads which contained successfully ligated cDNA library fragments, termed the Percent Templated ISPs. The minimum percentage suggested by the manufacturer is 10%. This percentage is derived from the fluorescence of two fluorescent probes, Alexa Fluor™ 647 and Alexa Fluor™ 488. Alexa Fluor™ 647 binds to the A adaptor sequences. Alexa Fluor™ 488 anneals to a site on the B primer sequence. For the Ion Proton system, less than 10% Templated ISPs could result in a lower amount of total sequenced reads for a small RNA-seq run (Appendix I). Ideally, each ISP was populated with only one cDNA library fragment. However, multi-templated ISPs do occur. In this study, the Percent Templated ISPs was not an ideal indicator of their occurrence or the total reads sequenced in a small-RNA seq run (Appendix I). Following the Qubit™ assay, the ISPs were loaded onto an Ion PI sequencing chip for small RNA-seq.

Small RNA-sequencing required the Ion PI™ Hi-Q™ Sequencing 200 kit (#A26772, Life Technologies) and the Ion PI™ Chip (#A26771, Life Technologies). The sequencing results underwent a preliminary analysis in the small RNA_analysis plugin provided by the Ion Torrent Suite v.5.10 software. The analysis plugin software provided the total small RNA reads and microRNA reads per patient sample. Raw sequences for those samples which yielded 200,000 or more microRNA reads were uploaded to Partek® Flow® (Partek Incorporated) to be converted to normalized microRNA read counts for analysis.
Preparing small RNA sequences for Analysis of Differential microRNA Expression

Raw small RNA sequences are not ready fodder for analysis of differential microRNA expression, because the identity and quantity of microRNAs in each sample is not available. Here, Partek® Flow® software (Partek Incorporated) was utilized to convert the raw sequences (available in .fastQ format) into a more readily analyzed format: normalized microRNA read counts. Small RNA sequences from the 21 tissue samples were analyzed separately from the small RNA sequences for the 25 serum samples.

Pre-Alignment Processing Removes Low Quality Bases from Raw Sequences

Partek® Flow® allows for trimming of bases from the 5’end or 3’ end of sequenced reads, in order to maximize the alignment of reads to a reference database. In the sequencing process, both ends of a read are the result of more cycles than the middle portion, which predisposes base calls at either end to be less reliable.

Standard practice is to trim sequences of bases with a Phred quality score below 20. A Phred quality score describes the likelihood of a miscalled base. A Phred quality score of 20 is equivalent to a 1% probability of a miscalled base. To facilitate downstream processes, read lengths of approximately 170 base pairs were required. In order to obtain this read length for reads in all cDNA libraries, trimming was based on a Phred quality score of 19 or a 1.26% probability of a miscalled base. Sequences derived from microRNA products were between 94-114bp. Parameters for base trimming for the serum and tissue datasets maintained a minimum read length of 15bp.
Aligning the Raw Sequences to miRBase v.21 eliminates non-microRNA sequences from Analysis

Following base trimming, the raw sequences were aligned to a genome reference. Since our interest was solely in microRNA expression, the genome reference was miRBase v.21, a database which has compiled the sequences of every known human microRNA [144, 145]. In line with current literature and recommendations from Partek, the aligner algorithm used was the Bowtie aligner [146, 147]. Bowtie was constructed specifically to align short DNA sequences to larger genome references.

Quantification of aligned reads to miRBase v.21 results in microRNA read counts

In order to convert the raw sequences to microRNA read counts, the aligned sequences must be quantified. Here, the sequences were mapped to known human microRNA sequences using a microRNA annotation file from miRBase v.21. Partek® Flow®’s modified expectation/maximization (E/M) quantification algorithm provided estimated expression for each microRNA. Partek® Flow® documentation indicates this modified algorithm can also correct for any sequences which map to multiple locations in the human genome. Here, minimum seed length was set to 10bp with only 1 mismatch allowed between a read and a known microRNA seed sequence.

Normalization of microRNA read counts by Trimmed Mean of M-values

A Trimmed Mean of M-values (TMM) normalization was applied to the microRNA read count datasets for the tissue and serum samples [148-150]. This is to account for the variability in total microRNA reads in each patient sample, which can vary due to any of
the many technical steps required for small RNA sequencing. TMM was advantageous, because it does not use microRNAs with very low or very high expression for normalization. The distribution of microRNA read counts in the serum and tissue datasets was skewed to the right, due to the presence of those microRNAs with very high expression. Instead, this method focuses only on the microRNAs for which the read count distribution in all samples approximates a Normal distribution.

A portion of microRNAs had zero read counts in some tissue and serum samples, therefore an offset value of 1.00 was added to each read count value to prevent non-zero errors in the calculation of M-values in TMM normalization. This offset value also prevents non-zero errors in the calculation of expression fold change in downstream analyses.

*Filtering based on TMM-normalized microRNA read count applied to Serum and Tissue datasets*

Both the serum and tissue datasets were filtered in Partek™ Flow™ for microRNAs which contained 100 or more TMM-normalized microRNA read counts in 25% of patient samples. All downstream analyses were performed on the filtered serum and tissue datasets. This 100 microRNA read count filter was intended to reduce the absolute noise caused by microRNAs with very low read counts in most samples in the datasets. In addition, this threshold alleviates the possibility that significant differential expression is due to sequencing anomalies [151].
The minimum number of microRNA read counts permitted for each sequenced sample included in data analysis was 200,000. After alignment and quantitation to miRbase v.21, the minimum average read count for each microRNA in each sample was 77.28. Thus, setting the microRNA count threshold to 100 microRNA read counts was not a particularly stringent filter. This read count filter allowed for low read counts in one pathology group, while retaining enough in the three other pathology groups to allow for ready detection of any differential microRNA expression.

Identification of Potential microRNA Signatures from Normalized microRNA Read Counts

Since only a total of 21-25 patients were included in the serum and tissue analyses, this severely limits the statistical methods which can be utilized to analyze the filtered TMM-normalized microRNA read count data. To identify microRNA signatures which could differentiate all the stages of BE pathogenesis, three general techniques were utilized: clustering analyses, Nearest Shrunken Centroids (NSC) classification, and the randomForest classifier.

Here, the clustering analyses used were hierarchical clustering and a combination of a heat map analysis with k-means clustering. Both analyses were performed in the Orange Data Mining version 3.20.1 software [152].

The R Statistical Programming software provides an interface for several statistical packages useful for analyzing large data sets [153]. The randomForest package in R is an excellent candidate to analyze the dataset in this study, as it allows for a large p and a
comparatively small n [154, 155]. Since BE pathogenesis follows a defined step-wise progression, classification randomForest and regression randomForest analyses were both deemed appropriate for this study. The regression component allows the regression randomForest to account for pathology order. For this analysis, the stages of BE pathogenesis were numbered sequentially with GERD being 1 and EAC being 4. Nearest Shrunken Centroids (NSC) analysis was performed using the pamr package in R [156].

**Pair-wise Comparisons Between Pathology Groups to Identify Candidate microRNAs**

Candidate microRNAs in this study are those microRNAs which are differentially expressed between two pathology groups. Identification and subsequent gene target analysis of these microRNAs was utilized to understand BE pathogenesis.

The Partek® Flow® software contains a Gene Specific Analysis (GSA) function which allows grouping of samples into categories and pair-wise comparisons between the sample groups. Here, the patient samples were grouped by their pathology: GERD, BE, LGD, and EAC. To execute the comparisons, the Log Normal with Shrinkage (LNS) statistical model was selected as the best model for this study. The Akaike Information Criterion corrected (AICc) values for each microRNA tended to be lowest when utilizing the LNS model in our study and in a 2004 study by Burnham and Anderson [157]. The AICc values are measures of how well a statistical model fits a data set [158, 159].

Partek® Flow® documentation regarding the use of LNS in the GSA suggests exclusion of features in the datasets with very low expression. To accomplish this, the lowest
average coverage value was increased from 1.0 to 4.0 in both the tissue and the serum analyses. This value was based on a visual examination of the LNS shrinkage plot generated in Partek ® Flow ®. This eliminated any microRNAs which had a geometric mean across all samples below 4.0. The comparisons made were according to the sequence followed by BE pathogenesis: BE vs GERD, LGD vs BE, EAC vs BE, and EAC vs LGD. The GSA algorithm then provided an uncorrected p-value and fold change for each microRNA in each sample.

Before application of GSA algorithm, the total number of microRNAs aligned and quantified to miRBase v.21 was 2,588. After adjusting the lowest average coverage to 4.0, the number of tissue microRNAs reduces to 1,193, while the number of serum microRNAs decreases to 718 microRNAs.

In order to identify candidate microRNA biomarkers from each comparison, a Bonferroni threshold was applied to the p-values generated by the GSA [160, 161]. Here, 0.05 was set as the desired α value, and the number of observations was equal to the total number of microRNAs.

**Gene Target Analysis for Selected Candidate microRNA biomarkers**

Two approaches exist to identify gene targets for a given microRNA from gene target databases. The first requires the use of a proprietary algorithm which predicts gene targets for a microRNA based on its seed sequence as well as other factors. Which factors are included in the algorithm differ based on the target database. This first approach is utilized by miRDB and TargetScan Human 7.2 (Table 2) [162-165]. The second approach
<table>
<thead>
<tr>
<th>Method</th>
<th>Target Type</th>
<th>Filtering of Gene Target Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRDB</td>
<td>Predicted</td>
<td>Target Prediction Score ≥ 80</td>
</tr>
<tr>
<td>TargetScan</td>
<td>Predicted</td>
<td>Cumulative weighted context ++ score &lt; 0</td>
</tr>
<tr>
<td>Human 7.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIANA</td>
<td>Experimentally-validated</td>
<td>Validated by ≥ 2 low-throughput experiments</td>
</tr>
<tr>
<td>TarBase v.8</td>
<td>Predicted, Experimentally-validated</td>
<td>Validated by ≥ 1 luciferase reporter assay experiments</td>
</tr>
<tr>
<td>miRNet</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
is experimental microRNA target identification. This can be achieved through a variety of high-throughput and low-throughput methods, including expression profiling following microRNA inhibition or luciferase reporter assays. TarBase v.8 from the DIANA toolbox provides a compilation of experimentally-validated gene targets for each microRNA, while miRNet includes gene targets from multiple target databases utilizing both approaches (Table 2) [166, 167].

Several reference databases exist which have compiled experimentally-validated gene targets for many of the known human microRNAs. TarBase v8 and miRNet are two reference databases utilized to uncover gene targets for the six candidate microRNAs [166, 167]. Low-throughput assays are utilized for validation of microRNA gene targets, which are often first identified by genome-wide high-throughput screening experiments. Results from TarBase v8 were filtered for low-throughput assays only for this study (Table 2). Luciferase reporter assays are considered one of the most reliable low-throughput assays to validate microRNA gene targets [168, 169]. The gene targets compiled by miRNet were filtered for those targets which were validated by luciferase reporter assay (Table 2).

A gene’s expression can be indirectly affected by a microRNA. It is important to verify that a potential gene target contains at least one 6-8mer target site for a microRNA. Various algorithms have been created to predict microRNA targets based on microRNA sequence and potential target gene sequences. Here, Target Scan 7.2 and miRDB are two target prediction algorithms used to predict targets of the six candidate microRNAs (Table 2) [162, 165]. miRDB generates a prediction rank for each gene target. According
to the algorithm’s creators, a rank score above 80 is deemed acceptable [165]. Results from miRDB are filtered for targets ranking 80 or above (Table 2).

**Quantitative Real-Time PCR validation of Candidate microRNAs in Human Patient Tissues**

Another set of human patient tissues was obtained for validation of the small RNA-seq experiments. This cohort included 26 formalin-fixed paraffin-embedded (FFPE) samples from patients diagnosed with either Barrett’s Esophagus or Gastroesophageal Reflux Disease. The n for these experiments was based on the normalized microRNA read counts for the chosen candidate microRNAs obtained by small RNA-seq. 80% statistical power was desired, thus a total of 13 samples was necessary for each of the two pathology groups of interest.

Total RNA was extracted from the FFPE tissues using the Covaris truXTRAC FFPE microtube RNA kit (Covaris, #520161). This method utilizes the Covaris Adaptive Focused Acoustics technology to remove the paraffin from the tissues prior to total RNA extraction. The protocol for the Covaris kit was followed with an additional wash with the provided ethanol-based buffer. Without the additional wash, guanidine ITC contamination was sometimes detected following elution off the Covaris-provided columns. Quality and quantity of total RNA yield was assessed via a ThermoFisher Nanodrop ONE spectrophotometer. The guanidine ITC contamination was indicated by a low 260/230 absorbance ratio.
cDNA synthesis and qPCR followed the protocol for the TaqMan Advanced microRNA assays (#A25576, #A28007). 10ng total RNA proved to be adequate starting material. The Applied Biosystems™ QuantStudio™ 7 Flex Real-Time PCR System was used to perform qPCR on four experimental plates. To assess differential expression of candidate microRNAs, the TaqMan® Advanced miRNA Assays for each candidate microRNA were selected (478230_mir, 478585_mir, 478769_mir). The ΔΔ Ct method was used to analyze the output of the qPCR experiments [170]. Since this is a relative quantitative method, an endogenous control microRNA was needed to provide the ΔCt values. Several endogenous control microRNAs are suggested by TaqMan®; however, there are no microRNAs which are known to keep constant expression in distal esophageal tissues. One of the suggested microRNAs, hsa-miR-423-5p (478090_mir) was chosen as the endogenous control microRNA as it did not display significant differential expression between the GERD and BE tissues in the small RNA-seq experiments.
RESULTS

Small RNA-sequencing of microRNAs from Patient Serum and Fresh-Frozen Tissue

*Construction of cDNA libraries for small RNA-sequencing*

Thirty-eight patients were included in this study: 12 GERD, 9 BE, 6 LGD, 5 EAC, 5 Normal, and 1 Esophageal Squamous Cell Carcinoma (ESCC). The ESCC sample initially had been identified as EAC. When it was accurately identified as ESCC it was excluded from all further analyses. At least one serum and one fresh-frozen tissue biopsy was collected from each patient at a routine upper endoscopy appointment. For small RNA-sequencing of microRNAs, cDNA libraries were constructed from small RNAs extracted from patient serum and fresh-frozen tissue. Total RNA was first extracted from each patient sample through a phenol extraction followed by a solid-phase extraction on glass-fiber filters. Small RNAs were enriched using a larger concentration of ethanol, thereby increasing their binding affinity for a second filter before elution into 40ul nuclease-free water.

In the samples extracted from fresh-frozen tissue, a DNase-treatment was applied to 30ul extracted small RNA to eliminate any potential DNA contamination. Effectiveness of DNase treatment was verified by lack of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) amplification via qPCR. Then total small RNA concentration and microRNA
concentration were assessed by an Agilent 2100 Bioanalyzer (Figure 1). Generally small RNA samples with over 30% microRNA were selected for cDNA library preparation (Table 3). 10ng microRNA was utilized to construct cDNA libraries for the tissue samples.

