

Washington University in St. Louis

## Washington University Open Scholarship

---

McKelvey School of Engineering Theses & Dissertations

McKelvey School of Engineering

---

Winter 12-15-2017

### Identification of Prognostic Cancer Biomarkers through the Application of RNA-Seq Technologies and Bioinformatics

Nathan Wong

*Washington University in St. Louis*

Follow this and additional works at: [https://openscholarship.wustl.edu/eng\\_etds](https://openscholarship.wustl.edu/eng_etds)



Part of the [Bioinformatics Commons](#), [Biomedical Engineering and Bioengineering Commons](#), and the [Oncology Commons](#)

---

#### Recommended Citation

Wong, Nathan, "Identification of Prognostic Cancer Biomarkers through the Application of RNA-Seq Technologies and Bioinformatics" (2017). *McKelvey School of Engineering Theses & Dissertations*. 288. [https://openscholarship.wustl.edu/eng\\_etds/288](https://openscholarship.wustl.edu/eng_etds/288)

This Dissertation is brought to you for free and open access by the McKelvey School of Engineering at Washington University Open Scholarship. It has been accepted for inclusion in McKelvey School of Engineering Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact [digital@wumail.wustl.edu](mailto:digital@wumail.wustl.edu).

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science  
Department of Biomedical Engineering

Dissertation Examination Committee:

Xiaowei Wang, Chair

Hong Chen

Christopher Maher

Jin-Yu Shao

Gary Stormo

Identification of Prognostic Cancer Biomarkers through the Application of RNA-Seq  
Technologies and Bioinformatics

by

Nathan William Wong

A dissertation presented to  
The Graduate School  
of Washington University in  
partial fulfillment of the  
requirements for the degree  
of Doctor of Philosophy

December 2017  
St. Louis, Missouri

© 2017, Nathan Wong

# Table of Contents

List of Figures .....	iv
List of Tables .....	vi
Acknowledgments.....	viii
Abstract.....	x
Chapter 1: Introduction.....	1
1.1 Biomarkers in Contemporary Personalized Medicine .....	1
1.2 MicroRNAs: History and Functions in Biological Processes and Disease.....	5
1.2.1 MicroRNA Discovery, Biogenesis, and Function.....	5
1.2.2 Prediction of MicroRNA Targets .....	7
1.2.3 MicroRNAs as Biomarkers for Disease .....	9
1.3 The Role of Human Papillomavirus in Tumor Formation .....	12
1.3.1 A Brief History of Oncoviruses.....	12
1.3.2 Function of Human Papillomavirus Proteins .....	19
1.3.3 Effects of Human Papillomavirus and Other Oncoviruses on MicroRNA Expression.....	25
1.4 Characterization of Cervical and Oropharyngeal Cancer in the Context of Human Papillomavirus.....	28
1.4.1 Tumor Source Site and Genomics Affect Cervical Cancer Prognosis.....	29
1.4.2 Human Papillomavirus Distinguishes Oropharyngeal Cancer from Other Head and Neck Tumors .....	31
1.4.3 Applications of Human Papillomavirus in the Diagnostic Setting.....	33
1.5 Project Aims .....	34
1.6 References .....	36
Chapter 2: Prognostic miRNA Signatures Derived from The Cancer Genome Atlas for Cancers of the Head and Neck and the Cervix .....	57
2.1 Abstract .....	57
2.2 Introduction .....	58
2.3 Materials and Methods .....	60
2.4 Results .....	63
2.5 Discussion .....	82

2.6 References .....	84
Chapter 3: OncomiR: An Online Resource for Exploring Pan-Cancer MicroRNA Dysregulation .....	88
3.1 Abstract .....	88
3.2 Introduction .....	89
3.3 Materials and Methods .....	90
3.4 Results .....	92
3.5 Discussion .....	101
3.6 References .....	104
Chapter 4: Pathway Analysis Identifies MicroRNA-Mediated Mechanisms of HPV-Induced Oncogenesis and Tumor Survival.....	107
4.1 Abstract .....	107
4.2 Introduction .....	108
4.3 Materials and Methods .....	110
4.4 Results .....	112
4.5 Discussion .....	119
4.6 References .....	124
4.7 Supplementary Tables .....	127
Chapter 5: Conclusions .....	139

# List of Figures

Figure 1.1: miRNA biogenesis and function.....	6
Figure 1.2: Structure of the HPV viroid and genome .....	20
Figure 1.3: Oncogenic activity of the HPV E6 and E7 proteins.....	22
Figure 2.1: Validation of an existing miRNA signature with TCGA data.....	64
Figure 2.2: Four significant miRNAs associated with overall survival of TCGA OPSCC patients .....	67
Figure 2.3: Kaplan-Meier survival analysis to evaluate the novel OPSCC 4-miRNA prognostic signature .....	69
Figure 2.4: Receiver operating characteristic (ROC) curves for the training and validation cohorts from TCGA.....	73
Figure 2.5: Kaplan-Meier survival analysis to evaluate the OSCC and LSCC miRNA prognostic models.....	75
Figure 2.6: Kaplan-Meier survival analysis to evaluate the miRNA prognostic signatures in other subtypes of HNSCC.....	76
Figure 2.7: Kaplan-Meier survival analysis to evaluate an existing OPSCC miRNA signature in OSCC and LSCC.....	77
Figure 2.8: Kaplan-Meier survival analysis to analyze an existing 2-miRNA prognostic signature in cervical cancer.....	79
Figure 2.9: Kaplan-Meier analysis for a novel 4-miRNA prognostic signature in cervical cancer.....	80
Figure 2.10: The receiver operating characteristic (ROC) curve for the novel CESC signature in the TCGA training cohort.....	80
Figure 2.11: Kaplan-Meier analysis for the novel CESC signature in cervical squamous cell carcinomas and cervical adenosquamous carcinomas and adenocarcinomas.....	81
Figure 2.12: Kaplan-Meier analysis of the novel CESC signature in an independent cervical cancer cohort.....	82
Figure 3.1: Database and server design for OncomiR.....	93
Figure 3.2: Database schematic for OncomiR.....	94

Figure 3.3: Search for miRNA biomarkers in the OncomiR database.....	95
Figure 3.4: Search results for survival-associated miRNAs.....	96
Figure 3.5: Search results for average miRNA expression levels.....	97
Figure 3.6: OncomiR search results for miRNA target prediction.....	98
Figure 3.7: <i>De novo</i> analysis in OncomiR for survival signature and tumor clustering.....	100
Figure 3.8: Overview of OncomiR's functionality.....	103
Figure 4.1: Mechanism for identifying miRNA-target interactions in cancers.....	115
Figure 4.2: A diagram of potential miRNA-mediated dysregulation in response to HPV.....	119

# List of Tables

Table 1.1: ICTV classification of viruses into major orders .....	13
Table 1.2: Baltimore classification of viruses.....	14
Table 1.3: Human oncoviruses and associated viral oncoproteins.....	16
Table 2.1: Characteristics of the HNSCC patients included in TCGA.....	66
Table 2.2: Multivariate Cox regression analysis to evaluate independence of the prognostic miRNA signatures from clinical parameters.....	70
Table 2.3: Characteristics of the OPSCC patients at Washington University.....	71
Table 2.4: Significantly dysregulated miRNAs associated with overall survival and used to develop prognostic models for OSCC and LSCC.....	74
Table 3.1: Summary of cancer types and patient counts from The Cancer Genome Atlas.....	91
Table 4.1: Patient characteristics of the HPV cancer cohorts.....	111
Table 4.2: HPV-dysregulated miRNAs in OPSCC.....	113
Table 4.3: HPV-dysregulated miRNAs in CESC.....	114
Table 4.4: miRNA-Target Interactions Conserved Between Cervical and Oropharyngeal Cancers in Response to HPV Status.....	118
Table 4.5: HPV types in cervical and oropharyngeal cancers, separated by tumor source site...	122
Supplementary Table 4.1: Significant biological processes in HPV(+) OPSCC, through initial identification of significantly dysregulated miRNAs and subsequent targets.....	127
Supplementary Table 4.2: Significant biological processes in HPV(+) CESC, through initial identification of significantly dysregulated miRNAs and subsequent targets .....	129
Supplementary Table 4.3: Significant biological processes in HPV(-) OPSCC, through initial identification of significantly dysregulated miRNAs and subsequent targets.....	131
Supplementary Table 4.4: Significant biological processes in HPV(-) CESC, through initial identification of significantly dysregulated miRNAs and subsequent targets.....	132



Supplementary Table 4.5: Significant biological processes in HPV(+) OPSCC, through initial identification of significantly dysregulated genes.....	133
Supplementary Table 4.6: Significant biological processes in HPV(+) CESC, through initial identification of significantly dysregulated genes.....	134
Supplementary Table 4.7: Significant biological processes in HPV(-) OPSCC, through initial identification of significantly dysregulated genes.....	136
Supplementary Table 4.8: Significant biological processes in HPV(-) CESC, through initial identification of significantly dysregulated genes.....	138

# Acknowledgments

First and foremost, I want to thank my mentor, Dr. Xiaowei Wang, for guiding me through this remarkable journey. He has been an excellent teacher and role model, and an inspiration for the kind of researcher I aspire to be.

I would like to thank the members of my thesis committee, Drs. Jin-Yu Shao, Hong Chen, Gary Stormo, and Christopher Maher, for helping steer me through this research and providing valuable insight and direction.

I would like to thank all the members of the Wang lab, past and present, especially (in no particular order) Weijun, Shuai, Callie, Paul, Yuhao, Ping, Wesley and Arlise, for making the lab so much more of an interesting and enjoyable place to work.

I would like to thank the Department of Biomedical Engineering, for accepting me into such an exciting program that encourages personal and scientific growth, and for introducing me to a group of peers of that is unique, diverse, and pleasantly able to transition seamlessly between scientific jargon and pop culture.

I want to thank all my science and math teachers, past and present, for encouraging me from an early age to dig in to problems and find answers to two fundamental questions in nature: “Why?” and “How?”

Many special thanks to the first two teachers in my life, my parents William and Felicia. You’ve inspired me to become not just a better student, but a better person, and also taught me how to cook so I wouldn’t starve when I left home. I also want to thank my brothers for being emotional supports and motivators, even if I didn’t always say so.

Above all, I want to thank my wife, Amy, for joining me through this strange, weird, amazing stage of life. For reasons that I can't always understand, you came out to St. Louis with me when I told you that I wasn't done with school, and you've made so many sacrifices to help us get through early marriage and parenthood. I can't express enough how you've helped me and supported me as this journey has reached this end, and how much I look forward to taking the next steps with you and our family.

Dedicated to my children, Richard and Adelaide.

Nathan Wong

*Washington University in St. Louis*

*December 2017*

## ABSTRACT OF THE DISSERTATION

Identification of Prognostic Cancer Biomarkers through the Application of RNA-Seq

Technologies and Bioinformatics

by

Nathan William Wong

Doctor of Philosophy in Biomedical Engineering

Washington University in St. Louis, 2017

Dr. Xiaowei Wang, Chair

MicroRNAs (miRNAs) are short single-stranded RNAs that function as the guide sequence of the post-transcriptional regulatory process known as the RNA-induced silencing complex (RISC), which targets mRNA sequences for degradation through complementary binding to the guide miRNA. Changes in miRNA expression have been reported as correlated with numerous biological processes, including embryonic development, cellular differentiation, and disease manifestation. In the latter case, dysregulation has been observed in response to infection by human papillomavirus (HPV), which has also been established as both oncogenic in cervical cancers and oropharyngeal cancers and favorable for overall patient survival after tumor formation. The identification of dysregulated miRNAs associated with both HPV infection and cancer survival requires large datasets of high-throughput sequencing data, which were obtained through The Cancer Genome Atlas. By analyzing this public data, we have identified a series of proposed mechanisms for cancer formation and survival that is mediated through the miRNA-RISC regulatory mechanism in response to HPV infection. We have also identified a diverse set of miRNA biomarkers that have been incorporated into linear expression-based risk signatures that are prognostic for overall patient survival after tumor diagnosis in HPV-related cancers. The

tools that were used to identify both miRNA biomarkers and proposed targets in public datasets, such as The Cancer Genome Atlas, have since been incorporated into an web-accessible resource, OncomiR.org, to streamline the process of biomarker identification for the cancer research community.

# **Chapter 1: Introduction**

The expanded role of RNA-sequencing platforms in the identification of cancer biomarkers has allowed for the research community to delve deeply into the genomic mechanisms behind tumor formation and survival, and subsequently identify coding and noncoding nucleotide sequences that can be applied to the clinical setting in determining patient risk of tumor formation and survival. Here, we will describe bioinformatics techniques and pipelines that have been applied to next-generation RNA-sequencing data used to identify such biomarkers, and various mechanisms that may be affected by the dysregulated biomarkers in tumor tissues.

## **1.1 Biomarkers in Contemporary Personalized Medicine**

Medicine has always consisted of two primary facets: the identification of an ailment based on presented symptoms, and the treatment of said ailment. In modern times, the two aspects have begun to overlap as the paradigm of personalized medicine has taken form. It is rare that two patients present with identical symptoms, due to variations in environment, behavior, and genetics; as such, therapies should be tailored as well, so as to maximize treatment efficacy. This has taken a number of different forms including the identification of specific therapeutic targets, the expansion of diagnostic criteria, and the intersection of the two fields.

The field of biomarker identification can also be separated into these two categories. Certain biomarkers can be utilized in the clinical setting to stratify patients by risk of disease development or progression. Comparatively, the identification of markers that can be acted upon therapeutically, either through drugs or other interventions, constitute the second category. Biomarkers can serve as both therapeutic interventions and diagnostic criteria. One of the better known examples is in the contemporary treatment of breast cancer. Upon presentation, a tumor

sample can be tested for three different cell surface receptors: estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor (HER2/neu, or HER2) (1). Breast cancer cells may express these markers at a significantly higher level than the surrounding healthy cells, as well as being crucial for tumor proliferation, making the markers strong candidates for intervention. The standard chemotherapeutic approach for ER/PR(+), also called hormone-positive, breast cancer is hormone therapy, which attacks the signaling pathway controlled by the receptors. One mechanism of preventing the receptor-mediated signaling cascade is through competitive inhibition, in which the drug competes with estrogen for binding to the receptor; such drugs include tamoxifen (trade name Nolvadex) and toremifene (Farestan), which modulate binding of estrogen through competition for the receptor, and fulvestrant (Faslodex), which destabilizes the receptor after binding to induce receptor degradation (2,3). Another approach is the reduction of estrogen available to activate the receptor signaling pathway, which can be achieved by preventing its production in the ovaries or through aromatase inhibitors, such as anastrozole (Arimidex), which prevent the aromatase enzyme from producing estrogen (4). HER2(+) breast cancers can be treated with monoclonal antibodies (mAbs) such as trastuzumab (Herceptin) and pertuzumab (Perjeta) (5,6). HER2 is an EGF receptor that initiates signaling in the MAPK, PI3K/Akt, and STAT pathways, among others, leading to cell proliferation (7). By targeting HER2, tumor growth and cellular replication can be arrested; additionally, some evidence suggests that trastuzumab also can activate the anti-proliferative protein p27 (8). It should also be noted that breast cancers may also test negative for all three markers, which constitutes a small but high-risk cohort of cancers which are candidates for more drastic interventions (9).

Diagnostic criteria in the clinic has typically included samples that can be easily observed or obtained from the patient through minimally invasive procedures; for example, prostate-specific antigen, a biomarker that can indicate the development of prostate cancer, can be isolated from a blood draw (10). However, such diagnostic tests can be imprecise in determining risk; the same warning holds true for evaluating behavioral and environmental risk factors such as smoking and alcohol consumption or exposure to high risk carcinogens. Additional environmental factors to consider can include the presence of foreign bodies such as viruses. Evidence has shown that certain viruses, including the Epstein-Barr virus, human papillomavirus, and hepatitis viruses B and C, are tumorigenic agents (11). As such, behavioral factors such as sexual practices may need to be considered by the diagnostician, since certain activities increase the likelihood of patient exposure.

Precision for risk stratification can be improved through genotyping, which has become increasingly more affordable in recent years, as evidenced by home genotyping test distributed by companies such as 23andMe (12). Diagnostic genomic markers can include protein-coding genes, which can be observed and measured at the proteomic level through procedures like in situ hybridization, the method used to identify the previously described markers for breast cancer. Genomic biomarkers can also include noncoding RNA sequences which can be shorter than 50 bases (e.g. microRNAs and PIWI-interacting RNAs) or as long as coding genes (e.g. large intergenic noncoding RNAs and circular RNAs) (13,14). Historically, noncoding RNA sequences were considered “junk” RNA, as they did not fit into the classical dogma of molecular biology, i.e. DNA is transcribed to RNA, which is translated to proteins (13). Relatively recently, some noncoding RNA species have been shown to have a greater purpose; for example, microRNAs serve an important role in post-transcriptional regulation, which will be described



later in this chapter in greater detail (15). Similar to protein-coding transcripts, noncoding transcripts may also be dysregulated in tumor tissues when compared to normal tissues, or even when comparing between poor and fair patient prognosis. Given the regulatory implications of some of these RNA species, their biological effects may be as drastic as a typical protein regulatory element.

The improved accessibility of genomic and sequencing platforms has also allowed for large-scale genomic studies for the characterization of various cancers. In 2004, the National Cancer Institute and the National Human Genome Research Institute launched a nationwide pilot program titled “The Cancer Genome Atlas” (TCGA) to perform such analyses on cancers of significant clinical interest (16). By pooling patient tumor samples from multiple treatment centers throughout North America, geographical bias could be eliminated in identifying tumorigenic and relevant clinical features. Additionally, TCGA was designed to provide a central data repository for both raw and processed data. The first publication from TCGA provided insights into the genetic makeup of glioblastoma that confirmed previous observations as well as identified novel tumor characteristics (17). Similar analyses were conducted on 32 other cancer types, including more common cancer species like breast, head and neck, and cervical cancers, and less common but higher risk types such as adrenocortical carcinoma and mesothelioma (16). The diversity of analysis types include whole genome sequencing, RNA-sequencing, short RNA-sequencing (also described as miRNA-sequencing), and methylation analysis (16). By providing both the raw and the processed data, the research community has access to a large dataset of genomic data that can be mined for diagnostic and therapeutic biomarkers in a wide variety of cancer types.

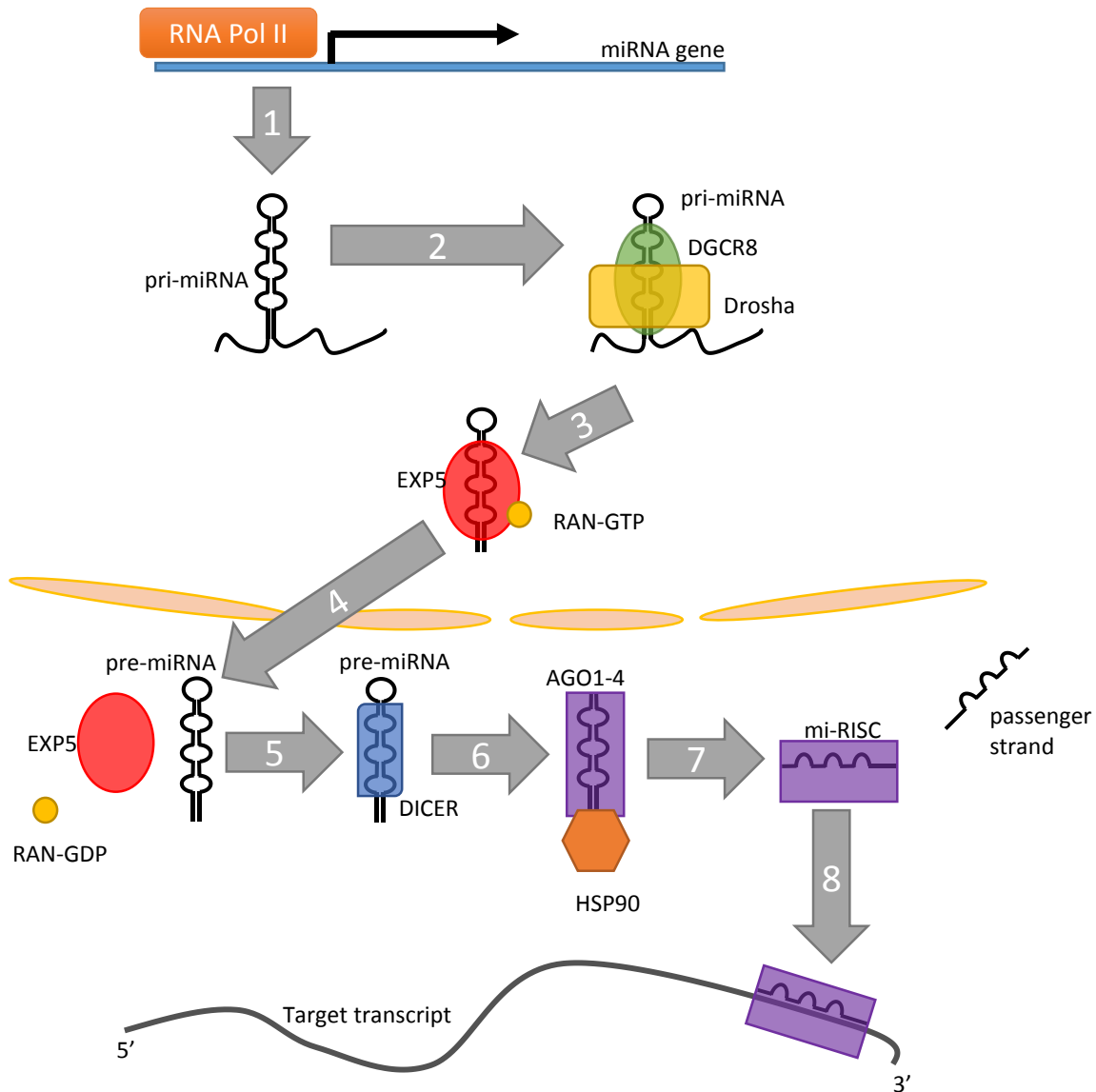
Biomarker identification requires large datasets such as those made available by TCGA, so as to identify significant features that stand out from the background. The process of identifying said biomarkers, as well as their biological relevance requires an understanding of bioinformatics, which utilizes statistical analysis for biological data. In this dissertation, I will be presenting the methods and principles of applying bioinformatics for the identification of transcript-based biomarkers in HPV-related cancers, and future directions for the clinical application of the results of this research.

## **1.2 MicroRNAs: History and Functions in Biological Processes and Diseases**

### **1.2.1 MicroRNA Discovery, Biogenesis, and Function**

MicroRNAs (miRNAs) are short (~22 nucleotide) single-stranded RNA sequences that function within the post-transcriptional regulatory process known as RNA interference (RNAi). MicroRNAs were first described in 1993 in the nematode *Caenorhabditis elegans*, when the single-stranded 22nt long gene *lin-4* was determined to regulate expression of the developmental gene *lin-14* through complementary binding in the 3' untranslated region (3'UTR) (18,19). A later study identified the developmentally miRNA *let-7* in *C. elegans*, which was found to also be conserved in multiple species, including *Homo sapiens*, indicating that microRNAs and their regulatory effects were not limited to nematodes (20,21). In the years since, miRNAs have been identified in 223 species, with 2588 high-confidence mature miRNA sequences in humans alone (22).

The biogenesis of microRNAs has been extensively reviewed by Ha and Kim (23). Canonically, after initial transcription by RNA-polymerase II, the pri-miRNA is processed



**Figure 1.1:** miRNA biogenesis and function. The miRNA gene is transcribed by RNA polymerase II [1] into the pri-miRNA, which is cleaved by the Drosha/DGCR8 complex to form the pre-miRNA hairpin structure [2]. The pre-miRNA is transported from the nucleus to the cytoplasm by exportin 5/RAN-GTP [3-4]. The hairpin is cleaved by Dicer [5] before the paired miRNA strands loaded into the Argonaute proteins [6]. HSP90 helps remove the passenger strand, resulting in the mature miRNA silencing complex [7], which can bind to the 3'UTR of target sequences and induce transcript degradation [8]. Image adapted from Ha and Kim (23).

within the nucleus by the Drosha-DGCR8 complex into the hairpin-shaped stem-loop pre-miRNA. The pre-miRNA is then exported to the cytoplasm by means of a transport complex comprised of exportin 5 and RAN-GTP. Once in the cytoplasm, the pre-miRNA is cleaved by

Dicer near the terminal loop, resulting in a small RNA duplex. The resulting strands form the 3p and 5p species of the mature miRNA. The duplex is loaded into the Argonaute RNA-induced silencing complex (RISC) with the aid of the HSP90 protein, after which the passenger strand is removed (24). The RISC is then guided to messenger RNA strands and prevents translation through inhibition or degradation of the target transcript (15,25) (Figure 1.1).

### **1.2.2 Prediction of MicroRNA Targets**

The selection of target mRNA sequences is driven primarily by complementary matching of bases 2-8 of the miRNA, or its seed sequence, to regions of the 3'UTR, which suggests that miRNAs are able to modulate the expression of multiple gene targets. This particular characteristic of miRNA-target interactions has driven the research community to identify likely miRNA-target pairings and additional distinguishing attributes, starting with TargetScan in 2005 (26). As initially developed as a standalone package by the Bartel group, TargetScan implemented the seed match as the primary factor in defining miRNA targets, and scored each potential target interaction based on the Gibbs free energy of the binding site; subsequent score ranking was used to define a cutoff that maximized signal-to-noise ratio (27). This first iteration identified species conservation as a strong contributing factor, as the ratio was increased when miRNA-target interactions were identified in the three species used for model training: humans, mice, and pufferfish. The second version of TargetScan, which accompanied the launch of the TargetScan web database, confirmed species conservation as a significant feature by expanding the training cohort to five vertebrate species, including dogs, chickens, and rats, but removing pufferfish. Subsequent editions built on this framework included features such as supplementary binding sites, target location in the 3'UTR, and GC content, as well as identifying potential

compensatory mechanisms for non-canonical (i.e. not based on seed-complementarity) binding, including target site abundance (28-30)

. The latest version, released in 2015, adds fourteen distinct features which further account for non-canonical binding, and implemented step-wise regression based on the Akaike information criterion (AIC), which characterizes data loss within models (31).

Although TargetScan is one of the better-known miRNA prediction algorithms, it is by far not the only one available to the research community. Additional resources have been created, including DIANA-microT, miRanda, RNA22, and MirTarget, which incorporate a variety of bioinformatics techniques to determine necessary and supplemental features in microRNA targeting (32-35). DIANA-microT also used stepwise-regression based on the AIC in 2012 to identify features used in non-canonical binding. miRanda utilized a base scoring system similar to genome alignment scoring: perfect matches scored highly, G:U wobbles were permitted and scored moderately, and alignment gaps are strongly penalized. RNA22 identified sites using pattern recognition, without directly implementing species conservation as a filter. MirTarget, which hosts its results in miRDB, performs feature selection using support vector machine (SVM), a form of supervised learning that maximizes separation between two groups in a multidimensional space. The latest version of MirTarget identified 50 relevant features for miRNA target prediction through recursive feature elimination, and weights for each feature were calculated by SVM to generate a score.

Across the majority of prediction algorithms, certain key features are identified as crucial for miRNA-targeted RNAi: seed sequence complementarity, species conservation, Gibbs free energy of RNA-RNA binding, and target site accessibility (36). Additional features include supplementary binding sites, as illustrated by TargetScan and MirTarget, and specific features

affecting site accessibility, such as nucleotide composition of the target site and neighboring regions, and location of the target site within the 3'UTR (35,37). Despite algorithmic differences, the identification of miRNA targets requires the integration of both wet and dry labs, i.e. the experimental detection of targets and the computational resources to determine the factors that influence true miRNA-target interactions. Experimental techniques used to both train and validate these prediction resources range from low-throughput methods such as luciferase assays to high-throughput methods that include microarrays and next-generation RNA sequencing after artificial miRNA dysregulation (38,39). The methods for manipulating miRNA expression can either increase miRNA levels, such as through miRNA overexpression, or decrease miRNA levels, which has been performed through miRNA sponges and, more recently, CRISPR-Cas9 gene editing (40,41). More recent techniques for identifying miRNA targets without overt manipulation of expression levels include CLIP-ligation and sequencing, in which the miRNA and target strands are cross-linked to the Argonaute protein, ligated to create a hybrid, and sequenced after Argonaute-immunoprecipitation (42,43).

### **1.2.3. MicroRNAs as Biomarkers for Disease**

Expression changes in miRNAs have been associated with developmental growth in numerous organisms, such as nematodes, fruit flies, and zebrafish, in addition to mammals, including humans and mice (18,44-46). This suggests that miRNAs can be temporally expressed specifically to regulate certain cellular and physiological functions. Subsequent changes in miRNA expression have also been associated with cellular differentiation into various tissue types (47). By the same token, dysregulation of miRNAs may result in atypical phenotypes and presentation of various diseases, such as Alzheimer's disease and cancer (48,49). Although

miRNA expression changes may not necessarily be the driving factors behind disease phenotypes, their roles as post-transcriptional regulators cannot be discounted in the propagation of disease.

The roles of miRNA in cancer development and progression have been extensively reviewed in the literature, with various focuses such as: the roles of miRNAs on general cellular dysregulation leading to tumor formation (50-53); diagnostic applications of miRNAs such as in tumor classification and patient prognosis (54-58); and their potential as therapeutic targets (54,56,58,59). In the context of tumor development, miRNAs can be expressed and dysregulated in a manner similar to known oncogenes or tumor suppressors. As a result, miRNAs that are overexpressed in tumor tissue compared to normal tissue have been described as “oncomiRs.” Considering the inhibitive regulatory mechanisms in which miRNAs are involved, oncomiRs typically target tumor suppressor genes, while tumor suppressive miRNAs target oncogenes. This sort of relationship has been observed with the better known cancer-related miRNAs. For example the classic oncomiR miR-21-5p, which has been reported as upregulated in glioblastoma, acute myeloid leukemia, breast cancer, and prostate cancer, among others, is known to target the tumor suppressor PTEN, PCDC4, TPM1 and TIMP3 (50). Comparatively, miR-34 acts as a tumor suppressor, targeting the cell cycle activators CDK4, CDK6, cyclin E2, EZF3, and met (50). It should be recognized, however, that the role of the miRNA as an oncomiR or a tumor suppressor can be tissue-dependent. miR-221 and miR-222 both are confirmed to be upregulated and target the oncogene KIT in erythroblastic leukemia, thereby functioning as tumor suppressor. However, in other tumor types, confirmed targets of miR-221 and miR-222 include the tumor suppressors p27, p57, PTEN and TIMP3, and upregulation of the miRNAs resulted in inhibition of expression (58). Additionally, the relevant categorization of the

miRNA as an oncomiR or tumor suppressor is dependent on its directional dysregulation in tumor tissue.

The expression levels of miRNAs can also be evaluated in a diagnostic context without requiring an in-depth understanding of potential targets and resulting cellular response. In this capacity, miRNAs can serve as biomarkers, providing insight into patient risk for developing cancer, or further risk of cancer death. A number of studies have identified single miRNAs as potential biomarkers for cancer prognosis. For example, increased miR-126 expression has been correlated to metastasis in renal clear cell carcinoma while upregulation of miR-31 in cervical and oropharyngeal cancers is negatively associated with patient survival (41,60,61). Some groups have also proposed using a panel of miRNAs as a diagnostic indicator, i.e. using the cumulative expression profile in the form of a prognostic signature. Such signatures have been derived for a variety of cancers, including cervical cancer (62,63), oropharyngeal cancer (64-66), and bladder cancer (67). The primary advantage of using multiple miRNAs in a clinical diagnostic panel is the ability to compensate for technical and biological variability both in screening techniques and patient cohorts. Even so, not all signatures can be validated despite the use of rigorous and comprehensive bioinformatics pipelines on large-scale patient cohorts.

The implementation of miRNAs as potential therapeutic targets is a relatively new field and can be considered a branch of gene therapy in the sense that therapeutic goals are the inhibition of undesirable miRNA transcripts or enhancement and restoration of pro-survival miRNAs (68). Inhibition of miRNAs has been performed *in vitro* through antagomiRs, locked nucleic acid constructs (LNAs), anti-sense nucleotides and sponges (69). Among animal models for *in vivo* miRNA inhibition, experiments have tested: the delivery of miR-10b antagomiRs to prevent metastasis in tumor bearing mice (70); LNA inhibition of miR-122 to improve hepatitis



C outlook in primates (71); and miRNA target saturation with sponges (72). Similarly, tumor suppressive and pro-survival miRNAs have been explored in the context of therapies that restore their functions. Mouse models have shown that delivery of let-7 miRNAs through lentiviral constructs and intravenous lipid emulsions were able to reduce tumor burden in mice (73,74), and restoration of the KRAS targeting miRNAs miR-143 and miR-145a with a nanovector delivery resulted in the reduction of xenografted pancreatic tumors (75). Recent research has shown that miRNAs can be inhibited *in vivo* through CRISPR constructs (41); given the latest research showing the CRISPR can be used to edit human embryos, there may be future applications of CRISPR to either inhibit or restore miRNA expression at the genomic level (76).

Despite their distinctively short sequence length, microRNAs function in a significant role in general cellular biology. As the crucial targeting member of the regulatory RISC body, miRNAs aid in controlling temporal cell growth and development, as well as later roles in maintenance of protein expression. Dysregulation of such a diverse controller can lead to disease formation, which has driven research to determine targets of miRNAs, mechanisms by which miRNA expression changes occur, and methods to adjust for these changes. Simultaneously, the relative expression of miRNAs can be explored within the clinical setting to aid diagnosticians in determining course of treatment, without requiring intervention in the miRNome. As a whole, the miRNome is a crucial aspect of the human transcriptome, and should not be overlooked in how it may be explored in the contexts of both general physiology and the course of disease.

## **1.3 The Role of Human Papillomavirus in Tumor Formation**

### **1.3.1 A Brief History of Oncoviruses**

Viruses are a class of infectious agents that rely on host cells for life cycle progression and replication. The official classification of viruses is the responsibility of the International Committee of the Taxonomy of Viruses (ICTV), which assigns viruses to one of eight orders or otherwise defines a virus order as “unclassified,” based on biological properties such as pathogenicity and epidemiology, and sequence relationships such as divergence phylogeny (77,78) (Table 1.3.1). Other classification systems include that proposed by David Baltimore in 1971, which classifies viruses based on nucleic acid (i.e. RNA or DNA), strandedness (single or double), transcription direction (sense or antisense), and method of replication (e.g. reverse transcription) (79,80). As of 2017, the eight major ICTV orders can be classified into three different Baltimore groups, which, in conjunction with the high number of unclassified viral families, suggests that further classification is ongoing (Table 1.2).

Viral replication can be described as either lytic or nonlytic. Lytic viral infection typically follows a five step process of adsorption, penetration, replication, assembly, and release (81). Adsorption and penetration describe the process by which the virus infects the host cell, and viral mRNAs are produced in the replication phase, either through viral enzymes or host transcription

**Table 1.1.** ICTV classification of viruses into major orders

<b>Order</b>	<b>Number of families</b>	<b>Number of species</b>	<b>Example family [f], genus [g], or species [s]</b>
<i>Bunyvirales</i>	9	157	[f] <i>Hantaviridae</i>
<i>Caudovirales</i>	3	956	[g] <i>T4virus</i>
<i>Herpesvirales</i>	3	103	[s] <i>Human gammaherpesvirus 4</i> (Epstein-Barr virus)
<i>Ligamenvirales</i>	2	11	[g] <i>Rudivirus</i>
<i>Mononegavirales</i>	9	212	[g] <i>Ebolavirus</i>
<i>Nidovirales</i>	4	64	[s] <i>SARS coronavirus</i>
<i>Picornavirales</i>	6	196	[s] <i>Foot-and-mouth disease virus</i>
<i>Tymovirales</i>	4	180	[g] <i>Trichovirus</i>
<b>Unassigned</b>	85	2525	[f] <i>Papillomaviridae</i>
<b>Total</b>	125	4404	

**Table 1.2.** Baltimore classification of viruses

Group	Name	Abbreviation	Known orders	Example viruses
<b>I</b>	Double-stranded DNA viruses	dsDNA viruses	<i>Caudovirales</i> <i>Herpesvirales</i> <i>Ligamenvirales</i>	Human papillomavirus, herpesviruses, adenoviruses
<b>II</b>	Single-stranded DNA viruses	ssDNA virus		Parvoviruses
<b>III</b>	Double-stranded RNA viruses	dsRNA viruses		Rotavirus
<b>IV</b>	Sense single-stranded RNA viruses	(+)ssRNA viruses	<i>Nidovirales</i> <i>Picornavirales</i> <i>Tymovirales</i>	Rubella virus
<b>V</b>	Antisense single-stranded RNA viruses	(-)ssRNA viruses	<i>Bunyavirales</i> <i>Mononegavirales</i>	Rabies virus
<b>VI</b>	Single-stranded RNA reverse-transcribing viruses	ssRNA-RT viruses		Human immunodeficiency virus
<b>VII</b>	Double-stranded DNA reverse-transcribing viruses	dsDNA-RT viruses		Hepatitis B virus

factors; in either case, the host translation machinery is utilized to translate viral mRNA transcripts. Viral progeny is then assembled from the translated proteins and released in the final phase, which causes cell lysis (81). Nonlytic viruses include retroviruses, a special class that are able to integrate their own genomic sequences into the host genome. Transcription by the host produces the viral progeny, which can be then be released through exocytosis without necessitating host death and lysis (81).

The degree to which various virus families hijack host cellular components for replication can also be used in classification. DNA viruses (Baltimore Classes I and II) require the host transcription machinery to produce viral mRNA transcripts before translation (79). Double stranded RNA viruses (Class III) contain enough information for both protein synthesis and replication while single-stranded RNA viruses (Classes IV and V) necessitate host production of template strands (79). Both RNA and DNA retroviruses (Classes VI and VII, respectively), integrate into the host genome through transcription, which makes viral clearance by the host much more difficult (79,80). For these two particular viral classes, integration is crucial to replication and survival; however, genomic integration of viral sequences has been observed by

other classes, such as the Class I human papillomavirus, as a result of chromosomal instability and recombination of cellular and viral genome fragments (82).

In addition to utilizing the transcription and translation machinery, viruses will also alter the cellular environment in order to favor viral replication. Examples include: the degradation of host mRNA via viral endoribonucleases by alpha herpesviruses, so as to reduce competition for translational machinery (83); viral stimulation of the cell cycle to improve environmental conditions for replication by polyomaviruses and adenoviruses (84,85); and competition for translational machinery through internal ribosome entry sites by hepatitis C virus (86). Additional viral responses may interfere with immune response or apoptotic signaling (87,88). The alteration of the internal environment to favor the viral life cycle is rarely to the benefit of the host, and in some organisms, can lead to abnormal cell growth and replication, and subsequent tumor formation. The subset of viruses that are capable of inducing such transition has accordingly been termed “oncoviruses.”

The first known oncovirus was described in 1911 by Peyton Rous, who discovered the Rous sarcoma virus in chickens (89). The first human virus linked to tumorigenesis was the Epstein-Barr virus (EBV) which was strongly correlated with Burkitt’s lymphoma in the 1960s, and later to nasopharyngeal carcinoma (90,91). Through the 1970s and 1980s, the list of oncoviruses grew to include human papillomavirus (HPV) as a potential causative factor in cervical cancers, hepatitis B virus (HBV) as associated with hepatocellular carcinoma, and human T-cell leukemia virus type 1 (HTLV-1) (92-96). The existence of hepatitis C virus (HCV) was initially proposed in 1975, but was not confirmed until almost 15 years later, as well as its potential role in the development of hepatocellular carcinoma (97-99). More recent technologies have led to the identification of Kaposi’s sarcoma associated herpesvirus in the 1990s and the

role of Merkel cell polyomavirus in Merkel cell carcinoma (100,101). To date, these seven viruses make up the breadth of known human oncoviruses. Additional oncogenic viruses have been discovered in other species, including Marek's disease virus in chickens, simian vacuolating virus 40 (SV40) in some animal models, and feline leukemia virus in chickens (102-104).

