NTNU
Norwegian University of
Science and Technology

Master's thesis

2019

Verena Leonie Lörks

Master's thesis

NTNU
Norwegian University of
Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

Verena Leonie Lörks

The transcriptional and post-transcriptional regulation in cancer

A gene expression profiling study of human head and neck cancers

June 2019

NTNU
Norwegian University of
Science and Technology

NTNU

Norwegian University of
Science and Technology

The transcriptional and post-transcriptional regulation in cancer

A gene expression profiling study of human head and neck cancers

A gene expression profiling study of human head and neck cancers

Master of Science in Molecular Medicine
Submission date: June 2019

Norwegian University of Science and Technology
Department of Clinical and Molecular Medicine

Contents

Acknowledgements

This thesis is part of the MSc. program in Molecular Medicine at the Norwegian University of Science and Technology and was conducted at The Ohio State University Comprehensive Cancer Center in Columbus, Ohio, USA under the supervision of Dr. Wayne Miles.

First, I would like to thank Dr. Miles for the opportunity to work in his lab and his continuous guidance and support. Second, I would like to thank the members in Dr. Wayne Miles's lab for sharing their knowledge and their incredible support throughout the entire time in and outside the lab.

I would like to thank the TRONDBUSS program for the financial, administrative and personal support which made this exchange possible. I also would like to thank Prof. Liv Thommesen for her invaluable and frequent feedback on this work and guidance throughout the writing process.

Lastly, I would like to thank my family and friends in the US, Norway and Germany who supported me in so many different ways.

Preface

This thesis encompasses two independent projects. The project on the interactome of putative PUMILIO interactors was terminated due to technical problems that could not be solved within the time frame of the thesis. In order to give a full overview of the accomplished work, both projects are included as separate parts in chronological order.

Abstract

Head and neck squamous cell carcinomas are a collection of tumors that arise from mucosal cells in the oral cavity and the upper airway and food passages. Despite their common mucosal origin, the group of HNSCCs is rather heterogeneous. Human papillomavirus (HPV) has emerged as the driving force of a subset of these malignancies. HPV positive and HPV negative HNSCCs largely differ in epidemiological, clinical and molecular features and are perceived as two different entities of the disease. Numerous studies have attempted to shed light on the transcriptional profiles of head and neck cancer, however, the field of HPV-related gene expression signatures in HNSCC is still evolving.

In this study, we highlight the differences in the coding and non-coding transcriptome of HPV^+ and HPV^- HNSCC. We utilized RNA sequencing data from six different HNSCC cell lines to confirm HPV status as the main driver of transcriptional differences. Subsequently, we mapped out the differentially expressed genes in HPV^+ and HPV^- cell lines with and without a normal control cell line. We also detected differential expression in HPV negative (*n=162*) and HPV positive (*n=32*) tumor samples from the TCGA database and compared the results to our findings from the HNSCC cell line dataset. The differences in gene expression were validated by qPCR in a subset of genes. The function of DE protein coding genes was assessed by GO Molecular Function overrepresentation analysis.

We identified *n=154* coding and *n=10* non-coding differentially expressed genes between HPV^+ and HPV^- groups across both datasets. A fraction of the identified differentially expressed non-coding transcripts has previously been linked to HNSCC or other cancer types. Among the protein coding differentially expressed genes, we found an enrichment of serine proteases, aldo-keto reductases and cytokines among others.

In sum, this work has contributed to elaborate the transcriptional profiles of HPV^+ and HPV^- head and neck cancer. We identified several non-coding genes that had not been linked to HPV-related subtypes of head and neck cancer so far. We were able to identify a set of differentially expressed coding genes and annotated their molecular function. Our contribution to the mapping of non-coding transcriptional profiles in HPV-related head and neck cancers may be valuable with respect to the identification of new biomarkers for HNSCC.

Abbreviations

3’UTR	3’ untranslated region
BAM	Binary alignment map
CPM	Counts per million
DE	Differentially expressed
DND1	Dead end protein homolog 1
eGFP	Enhanced GFP
FMRP	Fragile X mental retardation protein 1 (encoded by FMR1)
FPKM	Fragments per kilobase per million
FTP	File transfer protocol
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1
HNSCC	Head and neck squamous cell carcinoma
HPV	Human papillomavirus
HSP90AA1	Heat shock protein HSP90 alpha
L2FC	Log2 fold change
LAP tag	localization and purification tag
METTL3	N6-adensoine-methyltransferase catalytic subunit
NANOS2	Nanos homolog 2
NORAD	Non-coding RNA activated by DNA damage
OSCC	Oral squamous cell carcinoma
Padj	Adjusted p-value
PCA	Principal component analysis
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
RBFOX2	RNA binding protein fox-1 homolog 2
SAM68	KH domain-containing, RNA-binding, signal transduction-associated protein 1
SE	Standard error of the mean
TCGA	The Cancer Genome Atlas
TSCC	Tongue squamous cell carcinoma
vst	Variance stabilizing transformation

1. Introduction

1.1 PUMILIO-mediated post-transcriptional control

Gene expression is a highly regulated process. Beyond the layers of regulation prior to transcription such as transcription factors and chromatin structure, several processes regulate newly synthesized mRNA post transcription. Between transcription and translation, mRNAs undergo alternative splicing events and nuclear export, which are tightly controlled. Translation efficiency, mRNA stability and miRNA-mediated mRNA decay, to name a few, represent additional means of regulation which are largely controlled by RNA binding proteins.[1]

Human PUMILIO proteins, PUM1 and PUM2, are members of the PUF family of sequence-specific RNA binding proteins. Originally discovered in *Drosophila*, they characteristically bind to a conserved consensus RNA sequence, the Pumilio Response Element, located in the 3'UTR of their targets where they mainly act as translational repressors. The scope of PUM1/2 mediated regulation has not been fully elucidated, however, mRNAs and non-coding RNAs from 7822 genes were found to have at least one Pumilio Response Element in their 3'UTR.[2]

PUMILIO proteins do not act on their own but form functional complexes with other RNA binding proteins. However, the understanding of the PUMILIO interactome is still limited. In *Drosophila*, *pumilio* acts in complex with *nanos* and *brat*. Whereas NANOS proteins exist in humans, the human homolog of *brat* is yet to be confirmed. Both E3 ubiquitin-protein ligase family members TRIM71 or and TRIM32 have been suggested to be the human homologs of *brat*.[3,4] Beyond that, there is growing evidence for PUMILIOs interaction with several other proteins. One potential interactor of PUMILIO is DND1, which has been shown to interact with NANOS2, a known PUMILIO binding partner. It is believed that DND1 could mediate the binding of NANOS2 and Pumilo.[5] Another putative binding partner of PUMILIO is FMRP, which was found to act in complex with PUM1/2. FMRP and PUM1/2 were also found to regulate each other's mRNAs.[6] Besides regulating mRNA targets, PUM2 has been shown to bind to the abundant long non-coding RNA (lncRNA) NORAD. The PUM-NORAD binding is believed to be facilitated by the protein SAM68.[7] The alternative splicing regulator RBFOX2 binds to a sequence similar to the Pumilio Response Element in the 3'UTR of mRNA transcripts. As PUMILIO itself contains Pumilio Response Elements in its 3'UTR, RBFOX2 is a putative post-transcriptional regulator of PUMILIO.[8] Modifications within Pumilio Response Elements such as *N*6-methylation on adenosines (m6A) have recently been associated with a weakened PUM2 binding. M6A methylation is catalyzed by METTL3, which makes it a putative indirect PUMILIO regulator.[9]

In sum, PUMILIO is a highly conserved post-transcriptional regulator with an abundance of potential interactors. The large number of potential targets stresses its relevance for development and disease.

1.2 Head and neck squamous cell carcinoma

Head and neck squamous cell carcinoma (HNSCC) is one of the most widespread cancer types accounting for more than 650,000 incidences and 330,000 deaths per year worldwide.[10] In the US, head and neck cancer makes up for 3% of all new cancer diagnoses and 1.8% of cancer related deaths per year.[11] Tumors develop from mucosal epithelial cells in several anatomical structures in the oral cavity as well as the upper airway and food passages. An overview of the sites of HNSCC is given in Figure 1.

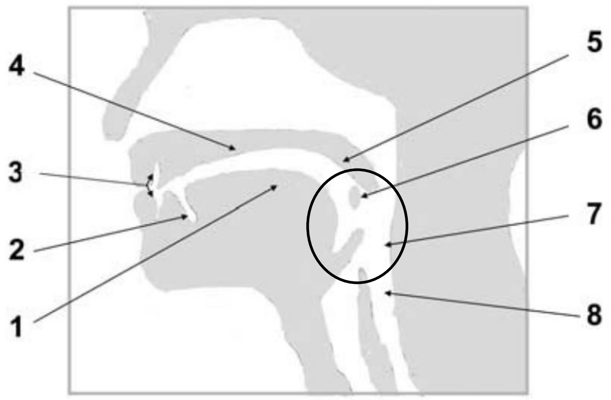


Figure 1. in head and neck squamous cell carcinoma (adapted from Lambert et al., 2011)[12]
1 tongue, **2** floor of the mouth, **3** lips, **4** hard palate in oral cavity, **5** soft palate in oral cavity, **6** tonsils, **7** oropharynx and **8** hypopharynx. All structures that are summarized under oropharyngeal tumors are indicated by the oval shape.

Risk factor profiles of HNSCC subtypes

Despite the seemingly homogenous cell population that gives rise to the tumors, HNSCC is a rather heterogeneous disease leading to large differences in response to treatment, prognosis and patient survival. The etiology of HNSCC has largely been associated with classical risk factors such as tobacco use and excessive alcohol consumption.[13] More recently, the infection with human papillomaviruses HPV-16 and HPV-18 has emerged as a pronounced risk factor as an increased number of HPV-related cases has been observed in western countries.[14,15] HPV-driven cancers mainly originate in the oropharynx which comprises the base of the tongue, the tonsils, the soft palate, and the walls of the pharynx (Figure 1). In oropharyngeal cancers, HPV infection accounts for 13-60% of the cases and based on current trends, this number is likely to increase.[16,17] HPV positive and negative oropharyngeal tumors represent two distinct clinical entities as they differ in several clinical and molecular aspects. HPV positive tumors tend to occur at an earlier age than HPV negative tumors. A history of alcohol and tobacco consumption is less common in the HPV positive group, and overall, HPV positive patients have a higher socioeconomic status.[16]

Molecular mechanisms in HPV positive and HPV negative HNSCC

The etiology of HPV positive HNSCC is mainly driven by two early genes in the viral genome. E6 and E7 act as viral oncoproteins by targeting and degrading the human tumor suppressors cellular tumor antigen p53 (p53) and Retinoblastoma-associated protein 1 (pRb). Upon infection, HPV-E6 binds to a host ubiquitin ligase leading to ubiquitin-mediated proteasomal degradation of p53.[18] P53 potentially is the most prominent tumor suppressor protein and is altered in more than 50% of human malignancies. In the event of DNA damage or cellular stress, wild-type p53 preserves DNA integrity by inducing DNA repair, cell cycle arrest, apoptosis or senescence.[19] Wild-type pRB acts as a suppressor of cell proliferation by preventing the transcription factor E2F from transcribing its targets. E2F regulated genes are involved in the G1-S phase transition of the cell cycle which is restricted when pRb is active. HPV-E7 inactivates pRb, leading to its functional loss and uncontrolled cell cycle progression mediated by unrestricted E2F binding to its target genes. The protein p16, encoded by cyclin-dependent kinase inhibitor 2A (*CDKN2A*), is a repressor of pRb and is indirectly regulated by pRb through a negative feedback loop. Thus, p16 levels increase when pRb is inactivated by E7.[20]

In contrast to HPV driven head and neck cancers, the etiology of HPV negative head and neck cancers is more complex. With alcohol- and tobacco use being the main risk factors, the disease is driven by chemical mutagens which explains the overall higher age at diagnosis. Many patients develop pre-cancerous lesions in the mucosal linings which may turn into malignant tumors as they become more dysmorphic and genetic changes accumulate. Mutations in p53 emerged as early predictors for malignant transformation.[21] Further, the loss of heterozygosity for the chromosomal regions 3p and 9p have been associated with a high risk of tumor development from oral premalignant lesions.[22] Not surprisingly, *CDKN2A,* encoded in the 9p21.3 locus, is frequently lost in HPV$^-$ tumors.

Mutational signatures in HPV$^+$ and HPV$^-$ tumors affect multiple cancer pathways

Overall, HPV positive tumors show low mutation rates for *TP53* and thereby differ significantly from most other tumors. Also, low mutation rates of *CDKN2A* and low expression levels of pRb and cyclin D1 have been observed. A comprehensive study on the genetic changes in head and neck cancer from 2015 found mutated TNF receptor-associated factor 3 (*TRAF3)*, activating mutations of *PIK3CA*, and amplification of *E2F1* to be characteristic for HPV positive tumors. TRAF3 has been linked to anti-viral responses whereas *PIK3CA* and *E2F1* aberrations point to a dysregulation of the cell cycle. [23,24]

In HPV$^-$ HNSCC, p53 aberrations were found to play an important role in the etiology of the disease. Inactivating *TP53* mutations are found in 50-80% of the cases, especially in tobacco-related HNSCC.[24] In contrast to HPV$^+$ tumors, p16 is frequently lost in connection with overexpression of cyclin D1, which results in G1-S checkpoint dysregulation. While gene amplifications have been found for cyclin D1 (*CCND1)* and growth factor receptors such as EGFR and FGFR1, multiple inactivating mutations were found in genes involved in Wnt-signaling (*FAT1* and *AJUBA*).[24,25]

Transcriptional profiles of HNSCC subtypes

Numerous studies have focused on identifying transcriptional profiles of HPV^+ and HPV^- head and neck cancers. Although different sample material from different anatomical sites has been used among the studies, several differentially expressed genes were found regardless of these influence factors. Differentially expressed genes were found to be involved in a large number of cellular processes.

In HPV^+ tumors, a frequent upregulation of cell cycle control genes (e.g. *CDKN2A*, *CDC7*, *MCM2*) as well as genes involved in DNA replication (e.g. *RCF4*) has been observed. Beyond that, genes that are known to be regulated by p53 or E2F were found to be expressed at higher levels in the HPV^+ group compared to normal tissues.[26–28] Related to previously mentioned mutational differences between the two groups, an elevated expression of genes in the 3q24 locus has been found compared to HPV^- tumors.[27] Transcripts that consistently were found to be downregulated in HPV^+ tumors versus normal controls are involved in multiple cellular processes, for example cell differentiation. Immune response genes such as interleukins (IL-10 and IL-13) as well as interferon-induced proteins (IFIT1, IFITM1-3, IFI6-16 and OAS2) were frequently downregulated in HPV^+ tumors compared to a HPV^- group.[28] In HPV^- tumors, differentially expressed genes were found to be involved in cell signaling and signal transduction such as endothelial cell growth factor 1 (*ECGF1*) or insulin-like growth factor binding protein 5 (*IGFBP5*) compared to normal adjacent tissue.[26]

Long non-coding RNAs in cancer

Non- coding RNAs are RNAs that are not translated into a protein sequence. Based on their length, non-coding RNAs longer than 200 bp are considered lncRNAs. Like mRNAs, lncRNAs are often transcribed by RNA polymerase II and can undergo splicing. The term lncRNAs refers to a fairly diverse group of transcripts which are often subclassified according to their length, function or other transcript properties.[29]

Although many functions of lncRNAs are still poorly understood, several well-characterized lncRNAs have shed light on the roles these molecules may play in human cells. LncRNAs have been found to be involved in chromatin remodeling which regulates gene expression by influencing accessibility of a certain genomic region. The most well-studied example for this functional process is X-inactive specific transcript (XIST) which is involved in X-chromosome inactivation in females. The role of XIST in cancer is controversial. The interaction of XIST with BRCA1 was believed to ensure proper X-inactivation which would be lost in the event of *BRCA1* mutations that are frequently seen in breast cancer. However, this functional interaction is highly debated as several studies could not confirm co-localizations of BRCA1 and XIST.[30] In mice, XIST was found to suppress hematological cancers.[31]

Another lncRNA that acts through chromatin remodeling is HOTAIR. First studied in breast cancer cells, it recruits Polycomb repressive complex 2 which leads to trimethylation on lysine 27 on histone 3

(H3K2) and thereby to epigenetic silencing of metastasis suppressing genes. Accordingly, HOTAIR overexpression has been associated with an increased tumor invasiveness and metastasis formation whereas downregulation has had a protective effect.[32]

Furthermore, lncRNAs can act as transcriptional co-regulators, contributing to either transcriptional activation or repression of their targets. LincRNA-p21 is a target gene of p53. When transcribed, it has been shown to repress other p53 target genes in association with the RNA binding protein hnRNA-K.[33] By mediating the transcriptional repression of p53 targets, it might play an important role in tumor suppression.[34]

Through regulation of the location and levels of splicing factors, certain lncRNAs such as MALAT1 have been shown to regulate alternative splicing events.[35] In non-small cell lung cancer, MALAT1 was associated with a poor prognosis and was shown to promote cell migration and growth.[36]

LncRNAs are able to regulate miRNA-mediated decay of mRNA transcripts. The most well-studied example for this process is Phosphatase and Tensin homolog *PTEN* and its pseudogene *PTENP1*. PTENP1 competes for binding of miRNAs that target PTEN mRNA and thereby prevents mRNA degradation. PTENP1 was found to be downregulated in several human cancers.[37]

lncRNAs in head and neck cancer

Several of the aforementioned lncRNAs have been linked to subsets of head and neck cancers. Associations between lncRNAs and cancer phenotypes have mostly been investigated for specific anatomical sites of HNSCC.

