# High-throughput bioinformatics approaches to understand gene expression regulation in head and neck tumors

by

Yanxiao Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2016

Doctoral Committee:

Associate Professor Maureen A. Sartor, Chair
Professor Thomas E. Carey
Assistant Professor Hui Jiang
Professor Ronald J. Koenig
Associate Professor Laura M. Rozek
Professor Kerby A. Shedden

I dedicate this thesis to my family.
For their unfailing love, understanding and support.

# ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Maureen Sartor for her guidance in my research and career development. She is a great mentor. She patiently taught me when I started new in this field, granted me freedom to explore and helped me out when I got lost. Her dedication to work, enthusiasm in teaching, mentoring and communicating science have inspired me to feel the excitement of research beyond novel scientific discoveries. I'm also grateful to have an interdisciplinary committee. Their feedback on my research progress and presentation skills is very valuable. In particular, I would like to thank Dr. Thomas Carey and Dr. Laura Rozek for insightful discussions on the biology of head and neck cancers and human papillomavirus, Dr. Ronald Koenig for expert knowledge on thyroid cancers, Dr. Hui Jiang and Dr. Kerby Shedden for feedback on the statistics part of my thesis.

I would like to thank all the past and current members of Sartor lab for making the lab such a lovely place to stay and work in. And for enormous help I received from them for my projects. In particular, I'd like to thank Yu-Hsuan Lin for developing the prototype pipeline for PePr, Lada Koneva, Shama Virani and Pelle Hall for identifying viral integration and RNA-seq mutation in the head and neck cancer project, and Chee Lee for the pathway analysis in the thyroid cancer project. I would also like to thank all of my friends here at University of Michigan for their company, support and inspirations. Ann Arbor is like my second hometown with a lot of cherishable memories.

Finally, I'm very grateful to my parents for encouraging and supporting me to chase my own dreams, to become who I am today. And to my dear Yuqi, who has always been a truthful friend and a loving company. Thank you so much for always having faith in me.

# TABLE OF CONTENTS

# LIST OF FIGURES

**Figure**

# LIST OF TABLES

**Table**

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **ALL** | acute lymphoblastic leukemia |
| **BAF** | B allele frequency |
| **ChIP-seq** | Chromatin immunoprecipitation followed by sequencing |
| **CML** | chronic meylogenous leukemia |
| **CNA** | copy number alteration |
| **CR** | combine replicates |
| **DBD** | DNA binding domain |
| **DMSO** | dimethyl sulfoxide |
| **EMT** | epithelial-mesemchymal transition |
| **ENCODE** | encyclopedia of DNA elements |
| **FA** | fanconi anemia |
| **FDR** | false discovery rate |
| **FISH** | fluorescence in situ hybridization |
| **FNA** | fine needle aspiration biopsy |
| **FTC** | follicular thyroid cancer |
| **GATK** | the genome analysis toolkit |
| **GEO** | gene expression omnibus |
| **GO** | gene ontology |
| **HNSCC** | head and neck squamous cell carcinomas |
| **HNC** | head and neck cancer |
| **HPV** | human papillomavirus |
| **IDR** | irreproducible discovery rate |
| **IP** | immunoprecipitation |
| **LRR** | log R ratio |
| **PCR** | polymerase chain reaction |
| **PePr** | peak-calling prioritization |
| **PPFP** | PAX8-PPARG fusion protein |
| **PSSM** | position specific score matrix |
| **PTC** | papillary thyroid carcinoma |
| **PTM** | post-translational modifications |
| **RNA-seq** | RNA sequencing |
| **Rb** | retinoblastoma |
| **SA** | separate analysis |
| **SCC** | squamous cell carcinomas |
| **SNP** | single nucleotide polymorphism |
| **TAG** | tumor associated gene |
| **TBHP** | tert-Butyl hydroperoxide |

| | |
|---|---|
| **TCGA** | the Cancer Genome Atlas |
| **TF** | transcription factor |
| **TMM** | trimmed mean of M values |
| **TSG** | tumor suppressor gene |
| **UM** | the University of Michigan |

# ABSTRACT

Cancer is defined as uncontrolled growth of abnormal cells bearing various molecular aberrations. With the aid of massively parallel DNA sequencing technology, we can now comprehensively characterize the genomic, epigenomic and transcriptomic landscapes of cancers. Subtypes of cancer are continually being uncovered, often by clustering expression profiles or determining driver mutations, the identification of which can be very important for prognosis and personalized treatment plans. The goal of this dissertation is to develop and apply bioinformatics algorithms to study subtypes of head and neck tumors. Using computational approaches, we both uncover new subtypes, and investigate the oncogenic mechanisms of driving molecular events in the tumor subtypes. My dissertation consists of three main chapters. In the first chapter, we present a software program (PePr) for conducting the differential binding analysis of replicated ChIP-seq data. PePr estimates the biological variation among samples and reports consistent changes across sample groups. We use PePr to characterize the difference in histone modifications between two human papillomavirus (HPV)-associated cancer cell lines and two non-HPV cell lines. In the second chapter, we identify two robust HPV(+) head and neck squamous cell carcinoma subtypes based on gene expression clustering. One subtype (HPV-KRT) shows more frequent genic viral integration and splicing of E6, and reduced viral oncogenic E6 activity. The HPV-KRT subtype also has more frequent copy number gains of chr3q, fewer losses of chr16q, and more PIK3CA mutations. These genomic changes could potentially lead to the differences in gene expression between the subtypes, including elevated immune response and mesenchymal differentiation in HPV-IMU subtype, and up-regulated keratinocyte differentiation and oxidation-reduction process

in HPV-KRT subtype. In the last chapter, we characterize the binding profile of a fusion oncogene, PPFP (fusion of PAX8/PPARG; observed in 30% of follicular thyroid cancer) using ChIP-seq data from a rat PPFP-transfected PCCL3 cell line. Our RNA-seq and ChIP-seq results suggest that PPFP regulates many pathways related to cancer, and a PPARG agonist, pioglitazone, may reverse the oncogenic effect of PPFP by altering oxidative stress. Altogether, we demonstrate how the integrative analysis of high-throughput data can guide subtype discovery and mechanistic research in cancer.

# CHAPTER I

# Introduction

## 1.1 Introduction

Cancer, which refers to any type of malignant tumor or neoplasm, is the second leading cause of death in the US, and is expected to exceed heart disease as the leading cause of death within the next few years (Rl et al., 2015). Approximately one in two men and one in three women will develop cancer during their lifetime (Howlader et al., 2015). Typically, cancer is characterized as uncontrolled division and growth of abnormal cells in a part of the body, and that has the potential to spread to other parts of the body. These abnormal cells grow out of control, metastasize or locally spread to other parts of the body, and gradually take over healthy tissues, causing the body to lose normal functions and thus finally leading to death.

Cancer is probably one of the most complicated diseases, and its complicated nature comes in many aspects. First, cancer can develop virtually anywhere in the body. Depending on the type(s) of cell at the location of origin, cancers can be classified as: carcinoma (epithelial cells), sarcoma (connective tissue), leukemia or lymphoma (hematopoietic cells), blastoma (embryonic tissue), etc. The fact that cancer can arise in various cell types and locations implies a consensus disease mechanism, whereas the dramatic difference in survival rates for different cancers illustrates its heterogeneous nature. Second, the causes of cancer can be a mixture of genetic and environmental factors, which cannot yet be fully disentangled. Known environmental risk factors include tobacco,

alcohol, obesity, viral infection, and radiation. Some people are genetically predisposed to develop certain types of cancer. For example, fanconi anemia (FA) is a rare genetic disease caused by a mutation in a cluster of proteins involved in DNA repair. The relative risk of cancer in FA is exceedingly high compared to the general population (Alter, 2014). Because oncogenesis is inherently a chronic and sporadic process, it is extremely difficult to know which of the factors directly causes cancer, thus compounding disease prevention.

A third reason is that at the molecular level, cancer is an orchestra of dysregulation of multiple pathways. As reviewed in (Hanahan and Weinberg, 2000), transformation of normal cells to cancer involves at least the following changes: (1) growth signal autonomy (2) apoptosis evasion, (3) unbounded replicative potential, (4) sustained angiogenesis, and (5) tissue invasion (metastasis). More recently, reprogramming of metabolism and immune evasion and were added as essential components of cancer (Hanahan and Weinberg, 2011). Furthermore, an enormous amount of genome instability is observed in cancer, which may be the culprit for cancer cells to evolve and acquire all of the aforementioned changes toward successful transformation (Hanahan and Weinberg, 2011).

Due to its formidable complexity, cancer has been perceived as incurable since its first appearance in historical records centuries ago. This is no longer true, however, owing to advances and innovations in treatments such as surgery, radiation and chemotherapy, and more importantly, biologic understanding of cancer. Tremendous progress has been made over the last few decades, giving birth to new cancer drugs that target specific molecules and pathways, which hopefully improve prognosis and reduce side effects. Driven by this promise of "precision medicine", the field is marching into the characterization of the molecular landscape of diverse cancers. Particularly with the aid of high-throughput biomedical technologies, it is now possible to make unbiased genome-wide high-throughput interrogations to begin to unveil the complex mysteries of cancer.

Since the introduction of DNA microarrays and subsequently second-generation massively-

parallel sequencing, huge amounts of data have been generated. These two powerful tools enable a repertoire of microarray-based and/or sequencing-based measurements of cellular states. These cellular states include the genome (the complete sequence of a sample, containing information such as single nucleotide polymorphisms (SNP), translocations and copy number alterations (CNAs)), transcriptome (expression levels of all genes), methylome (genome-wide methylation status of cytosines), and regulome (genome-wide binding sites of transcription factors and histone modifications). Thousands to millions of measurements are generated simultaneously, giving a data-intensive snapshot of the cellular states in any given sample. These high-dimensional data comprise a magnificent tool to study complex problems such as cancer biology. However, there is no free lunch. Many challenges remain to accurately analyze such high-dimensional data and to integrate several datasets to make insightful, clinically relevant inferences.

This dissertation focuses on the analysis of high-throughput molecular data from cancers that histologically originated from the head and neck, and an effort to develop a novel bioinformatics tool to analyze one type of these datasets. In this chapter, I will introduce the sequencing technologies used in this dissertation, and review the discoveries of characteristic cellular states of cancer revealed by high-throughput bioinformatics approaches, with an emphasis on head and neck tumors. Finally, I will briefly summarize the contribution of the following chapters.

## 1.2 Background

### 1.2.1 High-throughput technologies

DNA microarrays remain the technology of choice for large-scale gene expression profiling even now due to their cheap price, standardized analysis pipeline and small data volume. Microarrays have many applications beyond profiling transcriptomes, including array comparative genomic hybridization (quantify CNA) and SNP-arrays (genotype SNPs and CNA). SNP-arrays are a type of DNA microarray originally designed to genotype thousands of SNPs across the human genome.

Surprisingly after a decade of development, this platform's application has expanded to the detection of large-scale copy number variations and loss-of-heterozygosity, which is very common in cancer cells (LaFramboise, 2009). Microarrays, however, suffer from many limitations, despite their utility in clinical research. For example, they can only measure genes/probes that are included in the microarray, and their signal is sensitive to noise and various experimental conditions thus causing many of the observations to be not readily reproducible. As a result, deep sequencing based approaches were developed and are now preferred over microarray-based approaches.

Second-generation deep sequencing technologies can sequence millions of DNA molecules in parallel. It is widely applied to genome resequencing, transcriptome profiling (RNA-seq), DNA-protein interactions (ChIP-seq) and epigenome characterization (bisulfite-seq and ChIP-seq). RNA-seq and ChIP-seq are two essential tools to learn the regulatory patterns of gene expression. A typical analysis of both types of datasets is multilayered: the sequences are first mapped to a reference genome, and then the mapped reads are counted over certain features (gene, enhancer, or any genomic window). The resulting read counts are normalized and quantitatively compared among individuals or between sample groups. Because the read counts are discrete numbers, they are often modeled with Poisson or negative binomial distribution to account for additional variance. There have been many challenges for each of these steps, and therefore dozens of software tools have been developed. For instance, most sequencing datasets suffer from small sample size, thus small-sample inference and variance stabilization strategies are especially important. Common strategies for variance stabilization include borrowing information (variance) from genes that have similar expression levels. Despite these sophisticated and computationally intensive analysis steps, RNA-seq has many advantages over expression microarrays. For example, RNA-seq has a higher dynamic range of signals, is more sensitive to lowly expressed genes, and has expanded capabilities to discover and quantify unannotated transcripts and to detect expressed mutations (Wang et al., 2009). ChIP-seq, short for chromatin immunoprecipitation followed by sequencing,

uses antibodies to enrich DNA sequences from binding sites of a specific protein (transcription factor or post-transcriptionally modified histone) for sequencing. After the sequencing reads are aligned to the reference genome, reads will pile up at binding sites. The primary task of ChIP-seq analysis is to identify these binding sites by determining regions that are statistically enriched for reads compared to the background noise. A typical program would model the read counts as Poisson or negative binomial distribution and test if the mean of one genomic window is greater than that of the background. The signal strength for each binding site varies significantly, and the signal-to-noise ratio is highly dependent on the efficiency of the antibody, thus making it difficult to determine the optimal threshold and to separate the weak binding sites from the background noise. The analysis is further complicated by the fact that there are three types of peaks (sharp, broad and mixed) (Park, 2009). Typically, a ChIP-seq study of sufficient coverage (>50 million reads) can uncover tens to hundreds of thousands of binding sites of a protein of interest. In addition, a substantial portion of the transcription factor binding sites and histone marks are specific to cell types or developmental stages, making differential binding analysis a necessary yet challenging task.

Thousands of terabytes of sequencing data are being generated every year, presenting challenges for storage, computational power and novel analysis algorithms. Nevertheless, with the aid of this technology, we can now routinely conduct genomic, epigenomic and transcriptional profiling to explore the complex regulatory pathways in human biology and cancer.

### 1.2.2 Genomic aberrations and instability in cancer

Genomic instability is a hallmark of cancer, and can be manifested at different levels ranging from as large as whole genome copy number alterations, to as tiny as single base substitutions. The most profound change (even visible under a microscope) in cancer is aneuploidy, i.e., the amplification/deletion of whole chromosomes. Because it is a common feature of many cancers,

there has been an active area of research to identify its causes and roles in tumorigenesis (Hanks et al., 2004; Schvartzman et al., 2010). In addition, focal copy number alterations (CNA) are equally prevalent. It is estimated that in a typical cancer sample, 25% of the genome has arm-level CNAs and 10% has focal CNAs, with 2% overlap (Beroukhim et al., 2010). There are some arm-level "hotspot" CNAs, such as gains on 20q, 1q, 3q, 5p, 7q and 17q, and losses on 3p, 4q, 13q, 17p and 18q, but these recurrent changes are not necessarily present in all types of cancer (Baudis, 2007). The large size of arm-level CNAs makes it difficult to identify the essential genes for cancer transformation and outgrowth, however their recurrent nature across cancers suggests that some additive effects may confer the cancer cells growth advantages. By contrast, studies of the focal CNAs have pinpointed several oncogenes/tumor suppressors, suggesting that CNAs play critical roles in activating and inactivating oncogenes and tumor suppressors, respectively, by changing their expression dosages (Baudis, 2007; Eder et al., 2005; Lahortiga et al., 2007; Weir et al., 2007; Zender et al., 2006). Positively selected CNAs, which likely contribute to cancer progression, are suggested to recur across cancers at higher rates (Bignell et al., 2010; Kim et al., 2013). For example, a pan-cancer analysis found that the most frequent focal CNAs include MYC amplifications and CDKN2A/B deletions, each observed in 14% of all tumor samples (Beroukhim et al., 2010).

Another notable structural rearrangement event in cancer is chromosomal translocation, which leads to the juxtaposition of otherwise distant regulatory or coding DNA sequences between genes. Since their first discovery in the early 1980s, inter-chromosomal translocations have been routinely identified by guided approaches such as the chromosome banding analysis and fluorescence in situ hybridization (FISH), which were then complemented by the development of microarray and deep sequencing technologies due to their critical drawbacks (Mertens et al., 2015). Recently, more than 8,600 different fusion transcripts were reported by mining the transcriptomic data from 4,366 tumors, from 13 different neoplasm types, within the Cancer Genome Atlas (TCGA) network

(Yoshihara et al., 2015). Most of the events are passenger events (Mitelman et al., 2015). A few gene fusions that have clinical relevance and have been extensively studied revealed two major pathogenic mechanisms: (1) fusion of distant regulatory sequences to oncogenes leading to their elevated expression, for example, fusion of immunoglobulin gene loci to MYC (Leder et al., 1983), (2) creation of chimeric genes with abnormal activity (examples include chimeric BCR-ABL1 in chronic meylogenous leukemia (CML) (Shtivelman et al., 1985) and TCF3-PBX1 in acute lymphoblastic leukemia (ALL) (Kamps et al., 1990)). However, the oncogenic mechanisms of many more fusion genes (such as PAX8/PPARG in follicular thyroid cancer studied in this dissertation) remain unclear. Recurrent gene fusions are strongly correlated with tumor subtypes, making them ideal for diagnostic purposes in some types of cancer. Drugs that target gene fusions, such as imatinib (against BCR-ABL1), have significantly improved the survival and quality of life for patients (with CML, in the case of imatinib) (Druker, 2008).

The accumulation of single base substitutions and small insertions and deletions is also a common phenomenon in solid tumors. The source of mutations can be endogenous events, such as erroneous DNA replication and mismatch repair, or exogenous factors such chemical mutagens or radiation. On average, 33 to 66 genes show somatic mutations that alter their protein sequences in common solid tumors (Vogelstein et al., 2013). Lung tumors and melanomas, which are exposed to potent mutagens (cigarette and ultraviolet light), tend to harbor more mutations; tumors with defects in DNA repair also have extraordinarily high mutational loads (Gryfe and Gallinger, 2001). As with CNAs and gene fusions, the majority of the somatic mutations are passengers rather than drivers. The study of well-known oncogenes and tumor suppressor genes reveals nonrandom and highly characteristic patterns of mutations. That is, mutations in oncogenes tend to recur at the same amino acid positions, whereas tumor suppressor genes are often mutated by protein-truncating substitutions throughout their length. This rule can be used to accurately classify a driver gene as an oncogene or tumor suppressor gene. A total of 125 driver genes were designated by a

stringent rule described in Vogelstein et al. (2013). This list includes 71 tumor suppressor genes (*TP53, APC, RB1, WT1, etc*) and 54 oncogenes (*KRAS, BRAF, ID1, EGFR, PIK3CA, etc*). The mutated driver genes can be grouped into a few signaling pathways related to cell survival, cell fate determination and genome maintenance, all of which are vital for cancer progression. Progress made in characterizing cancer genome perturbations has spurred development of genome-based medicine that target certain activating oncogenes and pathways. A representative example is the use of EGFR kinase inhibitor to treat cancers with an EGFR mutation (Sharma et al., 2007). However, targeting tumor suppressors and oncogenes other than kinases, and avoiding the development of drug resistance, remain challenging tasks.

### 1.2.3 Epigenetic dysregulation in cancer: DNA methylation and histone modifications.

Beyond genomic instability, the epigenetic landscape is also profoundly altered in cancer. Epigenetic switches, including DNA methylation and histone modification, play pivotal roles in moderating nuclear structure, accessibility to DNA and gene activity. Tumor cells often undertake a massive global loss of DNA methylation, mainly in the gene body, intergenic regions and repetitive DNA sequences, but also frequently acquire hypermethylation at certain promoters and in CpG islands (Esteller, 2005; Feinberg and Tycko, 2004; Herman and Baylin, 2003). Silencing tumor-suppressor genes via promoter hypermethylation is an important mechanism in tumorigenesis. For instance, the cell cycle inhibitor CDKN2A, DNA repair genes MLH1, MGMT, and BRCA1, and dozens of other tumor-suppressor genes have been shown to be silenced by DNA hypermethyation in cancer (Esteller, 2007). In contrast, the cause and consequences of global hypomethylation is still poorly understood. There is evidence that active demethylation through a hemimethylated intermediate may play important role in DNA hypomethylation in cancer (Ehrlich, 2009). It has also been hypothesized that the hypomethylation could be linked to genomic instability, reactivation of transposable elements, and loss of imprinting (Esteller, 2005).

DNA wraps around histones and form nucleosomes, the basic unit for chromatin. There is a multitude of post-translational modifications (PTMs, including acetylation, methylation, phosphorylation and so on) at different amino acid locations of the tails of histone proteins (H2A, H2B, H3 and H4). Each histone modification can be dynamically added or removed by specific enzymes. Certain combinations of PTMs are often identified together; these signatures are called the "histone code", and serve as important platforms for controlling cellular processes such as gene expression, DNA replication, and chromosome condensation (Strahl and Allis, 2000; Turner, 2000). Deregulation of histone-modifying enzymes is a hallmark of human cancer, suggesting that altered histone PTMs have important roles in cancer development. For instance, global loss of acetylation of histone H4 at lysine 16 (H4K16ac) and trimethylation of histone H4 at lysine 20 (H4K20me3) were first reported at repetitive DNA sequences in multiple primary tumors (Fraga et al., 2005). With the aid of ChIP-seq assays, dysregulation of various other histone PTMs are being discovered and linked to tumor progression and prognosis, including H3K4me3, H3K27me3, H3K9ac/me3, H3K56ac, etc (Füllgrabe et al., 2011). Dysregulated histone modifications are associated with changes in gene expression, silencing at heterochromatin domains, cell cycle checkpoint instability and impaired DNA repair (Füllgrabe et al., 2011). It has also become evident recently that histone modification and DNA methylation can be dependent on each other during normal development and tumorigenesis, mediated through the interaction between histone and DNA methyltransferases (Cedar and Bergman, 2009).

### 1.2.4 Discovery of molecular cancer subtypes: distinct etiology and prognosis.

Cancer is heterogeneous in the sense that every patient, and even every tumor is a different molecular entity. Traditionally, cancer subtypes were defined by morphological and histological features. However, this classification regime can be very inaccurate and cannot sufficiently capture the heterogeneity underlying the tumor. The fact that every tumor harbors distinct somatic

mutations, CNAs, and epigenetic changes has driven the treatment of cancer into the new era of precision medicine (Garay and Gray, 2012). Some driver mutations and gene fusions occur so frequently in cancers that they are used to define cancer subtypes. Adjuvant therapies with chemicals specifically targeting these aberrations greatly improve the outcome of the patients bearing the mutations or gene fusions. For instance, ABL1 translocation (in CML and ALL), EGFR mutation/amplification (Glioma and lung cancer) and BRCA1/2 mutation (Breast, ovarian and pancreatic cancer) define distinct tumor subtypes that can be successfully targeted by therapeutic agents (Druker and Guilhot, 2006; Fong et al., 2009; Iyer and Bharthuar, 2010). Cancer subtypes can also be defined using high-throughput data features such as transcription profiling of coding and non-coding RNAs, methylation patterns, and protein levels. Subtype discovery using mRNA expression patterns is widely used in most cancer types, particularly for large datasets generated by the Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network, 2011; The Cancer Genome Atlas Network, 2008; Taylor et al., 2010). The subtypes defined by transcriptional profiles in some cancers are suggestive of distinct subtype etiology and strongly associated with disease outcome (van't Veer and Bernards, 2008).

### 1.2.5 The biology of head and neck tumors

Head and neck tumors include several categories of cancers with different origins. The two main categories investigated in this dissertation are (1) head and neck squamous cell carcinomas (HNSCC) of the oral cavity, oropharynx and larynx, and (2) thyroid gland tumors.

HNSCC is the sixth most common non-skin cancer worldwide with 600,000 incidences each year and a five-year mortality rate of 37 to 62 percent (Ferlay et al., 2010; Vokes et al., 1993). HNSCC is about three times more common in men than women (Ferlay et al., 2010). The major known risk factors for HNSCC are tobacco, alcohol, and infection with high risk human papillomavirus (HPV). Extensive studies of HPV in cervical cancer (99% of which are caused by HPV) have

shown that the oncogenic potential of HPV can mainly be attributed to two early viral genes, E6 and E7. E6 leads to the degradation of tumor suppressor protein p53, and E7 sequesters retinoblastoma (Rb), which cooperatively suppress apoptosis and promote tumor cell growth and proliferation (Moody and Laimins, 2010). HNSCC tumors associated with HPV have distinct molecular profiles and better treatment responses than non-HPV tumors (Leemans et al., 2011). HPV associated tumors are dominated by activating mutations of the oncogene PIK3CA, loss of TRAF3, elevated expression of p16 (CDKN2A), and amplifications of E2F1. The former two genes point to aberrant activation of NF-$\kappa$B pathway, whereas the latter two are linked to cell cycle regulation. In contrast, common features in smoking-related HNSCCs are loss-of-function TP53 mutations, inactivation/truncation of CDKN2A, and frequent copy number alterations (The Cancer Genome Atlas Research Network, 2015). Transcription and methylation profiling also demonstrate numerous differences between HPV and non-HPV tumors. Clustering based on these genome-wide profiles in multiple studies has identified four to five distinct molecular subtypes that may be useful in guiding novel biomarker development (Chung et al., 2004; The Cancer Genome Atlas Research Network, 2015; Seiwert et al., 2014).

Thyroid cancer is the most common endocrine malignancy and is three times more frequent in women than men (Rl et al., 2015). The incidence of thyroid cancer has nearly doubled worldwide since 2000, possibly due to combination of more frequent use of sensitive diagnostic tools such as ultrasound and FNA (fine needle aspiration biopsy) and a true increase as a consequence from increased exposure to radiation (Pellegriti et al., 2013; Zevallos et al., 2015). Thyroid cancers are transformed from thyroid follicular or parafollicular cells and are well-differentiated in most cases. Papillary thyroid carcinoma (PTC) accounts for 80% of all thyroid tumors, and most PTCs have good prognosis after treatment. In contrast, a subset of follicular thyroid carcinomas (FTCs) are often more aggressive and less responsive to therapy. 60-70% of PTCs harbor one of a few aberrations, such as RET-rearrangements, or RAS or BRAF mutations (Vu-Phan and Koenig, 2014).

For a subset of FTC ($\sim$ 30%), a notable tumorigenic event is the fusion protein (PPFP) of two transcription factors, PPARG and PAX8 (Kroll, 2000). Several in vitro studies have shown that PPFP acts as an oncogene (Au et al., 2006; Espadinha et al., 2007; Gregory Powell et al., 2004). However, additional studies are required to uncover its interacting partners and the population of genes and pathways it regulates.

## 1.3 Dissertation overview

Dozens of high-throughput datasets have been generated in the public domain and here at the University of Michigan to begin to understand the complex molecular architecture of head and neck tumors. As reviewed in the previous sections, each tumor cell population has a combination of a few frequent genetic or epigenetic aberrations, which can be grouped to define subtypes of cancers. This is also called inter-tumor heterogeneity. Discovery of the cancer subtypes is crucial to decomposing the complex oncogenic pathways leading to cancer, and will lay the foundation for developing biomarkers and therapies that are tailored to a specific cancer subtype. The goal of this dissertation is to develop and apply bioinformatics algorithms to uncover the inter-tumor heterogeneity (subtypes) of head and neck tumors, and investigate the oncogenic mechanisms of the tumor subtypes.

In chapter II, we develop a novel pipeline to enable the differential binding analysis of replicated ChIP-seq data. ChIP-seq is a relatively new innovation that can be used to map the locations of post-transcriptional modifications (PTM) of histones, the dysregulation of which are frequently observed in cancer. Identification of differential binding sites of histone PTMs between cancer and normal tissue or between subtypes of cancer is a critical step to characterize their epigenetic differences. There is substantial individual difference between cancer samples, which is not the essential change associated with each subtype. To remove sample biases, biological replicates are often required for ChIP-seq studies. Although more than a dozen ChIP-seq software tools exist, at the

time of publication of our software, none of them were specifically designed to analyze data with replicates. Consequently, ChIP-seq datasets with replicates could only be processed with suboptimal methods. Inspired by the successful application of the negative binomial model for RNA-seq data, we developed a similar model for ChIP-seq data with several innovated optimization steps specific for ChIP-seq, such as window size estimation, signal normalization, variance stabilization through kernel smoothing, and artifact removal. The whole pipeline is called Peak calling Prioritization (PePr) pipeline. We demonstrated our superior performance by comparing PePr to existing approaches. We also applied our method on histone PTM data from HNSCC cell lines to successfully characterize the differences of H3K27me3 binding profiles between HNSCC subtypes defined by HPV status. PePr is also applicable to other DNA-seq datasets, such as affinity-based DNA methylation and hydroxymethylation data. All together, this pipeline will be a useful tool for characterizing the epigenomic landscapes of cancers and cancer subtypes.

Chapter III concentrates on subtype discovery and characterization of HPV(+) HNSCCs. HPV associated HNSCCs are less characterized than their non-HPV counterparts, due to their common exclusion, unknown status, or small numbers in relevant studies. Although HPV(+) patients overall have better prognosis, heterogeneity in terms of clinical outcomes and biology remains. De-escalated therapies have been proposed for HPV(+) patients to reduce unnecessary treatment-induced morbidity, however, heterogeneity within this patient subgroup has to be examined to distinguish aggressive tumors from easily treated ones before clinical decisions are made. In addition, the incidence of HPV(+) HNSCC has been steadily increasing in developed countries such as United States, whereas the number of HPV(-) HNSCC cases has decreased. Although the use of HPV vaccine is expected to reduce the prevalence of HPV(+) HNSCC caused by the most common high-risk HPV types, it will be many years before the effects are seen. Instead of just relying on the vaccines, the research on HPV(+) HNSCC tumors is still clinically important. In this chapter, we aimed to characterize the heterogeneity within HPV(+) HNSCCs by mining RNA-seq and SNP-

array data. We identified two robust HPV(+) HNSCC subtypes using gene expression-based clustering, and characterized the differences in transcriptional and genomic profiles (including copy number alterations and genic mutations) and HPV characteristics between the two subtypes. We found that one subtype (HPV-KRT) has more keratinization, viral integration events, and spliced E6*, and less full length E6 activity, immune activity, and chr16q deletions. HPV-KRT also has more frequent chr3q amplifications and PIK3CA mutations. Literature concerning all of these features except keratinization suggest a worse outcome for this subtype, thus we hypothesized that HPV-KRT would have a lower survival rate. This expected trend was observed in The Cancer Genome Atlas survival analysis, although it was not statistically significant. Our study provides valuable insight into the key genetic events that likely drive two different paths to (or stages of) oncogenesis of HPV(+) HNSCCs, which will be important for the development of new biomarkers and therapies for HPV(+) patients.

In Chapter IV, we investigated the oncogenic mechanism of the fusion protein PPFP, which is associated with a subset of follicular thyroid cancer (FTC). The fusion protein PPFP is a fusion of two transcription factors, PPARG and PAX8. As reviewed in the previous section, PPFP is one of the major genetic changes found in FTC (accounting for 30% of the cases). PPFP shows oncogenic effect in vitro (inducing cell division and repressing apoptosis) (Au et al., 2006; Espadinha et al., 2007; Gregory Powell et al., 2004), and may be a potential therapeutic target for PPFP-associated FTC. Pioglitazone, a PPARG agonist, shows strong therapeutic effect in a mouse model of PPFP FTC, significantly shrinking the primary tumor and preventing metastasis (Dobson et al., 2011). However, the oncogenic function of PPFP and therapeutic effect of pioglitazone were poorly understood. In this chapter, we characterized the genomic binding sites of PPFP using ChIP-seq and showed that PPFP retains the binding ability of both original transcription factors, PPARG and PAX8, in the rat PPFP transfected PCCL3 cell line. Combined with RNA-seq data, we showed that PPFP binds to and regulates the expression of genes involved in multiple cancer-related pro-

cesses. PPFP binds to adipocyte genes in preference to macrophage genes, and the adipogenic effect was greatly enhanced in the presence of pioglitazone. In addition, enlightened by RNA-seq discovery, we designed experiments to confirm that PPFP induces oxidative stress in thyroid cells and pioglitazone further increases susceptibility to oxidative stress, which may eventually lead to cell death. Our data highlights the complexity of PPFP as a fusion transcription factor and various ways it regulates thyroid oncogenesis.

## Bibliography

Alter, B. P. Fanconi anemia and the development of leukemia. *Best practice & research. Clinical haematology*, 27(3-4):214–21, 2014.

Au, A. Y. M., McBride, C., Wilhelm, K. G., Koenig, R. J., Speller, B., Cheung, L., Messina, M., Wentworth, J., Tasevski, V., Learoyd, D., Robinson, B. G. and Clifton-Bligh, R. J. PAX8-peroxisome proliferator-activated receptor gamma (PPARgamma) disrupts normal PAX8 or PPARgamma transcriptional function and stimulates follicular thyroid cell growth. *Endocrinology*, 147(1):367–76, 2006.

Baudis, M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC cancer*, 7:226, 2007.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M.-S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R. and Meyerson, M. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, 2010.

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., Buck, G., Chen, L., Beare, D., Latimer, C., Widaa, S., Hinton, J., Fahey, C., Fu, B., Swamy, S., Dalgliesh,

G. L., Teh, B. T., Deloukas, P., Yang, F., Campbell, P. J., Futreal, P. A. and Stratton, M. R. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–8, 2010.

Cedar, H. and Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, 2009.

Chung, C. H., Parker, J. S., Karaca, G., Wu, J., Funkhouser, W. K., Moore, D., Butterfoss, D., Xiang, D., Zanation, A., Yin, X., Shockley, W. W., Weissler, M. C., Dressler, L. G., Shores, C. G., Yarbrough, W. G. and Perou, C. M. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell*, 5(5):489–500, 2004.

Dobson, M. E., Diallo-Krou, E., Grachtchouk, V., Yu, J., Colby, L. A., Wilkinson, J. E., Giordano, T. J. and Koenig, R. J. Pioglitazone induces a proadipogenic antitumor response in mice with PAX8-PPARgamma fusion protein thyroid carcinoma. *Endocrinology*, 152(11):4455–65, 2011.

Druker, B. and Guilhot, F. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *The New England Journal of Medicine*, 355:2408–17, 2006.

Druker, B. J. Translation of the Philadelphia chromosome into therapy for CML. *Blood*, 112(13):4808–17, 2008.

Eder, A. M., Sui, X., Rosen, D. G., Nolden, L. K., Cheng, K. W., Lahad, J. P., Kango-Singh, M., Lu, K. H., Warneke, C. L., Atkinson, E. N., Bedrosian, I., Keyomarsi, K., Kuo, W.-l., Gray, J. W., Yin, J. C. P., Liu, J., Halder, G. and Mills, G. B. Atypical PKCiota contributes to poor prognosis through loss of apical-basal polarity and cyclin E overexpression in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12519–24, 2005.

Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics*, 1(2):239–59, 2009.

Espadinha, C., Cavaco, B. M. and Leite, V. PAX8PPARgamma stimulates cell viability and modulates expression of thyroid-specific genes in a human thyroid cell line. *Thyroid : official journal of the American Thyroid Association*, 17(6):497–509, 2007.

Esteller, M. Aberrant Dna Methylation As a Cancer-Inducingmechanism. *Annual review of pharmacology and toxicology*, 45:629–56, 2005.

Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews. Genetics*, 8(4):286–298, 2007.

Feinberg, A. P. and Tycko, B. The history of cancer epigenetics. *Nature reviews. Cancer*, 4(2):143–153, 2004.

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C. and Parkin, D. M. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, 127(12):2893–2917, 2010.

Fong, P. C., Boss, D. S., Yap, T. A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M. J., Ashworth, A., Carmichael, J., Kaye, S. B., Schellens, J. H. and de Bono, J. S. Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers. *New England Journal of Medicine*, 361(2):123–134, 2009.

Fraga, M. F., Ballestar, E., Villar-Garea, A., Boix-Chornet, M., Espada, J., Schotta, G., Bonaldi, T., Haydon, C., Ropero, S., Petrie, K., Iyer, N. G., Pérez-Rosado, A., Calvo, E., Lopez, J. A., Cano, A., Calasanz, M. J., Colomer, D., Piris, M. A., Ahn, N., Imhof, A., Caldas, C., Jenuwein, T. and Esteller, M. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature genetics*, 37(4):391–400, 2005.

Füllgrabe, J., Kavanagh, E. and Joseph, B. Histone onco-modifications. *Oncogene*, 30(31):3391–3403, 2011.

Garay, J. P. and Gray, J. W. Omics and therapy A basis for precision medicine. *Molecular Oncology*, 6(2):128–139, 2012.

Gregory Powell, J., Wang, X., Allard, B. L., Sahin, M., Wang, X.-L., Hay, I. D., Hiddinga, H. J., Deshpande, S. S., Kroll, T. G., Grebe, S. K. G., Eberhardt, N. L. and McIver, B. The PAX8/PPARgamma fusion oncoprotein transforms immortalized human thyrocytes through a mechanism probably involving wild-type PPARgamma inhibition. *Oncogene*, 23(20):3634–41, 2004.

Gryfe, R. and Gallinger, S. Microsatellite instability, mismatch repair deficiency, and colorectal cancer. *Surgery*, 130(1):17–20, 2001.

Hanahan, D. and Weinberg, R. a. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.

Hanks, S., Coleman, K., Reid, S., Plaja, A., Firth, H., Fitzpatrick, D., Kidd, A., Méhes, K., Nash, R., Robin, N., Shannon, N., Tolmie, J., Swansbury, J., Irrthum, A., Douglas, J. and Rahman, N. Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. *Nature genetics*, 36(11):1159–61, 2004.

Herman, J. G. and Baylin, S. B. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med*, 349(21):2042–2054, 2003.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E. and Cronin, K. e. SEER Cancer Statistics Review, 1975-2012. 2015.

Iyer, R. and Bharthuar, A. A review of erlotinib an oral, selective epidermal growth factor receptor tyrosine kinase inhibitor. *Expert Opinion on Pharmacotherapy*, 2010.

Kamps, M. P., Murre, C., Sun, X.-h. and Baltimore, D. A new homeobox gene contributes the DNA binding domain of the t(1;19) translocation protein in pre-B all. *Cell*, 60(4):547–555, 1990.

Kim, T.-M., Xi, R., Luquette, L. J., Park, R. W., Johnson, M. D. and Park, P. J. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome research*, 23(2):217–27, 2013.