Small RNA samples extracted from patient serum were not treated with DNase or assessed by Agilent 2100 Bioanalyzer. It was determined that DNase treatment could detrimentally affect the already low yield of small RNA extraction from serum (Michael Zianni, Thermo Fisher Scientific, Personal Communication). Preliminary experiments with small RNA extraction from serum samples found that concentrations of small RNAs extracted from the serum samples were often below the quantitative range of the Agilent 2100 Bioanalyzer (50-2000 pg/ul). This is not unprecedented, and in previous studies, a fixed volume was utilized rather than a fixed microRNA concentration for cDNA library construction [171]. The 30ul eluant from small RNA extraction was concentrated to 4ul using a Savant DNA Speed Vac Concentrator.

For cDNA libraries constructed from either serum or tissue libraries, double stranded DNA/RNA adaptors from the Ion Adaptor Mix v2 (Ion Total RNAseq kit v2) were ligated to both ends of the small RNAs prior to reverse transcription (Figure 2). These adaptors contain degenerate bases which allow for first strand synthesis. Size-selection for desired RNA products was then implemented by another solid-phase extraction. Here, the cDNA library fragments were bound to nucleic acid binding magnetic beads. For the tissue samples, the protocol first bound larger RNA products, such as from mRNA and rRNA, to the magnetic beads. Then the small RNA products were bound to beads with an
**Overall Results for sample 5:**

- **Small RNA Concentration [pg/µl]:** 33,157.5
- **miRNA Concentration [pg/µl]:** 24,285.9
- **miRNA / Small RNA Ratio (%):** 73

**Result Flagging Color:**

- **73 % miRNA; Concentration: 24285.90 pg/µl**

**Region table for sample 5:**

<table>
<thead>
<tr>
<th>Name</th>
<th>From [nt]</th>
<th>To [nt]</th>
<th>Average Size [nt]</th>
<th>Size distribution in CV [%]</th>
<th>Conc. [pg/µl]</th>
<th>% of Total</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small RNA</td>
<td>0</td>
<td>248</td>
<td>45</td>
<td>55.0</td>
<td>33,157.5</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>miRNA</td>
<td>10</td>
<td>40</td>
<td>26</td>
<td>29.4</td>
<td>24,285.9</td>
<td>73</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Assessment of small RNA and microRNA Yields following small RNA Extraction

A representative small RNA sample extracted from patient tissue biopsy (Agilent Bioanalyzer small RNA chip). Peaks shown represent total small RNA from a 1 microliter aliquot. Peaks between 10-40 nucleotides were considered microRNAs.
Table 3: Small RNA Extracted from Fresh-Frozen Tissue Biopsies\(^1\)\(^2\)

<table>
<thead>
<tr>
<th>Patient Sample</th>
<th>Total small RNA (ng)</th>
<th>miRNA (ng)</th>
<th>Total small RNA (ng)</th>
<th>miRNA (ng)</th>
<th>Total small RNA (ng)</th>
<th>miRNA (ng)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T</td>
<td>1634.49</td>
<td>1063.08</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2T</td>
<td>1373.13</td>
<td>857.97</td>
<td>699.75</td>
<td>112.68</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3T</td>
<td>1355.25</td>
<td>983.835</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4T</td>
<td>205.38</td>
<td>9.9</td>
<td>170.64</td>
<td>21.87</td>
<td>158.4</td>
<td>9.45</td>
</tr>
<tr>
<td>5T</td>
<td>1043.73</td>
<td>192.24</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6T</td>
<td>281.205</td>
<td>51.435</td>
<td>310.41</td>
<td>49.05</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>7T</td>
<td>311.985</td>
<td>171.225</td>
<td>145.26</td>
<td>33.93</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>9T</td>
<td>977.85</td>
<td>694.17</td>
<td>343.08</td>
<td>97.56</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>11T</td>
<td>4027.68</td>
<td>2616.84</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>12T</td>
<td>856.35</td>
<td>222.66</td>
<td>1443.375</td>
<td>860.445</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>13T</td>
<td>1351.53</td>
<td>758.97</td>
<td>834.93</td>
<td>648.18</td>
<td>548.37</td>
<td>323.73</td>
</tr>
<tr>
<td>15T</td>
<td>1013.85</td>
<td>539.73</td>
<td>1274.7</td>
<td>930.36</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>16T</td>
<td>291.21</td>
<td>45.66</td>
<td>1874.85</td>
<td>1232.76</td>
<td>1558.92</td>
<td>1077.36</td>
</tr>
<tr>
<td>17T</td>
<td>1869.57</td>
<td>944.28</td>
<td>1468.86</td>
<td>723.18</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>18T</td>
<td>1270.89</td>
<td>690.21</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>19T</td>
<td>2087.19</td>
<td>1222.56</td>
<td>1769.22</td>
<td>1352.82</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>20T</td>
<td>2736.18</td>
<td>1633.5</td>
<td>1108.8</td>
<td>506.7</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>21T</td>
<td>2747.07</td>
<td>1585.08</td>
<td>1406.76</td>
<td>1089.9</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>22T</td>
<td>1292.49</td>
<td>885.15</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>23T</td>
<td>1314.45</td>
<td>829.71</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>24T</td>
<td>1849.23</td>
<td>1087.29</td>
<td>578.31</td>
<td>406.5</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>25T</td>
<td>1368.54</td>
<td>558.18</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>26T</td>
<td>3317.49</td>
<td>1990.53</td>
<td>406.53</td>
<td>141.39</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>27T</td>
<td>1807.74</td>
<td>823.23</td>
<td>718.92</td>
<td>472.59</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>28T</td>
<td>726.75</td>
<td>495</td>
<td>587.88</td>
<td>282.84</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>29T</td>
<td>1729.26</td>
<td>1032.39</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>30T</td>
<td>362.07</td>
<td>61.92</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>31T</td>
<td>1728.54</td>
<td>1109.16</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>32T</td>
<td>1674.36</td>
<td>889.65</td>
<td>2028.33</td>
<td>1523.565</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>33T</td>
<td>981.18</td>
<td>512.19</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>34T</td>
<td>529.83</td>
<td>221.49</td>
<td>1951.83</td>
<td>965.16</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>35T</td>
<td>1541.97</td>
<td>1224.27</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>36T</td>
<td>5867.46</td>
<td>3882.06</td>
<td>1812.6</td>
<td>1243.35</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>37T</td>
<td>986.4</td>
<td>402.345</td>
<td>1346.31</td>
<td>658.98</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>38T</td>
<td>8096.085</td>
<td>5249.16</td>
<td>3351.24</td>
<td>1851.03</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>39T</td>
<td>1016.505</td>
<td>390.645</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>40T</td>
<td>4171.14</td>
<td>1201.77</td>
<td>255.825</td>
<td>70.335</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

\(^1\) Agilent Bioanalyzer small RNA chip

\(^2\) Samples indicated in boldface were used for cDNA library preparation.
Small RNAs are hybridized with a mix of 5’ and 3’double-stranded DNA/RNA adaptors which contain 5’ and 3’ single-stranded extensions composed of degenerate bases. These extensions allow adaptor binding to all small RNAs within a sample. Following single-strand synthesis and a size-selection protocol, a series of adaptors are added to the 5’ and 3’ ends of the small RNA products. The A Adaptor and Key Sequence together are 30bp. The unique barcode sequence is 10bp. The 10 bp Internal Adaptor contains the 5’ primer binding site. The P1 Adaptor is 30bp in length. (Michael Zianni, Thermo Fisher Scientific, Personal Communication).
increased ethanol concentration. However, for the serum samples, this first step was deemed unnecessary due to the assumed low input RNA concentrations. Therefore, the small RNA library size selection protocol from the Total Exosome RNA and Protein Isolation kit (#4478545, Life Technologies) was utilized for the serum derived cDNA libraries. This protocol simply requires binding of the small RNA products to nucleic acid binding magnetic beads, a wash step to remove larger RNA products, and then elution into 10ul nuclease free water.

Following size-selection for small RNA products, three new adaptors were added during PCR amplification of the cDNA library (Figure 2). The 5’ A adaptor was 40 base pairs in length and included one of 16 unique barcodes. This 10bp barcode sequence was essential to differentiate among cDNA libraries following small RNA-seq. A 5’ 10 bp internal adaptor provides the binding site for the 5’ PCR primer. The P1 adaptor is 30bp in length and added to the 3’ ends of the library fragments. The A and P1 adaptors are necessary for small RNA-seq. After amplification of the cDNA libraries by PCR, another size-selection step identical to the first occurs. The yield and size distribution for each purified cDNA library is then assessed in a quality control step via an Agilent Bioanalyzer DNA1000 chip.

Selection of cDNA libraries for small RNA-sequencing

Concentration of microRNA products in each cDNA library was determined by an Agilent 2100 Bioanalyzer (Figure 3). Since the adaptors and barcode total to 80 base pairs (Figure 2), cDNA library fragments derived from microRNAs (14-24bp) are expected to be between 94 – 114 bp [172]. Ideally, only those cDNA libraries which
Figure 3: cDNA library Selection is Based on the size of the 94-114bp peak

Representative cDNA libraries constructed from patient tissue biopsy (Agilent Technologies DNA1000 chip). The first peak at 15bp and the last peak at 1500bp are an internal DNA standards provided in the Agilent Bioanalyzer DNA1000 chip kit. (A) A representative ideal cDNA library. Large peak at 94-114bp indicates a large concentration of microRNA products. The small peak adjacent to the 94-114bp peak was the adaptor-dimer peak. Any peaks above 300bp were other small RNAs. (B) A representative suboptimal library. The second largest peak at 89bp is considered the adaptor-dimer peak.
possessed a single large peak in this size range were selected for small RNA-seq. Peaks occurring before 94 base pairs were believed to belong to adaptor dimers. High concentrations of adaptor dimers, as depicted in Figure 3B, were thought to skew the sequencing results. Peaks after 114 base pairs generally were considered contamination by small RNAs, including tRNAs, indicating reduced effectiveness in size-selection during cDNA library construction.

To maintain an appropriate number of patients per pathology group for sufficient statistical power, suboptimal cDNA libraries were included in the small RNA-seq experiments (Appendix II). 93 cDNA libraries derived from 69 serum or tissue samples were sequenced. Of those, several cDNA libraries were considered suboptimal as they contained large adaptor-dimer peaks. It became evident that small RNA-seq of cDNA libraries with large adaptor dimer peaks and small peaks at 94-114bp did not necessarily result in reduced sequencing of microRNA products (Appendix II). Some cDNA libraries with only the one large peak between 94-114 base pairs still sequenced large amounts of adaptor dimers and were not guaranteed to have high microRNA read counts in each sample. In this study, the peak trace provided by the Agilent 2100 Bioanalyzer was not a useful predictor of how well a cDNA library would sequence, in terms of microRNA read counts.

Small RNA-seq of cDNA libraries prepared from Patient Serum and Fresh-Frozen Tissue

Small RNA-sequencing was performed on the Ion Proton Next Generation Sequencing instrument using the Ion PI Hi-Q emulsion PCR and sequencing kits. Up to 14 cDNA libraries were pooled for each small RNA-seq run. The selected cDNA libraries were first
diluted to 5nM. The pooled cDNA libraries were then diluted to a final concentration ranging from 5pM-20pM.

Even though the same concentration of cDNA library was included for each sample in each small RNA-seq run, it was observed that some samples sequenced substantially more microRNA read counts than others on the same Ion PI sequencing chip. In instances where this occurred, the samples that did not sequence well were sequenced on another chip and the pooled library concentration was increased to allow for more overall total reads per chip. Initially, the 5pM pooled library concentration was intended to limit the number of reads sequenced by adapter dimers. However, the percentage of adapter dimer sequenced did not increase dramatically when the pooled library concentrations were increased to 15pM or 20pM (Appendix I).

After sequencing 93 cDNA libraries over 16 small RNA-seq runs, a quality control step, the Ion Torrent smallRNA_analysis plugin, obtained a measure of microRNA reads sequenced for these cDNA libraries (Appendix II). This preliminary analysis aligned and quantitated the sequences for each sample to miRbase v20 [173, 174]. The number of microRNA reads for a cDNA library ranged from 16,829 – 13,538,600 reads (Appendix II). 21 cDNA libraries sequenced more than 1,000,000 microRNA reads, the threshold generally suggested by the manufacturer of the sequencing instrument. These libraries did not include enough of each pathology group to maintain sufficient statistical power in downstream analyses. To address this, the minimum microRNA read count threshold to select cDNA libraries for analysis was reduced to 200,000 microRNA reads. The 55 cDNA libraries that passed this threshold included 12 BE, 10 EAC, 14 GERD, 10 LGD,
and 9 Normal serum or tissue samples (Appendix II). Due to the high variability in microRNA expression within the Normal patient group (Appendix III), the nine Normal samples were excluded from the analysis of the small RNA-seq data. The remaining 46 cDNA libraries comprised of 25 serum samples and 21 tissue samples. Each stage of BE pathogenesis was represented by at least 5 patients (Table 4).

The small RNA sequences from the 46 samples selected from this quality control step were converted to microRNA read counts in Partek Flow®. The small RNA sequences could originate from a variety of small RNA products, not just microRNA products. Therefore, sequences were aligned and quantitated to miRBase v.21 [144, 145]. This provided the raw read counts for all 2,588 known human microRNAs in each patient sample (Table 5).

*Normalization of microRNA read counts by Trimmed Mean of M-values (TMM)*

Total microRNA reads in each patient sample is variable, and this can affect downstream statistical analysis. To ensure all samples are comparable, a normalization method is applied to the microRNA read counts obtained after alignment and quantification to miRBase v.21. The Trimmed Mean of M-values (TMM) normalization method has a precedent for use with RNA sequencing [148-150]. The majority of microRNAs in the 25 patient serum samples and the 21 patient tissue samples had 0-10 read counts (Figure 4). The similarities among the microRNA read count distributions for each sample indicated that most genes are not differentially expressed among the four pathology groups. This is an important observation, as TMM is only an appropriate normalization method for those data where a small fraction of genes are differentially expressed in an experimental group.
Table 4: Composition of Serum and Tissue Sequencing Data Sets

<table>
<thead>
<tr>
<th>Pathology</th>
<th>Serum Samples (n)</th>
<th>Tissue Samples (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERD</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>BE</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>LGD</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>EAC</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td><strong>25</strong></td>
<td><strong>21</strong></td>
</tr>
</tbody>
</table>