In laryngeal squamous cell carcinoma, high expression levels of HOTAIR were associated with poor differentiation and seemed to promote malignant progression.[38] Similar effects were seen in oral squamous cell carcinoma (OSCC)[39], however, no upregulation was seen in tongue squamous cell carcinoma (TSCC).[40] In addition to HOTAIR, two other well-characterized cancer-related lncRNAs seem to predict metastatic growth in OSCC: Nuclear enriched abundant transcript 1 (NEAT-1) and Urothelial cancer associated 1 (UCA1). The expression of these transcripts increased in with metastatic status and was detectable in saliva. Maternally expressed 3 (MEG-3), a lncRNA which is frequently lost in cancer, also had decreased levels in OSCC.[39] In TSCC, UCA1 expression also correlated with the presence of lymph node metastases whereas NEAT-1 and MEG3 do not seem to play a role in TSCC.[40]

CDKN2B antisense RNA 1 (ANRIL), like HOTAIR, acts on the Polycomb repressive complex. It was found to be highly expressed in HPV$^-$ HNSCC irrespective of anatomical site. When knocked down in HNSCC cells, it inhibited proliferation. An additional function for ANRIL was found as it is a competing endogenous RNA which modulates miR-125a-3p and its downstream target such as fibroblast growth factor 1 (FGF1). Through this mechanism, it was suggested to act on the MAP kinase pathway and eventually lead to tumor growth.[41]

The associations of functionally characterized lncRNAs with head and neck cancer are a promising spotlight on the non-coding transcriptome of these tumors. However, few large-scale transcriptomic studies have attempted to achieve a more comprehensive understanding of the non-coding landscape in head and neck cancer. A study by Salyakina and Tsinoremas characterized the non-coding transcriptomic profiles of 442 HNSCC samples according to their HPV status. Surprisingly, they could not identify any of the well characterized lncRNAs that had been associated with subtypes of HNSCC earlier.[42]

Long non-coding RNAs have several inherent properties that rise the potential to exploit them as biomarkers for cancer diagnostics, disease monitoring, prognosis and therapy as well as for research. The expression pattern of lncRNAs is highly tissue-specific compared with coding genes and they are often co-expressed with neighboring genes.[43] Thus, lncRNA profiles might be exploited to elucidate the composition of various cell populations in a tumor. In contrast to mRNA, lncRNAs exert their function without the necessity of translation which allows for direct correlations to the respective outcome. The number of studies linking lncRNA expression to patient outcome has risen dramatically. Even though a thorough validation is needed for a lncRNA to be used as a predictive biomarker, several studies have shown that certain lncRNAs such as HOTAIR reliably predict disease progression and patient outcome.[32] An example for a successful translation into the clinical setting is PCA3, a lncRNA highly associated with prostate cancer, which is now screened for in urine samples as a routine procedure making prostate biopsies largely obsolete.[44]

Collectively, the investigation of the non-coding transcriptome in head and neck cancer could yield significant advancements in the search for novel biomarkers.

2. Aim of study

This thesis covers two separate projects. The first project aimed to map out the interactome of a set of putative interactors and mediators of the post-transcriptional regulator PUMILIO.

The aim of the second and main project was to explore transcriptional signatures of different subtypes of HNSCC with a special focus on non-coding transcripts.

3. Materials and Methods

3.1. Cell lines and culture conditions

HeLa Kyoto (HeLa-K) strains stably transformed with human BACs were used to express LAP-tagged METTL3, FMR1, and HSP90AA1 at physiological levels. HeLa cells expressing non-fused eGFP as well as other BAC-transformed strains (CDKN2A, RC3H1, WTAP and HSPB1) were used as positive controls. HeLa wild-type cells were used as a negative control. All BAC constructs used in this study had the LAP-cassette fused to the C-terminus (see supplementary Figure S1). The BAC-transformed HeLa-K strains as well as HeLa-K wild-type cells were received through the TransgeneOmics project from Dr. Ina Poser at the Max Planck Institute for Molecular Cell Biology and Genetics (Dresden, Germany)[45]. HeLa-K cells expressing only the eGFP tag had been stably transformed through lentiviral infection at Dr. Wayne Miles's lab. HeLa-K cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% FBS, 1% L-Glutamine and 1% Penicillin/Streptomycin and kept at 37°C and 5% CO_2. BAC-transformed cells were kept under a maintenance dose of 400 µg/ml G418 to impose selection pressure. HeLa-K cells infected with the lentiviral expression construct were kept under a maintenance dose of 5 ng/µl Neomycin.

For the second project, six different patient-derived HNSCC cell lines, FaDu, Cal33, SCC2, SCC4, SCC90 and HSC3, were used. OKF6 was used as a non-cancer control cell line. The cells were a gift from Dr. James W. Rocco's lab at The Ohio State University Comprehensive Cancer Center (Columbus, OH, USA). All HNSCC cells were grown in Ham's F-12/DMEM (3:1) medium supplemented with 10% FBS and 1% penicillin/streptomycin at 37°C and 5% CO_2.

OKF6 cells were cultured in equal parts of Keratinocyte-SFM medium and DFK medium. Keratinocyte-SFM medium (1X) with L-glutamine and $CaCl_2$ was supplemented with 0.2 ng/mL human recombinant epidermal growth factor (EGF) 1-53, 25 µg/mL bovine pituitary extract (BPE) and 1% penicillin/streptomycin. DFK medium was made with equal parts of DMEM and Ham's F-12 supplemented with 0.2 ng/mL EGF 1-53, 25 µg/mL BPE, 2mM L-glutamine and 1% penicillin/streptomycin.

3.2 Isolation of genomic DNA and accomplishment of PCR

Genomic DNA was isolated from 3 x 10^6 cells for each transgenic HeLa-K cell line and WT HeLa-K cells. Each cell pellet was incubated with 100µl tissue digestion buffer (see appendix for details) at 56°C overnight. Then, samples were spun down briefly and 100µl phenol/chloroform/isoamyl alcohol (25:24:1) pH 6.7± 0.2 was added and briefly vortexed. Samples were spun down 5min at 18,000x*g*. The aqueous layer (ca. 100µl) was transferred to new tubes, thoroughly mixed with 10µl 3M NaAc pH 5.5 and spun down briefly. 250µl 100% ethanol were added to each sample, mixed and centrifuged at

18,000x*g* for 10min. The supernatant was discarded, 250µl 70% ethanol was added, the samples were centrifuged for 3min at 18,000x*g* and the supernatant was poured off. The pelleted DNA was dried and resuspended in 50µl TE buffer pH 8.0.

Forward primers for *METTL3, FMR1* and *HSP90AA1* were designed using the NCBI Primer Design tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast) restricting the region of interest to the last exon of the respective gene (Appendix A3). A forward primer for eGFP was designed based to match the start of the coding sequence. A reverse primer binding within the coding sequence of eGFP was used in all reactions. All primers were selected to match the reverse primer in terms of GC content, melting temperature and amplicon length. The conditions for genomic PCR are stated below (For the manufacturer – see Appendix A1).

PCR mix		**Cycling conditions**		
Per reaction				
10µl	5x One*Taq* GC buffer	94°C		2min
1µl	dNTP mix		94°C	30s
1µl	Forward primer 10µM	39x	55°C	30s
1µl	eGFP reverse primer 10µM		72°C	1min
150 µg	Genomic template DNA	72°C		5min
0.25µl	One*Taq* polymerase	12°C		∞
add	ddH$_2$O			
50µl	*total reaction volume*			

The resulting PCR products were loaded on 1% agarose gels. Amplicons of the expected size were confirmed using Sanger sequencing at university's genomics core facility and aligned with NCBI BLASTn.

3.3 Protein isolation and quantification

Cells were harvested at 80-90% confluency (~ 5 x 10^6 cells). After aspirating the culture media, cells were washed once with 5ml Phosphate-buffered saline (PBS). Thereafter, cells were scraped in 10ml PBS and centrifuged at 100x*g* for 5min. The pellet was transferred to a 1.5ml Eppendorf cup and pelleted again at 10,000x*g* for 1min to remove residual PBS. For whole cell lysis, two different protocols were used, involving cell lysis with RIPA buffer and a slightly harsher 4% SDS lysis. The RIPA lysis was performed by resuspending the cell pellet in 200µl RIPA buffer on ice and syringing the suspension 10x with an 18-gauge needle. For SDS lysis, the cell pellets were resuspended in 200µl 4% SDS on ice and subjected to seven pulses of sonication. After lysis, samples were kept on ice for 30min and thereafter centrifuged for 10min at 10,000x*g* and 4°C to separate cell debris. The protein concentration of the supernatant was assessed using the Pierce™ BCA Protein Assay Kit according to the user's manual with 2µl of whole cell lysate diluted in 18µl PBS. The Epoch™ Microplate Spectrophotometer (BioTek Instruments, Inc.) was used for assessing absorbance at wavelength 562nm.

3.4 Western Blot

Whole cell lysates were mixed with 3x SDS-PAGE loading buffer, incubated at 95°C for 5min and 40µg loaded on a 10% SDS gel. Samples were run at 100V throughout the stacking gel. The voltage was increased to 200V once the samples had entered the resolving gel.

Blotting was performed with the Trans-Blot® Turbo™ RTA Midi PVDF Transfer Kit. The membranes were soaked in 100% Ethanol for 2min, washed in H_2O for 5min and equilibrated in 1x Trans-Blot® buffer for 10min. For blotting, the BIO-RAD Trans-Blot Turbo transfer system with settings for midi gels was used. The membrane was blocked in blocking solution for 1h at room temperature. Thereafter, the membrane was incubated with primary antibody at 4°C overnight. The membrane was then washed three times with PBST for 10min per wash. Subsequently, an HRP-coupled secondary antibody was incubated for 1h at room temperature. The HRP signal was detected using the ECL™ Prime Western Blotting Detection Reagent according to the manufacturer's instructions and visualized with the LI-COR® Odyssey® Fc Imaging System (LI-COR, Inc.). A list of primary and secondary antibodies and the respective dilutions is given in Appendix A5.

Control stains were performed by re-probing with antibodies of a different species following the procedure as described above from the blocking step. In order to re-probe western blot membranes with antibodies of the same species, antibody signal was removed through a stripping protocol. Dried membranes were re-hydrated in PBST for 1h and incubated with 10ml Restore™ Stripping Buffer for 15min at 37°C. Thereafter, membranes were washed once and stored in PBST at 4°C until blocking.

3.5 Cell fixation for fluorescence microscopy

HeLa-K cells (~$3x10^5$ cells per well) were seeded onto glass cover slips in 6-well plates and grown until 50% confluency. Cells were washed once with PBS and incubated with 4% paraformaldehyde for 30min at room temperature. Thereafter, the coverslips were washed once and permeabilized with 0.5% Triton X-100 in PBS for 5min at room temperature. After permeabilization, cells were washed three times with PBS and incubated with 300 nM DAPI in PBS for 5min in a dark chamber at room temperature. The cells were then washed three times with PBS and mounted on microscopy slides with Immu-Mount™ mounting solution.

3.6 Gateway® cloning

Expression plasmids for establishing stable transgenic cell lines were cloned utilizing the Gateway® System. First, an entry vector is created by amplifying the insert of interest with a CACC-overhang at the 5' end and ligating it into the directional cloning vector pENTR™/D-TOPO® through a topoisomerase reaction. The entry vector is transformed into ultra-competent OneShot® TOP 10 *E.coli* cells, amplified and purified. Secondly, the insert of interest is switched into the destination vector through the LR recombination reaction which forms the expression vector.

Inserts of interest were amplified from plasmids carrying the respective coding sequence. A full list of manufacturers and plasmids is given in the Appendix A1 and A4 respectively. Primers for the amplification of inserts were designed according in line with the pENTR™ Directional TOPO®Cloning user's guide[46]. SnapGene® Viewer (version 4.2.4) and the NEB Tm calculator (version 1.10.4, http://tmcalculator.neb.com/#!/main) were used to match melting temperature and GC content. Due to several GC-rich inserts resulting in high primer melting temperatures, a two-step PCR protocol was used and/or a specific PCR buffer for GC-rich templates (GC-buffer).

Conditions for insert amplification PCR

PCR mix		**Cycling conditions**			**Cycling conditions (two-step)**		
Per reaction							
10µl	5x PCR buffer (HF or GC)	95°C		3min	98°C		3min
1µl	dNTP mix	34x	95°C	30s	35x	95°C	30s
1µl	Forward primer 10µM		T_a*	30s		72°C	30s per kb
1µl	Reverse primer 10µM		72°C	1min 10s per kb	72°C		10min
1µl	Template DNA 20ng/µl	72°C		5min	12°C		∞
1µl	Phusion® polymerase	12°C		∞			
35µl	ddH_2O						
50µl	*total reaction volume*						

*T_a = annealing temperature. Adjusted based on optimal annealing temperature for respective primer pair.

PCR products were mixed with 6x loading buffer and loaded on a 1% agarose gel. For size estimation, the 1kb HyperLadder™ was used. Bands with the expected size were cut out and the DNA was purified using the QIAquick® Gel Extraction Kit according to the user's manual.

The TOPO-ligation was performed using the pENTR™/D-TOPO® Cloning Kit according to the recommendations from the TOPO®Cloning user's guide. Briefly, 2µl of gel-purified PCR product was mixed with 1µl salt solution, 1µl pENTR vector, 2µl sterile H_2O to a final volume of 6µl and incubated at room temperature for 30min.

OneShot® TOP 10 *E.coli* cells were thawed on ice and gently spun down. 1µl of the TOPO-ligation reaction was added to 25µl of cell suspension and incubated for 5min on ice. Cells were heat shocked at 42°C for 30 seconds and immediately put on ice for 5min. To each reaction, 125µl S.O.C recovery medium pre-warmed to 37°C was added and samples were incubated at 37°C for 1h. Subsequently, the samples were plated on LB-*Kan* agar plates and incubated at 37°C overnight.

Transformants were screened by mini-prep and subsequent restriction digestion. 3ml of LB medium were inoculated per single colony and incubated in a shaker at 37°C and 250 rpm overnight. Plasmid

DNA was isolated from the culture using the QIAprep® Spin miniprep Kit according to the manufacturer's instructions. 30µl of each culture were saved for later use.

The concentration of eluted plasmid DNA was measured with the Epoch™ Microplate Spectrophotometer (BioTek Instruments, Inc.). For size verification, DNA was linearized with the single cutter *NotI* and loaded on a 1% agarose gel. The digestion was assembled according to NEBcloner®v1.3.12 (http://nebcloner.neb.com/#!/redigest) and scaled down to a total volume of 25µl, containing 500ng of plasmid DNA, 2.5µl 10x 3.1 buffer, 0.5µl *NotI* and the respective volume of dH_2O. The digestion was incubated for 1h at 37°C. Fragments which showed the correct size were confirmed through Sanger sequencing. Clones that showed the correct sequence were amplified by inoculating 100ml LB medium + *Kan* with 10µl of the retained mini-prep culture, shaking at 37°C and 250 rpm overnight. Plasmid DNA was isolated using the HiSpeed® Plasmid Midi Kit according to the manufacturer's instructions and the DNA concentration was measured.

150 ng of the entry vector and destination vector were mixed with 1µl LR Clonase™ II Enzyme Mix and filled up to a volume of 5µl with sterile H_2O. The reaction was incubated at room temperature overnight. Thereafter, 0.5µl Proteinase K was added and incubated for 10min at 37°C. 2µl of the recombination reaction was used to transform 25µl of chemically competent DH5α *E.coli* cells. The cell suspension was thawed on ice and incubated with the recombination reaction for 15min on ice. Cells were heat shocked at 42°C for 45 seconds and put on ice immediately after for 5min. 250µl pre-warmed S.O.C medium was added. Thereafter, transformation reactions were treated as described earlier.

3.7 RNA isolation

Cells were harvested for RNA isolation at 80-90% confluency. The culture medium was aspirated, and cells were washed once with ice-cold PBS while keeping the plate on ice. For each 10 cm dish, 450µl lysis buffer was dripped onto the cells. Cells were scraped, transferred to an Eppendorf tube and incubated on ice for 10min. After incubation, the cell suspension was syringed ten times through a 26-gauge needle. Cell debris was cleared by centrifugation for 10min at 20,000x*g* and 4°C. 100µl of the supernatant was removed for total RNA isolation using the RNeasy® Mini Kit according to the manufacturer's instructions. RNA concentration was measured with the Epoch™ Microplate Spectrophotometer and RNA quality was assessed on a Bioanalyzer. For library preparation, RNA with RIN numbers >9 was used.