The primary mechanisms of infection and tumor formation can vary between the seven human oncoviruses (reviewed in depth by White *et al.* (11) and Mesri *et al.* (105)). Epstein-Barr virus is well-characterized as the causative agent of classical acute infectious mononucleosis; transmission primarily occurs via saliva, as well as potentially through sexual contact (106). The hepatitis viruses, despite being of different families, are both transmitted through bodily fluids via interactions such as sexual intercourse or shared intravenous needles (107,108). HPV transmission is through mucosal and skin-to-skin contact, the latter of which is further

**Table 1.3.** Human oncoviruses and associated viral oncoproteins

<b>Virus</b>	<b>Abbreviation</b>	<b>Family and Baltimore classification</b>	<b>Associated cancers</b>	<b>Selected viral oncoproteins</b>
<b>Epstein-Barr virus</b>	EBV	<i>Herpesviridae</i> I (dsDNA)	Burkitt's lymphoma Nasopharyngeal carcinoma	EBNA1 LMP1 LMP2A LMP2B
<b>Kaposi's sarcoma-associated herpesvirus</b>	KSHV	<i>Herpesviridae</i> I (dsDNA)	Kaposi's sarcoma	LANA LAMP
<b>Hepatitis B virus</b>	HBV	<i>Hepadnaviridae</i> VII (dsDNA-RT)	Hepatocellular carcinoma	HBx
<b>Hepatitis C virus</b>	HCV	<i>Flaviviridae</i> IV ((+)ssRNA)	Hepatocellular carcinoma	HCV core protein NS3
<b>Human T-cell leukemia virus type 1</b>	HTLV-1	<i>Retroviridae</i> VI (ssRNA-RT)	Adult T-cell leukemia	Tax
<b>Merkel cell polyomavirus</b>	MCV	<i>Polyomaviridae</i> I (dsDNA)	Merkel cell carcinoma	Large T-antigen Small t-antigen
<b>Human papillomavirus</b>	HPV	<i>Papillomaviridae</i> I (dsDNA)	Cervical cancer Anal cancer Penile cancer Oropharyngeal cancer	E6 E7

compounded by epithelial microabrasions (109). Transmission of HTLV-1 is conducted primarily through mother-to-child interactions, sexual intercourse, or blood transfusions, the latter of which has been mostly controlled for by screening donated blood (110). KSHV is transmitted in children through saliva, and in adults through high-risk sexual activity (111). The least understood mechanism of oncovirus transmission is that of MCV; current hypotheses suggest that dermal fibroblasts are infected by MCV and are either transformed into Merkel cells or infect neighboring cells (112). The viral mechanisms leading to tumorigenesis are much more clearly understood, and encompass a variety of cellular modifications and responses (Table 1.3).

Both Epstein-Barr virus and Kaposi's sarcoma-associated herpesvirus are members of the *Herpesviridae* family (Baltimore Group I), which are characterized by their latency states. EBV expresses the oncoproteins EBV nuclear antigen (EBNA1) and latent membrane proteins 1 and 2 (LMP1 and LMP2A/B). The primary function of EBNA1 is to promote viral proliferation. By binding to host DNA and regulating host cellular transcription, EBNA1 encourages EBV episomal retention and segregation, as well as preventing cell death (113). LMP1 and LMP2 are two latency associated viral proteins that mimic oncogenic proliferative signals. LMP1 imitates an active CD40 receptor, which recruits TRAFs and causes NF- $\kappa$ B activation, a known factor in lymphoma formation (114). The LMP2 proteins mimic a cross-linked Ig receptor, which leads to the activation of the PI3K-Akt-MTOR pathways, which in turn promotes B-cell differentiation, prolonged survival, and cell growth (115,116). Similarly, KSHV expresses the KSHV latency associated nuclear antigen (LANA), which connects the KSHV episome to the host (117). In doing so, KSHV also compromises a number of cellular regulators, including p53, pRb, GSK-3 $\beta$ , and p300 (118). KSHV also expresses latency associated membrane protein (LAMP), which

resembles LMP1 in function and activates anti-apoptotic pathways, including Ras/MAPK, NF- $\kappa$ B, and vIRF3 (119).

The hepatitis viruses HBV and HCV are members of different viral families linked only by their targeted infection of the liver. HBV is a double-stranded DNA hepadnavirus (Baltimore Group VII) that encodes the viral oncoprotein HBx, which can stimulate cell cycle entry, as well as survival pathways such as Ras and NF- $\kappa$ B (120). Regarding apoptosis, HBx can either prevent apoptosis by blocking caspase activity and directly interacting with p53, or or promot TNG-mediated cell death, potentially to increase hepatocyte regeneration (120). HBx is also capable of activating the pRb-E2F1 oncogenic pathway through p16<sup>INK4a</sup> inactivation via phosphorylation (121). In comparison, HCV is a single-stranded sense RNA virus (Baltimore Group IV), and as such, is not integrated into host genome. The primary viral proteins behind HCV-driven oncogenesis are currently believed to be the HCV core protein and nonstructured protein 3 (NS3). HCV core protein has been shown to interact with a number of transcription factors and regulators, including p53, p21, and NF- $\kappa$ B, as well as the Ras/Raf/MAPK pathway, while NS3 has been shown to bind and inactivate p53 (122,123).

HTLV-1 is an RNA retrovirus (Baltimore Group VI) that encodes a variety of oncoproteins. The Tax protein is particularly effective in activating cell proliferation and survival through various mechanisms. Tax binds directly to CREB to induce and increase viral transcription, which is accentuated by the recruitment of p300 (124). Tax is also able to activate NF- $\kappa$ B to induce cellular transformation, as well advance cell cycle progression by activating cyclin E and Cdk2, stabilizing cyclins D2 and D3 and the cyclin D/Cdk4 complexes (124). Through these mechanisms, Tax is able to induce phosphorylation of pRb and E2F release (124). HTLV-1 also encodes HBZ, which is implied to activate the transcription of the pro-survival

genes JUND, JUN, and ATF, as well as promoting transcription of E2F to induce cell proliferation and hTERT expression to confer cell cycle immortality (124,125).

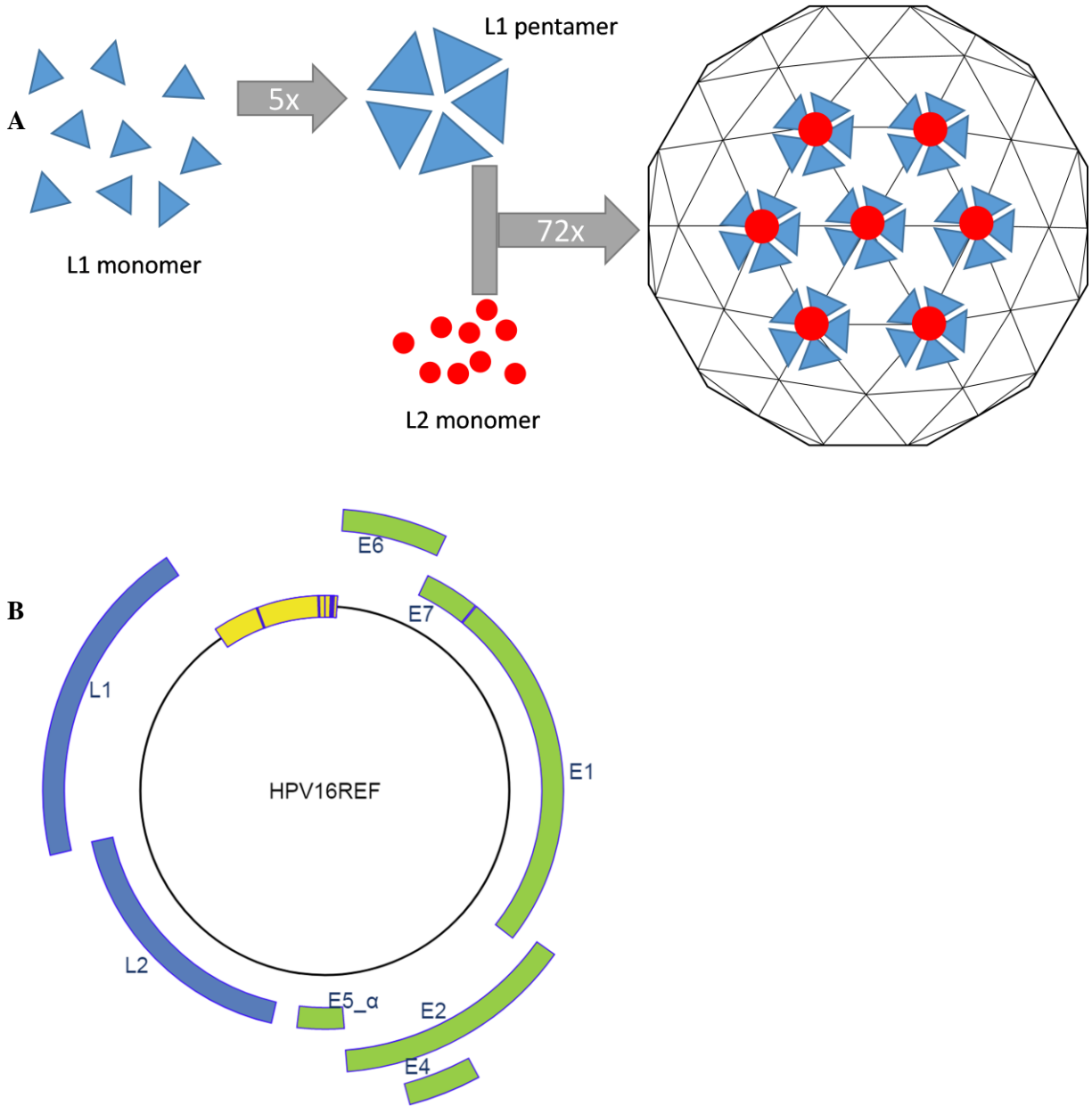
Merkel cell polyomavirus is a polyomavirus (Group I) that was discovered to be integrated into the genome in most Merkel cell carcinomas. Given the relatively recent discovery of MCV as a causative agent of MCC, the mechanisms by which MCV integration induces tumor formation are still under investigation. Another tumorigenic polyomavirus, SV40, has provided some guidance, particularly with a focus on large T- and small t-antigen functions (126). Co-immunoprecipitation experiments have shown that large T-antigen is bound to pRb, thereby inducing cell cycle progression, as well as disrupting lysosomal clustering by binding to Vam6p (127). The MCV small t-antigen also binds to and hyperphosphorylates the transcription initiation factor 4E-BP1 to dysregulate cap-dependent translation, and binds to the E3 ubiquitin ligase Fbw7, resulting in increased c-myc and cyclin E activity (128,129).

### **1.3.2 Functions of Human Papillomavirus Proteins**

Papillomaviruses (PVs) are a family of double-stranded DNA viruses with genomes that are approximately 8 kilobases long, encased in an icosahedral capsid (Figure 1.2) (130,131). Taxonomically, the Papillomaviridae family consists of 48 genera (with one “unclassified” genus), encompassing 123 clearly defined species, with an additional 18 species presently defined as “unclassified” (132). The PV species are divided into types, which can be further classified into subtypes and variants (130), 183 of which are human specific (i.e. human papillomavirus (HPVs)) and are members of the genera alphapapillomavirus, betapapillomavirus, gammapapillomavirus, mupapillomavirus, and nupapillomavirus. High-risk HPV types, specifically those associated with oncogenesis, are typically in the alpha-



papillomavirus genus. HPV infects the basal layer of mucosal and epithelial cells, normally through microlesions and microabrasions (133). The most common visible symptoms of HPV

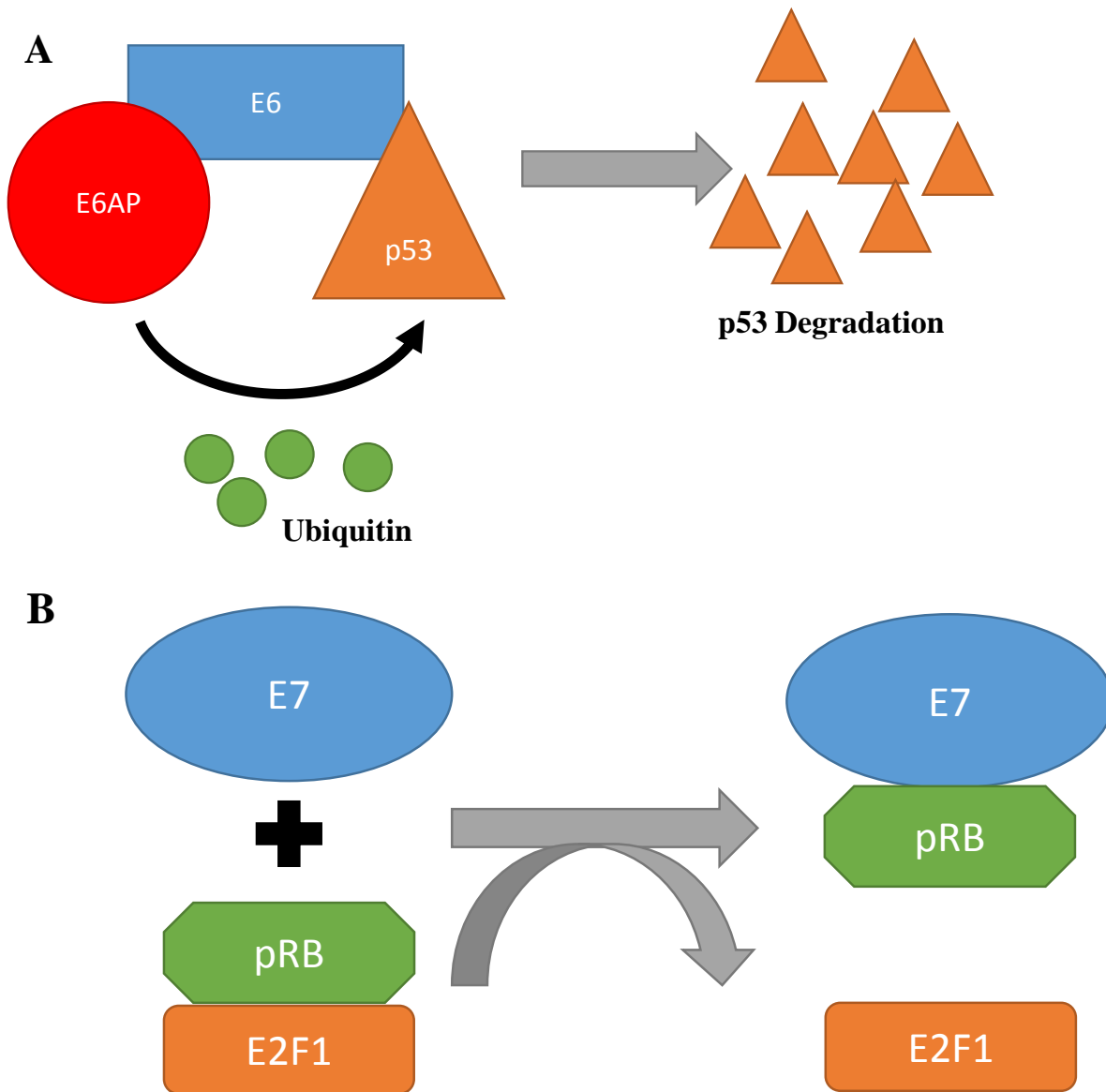


**Figure 1.2.** Structure of the HPV viroid and genome. (A) L1 protein monomers self-assemble into pentamers and L2 monomers reside in the center of the pentamers. 72 capsomers form the icosahedral HPV virion. Adapted from Schiller and Muller (209). (B) The HPV16 genome, as a representative HPV type, is circular and encodes for eight proteins: E1, E2, E4, E5, E6, E7, L1, and L2. Adapted from the Papillomavirus Episteme ([pave.niaid.nih.gov](http://pave.niaid.nih.gov)) (128).

are benign tumors, namely warts and papillomas. However, symptoms are often invisible to the naked eye, taking the form of microlesions. Overt lesions often are the result of immune suppression; otherwise, the majority of HPV types coexist silently with the host (133).

A typical papillomavirus encodes 7 proteins: the early proteins E1, E2, E4, E6, and E7, and the late proteins L1 and L2; some PV types, including those that infect humans, encode an additional early protein, E5 (134). The HPV replication cycle progresses through a number of well-characterized steps, which can be expounded on in the context of viral protein expression and function (reviewed concisely by Graham (134)). After infection of basal cells and successful translocation to the nucleus, the first viral genes to be expressed are E1 and E2. Functionally, E1 and E2 work in conjunction to bind to the viral origin of replication. E2 acts as a tether and recruits E1 to the binding site, and E1 recruits cellular transcription and replication factors for viral genome amplification (135,136). E2 also maintains viral genome levels by tethering the viral genome to host chromatin binding proteins, as well as temporally limiting viral transcription to prevent immune activation (136). E6 and E7 are transcribed initially during the early stages of the replication cycle, but transcript and protein expression levels are not likely to be depleted until the virus is cleared (134). Despite their primary recognition as the HPV oncogenes, their roles are vital for viral replication. E6 and E7 function in tandem to promote cell proliferation, and subsequently viral proliferation, without inducing apoptosis (137,138). The mechanisms by which this is performed can lead to oncogenesis when left unchecked. The E4 and E5 proteins are not as well studied as the other HPV proteins, but their general functions have been elucidated. E4 is encoded within E2; it contributes to genome amplification and capsid synthesis, and may arrest cell-cycle in G2 phase (139). E4 has also been noted for its abundant expression in upper epithelial layers, and consequently been proposed as a biomarker for HPV

infection (139). E5 is not expressed by all papillomaviruses, but in those that do, it is shown to have a variety of possible functions (140). It is weakly oncogenic, possibly supplementing E6 and E7 activity as a cofactor by upregulating EGFR signaling pathways, as well as possibly



**Figure 1.3.** Oncogenic activity of the HPV E6 and E7 proteins. (A) The E6 protein induces tumor formation by binding to the E6 associated protein (E6AP) and the tumor suppressor p53. E6AP then recruits ubiquitins to target p53 for degradation. (B) The E7 protein functions by binding to the tumor suppressor pRB, which releases the transcription factor E2F1. E2F1 is then free to transcribe genes associated with cell cycle progression. Adapted from Yim and Park (140).

impairing intercellular communication by binding with vacuolar ATPase. This latter property, along with inhibition of HLA-1 intracellular transport, may contribute to immune evasion (140).

The late proteins are the structural capsid proteins for HPV. The major capsid protein L1 is a protein of approximately 55 kilodaltons that can spontaneously self-assemble into virus-like particles (131). The L1 proteins form a pentameric capsomer; 72 capsomers self-assemble into the icosahedral virion (131). Each capsomer binds to an L2 minor capsid protein, in such a manner that the mature virion keeps most the L2 protein body concealed below the capsid surface (131,141). When the virus encounters the host cell surface, L1 interacts with heparin sulfate carbohydrates on basement membrane proteoglycans (131). This induces a conformational change in L1 that exposed L2. L2 cleavage by furin allows for a conformational change that in turn binds to a secondary receptor on the cellular plasma membrane (141). Subsequent cell entry resembles micropinocytosis, after which the virion is transported to the nucleus through membrane bound cellular components and tubulin transport (134). Nuclear entry can occur either through nuclear pores or in the midst of nuclear envelope dissolution during mitosis (134).

As mentioned previously, E6 and E7 contribute to the viral life cycle but are also the primary drivers of HPV-induced oncogenesis and have been extensively reviewed in the literature (Figure 1.3) (142-145). The classical mechanism by which E6 induces oncogenesis is also arguably the best described function of the protein. E6 targets and binds proteins with an LXXL motif (137). Included among these is the E6 associated protein (E6AP), an ubiquitin ligase encoded by the UBE3A gene. This interaction results in a conformational change in E6 that allows it to recruit the regulatory protein p53, forming a ternary complex. In doing so, ubiquitin peptides are transported from E6AP to p53, turning p53 into a target for degradation

(146). As a tumor suppressor, p53 is involved in promoting cell cycle arrest and apoptosis in the event of cellular stress; subsequently, p53 degradation can result in cell immortality and eventually tumor formation (147). E6 is also capable of targeting other cellular regulatory proteins, including p300/CBP, HIF-1 $\alpha$ /HIF-2 $\alpha$ , and MAML1 (148). It should be noted that these latter interactions are more often observed in lower-risk HPV types, and certain HPV types also exhibit preferential binding; experiments where E6 binding partners are coexpressed demonstrated that high-risk E6 proteins bind to E6AP more often than other candidates, and the low-risk HPV E6 prefers alternate candidates (148,149).

Another observed E6 interaction is the activation of the telomerase enzyme hTERT, which adds telomere repeats to the ends of chromosomes, and essentially conferring immortality (150). This particular response is notable for its independence from E6-p53 activity, but it does appear to require E6AP, at least in high-risk HPV types (151,152). E6 alterations of cellular transcription levels are mediated by its interaction with histone acetyltransferases, including degradation of Ada3 and Tip60, along with p300 (137). Further compromising cellular homeostasis are E6 effects on PDK1 and mTORC2, which lead to the activation of Akt and the mTORC1 signaling pathway, resulting in increased metabolism (153,154). E6 is also known to affect immune response by activating NF- $\kappa$ B, increase proliferation and prevent apoptosis by binding to and degrading Bak, and inhibit additional apoptotic signaling cascades by binding to procaspase 8 (155-157). The interferon signaling cascade is also inhibited by E6 binding Tyk2, a member of the Jak/Stat signaling pathway, and IRF3, thereby preventing activation of interferon responsive genes (158,159).

The primary oncogenic mechanism of E7 is through its interaction with the pRB pathway, by binding directly to the retinoblastoma protein (pRB) at an LXCXE domain, leading

to its phosphorylation and degradation (160,161). Reduced pRB level leads to the release of E2F, a transcription factor that controls for proliferative genes. An additional function of E2F is the activation of p53, which under normal circumstances would increase the likelihood of cell death; however, this is mitigated by E6-mediated p53 degradation (142). E7 has been proposed as the primary oncogenic mechanism of HPV, as transfection of the E7 gene initiated benign tumor growth, while E6 induced the conversion from benignity to malignancy (162). Recent research has also found that E7 conservation was crucial to tumor formation, supporting the hypothesis of E7 as the primary oncogenic factor; a significantly lower frequency of variants was identified in patients whose HPV infection progressed to cervical cancer (163). Supplementary E7 functions leading to tumor formation include binding and inhibiting cyclin-dependent kinase inhibitors, including p21<sup>Cip1</sup> and p27<sup>Kip1</sup> (164,165).

### **1.3.3 Human Papillomavirus and Other Oncoviruses Can Alter MicroRNA Expression**

Oncoviruses are not restricted to direct interaction with cellular regulatory proteins in tumorigenesis; further indirect mechanisms, such as RISC, can be affected as well. The first discovery of viral encoded miRNAs was reported in B-cells infected with EBV (166). In the years since, over 500 miRNAs have been identified as virally encoded and curated in miRBase (22). The majority of functional viral miRNAs have been identified in members of the herpesvirus family, some of which have been previously described as oncoviruses; other potential viral miRNAs have been identified in polyomaviruses and adenoviruses (reviewed by Roberts, Lewis, and Jopling (167), and Skalsky and Cullen (168)).

Five major species of herpesvirus have been identified as coding functional miRNAs. Of note are the gammaherpesviruses, EBV and KSHV. Since the first discovery of viral miRNAs in EBV, 25 total pre-miRNAs have been reported (169,170). Notable miRNAs in EBV are miR-BART2, miR-BART5, and miR-BHRF1-3. miR-BART5 targets the cellular proapoptotic protein PUMA while miR-BHRF1-3 targets the cellular T-cell attractant CXCL11 (171,172). miR-BART2 has been shown to target MICB, thereby improving immune evasion (173). Viral targets of EBV miRNAs include the lytic gene BALF5 by miR-BART2, presumably to stabilize viral latency, and LMP1 by three other EBV miRNAs (174,175). KSHV expresses a total of 12 pre-miRNAs (168). Viral-viral interactions include the targeting of RTA by miR-K12-9-5p, which prevents early entry into the lytic cycle (176). Host cellular target interactions with KSHV miRNAs include: MICB by miR-K12-7; the apoptotic protein BCLAF1 by miR-K12-5, -9 and -10; the T-cell attractant THBS1 by miR-K12-1, -3-3p, -6-3p, and -11; p21 by miR-K12-1; the transcriptional repressor MAF by miR-K12-6 and -11; and the transcriptional repressor BACH1 by miR-K12-11 (173,177-181).

As previously indicated, SV40 is a polyomavirus that has also been shown to be potentially oncogenic; it also encodes a single pre-miRNA that targets the viral large T-antigen, which improves immune invasion (182). Both the miRNA and target are conserved in other polyomaviruses, including MCV, human BK virus, and JC virus (183,184). Some adenoviruses have also been shown to produce non-coding RNAs, VAI and VAII, that interact with DICER and the miRNA RISC complex (185,186). The only known direct target is the proapoptotic RNA metabolism factor TIA-1, but no significant effects were observed (187).

Alterations in miRNA activity in response to viral infection is not limited to virally encoded miRNAs, as the host miRNome also undergoes dysregulation. The interactions can be

essentially be classified into one of two categories: viral regulation of cellular miRNAs, and host regulation of viruses through miRNA-RISC. It should be noted, though, that sometimes both interactions are observed simultaneously.

Virally induced dysregulation of miRNA expression typically results in a response favorable to viral survival, such as increased immune evasion or improved environment for replication. Among the known oncoviruses, EBV upregulates miR-155 and miR-146a expression, both of which have been shown to target TRAF and IRAK1, which results in inhibited immune response (188-190). EBV also upregulates miR-29b, which targets TCL1, affecting cell survival, and miR-21, which is upregulated in several cancer types (175,191). KSHV, as previously described, encodes miR-K12-11, which is a functional mimic of miR-155; both the host and the viral mimic miRNAs target transcription factors, and subsequently contribute to oncogenesis (181). This latter viral miRNA is also observed in other herpesviruses shown to cause cancer in other organisms, such as Marek's disease virus in chickens (192). HCMV downregulates miR-100 and miR-101, which have been shown to inhibit HCMV replication (193). Among non-herpesvirus miRNA effectors, HIV-1 was shown to promote replication by downregulating members of the miR-17/92 cluster (194).

Host miRNAs are also capable of targeting viral genes, although the infection agent may take advantage of the host response. This is observed in HCMV and HSV-1, as miR-200 targets the HCMV protein IE2 and miR-138 targets the HSV-1 protein ICP0; downregulation of these two proteins help promote viral latency and survival (195,196). Not all miRNA responses are to the detriment of the host, as miR-29a is upregulated in response to HIV infection, and specifically targets the HIV-1 transcript for degradation (197). Hepatitis C virus was also shown to indirectly induce the overexpression of a number of miRNAs, including miR-196, miR-296,



and miR-351 through increased IFN- $\beta$  activity, but these miRNAs appear to modulate HCV replication (198).

HPV has been shown to dysregulate a number of miRNAs, notably downregulating miR-143, miR-145, miR-34a, and miR-203, resulting in increased cell motility (199-202). miR-145 downregulation also results in genome amplification associated with cellular replication and growth (203). The downregulation of miR-218 by HPV results in increased translation of its target LAMB3, a protein specific to epithelial cells that may play a role in differentiation and oncogenesis (204). Upregulated miRNAs include miR-9, which has been shown to be upregulated in HPV(+) cervical cancers by the E6 protein, especially by HPV16 (205). HPV16 E6 also specifically represses miR-23b in cervical cancer, which is an apoptotic tumor suppressor that regulates the anti-apoptotic oncogene c-MET (206). Additional HPV-associated miRNAs in cervical cancer include the upregulated miR-16, miR-25, miR-92a, and miR-378, and the downregulated miR-22, miR-27a, miR-29a, and miR-100 (207). Within oropharyngeal cancers, HPV infection is associated with upregulation of miR-9, similar to the observation in cervical cancer, and miR-155, along with downregulation of miR-31, miR-223, and miR-18a (64). A few analyses have looked at the role of HPV in miRNA expression changes for both oropharyngeal and cervical cancers, and have concluded that some similarities exist between the two cancer types, particularly upregulation of miR-10b, miR-16, and miR-20b, along with downregulation of miR-145, miR-199a and miR-199b (208). Direct expression of E6 and E7 proteins from HPV16 in non-tumor tissues yielded similar results, as well as demonstrating how characteristic oncomiRs and tumor suppressive miRs such as miR-203a can influence the transcriptome (209).

## **1.4 Characterization of Cervical and Oropharyngeal Cancer in the Context of HPV**

In the context of cervical cancer, Harald zur Hausen proposed a link between tumor formation and HPV infection status in 1974; this research garnered him the Nobel Prize in medicine in 2008 (92,93). Cervical cancer is the most common gynecological tumor in the world, with approximately 528,000 new cases diagnosed annually (210). Of these cases, it is estimated that 95% are the result of HPV infection (211). This direct causation has spurred research into HPV vaccines with the intent of reducing overall risk of cervical and other anogenital cancers, leading to the first marketed vaccines: Gardasil, from Merck, inoculated against HPV types 6, 11, 16, and 18; and Cervarix from GlaxoSmithKline, protected against types 16 and 18 (212,213). HPV types 16 and 18 alone are estimated to be responsible for 70% of new cervical cancers. The next generation of Gardasil also protects against the high-risk HPV types 31, 33, 45, 52, and 58, thereby increasing the coverage to viral types responsible for up to 90% of potential cervical cancers (214).

### **1.4.1 Tumor Source Site and Genomics Affect Cervical Cancer Prognosis**

Despite its oncogenic properties, HPV infection has also been shown to be a positive prognostic marker for overall patient tumor outcome (215-218). The mechanisms behind this duality are not entirely understood, which has encouraged deeper research into the genomic alterations that result from and occur independently of HPV infection. Some of these differences can be attributed to variance in tissue source site, as cervical cancer can be categorized primarily into cervical squamous cell carcinomas (which constitute approximately 75-80% of cervical cancer diagnoses), adenosquamous cell carcinomas, and adenocarcinomas (219-221). EGFR

mutations have been reported as more frequent in squamous cell carcinomas, while adenocarcinomas demonstrated a significantly higher rate of KRAS mutations; both cervical cancer types also demonstrated a notable rate of PIK3CA mutations that were a marker for poorer patient prognosis (222). Interestingly, it has also been reported that adenocarcinomas and adenosquamous cell carcinomas are infected by alphapapillomavirus 7 types, specifically HPV 18, more frequently than squamous cell carcinomas; it should be noted that HPV16 is still the most prevalent infectious HPV type in cervical cancer independent of tumor source site (223). Additionally, infection by alphapapillomavirus 7 types has been indicated as a greater risk factor for survival than alphapapillomavirus 9 types (224-227). This variance may account for some of the reports that patients with squamous cell carcinomas generally have better outcomes than patients with adenosquamous cell and adenocarcinoma, although some literature also propose that there is no difference in patient outcome based on tumor source site (228-233).

Nonetheless, research has consistently shown that HPV(-) tumors have worse outcome than tumors with any sort of HPV infection (227,234). This may be attributable to some of the genomic alterations that have been identified. This includes PIK3CA, KRAS, and EGFR, as previously described (222). Additional large-scale genomic studies have also been performed to confirm previously described genomic alterations, as well as identify novel features. One study published in 2014 analyzing samples from 115 cervical cancer patients confirmed literature reports of PTEN and STK11 in squamous cell carcinomas, as well as describing novel mutations in EP300, FBXW7, HLA-B, MAPK1, and NFE2L2 (235). Many of these mutations have been identified in other cancer types. Specifically, FBX27 and EP300 have been identified as mutated in endometrial and head and neck cancers, HLA-B mutants correspond with HLA-A and B2M mutants in lung squamous cell carcinomas. The specific MAPK1 mutations identified in this

study was also found in an oropharyngeal cancer cell line, and the NFE2L2 mutation was previously described in lung squamous cancers. Another notable large-scale genomic study was published recently by The Cancer Genome Atlas Research Network that included 178 patient samples (236). This study confirmed the aforementioned mutants as occurring with significant frequency, as well as identifying the novel mutants SHKBP1, ERBB3, CASP8, HLA-A and TGFRB2; notably, mutations in HLA-A, HLA-B, NFE2L2, MAPK1, CASP8, SHKB1 and TGFRB2 were exclusive to squamous cell carcinomas. This study also examined copy number alterations and transcriptome levels for both coding and noncoding RNAs to identify three major clusters of cervical cancers: high-keratin squamous cell tumors, low-keratin squamous cell tumors, and adenocarcinomas. Alphapapillomavirus 7 types were confirmed to be enriched in the adenocarcinoma and low-keratin clusters. Meanwhile, HPV(-) tumors demonstrated higher rates of KRAS, ARID1, and PTEN mutations, which may indicate possible rationales for higher treatment failure rates (236).

#### **1.4.2 HPV Distinguishes Oropharyngeal Cancers from Other Head and Neck Tumors**

This pattern of HPV status positively affecting cancer prognosis extends beyond anogenital tumors. Oropharyngeal tumors are cancers of the oropharynx, a region that includes the base of the tongue, the tonsils, soft palate, and walls of the pharynx. Traditional causes of OPSCC include tobacco and alcohol consumption, but there has been an increase of HPV(+) oropharyngeal cancer diagnoses despite the overall decrease in total new OPSCC cases (237). HPV(+) cases now approximate 75% of new oropharyngeal cancer diagnoses; this rise has been attributed to an increase in frequency of oral sexual behaviors (238,239). The mechanism for

HPV-induced oncogenesis is unchanged; however, due to differences in tissue type, the genomic characterization of additional oncogenic behavior differs somewhat from what has been observed in cervical cancers.

Similarly to cervical cancers, large-scale genomic studies have been performed to analyze the host genome alterations associated with oropharyngeal tumor formation and progression. Two of the larger studies in the literature analyzed head and neck squamous cell carcinomas as a whole. The first study by Parfenov and colleagues analyzed the immediate effects of HPV infection in head and neck tumors (240). Of the 279 tumor samples retrieved, 35 were identified as HPV(+); 29 were HPV16(+), while the remainder were positive for HPV33 or HPV35, and of the 35, 25 showed genomic integration. Genomic integration tended to be near coding regions and may be associated with somatic mutations of genes near the integration sites, including the silencing mutations of the DNA repair protein RAD51B, the tumor suppressor ETS2, and the apoptotic gene PDL1, as well as amplification of the oncogene NR4A2 (240).

Utilizing the same cohort, The Cancer Genome Atlas Research Network identified 36 HPV(+) tumors (241). Additionally, 33 samples were identified as oropharyngeal cancers, 21 of which were in the HPV(+) cohort, indicating an enrichment of HPV(+) tumors in the oropharynx as compared to other tumor source sites in the head and neck. Throughout the HPV(+) cohort, there was a significant number of deletions and truncations of TRAF3, as well as amplification of E2F1. Comparatively, HPV(-) HNSCCs were noted for having deletions in NSD1, as well as tumor suppressors such as NOTCH1 and CDKN2A. In addition to this, HPV(-) tumors in the head and neck showed amplification of receptor tyrosine kinases such as EGFR and ERBB2, which promote cell proliferation, as well as activating mutations of the oncogene HRAS and inactivating mutations of the proapoptotic factor CASP8. Independently of HPV

status, TCGA identified amplifications of a chromosomal region containing the transcription factors TP63 and SOX2, as well as the oncogene PIK3CA. HPV(-) tumors also contained inactivating mutations of CDKN2A, TP 53, and FAT1 at a higher rate than HPV(+) tumors; comparatively, HPV(+) tumors were contained activating mutations of PIK3CA in addition to the HPV-independent regional amplification (241).

### **1.4.3 Applications of Human Papillomavirus in the Diagnostic Setting**

Beyond the obvious implications of HPV presenting as a possible factor for oncogenesis, HPV status is also being used as a diagnostic criterion in the clinic. A pair of studies including 111 patients in Canada and 323 patients in the United States both confirmed the utility of HPV status as an independent prognostic factor in head and neck cancers (216,217). As a treatment target, HPV vaccines have been described as a preventative measure. Additionally, some treatments have been designed to target HPV(-) tumors more specifically, such as the use of the hybrid human/mouse mAb cetuximab (trade name Erbitux) or the pure human mAb panitumumab (Vectibix) to target and inhibit EGFR, in addition to the current standard of cisplatin and radiotherapy (242-244). On the opposite end of the treatment spectrum, HPV(+) oropharyngeal cancer patients may qualify for de-escalation protocols, some which have been reviewed by Masterson and colleagues (245). In their review and meta-study, it was concluded that reduction in radiation intensity for lower-risk, i.e. HPV(+), patients merited continued investigation. Regarding the change in treatment modalities to replace cisplatin with EGFR inhibitors, reports were mixed, with one study suggesting that the treatment was more effective in the HPV(-) cohort, thereby increasing the risk for HPV(+) patients unnecessarily. Despite this,

authors still recommended continued investigation, as metastasis-free survival after EGFR-mAb treatment was not yet clearly defined.

In the years since HPV was hypothesized to be an oncovirus, and later confirmed to be a the primary cause of cervical cancer, there has been a concentrated study of how an external factor such as a virus can alter the cellular environment so drastically as to induce tumor formation. Simultaneously, clinical studies have focused on prevention and treatment, leading to the creation of the HPV vaccine, with the long-term goal of eliminating virally-induced cervical cancer. Until the time that universal vaccination eradicates HPV as a sexually-transmitted infection, there remains a need for biological studies to elucidate mechanisms of HPV-based tumor formation and survival. Such research is translatable to the clinical setting, where treatments can be designed to target and interfere with HPV-controlled functions, as well as determine appropriate courses of treatment in accordance with modern paradigm of personalized medicine. As such, further diagnostic parameters that can stratify patients based on risk of treatment failure, both independently and in concordance with HPV status, will continue to be in demand and relevant until widespread vaccination and prevention is attained.

## **1.5 Project Aims**

The goals of this project were threefold:

1. To develop a bioinformatics pipeline to identify biomarkers in HPV-related cancers available in TCGA, and design prognostic signatures for HPV-related cancers based on RNA sequencing and miRNA sequencing data;
2. To apply this bioinformatics pipeline to the cancer data from TCGA as a whole and identify potential biomarkers across multiple cancer types;

3. To obtain insight into how these biomarkers function within HPV-related cancers and the mechanisms effected in increasing or decreasing patient risk.

In order to conduct this first aim, we developed a comprehensive approach to identify HPV status of tumor samples in oropharyngeal and cervical cancer patients using RNA-Seq data, as well as determine the expression level of coding transcripts. A parallel pipeline was also designed to identify dysregulated miRNA transcripts from miRNA-Seq data. By combining the results of these transcriptomic datasets with clinical data provided by TCGA, we were able to determine the contribution of miRNA and RNA expression levels to overall patient survival. A rigorous process was then used to select a subset of these biomarkers in the design of prognostic survival signatures. Within oropharyngeal cancer, we were able to not only design an HPV-independent prognostic signature based on the expression level of four microRNAs, but also experimentally validate the signature in an independent dataset using quantitative reverse transcription polymerase chain reaction, thereby demonstrating that the signature can potentially be applied and evaluated within a clinical setting in a cost-effective manner.

The purpose of the second aim is to show that these bioinformatics pipelines can be extended beyond the scope of HPV-dependent cancers. By developing statistical programs to perform automated analysis, we were able to identify miRNAs in the TCGA dataset that were related to cancer development, progression, and cancer survival in a type specific manner. Additionally, we extended the miRNA analysis to incorporate target prediction, so as to identify potential genes that may be controlled by dysregulated miRNAs both within and between cancer types. The results of this analysis have been made publicly available at [oncomir.org](http://oncomir.org), a combined database and web server for cancer-related miRNAs. The server is also capable of performing de novo



analysis for miRNA-based survival signatures and the identification of miRNA-based clusters of cancer types.

The third aim of this project is to expand on the known biology of HPV-induced carcinogenesis and tumor survival. The mechanisms by which high-risk HPV types lead to tumor formation through the E6 and E7 proteins are extremely well-documented, and more recent studies have begun to examine the rationale behind HPV being a positive biomarker for disease-free and overall patient survival. Intermediate regulatory networks, such as those mediated by miRNAs, are less well-studied, but may provide a greater insight into properties of HPV-related tumor survival. By conducting a pathway analysis based on miRNA response to HPV that focuses on the miRNA-target regulatory interactions, we demonstrate which pathways supplement the traditional E6/E7 mechanisms in tumor formation, as well as the mutation-driven pathways in HPV(-) tumors that portend less favorable patient outcomes. This also provides guidance for future research, as the pathways identified may be candidates for pharmaceutical therapies in HPV-defined patient populations.

## 1.6 References

1. Onitilo, A.A., Engel, J.M., Greenlee, R.T. and Mukesh, B.N. (2009) Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. *Clin Med Res*, **7**, 4-13.
2. Riggs, B.L. and Hartmann, L.C. (2003) Selective estrogen-receptor modulators -- mechanisms of action and application to clinical practice. *N Engl J Med*, **348**, 618-629.
3. Osborne, C.K., Wakeling, A. and Nicholson, R.I. (2004) Fulvestrant: an oestrogen receptor antagonist with a novel mechanism of action. *Br J Cancer*, **90 Suppl 1**, S2-6.
4. Miller, W.R. (2003) Aromatase inhibitors: mechanism of action and role in the treatment of breast cancer. *Seminars in Oncology*, **30, Supplement 14**, 3-11.