Kroll, T. G. PAX8-PPARgamma 1 Fusion in Oncogene Human Thyroid Carcinoma. *Science*, 289(5483):1357–1360, 2000.

LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, 37(13):4181–93, 2009.

Lahortiga, I., De Keersmaecker, K., Van Vlierberghe, P., Graux, C., Cauwelier, B., Lambert, F., Mentens, N., Beverloo, H. B., Pieters, R., Speleman, F., Odero, M. D., Bauters, M., Froyen, G., Marynen, P., Vandenberghe, P., Wlodarska, I., Meijerink, J. P. P. and Cools, J. Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nature genetics*, 39(5):593–5, 2007.

Leder, P., Battey, J., Lenoir, G., Moulding, C., Murphy, W., Potter, H., Stewart, T. and Taub, R. Translocations among antibody genes in human cancer. *Science*, 222(4625):765–771, 1983.

Leemans, C. R., Braakhuis, B. J. M. and Brakenhoff, R. H. The molecular biology of head and neck cancer. *Nature reviews. Cancer*, 11(1):9–22, 2011.

Mertens, F., Johansson, B., Fioretos, T. and Mitelman, F. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, 15(6):371–381, 2015.

Mitelman, F., Johansson, B. and Mertens, F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2015). 2015.

Moody, C. a. and Laimins, L. a. Human papillomavirus oncoproteins: pathways to transformation. *Nature Reviews Cancer*, 10(8):550–560, 2010.

Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80, 2009.

Pellegriti, G., Frasca, F., Regalbuto, C., Squatrito, S. and Vigneri, R. Worldwide Increasing Incidence of Thyroid Cancer: Update on Epidemiology and Risk Factors. *Journal of Cancer Epidemiology*, 2013:1–10, 2013.

Rl, S., Kd, M. and Jemal, A. Cancer statistics , 2015 . *CA Cancer J Clin*, 65(1):21254, 2015.

Schvartzman, J.-M., Sotillo, R. and Benezra, R. Mitotic chromosomal instability and cancer: mouse modelling of the human disease. *Nature reviews. Cancer*, 10(2):102–15, 2010.

Seiwert, T. Y., Zuo, Z., Keck, M. K., Khattri, a., Pedamallu, C. S., Stricker, T. P., Brown, C. D., Pugh, T. J., Stojanov, P., Cho, J., Lawrence, M., Getz, G., Bragelmann, J., DeBoer, R., Weichselbaum, R. R., Langerman, a., Portugal, L. D., Blair, E. a., Stenson, K. M., Lingen, M. W., Cohen, E. E., Vokes, E. E., White, K. P. and Hammerman, P. S. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. *Clinical Cancer Research*, 21(3):632–641, 2014.

Sharma, S. V., Bell, D. W., Settleman, J. and Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nature reviews. Cancer*, 7(3):169–81, 2007.

Shtivelman, E., Lifshitz, B., Gale, R. P. and Canaani, E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature*, 315(6020):550–554, 1985.

Strahl, B. D. and Allis, C. D. The language of covalent histone modifications. *Nature*, 403(6765):41–5, 2000.

Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., Arora, V. K., Kaushik, P., Cerami, E., Reva, B., Antipin, Y., Mitsiades, N., Landers, T., Dolgalev, I., Major, J. E., Wilson, M., Socci, N. D., Lash, A. E., Heguy, A., Eastham, J. A., Scher, H. I., Reuter, V. E., Scardino, P. T., Sander, C., Sawyers, C. L. and Gerald, W. L. Integrative genomic profiling of human prostate cancer. *Cancer cell*, 18(1):11–22, 2010.

The Cancer Genome Atlas Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–582, 2015.

Turner, B. M. Histone acetylation and an epigenetic code. *BioEssays*, 22(9):836–845, 2000.

van't Veer, L. J. and Bernards, R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452(7187):564–70, 2008.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr., L. A. and Kinzler, K. W. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, 2013.

Vokes, E. E., Weichselbaum, R. R., Lippman, S. M. and Hong, W. K. Head and Neck Cancer. *New England Journal of Medicine*, 328(3):184–194, 1993.

Vu-Phan, D. and Koenig, R. J. Genetics and epigenetics of sporadic thyroid cancer. *Molecular and cellular endocrinology*, 386(1-2):55–66, 2014.

Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.

Weir, B. A., Woo, M. S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W. M., Province, M. A., Kraja, A., Johnson, L. A., Shah, K., Sato, M., Thomas, R. K., Barletta, J. A., Borecki, I. B., Broderick, S., Chang, A. C., Chiang, D. Y., Chirieac, L. R., Cho, J., Fujii, Y., Gazdar, A. F., Giordano, T., Greulich, H., Hanna, M., Johnson, B. E., Kris, M. G., Lash, A., Lin, L., Lindeman, N., Mardis, E. R., McPherson, J. D., Minna, J. D., Morgan, M. B., Nadel, M., Orringer, M. B., Osborne, J. R., Ozenberger, B., Ramos, A. H., Robinson, J., Roth, J. A., Rusch, V., Sasaki, H., Shepherd, F., Sougnez, C., Spitz, M. R., Tsao, M. S., Twomey, D., Verhaak, R. G., Weinstock, G. M., Wheeler, D. A., Winckler, W., Yoshizawa, A., Yu, S., Zakowski, M. F., Zhang, Q., Beer, D. G., Wistuba II, Watson, M. A., Garraway, L. A., Ladanyi, M., Travis, W. D., Pao, W., Rubin, M. A., Gabriel, S. B., Gibbs, R. A., Varmus, H. E., Wilson, R. K., Lander, E. S. and Meyerson, M. Characterizing the cancer genome in lung adenocarcinoma. *Nature*, 450(7171):893–898, 2007.

Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H. and Verhaak, R. G. W. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, 34(37):4845–54, 2015.

Zender, L., Spector, M. S., Xue, W., Flemming, P., Cordon-Cardo, C., Silke, J., Fan, S.-T., Luk, J. M., Wigler, M., Hannon, G. J., Mu, D., Lucito, R., Powers, S. and Lowe, S. W. Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell*, 125(7):1253–67, 2006.

Zevallos, J. P., Hartman, C. M., Kramer, J. R., Sturgis, E. M. and Chiao, E. Y. Increased thyroid cancer incidence corresponds to increased use of thyroid ultrasound and fine-needle aspiration: a study of the Veterans Affairs health care system. *Cancer*, 121(5):741–6, 2015.

# CHAPTER II

# PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data

## 2.1 Introduction

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) is the standard technique to identify the genome-wide occurrences of transcription factor (TF) binding sites and histone modifications in vivo. Over the past few years, there has been tremendous development of analysis methods for ChIP-seq data, with tens of 'peak finders' published (Blahnik et al., 2010; Boyle et al., 2008; Fejes et al., 2008; Jothi et al., 2008; Kornacker et al., 2012; Qin et al., 2010; Rashid et al., 2011; Rozowsky et al., 2009; Song and Smith, 2011; Valouev et al., 2008; Wang et al., 2013; Xu et al., 2010; Zang et al., 2009; Zhang et al., 2008). Over this course,several characteristics of ChIP-seq datasets, such as enrichment profile features (peak width, signal-to-noise ratio and location relative to genomic features) of different types of TFs and histone modifications, sources of artifacts and the commonly observed statistical distributions of read counts, have been gradually revealed (Park, 2009; Pepke et al., 2009; Rye et al., 2011). As sequencing cost decreases, use of biological replicates is emerging and may eventually become the standard practice for ChIP-seq studies. Most of the Encyclopedia of DNA Elements (ENCODE) consor-

The work presented in Chapter II is published as **Zhang Y***, Lin YH*, Johnson TD, Rozek LS, Sartor MA. "PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data." *Bioinformatics*. 2014 Sep 15;30(18):2568-75. doi: 10.1093/bioinformatics/btu372 (*equal contribution)

tium data were performed in duplicate (Landt et al., 2012). Furthermore, as researchers shift from performing ChIP-seq experiments that address mechanistic questions to those that hypothesize differential and/or context-specific binding in a disease, treatment or epidemiologic setting, the use of replicates to account for individual variability becomes crucial. We expect that this will lead to more analyses comparing a group of ChIP samples with a group of controls, or two groups of ChIP samples, with or without controls, run under different experimental conditions. ChIP-seq peak finders that perform direct group comparisons within the peak-calling pipeline are currently lacking. When biological replicates are available, researchers may choose to combine the replicates (CR) in each group and run one-ChIP-versus- one-control analysis to identify all possible peaks. Alternatively, they can pair ChIP and control samples, conduct a separate analysis (SA) for each pair and then stipulate rules to combine the peak-finding results, such as requiring the peaks to be found in all pairwise comparisons. The CR approach is often used in TF ChIP-seq studies to identify all possible binding sites. However, if the goal is to find consistent binding among replicates, many false positives may occur where binding is present in only one or a subset of the samples. The SA approach is more sophisticated in the sense that it does not lose all information regarding sample-to-sample variability and is more applicable to experiments that have a natural pairing of samples. However, because it evaluates peaks for each replicate separately, the effects of false negatives across replicates may become compounded. Thus, the SA approach is more likely to miss moderate, yet consistent, differences in binding. When there is no inherent pairing between test and control samples, as often occurs with differential binding analyses, the SA approach may call a peak because in each one-versus-one analysis, one sample has greater enrichment than its paired sample. Yet, if the pairs were constructed differently, some 'peaks' may no longer exist. An alternative approach is the irreproducible discovery rate (IDR) (Landt et al., 2012) approach recommended by ENCODE. IDR can be considered a sophisticated CR approach, which assesses the consistency of peak rankings in replicates to find an optimum significance cutoff for determin-

ing the final peak list. Correctly modeling the variation among samples in gene expression studies when testing for differential expression has been shown to be of great importance (Anders and Huber, 2010; Robinson et al., 2010; Sartor et al., 2006). For RNA-Seq analysis, several methods [for example, edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010)] use a negative binomial distribution instead of a Poisson distribution to capture the extra variance among replicates. These approaches can be used with ChIP-seq data; however, they do not perform the first several steps of the ChIP-seq analysis pipeline nor do they take advantage of local chromosomal information. An exact negative binomial test (diffReps) was recently introduced for ChIP-seq data and compared with edgeR and DESeq using two histone modification datasets (Shen et al., 2013). Other approaches to identify differential binding with replicates include the R packages DiffBind (Ross-Innes et al., 2012) and DBChIP (Liang and Keles, 2012a); although these programs take into account sample variation, they rely on other peak callers to generate peak sets for each individual sample first and conduct analysis on the candidate regions that fall within the peak sets.

Here, we introduce a ChIP-seq peak-finding and prioritization (PePr) pipeline that can analyze either a group of ChIP-seq samples together with controls or compare two groups of ChIP-seq samples, with or without controls. PePr uses a sliding window approach and models read counts across replicates and between groups with a local negative binomial model. Genomic regions with less variable read counts across replicates are ranked more favorably than regions with greater variability, thus prioritizing consistently enriched regions. We tested PePr on ChIP-seq data for activating transcription factor 4 (ATF4) (Han et al., 2013), seven ENCODE TF datasets and one histone modification data- set (H3K27 tri-methylation), and compared the performance of PePr to several ChIP-seq methods representing different statistical models and using different sources of information: MACS (Zhang et al., 2008),MACS2 and SPP (Kharchenko et al., 2008) with IDR (Landt et al., 2012), ZINBA (Rashid et al., 2011), SICER (Zang et al., 2009), diffReps (Shen et al., 2013), DiffBind (Ross-Innes et al., 2012) and edgeR (Robinson et al., 2010).We show that

PePr performs favorably compared with the other tested approaches, prioritizing peaks that reflect stronger enrichment fold and higher consistency among samples.

## 2.2 Methods

### 2.2.1 Datasets

ATF4 data

ATF4 data were previously published (Han et al., 2013). Briefly, samples were obtained from mouse embryonic fibroblasts from transgenic mice after 8h treatment with tunicamycin, including three ATF4 wild type ChIP samples and three ATF4 knockout ChIP samples, which served as the controls. Data were obtained from Gene Expression Omnibus (GEO) with the accession number GSE35681.

ENCODE TF data

Neuron-restrictive silencer factor (NRSF), CCCTC-binding factor (CTCF), GA-binding protein (GABP), nuclear respiratory factor 1 (NRF1), structure maintenance of chromosome 3 (SMC3), upstream stimulatory factor 1 (USF1) and USF2 were downloaded from the UCSC collection of ENCODE ChIP-seq data.

H3K27me3 data

ChIP-seq using two human papillomavirus (HPV)-positive and two HPV-negative squamous cell carcinoma (SCC) cell lines were performed. Cell lines were cultured as previously described (Sartor et al., 2011), and chromatin immunoprecipitation using HistonePathTM (Active Motif) for the commercial-quality antibody pull downs for H3K27 tri-methylation and library preparation were performed by GenPathway, Inc. (part of Active Motif, Carlsbad, CA). DNA was amplified according to the Illumina ChIP-seq library construction protocol, and a region of 250350bp was excised from the preparative Agarose gel. Sequencing of the four immunoprecipitated samples and

four input DNA samples was performed at the University of Michigan DNA sequencing core using the Illumina HiSeq with 50 base single-end reads. Data were deposited in GEO with accession number GSE38629. Raw reads were aligned to hg19 using BWA (Li and Durbin, 2009) with the default parameters. The numbers of peaks called are listed in Table 2.1.

### 2.2.2 PePr algorithm

preprossing of data

*Removal of duplicated reads (optional)*: For every sample, PePr offers the option to remove the duplicated reads mapped at the same genomic location. Sometimes the same DNA fragments can be sequenced repeatedly due to PCR amplification or library preparation and are over-represented in the library. Assuming each piece of DNA in the genome has equal probability of being sequenced, then the occurrence of the same sequence read multiple times would be low and would depend on the sequencing depth. Therefore, PePr removes extra duplicated reads that are beyond the expected maximum at each genomic location. The maximum is calculated using a binomial distribution as specified in (Zhang et al., 2008).

*Fragment length estimation*: PePr estimates the shift size (half of the DNA fragment length) for each ChIP sample and shifts all reads to their 3' direction by this amount. For single-end ChIP-seq data, since the sequencing read length is shorter than the DNA fragment length, the cluster of forward-strand reads and that of reverse-strand reads at the binding sites show a phase lag. Properly shifting both strands of reads towards the center of the DNA fragment can improve the power and precision of detecting binding sites. For each chromosome, PePr shifts all the reads by several attempted shift sizes (starting at zero and increasing base by base), and counts the overlap between reads from forward and reverse strands. The shift size which maximizes the overlap is the optimum shift size. For every ChIP sample, the median of the estimated shift sizes from five chromosomes (chr1 to chr5) is calculated and used. We use the median from five chromosomes to balance speed and robustness against potential outliers; the shift sizes estimated from these five

chromosomes have been consistent for all datasets tested thus far. If control samples are included in the analysis, the average shift size derived from the ChIP samples are applied to the controls.

*Window size estimation*: To divide the genome into windows, a recommended window size is calculated as the estimated average width of the peaks, allowing PePr to optimally capture the reads in peak regions. To achieve this goal, we first divide the genome into non-overlapping 20bp bins. For each chromosome, the bin with the largest number of reads is chosen as the seed and extended to the flanking bins until a bin is reached which has less than 10% of the reads in the seed bin. The combined width of these bins is recorded. The abovementioned process is repeated 100 times after which the median of the widths is calculated. The median of widths for all chromosomes is the recommended window size. The genome is then divided into windows of the chosen (either recommended or user-specified) size that overlap by 50% and the number of reads in each window is multiplied by the normalization constant for each sample.

*Normalization*: The total number of reads often varies among samples, and the immunoprecipitation efficiency can also differ substantially among ChIP samples, which may artificially increase the variation among samples if unnormalized, raw read counts are used. Currently, PePr uses the Normalization of ChIP-seq (NCIS) method (Liang and Keles, 2012b) to normalize input (control) samples and a modified Trimmed Mean of M values (TMM) method to normalize ChIP samples (Robinson et al., 2010). First, PePr splits the genome into 1000bp bins. The mean of all ChIP libraries is used as the reference sample, towards which every sample will be normalized. For every input sample $i$ versus the reference $r$, let $n_{ig}$ and $n_{rg}$ be the number of reads in the $g^{th}$ genomic bin for the input and reference samples, respectively. The normalization factor for the input sample is calculated as

$$\hat{r} = \frac{\sum_{g \in B} n_{rg}}{\sum_{g \in B} n_{ig}} \tag{2.1}$$

where $B$ represents the background bins (in which no enrichment by the antibody exists). Let $n_g = n_{rg} + n_{ig}$, Given background bins are more likely to have lower numbers of reads, we define

$B = \{g : n_g \leq \hat{t}\}$, where the count threshold $\hat{t}$ is the smallest $t$ for which $B$ consists of >0.75 of the genome; this percentage was used and tested in (Liang and Keles, 2012b), and works well as long as the DNA binding protein does not bind to >25% of the genome. Finally, the number of reads in each window for the input sample is multiplied by its normalization factor, $\hat{r}$. The process is then repeated for each input sample. To normalize the ChIP samples for different immunoprecipitation efficiencies, for each ChIP sample, $c$, versus the reference $r$, the bin-wise log fold change for the $g^{th}$ genomic bin is defined as

$$M_g = \log_2(\frac{n_{rg}}{n_{cg}}) \tag{2.2}$$

and the geometric mean of log read counts is defined as

$$A_g = \frac{1}{2}\log_2(n_{rg} \cdot n_{cg}) \tag{2.3}$$

Where $n_{rg}$ and $n_{cg}$ are the raw read counts in the $g^{th}$ bin of the reference sample and target ChIP sample, respectively. The trimmed mean of $M_g$ values (TMM) is calculated as the weighted average of $M_g$ after removing the upper and lower $x$ percentages of data (based on both $M_g$ and $A_g$) as described in (Robinson et al., 2010). The default trimming percentages for $M_g$ and $A_g$ are 20% and 5% respectively. For $M_g$, 20% is a conservative estimate to exclude the differential sites, whereas for Ag, 5% is used to remove the highest and lowest signal regions where there may be a high percent of artefacts. The log fold change, $M_g$, is weighted by the mean log read counts. Thus,

$$\log_2(TMM_c) = \frac{\sum_{g \in G^*} A_g M_g}{\sum_{g \in G^*} A_g} \tag{2.4}$$

where $G^*$ denotes the remaining bins after the trimming. Since we aim to normalize for the difference in antibody efficiency among the ChIP samples, the normalization constant should be estimated only from enriched regions. Inclusion of background bins will bias the estimator towards the library ratio (e.g., if all bins were used the estimator would equal the library total read count ratio). In practice, the number of enriched regions varies across different TFs and it may not be clear how

many bins should be included before we have formally called the peaks. To overcome this uncertainty, PePr sorts the bins by $n_g = n_{rg} + n_{cg}$ and estimates the TMM from the largest $N$ bins, where $N$ is a vector of values ranging from 1,000 to 50,000 (1000, 5000, 10000, 20000, 30000, 40000, 50000; the range was set based on the number of peaks observed for common TFs and histones). From the several TMMs estimated from the different Ns, the one that is most different from the library ratio is reported. This will be close to the optimal TMM because as N increases toward the true number of peaks, the TMMs trend away from the library ratio, approach the enrichment signal ratio, and then eventually return to converge to the library ratio as $N$ surpasses and grows beyond the true number of peaks. Plots illustrating the steps of this normalization process are available on our website at http://code.google.com/p/pepr-chip-seq/.

Detection of significant windows

Read counts in the test and control sample groups (or two ChIP sample groups) are modeled using the negative binomial distribution as described here. Let $Y_{ijk}$ denote the observed number of reads in the $i^{th}$ genomic window($i = 1, \ldots, I$), the $j^{th}$ replicate($j = 1, \ldots, J_k$) and $k^{th}$ group($k = 1, 2$). Assuming a negative binomial distribution, we have

$$Y_{ijk} \sim NB(\mu_{i\cdot k}, \psi) = \frac{\Gamma(y_{ijk} + \psi^{-1})}{\Gamma(\psi^{-1})\Gamma(y_{ijk} + 1)} \frac{{\psi^{-1}}^{\psi^{-1}} \mu_{i\cdot k}^{y_{ijk}}}{(\psi^{-1} + \mu_{i\cdot k})^{(\psi^{-1} + y_{ijk})}} \tag{2.5}$$

where$\mu_{i\cdot k} = E(Y_{ijk})$ and $\psi$ is the dispersion factor (as $\psi \to 0$, the distribution converges to a Poisson distribution). By parameterizing the means of read counts for each window $i$ as $\mu_{i\cdot 1} = \mu_{i\cdot}$ and $\mu_{i\cdot 2} = \gamma\mu_{i\cdot}$, we can test for a significant difference between two groups by testing the following hypothesis:

$$H_0 : \gamma \leq 1 vs H_1 : \gamma > 1$$

In the case of test (ChIP) versus control comparisons, the controls are assigned as group 1 and test samples are group 2 so only one direction of the hypothesis will be tested; whereas in the case of two ChIP group comparisons (*i.e.* differential binding), a sample/group swap is performed and

the hypothesis is tested both ways automatically. The local dispersion parameter is estimated for each window using a weighted average of initial dispersion estimates from local windows in order to gain more robust estimates as described here. The log-likelihood for a given window is:

$$l_i(\psi) = \sum_{k=1}^{2}\sum_{j=1}^{J_k}[\log\Gamma(y_{ijk}+\psi^{-1}) - \log\Gamma(\psi^{-1}) - \log\Gamma(y_{ijk}+1)+ \\ \psi^{-1}\log(\psi^{-1}) + y_{ijk}\log(\widehat{\mu_{i\cdot k}}) - (\psi^{-1}+y_{ijk})\log(\psi^{-1}+\widehat{\mu_i.k})] \tag{2.6}$$

$$where\ \widehat{\mu_i \cdot k} = \frac{\sum_{j=1}^{J_k} y_{ijk}}{J_k}$$

The local dispersion estimator $\hat{\psi}$ maximizes the log likelihood over $W$ nearby windows (including the current window) using the triangular weight:

$$L(\psi) = \sum_{x=-W}^{W}(1 - \frac{|x|}{w+1})l_{i+x}(\psi) \tag{2.7}$$

The use of a local dispersion estimator provides a stable estimator of the dispersion factor when the sample size is small. $W$ is one for the SHARP peak setting and ten for the BROAD peak setting, based on observations of autocorrelation in multiple datasets. To calculate the significance, we use an asymptotic Wald's test with log transformation. We can define:

$$Z_i = \frac{[g(\hat{\gamma}) - g(\gamma_0)]}{g'(\hat{\gamma})\widehat{\sigma_{\hat{\gamma}}}} = \frac{[\log(\hat{\gamma}) - \log(\gamma_0)]\hat{\gamma}}{\widehat{\sigma_{\hat{\gamma}}}} \tag{2.8}$$

Where $Z_i$ has an asymptotic standard normal distribution,$\hat{\gamma} = \hat{y}/\hat{x}, \gamma_0 = 1,$ and $\widehat{\sigma_{\hat{\gamma}}}$ is defined as

$$\widehat{\sigma_{\hat{\gamma}}} = \sqrt{\frac{\bar{y}[J_1\bar{x}(\hat{\psi}^{-1}+\bar{y}) + J_2\bar{y}(\hat{\psi}^{-1}+\bar{x})]}{J_1J_2\hat{\psi}^{-1}\bar{x}^3}} \tag{2.9}$$

$$where\ \bar{x} = \frac{\sum_{j=1}^{J_1} y_{ijk}}{J_1}\ and\ \bar{y} = \frac{\sum_{j=1}^{J_2} y_{ijk}}{J_2}$$

P-values are calculated using $Z_i$ as the test statistic. Windows satisfying the specified p-value cutoff (the default is1e-5) are called as significant windows. Benjamini-Hochberg FDR is also reported.

Defining peak regions and post-processing of peaks

The significant windows that are localized in the same genomic area are merged. PePr has two different settings for merging windows; the maximal merging distance is smaller for the SHARP peak setting and larger for the BROAD peak setting (to ensure that the broad histone peaks are not broken into multiple enrichment regions in a given area). Generally in an explorative analysis when the enrichment shape of the peak is unknown to the user, the latter BROAD peak setting is recommended.

Optionally, PePr can remove peaks due to a high level of PCR duplicates in ChIP samples. Those peaks show no strand lag between forward and reverse strand reads (Landt et al., 2012) and are very likely to be false positives; a high proportion of these peaks in the final peak list is an indicator of poor data quality. Removing these artifacts requires accurate estimation of the shift size, otherwise we will be risking removing true positives. Fortunately, these artifacts also occur in a properly prepared control sample, displaying similar read profiles. Thus, PePr tackles this issue by removing peaks that have similar shape in both the ChIP and input samples. Specifically, for each peak, let $\pi_{xk}$ be the proportion of reads in the peak at nucleotide position $x$ for group $k$, where $k = 1$ is the ChIP group and $k = 2$ is the input group. Reverse-strand reads are counted at their 3' end. Thus, for each group $k$, $\sum_{x \in P} \pi_{xk} = 1$, where $P$ is the entire set of positions in the peak region. The minimum of the ChIP and input proportion at each position is determined, and the resulting values are summed to define the value $R$ across all positions in the peak using the formula:

$$R = \sum_{x \in P} \min(\pi_{x1}, \pi_{x2}) \tag{2.10}$$

$R$ ranges from 0 to 1, and will have a high value when the peak shape is similar between ChIPs and controls; based on observations, technical artifacts typically have a high R value greater than 0.5, whereas R values for most (true) peaks are distributed between 0 and 0.2. PePr removes the

peaks having R value greater than 0.5.

Additionally, PePr evaluates the overlap between forward-strand reads and reverse-strand reads (counted at their 3' end) before and after shifting. A peak with strand-overlap-ratio that is high ($>$0.2) before shifting and decreases significantly after shifting (decrease $>$50% of the original level) is removed by PePr. Most PCR-duplicate peaks simultaneously meet both of the two criteria defined above. These removed peaks are reported in a separate file.

Finally, PePr offers the option to refine the peak width for sharp peaks. Typically for TFs, downstream analysis such as motif analysis works optimally with a fine resolution of the peaks (i.e. reduced to minimal width that may contain the core protein-protected binding region). In an ideal (hypothetical) ChIP experiment, the core DNA binding site would be between the last starting position of the forward-strand reads and the first starting position of the reverse-strand reads. However, real-life ChIP-seq experiments are "contaminated" substantially by the background sequences (the percentages were observed to vary from 30% to close to 100% of the library (Liang and Keles, 2012b) ) and complicated by other technical factors influencing mappability and sequencability. Therefore, we use a more robust method to narrow the peak width without losing the protected region by setting the left boundary to be at the 20% quantile of the starting position from the forward-strand reads and the right boundary to be at the 80% quantile of the starting position from the reversestrand reads. These percentages are conservatively chosen.

*Differential peak binding*: the differential binding analysis entails an extra step compared to the peak calling analysis. In addition to the routine pre-processing steps, the reads in each window of an input sample will be subtracted from its respective paired ChIP sample if they are matched. In the case of uneven number of ChIP/input samples within each group or unpaired ChIP and input samples, the mean input reads will be subtracted from each ChIP sample. Any negative resulting values are redefined as zero counts. As mentioned earlier, the hypothesis will be tested both ways, calling differential binding sites enriched in each group.

### 2.2.3 Motif analysis

MEME (Bailey and Elkan, 1994) was used to identify over-represented motifs in the binding sites. For TF datasets, peaks that were found in all programs were used, and the region within 150bp of the peak mode was used as input to MEME. The most significant motifs identified by MEME are listed in Figure 2.1, and were consistent with previous reports (Han et al., 2013; Jothi et al., 2008). Upon obtaining the motif position specific score matrix (PSSM) for each TF from MEME, FIMO (Grant et al., 2011) was used to find motif matches in the regions within 150bp of the peak mode found by each program to identify the motif occurrences in the peaks.

### 2.2.4 Unique peak analysis

The versions, parameters and significance cut-offs used for each program are provided in Table 2.2 and Appendix A. The unique peaks for each program were defined as the peaks not overlapping any peak from the alternative program being compared. Since the number of unique peaks was often highly imbalanced, we examined the same numbers of top unique peaks with a maximum of 500. If too few unique peaks (less than 150) were identified, then the top 500 peaks identified by each but with the highest difference in rank were used as a surrogate to unique peaks. The heatmaps of unique peaks were generated using HOMER (Heinz et al., 2010), and visualized with Java TreeView (Saldanha, 2004).

## 2.3 Results

### 2.3.1 Overview of the PePr method

A schematic overview of the PePr pipeline is shown in Figure 2.2. After shifting forward and reverse strand reads to achieve proper alignment, PePr estimates a recommended window width based on the median peak width among top pre-candidate peaks to optimize statistical power. This is in contrast to most peak-finders, which use a fixed or user-specified window size. Motivated

by the importance of modeling variation in RNA-Seq data, we model read counts with a negative binomial distribution to account for extra-variation beyond that of the Poisson distribution observed in replicated ChIP-seq data (Figure 2.3). Unlike the RNA-Seq methods, however, we estimate the dispersion parameter $\psi$ (which accounts for extra-variation) from the local genomic area. After calculation of *p*-values, PePr merges adjacent significant windows to form continuous peak regions, which then undergo multiple post-processing steps to generate the final peak calls.

The motivation for estimating $\psi$ using local genomic information is that estimates of $\psi$ from one window are often unstable due to the small sample sizes commonly used in ChIP-seq studies. This problem of unstable variance estimates in experiments with small sample size has been studied extensively in the context of microarray data analysis, where using information from other genes hasshown significant improvement (Sartor et al., 2006; Smyth, 2004). For ChIP-seq data, we conjectured that close genomic regions share a similar microenvironment, and thus their behavior across samples may be correlated. Especially for histone marks like H3K27me3, that result in broad peaks, the estimated dispersion parameters from adjacent regions show strong correlation (Figure 2.4). Given the high auto-correlation observed for $\psi$ estimates along the genome, including information from nearby windows effectively increases the sample size, improving the robustness of the estimator.

### 2.3.2 Comparison to other methods

We assess the performance of PePr by applying it to eight TF ChIP-seq datasets: NRSF, ATF4, CTCF, GABP, NRF1, SMC3, USF1 and USF2, and to a histone modification dataset: trimethylation of Histone 3 lysine 27 (H3K27me3). The performance was compared with MACS, MACS2, SPP, ZINBA, edgeR and diffReps for TF data, and to SICER, ZINBA, edgeR, DiffBind and diffReps for histone data. MACS and SICER are among the favorite choices for sharp peaks and broad peaks respectively. ZINBA uses a similar sliding window approach, and is geared

towards either sharp or broad peaks. MACS and ZINBA were run with both the Combine the Replicates (CR) and Separate Analyses (SA) approaches described above. IDR was incorporated with MACS2 and SPP, as is recommended by the ENCODE project. EdgeR was performed in two ways in order to distinguish the effects of the core statistical model from the effects of the pre- and post-processing steps: adopting all of PePr's pre- and post-processing steps, and following a basic processing procedure. We denote them as edgeR-plus and edgeR-basic, respectively, in the main text.

Comparison of PePr and alternative methods using NRSF ChIP-seq data

The NRSF data consists of two ChIP and two input DNA samples, each sample having 14.3-26.6 million mapped reads. PePr identified a total of 5,284 peaks, comparable to diffReps, edgeR, SA and IDR-based approaches (Figure 2.5 B-F(i), Figure 2.6). CR-based approaches identified significantly more peaks, as expected since they gained coverage by pooling samples in the same group. This trend in number of peaks was also observed for the other TFs (Table 2.1). Comparing the ranks of peaks among the methods, we observed high correlation between PePr and MACS-CR (Pearson's r = 0.73), MACS-SA (r = 0.79), SPP-IDR (r=0.93), MACS2-IDR (r= 0.65), edgeR-basic (r=0.78), and edgeR-plus (r = 0.84), but much lower correlation between PePr and ZINBA-CR (r=0.14), ZINBA-SA (r=0.16), and diffReps (r=-0.25) (Figure 2.7). This trend in rank correlations between PePr and the other methods was also observed for the other TFs (see Figure 2.8 for ATF4).

The most direct assessment of peak-calling results that has been used is visual inspection of the shape and read coverage of the peak regions (Landt et al., 2012; Rye et al., 2011), however because this evaluation process cannot be fully automated, it is often overlooked in the evaluation of ChIP-seq methods. Instead, much of the literature depends on motif occurrence rate as the main performance measure, which can be inaccurate when the goal of the analysis is to identify differential or consistent binding sites under a specific biological context. Thus, we present visual

inspections of the peak profile results, as well as the motif occurrence rates in light of these results.

For each comparison between PePr and an alternative approach, we examined the peaks uniquely found by each (see Methods). In most of the comparisons (except for comparing to diffReps), PePr-unique peaks were more consistent between replicates and showed stronger read intensity (Figure 2.5 A-E (iii) and Figure 2.6) than the alternative program. Examining each peak individually (Figure 2.5 A-F (ii)), PePr-unique peaks exhibited a smooth peak shape and a strand lag, whereas unique peaks found by other approaches either had low read count which formed ambiguous shapes (MACS-SA, ZINBA-SA and edgeR-plus), peak profile shapes suggesting inconsistent binding (MACS-CR, ZINBA-CR, SPP-IDR and MACS2-IDR) or severe PCR-duplications (most notably diffReps and edgeR-basic). As expected by the limitation of the CR approach (including IDR), many of their unique peaks were only observed in one replicate (Figure 2.5 A,B(ii)).

In the average signal intensity plots, the mode height of MACS-CR, MACS2-IDR and diffReps unique peaks were higher than PePr-unique peaks (Figure 2.5 A,B,F(iv)); however, they were not the expected peak shape, but rather strongly spiked with width close to the read length. This suggests that the reads forming these peaks were mostly PCR duplicates from a limited number of sequences. Some diffReps-unique peaks even had the same peak shape in the control samples, but with fewer reads (Figure 2.5 F(ii,iv)). Whereas the narrow spiked modes are likely false positives with no shift size between strands, the signal levels of the shoulders of these plots likely represent real binding sites, with the expected shift size between strands. The signal in these shoulder regions are higher in PePr than the alternatives (Figure 2.5 A,B(iv)).

We compared the motif rates for peaks uniquely identified by PePr or an alternative approach (Table 2.3). The peaks uniquely found by PePr had comparable motif occurrence rate to MACS-CR, ZINBA-CR and MACS2-IDR, and had substantially higher motif occurrence rate than MACS-SA, ZINBA-SA, SPP-IDR, diffReps and edgeR-basic (Table 2.3). However, the motif rate of PePr's unique peaks is lower than edgeR-plus for NRSF, contrary to their stronger read signals

(Figure 2.5 E (iv)). The difference between edgeR-plus and edgeR-basic suggests that adopting PePr's processing steps results in a marked improvement. The CR approaches gained coverage by pooling the samples, resulting in a motif occurrence rate similar to that of PePr. Although the motif is often present, many of their unique peaks only showed enrichment in one replicate, and thus would likely be false positives for identification of consistent binding sites in a specific biological context under study. In the case of diffReps, the low motif occurrence rate in the unique peaks suggests that those peaks with a narrow spiked shape were likely not true NRSF binding sites.

Comparison of PePr and alternative methods using ATF4 and additional ENCODE ChIP-seq data

We repeated the comparison among methods on ATF4 ChIP-seq data, which had three samples each of ChIP and control (each having 26.8-30.2 million mapped reads), and the control samples were from chromatin immunoprecipitated ATF4 knock-out mice. All methods identified nearly twice as many or more peaks for ATF4 as for NRSF, except diffReps, which identified substantially fewer peaks than all other methods (Table 2.1). Again, we examined the unique peaks found by PePr versus the other programs. In all comparisons, we observed PePr unique (or ranked higher) peaks had higher read intensities (if we remove the high middle spike which is likely due to PCR duplications) and higher motif occurrence rate than the alternative programs, including edgeR-plus (Figure 2.9, Figure 2.10, and Table 2.3).

We analyzed six additional ENCODE TF datasets (CTCF, GABP, NRF1, SMC3, USF1 and USF2), which had conserved motifs and both duplicated ChIP and control samples. Since we showed in the NRSF comparison that the CR-based methods identify many sites that are inconsistent among samples, in these additional datasets we compared PePr to each of the alternative methods that take into account variation/differences among the replicates: diffReps, edgeR-basic, edgeR-plus, MACS-SA, and ZINBA-SA, with the same motif analysis (Table 2.4). In 24 of 30

comparisons, PePr-unique peaks had higher motif occurrence rate than the alternative method.

Comparison of PePr with alternative methods using a histone modification dataset

Oncogenic human papillomavirus (HPV) infection and tobacco-use are associated with the etiology of two subtypes of oropharyngeal squamous cell carcinomas (SCCs) (Chung and Gillison, 2009). We generated H3K27me3 ChIP-seq data from two HPV(+) and two age and gender matched HPV(-) SCC cell lines. The aim of the study was to identify candidate differential H3K27me3 sites by HPV status. The H3K27me3 mark exhibits very broadly enriched regions in ChIP-seq data, which we observed to be often highly variable between samples (Figure 2.11). Due to the high variation among samples and the goal of identifying consistent differences between HPV(+) and HPV(-) tumors, the CR approach would not be suitable; thus, we compared PePr to the SA approach using two peak-callers developed for broad peaks: ZINBA and SICER, as well as diffReps, DiffBind and edgeR-plus.

To find HPV(-) specific peaks, we used HPV(-) cell lines as the test samples, and compared them to the HPV(+) cell lines. For ZINBA and SICER, age and gender matched samples were used in each pair. SICER identified 35403 HPV(-) specific peaks for the first pair, 20207 peaks for the second pair, and 8823 regions (19%) were found in both. ZINBA identified 13814 peaks and 22701 peaks respectively, and only 1878 regions (5%) were found in both, illustrating the substantial level of variation among the samples in each group. PePr, diffReps, and edgeR-plus identified 1015, 17924 and 181 peaks, respectively. EdgeR-plus identified very few significant peaks, possibly due to the high global dispersion parameter estimated from the data, whereas PePr estimated it locally. DiffReps was much more sensitive than edgeR as previously shown for broad peaks (Shen, et al., 2013). For DiffBind, SICER peaksets were generated for each cell line and used as input to the program. The peaksets were merged and a total of 29510 regions were tested. DiffBind has two built-in analysis methods: edgeR and DESeq. DiffBind with edgeR reported no

significant peaks (possibly due to the same reason explained above). With DESeq it identified 918 HPV(-) specific sites, which we use for DiffBind below.