1) 20 Serum and Tissue came from the same patients.
<table>
<thead>
<tr>
<th>Sample</th>
<th>Pathology</th>
<th>Total microRNA reads¹</th>
<th>Sample</th>
<th>Pathology</th>
<th>Total microRNA reads¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>13S</td>
<td>EAC</td>
<td>204,003</td>
<td>13T</td>
<td>EAC</td>
<td>449,663</td>
</tr>
<tr>
<td>15S</td>
<td>BE</td>
<td>413,519</td>
<td>15T</td>
<td>BE</td>
<td>345,976</td>
</tr>
<tr>
<td>16S</td>
<td>LGD</td>
<td>183,935</td>
<td>16T</td>
<td>LGD</td>
<td>335,729</td>
</tr>
<tr>
<td>17S</td>
<td>LGD</td>
<td>282,797</td>
<td>17T</td>
<td>LGD</td>
<td>1,026,395</td>
</tr>
<tr>
<td>18S</td>
<td>BE</td>
<td>556,967</td>
<td>1T</td>
<td>BE</td>
<td>300,635</td>
</tr>
<tr>
<td>19S</td>
<td>GERD</td>
<td>223,628</td>
<td>20T</td>
<td>BE</td>
<td>1,400,802</td>
</tr>
<tr>
<td>1S</td>
<td>BE</td>
<td>883,503</td>
<td>21T</td>
<td>GERD</td>
<td>309,603</td>
</tr>
<tr>
<td>20S</td>
<td>BE</td>
<td>1,227,629</td>
<td>23T</td>
<td>GERD</td>
<td>176,626</td>
</tr>
<tr>
<td>21S</td>
<td>GERD</td>
<td>2,420,358</td>
<td>26T</td>
<td>EAC</td>
<td>2,023,514</td>
</tr>
<tr>
<td>23S</td>
<td>GERD</td>
<td>1,277,820</td>
<td>27T</td>
<td>EAC</td>
<td>397,774</td>
</tr>
<tr>
<td>24S</td>
<td>BE</td>
<td>537,818</td>
<td>28T</td>
<td>LGD</td>
<td>1,246,450</td>
</tr>
<tr>
<td>26S</td>
<td>EAC</td>
<td>161,492</td>
<td>29T</td>
<td>BE</td>
<td>165,614</td>
</tr>
<tr>
<td>27S</td>
<td>EAC</td>
<td>135,689</td>
<td>2T</td>
<td>GERD</td>
<td>429,997</td>
</tr>
<tr>
<td>28S</td>
<td>LGD</td>
<td>152,250</td>
<td>32T</td>
<td>GERD</td>
<td>825,618</td>
</tr>
<tr>
<td>29S</td>
<td>BE</td>
<td>553,629</td>
<td>33T</td>
<td>BE</td>
<td>317,549</td>
</tr>
<tr>
<td>2S</td>
<td>GERD</td>
<td>302,725</td>
<td>34T</td>
<td>EAC</td>
<td>411,163</td>
</tr>
<tr>
<td>32S</td>
<td>GERD</td>
<td>525,009</td>
<td>36T</td>
<td>GERD</td>
<td>254,886</td>
</tr>
<tr>
<td>33S</td>
<td>BE</td>
<td>735,174</td>
<td>37T</td>
<td>LGD</td>
<td>2,881,190</td>
</tr>
<tr>
<td>34S</td>
<td>EAC</td>
<td>177,121</td>
<td>38T</td>
<td>EAC</td>
<td>1,977,969</td>
</tr>
<tr>
<td>36S</td>
<td>GERD</td>
<td>1,236,167</td>
<td>39T</td>
<td>LGD</td>
<td>512,297</td>
</tr>
<tr>
<td>37S</td>
<td>LGD</td>
<td>4,239,575</td>
<td>6T</td>
<td>GERD</td>
<td>1,089,611</td>
</tr>
<tr>
<td>38S</td>
<td>EAC</td>
<td>371,094</td>
<td>39S</td>
<td>LGD</td>
<td>1,109,802</td>
</tr>
<tr>
<td>3S</td>
<td>GERD</td>
<td>664,552</td>
<td>7S</td>
<td>GERD</td>
<td>2,969,782</td>
</tr>
</tbody>
</table>

¹Raw microRNA reads generated in Partek Flow
Figure 4: Distribution of Raw MicroRNA Read Counts in Serum and Tissue Samples

Processing of small RNA sequencing and conversion to microRNA read counts was performed in Partek Flow™. Each barplot was generated in Partek Flow™. Each cDNA library sample was coded with a unique color. The y-axis is the number of known human microRNAs which occur in a sample for each read count category (x-axis). (A) Serum dataset (25 samples) following base trimming, alignment to miRbase v21, and quantification to miRbase v21. (B) Tissue dataset (21 samples) following base trimming, alignment to miRbase v21, and quantification to miRbase v21.
TMM equates overall expression of all genes within the libraries, thereby decreasing the rates of false positives when analyzing the data for differential microRNA expression \cite{149}. First, for a given microRNA, a ratio \(x\) between the raw microRNA read counts in each sample to the total number of read counts in that same sample were calculated:

\[
\text{For a given microRNA in a sample: } x_{\text{sample}} = \frac{\text{microRNA read counts}}{\text{total reads}}
\]

Log-fold changes for each microRNA (M-values) compared these ratios between a given sample and a randomly selected reference sample \cite{149}:

\[
M_{\text{microRNA}} = \log_2 \left( \frac{x_{\text{sample}}}{x_{\text{reference}}} \right)
\]

Absolute expression levels (A-values) were also calculated for each microRNA in each sample \cite{149}:

\[
A_{\text{microRNA}} = \frac{1}{2} \log_2 \left( x_{\text{sample}} \times x_{\text{reference}} \right)
\]

Then, the M-values were trimmed by 30% and the A-values by 5% on both ends. The TMM normalization factor applied to the data set is derived from the weighted mean of trimmed M-values \cite{149}. Since some of the 2,588 microRNAs displayed no expression in some of the serum and tissue samples, it was necessary to add an offset value of 1.00 to prevent non-zero errors in the TMM normalization.

All subsequent data analyses were performed on the TMM-normalized microRNA read counts for the patient serum and patient tissue samples.
Applying a Minimum Normalized microRNA Read Count Filter to Serum and Tissue Datasets

To ensure the candidate microRNAs revealed by these analyses were selected based on sufficient statistical power, a filter based on normalized microRNA read counts was applied to the serum and tissue datasets. For a given microRNA, 25% of the patient samples needed to contain 100 or more normalized microRNA read counts. microRNAs with 100 or more normalized read counts would be more likely to be detected by qPCR during validation experiments [151]. This filter removed any microRNAs that had low expression in every pathology group. Allowing for 25% of the samples to exhibit low normalized read counts enabled capture of any candidate microRNAs which might be lowly expressed in only one pathology group. For the serum and tissue datasets, the normalized 100 microRNA read count filter removed all but 187 and 272 microRNAs, respectively.

Identification of Candidate microRNA Biomarkers for BE Pathogenesis in Patient Serum and Tissues

The filtered TMM-normalized microRNA read counts for the serum and tissue data sets were analyzed separately by three different methods: unsupervised hierarchical clustering, Nearest Shrunken Centroids classification, and the randomForest classifier. The objective of these analyses was to reveal any potential microRNA signatures in the serum or tissue datasets which could differentiate all four stages of BE pathogenesis.
Hierarchical Clustering

In hierarchical clustering, samples are clustered based on the extent of their similarities to each other. Similarities are determined by the Euclidean distances between microRNA read counts for each sample [175]. Clustering was based on the average Euclidean distances for all data points between two clusters, a method called average linkage [176]. Hierarchical clustering for this study was performed in Orange Dating Mining Tool Box v. 3.20.1 and visualized as a dendrogram [152]. For the serum dataset, all 187 filter microRNAs were included in four distinct clusters (Figure 5A). In terms of pathology, the four clusters of serum samples were very heterogeneous and a relationship between BE pathogenesis and microRNA expression was not apparent. When hierarchical clustering is applied to the 25 tissue samples, two larger clusters can be observed (Figure 5B). However, these clusters are also heterogeneous in terms of pathology, and do not include six of the tissue samples. These six tissue samples consist of 1 GERD, 3 LGD, and 2 EAC. It appears that normalized microRNA read counts cannot be used to group tissue samples based on pathology, according to this hierarchical clustering analysis.

Visualization of the serum and tissue data sets by heat mapping demonstrates the difficulty of identifying microRNA signatures from 187 microRNAs or 272 microRNAs, respectively (Figure 6). In the heat maps, k-means clustering separates the microRNAs into a pre-determined number of clusters, based on proximity to calculated centroids [177, 178]. The centroid of a cluster is a point where the Euclidean distances among the data points are minimized. If a microRNA signature existed in either the serum or tissue
Figure 5: Hierarchical Clustering of Serum and Tissue Samples
Hierarchical Clustering was performed and visualized in Orange Data Mining version 3.20.1. Metrics used: Euclidean Distances and Average Linkage. (A) 25 serum samples clustered based on normalized microRNA read count data for 187 filtered microRNAs. (B) 21 tissue samples clustered based on normalized microRNA read count data for 272 filtered microRNAs.
A

hsa-let-7b-3p, hsa-let-7c-5p (158 more)
hsa-miR-451a
hsa-miR-223-3p
hsa-miR-19a-3p
hsa-miR-144-3p
hsa-miR-486-5p
hsa-let-7a-5p, hsa-miR-106b-5p (7 more)
hsa-miR-19b-3p
hsa-let-7g-5p (13 more)
hsa-miR-222-3p

B

hsa-let-7b-3p, hsa-let-7b-5p (217 more)
hsa-miR-451a
hsa-miR-145-5p
hsa-miR-203a-3p, hsa-miR-205-5p
hsa-let-7a-5p
hsa-miR-148a-3p, hsa-miR-23c (2 more)
hsa-let-7c-5p, hsa-let-7g-5p (40 more)
hsa-miR-223-3p
hsa-miR-19a-3p, hsa-miR-19b-3p (2 more)
hsa-miR-21-5p
Figure 6: Heat Maps of Serum and Tissue Samples
Heat Maps were constructed in Orange Data Mining version 3.20.1. k-means clustering (10 clusters) was used to group microRNAs. (A) 25 serum samples (x-axis) vs normalized microRNA read count data for 187 filtered microRNAs (y-axis). (B) 21 tissue samples (x-axis) vs normalized microRNA read count data for 272 filtered microRNAs (y-axis).
data sets, it was expected that a cluster of microRNAs would display distinct differential expression in the heat map among the four stages of BE pathogenesis.

*Nearest Shrunken Centroids*

A classifier termed Nearest Shrunken Centroids (NSC) was used as an alternative to hierarchical clustering. NSC has been utilized previously to identify several genes which differentiate multiple types of round blue cell tumors [156]. NSC is advantageous in analyzing data with numerous variables and a significantly smaller number of samples, which described the serum and tissue small RNA-seq datasets.

NSC was implemented through the PAMR package available for R. For each microRNA in each sample, a centroid was calculated for each pathology group by dividing the average microRNA expression by its standard deviation. NSC is a supervised classifier, in that the groups must be defined beforehand. Then for each pathology group, the centroids were “shrunk” towards zero by a certain threshold value [156]. This threshold value is associated with a misclassification error percentage. Each microRNA was given a score based on their distance from the centroid after shrinkage. NSC outputs a combination, or signature, of microRNAs which can differentiate the predefined pathology groups from each other. If a microRNA possessed a non-zero score for a pathology group, that microRNA contributed to the signature’s ability to differentiate that pathology group from the other three pathology groups.

When NSC was applied to the serum dataset, 185 of the 187 microRNAs showed non-zero scores for at least one pathology (Figure 7). This indicates that these 185
Figure 7: Serum microRNAs cannot distinguish among Four Stages of BE Pathogenesis

Nearest Shrunken Centroids (PAMR package in R) analysis applied to 25 Serum samples, based on TMM-normalized read counts for the 187 filtered microRNAs. A representative sample of 185 microRNAs are shown here. These 185 microRNAs were associated with the least number of misclassification errors for each pathology group in the NSC analysis and an overall misclassification rate of 44.6%.
microRNAs could operate together as a microRNA signature to differentiate every pathology group in BE pathogenesis. However, the overall misclassification rate for these microRNAs was 44.6%. The misclassification rates for each pathology group ranged from 20% - 80%. While many serum microRNAs may be differentially expressed in BE pathogenesis according to the NSC analysis, the misclassification rate is too high for these microRNAs to function as a signature for every stage of BE pathogenesis.

In the tissue dataset, 23 of the 272 microRNAs have non-zero scores for one or more of the four pathology groups (Figure 8). The overall misclassification rate for this analysis was 47.5%. Most microRNAs in this potential tissue microRNA signature appear to be specific to LGD or EAC. However, the misclassification rates associated with these pathology groups was 40% and 80%, respectively. One microRNA had a non-zero score for the GERD pathology group, miR-126-5p. No microRNAs were able to differentiate the BE group from the other pathology groups, Therefore, this 23 microRNA signature could not function as a biomarker panel for BE pathogenesis.

**randomForest**

A decision tree can separate tissue or serum samples into the pathology groups based on the TMM-normalized microRNA read count data (Figure 9). A single decision is prone to over-fitting [179]. randomForest is a supervised machine learning method which samples with replacement from the dataset to build a “forest” of decision trees [154, 155]. Two different randomForest techniques were utilized in this study: randomForest classification and randomForest regression. The classification randomForest separates the samples into previously assigned groups using 1 million decision trees created from the microRNA
hsa-miR-181a-5p
hsa-miR-214-3p
hsa-miR-19b-3p
hsa-miR-199a-5p
hsa-miR-15a-5p
hsa-miR-130a-3p
hsa-miR-18b-5p
hsa-miR-19a-3p
hsa-miR-18a-5p
hsa-miR-139-5p
hsa-miR-92a-3p
hsa-miR-199b-5p
hsa-miR-126-3p
hsa-miR-17-5p
hsa-miR-106b-5p
  hsa-let-7c-5p
  hsa-miR-93-5p
hsa-miR-20a-5p
hsa-miR-192-5p
hsa-miR-142-3p
hsa-miR-223-3p
hsa-miR-126-5p
hsa-miR-29b-3p

Scores: Shrunken Distances to Centroids

-0.2 0 0.2 0.4 0.6 0.8 1

LGD-T-score
GERD-T-score
EAC-T-score
BE-T-score
Figure 8: Tissue microRNAs cannot distinguish among Four Stages of BE Pathogenesis

Nearest Shrunken Centroids (PAMR package in R) analysis applied to 21 Tissue samples, based on TMM-normalized read counts for the 272 filtered microRNAs. These 23 microRNAs were associated with the least number of misclassification errors for each pathology group in the NSC analysis and an overall misclassification rate of 47.5%.
Figure 9: Decision Trees Can Separate Tissues into the Four Pathology Groups

Representative decision tree was generated in Orange Data Mining version 3.20.1. from TMM-normalized read counts for six select microRNAs from the Tissue dataset (21 total samples).
read count data. The microRNAs are ranked by mean decrease accuracy, which indicates the importance of a microRNA in constructing the randomForest. A high mean decrease accuracy occurs if the accuracy of the model decreases when the microRNA is removed from the randomForest. A measure of predictive error for this analysis is the out of bag (OOB) estimate of error rate.

RandomForest regression includes an extra component of sample order. Since BE pathogenesis is a step-wise malignant progression, the pathology groups were numbered sequentially with GERD being 1 and EAC being 4. Instead of Mean Decrease Accuracy, a microRNA’s importance in the RandomForest regression is indicated by the percentage the mean squared error (%IncMSE) of the predictions increases as a result of removing the microRNA. There is not an equivalent metric of OOB estimate of error rate for the RandomForest regression

RandomForest classification

As can be observed in Figure 10A, miR-190a-5p is considered the most important microRNA in classifying the 25 serum samples into the four pathology groups. The 64% OOB estimate of error rate for the serum RandomForest classification was not due to a misclassification of any particular group. All the pathology groups had at least 50% of their samples misclassified into the incorrect pathology group. Only LGD had a 100% misclassification rate; all 5 samples misclassified as either EAC, BE, or GERD.

miR-196b-5p and miR-196a-5p are the top two most important microRNAs for classifying the four pathology groups, according to the tissue RandomForest classification
analysis (Figure 10B). Here, the OOB estimate of error rate was 67%. All 5 EAC tissue samples misclassified as either BE or LGD, and four out of six GERD samples misclassified as either EAC, LGD, or BE. As in the serum analysis, no single pathology group contributed to the large OOB estimate of error rate.