3.8 Library preparation and RNA seq

Two independent RNA libraries were prepared from different cell batches resulting in two biological replicates per cell line**.** RNA-seq libraries were prepared from using the NEBNext® UltraTM II

Directional RNA Library Prep Kit for Illumina® according to the provided manual. First, poly-adenylated RNA was isolated with the NEBNextPoly(A) mRNA Magnetic Isolation Module utilizing magnetic oligo-dT beads. The RNA was then fragmented and primed with random primers. Subsequently, the first and second strand of cDNA were synthesized, and double-stranded cDNA was purified with sample purification beads. cDNA ends were prepped to make them compatible for ligation with NEBNext Adapters. Ligated fragments were again purified with sample purification beads and enriched by PCR. After purification of the PCR product with sample purification beads, the library quality was assessed on a Bioanalyzer. RNA sequencing was carried out by Novogene Inc. on an Illumina HiSeq 4000 platform.

3.9 Quantitative real-time PCR

Primers for quantitative PCR were designed using the NCBI primer design tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast). Templates for the primer design were defined by the respective RefSeq ID of each gene given in table 6. Primers were designed to amplify a 100 -150 bp fragment within the coding sequence (if applicable). For cDNA synthesis, RNA from both biological replicates used for RNA sequencing was diluted to 10µl aliquots containing 250ng of RNA. 10µl master mix was added to each sample and reverse transcription was performed. Components for the master mix are described below. For manufacturer information, see Appendix A1.

cDNA synthesis mix		**Cycling conditions**	
Per reaction			
10µl	250ng RNA	25°C	10min
2µl	10x RT buffer	37°C	120min
0.8µl	25x dNTP mix (100mM)	85°C	5min
2µl	10x RT random primers	4°C	∞
1µl	MultiScribe® Reverse Transcriptase		
4.2µl	dH_2O		
20µl	*total volume*		

cDNA was diluted 1:10 for subsequent use in qPCR. qPCR was carried out in 96-well plates with a total reaction volume of 15µl. Per well, 7.5µl SYBR® Green Master (2x), 1µl primer mix (forward and reverse 1:1, each at a concentration of 10µM), 2.5µl of diluted cDNA and 3µl dH_2O were mixed and run on the StepOnePlus™ Real-Time PCR System with 40 cycles.

All reactions were carried out in duplicates of biological replicates and considered quadruplicates for quantification. *GAPDH* was used as a housekeeping gene for normalization. Selected genes that showed a negative log2-Fold change in the RNA sequencing data were tested in HPV negative cell lines and normalized to one HPV^{+} control and the normal control. Vice versa, genes with a positive log2-Fold change were tested in HPV^{+} cells and normalized to one HPV^{-} control and the normal control.

3.10 Pre-processing of RNA sequencing data from HNSCC cell lines

All pre-processing steps of raw RNA sequencing reads from six HNSCC cell lines and the non-cancer control cell line were carried out on Galaxy, an open source, web-based platform for data research in biomedical sciences (https://usegalaxy.org/)[47]. Subsequent analyses were carried out in R (version 3.5.2, Platform: x86_64-w64-mingw32/x64, 64-bit).

FASTQ files containing raw paired-end sequencing reads from both sequencing runs were made accessible by Novogene Inc. and uploaded to Galaxy. All pre-processing steps were carried out with the default settings for paired-end reads implemented on the platform unless specified otherwise. Read quality and adapter contamination was assessed using fastQC (Galaxy Version 0.72). Adapter sequences were clipped, and low-quality reads removed using Trim Galore! (Galaxy Version 0.4.3.1) with the adapter sequence specified as "Illumina universal" and otherwise default settings. Subsequently, all samples were depleted for rRNA and tRNA sequences using bowtie2 (Galaxy Version 2.3.4.2, files available upon request). The remaining reads were mapped to the human genome (hg38) using HISAT2 (Galaxy Version 2.1.0+galaxy3). The resulting BAM file was converted into a matrix of raw counts annotated to Entrez Gene IDs through featureCounts (Galaxy Version 1.6.3+galaxy2), downloaded and saved as a CSV file. The pre-processing workflow can be accessed through https://usegalaxy.org/u/lorks.1/w/complete-mapping. Entrez IDs from the featureCounts output were annotated to the respective gene symbols, full gene names and RefSeq IDs with the AnnotateMyIDs tool (Galaxy Version 3.5.0.1).

3.11 Principal component analysis and differential expression analysis

Principal component analysis and differential expression analysis was carried out using the DESeq2 package (version 1.22.2)[48]. The ggplot2 package (version 3.3.1)[49] was used for visualization of the data. The codes were initially taken from the vignettes of the respective packages but had to be optimized for the purposes of this study. Explanations for each section of code are led by a hashtag. For each independent analysis, slight modifications were made to match the file- and feature names. Examples are included in the code sections below.

The featureCounts output and the categorized cell line information as presented in table 1, were imported into R and incorporated into a DESeq data set. The DESeq data set requires the specification of an experimental design. Since the experimental design was supposed to be based on the following principal component analysis, the design formula was later re-adjusted according to the findings.

```
#import CSV files
featureCounts <- read.csv(file, row.names=1, sep=";")
metadata <- read.csv(file, row.names=1, sep=";")

#create DESeq2 data set
library(DESeq2)
dds <- DESeqDataSetFromMatrix(countData = featureCounts, colData =
metadata, design = ~ cell_line)
```

For explorative data analysis, counts were transformed through the variance stabilizing transformation function vst() blind for experimental design and the first two principal components (PC) were plotted.Color and shape of the data points were specified to show a desired combination of features. In the cell line dataset, “cell line” was indicated by color and the shape indicated the desired category from table 1. In the TCGA dataset, HPV status was indicated by shape whereas other features were indicated by color. The different PCA plots were generated to determine which factors account for cluster formation. These factors were later included in the DESeq design formula.

```
#variance stabilizing transformation regardless of design formula
vsd <- vst(dds, blind = TRUE)

#Principal component analysis and PCA plot
pcaData <- plotPCA(vsd, intgroup = c( "HPV_status", "Tissue",
"Mutations", "run"), returnData = TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"))
ggplot(pcaData, aes(x = PC1, y = PC2, color = Cell_line, shape =
HPV_status)) +  geom_point(size =3) +
  xlab(paste0("PC1: ", percentVar[1], "% variance")) +
  ylab(paste0("PC2: ", percentVar[2], "% variance")) +
  coord_fixed()
```

Prior to running the differential expression analysis, the experimental design was adjusted and a reference level for the dependent variable was set. Differential expression was calculated by running the DESeq() function on the untransformed counts from the DESeq data set with default settings. Results were extracted with the results() function.

The contrast displayed in the results file was specified by naming the respective category and the levels that should be contrasted against each other. Although the contrast was specified earlier by setting the reference level, this step assured the correct comparison. Results were exported for further processing using MS Excel.

```r
#re-adjust design formula
design(dds) <- formula(~tissue + HPV_status)
#set reference level
dds$HPV_status <- relevel(dds$HPV_status, ref = "control")
#run DESeq2
dds <- DESeq2(dds)

#save results in results object
res <- results(dds)

#save results with specified contrast
res <- results(dds, contrast = c("HPV_status", "positive",
"negative"))

# save results in .csv file
write.csv2(res, "file path/filename.csv")
```

The cutoffs for differential expression were set to an adjusted p-value of *padj ≤ 0.05* and an absolute log2-fold change of |*l2fc*| ≥ *2.* Rows without gene symbol or gene name were removed.

Validation of results in TCGA dataset

Annotated RNA seq read counts from 263 head and neck cancer samples were retrieved from FireBrowse (http://firebrowse.org/). Clinical information for these individuals was downloaded from the TCGA Research Network (https://portal.gdc.cancer.gov/)[24]. Read counts and clinical data were subjected to explorative data analysis and differential expression analysis as described above. After differential expression analysis, the exported DE results tables from the cell line dataset and the TCGA dataset were merged and EntrezIDs that were not contained in both datasets were removed. The merged table were then filtered for common directionality of differential expression and classified into coding and non-coding genes.

GO term overrepresentation analysis

Overrepresentation of GO Molecular Function terms within the validated DE genes was examined using the enrichGO() function from clusterProfiler (version 3.10.1)[50]. Results were visualized in a heat plot using enrichplot (version 1.2.0)[51].

```r
# define and sort gene list
d <- commonDE_CL_TCGA
mygeneList = d[,2]
names(mygeneList)=as.character(d[,1])
mygeneList = sort(mygeneList, decreasing = TRUE)

library(clusterProfiler)
library(enrichplot)
library(org.Hs.eg.db)

#specify organism (homo sapiens) and ontology (molecular function)
MF <- enrichGO(names(mygeneList), OrgDb = org.Hs.eg.db, ont = "MF",
readable = TRUE)
resMF <- as.data.frame(MF)

#plot and export results
heatplot(MF)
write.csv2(resMF, "file path/filename.csv")
```

Pre-selection of genes to be validated by qPCR

Coding and non-coding genes that were shown to be differentially expressed in both cell line and patient data were considered for further validation. To assure that transcripts would be detectable in a qPCR experiment, the pre-selection was based on FPKM and CPM values obtained through edgeR (version 3.24.3).[52]

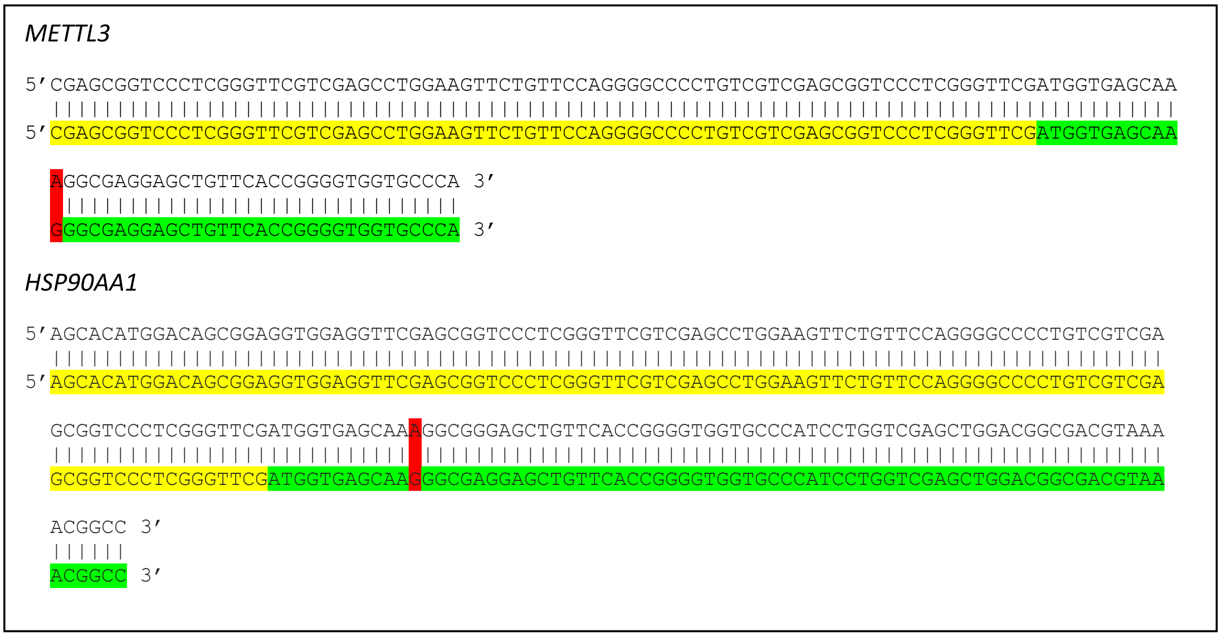
```r
#calculate CPM values
cpm <- cpm(featureCounts, log = FALSE)
write.csv2(cpm, "file path/filename.csv”)

#calculate RPKM values
RPKM <- rpkm(featureCounts, gene.length = featureCounts$genelength,
log = FALSE)
write.csv2(RPKM, "file path/filename.csv")
```

For each gene, the median CPM and FPKM value was calculated based on the annotated counts from the cell line RNA sequencing data. Secondly, the median of all median CPM and FPKM values was determined. Only genes that had a median CPM and FPKM value above the median(median CPM) and median(median FPKM) respectively were considered for qPCR.

4. Results

4.1 BAC-transformed cell lines are not suited for protein complex precipitation

In order to verify the insertion of the eGFP-tagged protein in the BAC-transformed cell lines, we performed genomic PCR with primers spanning across the junction of the last exon of the gene of interest and the LAP tag. The expected fragment lengths of ca. 600bp (*METTL3-eGFP:* 639 bp, *HSP90AA1-eGFP:* 599 bp, *FMR1-eGFP*: 625 bp) were confirmed by gel electrophoresis (data not shown) and sequenced. Figure 2 shows two representative examples for aligned sequencing results from the gel-purified fragments with highlighted sequence features. We were able to confirm the presence of the LAP-tag sequence for all BAC-transformed cell lines tested.



METTL3

```
5’CGAGCGGTCCCTCGGGTTCGTCGAGCCTGGAAGTTCTGTTCCAGGGGCCCCTGTCGTCGAGCGGTCCCTCGGGTTCGATGGTGAGCAA
  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5’CGAGCGGTCCCTCGGGTTCGTCGAGCCTGGAAGTTCTGTTCCAGGGGCCCCTGTCGTCGAGCGGTCCCTCGGGTTCGATGGTGAGCAA

  AGGCGAGGAGCTGTTCACCGGGGTGGTGCCCA 3’
   |||||||||||||||||||||||||||||||
  GGGCGAGGAGCTGTTCACCGGGGTGGTGCCCA 3’
```

HSP90AA1

```
5’AGCACATGGACAGCGGAGGTGGAGGTTCGAGCGGTCCCTCGGGTTCGTCGAGCCTGGAAGTTCTGTTCCAGGGGCCCCTGTCGTCGA
  |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5’AGCACATGGACAGCGGAGGTGGAGGTTCGAGCGGTCCCTCGGGTTCGTCGAGCCTGGAAGTTCTGTTCCAGGGGCCCCTGTCGTCGA

  GCGGTCCCTCGGGTTCGATGGTGAGCAAAGGCGGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAA
  |||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  GCGGTCCCTCGGGTTCGATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAA

  ACGGCC 3’
  ||||||
  ACGGCC 3’
```

Figure 2. Alignment of sequencing results with template sequence. Example for sequencing results obtained from genomic PCR for *METTL3-eGFP* and *HSP90AA1-eGFP*. The top sequence is taken from the sequencing output while the bottom strand represents the template sequence at the junction between the last exon (yellow) of the respective gene and the start of eGFP (green). A mismatch (red) was found within eGFP which does not affect the function.

EGFP-tagged proteins could not be detected by immunofluorescence or western blots

In addition to verifying the integration of transgenic constructs on sequence level, we qualitatively assessed eGFP expression by fluorescence microscopy. Among the transgenic cells tested, only cells infected with the lentiviral construct carrying the non-fused eGFP tag gave green fluorescence signal (data not shown). In the BAC-transformed cells, no fluorescent signal could be detected.

Next, we performed western blot analysis to verify the expression of eGFP-tagged proteins. Results are presented in Figure 3. No signal could be detected when probing against eGFP (Fig. 3A and B, left panel). The β-actin control showed a strong band at 45 kDa indicating that cells were properly lysed and there was no overall protein degradation (Fig. 3B, right panel). In a control experiment, we tested whole cell lysates from other BAC transformed HeLa-K cell lines for eGFP expression. Again, no specific signal could be detected (data not shown) but a β-actin signal was seen in the control staining (Fig. 3C). In a subsequent anti-p16 staining, we detected bands from the endogenous protein at ~16 kDa with residual β-actin signal still visible at ~45 kDa (Fig. 3D) The signal for p16-eGFP was expected at ~42 kDa in the p16 lane (asterisk, Fig. 3D) however, no additional band was observed. The aliquot of anti-GFP antibody was used on a regular basis by other lab members with no impairment in signal quality (data not shown).