Since the number of peaks varied substantially among the programs tested, we evaluated how each program prioritized the peak findings. The top 900 peaks from each program were chosen based on their significance and compared (edgeR-plus was excluded due to finding so few peaks; 900 was chosen because all other methods identified 900 sites). Figure 2.13A shows the overlap between PePr and each of the other five programs. We examined the top ranking peaks that were uniquely identified by PePr, SICER, ZINBA, DiffBind or diffReps. The peaks uniquely identified by PePr were consistent between replicates, while the peaks uniquely found by SICER, ZINBA, or DiffBind often showed large differences in coverage between samples in the same group (Figure 2.13 B,C, D). DiffReps unique peaks seemed consistent in coverage between samples in the same group, however, the ratio in coverage between the test and control groups was smaller than that of PePr (Figure 2.13 E). In addition, when looking at each peak individually, the top diffReps unique peaks had average peak width less than 2 kb, which is much narrower than that expected for H3K27me3.

To further assess the robustness of the peak-calling methods (as opposed to differential binding methods) in identifying broad peaks, we conducted a scaling FDR analysis as described in (Zang et al., 2009) for all four ChIP-seq versus four matching input controls. Briefly, for each replicate, we randomly sample half of the reads to produce several pseudo half-size libraries. The proportion of peaks called only in the half-size library but not in the full-size library is defined as the scaling FDR. Performing this for ten simulations, we observed that PePr had a smaller scaling FDR (mean = 1.66%) than SICER (mean = 4.56%), ZINBA (mean = 11.38%), diffReps (mean=12.54%) and edgeR-plus (mean=5.83%), and thus PePr's peak prediction was most robust to differences in coverage levels (Figure 2.12).

## 2.4 Discussion

Currently, there is a lack of ChIP-seq analysis programs that account for biological variability within the peak-finding process. We have developed a method and tool, PePr, which uses a local negative binomial model to identify consistent or differential binding sites among ChIP-seq data, and that additionally estimates the optimal moving window size and offers post-processing steps to reduce false positives and refine peak resolution.

Variation among samples in ChIP-seq data can sometimes be quite large, such that some binding sites, even for transcription factors, are not reproducible (Landt et al., 2012). Inconsistent TF peaks among biological samples can exist for many reasons, including differences in accessibility of chromatin regions (e.g., due to histone tail modifications or DNA methylation), common sequence variants, competitive or cooperative binding differences with another TF (Whitfield et al., 2012), or technical artifacts that only occurred in one of the replicates. However, all but the last of these reasons are not significant concerns for most peak-finder programs, the goal of which is to identify all potential binding sites rather than consistent or differential binding sites. In addition, as public datasets from large consortiums such as ENCODE (Consortium, 2012) more comprehensively cover known TF binding in commonly used cell and tissue types, there will be less incentive for individual laboratories to identify all of the potential binding sites for a protein, as many will be available. A more refined hypothesis may be "where does this TF (or histone modification) bind consistently in this specific context (a specific disease, developmental stage, exposure, or treatment)?" Accurately modeling the variation is highly important in population epigenomics studies where substantial variation exists among samples, not only among individuals but also between tissue types (Cui, et al., 2009), developmental time points (Rugg-Gunn, et al., 2010; Sarmento, et al., 2004), and during disease progression (Conte and Altucci, 2014; Jakopovic et al., 2013).

We compared PePr to five commonly used single-sample peak-finders that use different under-

lying statistical models (MACS, MACS2, SPP, ZINBA, and SICER), as well as three programs that were designed for replicates (diffReps, DiffBind and edgeR), and found that PePr performed favorably in terms of consistently enriched read counts, motif occurrence rate and known characteristics of TF binding based on visual inspection. For comparison with MACS, ZINBA or SICER, we either performed separate paired analyses and then called peaks in the overlapping regions (SA) or combined the reads for the replicates and called peaks from the concatenated lists (CR). IDR was incorporated with MACS2 and SPP to determine the peak list cut-off, as recommended by the ENCODE consortium. Visual inspection of the peak shape and summarizing the read counts in peaks were extremely valuable in characterizing the unique tendencies of each approach. In particular, MACS was sensitive to detecting regions that had very low background and tended to miss peaks that had a relatively high background (Figure 2.5A(iii)); visual inspection of the ZINBA and diffReps unique peaks revealed that many had similar peak shape in both ChIP and control samples; SPP had severe false negatives for the NRSF data, which is possibly due to the removal of true binding sites that SPP mistakenly assumed to be artifacts due to having unexpectedly small shift size (i.e., in the "phantom peak" as defined in (Landt et al., 2012)). When we compared PePr to SPP-IDR and MACS2-IDR, we observed PePr-unique peaks (that are missed by the other two) had high read counts and motif rate.

Although motif occurrence rate is a useful marker for DNA binding, its value as a marker for consistent or differential DNA binding is not as clear. For identification of all DNA binding sites, motif analysis is expected to be specific (if the motif is found within a peak, it is assumed that binding occurs) but not necessarily highly sensitive (indirect binding cooperatively with other protein(s) may not result in a motif occurrence). Because the percent of binding sites without a motif is only expected to vary by DNA binding protein, and not by peak-caller, this is often ignored when comparing peak-callers. However, for consistent or differential binding experiments we can no longer assume specificity; a peak-finder that identifies fewer overall peaks with a motif

than an alternative may be correct in not calling the additional peaks as consistently bound or differentially bound. Given these caveats, we nonetheless found motif occurrence rate informative for interpreting our results when used in conjunction with visual inspection of peaks. The large improvement in motif occurrence rates for PePr-unique peaks compared to peaks identified by the Separate Analyses (SA) approaches and edgeR-basic suggests that peaks with higher read intensities and the expected smooth peak shape are more likely to contain a motif (Table 2.3). The CR approaches, on the other hand, were comparable in motif occurrence rate to PePr, but many of these were only bound in one replicate upon visual inspection, and thus are likely false positives for identification of consistent binding sites in the biological system under study.

Although PePr and edgeR use a similar underlying negative binomial model, edgeR lacks initial steps required for ChIP-seq peak finding (shifting opposite strand reads, defining and summarizing reads per window, etc), does not incorporate information from neighboring windows which especially benefits histone modification analyses, and does not offer post-processing steps to improve peak resolution or reduce false positives. In 5 of the 8 TF datasets, PePr performed better in motif rate than edgeR if the same PePr-processing steps are performed for edgeR; with the histone data, PePr was more sensitive than edgeR due to estimation of the dispersion parameters locally. PePr's post-processing steps improved edgeR's performance when there is a high proportion of PCR-duplicate peaks (6% in NRSF, 1.6% in ATF4, and $<$0.5% in other datasets). DiffBind was previously shown to work well with differential binding in TF data (Ross-Innes, et al., 2012), however, in H3K27me3 data with broad and highly variable peaks, DiffBind's edgeR module had very low detection power, while its DESeq module identified 918 peaks, many of which were inconsistent among samples. DiffReps resulted in unpredictable numbers of peaks across the datasets we tested, for example, it identified substantially fewer peaks than all other programs for ATF4 and SMC3 (Table 2.1), but many more for H3K27me3.

Due to the lack of benchmarks in histone modification datasets, in this dissertation we mainly

relied on transcription factor datasets to compare methods. However, PePr is adaptable to datasets with either sharp or broad peaks due to its empirical estimation of the optimal sliding window size, and thus is equally relevant for analysis of histone modification ChIP-seq datasets as illustrated with our H3K27me3 data. H3K27me3 tends to occur in broad regions several kilobases in length, making consistent peak calling more difficult. Based on our visual inspection of peaks and scaling FDR analysis for the approaches compared, we showed that PePr identified binding regions consistent between groups without being sensitive to changes in read coverage.

One limitation of PePr is that it currently does not perform paired analysis, similar to the limitation of multiple RNA-Seq differential analysis programs (Anders and Huber, 2010; Trapnell et al., 2013); thus, for example, the paired nature of tumor and patient-matched normal samples, could not be taken into account. For experiments requiring covariates, we currently recommend edgeR. PePr also makes the assumption that the quality of data for each ChIP-seq experiment is approximately equal, similar to most methods for other types of high-throughput analysis. When this assumption is violated, the result may be a high false negative rate due to missing peak regions in the lower quality sample(s); this may especially be true for experiments with very small sample size. In this case, users may obtain better performance using a different peak finder on individual samples, and a secondary method to explore options to combine results. Future versions of peak-finders for replicated ChIP-seq data could take into account quality, for example by assigning a weight to each sample.

# Figures

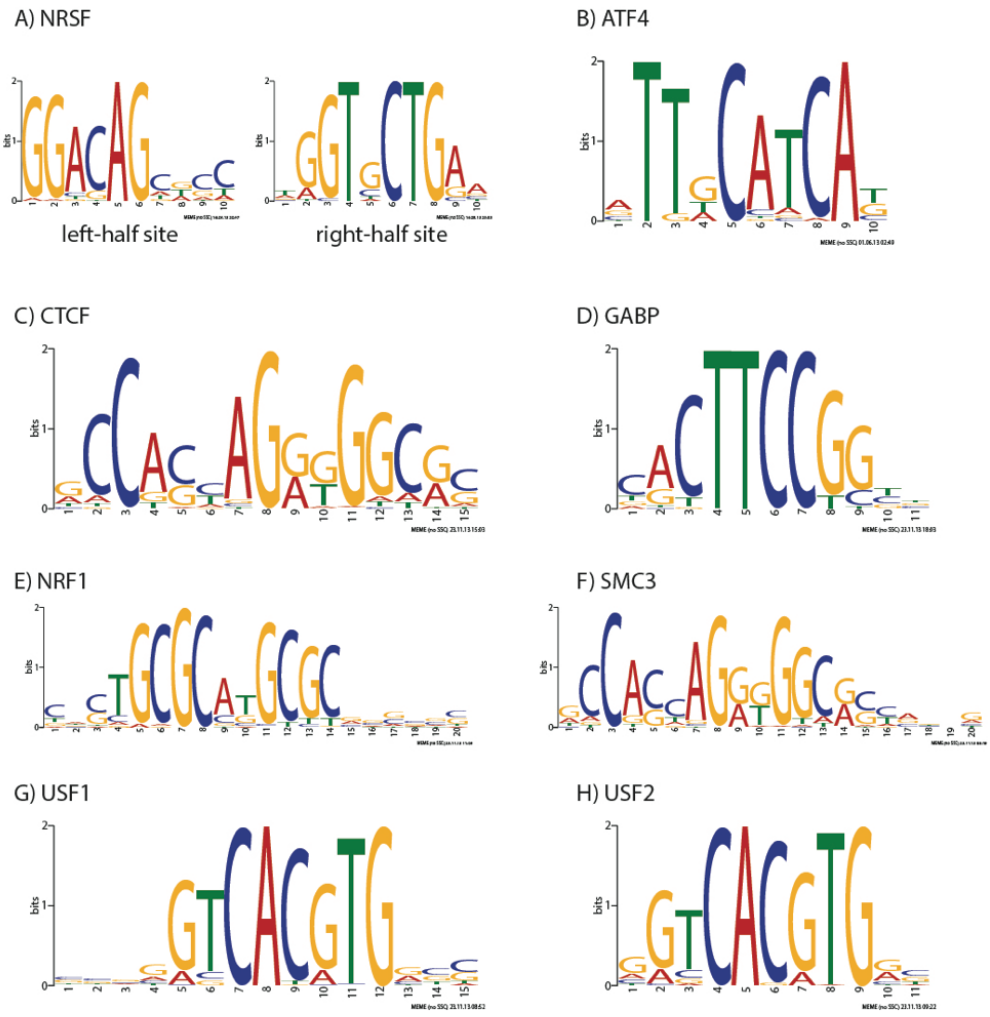



Figure 2.1: **Motif logos for all TFs used in our comparisons.** (A) Motif logo identified by MEME for NRSF data. NRSF binding sites have variable spacing between the two halves of the motif. (B-H) Motif logo identified by MEME for ATF4 (B), CTCF (C), GABP (D), NRF1 (E), SMC3 (F), USF1 (G) and USF2 (H).

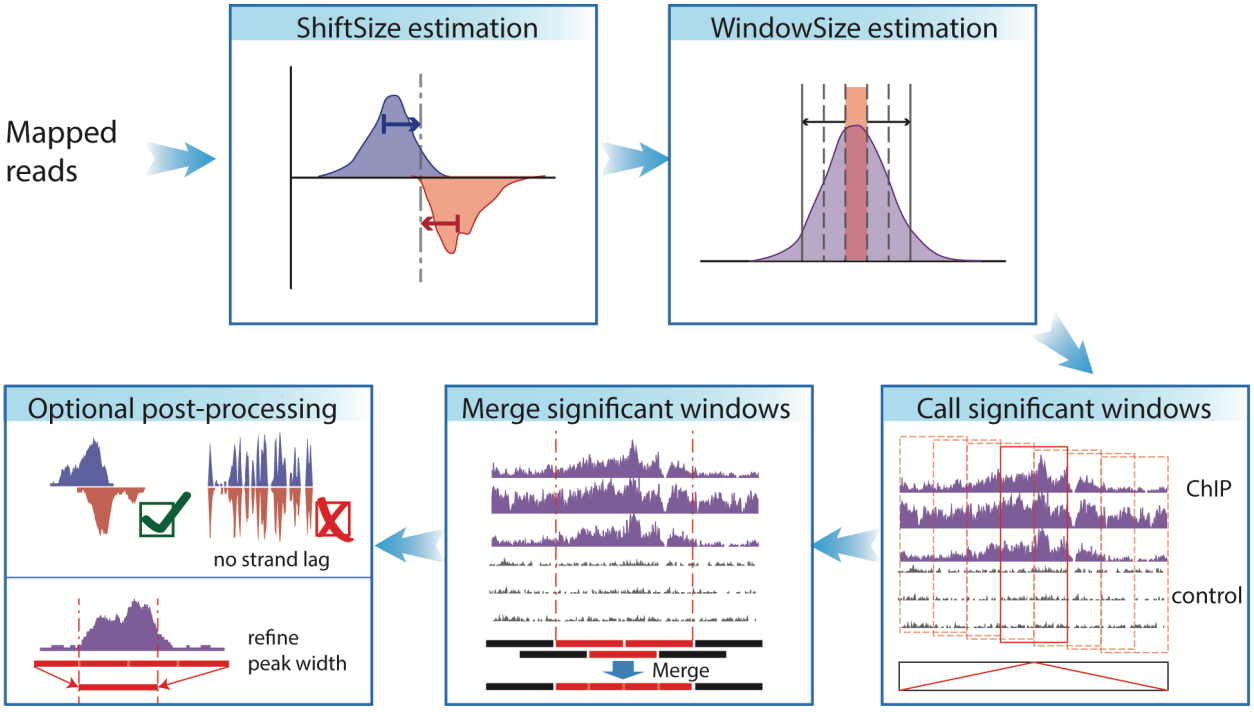



Figure 2.2: **Workflow of PePr**

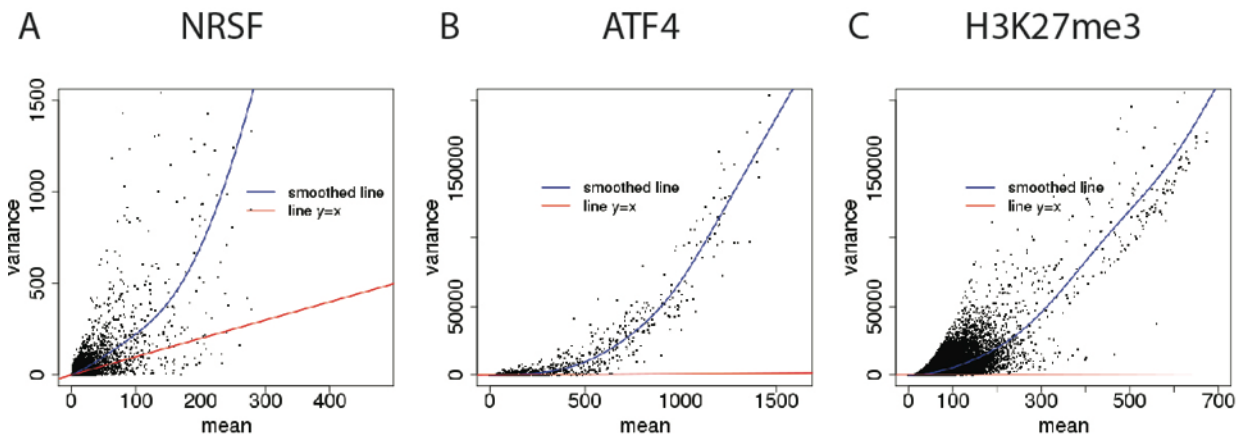



Figure 2.3: **Extra-variance beyond that of the Poisson distribution is observed in ChIP-seq data.** Plot of mean versus variance estimates for windows across the genome in (A) NRSF ChIP-seq data with two replicates (window size of 200bp), (B) ATF4 ChIP-seq data with three biological replicates (window size of 160bp), and (C) H3K27me3 ChIP-seq data from squamous cell carcinoma cell lines, with four replicates (window size of 340bp). The red line indicates the expected fit based on the Poisson distribution. The blue line is the fitted curve estimated using cubic smoothing spline.

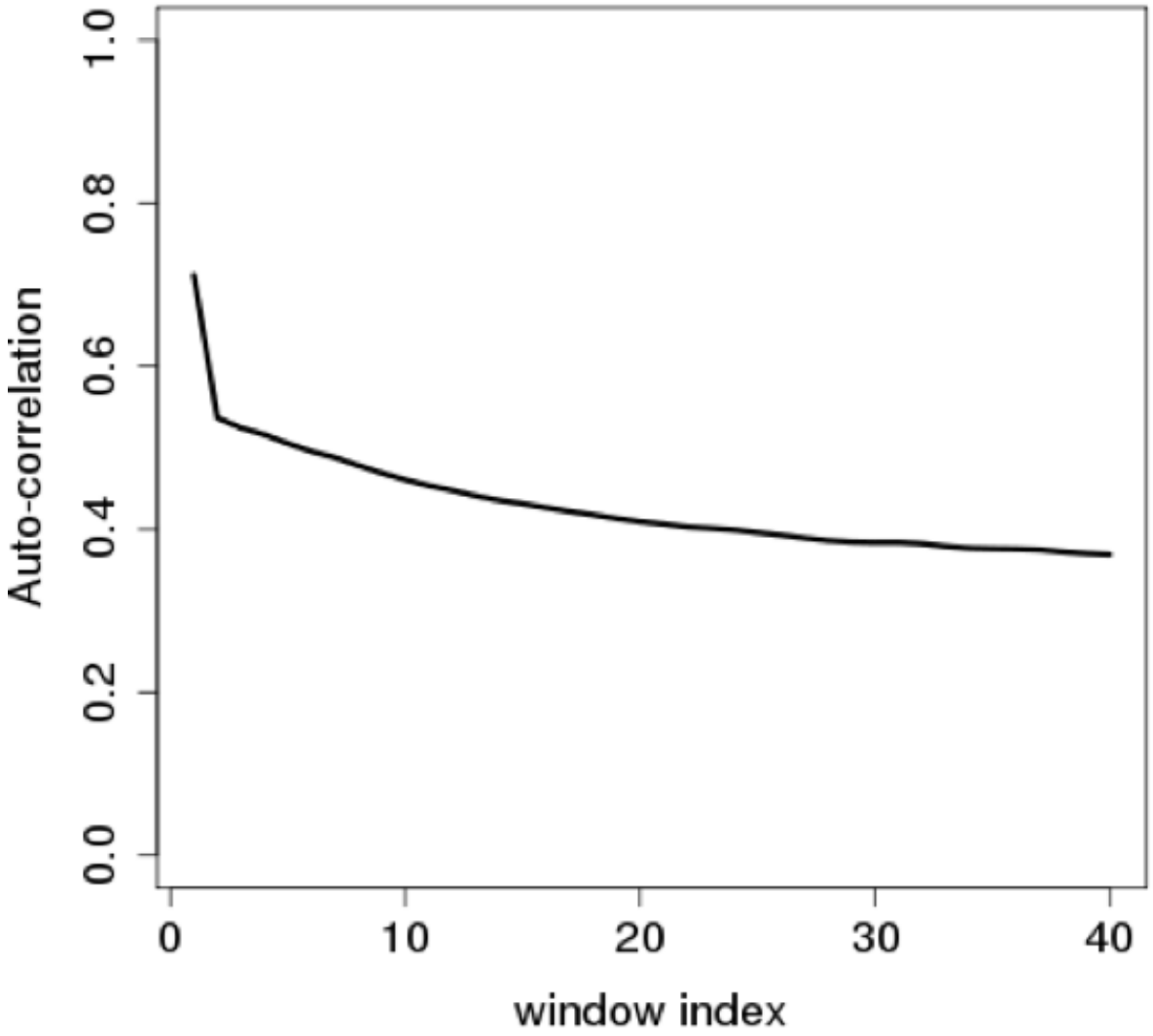



Figure 2.4: **H3K27me3 data show a high autocorrelation of the dispersion parameters estimated for nearby windows.** The genome was split into non-overlapping windows of 336 bp (Optimal window size estimated by PePr) and the dispersion parameter for each window was estimated. The autocorrelation of the dispersion parameters of the windows separated by (i-1) windows showed a correlation coefficient greater than 0.4 for to the range of 10 - 20 windows apart.

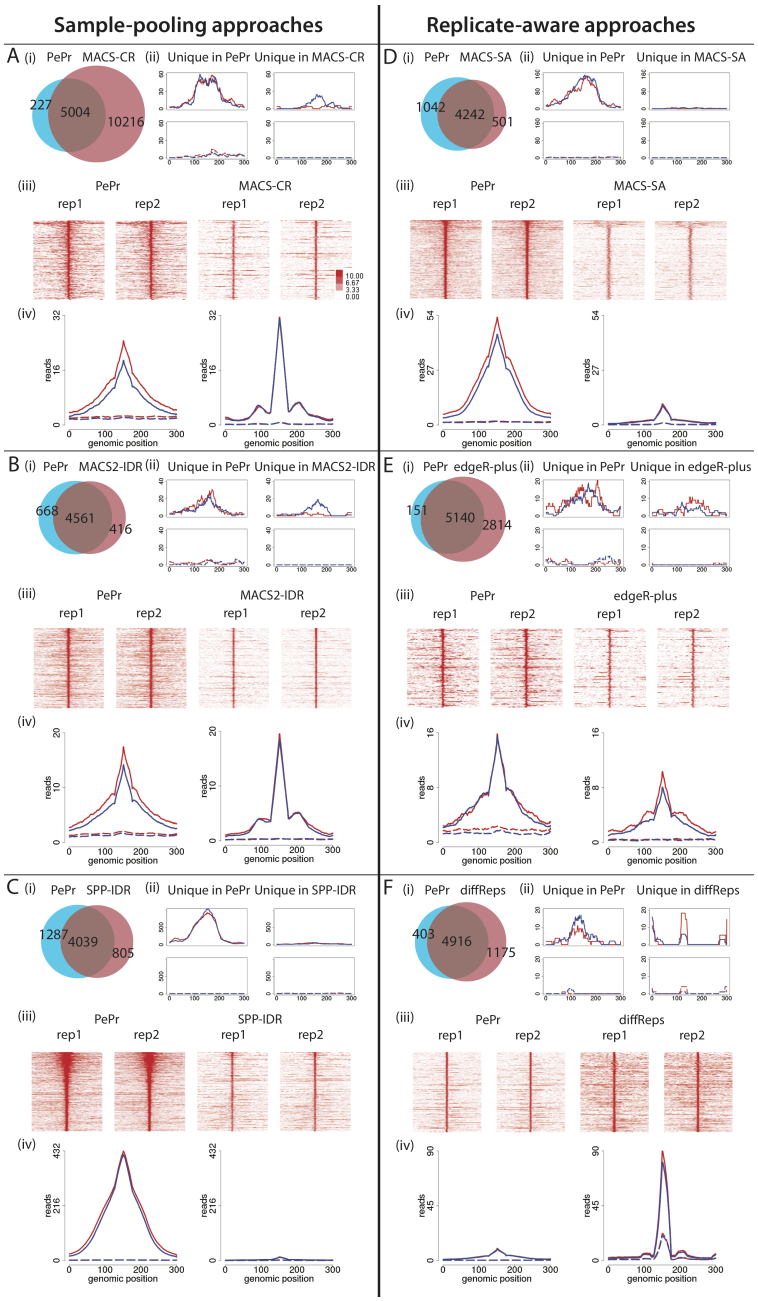



Figure 2.5: **Comparison of PePr to other approaches on NRSF data.** Other approaches are: MACS-CR (A); MACS2-IDR (B); SPP-IDR (C); MACS-SA (D); edgeR-plus (E); diffReps (F). The subplots in each panel are: (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group. ZINBA and edgeR-basic results are presented in Figure S4.

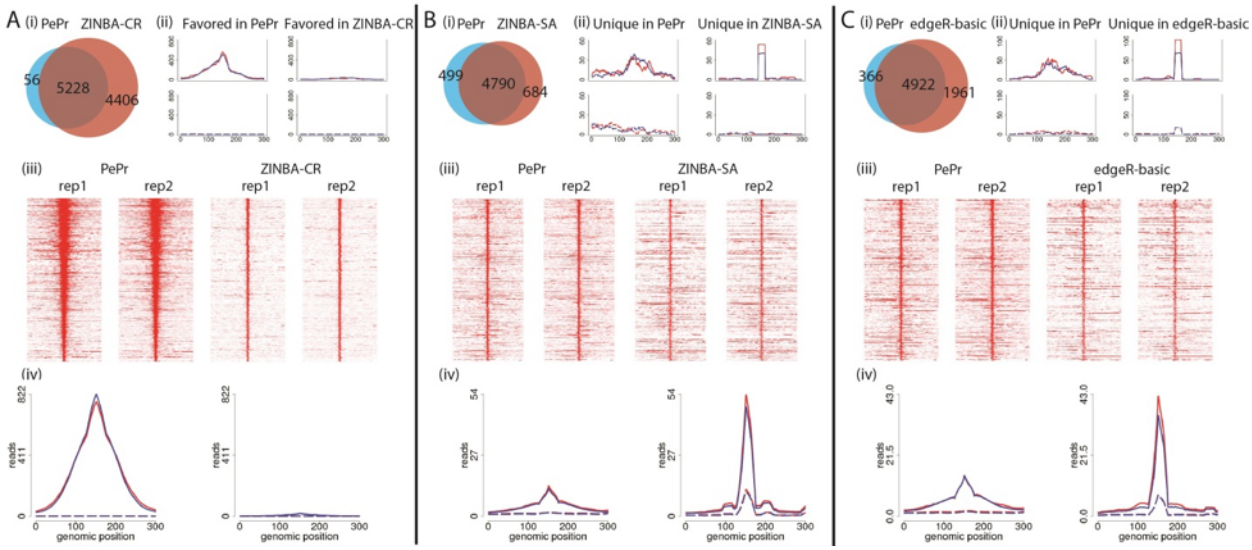



Figure 2.6: **Comparison of PePr to ZINBA-CR (A) ZINBA-SA (B) and edgeR-basic (C) on NRSF data.** (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group.

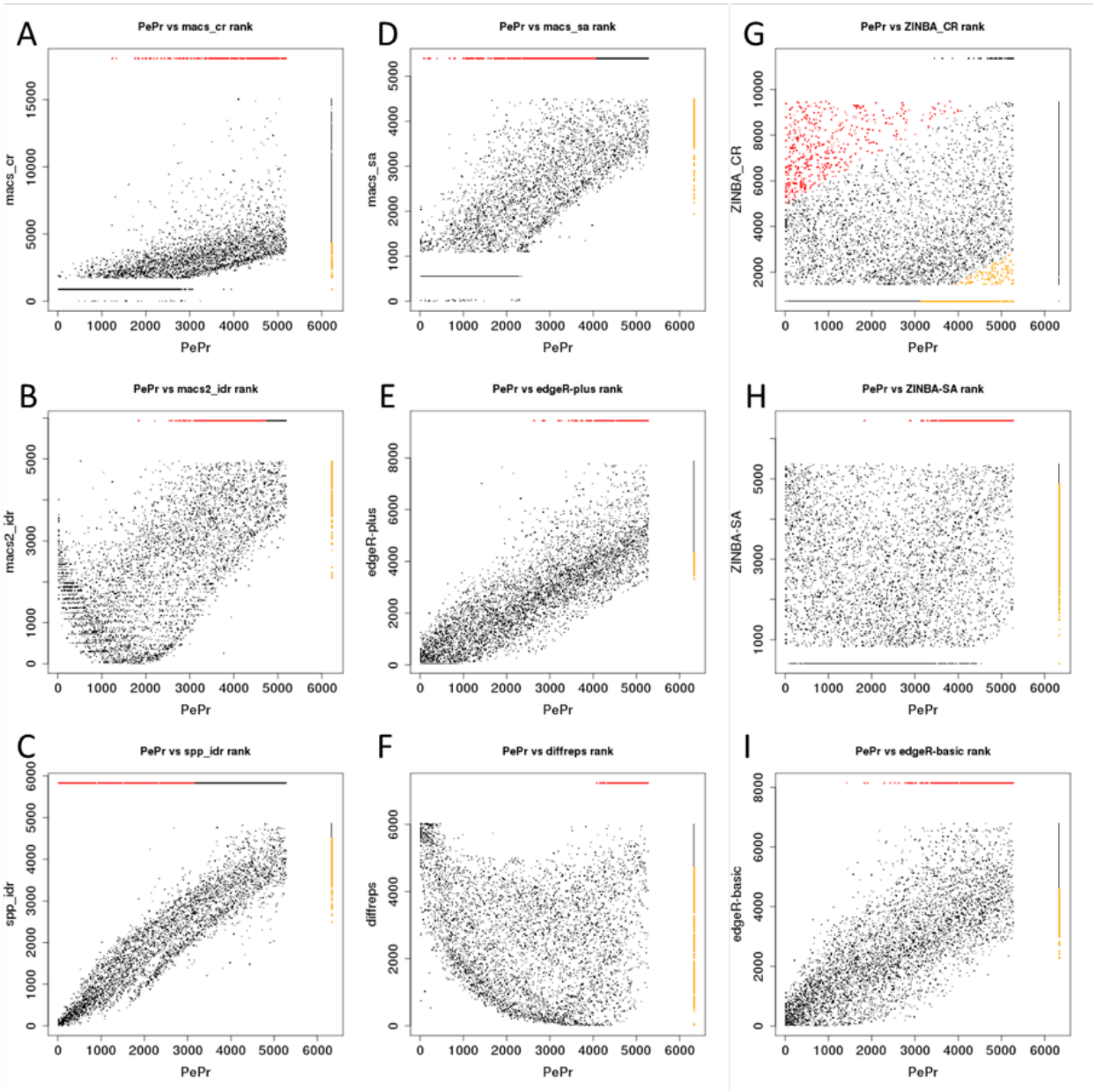



Figure 2.7: **Rank comparisons between PePr and the alternative approaches on NRSF data.** Rank comparisons between PePr and (A) MACS-CR (Pearson's r=0.73), (B) MACS2-IDR (r=0.65), (C) SPP-IDR (r=0.93), (D) MACS-SA (r=0.79), (E) edgeR-plus (r=0.84), (F) diffReps (r=-0.25), (G) ZINBA-CR (r=0.14), (H) ZINBA-SA (r=0.16), and (I) edgeR-basic (r=0.78). The peaks are ranked by the significance for each program. The points located at the top of each plot are PePr-unique peaks, and the points on the right margin of each plot are unique peaks for the alternative approach. Red and orange points refer to subsets of PePr-unique and alternative-approach-unique peaks which were used in the enrichment signal and motif occurrence comparisons.

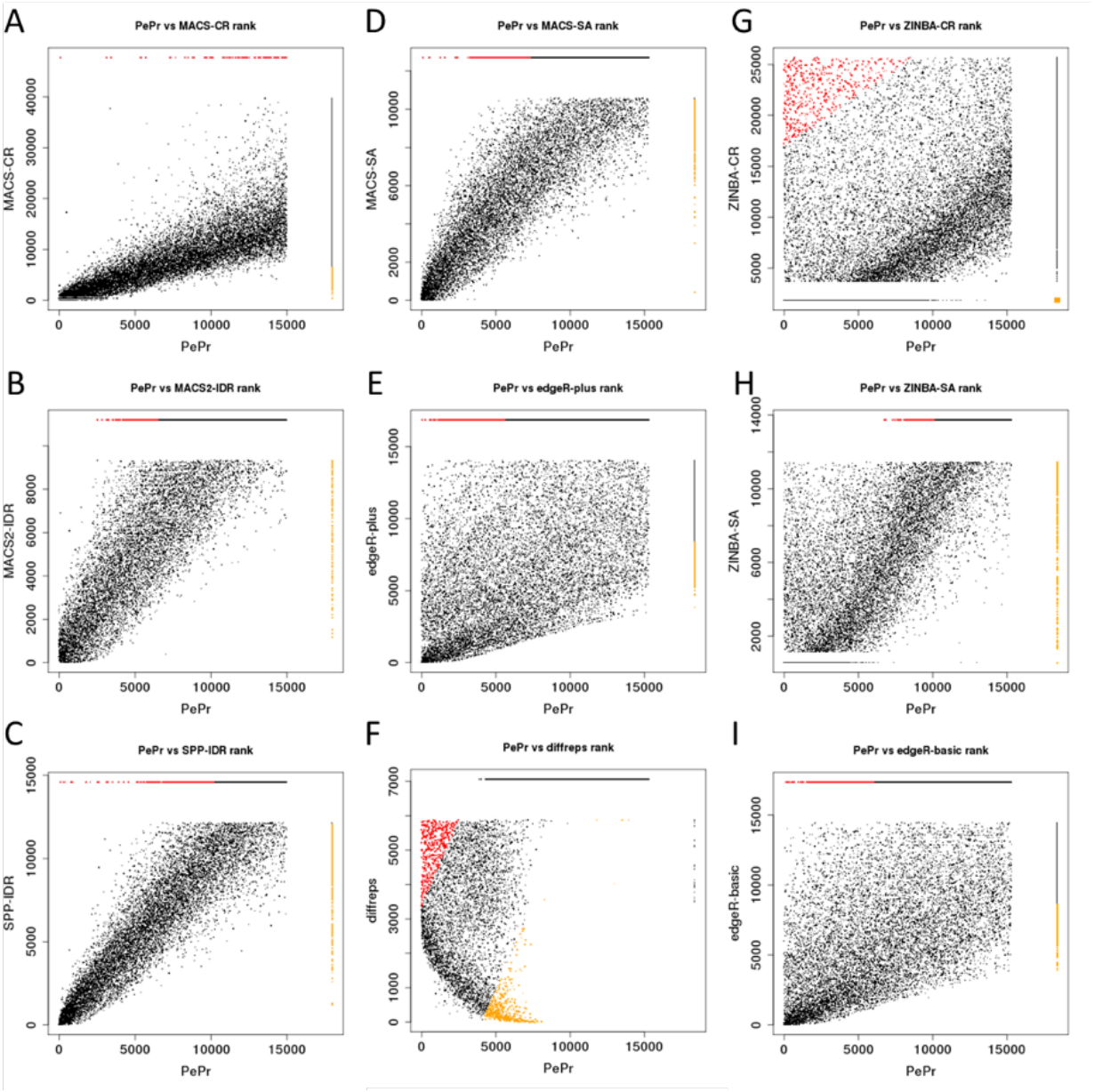



Figure 2.8: **Rank comparisons between PePr and the alternative approaches on ATF4 data.** Rank comparisons between PePr and (A) MACS-CR (Perason's r=0.82), (B) MACS2-IDR (r=0.82), (C) SPP-IDR (r=0.90), (D) MACS-SA (r=0.86), (E) edgeR-plus (r=0.51), (F) diffReps (r=-0.14), (G) ZINBA-CR (r=0.28), (H) ZINBA-SA (r=0.68), and (I) edgeR-basic (r=0.56). The peaks are ranked by the significance for each program. The points located at the top of each plot are PePr-unique peaks, and the points on the right margin of each plot are unique peaks for the alternative approach. Red and orange points refer to subsets of PePr-unique and alternative-approach-unique peaks which were used in the enrichment signal and motif occurrence comparisons.