Since randomForest separates data into groups using decision trees, it will tend towards a focus on all variability in the data, rather than just the differences between pathology groups. This is especially true when less than 5 groups are assigned in a dataset. This may account for why even the most important microRNAs from the serum or tissue datasets were not good classifiers of BE pathogenesis.

**randomForest regression**

In the serum analysis, miR-190a-5p, miR-194-5p, and miR-106b-3p had the highest %IncMSE values. These microRNAs appeared to be the most important microRNAs in accurately predicting a sample’s pathology and its order in BE pathogenesis (Figure 11A). In tissue, miR-196b-5p, miR-223-3p, and miR-196a-5p rank most highly, according to their %IncMSE values (Figure 11B).

**Conclusions**

It was clear from the hierarchical clustering and NSC analyses that a microRNA signature for BE pathogenesis was not detectable in the serum data sets. The NSC analysis of the tissue data set did provide a large potential signature of microRNAs which could differentiate all four stages of BE pathogenesis. However, this was associated with high misclassification rates for each pathology group. While several serum and tissue
Figure 10: randomForest classification on Serum and Tissue datasets

randomForest classification was performed in the randomForest package available for R. (A) 187 Filtered Serum Samples. (B) 272 Filtered Tissue Samples
randomForest regression was performed in the randomForest package available for R. (A) 187 Filtered Serum Samples. (B) 272 Filtered Tissue Samples
microRNAs ranked as the most important classifiers for BE pathogenesis in the randomForest analyses, misclassification rates for each pathology were still not ideal. For this study, a microRNA signature for BE pathogenesis was not readily apparent in either the serum or tissue samples when analyzed by hierarchical clustering, NSC, or randomForest.

**Gene Specific Analysis Identifies Single Candidate microRNAs through Pair-Wise Comparisons**

Gene Specific Analysis (GSA) in Partek Flow allows for comparisons between two groups in a data set and generates a p-value for each of the 2,588 microRNAs aligned to miRBase v21, based on a selected statistical model. This GSA tested differential expression between BE vs GERD, LGD vs BE, EAC vs BE, and EAC vs LGD. The EAC vs BE comparison was included as BE patients can progress quickly through the LGD step to develop EAC.

Since GSA entails multiple comparisons, a Bonferroni correction is necessary to establish a limit to Type I statistical errors [180]. This correction entails dividing an $\alpha$ value by the number of pair-wise comparisons. The $\alpha$ describes the permissible false positive rate, while the number of pair-wise comparisons are those microRNAs which passed the minimum read count filter.

*Serum*

When a false positive rate of 5% was permitted ($\alpha = 0.05$), none of the microRNAs passed the Bonferroni threshold applied to the GSA results. When $\alpha$ was increased to 0.1,
only miR-194-5p showed significant downregulation in serum of EAC patients in comparison to BE patients (Table 6). This indicates that differential microRNA expression among the four stages of BE pathogenesis cannot be detected in the 24 patient serum samples. miR-194-5p is in the top four and the top two most important microRNAs in the serum randomForest classification and regression analyses, respectively. miR-194-5p has previously been found to have increased expression in tissue and serum of BE patients [94].

Tissue

Of the 272 microRNAs, only three passed the Bonferroni threshold with $\alpha$ set to 0.05 (Table 6). Three additional microRNAs were included when $\alpha$ was reduced to 0.1 (Table 6). miR-196a-5p, miR-196b-5p, and miR-215-3p were all upregulated in BE patient tissues when compared to the 6 GERD patient tissues. miR-596 showed downregulation in EAC patient tissues, in comparison to LGD. However, miR-596 did not pass the Bonferroni threshold in the BE vs EAC comparison. Expression of miR-223-3p was 15-fold higher in the EAC vs BE patients. miR-4655-3p was downregulated in EAC vs BE. Yet, neither miR-223-3p nor miR-4655-3p were found to be significantly differentially expressed in either the BE vs LGD or EAC vs BE comparisons. In previous studies, miR-196a-5p, miR-196b, miR-215, and miR-223 have all been shown to be upregulated in BE tissues [94].
Table 6: Seven microRNAs display Differential Expression in Stages of BE pathogenesis

<table>
<thead>
<tr>
<th>α</th>
<th>microRNA</th>
<th>Adjusted p-value</th>
<th>Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum (BE vs EAC)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>miR-194-5p</td>
<td>0.05928</td>
<td>-6.74</td>
</tr>
<tr>
<td>Tissue (BE vs GERD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>miR-196b-5p</td>
<td>0.00048</td>
<td>23.20</td>
</tr>
<tr>
<td>0.05</td>
<td>miR-196a-5p</td>
<td>0.01994</td>
<td>18.30</td>
</tr>
<tr>
<td>0.1</td>
<td>miR-215-3p</td>
<td>0.05930</td>
<td>37.90</td>
</tr>
<tr>
<td>Tissue (LGD vs BE)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>None Passed Threshold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tissue (EAC vs BE)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>miR-223-3p</td>
<td>0.00588</td>
<td>15.00</td>
</tr>
<tr>
<td>0.1</td>
<td>miR-4655-3p</td>
<td>0.05930</td>
<td>-15.10</td>
</tr>
<tr>
<td>Tissue (EAC vs LGD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>miR-596</td>
<td>0.06174</td>
<td>-13.00</td>
</tr>
</tbody>
</table>

1 p-values Bonferroni-corrected for multiple comparisons.
**randomForest analysis of BE and GERD Tissue Samples support Findings from Gene Specific Analysis**

miR-196a-5p and miR-196b-5p were identified as important microRNAs for overall classification of the tissue samples into the four pathology groups, according to both randomForest analyses (Figures 10 and 11). miR-223-3p, miR-4655-3p and miR-596 also appear to be the top 20 most important microRNAs in the tissue randomForest analyses.

To determine how well the microRNA candidates from the GSA classify the samples into their corresponding pathology groups, miR-196a-5p, miR-196b-5p, and miR-215-3p were analyzed again by randomForest classification. Instead of including all four pathology groups, only BE and GERD tissues were included in this follow-up analysis. miR-196a-5p and miR-196b-5p ranked first and fourth among the 272 microRNAs, indicating their ability to differentiate BE and GERD. When the input data is limited to only the three microRNAs of interest, 1 out of 5 BE samples and 0 of 6 GERD samples were misclassified. The overall error rate for these microRNAs was 9.09%. As only 5-6 patient samples are included in each pathology group, the randomForest analyses were not particularly robust validation methods for the GSA results. Analysis of only these three microRNAs in the BE and GERD groups by hierarchical clustering or heat mapping did not reveal clustering based on pathology. The randomForest classifier does lend further support to the candidacy of these microRNAs as biomarkers. Since the small sample size of the small RNA-seq study prevents validation by *in silico* methods, a qPCR study on a larger sample set is necessary to validate the candidate microRNA biomarkers.
Validation of Small RNA-Sequencing Results by Quantitative real-time PCR

Preparation of cDNA libraries for small RNA-seq and the sequencing technology itself can introduce technical biases to data [181]. Therefore, it is necessary to confirm small RNA-sequencing results by another technique [182]. qPCR is a technique renowned for its accuracy and specificity in RNA expression [183]. It is considered the obligatory technique for validation experiments for small RNA-sequencing studies.

The three BE vs GERD microRNA candidates were chosen for qPCR validation in this study. miR-196b-5p, miR-196a-5p, and miR-215-3p have the lowest p-values of all the microRNAs which passed either of the Bonferroni correction thresholds. All three of these microRNAs are significantly upregulated in the BE patient tissues when compared to the GERD patient tissues (Table 6). GERD is estimated to affect 20% patients in the United States [184]. Since BE lesions do not present additional symptoms, the actual number of GERD patients with BE is unknown. GERD patients with undiagnosed BE still have the same heightened risks of dysplasia and EAC but do not undergo surveillance. The three BE vs GERD candidates could become confirmatory biomarkers for BE in GERD patients.

Since the small RNA-seq study included a small number of human patient samples, the variance in microRNA read counts, even in the same pathology group, could be large. For example, the variance among microRNA read counts in the BE patients was noticeably larger than the variance in the GERD patients, especially for miR-215-3p (Figure 12). The heightened variability in the BE tissues for miR-215-3p may account for why this microRNA did not pass the Bonferroni threshold when $\alpha = 0.05$. 

77
miR-196a-5p
miR-196b-5p
miR-215-3p

TMM-Normalized microRNA read counts

GERD.
BE.

GERD.
BE.

GERD.
BE.

GERD.
BE.
Figure 12: Three microRNAs show significant increased expression in BE Tissues vs GERD Tissues

TMM-normalized microRNA reads counts obtained by small RNA-seq for the three BE vs GERD candidate microRNAs show increased expression and larger variance in the BE patient tissues.
Power Analyses of Tissue microRNA Candidates

A separate cohort of formalin-fixed paraffin-embedded (FFPE) tissues from 13 GERD and 13 BE patients was utilized for the qPCR validation studies. The sample number (n) of these studies was selected based on power analyses performed for each of the three candidate microRNAs (Table 7). These analyses were done using the GERD vs. BE microRNA read counts obtained from the small RNA sequencing libraries (Table 7). Though miR-196a-5p and miR-196b-5p required eight and seven samples to achieve 80% power, 13 samples per pathology group were needed to achieve sufficient statistical power for miR-215-3p.

Total RNA Extraction from Formalin-Fixed Paraffin-Embedded Tissues required Optimization

Total RNA was extracted from the FFPE tissues by the Covaris truXTRAC FFPE RNA microtube kit which utilizes the Covaris M220 Acoustic Focused ultra-sonicator. Initial total RNA extractions contained guanidine ITC contamination, noted upon assessment by NanoDrop ONE. Guanidine salts are a component of the Covaris kit necessary for the column chemistry. Total RNA was re-precipitated from these samples using a sodium hydroxide/ethanol precipitation protocol. Subsequent total extractions included a third wash with the Covaris ethanol wash buffer, which appeared to resolve the guanidine salt contamination. Total RNA concentration was obtained by NanoDrop ONE (Table 8).
Table 7: Power analysis for qPCR validation

<table>
<thead>
<tr>
<th>microRNAs¹</th>
<th>TMM-Normalized microRNA Read Counts</th>
<th>Samples Required²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GERD (mean)</td>
<td>BE (mean)</td>
</tr>
<tr>
<td>miR-196a-5p</td>
<td>26.43</td>
<td>484.09</td>
</tr>
<tr>
<td>miR-196b-5p</td>
<td>18.89</td>
<td>438.03</td>
</tr>
<tr>
<td>miR-215-3p</td>
<td>12.77</td>
<td>483.61</td>
</tr>
</tbody>
</table>

¹ MicroRNAs identified as differentially regulated in the GERD vs. BE comparison in tissue samples based on small RNA-seq data.

² Minimum number of samples required to obtain power of 0.8 with an $\sigma$ of 0.05. Power analyses based on TMM-normalized microRNA read counts from tissue small RNA-seq data.
Table 8: Total RNA Extracted from Formalin Fixed Paraffin Embedded Tissues

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Pathology</th>
<th>Total RNA (ng)</th>
<th>Absorbance 260/280</th>
<th>Absorbance 260/230</th>
</tr>
</thead>
<tbody>
<tr>
<td>3524B</td>
<td>BE</td>
<td>1,152</td>
<td>1.83</td>
<td>1.03</td>
</tr>
<tr>
<td>2925B</td>
<td>BE</td>
<td>826</td>
<td>1.83</td>
<td>1.41</td>
</tr>
<tr>
<td>2427F</td>
<td>BE</td>
<td>1,082</td>
<td>1.86</td>
<td>1.59</td>
</tr>
<tr>
<td>2401B</td>
<td>BE</td>
<td>1,344</td>
<td>1.88</td>
<td>1.73</td>
</tr>
<tr>
<td>505E</td>
<td>BE</td>
<td>2,763</td>
<td>1.93</td>
<td>1.94</td>
</tr>
<tr>
<td>4176B</td>
<td>BE</td>
<td>1,050</td>
<td>1.90</td>
<td>1.26</td>
</tr>
<tr>
<td>3983D</td>
<td>BE</td>
<td>1,180</td>
<td>1.87</td>
<td>1.68</td>
</tr>
<tr>
<td>3694C</td>
<td>BE</td>
<td>385</td>
<td>1.71</td>
<td>1.98</td>
</tr>
<tr>
<td>766E</td>
<td>BE</td>
<td>749</td>
<td>1.82</td>
<td>1.29</td>
</tr>
<tr>
<td>5138C</td>
<td>BE</td>
<td>1,306</td>
<td>1.86</td>
<td>1.53</td>
</tr>
<tr>
<td>1507F</td>
<td>BE</td>
<td>1,054</td>
<td>1.90</td>
<td>1.79</td>
</tr>
<tr>
<td>2244C</td>
<td>BE</td>
<td>1,985</td>
<td>1.93</td>
<td>1.78</td>
</tr>
<tr>
<td>1689G</td>
<td>BE</td>
<td>1,750</td>
<td>1.90</td>
<td>0.57</td>
</tr>
<tr>
<td>1823H</td>
<td>BE</td>
<td>529</td>
<td>1.72</td>
<td>2.16</td>
</tr>
<tr>
<td>1976B</td>
<td>BE</td>
<td>1,432</td>
<td>1.90</td>
<td>1.64</td>
</tr>
<tr>
<td>2374C</td>
<td>GERD</td>
<td>2,114</td>
<td>1.82</td>
<td>2.16</td>
</tr>
<tr>
<td>4268E</td>
<td>GERD</td>
<td>977</td>
<td>1.69</td>
<td>2.25</td>
</tr>
<tr>
<td>2493-</td>
<td>GERD</td>
<td>830</td>
<td>1.76</td>
<td>2.05</td>
</tr>
<tr>
<td>1849B</td>
<td>GERD</td>
<td>1,050</td>
<td>1.65</td>
<td>2.09</td>
</tr>
<tr>
<td>2870A</td>
<td>GERD</td>
<td>221</td>
<td>1.64</td>
<td>1.67</td>
</tr>
<tr>
<td>2024A</td>
<td>GERD</td>
<td>1,428</td>
<td>1.65</td>
<td>1.95</td>
</tr>
<tr>
<td>5791B</td>
<td>GERD</td>
<td>1,876</td>
<td>1.67</td>
<td>1.22</td>
</tr>
<tr>
<td>5281D</td>
<td>GERD</td>
<td>567</td>
<td>1.60</td>
<td>1.23</td>
</tr>
<tr>
<td>1927B</td>
<td>GERD</td>
<td>1,687</td>
<td>1.81</td>
<td>1.54</td>
</tr>
<tr>
<td>1923E</td>
<td>GERD</td>
<td>1,316</td>
<td>1.70</td>
<td>1.35</td>
</tr>
<tr>
<td>1833A</td>
<td>GERD</td>
<td>2,027</td>
<td>1.85</td>
<td>1.74</td>
</tr>
<tr>
<td>6302D</td>
<td>GERD</td>
<td>872</td>
<td>1.74</td>
<td>1.36</td>
</tr>
<tr>
<td>5751A</td>
<td>GERD</td>
<td>557</td>
<td>1.64</td>
<td>1.06</td>
</tr>
<tr>
<td>5307C</td>
<td>GERD</td>
<td>532</td>
<td>1.67</td>
<td>0.90</td>
</tr>
</tbody>
</table>

1 Extracted by Covaris TruXTRAC FFPE RNA microTUBE kit (Covaris M220 Focused Ultrasonicator)

2 Measured by NanoDrop ONE spectrophotometer
Quantitative real-time PCR for 13 GERD and 13 BE Tissue Samples

cDNA for the qPCR experiments was synthesized from 10ng total RNA by the TaqMan Advanced microRNA cDNA synthesis kit. Since no quality control step is included to assess cDNA quantity and quality prior to qPCR, the 10ng starting amount is necessary to ensure a common baseline among the 26 samples.