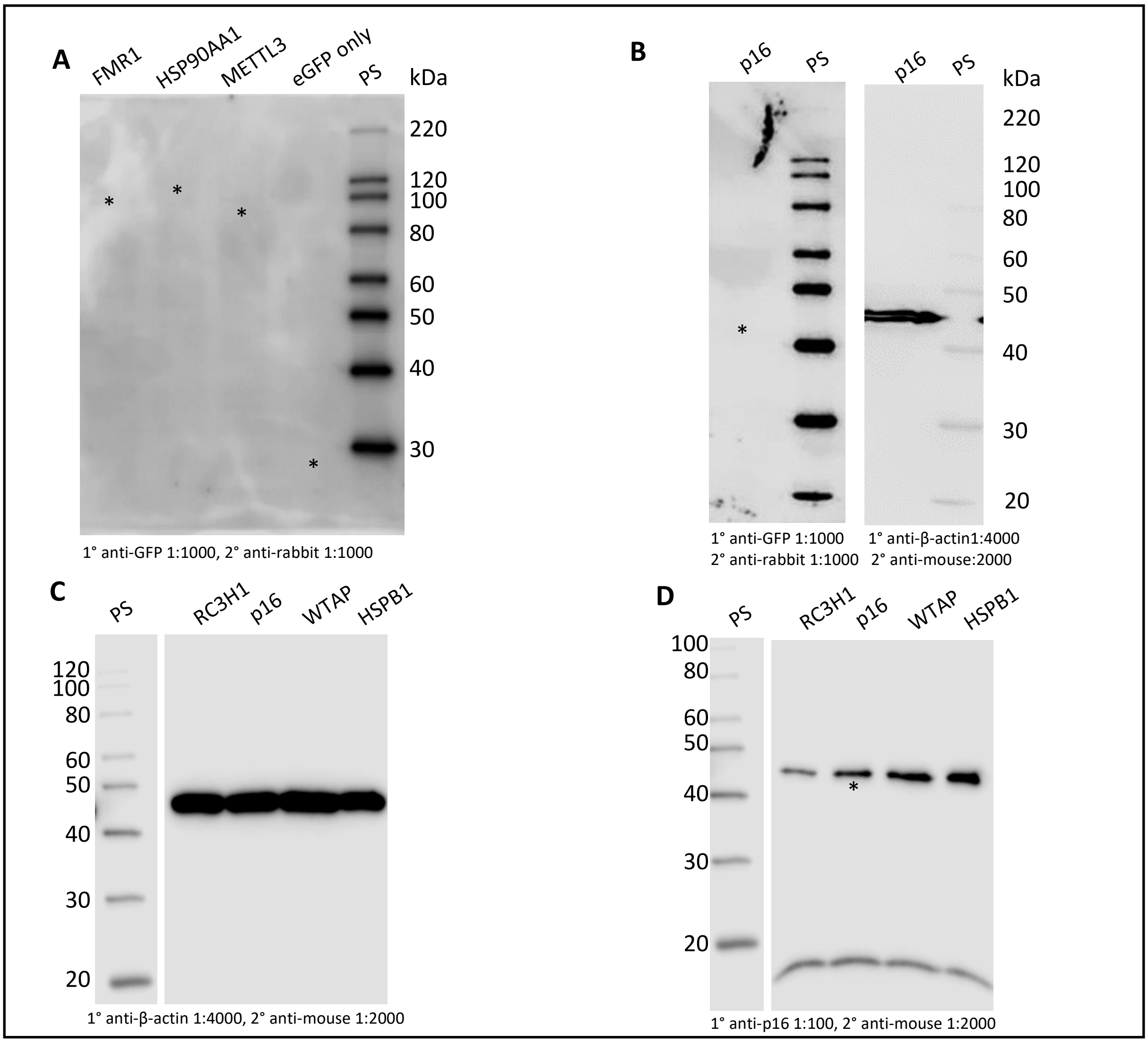


Figure 3. Western blot results for eGFP-tagged proteins. (**A** and **B**) No specific signal could be detected with an anti-eGFP antibody for any of the tested cell lysates. Asterisks indicate where protein signal was expected based on the protein standard. (**B** and **C**) The control stain yielded strong β-actin signals. (**D**) The following anti-p16 staining did not yield any band for p16-eGFP at ~42kDa (asterisk). Endogenous p16 (~16 kDa) and residual β-actin signal from the previous stain is visible. PS = MagicMark protein standard

4.2 Incomplete collection of expression vectors for GFP-tagged RBPs

In order to establish transgenic cell lines expressing eGFP tagged versions of putative PUMILIO interactors, we tried to insert the respective coding sequences into the expression vector pEZY-hPGK-eGFP via Gateway cloning. We successfully obtained PCR amplicons for all inserts of interest, namely *DND1, RBFOX2, TRIM71, HNRNPA1, TRIM32* and *Sam68.* Examples for successfully amplified *DND1*, *TRIM71* and *Sam68* are shown in Figure 4. Entry- and expression plasmids were successfully cloned for *TRIM32* and *HNRNPA1*. Figure 5 illustrates the alignment of the pEZY-hPGK-eGFP-TRIM32 sequencing result to the predicted plasmid sequence. We did not obtain entry clones for *DND1, TRIM71* and *Sam68* since the amplicons failed to ligate with pENTR. An entry clone was obtained for *RBFOX2* however the insert was found to be in the wrong orientation.

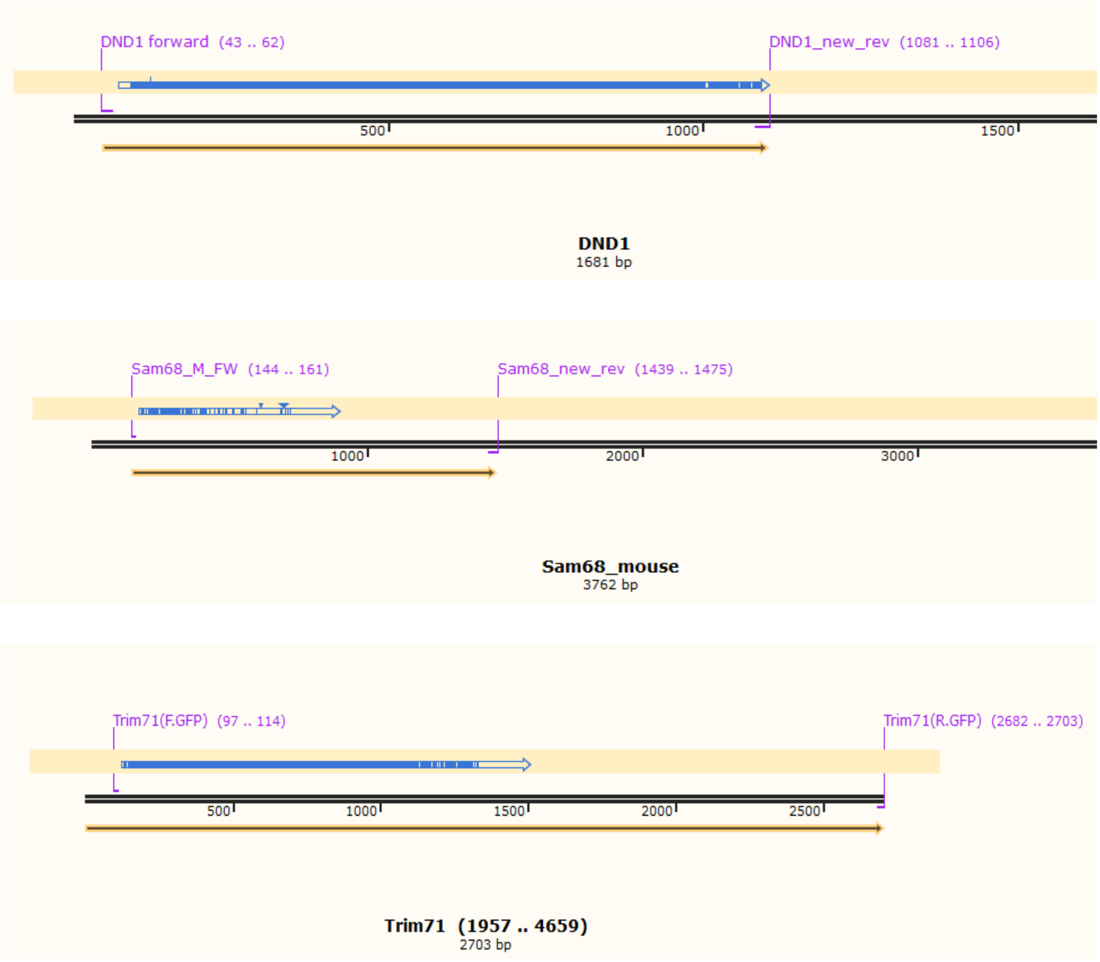
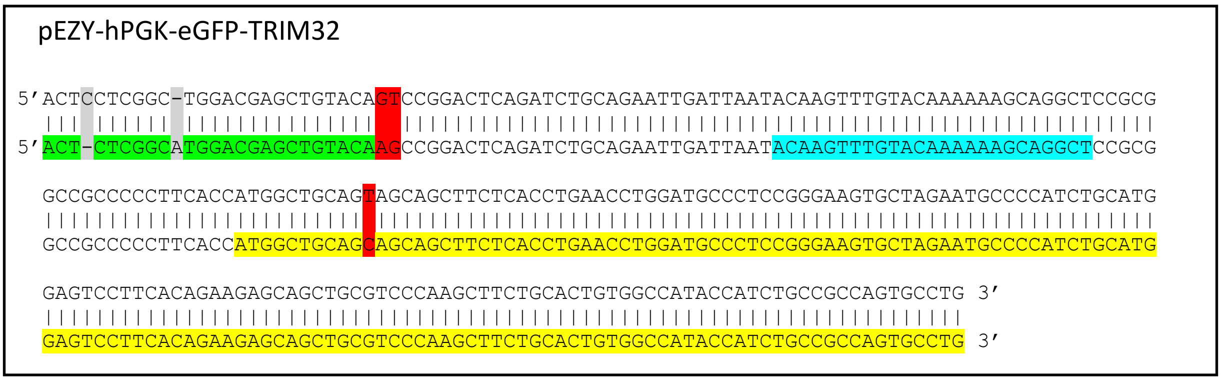


Figure 4. Sequencing results from PCR products of DND1, Sam68 and TRIM71 aligned to the gene sequence. The coding sequence is indicated by the orange arrow. Sequence alignment is represented by the blue arrow. Matches are indicated by the blue color.



pEZY-hPGK-eGFP-TRIM32

```
5’ACTCCTCGGC-TGGACGAGCTGTACAGTCCGGACTCAGATCTGCAGAATTGATTAATACAAGTTTGTACAAAAAAGCAGGCTCCGCG
  ||| |||||| |||||||||||||||  |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5’ACT-CTCGGCATGGACGAGCTGTACAAGCCGGACTCAGATCTGCAGAATTGATTAATACAAGTTTGTACAAAAAAGCAGGCTCCGCG

  GCCGCCCCCTTCACCATGGCTGCAGTAGCAGCTTCTCACCTGAACCTGGATGCCCTCCGGGAAGTGCTAGAATGCCCCATCTGCATG
  ||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  GCCGCCCCCTTCACCATGGCTGCAGCAGCAGCTTCTCACCTGAACCTGGATGCCCTCCGGGAAGTGCTAGAATGCCCCATCTGCATG

  GAGTCCTTCACAGAAGAGCAGCTGCGTCCCAAGCTTCTGCACTGTGGCCATACCATCTGCCGCCAGTGCCTG 3’
  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  GAGTCCTTCACAGAAGAGCAGCTGCGTCCCAAGCTTCTGCACTGTGGCCATACCATCTGCCGCCAGTGCCTG 3’
```

Figure 5. Partial sequence alignment of pEZY-hPGK-TRIM32. Top sequence was retrieved from the sequencing result and aligned to the predicted plasmid sequence. This section covers the end of the eGFP coding sequence (green), the Gateway® *att*B1 recombination site (blue) and the start of the *TRIM32* coding sequence (yellow). Mismatches are marked red and insertions/deletions are marked grey.

4.3 Conclusion and experimental improvements

The interactomes of putative PUMILIO interacting proteins were attempted to be investigated through immunoprecipitation assays and subsequent LC-MS analysis of eGFP tagged proteins. BAC-transformed transgenic cell lines were used to map out the experimental conditions, however, neither fluorescence microscopy nor western blot using GPF antibody could confirm the expression of eGFP-tagged protein. A control experiment showed the endogenous protein p16 at ~16 kDa as expected, but not the eGFP-tagged version at approximately ~42 kDa. The bands at ~45kDa represent residual β-actin signal from a previous stain that despite the applied stripping procedure could not be removed. The β-actin signal could have masked the expected p16-eGFP signal at ~42 kDa. However, this should have led to an overall stronger band in the p16 lane compared to the control cell lines. In sum, the results suggest that the eGFP-tagged proteins are not expressed. A potential reason for this outcome could be the epigenetic silencing of the BAC-sequence. Conclusively, immunoprecipitation assays could not be performed.

In order to investigate PUMILIO interaction in different cell lines, expression constructs for multiple eGFP-tagged putative PUMILIO interactors were designed including TRIM32, TRIM71, DND1, RBFOX2 and Sam68. Additionally, an expression construct for the non-specific RNA binding protein HNRNP-A1 was designed to serve as a positive control for RNA binding. These constructs will be used to establish transgenic cell lines expressing the tagged proteins of interest at physiological levels and subsequently investigate their interactome. However, the experimental conditions for molecular cloning of these constructs need to be optimized for every insert of interest. Furthermore, a different system for molecular cloning will be utilized to overcome ligation problems.

4.4 Explorative data analysis of HNSCC cell line RNA sequencing data

With the collection of cell lines used in this study we aimed to represent a broad spectrum of head and neck cancer phenotypes and risk factors. Explorative data analysis was performed to identify relevant mediators of the transcriptional differences in this dataset. The available information from the cell lines was categorized as outlined in Table 1. These categories were then tested for their influence on gene expression in the HNSCC cell line dataset. The full table of clinical characteristics is given in Supplementary table S2.

Table 1. Characterization of patients from whom the different HNSCC cell lines have been established.

cell line	age (years)	category	mutations	tissue	smoker	HPV status
OKF6[53]	57	control	none	floor of mouth	no	negative
SCC2[54]	58	cancer	PTEN	hypopharynx	yes	positive
SCC90[55]	46	cancer	PTEN, PIK3CA	tongue	yes	positive
HSC3[56]	63	cancer	p53	tongue	no	negative
FaDu[57]	56	cancer	p53, p16	pharynx	no	negative
Cal33[58]	69	cancer	p53, PIK3CA	tongue	no	negative
SCC4[59]	55	cancer	p53	tongue	no	negative

For each cell line, two biological replicates were sequenced in two independent runs and included in the dataset. On average, 99.78 % of the reads per sample were recovered after rRNA and tRNA depletion and 44,471,238 sequencing reads per sample mapped to the human genome (hg38). Aligned reads were annotated to a total of 28,395 coding and non-coding genes by featureCounts of which 28,010 could be matched to gene symbols and were considered for further analysis.

Gene expression profiles are almost identical across biological replicates

Principal component analysis (PCA) was used to visualize sample distances. This statistical tool simplifies and structures large datasets by introducing two new variables (principal components). The principal components are represented through two orthogonal axes which are fitted to the data in a way that they reduce the variance of the data. In the principal component plot, these two axes are the x- and y- axis. The variance that principal components can explain tends to be lower with increasingly complex datasets.

In the context of RNA sequencing data, data points in close proximity to each other indicate a high similarity of gene expression profiles. Figure 6 shows that in our analysis, the first two principal components accounted for 54% of the total variance (PC1: 32%, PC2: 22%). Both sequencing runs overlapped very closely in the principal component plot which indicates that their expression profiles are only marginally different.

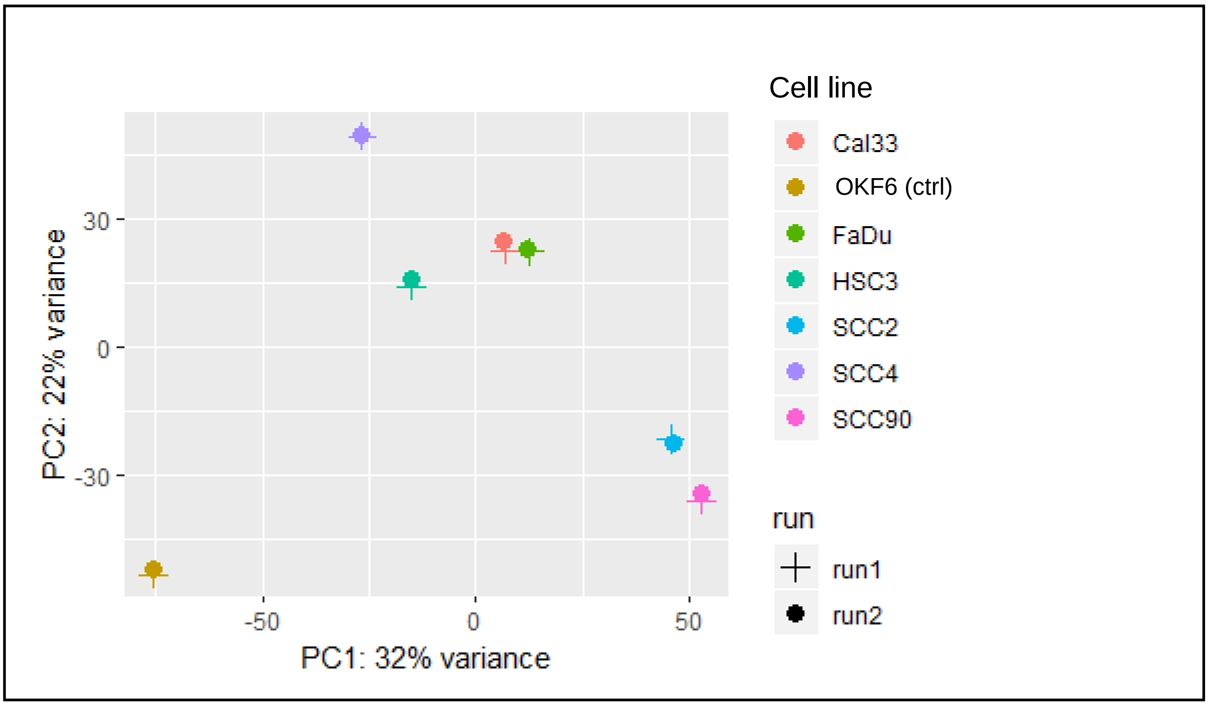


Figure 6. PCA plot for two biological replicates of HNSCC sequencing data. Different cell lines are represented by color while independent sequencing runs are indicated by shape. The closer both shapes are for a specific color, the closer is the overlap between two datasets

HPV status discriminates the transcriptional profile of head and neck cancer cell lines

Besides a close overlap of biological replicates, the PCA plot showed two main clusters separated from the control cell line OKF6. In Figure 7, cell lines were grouped according to the categories listed in Table 1 to examine whether tissue of origin, HPV status or mutations would influence cluster formation. While tissue of origin and mutational background do not seem to be critical for cluster formation, HPV status was found to be decisive for most of the differences in gene expression profiles. As seen in Figure 7B, SCC4 was found to be slightly farther apart from the rest of the HPV$^-$ cluster but was not excluded from the analysis. Based on these findings, the following analyses focused on exploring transcriptional differences between HPV positive and HPV negative groups.