5. Romond, E.H., Perez, E.A., Bryant, J., Suman, V.J., Geyer, C.E., Davidson, N.E., Tan-Chiu, E., Martino, S., Paik, S., Kaufman, P.A. *et al.* (2005) Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med*, **353**, 1673-1684.
6. Blumenthal, G.M., Scher, N.S., Cortazar, P., Chattopadhyay, S., Tang, S., Song, P., Liu, Q., Ringgold, K., Pilaro, A.M., Tilley, A. *et al.* (2013) First FDA approval of dual anti-HER2 regimen: pertuzumab in combination with trastuzumab and docetaxel for HER2-positive metastatic breast cancer. *Clin Cancer Res*, **19**, 4911-4916.
7. Yarden, Y. (2001) Biology of HER2 and its importance in breast cancer. *Oncology*, **61 Suppl 2**, 1-13.
8. Hudis, C.A. (2007) Trastuzumab - Mechanism of action and use in clinical practice. *N Engl J Med*, **357**, 39-51.
9. Foulkes, W.D., Smith, I.E. and Reis-Filho, J.S. (2010) Triple-negative breast cancer. *N Engl J Med*, **363**, 1938-1948.
10. Stamey, T.A., Yang, N., Hay, A.R., McNeal, J.E., Freiha, F.S. and Redwine, E. (1987) Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. *N Engl J Med*, **317**, 909-916.
11. White, M.K., Pagano, J.S. and Khalili, K. (2014) Viruses and human cancers: a long road of discovery of molecular paradigms. *Clin Microbiol Rev*, **27**, 463-481.
12. Annas, G.J. and Elias, S. (2014) 23andMe and the FDA. *N Engl J Med*, **370**, 985-988.
13. Esteller, M. (2011) Non-coding RNAs in human disease. *Nat Rev Genet*, **12**, 861-874.
14. Hansen, T.B., Kjems, J. and Damgaard, C.K. (2013) Circular RNA and miR-7 in cancer. *Cancer Res*, **73**, 5609-5612.
15. Ambros, V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350-355.
16. The Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, **45**, 1113-1120.
17. The Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.
18. Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, **75**, 843-854.
19. Wightman, B., Ha, I. and Ruvkun, G. (1993) Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*, **75**, 855-862.

20. Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86-89.
21. Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R. and Ruvkun, G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901-906.
22. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, **42**, D68-D73.
23. Ha, M. and Kim, V.N. (2014) Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol*, **15**, 509-524.
24. Meister, G. (2013) Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet*, **14**, 447-459.
25. Djuranovic, S., Nahvi, A. and Green, R. (2012) miRNA-Mediated Gene Silencing by Translational Repression Followed by mRNA Deadenylation and Decay. *Science*, **336**, 237.
26. Lewis, B.P., Burge, C.B. and Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15-20.
27. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.
28. Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, **27**, 91-105.
29. Bartel, D.P. (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, **136**, 215-233.
30. Shin, C., Nam, J.W., Farh, K.K., Chiang, H.R., Shkumatava, A. and Bartel, D.P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell*, **38**, 789-802.
31. Agarwal, V., Bell, G.W., Nam, J.-W. and Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**, e05005.
32. Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T. and Hatzigeorgiou, A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research*, **41**, W169-W173.

33. Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D.S. (2003) MicroRNA targets in *Drosophila*. *Genome Biology*, **5**, R1.
34. Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A.M., Lim, B. and Rigoutsos, I. (2006) A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell*, **126**, 1203-1217.
35. Wang, X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, **32**, 1316-1322.
36. Peterson, S.M., Thompson, J.A., Ufkin, M.L., Sathyanarayana, P., Liaw, L. and Congdon, C.B. (2014) Common features of microRNA target prediction tools. *Front Genet*, **5**, 23.
37. Wong, N. and Wang, X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*, **43**, D146-D152.
38. Wang, X. and Wang, X. (2006) Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res*, **34**, 1646-1652.
39. Linsley, P.S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M.M., Bartz, S.R., Johnson, J.M., Cummins, J.M., Raymond, C.K., Dai, H. *et al.* (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol Cell Biol*, **27**, 2240-2252.
40. Ebert, M.S., Neilson, J.R. and Sharp, P.A. (2007) MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat Methods*, **4**, 721-726.
41. Liu, W., Chen, H., Wong, N., Haynes, W., Baker, C.M. and Wang, X. (2017) Pseudohypoxia induced by miR-126 deactivation promotes migration and therapeutic resistance in renal cell carcinoma. *Cancer Letters*, **394**, 65-75.
42. Grosswendt, S., Filipchuk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E. and Rajewsky, N. (2014) Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. *Molecular Cell*, **54**, 1042-1054.
43. Helwak, A., Kudla, G., Dudnakova, T. and Tollervey, D. (2013) Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell*, **153**, 654-665.
44. Sempere, L.F., Sokol, N.S., Dubrovsky, E.B., Berger, E.M. and Ambros, V. (2003) Temporal regulation of microRNA expression in *Drosophila melanogaster* mediated by hormonal signals and Broad-Complex gene activity. *Developmental Biology*, **259**, 9-18.

45. Suh, M.-R., Lee, Y., Kim, J.Y., Kim, S.-K., Moon, S.-H., Lee, J.Y., Cha, K.-Y., Chung, H.M., Yoon, H.S., Moon, S.Y. *et al.* (2004) Human embryonic stem cells express a unique set of microRNAs. *Developmental Biology*, **270**, 488-498.
46. Bazzini, A.A., Lee, M.T. and Giraldez, A.J. (2012) Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science*, **336**, 233.
47. Ivey, K.N. and Srivastava, D. (2010) MicroRNAs as Regulators of Differentiation and Cell Fate Decisions. *Cell Stem Cell*, **7**, 36-41.
48. Hébert, S.S., Horré, K., Nicolai, L., Papadopoulou, A.S., Mandemakers, W., Silahdaroglu, A.N., Kauppinen, S., Delacourte, A. and De Strooper, B. (2008) Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/ $\beta$ -secretase expression. *Proceedings of the National Academy of Sciences*, **105**, 6415-6420.
49. Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M. *et al.* (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 2999-3004.
50. Croce, C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet*, **10**, 704-714.
51. Garzon, R., Calin, G.A. and Croce, C.M. (2009) MicroRNAs in Cancer. *Annu Rev Med*, **60**, 167-179.
52. Di Leva, G., Garofalo, M. and Croce, C.M. (2014) MicroRNAs in cancer. *Annu Rev Pathol*, **9**, 287-314.
53. Jacobsen, A., Silber, J., Harinath, G., Huse, J.T., Schultz, N. and Sander, C. (2013) Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol*, **20**, 1325-1332.
54. Cho, W.C. (2010) MicroRNAs: Potential biomarkers for cancer diagnosis, prognosis and targets for therapy. *Int J of Biochem Cell Biol*, **42**, 1273-1281.
55. Shea, A., Harish, V., Afzal, Z., Chijjoke, J., Kedir, H., Dusmatova, S., Roy, A., Ramalinga, M., Harris, B., Blancato, J. *et al.* (2016) MicroRNAs in glioblastoma multiforme pathogenesis and therapeutics. *Cancer Med*, **5**, 1917-1946.
56. Schoof, C.R.G., Botelho, E.L.d.S., Izzotti, A. and Vasques, L.d.R. (2012) MicroRNAs in cancer treatment and prognosis. *Am J Cancer Res*, **2**, 414-433.
57. Hui, A., How, C., Ito, E. and Liu, F.F. (2011) Micro-RNAs as diagnostic or prognostic markers in human epithelial malignancies. *BMC Cancer*, **11**, 500.
58. Iorio, M.V. and Croce, C.M. (2012) MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol Med*, **4**, 143-159.

59. Melo, S.A. and Kalluri, R. (2012) Molecular pathways: microRNAs as cancer therapeutics. *Clin Cancer Res*, **18**, 4234-4239.
60. Lindenbergh-van der Plas, M., Martens-de Kemp, S.R., de Maaker, M., van Wieringen, W.N., Ylstra, B., Agami, R., Cerisoli, F., Leemans, C.R., Braakhuis, B.J. and Brakenhoff, R.H. (2013) Identification of lethal microRNAs specific for head and neck cancer. *Clin Cancer Res*, **19**, 5647-5657.
61. Wang, N., Zhou, Y., Zheng, L. and Li, H. (2014) MiR-31 is an independent prognostic factor and functions as an oncomir in cervical cancer via targeting ARID1A. *Gynecol Oncol*, **134**, 129-137.
62. Hu, X., Schwarz, J.K., Lewis, J., James S, Huettner, P.C., Rader, J.S., Deasy, J.O., Grigsby, P.W. and Wang, X. (2010) A microRNA expression signature for cervical cancer prognosis. *Cancer Res*, **70**, 1441-1448.
63. How, C., Pintilie, M., Bruce, J.P., Hui, A.B., Clarke, B.A., Wong, P., Yin, S., Yan, R., Waggott, D., Boutros, P.C. *et al.* (2015) Developing a prognostic micro-RNA signature for human cervical carcinoma. *PLoS One*, **10**, e0123946.
64. Gao, G., Gay, H.A., Chernock, R.D., Zhang, T.R., Luo, J., Thorstad, W.L., Lewis, J., James S and Wang, X. (2013) A microRNA expression signature for the prognosis of oropharyngeal squamous cell carcinoma. *Cancer*, **119**, 72-80.
65. Hui, A.B., Lin, A., Xu, W., Waldron, L., Perez-Ordenez, B., Weinreb, I., Shi, W., Bruce, J., Huang, S.H., O'Sullivan, B. *et al.* (2013) Potentially prognostic miRNAs in HPV-associated oropharyngeal carcinoma. *Clin Cancer Res*, **19**, 2154-2162.
66. Miller, D.L., Davis, J.W., Taylor, K.H., Johnson, J., Shi, Z., Williams, R., Atasoy, U., Lewis Jr, J.S. and Stack, M.S. (2015) Identification of a Human Papillomavirus-Associated Oncogenic miRNA Panel in Human Oropharyngeal Squamous Cell Carcinoma Validated by Bioinformatics Analysis of The Cancer Genome Atlas. *The American Journal of Pathology*, **185**, 679-692.
67. Sapre, N., Macintyre, G., Clarkson, M., Naeem, H., Cmero, M., Kowalczyk, A., Anderson, P.D., Costello, A.J., Corcoran, N.M. and Hovens, C.M. (2016) A urinary microRNA signature can predict the presence of bladder urothelial carcinoma in patients undergoing surveillance. *Br J Cancer*, **114**, 454-462.
68. Niidome, T. and Huang, L. (2002) Gene therapy progress and prospects: nonviral vectors. *Gene Ther*, **9**, 1647-1652.
69. Garzon, R., Marcucci, G. and Croce, C.M. (2010) Targeting microRNAs in cancer: rationale, strategies and challenges. *Nat Rev Drug Discov*, **9**, 775-789.
70. Ma, L., Reinhardt, F., Pan, E., Soutschek, J., Bhat, B., Marcusson, E.G., Teruya-Feldstein, J., Bell, G.W. and Weinberg, R.A. (2010) Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model. *Nat Biotechnol*, **28**, 341-347.

71. Elmén, J., Lindow, M., Schütz, S., Lawrence, M., Petri, A., Obad, S., Lindholm, M., Hedtjärn, M., Hansen, H.F., Berger, U. *et al.* (2008) LNA-mediated microRNA silencing in non-human primates. *Nature*, **452**, 896-899.
72. Ebert, M.S. and Sharp, P.A. (2010) MicroRNA sponges: progress and possibilities. *RNA*, **16**, 2043-2050.
73. Trang, P., Medina, P.P., Wiggins, J.F., Ruffino, L., Kelnar, K., Omotola, M., Homer, R., Brown, D., Bader, A.G., Weidhaas, J.B. *et al.* (2010) Regression of murine lung tumors by the let-7 microRNA. *Oncogene*, **29**, 1580-1587.
74. Trang, P., Wiggins, J.F., Daige, C.L., Cho, C., Omotola, M., Brown, D., Weidhaas, J.B., Bader, A.G. and Slack, F.J. (2011) Systemic delivery of tumor suppressor microRNA mimics using a neutral lipid emulsion inhibits lung tumors in mice. *Mol Ther*, **19**, 1116-1122.
75. Pramanik, D., Campbell, N.R., Karikari, C., Chivukula, R., Kent, O.A., Mendell, J.T. and Maitra, A. (2011) Restitution of tumor suppressor microRNAs using a systemic nanovector inhibits pancreatic cancer growth in mice. *Mol Cancer Ther*, **10**, 1470-1480.
76. Ma, H., Marti-Gutierrez, N., Park, S.W., Wu, J., Lee, Y., Suzuki, K., Koski, A., Ji, D., Hayama, T., Ahmed, R. *et al.* (2017) Correction of a pathogenic gene mutation in human embryos. *Nature*, **548**, 413-419.
77. Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B. *et al.* (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol*, **15**, 161-168.
78. Adams, M.J., Lefkowitz, E.J., King, A.M.Q., Harrach, B., Harrison, R.L., Knowles, N.J., Kropinski, A.M., Krupovic, M., Kuhn, J.H., Mushegian, A.R. *et al.* (2017) Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). *Archives of Virology*, **162**, 2505-2538.
79. Baltimore, D. (1971) Expression of animal virus genomes. *Bacteriol Rev*, **35**, 235-241.
80. Temin, H.M. (1985) Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol Biol Evol*, **2**, 455-468.
81. Lodish, H., Berk, A., Kaiser, C.A., Krieger, M., Bretscher, A., Ploegh, H. and Amon, A. (2013), *Molecular Cell Biology*. 7th ed. W.H. Freeman and Co., New York, pp. 160-166.
82. Duensing, S. and Münger, K. (2002) Human papillomaviruses and centrosome duplication errors: modeling the origins of genomic instability. *Oncogene*, **21**, 6241-6248.
83. Smiley, J.R. (2004) Herpes simplex virus virion host shutoff protein: immune evasion mediated by a viral RNase? *J Virol*, **78**, 1063-1068.

84. Khalili, K., Sariyer, I.K. and Safak, M. (2008) Small tumor antigen of polyomaviruses: role in viral life cycle and cell transformation. *J Cell Physiol*, **215**, 309-319.
85. Blackford, A.N. and Grand, R.J. (2009) Adenovirus E1B 55-kilodalton protein: multiple roles in viral infection and cell transformation. *J Virol*, **83**, 4000-4012.
86. Balvay, L., Soto Rifo, R., Ricci, E.P., Decimo, D. and Ohlmann, T. (2009) Structural and functional diversity of viral IRESes. *Biochim Biophys Acta*, **1789**, 542-557.
87. Hansen, T.H. and Bouvier, M. (2009) MHC class I antigen presentation: learning from viral evasion strategies. *Nat Rev Immunol*, **9**, 503-513.
88. Galluzzi, L., Brenner, C., Morselli, E., Touat, Z. and Kroemer, G. (2008) Viral control of mitochondrial apoptosis. *PLoS Pathog*, **4**, e1000018.
89. Rous, P. (1911) A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J Exp Med*, **13**, 397-411.
90. Epstein, M.A., Achong, B.G. and Barr, Y.M. (1964) Virus particles in cultured lymphoblasts from Burkitt's lymphoma. *Lancet*, **1**, 702-703.
91. Henle, W., Henle, G., Zajac, B.A., Pearson, G., Waubke, R. and Scriba, M. (1970) Differential reactivity of human serums with early antigens induced by Epstein-Barr virus. *Science*, **169**, 188-190.
92. zur Hausen, H., Meinhof, W., Scheiber, W. and Bornkamm, G.W. (1974) Attempts to detect virus-specific DNA in human tumors. I. Nucleic acid hybridizations with complementary RNA of human wart virus. *Int J Cancer*, **13**, 650-656.
93. zur Hausen, H., Schulte-Holthausen, H., Wolf, H., Dörries, K. and Egger, H. (1974) Attempts to detect virus-specific DNA in human tumors. II. Nucleic acid hybridizations with complementary RNA of human herpes group viruses. *Int J Cancer*, **13**, 657-664.
94. Beasley, R.P., Hwang, L.Y., Lin, C.C. and Chien, C.S. (1981) Hepatocellular carcinoma and hepatitis B virus. A prospective study of 22 707 men in Taiwan. *Lancet*, **2**, 1129-1133.
95. Yoshida, M., Miyoshi, I. and Hinuma, Y. (1982) Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. *Proc Natl Acad Sci U S A*, **79**, 2031-2035.
96. Seiki, M., Hattori, S., Hirayama, Y. and Yoshida, M. (1983) Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proc Natl Acad Sci U S A*, **80**, 3618-3622.
97. Feinstone, S.M., Kapikian, A.Z., Purcell, R.H., Alter, H.J. and Holland, P.V. (1975) Transfusion-associated hepatitis not due to viral hepatitis type A or B. *N Engl J Med*, **292**, 767-770.



98. Choo, Q.L., Kuo, G., Weiner, A.J., Overby, L.R., Bradley, D.W. and Houghton, M. (1989) Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*, **244**, 359-362.
99. Saito, I., Miyamura, T., Ohbayashi, A., Harada, H., Katayama, T., Kikuchi, S., Watanabe, Y., Koi, S., Onji, M. and Ohta, Y. (1990) Hepatitis C virus infection is associated with the development of hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, **87**, 6547-6549.
100. Chang, Y., Cesarman, E., Pessin, M.S., Lee, F., Culpepper, J., Knowles, D.M. and Moore, P.S. (1994) Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, **266**, 1865-1869.
101. Feng, H., Taylor, J.L., Benos, P.V., Newton, R., Waddell, K., Lucas, S.B., Chang, Y. and Moore, P.S. (2007) Human Transcriptome Subtraction by Using Short Sequence Tags To Search for Tumor Viruses in Conjunctival Carcinoma. *Journal of Virology*, **81**, 11332-11340.
102. Biggs, P.M. (2001) The history and biology of Marek's disease virus. *Curr Top Microbiol Immunol*, **255**, 1-24.
103. Cicala, C., Pompetti, F. and Carbone, M. (1993) SV40 induces mesotheliomas in hamsters. *Am J Pathol*, **142**, 1524-1533.
104. Hardy, W.D., Hess, P.W., MacEwen, E.G., McClelland, A.J., Zuckerman, E.E., Essex, M., Cotter, S.M. and Jarrett, O. (1976) Biology of feline leukemia virus in the natural environment. *Cancer Res*, **36**, 582-588.
105. Mesri, E.A., Feitelson, M.A. and Munger, K. (2014) Human viral oncogenesis: a cancer hallmarks analysis. *Cell Host Microbe*, **15**, 266-282.
106. Papesch, M. and Watkins, R. (2001) Epstein-Barr virus infectious mononucleosis. *Clinical Otolaryngology & Allied Sciences*, **26**, 3-8.
107. Trépo, C., Chan, H.L.Y. and Lok, A. Hepatitis B virus infection. *The Lancet*, **384**, 2053-2063.
108. Webster, D.P., Klenerman, P. and Dusheiko, G.M. Hepatitis C. *The Lancet*, **385**, 1124-1135.
109. Crosbie, E.J., Einstein, M.H., Franceschi, S. and Kitchener, H.C. Human papillomavirus and cervical cancer. *The Lancet*, **382**, 889-899.
110. Ishitsuka, K. and Tamura, K. (2014) Human T-cell leukaemia virus type I and adult T-cell leukaemia-lymphoma. *The Lancet Oncology*, **15**, e517-e526.
111. Schulz, T.F. (2000) Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8): epidemiology and pathogenesis. *J Antimicrob Chemother*, **45 Suppl T3**, 15-27.

112. Liu, W., MacDonald, M. and You, J. (2016) Merkel cell polyomavirus infection and Merkel cell carcinoma. *Curr Opin Virol*, **20**, 20-27.
113. Kirchmaier, A.L. and Sugden, B. (1997) Dominant-negative inhibitors of EBNA-1 of Epstein-Barr virus. *J Virol*, **71**, 1766-1775.
114. Mosialos, G., Birkenbach, M., Yalamanchili, R., VanArsdale, T., Ware, C. and Kieff, E. (1995) The Epstein-Barr virus transforming protein LMP1 engages signaling proteins for the tumor necrosis factor receptor family. *Cell*, **80**, 389-399.
115. Merchant, M., Caldwell, R.G. and Longnecker, R. (2000) The LMP2A ITAM is essential for providing B cells with development and survival signals in vivo. *J Virol*, **74**, 9115-9124.
116. Portis, T. and Longnecker, R. (2004) Epstein-Barr virus (EBV) LMP2A mediates B-lymphocyte survival through constitutive activation of the Ras/PI3K/Akt pathway. *Oncogene*, **23**, 8619-8628.
117. Ganem, D. (2010) KSHV and the pathogenesis of Kaposi sarcoma: listening to human biology and medicine. *J Clin Invest*, **120**, 939-949.
118. Verma, S.C., Lan, K. and Robertson, E. (2007) Structure and function of latency-associated nuclear antigen. *Curr Top Microbiol Immunol*, **312**, 101-136.
119. Mesri, E.A., Cesarman, E. and Boshoff, C. (2010) Kaposi's sarcoma and its associated herpesvirus. *Nat Rev Cancer*, **10**, 707-719.
120. Bouchard, M.J. and Schneider, R.J. (2004) The Enigmatic X Gene of Hepatitis B Virus. *Journal of Virology*, **78**, 12725-12734.
121. Jung, J.K., Arora, P., Pagano, J.S. and Jang, K.L. (2007) Expression of DNA methyltransferase 1 is activated by hepatitis B virus X protein via a regulatory circuit involving the p16INK4a-cyclin D1-CDK 4/6-pRb-E2F1 pathway. *Cancer Res*, **67**, 5771-5778.
122. Gale, M., Kwieciszewski, B., Dossett, M., Nakao, H. and Katze, M.G. (1999) Antiapoptotic and oncogenic potentials of hepatitis C virus are linked to interferon resistance by viral repression of the PKR protein kinase. *J Virol*, **73**, 6506-6516.
123. Deng, L., Nagano-Fujii, M., Tanaka, M., Nomura-Takigawa, Y., Ikeda, M., Kato, N., Sada, K. and Hotta, H. (2006) NS3 protein of Hepatitis C virus associates with the tumour suppressor p53 and inhibits its function in an NS3 sequence-dependent manner. *J Gen Virol*, **87**, 1703-1713.
124. Matsuoka, M. and Jeang, K.-T. (2007) Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. *Nat Rev Cancer*, **7**, 270-280.

125. Satou, Y., Yasunaga, J., Yoshida, M. and Matsuoka, M. (2006) HTLV-I basic leucine zipper factor gene mRNA supports proliferation of adult T cell leukemia cells. *Proc Natl Acad Sci U S A*, **103**, 720-725.
126. Shuda, M., Feng, H., Kwun, H.J., Rosen, S.T., Gjoerup, O., Moore, P.S. and Chang, Y. (2008) T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. *Proc Natl Acad Sci U S A*, **105**, 16272-16277.
127. Liu, X., Hein, J., Richardson, S.C., Basse, P.H., Toptan, T., Moore, P.S., Gjoerup, O.V. and Chang, Y. (2011) Merkel cell polyomavirus large T antigen disrupts lysosome clustering by translocating human Vam6p from the cytoplasm to the nucleus. *J Biol Chem*, **286**, 17079-17090.
128. Shuda, M., Kwun, H.J., Feng, H., Chang, Y. and Moore, P.S. (2011) Human Merkel cell polyomavirus small T antigen is an oncoprotein targeting the 4E-BP1 translation regulator. *J Clin Invest*, **121**, 3623-3634.
129. Kwun, H.J., Shuda, M., Feng, H., Camacho, C.J., Moore, P.S. and Chang, Y. (2013) Merkel cell polyomavirus small T antigen controls viral replication and oncoprotein expression by targeting the cellular ubiquitin ligase SCFFbw7. *Cell Host Microbe*, **14**, 125-135.
130. de Villiers, E.M., Fauquet, C., Broker, T.R., Bernard, H.U. and zur Hausen, H. (2004) Classification of papillomaviruses. *Virology*, **324**, 17-27.
131. Buck, C.B., Day, P.M. and Trus, B.L. (2013) The papillomavirus major capsid protein L1. *Virology*, **445**, 169-174.
132. Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., Sun, Q., Kaur, R., Huyen, Y. and McBride, A.A. (2017) The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res*, **45**, D499-D506.
133. Cubie, H.A. (2013) Diseases associated with human papillomavirus infection. *Virology*, **445**, 21-34.
134. Graham, S.V. (2017) The human papillomavirus replication cycle, and its links to cancer progression: a comprehensive review. *Clin Sci (Lond)*, **131**, 2201-2221.
135. Bergvall, M., Melendy, T. and Archambault, J. (2013) The E1 proteins. *Virology*, **445**, 35-56.
136. McBride, A.A. (2013) The papillomavirus E2 proteins. *Virology*, **445**, 57-79.
137. Vande Pol, S.B. and Klingelutz, A.J. (2013) Papillomavirus E6 oncoproteins. *Virology*, **445**, 115-137.
138. Roman, A. and Munger, K. (2013) The papillomavirus E7 proteins. *Virology*, **445**, 138-168.

139. Doorbar, J. (2013) The E4 protein; structure, function and patterns of expression. *Virology*, **445**, 80-98.
140. DiMaio, D. and Petti, L.M. (2013) The E5 proteins. *Virology*, **445**, 99-114.
141. Wang, J.W. and Roden, R.B. (2013) L2, the minor capsid protein of papillomavirus. *Virology*, **445**, 175-186.
142. zur Hausen, H. (2002) Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer*, **2**, 342-350.
143. Munger, K., Baldwin, A., Edwards, K.M., Hayakawa, H., Nguyen, C.L., Owens, M., Grace, M. and Huh, K. (2004) Mechanisms of human papillomavirus-induced oncogenesis. *J Virol*, **78**, 11451-11460.
144. Yim, E.K. and Park, J.S. (2005) The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. *Cancer Res Treat*, **37**, 319-324.
145. Hoppe-Seyler, K., Bossler, F., Braun, J.A., Herrmann, A.L. and Hoppe-Seyler, F. (2017) The HPV E6/E7 Oncogenes: Key Factors for Viral Carcinogenesis and Therapeutic Targets. *Trends Microbiol.*
146. Scheffner, M., Huibregtse, J.M., Vierstra, R.D. and Howley, P.M. (1993) The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell*, **75**, 495-505.
147. Polager, S. and Ginsberg, D. (2009) p53 and E2f: partners in life and death. *Nat Rev Cancer*, **9**, 738-748.
148. White, E.A., Kramer, R.E., Tan, M.J., Hayes, S.D., Harper, J.W. and Howley, P.M. (2012) Comprehensive analysis of host cellular interactions with human papillomavirus E6 proteins identifies new E6 binding partners and reflects viral diversity. *J Virol*, **86**, 13174-13186.
149. Brimer, N., Lyons, C., Wallberg, A.E. and Vande Pol, S.B. (2012) Cutaneous papillomavirus E6 oncoproteins associate with MAML1 to repress transactivation and NOTCH signaling. *Oncogene*, **31**, 4639-4646.
150. Van Doorslaer, K. and Burk, R.D. (2012) Association between hTERT activation by HPV E6 proteins and oncogenic risk. *Virology*, **433**, 216-219.
151. Kiyono, T., Hiraiwa, A., Fujita, M., Hayashi, Y., Akiyama, T. and Ishibashi, M. (1997) Binding of high-risk human papillomavirus E6 oncoproteins to the human homologue of the Drosophila discs large tumor suppressor protein. *Proc Natl Acad Sci U S A*, **94**, 11612-11616.
152. Klingelhutz, A.J., Foster, S.A. and McDougall, J.K. (1996) Telomerase activation by the E6 gene product of human papillomavirus type 16. *Nature*, **380**, 79-82.

153. Spangle, J.M. and Münger, K. (2010) The human papillomavirus type 16 E6 oncoprotein activates mTORC1 signaling and increases protein synthesis. *J Virol*, **84**, 9398-9407.
154. Spangle, J.M., Ghosh-Choudhury, N. and Munger, K. (2012) Activation of cap-dependent translation by mucosal human papillomavirus E6 proteins is dependent on the integrity of the LXXLL binding motif. *J Virol*, **86**, 7466-7472.
155. Havard, L., Rahmouni, S., Boniver, J. and Delvenne, P. (2005) High levels of p105 (NFKB1) and p100 (NFKB2) proteins in HPV16-transformed keratinocytes: role of E6 and E7 oncoproteins. *Virology*, **331**, 357-366.
156. Thomas, M. and Banks, L. (1999) Human papillomavirus (HPV) E6 interactions with Bak are conserved amongst E6 proteins from high and low risk HPV types. *J Gen Virol*, **80** ( Pt 6), 1513-1517.
157. Tungteakkhun, S.S. and Duerksen-Hughes, P.J. (2008) Cellular binding partners of the human papillomavirus E6 protein. *Arch Virol*, **153**, 397-408.
158. Li, S., Labrecque, S., Gauzzi, M.C., Cuddihy, A.R., Wong, A.H., Pellegrini, S., Matlashewski, G.J. and Koromilas, A.E. (1999) The human papilloma virus (HPV)-18 E6 oncoprotein physically associates with Tyk2 and impairs Jak-STAT activation by interferon-alpha. *Oncogene*, **18**, 5727-5737.
159. Ronco, L.V., Karpova, A.Y., Vidal, M. and Howley, P.M. (1998) Human papillomavirus 16 E6 oncoprotein binds to interferon regulatory factor-3 and inhibits its transcriptional activity. *Genes Dev*, **12**, 2061-2072.
160. Dyson, N., Howley, P.M., Münger, K. and Harlow, E. (1989) The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science*, **243**, 934-937.
161. Dyson, N., Guida, P., Münger, K. and Harlow, E. (1992) Homologous sequences in adenovirus E1A and human papillomavirus E7 proteins mediate interaction with the same set of cellular proteins. *J Virol*, **66**, 6893-6902.
162. Song, S., Liem, A., Miller, J.A. and Lambert, P.F. (2000) Human papillomavirus types 16 E6 and E7 contribute differently to carcinogenesis. *Virology*, **267**, 141-150.
163. Mirabello, L., Yeager, M., Yu, K., Clifford, G.M., Xiao, Y., Zhu, B., Cullen, M., Boland, J.F., Wentzensen, N., Nelson, C.W. *et al.* (2017) HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell*, **170**, 1164-1174.e1166.
164. Shin, M.K., Balsitis, S., Brake, T. and Lambert, P.F. (2009) Human papillomavirus E7 oncoprotein overrides the tumor suppressor activity of p21Cip1 in cervical carcinogenesis. *Cancer Res*, **69**, 5656-5663.

165. Zerfass-Thome, K., Zwerschke, W., Mannhardt, B., Tindle, R., Botz, J.W. and Jansen-Dürr, P. (1996) Inactivation of the cdk inhibitor p27KIP1 by the human papillomavirus type 16 E7 oncoprotein. *Oncogene*, **13**, 2323-2330.
166. Pfeffer, S., Zavolan, M., Grässer, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C. *et al.* (2004) Identification of virus-encoded microRNAs. *Science*, **304**, 734-736.
167. Roberts, A.P., Lewis, A.P. and Jopling, C.L. (2011) The role of microRNAs in viral infection. *Prog Mol Biol Transl Sci*, **102**, 101-139.
168. Skalsky, R.L. and Cullen, B.R. (2010) Viruses, microRNAs, and host interactions. *Annu Rev Microbiol*, **64**, 123-141.
169. Cai, X., Schäfer, A., Lu, S., Bilello, J.P., Desrosiers, R.C., Edwards, R., Raab-Traub, N. and Cullen, B.R. (2006) Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed. *PLoS Pathog*, **2**, e23.
170. Grundhoff, A., Sullivan, C.S. and Ganem, D. (2006) A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA*, **12**, 733-750.
171. Choy, E.Y., Siu, K.L., Kok, K.H., Lung, R.W., Tsang, C.M., To, K.F., Kwong, D.L., Tsao, S.W. and Jin, D.Y. (2008) An Epstein-Barr virus-encoded microRNA targets PUMA to promote host cell survival. *J Exp Med*, **205**, 2551-2560.
172. Xia, T., O'Hara, A., Araujo, I., Barreto, J., Carvalho, E., Sapucaia, J.B., Ramos, J.C., Luz, E., Pedrosa, C., Manrique, M. *et al.* (2008) EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-mir-BHRF1-3. *Cancer Res*, **68**, 1436-1442.
173. Nachmani, D., Stern-Ginossar, N., Sarid, R. and Mandelboim, O. (2009) Diverse Herpesvirus MicroRNAs Target the Stress-Induced Immune Ligand MICB to Escape Recognition by Natural Killer Cells. *Cell Host & Microbe*, **5**, 376-385.
174. Barth, S., Pfuhl, T., Mamiani, A., Ehses, C., Roemer, K., Kremmer, E., Jäker, C., Höck, J., Meister, G. and Grässer, F.A. (2008) Epstein-Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5. *Nucleic Acids Res*, **36**, 666-675.
175. Lo, A.K., To, K.F., Lo, K.W., Lung, R.W., Hui, J.W., Liao, G. and Hayward, S.D. (2007) Modulation of LMP1 protein expression by EBV-encoded microRNAs. *Proc Natl Acad Sci U S A*, **104**, 16164-16169.
176. Bellare, P. and Ganem, D. (2009) Regulation of KSHV lytic switch protein expression by a virus-encoded microRNA: an evolutionary adaptation that fine-tunes lytic reactivation. *Cell Host Microbe*, **6**, 570-575.

177. Samols, M.A., Skalsky, R.L., Maldonado, A.M., Riva, A., Lopez, M.C., Baker, H.V. and Renne, R. (2007) Identification of cellular genes targeted by KSHV-encoded microRNAs. *PLoS Pathog*, **3**, e65.
178. Ziegelbauer, J.M., Sullivan, C.S. and Ganem, D. (2009) Tandem array-based expression screens identify host mRNA targets of virus-encoded microRNAs. *Nat Genet*, **41**, 130-134.
179. Hansen, A., Henderson, S., Lagos, D., Nikitenko, L., Coulter, E., Roberts, S., Gratrix, F., Plaisance, K., Renne, R., Bower, M. *et al.* (2010) KSHV-encoded miRNAs target MAF to induce endothelial cell reprogramming. *Genes Dev*, **24**, 195-205.
180. Gottwein, E., Mukherjee, N., Sachse, C., Frenzel, C., Majoros, W.H., Chi, J.T., Braich, R., Manoharan, M., Soutschek, J., Ohler, U. *et al.* (2007) A viral microRNA functions as an orthologue of cellular miR-155. *Nature*, **450**, 1096-1099.
181. Skalsky, R.L., Samols, M.A., Plaisance, K.B., Boss, I.W., Riva, A., Lopez, M.C., Baker, H.V. and Renne, R. (2007) Kaposi's sarcoma-associated herpesvirus encodes an ortholog of miR-155. *J Virol*, **81**, 12836-12845.
182. Sullivan, C.S., Grundhoff, A.T., Tevethia, S., Pipas, J.M. and Ganem, D. (2005) SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature*, **435**, 682-686.
183. Seo, G.J., Fink, L.H., O'Hara, B., Atwood, W.J. and Sullivan, C.S. (2008) Evolutionarily conserved function of a viral microRNA. *J Virol*, **82**, 9823-9828.
184. Seo, G.J., Chen, C.J. and Sullivan, C.S. (2009) Merkel cell polyomavirus encodes a microRNA with the ability to autoregulate viral gene expression. *Virology*, **383**, 183-187.
185. Aparicio, O., Razquin, N., Zaratiegui, M., Narvaiza, I. and Fortes, P. (2006) Adenovirus virus-associated RNA is processed to functional interfering RNAs involved in virus production. *J Virol*, **80**, 1376-1384.
186. Xu, N., Segerman, B., Zhou, X. and Akusjärvi, G. (2007) Adenovirus virus-associated RNAII-derived small RNAs are efficiently incorporated into the rna-induced silencing complex and associate with polyribosomes. *J Virol*, **81**, 10540-10549.
187. Aparicio, O., Carnero, E., Abad, X., Razquin, N., Guruceaga, E., Segura, V. and Fortes, P. (2010) Adenovirus VA RNA-derived miRNAs target cellular genes involved in cell growth, gene expression and DNA repair. *Nucleic Acids Res*, **38**, 750-763.
188. Yin, Q., McBride, J., Fewell, C., Lacey, M., Wang, X., Lin, Z., Cameron, J. and Flemington, E.K. (2008) MicroRNA-155 is an Epstein-Barr virus-induced gene that modulates Epstein-Barr virus-regulated gene expression pathways. *J Virol*, **82**, 5295-5306.

189. Cameron, J.E., Yin, Q., Fewell, C., Lacey, M., McBride, J., Wang, X., Lin, Z., Schaefer, B.C. and Flemington, E.K. (2008) Epstein-Barr virus latent membrane protein 1 induces cellular MicroRNA miR-146a, a modulator of lymphocyte signaling pathways. *J Virol*, **82**, 1946-1958.
190. Xiao, C. and Rajewsky, K. (2009) MicroRNA control in the immune system: basic principles. *Cell*, **136**, 26-36.
191. Mrázek, J., Kreutmayer, S.B., Grässer, F.A., Polacek, N. and Hüttenhofer, A. (2007) Subtractive hybridization identifies novel differentially expressed ncRNA species in EBV-infected human B cells. *Nucleic Acids Res*, **35**, e73.
192. Zhao, Y., Yao, Y., Xu, H., Lambeth, L., Smith, L.P., Kgosana, L., Wang, X. and Nair, V. (2009) A functional MicroRNA-155 ortholog encoded by the oncogenic Marek's disease virus. *J Virol*, **83**, 489-492.
193. Wang, F.Z., Weber, F., Croce, C., Liu, C.G., Liao, X. and Pellett, P.E. (2008) Human cytomegalovirus infection alters the expression of cellular microRNA species that affect its replication. *J Virol*, **82**, 9065-9074.
194. Triboulet, R., Mari, B., Lin, Y.L., Chable-Bessia, C., Bennasser, Y., Lebrigand, K., Cardinaud, B., Maurin, T., Barbry, P., Baillat, V. *et al.* (2007) Suppression of microRNA-silencing pathway by HIV-1 during virus replication. *Science*, **315**, 1579-1582.
195. O'Connor, C.M., Vanicek, J. and Murphy, E.A. (2014) Host microRNA regulation of human cytomegalovirus immediate early protein translation promotes viral latency. *J Virol*, **88**, 5524-5532.
196. Pan, D., Flores, O., Umbach, J.L., Pesola, J.M., Bentley, P., Rosato, P.C., Leib, D.A., Cullen, B.R. and Coen, D.M. (2014) A neuron-specific host microRNA targets herpes simplex virus-1 ICP0 expression and promotes latency. *Cell Host Microbe*, **15**, 446-456.
197. Nathans, R., Chu, C.Y., Serquina, A.K., Lu, C.C., Cao, H. and Rana, T.M. (2009) Cellular microRNA and P bodies modulate host-HIV-1 interactions. *Mol Cell*, **34**, 696-709.
198. Pedersen, I.M., Cheng, G., Wieland, S., Volinia, S., Croce, C.M., Chisari, F.V. and David, M. (2007) Interferon modulation of cellular microRNAs as an antiviral mechanism. *Nature*, **449**, 919-922.
199. Wang, X., Tang, S., Le, S.Y., Lu, R., Rader, J.S., Meyers, C. and Zheng, Z.M. (2008) Aberrant expression of oncogenic and tumor-suppressive microRNAs in cervical cancer is required for cancer cell growth. *PLoS One*, **3**, e2557.
200. Wang, X., Wang, H.K., McCoy, J.P., Banerjee, N.S., Rader, J.S., Broker, T.R., Meyers, C., Chow, L.T. and Zheng, Z.M. (2009) Oncogenic HPV infection interrupts the



- expression of tumor-suppressive miR-34a through viral oncoprotein E6. *RNA*, **15**, 637-647.
201. Melar-New, M. and Laimins, L.A. (2010) Human papillomaviruses modulate expression of microRNA 203 upon epithelial differentiation to control levels of p63 proteins. *J Virol*, **84**, 5212-5221.
  202. McKenna, D.J., McDade, S.S., Patel, D. and McCance, D.J. (2010) MicroRNA 203 expression in keratinocytes is dependent on regulation of p53 levels by E6. *J Virol*, **84**, 10644-10652.
  203. Gunasekharan, V. and Laimins, L.A. (2013) Human papillomaviruses modulate microRNA 145 expression to directly control genome amplification. *J Virol*, **87**, 6037-6043.
  204. Martinez, I., Gardiner, A.S., Board, K.F., Monzon, F.A., Edwards, R.P. and Khan, S.A. (2008) Human papillomavirus type 16 reduces the expression of microRNA-218 in cervical carcinoma cells. *Oncogene*, **27**, 2575-2582.
  205. Liu, W., Gao, G., Hu, X., Wang, Y., Schwarz, J.K., Chen, J.J., Grigsby, P.W. and Wang, X. (2014) Activation of miR-9 by human papillomavirus in cervical cancer. *Oncotarget*, **5**, 11583-11593.
  206. Yeung, C.L., Tsang, T.Y., Yau, P.L. and Kwok, T.T. (2017) Human papillomavirus type 16 E6 suppresses microRNA-23b expression in human cervical cancer cells through DNA methylation of the host gene C9orf3. *Oncotarget*, **8**, 12158-12173.
  207. Wang, X., Wang, H.K., Li, Y., Hafner, M., Banerjee, N.S., Tang, S., Briskin, D., Meyers, C., Chow, L.T., Xie, X. *et al.* (2014) microRNAs are biomarkers of oncogenic human papillomavirus infections. *Proc Natl Acad Sci U S A*, **111**, 4262-4267.
  208. Lajer, C., Garnaes, E., Friis-Hansen, L., Norrild, B., Therkildsen, M., Glud, M., Rossing, M., Lajer, H., Svane, D., Skotte, L. *et al.* (2012) The role of miRNAs in human papillomavirus (HPV)-associated cancers: bridging between HPV-related head and neck cancer and cervical cancer. *Br J Cancer*, **106**, 1526-1534.
  209. Harden, M.E., Prasad, N., Griffiths, A. and Munger, K. (2017) Modulation of microRNA-mRNA Target Pairs by Human Papillomavirus 16 Oncoproteins. *MBio*, **8**.
  210. Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D. and Bray, F. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, **136**, E359-386.
  211. Schiffman, M., Wentzensen, N., Wacholder, S., Kinney, W., Gage, J.C. and Castle, P.E. (2011) Human papillomavirus testing in the prevention of cervical cancer. *J Natl Cancer Inst*, **103**, 368-383.