In this study, each microRNA was assessed in triplicate for each BE and GERD sample. Ct values for the biological replicates were the average of the technical replicates. Since the analysis utilized, the ΔΔCt method, is a relative quantitative method, an endogenous control microRNA was chosen to normalize the data from the three experimental microRNAs. miR-423-5p did not show significant differential expression between the GERD and BE patient tissues in the small RNA-seq experiments (Figure 13). In the qPCR experiments, miR-423-5p expression significantly differed in the BE FFPE tissues when compared to the GERD FFPE tissues. However, this difference was in the same direction as the expression difference for the three experimental microRNAs (Figure 14). While this indicated that miR-423-5p was not a perfect control microRNA for qPCR in distal esophageal tissues, its effect on the overall results of the ΔΔCt calculations appeared to have been minimal.

After the average Ct values for each experimental microRNA are normalized to miR-423-5p, a Student’s T-test was performed on the ΔCt values. It should be noted that the lower the Ct value, the more abundant and more highly expressed the microRNA in the tissue sample. As shown in Figure 15, miR-196b-5p showed the most significant difference between GERD and BE tissues, followed closely by miR-215-3p. In the small RNA-seq
miR-423-5p: Tissue small RNA-seq

TMM-Normalized microRNA read counts (Tissue)

GERD.T
BET
LGD.T
EACT.

p-value = 0.8919
Figure 13: miR-423-5p is not significantly differentially expressed between GERD and BE tissues

TMM-normalized microRNA reads counts for miR-423-5p were obtained by small RNA-seq. A Student’s t-test performed between GERD patient tissues and the BE patient tissues for miR-423-5p produced a p-value of 0.8919. Box plot and Student’s T-test were generated in R.
Figure 14: miR-423-5p does not display constant expression between GERD and BE FFPE tissues

Average Ct values for miR-423-5p were obtained by qPCR. Student’s t-tests performed between GERD patient tissues and the BE patient tissues for miR-423-5p and three experimental microRNAs produced the following p-values: *: p-value = 4.87 x 10^{-2} **: p-value = 6.84 x 10^{-3} ***: p-value = 6.01 x 10^{-6} ****: p-value = 3.42 x 10^{-6}
Figure 15: miR-196a-5p, miR-196b-5p, and miR-215-3p all display significantly increased expression in BE tissues in comparison to GERD Tissues

Average Ct values for each microRNA were obtained by qPCR. ΔCt values for each microRNA were normalized to miR-423-5p. A Student’s t-test was used to calculate p-values for pair-wise comparisons between GERD patient tissues and the BE patient tissues.
and randomForest results, miR-196b-5p was also the most significantly changed between GERD and BE, thus this microRNA is a compelling candidate tissue biomarker for BE in GERD patients.

Expression fold change ($2^{-\Delta\Delta C_t}$) was calculated for each microRNA. As expected, all were upregulated in BE vs. GERD (Table 9). The significant changes in the expression levels of these microRNAs indicate that miR-196b-5p, miR-215-3p, and miR-196α-5p are potential tissue biomarkers for BE in GERD patients.

**Gene Target Analysis Reveals Potential Regulation of Homeotic Genes by Candidate microRNAs**

MicroRNAs can be utilized as clinical biomarkers to diagnose disease; however, they also can be utilized to increase understanding of a disease’s molecular mechanisms. For each candidate microRNA, lists of predicted gene targets from TargetScan Human 7.2 and miRDB were compared to experimentally-validated targets found through TarBase v8 and miRNet [145, 162, 165-167, 185]. Preference was showed for those predicted gene targets which were also experimentally-validated. Additional insight into the impact of the candidate microRNAs on the molecular mechanisms behind BE development was gained through PubMed.gov literature search.

*BE vs GERD candidate microRNAs: miR-196a-5p, miR-196b-5p, and miR-215-3p*

miR-196a-5p and miR-196b-5p show similar but not exact microRNA read counts in the 5 BE and 6 GERD patient tissue samples. The sequences for these two mature microRNAs differ by only a single nucleotide [186]. However, the two genes, MIR-196a-
Table 9: Change in Expression of selected microRNAs in BE vs. GERD FFPE tissues.

<table>
<thead>
<tr>
<th>microRNA</th>
<th>ΔΔCt&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Expression Fold Change&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-196a-5p</td>
<td>-2.49</td>
<td>+5.62</td>
</tr>
<tr>
<td>miR-196b-5p</td>
<td>-4.92</td>
<td>+30.33</td>
</tr>
<tr>
<td>miR-215-3p</td>
<td>-4.38</td>
<td>+20.79</td>
</tr>
</tbody>
</table>

<sup>1</sup> ΔΔCt = ΔCt<sub>BE</sub> – ΔCt<sub>GERD</sub>

<sup>2</sup> Expression Fold Change = 2<sup>ΔΔCt</sup>
1 and MIR-196a-2, which can generate miR-196a-5p are located on Chromosome 12 and 17. miR-196b-5p is a product of the MIR-196b gene on Chromosome 7.

The gene target analysis for both microRNAs showed midcluster and posterior HOX genes from the A, B, and C clusters as common targets (Table 10). miR-196a-1 and miR-196b-1 are located intergenic to the HOXB and HOXC clusters, respectively [186]. The genes are both located in between the HOX9 and HOX10 genes. Due to the collinearity of HOX genes, their position could indicate that the microRNAs are transcribed when posterior HOXB and HOXC genes are activated. miR-196b is located intergenic to the posterior genes of the HOXA cluster.

Gene targets for miR-215-3p have not been well-studied. Both target prediction algorithms did not return results for miR-215-3p. Investigations into TarBase v8 and miRNet only produced one target, FOXO1, which was validated to be a target of miR-215-3p by luciferase reporter assay [187]. FOXO1 is also an experimentally-validated target of miR-196a-5p [188]. An additional literature search revealed another target validated by luciferase reporter assay: BMI1 [189]. BMI is considered a catalytic subunit of the Polycomb Repressive Complex 1 (PRC1), a well-known silencer of HOX and other developmental genes [190].

**EAC vs BE candidate microRNAs: miR-223-3p and miR-4655-3p**

miR-223-3p is a microRNA known to be dysregulated in numerous cancers, including hepatocellular carcinoma, leukemia, lymphoma, and EAC [191-196]. The role of miR-223-3p in cancer seems to vary depending on tissue type. In hepatocellular carcinoma,
Table 10: Selected Gene Targets for the Three Candidate Tissue microRNA Biomarkers

<table>
<thead>
<tr>
<th>miR-196a-5p targets</th>
<th>miR-196b-5p targets</th>
<th>miR-215-3p targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 3' UTR seeds</td>
<td>Gene 3' UTR seeds</td>
<td>Gene 3' UTR seeds</td>
</tr>
<tr>
<td>HOXB7(^{1,2,4})</td>
<td>HOXB7(^{1,2,4})</td>
<td>FOXO1(^{2,4})</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HOXC8(^{1,2,3,4})</td>
<td>RDX(^{1,4})</td>
<td>BMI1(^{5})</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>(\delta)</td>
</tr>
<tr>
<td>HOXA7(^{1,2,4})</td>
<td>HOXC8(^{1,2,4})</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>HOXB8(^{2,4})</td>
<td>HOXA9(^{1,2,3,4})</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>FOXO1(^{3,4})</td>
<td>FAS(^{3,4})</td>
<td></td>
</tr>
<tr>
<td>(\delta)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HOXA5(^{3,4})</td>
<td>HOXA7(^{1,2})</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>HOXA9(^{1})</td>
<td>HOXA5(^{1,2})</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HOXB6(^{1})</td>
<td>HOXB6(^{1,2})</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HOXB1(^{2})</td>
<td>HOXB8(^{2})</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HOXB1(^{2})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\(^1\)miRDB

\(^2\)TargetScan Human 7.2

\(^3\)DIANA TarBase v8

\(^4\)miRNet

\(^5\) Jones et al., 2015 [186]

\(\delta\) Partial seeds or no seeds were indicated
miR-223-3p is downregulated and is thought to take on a tumor suppressive role [191]. Streppel et al (2013) found that this microRNA displays a step-wise increase in BE pathogenesis [196]. In that same study, overexpression of miR-223-3p in EAC cell lines promotes migration and invasion. In the small RNA-seq data, miR-223-3p showed significantly increased expression in the 5 EAC patients, in comparison to the 5 BE patients (Table 6). However, significant differential expression was not observed between LGD and BE or EAC and LGD.

miR-223-3p has hundreds of predicted and experimentally-validated gene targets. Input of its gene targets found via the four methods into DAVID 6.8 Functional Annotation tool indicated these genes play roles in FOXO, AMPK, and PI3K-Akt signaling [197]. miR-223-3p does not appear to directly target HOX genes from any of the four clusters.

miR-4655-3p is not a well-studied microRNA. None of its predicted targets from miRDB or TargetScan Human 7.2 have been experimentally validated. Two studies have shown that miR-4655-3p is differentially expressed in colorectal and triple negative breast cancer cell lines [198, 199]. This may be due to the location of miR-4655, the gene encoding miR-4655-3p, on chromosome 7. This gene is intragenic to the mitotic arrest deficient 1 like 1 [200], the human homolog for the Drosophila MAD1 gene [200, 201]. MAD1L1 is a mitotic check point whose function is affected by chromatin instability in cancer [200]. In colon cancer cell lines, this gene has been found to promote resistance to doxorubicin [202]. The downregulation of miR-4655-3p in EAC may be due to decreased transcription of the MAD1L1 gene in those tissues.
**EAC vs LGD candidate microRNAs: miR-596**

The gene encoding miR-596 is in a portion of Chromosome 8 which is associated with breakage in cancer [203, 204]. According to the initial gene target analysis, the only miR-596 gene target validated by luciferase reporter assay has been LGALS3BP, a secreted galectin-3 ligand [205]. Increased expression of this ligand is associated with poor prognosis in cancers [205]. LGALS3BP is thought to activate components of ERK1/2 signaling, thereby promoting cell proliferation and inhibiting apoptosis [205]. A further literature review concerning miR-596 procured another experimentally-validated target: Smurf1 [206]. Smurf1 is a known negative regulator of p53, as it binds to and stabilizes MDM2, a prominent p53 inhibitor [206]. The downregulation of miR-596 in EAC patient tissues would lead to released MDM2 inhibition of p53 and increased expression of LGALS3BP and its action on ERK1/2 signaling.
DISCUSSION

Limitations in Identifying Biomarkers for BE pathogenesis

Identification of microRNA biomarkers in serum and tissue for clinical use faces numerous challenges as there is no standardized methods for RNA extraction, cDNA synthesis, or statistical analyses of differential microRNA expression [171, 207]. In this study, no combination of microRNA biomarkers was found in the analysis of the small RNA-seq data from either patient tissue or serum. This may be for multiple reasons.

**Serum**

Only one serum microRNA, miR-194-5p, could differentiate between two of the four pathology groups in this study. miR-194-5p was significantly decreased in the serum of EAC patients when compared to BE patients (Table 6). miR-194-5p was an important microRNA in the classification and regression randomForest analyses (Figures 10 and 11) for the serum data set, indicating it may differentiate all four pathology groups from each other. However, in the GSA for the serum data set, its expression was not significantly different between the LGD and EAC groups or the BE and LGD groups. miR-194-5p has been identified as an upregulated microRNA in tissue and serum of BE patients in numerous previous studies [94].

Circulating microRNAs from serum or plasma have been used as diagnostic biomarkers for many diseases and conditions, such as adult-onset Still’s disease, ectopic pregnancy,
and hepatocellular carcinoma [208-210]. Success in identifying these microRNAs as biomarkers may be based on the characteristics of the conditions studied. Some of these conditions, such as adult-onset Still’s disease and ectopic pregnancy, are systemic. Changes which occur throughout the body, rather than in a localized region, would be more likely to generate a unique microRNA signature which could be associated with that condition. Cancers like hepatocellular carcinoma can be localized to a single tissue type. However, tumor-derived exosomes have been shown to provide a unique signature in the bloodstream [211].

Chronic atrophic gastritis (CAG) bears many similarities to BE, as both diseases are precancerous lesions characterized as an intestinal metaplasia in an anterior organ [112, 212]. Ten differentially expressed serum microRNAs have been detected in CAG, when compared to chronic non-atrophic gastritis. Of these ten microRNAs, miR-148a-3p, miR-320a, miR-451a, miR-486-3p, miR-486-5p, and miR-92a-3p showed similar though not statistically significant differential expression in BE tissue and serum in comparison to GERD patients. The difficulty in studying serum microRNAs in both diseases is that the aberrant cell types are also present in the small intestine. Any microRNA signature exported from the metaplastic lesions into the blood likely resembles the microRNAs exported from the small intestine. This could obscure any microRNA biomarkers present in the serum samples. CAG can proceed continuously through dysplasia to develop gastric cancer [212]. The study that identified the ten serum microRNA biomarkers for this disease did not include every stage in this progression. Instead, patients with CAG were compared against patients with a related condition, chronic non-atrophic gastritis
This allowed for a binary decision to occur without ambiguity among the groups being compared.

BE pathogenesis is diagnosed as a step by step progression; however, it is a continuous process. Although BE pathogenesis is separated into four pathology groups for this study, the groups are not well delineated. The pathology groups do not describe how close each patient was to entering the next stage of BE pathogenesis. BE is often not diagnosed immediately upon formation of the lesions, and it is impossible to predict when each patient will begin the transition to the next stage. Thus, using serum microRNAs to cluster patients into four groups based on pathology was not successful in this study.

It was expected that, of the four pathology groups, EAC would have yielded significantly differently expressed serum microRNAs. In the GSA, miR-194-5p was significantly downregulated in EAC vs BE tissues. Eight microRNAs known to be differently regulated in various cancers were included in the top twenty most important microRNAs for the serum randomForest classification, randomForest regression, or both analyses (Figures 10 and 11). None of these eight microRNAs were significantly differentially expressed in any comparison in the GSA.

Serum is not the only biofluid available for non-invasive diagnosis of BE pathogenesis. Four differentially expressed microRNAs have been detected in saliva of esophageal cancer patients [213]. Three of these microRNAs, miR-144, miR-21-3p, and miR-451a, showed similar, if not statistically significant, fold changes in the tissue of EAC patients when compared to BE patients in this study. However, these trends were not observed in the serum of EAC patients when compared to BE patients.
Despite the limitations imposed by the biology of BE pathogenesis, serum microRNA biomarkers for this disease may still be revealed if the technical limitations of this study are addressed. This study was unique in that four stages of BE pathogenesis were examined in serum samples. Five studies have identified serum microRNA biomarkers for BE or EAC when compared against controls [129, 130, 214-216]. All these studies utilized either qPCR on select microRNAs or microRNA qPCR arrays to detect differential expression of serum microRNAs. In this study, qPCR was reserved for validation of those serum microRNAs discovered by small RNA-seq.