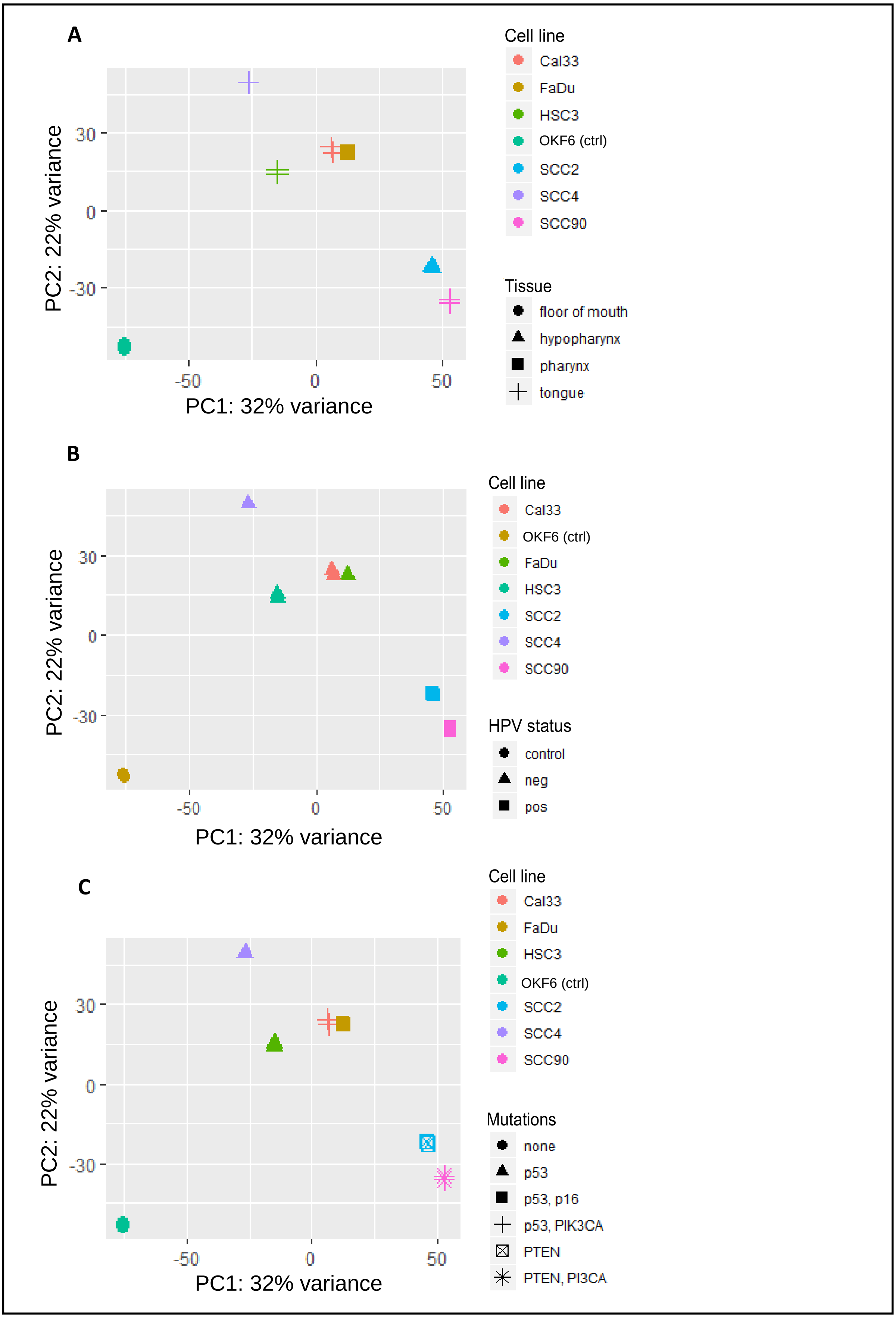


Figure 7.PCA plots of potential influence factors on gene expression. (**A**) Tissue type influences clustering to a certain extent with one outlier in the tongue group. All other groups only consist of one cell line. (**B**) HPV status predicts cluster formation in cancer cell lines. (**C**) PTEN and p53 mutations appear to influence cluster formation whereas PIK3A mutations does not influence clustering.

4.5 Differentially expressed transcripts based on HPV status in HNSCC cell lines

Based on the findings from principal component analysis, we assessed differential gene expression according to HPV status with two cell lines in the HPV$^+$ group and four cell lines in the HPV$^-$ group. We approached the transcriptional differences with two different comparisons.

The first differential expression analysis compared gene expression in HPV$^+$ and HPV$^-$ cell lines to the control cell line. The number of differentially expressed (DE) genes categorized into coding and non-coding transcripts are summarized in Table 2.

Table 2. DE transcripts in HPV$^+$/$^-$ cell lines versus the control cell line

	DE *total*	Upregulated	Downregulated
HPV$^+$/control			
Coding	3934	2156	1778
Non-coding	768	364	404
total	*4702*	*2520*	*2182*
HPV$^-$/control			
Coding	3657	2348	1309
Non-coding	729	447	282
total	*4386*	*2795*	*1591*

Compared to the control, we found a comparable number of genes to be DE in both HPV$^+$ and HPV$^-$ cell lines. In order to identify the transcripts that were found in both groups, we combined the results from the HPV$^+$/control and HPV$^-$/control comparisons. Thereafter, we filtered for genes that were DE in both the HPV$^+$/control and the HPV$^-$/control comparison. Based on the differences between HPV$^+$ and HPV$^-$ cell lines that we observed in the principal component analysis, we were particularly interested in inversely expressed transcripts since they may contribute to the transcriptional differences between the two groups. However, the majority of both coding and non-coding genes was found to be up- or downregulated synchronously instead of following an inverse trend (Table 3).

Table 3. Common DE genes in HPV$^+$ and HPV$^-$ cell lines versus the control cell line

	Trend				***total***
HPV$^+$/control	**Up**	**Down**	**Up**	**Down**	
HPV$^-$/control	**Up**	**Down**	**Down**	**Up**	
Coding	1391	849	18	40	*2298*
Non-coding	267	201	1	4	*473*
total	*1658*	*1050*	*19*	*44*	*2771*

Since a notable number of genes following the same trend in both HPV$^+$ and HPV$^-$ cells when compared to the control, we performed another analysis to uncover transcriptional differences between the two groups irrespective of the control. We therefore compared gene expression of HPV$^+$ versus HPV$^-$ cell lines using the HPV$^-$ group as the baseline. The results are summarized in Table 4. All in all, we found a higher number of both coding and non-coding genes to be downregulated than upregulated.

Table 4. DE transcripts in HPV$^+$ versus HPV$^-$ cell lines

	DE *total*	**Upregulated***	**Downregulated***
HPV$^+$ vs HPV$^-$			
Coding	2474	875	1599
Non-coding	421	102	319
total	*2895*	*977*	*1918*

* Up/Downregulated in HPV$^+$ compared to HPV$^-$

Comparing the two approaches of identifying DE genes between HPV$^+$ and HPV$^-$ cancer cell lines, we found that the two methods identify a largely different set of genes. Of 473 DE non-coding genes identified through the HPV$^+$/control and HPV$^-$/control comparisons, we found 49 in the 421 DE non-coding genes from the HPV$^+$ vs HPV$^-$ comparison. Among these, we identified the five transcripts that were inversely regulated in the HPV$^+$/control and HPV$^-$/control comparison (see Table 3).

4.6 Head and neck cancer data from The Cancer Genome Atlas

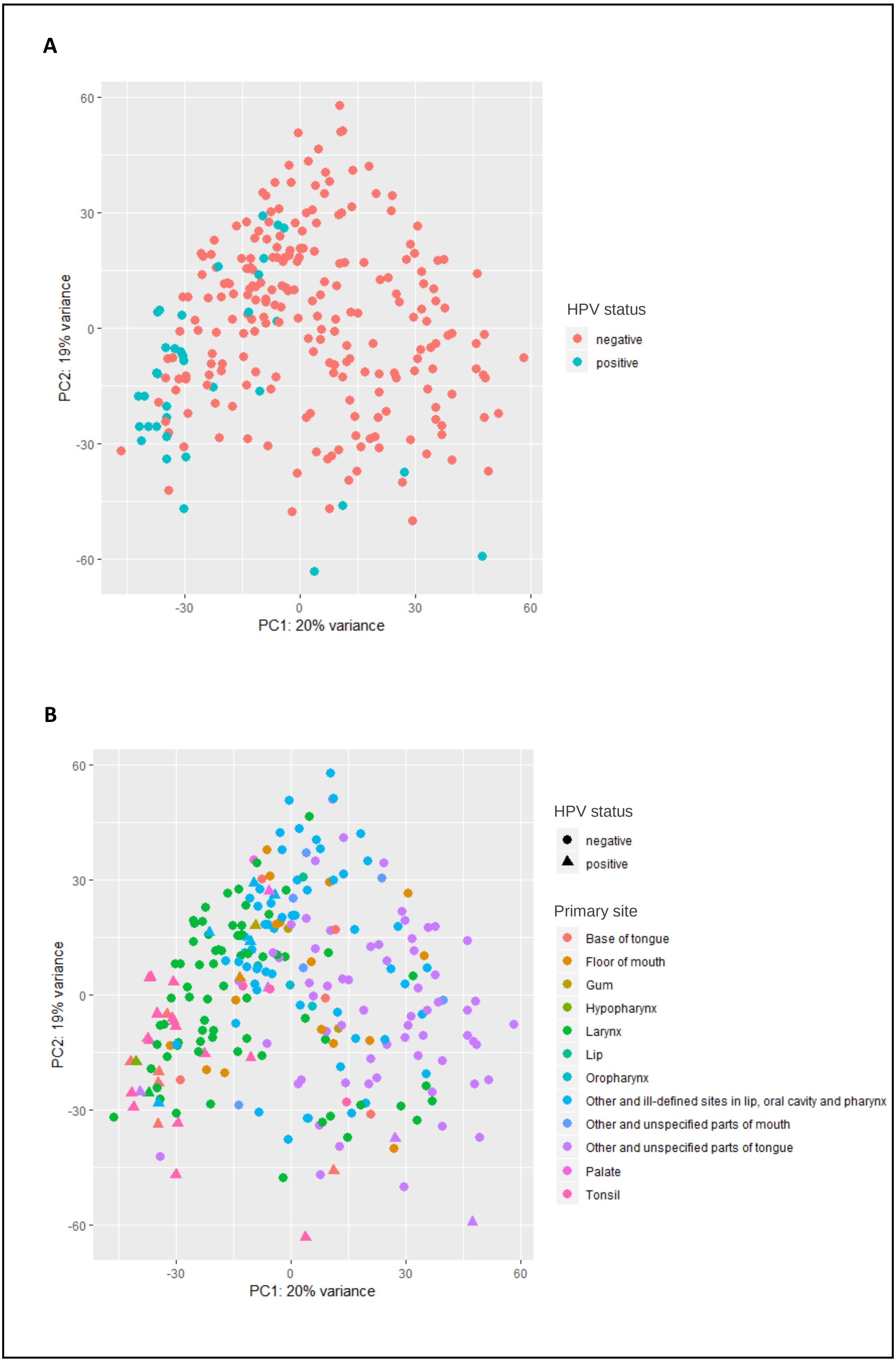
RNA sequencing data as well as clinical data was obtained for 263 head and neck cancer patients from the TCGA Research Network. One sample was removed from the cohort due to missing information on HPV status. Relevant clinical information on the remaining 262 individuals is presented in Table 5. We performed a Chi^2 test for independence to see which features show a significant interaction with HPV status. Significant associations are indicated by asterisks in Table 5.

Table 5. Clinical data on patient samples. NA= no data available

Feature	**Number of individuals**		
	total	HPV$^+$	HPV$^-$
Total cohort	***262***	*36*	*226*
Gender*			
male	*194*	32	162
female	*68*	4	64
Age at diagnosis (years)			
Mean	*61.55*	61.62	61.50
Median	*61.95*	62.00	61.90
Tumor stage			
I	*13*	2	11
II	*43*	5	38
III	*34*	4	30
IVa	*128*	11	117
IVb	*4*	0	4
NA	*40*	14	26
Tissue of origin*			
Base of tongue	*11*	6	5
Floor of mouth	*17*	1	16
Gum	*3*	1	2
Hypopharynx	*1*	0	1
Larynx	*72*	1	71
Lip	*1*	0	1
Oropharynx	*2*	0	2
Other and ill-defined sites in lip, oral cavity and pharynx	*61*	5	56
Other and unspecified parts of mouth	*6*	0	6
Other and unspecified parts of tongue	*67*	4	63
Palate	*2*	1	1
Tonsil	*19*	16	3
Alcohol history*			
Yes	*179*	31	148
No	*77*	4	73
NA	*6*	1	5

*Significant differences between HPV$^+$ and HPV$^-$ (*p <0.05* Chi^2 test)

The PCA of the TCGA dataset resulted in a rather scattered PCA plot indicating a high dimensionality of the data (Fig. 8). In total, 39% of the variance in the TCGA dataset was explained by the two first principal components (PC1: 20%, PC2: 19%). Figure 8A shows that HPV positive tumors group together to a certain extent even though cluster formation is not as distinct. In the PCA plot displaying primary tumor site and HPV status displayed in Figure 8B, data points of a respective primary site group together, however, the clusters overlap. Gender, alcohol history and tumor stage do not show a distinctive pattern on the PCA plot (Fig. 8C-E), indicating that these factors are less important for the differences in gene expression.