212. Schiller, J.T., Castellsagué, X. and Garland, S.M. (2012) A review of clinical trials of human papillomavirus prophylactic vaccines. *Vaccine*, **30 Suppl 5**, F123-138.
213. Schiller, J.T. and Müller, M. (2015) Next generation prophylactic human papillomavirus vaccines. *Lancet Oncol*, **16**, e217-225.
214. Serrano, B., Alemany, L., Tous, S., Bruni, L., Clifford, G.M., Weiss, T., Bosch, F.X. and de Sanjosé, S. (2012) Potential impact of a nine-valent vaccine in human papillomavirus related cervical disease. *Infect Agent Cancer*, **7**, 38.
215. Lombard, I., Vincent-Salomon, A., Validire, P., Zafrani, B., de la Rochefordière, A., Clough, K., Favre, M., Pouillart, P. and Sastre-Garau, X. (1998) Human papillomavirus genotype as a major determinant of the course of cervical cancer. *J Clin Oncol*, **16**, 2613-2619.
216. Shi, W., Kato, H., Perez-Ordóñez, B., Pintilie, M., Huang, S., Hui, A., O'Sullivan, B., Waldron, J., Cummings, B., Kim, J. *et al.* (2009) Comparative prognostic value of HPV16 E6 mRNA compared with in situ hybridization for human oropharyngeal squamous carcinoma. *J Clin Oncol*, **27**, 6213-6221.
217. Ang, K.K., Harris, J., Wheeler, R., Weber, R., Rosenthal, D.I., Nguyen-Tân, P.F., Westra, W.H., Chung, C.H., Jordan, R.C., Lu, C. *et al.* (2010) Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*, **363**, 24-35.
218. Meulendijks, D., Tomaso, N.B., Dewit, L., Smits, P.H., Bakker, R., van Velthuysen, M.L., Rosenberg, E.H., Beijnen, J.H., Schellens, J.H. and Cats, A. (2015) HPV-negative squamous cell carcinoma of the anal canal is unresponsive to standard treatment and frequently carries disruptive mutations in TP53. *Br J Cancer*, **112**, 1358-1366.
219. Sasieni, P. and Adams, J. (2001) Changing rates of adenocarcinoma and adenosquamous carcinoma of the cervix in England. *Lancet*, **357**, 1490-1493.
220. Anderson, G.H., Benedet, J.L., Le Riche, J.C., Maticic, J.P. and Thompson, J.E. (1992) Invasive cancer of the cervix in British Columbia: a review of the demography and screening histories of 437 cases seen from 1985-1988. *Obstet Gynecol*, **80**, 1-4.
221. Smith, H.O., Tiffany, M.F., Qualls, C.R. and Key, C.R. (2000) The rising incidence of adenocarcinoma relative to squamous cell carcinoma of the uterine cervix in the United States--a 24-year population-based study. *Gynecol Oncol*, **78**, 97-105.
222. Wright, A.A., Howitt, B.E., Myers, A.P., Dahlberg, S.E., Palescandolo, E., Van Hummelen, P., MacConaill, L.E., Shoni, M., Wagle, N., Jones, R.T. *et al.* (2013) Oncogenic mutations in cervical cancer: genomic differences between adenocarcinomas and squamous cell carcinomas of the cervix. *Cancer*, **119**, 3776-3783.
223. Bulk, S., Berkhof, J., Bulkman, N.W., Zielinski, G.D., Rozendaal, L., van Kemenade, F.J., Snijders, P.J. and Meijer, C.J. (2006) Preferential risk of HPV16 for squamous cell

- carcinoma and of HPV18 for adenocarcinoma of the cervix compared to women with normal cytology in The Netherlands. *Br J Cancer*, **94**, 171-175.
224. Smith, J.S., Lindsay, L., Hoots, B., Keys, J., Franceschi, S., Winer, R. and Clifford, G.M. (2007) Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update. *Int J Cancer*, **121**, 621-632.
  225. Lai, C.H., Chang, C.J., Huang, H.J., Hsueh, S., Chao, A., Yang, J.E., Lin, C.T., Huang, S.L., Hong, J.H., Chou, H.H. *et al.* (2007) Role of human papillomavirus genotype in prognosis of early-stage cervical cancer undergoing primary surgery. *J Clin Oncol*, **25**, 3628-3634.
  226. Clifford, G. and Franceschi, S. (2008) Members of the human papillomavirus type 18 family (alpha-7 species) share a common association with adenocarcinoma of the cervix. *Int J Cancer*, **122**, 1684-1685.
  227. Wang, C.C., Lai, C.H., Huang, H.J., Chao, A., Chang, C.J., Chang, T.C., Chou, H.H. and Hong, J.H. (2010) Clinical effect of human papillomavirus genotypes in patients with cervical cancer undergoing primary radiotherapy. *Int J Radiat Oncol Biol Phys*, **78**, 1111-1120.
  228. Wentz, W.B. and Reagan, J.W. (1959) Survival in cervical cancer with respect to cell type. *Cancer*, **12**, 384-388.
  229. Nakanishi, T., Ishikawa, H., Suzuki, Y., Inoue, T., Nakamura, S. and Kuzuya, K. (2000) A comparison of prognoses of pathologic stage Ib adenocarcinoma and squamous cell carcinoma of the uterine cervix. *Gynecol Oncol*, **79**, 289-293.
  230. Galic, V., Herzog, T.J., Lewin, S.N., Neugut, A.I., Burke, W.M., Lu, Y.S., Hershman, D.L. and Wright, J.D. (2012) Prognostic significance of adenocarcinoma histology in women with cervical cancer. *Gynecol Oncol*, **125**, 287-291.
  231. Rose, P.G., Java, J.J., Whitney, C.W., Stehman, F.B., Lanciano, R. and Thomas, G.M. (2014) Locally advanced adenocarcinoma and adenosquamous carcinomas of the cervix compared to squamous cell carcinomas of the cervix in gynecologic oncology group trials of cisplatin-based chemoradiation. *Gynecol Oncol*, **135**, 208-212.
  232. Yokoi, E., Mabuchi, S., Takahashi, R., Matsumoto, Y., Kuroda, H., Kozasa, K. and Kimura, T. (2017) Impact of histological subtype on survival in patients with locally advanced cervical cancer that were treated with definitive radiotherapy: adenocarcinoma/adenosquamous carcinoma versus squamous cell carcinoma. *J Gynecol Oncol*, **28**, e19.
  233. Shingleton, H.M., Bell, M.C., Fremgen, A., Chmiel, J.S., Russell, A.H., Jones, W.B., Winchester, D.P. and Clive, R.E. (1995) Is there really a difference in survival of women with squamous cell carcinoma, adenocarcinoma, and adenosquamous cell carcinoma of the cervix? *Cancer*, **76**, 1948-1955.

234. Rodriguez-Carunchio, L., Soveral, I., Steenbergen, R., Torne, A., Martinez, S., Fuste, P., Pahisa, J., Marimon, L., Ordi, J. and del Pino, M. (2014) HPV-negative carcinoma of the uterine cervix: a distinct type of cervical cancer with poor prognosis. *BJOG*, **122**, 119-127.
235. Ojesina, A.I., Lichtenstein, L., Freeman, S.S., Peadarallu, C.S., Imaz-Rosshandler, I., Pugh, T.J., Cherniack, A.D., Ambrogio, L., Cibulskis, K., Bertelsen, B. *et al.* (2014) Landscape of genomic alterations in cervical carcinomas. *Nature*, **506**, 371-375.
236. Network, C.G.A.R., Medicine, A.E.C.o., Services, A.B., Hospital, B.C., Medicine, B.C.o., Hope, B.R.I.o.C.o., Aging, B.I.f.R.o., Centre, C.s.M.S.G.S., School, H.M., Services, H.F.G.C.C.R.I.a.C.C.H. *et al.* (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**, 378-384.
237. Warnakulasuriya, S. (2009) Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol*, **45**, 309-316.
238. D'Souza, G., Agrawal, Y., Halpern, J., Bodison, S. and Gillison, M.L. (2009) Oral sexual behaviors associated with prevalent oral human papillomavirus infection. *J Infect Dis*, **199**, 1263-1269.
239. Moore, K.A. and Mehta, V. (2015) The Growing Epidemic of HPV-Positive Oropharyngeal Carcinoma: A Clinical Review for Primary Care Providers. *J Am Board Fam Med*, **28**, 498-503.
240. Parfenov, M., Peadarallu, C.S., Gehlenborg, N., Freeman, S.S., Danilova, L., Bristow, C.A., Lee, S., Hadjipanayis, A.G., Ivanova, E.V., Wilkerson, M.D. *et al.* (2014) Characterization of HPV and host genome interactions in primary head and neck cancers. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 15544-15549.
241. The Cancer Genome Atlas, N. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**, 576-582.
242. Roda, J.M., Joshi, T., Butchar, J.P., McAlees, J.W., Lehman, A., Tridandapani, S. and Carson, W.E. (2007) The activation of natural killer cell effector functions by cetuximab-coated, epidermal growth factor receptor positive tumor cells is enhanced by cytokines. *Clin Cancer Res*, **13**, 6419-6428.
243. Bellone, S., Frera, G., Landolfi, G., Romani, C., Bandiera, E., Tognon, G., Roman, J.J., Burnett, A.F., Pecorelli, S. and Santin, A.D. (2007) Overexpression of epidermal growth factor type-1 receptor (EGF-R1) in cervical cancer: implications for Cetuximab-mediated therapy in recurrent/metastatic disease. *Gynecol Oncol*, **106**, 513-520.
244. Vermorken, J.B., Stöhlmacher-Williams, J., Davidenko, I., Licitra, L., Winkvist, E., Villanueva, C., Foa, P., Rottey, S., Skladowski, K., Tahara, M. *et al.* (2013) Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic

squamous-cell carcinoma of the head and neck (SPECTRUM): an open-label phase 3 randomised trial. *Lancet Oncol*, **14**, 697-710.

245. Masterson, L., Moualed, D., Liu, Z.W., Howard, J.E., Dwivedi, R.C., Tysome, J.R., Benson, R., Sterling, J.C., Sudhoff, H., Jani, P. *et al.* (2014) De-escalation treatment protocols for human papillomavirus-associated oropharyngeal squamous cell carcinoma: a systematic review and meta-analysis of current clinical trials. *Eur J Cancer*, **50**, 2636-2648.

# Chapter 2: Prognostic miRNA Signatures Derived from The Cancer Genome Atlas for Cancers of the Head and Neck and the Cervix

*This chapter is adapted from and expanded upon the following publication (1):*

Wong, Nathan, Shariq S Khwaja, Callie M Baker, Hiram A Gay, Wade L Thorstad, Mackenzie D Daly, James S Lewis, Xiaowei Wang. (2016) Prognostic microRNA signatures derived from The Cancer Genome Atlas for head and neck squamous cell carcinomas. *Cancer Medicine* 5(7): 1619-1628.

## **2.1 Abstract**

**BACKGROUND:** Identification of novel prognostic biomarkers typically requires a large dataset which provides sufficient statistical power for discovery research. To this end, we took advantage of the high-throughput data from The Cancer Genome Atlas (TCGA) to identify a set of prognostic biomarkers in head and neck squamous cell carcinomas (HNSCC) including oropharyngeal squamous cell carcinoma (OPSCC) and other subtypes, and cervical squamous cell carcinomas (CESC).

**METHODS:** In this study we analyzed miRNA-seq data obtained from TCGA patients to identify prognostic biomarkers for OPSCC. The identified miRNAs were further tested with an independent cohort. miRNA-seq data from TCGA was also analyzed to identify prognostic miRNAs in oral cavity squamous cell carcinoma (OSCC), laryngeal squamous cell carcinoma (LSCC), and cervical squamous cell and endocervical carcinoma (CESC).

**RESULTS:** Our study identified that miR-193b-3p and miR-455-5p were positively associated with survival, and miR-92a-3p and miR-497-5p were negatively associated with survival in OPSCC. A combined expression signature of these four miRNAs was prognostic of overall survival in OPSCC, and more importantly, this signature was validated in an independent OPSCC cohort. Furthermore, we identified four miRNAs each in oral squamous cell carcinoma (OSCC) and laryngeal squamous cell carcinoma (LSCC) that were prognostic of survival, and combined signatures were specific for subtypes of HNSCC. An additional signature was developed for CESC that was significant across cervical tumor subtypes.

**CONCLUSIONS:** A robust 4-miRNA prognostic signature in OPSCC, as well as prognostic signatures in other subtypes of HNSCC and within CESC, was developed using sequencing data from TCGA as the primary source. This demonstrates the power of using TCGA as a potential resource to develop prognostic tools for improving individualized patient care.

## **2.2 Introduction**

Head and neck squamous cell carcinoma (HNSCC) constitutes approximately 3% of all cancer diagnoses in the United States, with about 45,000 new cases in 2015 (2). Among head and neck cancers, oral cavity, oropharyngeal and laryngeal cancers are the most common, accounting for 24%, 23% and 27% of all diagnosed cases, respectively (3).

Due to the heterogeneity of these subtypes of HNSCC, a single prognostic signature identifying high- and low-risk patients cannot be generated to cover all types of HNSCC. However, multiple studies have indicated that individual biomarkers can stratify high-risk and low-risk patients within the various subtypes (4-6). These biomarkers are not limited to coding genes. Included among the proposed biomarkers are microRNAs (miRNAs), which are short single-stranded RNA sequences (~22 n.t.) that function in post-transcriptional regulation. Further

studies have shown that within oropharyngeal cancer, infection by human papillomavirus (HPV) is a favorable prognostic marker (7). Greater prognostic power has been attained by combining groups of biomarkers into a single signature for different subtypes of HNSCC, with various degrees of success (8,9).

The other major cancer type associated with HPV infection is cervical cancer (CESC). An estimated 528,000 new cases of cervical cancer worldwide are diagnosed annually, 95% of which are caused by HPV (10,11). Cervical cancers can be further categorized by their tumor source site, the majority of which can be classified as either cervical squamous cell carcinomas (CSCC), or cervical adenosquamous carcinomas or adenocarcinomas (cervical adeno-type cancers or CASC). As with oropharyngeal cancer, miRNAs have been examined as potential biomarkers for determining patient prognosis, with several signatures having been published in the literature (12,13).

The identification of novel biomarkers and subsequent development of prognostic signatures requires in-depth analysis of genetic profiles. For example, high-throughput gene expression profiling data have been made available by The Cancer Genome Atlas (TCGA), a joint effort of the National Cancer Institute and the National Human Genome Research Institute to provide a comprehensive set of patient genetic profiles across multiple cancer types (14). This has extended to HNSCC, with a total of 529 HNSCC and 304 CESC samples being made available (15,16). Included in the available data are RNA-seq and miRNA-seq profiles for the majority of the provided patient samples.

In the current study, we investigated the prognostic value of miRNA biomarkers for oropharyngeal squamous cell carcinoma (OPSCC), oral squamous cell carcinoma (OSCC), and laryngeal squamous cell carcinoma (LSCC), using profiling data obtained from TCGA. These



biomarkers were then used to develop unique prognostic signatures that robustly predicted overall survival in the respective subsets of HNSCC. An additional signature was identified through TCGA analysis for cervical cancer that maintained significance in both squamous cell carcinoma (CSCC) and cervical adenosquamous carcinomas and adenocarcinomas (CASC). We further demonstrate that use of the TCGA public dataset can provide a more general picture of head and neck cancer as the prediction models obtained can be applied to an independently obtained dataset. Through a combined analysis of TCGA data and independently generated data, we have provided an additional set of biomarker tools for the clinical setting that can assist in determining the best course of treatment for patients with head and neck cancer.

## **2.3 Materials and Methods**

### *Retrieval of Public Data*

A total of 523 anonymized patients in the TCGA database were identified as having primary HNSCC. The clinical patient files were downloaded from TCGA Data Portal ([tcga-data.nci.nih.gov](http://tcga-data.nci.nih.gov)). Of the 523 HNSCC patients, 82 patients had a primary tumor in the oropharynx, 313 patients had a primary tumor of the oral cavity, and 115 patients had primary tumors in the larynx. A total of 304 patients were identified as having primary CESC. Of the 307, 254 patients had a primary CSCC, 50 had a primary CACC, and 3 had a primary endothelial carcinoma. A cutoff of five years was applied to all patient survival data.

All gene sequences were downloaded from the UCSC Genome Browser (17). Index files mapping transcript accessions to NCBI Gene IDs were downloaded from the NCBI ftp site (18). All mature miRNA sequences were downloaded from miRBase (19). Raw miRNA-seq data was obtained for 81 of the 82 OPSCC patients, 311 of 313 OSCC patients, and all of the laryngeal

cancer patients. Raw RNA-seq data was obtained for 72 of the 82 oropharyngeal cancer patients. In the cervical cancer cohort, miRNA-Seq and RNA-Seq data were obtained for 227 C5CC patients and 45 CACC patients. All raw RNA-seq and miRNA-seq files were downloaded through the Cancer Genomics Hub (20).

### *TCGA Sequence Analysis*

Sequence alignment was performed using the Bowtie program (21). Raw miRNA-seq reads were aligned to the human miRNome. The read counts were then normalized to reads per million reads mapped per sample and set to a floor value of 1 for lowly expressed miRNAs. Raw RNA-seq reads were aligned sequentially to human RefSeq annotated sequences, the human reference genome, and the virome. The read counts were normalized to reads per kilobase per million mapped reads, then to the 2000<sup>th</sup> gene before being set to a floor of 5 normalized reads for lowly expressed mRNAs. Both miRNA-seq and RNA-seq reads were subsequently log<sub>2</sub> normalized.

### *Statistical Analysis for Survival and Correlation*

Overall survival analysis was conducted using the ‘survival’ package in R (<http://www.r-project.org>). Correlation and covariance analysis was conducted in MATLAB (22). Univariate Cox proportional hazards regression analyses were performed to evaluate the correlation between individual miRNAs or mRNAs with overall survival. The p-values for outcome correlation were calculated using the Wald test. The final prognostic signatures were also evaluated in this manner. Multivariate Cox proportional hazards analyses were conducted to evaluate the independent prognostic value of the miRNA signature after controlling for common

clinical variables. The Kaplan-Meier estimator was used to determine the empirical survival probabilities and p-values from the log-rank test indicated the significance of the miRNA prediction outcome model.

#### *Collection of Independent Validation Data Sets*

A total of 95 OPSCC cases were included in this study for validation. Patients were treated at Washington University School of Medicine with definitive chemoradiation, or with primary surgery followed by radiation therapy with or without chemotherapy. Clinical data were collected from the patients and then updated retrospectively after follow-up review.

For all 95 of the patients, formalin-fixed, paraffin-embedded (FFPE) tumor tissues were collected for pathological analysis before radiotherapy or chemotherapy. Sections from each case were stained with hematoxylin and eosin and reviewed by a study pathologist at Washington University to confirm the diagnoses. Tumor regions from each section were identified and macrodissection was conducted. Total RNA was extracted from the identified tumor regions using the miRNeasy FFPE kit (Qiagen) according to the manufacturer's protocol. 66 patients were used for the validation of the OPSCC miRNA prognostic model, and 39 patients for the validation of the OPSCC mRNA model.

#### *Quantitative Reverse Transcription PCR for miRNA model validation*

Quantitative reverse transcription polymerase chain reaction (qRT-PCR) was used to profile the miRNAs identified as significant in OPSCC and CESC. The details of this experimental procedure have been described previously (23). Briefly, the RT reaction was performed with the High Capacity cDNA Reverse Transcription Kit (Life Technologies). Each

RT reaction included 100 ng of tumor RNA and a pool of RT primers for selected miRNAs and control RNAs. Quantitative PCR was performed with Power SYBR Green PCR Master Mix (Life Technologies) and specific PCR primers for selected miRNAs or control RNAs. miRNA raw profiling data for individual samples were normalized with four small RNA controls (SNORD48, SNORD47, RNA5-8S5 and RNU6-1). Specifically, the expression levels of the four small RNAs were averaged and used as the reference to control for sample variations during miRNA profiling analysis.

The expression of p16 protein was determined by immunohistochemistry as previously described (24). The expression profiles of E6 and E7 transcripts from six oncogenic HPV types were determined by qRT-PCR, including types 16, 18, 33, 39, 56, and 59. The details of the HPV assays and the experimental protocol have been described previously (24). In brief, primer sequences for the assays were selected from the E6 and E7 coding regions of the high-risk HPV genomes. The expression profiles of GAPDH and  $\beta$ -actin were used as reference controls for data normalization.

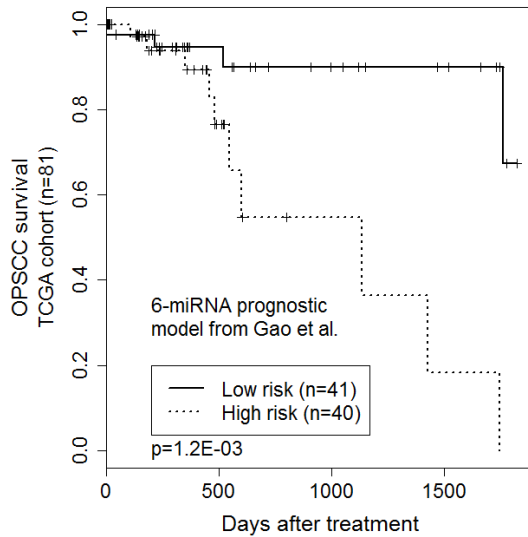
## **2.4 Results**

### *Validation of an existing miRNA prognostic signature*

To verify that the miRNA data obtained from The Cancer Genome Atlas (TCGA) could be used in further biomarker identification, we evaluated our previously published prognostic model for OPSCC (9) with TCGA data. Briefly, this model identified miR-24-3p, miR-31-5p, and miR-193b-3p as negatively associated with survival, and miR-26b-5p, miR-142-3p, and miR-146a-5p as positively associated with survival. The expression levels of these 6 miRNAs were then combined to create a single prognostic model as described previously (9):

**A**

<b>miRNA Name</b>	<b>Fold change</b>	<b>P-value</b>
miR-24-3p	0.08	0.62
miR-31-5p	0.14	0.11
miR-193b-3p	0.83	4.6E-03
miR-26b-5p	-0.33	0.097
miR-142-3p	-0.31	0.065
miR-146a-5p	-0.30	0.05

**B**

**Figure 2.1.** Validation of an existing miRNA signature with TCGA data. **(A)** The six miRNAs from our previously published prognostic model for OPSCC were examined for association with the overall survival of TCGA patients. Fold change values were  $\log_2$  transformed and represent the average expression difference of the miRNAs in the deceased patient group compared to the living patient group. Statistical significance was determined with the logrank test in Cox regression analysis. **(B)** Kaplan-Meier survival analysis to evaluate the prognostic performance of the six-miRNA signature for predicting overall survival in OPSCC.

$$S = 2.62E_{\text{miR-24-3p}} + 3.16E_{\text{miR-31-5p}} + 2.45E_{\text{miR-193b-3p}} - 2.69E_{\text{miR-26b-5p}} - 3.34E_{\text{miR-142-3p}} - 2.81E_{\text{miR-146a-5p}}$$

146a-5p

We examined each miRNA individually with TCGA data and found that two miRNAs were significantly associated with survival, two miRNAs maintained borderline significance ( $p < 0.1$ ), and two miRNAs was not found to be significant ( $p > 0.1$ ) (Figure 2.1A). It should be noted, however, that the directions of expression changes in relation to survival outcome were maintained for all six miRNAs (i.e. positive correlations for miR-26b-5p, miR-142-3p, miR-146a-5p, and negative correlations for miR-31-5p, miR-193b-3p, miR-26b-5p) (Figure 2.1A). When we analyzed this prognostic model as whole, it was able to significantly differentiate between high- and low-risk OPSCC patients from TCGA (Figure 2.1B). This demonstrated that this previously published model was robust and could be applied to patient miRNA profiles obtained from other institutions, while also indicating that the data from TCGA was a valuable resource for further biomarker identification and analysis.

#### *Unique TCGA miRNA expression profiles correlated with overall survival in OPSCC*

miRNA expression analysis was performed for the 81 OPSCC patients obtained from TCGA. The characteristics of these patients are summarized in Table 1. The miRNAs were examined individually using Cox univariate proportional hazards analysis to determine which miRNAs were correlated with overall survival. This analysis provides a log-rank p-value, which indicates the significance of the miRNA in relation to survival, as well as a Wald coefficient, which indicates the weight associated with the expression level of the miRNA.

We then implemented recursive feature elimination (RFE) technique to determine the relative prognostic performance of individual miRNAs. In this process, a regression model was generated using the given miRNA features and outcomes (i.e. miRNA expression and overall survival, respectively), and the least impactful feature was eliminated. The process was then

**Table 2.1.** Characteristics of the HNSCC patients included in TCGA.

<b>CHARACTERISTICS</b>	<b>OPSCC (n=81)</b>	<b>OSCC (n=311)</b>	<b>LSCC (n=115)</b>
<b>Age at diagnosis (mean <math>\pm</math> SD, y)</b>	55.9 $\pm$ 9.3	61.9 $\pm$ 13.2	61.9 $\pm$ 9.1
<b>Sex</b>			
Male	69 (85.2%)	206 (66.2%)	95 (82.6%)
Female	12 (14.8%)	105 (33.8%)	20 (17.4%)
<b>Race</b>			
White	75 (92.6%)	268 (86.2%)	91 (79.1%)
Other	6 (7.4%)	43 (13.8%)	24 (20.9%)
<b>Smoking<sup>a</sup></b>			
Unreported	1 (1.2%)	10 (3.2%)	4 (3.5%)
Non-smoker	25 (30.8%)	88 (28.3%)	6 (5.2%)
Long-term former smoker	8 (9.9%)	51 (16.4%)	11 (9.6%)
Other former smoker	25 (30.9%)	68 (21.9%)	36 (31.3%)
Current smoker	22 (27.2%)	94 (30.2%)	58 (50.4%)
<b>T classification</b>			
T1	13 (16.0%)	29 (9.3%)	7 (6.1%)
T2	36 (44.4%)	102 (32.8%)	20 (17.4%)
T3	20 (24.7%)	64 (20.6%)	33 (28.7%)
T4	12 (14.8%)	116 (37.3%)	55 (47.8%)
<b>N Classification</b>			
NX	0 (0.0%)	4 (1.3%)	2 (1.7%)
N0	21 (25.9%)	142 (45.7%)	52 (45.2%)
N1	52 (64.2%)	52 (16.7%)	12 (10.4%)
N2	3 (3.7%)	110 (35.4%)	46 (40.0%)
N3	5 (6.2%)	3 (1.0%)	3 (2.6%)
<b>Stage</b>			
I	5 (6.2%)	19 (6.1%)	2 (1.7%)
II	11 (13.6%)	62 (19.9%)	15 (13.0%)
III	12 (14.8%)	57 (18.3%)	18 (15.7%)
IV	53 (65.4%)	173 (55.6%)	80 (69.6%)
<b>Deceased</b>	14 (17.2%)	109 (35.0%)	33 (28.7%)

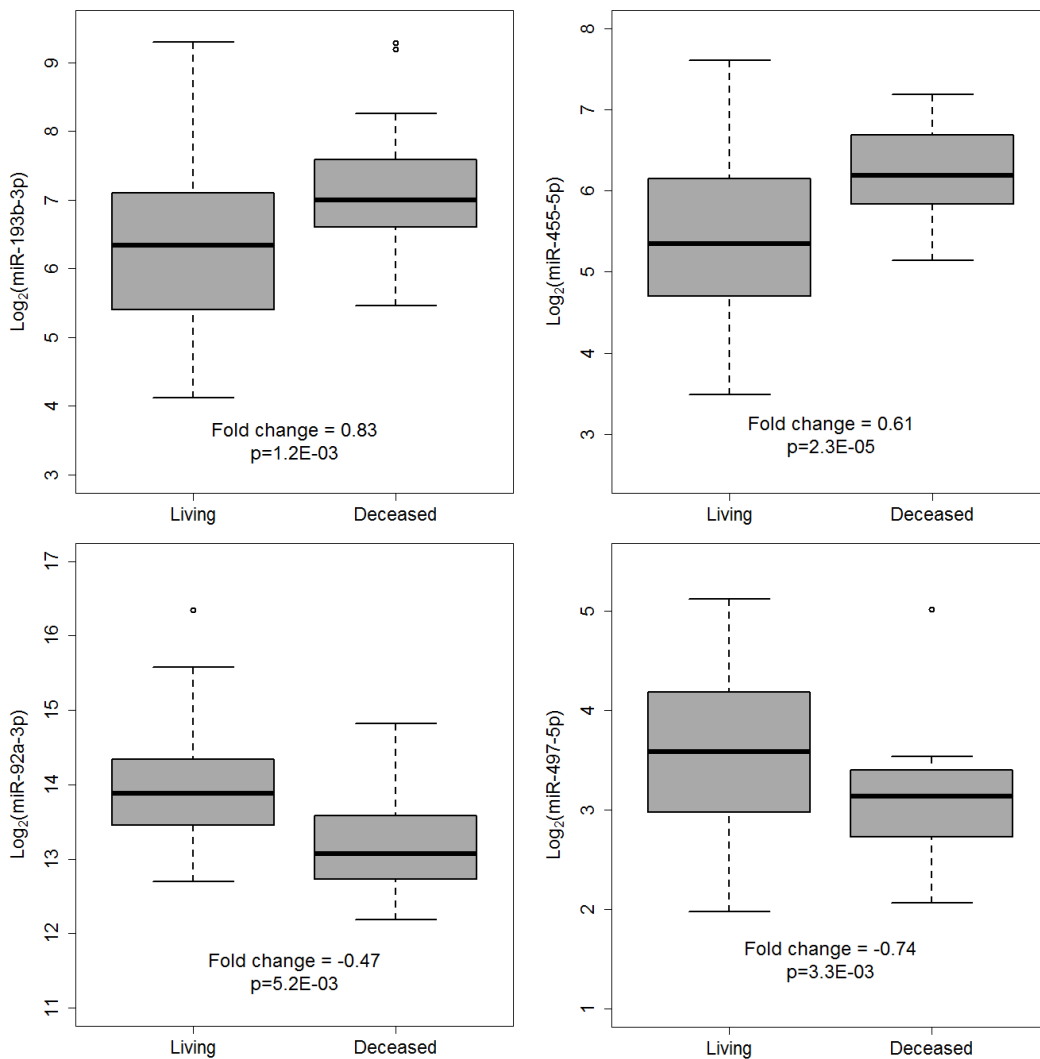
Abbreviations: OPSCC, oropharyngeal squamous cell carcinoma; OSCC, oral cavity squamous cell carcinoma; LSCC, laryngeal squamous cell carcinoma; SD, standard deviation

<sup>a</sup> Smoking was defined as no history of smoking, a former smoker of  $\geq$  15 years, other former smoker of  $<$ 15 years, or a current smoker.

repeated until the final iteration identified the most significant feature associated with the classifier. This was performed on a subset of top-ranking 189 miRNAs in OPSCC ordered by log-rank p-value while maintaining a Wald coefficient greater than or equal to one, and an

average expression across all samples greater than 1.414 (i.e. a  $\log_2$ Expression of 0.5). In this way, we were able to initially identify a set of promising miRNA candidates for further model development.

In examining the 50 most significant miRNAs in accordance with the RFE, miR-193b-3p, miR-455-5p, miR-92a-3p, and miR-497-5p were identified as maintaining a high RFE ranking after 10-fold cross-validation, as well as being statistically significant in the univariate Cox



**Figure 2.2.** Four significant miRNAs associated with overall survival of TCGA OPSCC patients. Fold change values were  $\log_2$  transformed and represent the average expression difference of the miRNAs in the deceased patient group compared to the living patient group. Significance was determined with the logrank test in Cox regression analysis.



proportional hazards analysis (Figure 2.2). All four of these miRNAs have been reported as dysregulated in other cancer types, including colorectal cancer and pancreatic cancer (25-28). We further confirmed the validity of miR-193b-3p as a prognostic marker in OPSCC, which we had previously reported and incorporated in our previous model for outcome prediction (9).

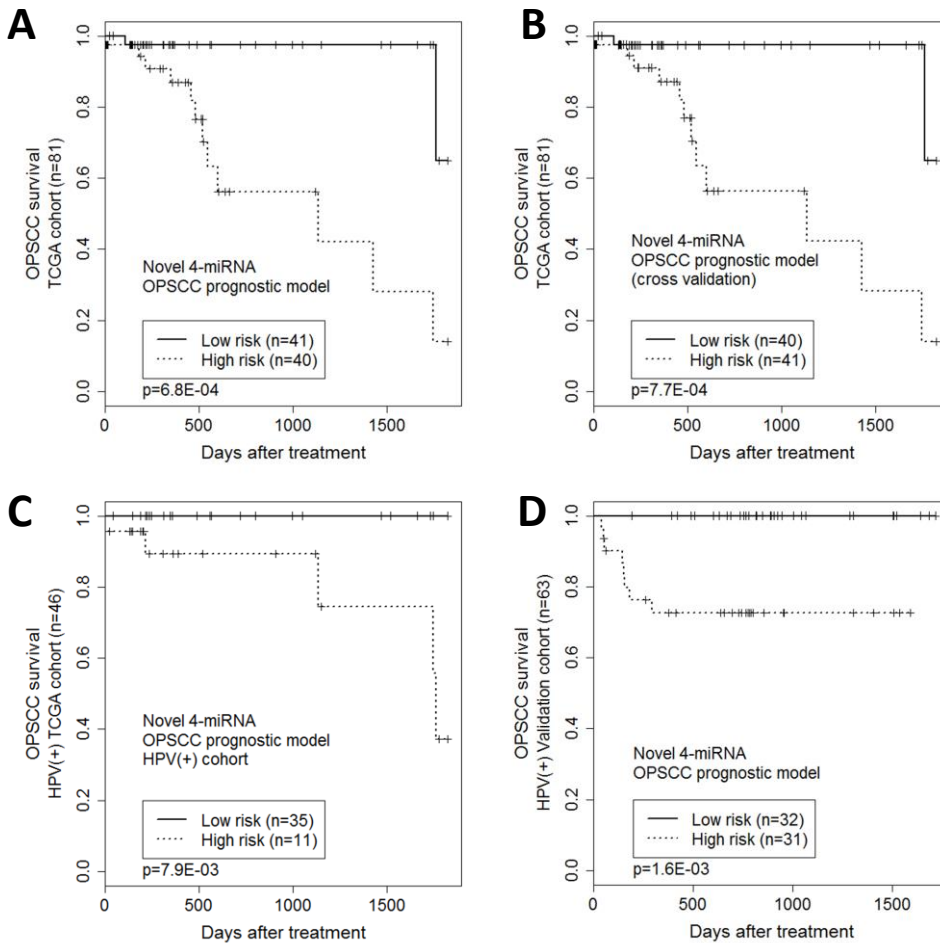
*A combined miRNA prognostic signature predicts overall survival in OPSCC*

We further hypothesized that a combination of prognostic miRNAs within OPSCC could be effectively used to predict overall survival. The miRNAs chosen in the aforementioned analysis were used to build the following prognostic model:

$$S_{\text{OPSCC}} = 11.31E_{\text{miR-193b-3p}} + 13.53E_{\text{miR-455-5p}} - 7.25E_{\text{miR-92a-3p}} - 7.3E_{\text{miR-497-5p}},$$

where S indicates the risk score for each patient and E represents the normalized expression level of the identified miRNA from the primary tumor. The coefficients in this equation are the Wald scores from the Cox regression analysis and are representative of the relative importance of the miRNA towards survival status.

In this prediction model, higher scores indicate higher risk and predict a poor survival outcome for the patient. The patients were stratified internally by median risk score to produce 2 cohorts of similar size, so as to determine the validity of the prognostic model. By this method, 40 OPSCC patients were predicted to be high-risk (with > median score) and 41 patients were predicted to be low-risk (i.e. with  $\leq$  median score); significantly different risks of death were observed based on this classification ( $p = 6.8E-04$ ) (Figure 2.3A).



**Figure 2.3.** Kaplan-Meier survival analysis to evaluate the novel OPSCC 4-miRNA prognostic signature. Patients were stratified into the low risk group or high risk group based on risk score. **(A)** The signature was evaluated for overall survival in the training set from TCGA. Significance was determined using the logrank test. **(B)** Leave-one-out cross-validation to evaluate the miRNA modeling strategy. The cross-validated results from all rounds were combined for prognostic evaluation of overall survival. **(C)** Independence of the miRNA signature in HPV(+) patients. **(D)** Survival analysis to evaluate the miRNA signature for overall survival in the validation cohort.

One primary concern for prognostic model development is the risk of overtraining. To address this issue, we performed leave-one-out cross-validation. For this cross-validation, within each iteration, we removed one sample from the training set and trained a model with the miRNA profiles from the remaining samples. The removed sample was then used for independent model testing. The process was repeated until all the samples had been used independently for model testing. For each validation round, the Wald coefficient for the

candidate miRNAs were calculated based on the training set and used to generate a slightly different model for testing. Cross-validation still yielded a significant separation of high- and low-risk patients (Figure 2.3B), indicating that the model is robust within the training data.

*The miRNA prognostic signature was independent of clinical features*

We assessed if the miRNA signature maintained its prognostic value within the context of commonly used clinical parameters, including age at diagnosis, gender, race, smoking history, initial tumor staging, and treatment type. This analysis was conducted through multivariate Cox hazards analysis. This miRNA signature was found to maintain statistical significance, with a hazards ratio of 11.85 and p-value of 3.9E-03 (Table 2.2).

*The OPSCC miRNA signature maintained its prognostic value independent of HPV status*

Previous work has shown that HPV positivity is a favorable prognostic marker in OPSCC, and thus we extended our miRNA signature to explore whether the prognostic significance was maintained independently of HPV status. OPSCC patients were identified as HPV-positive if sequencing reads from the RNA-seq data that did not align to the human

**Table 2.2.** Multivariate Cox regression analysis to evaluate independence of the prognostic miRNA signatures from clinical parameters.

Parameter	OPSCC		OSCC		LSCC	
	HR	P-value <sup>a</sup>	HR	P-value <sup>a</sup>	HR	P-value <sup>a</sup>
miRNA signature	11.847	0.0039	1.88	1.8E-03	2.843	1.3E-02
Age	1.056	0.054	1.017	6.1E-02	0.978	0.32
Sex	0.741	0.68	1.129	0.6	0.455	5.6E-02
Stage (I/II/III vs IV)	2.936	0.11	2.155	6.3E-04	1.013	0.91
Tobacco	1.386	0.31	1.052	0.54	1.17	0.40
Treatment (chemotherapy vs radiotherapy vs combined)	0.637	0.034	0.906	0.96	0.866	0.17
Race (White vs. other)	7.906	0.11	0.919	0.95	2.369	2.6E-02

<sup>a</sup> P-values were calculated using the Wald test.