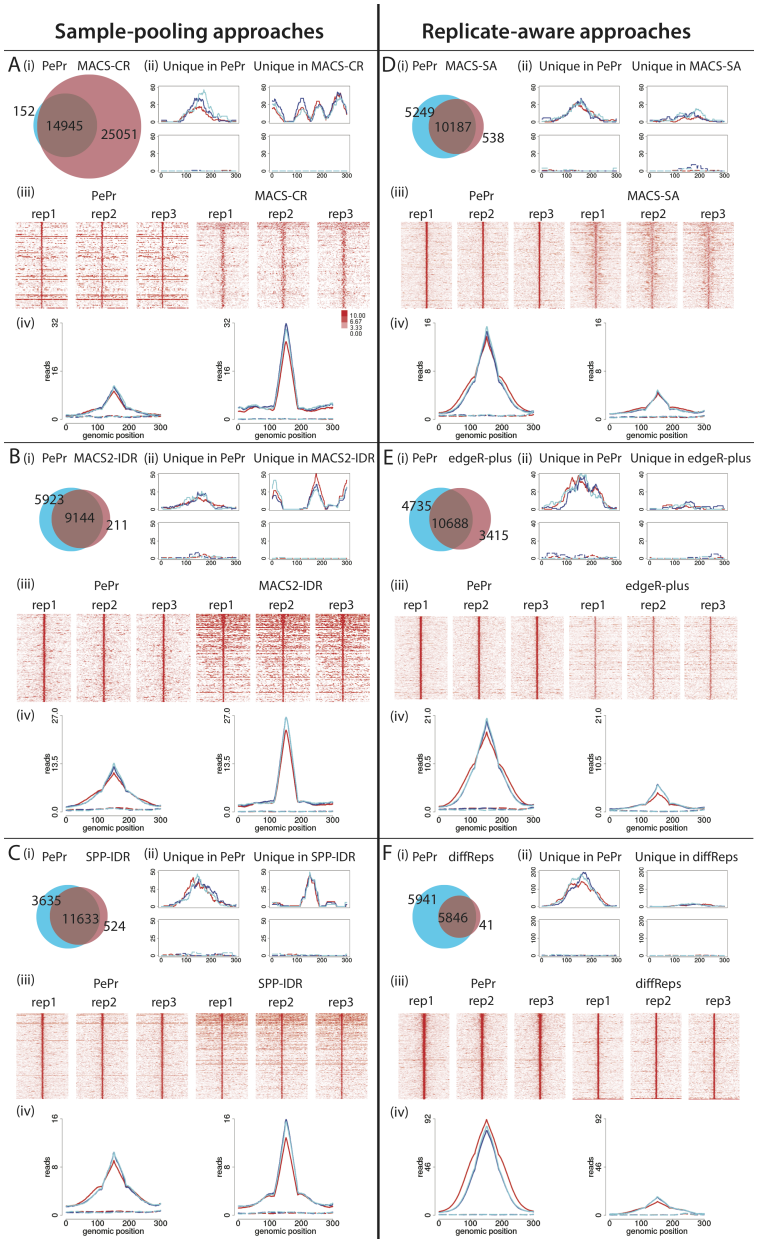



Figure 2.9: **Comparison of PePr to other approaches on ATF4 data.** Other approaches are: MACS-CR (A); MACS2-IDR (B); SPP-IDR (C); MACS-SA (D); edgeR-plus (E); diffReps (F). The subplots in each panel are: (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group. ZINBA and edgeR-basic results are presented in Figure S9.

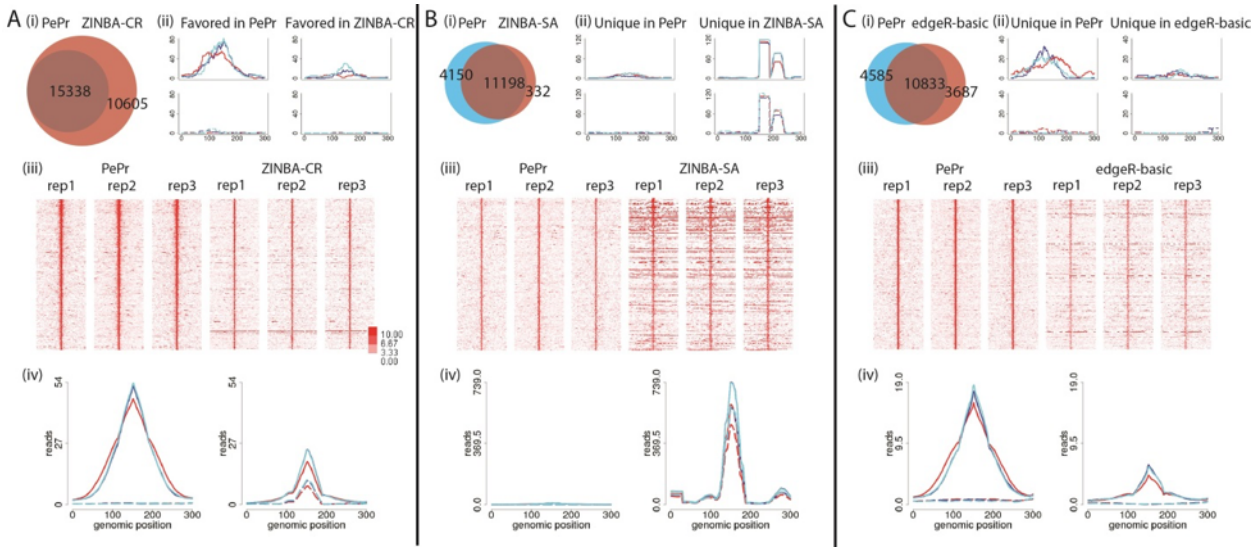



Figure 2.10: **Comparison of PePr to ZINBA-CR (A), ZINBA-SA (B) and edgeR-basic (C) on ATF4 data.** (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group.

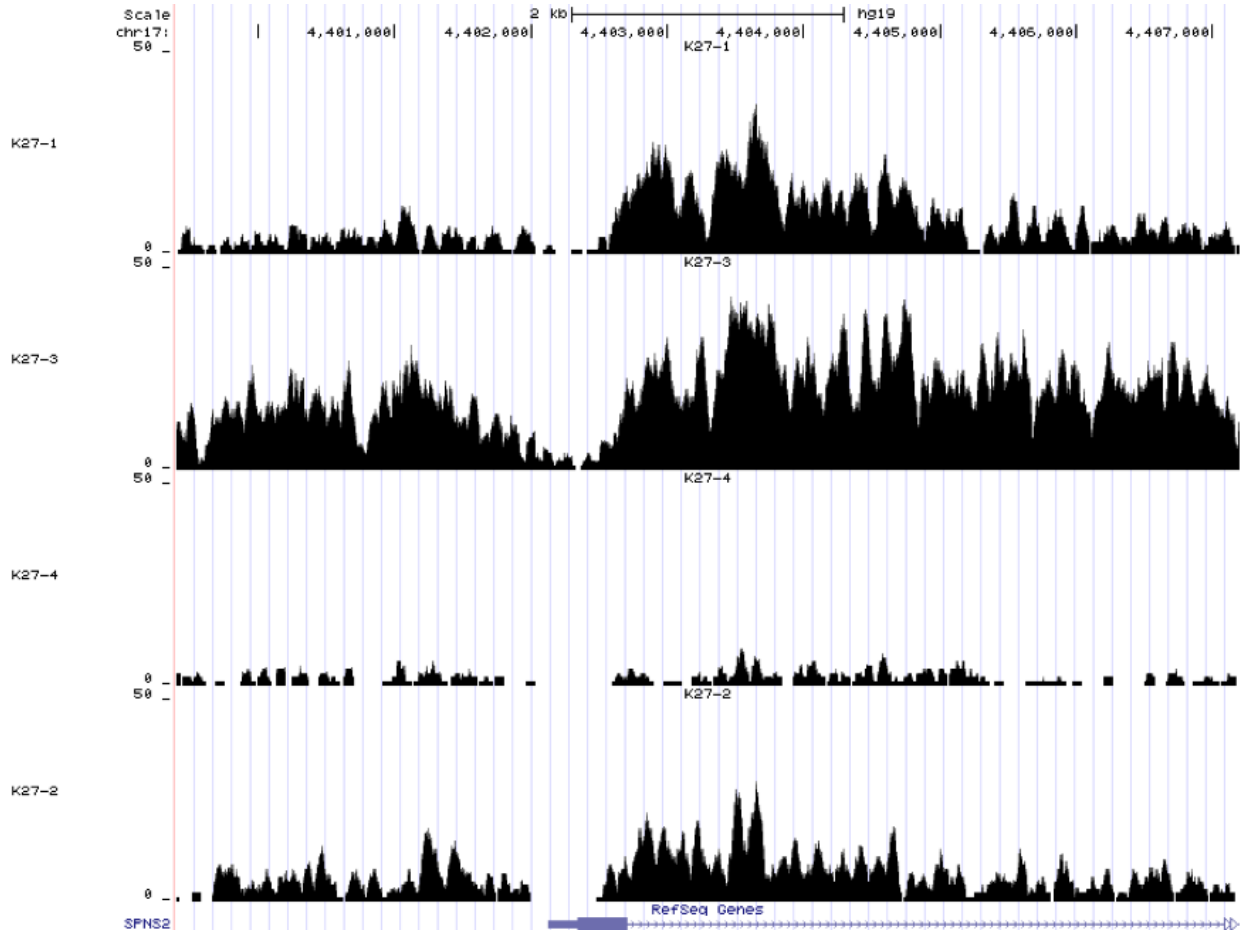



Figure 2.11: **Example of an H3K27me3 enriched region showing high variation of ChIP-seq signals across samples.** Each profile represents one ChIP-seq sample, with the x-axis and y-axis denoting chromosomal location and read coverage respectively.

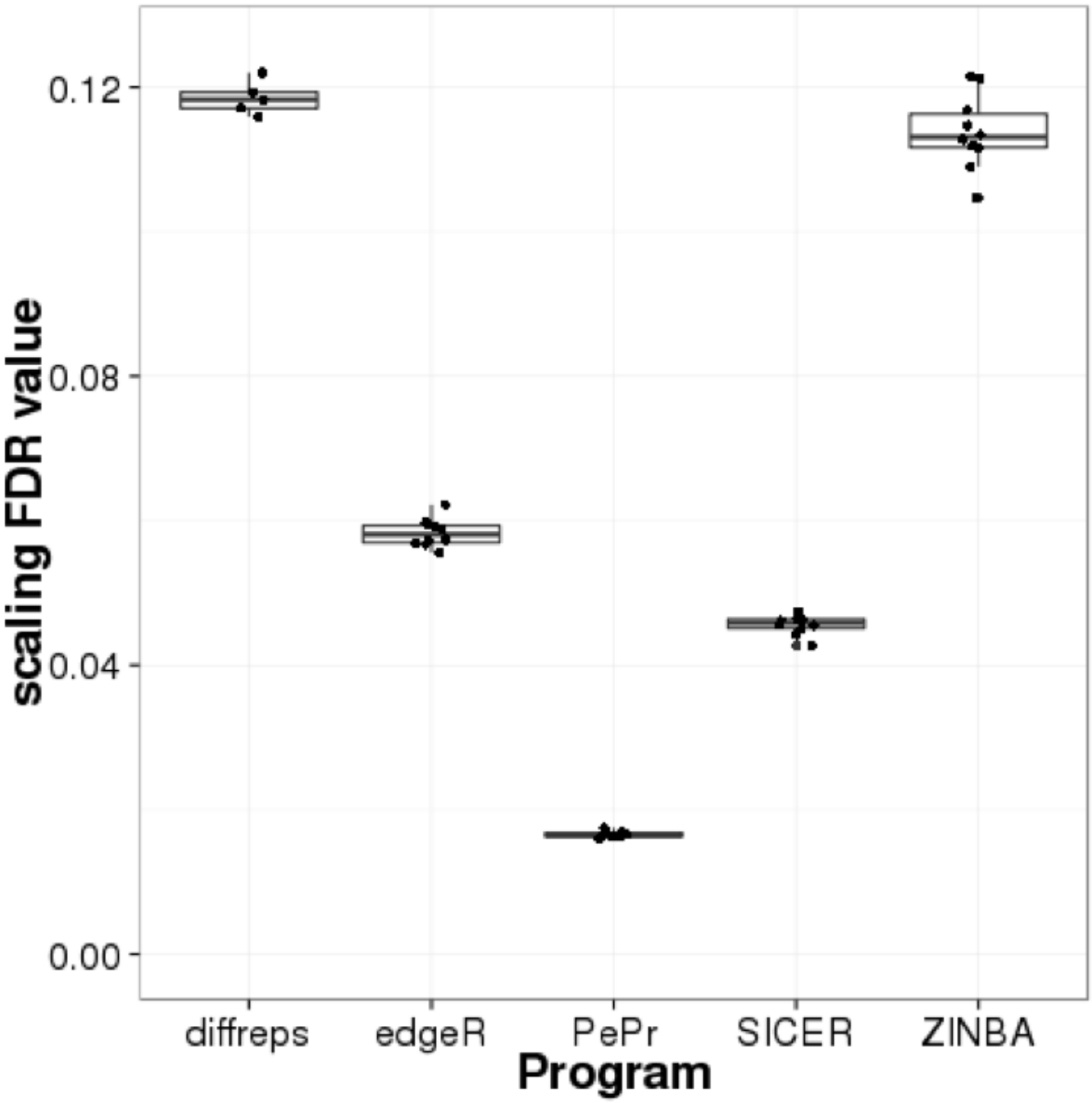



Figure 2.12: **A scaling FDR analysis of the H3k27me3 dataset shows PePr was most robust to differences in read coverage level.** The scaling FDR was calculated for PePr, ZINBA, SICER, diffReps, and edgeR on the H3K27me3 data as described in the main text. PePr had the lowest scaling FDR estimate of the methods tested.

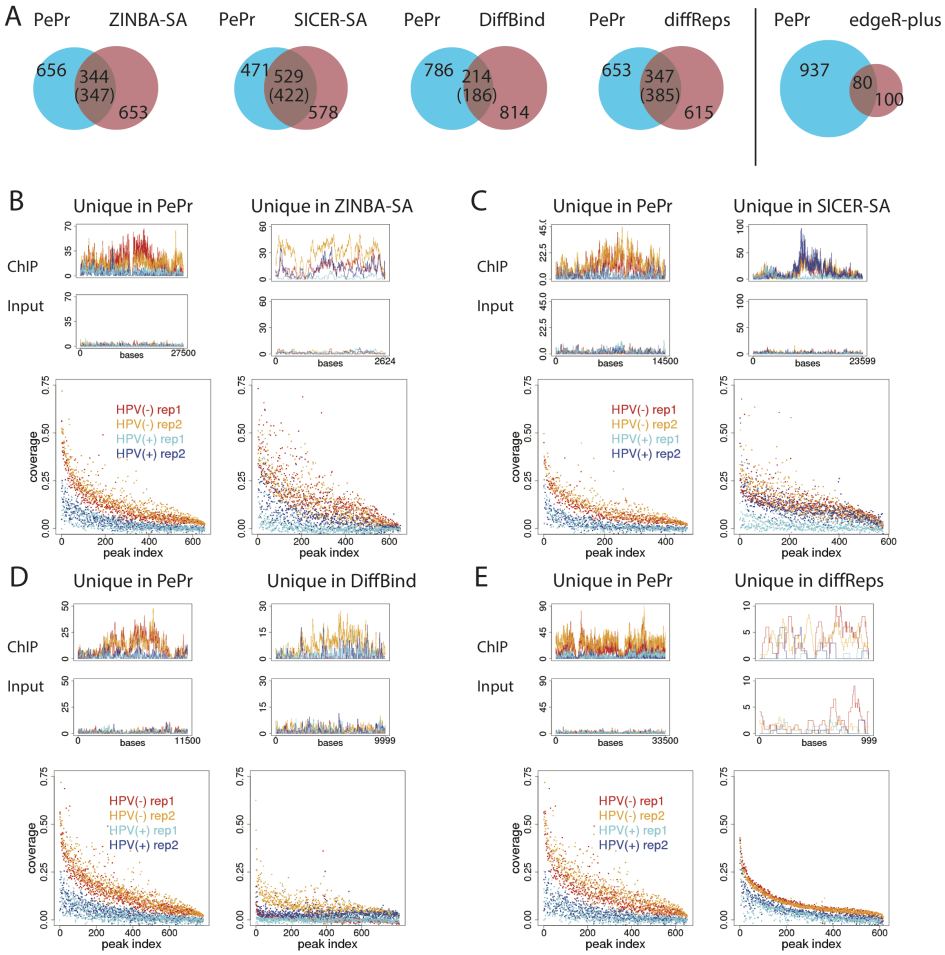



Figure 2.13: **Comparison of PePr to other approaches for H3K27me3 data.** (A) Venn diagrams showing the overlap between the top 900 peaks from PePr and alternative approaches (the number in parenthesis shows the number of peaks in the alternative program that overlap with PePr peaks). (B, C, D, E) Each plot on top shows the genomic view of top ranking peaks uniquely found by PePr or the alternative approach. The bottom plots show the normalized coverage of reads in unique peaks, sorted by the average coverage of both HPV(-) samples.

# Tables

Table 2.1: **Total number of peaks identified in each TF dataset.**

| | NRSF | ATF4 | CTCF | GABP | NRF1 | SMC3 | USF1 | USF2 |
|---|---|---|---|---|---|---|---|---|
| PePr | 5,284 | 15,338 | 34,548 | 5,158 | 4,729 | 25,789 | 6,837 | 5,025 |
| MACS-CR | 15,068 | 39,774 | 50,286 | 5,920 | 13,052 | 48,945 | 36,517 | 26,755 |
| ZINBA-CR | 9,468 | 25,684 | 57,398 | 5,880 | 14,052 | 62,044 | 12,343 | 23,376 |
| MACS-SA | 4,495 | 10,592 | 38,576 | 3,122 | 5,344 | 15,861 | 5,777 | 7,476 |
| ZINBA-SA | 5,374 | 11,453 | 41,675 | 4,613 | 6,286 | 21,912 | 5,706 | 9,060 |
| MACS2-IDR | 4,946 | 9,337 | 35,033 | 3,991 | 5,584 | 23,274 | 6,364 | 6,078 |
| SPP-IDR | 4,861 | 12,160 | 40,006 | 5,095 | 5,042 | 25,470 | 7,074 | 6,794 |
| diffReps | 6,030 | 5,781 | 29,317 | 3,992 | 3,474 | 3,499 | 4,270 | 3,642 |
| edgeR-basic | 6,790 | 14,463 | 43,443 | 6,962 | 6,643 | 15,731 | 7,581 | 16,397 |
| edgeR-plus | 7,868 | 14,057 | 40,841 | 8,116 | 9,667 | 13,303 | 7,315 | 12,426 |

Table 2.2: **Significance cut-offs for ChIP-seq programs involved.** *see program parameters for details.

| Program | Significance cuttoff |
|---|---|
| PePr | p-value <1e-5 |
| MACS | p-value <1e-5 |
| ZINBA | Posterior Probability >0.95 |
| SICER | FDR <1e-2 |
| MACS2-IDR | *optimum set |
| SPP-IDR | *optimum set |
| edgeR | p-value <1e-4 |
| diffReps | p-value <1e-4 |
| DiffBind | FDR <0.1 |

Table 2.3: **Motif occurrence rate in unique peaks called by PePr or alternative programs for NRSF and ATF4**.*peaks with highest rank difference were used, as explained in Methods.

| Program compared | NRSF | | | ATF4 | | |
|---|---|---|---|---|---|---|
| | | %motif | | | %motif | |
| | # peaks | PePr | alternative | # peaks | PePr | alternative |
| MACS-CR | 227 | 77.1 | 83.2 | 152 | 48.0 | 26.3 |
| ZINBA-CR | 500* | 97.8 | 92.6 | 500* | 67.0 | 51.4 |
| MACS2-IDR | 416 | 76.2 | 78.6 | 211 | 40.3 | 20.0 |
| SPP-IDR | 500 | 95.2 | 66.4 | 500 | 43.0 | 39.2 |
| MACS-SA | 500 | 89.0 | 70.4 | 500 | 45.8 | 29.8 |
| ZINBA-SA | 499 | 76.5 | 63.9 | 332 | 56.0 | 16.9 |
| edgeR-basic | 366 | 74.8 | 67.5 | 500 | 56.2 | 39.4 |
| edgeR-plus | 151 | 67.5 | 82.7 | 500 | 59.0 | 45.8 |
| diffReps | 403 | 79.9 | 25.6 | 500* | 75.2 | 61.6 |

Table 2.4: **Motif results for ENCODE TF data**. *peaks with highest rank difference were used, as explained in Methods.

| programs | MACS-SA | | | ZINBA-SA | | | edgeR-plus | | | edgeR-basic | | | diffReps | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | %motif | | | %motif | | | %motif | | | %motif | | | %motif | |
| | # peaks | PePr | alter. | # peaks | PePr | alter. | # peaks | PePr | alter. | # peaks | PePr | alter. | # peaks | PePr | alter. |
| CTCF | 500 | 76.4 | 81.2 | 500* | 96.2 | 87.6 | 500 | 78.2 | 78.4 | 278 | 73.0 | 80.2 | 500* | 95.8 | 87.2 |
| GABP | 500* | 85.8 | 70.2 | 312 | 55.4 | 22.7 | 172 | 54.6 | 8.1 | 370 | 55.6 | 24.6 | 500* | 89.2 | 74.8 |
| NRF1 | 500 | 82.4 | 74.4 | 500* | 99.0 | 95.2 | 500* | 89.0 | 96.0 | 500* | 90.4 | 96.0 | 500* | 99.6 | 95.6 |
| SMC3 | 500 | 96.0 | 62.6 | 500 | 90.6 | 24.2 | 500 | 96.8 | 50.4 | 500 | 97.8 | 58.8 | 165 | 97.6 | 7.9 |
| USF1 | 500 | 68.2 | 78.6 | 500 | 69.2 | 56.8 | 500 | 71.8 | 44.4 | 500 | 74.0 | 46.0 | 202 | 73.8 | 29.7 |
| USF2 | 500 | 64.6 | 49.6 | 500* | 79.2 | 73.8 | 173 | 52.0 | 43.3 | 500* | 67.4 | 60.8 | 216 | 69.9 | 19.0 |

## Bibliography

Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

Bailey, T. L. and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

Blahnik, K. R., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A. R., Iyengar, S., Nicolet, C. M., Ludascher, B., Korf, I. and Farnham, P. J. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic acids research*, 38(3):e13, 2010.

Boyle, A. P., Guinney, J., Crawford, G. E. and Furey, T. S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, 24(21):2537–2538, 2008.

Chung, C. H. and Gillison, M. L. Human papillomavirus in head and neck cancer: its role in pathogenesis and clinical implications. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15(22):6758–6762, 2009.

Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Conte, M. and Altucci, L. Functions, aberrations, and advances for chromatin modulation in cancer. *Cancer treatment and research*, 159:227–239, 2014.

Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M. and Jones, S. J. M. FindPeaks

3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics (Oxford, England)*, 24(15):1729–1730, 2008.

Grant, C. E., Bailey, T. L. and Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, 27(7):1017–1018, 2011.

Han, J., Back, S. H., Hur, J., Lin, Y.-H., Gildersleeve, R., Shan, J., Yuan, C. L., Krokowski, D., Wang, S., Hatzoglou, M., Kilberg, M. S., Sartor, M. A. and Kaufman, R. J. ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nature cell biology*, 15(5):481–490, 2013.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–589, 2010.

Jakopovic, M., Thomas, A., Balasubramaniam, S., Schrump, D., Giaccone, G. and Bates, S. E. Targeting the epigenome in lung cancer: expanding approaches to epigenetic therapy. *Frontiers in oncology*, 3:261, 2013.

Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic acids research*, 36(16):5221–5231, 2008.

Kharchenko, P. V., Tolstorukov, M. Y. and Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology*, 26(12):1351–1359, 2008.

Kornacker, K., Rye, M. B., Handstad, T. and Drablos, F. The Triform algorithm: improved sensitivity and specificity in ChIP-Seq peak finding. *BMC bioinformatics*, 13:176, 2012.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J. and Snyder, M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–1831, 2012.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, 2009.

Liang, K. and Keles, S. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics (Oxford, England)*, 28(1):121–122, 2012a.

Liang, K. and Keles, S. Normalization of ChIP-seq data with control. *BMC bioinformatics*, 13:199, 2012b.

Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–680, 2009.

Pepke, S., Wold, B. and Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nature methods*, 6(11 Suppl):S22–32, 2009.

Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., Yu, J. and Chinnaiyan, A. M. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC bioinformatics*, 11:369, 2010.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W. and Lieb, J. D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12(7):R67, 2011.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., Ali, S., Chin, S.-F., Palmieri, C., Caldas, C. and Carroll, J. S. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389–393, 2012.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M. B. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, 2009.

Rye, M. B., Saetrom, P. and Drablos, F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic acids research*, 39(4):e25, 2011.

Saldanha, A. J. Java Treeview–extensible visualization of microarray data. *Bioinformatics (Oxford, England)*, 20(17):3246–3248, 2004.

Sartor, M. A., Dolinoy, D. C., Jones, T. R., Colacino, J. A., Prince, M. E. P., Carey, T. E. and Rozek, L. S. Genome-wide methylation and expression differences in HPV(+) and HPV(-) squamous cell carcinoma cell lines are consistent with divergent mechanisms of carcinogenesis. *Epigenetics*, 6(6):777–787, 2011.

Sartor, M. A., Tomlinson, C. R., Wesselkamper, S. C., Sivaganesan, S., Leikauf, G. D. and Medvedovic, M. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC bioinformatics*, 7:538, 2006.

Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J. and Nestler, E. J. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PloS one*, 8(6):e65598, 2013.

Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:Article3, 2004.

Song, Q. and Smith, A. D. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics (Oxford, England)*, 27(6):870–871, 2011.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, 2013.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. and Sidow, A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9):829–834, 2008.

Wang, J., Lunyak, V. V. and Jordan, I. K. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics (Oxford, England)*, 29(4):492–493, 2013.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M. and Weng, Z. Functional analysis of transcription factor binding sites in human promoters. *Genome biology*, 13(9):R50, 2012.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.-L., Lin, F. and Sung, W.-K. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics (Oxford, England)*, 26(9):1199–1204, 2010.

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K. and Peng, W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics (Oxford, England)*, 25(15):1952–1958, 2009.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. and Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137, 2008.

# CHAPTER III

# Subtypes of HPV-positive head and neck cancers are associated with HPV characteristics, copy number alterations, PIK3CA mutation, and pathway signatures

## 3.1 Introduction

Head and neck cancer (HNC) is the 6$^{th}$ most prevalent non-skin cancer in the world, affecting approximately 600,000 patients per year, and with five-year survival rates ranging from 37 to 62 percent(Ferlay et al., 2010; Vokes et al., 1993). The majority of HNC cases have been historically attributed to excessive exposure of carcinogens such as tobacco and alcohol, but a significant and increasing proportion of cases are associated with high-risk human papillomavirus (HPV) infection, with HPV type 16 being the most common, accounting for 87% of HPV(+) cases in oropharyngeal HNC, 68% in oral HNC and 69% in laryngeal HNC carcinomas (Kreimer et al., 2005). Currently, approximately 75% of oropharyngeal tumors are associated with HPV; the HPV prevalence is relatively rare but still occurs in non-oropharyngeal sites such as oral cavity and larynx (Ang et al., 2010; Chung et al., 2014). Overall, HPV(+) HNC patients tend to have more favorable prognosis and treatment response rates, and different patient characteristics such as younger age at diagnosis, lower smoking rate, and higher intake of beneficial micronutrients compared to

The work in Chapter III will be submitted as Zhang Y, Koneva LA, Virani S, Arthur AE, Virani A, Hall PB, Warden CD, Carey TE, Chepeha DB, McHugh JB, Wolf GT, Rozek LS, Sartor MA. "Subtypes of HPV-positive head and neck cancers are associated with HPV characteristics, copy number alterations, PIK3CA mutation, and pathway signatures."

their HPV(-) counterparts (Arthur et al., 2011; Dayyani et al., 2010; Duray et al., 2014). They also have key molecular differences, such as the observed high expression of p16 (CDKN2A) in HPV(+) tumors and loss of p16 expression in HPV(-) tumors, and HPV(+) tumors are generally less-differentiated (Syrjanen, 2010) and have different copy number profiles (Hayes et al., 2015) than HPV(-) tumors.

HPV normally infects the basal layer of the epithelium, and then exploits the epithelial-to-keratinocyte proliferation and differentiation pathways in order to complete the viral life cycle. HPV expresses two main viral oncoproteins, E6 and E7, which cooperatively inhibit apoptosis and enhance tumor cell growth and proliferation by inducing degradation of tumor suppressor p53 and disruption of function of Rb, respectively (Moody and Laimins, 2010). Alteration of additional pathways, such as suppression of immune response (Tindle, 2002) and cell adhesion (Whiteside et al., 2008), induction of DNA damage(Duensing and Münger, 2002), centrosome amplification(Duensing et al., 2000) and oxidative stress (Williams et al., 2014), may also be important for tumor transformation.

High-risk HPV E6 is expressed in cells at two main isoforms: a full-length variant (E6) and a few truncated variants often collectively referred to as E6*. It was shown that E6* inversely regulate the ability of E6 to degrade p53 (Pim and Banks, 1999). Full-length E6 and E6* also bind to different sites of procaspase 8 and alternatively modulate its stability (destabilizing and stabilizing, respectively) (Tungteakkhun et al., 2009). Altogether, these studies suggests that E6* has distinct functions from full-length E6. In addition to the combined oncogenic potential of HPV E6 and E7, the integration of part or all of the HPV genome into the host genome is suggested to be a driver of the neoplastic process, and is estimated to occur in 75% of HNC cases, of which 54% are integrated into a known gene (Parfenov et al., 2014). Integrated viral transcripts are more stable than those derived from episomal HPV (Jeon et al., 1995), and integration confers an increased proliferative capacity and selective growth advantage (Jeon et al., 1995; Moody and

Laimins, 2010). Expression of these transcripts is key, as it has been observed that tumors positive for HPV DNA, but negative for HPV RNA, are similar to HPV(-) tumors with respect to gene expression and TP53 mutation frequencies (Wichmann et al., 2015).

HPV(+) HNC tumors represent a different molecular entity to HPV(-) tumors, distinct in their disease etiology and response to treatment. Several studies have investigated the differences in gene expression and copy number changes between HPV(+) and HPV(-) tumors (Pyeon et al., 2007; Slebos, 2006), or have defined expression subtypes irrespective of HPV status (Chung et al., 2004; The Cancer Genome Atlas Network, 2015). However, in most genome-wide studies, the number of HPV(+) cases are relatively low, therefore obscuring the discovery of subtypes within the HPV(+) population. In one study, Pyeon *et al* found two distinct subgroups of HPV(+) cancers, which were differentiated by a few key genes (Pyeon et al., 2007). Keck *et al* also independently identified two HPV(+) subtypes (Keck et al., 2015), and characterized differences in morphology (differentiation and proliferation), expression (mesenchymal and immune response) and copy number of a few key neoplasm genes (PIK3CA, TP63, SOX2) between subtypes. However, both studies used microarray data, thus limiting their ability to comprehensively characterize the differences between subtypes and identify the underlying causes of the expression differences.

Here, by deep analyses of 36 HNC tumors (18 HPV(+); 18 HPV(-)) collected at the University of Michigan with transcriptome and whole genome copy number alterations (CNA) data, we define two robust HPV(+) subtypes distinguished by gene expression patterns. The two clusters correlate with genic viral integration status, E2/E4/E5 expression and E6 splicing ratio. In addition, the clusters show different CNA patterns, mutation frequencies in PIK3CA and expression of cancer-relevant pathways, such as host immune response and keratinocyte differentiation. Similar analysis carried out on 66 additional HPV(+) samples from The Cancer Genome Atlas (TCGA) yielded the same findings, demonstrating the robustness of the subtypes and their related characteristics.

## 3.2 Methods

### 3.2.1 Tumor tissue acquisition, DNA and RNA extraction.

As a part of our ongoing survivorship cohort, we identified incident cases of HNC patients at University of Michigan hospital with untreated oropharynx or oral cavity squamous cell carcinoma between 2011  2013 were screened for eligibility. Written informed consent was obtained. The study was approved by the University of Michigan Institutional Review Board. Pretreatment Tumor tissue and blood were collected into a cryogenic storage tube and flash frozen in liquid nitrogen by surgical staff until storage at -80C. The flash frozen tissues were embedded in OCT media in vinyl cryomolds on dry ice and stored in -80C until prepared for histology. H&E slides were sectioned from each frozen tumor specimen on a cryostat and assessed by a board certified pathologist for degrees of cellularity and necrosis. Criteria used for inclusion in the study were a minimum of 70% cellularity and less than 10% necrosis. The first 36 tumors meeting these criteria were selected. Using a sterile scalpel, surface scrapings were taken directly from the frozen tissue blocks from the region of tissue identified as having at least 70% tumor cellularity, over dry ice to allow the tissue to remain frozen. Frozen scrapings were placed into pre-chilled tubes on dry ice and processed using the Qiagen AllPrep DNA/RNA/Protein Mini Kit (Valencia, CA, USA) as per manufacturer protocol. Blood DNA from the same patients was isolated using the Qiagen QIAamp Blood DNA Mini Kit.

### 3.2.2 RNA-seq and SNP-array protocol

RNA library construction and sequencing on Illumina HiSeq using 100 nt paired-end reads were performed by the University of Michigan DNA sequencing Core Facility. Samples were multiplexed to avoid lane variations. DNA from all tumors and matched blood samples was run on the Ilumina HumanOmniExpress BeadChip SNP-array. The raw microarray images were processed by Illumina Genome Studio to yield the log R ratio (LRR) values and B allele frequency (BAF)

values. Raw and processed RNA-seq and SNP-array data can be accessed from GEO with the accession number GSE74956.

### 3.2.3 RNA-seq analysis of the host gene expression

The RNA-seq library sizes ranged from 33 to 78 million reads (average of 47 million). The raw sequences were aligned to hg19 using Tophat2 v2.0.11 (Kim et al., 2013) with default parameters, resulting in an alignment rate of 48%-92% (average 86%). Quality control was performed using FastQC (Andrews, 2010) and RSeQC (Wang et al., 2012) before and after alignment. Gene expression levels were quantified using HTSeq v0.6.1p1 with the 'interFsection-strict' option (Anders et al., 2014). Genes expressed at least 1 count per million (CPM) in at least 10% of the samples were kept for downstream analysis. Normalization and differential expression testing were performed using the Bioconductor package edgeR v3.8.5 (edgeR-robust) (Robinson et al., 2010; Zhou et al., 2014), and adjustments for multiple testing were made using the False Discovery Rate approach.

### 3.2.4 Measuring HPV gene expression, and detection of HPV subtypes and genic integration

The libraries were also aligned to HPV genomes (downloaded from NCBI) using rnaSTAR v2.3.0 (Dobin et al., 2013) to allow gapped alignment over splicing junctions. We aligned to the HPV genomes of all high-risk types: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 and 68. STAR first-pass alignment detected the most abundant splice junctions, which were then used in the second-pass alignment. All of the quantification and analyses below were based on the second-pass alignment. Samples were classified as HPV(+) if they had more than 500 read pairs aligned to any HPV genome, and HPV type was determined as the type which had the most reads aligned. No samples had between 50 and 1000 reads aligned to HPV, resulting in a very clear separation between HPV positive and HPV negative. Overall we identified 14 HPV type16, one type18, one type33, and two type35. Virus integration events resulting in HPV-host fusion transcripts were

identified using VirusSeq (Chen et al., 2013b) with the default parameters. A positive integration event was defined as having at least four discordant read pairs and at least one junction spanning read. A tumor sample was called genic integration positive if it contained at least one identified integration event.

HPV is a small genome with short oncogenes, therefore in order to optimize the comparability between UM data and TCGA data in terms of HPV gene expression and splicing, all UM libraries were trimmed to the first 48 base pairs (as 48bp was the length of the TCGA RNA-seq reads) for the alignment to HPV genomes. This was implemented after observing differences between cohorts when not trimming UM sample reads; the trimming removed these artefactual differences. HPV gene expression levels were represented in the form of "count per million (CPM) value", which were calculated as the number of read pairs aligned to the HPV genome and intersected with each gene, and then divided by the total library size and multiplied by one million.

### 3.2.5 Computing full-length E6 percentages

E6 is expressed in two main forms: (1) a full-length E6 isoform and (2) various spliced isoforms of E6, collectively referred to as E6*. We define full-length E6 percentage as the ratio of full-length E6 and all E6 transcripts. Because all E6* skip the first intron, we can compute full-length E6 percentage as the ratio of the average coverage level in the first intron (approximate full-length E6 level) divided by the average coverage level in the first exon (approximate all E6 transcript level). We used rnaSTAR to identify the first introns for E6*, which are, in the format of [donor-acceptor], [227-408], [234-415], [232-508] and [233-414] for type16, 18, 33 and 35, accordingly. The average coverage of read pairs within the intron (i.e. 227-408 for HPV-16) was then divided by the average coverage of read pairs within exon1 (i.e. 82-227 for HPV-16), where coverage was computed by dividing the number of reads by the feature width (Fig 3.8D).

### 3.2.6 Finding unique pathways in each HPV(+) cluster

To understand how the pathways were dysregulated in the context of all UM HNC tumors, we ran edgeR and LRpath to find uniquely up- or down-regulated gene sets in each cluster (non-HPV, HPV-KRT and HPV-IMU). To do this, we constructed a model matrix for edgeR with three group labels: non-HPV, HPV-KRT and HPV-IMU. And then we ran edgeR GLM with the contrast matrix (HPV-KRT vs non-HPV + HPV-IMU) to find genes uniquely up- or down-regulated in HPV-KRT (the same step is repeated on HPV-IMU). Finally, we ran LR-path directional test to find enriched pathways.

### 3.2.7 Unsupervised clustering of gene expression values

Standard hierarchical clustering was performed on the median-centered log-transformed count per million (log-cpm) values. Genes used for clustering were chosen based on the following procedure. First, we calculated the rank of the variance and the rank of mean log-cpm for each gene analyzed. Then we used the rank difference between variance and mean to sort the gene list. Genes with a higher rank of variance than rank of the mean were considered more variable. Different thresholds (top 5000 genes to all genes) were used to choose the gene list, but the clusters were robust to these differences, with at most one sample switching clusters. In addition, different distance measures (Euclidean distance, centered and uncentered correlation), linkages (complete, average and single) were also tested; again, the clusters were robust to the different parameters used. We also performed consensus clustering (Monti et al., 2003) using recommended parameters provided in the user manual, which produced the same clustering results. The final clusters reported were using genes with a positive rank difference (6922 genes), uncentered Pearson's correlation and average linkage. The same clustering method was performed on UM+TCGA HPV(+) samples, using genes with a positive rank difference (6780 genes), uncentered Perason's correlation and average linkage. The consensus clustering results are presented in Fig 3.6A,B.

### 3.2.8 Pathway scores

Sample-wise pathway scores were calculated for four selected representative gene sets: E6 regulated genes, EMT genes, T-cell activation (GO: 0042110) and keratinocyte differentiation (GO: 0030216). E6 regulated genes were derived from table 1 in Duffy et al (Duffy et al., 2003). EMT genes were cancer EMT markers defined in table 1 in Zeisberg et al (Zeisberg and Neilson, 2009). First, for each gene in the pathway, we ranked the samples according to their expression levels. For each sample, the ranks of the genes were summed, and the resulting values are then centered by mean and scaled by standard deviation across samples to give the final scores. For E6 negatively regulated genes, we ranked the expression levels in descending order, because the direction of regulation is known to be opposite.