A
PC2: 19% variance
PC1: 20% variance
60
30
0
-30
-60
-30
0
30
60
HPV status
negative
positive
B
PC2: 19% variance
PC1: 20% variance
60
30
0
-30
-60
-30
0
30
60
HPV status
negative
positive
Primary site
Base of tongue
Floor of mouth
Gum
Hypopharynx
Larynx
Lip
Oropharynx
Other and ill-defined sites in lip, oral cavity and pharynx
Other and unspecified parts of mouth
Other and unspecified parts of tongue
Palate
Tonsil

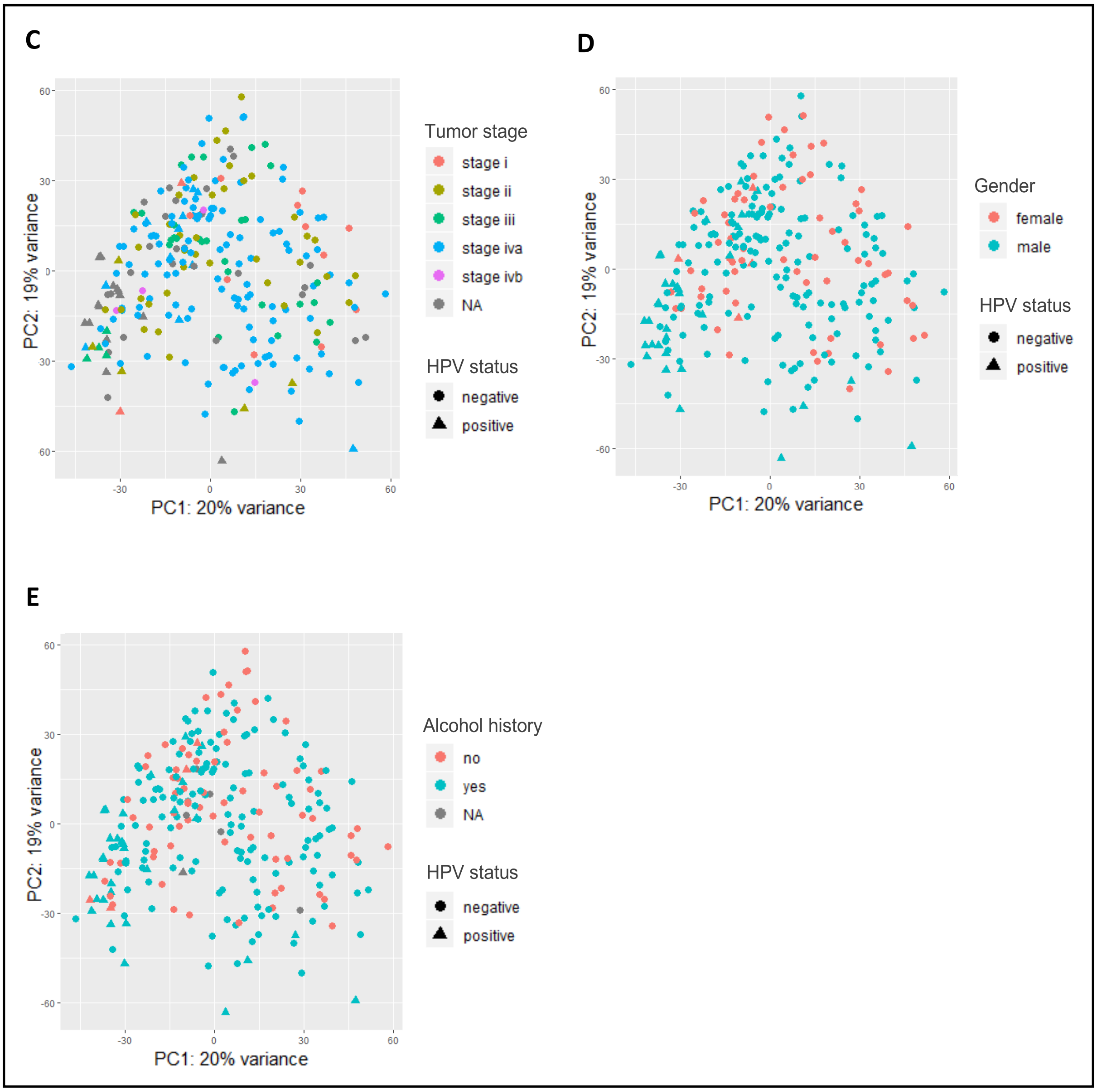


Figure 8. PCA plots from the TCGA data set. (**A**) HPV positive tumors are loosely clustered together. Four outliers are located to the bottom right which potentially could be a separate cluster. (**B**) Tumors from the same site largely group together. (**C-E**) Tumor stage, gender and alcohol history do not seem to influence transcriptional differences in a systematic way that results in distinct clusters.

Since tumor samples from the same primary site appeared group together to a certain extent and showed a significant interaction with HPV status in the Chi^2 test, we included it as a co-variate in our subsequent analysis. Apart from primary site, we found significant interactions of HPV status with gender and alcohol history in the Chi^2 test. However, these factors did not seem to be responsible for the formation of clusters, indicating no significant influence on differential expression. As the tumor dataset was used for validation of the findings from the cell line dataset, we tried to reduce potential bias through confounding variables. Since the cell lines used in this study were exclusively derived from male patients, we excluded female samples from the analysis.

Differential expression analysis was carried out with RNA sequencing data of *n=194* male cancer patients of which *n=32* were HPV positive (see Table 5). A total number of 466 genes (13 non-coding) were found to be differentially regulated, of which 228 were upregulated (8 non-coding) and 238 downregulated (5 non-coding) in the HPV$^+$/HPV$^-$ comparison.

4.7 Common trends of differential expression in cell line data and patient data

Next, we compared the differential expression data for HPV positive and HPV negative samples from the cell line dataset and the TCGA dataset. For that, we merged the differential expression analysis results from both datasets and filtered for common DE genes. Due to slightly incompatible gene annotation, the total number of genes that could be analyzed was reduced. As illustrated in Figure 9, a proportion of transcripts was not annotated in either the cell line or the TCGA dataset and could therefore not be considered for validation. Especially for non-coding genes, a significantly lower number of transcripts was found in the TCGA dataset, limiting the number of non-coding genes to 1,282.

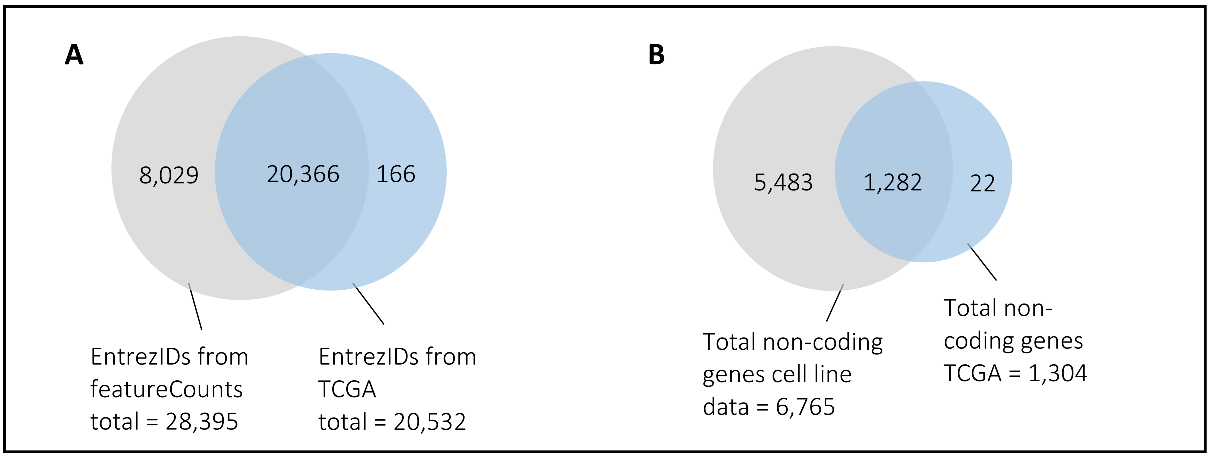

A
8,029
20,366
166
EntrezIDs from featureCounts total = 28,395
EntrezIDs from TCGA total = 20,532
B
5,483
1,282
22
Total non-coding genes cell line data = 6,765
Total non-coding genes TCGA = 1,304

Figure 9. Overlap of annotated genes between cell line and TCGA data set. (**A**) The overlap between the total number of EntrezIDs contained in the featureCounts output (red) and the annotated raw counts from TCGA (blue) shows the total number of genes considered for further analysis. (**B**). The number of non-coding genes contained in the TCGA data was significantly lower compared to cell line data. In total, 1,282 non-coding genes were considered for analysis.

In total, we found 191 DE genes of which 10 were non-coding. Of these, 167 followed the same trend with 80 genes being commonly upregulated (six non-coding) and 87 (four non-coding) commonly downregulated in both datasets (Supplementary table S2). The remaining 24 genes showed an inversed directionality, yet no non-coding gene fell into this category

4.8 GO term analysis highlights molecular function of differentially expressed genes

In order to investigate the protein function of the 164 DE genes, we performed GO Molecular Function (MF) overrepresentation analysis. Results are summarized in Figure 10. In total, 17 GO MF terms were found to be overrepresented in the gene set at a significance level of *p<0.05.* However, several overrepresented terms are redundant. Serine-related enzyme activity, oxidoreductase activity and receptor ligand binding were found to be highly associated. In sum, 41 genes could be linked to GO-MF terms in the overrepresentation analysis. Detailed information on p-values and category sizes is displayed in Supplementary table S3.

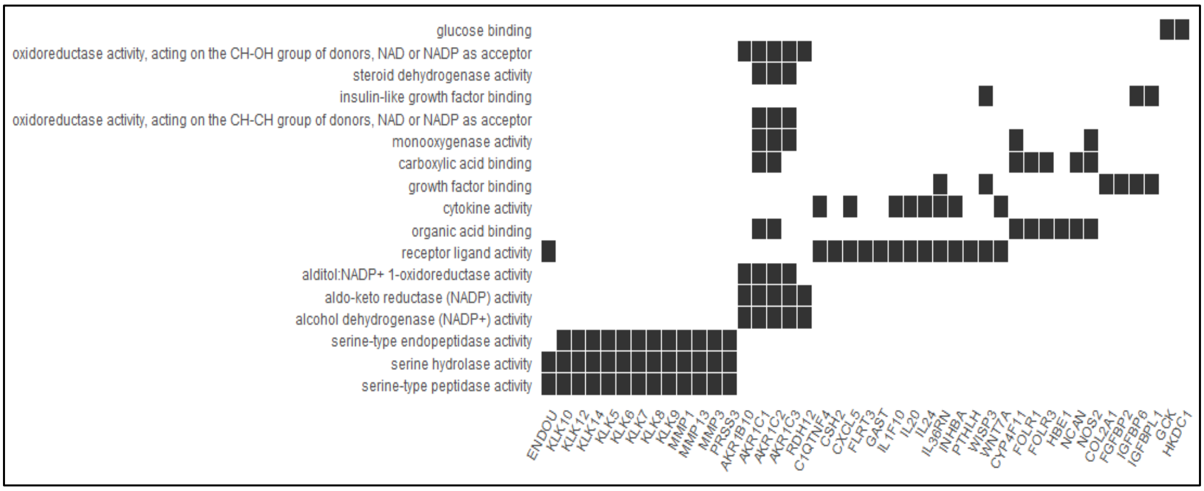


Figure 10. GO Molecular Function terms overrepresented in DE genes. The black boxes indicate significant GO term association of a respective gene. In total, 43 genes were assigned to 17 GO MF terms (adjusted *p<0.05*). Redundancies were found especially for serine-related enzyme activity and oxidoreductase activity.

4.9 Differences in gene expression confirmed by qPCR

In total, 76 genes (seven non-coding) sustained the pre-selection process for qPCR validation (see Table 6). 11 genes were selected for qPCR validation of which were five coding and six non-coding genes. In the non-coding group, three pseudogenes, two antisense-RNAs and one miRNA host gene were selected. Primers designed for ADORA antisense RNA 1 (*ADORA2A-AS1)* and ATP binding cassette subfamily A member 17, pseudogene (*ABCA17P)* did not yield reliable melting curves in a test run and were excluded from the validation. Overall low Ct values were observed in OKF6 with no signal for Wnt family member 7A (WNT7A) and CDKN2A.

Table 6. Pre-selection of DE genes for qPCR based on RPKM and CPM values. Genes selected for qPCR are marked grey.

RefSeq ID	gene symbol	cell line data				TCGA data	
		log2-Fold change	adjusted p-value	median RPKM*	median CPM#	log2-Fold change	adjusted p-value
NR_027082	SFTA1P	-11,831	1,23E-10	2,432	1,764	-2,6564	3,59E-05
NM_001281431	KLK8	-10,987	1,67E-31	10,683	10,159	-2,7435	7,19E-07
NM_002422	MMP3	-10,807	1,70E-06	0,160	0,309	-2,5290	2,10E-05
NM_001303419	TRIML2	-10,339	3,02E-05	0,136	0,319	-2,0365	0,03780108
NM_001077491	KLK5	-10,007	1,36E-29	49,499	79,164	-4,9964	3,90E-15
NM_032654	AFAP1-AS1	-9,416	3,09E-07	0,318	2,177	-4,8253	6,49E-11
NM_001207053	KLK7	-9,234	3,54E-12	0,547	1,051	-2,6319	9,97E-06
NM_000329	RPE65	-9,171	2,60E-05	2,209	6,045	-3,2283	8,09E-05
NM_012315	KLK9	-8,816	9,91E-19	5,484	7,338	-3,0330	3,04E-07
NM_001077500	KLK10	-7,869	1,61E-24	17,520	62,407	-2,3104	3,28E-05
NM_001785	CDA	-7,315	5,19E-11	3,338	3,114	-3,2416	7,07E-10
NM_001137556	FAM25BP	-7,121	0,00015458	0,716	0,251	-3,7396	3,25E-08
NM_002178	IGFBP6	-6,848	3,88E-20	165,2	166,9	-2,4643	8,77E-06
NM_001135639	CNGB1	-6,622	8,56E-06	0,114	0,827	-2,9295	4,77E-08
NM_002192	INHBA	-6,576	8,48E-06	1,442	3,082	-2,2473	2,06E-06
NM_000872	HTR7	-6,457	4,85E-08	0,197	0,697	-2,1405	4,91E-06
NM_001145938	MMP1	-6,340	1,68E-14	29,638	61,315	-2,9715	1,03E-08
NM_001126063	KHDC1L	-6,299	1,18E-06	0,833	0,523	-2,2369	0,00209613
NM_002427	MMP13	-6,004	1,71E-13	9,014	22,667	-3,5838	3,25E-06
NM_001012964	KLK6	-5,919	1,57E-07	10,692	17,330	-2,7918	2,56E-05
NM_012275	IL36RN	-5,642	2,56E-07	0,857	2,419	-2,1327	0,00056742
NM_001253908	AKR1C3	-5,088	5,02E-17	12,587	26,931	-2,2440	0,00044961
NM_004625	WNT7A	-5,049	1,06E-05	2,870	6,764	-2,1132	0,00019856
NM_001128932	CYP4F11	-4,955	1,14E-65	6,916	21,864	-2,0050	0,00435798
NM_020299	AKR1B10	-4,947	0,00017585	1,889	4,753	-2,5132	3,08E-05
NM_001103160	SH2D5	-4,904	0,00312598	0,582	2,992	-2,6371	7,41E-08
NM_002820	PTHLH	-4,735	0,00017406	9,299	28,702	-2,4837	1,09E-07
NM_001201325	PDZK1	-4,234	0,0014474	0,167	0,492	-2,3046	3,40E-05
NM_152611	LRRN4	-4,180	0,00047206	0,085	0,337	-2,5872	1,01E-05
NM_000526	KRT14	-4,126	2,60E-05	223,2	397,5	-2,2713	2,89E-05
NM_001287758	COL4A6	-3,717	6,45E-05	1,482	10,951	-2,4835	1,16E-07
NM_001135241	AKR1C2	-3,563	1,59E-12	15,864	70,785	-2,4388	0,00013641
NM_001145106	FIBCD1	-3,124	1,68E-05	0,821	2,862	-2,3417	0,00096429
NM_001300845	SLC35F3	-2,742	0,01075204	0,357	1,324	-3,1091	2,66E-10
NM_152443	RDH12	-2,693	0,00339528	0,176	0,349	-4,0139	8,53E-12
NM_001311182	KLK14	-2,680	0,02752969	0,360	0,435	-3,3174	5,91E-07
NM_001165960	ALOXE3	-2,624	0,03273011	0,157	0,569	-2,1148	4,27E-05
NM_001353	AKR1C1	-2,574	0,0107339	26,131	38,086	-2,4852	0,00017684
NM_001015886	HMGA2	-2,557	0,02339872	8,752	46,724	-2,7621	2,18E-08
NM_000805	GAST	-2,281	0,04106282	0,666	0,322	-2,8100	2,96E-05
NM_001102658	CT62	-2,258	0,02390281	0,207	0,430	-2,0623	0,00982914
NM_001031692	LRRC17	-2,208	0,03353162	0,109	0,298	-2,4942	1,14E-05
NM_004948	DSC1	-2,152	0,02980881	0,108	0,478	-2,9881	7,14E-07
NR_024391	MIR924HG	2,020	4,12E-07	1,133	2,443	2,2747	2,30E-06
NM_001105578	SYCE2	2,165	1,95E-13	1,003	1,049	3,1877	2,50E-26
NM_001039780	CCNI2	2,350	0,00068796	0,079	0,394	2,2961	3,96E-06
NM_020309	SLC17A7	2,380	3,40E-09	0,128	0,393	2,3086	1,49E-06
NM_004209	SYNGR3	2,748	3,77E-05	2,452	4,466	2,9903	9,98E-15
NM_000077	CDKN2A	2,915	0,00045925	12,410	61,780	2,3529	0,00035551
NM_001039784	ADORA2A-AS1	3,375	0,03435945	0,096	0,350	2,3318	1,25E-20
NM_001322799	KCNS1	3,391	0,04082416	1,300	5,387	3,5187	1,78E-06
NM_013356	SLC16A8	3,534	0,00177614	1,017	2,053	2,0722	8,74E-06
NM_001099652	GPR137C	3,570	1,15E-12	0,221	0,860	2,1267	3,87E-15
NM_001308165	SOX30	3,823	3,68E-05	0,177	1,084	3,6754	5,81E-14
NM_015063	SLC8A2	3,859	0,0002245	0,071	0,304	2,3805	0,00042382
NM_001345843	C19orf57	3,881	0,00035134	0,898	3,258	2,2735	1,04E-15

Table 6 cont.