**Table 2.3.** Characteristics of the OPSCC patients at Washington University

	<b>OPSCC miRNA validation cohort (n=66)</b>	<b>OPSCC mRNA validation cohort (n=39)</b>
<b>Age at diagnosis (mean <math>\pm</math> SD, y)</b>	58.5 $\pm$ 10.2	55.8 $\pm$ 10.3
<b>Sex</b>		
Male	54 (81.8%)	36 (92.3%)
Female	12 (18.2%)	3 (7.7%)
<b>Race</b>		
White	65 (98.5%)	36 (92.3%)
Other	1 (1.5%)	3 (7.7%)
<b>Smoking</b>		
Unreported	1 (1.5%)	4 (10.3%)
Non-smoker	22 (33.3%)	12 (30.8%)
Former smoker	27 (40.9%)	20 (51.3%)
Current smoker	16 (24.2%)	3 (7.7%)
<b>T Classification</b>		
Tx	6 (9.1%)	6 (15.4%)
T1	28 (42.4%)	14 (34.9%)
T2	15 (22.7%)	9 (23.1%)
T3	7 (10.6%)	4 (10.3%)
T4	10 (15.2%)	6 (15.4%)
<b>N Classification</b>		
NX	0 (0.0%)	6 (15.4%)
N0	12 (18.2%)	2 (5.1%)
N1	13 (19.7%)	4 (10.2%)
N2	37 (56.1%)	24 (61.5%)
N3	4 (6.1%)	3 (7.7%)
<b>Stage</b>		
Unreported	0 (0.0%)	6 (15.4%)
I	4 (6.1%)	1 (2.6%)
II	5 (7.6%)	0 (0.0%)
III	15 (22.7%)	5 (12.8%)
IV	42 (63.6%)	27 (69.2%)
<b>Deceased</b>	10 (15.2%)	14 (35.9%)

genome aligned to any of the 143 types of HPV. Of the 72 OPSCC patients with RNA-seq data, 46 were identified as having reads aligned to one of three types of HPV. Specifically, 39 patients were positive for HPV16, four for HPV33, and three for HPV35, leaving 26 patients as HPV-negative.

Of the 46 patients who were identified as HPV positive, 35 were identified as low-risk and 11 as high-risk by the miRNA prognostic signature. Kaplan-Meier survival analysis indicated that the high-risk group had poor survival as compared to the low-risk group ( $p = 7.9E-03$ ) (Figure 2.3C). The model was not statistically significant when applied to HPV-negative patients (data not shown); however, it should be noted that the HPV-negative set was a much smaller cohort ( $n=26$ ), which significantly reduced the power of the model.

#### *Validation of the OPSCC miRNA signature with an independent cohort*

To confirm the validity of the 4-miRNA model for OPSCC prognosis, we applied our miRNA signature to an independent cohort of 66 OPSCC patients treated at the Washington University School of Medicine in St. Louis. The clinical characteristics of these patients are outlined in Table 2.3. We hypothesized that the miRNA signature provides independent prognostic value from HPV biomarker. Since HPV positivity is a favorable prognostic marker for OPSCC, we were interested to know whether the new miRNA signature maintains its prognostic value by further risk-stratifying HPV-positive patients.

All 66 patients were pre-selected to be p16 positive by immunohistochemistry, as p16 is a robust surrogate biomarker for HPV expression (24). HPV expression in these tumors was further validated by quantitative reverse-transcription PCR (qRT-PCR, see Methods for details). Of the 66 tumors, 61 were HPV16 positive and two were HPV18 positive. HPV transcripts were not detected in the remaining three samples.

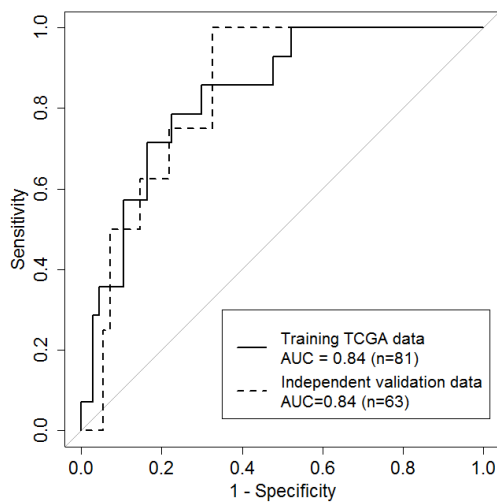
Furthermore, qRT-PCR was conducted on the tumor samples for the four miRNAs included in the signature. miRNA expression readings were normalized using four internal small RNA controls (see Methods for details). The risk score was then calculated for each of these

patients based on the miRNA signature. The patients were then stratified into high-risk and low-risk groups by the median risk score. Kaplan-Meier survival analysis indicated that the miRNA model was significantly predictive of survival outcome for the 63 HPV-positive cases ( $p = 1.6E-03$ , Figure 2.3D). The miRNA signature had a similar prognostic performance when applied to all 66 p16-positive cases ( $p = 2.8E-03$ ).

We also analyzed the signature prediction scores with a receiver operating characteristic (ROC) curve, which evaluated both the true positive rate (sensitivity) and the false positive rate (specificity). In the training and validation sets, the areas under the curve were 0.84 and 0.84, respectively, indicating robust performance of the model for both sensitivity and specificity when applied to independent cohorts (Figure 2.4).

#### *Unique miRNA expression profiles correlated with distinct subtypes of head and neck cancer*

We extended our miRNA expression profiling analysis to the 311 OSCC and 115 LSCC patients obtained from TCGA. The characteristics of these patients are summarized in Table 1.



**Figure 2.4.** Receiver operating characteristic (ROC) curves for the training and validation cohorts from TCGA. The high Area Under the Curve (AUC) values indicate strong sensitivity and specificity.

**Table 2.4.** Significantly dysregulated miRNAs associated with overall survival and used to develop prognostic models for OSCC and LSCC, respectively.

	miRNA name	Fold change <sup>a</sup>	p-value <sup>b</sup>
OSCC	hsa-miR-337-3p	0.220	8.6E-04
	hsa-miR-369-5p	0.428	5.5E-03
	hsa-miR-218-5p	0.197	1.4E-02
	hsa-miR-127-5p	0.381	7.0E-03
LSCC	hsa-let-7a-3p	-0.710	5.2E-04
	hsa-miR-145-5p	-0.440	6.2E-03
	hsa-miR-129-5p	1.349	3.8E-02
	hsa-miR-26b-5p	-0.333	8.4E-03

<sup>a</sup> Fold change is log2 normalized.

<sup>b</sup> P-values are from the logrank score from Cox univariate analysis.

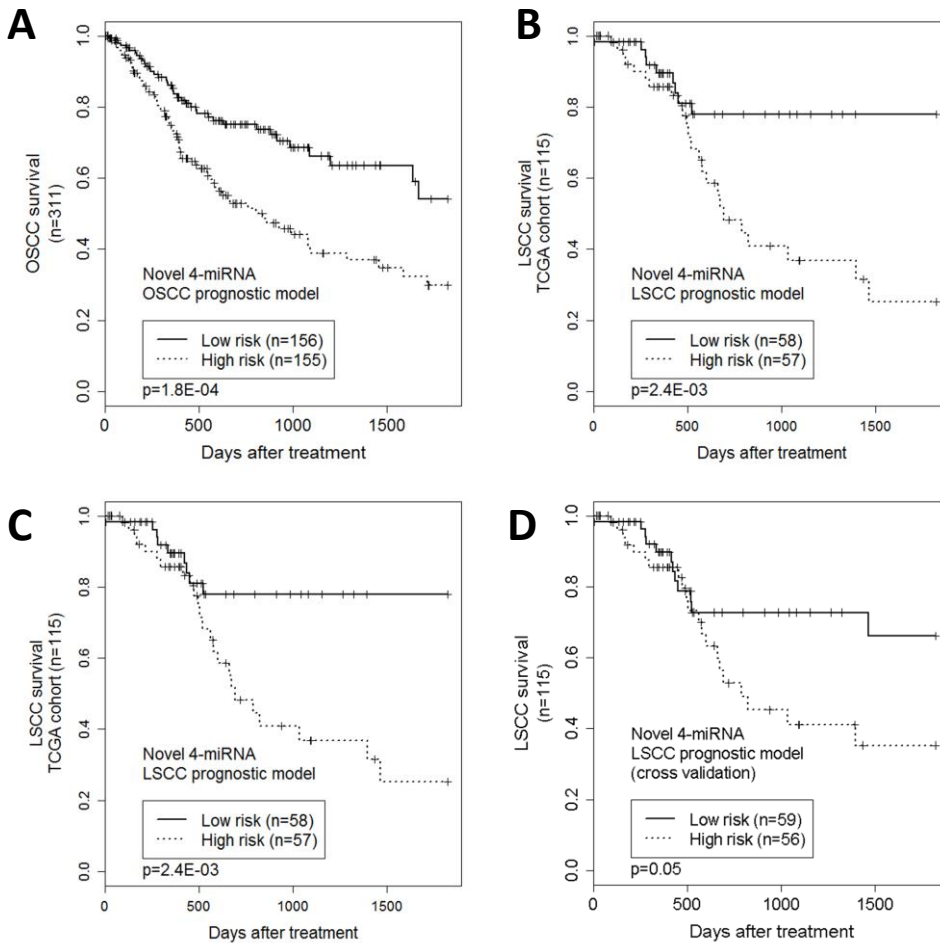
For each additional subtype of HNSCC, we conducted similar analyses as described for OPSCC and identified 4 miRNAs in each subset that were predictive of overall survival (Table 2.4).

These miRNAs were then combined to generate the following prognostic models:

$$S_{OSCC} = 10.73E_{miR-337-3p} + 7.82E_{miR-369-5p} + 6.21E_{miR-218-5p} + 7.01E_{miR-127-5p},$$

$$S_{LSCC} = -10.70E_{let-7a-3p} - 6.96E_{miR-145-5p} + 4.59E_{miR-129-5p} - 6.43E_{miR-26b-5p},$$

As described earlier, the median score was used within each subset to separate patients into high- and low-risk groups, which were also found to have significantly different risks of death ( $p = 1.8E-04$  in OSCC and  $p = 2.4E-03$  for LSCC) (Figures 2.5A and B). We also conducted leave-one-out cross-validation analysis for these two signatures and found a significantly different risk of survival in the miRNA-stratified groups of OSCC and a borderline significance for LSCC (Figures 2.5C and D). Despite borderline significance of the LSCC model in cross-validation analysis, the LSCC prognostic miRNA model may still be useful for prediction of patient survival. In particular, these models maintained statistical significance

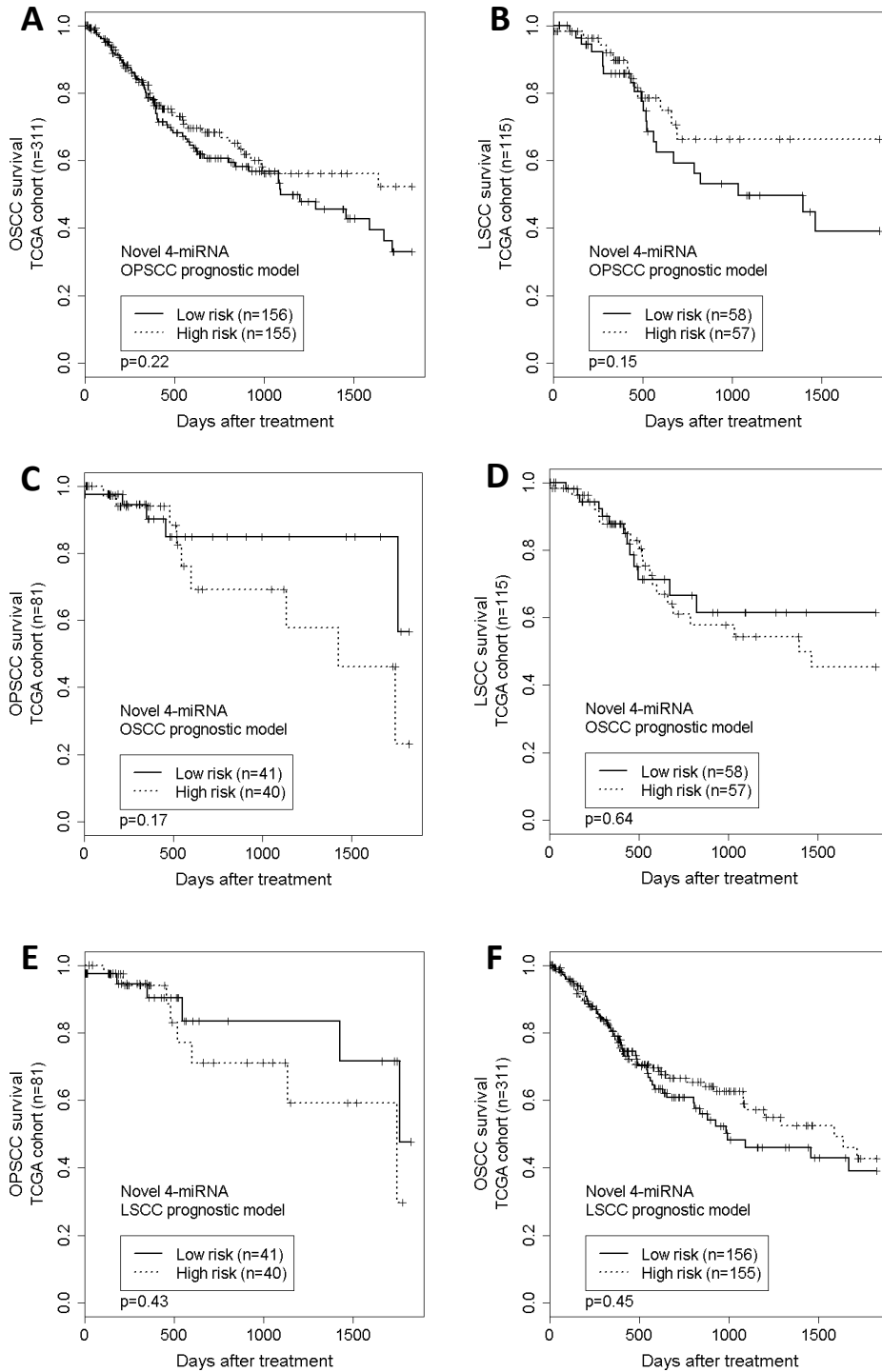


**Figure 2.5.** Kaplan-Meier survival analysis to evaluate the OSCC and LSCC miRNA prognostic models. (A, B) The models were evaluated in the respective training sets. (C, D) Leave-one-out cross-validation results were combined for prognostic evaluation.

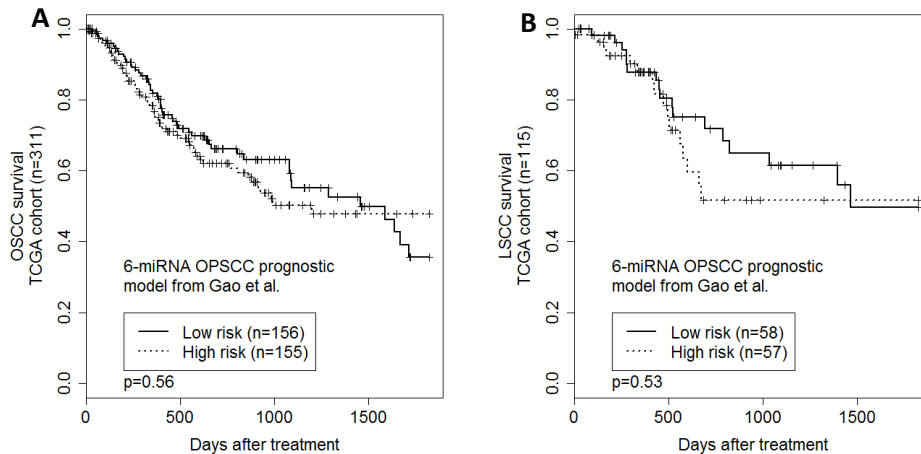
independently of clinical features when analyzed with multivariate Cox analysis, with the OSCC model having a hazards ratio of 1.88 and a p-value of 1.8E-03, and the LSCC model having a hazards ratio of 2.84 and a p-value of 1.3E-02 (Table 2.2).

It is noteworthy that each miRNA signature carried prognostic significance when applied to the HNSCC subtype where it was derived. On the other hand, when applied to other subtypes of HNSCC, none of the signatures were able to effectively distinguish high-risk and low-risk patients (Figure 2.6). We also observed this phenomenon when we applied the





**Figure 2.6.** Kaplan-Meier survival analysis to evaluate the miRNA prognostic signatures in other subtypes of HNSCC. (**A, B**) Survival analysis of the OPSCC miRNA signature in OSCC (**A**) and LSCC (**B**). (**C, D**) Survival analysis of the OSCC miRNA signature in OPSCC (**C**) and LSCC (**D**). (**E, F**) Survival analysis of the LSCC miRNA signature in OSCC (**E**) and LSCC (**F**).



**Figure 2.7.** Kaplan-Meier survival analysis to evaluate an existing OPSCC miRNA signature in OSCC (A) and LSCC (B).

previously developed 6-miRNA prognostic model to OSCC and LSCC (Figure 2.7). In conjunction with previous studies indicating significant genetic heterogeneity between subtypes of HNSCC (15), our results indicate that the miRNome is just as unique for each HNSCC subtype.

#### *Cervical cancer miRNA signatures maintain significance across tumor source sites*

The Cancer Genome Atlas also provided samples for cervical cancer. We were able to analyze 276 total cancers that were identified to have sufficient miRNA-Seq results and appropriate follow-up data (Table 2.5). Through similar analyses, we first aimed to validate a previously described survival signature for cervical cancer based on 2 miRNAs (12):

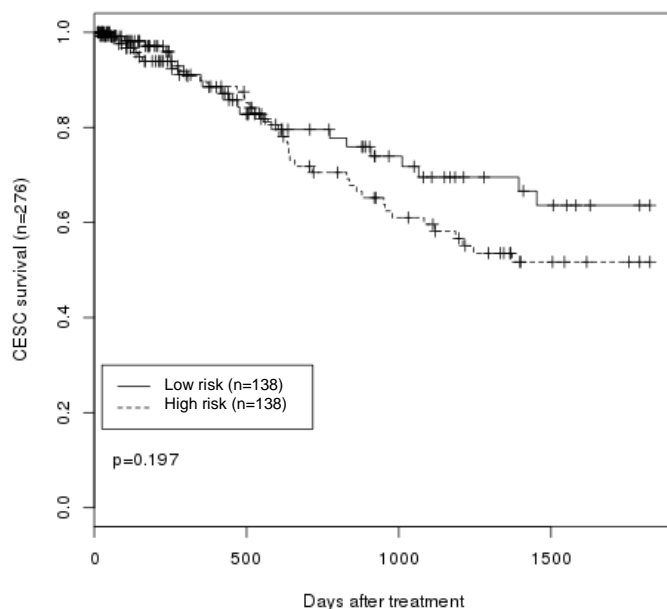
$$S = 17.9 - 0.284E_{\text{miR-9-5p}} - 0.376E_{\text{miR-200a-3p}},$$

where 17.9 was chosen as the zeroing coefficient. Given the change in quantification platforms from qRT-PCR to RNA-Seq, the signature was modified by removing the coefficient and using

**Table 2.5.** Characteristics of the CESC patients included in TCGA.

<b>CHARACTERISTICS</b>	<b>Total Cervical Cancer (n=276)</b>	<b>Squamous Cell Carcinoma (n=227)</b>	<b>Adenocarcinoma and Adenosquamous Cell Carcinoma (n=45)</b>
<b>Age at diagnosis <math>\pm</math> SD</b>	48.2 $\pm$ 13.9	48.7 $\pm$ 14.2	45.1 $\pm$ 12.2
<b>Median follow up time (days)</b>	469	471	306
<b>Race</b>			
White	197	158	35
Black or African American	28	26	2
Other	22	19	3
Unreported	29	24	5
<b>HPV (+)</b>	254	216	37
<b>Smoking<sup>a</sup></b>			
Nonsmoker	137	109	24
Long-term former smoker	8	7	1
Other former smoker	42	34	8
Current smoker	60	51	9
Unreported	29	26	3
<b>T classification</b>			
TX	17	15	2
Tis	1	1	0
T1	131	103	26
T2	63	51	11
T3	17	14	2
T4	10	9	1
Unreported	37	34	3
<b>N classification</b>			
NX	65	54	10
N0	122	95	24
N1	52	44	8
N2	0	0	0
N3	0	0	0
Unreported	37	34	3
<b>Stage</b>			
I	152	117	33
II	61	55	5
III	37	34	2
IV	20	15	5
Unreported	6	6	0
<b>Deceased in study</b>	63	53	10

<sup>a</sup>Smoking was defined as no history of smoking, a former smoker of  $\geq$  15 years, other former smoker of  $<$ 15 years, or a current smoker.



**Figure 2.8.** Kaplan-Meier survival analysis to analyze an existing 2-miRNA prognostic signature in cervical cancer.

the median scores across all TCGA samples as the cutoff value to distinguish high- and low-risk groups. The trend was accurate, but was unable to reach statistical significance (Figure 8).

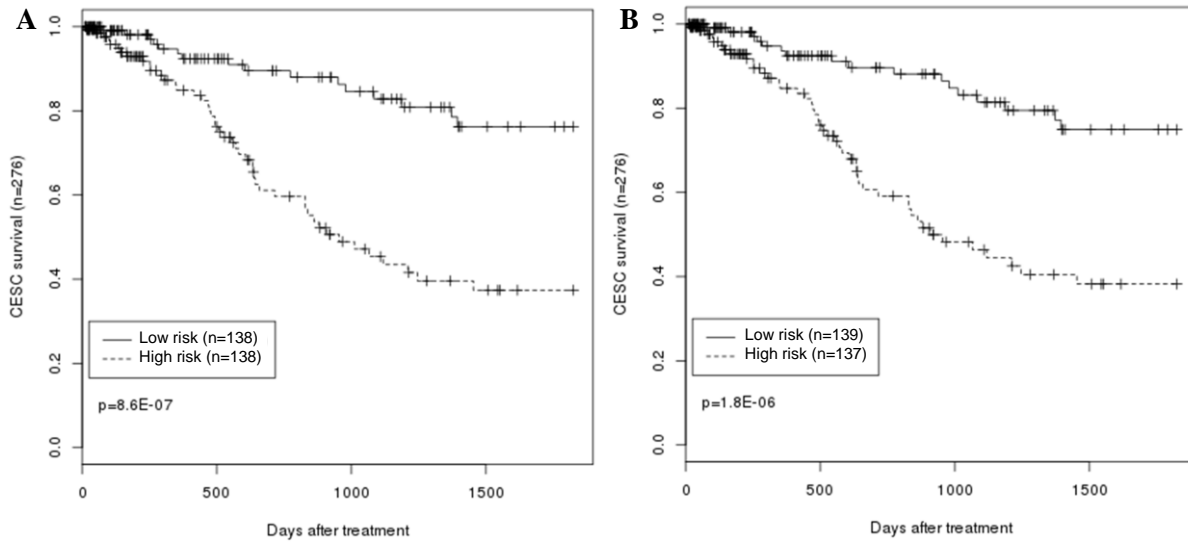
We then aimed to develop a novel miRNA-expression prognostic signature, using TCGA as the primary training set. The analysis identified four miRNAs as statistically significant for survival: miR-361-3p, miR-532-3p, and miR-150-5p are positively associated with survival, and miR-335-3p is negatively associated (Table 2.6). By using the z-scores obtained from univariate Cox survival analysis as coefficients, the resulting signature is as follows:

**Table 2.6.** Significantly dysregulated miRNAs associated with overall survival and used to develop prognostic models for CESC.

miRNA name	Fold change <sup>a</sup>	P-value <sup>b</sup>	z-score
hsa-miR-361-3p	-0.481	3.8E-07	-5.08
hsa-miR-532-5p	-0.514	8.7E-05	-3.92
hsa-miR-150-5p	-0.911	2.7E-03	-4.90
hsa-miR-335-3p	0.544	1.8E-02	3.00

<sup>a</sup> Fold change is log<sub>2</sub> normalized.

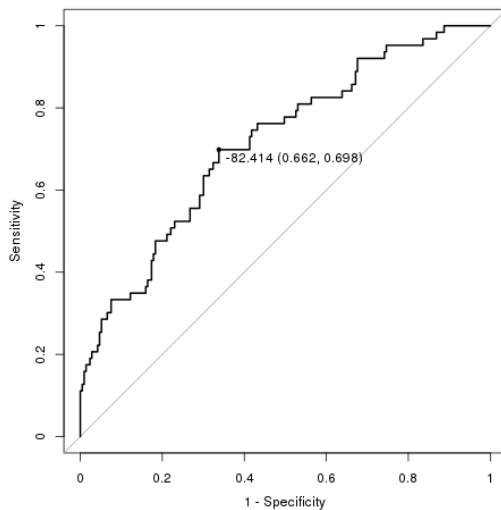
<sup>b</sup> P-values are from the logrank score from Cox univariate analysis.



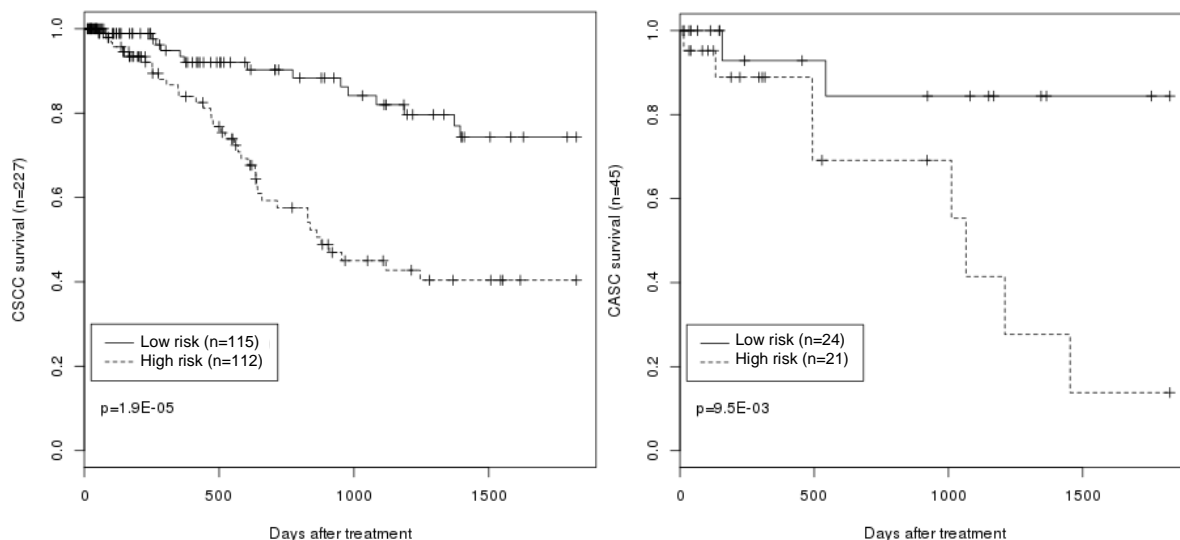
**Figure 2.9.** Kaplan-Meier analysis for a novel 4-miRNA prognostic signature in cervical cancer. The model was evaluated in **(A)** the training data from TCGA and **(B)** with 10-fold cross-validation. Significance is determined using the logrank p-value from Cox univariate survival analysis.

$$S = -5.08E_{miR-361-3p} - 3.92E_{miR-532-3p} - 4.90E_{miR-150-5p} + 3.00E_{miR-355-3p}$$

In the training data, this signature was confirmed to significantly separate high-risk and low-risk

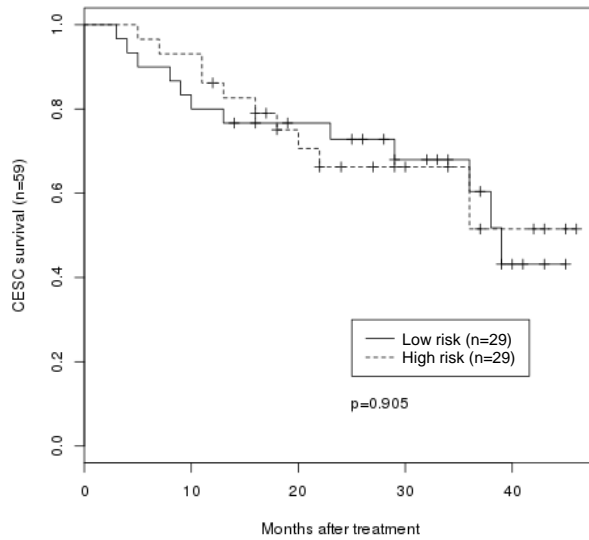


**Figure 2.10:** The receiver operating characteristic (ROC) curve for the novel CESC signature in the TCGA training cohort. The area under the curve was 0.716.



**Figure 2.11:** Kaplan-Meier analysis for the novel CESC signature in (A) cervical squamous cell carcinomas and (B) cervical adenosquamous carcinomas and adenocarcinomas. Significance is determined using the logrank p-value from Cox univariate survival analysis.

patients ( $p = 8.6E-07$ ) without overtraining, as shown in 10-fold cross-validation ( $1.8E-06$ ) (Figures 2.9A and B). ROC analysis also indicated model robustness in the context of sensitivity and specificity, with an area under the curve of 0.716 (Figure 2.10). More importantly, Kaplan-Meier analysis also showed that this signature was able to distinguish between high-risk and low-risk patients in both squamous cell carcinomas and adeno-type carcinomas. High-risk and low-risk cohorts in the squamous cell patient group were separated with a p-value of  $1.9E-05$ , and in the adeno-type carcinomas with a p-value of  $9.5E-03$  (Figure 2.11). However, this signature could not be validated in an independent cohort of 59 cervical cancer patients treated at Washington University, likely due to significant patient-to-patient variations from different cohorts or treatment regimens (Figure 2.12).



**Figure 2.12:** Kaplan-Meier analysis of the novel CESC signature in an independent cervical cancer cohort. Significance was determined using the logrank test from univariate Cox survival analysis.

## 2.5 Discussion

Identification of novel prognostic biomarkers typically requires a large dataset which provides sufficient statistical power for discovery research. To this end, we took advantage of the high-throughput data from TCGA for biomarker analysis. The TCGA consortium has published many studies identifying the mutations and dysregulations associated with tumors in comparison to matched normal tissue samples. There are also a number of studies that used TCGA data for independent validation of existing biomarkers (13,29,30). Additionally, many studies exploring miRNA biomarkers in head and neck cancer, including the miR-34 family and miR-200a species, have indicated their roles in oncogenesis (31). However, few studies have utilized TCGA data in systematically identifying biomarkers associated with patient outcome.

In this study, we have presented a new strategy to identify prognostic miRNA biomarkers by analyzing TCGA data directly, followed by experimental validation using an independent cohort. As the first step, we utilized TCGA data as the primary source to identify biomarkers and develop prognostic models for OPSCC. Within OPSCC, infection by HPV has already been

indicated as a favorable prognostic factor (7). Our model was able to further improve the prognostic value of HPV positivity by identifying a high-risk cohort among HPV(+) patients. Next, we were able to validate the robustness of this signature using an independent cohort that consisted only of HPV(+) OPSCC patients. This confirmed that the miRNA signature was able to further distinguish high- and low-risk patients within HPV(+) OPSCC patients.

Among the subtypes of HNSCC, OPSCC has unique characteristics as HPV infection is associated with most OPSCC cases. Although the total number of HNSCC cases has decreased steadily on a yearly basis, the number of reported OPSCC cases has increased significantly as a result of rapid rise in HPV(+) OPSCC cases (32,33). Our clinical goal of building a powerful prognostic model is to reliably stratify OPSCC patients for treatment failures after standard therapy. The availability of such a reliable prognostic model is critical for providing individualized cancer therapy, including both de-intensifying treatment for low-risk patients as well as intensification for high-risk patients. In particular, there is currently significant clinical interest in identifying a subset of OPSCC patients who have low-risk of treatment failures, in order to de-intensify their overall treatment. As present, multi-institutional de-escalation clinical trials are underway for HPV(+) OPSCC patients (34,35). However, there is still a significant portion of HPV(+) OPSCC cases that have poor outcome. For these cases, de-escalation treatment should not be applied and instead the treatment should be intensified. Thus, there is a critical need to develop robust prognostic models to further stratify HPV(+) OPSCC patients for enrollment in de-escalation trials. To this end, our proposed miRNA-based prognostic model will fill in a critical need by selecting HPV(+) OPSCC patients who will most likely benefit from de-escalation treatment. Further work would be required to bring this signature fully to the clinical



setting, such as the inclusion of reference genes to standardize the signature score and allow clinicians to determine the appropriate treatment modality.

Our analysis involving cervical cancer demonstrated that the methods for identifying miRNA biomarkers in TCGA data, or other large public –omics datasets, can be extended beyond head and neck cancers. The signature we described was robust within the training data from TCGA, as demonstrated through cross-validation and ROC analysis, but was unable to maintain significance in an independent dataset. Despite this drawback, it was notable that the signature was able to separate high- and low-risk patients across cervical cancer species. It has been noted in the literature that patients with adeno-type cervical carcinomas are genetically distinct from squamous type carcinomas, as well as conferring higher risk (16,36). Therefore, the possibility of using a single prognostic signature to stratify patients prior to determination of tumor source in cervical cancers merits further investigation. Additional work in this field would incorporate more individualized approaches, as genomic heterogeneity may also require unique miRNomic profiles for the differing cervical tissues.

Besides OPSCC, we have also shown that our strategy on TCGA-based biomarker discovery can be extended to the study of other subtypes of HNSCC, as well as CESC. In this way, we have demonstrated that TCGA represents a rich resource for cancer prognostic studies. We expect that prognostic tools developed using TCGA data, with proper validation, will significantly expand our ability to more precisely manage cancer patients by applying individualized treatment plans.

## 2.6 References

1. Wong, N., Khwaja, S.S., Baker, C.M., Gay, H.A., Thorstad, W.L., Daly, M.D., Lewis, J.S. and Wang, X. (2016) Prognostic microRNA signatures derived from The Cancer Genome Atlas for head and neck squamous cell carcinomas. *Cancer Medicine*, **5**, 1619-1628.

2. Siegel, R.L., Miller, K.D. and Jemal, A. (2015) Cancer Statistics, 2015. *CA Cancer J Clin*, **64**, 5-29.
3. Carvalho, A.L., Nishimoto, I.N., Califano, J.A. and Kowalski, L.P. (2005) Trends in incidence and prognosis for head and neck cancer in the United States: A site-specific analysis of the SEER database. *Int J Cancer*, **114**, 806-816.
4. Hu, A., Huang, J.-J., Xu, W.-H., Jin, X.-J., Li, J.-P., Tang, Y.-J., Huang, X.-F., Cui, H.-J., Sun, G.-B., Li, R.-L. *et al.* (2015) MiR-21/miR-375 ratio is an independent prognostic factor in patients with laryngeal squamous cell carcinoma. *Am J of Cancer Res*, **5**, 1775-1785.
5. Peng, S.-C., Liao, C.-T., Peng, C.-H., Cheng, A.-J., Chen, S.-J., Huang, C.-G., Hsieh, W.-P. and Yen, T.-C. (2014) MicroRNAs miR-218, miR-125b, and let-7g predict prognosis in patients with oral cavity squamous cell carcinoma. *PLoS ONE*, **9**, e102403.
6. Long, X.-B., Sun, G.-B., Hu, S., Liang, G.-T., Wang, N., Zhang, X.-H., Cao, P.-P., Zhen, H.-T., Cui, Y.-H. and Liu, Z. (2009) Let-7a microRNA functions as a potential tumor suppressor in human laryngeal cancer. *Oncol Rep*, **22**, 1189-1195.
7. D'Souza, G., Kreimer, A.R., Viscidi, R., Pawlita, M., Fakhry, C., Koch, W.M., Westera, W.H. and Gillison, M.L. (2007) Case-control study of human papillomavirus and oropharyngeal cancer. *N Engl J Med*, **356**, 1944-1956.
8. Hui, A.B., Lin, A., Xu, W., Waldron, L., Perez-Ordóñez, B., Weinreb, I., Shi, W., Bruce, J., Huang, S.H., O'Sullivan, B. *et al.* (2013) Potentially prognostic miRNAs in HPV-associated oropharyngeal carcinoma. *Clin Cancer Res*, **19**, 2154-2162.
9. Gao, G., Gay, H.A., Chernock, R.D., Zhang, T.R., Luo, J., Thorstad, W.L., Lewis, J., James S and Wang, X. (2013) A microRNA expression signature for the prognosis of oropharyngeal squamous cell carcinoma. *Cancer*, **119**, 72-80.
10. Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D. and Bray, F. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, **136**, E359-386.
11. Schiffman, M., Wentzensen, N., Wacholder, S., Kinney, W., Gage, J.C. and Castle, P.E. (2011) Human papillomavirus testing in the prevention of cervical cancer. *J Natl Cancer Inst*, **103**, 368-383.
12. Hu, X., Schwarz, J.K., Lewis, J., James S, Huettner, P.C., Rader, J.S., Deasy, J.O., Grigsby, P.W. and Wang, X. (2010) A microRNA expression signature for cervical cancer prognosis. *Cancer Res*, **70**, 1441-1448.
13. How, C., Pintilie, M., Bruce, J.P., Hui, A.B., Clarke, B.A., Wong, P., Yin, S., Yan, R., Waggott, D., Boutros, P.C. *et al.* (2015) Developing a prognostic micro-RNA signature for human cervical carcinoma. *PLoS ONE*, **10**, e0123946.

14. The Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.
15. The Cancer Genome Atlas, N. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**, 576-582.
16. Network, C.G.A.R., Medicine, A.E.C.o., Services, A.B., Hospital, B.C., Medicine, B.C.o., Hope, B.R.I.o.C.o., Aging, B.I.f.R.o., Centre, C.s.M.S.G.S., School, H.M., Services, H.F.G.C.C.R.I.a.C.C.H. *et al.* (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**, 378-384.
17. Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic acids research*, **42**, D764-770.
18. Coordinators, N.R. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **43**, D6-D17.
19. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, **42**, D68-D73.
20. Wilks, C., Cline, M.S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D. *et al.* (2014) The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database*, **2014**.
21. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
22. (2012). The Mathworks, Natick, MA.
23. Wang, X. (2009) A PCR-based platform for microRNA expression profiling studies. *RNA (New York, N.Y)*, **15**, 716-723.
24. Gao, G., Chernock, R.D., Gay, H.A., Thorstad, W.L., Zhang, T.R., Wang, H., Ma, X.J., Luo, Y., Lewis, J.S., Jr. and Wang, X. (2013) A novel RT-PCR method for quantification of human papillomavirus transcripts in archived tissues and its application in oropharyngeal cancer prognosis. *International journal of cancer*, **132**, 882-890.
25. Xu, J.-W., Wang, T.-X., You, L., Zheng, L.-F., Shu, H., Zhang, T.-P. and Zhao, Y.-P. (2014) Insulin-Like Growth Factor 1 Receptor (IGF-1R) as a Target of MiR-497 and Plasma IGF-1R Levels Associated with TNM Stage of Pancreatic Cancer. *PLoS ONE*, **9**, e92847.
26. Li, J., Kong, F., Wu, K., Song, K., He, J. and Sun, W. (2014) miR-193b directly targets STMN1 and uPA genes and suppresses tumor growth and metastasis in pancreatic cancer. *Mol Med Rep*, **10**, 2613-2620.

27. Zheng, G., Du, L., Yang, X., Zhang, X., Wang, L., Yang, Y., Li, J. and Wang, C. (2014) Serum microRNA panel as biomarkers for early diagnosis of colorectal adenocarcinoma. *Br J Cancer*, **111**, 1985-1992.
28. Chai, J., Wang, S., Han, D., Dong, W., Xie, C. and Guo, H. (2015) MicroRNA-455 inhibits proliferation and invasion of colorectal cancer by targeting RAF proto-oncogene serine/threonine-protein kinase. *Tumour Biol*, **36**, 1313-1321.
29. Miller, D.L., Davis, J.W., Taylor, K.H., Johnson, J., Shi, Z., Williams, R., Atasoy, U., Lewis, J.S., Jr. and Stack, M.S. (2015) Identification of a human papillomavirus-associated oncogenic miRNA panel in human oropharyngeal squamous cell carcinoma validated by bioinformatics analysis of the Cancer Genome Atlas. *The American journal of pathology*, **185**, 679-692.
30. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, **2**, 401-404.
31. Sethi, N., Wright, A., Wood, H. and Rabbitts, P. (2014) MicroRNAs and head and neck cancer: reviewing the first decade of research. *Eur J Cancer*, **50**, 2619-2635.
32. Ernster, J.A., Sciotto, C.G., O'Brien, M.M., Finch, J.L., Robinson, L.J., Willson, T. and Mathews, M. (2007) Rising incidence of oropharyngeal cancer and the role of oncogenic human papilloma virus. *The Laryngoscope*, **117**, 2115-2128.
33. Syrjanen, S. (2005) Human papillomavirus (HPV) in head and neck cancer. *J Clin Virol*, **32 Suppl 1**, S59-66.
34. Mirghani, H., Amen, F., Blanchard, P., Moreau, F., Guigay, J., Hartl, D.M. and Lacau St Guily, J. (2015) Treatment de-escalation in HPV-positive oropharyngeal carcinoma: ongoing trials, critical issues and perspectives. *International journal of cancer*, **136**, 1494-1503.
35. Masterson, L., Moualed, D., Liu, Z.W., Howard, J.E., Dwivedi, R.C., Tysome, J.R., Benson, R., Sterling, J.C., Sudhoff, H., Jani, P. *et al.* (2014) De-escalation treatment protocols for human papillomavirus-associated oropharyngeal squamous cell carcinoma: a systematic review and meta-analysis of current clinical trials. *Eur J Cancer*, **50**, 2636-2648.
36. Yokoi, E., Mabuchi, S., Takahashi, R., Matsumoto, Y., Kuroda, H., Kozasa, K. and Kimura, T. (2017) Impact of histological subtype on survival in patients with locally advanced cervical cancer that were treated with definitive radiotherapy: adenocarcinoma/adenosquamous carcinoma versus squamous cell carcinoma. *J Gynecol Oncol*, **28**, e19.