### 3.2.9 RNA-seq mutation calling

Pre-processing and variant calling were performed following GATK Best Practices for RNA-seq data (DePristo et al., 2011) for identification of single nucleotide polymorphisms (SNPs) and small indels. Briefly, paired-end reads were aligned to the human genome using STAR v2.3.0 (Dobin et al., 2013). GATK v3.2-2 was used for indel realignment and base recalibration. Variants were called for each sample using HaplotypeCaller. Variants which fell under any of the following criteria were filtered out: quality scores less than 25 (QUAL$<$25), strong strand bias (FS $>$30), normalized quality score (QD $<$2.0), or variants part of a SNP cluster (defined as 2 SNPs within 35 bps) indicating a false positive. Resulting variants were filtered by SnpEff v4.0e (DePristo et al., 2011) to retain variants predicted to disrupt the primary structure of the protein. Rare and damaging variants were identified using ANNOVAR (version released Nov12, 2014) (Wang et al., 2010). Rare variants were defined as variants that are present in less than 5% of 1000 Genomes (The 1000 Genome Project Consortium, 2012) and NHLBI Exome Sequencing Project subjects (Fu et al., 2012). Damaging variants were those predicted to be damaging by either PolyPhen-2

(Adzhubei et al., 2010) or SIFT (Pauline C. Ng and Steven Henikoff, 2003), stored in the ljb23 database. RNA-seq variant calling has the limitation that it can only uncover expressed mutations that is greater than a certain coverage. For PIK3CA showcased in the main text, all of the samples had an average coverage level that satisfied the minimum requirement (8X) to call variants. A gene is denoted as mutated if it contains at least one predicted damaging mutation. Gene mutation percent in a group is defined as the percent of samples containing mutated copies. The mutation percent is calculated for each subtype (HPV-KRT and HPV-IMU) in each cohort (UM and TCGA) separately.

#### 3.2.10 TCGA RNA-seq analysis

RNA-seq fastq files of 66 TCGA HPV+ tumor samples were downloaded from cghub (Wilks et al., 2014). The data were re-aligned and analyzed in the same way as UM RNA-seq data described above, except for variant calling. Gene somatic mutations for TCGA samples were instead downloaded from Xena (Goldman et al., 2015).

#### 3.2.11 CNA analysis

OncoSNP v2.1 was run on every tumor and matched normal, with the setting that the maximal possible stromal contamination is 0.5 (Yau et al., 2010). The level 1 through level 5 CNA output data from OncoSNP were overlayed to provide the final CNA calls. The copy number for each gene was calculated as the average of copy number of the segments overlapping the gene rounded to the nearest integer. CNA data for TCGA samples were downloaded from TCGA data portal.

## 3.3 Results

#### 3.3.1 Overview of differential expression results from HPV(+) and HPV(-) tumors

We performed transcriptomic analysis via RNA-deep sequencing on 36 tumor samples (18 HPV+ and 18 HPV-) to define gene expression levels. Supervised differential expression analysis

using HPV status as the group variable identified 1887 and 1644 genes significantly up-regulated and down-regulated in HPV(+) samples, respectively (FDR<0.05 and fold change>2). We annotated the genes to neoplasm-related terms downloaded from Gene2Mesh (Ade, AS; Wright, ZC; States, 2007), and reidentified several important differentially expressed genes: TP53, CDKN2A, BRCA2, CYP2E1, KIT and EZH2 were significantly up-regulated in HPV(+) tumors, and CCND1, GSTM1, HIF1A, MMP2, CD44 and MET were down-regulated. We performed Gene Ontology enrichment analysis with LRpath (Sartor et al., 2009) and found that "immune response", "cell cycle", and "DNA replication" were up-regulated in HPV(+) samples compared to HPV(-), whereas "extracellular matrix" and "epithelium development" were up-regulated in HPV(-) samples. These findings are consistent with what has been previously reported (Pyeon et al., 2007). Enrichment analysis with cytobands identified several locations on 11q (11q13, 11q22.3 and 11q23.3) as enriched for genes up-regulated in HPV(-) samples. This may be driven by frequent focal amplification of 11q13 and 11q22 and deletion of the far end of chr11q in HPV(-) samples (Fig 3.6) (The Cancer Genome Atlas Network, 2015).

### 3.3.2 Unsupervised clustering revealed two HPV(+) subgroups

Unsupervised clustering using the most variably expressed genes among all samples revealed three distinct groups (Fig 3.1A and Fig 3.7A). First, HPV(-) samples were distinguished from HPV(+) samples, except for one HPV(-) sample which clustered with HPV(+) samples. The 18 HPV(+) samples separated into two clusters of 8 and 10 HPV(+) samples. These clustering results were robust to various clustering metrics (Eucledian distance, centered or uncentered Pearson's correlation), linkages (single, average or robust) and different numbers of top variable genes used. To assure that the clusters we found are generalizable, not limited to our cohort, we included 66 additional HPV(+) samples from The Cancer Genome Atlas (TCGA) HNC cohort, and then performed unsupervised clustering on the 84 HPV(+) samples combined. Again, two clusters were

robustly identified [Fig 3.1B and Figure 3.7B]. The clusters were not correlated with smoking history, anatomical site, or tumor stage, but were correlated with gender and HPV type ($\alpha$-level = 0.05) (Fig 3.1A and Table 3.1).

### 3.3.3 Differentially regulated genes and pathways between HPV(+) subgroups

We next characterized the molecular differences between the HPV(+) subgroups. Differential expression analysis between the two HPV(+) clusters found 3515 genes significantly differentially expressed (absolute fold change$>$2 and FDR$<$0.05). Up-regulated genes in one cluster were enriched for "immune response", "mesenchymal cell differentiation" and various differentiation and development-related terms; up-regulated genes in the other cluster were most significantly enriched for "keratinocyte differentiation" and "oxidative reduction process". Therefore, we name the clusters HPV-IMU and HPV-KRT respectively. The top differentially expressed genes from each relevant gene set are shown in Fig 3.1C, including BCL2 for mesenchymal differentiation, CDH3 and TP63 for keratinization, and CDH1 and KRT16 for cell adhesion. To understand how the pathways were dysregulated in the context of all HNC tumors, we compared each HPV(+) cluster to the other cluster and HPV(-) samples, and identified pathways that were uniquely up- or down-regulated (see methods). Enrichment testing results showed remarkably elevated immune response in HPV-IMU consisting of increased T-cell activation, B-cell activation, lymphocyte activation, uniquely repressed mesenchymal differentiation and extracellular matrix-related expression in HPV-KRT; it also showed increased keratinization/epidermal differentiation and oxidative-reduction process gene expression in HPV-KRT relative to HPV-IMU, with mixed expression in the HPV(-) samples (Fig 3.1D). Multidimensional scaling analysis revealed that the HPV-KRT subgroup was overall more similar to the HPV(-) samples (Fig 3.1B). Although we were underpowered to compare most epidemiologic characteristics, we observed that HPV-KRT patients were more likely female and more likely to be type HPV16 compared to HPV-IMU patients.

### 3.3.4 HPV(+) subgroups correlate with HPV integration, E2/E4/E5 expression levels, full-length E6 percent and E6 activity.

We described HPV genic integration status and integration sites with VirusSeq (Chen et al., 2013b) using RNA-seq data for the HPV(+) tumors. Nine of the 18 UM and 41 of the 66 TCGA HPV+ samples were found to contain at least one genic HPV integration site, hereafter denoted as genic-integration. Surprisingly, cluster HPV-KRT had more samples with genic-integration (7 out of 10; 70%) than cluster HPV-IMU (2 out of 8; 25%). The same difference was observed for TCGA data, where cluster HPV-KRT had 32 out of 41 (78%) with a genic-integration and HPV-IMU had 9 out of 25 (36%) (Fig 3.2A). This difference in genic-integration was significant between subgroups (combined p-value=0.0001; Fisher's exact test).

We next tested whether any of the early HPV genes (E1, E2, E4, E5, E6, E7) were differentially expressed between the subgroups. Of these we found that E2, E4, and E5 had significantly lower expression in the HPV-KRT subgroup (E5 in Fig 3.2B, and E2 and E4 in Fig 3.8B,C). HPV integration frequently associates with loss of E2, E4, and/or E5 (zur Hausen, 2002). Since the HPV-KRT cluster had more genic integration events than HPV-IMU, this result is consistent with the difference in genic integration events. Upon closer examination, 4 of the 9 UM genic-integration-positive samples and 22 of the 41 TCGA samples displayed lost expression of E2/E4/E5 (Fig 3.8A).

It is known that E6 may be expressed in either of two main forms: a full-length E6 isoform and various spliced isoforms, together referred to as E6*, which have different functions not yet fully understood. We next asked whether the ratio of full-length E6 to total E6 expression correlates with subgroup membership (Fig 3.8D; see supplementary methods for details). Because the full-length E6 percentages differed significantly by HPV types (Fig 3.8E), we restricted our analysis to only the most prevalent HPV type, HPV16 (82% of all cases). HPV-KRT has significantly lower full-length E6 percent than HPV-IMU (Fig 3.2C) (Wilcox rank sum test p =0.001), suggesting

that HPV-KRT expresses less full-length E6 transcript and more spliced form E6*. Although the full-length E6 percent was different between subgroups, total E6 expression levels measured by RNA-seq was not (Fig 3.2D). Since the expression levels were quantified using RNA levels, it may not reflect the actual E6 protein activity level in the cell. We took advantage of a published study by Duffy et al (Duffy et al., 2003) of 51 genes (35 down and 16 up) regulated by E6, to calculate an E6 activity score for each tumor sample (see supplementary methods). Overall, the E6 score was significantly higher in HPV-IMU, indicating elevated E6 activity in HPV-IMU (Fig 3.2E,F; p=2.6e-7). Interestingly, the genes down-regulated by E6 in Duffy's study were especially more repressed in HPV-IMU than HPV-KRT (Fig 3.2E).

### 3.3.5 Correlation of subgroups with copy number alterations and PIK3CA mutation

Examining genomic properties of the two HPV(+) subgroups allowed us to further compare them with the HPV(-) HNC samples. Somatic copy numbers were obtained for all samples by analyzing SNP-array data from tumors and blood (see Methods). In our sample of 36 tumors, HPV(-) samples tended to harbor more copy number gains than HPV(+) samples (Fig 3.3A). In addition to the commonly observed copy number changes associated with HPV status (gain of 3q, 5p, 8q in HPV(-) tumors and loss of 11q and 13q in HPV(+) tumors) (The Cancer Genome Atlas Network, 2015), we found substantial differences between the two HPV(+) subgroups. Overall, HPV-KRT tumors tended to have more amplifications than HPV-IMU. Particularly at the chromosomal arm level, HPV-KRT had more amplifications on all or a significant portion of chr3q than subgroup HPV-IMU (p=1.7e-5, Wilcoxon test). In addition, HPV-IMU had frequent copy number loss on chr16q, which was completely absent in HPV-KRT and HPV(-) samples. Differences in chr3q CNA gain and chr16q CNA loss were also observed in the respective HPV(+) TCGA subgroups (Fig 3.3B). To determine if the CNAs affected gene expression differences, we asked what percent of the genes on chr3q and chr16q were up versus down-regulated. Out of the 693 and

427 genes on chr3q and chr16q, 148 (21%) and 132 (31%) genes were significantly up-regulated in HPV-KRT compared to HPV-IMU, respectively; in contrast, a much smaller percent of genes were down-regulated (7% and 5%) (Fig 3.3C,D). This is quite distinct from the opposite arms of these chromosomes, 3p and 16p, where nearly equal percentages of genes were up and down regulated. The same trends were observed for TCGA data (Fig 3.3C,D). This suggests that the gain and loss of copy numbers in part drove the expression differences between the two subgroups. Using neoplasm-related gene annotations from Gene2Mesh (Ade, AS; Wright, ZC; States, 2007), we found 33 oncology-related genes, including TNSF10, PIK3CA, TP63 and MUC4 on chr3q, and MMP2, CDH1, NQO1 and CDH13 on chr16q (Table 3.2). Nineteen of the 33 genes were also differentially expressed between the clusters.

To investigate if there were any differences in gene mutation frequencies between the two subgroups, we analyzed expressed, non-synonymous mutations from the RNA-seq data. We also obtained TCGA gene-level somatic mutation data from Xena UCSC (Goldman et al., 2015). None of the genes had a mutation difference greater than 20% in both cohorts, except for oncogene PIK3CA (Fig 3.4A). PIK3CA had a mutation in a striking 60% of samples from the HPV-KRT subgroup and 0% in HPV-IMU for the UM cohort. In TCGA, the difference was smaller (37% in HPV-KRT versus 16% in HPV-IMU), however still appreciable (Fig 3.4B). Five of the 6 UM PIK3CA mutations were known activating mutations (E545K, E545G and E542K), while the remaining mutation has unknown effect (E81K). Similarly, 17 out of the 19 mutations from TCGA were known activating mutations. Additionally, PIK3CA is located on chr3q, which was also found to have more amplifications in HPV-KRT. Not surprisingly, the copy numbers of PIK3CA were higher in HPV-KRT (Fig 3.4C). Together, these two results strongly suggest up-regulated PI3-kinase activity in the HPV-KRT tumors.

### 3.3.6 Characteristics associated with HPV(+) subgroups and patient survival

Summarizing and visualizing all of the above findings, we observe that although the clusters can be robustly separated by expression patterns, there remains substantial heterogeneity within each subgroup for each variable (Fig 3.5A). For example, we created an epithelial-to-mesenchymal transition (EMT) score for each tumor that combines the expression levels of the epithelial and mesenchymal differentiation genes (see Methods). While the HPV-KRT subgroup clearly had lower EMT scores overall, there was substantial variation in scores, especially among the TCGA cohort. HPV-IMU has higher levels of immune response (represented in Fig 3.5A by an overall T cell score), frequent chr16q loss, less viral integration in an expressed genic region, more E2/E5, and higher BCL2 gene expression. Except for immune response and a stronger EMT signature, all other attributes for HPV-IMU were correlated with better prognosis in the literature (Fig 3.5B and Discussion). Because the elevated immune response in HPV-IMU could be an indication of higher infiltration of cytotoxic T-cells, we calculated a T-cell activation score for each tumor based on 196 genes from the GO term "T-cell activation" (see Methods). Using overall survival data from TCGA, we found that TCGA HPV(+) tumors with high T-cell activation scores had better overall survival than those with low T-cell activation scores (Fig 3.5C; $p<0.05$). Although we did not find a significant difference in overall survival between HPV-KRT and HPV-IMU with the same TCGA data, HPV-KRT tended to have worse overall survival than HPV-IMU (Fig 3.5D), consistent with the general predictions from literature. Lack of a significant difference could be due to the poor follow-up data available from TCGA, or due to truly similar overall survival between subgroups. In either case, the many above-described differences strongly suggest that different therapies may best benefit patients within each subgroup.

## 3.4 Discussion

Head and neck squamous cell carcinomas represent a heterogeneous disease that consists of two molecular and clinically distinct entities distinguished by HPV infection. Due to a lack of HPV information and/or a small number of HPV(+) cases in most previous studies, attempts to identify tumor subtypes in HNCs have often neglected HPV as an important variable. Two published studies that have identified within-HPV(+) subtypes with high-dimensional genomics data (Keck et al., 2015; Pyeon et al., 2007) were unable to comprehensively describe molecular differences between the HPV(+) subtypes. Particularly, it remained unclear how these subtypes correlated with HPV characteristics, and the likely cause for their distinct behaviors. With unsupervised clustering analysis, we independently identified two robust HPV(+) subgroups with our cohort that persisted when we applied the same algorithm to the TCGA cohort. Although our clusters are constructed on RNA-seq data which has a better signal-to-noise ratio than microarrays, they share great similarity with the two HPV(+) clusters found in Keck et al. (Keck et al., 2015). We observed similar differences between the two clusters in immune response, mesenchymal differentiation and keratinization, and copy number changes in PIK3CA and TP63. In addition to providing an independent source validating the existence of two HPV(+) subtypes, we greatly expanded the depth of characterization by comprehensively profiling the expressed mutations, whole-genome CNAs and exploring HPV characteristics. Most importantly, our findings highlight two oncogenic paths in HPV(+) tumors that are likely driven by differences in HPV characteristics, chromosome-arm level CNAs, and PI3K pathway activity.

The most significant difference in expression between the HPV(+) clusters is the up-regulation of mesenchymal and immune-response genes in the HPV-IMU group, and keratinization and oxidation-reduction process in HPV-KRT. In fact, all of these pathways are explained by the biology of HPV carcinogenesis. The HPV E6 oncoprotein is reported to down-regulate a large num-

ber of genes involved in keratinocyte differentiation, and up-regulate genes normally expressed in mesenchymal lineages (Duffy et al., 2003). In addition, HPV type16 spliced variant E6* induces oxidative stress (Williams et al., 2014). Our analysis revealed that the HPV-IMU group had higher E6 activity whereas HPV-KRT had higher levels of the spliced forms of E6, E6*, which is concordant with the repressed keratinization and induced mesenchymal differential pathways in HPV-IMU and higher oxidation-reduction response in HPV-KRT. The HPV-KRT group was also more likely to have a detected HPV integration event, which may drive the expression of more E6* and less full-length E6. Our results regarding E6 activity scores also imply that the RNA-seq quantification of E6 expression levels may not adequately reflect E6 activity in the cells.

The higher immune response in HPV-IMU may be particularly relevant, as it was shown that HPV(+) HNCs have a different immune profile than their HPV(-) counterparts, featuring higher numbers of infiltrating CD8+ T lymphocytes, myeloid dendritic cell and proinflammatory chemokines (Partlová et al., 2015). Together, these are hypothesized to promote better response to treatment in HPV(+) patients (Partlová et al., 2015). In our study, we showed that stronger immune response could indeed predict better overall survival (Fig 3.5C). We hypothesize that when HPV shifts from the initial episomal form to an integrated transcribed form, the inflammatory/immune response towards HPV concurrently weakens. That may explain why HPV-IMU, having fewer HPV-integration events, have a stronger immune response. The stronger inflammatory/immune response in HPV-IMU group may partially explain why HPV(+) HNCs overall have better prognosis.

Another key finding of our study is that more chr3q amplifications were observed in HPV-KRT, whereas frequent chr16q deletions were found in HPV-IMU. These two signature copy number changes were evident in the independent UM and TCGA cohorts, validating these findings. However, the functional significance behind the associations of the copy number signatures with subgroups is unclear. One possible explanation could be that the change in copy number of a

few key genes on chr3q and chr16q favors the survival/growth of tumor cells in each respective subgroup, thus they were each positively selected in the tumor evolution of each subtype. We queried a tumor associated gene (TAG) database (Chen et al., 2013a) for the percent of oncogenes and tumor suppressor genes (TSG) on chr3q and chr16q, and interestingly, the majority of TAGs on chr3q are oncogenes (14 oncogenes, 4 TSGs, and 9 unknown) whereas most TAGs on chr16q are TSGs (12 TSGs, 1 oncogene, and 2 unknown). This preliminary result suggests that the amplification of chr3q and deletion of chr16q are both more likely to promote tumor growth either by increasing the copies of oncogenes or decreasing that of TSGs. Examining the genes more closely, we found that on chr3q, several key cancer genes stood out as having anti-apoptotic function, such as PIK3CA, TP63, MUC4, which may promote cancer cell survival for HPV-KRT as a result of chr3q amplification. Particularly Np63 (an isoform of p63) is dominantly over expressed in HNC and plays a pivotal anti-differentiation and anti-apoptosis role in the formation of HNC (Rothenberg and Ellisen, 2012). On chr16q, CDH1, CDH13, and BCAR1 were associated with cell adhesion, thus the attenuation of their expression by copy number loss in HPV-IMU likely strengthens EMT properties in this subtype.

When we investigated the mutation difference between the clusters, PIK3CA was found to be the top hit. PI3-kinase signaling pathway has previously been implicated in tumorigenesis and PIK3CA activating mutations have been found in various cancer types (Yuan and Cantley, 2008). We discovered that HPV-KRT subtype not only has more PIK3CA activating mutations, but also higher copy numbers of PIK3CA, suggesting elevated PIK3CA activity is important especially in the HPV-KRT group, and which may imply a difference in tumorigenesis between the two subgroups.

One limitation of our study is the lack of patient survival data. UM HPV(+) tumors were collected between 2011 - 2013, and due to high overall survival in this group, we are not yet able to perform meaningful risk analysis on these patients. The TCGA cohort has 66 HPV(+)

samples, however most of the patients were lost to follow-up after 12 months, therefore effectively reducing the sample size and conferring the analysis a much lower power. Nevertheless, HPV-KRT patients are thought to have worse outcome based on several lines of evidence. First, HPV-KRT's expression profiles (Fig3.1B) and copy number profiles, particularly gain on chr3q (Fig 3.3A), are more similar to those of HPV(-) patients, who are known to have worse treatment response and survival. HPV-IMU has more frequent chr16q loss, which was reported to correlate with better survival for patients with oropharyngeal SCCs (Klussmann et al., 2009). In addition, HPV-KRT has more detected virus integration events, which may be partially caused by a higher level of genomic instability (or vice versa). In cervical cancer, it was found that patients who only had integrated HPV expression had a trend towards worse disease free survival compared to patients with both integrated and episomal HPV forms (Shin et al., 2014). Studies on HNCs showed worse recurrence-free survival for HPV(+) tumors with low levels of E2 (Ramqvist et al., 2015) and E5 (Um et al., 2014), markers for integrated HPV. HPV-IMU has much higher levels of BCL2 expression than HPV-KRT, which has been reported to correlate with more favorable outcome in HNCs (Camisasca et al., 2009; Wilson et al., 2001). Altogether, this network of correlated events and pathways imply that the HPV-IMU and HPV-KRT subtypes proceed through carcinogenesis using different driving forces, and are thus likely to benefit from different treatment strategies. While patients falling in the HPV-KRT subgroup may benefit more from future immunotherapies, HPV-IMU may benefit from treatment deterring metastasis associated with EMT. Furthermore, our results strongly suggest that most of these differences are driven by the mode of HPV infection itself.

In conclusion, we identified two subtypes of HNC HPV(+) tumors based initially on expression patterns from RNA-seq data, but found to strongly correlate with a substantial number of important molecular markers, including viral characteristics, CNAs, and oncogenic PIK3CA mutations. This study takes a significant step towards decoding the molecular heterogeneity among HPV-associated

HNCs. Our work has important translational implications that could guide biomarker development and precision medicine for HPV(+) HNC patients.

# Figures

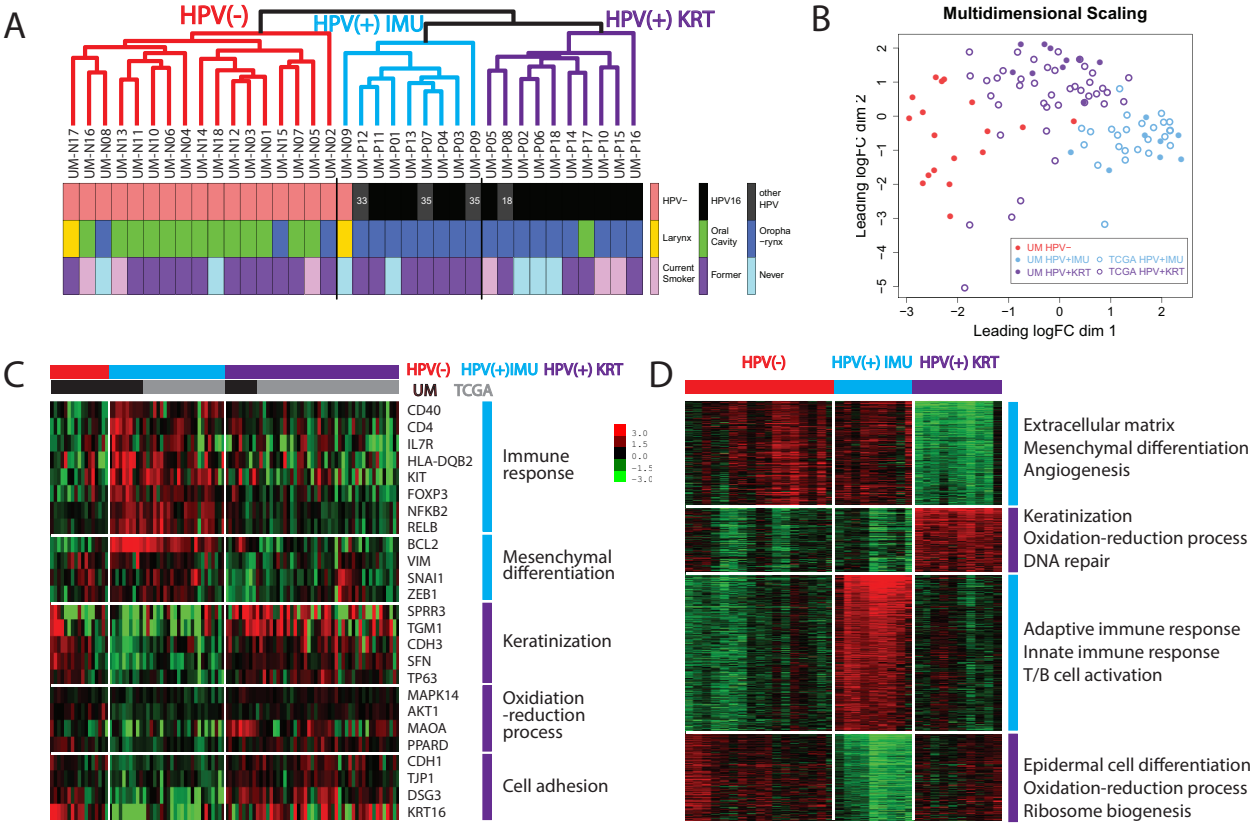



Figure 3.1: **Identification of two HPV(+) subgroups and pathway differences between them**. (A) Hierarchical clustering (shown here) and consensus clustering (Fig S1A) revealed two distinct HPV(+) clusters in the UM cohort. The clusters were not correlated with anatomical site or smoking. (B) Multi-Dimensional Scaling plot displaying the relationship among combined UM (n=36) and TCGA (n=66) samples by subgroup. The HPV-KRT subgroup is more similar to the non-HPV samples than is the HPV-IMU subgroup. (C) Heatmap showing representative genes/pathways different between HPV-KRT and HPV-IMU. (D) Heatmap showing the top differentially expressed genes (and pathways) among the three clusters. The genes were grouped by their expression signatures across the three clusters.

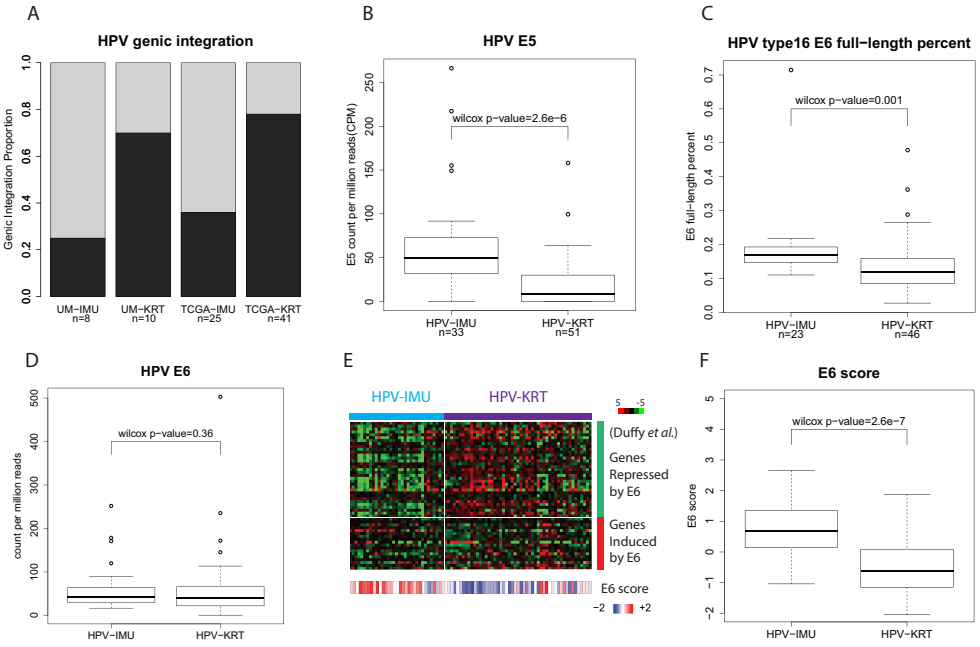



Figure 3.2: **HPV(+) subgroups correlate with several HPV characteristics**. (A) Barplot showing HPV-KRT tumors were more likely to have a detected genic-integration than HPV-IMU. (B) boxplot of HPV E5 expression levels; HPV-KRT had lower E5 expression than HPV-IMU. (See Fig S2B for plots of E2 and E4.) (C) boxplot of HPV E6 full-length percent for HPV type 16 samples. HPV-KRT had significantly lower E6 full-length percent than HPV-IMU. (D) barplot of HPV E6 expression levels. (E) Heatmap of the E6-regulated genes from Duffy et al., showing genes repressed by E6 were lower expressed in HPV-IMU. (F) boxplot of E6 activity score shows overall higher E6 activity in HPV-IMU.

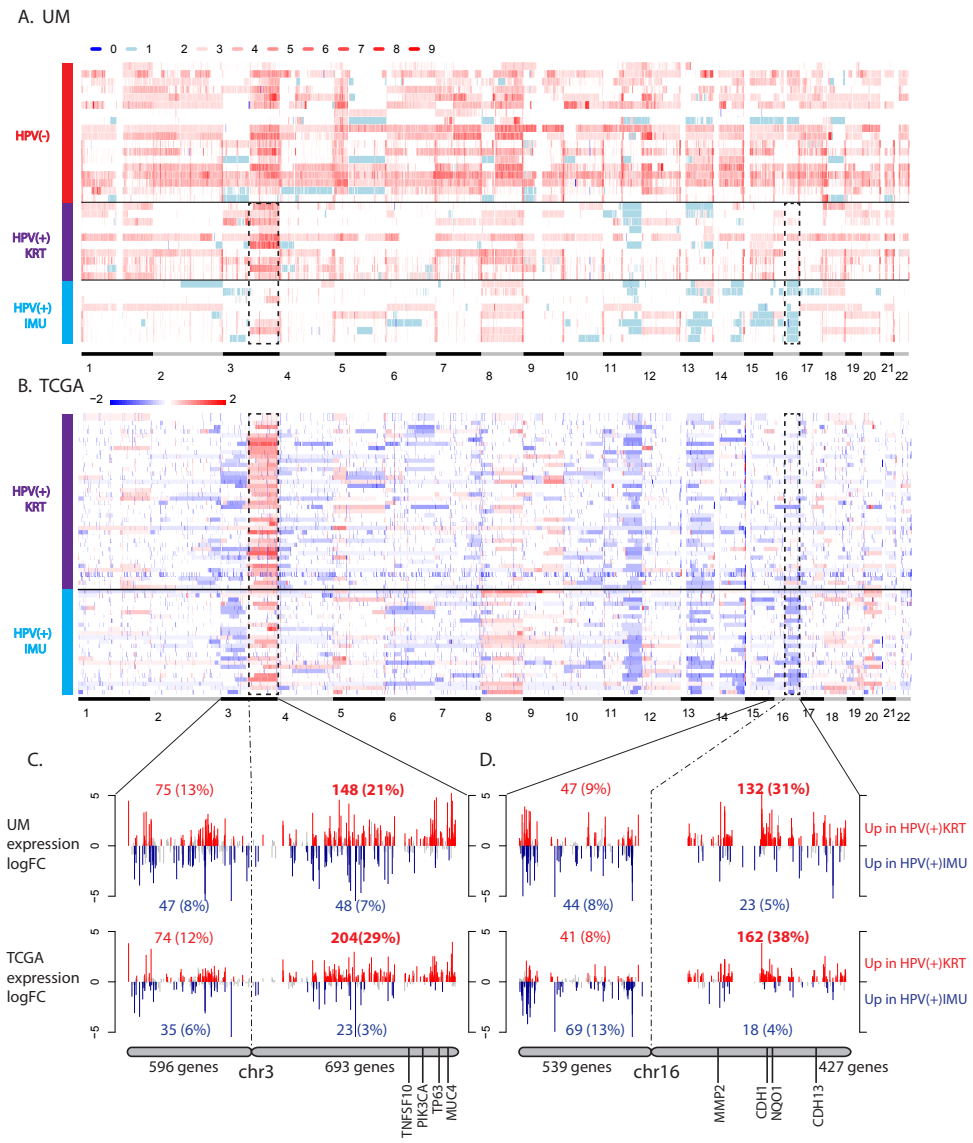



Figure 3.3: **HPV(+) subgroups differ by copy number alterations**. (A,B) Heatmaps of somatic copy numbers for UM (A) and TCGA (B) samples. The dark dashed lines highlight the regions that differ by HPV(+) subgroup (chr3q gain and chr16q loss). (C,D) Plots showing the expression fold changes of genes on chr3q (C) and chr16q (D). There were more up-regulated genes (shown in red) than down-regulated genes (shown in blue) in HPV-KRT compared to HPV-IMU due to the CNA difference. Known differentially expressed cancer-related genes are noted for each. Significance was defined byq-value $<0.05$.

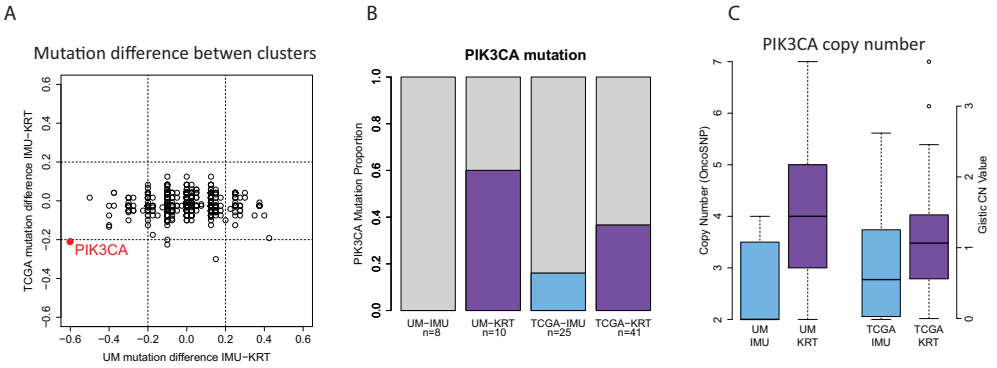
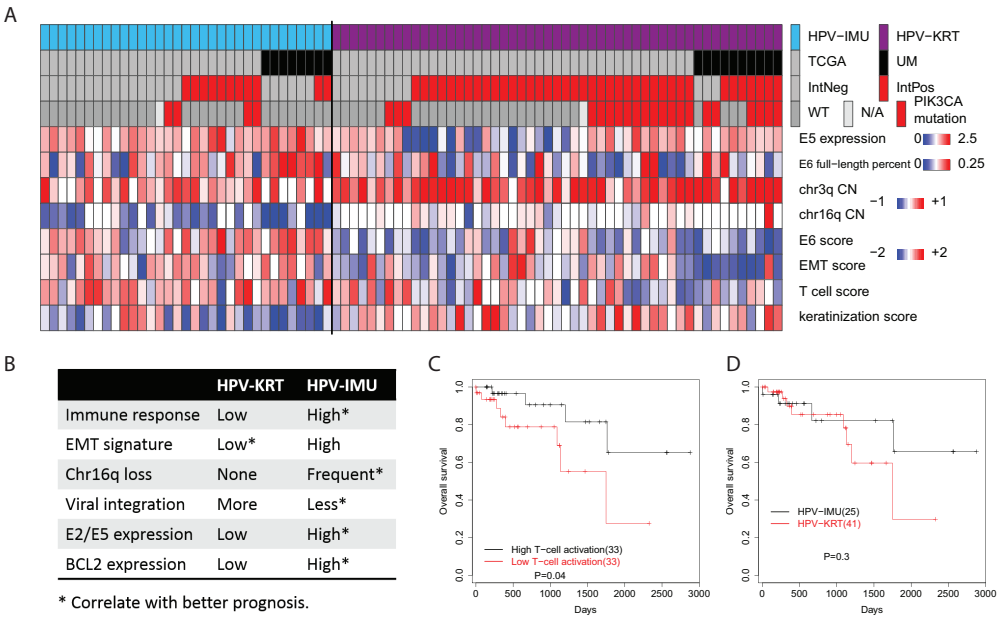



Figure 3.4: **The HPV-KRT subgroup had more PIK3CA mutations and amplifications**. (A) Scatterplot showing the difference in mutation rates between subgroups for each gene. PIK3CA was the only gene that had >20% difference in both cohorts. The difference was calculated by subtracting the mutation rate in HPV-KRT from that of HPV-IMU. (B) Barplot showing the PIK3CA mutation rates for each subgroup and cohort. (C) Boxplot showing that in both cohorts, HPV-KRT had more PIK3CA copy number amplifications than HPV-IMU.




| | HPV-KRT | HPV-IMU |
|---|---|---|
| Immune response | Low | High* |
| EMT signature | Low* | High |
| Chr16q loss | None | Frequent* |
| Viral integration | More | Less* |
| E2/E5 expression | Low | High* |
| BCL2 expression | Low | High* |

* Correlate with better prognosis.

Figure 3.5: **Summary of characteristics that differ by HPV(+) subgroup, and prognosis**. (A) Heatmap of variables that correlated with HPV(+) subgroup. The columns represent samples which are sorted by cluster, cohort, viral integration and PIK3CA mutation. E5 expression is the log10 transformed CPM values. (B) A table of observed features and associated publications suggesting better prognosis for HPV-IMU for all but EMT. (C) Overall survival for TCGA tumors with high and low T-cell activation scores. (D) Overall survival for HPV-KRT and HPV-IMU tumors from TCGA cohort. (C,D) P-values were calculated using a univariate Kaplan-Meier log rank test.