For the serum cDNA libraries, small RNAs were extracted from 600ul serum for each patient using the mirVana Paris RNA and Protein Isolation kit. The five previous studies were able to extract enough microRNAs from 200-900ul serum for detection of differential expression by qPCR [129, 130, 214-216]. In this study, the microRNA concentration for each patient serum sample following small RNA extraction was unknown. It is possible that extraction from 600ul serum simply did not provide enough starting microRNA material for optimal small RNA-seq results. A kit optimized for extraction of exosomes from serum samples may be more suited for this type of study.

The threshold chosen for the quality control step following small RNA-seq was 200,000 microRNA reads, as assessed by the Ion Torrent smallRNA_analysis plugin. This threshold was selected in order to increase the number of tissue and serum samples included in each pathology group for data analysis. However, a 200,000 microRNA reads threshold may have been too low and allowed for numerous Type II statistical errors. If the cDNA libraries each had possessed 1 million or more microRNA reads to increase
statistical power, more significantly differentially expressed serum microRNAs may have been revealed.

*Tissue*

Six microRNAs were found to be differentially expressed in BE pathogenesis (Table 6). Of these six, only two microRNAs, miR-596 and miR-4655-3p, were novel. In previous studies, over 90 microRNAs have been found to be differentially expressed in BE and EAC tissues [94]. While not all 90 microRNAs passed the Bonferroni threshold and thus were not considered statistically significant, many serum and tissues still appeared to be differentially regulated in this study in the same manner as in literature (Table 11). In this study, only three of the 90 microRNAs appeared as significantly differentially expressed in the fresh-frozen patient tissues (Table 11). This may be due to small sample size; only 5-8 patients were available for each pathology group. This small sample number perhaps could not overcome the variability within the pathology groups caused by the heterogeneity inherent to BE and its associated dysplasia and EAC.

To identify differential expression of microRNAs in BE pathogenesis, previous studies have employed targeted experimental techniques such as microarrays and qPCR [94]. Prior to this study, next generation sequencing techniques, which capture global microRNA expression, had been utilized in three studies for BE pathogenesis [217-219]. Expression fold changes for tissue microRNAs identified in these studies generally match the direction of differential expression observed in this study for comparisons between BE and GERD and EAC and GERD.
Table 11: Serum and Tissue microRNAs Differentially Expressed in BE Literature

<table>
<thead>
<tr>
<th>microRNA</th>
<th>Serum Adjusted p-value</th>
<th>Serum Fold Change</th>
<th>Tissue Adjusted p-value</th>
<th>Tissue Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-196a-5p</td>
<td>2.58E-04</td>
<td>1.87</td>
<td>2.70E-07</td>
<td>18.31</td>
</tr>
<tr>
<td>miR-3613-5p</td>
<td>4.99E-03</td>
<td>1.49</td>
<td>1.83E-03</td>
<td>1.21</td>
</tr>
<tr>
<td>miR-375</td>
<td>2.70E-07</td>
<td>1.39E-04</td>
<td>1.21</td>
<td>2.67</td>
</tr>
<tr>
<td>miR-135b-3p</td>
<td>2.70E-07</td>
<td>1.39E-04</td>
<td>1.21</td>
<td>2.67</td>
</tr>
<tr>
<td>miR-192-5p</td>
<td>3.03E-03</td>
<td>1.21</td>
<td>1.04</td>
<td>2.06</td>
</tr>
<tr>
<td>miR-194-5p</td>
<td>1.85E-03</td>
<td>1.21</td>
<td>1.04</td>
<td>2.06</td>
</tr>
<tr>
<td>miR-196b-5p</td>
<td>2.85E-04</td>
<td>1.21</td>
<td>1.04</td>
<td>2.06</td>
</tr>
<tr>
<td>miR-199a-5p</td>
<td>1.77E-06</td>
<td>1.21</td>
<td>1.04</td>
<td>2.06</td>
</tr>
<tr>
<td>miR-215-3p</td>
<td>2.13E-04</td>
<td>1.21</td>
<td>1.04</td>
<td>2.06</td>
</tr>
<tr>
<td>miR-542-3p</td>
<td>1.25E-03</td>
<td>1.21</td>
<td>1.04</td>
<td>2.06</td>
</tr>
<tr>
<td>miR-136-3p</td>
<td>2.41E-03</td>
<td>1.43</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-29c-3p</td>
<td>2.01E-03</td>
<td>2.04</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-30a-3p</td>
<td>2.34E-03</td>
<td>1.41</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-376c-3p</td>
<td>7.24E-03</td>
<td>3.39</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-409-3p</td>
<td>8.62E-04</td>
<td>1.69</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-551b-3p</td>
<td>2.12E-03</td>
<td>1.17</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-133a-3p</td>
<td>1.39E-03</td>
<td>-1.09</td>
<td>1.85E-03</td>
<td>-1.35</td>
</tr>
<tr>
<td>miR-33a-3p</td>
<td>3.37E-03</td>
<td>-1.18</td>
<td>1.18E-03</td>
<td>-1.64</td>
</tr>
<tr>
<td>miR-99a-3p</td>
<td>2.50E-03</td>
<td>-1.43</td>
<td>5.96E-04</td>
<td>-2.31</td>
</tr>
<tr>
<td>miR-543</td>
<td>3.09E-03</td>
<td>-1.43</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-4462</td>
<td>5.83E-04</td>
<td>-1.65</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-125b-2-3p</td>
<td>3.39E-03</td>
<td>-1.65</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-127-3p</td>
<td>1.73E-03</td>
<td>-1.37</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-133b</td>
<td>2.38E-03</td>
<td>-1.03</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-136-5p</td>
<td>7.77E-04</td>
<td>-1.73</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-149-5p</td>
<td>2.29E-04</td>
<td>-3.93</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-154-5p</td>
<td>8.39E-04</td>
<td>-2.02</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-205-5p</td>
<td>6.14E-05</td>
<td>-12.27</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-224-5p</td>
<td>1.23E-03</td>
<td>-3.60</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-382-5p</td>
<td>1.89E-03</td>
<td>-1.66</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-708-5p</td>
<td>6.39E-04</td>
<td>-2.53</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-10a-5p</td>
<td>2.71E-04</td>
<td>-2.26</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-1-3p</td>
<td>1.82E-04</td>
<td>-3.35</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>miR-18a-3p</td>
<td>3.56E-03</td>
<td>-1.02</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

1Shows similar trends in expression levels in BE literature and in BE vs GERD comparison from small RNA-seq
2Differentially Expressed in both Serum and Tissue data sets from small RNA-seq
3p-values Bonferroni-corrected for multiple comparisons.
4Determined to be significantly differentially expressed after Bonferroni correction.
As previously discussed, BE pathogenesis is continuous, rather than consisting of discrete stages. Perhaps if 1 million or more microRNA reads for each library had been obtained or if more samples per pathology group had been collected, the biological limitations to this study could have been surmounted.

**HOX microRNAs may regulate posterior HOX genes as part of a Negative Feedback Mechanism**

Two of the candidate tissue microRNAs from this study, miR-196a-5p and miR-196b-5p, are considered HOX microRNAs, as they are transcribed from genes intergenic to HOX genes [220]. HOX genes are homeotic genes, a set of transcription factors that contain a DNA-binding homeobox domain. These genes are responsible for gene expression patterning along the AP axis and as such are essential for establishing and maintaining the identities of the foregut, midgut, and hindgut in embryonic development [74]. In the human, thirty-nine HOX genes are divided into four gene clusters, the results of past genome duplications [80]. In the A, B, C, and D clusters, the HOX genes are organized on the chromosome in order of their expression on the AP axis (Figure 16). Correct HOX gene function is dependent on the place and time of its expression. Generally, the HOX genes are grouped into three groups: anterior, central, and posterior (Figure 16) [71, 72]. Central HOX genes from the HOXB cluster have been found to be upregulated in BE tissues [69]. No studies to date have examined the differential expression of microRNAs and their regulation of HOX genes in BE pathogenesis.

Five HOX microRNAs can be transcribed from within the human HOX clusters [220]. miR-10a-5p, miR-10b-5p, and miR-615-3p can be transcribed from genes intergenic to
Four HOX gene clusters in humans are a result of multiple genome duplications. Thirty-nine HOX genes are divided across these clusters. Not all HOX clusters contain all 13 gene homologs. Genes are grouped into three general categories: anterior (HOX1-3), central (HOX4-8), and posterior (HOX 9-13). These categories define the positional expression for those HOX genes. Arrows denote locations of genes encoding HOX microRNAs: (A) MIR-196b encodes miR-196b-5p. (B) MIR-196a-1 encodes miR-196a-5p. (C) MIR-196a-2 encodes miR-196a-5p (D) MIR-10a encodes miR-10a-5p (E) MIR-10b encodes miR-10b-5p (F) MIR-615 encodes miR-615-3p.
anterior HOXA, HOXD, and HOXC genes, respectively (Figure 16) [220]. In this study, these three microRNAs all displayed increased expression in the BE vs GERD comparison for the tissue dataset. However, the differential expression did not pass the Bonferroni threshold and therefore was not considered statistically significant. miR-10a-5p and miR-10b-5p are among the twenty most important microRNAs in the tissue randomForest classification analysis (Figure 10). miR-10a-5p and miR-615-3p are among the twenty most important microRNAs in the tissue randomForest regression analysis (Figure 11).

miR-196a-5p can be transcribed from genes intergenic to either posterior HOXB or HOXC genes, while miR-196b-5p originates from a gene adjacent to HOXA9 [220]. HOX microRNAs are often co-expressed with adjacent HOX genes. Both microRNAs had significantly increased expression in BE tissues when compared to GERD tissues in this study. In Huntington’s Disease, miR-196a-5p upregulation has been correlated with increased transcription of HOXC10 or HOXB9, while upregulation of HOXA10, HOXA11, or HOXA13 have been associated with increased expression of miR-196b-5p [186]. In this study, the comparative upregulation of these two microRNAs in the BE patients may indicate an initial increase in central and posterior HOX gene expression in BE development (Figure 17).

Gene target analysis for miR-196a-5p and miR-196b-5p demonstrated regulation of central and posterior HOX genes from the HOXA, HOXB, and HOXC clusters (Table 10). This may be evidence of a negative feedback mechanism where posterior HOX genes are targeted by co-expressed HOX microRNAs (Figure 17).
Figure 17: A Model for Homeotic Gene Regulation in BE Development

In BE development, increased expression of miR-196a-5p, miR-196b-5p, and miR-215-3p may indicate a mechanism for fine tuning posterior HOX gene expression in the distal esophagus. miR-196a-5p and miR-196b-5p can both be transcribed from genes adjacent to posterior HOX genes. miR-196b-5p and miR-196a-5p are transcribed from genes intergenic to the HOXA and HOXB and HOXC clusters. Their co-expression with posterior HOX genes allows for their targeting of central and posterior genes in the other HOX gene clusters, a negative feedback mechanism. miR-196b-5p is also known to target MEIS1, a HOX co-factor. MLL complex activates HOX genes by methylation of histone 3 at Lysine 4 (H3K4me) and opening chromatin structure. MLL complex can also methylate histone3 at Lysine 79 (H3K79me) at MIR-196b in the HOXA cluster, leading to miR-196b-5p expression. CDX1 can bind to HOX gene promoters to activate HOX gene expression. CDX1 also activates miR-215-3p, allowing this microRNA to target BMI1, a component of PRC1. PRC1 and PRC2 work in conjunction to trimethylate histone 3 at Lysine 27 (H3K27me3) and compact chromatin structure. Thus, posterior HOX genes may be aberrantly expressed in BE pathogenesis under indirect and direct regulation by the candidate microRNAs of this study.
miR-215-3p and miR-196b-5p are involved in indirect regulation of HOX gene expression

In humans, polycomb repressive complexes (PRC) and complexes of Mixed Lineage Leukemia (MLL) proteins work in opposition to regulate HOX gene expression [81, 84]. PRC2 and PRC1 act in conjunction to repress HOX gene expression through methylation of histone H3 at Lysine 27 (H3K27me) and compaction of chromatin structure, respectively [221]. Transcription machinery access to HOX genes located in PRC1-catalyzed compacted chromatin would be restricted [222]. Thus, PRC1 can repress central or posterior HOX gene expression in anterior regions of the body. MLL is a histone methyltransferase that binds to promoters of HOX genes and methylates histone H3 at Lysine 4 (H3K4me) to activate HOX gene expression [81]. The H3K4me epigenetic mark can recruit chromatin remodeling factors which open chromatin and allow active transcription [223].

miR-215-3p targets BMI1, a component of PRC1

PRC2 trimethylates histone H3 at Lysine 27, a signal which recruits PRC1 through one of its core components, BMI1 [224]. The exact mechanism behind PRC1’s ability to compact chromatin is still unclear. BMI1 is required for PRC1-catalyzed chromatin compaction and subsequent gene silencing [225]. BMI1 is an experimentally-validated target of miR-215-3p [189]. Expression of posterior HOXB genes in BE tissues was associated with loss of the H3K27 methylation and decompaction of chromatin at the 5’ end of the HOXB gene cluster [69]. This would suggest that PRC1 repression of central and posterior HOXB genes in the distal esophagus is released in BE development.
CDX1 has been shown to promote expression of miR-215-3p (Figure 17) [189]. CDX genes, including CDX1, are known regulators of HOX genes [39]. CDX gene mutations in humans can lead to homeotic transformations in the central regions of the body due to dysregulation of posterior HOX genes [47]. CDX binding sites have been found in enhancer elements of HOX genes, where CDX genes are thought to acetylate H3K27 to activate central HOX genes [93]. CDX1 expression in the distal esophagus may precede central and posterior HOX gene expression in BE.

*Multiple Lineage Leukemia (MLL) Complex promotes mir-196b-5p expression*

MLL1 is a human homolog to Trithorax originally identified in *Drosophila* [226]. In embryonic development, MLL1 works in a complex with other MLL proteins to upregulate central and posterior HOX genes in the posterior segments of the body [82]. MLL complexes directly bind to HOX promoters to apply an activating epigenetic marker, H3K4 trimethylation [81] [84, 227, 228]. The presence of H3K4me3 recruits the transcription factor TFIID, an initiating player in gene transcription [228-230]. No study to date has examined differential expression of MLL in the human esophagus.

MLL complex has been shown to promote miR-196b-5p expression [231]. This complex methylates Lysine 79 on histone H3 (H3K79me) in certain regions of the HOXA cluster [231]. This epigenetic modification is associated with transcription activation [232]. One region that is affected by this modification contains MIR-196b, the gene which encodes miR-196b-5p. The observed increased expression of miR-196b-5p in the BE patient tissues in this study may have been due to the action of MLL complex in BE development. miR-196b-5p, in turn, can target the HOX cofactor, myeloid ecotropic viral
integration site 1 (MEIS1) [233, 234]. MEIS1 is a member of the three amino acid loop extension (TALE) homeobox gene cluster, which all encode atypical homeodomains [235]. These genes act as cofactors for HOX genes by augmenting their DNA binding, thereby enhancing transcriptional regulation by HOX genes [236]. Thus, increased expression of miR-196b-5p has a detrimental effect on the activity of HOX genes.