# **Chapter 3: Development of an Online Resource for Exploring Pan-Cancer MicroRNA Dysregulation**

*This chapter is adapted from and expanded upon the following publication (1):*

Wong, Nathan, Yuhao Chen, Shuai Chen and Xiaowei Wang. (2017). *OncomiR*: An online resource for pan-cancer microRNA dysregulation. *Bioinformatics*, btx627. DOI: 10.1093/bioinformatics/btx627.

## **3.1 Abstract**

Dysregulation of microRNAs (miRNAs) is extensively associated with cancer development and progression. miRNAs have been shown to be biomarkers for predicting tumor formation and outcome. However, identification of the relationships between miRNA expression and tumor characteristics can be difficult and time-consuming without appropriate bioinformatics expertise. To address this issue, we present *OncomiR* (<http://oncomir.org>), an online resource for exploring miRNA dysregulation in cancer. Using combined miRNA-Seq, RNA-Seq, and clinical data from The Cancer Genome Atlas, we systematically performed statistical analyses to identify dysregulated miRNAs that are associated with tumor development and progression in most major cancer types. Additional analyses further identified potential miRNA-gene target interactions in tumors. These results are stored in a backend database and are presented through a web server interface. Moreover, through a backend bioinformatics pipeline, *OncomiR* can perform dynamic analysis with custom input data for in-depth characterization of miRNAs in cancer.

## 3.2 Introduction

MicroRNAs (miRNAs) are short, single-stranded RNA sequences of approximately 22 nucleotides that function in post-transcriptional regulation of gene expression. By targeting RNA transcripts for degradation or inhibition of translation, miRNAs are actively involved in controlling downstream proteomic profiles. This phenomenon is observed in numerous physiological or disease processes, such as embryonic development, tissue differentiation, immune response, and tumor progression (2). Furthermore, miRNAs can serve as biomarkers for various diseases, with particular clinical interest in predicting likelihood of cancer development and progression.

miRNA biomarkers have been discovered in nearly all cancer types. For example, in breast cancer, miR-21-5p and miR-155-5p, among others, have been reported as upregulated as compared to normal tissue, while miR-34a-5p and miR-145-5p are downregulated (3). However, it has also been noted that miR-221-3p and miR-222-3p are both downregulated in erythroblastic leukemia but upregulated in thyroid carcinoma and hepatocellular carcinoma, which suggests that the mechanisms driving tumor formation and progression are not uniform across all cancer types (3). The majority of studies that identify miRNA biomarkers focus on a single or a subset of miRNAs; however, with the growth of affordable high-throughput sequencing technologies, it is possible to analyze the complete miRNomes from many patients across multiple cancer types.

Analysis of high-throughput data is difficult for researchers with little computational expertise. To address this issue, a number of databases have been previously established to characterize miRNA functions in cancer. For example, miRCancer, miR2Disease, and OncomiRDB present experimentally validated relationships between miRNA expression and cancer development, based on literature reports (4-6). Other resources, including starBase,

cBioPortal and FireBrowse, among others, provide results of statistical analyses on high-throughput sequencing studies of cancer genomics as a whole, but not specifically focused on miRNA analysis (7-11).

Incorporating all the facets of miRNA biology into a comprehensive user-friendly toolset is a daunting task, as it requires identifying potential miRNA biomarkers and their targets as well as establishing their functional relationships in the context of cancer biology. To address this need, we present OncomiR, an online pan-cancer resource for analysis of miRNA dysregulation. OncomiR contains three major features: 1) A database of statistically dysregulated miRNAs associated with clinical characteristics of cancer; 2) miRNA-target expression correlation and prediction across cancer types; and 3) tools for dynamic analysis of miRNA-derived survival signatures and clustering of cancer types. The diverse functionality of OncomiR would make it a valuable resource to the miRNA and cancer research community.

### **3.3 Materials and Methods**

#### *Data Retrieval*

Anonymized patient clinical data, normalized mature miRNA-Seq read counts, and normalized RNA-Seq read counts were obtained from The Cancer Genome Atlas data portal ([tcga-data.nci.nih.gov](http://tcga-data.nci.nih.gov), [gdc-portal.nci.nih.gov](http://gdc-portal.nci.nih.gov)). Patients were excluded from subsequent analysis if the clinical data indicated no follow up time, the tissue specimens were not obtained from primary tumors, or the sample lacked either miRNA-Seq or RNA-Seq data. In total, 9,498 patients were analyzed across 30 cancer types (Table 3.1). miRNA-Seq read counts less than 1 read per million reads mapped (RPM) were fixed to a floor value of 1 RPM and log<sub>2</sub> transformed. RNA-Seq read

counts less than 5 reads per kilobase per million reads mapped (RPKM) were fixed to a floor value of 5 RPKM prior to log<sub>2</sub> transformation.

Data Analysis

Patient survival time is defined as the days between the start of treatment and the most recent follow up appointment or patient death. Follow up time was truncated to five years in order to determine the five-year survival status.

**Table 3.1.** Summary of cancer types and patient counts from The Cancer Genome Atlas.

<b>Cancer Type (abbreviation)</b>	<b>Total patients</b>	<b>Patients analyzed (patients excluded)</b>
Adrenocortical carcinoma (ACC)	80	80 (0)
Bladder urothelial carcinoma (BLCA)	412	407 (5)
Brain lower grade glioma (LGG)	515	508 (3)
Breast invasive carcinoma (BRCA)	1098	1065 (33)
Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)	308	289 (19)
Cholangiocarcinoma (CHOL)	36	36 (0)
Colon adenocarcinoma (COAD)	459	424 (35)
Esophageal carcinoma (ESCA)	185	184 (1)
Head and neck squamous cell carcinoma (HNSC)	528	522 (6)
Kidney chromophobe (KICH)	66	65 (1)
Kidney renal clear cell carcinoma (KIRC)	536	514 (22)
Kidney renal papillary cell carcinoma (KIRP)	291	288 (3)
Liver hepatocellular carcinoma (LIHC)	377	366 (11)
Lung adenocarcinoma (LUAD)	521	500 (21)
Lung squamous cell carcinoma (LUSC)	504	472 (32)
Mesothelioma (MESO)	87	86 (1)
Ovarian serous cystadenocarcinoma (OV)	585	484 (101)
Pancreatic adenocarcinoma (PAAD)	185	177 (8)
Pheochromocytoma and paraganglioma (PCPG)	179	179 (0)
Prostate adenocarcinoma (PRAD)	498	494 (4)
Rectal adenocarcinoma (READ)	171	155 (16)
Sarcoma (SARC)	261	259 (0)
Skin cutaneous melanoma (SKCM)	570	96 (474)
Stomach adenocarcinoma (STAD)	443	413 (13)
Testicular germ cell tumor (TGCT)	150	134 (16)
Thyroid carcinoma (THCA)	507	505 (2)
Thymoma (THYM)	124	123 (1)
Uterine corpus endometrial carcinoma (UCEC)	548	536 (12)
Uterine carcinosarcoma (UCS)	57	56 (1)
Uveal melanoma (UVM)	80	80 (0)
<b>Total</b>	<b>10,361</b>	<b>9,497 (864)</b>



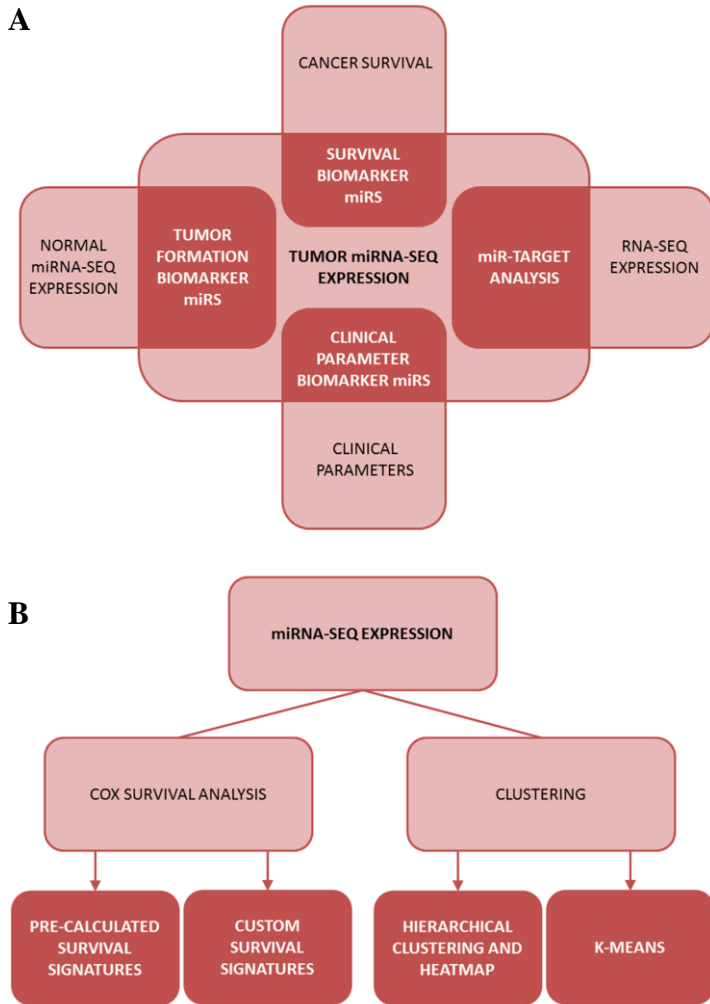
Paired tumor and non-tumor samples were obtained from 670 patients in the data set. The relationship between tumor formation and miRNA expression within these cancer types were evaluated using paired Student's t-test. For each clinical feature, analysis of variation (ANOVA) was used to determine the association between miRNA expression and relevant feature values. Survival analysis was performed using Cox proportional hazards analysis. Univariate analysis was conducted to determine the influence of the expression of a single miRNA on survival time; multivariate analysis was implemented to determine if the effect of the miRNA was independent of clinical characteristics. Additionally, the unpaired Student's t-test was employed to evaluate the difference in the average miRNA expression between living and deceased patients.

The likelihoods of miRNA-gene target pairings were evaluated using Pearson's correlation analysis based on the expression profiles of both miRNAs and mRNAs, in conjunction with target prediction scores obtained from miRDB (12). For each individual cancer type, all available tumor samples were incorporated in the correlation analysis. Additionally, all paired tumor/normal tissue types were evaluated as a single set, so as to provide a miRNA-target interactome specific to tumor formation by comparing to normal tissues. All statistical analysis was conducted using the R statistical program ([www.r-project.org](http://www.r-project.org)).

## **3.4 Results**

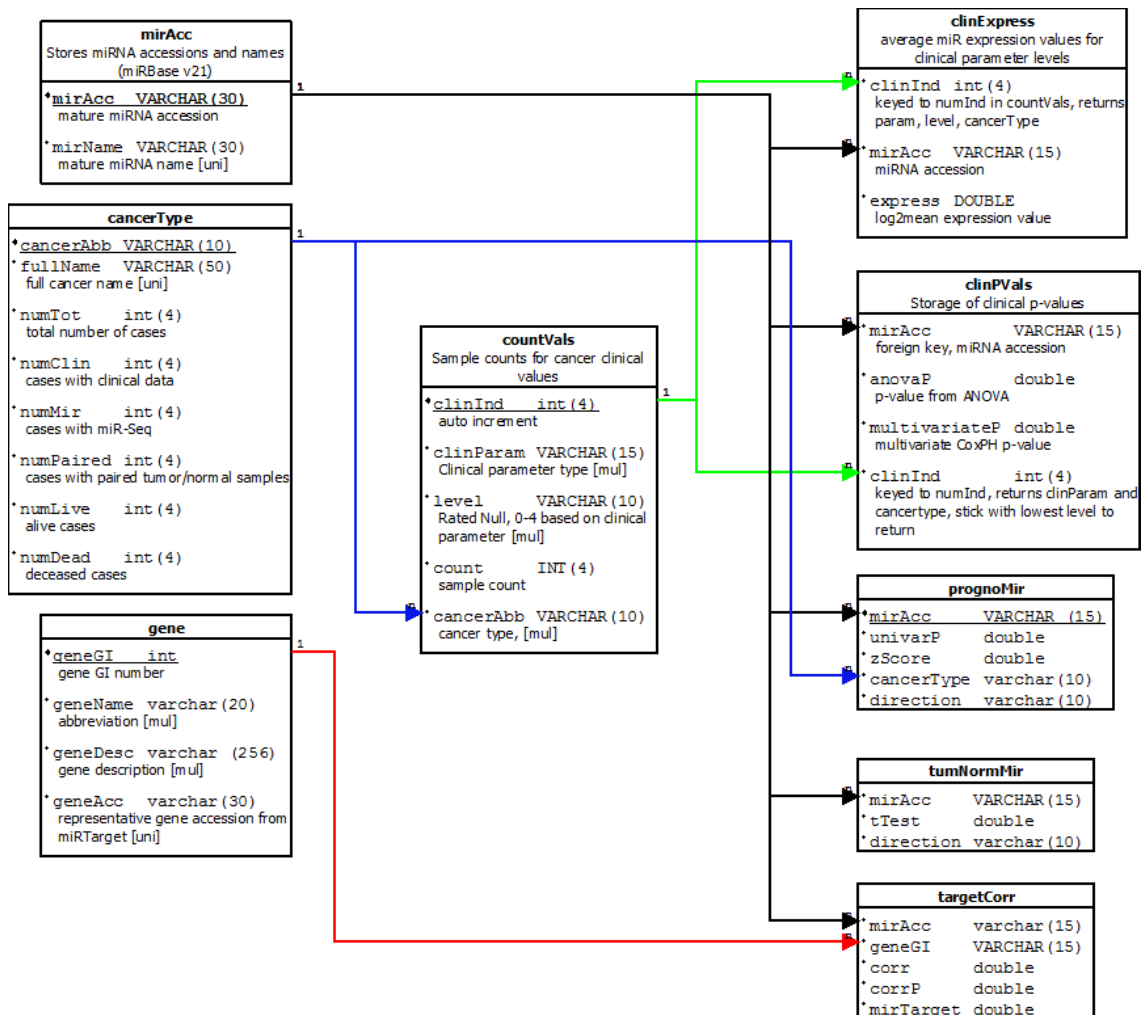
### *Database and Web Server Construction*

OncomiR consists of a primary backend database and a dynamic web server. Results from the statistical analyses are stored in a MySQL database accessible through Perl CGI and Perl DBI (Figure 3.1A). The OncomiR web server implements Perl CGI in conjunction with the R statistical program in order to conduct ad hoc backend analysis (Figure 3.1B).



**Figure 3.1.** Database and server design for OncomiR. **(A)** The OncomiR database contains the results of statistical analysis of miRNA expression in relation to patient characteristics for biomarker identification, and with mRNA expression to identify potential gene targets. **(B)** The OncomiR web server can perform *de novo* analysis for miRNA survival signatures and miRNA expression-based clustering for most cancer types.

The primary database design uses keys based on miRNA accession, cancer type, gene GI, and clinical diagnostic parameter (Figure 3.2). The resulting combinations can be used to search for: miRNAs associated with tumor formation; miRNAs dysregulated between cancer diagnostic stages; miRNAs correlated patient overall survival in individual cancer types; average miRNA expression levels in cancers; and miRNA-target interactions specific to cancer types.



**Figure 3.2.** Database schematic for OncomiR. Each box represents an individual table in the OncomiR database. Arrows indicate the keys used by MySQL database. Arrow sources show the home tables of the key, and the destinations show how the keys can be used in combination for rapid data retrieval.

*miRNA Dysregulation in Cancer Development and Progression*

The identification of novel RNA-derived molecular biomarkers in the clinical setting requires a combination of high-throughput data, such as next-generation sequencing, and robust statistical analysis. With this in mind, we have conducted such analyses using miRNA-Seq data from TCGA and evaluated the significance of miRNA expression in relation to clinical

parameters. The results of these studies have been included in OncomiR, encompassing 30 major cancer types and 1,171 distinct mature miRNA sequences.

OncomiR features a web query interface for the retrieval of miRNA associations to three primary clinical features: tumor development, tumor staging and grade, and overall patient

**A**

**B**

There are 33 clinical parameters in the selected cancer(s) associated with hsa-miR-92a-3p

miRNA Name	Cancer Abbreviation	Clinical Parameter	ANOVA P-value	ANOVA FDR	Multivariate Log Rank P-value	Multivariate Log Rank FDR
hsa-miR-92a-3p	ACC	Clinical M Status	6.23e-03	7.89e-02	2.78e-01	5.18e-01
hsa-miR-92a-3p	ACC	Pathologic N Status	9.84e-01	9.98e-01	3.03e-02	1.10e-01
hsa-miR-92a-3p	ACC	Sex	3.83e-01	8.00e-01	3.71e-02	1.21e-01
hsa-miR-92a-3p	BLCA	Pathologic N Status	1.69e-02	1.02e-01	1.85e-01	5.46e-01
hsa-miR-92a-3p	BRCA	Pathologic N Status	4.28e-06	1.29e-03	8.98e-01	9.98e-01
hsa-miR-92a-3p	BRCA	Pathologic Stage	2.55e-03	7.00e-02	9.06e-01	9.98e-01
hsa-miR-92a-3p	BRCA	Pathologic T Status	3.87e-04	5.48e-03	9.62e-01	9.98e-01
hsa-miR-92a-3p	COAD	Pathologic M Status	2.90e-04	3.95e-03	3.72e-01	9.97e-01
hsa-miR-92a-3p	HNSC	Clinical M Status	3.10e-02	1.52e-01	2.58e-01	5.81e-01
hsa-miR-92a-3p	HNSC	Histologic Grade	7.07e-03	2.52e-02	1.74e-01	4.68e-01
hsa-miR-92a-3p	HNSC	Sex	1.52e-02	2.25e-01	3.35e-01	6.63e-01
hsa-miR-92a-3p	KICH	Pathologic Stage	4.44e-02	9.49e-02	2.25e-01	6.34e-01

**Figure 3.3.** Search for miRNA biomarkers in the OncomiR database. A screenshot of a miRNA search in relation to clinical parameters (A) produces a table of results containing the miRNA and related cancer types, and p-values from relevant statistical tests (B).

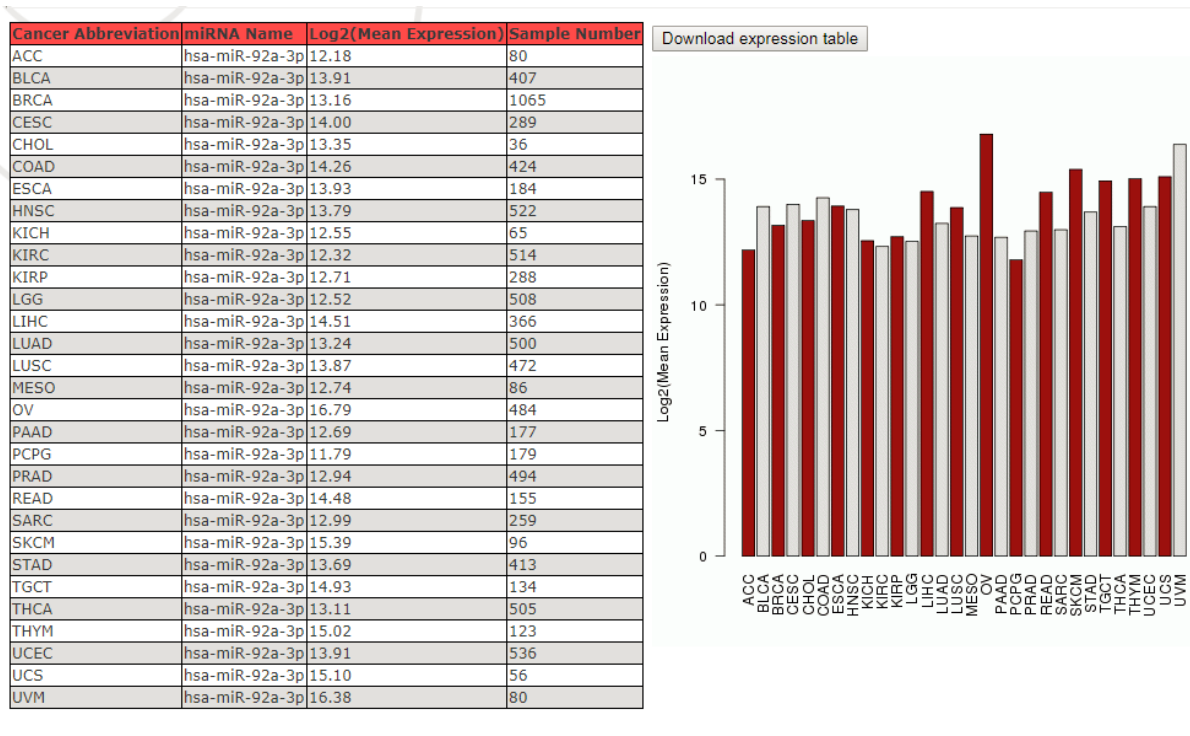
survival. For all these categories, users can search by miRNA name as assigned in miRBase Release 21 (13); users are also able to filter results by selecting one or more cancer types (Figure 3.3A). In this way, users can retrieve lists of miRNAs associated with specific clinical features in selected tumor types.

The results are presented in a tabular format, where each row shows the paired miRNA and specified cancer types and clinical features, with relevant statistics indicating the significance of the miRNA interaction. A screenshot of search results for diagnostic parameters associated with a specific miRNA is shown in Figure 3.3B: each row in the results shows the miRNA, cancer type, diagnostic criteria, and relevant p-values. Such results are similar across different search parameters. For example, searching for miRNAs associated with survival in different cancer types will return a list of miRNAs, with raw and adjusted p-values from unpaired Student's t-test comparing expression between living and deceased patients, as well as univariate Cox proportional hazards analysis, which includes survival time as a factor (Figure

There are 66 miRNAs significantly associated with survival in BRCA

miRNA Name	Cancer Abbreviation	Log Rank P-value	Log Rank FDR	Z-score	Upregulated in:	Deceased Log2 Mean Expression	Living Log2 Mean Expression	T-Test P-value	T-Test FDR
hsa-miR-874-3p	BRCA	3.21e-04	7.20e-01	3.586	Deceased	5.08	4.90	7.93e-02	6.95e-01
hsa-miR-1307-3p	BRCA	1.92e-03	7.20e-01	3.159	Deceased	10.22	10.05	1.24e-01	6.95e-01
hsa-miR-101-3p	BRCA	2.25e-03	9.19e-01	3.085	Living	13.29	13.42	5.01e-02	6.71e-01
hsa-miR-30a-5p	BRCA	2.80e-03	6.61e-01	3.036	Living	15.66	15.82	2.41e-01	6.95e-01
hsa-miR-30a-3p	BRCA	3.09e-03	9.19e-01	3.010	Living	13.77	13.93	1.99e-01	6.95e-01
hsa-miR-148b-5p	BRCA	3.61e-03	8.44e-01	2.983	Deceased	1.71	1.58	2.76e-01	6.95e-01
hsa-miR-181c-5p	BRCA	4.64e-03	6.54e-01	2.811	Deceased	5.90	5.76	6.94e-01	8.96e-01
hsa-miR-1224-5p	BRCA	6.26e-03	7.20e-01	2.968	Deceased	1.43	1.23	3.46e-01	7.12e-01
hsa-miR-2115-3p	BRCA	6.62e-03	9.27e-01	3.002	Deceased	1.33	1.00	2.69e-01	6.95e-01
hsa-miR-3907	BRCA	6.94e-03	7.20e-01	4.453	Deceased	0.02	0.00	3.20e-01	6.95e-01
hsa-miR-3074-3p	BRCA	7.91e-03	6.39e-01	1.000	Living	0.00	0.02	1.21e-03	4.05e-01
hsa-miR-3680-5p	BRCA	8.17e-03	8.44e-01	4.306	Deceased	0.01	0.00	4.13e-01	7.34e-01
hsa-miR-302a-5p	BRCA	8.49e-03	7.24e-01	4.305	Deceased	0.01	0.00	3.24e-01	7.00e-01
hsa-miR-3140-3p	BRCA	9.21e-03	9.27e-01	3.705	Deceased	0.02	0.01	7.49e-01	9.06e-01
hsa-miR-148b-3p	BRCA	9.44e-03	6.39e-01	2.587	Deceased	7.97	7.76	2.47e-02	6.71e-01

**Figure 3.4.** Search results for survival-associated miRNAs. Results indicate the relevant miRNAs in a given cancer type, their associated statistical significance with survival, and the cohort in which the miRNA is upregulated.



**Figure 3.5.** Search results for average miRNA expression levels. The results are presented in the form of log<sub>2</sub> mean expression in each cancer type, as well as a bar graph for a visual comparison of mean expression.

3.4).

OncomiR also offers the users to compare the mean expression values of miRNAs across two or more cancer types. When selecting a single miRNA, the results are presented both as a table of mean values and as a bar graph (Figure 3.5), while a search for multiple miRNAs will produce a table containing the search results.

### *miRNA-Target Prediction for Tumor Samples*

Since miRNAs function as post-transcriptional regulators of gene expression, dysregulation of miRNAs implies that downstream regulation of mRNA targets would also be affected. To address this within the context of cancer biology, OncomiR offers the ability to

search for significant expression correlations between miRNAs and potential gene targets in tumor samples. The likelihood of a miRNA targeting a specific transcript is dependent on multiple features, such as the miRNA seed sequence (nucleotides 2-8) and target site accessibility. OncomiR combines the results of expression correlation analysis between miRNAs and mRNAs with the results of the recently updated MirTarget algorithm (version 3) to identify likely targeting effects within specific cancer types (14). All target prediction data were retrieved from miRDB.org (12).

Users are able to search for potential interactions by querying either for miRNA or gene target and selecting one or more cancer types. The inclusion of cancer types as a search parameter is necessary, considering the genomic heterogeneity between different tissues. Each miRNA-target pair within the set is presented with the correlation coefficient and p-value from Pearson’s correlation analysis, along with MirTarget prediction score, as obtained from miRDB (Figure 3.6). The most likely miRNA-target pairs have both strongest negative correlation

NFAT5 has 143 potential miRNA-target pairs.

Cancer Type	miRNA Name	Gene	Gene Description	Correlation	Correlation P-value	Correlation FDR	miRDB Score
ESCA	hsa-miR-216b-5p	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.1905	1.25e-02	8.19e-01	50
ESCA	hsa-miR-217	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.1829	1.67e-02	8.19e-01	71
ESCA	hsa-miR-4310	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.1711	2.52e-02	8.19e-01	67
BRCA	hsa-miR-1299	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0906	8.40e-03	9.46e-01	71
BRCA	hsa-miR-3163	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0880	1.05e-02	9.46e-01	100
BRCA	hsa-miR-3185	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0876	1.09e-02	9.46e-01	91
BRCA	hsa-miR-4310	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0870	1.14e-02	9.46e-01	67
BRCA	hsa-miR-3646	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0863	1.21e-02	9.46e-01	100
BRCA	hsa-miR-548w	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0855	1.29e-02	9.46e-01	100
BRCA	hsa-miR-518c-5p	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0842	1.44e-02	9.46e-01	59
BRCA	hsa-miR-548a-3p	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0836	1.51e-02	9.46e-01	97
BRCA	hsa-miR-516b-5p	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0826	1.64e-02	9.46e-01	83
BRCA	hsa-miR-519d-3p	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0816	1.76e-02	9.46e-01	92
BRCA	hsa-miR-1323	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0814	1.79e-02	9.46e-01	54
BRCA	hsa-miR-3674	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0812	1.82e-02	9.46e-01	61
BRCA	hsa-miR-449c-3p	NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	-0.0812	1.82e-02	9.46e-01	65

**Figure 3.6.** OncomiR search results for miRNA target prediction. The paired miRNA-target interactions are evaluated in individual cancer types for directional correlation and prediction score, obtained from the MirTarget algorithm.

coefficients among the tumors and highest target prediction scores by MirTarget. The selection criteria can be further loosened to include more miRNA-target pairs that have lower Pearson's correlations or MirTarget scores.

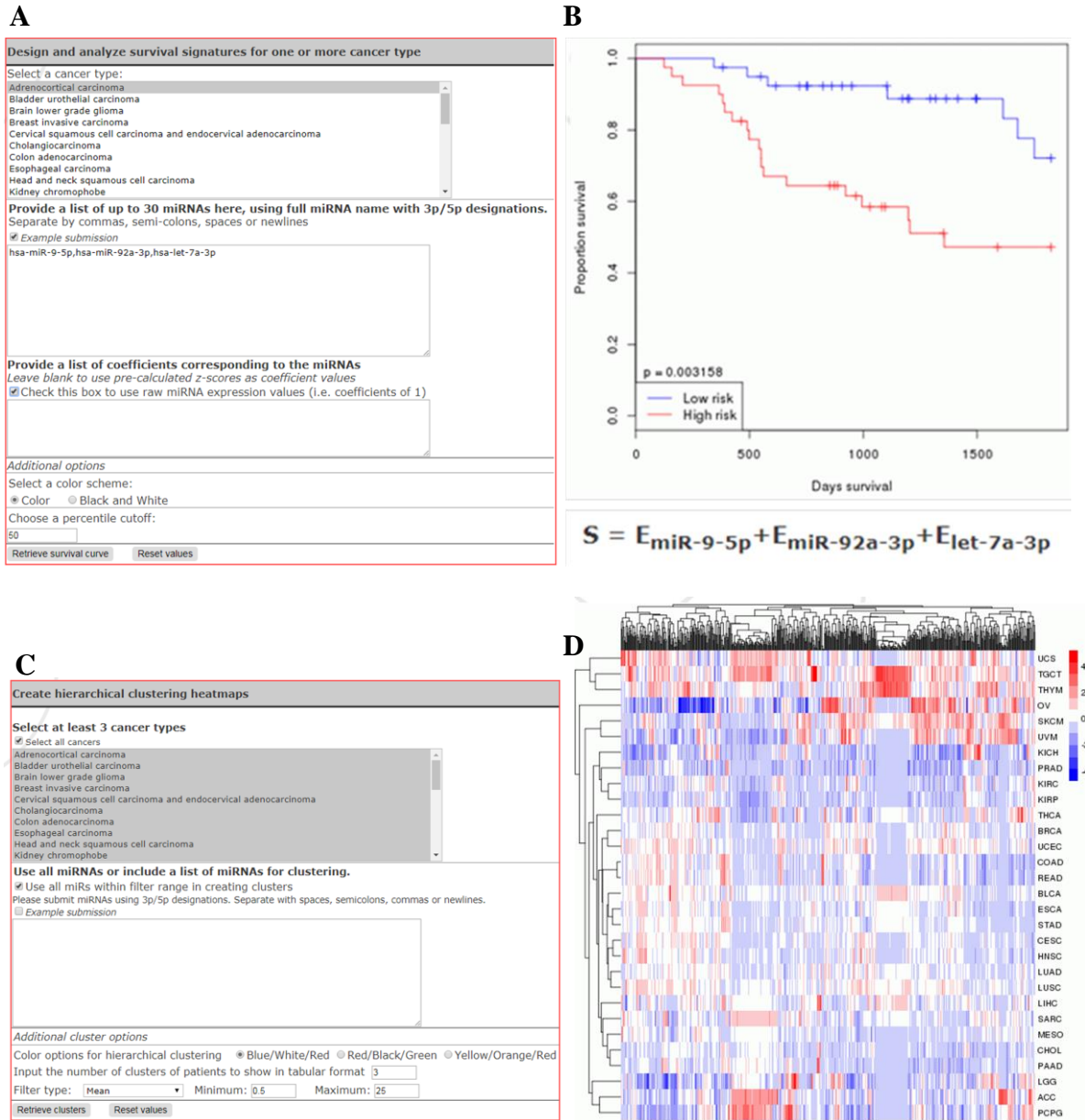
### *Server Interface for Custom miRNA Analysis*

One of the most significant clinical applications of cancer biomarker research is the stratification of patients for individualized therapy based on treatment outcome. To make our analysis more accessible to the clinical research community at large, OncomiR can analyze miRNA-derived survival outcome signatures dynamically for one or more cancer types. Users are able to input their selected miRNAs with pre-determined coefficients, as well as a percentile cutoff to determine the sizes of high- and low-risk cohorts (Figure 3.7A). The coefficient options also include using raw miRNA expression levels, or using z-scores resulting from a preliminary univariate Cox proportional hazards analysis. The results are presented in the form of a Kaplan-Meier survival curve and the logrank p-value to indicate the significance of cohort separation (Figure 3.7B). This feature is particularly useful for the evaluation of new biomarker signatures discovered with TCGA data, or validation of existing signatures derived from other independent studies.

Another feature of OncomiR is the ability to dynamically cluster major cancer types by miRNA expression. To this end, OncomiR offers dynamic clustering, whereby users can evaluate the suitability of using miRNA subsets to distinguish different cancer types. The mean expression of each miRNA was calculated within the individual cancer types; this average miRNomic profile can then be used as the basis of k-means or hierarchical clustering. (Figure 3.7C). Briefly, k-means clustering requires a predetermined number of groups, or clusters, into



which patients are sorted on the basis of expression similarity; hierarchical clustering initially treats each sample as its own individual cluster and builds a dendrogram by connecting clusters that are most closely related (15). Both clustering options return a list of cancer types that cluster



**Figure 3.7.** *De novo* analysis in OncomiR for survival signature and tumor clustering. (A) Survival analysis is conducted by selecting one or more cancer types, a list of miRNAs, and a list of coefficients. (B) The results of survival analysis are shown as a Kaplan-Meier curve. (C) Clustering of cancer types can be conducted using all miRNAs or a user-defined subset of miRNAs. (D) An example of hierarchical clustering is shown as a heat map.

together, and hierarchical clustering also produces a heatmap to visualize the similarities and differences between clusters (Figure 3.7D).

### **3.5 Discussion**

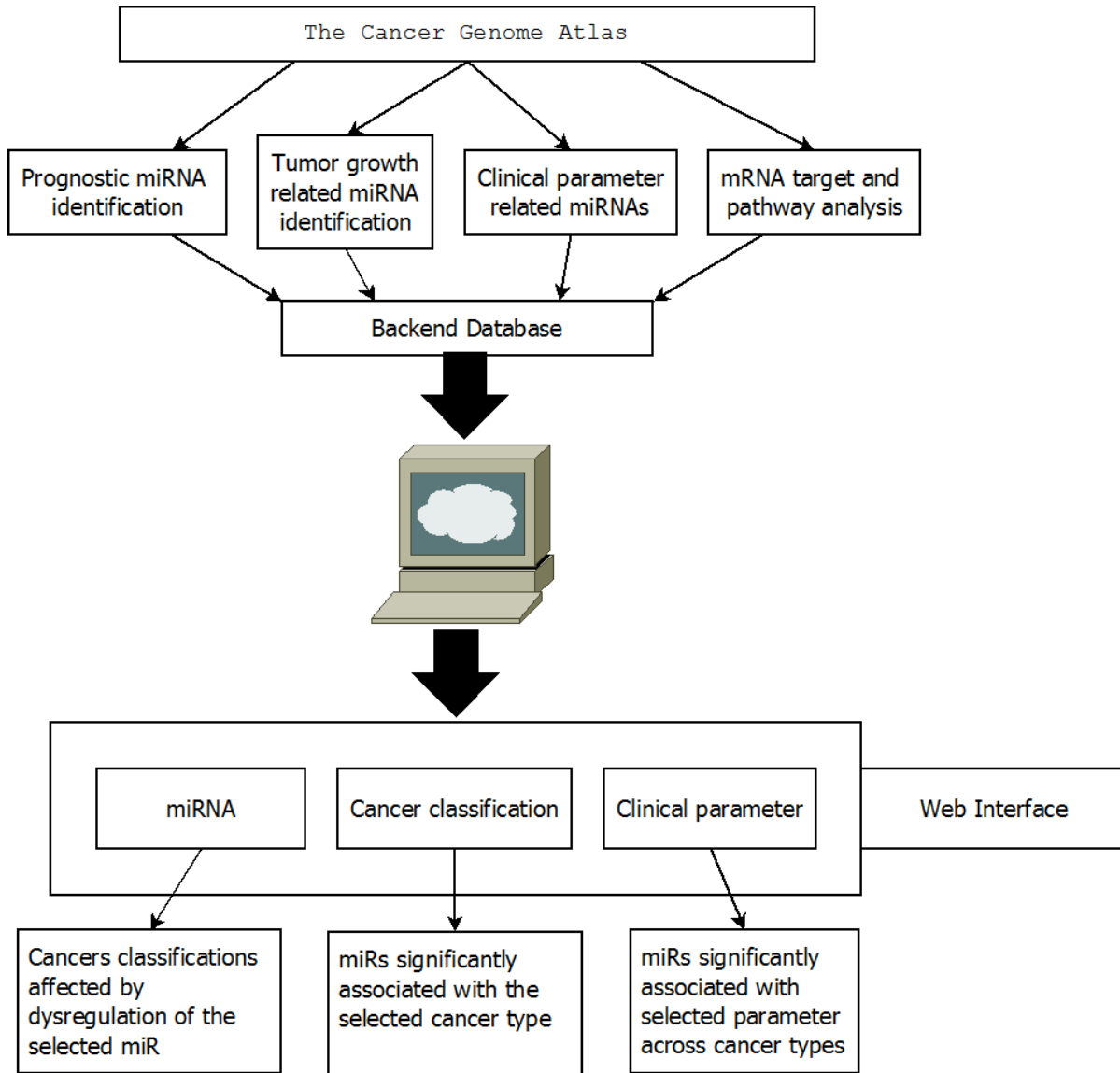
The identification of novel molecular biomarkers often requires comprehensive high-throughput datasets that are sufficiently large to minimize potential noise from patient-to-patient variations while also being thoroughly inclusive of less well-studied genes. Through TCGA, we obtained high-throughput miRNA-Seq and RNA-Seq data across 30 cancer types, with corresponding clinical profiles from thousands of patients. To systematically analyze miRNA-related TCGA data, we established a comprehensive bioinformatics pipeline to evaluate miRNA expression changes in relation to various clinical parameters such as tumor staging and patient survival status. In this way, we identified many dysregulated miRNAs involved in tumor formation, progression, and survival, and we have presented these results in OncomiR, a web accessible database.

By focusing primarily on miRNAs and miRNA-mediated biological functions, OncomiR can provide a greater insight into the miRNomic effects on tumor biology. Navigating OncomiR for miRNA biomarkers is designed to be both intuitive and informative. Users are able to select a preliminary search criterion, and navigate through a single search to find the desired results. In addition, flexible options are provided for more advanced analyses. In combination, these analyses can identify dysregulated miRNAs and targets in specific tumor types, and subsequently suggest potential pathways involved in the observed clinical phenomena, such as metastatic staging or overall survival. Multiple well-established miRNAs have been rediscovered through our analyses as being involved in cancer, consistent with previously reported studies. For

example, miR-92a-3p functions as an oncogenic miRNA, i.e. is overexpressed in tumor tissue as compared to normal tissue (16). More importantly, many new miRNA/cancer associations have been identified through our systematic analysis, especially in the context of specific cancer types. These new data provide useful clues for further characterization of miRNA functions in various types of cancer.

By incorporating the R statistical program, OncomiR performs clustering analysis of most known cancer types based on miRNA expression profiles. Subsets of miRNAs may be able to provide insight into similarities between cancer types. For example, our miRNA clustering analysis reveals that two subtypes of lung cancer, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) cluster together as expected, while a less intuitive similarity is observed as prostate adenocarcinoma (PRAD) clusters with three subtypes of kidney cancer (chromophobe, KICH; renal cell carcinoma, KIRC; and renal papillary carcinoma, KIRP) (Figure 3.7D). By evaluating the similarities as well as differences in miRNA expression in various cancer types, a greater understanding of how these cancers, and by extension, their original tissue sources, could in turn lead to improved clinical interpretations and subsequent interventions.