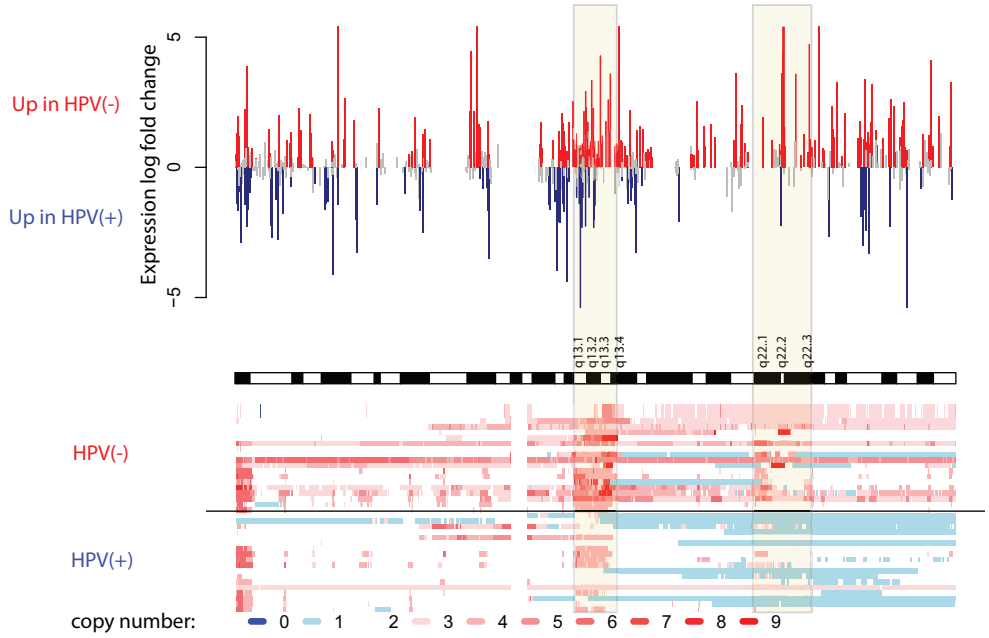



Figure 3.6: **Focal amplification of chr11q13 and chr11q22 in HPV(-) tumors and far-end deletion of chr11q in HPV(+) tumors**. The top panel shows the gene expression fold changes (log transformed) between HPV(-) and HPV(+) samples. Up-regulated genes in HPV(-) tumors is colored in red and up-regulated in HPV(+) tumors is in blue. The bottom panel shows the CNA of the tumors (divided in the middle by HPV status).

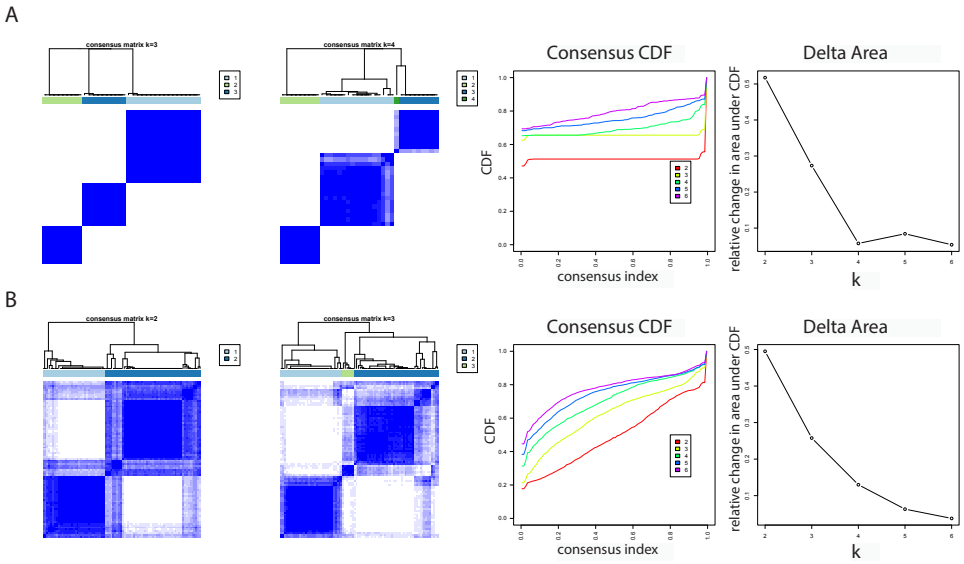



Figure 3.7: **ConsensusCluster Plus output for (A) UM and (B) UM+TCGA HPV(+) samples**. (A, B) The left-most figures show the tree structure for the selected k (k=3 for UM and k=2 for UM+TCGA). The second-to-the-left figures show the tree structure for k=(optimal K)+1. In both cases, the additional cluster only picks out trivial number of samples (one and five, respectively), providing strong evidence that the selected k is most appropriate. The 3rd figures (second to the right) show the consensus CDF for different k values, and the 4th (right-most) figures show the relative change in the area under CDF when k increases.

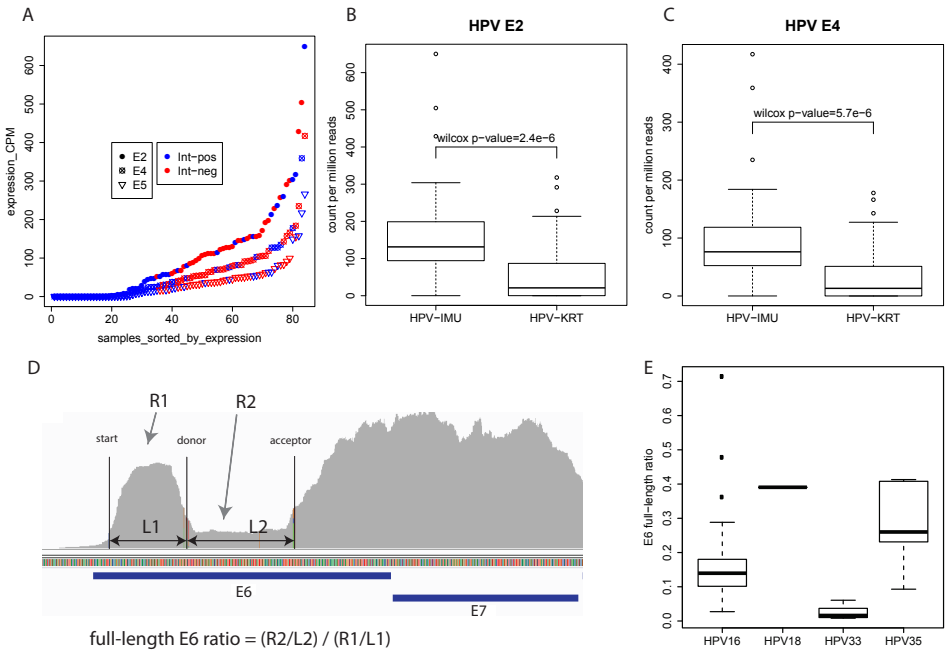



Figure 3.8: **Correlation of clusters with HPV characteristics**. (A) Plot of the expression levels of E2/E4/E5, using color to represent viral genic integration status. Each gene is sorted across samples by its expression value from low to high. (B,C) boxplots of HPV E2/E4 expression levels by subgroup. (D) Illustration of how full-length E6 percentage is calculated. R1 and R2 are the numbers of reads aligned to each interval. L1 and L2 are the lengths of the interval. R2/L2 is an approximation of full-length E6 transcript abundance, whereas R1/L1 is an approximation of the abundance of all E6 isoform. Full-length E6 percentage is calculated as in the formula shown. (E) Boxplot showing the full-length E6 percentage by HPV type.

# Tables

Table 3.1: **Patient demographics**. *tests were performed between HPV-KRT and HPV-IMU for both cohorts combined. Wilcoxon rank sum test was performed for age, whereas Fisher's exact test was used for other variables. For HPV type, HPV16 versus others combined was tested. For anatomical sites, only oropharynx versus oral cavity was tested due to insufficient tumors from other sites.

| | UM tumors | | | | TCGA tumors | | | p-value* |
|---|---|---|---|---|---|---|---|---|
| | Total | Non-HPV | HPV-KRT | HPV-IMU | Total | HPV-KRT | HPV-IMU | |
| | 36 | 18 | 10 | 8 | 66 | 41 | 25 | |
| **Age** | | | | | | | | |
| Median(std) | | 56.5(10) | 55.5(6.8) | 62.5(6.7) | | 58(10) | 57(7.8) | 0.38 |
| **Gender** | | | | | | | | |
| Male | 26 | 9 | 9 | 8 | 60 | 35 | 25 | 0.039 |
| Female | 10 | 9 | 1 | 0 | 6 | 6 | 0 | |
| **HPV type** | | | | | | | | |
| HPV16 | 14 | | 9 | 5 | 55 | 37 | 18 | 0.022 |
| HPV18 | 1 | | 1 | 0 | 0 | 0 | 0 | |
| HPV33 | 1 | | 0 | 1 | 8 | 3 | 5 | |
| HPV35 | 2 | | 0 | 2 | 3 | 1 | 2 | |
| **Anatomical Site** | | | | | | | | |
| Oropharynx | 20 | 3(17%) | 9(90%) | 8(100%) | 47 | 26(63%) | 21 (84%) | 0.051 |
| Oral cavity | 14 | 13(72%) | 1(10%) | 0 | 16 | 13(32%) | 3(12%) | |
| Larynx | 2 | 2(11%) | 0 | 0 | 1 | 0 | 1(4%) | |
| Hypopharynx | 0 | 0 | 0 | 0 | 2 | 2(5%) | 0 | |
| **Tumor stage** | | | | | | | | |
| I-II | 5 | 4 | 0 | 1 | 10 | 6 | 4 | 0.33 |
| III | 3 | 1 | 1 | 1 | 7 | 3 | 4 | |
| IV | 28 | 13 | 9 | 6 | 23 | 13 | 10 | |
| Unknown | | | | | 23 | 13 | 10 | |
| **Smoking** | | | | | | | | |
| Never | 7 | 3(17%) | 3(30%) | 1(13%) | 22 | 11(27%) | 11(44%) | 0.66 |
| Former | 23 | 12(66%) | 4(40%) | 7(87%) | 30 | 23(56%) | 7(28%) | |
| Current | 6 | 3(17%) | 3(30%) | 0 | 13 | 7(17%) | 6(24%) | |
| Unknown | | | | | 1 | | 1 | |

Table 3.2: **Neoplasm-associated genes on chr3q and chr16q (identified by gene2Mesh)**. The 'logFC', 'p-value' and 'q-value' columns are from testing the expression difference between HPV-KRT and HPV-IMU tumors (UM and TCGA combined).

| Symbol | chr | start | end | neoplasm count | logFC | p-value | q-value |
|---|---|---|---|---|---|---|---|
| PIK3CA | chr3 | 178866310 | 178952497 | 6 | 0.611423753 | 9.47E-05 | 0.000438548 |
| TP63 | chr3 | 189349215 | 189615068 | 4 | 1.008705532 | 9.37E-07 | 7.01E-06 |
| TNFSF10 | chr3 | 172235144 | 172241265 | 3 | 0.902624175 | 3.43E-05 | 0.000177248 |
| UMPS | chr3 | 124449212 | 124468119 | 3 | 0.810107441 | 6.98E-11 | 1.18E-09 |
| MUC4 | chr3 | 195473637 | 195538844 | 2 | 3.134042961 | 3.55E-10 | 5.33E-09 |
| MFI2 | chr3 | 196728611 | 196756687 | 2 | -0.197088928 | 0.379116911 | 0.492929089 |
| MME | chr3 | 154797435 | 154901518 | 2 | 0.588142839 | 0.0913372 | 0.157720074 |
| SOX2 | chr3 | 181429711 | 181432223 | 1 | -0.036159934 | 0.884715468 | 0.920307803 |
| CLDN1 | chr3 | 190023489 | 190040235 | 1 | -0.318512837 | 0.205996477 | 0.304042693 |
| TRPC1 | chr3 | 142443265 | 142526729 | 1 | -0.497394572 | 0.050235418 | 0.096794332 |
| GCSAM | chr3 | 111839687 | 111852152 | 1 | NA | NA | NA |
| CLDN18 | chr3 | 137729005 | 137752494 | 1 | NA | NA | NA |
| CCNL1 | chr3 | 156865585 | 156878482 | 1 | 0.477053726 | 0.000723821 | 0.002610024 |
| ABCC5 | chr3 | 183637723 | 183735727 | 1 | 1.077229262 | 3.59E-06 | 2.34E-05 |
| THPO | chr3 | 184089772 | 184095932 | 1 | NA | NA | NA |
| ACPP | chr3 | 132036210 | 132077690 | 1 | 1.432135514 | 3.05E-07 | 2.55E-06 |
| UPK1B | chr3 | 118892424 | 118924000 | 1 | -1.514829696 | 0.000973055 | 0.003368935 |
| TFG | chr3 | 100428174 | 100467811 | 1 | 0.312782591 | 0.006510867 | 0.017190072 |
| TM4SF1 | chr3 | 149086804 | 149095568 | 1 | 1.218675059 | 8.72E-09 | 1.00E-07 |
| CDH1 | chr16 | 68771194 | 68869444 | 13 | 1.038000997 | 8.17E-12 | 1.61E-10 |
| MMP2 | chr16 | 55515473 | 55540586 | 7 | -0.163938831 | 0.53907379 | 0.643135375 |
| NQO1 | chr16 | 69743303 | 69760533 | 4 | 2.290590728 | 1.80E-12 | 4.01E-11 |
| WWOX | chr16 | 78133326 | 79246564 | 2 | -0.563230306 | 2.52E-05 | 0.000134277 |
| CDH13 | chr16 | 82660398 | 83830215 | 2 | -0.605368955 | 0.038283574 | 0.077286453 |
| CDH11 | chr16 | 64980682 | 65155919 | 1 | -0.663114206 | 0.027828681 | 0.059312651 |
| ZFHX3 | chr16 | 72816785 | 73082274 | 1 | 0.078619508 | 0.666940877 | 0.75105827 |
| BCAR1 | chr16 | 75262927 | 75285526 | 1 | 1.217312903 | 1.38E-19 | 9.40E-18 |
| TUBB3 | chr16 | 89988416 | 90002505 | 1 | 1.270361004 | 9.00E-06 | 5.36E-05 |
| CYLD | chr16 | 50775960 | 50835846 | 1 | 0.302150962 | 0.027459531 | 0.058596546 |
| HPR | chr16 | 72097124 | 72111145 | 1 | NA | NA | NA |
| TOX3 | chr16 | 52471917 | 52581714 | 1 | -2.167580909 | 0.001362538 | 0.004512094 |
| CBFB | chr16 | 67063049 | 67134958 | 1 | 0.571606939 | 5.00E-08 | 4.94E-07 |
| MC1R | chr16 | 89984286 | 89987385 | 1 | 1.162756786 | 6.32E-09 | 7.51E-08 |

## Bibliography

Ade, AS; Wright, ZC; States, D. Gene2MeSH. 2007.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

Anders, S., Pyl, P. T. and Huber, W. HTSeq A Python framework to work with high-throughput sequencing data. *bioRxiv*, 31(2):002824, 2014.

Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. 2010.

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., Kim, H., Axelrod, R., Silverman, C. C., Redmond, K. P. and Gillison, M. L. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. *New England Journal of Medicine*, 363(1):24–35, 2010.

Arthur, A. E., Duffy, S. A., Sanchez, G. I., Gruber, S. B., Terrell, J. E., Hebert, J. R., Light, E., Bradford, C. R., D'Silva, N. J., Carey, T. E., Wolf, G. T., Peterson, K. E. and Rozek, L. S. Higher micronutrient intake is associated with human papillomavirus-positive head and neck cancer: a case-only analysis. *Nutrition and cancer*, 63(5):734–42, 2011.

Camisasca, D. R., Honorato, J., Bernardo, V., da Silva, L. E., da Fonseca, E. C., de Faria, P. A. S., Dias, F. L. and Lourenço, S. d. Q. C. Expression of Bcl-2 family proteins and associated clinicopathologic factors predict survival outcome in patients with oral squamous cell carcinoma. *Oral oncology*, 45(3):225–33, 2009.

Chen, J.-S., Hung, W.-S., Chan, H.-H., Tsai, S.-J. and Sun, H. S. In silico identification of onco-

genic potential of fyn-related kinase in hepatocellular carcinoma. *Bioinformatics (Oxford, England)*, 29(4):420–7, 2013a.

Chen, Y., Yao, H., Thompson, E. J., Tannir, N. M., Weinstein, J. N. and Su, X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics (Oxford, England)*, 29(2):266–7, 2013b.

Chung, C. H., Parker, J. S., Karaca, G., Wu, J., Funkhouser, W. K., Moore, D., Butterfoss, D., Xiang, D., Zanation, A., Yin, X., Shockley, W. W., Weissler, M. C., Dressler, L. G., Shores, C. G., Yarbrough, W. G. and Perou, C. M. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell*, 5(5):489–500, 2004.

Chung, C. H., Zhang, Q., Kong, C. S., Harris, J., Fertig, E. J., Harari, P. M., Wang, D., Redmond, K. P., Shenouda, G., Trotti, A., Raben, D., Gillison, M. L., Jordan, R. C. and Le, Q.-T. p16 Protein Expression and Human Papillomavirus Status As Prognostic Biomarkers of Nonoropharyngeal Head and Neck Squamous Cell Carcinoma. *Journal of Clinical Oncology*, 32(35):3930–3938, 2014.

Dayyani, F., Etzel, C. J., Liu, M., Ho, C.-H., Lippman, S. M. and Tsao, A. S. Meta-analysis of the impact of human papillomavirus (HPV) on cancer risk and overall survival in head and neck squamous cell carcinomas (HNSCC). *Head & Neck Oncology*, 2(1):15, 2010.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. and Daly, M. J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, 2011.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson,

M. and Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29:15–21, 2013.

Duensing, S., Lee, L. Y., Duensing, A., Basile, J., Piboonniyom, S., Gonzalez, S., Crum, C. P. and Munger, K. The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):10002–7, 2000.

Duensing, S. and Münger, K. The human papillomavirus type 16 E6 and E7 oncoproteins independently induce numerical and structural chromosome instability. *Cancer research*, 62(23):7075–82, 2002.

Duffy, C. L., Phillips, S. L. and Klingelhutz, A. J. Microarray analysis identifies differentiation-associated genes regulated by human papillomavirus type 16 E6. *Virology*, 314(1):196–205, 2003.

Duray, A., Lacremans, D., Demoulin, S., Delvenne, P. and Saussez, S. Prognosis of HPV-positive head and neck cancers: implication of smoking and immunosuppression. *Advances in Cellular and Molecular Otolaryngology*, 2, 2014.

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C. and Parkin, D. M. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, 127(12):2893–2917, 2010.

Fu, W., OConnor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., NHLBI Exome Sequencing Project and Akey, J. M. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, 2012.

Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., Haussler, D. and Zhu, J. The UCSC Cancer Genomics Browser: update 2015. *Nucleic acids research*, 43(Database issue):D812–7, 2015.

Hayes, D. N., Van Waes, C. and Seiwert, T. Y. Genetic Landscape of Human Papillomavirus-Associated Head and Neck Cancer and Comparison to Tobacco-Related Tumors. *Journal of Clinical Oncology*, 33(29):3227–3234, 2015.

Jeon, S., Allen-Hoffmann, B. L. and Lambert, P. F. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol*, 69:2989–2997, 1995.

Keck, M. K., Zuo, Z., Khattri, A., Stricker, T. P., Brown, C. D., Imanguli, M., Rieke, D., Endhardt, K., Fang, P., Bragelmann, J., DeBoer, R., El-Dinali, M., Aktolga, S., Lei, Z., Tan, P., Rozen, S. G., Salgia, R., Weichselbaum, R. R., Lingen, M. W., Story, M. D., Ang, K. K., Cohen, E. E. W., White, K. P., Vokes, E. E. and Seiwert, T. Y. Integrative Analysis of Head and Neck Cancer Identifies Two Biologically Distinct HPV and Three Non-HPV Subtypes. *Clinical Cancer Research*, 21(4):870–881, 2015.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013.

Klussmann, J. P., Mooren, J. J., Lehnen, M., Claessen, S. M. H., Stenner, M., Huebbers, C. U., Weissenborn, J. S., Wedemeyer, I., Preuss, S. F., Straetmans, J. M. J. a. a., Manni, J. J., Hopman, A. H. N. and Speel, E. J. M. Genetic signatures of HPV-related and unrelated oropharyngeal carcinoma and their prognostic implications. *Clinical Cancer Research*, 15(5):1779–1786, 2009.

Kreimer, A. R., Clifford, G. M., Boyle, P. and Franceschi, S. Human papillomavirus types in head

and neck squamous cell carcinomas worldwide: a systematic review. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 14(2):467–75, 2005.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(i):91–118, 2003.

Moody, C. a. and Laimins, L. a. Human papillomavirus oncoproteins: pathways to transformation. *Nature Reviews Cancer*, 10(8):550–560, 2010.

Parfenov, M., Sekhar, C., Gehlenborg, N., Freeman, S. S., Danilova, L., Pedamallu, C. S., Gehlenborg, N., Freeman, S. S., Danilova, L., Bristow, C. a., Lee, S., Hadjipanayis, a. G., Ivanova, E. V., Wilkerson, M. D., Protopopov, a., Yang, L., Seth, S., Song, X., Tang, J., Ren, X., Zhang, J., Pantazi, a., Santoso, N., Xu, a. W., Mahadeshwar, H. S., Wheeler, D. a., Haddad, R. I., Jung, J., Ojesina, a. I., Issaeva, N., Yarbrough, W. G., Hayes, D. N., Grandis, J. R., El-Naggar, a. K., Meyerson, M., Park, P. J., Chin, L., Seidman, J. G., Hammerman, P. S., Kucherlapati, R. and the TCGA network. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proceedings of the National Academy of Sciences*, 111:15544–15549, 2014.

Partlová, S., Bouček, J., Kloudová, K., Lukešová, E., Zábrodský, M., Grega, M., Fučíková, J., Truxová, I., Tachezy, R., Špíšek, R. and Fialová, A. Distinct patterns of intratumoral immune cell infiltrates in patients with HPV-associated compared to non-virally induced head and neck squamous cell carcinoma. *OncoImmunology*, 4(1):e965570, 2015.

Pauline C. Ng and Steven Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

Pim, D. and Banks, L. HPV-18 E6*I protein modulates the E6-directed degradation of p53 by binding to full-length HPV-18 E6. *Oncogene*, 18(52):7403–7408, 1999.

Pyeon, D., Newton, M. a., Lambert, P. F., Den Boon, J. a., Sengupta, S., Marsit, C. J., Woodworth, C. D., Connor, J. P., Haugen, T. H., Smith, E. M., Kelsey, K. T., Turek, L. P., Ahlquist, P., Den, B. J. A., Sengupta, S., Marsit, C. J., Woodworth, C. D., Connor, J. P., Haugen, T. H., Smith, E. M., Kelsey, K. T., Turek, L. P. and Ahlquist, P. Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Research*, 67(10):4605–4619, 2007.

Ramqvist, T., Mints, M., Tertipis, N., Näsman, A., Romanitan, M. and Dalianis, T. Studies on human papillomavirus (HPV) 16 E2, E5 and E7 mRNA in HPV-positive tonsillar and base of tongue cancer in relation to clinical outcome and immunological parameters. *Oral oncology*, 2015.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

Rothenberg, S. M. and Ellisen, L. W. The molecular pathogenesis of head and neck squamous cell carcinoma. *Journal of Clinical Investigation*, 122(6):1951–1957, 2012.

Sartor, M. A., Leikauf, G. D. and Medvedovic, M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics (Oxford, England)*, 25(2):211–7, 2009.

Shin, H.-J., Joo, J., Yoon, J. H., Yoo, C. W. and Kim, J.-Y. Physical Status of Human Papillomavirus Integration in Cervical Cancer Is Associated with Treatment Outcome of the Patients Treated with Radiotherapy. *PLoS ONE*, 9(1):e78995, 2014.

Slebos, R. J. Gene Expression Differences Associated with Human Papillomavirus Status in Head and Neck Squamous Cell Carcinoma. *Clinical Cancer Research*, 12(3):701–709, 2006.

Syrjanen, S. The role of human papillomavirus infection in head and neck cancers. *Annals of Oncology*, 21(Supplement 7):vii243–vii245, 2010.

The 1000 Genome Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–582, 2015.

Tindle, R. W. Immune evasion in human papillomavirus-associated cervical cancer. *Nature Reviews Cancer*, 2(1):59–64, 2002.

Tungteakkhun, S. S., Filippova, M., Fodor, N. and Duerksen-Hughes, P. J. The Full-Length Isoform of Human Papillomavirus 16 E6 and Its Splice Variant E6* Bind to Different Sites on the Procaspase 8 Death Effector Domain. *Journal of Virology*, 84(3):1453–1463, 2009.

Um, S. H., Mundi, N., Yoo, J., Palma, D. a., Fung, K., MacNeil, D., Wehrli, B., Mymryk, J. S., Barrett, J. W. and Nichols, A. C. Variable expression of the forgotten oncogene E5 in HPV-positive oropharyngeal cancer. *Journal of Clinical Virology*, 61(1):94–100, 2014.

Vokes, E. E., Weichselbaum, R. R., Lippman, S. M. and Hong, W. K. Head and Neck Cancer. *New England Journal of Medicine*, 328(3):184–194, 1993.

Wang, K., Li, M. and Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164, 2010.

Wang, L., Wang, S. and Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28:2184–2185, 2012.

Whiteside, M. A., Siegel, E. M. and Unger, E. R. Human papillomavirus and molecular considerations for cancer risk. *Cancer*, 113(S10):2981–2994, 2008.

Wichmann, G., Rosolowski, M., Krohn, K., Kreuz, M., Boehm, A., Reiche, A., Scharrer, U., Halama, D., Bertolini, J., Bauer, U., Holzinger, D., Pawlita, M., Hess, J., Engel, C., Hasenclever, D., Scholz, M., Ahnert, P., Kirsten, H., Hemprich, A., Wittekind, C., Herbarth, O., Horn, F., Dietz, A. and Loeffler, M. The role of HPV RNA transcription, immune response-related gene expression and disruptive TP53 mutations in diagnostic and prognostic profiling of head and neck cancer. *International journal of cancer. Journal international du cancer*, 00, 2015.

Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D., Litzinger, B., Hatton, T., Maltbie, L., Ainsworth, M., Allen, P., Rosewood, L., Mitchell, E., Smith, B., Warner, J., Groboske, J., Telc, H., Wilson, D., Sanford, B., Schmidt, H., Haussler, D. and Maltbie, D. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database*, 2014(0):bau093–bau093, 2014.

Williams, V. M., Filippova, M., Filippov, V., Payne, K. J. and Duerksen-Hughes, P. Human Papillomavirus Type 16 E6* Induces Oxidative Stress and DNA Damage. *Journal of Virology*, 88(12):6751–6761, 2014.

Wilson, G. D., Saunders, M., Dische, S., Richman, P., Daley, F. and Bentzen, S. M. bcl-2 expression in head and neck cancer: an enigmatic prognostic marker. *International Journal of Radiation Oncology*Biology*Physics*, 49(2):435–441, 2001.

Yau, C., Mouradov, D., Jorissen, R. N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O. and Holmes, C. C. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome biology*, 11(9):R92, 2010.

Yuan, T. L. and Cantley, L. C. PI3K pathway alterations in cancer: variations on a theme. *Oncogene*, 27(41):5497–510, 2008.

Zeisberg, M. and Neilson, E. G. Biomarkers for epithelial-mesenchymal transitions. *Journal of Clinical Investigation*, 119(6):1429–1437, 2009.

Zhou, X., Lindsay, H. and Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research*, 42(11):1–10, 2014.

zur Hausen, H. Papillomaviruses and cancer: from basic studies to clinical application. *Nature reviews. Cancer*, 2(5):342–50, 2002.

# CHAPTER IV

# Genomic binding and regulation of gene expression by the thyroid carcinoma-associated PAX8-PPARG fusion protein

## 4.1 Introduction

Thyroid carcinoma is the most common endocrine malignancy, and its incidence has increased nearly 3-fold since 1990 (Enewold et al., 2009; Howlader et al., 2015). The majority of thyroid carcinomas contain one of a small number of driver mutations, such as BRAF or RAS mutations, gene fusions involving RET, or gene fusions between PAX8 and PPARG (reviewed in (Vu-Phan and Koenig, 2014)). The PAX8-peroxisome proliferator-activated receptor gamma (PPARG) gene fusion is a consequence of a translocation between chromosomes 2 and 3, and is found in $\sim$ 30% of follicular thyroid carcinomas and $\sim$ 5% of follicular variant papillary carcinomas. The resulting PAX8-PPARG fusion protein (PPFP) is unusual in that it is the fusion of two transcription factors and it retains the DNA binding domains (DBDs) of both parent proteins (Kroll et al., 2000). Thus, at least in principle, PPFP should be capable of binding to PAX8 and PPARG response elements and potentially regulating target genes of both transcription factors. However, no data exist to define the genomic binding sites of PPFP, and the largest study characterizing global gene expression patterns in human PPFP carcinomas consisted of only 7 cases (Giordano et al., 2006).

The work in Chapter IV is published as **Zhang Y***, Yu J*, Lee C*, Xu B, Sartor MA, Koenig RJ. "Genomic binding and regulation of gene expression by the thyroid carcinoma-associated PAX8-PPARG fusion protein." *Oncotarget*. 2015 Dec 1;6(38):40418-32. doi: 10.18632/oncotarget.6340.(*equal contribution)

Given these limited data, the mechanism of oncogenesis is poorly understood (reviewed in (Raman and Koenig, 2014)).

PAX8 is a member of the paired box family of transcription factors and is essential for thyroid gland development (Macchia et al., 1998; Pasca di Magliano et al., 2000). In the mature thyroid, PAX8 drives the expression of numerous thyroid-specific genes (Pasca di Magliano et al., 2000). PPARG is a member of the nuclear receptor family of transcription factors. It has no identified role in the normal thyroid and is expressed at extremely low levels in that organ. PPARG is the master regulator of adipogenesis (Rosen et al., 1999), and also plays an important role in macrophage development, where it promotes an anti-inflammatory phenotype (Corzo and Griffin, 2013). Synthetic agonist ligands for PPARG such as pioglitazone are insulin sensitizers and hence are used to treat type 2 diabetes. PPARG ligands also are ligands for PPFP. In a mouse model of PPFP thyroid carcinoma, pioglitazone was highly therapeutic, greatly shrinking thyroid size and preventing metastatic disease (Dobson et al., 2011). Pioglitazone was strongly pro-adipogenic in these murine thyroid tumors, converting the thyroid cells into lipid-laden adipocyte-like cells. Although this indicates that PPFP is strongly PPARG-like in the presence of pioglitazone, the mechanism underlying the therapeutic efficacy of pioglitazone in this mouse model of PPFP thyroid carcinoma is not known.

There are no existing cell lines from PPFP thyroid carcinomas. However, PPFP has been stably expressed in the PCCL3 rat thyroid cell line at a level comparable to that in human thyroid cancers, herein denoted PPFP cells (Vu-Phan et al., 2013). PPFP expression confers upon PCCL3 cells an increased ability to invade through Matrigel and to form colonies in soft agar, both signs of cellular transformation (Vu-Phan et al., 2013). Thus, PPFP cells are a useful cell culture model to study PPFP-dependent oncogenesis, and potentially, the response to pioglitazone. PCCL3 cells also have been used to create cell culture models of thyroid carcinomas caused by oncogenic driver mutations in BRAF (Mitsutake et al., 2005) and RAS (Vitagliano et al., 2006), and RET gene fusions (Croyle

et al., 2008).

Here, we have used RNA deep sequencing (RNA-seq) to study the gene expression of PPFP cells versus control empty vector (EV) cells, cultured with and without pioglitazone. We also performed chromatin immunoprecipitation-deep sequencing (ChIP-seq) to identify the PPFP binding sites within the PCCL3 cell genome, and integrated the results with the gene expression data and publicly-available PAX8 and PPARG ChIP-seq data. The results provide novel insights into the transcriptional regulatory activity of PPFP, its oncogenic actions, and the response to pioglitazone.

## 4.2 Materials and methods

### 4.2.1 Cell culture

PCCL3-PPFP cells stably express human PPFP with a 3xMyc tag at the amino terminus, and PCCL3-EV cells have been stably transfected with the empty vector (Vu-Phan et al., 2013). PPFP and EV cells were cultured as previously described (Vu-Phan et al., 2013). In some experiments, the cells were treated with 1 M pioglitazone (from a 1 mM stock solution in DMSO) or vehicle for the times indicated prior to harvest.

### 4.2.2 Antibodies

Antibodies were obtained from the following sources as indicated: Cell Signaling Technology (Danvers, MA) beta actin #8457, Ccnb1 #12231, Hes1 #11988, Myc tag #2276, Notch1 #3608 and Plk3 #4896; Proteintech (Chicago, IL) Acaa2 #11111-1-AP and Icam5 #12759-1-AP; Sigma (St. Louis, MO) Foxe1 #SAB2100840; and Life Technologies (Grand Island, NY) Cdk1 #MA5-11472 and Plin1 #PA1-1051.

### 4.2.3 Flow cytometry

DNA content was analyzed by propidium iodide staining and oxidative stress was analyzed with CellROX Deep Red per the vendor's protocols (Life Technologies). For the oxidative stress

experiments, the cells were cultured $\pm$ 1 $\mu M$ pioglitazone for 2 days and $\pm$ 50 $\mu M$ TBHP for the final hour. Approximately 10,000 cells per condition were analyzed in the University of Michigan Comprehensive Cancer Center Flow Cytometry Core on a MACSQuant cytometer.

### 4.2.4 ChIP-seq assay

PPFP cells were cultured with pioglitazone for 16 hours, crosslinked with formaldehyde, sonicated to an average DNA size of 300 to 500 bp, and immunoprecipitated with anti-Myc antibody at 1:500 overnight at 4C using the protocol of Upstate Biotechnology (Lake Placid, NY), except that immunoprecipitation was performed with Dynabeads G (Life Technologies). ChIP and input DNA were used for next generation library construction and DNA sequencing on an Illumina HiSeq 2000 per the manufacturer's protocol using 50 nt single-end reads, performed by the University of Michigan DNA Sequencing Core. Four samples were barcoded and run on one lane, obtaining an average of 31 million reads per sample.

### 4.2.5 RNA-seq assay

PPFP and EV cells were treated with 1 $\mu M$ pioglitazone or vehicle for 16 hours. Total RNA was prepared using an RNeasy Mini Kit (Qiagen). Three independent experiments were performed. Library construction (Illumina TruSeq RNA) and sequencing on an Illumina HiSeq2000 using 50 nt paired end reads were per the manufacturer's protocols, performed by the University of Michigan DNA Sequencing Core. The samples were barcoded and loaded onto the same run, with all samples from each experiment run on the same two lanes; an average of 88 million reads were obtained per sample. PPFP ChIP-seq and RNA-seq data were deposited in Gene Expression Omnibus (GEO) with the accession ID GSE70354.

### 4.2.6 ChIP-seq data analysis

The quality of reads were assessed using FastQC (Andrews, 2015). There were 29.3 and 33.5 million reads sequenced for PPFP ChIP and input samples, respectively. ChIP-seq and input reads were aligned to the rat reference genome (rn4) using BWA (version 0.5.9-r16) with default options. MACS2 (2.0.10.20131216beta) was used to call peaks using a q-value $<0.05$ cutoff (Feng et al., 2012). Peaks were then filtered by PePr (1.0.5) (Zhang et al., 2014) to remove artifacts due to high PCR duplications. Peak boundaries were re-defined as 150 bp from the peak mode, and over-represented motifs were identified from the peaks by MEME(4.9.1), searching for the top 10 motifs with minimal width of 10bp and maximal width of 18bp. The most over-represented motif (shown in Fig 4.5A) was very close to PPARG motif previously reported (Lefterova et al., 2008; 2010; Nielsen et al., 2008) and was used as the position weight matrix (PWM) for the PPARG motif in all following analyses. Motif occurrences in the peaks were detected by FIMO(4.9.1) using default parameters and the PWM output from MEME. The presence of PPARG and PAX8 motifs in PPFP peaks was detected with the MEME/FIMO suite, and was also independently discovered using another motif discovery tool, HOMER (Heinz et al., 2010), as the top 2 known enriched motifs. Of the top 35 motifs found by HOMER, none were related to Myc, assuring that potential nonspecific binding caused by using a Myc tag antibody should be negligible.

Peaks were annotated to the genome with respect to gene features using an adapted HOMER script. If a peak had two or more annotations, a priority was assigned based on the order from left to right in Figure 4.4, as follows: -1 to + 1kb (relative to the TSS), -1 to -5kb, -5 to -10kb, exon, UTR, intron, and intergenic. We define "intergenic" as outside of the region between 10kb upstream from a TSS and its 3'UTR.

PAX8 ChIP-seq raw read data (GSE26871) were downloaded and analyzed as described above. The PAX8 motif (shown in Fig 4.5A) was found as the top hit by MEME searching for the top 10 overrepresented motifs with minimal width of 10bp and maximal width of 15bp. The

PAX8 motif identified here closely matches that previously published (Ruiz-Llorente et al., 2012). PPARG ChIP-seq peaks from mouse adipocyte and macrophage cells were downloaded from GEO (GSE21314).

### 4.2.7 RNA-seq data analysis

Quality checks were performed on RNA-seq reads with RSeQC (2.3.9) (Wang et al., 2012). The reads were aligned to rn4 with tophat2 (v2.0.11) and gene read counts were quantified by HTseq (0.6.1p1) (Anders et al., 2015) with option "-m intersection-strict" and normalized using the edgeR (3.2.4) Bioconductor package (McCarthy et al., 2012). Differential expression analysis was performed using edgeR with tagwise dispersion for each pairwise comparison of PPFP cells or EV cells treated with and without pioglitazone (four comparisons total). False discovery rate (FDR) was controlled using the Benjamini-Hochberg method (Yoav Benjamini, 1995).

### 4.2.8 Gene set enrichment testing

P-values from differential expression analysis from edgeR using the RNA-seq data were input into LRpath (Kim et al., 2012; Sartor et al., 2009) for gene set enrichment testing. LRpath is a logistic-regression-based method that models the relationship between the log-odds of genes belonging to a gene set and their -log(p-values). We used the directional test option in LRpath, and tested GO terms and KEGG pathways with each pairwise comparison of PPFP cells or EV cells treated with and without pioglitazone. Gene sets satisfying FDR $\leq 0.05$ were considered to be significant. Gene set enrichment testing of the ChIP-seq data was performed with ChIP-Enrich, a logistic-regression-based method that uses a smoothing spline to empirically adjust for gene locus length and mappability (Welch et al., 2014). Gene set enrichment results were corrected for multiple testing using the Benjamini-Hochberg FDR correction. Only gene sets with $\leq 500$ genes were reported, as gene sets with larger numbers of genes are more general and provide limited biological insight.