A Model for Homeotic Gene Regulation in BE Development

It is unknown what molecular mechanisms contribute to the formation of BE lesions in GERD patients. Three candidate microRNAs were identified by small RNA-seq in fresh-frozen BE patient tissues and validated in FFPE BE tissues by qPCR. Upregulation of miR-196a-5p, miR-196b-5p, and miR-215-3p posits a model for direct and indirect regulation of central and posterior HOX genes in BE development (Figure 17).

The upregulation of miR-196a-5p and miR-196b-5p in BE tissues indicates that posterior HOX genes are expressed in BE development, as these HOX microRNAs can be co-expressed with their adjacent HOX genes. Posterior HOX genes may exist in a bivalent state in the distal esophagus. Bivalency describes the presence of active (H3K4me) and repressive (H3K27me) epigenetic markers in the same gene promoter [237, 238]. This allows affected genes to be silenced but be “poised” for activation. The actions of the PRC1/2 complexes and the MLL complex may be in balance to keep posterior HOX genes in a bivalent state. The onset of BE development, increased expression of miR-215-3p and other factors may upset this balance. Removal of PRC1 repression of posterior HOX genes could be achieved by the observed upregulation of miR-215-3p in BE tissues, as this microRNA targets BMI1, a component of PRC1. miR-215-3p is
upregulated by CDX1, a well-known biomarker of BE lesions and regulator of central and posterior HOX genes[39, 93, 189]. This would allow for co-expression of miR-196a-5p, miR-196b-5p and their respective adjacent HOX genes. MLL complex can promote expression of miR-196b-5p as well [231, 233]. miR-196a-5p and miR-196b-5p are then able to target central and posterior HOX gene expression and activity [220, 233].

Increased expression of miR-196a-5p, miR-196b-5p, and miR-215-3p in BE development may allow for fine-tuning of central and posterior HOX gene expression in the distal esophageal epithelium (Figure 17). To test this model, modulation of HOX gene and microRNA expression in an experimental construct of BE development would need to be assessed. Unfortunately, a technical limitation that has impeded the understanding of BE pathogenesis is the lack of an adequate experimental model of BE. Human esophagus is lined with a non-keratinized squamous epithelium with submucosal glands. Mice and rats, commonly utilized model organisms, have keratinized-lined esophagi which lack submucosal glands [239]. It is known that BE can occur in the ducts of the esophageal submucosal glands, termed ductal metaplasia or buried BE [240]. This ductal metaplasia, like BE, is capable of dysplasia and can produce EAC [241]. The position of the gastroesophageal junction is also of import in BE development, as BE is confined to the distal esophagus. In mice, induced BE-like lesions can occur in the esophagus and proximal stomach, because this junction is positioned further down the GI tract[239]. However, the most important factor which separates humans from commonly used animal models is that gastroesophageal reflux is a natural occurrence. In other non-primate mammals, including mice, dogs, and pigs, reflux must be surgically induced in order for BE-like lesions to develop [239]. Due to the lack of a practical and biologically
relevant animal model, cell lines can be utilized to explore BE development in vitro. However, since BE is a metaplasia consisting of both columnar epithelia and intestinal-type cells, molecular signaling in a single esophageal cell line does not necessarily reflect the molecular signaling in the tissue. One technique to address the limitations of a single 2-D cell line has been a co-culture of multiple esophageal epithelial cell lines in a collagen matrix [242]. Though this 3-D culture system better approximates the BE physiology, it did not include the intestinal-type cells characteristic of BE. There is a distinct need for a model system derived from human esophageal tissue which can be cultured and maintained long-term, while still reflecting all the hallmarks of the BE phenotype.

If an appropriate experimental construct of BE development can be identified, modulation of central and posterior HOX gene transcription and translation would need to be explored. While qPCR is a standard assay technique for gene transcription, an array of techniques could be utilized to assay for HOX protein expression. In their 2012 study, Di Pietro et al utilized in situ hybridization and western blots to probe for HOX gene expression in BE tissues [69]. To further verify that the three candidate microRNAs target central and posterior HOX genes, BMI1, and MEIS1, a biotin-based pull-down assay may be utilized [243]. This method has been shown to be more specific than other techniques to validate microRNA targets, such as microRNA overexpression experiments [244]. microRNAs can block association of an mRNA with ribosomes to block translation, thus polyribosome (polysome) fractionation analysis can determine the strength of microRNA targeting [245, 246].
The model presented in Figure 17 is only possible through modulation of chromatin structure enacted through histone modifications by PRC1 and MLL complexes. Chromatin Immunoprecipitation (ChIP) can be utilized to assess methylation and acetylation at Lysines 27 and 79 on histone H3 at HOX gene promoters [247]. This technique could also validate binding of MLL1 and CDX1 to intergenic regions of central and posterior HOX genes.
LITERATURE CITED


195. Fang, G., J. Liu, Q. Wang, X. Huang, R. Yang, Y. Pang, and M. Yang, MicroRNA-223-3p Regulates Ovarian Cancer Cell Proliferation and Invasion by


131

### Appendix I: small RNA sequencing of Serum and Tissue cDNA libraries

<table>
<thead>
<tr>
<th>Ion PI Chip Number</th>
<th>Number of Samples</th>
<th>Concentration of Pooled Libraries</th>
<th>Percent Templated ISPs&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Total Reads Per Chip</th>
<th>Percent Adaptor Dimer</th>
<th>Percent Multi-Templated ISPs&lt;sup&gt;1&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;2&lt;/sup&gt;</td>
<td>9</td>
<td>5pM</td>
<td>11%</td>
<td>26,165,642</td>
<td>58%</td>
<td>24%</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>5pM</td>
<td>11%</td>
<td>33,467,142</td>
<td>61%</td>
<td>14%</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>5pM</td>
<td>10%</td>
<td>24,327,612</td>
<td>76%</td>
<td>16%</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>7.5pM</td>
<td>10%</td>
<td>24,983,549</td>
<td>78%</td>
<td>12%</td>
</tr>
<tr>
<td>5&lt;sup&gt;3&lt;/sup&gt;</td>
<td>4</td>
<td>5pM</td>
<td>29%</td>
<td>17,460,404</td>
<td>16%</td>
<td>28%</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5pM</td>
<td>8%</td>
<td>23,737,059</td>
<td>71%</td>
<td>26%</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>5pM</td>
<td>8%</td>
<td>12,794,097</td>
<td>70%</td>
<td>24%</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>15pM</td>
<td>13%</td>
<td>25,780,881</td>
<td>74%</td>
<td>16%</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>15pM</td>
<td>19%</td>
<td>42,018,947</td>
<td>62%</td>
<td>17%</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>15pM</td>
<td>15%</td>
<td>28,952,314</td>
<td>77%</td>
<td>12%</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>20pM</td>
<td>32%</td>
<td>35,833,666</td>
<td>65%</td>
<td>20%</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>15pM</td>
<td>21%</td>
<td>36,567,578</td>
<td>66%</td>
<td>21%</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
<td>15pM</td>
<td>28%</td>
<td>46,417,922</td>
<td>53%</td>
<td>25%</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>15pM</td>
<td>19%</td>
<td>38,863,719</td>
<td>48%</td>
<td>24%</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>15pM</td>
<td>28%</td>
<td>43,534,071</td>
<td>61%</td>
<td>16%</td>
</tr>
<tr>
<td>16</td>
<td>14</td>
<td>15pM</td>
<td>23%</td>
<td>45,600,035</td>
<td>55%</td>
<td>25%</td>
</tr>
</tbody>
</table>

<sup>1</sup> ISP: Ion Sphere<sup>TM</sup> Particles

<sup>2</sup> Inefficient loading of the Ion PI chip affected total and microRNA reads. Samples were rerun on Chip 2

<sup>3</sup>Inefficient loading of the Ion PI chip affected total reads but not microRNA reads. Small RNA control from kit was included.
### Appendix II: cDNA libraries from Serum and Tissue small RNAs