Currently, the strength of OncomiR lies in its ability to identify significant miRNAs based on clinical parameters shared across multiple cancer types, such as diagnostic staging and patient survival (Figure 3.8). Future updates of the database would benefit greatly from identifying biomarkers associated with cancer-specific traits. One example is the association of miRNAs with oncogenic viral infection, such as human papillomavirus in cervical and oropharyngeal cancers or hepatitis infection leading to liver cancer (17-19). Additional work may also include potential pathways mediated by miRNA dysregulation. Tools for analyzing



**Figure 3.8.** Overview of OncomiR’s functionality. The data was obtained from TCGA and analyzed before being stored in a backend database. The database is accessible through a web interface that allows users to search for and identify miRNAs associated with cancer classification and clinical diagnostic parameters.

comprehensive gene sets, as compared to individual genes, include PANTHER and Gene Set Enrichment Analysis (20,21). Such studies would be conducted on the gene transcripts regulated directly by the dysregulated miRNAs, as shown in the current target analysis results.

In summary, OncomiR is a user-friendly web resource for exploring miRNA dysregulation in cancer. We have conducted statistical analyses on miRNomes from TCGA to provide a readily accessible repository of miRNA associations with cancer characteristics. Additionally, correlation and target analysis were conducted to provide insights into possible miRNA-mediated mechanisms leading to cancer development and progression. Moreover, OncomiR also provides a set of dynamic tools for researchers to conduct custom miRNA analyses. Thus, OncomiR is a comprehensive tool that allows and encourages flexible miRNomic analysis across many cancer types.

### 3.6 References

1. Wong, N.W., Chen, Y., Chen, S. and Wang, X. (2017) OncomiR: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics*, btx627-btx627.
2. Ambros, V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350-355.
3. Croce, C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet*, **10**, 704-714.
4. Xie, B., Ding, Q., Han, H. and Wu, D. (2013) miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics*, **29**, 638-644.
5. Wang, D., Gu, J., Wang, T. and Ding, Z. (2014) OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics*, **30**, 2237-2238.
6. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G. and Liu, Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res*, **37**, D98-104.
7. firebrowse.org. (2016). Broad Institute of MIT and Harvard.
8. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, **2**, 401-404.

9. Lee, H., Palm, J., Grimes, S.M. and Ji, H.P. (2015) The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical-genomic driver associations. *Genome Med*, **7**, 112.
10. Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H. and Teschendorff, A.E. (2017) dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res*, **45**, D812-D818.
11. Li, J.H., Liu, S., Zhou, H., Qu, L.H. and Yang, J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*, **42**, D92-97.
12. Wong, N. and Wang, X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*, **43**, D146-D152.
13. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, **42**, D68-D73.
14. Wang, X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, **32**, 1316-1322.
15. Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data clustering: a review. *ACM computing surveys (CSUR)*, **31**, 264-323.
16. Zheng, G., Du, L., Yang, X., Zhang, X., Wang, L., Yang, Y., Li, J. and Wang, C. (2014) Serum microRNA panel as biomarkers for early diagnosis of colorectal adenocarcinoma. *Br J Cancer*, **111**, 1985-1992.
17. Wang, X., Wang, H.K., Li, Y., Hafner, M., Banerjee, N.S., Tang, S., Briskin, D., Meyers, C., Chow, L.T., Xie, X. *et al.* (2014) microRNAs are biomarkers of oncogenic human papillomavirus infections. *Proc Natl Acad Sci U S A*, **111**, 4262-4267.
18. Beasley, R.P., Hwang, L.Y., Lin, C.C. and Chien, C.S. (1981) Hepatocellular carcinoma and hepatitis B virus. A prospective study of 22 707 men in Taiwan. *Lancet*, **2**, 1129-1133.
19. Saito, I., Miyamura, T., Ohbayashi, A., Harada, H., Katayama, T., Kikuchi, S., Watanabe, Y., Koi, S., Onji, M. and Ohta, Y. (1990) Hepatitis C virus infection is associated with the development of hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, **87**, 6547-6549.
20. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D. and Thomas, P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, **45**, D183-D189.
21. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set

enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

# **Chapter 4: Pathway Analysis Identifies MicroRNA-Mediated Mechanisms of HPV-Induced Oncogenesis and Tumor Survival**

## **4.1 Abstract**

Human papillomavirus (HPV) is the primary cause of 95% of new cervical cancer diagnoses and 75% of new oropharyngeal cancer diagnoses. Despite its role in causing tumor formation, HPV is also established as a positive prognostic marker for tumor survival. Through infection HPV can induce expression changes in the host transcriptome, as well as regulatory elements such as microRNAs (miRNAs). The responsiveness of regulatory elements, such as those controlled by miRNAs, may explain the tumorigenic/pro-survival dichotomy of HPV infection. To determine how these effects are induced from both the host miRNome and host transcriptome, we have identified HPV infection status in 301 cervical cancers and 79 oropharyngeal cancers obtained from The Cancer Genome Atlas. Based on HPV status, we performed comprehensive statistical analysis to identify dysregulated miRNAs and gene transcripts. Potential gene targets were subjected to pathway analysis using the PANTHER database, so as to observe the cumulative effects in the context of Gene Ontology-defined biological processes. Pathway analysis revealed that significantly upregulated genes in both HPV(+) and HPV(-) tumors favored cellular reproduction and growth processes; HPV(-) tumors indicated underrepresentation of cellular adhesion pathways, which hint at possible mechanisms for tumor migration and poorer survival. More focused miRNA-target analysis showed



upregulation of both tumor suppressors and oncogenes, indicating that miRNA mediation may partly explain the dichotomy of HPV infection inducing tumor formation yet encouraging overall patient survival.

## 4.2 Introduction

With his Nobel Prize winning hypothesis that human papillomavirus is a correlative, and ultimately causative, factor in cervical cancer, Harald zur Hausen demonstrated that cancer may not only be treated after diagnosis, but that the potential existed that if the cause is known, cancer as a disease may also be preventable (1,2). In the years since, research has shown that the primary mechanisms of HPV-induced oncogenesis are based in the activities of the E6 and E7 viral proteins. E6 induces the ubiquitination of the regulatory protein p53, which is also known to function as a tumor suppressor (3). At the same time, E7 binds to and inactivates the tumor suppressor pRB, which in turn leads to the release of the transcription factor E2F1, which controls for a number of transcription factors associated with cell growth and proliferation (4).

HPV has also been shown to induce alteration in the expression profiles of microRNAs (miRNAs), short non-coding RNA transcripts of approximately 22 bases. miRNAs function as the guide sequence of the RNA-induced silencing complex (RISC), a post-transcriptional regulatory body that targets and degrades coding sequences, thereby preventing the translation of the protein. By controlling regulatory complexes such as RISC, in addition to altering cellular response through viral proteins, HPV may be able to prolong its own replicative cycle, while also making more subtle changes that can assist in the development of tumor growth. HPV does not encode any of its own miRNAs, but has been shown to induce changes in host miRNA expression level. Examples include miR-9-5p, which has been reported as upregulated in

HPV(+) cervical and oropharyngeal cancers, miR-145-5p, which has been shown as downregulated by HPV33, and miR-23b, which has been shown to be downregulated by HPV16 (5-8). The subsequent cellular responses can include increased motility, as associated with miR-9-5p targets, genome amplification through the suppression of miR-145-5p, and increased expression of the oncogene cMet, through miR-23b downregulation (5,7,8).

Interestingly, HPV has also been shown to be a powerful prognostic marker for improved cancer survival. In this context, a number of studies have also examined miRNA-mediated responses, and the potential mechanisms by which miRNAs can function as biomarkers for survival. The former has been demonstrated in the development of miRNA-profile derived survival signatures in cervical and oropharyngeal cancers, which may be used in the clinical setting to aid in determining course of treatment (9-13). Due to the varying survival rates based on HPV infection status, a common focus of these studies, especially in oropharyngeal cancers, is to demonstrate that these miRNA-based signatures can maintain significance in both HPV(+) and HPV(-) cohorts. However, the mechanisms by which HPV is capable of both inducing tumor formation as well as improving patient survival after diagnosis are unclear.

With this in mind, we have aimed to identify distinct miRNA-mediated pathways that indicate either HPV-induced tumor growth or HPV-related tumor survival. Using 79 oropharyngeal cancer samples and 301 cervical cancer samples obtained from The Cancer Genome Atlas, we have identified HPV-dysregulated microRNAs and coding transcripts. Further, we have identified potential miRNA targets, and by implementing pathway analysis, we have also identified a number of distinct mechanisms as potentially controlled through the miRNA regulatory mechanism. In summary, we have identified miRNA-controlled mechanisms

that supplement HPV-induced oncogenesis, as well as revealed potential genomic and miRNomic responses that may result in improved tumor survival.

## 4.3 Materials and Methods

### *Data Retrieval*

A total of 81 oropharyngeal cancer patients and 303 cervical cancer patients were identified in The Cancer Genome Atlas (Table 4.1). Through the Genomic Data Commons Data Portal ([portal.gdc.cancer.gov](http://portal.gdc.cancer.gov)), raw RNA-seq and miRNA-seq data were obtained for 79 patients in the oropharyngeal cancer group, and 301 cervical cancer patients (14). All gene sequences were downloaded from the UCSC Genome Browser (15). Index files mapping transcript accessions to NCBI Gene IDs were downloaded from the NCBI ftp site (16). Complete HPV genomes were downloaded from the Papillomavirus Episteme (17). All mature miRNA sequences were downloaded from miRBase (18).

### *TCGA Sequence Analysis*

Sequence alignment was performed using the Bowtie program (19). Raw miRNA-Seq reads were aligned to the human miRNome. The read counts were then normalized to reads per million reads mapped per sample and set to a floor value of 1 for lowly expressed miRNAs before being log<sub>2</sub> normalized. Raw RNA-seq reads were aligned sequentially to human RefSeq annotated sequences, the human reference genome, and the virome. The read counts were normalized to reads per kilobase per million mapped reads (RPKM), then to the 200<sup>th</sup> gene before being set to a floor of 5 normalized reads for lowly expressed transcripts.

Statistical Analysis for miRNA Correlation to HPV Status

miRNA and coding transcript expression levels were examined for significance in

**Table 4.1:** Patient characteristics of the HPV cancer cohorts

	All Cervical Cancer	Cervical Squamous Cell Carcinoma	Cervical Adeno-type Carcinoma	Oropharyngeal Squamous Cell Carcinoma
<b>Total patient count</b>	307	254	50	81
<b>Patients included</b>	301	252	49	79
<b>Age at diagnosis <math>\pm</math> SD</b>	48.2 $\pm$ 13.9	48.8 $\pm$ 14.1	45.3 $\pm$ 12.3	55.9 $\pm$ 9.3
<b>Median follow up time (days)</b>	350	471	241	637
<b>Sex</b>				
Female	301	252	49	11
Male	0	0	0	68
<b>Race</b>				
White	206	169	37	73
Black or African American	30	28	2	6
Other	29	24	5	0
Unreported	36	31	5	0
<b>HPV (+)</b>	281 (93.4%)	240 (95.2%)	41 (83.6%)	52 (65.8%)
<b>Smoking<sup>a</sup></b>				
Nonsmoker	142	114	28	23
Long-term former smoker	9	8	1	8
Other former smoker	44	36	8	25
Current smoker	63	54	9	22
Unreported	43	40	3	1
<b>T classification</b>				
TX	17	15	2	4
Tis	1	1	0	0
T1	137	110	27	13
T2	71	58	13	31
T3	20	18	2	19
T4	10	9	1	4
Unreported	45	41	4	0
<b>N classification</b>				
NX	66	56	10	3
N0	130	104	26	21
N1	60	51	9	8
N2	0	0	0	43
N3	0	0	0	4
Unreported	45	41	4	0
<b>Stage</b>				
I	159	125	34	5
II	76	69	7	10
III	38	35	3	13
IV	21	16	5	49
Unreported	7	7	0	2
<b>Deceased in study</b>	72	61	11	22

<sup>a</sup> Smoking was defined as no history of smoking, a former smoker of  $\geq$  15 years, other former smoker of  $<$ 15 years, or a current smoker.

relation to HPV status in individual tumor types. Coding transcripts were evaluated using the Student's t-test. miRNA expression levels were evaluated using a permutation test, also called label shuffling. Specifically, HPV status was randomly assigned to samples in the cohort while maintaining the same proportion of HPV positive patients. Expression fold change for the miRNA was calculated for the shuffled cohort. After one million iterations of this permutation, the true expression fold change was ranked in comparison to the randomly determined fold changes; its position in relation to the shuffled set constituted its significance.

#### *Target and Pathway Analysis*

Likely miRNA-target interactions were initially identified using the MirTarget (version 3) algorithm, the results of which were obtained from miRDB.org (20,21). Additional filters were used to confirm that the results were properly associated, i.e. that miRNA expression was negatively correlated with target expression in the given cohort.

Pathway analysis was conducted on significantly dysregulated targets using the PANTHER algorithm (22). Raw p-values calculated using the binomial test were used to identify significantly overrepresented or underrepresented pathways associated with the Gene Ontology Consortium definitions (23,24).

## **4.4 Results**

### *Dysregulation of miRNAs in Cervical and Oropharyngeal Cancers in Response to HPV Infection*

Within oropharyngeal cancers, 440 miRNAs were identified as being expressed, i.e. having a log<sub>2</sub> mean expression greater than 0.5 across all samples. Of these 440 expressed miRNAs, permutation analysis using a 2-tailed curve identified 207 as significantly

**Table 4.2: HPV-dysregulated miRNAs in OPSCC**

miRNA	raw p-value	FWER	log2 fold change	miRNA	raw p-value	FWER	log2 fold change
hsa-miR-20b-5p	0	0	3.517471	hsa-miR-30e-5p	1E-06	0.00044	0.829569
hsa-miR-9-5p	0	0	3.28475	hsa-miR-625-5p	9E-06	0.00396	0.816012
hsa-miR-363-3p	0	0	2.995527	hsa-miR-1295a	4E-06	0.00176	0.800879
hsa-miR-106a-5p	0	0	2.31328	hsa-miR-30e-3p	0	0	0.751577
hsa-miR-20b-3p	0	0	2.282462	hsa-miR-7-1-3p	1.4E-05	0.00616	0.725956
hsa-miR-9-3p	0	0	2.0645	hsa-miR-548b-3p	2.1E-05	0.00924	0.698775
hsa-miR-99a-5p	0	0	1.955681	hsa-miR-25-3p	2E-06	0.00088	0.688253
hsa-miR-125b-2-3p	1.6E-05	0.00704	1.754344	hsa-miR-107	2.5E-05	0.011	0.634541
hsa-miR-150-3p	1.1E-05	0.00484	1.514754	hsa-miR-34a-5p	3.8E-05	0.01672	0.627293
hsa-let-7c-5p	1.1E-05	0.00484	1.457579	hsa-miR-2277-5p	6E-06	0.00264	0.607582
hsa-miR-378c	0	0	1.415099	hsa-miR-3610	2.8E-05	0.01232	0.593877
hsa-miR-1266-5p	2.4E-05	0.01056	1.302272	hsa-miR-22-3p	0.000021	0.00924	-0.55274
hsa-miR-148a-5p	3E-06	0.00132	1.214824	hsa-miR-22-5p	0.000015	0.0066	-0.7918
hsa-miR-29c-3p	2.3E-05	0.01012	1.214248	hsa-miR-365a-3p	0.000016	0.00704	-0.87431
hsa-miR-598-3p	4.2E-05	0.01848	1.164478	hsa-miR-655-3p	0.00002	0.0088	-0.90531
hsa-miR-378a-3p	1.2E-05	0.00528	1.158094	hsa-miR-455-5p	0.000013	0.00572	-0.96418
hsa-miR-378a-5p	4E-06	0.00176	1.057344	hsa-miR-193b-5p	0.000016	0.00704	-0.98032
hsa-miR-15b-5p	0	0	1.049673	hsa-miR-199b-5p	0.000023	0.01012	-1.04124
hsa-miR-148a-3p	6E-06	0.00264	1.049506	hsa-miR-493-3p	0.000045	0.0198	-1.1074
hsa-miR-101-3p	0	0	1.036463	hsa-miR-369-3p	0.000024	0.01056	-1.13486
hsa-miR-16-2-3p	2E-05	0.0088	1.018546	hsa-miR-214-3p	0.00002	0.0088	-1.13489
hsa-miR-582-3p	2E-06	0.00088	1.007475	hsa-miR-2355-5p	0.000001	0.00044	-1.17311
hsa-miR-3917	0	0	0.993058	hsa-miR-376c-3p	0.000009	0.00396	-1.1785
hsa-miR-15b-3p	2E-06	0.00088	0.960788	hsa-miR-493-5p	0.000049	0.02156	-1.18468
hsa-miR-200b-3p	1E-06	0.00044	0.936816	hsa-miR-214-5p	0.000006	0.00264	-1.23375
hsa-miR-30d-3p	0	0	0.935208	hsa-miR-299-5p	0.000001	0.00044	-1.2757
hsa-miR-200b-5p	1E-05	0.0044	0.934769	hsa-miR-432-5p	0.000013	0.00572	-1.32626
hsa-miR-200a-5p	1.7E-05	0.00748	0.923817	hsa-miR-193b-3p	0	0	-1.43176
hsa-miR-625-3p	1E-06	0.00044	0.886313	hsa-miR-2355-3p	0.000001	0.00044	-1.44537
hsa-miR-30d-5p	0	0	0.852316	hsa-miR-584-5p	0	0	-1.6966
hsa-miR-16-5p	3E-06	0.00132	0.833649	hsa-miR-31-3p	0.000026	0.01144	-2.02521

dysregulated; 62 miRNAs maintained significance after Bonferroni multiple testing correction (Table 4.2). Similar numbers were observed in the cervical cancer cohort. Using the entire patient cohort across all HPV types, 426 miRNAs were defined as expressed. By raw p-value, 191 miRNAs were significantly dysregulated in response to HPV, 44 of which were maintained after Bonferroni correction (11 upregulated, 33 downregulated) (Table 4.3).

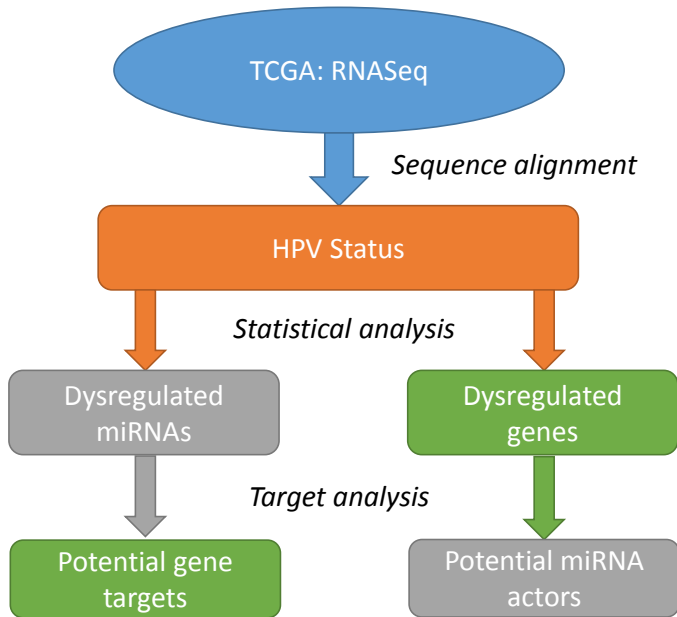
**Table 4.3: HPV-dysregulated miRNAs in CESC**

miRNA	raw p-value	FWER	log2 fold change	miRNA	raw p-value	FWER	log2 fold change
hsa-miR-944	0	0	3.397764	hsa-miR-425-3p	0	0	-1.02183
hsa-miR-205-5p	3E-06	0.001278	3.127832	hsa-miR-181c-3p	0.000004	0.001704	-1.03397
hsa-miR-31-5p	0	0	2.597114	hsa-miR-539-5p	0.00004	0.01704	-1.05188
hsa-miR-31-3p	0	0	2.196549	hsa-miR-130b-5p	0.000002	0.000852	-1.06675
hsa-miR-224-5p	3.7E-05	0.015762	1.569408	hsa-miR-96-5p	0.00001	0.00426	-1.10145
hsa-miR-205-3p	8E-06	0.003408	1.49973	hsa-miR-369-3p	0.000034	0.014484	-1.10919
hsa-miR-224-3p	3.1E-05	0.013206	1.063936	hsa-miR-744-3p	0.000035	0.01491	-1.11066
hsa-miR-221-3p	5.8E-05	0.024708	1.053583	hsa-miR-1468-5p	0	0	-1.11834
hsa-miR-21-3p	6E-06	0.002556	0.928698	hsa-miR-191-5p	0	0	-1.12185
hsa-let-7b-5p	1.8E-05	0.007668	0.677123	hsa-miR-425-5p	0.000001	0.000426	-1.12377
hsa-miR-21-5p	4.6E-05	0.019596	0.478553	hsa-miR-501-3p	0.000005	0.00213	-1.15102
hsa-miR-151a-5p	0.000015	0.00639	-0.61152	hsa-miR-323b-3p	0.000007	0.002982	-1.16143
hsa-miR-148b-5p	0.000022	0.009372	-0.6658	hsa-miR-495-3p	0.000025	0.01065	-1.16932
hsa-miR-324-5p	0.000036	0.015336	-0.77744	hsa-miR-183-5p	0.00001	0.00426	-1.18963
hsa-miR-744-5p	0.000039	0.016614	-0.78916	hsa-miR-432-5p	0.000029	0.012354	-1.21449
hsa-miR-93-3p	0.000011	0.004686	-0.79362	hsa-miR-483-3p	0.000054	0.023004	-1.2179
hsa-miR-340-3p	0.000001	0.000426	-0.80925	hsa-miR-323a-3p	0.000006	0.002556	-1.33359
hsa-miR-324-3p	0.000041	0.017466	-0.81764	hsa-miR-191-3p	0	0	-1.36666
hsa-miR-887-3p	0.000056	0.023856	-0.82948	hsa-miR-431-3p	0.000002	0.000852	-1.57704
hsa-miR-874-3p	0.000039	0.016614	-0.90159	hsa-miR-3200-3p	0	0	-1.84847
hsa-miR-103a-2-5p	0.000022	0.009372	-0.94129	hsa-miR-767-5p	0.000036	0.015336	-2.47921
hsa-miR-500a-3p	0.000008	0.003408	-0.94769	hsa-miR-105-5p	0.000026	0.011076	-2.73646

Of the two Bonferroni corrected sets, there are only three miRNAs that are significant in both oropharyngeal and cervical cancers: miR-31-3p, miR-369-3p, and miR-432-5p. The most notable of the three overlapping miRNAs is miR-31-3p, which was identified as upregulated by HPV in cervical cancer but downregulated by HPV in oropharyngeal cancer. The remaining two miRNAs in this subset were both downregulated by HPV.

*Pathway Analysis of Dysregulated miRNA Targets Indicates miRNA Effects Supplement Basal HPV Activity*

Identification of distinct microRNA-target interactions was performed independently through the significantly dysregulated miRNAs and the significantly dysregulated gene



**Figure 4.1:** Mechanism for identifying miRNA-target interactions in cancers. After determining HPV status through alignment to the virome, miRDB was used to identify targets of dysregulated miRNAs were identified through miRDB (left fork) and potential miRNA regulators of dysregulated transcripts (right fork).

transcripts (Figure 4.1). In the first method, targets of miRNAs significantly associated to HPV status were determined to be affected if the target fulfilled the following criteria: 1) the target is identified in miRDB as associated with the miRNA with a score greater than 50; 2) the target has an average expression greater than 6 RPKM; and 3) the transcript expression is negatively correlated with miRNA expression. In this technique, the significance of the target expression in relation to HPV status was not taken into account. A total of 2607 potential targets were identified as being targeted by dysregulated miRNAs in oropharyngeal cancer: 2333 in the HPV(+) cohort and 274 in the HPV(-) cohort. In the cervical cancer cohort, a total of 2638 potential miRNA targets were identified: 2051 in the HPV(+) cohort and 587 in the HPV(-) group. The inverse relationship was also explored; specifically, the significance of gene transcripts in response to HPV as the experimental factor was determined, and potential miRNA regulators were identified using the same criteria including miRDB score, expression as defined



previously, and negative correlation between miRNA and target. In OPSCC, this yielded 500 significant transcripts in the HPV(+) cohort and 354 transcripts in the HPV(-) cohort; in CESC, this yielded 1583 transcripts in the HPV(+) group and 384 in the HPV(-) group.

The targets of the potential miRNA interactions were then analyzed using the PANTHER database, available at pantherdb.org (22). PANTHER implements the Gene Ontology (GO) database to determine if certain gene sets, such as those defined by GO as members in the same biological process, are overrepresented in a submitted set of genes. Significance is determined with the binomial test. When using significant miRNAs as the initial feature, a total of 55 biological processes were identified as significantly overrepresented by upregulated genes in the HPV(+) oropharyngeal cancer dataset, while 53 processes were significantly underrepresented (Supplementary Table 4.1). In the HPV(+) cervical cancer datasets, 43 biological processes were overrepresented and 23 processes underrepresented (Supplementary Table 4.2). Among the HPV(-) datasets, 24 biological processes were significantly overrepresented by potential miRNA targets among oropharyngeal cancers, and 10 were underrepresented (Supplementary Table 4.3). The HPV(-) cervical cancer set, although limited in scope, showed 22 overrepresented biological processes and 31 underrepresented processes (Supplementary Table 4.4). Notable processes that were upregulated in HPV(+) tumors of both species include DNA processes of replication, recombination and repair, metabolic processes, cellular component organization, mitosis, and stress response; while downregulated pathways include cell recognition, complement activation, and GPCR signaling pathways. Processes upregulated in both HPV(-) cancer types include cellular component organization and rRNA metabolic processes. The only process downregulated in both HPV(-) tumor types was immune response.

Using dysregulated gene transcripts as the initial focus yielded similar results. After filtering by using potential miRNA dysregulation and interaction, 28 pathways were overrepresented in HPV(+) oropharyngeal tumors and 9 pathways were downregulated (Supplementary Table 4.5), while 45 pathways were upregulated in HPV(+) cervical cancer and 23 were downregulated (Supplementary Table 4.6). In HPV(-) oropharyngeal tumors, 37 biological processes were overrepresented and 27 underrepresented (Supplementary Table 4.7). Comparatively in HPV(-) cervical cancers, 9 processes were upregulated and 23 were underrepresented (Supplementary Table 4.8). The HPV(+) cohorts shared an overrepresentation of cellular defense response, immune responses, signaling cascades, and metabolic processes, while also demonstrating an underrepresentation of translation and defense response to bacterium. In the HPV(-) cohort, both cancer types showed an overrepresentation of cellular component biogenesis, metabolic processes, organelle organization, and translation, along with an underrepresentation of general biological regulation, immune response, and both intercellular and intracellular communication. Similar results were observed when comparing the pathway analyses generated with the two different methods of identifying potential targets, suggesting that the observed biological processes may be supplemented, rather than exclusively moderated, by miRNA dysregulation.

#### *Individual miRNA-Target Interactions are Conserved in HPV-Related Cancers*

By evaluating statistical significance of downregulated targets, in addition to statistical significance of miRNAs, a small subset of potential miRNA-target interactions were identified to be conserved between the two HPV-related cancer types (Table 12). These interactions were identified using significance for both miRNA expression and target expression ( $p < 0.05$ ), in addition to the same criteria as described previously: miRDB score, target expression, and

**Table 4.4:** miRNA-Target Interactions Conserved Between Cervical and Oropharyngeal Cancers in Response to HPV Status

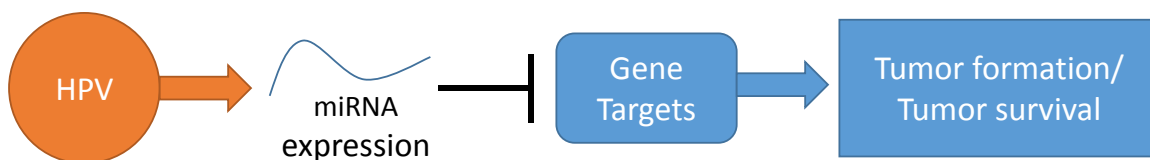
Targets of miRNAs Downregulated in HPV(+) Tumors				Targets of miRNAs Upregulated in HPV(+) Tumors	
hsa-miR-105-5p	FCER1A	hsa-miR-483-3p	MECP2	hsa-miR-101-3p	XPO5
hsa-miR-105-5p	MECP2	hsa-miR-485-3p	PIGK	hsa-miR-101-3p	PMPCB
hsa-miR-105-5p	TAF9B	hsa-miR-485-3p	KLF6	hsa-miR-142-5p	CNOT11
hsa-miR-105-5p	SNIP1	hsa-miR-485-3p	MAPKBP1	hsa-miR-16-1-3p	NLN
hsa-miR-105-5p	MED14	hsa-miR-485-3p	MAT2B	hsa-miR-205-3p	CIAO1
hsa-miR-136-5p	BTN3A2	hsa-miR-493-3p	KLF6	hsa-miR-29c-3p	GCSH
hsa-miR-136-5p	IFNGR1	hsa-miR-493-5p	IRF2		
hsa-miR-154-3p	KDM6A	hsa-miR-493-5p	LAMP2		
hsa-miR-154-5p	PCNA	hsa-miR-493-5p	KDM6A		
hsa-miR-154-5p	CLOCK	hsa-miR-493-5p	SNN		
hsa-miR-181a-2-3p	TMEM173	hsa-miR-493-5p	CLOCK		
hsa-miR-181a-2-3p	IL13RA1	hsa-miR-495-3p	IRF2		
hsa-miR-181a-2-3p	LY75	hsa-miR-495-3p	DCLRE1B		
hsa-miR-181a-2-3p	MED14	hsa-miR-495-3p	STAT3		
hsa-miR-181b-5p	COL16A1	hsa-miR-495-3p	UGCG		
hsa-miR-181b-5p	KLF6	hsa-miR-495-3p	SNIP1		
hsa-miR-181b-5p	GANC	hsa-miR-495-3p	CARD6		
hsa-miR-181b-5p	TAF9B	hsa-miR-495-3p	RUNX3		
hsa-miR-181b-5p	ZFP36L1	hsa-miR-495-3p	CLOCK		
hsa-miR-181b-5p	SNN	hsa-miR-514a-3p	KLF6		
hsa-miR-3127-5p	CLOCK	hsa-miR-514a-3p	SNIP1		
hsa-miR-323a-3p	PIGK	hsa-miR-539-5p	CCDC50		
hsa-miR-323a-3p	STAT3	hsa-miR-539-5p	FAM120C		
hsa-miR-323a-3p	KLF11	hsa-miR-539-5p	CPPED1		
hsa-miR-337-3p	CCDC50	hsa-miR-539-5p	DCLRE1B		
hsa-miR-337-3p	IL13RA1	hsa-miR-539-5p	ZFP36L1		
hsa-miR-337-3p	STAT3	hsa-miR-539-5p	KDM6A		
hsa-miR-369-3p	PIGK	hsa-miR-539-5p	CLOCK		
hsa-miR-369-3p	DPYD	hsa-miR-654-3p	CENPI		
hsa-miR-369-3p	UGCG	hsa-miR-654-3p	FUNDC2		
hsa-miR-370-3p	MESDC2	hsa-miR-654-3p	MED14		
hsa-miR-370-3p	GJB3	hsa-miR-654-3p	LITAF		
hsa-miR-370-3p	PMAIP1	hsa-miR-654-5p	EDARADD		
hsa-miR-370-3p	DCLRE1B	hsa-miR-655-3p	PIGK		
hsa-miR-370-3p	STAT3	hsa-miR-655-3p	UBD		
hsa-miR-370-3p	N4BP2L1	hsa-miR-655-3p	DPYD		
hsa-miR-382-5p	FAM120C	hsa-miR-655-3p	MOB3B		
hsa-miR-382-5p	PRPS2	hsa-miR-675-3p	ATG4A		
hsa-miR-409-3p	DCLRE1B	hsa-miR-675-3p	FAM120C		
hsa-miR-409-3p	MED14	hsa-miR-675-3p	MED14		
hsa-miR-410-3p	KLF6	hsa-miR-758-3p	PLP2		
hsa-miR-410-3p	HS3ST1	hsa-miR-758-3p	KDM6A		
hsa-miR-432-5p	MECP2	hsa-miR-758-3p	MOB3B		
hsa-miR-432-5p	N4BP2L1	hsa-miR-767-5p	KLF6		
hsa-miR-450b-5p	ENOX2	hsa-miR-767-5p	CCDC50		
hsa-miR-450b-5p	ADRB2	hsa-miR-767-5p	NASP		
hsa-miR-483-3p	ICAM1	hsa-miR-767-5p	N4BP2L1		

directional correlation. In doing so, 96 target interactions resulting from HPV-downregulated

miRNAs were identified as conserved between OPSCC and CESC, spanning 30 unique miRNAs and 52 unique targets. A number of gene transcripts were targeted by multiple miRNAs in this set; considering that the respective miRNAs are downregulated, these particular transcripts were upregulated. Of note are the tumor suppressor KLF6, the circadian rhythm gene CLOCK, and STAT3. Interestingly, increased activation of STAT3, a vital component of the JAK/STAT signaling pathway, promotes cancer growth and angiogenesis, while overexpression of CLOCK and KLF6 are both associated with tumor survival and reduction in tumor size (25-28). Only 6 total miRNA-target interactions were conserved by HPV-upregulated miRNAs; none of the targets of the identified interactions are noted in the literature for their involvement in cancer.

## 4.5 Discussion

Human papillomavirus infection causes approximately 95% of all cervical cancers, and may be responsible for up to 75% of new oropharyngeal cancer diagnoses (29,30). The role of HPV in tumor formation has been well-characterized, along with its prognostic significance after diagnosis (31,32). However, the mechanisms that result in this unusual dichotomy are not well characterized and may be controlled through some more subtle regulatory mechanisms, such as the RNAi mechanism in which miRNAs are involved.



**Figure 4.2:** A diagram of potential miRNA-mediated dysregulation in response to HPV. HPV infection can induce changes in miRNA expression levels, which result in the opposite effect of the potential gene transcript targets. Through the miRNA regulatory network, mechanisms of tumor formation may be supplemented, and additional mechanisms of tumor survival may be characterized.

The dysregulation of miRNA expression in response to HPV infection can result in downstream effects in the transcriptome prior to translation (Figure 2). Therefore, we aimed to identify both significantly dysregulated miRNA regulators and potential gene targets between HPV(+) and HPV(-) tumors. Contemporary pathway analysis, including PANTHER and Gene Set Enrichment Analysis, focuses primarily on coding transcripts and occasionally long non-coding RNAs (22,33). Therefore, we employed the PANTHER database to analyze the gene transcripts that we identified as potential targets of dysregulated miRNAs; to confirm our results, we also performed the analysis based on dysregulated transcripts, and performed target analysis to identify potential miRNA actors. Many of the biological processes that were determined to be significantly selected through both techniques could be inferred as associated with HPV-induced oncogenesis. For example, overrepresented processes in HPV(+) tumors included processes that are conceivably associated with cell growth and replication, such as mitosis, cellular component organization, metabolism, and cell cycle, while underrepresented processes included immune responses such as cell recognition. The underrepresented immune response in HPV(+) tumors suggest immune evasion, which has been indicated in the literature as one of the roles of the E2 protein (34). The HPV(-) tumors also show overrepresentation of the processes that can be inferred as relevant in oncogenesis, as well as a lack of immune response, which in this context may simply be unnecessary; however, HPV(-) tumors also indicate that biological processes associated with cell adhesion and cell communication are significantly underrepresented; this may account for poorer prognosis in the sense of encouraging metastasis and unchecked cell replication and growth. One of the primary weaknesses of PANTHER analysis is its inability to account for expression level, especially when compared to GSEA; however, PANTHER is more effective than GSEA when examining smaller gene sets.

Individual miRNA-target interactions were not ignored in this analysis; although a top-level pathway analysis can provide guidance as to how miRNAs may operate in response to HPV infection, individual miRNA-target interactions can also indicate potential therapeutic targets. This type of analysis has been proposed previously using HPV-infected NIKS cells (35), but to the best of our knowledge, this study is the first to examine such interactions using transcriptome and miRNome data from patient tissue samples. CLOCK, KLF6, and STAT3 have already been highlighted. Other targets that were overexpressed in HPV(+) tumors and targeted by downregulated miRNAs include the pro-survival gene CCDC50, the tumor suppressor KDM6A, and the oncogene SNIP1 (36-38). This combination of pro-survival and tumor suppressor genes with oncogenes among targets of dysregulated miRNAs indicate possible divergent mechanisms for tumor formation and later survival; these particular interactions are strong candidates for direct analysis through cellular and in-house experiments.

Adenocarcinomas and adenosquamous cell carcinomas of the cervix (cervical adeno-type carcinomas) are both histologically and genetically distinct from cervical squamous cell carcinomas (39,40). Additionally, there have been reports that members of the alphapapillomavirus 9 species enriched in adenosquamous cell carcinomas and adenocarcinomas of the cervix, in comparison to squamous cell carcinomas of the cervix (41). In conducting our HPV sequencing analysis, we determined that these previous findings are consistent within the cervical cancer cohort from TCGA (Table 4.5). The distribution of HPV types in cervical adeno-type carcinomas significantly favors members of the Alphapapillomavirus type 9 species, as well as HPV(-) tumors, at the expense of the more common Alphapapillomavirus type 7 species ( $\chi^2 p= 0.027$ ). This observation is not observed in cervical squamous cell carcinomas ( $\chi^2 p= 0.71$ ). Interestingly, HPV(+) OPSCCs from TCGA are all infected by members of the

**Table 4 5:** HPV types in cervical and oropharyngeal cancers, separated by tumor source site

HPV species	HPV type	All Cervical Cancer	Cervical Squamous Cell Carcinoma	Cervical Adeno-type Carcinoma	Oropharyngeal Squamous Cell Carcinoma
Alpha9	HPV16	163	136	27	45
	HPV31	7	7	0	0
	HPV33	9	9	0	4
	HPV35	6	6	0	3
	HPV52	8	8	0	0
	HPV58	7	7	0	0
	Total	200	173	27	52
Alpha7	HPV18	37	27	10	0
	HPV39	6	6	0	0
	HPV45	22	19	3	0
	HPV59	3	3	0	0
	HPV68	2	2	0	0
	Total	70	57	13	0
Other	HPV26	1	1	0	0
	HPV30	1	1	0	0
	HPV51	1	1	0	0
	HPV56	1	1	0	0
	HPV69	1	1	0	0
	HPV70	2	2	0	0
	HPV73	2	2	0	0
	Total	9	9	0	0
Multiple		2	1	1	0
HPV(-)		20	12	8	27

Alphapapillomavirus 9 species; considering that such tumors can cluster together with cervical squamous tumors, this preferential infectivity is not entirely unexpected.

Additional analysis can also be performed on the basis of HPV species. It has been shown that miR-9-5p is upregulated to a greater extent by HPV16, a member of the alphapapillomavirus 9 species, than HPV18, a member of the higher risk alphapapillomavirus 7 species (5).

However, the effect of HPV type on the dysregulation of the miRNome as a whole is unclear, especially considering that there is a preference for infection of adeno-type carcinomas by the higher risk species (41-43). This correlation begs the question of whether the higher risk of

adeno-type cancers results from the higher risk papillomaviruses, and the role of regulatory miRNAs in promoting patient survival.

Through these analyses, we have laid the foundation for examining miRNA-controlled mechanisms for supplementing HPV-derived tumor formation, as well as HPV-related tumor survival. By modulating and altering the host miRNome, HPV is able to prolong cell survival and replication, as well as activate various pro-survival and tumor suppressors after tumor formation. We have demonstrated some of the miRNA-based mechanisms that may be induced by HPV, and believe that a continued focus on how regulatory systems such as RNAi will be able to elucidate the unusual behavior of HPV that causes infection to be both oncogenic and pro-survival.

## 4.6 References

1. zur Hausen, H., Meinhof, W., Scheiber, W. and Bornkamm, G.W. (1974) Attempts to detect virus-specific DNA in human tumors. I. Nucleic acid hybridizations with complementary RNA of human wart virus. *Int J Cancer*, **13**, 650-656.
2. zur Hausen, H., Schulte-Holthausen, H., Wolf, H., Dörries, K. and Egger, H. (1974) Attempts to detect virus-specific DNA in human tumors. II. Nucleic acid hybridizations with complementary RNA of human herpes group viruses. *Int J Cancer*, **13**, 657-664.
3. Vande Pol, S.B. and Klingelutz, A.J. (2013) Papillomavirus E6 oncoproteins. *Virology*, **445**, 115-137.
4. Roman, A. and Munger, K. (2013) The papillomavirus E7 proteins. *Virology*, **445**, 138-168.
5. Liu, W., Gao, G., Hu, X., Wang, Y., Schwarz, J.K., Chen, J.J., Grigsby, P.W. and Wang, X. (2014) Activation of miR-9 by human papillomavirus in cervical cancer. *Oncotarget*, **5**, 11583-11593.
6. Gao, G., Chernock, R.D., Gay, H.A., Thorstad, W.L., Zhang, T.R., Wang, H., Ma, X.J., Luo, Y., Lewis, J.S. and Wang, X. (2013) A novel RT-PCR method for quantification of human papillomavirus transcripts in archived tissues and its application in oropharyngeal cancer prognosis. *Int J Cancer*, **132**, 882-890.