NM_001145451	ARHGEF33	3,908	2,55E-05	0,110	0,450	3,6982	9,01E-25
NR_003574	ABCA17P	4,418	0,00075759	0,190	0,792	4,6710	4,49E-21
NM_001080448	EPHA6	4,511	0,00093939	0,088	0,696	2,8119	0,00126095
NM_014258	SYCP2	4,865	1,42E-12	0,444	2,562	5,1219	9,19E-37
NR_015411	MIR9-3HG	4,874	6,57E-05	0,637	2,818	3,3918	6,23E-14
NM_000835	GRIN2C	4,937	7,13E-05	0,186	1,272	3,1842	6,81E-31
NM_001308245	BTNL9	5,185	0,00013099	0,392	2,103	4,2534	6,34E-16
NM_001039548	KLHL35	5,526	1,10E-05	1,038	2,447	3,3441	4,82E-20
NM_001168465	MAP7D2	5,676	4,18E-19	0,146	0,685	2,8366	4,06E-05
NM_015982	YBX2	5,802	1,93E-07	0,489	0,757	3,2543	5,04E-11
NM_001100411	FAM184A	5,831	1,17E-09	0,107	0,467	2,0666	7,00E-06
NM_001089	ABCA3	6,173	5,55E-14	0,261	1,813	2,3092	4,56E-06
NM_001321525	GPAT2	6,186	2,08E-13	0,099	0,336	2,8050	1,54E-09
NM_016170	TLX2	6,197	3,92E-27	0,135	0,266	3,0557	1,59E-06
NR_002947	TCAM1P	6,745	2,09E-06	0,117	0,382	3,4499	7,07E-42
NM_001291501	SMC1B	6,787	8,59E-36	0,122	0,538	6,8480	5,05E-45
NM_138370	PKDCC	7,646	2,79E-06	1,157	3,034	2,6371	1,06E-05
NM_001007563	IGFBPL1	8,084	9,47E-07	0,244	0,837	2,1932	0,00212843
NM_015720	PODXL2	8,811	2,02E-61	1,856	4,234	2,4675	6,19E-07
NM_022897	RANBP17	11,262	6,47E-36	0,652	4,007	2,2045	2,28E-07

*The cutoff for median RPKMs was set to RPKM >0.07.

#The cutoff for median CPMs was set to CPM >0.25.

The RNA seq data in HPV^{+} versus HPV^{-} group samples could be confirmed in the qPCR analysis. We compared the expression of selected HPV^{+}-upregulated transcripts in HPV^{+} cell lines to an HPV^{-} control cell line (FaDu) as well as the non-cancer control OKF6, which is illustrated in Figure 11. Likewise, we compared gene expression of HPV^{+}-downregulated transcripts in HPV^{-} cell lines to an HPV^{+} control cell line (SCC2) and OKF6, shown in Figure 12.

The selected HPV^{+}-upregulated transcripts showed a significantly higher relative expression in HPV^{+} cell lines compared to FaDu and OKF6 (Fig. 11A, C). Especially for CDK2N, high fold changes were found in the HPV^{+}/HPV^{-} comparison. Overall, the expression differences were lower when compared to OKF6 (Fig. 11B, D).

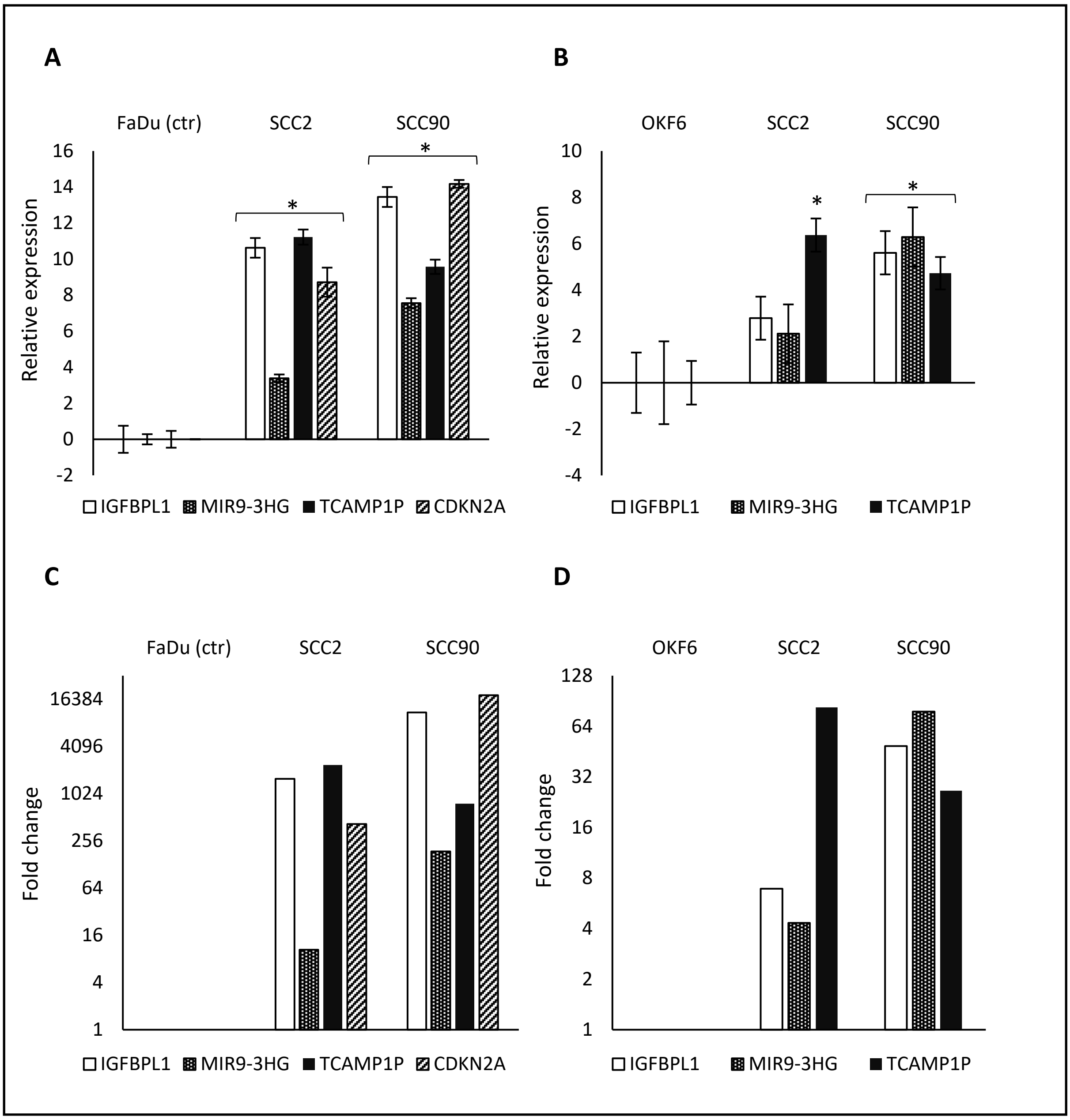
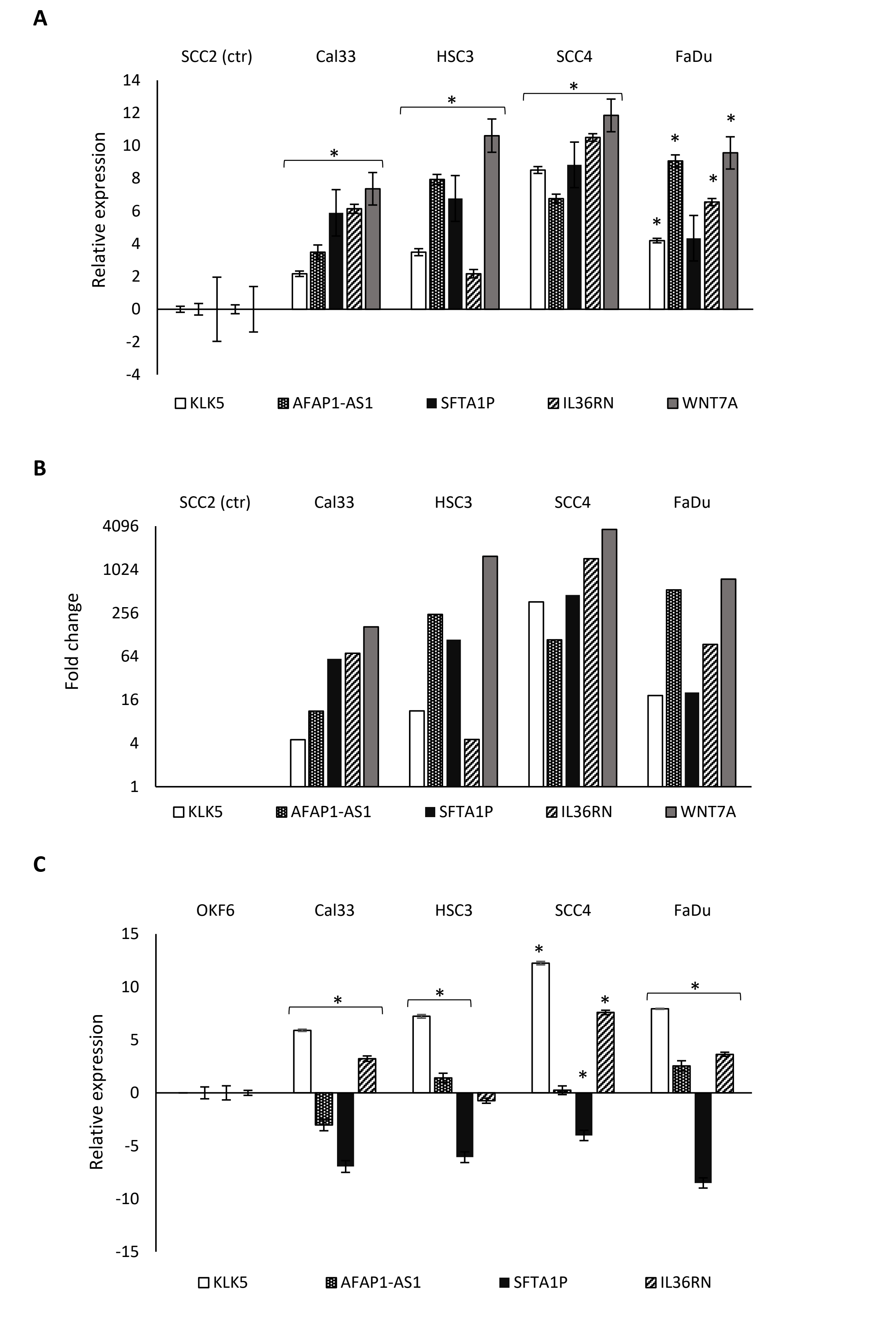


Figure 11. qPCR results for genes upregulated in HPV$^+$ group from HPV$^+$cell lines (**A**) Relative expression in HPV$^+$ cells compared to HPV$^-$ control cell line FaDu (mean ± SE, **p<0.05*). (**B**) Relative eypression in HPV$^+$ cells compared to non-cancer control cell line OKF6 (mean ± SE, **p<0.05*). (**C**) Fold change in HPV$^+$ cells compared to HPV$^-$ control cell line FaDu. (**D**) Fold change in HPV$^+$ cells compared to non-cancer control cell line OKF6.

Comparing gene expression in HPV$^-$ cell lines to the HPV$^+$ SCC2, an overall increased expression for the selected genes was observed (Fig. 12A, B). The largest differences were observed for WNT7A (Fig. 12B). When normalized to the non-cancer cell line OKF6, expression patterns diverged. Kallikrein related peptidase 5 (KLK5) and interleukin 36 receptor antagonist (IL36RN) expression levels remained significantly elevated in all cell lines when compared to the non-cancer control. Surfactant associated 1 pseudogene (SFTA1P) was consistently expressed at lower levels in all HPV$^-$ cell lines. AFAP1 antisense RNA 1 (AFAP1-AS1) expression was found to be slightly higher in FaDu and almost unchanged in HSC3 and SCC4. In Cal33, AFAP1-AS1 expression was lower compared to OKF6.


A
SCC2 (ctr)
Cal33
HSC3
SCC4
FaDu
Relative expression
14
12
10
8
6
4
2
0
-2
-4
*
□ KLK5
▨ AFAP1-AS1
■ SFTA1P
▨ IL36RN
■ WNT7A
B
SCC2 (ctr)
Cal33
HSC3
SCC4
FaDu
Fold change
4096
1024
256
64
16
4
1
□ KLK5
▨ AFAP1-AS1
■ SFTA1P
▨ IL36RN
■ WNT7A
C
OKF6
Cal33
HSC3
SCC4
FaDu
Relative expression
15
10
5
0
-5
-10
-15
*
□ KLK5
▨ AFAP1-AS1
■ SFTA1P
▨ IL36RN

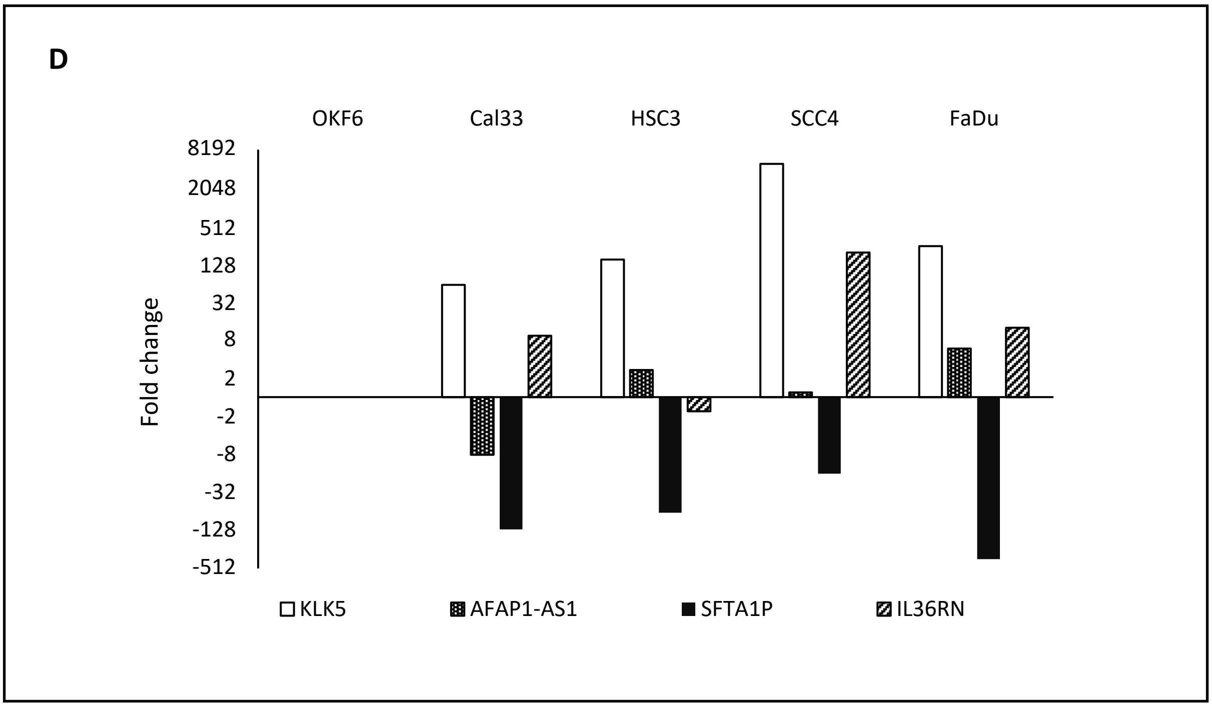


Figure 12. qPCR results for genes upregulated in HPV$^-$ group from HPV-cell lines (**A**) Relative expression in HPV- cells compared to HPV$^+$ control cell line SCC2 (mean ± SE, **p<0.05*). (**B**) Fold change in HPV$^-$ cells compared to HPV$^-$ control cell line SCC2. (**C**) Relative expression in HPV$^-$ cells compared to non-cancer control cell line OKF6 (mean ± SE, **p<0.05*). (**D**) Fold change in HPV$^-$ cells compared to non-cancer control cell line OKF6.

5. Discussion

Head and neck squamous cell carcinoma is a complex disease with a heterogeneous etiology. HPV infection has been established as the driving force of malignancies for a subset of patients, giving rise to a separate entity of the disease. Numerous studies have attempted to shed light on the transcriptional profiles of head and neck cancer, however, the field of HPV-related gene expression signatures in HNSCC is still evolving.

In this study, we highlight the differences in the coding and non-coding transcriptome of HPV$^+$ and HPV$^-$ HNSCC. We utilized RNA sequencing data from HNSCC cell lines to confirm HPV status as the main driver of transcriptional differences. Subsequently, we mapped out the DE genes in HPV$^+$ and HPV$^-$ groups and compared our findings to a validation sample of HPV$^+$ and HPV$^-$ tumors. Thereby, we identified *n=154* coding and *n=10* non-coding DE genes between HPV$^+$ and HPV$^-$ groups across both samples and validated a subset by qPCR. We categorized the coding DE genes according GO Molecular Function terms and found an enrichment of serine proteases, aldo-keto reductases and cytokines among others. Several of the identified DE non-coding transcripts have previously been linked to head and neck cancer or other cancer types.