<table>
<thead>
<tr>
<th>Patient Sample</th>
<th>Pathology</th>
<th>50-300bp cDNA (nmol/ul)</th>
<th>94-114bp cDNA (nmol/ul)</th>
<th>Barcode</th>
<th>microRNA reads$^{2,3}$</th>
<th>microRNA reads$^{2,3}$</th>
<th>microRNA reads$^{2,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1S</td>
<td>BE</td>
<td>148.6</td>
<td>28</td>
<td>2</td>
<td>1,165,368</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1T</td>
<td>BE</td>
<td>82.8</td>
<td>51</td>
<td>7</td>
<td>409,317</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2S</td>
<td>GERD</td>
<td>123.2</td>
<td>49.2</td>
<td>3</td>
<td>493,212</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2T</td>
<td>GERD</td>
<td>115.2</td>
<td>66.8</td>
<td>14</td>
<td>587,689</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3S</td>
<td>GERD</td>
<td>34.8</td>
<td>10</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3T</td>
<td>GERD</td>
<td>95</td>
<td>18.05</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3S</td>
<td>GERD</td>
<td>81.8</td>
<td>26.5</td>
<td>1</td>
<td>857,667</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3T</td>
<td>GERD</td>
<td>205.2</td>
<td>71</td>
<td>9</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3T</td>
<td>GERD</td>
<td>178.15</td>
<td>72.45</td>
<td>8</td>
<td>65,358</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4S</td>
<td>Normal</td>
<td>365.5</td>
<td>2.36</td>
<td>7</td>
<td>2,058,239</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4T</td>
<td>Normal</td>
<td>46.9</td>
<td>22.7</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4T</td>
<td>Normal</td>
<td>160.4</td>
<td>94.9</td>
<td>1</td>
<td>820,424</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>5S</td>
<td>GERD</td>
<td>95.7</td>
<td>11.1</td>
<td>11</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>5S</td>
<td>GERD</td>
<td>160.1</td>
<td>23.9</td>
<td>12</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>5T</td>
<td>GERD</td>
<td>No library made</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6S</td>
<td>GERD</td>
<td>257.9</td>
<td>186.9</td>
<td>8</td>
<td>67,564</td>
<td>65,239</td>
<td>45,648</td>
</tr>
<tr>
<td>6S</td>
<td>GERD</td>
<td>117.6</td>
<td>70.6</td>
<td>14</td>
<td>39,076</td>
<td>24,837</td>
<td>---</td>
</tr>
<tr>
<td>6T</td>
<td>GERD</td>
<td>87.5</td>
<td>59.3</td>
<td>2</td>
<td>1,538,959</td>
<td>675,776</td>
<td>619,922</td>
</tr>
<tr>
<td>7S</td>
<td>GERD</td>
<td>425.30</td>
<td>275.1</td>
<td>9</td>
<td>3,939,615</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>7T</td>
<td>GERD</td>
<td>41.3</td>
<td>29.9</td>
<td>3</td>
<td>67,000</td>
<td>50,000</td>
<td>---</td>
</tr>
<tr>
<td>8S</td>
<td>GERD</td>
<td>Not received</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>8T</td>
<td>GERD</td>
<td>Not received</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>9S</td>
<td>Normal</td>
<td>205.2</td>
<td>131.35</td>
<td>15</td>
<td>52,955</td>
<td>20,425</td>
<td>20,294</td>
</tr>
<tr>
<td>9S</td>
<td>Normal</td>
<td>283.6</td>
<td>62.3</td>
<td>10</td>
<td>259,675</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>9T</td>
<td>Normal</td>
<td>109</td>
<td>76.2</td>
<td>4</td>
<td>403,891</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>9T</td>
<td>Normal</td>
<td>20.5</td>
<td>18.8</td>
<td>4</td>
<td>478,133</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>10S</td>
<td>GERD</td>
<td>Not received</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>10T</td>
<td>GERD</td>
<td>Not received</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>11S</td>
<td>BE</td>
<td>259.35</td>
<td>158.25</td>
<td>16</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>11S</td>
<td>BE</td>
<td>315</td>
<td>173</td>
<td>10</td>
<td>24,928</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>11T</td>
<td>BE</td>
<td>166.8</td>
<td>59.7</td>
<td>6</td>
<td>89,451</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12S</td>
<td>LGD</td>
<td>122.9</td>
<td>0.7</td>
<td>9</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>12S</td>
<td>LGD</td>
<td>40.3</td>
<td>18.6</td>
<td>15</td>
<td>47,583</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>12T</td>
<td>LGD</td>
<td>99.1</td>
<td>71.7</td>
<td>12</td>
<td>157,235</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>13S</td>
<td>EAC</td>
<td>334.2</td>
<td>226.6</td>
<td>11</td>
<td>267,217</td>
<td>162,316</td>
<td>77,946</td>
</tr>
<tr>
<td>13T</td>
<td>EAC</td>
<td>135.2</td>
<td>98.85</td>
<td>5</td>
<td>84,000</td>
<td>40,000</td>
<td>31,000</td>
</tr>
<tr>
<td>13T</td>
<td>EAC</td>
<td>214.9</td>
<td>150.4</td>
<td>5</td>
<td>336,937</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>13T</td>
<td>EAC</td>
<td>241.1</td>
<td>153</td>
<td>1</td>
<td>733,000</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>14S</td>
<td>GERD</td>
<td>10.3</td>
<td>2.05</td>
<td>13</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>14T</td>
<td>GERD</td>
<td>Not received</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>15S</td>
<td>BE</td>
<td>227.5</td>
<td>130.7</td>
<td>12</td>
<td>565,053</td>
<td>367,094</td>
<td>65,239</td>
</tr>
<tr>
<td>15T</td>
<td>BE</td>
<td>24.2</td>
<td>9.1</td>
<td>6</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>15T</td>
<td>BE</td>
<td>63.5</td>
<td>42.3</td>
<td>6</td>
<td>130,000</td>
<td>105,000</td>
<td>---</td>
</tr>
<tr>
<td>15T</td>
<td>BE</td>
<td>247.6</td>
<td>174.6</td>
<td>4</td>
<td>454,000</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>16S</td>
<td>LGD</td>
<td>87.8</td>
<td>33.7</td>
<td>13</td>
<td>225,876</td>
<td>110,003</td>
<td>---</td>
</tr>
<tr>
<td>16T</td>
<td>LGD</td>
<td>128.85</td>
<td>70.1</td>
<td>7</td>
<td>373,289</td>
<td>303,889</td>
<td>---</td>
</tr>
<tr>
<td>16T</td>
<td>LGD</td>
<td>28</td>
<td>12</td>
<td>5</td>
<td>171,798</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>17S</td>
<td>LGD</td>
<td>71.3</td>
<td>19.2</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>17S</td>
<td>LGD</td>
<td>368.8</td>
<td>151.7</td>
<td>1</td>
<td>261,515</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>17S</td>
<td>LGD</td>
<td>165</td>
<td>110.8</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>17T</td>
<td>LGD</td>
<td>41.2</td>
<td>28.15</td>
<td>7</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>17T</td>
<td>LGD</td>
<td>331.9</td>
<td>215.7</td>
<td>2</td>
<td>1,494,000</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>18S</td>
<td>BE</td>
<td>48.1</td>
<td>27.8</td>
<td>2</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>18S</td>
<td>BE</td>
<td>338.1</td>
<td>162</td>
<td>2</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>18S</td>
<td>BE</td>
<td>244</td>
<td>183.6</td>
<td>2</td>
<td>844,152</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>18T</td>
<td>BE</td>
<td>120.05</td>
<td>91.35</td>
<td>8</td>
<td>119,000</td>
<td>76,000</td>
<td>62,000</td>
</tr>
<tr>
<td>18T</td>
<td>BE</td>
<td>120.05</td>
<td>91.35</td>
<td>8</td>
<td>37,997</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>19S</td>
<td>GERD</td>
<td>67.1</td>
<td>15.9</td>
<td>3</td>
<td>121,174</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>19S</td>
<td>GERD</td>
<td>374.3</td>
<td>196.3</td>
<td>3</td>
<td>195,120</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>19S</td>
<td>GERD</td>
<td>447</td>
<td>258.1</td>
<td>3</td>
<td>299,691</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>19T</td>
<td>GERD</td>
<td>74.75</td>
<td>60.45</td>
<td>9</td>
<td>123,113</td>
<td>120,310</td>
<td>79,137</td>
</tr>
<tr>
<td>19T</td>
<td>GERD</td>
<td>257.4</td>
<td>194.9</td>
<td>6</td>
<td>155,621</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>20S</td>
<td>BE</td>
<td>57.9</td>
<td>24.5</td>
<td>4</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>20S</td>
<td>BE</td>
<td>587.2</td>
<td>236.1</td>
<td>4</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>20S</td>
<td>BE</td>
<td>79.4</td>
<td>56.1</td>
<td>4</td>
<td>1,620,830</td>
<td>617,535</td>
<td>540,092</td>
</tr>
<tr>
<td>20T</td>
<td>BE</td>
<td>166.7</td>
<td>129.1</td>
<td>10</td>
<td>100,000</td>
<td>90,000</td>
<td>79,000</td>
</tr>
<tr>
<td>20T</td>
<td>BE</td>
<td>187.7</td>
<td>131.8</td>
<td>10</td>
<td>2,009,572</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>21S</td>
<td>GERD</td>
<td>489.2</td>
<td>84.6</td>
<td>5</td>
<td>63,046</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>21S</td>
<td>GERD</td>
<td>366.8</td>
<td>177.9</td>
<td>5</td>
<td>64,190</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>21S</td>
<td>GERD</td>
<td>371.6</td>
<td>225.7</td>
<td>5</td>
<td>2,917,007</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GERD</td>
<td>BE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21T</td>
<td>92.4</td>
<td>57.9</td>
<td>11</td>
<td>77,965</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>21T</td>
<td>199.9</td>
<td>157.1</td>
<td>7</td>
<td>403,686</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>21T</td>
<td>9.6</td>
<td>9</td>
<td>11</td>
<td>103,850</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>22S</td>
<td>53</td>
<td>12.6</td>
<td>6</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>22S</td>
<td>376.7</td>
<td>153.6</td>
<td>6</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>22S</td>
<td>143.6</td>
<td>52.7</td>
<td>6</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>22S</td>
<td>315.1</td>
<td>91.4</td>
<td>6</td>
<td>105,306</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>22T</td>
<td>129.8</td>
<td>92.2</td>
<td>12</td>
<td>60,886</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>23S</td>
<td>211.2</td>
<td>117.2</td>
<td>13</td>
<td>1,731,449</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>23S</td>
<td>79.05</td>
<td>51.95</td>
<td>13</td>
<td>240,323</td>
<td>226,658</td>
<td>172,304</td>
<td></td>
</tr>
<tr>
<td>24S</td>
<td>219.4</td>
<td>130.1</td>
<td>14</td>
<td>660,978</td>
<td>460,448</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>24T</td>
<td>64.55</td>
<td>41.15</td>
<td>14</td>
<td>135,218</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>24T</td>
<td>230.5</td>
<td>174.7</td>
<td>8</td>
<td>156,000</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>25S</td>
<td>56.8</td>
<td>37.5</td>
<td>15</td>
<td>2,056,656</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>25S</td>
<td>304.2</td>
<td>199.3</td>
<td>15</td>
<td>65,373</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>25T</td>
<td>123.4</td>
<td>73.6</td>
<td>15</td>
<td>121,540</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>26S</td>
<td>274.9</td>
<td>148.1</td>
<td>16</td>
<td>221,207</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>26T</td>
<td>173</td>
<td>132.4</td>
<td>16</td>
<td>130,000</td>
<td>98,000</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>26T</td>
<td>215.3</td>
<td>148.9</td>
<td>10</td>
<td>2,762,128</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>27S</td>
<td>370</td>
<td>276.4</td>
<td>10</td>
<td>201,209</td>
<td>115,318</td>
<td>61,062</td>
<td></td>
</tr>
<tr>
<td>27T</td>
<td>54</td>
<td>32.5</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>27T</td>
<td>28.05</td>
<td>17.5</td>
<td>13</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>27T</td>
<td>238.7</td>
<td>172.7</td>
<td>3</td>
<td>575,000</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>24T</td>
<td>64.55</td>
<td>41.15</td>
<td>14</td>
<td>135,218</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>24T</td>
<td>230.5</td>
<td>174.7</td>
<td>8</td>
<td>156,000</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>25S</td>
<td>56.8</td>
<td>37.5</td>
<td>15</td>
<td>2,056,656</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>25S</td>
<td>304.2</td>
<td>199.3</td>
<td>15</td>
<td>65,373</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>25T</td>
<td>123.4</td>
<td>73.6</td>
<td>15</td>
<td>121,540</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>26S</td>
<td>274.9</td>
<td>148.1</td>
<td>16</td>
<td>221,207</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>26T</td>
<td>173</td>
<td>132.4</td>
<td>16</td>
<td>130,000</td>
<td>98,000</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>26T</td>
<td>215.3</td>
<td>148.9</td>
<td>10</td>
<td>2,762,128</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>27S</td>
<td>370</td>
<td>276.4</td>
<td>10</td>
<td>201,209</td>
<td>115,318</td>
<td>61,062</td>
<td></td>
</tr>
<tr>
<td>27T</td>
<td>54</td>
<td>32.5</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>27T</td>
<td>28.05</td>
<td>17.5</td>
<td>13</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>27T</td>
<td>238.7</td>
<td>172.7</td>
<td>3</td>
<td>575,000</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>28S</td>
<td>242</td>
<td>86.25</td>
<td>2</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>28S</td>
<td>131.4</td>
<td>102.5</td>
<td>2</td>
<td>217,622</td>
<td>130,048</td>
<td>73,585</td>
<td></td>
</tr>
<tr>
<td>28T</td>
<td>80.1</td>
<td>60.2</td>
<td>2</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>28T</td>
<td>272.1</td>
<td>171.1</td>
<td>9</td>
<td>1,664,468</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>29S</td>
<td>90.35</td>
<td>46</td>
<td>3</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>29S</td>
<td>455.8</td>
<td>309.5</td>
<td>3</td>
<td>836,870</td>
<td>473,280</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>Sample</td>
<td>Tissue</td>
<td>R2</td>
<td>R1</td>
<td>Mean</td>
<td>S.D.</td>
<td>Median</td>
<td>Count</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>29T</td>
<td>BE</td>
<td>118.1</td>
<td>83.7</td>
<td>3</td>
<td>116,949</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>29T</td>
<td>BE</td>
<td>37.95</td>
<td>24.95</td>
<td>15</td>
<td>236,985</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>30S</td>
<td>Normal</td>
<td>129.8</td>
<td>65.7</td>
<td>4</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>30S</td>
<td>Normal</td>
<td>77.3</td>
<td>31.2</td>
<td>16</td>
<td>754,897</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>30T</td>
<td>Normal</td>
<td>106.3</td>
<td>69.4</td>
<td>4</td>
<td>1,031,683</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>31S</td>
<td>ESCC</td>
<td>Library not made</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>31T</td>
<td>ESCC</td>
<td>64.3</td>
<td>43.5</td>
<td>5</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>32S</td>
<td>GERD</td>
<td>231.6</td>
<td>158.8</td>
<td>14</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>32S</td>
<td>GERD</td>
<td>136.9</td>
<td>64.2</td>
<td>4</td>
<td>602,971</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>32T</td>
<td>GERD</td>
<td>20.6</td>
<td>8.1</td>
<td>9</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>32T</td>
<td>GERD</td>
<td>93.1</td>
<td>49.7</td>
<td>16</td>
<td>1,115,086</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>33S</td>
<td>BE</td>
<td>119.2</td>
<td>57.2</td>
<td>5</td>
<td>910,976</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>33T</td>
<td>BE</td>
<td>120.3</td>
<td>77.5</td>
<td>11</td>
<td>460,777</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>34S</td>
<td>EAC</td>
<td>461.55</td>
<td>229.05</td>
<td>8</td>
<td>247,384</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>34T</td>
<td>EAC</td>
<td>78.05</td>
<td>42.6</td>
<td>8</td>
<td>79,052</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>35S</td>
<td>GERD</td>
<td>374</td>
<td>243.7</td>
<td>2</td>
<td>109,889</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>35T</td>
<td>GERD</td>
<td>350.8</td>
<td>259.1</td>
<td>8</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>36S</td>
<td>GERD</td>
<td>404.3</td>
<td>235</td>
<td>7</td>
<td>1,670,527</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>36T</td>
<td>GERD</td>
<td>301.8</td>
<td>205</td>
<td>14</td>
<td>101,000</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>36T</td>
<td>GERD</td>
<td>324.7</td>
<td>230.6</td>
<td>1</td>
<td>330,521</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>37S</td>
<td>LGD</td>
<td>494.00</td>
<td>290.45</td>
<td>11</td>
<td>5,969,490</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>37T</td>
<td>LGD</td>
<td>82.45</td>
<td>46.25</td>
<td>11</td>
<td>98,000</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>37T</td>
<td>LGD</td>
<td>82.45</td>
<td>46.25</td>
<td>11</td>
<td>48,000</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>37T</td>
<td>LGD</td>
<td>195</td>
<td>109</td>
<td>5</td>
<td>4,060,390</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>38S</td>
<td>EAC</td>
<td>294.8</td>
<td>116.85</td>
<td>12</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>38S1</td>
<td>EAC</td>
<td>229.8</td>
<td>105.75</td>
<td>12</td>
<td>578,011</td>
<td>58,714</td>
<td>---</td>
</tr>
<tr>
<td>38T</td>
<td>EAC</td>
<td>65.1</td>
<td>45.45</td>
<td>12</td>
<td>23,310</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>38T</td>
<td>EAC</td>
<td>65.1</td>
<td>45.45</td>
<td>12</td>
<td>16,829</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>38T</td>
<td>EAC</td>
<td>158.9</td>
<td>85.4</td>
<td>12</td>
<td>2,705,433</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>39S</td>
<td>LGD</td>
<td>361.88</td>
<td>176.4</td>
<td>13</td>
<td>1,651,601</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>39S</td>
<td>LGD</td>
<td>57.7</td>
<td>41.6</td>
<td>13</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>39T</td>
<td>LGD</td>
<td>64.6</td>
<td>45</td>
<td>13</td>
<td>769,308</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>40S1</td>
<td>Normal</td>
<td>108.4</td>
<td>59.5</td>
<td>6</td>
<td>13,538,600</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>40T</td>
<td>Normal</td>
<td>34.1</td>
<td>17.5</td>
<td>5</td>
<td>2,620,041</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

1 Library was suboptimal (contained adaptor dimers and/or t-RNA peaks)

2 MicroRNA reads were provided by Ion Torrent smallRNA_analysis plugin, a quality control step. Read counts herein differ from those reported in Table 5 as the Ion Torrent and Partek Flow® map raw reads to different versions of miRBase, v20 and v21, respectively.

3 Libraries in bold-type had over 200,000 microRNAs and were included in further analyses
### Appendix III: Variance of TMM-normalized microRNA reads of Selected microRNAs

<table>
<thead>
<tr>
<th>microRNA</th>
<th>Normal</th>
<th>GERD</th>
<th>BE</th>
<th>LGD</th>
<th>EAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-106b-5p</td>
<td><strong>1.60E+08</strong></td>
<td>4.58E+07</td>
<td>1.04E+07</td>
<td>1.52E+08</td>
<td>6.01E+05</td>
</tr>
<tr>
<td>hsa-miR-15a-5p</td>
<td><strong>4.15E+08</strong></td>
<td>1.91E+08</td>
<td>3.19E+07</td>
<td>1.30E+08</td>
<td>4.79E+06</td>
</tr>
<tr>
<td>hsa-miR-181a-5p</td>
<td><strong>2.07E+07</strong></td>
<td>2.40E+06</td>
<td>1.00E+07</td>
<td>1.05E+07</td>
<td>5.82E+06</td>
</tr>
<tr>
<td>hsa-miR-194-5p</td>
<td><strong>6.01E+04</strong></td>
<td>1.96E+04</td>
<td>4.65E+03</td>
<td>1.31E+04</td>
<td>1.80E+02</td>
</tr>
<tr>
<td>hsa-miR-6131</td>
<td>2.17E+06</td>
<td>2.68E+05</td>
<td>4.41E+06</td>
<td>3.53E+05</td>
<td>7.22E+04</td>
</tr>
<tr>
<td>hsa-miR-92a-3p</td>
<td><strong>7.81E+07</strong></td>
<td>4.36E+07</td>
<td>3.37E+07</td>
<td>3.03E+07</td>
<td>1.21E+06</td>
</tr>
<tr>
<td>hsa-miR-93-5p</td>
<td><strong>3.14E+08</strong></td>
<td>7.50E+06</td>
<td>7.57E+06</td>
<td>6.56E+07</td>
<td>2.10E+06</td>
</tr>
<tr>
<td>Tissue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-106b-5p</td>
<td>1.75E+06</td>
<td>3.52E+04</td>
<td>3.02E+05</td>
<td>5.15E+05</td>
<td>1.29E+07</td>
</tr>
<tr>
<td>hsa-miR-15a-5p</td>
<td>2.85E+06</td>
<td>1.56E+05</td>
<td>1.57E+05</td>
<td>8.28E+05</td>
<td>2.63E+07</td>
</tr>
<tr>
<td>hsa-miR-181a-5p</td>
<td><strong>1.67E+05</strong></td>
<td>5.56E+03</td>
<td>6.47E+04</td>
<td>1.60E+05</td>
<td>1.32E+05</td>
</tr>
<tr>
<td>hsa-miR-194-5p</td>
<td>1.87E+02</td>
<td>1.61E+05</td>
<td>2.53E+05</td>
<td>1.84E+06</td>
<td>4.74E+06</td>
</tr>
<tr>
<td>hsa-miR-6131</td>
<td>4.83E+02</td>
<td>9.79E+02</td>
<td>1.76E+03</td>
<td>1.01E+04</td>
<td>2.31E+04</td>
</tr>
<tr>
<td>hsa-miR-92a-3p</td>
<td>2.97E+05</td>
<td>2.73E+03</td>
<td>2.91E+03</td>
<td>4.20E+04</td>
<td>8.76E+05</td>
</tr>
<tr>
<td>hsa-miR-93-5p</td>
<td>3.09E+06</td>
<td>7.44E+04</td>
<td>1.45E+04</td>
<td>2.03E+05</td>
<td>4.93E+06</td>
</tr>
</tbody>
</table>

1 Selected microRNAs were chosen based on a GSA performed on serum or tissue datasets composed of the same 24 patients (unshown). These microRNAs were unique to a specific pair-wise comparison between two of the five patient groups.

2 TMM-normalized microRNA reads from those 24 patients for which both serum and tissue samples passed the 200,000 microRNA read count threshold.

3 Instances when variance in the Normal patient group exceeded variance in other patient groups are in boldtype.