7. Gunasekharan, V. and Laimins, L.A. (2013) Human papillomaviruses modulate microRNA 145 expression to directly control genome amplification. *J Virol*, **87**, 6037-6043.
8. Yeung, C.L., Tsang, T.Y., Yau, P.L. and Kwok, T.T. (2017) Human papillomavirus type 16 E6 suppresses microRNA-23b expression in human cervical cancer cells through DNA methylation of the host gene C9orf3. *Oncotarget*, **8**, 12158-12173.
9. Gao, G., Gay, H.A., Chernock, R.D., Zhang, T.R., Luo, J., Thorstad, W.L., Lewis, J., James S and Wang, X. (2013) A microRNA expression signature for the prognosis of oropharyngeal squamous cell carcinoma. *Cancer*, **119**, 72-80.
10. Hu, X., Schwarz, J.K., Lewis, J., James S, Huettner, P.C., Rader, J.S., Deasy, J.O., Grigsby, P.W. and Wang, X. (2010) A microRNA expression signature for cervical cancer prognosis. *Cancer Res*, **70**, 1441-1448.
11. Hui, A.B., Lin, A., Xu, W., Waldron, L., Perez-Ordenez, B., Weinreb, I., Shi, W., Bruce, J., Huang, S.H., O'Sullivan, B. *et al.* (2013) Potentially prognostic miRNAs in HPV-associated oropharyngeal carcinoma. *Clin Cancer Res*, **19**, 2154-2162.
12. How, C., Pintilie, M., Bruce, J.P., Hui, A.B., Clarke, B.A., Wong, P., Yin, S., Yan, R., Waggott, D., Boutros, P.C. *et al.* (2015) Developing a prognostic micro-RNA signature for human cervical carcinoma. *PLoS One*, **10**, e0123946.
13. Wong, N., Khwaja, S.S., Baker, C.M., Gay, H.A., Thorstad, W.L., Daly, M.D., Lewis, J.S. and Wang, X. (2016) Prognostic microRNA signatures derived from The Cancer Genome Atlas for head and neck squamous cell carcinomas. *Cancer Medicine*, **5**, 1619-1628.
14. Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A. and Staudt, L.M. (2016) Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*, **375**, 1109-1112.
15. Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res*, **42**, D764-770.
16. Coordinators, N.R. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **43**, D6-D17.
17. Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., Sun, Q., Kaur, R., Huyen, Y. and McBride, A.A. (2017) The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res*, **45**, D499-D506.
18. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, **42**, D68-D73.
19. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

20. Wong, N. and Wang, X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*, **43**, D146-D152.
21. Wang, X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, **32**, 1316-1322.
22. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D. and Thomas, P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, **45**, D183-D189.
23. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.
24. The Gene Ontology Consortium. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*, **45**, D331-D338.
25. Yu, H., Lee, H., Herrmann, A., Buettner, R. and Jove, R. (2014) Revisiting STAT3 signalling in cancer: new and unexpected biological functions. *Nat Rev Cancer*, **14**, 736-746.
26. Cadenas, C., van de Sandt, L., Edlund, K., Lohr, M., Hellwig, B., Marchan, R., Schmidt, M., Rahnenführer, J., Oster, H. and Hengstler, J.G. (2014) Loss of circadian clock gene expression is associated with tumor progression in breast cancer. *Cell Cycle*, **13**, 3282-3291.
27. Sangodkar, J., Shi, J., DiFeo, A., Schwartz, R., Bromberg, R., Choudhri, A., McClinch, K., Hatami, R., Scheer, E., Kremer-Tal, S. *et al.* (2009) Functional role of the KLF6 tumour suppressor gene in gastric cancer. *Eur J Cancer*, **45**, 666-676.
28. Masilamani, A.P., Ferrarese, R., Kling, E., Thudi, N.K., Kim, H., Scholtens, D.M., Dai, F., Hadler, M., Unterkircher, T., Platania, L. *et al.* (2017) KLF6 depletion promotes NF- $\kappa$ B signaling in glioblastoma. *Oncogene*, **36**, 3562-3575.
29. Schiffman, M., Wentzensen, N., Wacholder, S., Kinney, W., Gage, J.C. and Castle, P.E. (2011) Human papillomavirus testing in the prevention of cervical cancer. *J Natl Cancer Inst*, **103**, 368-383.
30. Moore, K.A. and Mehta, V. (2015) The Growing Epidemic of HPV-Positive Oropharyngeal Carcinoma: A Clinical Review for Primary Care Providers. *J Am Board Fam Med*, **28**, 498-503.
31. Lombard, I., Vincent-Salomon, A., Validire, P., Zafrani, B., de la Rochefordière, A., Clough, K., Favre, M., Pouillart, P. and Sastre-Garau, X. (1998) Human papillomavirus genotype as a major determinant of the course of cervical cancer. *J Clin Oncol*, **16**, 2613-2619.

32. Ang, K.K., Harris, J., Wheeler, R., Weber, R., Rosenthal, D.I., Nguyen-Tân, P.F., Westra, W.H., Chung, C.H., Jordan, R.C., Lu, C. *et al.* (2010) Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*, **363**, 24-35.
33. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.
34. McBride, A.A. (2013) The papillomavirus E2 proteins. *Virology*, **445**, 57-79.
35. Harden, M.E., Prasad, N., Griffiths, A. and Munger, K. (2017) Modulation of microRNA-mRNA Target Pairs by Human Papillomavirus 16 Oncoproteins. *MBio*, **8**.
36. Farfsing, A., Engel, F., Seiffert, M., Hartmann, E., Ott, G., Rosenwald, A., Stilgenbauer, S., Döhner, H., Boutros, M., Lichter, P. *et al.* (2009) Gene knockdown studies revealed CCDC50 as a candidate gene in mantle cell lymphoma and chronic lymphocytic leukemia. *Leukemia*, **23**, 2018-2026.
37. Nickerson, M.L., Dancik, G.M., Im, K.M., Edwards, M.G., Turan, S., Brown, J., Ruiz-Rodriguez, C., Owens, C., Costello, J.C., Guo, G. *et al.* (2014) Concurrent alterations in TERT, KDM6A, and the BRCA pathway in bladder cancer. *Clin Cancer Res*, **20**, 4935-4948.
38. Liang, X., Zheng, M., Jiang, J., Zhu, G., Yang, J. and Tang, Y. (2011) Hypoxia-inducible factor-1 alpha, in association with TWIST2 and SNIP1, is a critical prognostic factor in patients with tongue squamous cell carcinoma. *Oral Oncol*, **47**, 92-97.
39. Wentz, W.B. and Reagan, J.W. (1959) Survival in cervical cancer with respect to cell type. *Cancer*, **12**, 384-388.
40. Network, C.G.A.R., Medicine, A.E.C.o., Services, A.B., Hospital, B.C., Medicine, B.C.o., Hope, B.R.I.o.C.o., Aging, B.I.f.R.o., Centre, C.s.M.S.G.S., School, H.M., Services, H.F.G.C.C.R.I.a.C.C.H. *et al.* (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**, 378-384.
41. Bulk, S., Berkhof, J., Bulkman, N.W., Zielinski, G.D., Rozendaal, L., van Kemenade, F.J., Snijders, P.J. and Meijer, C.J. (2006) Preferential risk of HPV16 for squamous cell carcinoma and of HPV18 for adenocarcinoma of the cervix compared to women with normal cytology in The Netherlands. *Br J Cancer*, **94**, 171-175.
42. Clifford, G. and Franceschi, S. (2008) Members of the human papillomavirus type 18 family (alpha-7 species) share a common association with adenocarcinoma of the cervix. *Int J Cancer*, **122**, 1684-1685.
43. Galic, V., Herzog, T.J., Lewin, S.N., Neugut, A.I., Burke, W.M., Lu, Y.S., Hershman, D.L. and Wright, J.D. (2012) Prognostic significance of adenocarcinoma histology in women with cervical cancer. *Gynecol Oncol*, **125**, 287-291.

## 4.7 Supplementary Tables

**Supplementary Table 4.1:** Significant biological processes in HPV(+) OPSCC, through initial identification of significantly dysregulated miRNAs and subsequent targets

Overrepresented Processes		Underrepresented Processes	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
apoptotic process (GO:0006915)	2.15E-02	system process (GO:0003008)	1.41E-20
biosynthetic process (GO:0009058)	2.03E-04	angiogenesis (GO:0001525)	8.19E-06
carbohydrate metabolic process (GO:0005975)	3.86E-03	anion transport (GO:0006820)	5.56E-04
catabolic process (GO:0009056)	5.83E-10	B cell mediated immunity (GO:0019724)	1.09E-03
cell cycle (GO:0007049)	8.91E-10	behavior (GO:0007610)	2.50E-02
cell death (GO:0008219)	3.80E-02	biological adhesion (GO:0022610)	2.20E-03
cellular amino acid metabolic process (GO:0006520)	2.44E-02	biological regulation (GO:0065007)	8.28E-05
cellular component organization (GO:0016043)	6.24E-07	blood circulation (GO:0008015)	4.73E-06
cellular component organization or biogenesis (GO:0071840)	4.68E-06	cation transport (GO:0006812)	4.91E-02
cellular process (GO:0009987)	1.87E-04	cell adhesion (GO:0007155)	2.20E-03
cellular protein modification process (GO:0006464)	4.44E-04	cell communication (GO:0007154)	1.93E-03
chromatin organization (GO:0006325)	1.38E-09	cell differentiation (GO:0030154)	1.55E-03
chromosome segregation (GO:0007059)	2.15E-03	cell recognition (GO:0008037)	3.66E-02
cytoskeleton organization (GO:0007010)	1.22E-04	cell surface receptor signaling pathway (GO:0007166)	3.98E-08
death (GO:0016265)	3.80E-02	cell-cell adhesion (GO:0016337)	1.61E-02
DNA metabolic process (GO:0006259)	4.20E-13	cell-cell signaling (GO:0007267)	1.84E-05
DNA recombination (GO:0006310)	3.39E-04	cell-matrix adhesion (GO:0007160)	4.52E-03
DNA repair (GO:0006281)	3.96E-08	complement activation (GO:0006956)	9.89E-03
DNA replication (GO:0006260)	1.44E-07	defense response to bacterium (GO:0042742)	1.80E-03
fatty acid beta-oxidation (GO:0006635)	2.06E-03	developmental process (GO:0032502)	1.77E-04
induction of apoptosis (GO:0006917)	3.83E-02	digestive tract mesoderm development (GO:0007502)	2.72E-02
lysosomal transport (GO:0007041)	1.08E-02	ectoderm development (GO:0007398)	1.28E-05
meiosis (GO:0007126)	2.76E-04	female gamete generation (GO:0007292)	3.51E-02
metabolic process (GO:0008152)	5.24E-24	fertilization (GO:0009566)	4.47E-02
mitosis (GO:0007067)	2.35E-05	G-protein coupled receptor signaling pathway (GO:0007186)	2.41E-07
mRNA processing (GO:0006397)	3.86E-16	heart development (GO:0007507)	1.02E-03
mRNA splicing, via spliceosome (GO:0000398)	5.69E-13	immune response (GO:0006955)	2.17E-04
nitric oxide biosynthetic process (GO:0006809)	4.32E-02	immune system process (GO:0002376)	6.70E-03
nitrogen compound metabolic process (GO:0006807)	2.01E-12	ion transport (GO:0006811)	4.44E-05
nuclear transport (GO:0051169)	1.14E-02	macrophage activation (GO:0042116)	3.31E-02
nucleobase-containing compound metabolic process (GO:0006139)	1.63E-25	mesoderm development (GO:0007498)	5.97E-04

nucleobase-containing compound transport (GO:0015931)	3.03E-02	multicellular organismal process (GO:0032501)	4.96E-19
organelle organization (GO:0006996)	1.06E-14	muscle contraction (GO:0006936)	8.34E-06
phosphate-containing compound metabolic process (GO:0006796)	4.39E-10	muscle organ development (GO:0007517)	9.76E-04
phospholipid metabolic process (GO:0006644)	5.58E-03	natural killer cell activation (GO:0030101)	1.56E-02
primary metabolic process (GO:0044238)	1.88E-19	nervous system development (GO:0007399)	8.25E-05
protein acetylation (GO:0006473)	1.95E-03	neurological system process (GO:0050877)	5.40E-15
protein localization (GO:0008104)	2.36E-03	pattern specification process (GO:0007389)	4.41E-02
protein methylation (GO:0006479)	4.92E-02	regulation of biological process (GO:0050789)	8.62E-05
protein targeting (GO:0006605)	5.86E-03	regulation of vasoconstriction (GO:0019229)	1.75E-02
pyrimidine nucleobase metabolic process (GO:0006206)	7.55E-03	response to biotic stimulus (GO:0009607)	1.41E-02
regulation of cell cycle (GO:0051726)	1.27E-05	response to stimulus (GO:0050896)	5.17E-07
regulation of nucleobase-containing compound metabolic process (GO:0019219)	2.45E-05	sensory perception (GO:0007600)	2.19E-15
regulation of phosphate metabolic process (GO:0019220)	2.65E-02	sensory perception of chemical stimulus (GO:0007606)	2.12E-19
regulation of transcription from RNA polymerase II promoter (GO:0006357)	6.10E-05	sensory perception of smell (GO:0007608)	2.22E-14
regulation of translation (GO:0006417)	1.17E-02	sensory perception of sound (GO:0007605)	2.74E-02
response to stress (GO:0006950)	4.78E-02	sensory perception of taste (GO:0050909)	3.95E-02
RNA catabolic process (GO:0006401)	2.29E-03	signal transduction (GO:0007165)	3.60E-03
RNA localization (GO:0006403)	1.75E-02	single-multicellular organism process (GO:0044707)	1.58E-18
RNA metabolic process (GO:0016070)	2.07E-12	skeletal system development (GO:0001501)	3.97E-04
RNA splicing, via transesterification reactions (GO:0000375)	6.02E-11	steroid metabolic process (GO:0008202)	2.37E-02
transcription from RNA polymerase II promoter (GO:0006366)	1.73E-06	synaptic transmission (GO:0007268)	5.54E-04
transcription initiation from RNA polymerase II promoter (GO:0006367)	1.01E-02	system development (GO:0048731)	4.03E-07
transcription, DNA-dependent (GO:0006351)	4.86E-04		
tRNA aminoacylation for protein translation (GO:0006418)	1.72E-03		

**Supplementary Table 4.2:** Significant biological processes in HPV(+) CESC, through initial identification of significantly dysregulated miRNAs and subsequent targets

<b>Overrepresented Processes</b>		<b>Underrepresented Processes</b>	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
7-methylguanosine mRNA capping (GO:0006370)	3.15E-02	cell recognition (GO:0008037)	5.00E-03
antigen processing and presentation (GO:0019882)	1.27E-02	complement activation (GO:0006956)	1.49E-02
biological adhesion (GO:0022610)	3.06E-03	defense response to bacterium (GO:0042742)	6.39E-04
catabolic process (GO:0009056)	1.80E-02	gamete generation (GO:0007276)	4.17E-02
cell adhesion (GO:0007155)	3.06E-03	G-protein coupled receptor signaling pathway (GO:0007186)	3.27E-02
cell cycle (GO:0007049)	1.23E-05	mitochondrion organization (GO:0007005)	1.30E-02
cellular component morphogenesis (GO:0032989)	4.98E-04	multicellular organismal process (GO:0032501)	1.88E-05
cellular component movement (GO:0006928)	3.29E-03	muscle organ development (GO:0007517)	3.49E-02
cellular component organization (GO:0016043)	7.51E-04	neurological system process (GO:0050877)	1.27E-06
cellular component organization or biogenesis (GO:0071840)	9.96E-03	oxidative phosphorylation (GO:0006119)	1.11E-02
cellular defense response (GO:0006968)	2.86E-02	phagocytosis (GO:0006909)	1.17E-02
cellular process (GO:0009987)	2.25E-04	protein metabolic process (GO:0019538)	1.44E-02
chromatin assembly (GO:0031497)	1.34E-02	response to biotic stimulus (GO:0009607)	1.56E-02
chromatin organization (GO:0006325)	2.22E-02	RNA metabolic process (GO:0016070)	3.60E-02
chromosome segregation (GO:0007059)	2.17E-04	rRNA metabolic process (GO:0016072)	4.29E-02
cytoskeleton organization (GO:0007010)	4.22E-04	sensory perception (GO:0007600)	1.09E-11
DNA metabolic process (GO:0006259)	5.53E-09	sensory perception of chemical stimulus (GO:0007606)	3.30E-16
DNA recombination (GO:0006310)	1.12E-02	sensory perception of smell (GO:0007608)	3.44E-12
DNA repair (GO:0006281)	1.73E-04	single-multicellular organism process (GO:0044707)	2.21E-05
DNA replication (GO:0006260)	3.77E-07	system process (GO:0003008)	1.99E-06
I-kappaB kinase/NF-kappaB cascade (GO:0007249)	4.39E-03	translation (GO:0006412)	6.14E-03
immune system process (GO:0002376)	4.81E-03	tRNA metabolic process (GO:0006399)	9.60E-03
intracellular signal transduction (GO:0035556)	1.71E-03	Unclassified (UNCLASSIFIED)	2.48E-04
locomotion (GO:0040011)	5.03E-03		
MAPK cascade (GO:0000165)	4.22E-02		
metabolic process (GO:0008152)	1.34E-02		
mitosis (GO:0007067)	4.37E-02		
negative regulation of apoptotic process (GO:0043066)	1.77E-02		
nitrogen compound metabolic process (GO:0006807)	4.51E-02		
nucleobase-containing compound metabolic process (GO:0006139)	1.44E-03		
organelle organization (GO:0006996)	4.73E-03		
phosphate-containing compound metabolic process (GO:0006796)	2.83E-03		
primary metabolic process (GO:0044238)	3.02E-02		

protein localization (GO:0008104)	3.58E-02
pyrimidine nucleobase metabolic process (GO:0006206)	4.25E-02
regulation of catalytic activity (GO:0050790)	1.89E-02
regulation of molecular function (GO:0065009)	2.54E-02
regulation of nucleobase-containing compound metabolic process (GO:0019219)	3.98E-02
regulation of transcription from RNA polymerase II promoter (GO:0006357)	2.43E-02
response to abiotic stimulus (GO:0009628)	3.56E-02
response to external stimulus (GO:0009605)	3.35E-02
response to interferon-gamma (GO:0034341)	3.32E-03
response to stress (GO:0006950)	1.16E-02

---

**Supplementary Table 4.3:** Significant biological processes in HPV(-) OPSCC, through initial identification of significantly dysregulated miRNAs and subsequent targets

<b>Overrepresented Processes</b>		<b>Underrepresented Processes</b>	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
anion transport (GO:0006820)	3.00E-02	gamete generation (GO:0007276)	4.32E-02
cellular component morphogenesis (GO:0032989)	9.32E-10	immune response (GO:0006955)	1.54E-03
cellular component movement (GO:0006928)	2.15E-06	regulation of nucleobase-containing compound metabolic process (GO:0019219)	1.90E-02
cellular component organization (GO:0016043)	4.29E-06	regulation of transcription from RNA polymerase II promoter (GO:0006357)	3.33E-02
cellular component organization or biogenesis (GO:0071840)	1.38E-07	reproduction (GO:0000003)	1.30E-02
cellular process (GO:0009987)	4.89E-06	response to stimulus (GO:0050896)	1.57E-02
chromosome segregation (GO:0007059)	1.31E-02	sensory perception (GO:0007600)	4.33E-02
developmental process (GO:0032502)	3.36E-02	sensory perception of chemical stimulus (GO:0007606)	4.01E-03
gluconeogenesis (GO:0006094)	3.04E-02	sensory perception of smell (GO:0007608)	2.47E-02
intracellular protein transport (GO:0006886)	1.29E-03	Unclassified (UNCLASSIFIED)	6.04E-04
ion transport (GO:0006811)	1.21E-02		
localization (GO:0051179)	7.59E-05		
locomotion (GO:0040011)	2.18E-03		
mesoderm development (GO:0007498)	2.55E-03		
muscle organ development (GO:0007517)	3.31E-02		
nuclear transport (GO:0051169)	4.51E-02		
pentose-phosphate shunt (GO:0006098)	1.25E-02		
polysaccharide metabolic process (GO:0005976)	1.83E-02		
protein targeting (GO:0006605)	4.69E-02		
protein transport (GO:0015031)	1.83E-03		
regulation of carbohydrate metabolic process (GO:0006109)	4.32E-03		
rRNA metabolic process (GO:0016072)	2.90E-02		
secondary metabolic process (GO:0019748)	6.18E-03		
transport (GO:0006810)	2.93E-04		



**Supplementary Table 4.4:** Significant biological processes in HPV(-) CESC, through initial identification of significantly dysregulated miRNAs and subsequent targets

<b>Overrepresented Processes</b>		<b>Underrepresented Processes</b>	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
biosynthetic process (GO:0009058)	2.61E-04	anatomical structure morphogenesis (GO:0009653)	4.26E-02
cellular amino acid metabolic process (GO:0006520)	5.32E-03	B cell mediated immunity (GO:0019724)	5.14E-03
cellular component biogenesis (GO:0044085)	4.24E-06	biological regulation (GO:0065007)	1.28E-04
cellular component organization or biogenesis (GO:0071840)	2.85E-03	cell communication (GO:0007154)	5.92E-05
cellular protein modification process (GO:0006464)	2.69E-02	cell differentiation (GO:0030154)	6.38E-03
metabolic process (GO:0008152)	8.60E-08	cell surface receptor signaling pathway (GO:0007166)	1.04E-03
mitochondrial transport (GO:0006839)	5.25E-03	cell-cell signaling (GO:0007267)	1.09E-05
mitochondrion organization (GO:0007005)	7.35E-04	complement activation (GO:0006956)	3.73E-02
nitrogen compound metabolic process (GO:0006807)	2.14E-02	cytokine-mediated signaling pathway (GO:0019221)	3.43E-02
nucleobase-containing compound metabolic process (GO:0006139)	7.53E-03	defense response to bacterium (GO:0042742)	2.99E-02
organelle organization (GO:0006996)	1.17E-03	developmental process (GO:0032502)	1.69E-02
oxidative phosphorylation (GO:0006119)	1.19E-02	G-protein coupled receptor signaling pathway (GO:0007186)	4.43E-03
porphyrin-containing compound metabolic process (GO:0006778)	3.88E-02	immune response (GO:0006955)	1.07E-05
primary metabolic process (GO:0044238)	4.39E-07	immune system process (GO:0002376)	1.11E-02
protein acetylation (GO:0006473)	2.13E-02	intracellular signal transduction (GO:0035556)	4.44E-02
protein metabolic process (GO:0019538)	2.76E-07	macrophage activation (GO:0042116)	4.28E-02
respiratory electron transport chain (GO:0022904)	4.99E-02	multicellular organismal process (GO:0032501)	2.72E-09
RNA metabolic process (GO:0016070)	1.12E-03	muscle organ development (GO:0007517)	4.56E-02
rRNA metabolic process (GO:0016072)	9.50E-05	nervous system development (GO:0007399)	4.51E-05
translation (GO:0006412)	6.34E-13	neurological system process (GO:0050877)	8.23E-08
tRNA aminoacylation for protein translation (GO:0006418)	4.23E-05	regulation of biological process (GO:0050789)	1.35E-04
tRNA metabolic process (GO:0006399)	1.08E-03	response to biotic stimulus (GO:0009607)	1.05E-02
		response to stimulus (GO:0050896)	1.76E-07
		sensory perception (GO:0007600)	6.72E-06
		sensory perception of chemical stimulus (GO:0007606)	8.01E-06
		sensory perception of smell (GO:0007608)	3.82E-04
		signal transduction (GO:0007165)	7.25E-05
		single-multicellular organism process (GO:0044707)	3.95E-09
		synaptic transmission (GO:0007268)	6.12E-04
		system development (GO:0048731)	4.36E-05
		system process (GO:0003008)	9.43E-08

**Supplementary Table 4.5:** Significant biological processes in HPV(+) OPSCC, through initial identification of significantly dysregulated genes

<b>Overrepresented Processes</b>		<b>Underrepresented Processes</b>	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
apoptotic process (GO:0006915)	2.26E-02	multicellular organismal process (GO:0032501)	3.02E-07
biosynthetic process (GO:0009058)	9.61E-03	single-multicellular organism process (GO:0044707)	3.22E-05
cell communication (GO:0007154)	3.46E-02	anatomical structure morphogenesis (GO:0009653)	9.70E-04
cell death (GO:0008219)	1.56E-02	biological regulation (GO:0065007)	4.86E-03
cellular defense response (GO:0006968)	2.98E-02	cell communication (GO:0007154)	6.93E-03
cellular process (GO:0009987)	2.07E-03	cell surface receptor signaling pathway (GO:0007166)	2.51E-02
cytokine-mediated signaling pathway (GO:0019221)	3.13E-02	cell-cell signaling (GO:0007267)	3.74E-02
death (GO:0016265)	1.56E-02	developmental process (GO:0032502)	3.97E-02
developmental process (GO:0032502)	1.61E-02	ectoderm development (GO:0007398)	4.54E-02
endoderm development (GO:0007492)	4.35E-03		
hemopoiesis (GO:0030097)	1.91E-02		
I-kappaB kinase/NF-kappaB cascade (GO:0007249)	7.30E-04		
immune response (GO:0006955)	1.33E-02		
immune system process (GO:0002376)	3.10E-02		
intracellular protein transport (GO:0006886)	3.48E-03		
JNK cascade (GO:0007254)	1.96E-04		
macrophage activation (GO:0042116)	2.29E-02		
MAPK cascade (GO:0000165)	4.94E-02		
metabolic process (GO:0008152)	4.71E-02		
mitosis (GO:0007067)	3.59E-02		
negative regulation of apoptotic process (GO:0043066)	3.16E-02		
phospholipid metabolic process (GO:0006644)	8.08E-03		
protein localization (GO:0008104)	4.21E-02		
protein transport (GO:0015031)	3.05E-03		
proteolysis (GO:0006508)	4.83E-02		
pyrimidine nucleobase metabolic process (GO:0006206)	1.96E-02		
regulation of sequence-specific DNA binding transcription factor activity (GO:0051090)	2.33E-02		
response to interferon-gamma (GO:0034341)	1.01E-04		

**Supplementary Table 4.6:** Significant biological processes in HPV(+) CESC, through initial identification of significantly dysregulated genes

<b>Overrepresented Processes</b>		<b>Underrepresented Processes</b>	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
antigen processing and presentation (GO:0019882)	1.02E-02	cell recognition (GO:0008037)	4.20E-04
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (GO:0002504)	2.60E-02	complement activation (GO:0006956)	1.80E-02
biological adhesion (GO:0022610)	5.47E-03	defense response to bacterium (GO:0042742)	5.68E-04
catabolic process (GO:0009056)	4.41E-02	gamete generation (GO:0007276)	3.98E-02
cell adhesion (GO:0007155)	5.47E-03	mitochondrion organization (GO:0007005)	3.13E-02
cell cycle (GO:0007049)	1.34E-05	multicellular organismal process (GO:0032501)	2.36E-04
cell proliferation (GO:0008283)	3.50E-02	muscle organ development (GO:0007517)	3.01E-02
cell-matrix adhesion (GO:0007160)	4.38E-02	neurological system process (GO:0050877)	6.59E-05
cellular component morphogenesis (GO:0032989)	7.20E-04	oxidative phosphorylation (GO:0006119)	2.31E-02
cellular component movement (GO:0006928)	4.87E-03	phagocytosis (GO:0006909)	1.15E-02
cellular component organization (GO:0016043)	9.28E-04	protein folding (GO:0006457)	3.39E-02
cellular component organization or biogenesis (GO:0071840)	9.55E-03	protein metabolic process (GO:0019538)	2.23E-02
cellular defense response (GO:0006968)	6.25E-03	reproduction (GO:0000003)	4.45E-02
cellular process (GO:0009987)	6.05E-04	response to biotic stimulus (GO:0009607)	2.95E-02
chromosome segregation (GO:0007059)	3.39E-04	RNA metabolic process (GO:0016070)	2.63E-02
cytokine-mediated signaling pathway (GO:0019221)	3.17E-02	sensory perception (GO:0007600)	2.85E-09
cytoskeleton organization (GO:0007010)	2.46E-04	sensory perception of chemical stimulus (GO:0007606)	1.64E-13
DNA metabolic process (GO:0006259)	1.31E-06	sensory perception of smell (GO:0007608)	2.48E-10
DNA recombination (GO:0006310)	9.02E-03	single-multicellular organism process (GO:0044707)	3.61E-04
DNA repair (GO:0006281)	2.41E-04	system process (GO:0003008)	9.41E-05
DNA replication (GO:0006260)	9.62E-05	translation (GO:0006412)	2.69E-03
ectoderm development (GO:0007398)	2.86E-02	tRNA metabolic process (GO:0006399)	3.09E-02
endoderm development (GO:0007492)	3.85E-02	Unclassified (UNCLASSIFIED)	1.02E-03
glycolysis (GO:0006096)	4.97E-02		
hemopoiesis (GO:0030097)	3.71E-02		
I-kappaB kinase/NF-kappaB cascade (GO:0007249)	2.86E-03		
immune system process (GO:0002376)	1.40E-02		
intracellular signal transduction (GO:0035556)	7.80E-04		
locomotion (GO:0040011)	5.76E-03		
MAPK cascade (GO:0000165)	2.73E-02		
metabolic process (GO:0008152)	2.73E-02		
mitosis (GO:0007067)	4.73E-02		
negative regulation of apoptotic process (GO:0043066)	2.59E-02		

nervous system development (GO:0007399)	3.41E-02
nucleobase-containing compound metabolic process (GO:0006139)	6.45E-03
organelle organization (GO:0006996)	2.77E-03
phosphate-containing compound metabolic process (GO:0006796)	4.57E-03
protein localization (GO:0008104)	4.47E-02
pyrimidine nucleobase metabolic process (GO:0006206)	4.60E-02
regulation of nucleobase-containing compound metabolic process (GO:0019219)	4.70E-02
regulation of transcription from RNA polymerase II promoter (GO:0006357)	1.28E-02
response to abiotic stimulus (GO:0009628)	3.44E-02
response to external stimulus (GO:0009605)	2.97E-02
response to interferon-gamma (GO:0034341)	7.20E-04
response to stress (GO:0006950)	3.88E-02

---

**Supplementary Table 4.7:** Significant biological processes in HPV(-) OPSCC, through initial identification of significantly dysregulated genes

<b>Overrepresented Processes</b>		<b>Underrepresented Processes</b>	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
biosynthetic process (GO:0009058)	6.28E-03	multicellular organismal process (GO:0032501)	6.57E-06
catabolic process (GO:0009056)	3.73E-04	single-multicellular organism process (GO:0044707)	8.26E-06
cellular amino acid catabolic process (GO:0009063)	1.69E-02	anatomical structure morphogenesis (GO:0009653)	4.40E-02
cellular amino acid metabolic process (GO:0006520)	7.68E-03	biological regulation (GO:0065007)	5.58E-03
cellular component biogenesis (GO:0044085)	7.83E-10	cell communication (GO:0007154)	6.49E-04
cellular component organization (GO:0016043)	1.92E-02	cell surface receptor signaling pathway (GO:0007166)	7.84E-03
cellular component organization or biogenesis (GO:0071840)	3.54E-06	cell-cell signaling (GO:0007267)	6.03E-03
cellular process (GO:0009987)	6.03E-03	developmental process (GO:0032502)	4.11E-03
cellular protein modification process (GO:0006464)	2.23E-02	ectoderm development (GO:0007398)	1.26E-02
coenzyme metabolic process (GO:0006732)	1.06E-02	G-protein coupled receptor signaling pathway (GO:0007186)	2.03E-02
generation of precursor metabolites and energy (GO:0006091)	1.27E-04	immune response (GO:0006955)	6.43E-04
glycolysis (GO:0006096)	3.08E-03	intracellular signal transduction (GO:0035556)	1.78E-02
metabolic process (GO:0008152)	7.42E-08	mesoderm development (GO:0007498)	6.93E-03
mitochondrial transport (GO:0006839)	9.70E-03	muscle organ development (GO:0007517)	1.71E-02
mitochondrion organization (GO:0007005)	1.42E-03	neurological system process (GO:0050877)	8.61E-05
nitrogen compound metabolic process (GO:0006807)	4.79E-04	regulation of biological process (GO:0050789)	9.29E-03
nuclear transport (GO:0051169)	4.99E-04	regulation of nucleobase-containing compound metabolic process (GO:0019219)	1.87E-02
nucleobase-containing compound metabolic process (GO:0006139)	1.77E-03	regulation of phosphate metabolic process (GO:0019220)	2.82E-02
nucleobase-containing compound transport (GO:0015931)	1.77E-02	response to stimulus (GO:0050896)	6.47E-05
organelle organization (GO:0006996)	2.89E-03	sensory perception (GO:0007600)	2.10E-05
phosphate-containing compound metabolic process (GO:0006796)	1.33E-02	sensory perception of chemical stimulus (GO:0007606)	5.99E-04
primary metabolic process (GO:0044238)	1.29E-07	sensory perception of smell (GO:0007608)	6.89E-03
protein complex assembly (GO:0006461)	7.60E-03	signal transduction (GO:0007165)	3.01E-03
protein complex biogenesis (GO:0070271)	7.87E-03	skeletal system development (GO:0001501)	4.63E-02
protein folding (GO:0006457)	2.43E-02	system development (GO:0048731)	4.50E-03
protein metabolic process (GO:0019538)	4.56E-06	system process (GO:0003008)	2.19E-05
protein methylation (GO:0006479)	1.57E-02	Unclassified (UNCLASSIFIED)	4.79E-03
protein targeting (GO:0006605)	1.45E-03		
regulation of translation (GO:0006417)	2.77E-03		
respiratory electron transport chain (GO:0022904)	1.45E-02		
RNA catabolic process (GO:0006401)	1.90E-02		
RNA metabolic process (GO:0016070)	6.67E-07		

rRNA metabolic process (GO:0016072)	4.54E-06
translation (GO:0006412)	7.88E-06
tricarboxylic acid cycle (GO:0006099)	1.08E-02
tRNA aminoacylation for protein translation (GO:0006418)	4.41E-02
tRNA metabolic process (GO:0006399)	2.82E-05

---

**Supplementary Table 4.8:** Significant biological processes in HPV(-) CESC, through initial identification of significantly dysregulated genes

<b>Overrepresented Processes</b>		<b>Underrepresented Processes</b>	
<i>Process Name (GO Number)</i>	<i>P-value</i>	<i>Process Name (GO Number)</i>	<i>P-value</i>
cellular component biogenesis (GO:0044085)	2.81E-02	B cell mediated immunity (GO:0019724)	2.76E-02
metabolic process (GO:0008152)	1.24E-04	biological regulation (GO:0065007)	2.27E-02
mitochondrion organization (GO:0007005)	1.00E-02	cell communication (GO:0007154)	4.75E-04
nucleobase-containing compound metabolic process (GO:0006139)	2.21E-02	cell differentiation (GO:0030154)	3.36E-02
organelle organization (GO:0006996)	2.36E-02	cell surface receptor signaling pathway (GO:0007166)	1.30E-03
oxidative phosphorylation (GO:0006119)	1.39E-02	cell-cell signaling (GO:0007267)	4.16E-04
primary metabolic process (GO:0044238)	8.46E-05	G-protein coupled receptor signaling pathway (GO:0007186)	1.23E-02
protein metabolic process (GO:0019538)	4.66E-04	immune response (GO:0006955)	2.97E-04
translation (GO:0006412)	7.18E-05	immune system process (GO:0002376)	3.31E-02
		multicellular organismal process (GO:0032501)	1.02E-05
		nervous system development (GO:0007399)	8.00E-04
		neurological system process (GO:0050877)	2.83E-05
		regulation of biological process (GO:0050789)	1.16E-02
		response to biotic stimulus (GO:0009607)	4.49E-02
		response to stimulus (GO:0050896)	7.50E-05
		sensory perception (GO:0007600)	7.50E-04
		sensory perception of chemical stimulus (GO:0007606)	3.38E-04
		sensory perception of smell (GO:0007608)	4.70E-03
		signal transduction (GO:0007165)	7.09E-04
		single-multicellular organism process (GO:0044707)	1.29E-05
		synaptic transmission (GO:0007268)	6.47E-03
		system development (GO:0048731)	6.52E-04
		system process (GO:0003008)	2.45E-05

# **Chapter 5: Conclusions**

In this dissertation, we set out to identify transcript-based biomarkers in HPV-related cancers. Human papillomavirus infection is a distinctive biomarker in cancer, due to its dual roles as a tumorigenic factor, as well as a positive biomarker for patient survival. Consequently, there is a demand for additional biomarkers to supplement the existing diagnostic role of HPV in the clinical setting. To do so, we developed a set of comprehensive bioinformatics tools to identify transcript-based biomarkers from RNA-seq expression data.

We first applied these bioinformatics techniques to HPV-related cancers in the head and neck and cervix, using data obtained from The Cancer Genome Atlas. In head and neck squamous cell carcinomas, we identified a novel set of miRNAs associated with overall survival in subtypes based on tumor source site; these miRNA biomarkers were also combined to create an expression-based survival signature that could accurately distinguish between high- and low-risk patients. Of note, the oropharyngeal cancer signature was able to differentiate patients based on risk even within the HPV(+) cohort, which can further the goal of personalized medicine in the treatment of oropharyngeal cancers. This signature was also validated in an independent dataset, using a different quantification technique, which indicates the robustness of the miRNA expression signature and its potential applicability within the clinical setting. When comparing the miRNA signatures to other subtypes of head and neck cancers, these miRNAs were determined to be subtype-specific, demonstrating the genomic heterogeneity between tumor source sites also extends to the miRNome. Consequently, the origin of the tumor should also be considered when determining course of treatment. This tissue speciation in terms of treatment modality has been previously observed in cervical cancer, as squamous cell carcinomas have better prognosis than adenocarcinomas and adenosquamous carcinomas. In spite of the genomic



variability, we were able to identify four distinct miRNAs related to overall survival that were prognostic in both squamous cell and adeno-type cervical cancers, and formulate an expression-based signature that was significant independently of tumor source site. This signature could not be validated in an independent sample cohort; however, the potential of a subtype-independent signature in cervical cancer shows that these results should not dissuade further research.

With the head and neck cancer cohort, we demonstrated that the techniques for identifying significant miRNA biomarkers associated with cancer diagnostic parameters could be extended beyond HPV-related cancers, as well as beyond survival. We obtained miRNA- and RNA-sequencing data for 30 different cancer types from TCGA and applied the bioinformatics pipelines to determine the relevance of miRNA expression to tumor formation, diagnostic staging parameters, and patient survival. The role of differential miRNA expression in various tumor types was also explored by combining correlation analysis with target prediction analysis within different tumor types. In addition to providing the static results of these analyses in a web-accessible database, we created a web server that could produce dynamic results for custom survival signature analysis and clustering analysis to classify the major cancer types. These tools are all publicly accessible at the website [www.oncomir.org](http://www.oncomir.org).

The underlying biological role of miRNA biomarkers in HPV-related cancers was then analyzed through a combination of target and correlation analysis integrated with pathway analysis. By identifying miRNAs dysregulated between HPV(+) and HPV(-) cohorts and subsequent targets, it was shown that the oncogenic aspect of HPV was supplemented by the miRNA-guided regulatory mechanism; HPV(-) tumors also demonstrated overrepresentation of similar biological pathways associated with tumor growth. However, the pathway analysis also indicated that HPV(-) tumors significantly disfavored biological processes that may prevent

metastasis. Such insights can guide further research into reason for poorer prognosis in HPV(-) tumors.

Through this dissertation, we have highlighted clinical and biological applications of miRNA biomarkers in cancer. The miRNA-mediated mechanisms for HPV-influenced tumor formation and survival are still under investigation, but the immediate applicability of miRNA expression levels in the diagnostic setting have been demonstrated. We have also demonstrated that the tools for identifying transcript biomarkers are applicable across all cancers, and have made the results of the analysis publicly available. In summary, tools for transcript biomarker identification have been developed, broadly applied, and produced actionable results for the research community at large.