## 4.3 Results

### 4.3.1 Overview of genes regulated by PPFP in the absence and presence of pioglitazone

An RNA-seq analysis was performed on RNA from PPFP cells versus EV cells treated with or without pioglitazone. PPFP regulated the expression of 1541 genes (628 up, 913 down) in the comparison of PPFP cells versus EV cells without pioglitazone (FDR $<$0.05 and fold change $>$2). When both cell lines were cultured with pioglitazone, slightly more genes were differentially expressed (2078; 877 up, 1201 down). In a comparison of PPFP cells cultured with versus without pioglitazone, 250 genes were differentially expressed (95 up, 155 down). The differentially expressed genes in all of these comparisons are highly overlapping (Figure 4.1). In contrast, there were no differentially expressed genes in EV cells cultured with versus without pioglitazone, consistent with the very low expression level of endogenous PPARG in thyroid cells and the specificity of pioglitazone. Figure 4.1 shows that 156 of the 250 genes differentially expressed in PPFP cells cultured with versus without pioglitazone also are differentially expressed in PPFP cells versus EV cells cultured without pioglitazone. The PPFP and pioglitazone-induced changes are in the same direction for 130 (83%) of these 156 genes (48 up, 82 down), indicating that pioglitazone reinforces most of the PPFP-induced changes. However, for 26 genes (17%), the changes were in opposite directions such that pioglitazone partially or completely reversed the effects of PPFP.

### 4.3.2 PPFP regulates processes related to oncogenesis

Gene expression changes in PPFP cells versus EV cells in the absence of pioglitazone potentially are relevant to the oncogenic actions of PPFP. We subjected this comparison to a functional enrichment analysis using Gene Ontology (GO) terms and KEGG pathways (Kim et al., 2012; Sartor et al., 2009). We identified 162 enriched gene sets (FDR$<$0.05), 55 of which were induced by PPFP and 107 repressed. The 15 induced and 15 repressed gene sets with the lowest q-values are shown in Table 4.1. Many of the induced gene sets involve processes directly related to cancer biol-

ogy. For example, gene sets related to the cell cycle include MCM complex, deoxyribonucleotide biosynthetic process, DNA replication, and others. Three cell cycle-related genes within these gene sets (Ccnb1, Cdk1 and Plk3) were investigated further and were also found to be induced at the protein level by PPFP (Figure 4.2). Consistent with the enrichment of cell cycle-related gene sets, cellular DNA content analysis by flow cytometry demonstrated that a greater fraction of PPFP cells than EV cells are in the S and G2/M phases of the cell cycle, and a lesser fraction are in G1 (Figure 4.3).

Other highly significant cancer-related processes were related to proteasome/protein folding, immune function and oxidative stress (Table 4.1). Gene sets related to mitochondria/lipids also were enriched, consistent with PPARG-like activity of PPFP. We confirmed the induction by PPFP of two such PPARG target genes, Acaa2 and Plin1, at the protein level (Figure 4.2). Twenty-two of the 107 repressed gene sets contain the word morphogenesis, differentiation or development (Table 4.1), consistent with the expectation that, as an oncogene, PPFP enforces a less differentiated state. The repressed genes include several involved in thyroid differentiation, including Fgfr2 (Celli et al., 1998), Hhex (Martinez Barbera et al., 2000), Foxe1 (Clifton-Bligh et al., 1998), Hes1 (Carre et al., 2011) and Notch1 (Porazzi et al., 2012), the latter three of which were confirmed at the protein level (Figure 4.2). The repressed gene sets also include cell adhesion, extracellular matrix, and several related terms. Repressing these gene sets could facilitate invasion and metastases. We confirmed PPFP-dependent repression of the cell adhesion protein Icam5 at the protein level (Figure 4.2).

### 4.3.3 PPFP can induce or repress PAX8-regulated genes

PAX8 induces thyroid-specific genes such as Tg, but only limited data exist to define PAX8-responsive genes more broadly. In a previous publication (Ruiz-Llorente et al., 2012), siRNA knockdown of PAX8 in PCCL3 cells yielded 601 differentially expressed genes that also were tested in our data set. In general the magnitude of change was modest in the siRNA experiment,

with 296 genes showing a fold change >1.2 (siPAX8 induced 175 genes and repressed 121). We determined what fraction of these siPAX8-responsive genes was differentially expressed in PPFP cells versus EV cells cultured without pioglitazone (using FDR<0.05 and fold change >1.5 as cut-offs). As shown in Table 4.2, slightly more than half of siPAX8-regulated genes are regulated by PPFP, and the direction of regulation is discordant about 2/3 of the time (p=0.00015 for observing this level of discordance by chance; Fisher's exact test). Since induction by siPAX8 implies repression by PAX8 and vice versa, the data indicate that PAX8 and PPFP regulate gene expression in the same direction for $\sim$ 2/3 of the genes, and in opposite directions for $\sim$ 1/3.

PPFP regulates genes related to fatty acid metabolism and mitochondrial function, especially in the presence of pioglitazone The 55 induced gene sets in the comparison of PPFP cells versus EV cells in the absence of pioglitazone include several related to mitochondria, fatty acids and lipids, such as mitochondrial envelope, lipid particle, and cellular response to fatty acid (Table 4.1). The induced genes in these gene sets include adipocyte PPARG target genes such as Acaa2 and Plin1, demonstrating that PPFP is PPARG-like on a subset of target genes. This is consistent with the fact that several PPARG target genes have been shown to be induced in human PPFP thyroid carcinomas (Giordano et al., 2006). However, the PPARG-like activity of PPFP is much more striking in the presence of pioglitazone. We found that 117 gene sets are enriched in the comparison of PPFP cells cultured with versus without pioglitazone (52 induced, 65 repressed). The 10 most significant gene sets are all induced by pioglitazone, and all relate to fatty acid metabolism and PPAR activity (Table 4.3). Among the PPARG target genes in these gene sets, we confirmed the inductions of Acaa2 and Plin1 at the protein level, as noted previously (Figure 4.2).

### 4.3.4 Overview of the PPFP cistrome

We performed ChIP-seq analysis on PPFP to begin to understand the DNA binding properties of PPFP and the genes it is likely to regulate through direct interactions. Using an FDR <0.05, we

identified 20,277 PPFP peaks in the PCCL3 cell genome. As has been found previously for PAX8 (Ruiz-Llorente et al., 2012) and PPARG (Lefterova et al., 2008), most PPFP peaks are intergenic. However, we observed an enrichment of PPFP peaks in genic regions, and most strikingly within 1 kb of transcription start sites (TSS's) (2.9-fold enriched) (Figure 4.4). PPFP peaks also are enriched 2-fold from -5 to -1 kb of TSS's and 1.7-fold in first introns, and are under-represented (0.8-fold) in intergenic regions. Since PPFP contains DBDs from both PAX8 and PPARG, in principle it could bind to the DNA motifs recognized by both transcription factors. This is what was observed, as we identified the PPARG and PAX8 motifs de novo as the top two most overrepresented sequences within the peak regions using HOMER (Heinz et al., 2010). Overall, 65% of PPFP peaks contain a PPARG motif and/or a PAX8 motif, and these partially overlap (Figure 4.5A). An unexpected finding was that 50% of the peaks with a PAX8 motif also contain a PPARG motif. This is interesting because a much lower rate of PPARG motifs would be expected near PAX8 motifs if PPFP uses only one of the two DBDs for every binding site. To investigate if the co-localizations of the two motifs were due to false-positive matches of the motif position weight matrices to the sequences, we examined the motif locations relative to the peak centers. The results show that both the PAX8 and PPARG motifs are centered within these PPFP peaks, as would be expected if both are functionally relevant to DNA binding (Figure 4.5B). This suggests that PPFP prefers to bind to the subset of PAX8 motifs that have nearby PPARG motifs. To evaluate this further, we took advantage of the fact that a ChIP-seq analysis has been published for PAX8 in PCCL3 cells (Ruiz-Llorente et al., 2012). We first filtered the PAX8 peaks by whether they contained at least one PAX8 motif, to remove potential false positive peaks that could confound the analysis. We then asked what fraction of the PAX8 peaks that contain a PAX8 motif and do or do not overlap with our PPFP peaks also contain a PPARG motif. As shown in Figure 4.5C, PPARG motifs are enriched in the PAX8 peaks to which PPFP also binds (odds ratio=1.9, p-value$<$2.2e-16, Fisher's exact test), confirming that PPFP preferentially binds to the subset of PAX8 peaks that also contain

a PPARG motif.

### 4.3.5 Overview of genes and gene sets containing PPFP peaks

PPFP peaks were found to encompass a number of known functional response elements in classic PAX8 and PPARG responsive genes. For example, PPFP peaks encompass the PAX8 response element in the Tg promoter (Zannini et al., 1992) and the PPARG response element in the Aqp7 promoter (Kishida et al., 2001). One hundred sixty eight GO terms were identified as enriched with PPFP peaks (FDR$<$0.05), after associating peaks with the gene having the nearest TSS. Nine of the 15 GO terms with the lowest q-values are related to immune function, development/differentiation, or lipid metabolism, and the GO terms in the full list include additional cancer-related concepts such as negative regulation of programmed cell death, regulation of cell migration, G1/S transition of mitotic cell cycle, and Wnt receptor signaling pathway. In subsequent analyses, we focused on gene sets that are both enriched in the ChIP-seq analysis and enriched among the differentially expressed genes by RNA-seq analysis. Eight such gene sets were induced in PPFP cells versus EV cells cultured without pioglitazone, and 13 were repressed (Table 4.4). Seven of the 8 induced gene sets relate to mitochondria, and include direct PPARG target genes such as Plin1. The one additional induced gene set, G1/S transition of mitotic cell cycle, includes genes such as Plk3 that promote progression through the cell cycle and cell division. The inductions of Plin1 and Plk3 were confirmed at the protein level, as noted previously (Figure 4.2). In contrast, the 13 repressed gene sets relate mostly to protein signaling (4 gene sets), morphogenesis/development/differentiation (4 gene sets), and cell communication/extracellular matrix/adhesion (4 gene sets). The repressed genes in these gene sets include the thyroid development genes Fgfr2 (Celli et al., 1998) and Hhex (Martinez Barbera et al., 2000).

PPFP functions through its PAX8 DBD to repress gene sets when bound at a distance from transcription start sites, but to induce gene sets when bound close to transcriptional start sites We

classified PPFP peaks as to whether they contain a PAX8 motif or a PPARG motif, and whether they are $\leq$10kb from a TSS or $>$10kb upstream from a TSS. We then performed gene set enrichment analyses on these 4 groups of PPFP peaks using ChIP-Enrich (Welch et al., 2014) and compared the results with the RNA-seq analysis of PPFP cells versus EV cells without pioglitazone. Analysis of PPFP peaks with PAX8 motifs $>$10kb upstream from TSS's yielded no gene sets that were enriched with ChIP-seq peaks and induced by PPFP. However, 16 gene sets were enriched and repressed by PPFP, encompassing 59 unique genes. This suggests that, when regulating gene sets through PAX8 motifs distant from the TSS, PPFP tends to act in a repressive manner. Furthermore, 11 of these 16 gene sets contain the words morphogenesis, development or organ formation (Table 4.5), implying that the effects of PAX8 binding at a distance $>$10kb upstream from TSS's are primarily anti-differentiation effects of PPFP. In contrast, analysis of PPFP peaks with PAX8 motifs $\leq$10kb from TSS's yielded 4 gene sets that were enriched in the ChIP and induced by PPFP, and no gene sets that were enriched and repressed. Thus, at the gene set level, there is a complete separation of PPFP as a repressor when the target genes have peaks with PAX8 motifs $>$10kb upstream from the TSS, versus an activator when the PAX8 peaks are $\leq$10kb. The 4 induced gene sets all relate to mitochondria and lipids (Table 4.6), and contain 24 unique genes. In contrast to the analysis of PPFP peaks with PAX8 motifs, analysis of peaks with PPARG motifs did not identify differences in gene set activation versus repression based on distance from the TSS (data not shown).

### 4.3.6 Why is pioglitazone adipogenic in PPFP-expressing cells?

When mice with PPFP thyroid carcinomas are treated with pioglitazone, metastatic disease is prevented and the primary thyroid tumors shrink markedly (Dobson et al., 2011). The most striking part of the response is that pioglitazone is highly adipogenic, causing large accumulations of intracellular lipid and the induction of numerous adipocyte PPARG target genes in the thyroids. In contrast, pioglitazone has no effect on the thyroid glands of control mice. The induction of adipocyte

genes is a hallmark of the pioglitazone response in cultured PPFP cells as shown here, too (Table 4.3). Since PPARG is the master regulator of adipogenesis, the data indicate that pioglitazone turns PPFP into a strongly PPARG-like transcription factor. However, PPARG also is expressed in macrophages, where it plays an important role in promoting an anti-inflammatory "M2" phenotype (Corzo and Griffin, 2013). Why does PPFP favor the induction of an adipocyte phenotype over a macrophage phenotype in the thyroid? To begin to understand this, we took advantage of the fact that a PPARG ChIP-seq analysis has been published comparing a mouse adipocyte cell line with mouse macrophages (Lefterova et al., 2010). This study thus identified genes with PPARG peaks in mouse adipocytes but not macrophages, and vice versa. Using HomoloGene, we assessed the overlap between mouse genes with a nearby ($\leq$10kb from TSS) PPARG peak and rat genes with at least one nearby PPFP peak. We found that PPFP binds near 34% of homologs with an adipocyte PPARG peak versus only 25% of homologs with a macrophage PPARG peak (Supplemental Table 4.7, p=0.0022). The fact that PPFP in the thyroid preferentially binds to adipocyte PPARG target genes likely underlies the observation that the pioglitazone response is adipocyte-like.

### 4.3.7 Why is pioglitazone therapeutic in the mouse model of PPFP thyroid carcinoma?

We reasoned that genes or pathways regulated in opposite directions by PPFP without pioglitazone versus PPFP with pioglitazone may be involved in the therapeutic efficacy of this drug. We therefore identified GO or KEGG terms in our RNA-seq data that are both induced (or repressed) in PPFP cells versus EV cells without pioglitazone, and repressed (or induced) in PPFP cells with pioglitazone versus PPFP cells without pioglitazone. Only three gene sets qualified, and all were induced by PPFP without pioglitazone and repressed by pioglitazone. The 3 gene sets relate to oxidative stress: Glutathione metabolism (KEGG), peroxidase activity, (GO) and arachidonic acid metabolism (KEGG). There are 49 unique, differentially expressed genes within these gene sets, including multiple glutathione peroxidases, glutathione reductase, glutathione synthase,

glutathione S-transferases, and peroxiredoxins (Supplemental Table 4.8). Based on these data, we hypothesized that PPFP induces oxidative stress and that pioglitazone impairs the ability of the cell to mount an appropriate antioxidant response.

To assess this, we evaluated oxidative stress in PPFP and EV cells by flow cytometry after treatment with the reactive oxygen species (ROS)-sensitive dye CellROX Deep Red. PPFP cells had greater ROS than EV cells, cultured in the absence of pioglitazone (Figure 4.6A). Analysis of PPFP cells cultured a low dose (50 $\mu M$) of the reactive peroxide tert-butyl hydroperoxide (TBHP) and pioglitazone showed that pioglitazone increased ROS in the presence of TBHP (Figure 4.6B) but not in its absence (Figure 4.6C). In contrast, in EV cells, pioglitazone did not increase ROS regardless of the presence or absence of TBHP (Figures 4.6D,E). These data support the hypothesis that PPFP induces oxidative stress, and that pioglitazone increases the susceptibility of PPFP cells to further oxidative stress. The data suggest that the therapeutic efficacy of pioglitazone in PPFP thyroid cancer may at least in part relate to synergism with cellular stressors to induce oxidative stress, cytotoxicity and ultimately, cell death.

### 4.3.8 Similarity of gene regulation by PPFP in PCCL3 cells and human thyroid carcinomas

To judge the clinical relevance of the PPFP cell ChIP-seq and RNA-seq analyses, it would be ideal to compare these data to similar data from human PPFP carcinomas. However, human ChIP-seq data would be difficult if not impossible to obtain due to the fact that antibodies to endogenous PPFP also bind to PAX8 or PPARG (our PPFP is epitope-tagged), as well as the fact that PPFP thyroid carcinomas are uncommon. Furthermore, the largest gene expression profiling study of human PPFP thyroid carcinomas included only 7 cases (Giordano et al., 2006). This study identified 275 genes that were differentially expressed in 7 PPFP follicular carcinomas versus 82 non-PPFP thyroid tumors and 4 normal thyroids. Of those 275 genes, 264 have rat homologs, and we found a 31% overlap with differential expression in our RNA-seq data and 22% overlap with our ChIP-

seq data. This overlap includes numerous PPARG target genes, including ANGPTL4 and AQP7, which were two of the six most highly induced genes in the human PPFP carcinomas. We also found that the set of 49 ROS-related genes described above is induced in the 7 PPFP follicular carcinomas versus the non-PPFP follicular carcinomas (p= 0.0007, Fisher's exact test, Supplemental Table 4.9), suggesting that PPFP also causes increased ROS in human thyroid carcinomas.

## 4.4 Discussion

PPFP is an unusual oncoprotein in that it is the fusion of two transcription factors, PAX8 and PPARG, and it retains the DBDs of both parent proteins. However, until now there were no studies to determine whether PPFP actively bound both PAX8 and PPARG target genes. Gene expression data in PPFP thyroid carcinomas also are very limited – the largest study included only 7 PPFP carcinomas (Giordano et al., 2006), and given patient heterogeneity, only limited conclusions can be drawn.

Here, we identified $\sim$ 20,000 putative PPFP binding sites in the rat PCCL3 cell line genome, and found that these binding sites are enriched within -5 kb of transcription start sites and in first introns. PPFP peaks encompass known PAX8 and PPARG binding sites, indicating that both DBDs within PPFP are functional. Interestingly, PPFP preferentially binds to the subset of PAX8 peaks that also contain PPARG motifs, implying that the PPARG portion of PPFP is particularly important in directing PPFP to its target genes. The fact that many PAX8 binding sites have a nearby PPARG motif also suggests PPARG may play a role in normal thyroid biology. Thyroid-specific deletion of murine PPARG has not been reported, but would provide a means to address this question.

RNA-seq analysis shows that PPFP regulates the expression of $\sim$ 1500 genes in the absence of pioglitazone and $\sim$ 2000 genes in its presence. In general, pioglitazone reinforces PPFP-dependent gene expression, but in 17% of cases it reverses the effects of PPFP. PPFP regulates many genes

known to be regulated by PAX8 in thyrocytes or PPARG in adipocytes. For the latter, the gene regulation by PPFP is particularly striking in the presence of pioglitazone. Although gene expression data in human PPFP thyroid carcinomas are very limited, there is excellent overlap between the human data and our data, including the induction of PPARG target genes.

We identified an unusual dichotomy in the function of PPFP for genes with peaks containing PAX8 motifs. When such motifs are located >10kb upstream from the TSS, the functional consequence tends to be gene set repression, including processes related to morphogenesis and development. In contrast, when such peaks are close to the TSS, the consequence tends to be gene set induction. This dichotomy likely reflects the summed activity of transcriptional activators and repressors brought to the target gene, but the factors that determine these differences are unknown.

The data are relevant to the biology of PPFP in human thyroid cancer. For example, PPFP induces gene sets related to the cell cycle, and represses gene sets related to differentiation. In a transgenic mouse model of PPFP thyroid carcinoma, pioglitazone was highly therapeutic (Dobson et al., 2011), and this has led to a clinical trial in patients (clinicaltrials.gov identifier NCT01655719). A remarkable aspect of this response is that the drug trans-differentiated the surviving thyroid cancer cells into adipocyte-like cells. Our ChIP-seq data show that, in PCCL3 thyroid cells, PPFP binds to adipocyte PPARG target genes in preference to macrophage PPARG target genes, likely explaining why pioglitazone is specifically pro-adipogenic.

It is plausible that the therapeutic efficacy of pioglitazone is at least in part due to its adipogenic pro-differentiation effects. However, our data uncovered another potential contributing factor. We found that PPFP cells have higher expression of ROS-related genes than EV cells, and this reflects higher levels of oxidative stress. Pioglitazone caused the PPFP cells to develop even greater levels of ROS when exposed to a low dose of the reactive peroxide TBHP, indicating that pioglitazone sensitizes PPFP cells to potential oxidant stressors. These data suggest pioglitazone might sensitize PPFP thyroid cancers in vivo to oxidative stressors, leading to increased cytotoxicity and cell death.

Substantiation of this hypothesis could lead to approaches to further enhance the efficacy of this drug.

## Figures

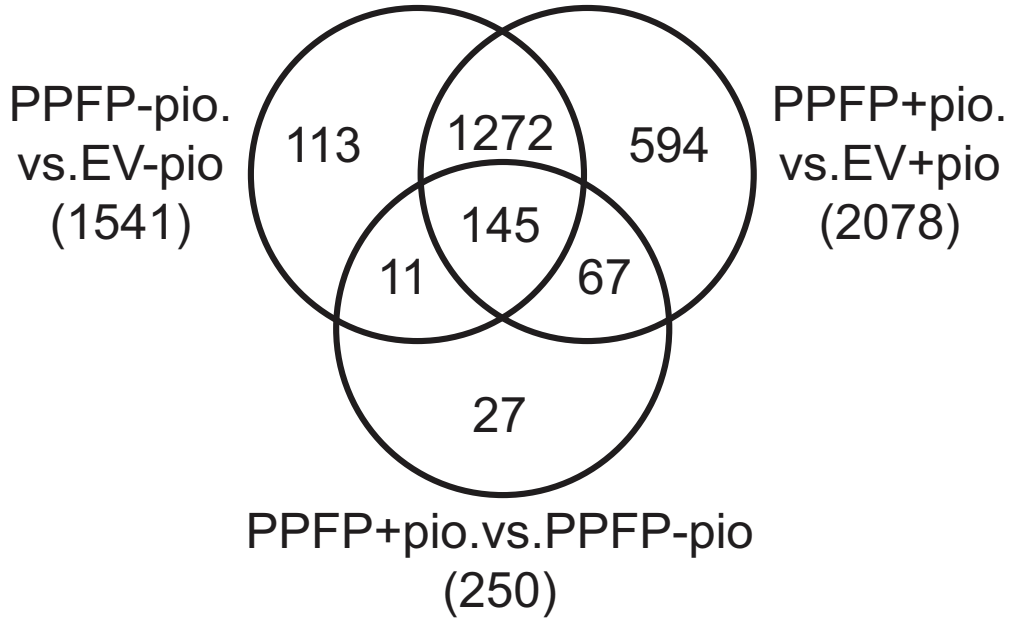



Figure 4.1: **Venn diagram illustrating the overlap of genes regulated by PPFP in comparisons of PPFP and EV cells cultured with and without pioglitazone**. The total number of differentially expressed genes for each comparison is shown in parentheses, using FDR$<$.05 and fold change $>$2 as cut-offs.

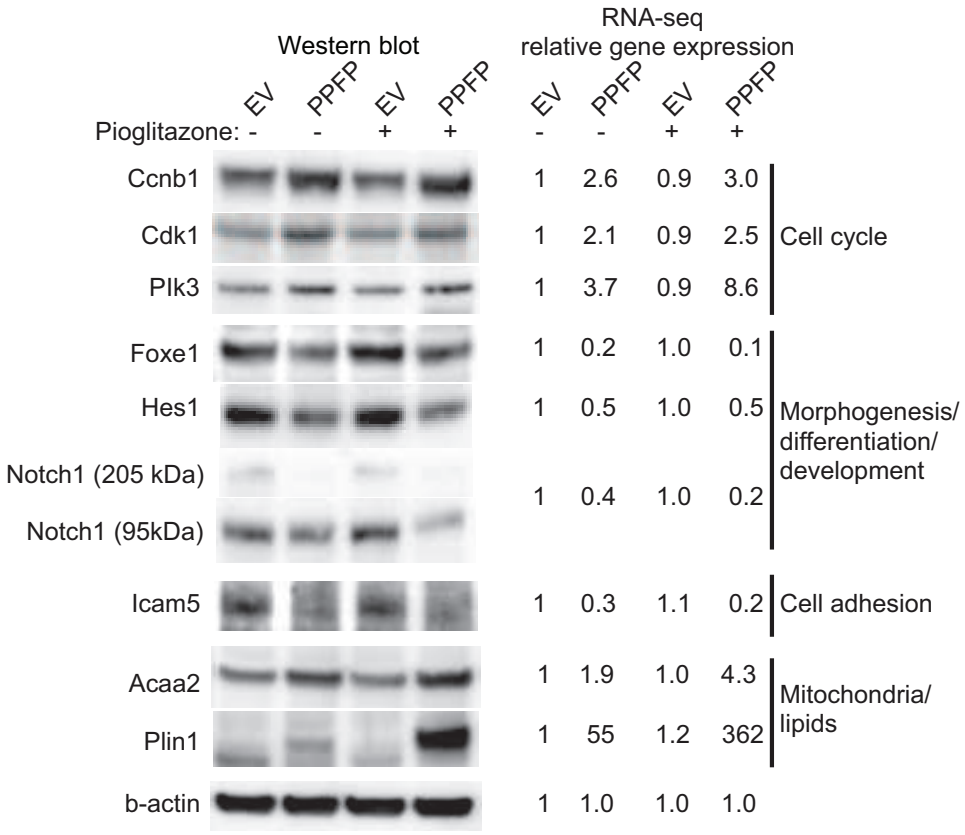



Figure 4.2: **Western blot analysis and RNA-seq expression data of selected genes in PPFP and EV cells cultured without and with pioglitazone**. RNA-seq data are normalized relative to EV cells cultured without pioglitazone. The genes are organized by concepts related to the gene set names with which they are associated, as described in Results. Although grouped together in one figure, the expression of these proteins without and with pioglitazone is presented at several different places within Results. (data generated by Bin Xu)

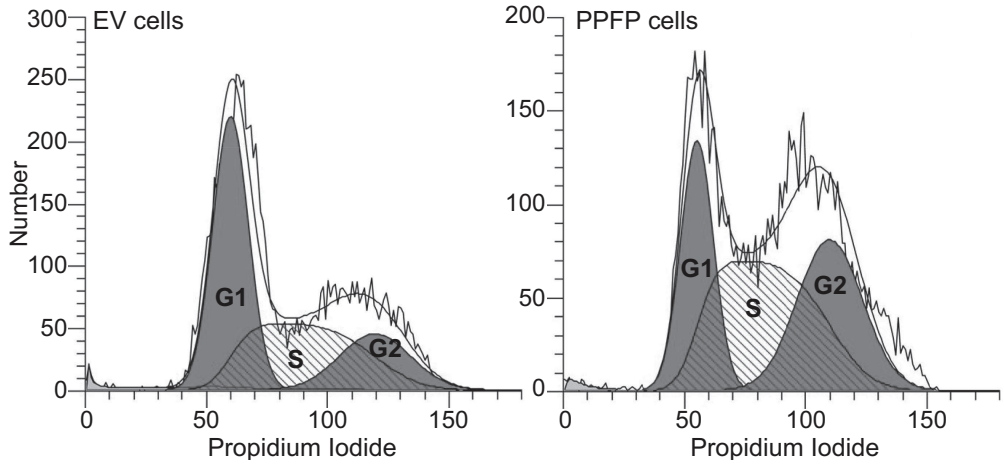



Figure 4.3: **DNA content analysis of EV and PPFP cells**. Cells were fixed, stained with propidium iodide and analyzed by flow cytometry using ModFit LT version 4.1 software. The graphs show the histograms and the derived areas of the G1, S and G2/M (labeled G2) phases of the cell cycle. These are quantified for EV cells as G1 45%, S 36% and G2/M 19%; and for PPFP cells, G1 26%, S 43% and G2/M 31% of cells. (data generated by Jingcheng Yu)

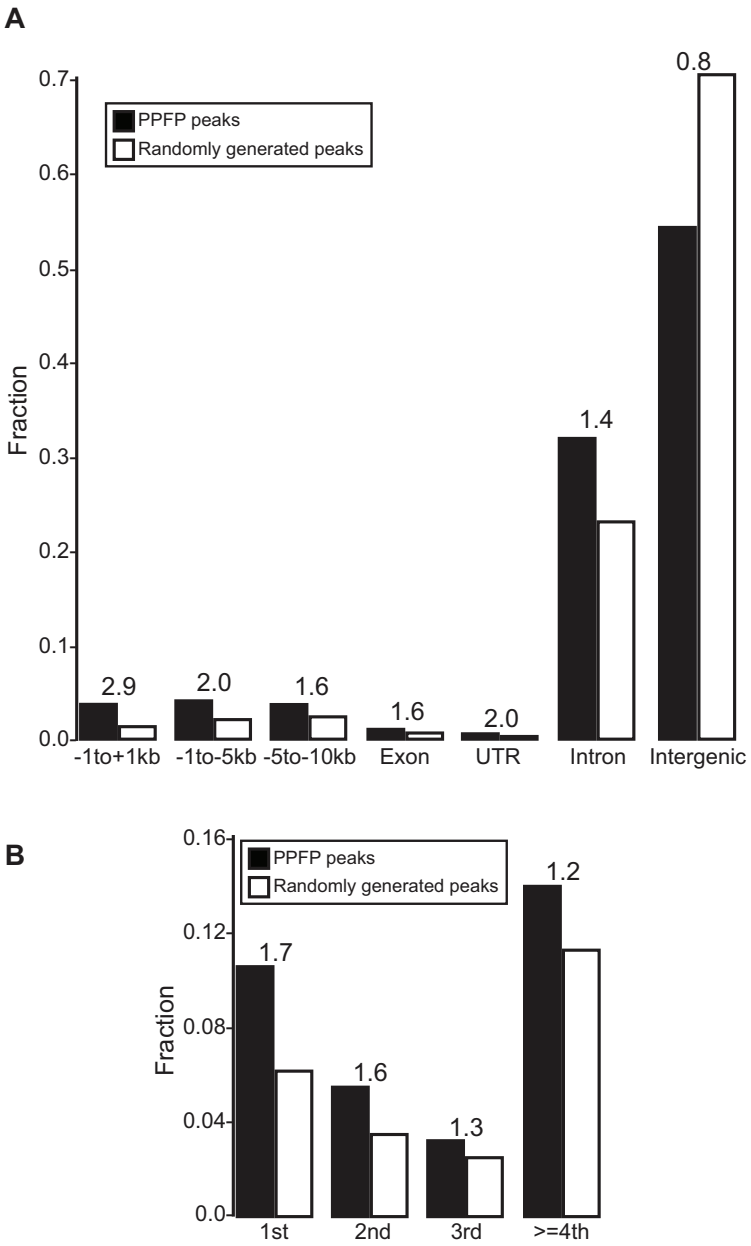
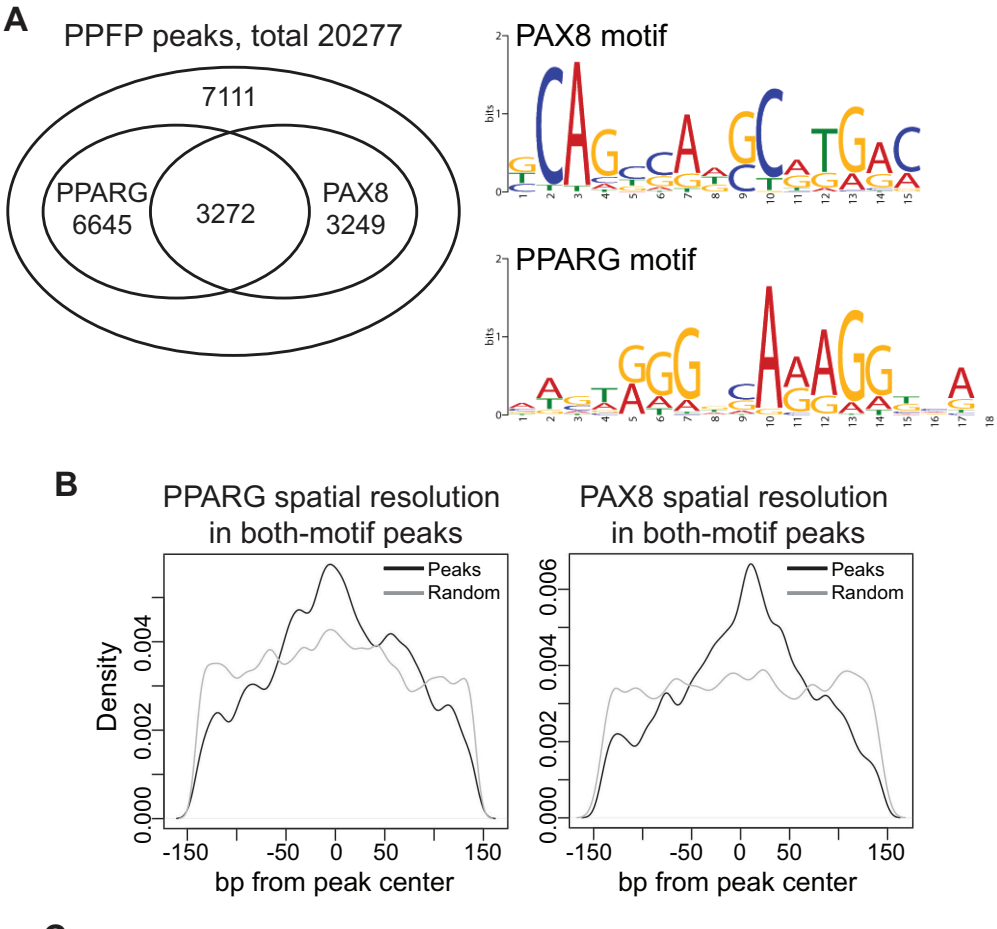



Figure 4.4: **Annotation of PPFP peaks versus randomly generated peaks, relative to genic and intergenic regions**. A. Peaks were assigned to one region only with the prioritization going from left to right. B. The intron group of A is divided into individual introns. The numbers above the bars indicate the ratios of PPFP to randomly generated peaks.




C

| | | Have PPARG motif | No PPARG motif |
|---|---|---|---|
| PAX8 peaks with PAX8 motif | Overlap with PPFP | 984 (52%) | 1457 (36%) |
| | No overlap | 921 (48%) | 2619 (64%) |
| | Total | 1905 | 4076 |

Figure 4.5: **PPFP peaks contain PAX8 and/or PPARG motifs**. A. Venn diagram showing the overlap of PAX8 and PPARG motifs within PPFP peaks, and the logos for PAX8 and PPARG motifs. B. Spatial resolution analysis showing that both the PAX8 and PPARG motifs are centered in the PPFP peaks that contain both motifs (black lines). The grey lines show the flat distribution of each motif in randomly sampled 300bp regions across genome, serving as negative controls. C. The peaks with PAX8 motifs identified in a previously published [23] PAX8 ChIP-seq analysis of PCCL3 cells were divided into those that overlap or not with the PPFP peaks identified in this study. These were then subdivided into peaks with or without PPARG motifs. PPARG motifs are enriched in the PAX8 peaks that overlap with PPFP peaks, $p<2.2e-16$, Fisher's exact test.

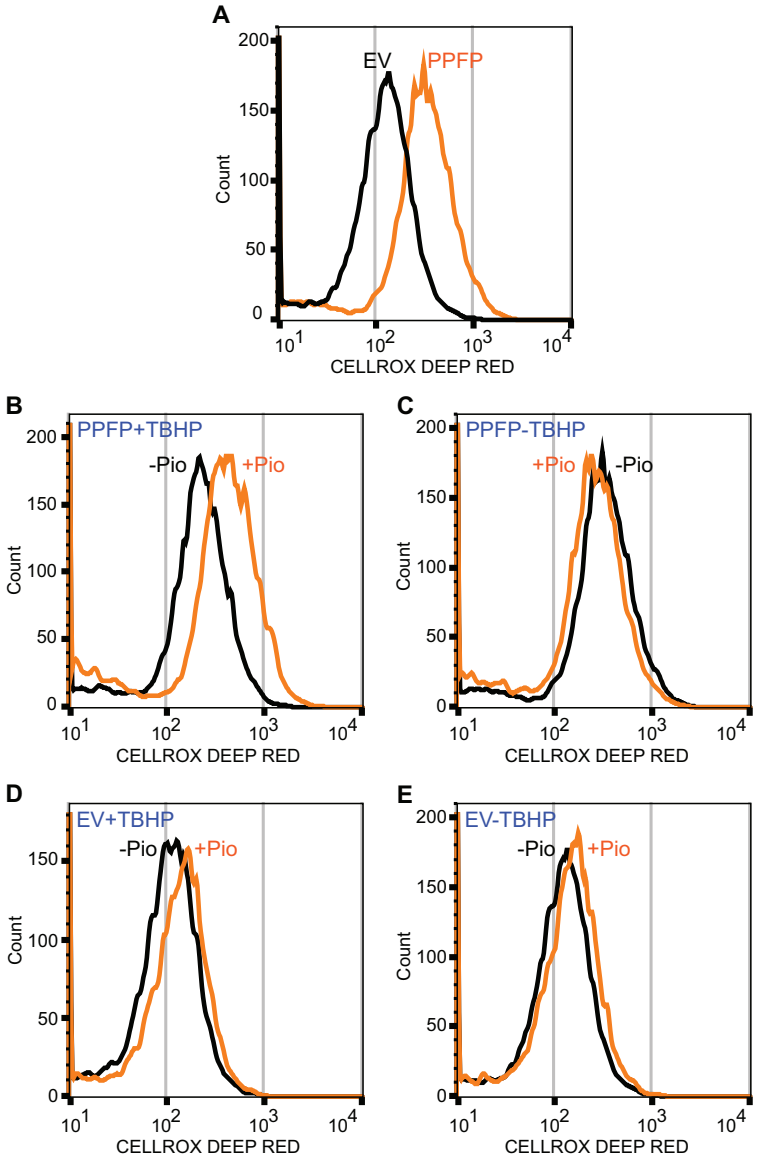



Figure 4.6: **Analysis of oxidative stress in PPFP and EV cells**. The cells were incubated with the reactive oxygen species-sensitive dye CellROX Deep Red and analyzed by flow cytometry. A. PPFP cells have increase ROS relative to EV cells (cultured without pioglitazone). B. Pioglitazone (pio) increases ROS in PPFP cells cultured with tert-butyl hydroperoxide (TBHP). C. Pioglitazone does not increase ROS in PPFP cells cultured without TBHP. D, E. Pioglitazone does not increase ROS in EV cells. All cells (A-E) were cultured and analyzed at the same time. The EV cell tracing in A is the same preparation of cells as shown in E without pioglitazone, and the PPFP cell tracing in A is the same as that in C without pioglitazone. (data generated by Jingcheng Yu)

# Tables

Table 4.1: **Fifteen induced and 15 repressed gene sets with the lowest q-values in PPFP cells versus EV cells cultured without pioglitazone**

| Concept.ID | Concept.name | p-value | q-value | Status PPFP vs EV |
|---|---|---|---|---|
| GO:0071346 | cellular response to interferon-gamma | 6.17E-07 | 1.37E-04 | up |
| GO:0030529 | ribonucleoprotein complex | 8.09E-06 | 1.71E-04 | up |
| GO:0005740 | mitochondrial envelope | 5.07E-05 | 6.49E-04 | up |
| GO:0005811 | lipid particle | 8.27E-05 | 8.87E-04 | up |
| GO:0046689 | response to mercury ion | 1.72E-05 | 1.05E-03 | up |
| GO:0005730 | nucleolus | 2.08E-04 | 2.04E-03 | up |
| GO:0004364 | glutathione transferase activity | 3.24E-05 | 2.25E-03 | up |
| rno00480 | Glutathione metabolism | 1.29E-05 | 2.48E-03 | up |
| GO:0042555 | MCM complex | 2.92E-04 | 2.54E-03 | up |
| GO:0071219 | cellular response to molecule of bacterial origin | 6.99E-05 | 2.75E-03 | up |
| GO:0071398 | cellular response to fatty acid | 1.02E-04 | 3.35E-03 | up |
| GO:0006396 | RNA processing | 1.22E-04 | 3.56E-03 | up |
| GO:0071384 | cellular response to corticosteroid stimulus | 1.23E-04 | 3.56E-03 | up |
| GO:0000502 | proteasome complex | 5.06E-04 | 3.95E-03 | up |
| GO:0006457 | protein folding | 2.13E-04 | 5.08E-03 | up |
| GO:0008021 | synaptic vesicle | 1.36E-14 | 6.25E-12 | down |
| GO:0007267 | cell-cell signaling | 1.98E-09 | 7.03E-06 | down |
| GO:0004888 | transmembrane signaling receptor activity | 3.96E-08 | 2.75E-05 | down |
| GO:0048858 | cell projection morphogenesis | 3.74E-08 | 2.95E-05 | down |
| GO:0048730 | epidermis morphogenesis | 4.75E-08 | 2.95E-05 | down |
| GO:0044306 | neuron projection terminus | 1.12E-06 | 3.45E-05 | down |
| GO:0050808 | synapse organization | 1.10E-07 | 4.87E-05 | down |
| GO:0048667 | cell morphogenesis involved in neuron differentiation | 1.38E-07 | 4.92E-05 | down |
| GO:0004872 | receptor activity | 1.56E-07 | 5.40E-05 | down |
| GO:0043534 | blood vessel endothelial cell migration | 2.87E-07 | 8.80E-05 | down |
| GO:1901342 | regulation of vasculature development | 5.06E-07 | 1.28E-04 | down |
| GO:0045995 | regulation of embryonic development | 1.15E-06 | 1.77E-04 | down |
| GO:0090288 | negative regulation of cellular response to growth factor stimulus | 2.02E-06 | 2.55E-04 | down |
| GO:0007167 | enzyme linked receptor protein signaling pathway | 2.91E-06 | 3.13E-04 | down |
| GO:0051960 | regulation of nervous system development | 3.10E-06 | 3.24E-04 | down |

Table 4.2: **Regulation of PAX8-responsive genes by PPFP**. In a previously published study of PCCL3 cells (Ruiz-Llorente et al., 2012), siPAX8 regulated the expression of 296 genes by at least 1.2-fold. This Table indicates how many of those genes are regulated by PPFP and in what direction (PPFP cells versus EV cells, cultured without pioglitazone).

| | PPFP induces | PPFP represses | PPFP no change | Total |
|---|---|---|---|---|
| **siPAX8 induces** | 32 (18%) 62 (35%) | 81 (46%) | 175 (100%) | |
| **siPAX8 represses** | 47 (39%) | 24 (20%) 50 (41%) | 121 (100%) | |
| **Total** | 79 | 86 | 131 | 296 |

Table 4.3: **The 10 induced gene sets with the lowest q-values in the comparison of PPFP cells cultured with versus without pioglitazone all relate to fatty acid metabolism, mitochondria and PPAR activity.**

| **Gene set ID** | **Description** | **q-value** |
|---|---|---|
| GO:0009062 | fatty acid catabolic process | 7.00E-11 |
| GO:0019395 | fatty acid oxidation | 1.53E-10 |
| GO:0004091 | carboxylesterase activity | 1.93E-09 |
| rno03320 | PPAR signaling pathway | 4.55E-08 |
| GO:0006637 | acyl-CoA metabolic process | 2.15E-07 |
| GO:0006641 | triglyceride metabolic process | 4.81E-07 |
| GO:0005777 | peroxisome | 5.61E-07 |
| GO:0005740 | mitochondrial envelope | 5.96E-07 |
| GO:0005759 | mitochondrial matrix | 1.03E-06 |
| GO:0071398 | cellular response to fatty acid | 2.74E-06 |

Table 4.4: **Gene sets enriched in PPFP peaks by ChIP-seq analysis and differentially expressed in PPFP cells versus EV cells cultured without pioglitazone.**

| Gene set ID | Description | q-value ChIP | q-value RNA-seq PPFP vs EV without pioglitazone | Status PPFP vs EV |
|---|---|---|---|---|
| GO:0005740 | mitochondrial envelope | 0.024 | 6.49E-04 | induced |
| GO:0044429 | mitochondrial part | 8.30E-04 | 7.19E-04 | induced |
| GO:0005743 | mitochondrial inner membrane | 0.029 | 7.19E-04 | induced |
| GO:0031966 | mitochondrial membrane | 0.029 | 8.51E-04 | induced |
| GO:0005811 | lipid particle | 0.0088 | 8.87E-04 | induced |
| GO:0019866 | organelle inner membrane | 0.024 | 9.99E-04 | induced |
| GO:0019915 | lipid storage | 0.03 | 0.01 | induced |
| GO:0000082 | G1/S transition of mitotic cell cycle | 0.01 | 0.042 | induced |
| GO:0007167 | enzyme linked receptor protein signaling pathway | 3.50E-04 | 3.13E-04 | repressed |
| GO:0007169 | transmembrane receptor protein tyrosine kinase signaling pathway | 0.011 | 0.0016 | repressed |
| GO:0022603 | regulation of anatomical structure morphogenesis | 7.16E-04 | 0.0019 | repressed |
| GO:0010648 | negative regulation of cell communication | 6.59E-04 | 0.0028 | repressed |
| GO:0045664 | regulation of neuron differentiation | 0.035 | 0.0053 | repressed |
| GO:0023057 | negative regulation of signaling | 0.0014 | 0.0072 | repressed |
| GO:0009968 | negative regulation of signal transduction | 3.08E-04 | 0.008 | repressed |
| GO:0031012 | extracellular matrix | 0.028 | 0.0081 | repressed |
| GO:0035295 | tube development | 0.034 | 0.014 | repressed |
| GO:0050867 | positive regulation of cell activation | 0.049 | 0.035 | repressed |
| GO:0030155 | regulation of cell adhesion | 0.048 | 0.04 | repressed |
| GO:0016331 | morphogenesis of embryonic epithelium | 0.042 | 0.043 | repressed |
| GO:0005539 | glycosaminoglycan binding | 0.022 | 0.048 | repressed |

Table 4.5: **Gene sets enriched in PPFP peaks with PAX8 motifs >10kb upstream from TSS's and differentially expressed in PPFP cells versus EV cells without pioglitazone**. All 16 gene sets are repressed by PPFP.

| Gene set ID | Description | q-value ChIP | q-value RNA-seq PPFP vs EV without pioglitazone | Status PPFP vs EV |
|---|---|---|---|---|
| GO:0048598 | embryonic morphogenesis | 0.017 | 0.0014 | repressed |
| GO:0022603 | regulation of anatomical structure morphogenesis | 0.013 | 0.0019 | repressed |
| GO:0003007 | heart morphogenesis | 0.04 | 0.0023 | repressed |
| GO:0072358 | cardiovascular system development | 0.029 | 0.0036 | repressed |
| GO:0035239 | tube morphogenesis | 0.019 | 0.0037 | repressed |
| GO:0060562 | epithelial tube morphogenesis | 0.046 | 0.0049 | repressed |
| GO:0031330 | negative regulation of cellular catabolic process | 0.029 | 0.01 | repressed |
| GO:0007507 | heart development | 0.016 | 0.013 | repressed |
| GO:0035295 | tube development | 0.009 | 0.014 | repressed |
| GO:0010463 | mesenchymal cell proliferation | 0.0041 | 0.017 | repressed |
| GO:0048645 | organ formation | 0.025 | 0.019 | repressed |
| GO:0061061 | muscle structure development | 0.037 | 0.022 | repressed |
| GO:0051240 | positive regulation of multicellular organismal process | 0.038 | 0.023 | repressed |
| GO:0009895 | negative regulation of catabolic process | 0.041 | 0.028 | repressed |
| GO:0010464 | regulation of mesenchymal cell proliferation | 0.016 | 0.031 | repressed |
| GO:0009887 | organ morphogenesis | 0.012 | 0.035 | repressed |

Table 4.6: **Gene sets enriched in PPFP peaks with PAX8 motifs ≤10 kb from TSS's and differentially expressed in PPFP cells versus EV cells without pioglitazone**. All 4 gene sets are induced by PPFP.

| **Gene set ID** | **Description** | **q-value ChIP** | **q-value RNA-seq PPFP vs EV without pioglitazone** | **Status PPFP vs EV** |
|---|---|---|---|---|
| GO:0005740 | mitochondrial envelope | 0.047 | 6.49E-04 | induced |
| GO:0044429 | mitochondrial part | 0.033 | 7.19E-04 | induced |
| GO:0031966 | mitochondrial membrane | 0.047 | 8.51E-04 | induced |
| GO:0005811 | lipid particle | 0.011 | 8.87E-04 | induced |

Table 4.7: **Overlap of PCCL3 PPFP peaks with PPARG peaks in mouse adipocytes and macrophages**. The PPARG ChIP-seq data are from Lefterova et al. (2010). P=0.0022, Fisher's exact test, two tailed.

| Sample | # of peaks | # of peaks ≤10kb to TSS | # of peaks with homologs | # overlap with PPFP (rat) |
|---|---|---|---|---|
| PPFP (rat) | 20277 | 3965 | 2809 | - |
| PPARG (mouse adipocyte) | 2634 | 870 | 537 | 186 (34%) |
| PPARG (mouse macrophage) | 1961 | 661 | 411 | 104 (25%) |

Table 4.8: **Genes in gene sets that are induced by PPFP and repressed by pioglitazone**. Three gene sets were induced in the comparison of PPFP cells versus EV cells cultured without pioglitazone, and repressed in the comparison of PPFP cells with pioglitazone versus PPFP cells without pioglitazone. The gene sets are glutathione metabolism (KEGG), peroxidase activity (GO), and arachidonic acid metabolism (KEGG). The 49 unique, significant genes within these gene sets are listed in this table.

| **Symbol** | **Description** |
|---|---|
| Cbr1 | carbonyl reductase 1 |
| Cbr3 | carbonyl reductase 3 |
| Cyp4a8 | cytochrome P450, family 4, subfamily a, polypeptide 8 |
| Cyp4f17 | cytochrome P450, family 4, subfamily f, polypeptide 17 |
| Cyp4f5 | cytochrome P450, family 4, subfamily f, polypeptide 5 |
| Duox2 | dual oxidase 2 |
| Ephx2 | epoxide hydrolase 2, cytoplasmic |
| Gclc | glutamate-cysteine ligase, catalytic subunit |
| Gclm | glutamate cysteine ligase, modifier subunit |
| Ggt6 | gamma-glutamyl transferase 6 |
| Ggt7 | gamma-glutamyltransferase 7 |
| Gpx2 | glutathione peroxidase 2 |
| Gpx4 | glutathione peroxidase 4 |
| Gpx8 | glutathione peroxidase 8 |
| Gsr | glutathione reductase |
| Gss | glutathione synthetase |
| Gsta4 | glutathione S-transferase mu 2 |
| Gstm2 | glutathione S-transferase mu 2 |
| Gstm7 | glutathione S-transferase, mu 7 |

| | |
|---|---|
| Gsto1 | glutathione S-transferase omega 1 |
| Gstp1 | glutathione S-transferase pi 1 |
| Gstt1 | glutathione S-transferase theta 1 |
| Hpgds | hematopoietic prostaglandin D synthase |
| Idh1 | isocitrate dehydrogenase 1 (NADP+), soluble |
| Idh2 | isocitrate dehydrogenase 2 (NADP+), mitochondrial |
| Iyd | iodotyrosine deiodinase |
| LOC501110 | similar to Glutathione S-transferase A1 (GTH1) (HA subunit 1) (GST-epsilon) (GSTA1-1) (GST class-alpha) |
| Lta4h | leukotriene A4 hydrolase |
| Mgst1 | microsomal glutathione S-transferase 1 |
| Mgst2 | microsomal glutathione S-transferase 2 |
| Odc1 | ornithine decarboxylase 1 |
| Park7 | parkinson protein 7 |
| Pla2g2d | phospholipase A2, group IID |
| Pla2g4a | phospholipase A2, group IVA (cytosolic, calcium-dependent) |
| Pla2g5 | phospholipase A2, group V |
| Prdx1 | peroxiredoxin 1 |
| Prdx2 | peroxiredoxin 2 |
| Prdx3 | peroxiredoxin 3 |
| Prdx4 | peroxiredoxin 4 |
| Prdx6 | peroxiredoxin 6 |
| Ptgs1 | prostaglandin-endoperoxide synthase 1 |
| Ptgs2 | prostaglandin-endoperoxide synthase 2 |

| | |
|---|---|
| Rrm1 | ribonucleotide reductase M1 |
| Rrm2 | ribonucleotide reductase M2 |
| Rrm2b | ribonucleotide reductase M2 B (TP53 inducible) |
| 15-Sep | selenoprotein 15 |
| Srm | spermidine synthase |
| Tpo | thyroid peroxidase |
| Txndc17 | thioredoxin domain containing 17 |

Table 4.9: **A set of ROS-related genes is induced in human PPFP follicular carcinomas versus non-PPFP follicular carcinomas**. Three gene sets in our RNA-seq data were induced in PPFP cells versus EV cells without pioglitazone, and repressed in PPFP cells with pioglitazone versus PPFP cells without pioglitazone: glutathione metabolism (KEGG), peroxidase activity (GO), and arachidonic acid metabolism (KEGG). There are 49 unique, differentially expressed genes within these gene sets. We tested whether this set of genes is induced in human PPFP follicular carcinomas versus non-PPFP follicular carcinomas by comparing the expression of all probesets for the 49 genes versus the probesets for all other genes in the human thyroid carcinoma Affymetrix study of Giordano (Giordano et al., 2006).

| | PPFP carcinomas expression greater than non-PPFP carcinomas | PPFP carcinomas expression less than non-PPFP carcinomas | Total |
|---|---|---|---|
| Probesets for set of ROS-related genes | 35 | 13 | 48 |
| All other probesets | 9685 | 10248 | 19933 |
| Total | 9720 | 10261 | 19981 |

## Bibliography

Anders, S., Pyl, P. T. and Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2):166–169, 2015.

Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. 2015.

Carre, A., Rachdi, L., Tron, E., Richard, B., Castanet, M., Schlumberger, M., Bidart, J.-M., Szinnai, G. and Polak, M. Hes1 is required for appropriate morphogenesis and differentiation during mouse thyroid gland development. *PloS one*, 6(2):e16752, 2011.

Celli, G., LaRochelle, W. J., Mackem, S., Sharp, R. and Merlino, G. Soluble dominant-negative receptor uncovers essential roles for fibroblast growth factors in multi-organ induction and patterning. *The EMBO journal*, 17(6):1642–1655, 1998.

Clifton-Bligh, R. J., Wentworth, J. M., Heinz, P., Crisp, M. S., John, R., Lazarus, J. H., Ludgate, M. and Chatterjee, V. K. Mutation of the gene encoding human TTF-2 associated with thyroid agenesis, cleft palate and choanal atresia. *Nature genetics*, 19(4):399–401, 1998.

Corzo, C. and Griffin, P. R. Targeting the Peroxisome Proliferator-Activated Receptor-gamma to Counter the Inflammatory Milieu in Obesity. *Diabetes & metabolism journal*, 37(6):395–403, 2013.

Croyle, M., Akeno, N., Knauf, J. A., Fabbro, D., Chen, X., Baumgartner, J. E., Lane, H. A. and Fagin, J. A. RET/PTC-induced cell growth is mediated in part by epidermal growth factor receptor (EGFR) activation: evidence for molecular and functional interactions between RET and EGFR. *Cancer research*, 68(11):4183–4191, 2008.

Dobson, M. E., Diallo-Krou, E., Grachtchouk, V., Yu, J., Colby, L. A., Wilkinson, J. E., Giordano,

T. J. and Koenig, R. J. Pioglitazone induces a proadipogenic antitumor response in mice with PAX8-PPARgamma fusion protein thyroid carcinoma. *Endocrinology*, 152(11):4455–65, 2011.

Enewold, L., Zhu, K., Ron, E., Marrogi, A. J., Stojadinovic, A., Peoples, G. E. and Devesa, S. S. Rising thyroid cancer incidence in the United States by demographic and tumor characteristics, 1980-2005. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 18(3):784–791, 2009.

Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nature protocols*, 7(9):1728–1740, 2012.

Giordano, T. J., Au, A. Y. M., Kuick, R., Thomas, D. G., Rhodes, D. R., Wilhelm, K. G. J., Vinco, M., Misek, D. E., Sanders, D., Zhu, Z., Ciampi, R., Hanash, S., Chinnaiyan, A., Clifton-Bligh, R. J., Robinson, B. G., Nikiforov, Y. E. and Koenig, R. J. Delineation, functional validation, and bioinformatic evaluation of gene expression in thyroid follicular carcinomas with the PAX8-PPARG translocation. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 12(7 Pt 1):1983–1993, 2006.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–589, 2010.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D. and Chen, H. SEER Cancer Statistics Review, 1975-2012, National Cancer Institute. Bethesda, MD. Technical report, 2015.

Kim, J. H., Karnovsky, A., Mahavisno, V., Weymouth, T., Pande, M., Dolinoy, D. C., Rozek, L. S.

and Sartor, M. A. LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. *BMC genomics*, 13:526, 2012.

Kishida, K., Shimomura, I., Nishizawa, H., Maeda, N., Kuriyama, H., Kondo, H., Matsuda, M., Nagaretani, H., Ouchi, N., Hotta, K., Kihara, S., Kadowaki, T., Funahashi, T. and Matsuzawa, Y. Enhancement of the aquaporin adipose gene expression by a peroxisome proliferator-activated receptor gamma. *The Journal of biological chemistry*, 276(51):48572–48579, 2001.

Kroll, T. G., Sarraf, P., Pecciarini, L., Chen, C. J., Mueller, E., Spiegelman, B. M. and Fletcher, J. A. PAX8-PPARgamma1 fusion oncogene in human thyroid carcinoma [corrected]. *Science (New York, N.Y.)*, 289(5483):1357–1360, 2000.

Lefterova, M. I., Steger, D. J., Zhuo, D., Qatanani, M., Mullican, S. E., Tuteja, G., Manduchi, E., Grant, G. R. and Lazar, M. A. Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Molecular and cellular biology*, 30(9):2078–2089, 2010.

Lefterova, M. I., Zhang, Y., Steger, D. J., Schupp, M., Schug, J., Cristancho, A., Feng, D., Zhuo, D., Stoeckert, C. J. J., Liu, X. S. and Lazar, M. A. PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. *Genes & development*, 22(21):2941–2952, 2008.

Macchia, P. E., Lapi, P., Krude, H., Pirro, M. T., Missero, C., Chiovato, L., Souabni, A., Baserga, M., Tassi, V., Pinchera, A., Fenzi, G., Gruters, A., Busslinger, M. and Di Lauro, R. PAX8 mutations associated with congenital hypothyroidism caused by thyroid dysgenesis. *Nature genetics*, 19(1):83–86, 1998.

Martinez Barbera, J. P., Clements, M., Thomas, P., Rodriguez, T., Meloy, D., Kioussis, D. and Beddington, R. S. The homeobox gene Hex is required in definitive endodermal tissues for nor-

mal forebrain, liver and thyroid formation. *Development (Cambridge, England)*, 127(11):2433–2445, 2000.

McCarthy, D. J., Chen, Y. and Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

Mitsutake, N., Knauf, J. A., Mitsutake, S., Mesa, C., Zhang, L. and Fagin, J. A. Conditional brafv600e expression induces dna synthesis, apoptosis, dedifferentiation, and chromosomal instability in thyroid pccl3 cells. *Cancer Research*, 65(6):2465–2473, 2005.

Nielsen, R., Pedersen, T. A., Hagenbeek, D., Moulos, P., Siersbaek, R., Megens, E., Denissov, S., Borgesen, M., Francoijs, K.-J., Mandrup, S. and Stunnenberg, H. G. Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes & development*, 22(21):2953–2967, 2008.

Pasca di Magliano, M., Di Lauro, R. and Zannini, M. Pax8 has a key role in thyroid cell differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 97(24):13144–13149, 2000.

Porazzi, P., Marelli, F., Benato, F., de Filippis, T., Calebiro, D., Argenton, F., Tiso, N. and Persani, L. Disruptions of global and JAGGED1-mediated notch signaling affect thyroid morphogenesis in the zebrafish. *Endocrinology*, 153(11):5645–5658, 2012.

Raman, P. and Koenig, R. J. Pax-8-PPAR-gamma fusion protein in thyroid carcinoma. *Nature reviews. Endocrinology*, 10(10):616–623, 2014.

Rosen, E. D., Sarraf, P., Troy, A. E., Bradwin, G., Moore, K., Milstone, D. S., Spiegelman, B. M.

and Mortensen, R. M. PPAR gamma is required for the differentiation of adipose tissue in vivo and in vitro. *Molecular cell*, 4(4):611–617, 1999.

Ruiz-Llorente, S., Carrillo Santa de Pau, E., Sastre-Perona, A., Montero-Conde, C., Gomez-Lopez, G., Fagin, J. A., Valencia, A., Pisano, D. G. and Santisteban, P. Genome-wide analysis of Pax8 binding provides new insights into thyroid functions. *BMC genomics*, 13:147, 2012.

Sartor, M. A., Leikauf, G. D. and Medvedovic, M. Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, 2009.

Vitagliano, D., Portella, G., Troncone, G., Francione, A., Rossi, C., Bruno, A., Giorgini, A., Coluzzi, S., Nappi, T. C., Rothstein, J. L., Pasquinelli, R., Chiappetta, G., Terracciano, D., Macchia, V., Melillo, R. M., Fusco, A. and Santoro, M. Thyroid targeting of the N-ras(Gln61Lys) oncogene in transgenic mice results in follicular tumors that progress to poorly differentiated carcinomas. *Oncogene*, 25(39):5467–5474, 2006.

Vu-Phan, D., Grachtchouk, V., Yu, J., Colby, L. A., Wicha, M. S. and Koenig, R. J. The thyroid cancer PAX8-PPARG fusion protein activates Wnt/TCF-responsive cells that have a transformed phenotype. *Endocrine-related cancer*, 20(5):725–739, 2013.

Vu-Phan, D. and Koenig, R. J. Genetics and epigenetics of sporadic thyroid cancer. *Molecular and cellular endocrinology*, 386(1-2):55–66, 2014.

Wang, L., Wang, S. and Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28:2184–2185, 2012.

Welch, R. P., Lee, C., Imbriano, P. M., Patil, S., Weymouth, T. E., Smith, R. A., Scott, L. J. and Sartor, M. A. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic acids research*, 42(13):e105, 2014.

Yoav Benjamini, Y. H. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Zannini, M., Francis-Lang, H., Plachov, D. and Di Lauro, R. Pax-8, a paired domain-containing protein, binds to a sequence overlapping the recognition site of a homeodomain and activates transcription from two thyroid-specific promoters. *Molecular and cellular biology*, 12(9):4230–4241, 1992.

Zhang, Y., Lin, Y.-H., Johnson, T. D., Rozek, L. S. and Sartor, M. A. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics (Oxford, England)*, 30(18):2568–2575, 2014.

# CHAPTER V

# Conclusions and future directions

## 5.1 Conclusions

High-throughput sequencing technologies have matured over the past decade and found their applications in various research fields, including cancer biology. Although bioinformaticians are still exploring novel analysis algorithms and building efficient software tools to meet the expanding needs, the power of this technology has already brought cancer research into the new "genomics" era. In this dissertation, I have contributed to the role bioinformatics plays in cancer research both in terms of methods and tool development (by introducing a new ChIP-seq analysis pipeline, PePr) and in terms of analysis of diverse omics data (by using PePr and other open-source bioinformatics software tools to study molecular subtypes in head and neck cancers).

In Chapter 2, I introduced a ChIP-seq pipeline (PePr) to analyze replicated ChIP-seq data, with the primary goal for differential binding analysis. PePr introduced many new functionalities into the ChIP-seq analysis pipeline, such as a window-size estimation method to automatically accommodate for different peak types (sharp or broad), a normalization strategy to account for differences in IP efficiencies between libraries, and post-processing steps to remove false-positive peaks. PePr has superior performance on both replicated transcription factor and histone modification datasets when compared to the most widely used software tools at the time of publication, and can also be used on pull-down DNA methylation datasets, such as those generated with MeDIP-seq

or hmeDIP-seq.

In Chapter 3, I performed integrative analysis of RNA-seq and SNP-array datasets on head and neck squamous cell carcinomas. We identified two robust HPV subtypes using gene expression-based clustering. One subtype (HPV-KRT) has more keratinization, higher copy number of PIK3CA and TP63, and lower immune response and mesenchymal differentiation compared to the other subtype (HPV-IMU), suggesting our HPV subtype findings are similar to what has been reported in Keck et al. (2015). However, unlike Keck et al, who used microarray data, we were able to more deeply characterize the difference between the two subgroups by analyzing SNP-array data for whole-genome CNAs and mining the RNA-seq data for HPV related information and expressed single-base mutations. We found that HPV-KRT has more genic viral integration events (identified from host-virus fusion transcripts), more spliced E6*, less full-length E6 activity and less E2/E4/E5 expression than HPV-IMU. In fact, based on a previous study of E6-regulated genes (Duffy et al., 2003), the differential expression of genes (between the subgroups) involved in mesenchymal differentiation, keratinization and oxidation reduction pathways could all be explained by the difference in E6 activity and splicing. In addition, we show that HPV-KRT has more chr3q amplification and PIK3CA activating mutations, whereas HPV-IMU has more chr16q deletions. Our preliminary analysis suggests that the two arm-level CNAs likely promote tumor survival through two different mechanisms: duplication of oncogenes on chr3q versus deletion of tumor suppressors on chr16q. Combined with other observed differences, we hypothesize the two subtypes likely undergo two different paths of oncogenesis and will benefit from different personalized treatment plans. For example, HPV-KRT may respond better to immunotherapy and drugs targeting oncogenes on chr3q such as PIK3CA and TP63, whereas HPV-IMU may benefit more from drugs that suppress epithelial-to-mesenchymal transition and/or E6 activity.

In Chapter 4, we characterized the binding profiles of the fusion oncogene PPFP using ChIP-seq data, demonstrating that binding domains of both the original proteins, PAX8 and PPARG, are

functional in the rat PPFP-transfected PCCL3 cell line. Our integrative analysis of RNA-seq and ChIP-seq data in the same cell line suggests that PPFP regulates genes in many pathways related to cancer. The protein expression levels of representative genes selected from these pathways were validated using Western blot. Several of our analyses could not have been possible without the use of previously published datasets. For instance, by incorporating the ChIP-seq data from Lefterova et al. (2010), we demonstrated that PPFP prefers to bind to adipocyte genes rather than macrophage genes, explaining the strong adipocyte phenotype in the presence of pioglitazone; we also showed PPFP regulates PAX8 response genes using a published siPAX8 study (Ruiz-Llorente et al., 2012). Finally, we explored the mechanism of the therapeutic effect of pioglitazone in PPFP tumors, and found that pioglitazone may reverse PPFP's oncogenic effect by altering the cell's oxidative stress.

## 5.2 Future directions

The software (PePr) and studies on head and neck cancers presented in this dissertation are parts of ongoing research efforts, and will open up many opportunities for future research.

### 5.2.1 Chapter II

As the sequencing depth and sample sizes keep increasing for ChIP-seq and other types of pull down experiments, the run time and memory usage of PePr will also linearly increase. Therefore, improvements on memory management and run time should be made in order for this tool to remain useful for the community. To reduce the memory usage, data could be preprocessed into trunks instead of loading into memory simultaneously. There are two potential strategies to reduce the run time. First, we could incorporate the python package "multiprocessing" to parallelize the program. Secondly, we could optimize the most computationally intensive module (the dispersion estimation) by finding a more efficient solver for finding the root for where the derivative of the maximum likelihood function (equation 2.7) equals zero.

In terms of models, we could expand the differential binding analysis to allow a design ma-

trix for multiple groups and/or covariates. Additional modules could be included to allow more comprehensive analysis of the data. For example, a preprocess module could be added to assess the data quality and immunoprecipitation efficiency; in addition to splitting the genome into tiling windows, we could also support testing pre-defined genomic regions such as promoters, enhancers and CpG islands. We can also expand our pipeline beyond supervised analysis to unsupervised data exploration where group labels are not used. For example, we could first identify a list of peaks from each sample, and then combine the peak list and count the reads in these regions, and then perform unsupervised clustering on these samples and peaks to discover patterns.

### 5.2.2 Chapter III

We have identified and characterized the transcriptomic and genomic properties of two HPV(+) HNSCC subtypes. However, the epigenomic landscape of HPV(+) tumors is also very interesting and should be examined in light of our identified subtypes when such data become available. Another limitation of our study is its small sample size, which is due to the difficulty in collecting samples that pass the quality requirements for sequencing. In order to expand our subtype study to a larger patient cohort and eliminate the need to deep sequence all of the samples, we could build a classification rule on a few key genetic and/or transcriptomic markers using machine learning algorithms to predict the two HNSCC subtypes. For instance, we could build a custom chip to measure these markers for the new tumor samples and predict their subtypes. Then using the predicted labels in the larger patient cohort, we can validate our subtype findings such as the difference in E2/E4/E5 expression, E6 activity, chr3q and chr16q CNA, and core pathway differences. With a larger patient cohort, we can also examine the correlation of the cluster membership with other important clinical variables, such as N stage, lymph node metastasis and disease recurrence.

### 5.2.3 Chapter IV

The PPFP protein contains both DNA binding domains (DBD) of PPARG and PAX8 and uses both to bind to 20,000 sites over the rat genome. It would be interesting to know to what extent each of the DBDs are contributing to its oncogenic effect. We could perform similar ChIP-seq and RNA-seq analysis on the same cell line transfected with PPFP protein with mutations on either of the DBDs, and investigate the change in the binding sites and expression profiles after the mutation. A tumor is more than a homogenous population of abnormal cells, but a complex mixture defined as the tumor microenvironment, consisting of the tumor cells and multiple normal cell types (Hanahan and Weinberg, 2011). The conclusions drawn from the cell line above may not apply to the complex tumor in vivo; and there may be novel effects in a tumor microenvironment, such as infiltrating immune cells, that are absent in cell lines. Although our findings in the rat cell line are mechanistically interesting and largely concordant with studies conducted in human and mouse, we should validate and expand upon them in a more complex model system, such as a mouse xenograft model and human primary tumors (though it is difficult to collect such tumors in human), which would also help eliminate species-specific effects.

## Bibliography

Duffy, C. L., Phillips, S. L. and Klingelhutz, A. J. Microarray analysis identifies differentiation-associated genes regulated by human papillomavirus type 16 E6. *Virology*, 314(1):196–205, 2003.

Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.

Keck, M. K., Zuo, Z., Khattri, A., Stricker, T. P., Brown, C. D., Imanguli, M., Rieke, D., Endhardt, K., Fang, P., Bragelmann, J., DeBoer, R., El-Dinali, M., Aktolga, S., Lei, Z., Tan, P., Rozen, S. G., Salgia, R., Weichselbaum, R. R., Lingen, M. W., Story, M. D., Ang, K. K., Cohen, E. E. W., White, K. P., Vokes, E. E. and Seiwert, T. Y. Integrative Analysis of Head and Neck Cancer Identifies Two Biologically Distinct HPV and Three Non-HPV Subtypes. *Clinical Cancer Research*, 21(4):870–881, 2015.

Lefterova, M. I., Steger, D. J., Zhuo, D., Qatanani, M., Mullican, S. E., Tuteja, G., Manduchi, E., Grant, G. R. and Lazar, M. A. Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Molecular and cellular biology*, 30(9):2078–2089, 2010.

Ruiz-Llorente, S., Carrillo Santa de Pau, E., Sastre-Perona, A., Montero-Conde, C., Gomez-Lopez, G., Fagin, J. A., Valencia, A., Pisano, D. G. and Santisteban, P. Genome-wide analysis of Pax8 binding provides new insights into thyroid functions. *BMC genomics*, 13:147, 2012.

# APPENDIX A

# Supplemental material for Chapter II

**Versions and detailed parameters of the programs used**: All programs were run under default parameters if possible. Significance cut-offs for all programs are provided in Table S4. The shift size and window size estimated for each dataset with PePr are provided in Table S5. For Separate Analysis (SA) approaches, the final peak regions were defined as the intersection of the peaks generated from all separate runs; the significance of each peak was defined as the average of the ranks in all separate analyses. See details below:

**PePr version 1.0.1**: default parameters were used. For TFs:

```
PePr c chip_file i control_file f file_format
--peaktype=SHARP remove_artefacts
```

For H3K27me3:

```
PePr c chip_file i control_file f file_format --peaktype=BROAD
```

**MACS version1.4.0rc2**: default parameters were used.

```
macs14 -t chip_file -c control_file
```

**SPP version 1.10.1**:

SPP was run with the ENCODE project IDR guidelines. IDR thresholds of 0.01 and 0.0025 were chosen for the original replicate threshold and pooled-pseudoreplicate threshold, respectively. The optimum set was reported.

```
Rscript run_spp.R -c=chipSampleRep1.tagAlign.gz
-i=controlSampleRep0.tagAlign.gz
-npeak=300000 -odir=/peaks/reps -savr -savp -rf
-out=/stats/phantomPeakStatsReps.tab
```

**MACS2 version 2.0.10.09132012**:

MACS2 was run with the ENCODE project IDR guidelines. IDR thresholds of 0.01 and 0.0025 were chosen for the original replicate threshold and pooled-pseudoreplicate threshold, respectively. The optimum set was reported.

```
macs2 callpeak -t chipSampleRep1.tagAlign.gz
-ccontrolSampleRep0.tagAlign.gz
-f BED -nchipSampleRep1_VS_controlSampleRep0 -g hs -p 1e-3
--to-large
```

**ZINBA version 2.01**:

The alignability function was run with the corresponding genome and read mappability file. A default read extension of 90 was used. For histone data, the "broad" argument was given.

```
generateAlignability(,athresh=1,extension=90),
zinba(, refinepeaks=0,seq=chip_file, input=control_file,
filetype=''bed'', extension=90)
```

**SICER version 1.1**:

Parameters recommended by the manual were used. The corresponding genome was used for each experiment. A window size of 200 and gap size of 600 were used for broad peaks. The fragment size was set to 150.

```
SICER.sh chip_file control_file .genome 2 200 150 0.8 600 1E-2
```

**edgeR version 3.2.4**:

edgeR-basic: First the reads were shifted (45bp) and counted in non-overlapping windows (200bp). The read counts and group assignments were prepared in edgeR specified format and then the following commands were applied. Windows passing the significance cut-off were deemed eligible and then adjacent windows were merged to form a final peak list.

edgeR-plus: All of PePrs pre-processing (shift size and window size estimates) and post-processing steps (removing artefacts) were applied using our default parameter settings.

```
y = DGEList(counts=counts,group = group)
y = calcNormFactors(y)
y<-estimateCommonDisp(y, rowsum.filter=5)
y<-estimateGLMTagwiseDisp(y,design)
fit_tag<-glmFit(y,design)
lrt.tagwise<-glmLRT(fit_tag,coef=2)
```

**DiffBind version 1.10.0**:

First, SICER was used to call peaks from each sample using the matching input samples as controls. The resulting peak lists from all four samples were input to DiffBind, which generated 29510 pre-candidate regions. The following commands were then executed to search for differential binding regions:

```
hpv = dba(sampleSheet="diffbind_sample.csv")
```

```
hpv = dba.count(hpv)
hpv = dba.contrast(hpv,hpv$masks$`HPV-`,
hpv$masks$`HPV+`, "HPV-", "HPV+")
For DESeq:
hpv =dba.analyze(hpv,method=DBA_DESEQ)
hpv.DB = dba.report(hpv,method=DBA_DESEQ)
And for edgeR:
hpv =dba.analyze(hpv,bReduceObjects=F)
hpv.DB = dba.report(hpv)
```

**diffReps version 1.55.4**:

Default parameters. An exact negative binomial test was used. Settings were slightly different between TFs and H3K27me3. For transcription factors, the sharp option was enabled and a window size of 200 was used:

```
diffReps.pl tr chip_file1 chip_file2 ...
-co control_file1 control_file2 ... -gname genome me nb
nsd sharp window 200
```

For H3K27me3, the default parameters were used (nsd="broad" and window size 1000):

```
diffReps.pl tr chip_file1 chip_file2 ...
-co control_file1 control_file2 ... -gname genome me nb nsd
```