While the coding landscape of HPV-discriminative gene expression in head and neck cancer has been studied extensively, fewer studies have attempted to identify the non-coding profiles of HPV$^+$ and HPV$^-$ head and neck cancers. In this study, we found six lncRNAs to be upregulated in the HPV$^+$ group, in total three pseudogenes (*TCAM1P*, *ABCA17P*, *UOX*), two miRNA host genes (*MIR9-3HG, MIR924HG*) and one antisense-RNA (*ADORA2A-AS1*). Two pseudogenes (*MT1L*, *SFTA1P*), one antisense RNA (*AFAP-AS1*) and one lincRNA (*PICSAR*) were downregulated in the HPV$^+$ group compared to the HPV$^-$ group.

Several of the differentially regulated lncRNAs in our dataset have been mentioned in the context of human cancers. A non-coding gene that has been reported to be elevated in HPV$^+$ head and neck cancers is testicular cell adhesion molecule 1, pseudogene *TCAM1P*. Besides the upregulation in HPV$^+$ head and neck cancer, it was also found to be upregulated in HPV$^+$ cervical cancer, suggesting a link to HPV infection.[27,60] Physiologically, TCAMP1P expression is restricted to germ line cells. When aberrantly expressed, it can serve as an early biomarker for HPV-related cancers.[61]

Micro RNA 9-3 host gene *MIR9-3HG* gives rise to mir9-3 via the microRNA biogenesis pathway. Mir9-3 is one of three identical versions of miR9 which are encoded in different genomic loci. The mir9 family has been extensively studied in the context of different cancers including head and neck cancer. A recent miRNA profiling study of HPV$^+$ and HPV$^-$ oral- and oropharyngeal squamous cell carcinomas by Božinović et al. found miR9 to be consistently upregulated in HPV$^+$ groups of tumor samples as well as in a TCGA replication sample. Moreover, they noticed a complete absence of miR9 expression in HPV$^-$

samples.[62] In fact, mir9 has been found to be frequently methylated in oral and oropharyngeal tumors which lead to a reduced expression.[63] This study did not report HPV status, however, it is possible that the observed epigenetic silencing is the predominant phenotype in HPV$^-$ HNSCC. The role of miR9 in tumor progression is debated because of contradictory findings. For example, Hersi et al. linked miR9 downregulation to poor treatment outcome in patients and aggressive tumor cell growth in cell culture whereas its overexpression decreased proliferation of HNSCC cell lines.[64] An article that reviewed the role of miR9 in various cancer types suggested that miR9 leads to elevated proliferation and migration. The same review presented evidence for the activation of miR9 expression through HPV-E6 in HPV$^+$ cervical cancer and head and neck cancer, which corresponds to our findings as well as the Božinović study.[62,65]

LncRNAs have a tendency to be co-expressed with their neighboring genes.[43] Several lncRNAs that we identified to be differentially expressed have not yet been associated with cancer, however, we found that their parent genes or genes in the same genomic locus were linked to cancer. For example, MT1L, a metallothionein pseudogene, is found in the same locus as the four functional isoforms. Metallothioneins are believed to be involved in tumor growth, differentiation, angiogenesis, metastasis, immune escape, and drug resistance.[66] *ADORA2A-AS1* which was upregulated in the HPV$^+$ group in our study, is located on the antisense strand of the protein-coding gene *ADORA2A*. *ADORA2A* has recently been shown to have higher mRNA levels in HPV$^+$ HNSCC patients than in HPV$^-$ patients.[67] In fact, lncRNA genes in anti-sense orientation to protein-coding genes have the capacity to interfere with transcription and modulate mRNA processing of the respective protein-coding gene.[68] In this context, we speculate that ADORA2A-AS1 might have an enhancing effect on ADORA2A expression.

Although lncRNA expression was found to be highly specific in terms of cancer type[69], we note that several of the here detected DE ncRNAs have been discovered in other cancers than HNSCC. Just as the name suggests, P38 inhibited cutaneous squamous cell carcinoma associated lincRNA (PICSA) has been identified in cutaneous squamous cell carcinoma where it was shown to have tumor promoting effects.[70] AFAP1-AS1 expression has been linked to worse clinical measures in several different cancers, such as colorectal cancer and esophageal cancer.[71]

The expression of lncRNAs is frequently dysregulated in cancer. Despite the fact that the function of a majority of lncRNAs is still unknown, a few well-characterized examples have emerged as regulators of gene expression.[68] Since lncRNAs function irrespective of translation, expression levels can be linked directly to a certain phenotype. As opposed to many protein-coding genes, the expression pattern of lncNRAs is highly tissue-specific and even specific to certain types of cancer. The frequently observed secondary structure formation in lncRNAs increases their stability and facilitates their detection in bodily fluids.[69] Taken together, these properties highlight the attractiveness of lncRNAs as biomarkers.

Apart from characterizing the non-coding gene expression profile of HPV$^+$ and HPV$^-$ HNSCC, we also explored DE protein coding genes. While the distinct molecular functions of lncRNAs are often unknown, we investigated whether the identified coding DE genes could be classified according to their molecular function.

GO Molecular Function analysis of the *n=164* DE genes showed the most significant enrichment for serine proteases and aldo-keto reductases as well as molecules with growth factor binding activity, cytokine activity and receptor ligand activity. The serine protease enrichment mainly stems from proteins of the Kallikrein-related peptidase (KLK) family which were found to be significantly downregulated in the HPV$^+$ group compared to the HPV$^-$ group. This is in line with previous findings from several transcriptional profiling studies.[26,28] Many KLK family members are zymogens and participate in proteolytic cascades. They are involved in several physiological processes such as extracellular-matrix remodeling, skin desquamation and signal transduction pathways though the activation of cell-surface receptors. Besides their involvement in head and neck cancer, members of the KLK family have been linked to several other cancer types such as prostate cancer or cervical cancer.[72] KLK expression has been associated with aggressive types of squamous cell carcinomas. Especially KLK5 is believed to promote loss of tissue integrity through destruction of desmosomes and thereby promote metastasis formation[73]. This finding could partially explain the overall worse prognosis in HPV$^-$ head and neck cancer as opposed to the HPV$^+$ group.

Aldo-keto-reductase expression was significantly lower in the HPV$^+$ group which also has been reported in previous studies.[26] Functionally, the upregulation of aldo-keto reductases has been associated with combustible tobacco exposure in oral squamous cell carcinoma cell lines in the context of a toxin metabolizing response.[74] This association reflects the distinct risk factor profiles of HPV$^+$ and HPV$^-$ head and neck cancers where smoking emerges as one of the most important risk factors for HPV$^-$ tumors. However, a tobacco-independent effect is possible since the HPV$^+$ cell lines in our cell line dataset were derived from smokers whereas the HPV$^-$ cell lines were derived from non-smokers.

Most genes that were linked to cytokine activity counterintuitively were found to have a lower expression in the HPV$^+$ group relative to the HPV$^-$ group. In a microarray study, Schlecht et al. observed a similar trend of downregulated interleukins (IL-10 and IL-13) in HPV$^+$ positive tumors.[28] As immune response mediators, one would expect an upregulation of inflammatory mediators in HPV infected tissue. However, the downregulation could be related to the causes of persistent HPV infection in these cells. A compromised immune response in a subset of cells potentially could increase the vulnerability for transformative HPV infections and eventually lead to cancer.

Another transcript that we found to be expressed at significantly higher levels in the HPV$^+$ group was CDKN2A. As described earlier, the gene product of *CDKN2A*, p16, characteristically is upregulated in HPV$^+$ cancers as a response to functionally inactivation of pRb.[18,27,28] In clinical settings, p16 is used as a surrogate marker for HPV infection[21]. Thus, the upregulation of *CDKN2A* in the HPV$^+$ group in our dataset reflects previously published results. Collectively, we found a strong overlap of the DE genes in our dataset with the gene expression profiles of HPV$^+$ and HPV$^-$ head and neck cancers from previous studies.

We assessed differential gene expression with two different comparisons in the cell line dataset; a direct comparison between HPV$^+$ and HPV$^-$ cell lines and a combination of HPV$^+$/$^-$ to normal control comparisons. When compared to the normal control, we hypothesized that only genes which are downregulated in one subtype should be upregulated in the other to contribute to a unique transcriptional signature. However, only 63 DE genes were detected. As cancer cells are significantly altered compared to normal cells, we concluded that the general differences between normal- and cancer cells masked the differences between the two cancer subtypes.

Transcriptional signatures were initially discovered in a very small collection of HNSCC cell lines, with two HPV$^+$ and four HPV$^-$ cell lines. We attempted to find larger sets of publicly available RNA sequencing data from HNSCC cell lines to enlarge the discovery sample. However, RNA sequencing data is made publicly available in form of FPKM values for the most part which was incompatible with the presented computational pipeline. Another challenge using publicly available sequencing data is the use of different identifiers. The conversion between different identifiers interferes with the reliable attribution of annotated counts to the respective gene. As a consequence, several rows from the TCGA dataset had to be removed in this study due to deprecated gene annotations.

We noted a low number of DE non-coding genes in the TCGA dataset (*n=13).* A potential explanation could be the diverse cellular composition of tumors. TCGA requires 60% tumor nuclei content to qualify as a tumor sample. As lncRNA expression is highly tissue- and even cell type specific, a fraction of DE lncRNAs might not have reached significance. The low number of non-coding genes contained in the TCGA dataset limited our ability to independently test our findings from the cell line dataset. Thereby, transcripts that significantly contribute to the gene expression profiles of the two groups may not have been detected. A validation sample of RNA sequencing data from HPV$^+$ and HPV$^-$ tumors analyzed with the same computational pipeline as the cell line dataset would provide more insight into the non-coding transcriptome.

The identification of HPV status as the main influence on differential expression as well as the identification of primary tissue as a secondary factor was based on cluster estimation on PCA plots. Even though PCA is a tool to structure highly complex datasets, it did not lead to distinct clusters in the 2D- visualization. Especially in the highly scattered TCGA PCA plot, cluster formation may have been overinterpreted. However, due to the strong evidence for HPV-mediated differences in gene expression[42,75] and the importance of the tissue of origin, our analysis represents a reasonable and impartial foundation.

Following up on both coding and non-coding genes that have been identified to be differentially expressed among the two groups, a more exhaustive analysis is needed to put them into a biological context. Our aim is to get a better understanding of the long non-coding RNAs that are essential in the context of head and neck cancer which we plan to address in a comprehensive lncRNA drop-out screen using a CRISPR-dCas9-KRAB system. Furthermore, a single-cell transcriptomic approach could give a more detailed picture of the non-coding signatures in different tumor cell types. An investigation of the regulatory network behind the DE genes found in this study might yield a stronger link to established molecular signatures of the disease. Together with a more comprehensive dataset in connection with more clinical data, these analyses could allow for a more refined subclassification of HPV$^+$ and HPV$^-$ head and neck cancers.

This work has contributed to elaborate the transcriptional profiles of HPV$^+$ and HPV$^-$ head and neck cancer. We identified several non-coding genes that had not been linked to HPV-related subtypes of head and neck cancer so far. With respect to the identification of new biomarkers for early detection of tumors, disease monitoring, prediction of outcome and adjustment of treatment strategy, the precise mapping of non-coding expressional profiles becomes highly relevant. In addition, we were able to identify a set of DE coding genes and annotated their molecular function. Both the coding and the non-coding landscape of head and neck cancer remain to be fully understood and future work will be needed to translate this knowledge into clinical applications.

References

1. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).
2. Bohn, J. A. *et al.* Identification of diverse target RNAs that are functionally regulated by human Pumilio proteins. *Nucleic Acids Research* **46**, 362–386 (2018).
3. Loedige, I., Gaidatzis, D., Sack, R., Meister, G. & Filipowicz, W. The mammalian TRIM-NHL protein TRIM71/LIN-41 is a repressor of mRNA function. *Nucleic Acids Res.* **41**, 518–532 (2013).
4. Schwamborn, J. C., Berezikov, E. & Knoblich, J. A. The TRIM-NHL Protein TRIM32 Activates MicroRNAs and Prevents Self-Renewal in Mouse Neural Progenitors. *Cell* **136**, 913–925 (2009).
5. Suzuki, A. *et al.* Dead end1 is an essential partner of NANOS2 for selective binding of target RNAs in male germ cell development. *EMBO Rep.* **17**, 37–46 (2016).
6. Zhang, M. *et al.* Post-transcriptional regulation of mouse neurogenesis by Pumilio proteins. *Genes Dev.* **31**, 1354–1369 (2017).
7. Tichon, A., Perry, R. B.-T., Stojic, L. & Ulitsky, I. SAM68 is required for regulation of Pumilio by the NORAD long noncoding RNA. *Genes Dev.* **32**, 70–78 (2018).
8. Carreira-Rosario, A. *et al.* Repression of Pumilio Protein Expression by Rbfox1 Promotes Germ Cell Differentiation. *Dev. Cell* **36**, 562–571 (2016).
9. Vaidyanathan, P. P., AlSadhan, I., Merriman, D. K., Al-Hashimi, H. M. & Herschlag, D. Pseudouridine and N6-methyladenosine modifications weaken PUF protein/RNA interactions. *RNA* **23**, 611–618 (2017).
10. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68**, 394–424 (2018).
11. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019: Cancer Statistics, 2019. *CA: A Cancer Journal for Clinicians* **69**, 7–34 (2019).
12. Lambert, R., Sauvaget, C., de Camargo Cancela, M. & Sankaranarayanan, R. Epidemiology of cancer from the oral cavity and oropharynx: *European Journal of Gastroenterology & Hepatology* **23**, 633–641 (2011).
13. Gillison, M. L. *et al.* Distinct Risk Factor Profiles for Human Papillomavirus Type 16–Positive and Human Papillomavirus Type 16–Negative Head and Neck Cancers. *JNCI: Journal of the National Cancer Institute* **100**, 407–420 (2008).
14. Chaturvedi, A. K. *et al.* Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States. *JCO* **29**, 4294–4301 (2011).
15. Näsman, A. *et al.* Incidence of human papillomavirus (HPV) positive tonsillar carcinoma in Stockholm, Sweden: An epidemic of viral-induced carcinoma? *Int. J. Cancer* **125**, 362–366 (2009).

16. Dok, R. & Nuyts, S. HPV Positive Head and Neck Cancers: Molecular Pathogenesis and Evolving Treatment Strategies. *Cancers (Basel)* **8**, (2016).

17. Craig, S. G. *et al.* Recommendations for determining HPV status in patients with oropharyngeal cancers under TNM8 guidelines: a two-tier approach. *Br J Cancer* **120**, 827–833 (2019).

18. Zaravinos, A. An updated overview of HPV-associated head and neck carcinomas. *Oncotarget* **5**, 3956–3969 (2014).

19. Perri, F., Pisconti, S. & Della Vittoria Scarpati, G. P53 mutations and cancer: a tight linkage. *Ann Transl Med* **4**, 522 (2016).

20. Romagosa, C. *et al.* p16Ink4a overexpression in cancer: a tumor suppressor gene associated with senescence and high-grade tumors. *Oncogene* **30**, 2087 (2011).

21. Leemans, C. R., Snijders, P. J. F. & Brakenhoff, R. H. The molecular landscape of head and neck cancer. *Nature Reviews Cancer* **18**, 269–282 (2018).

22. Zhang, L. *et al.* Loss of heterozygosity (LOH) profiles--validated risk predictors for progression to oral cancer. *Cancer Prev Res (Phila)* **5**, 1081–1089 (2012).

23. Kiessling, S.-Y., Broglie, M. A., Soltermann, A., Huber, G. F. & Stoeckli, S. J. Comparison of PI3K Pathway in HPV-Associated Oropharyngeal Cancer With and Without Tobacco Exposure: PI3K pathway in OPSCC. *Laryngoscope Investigative Otolaryngology* **3**, 283–289 (2018).

24. The Cancer Genome Atlas Network *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576 (2015).

25. Seiwert, T. Y. *et al.* Integrative and Comparative Genomic Analysis of HPV-Positive and HPV-Negative Head and Neck Squamous Cell Carcinomas. *Clinical Cancer Research* **21**, 632–641 (2015).

26. Martinez, I., Wang, J., Hobson, K. F., Ferris, R. L. & Khan, S. A. Identification of differentially expressed genes in HPV-positive and HPV-negative oropharyngeal squamous cell carcinomas. *Eur. J. Cancer* **43**, 415–432 (2007).

27. Slebos, R. J. C. *et al.* Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma. *Clin. Cancer Res.* **12**, 701–709 (2006).

28. Schlecht, N. *et al.* Gene expression profiles in HPV-infected head and neck cancer. *J. Pathol.* **213**, 283–293 (2007).

29. Jarroux, J., Morillon, A. & Pinskaya, M. History, Discovery, and Classification of lncRNAs. in *Long Non Coding RNA Biology* (ed. Rao, M. R. S.) 1–46 (Springer Singapore, 2017). doi:10.1007/978-981-10-5203-3_1

30. Sirchia, S. M. *et al.* Misbehaviour of XIST RNA in breast cancer cells. *PLoS ONE* **4**, e5559 (2009).

31. Yildirim, E. *et al.* Xist RNA Is a Potent Suppressor of Hematologic Cancer in Mice. *Cell* **152**, 727–742 (2013).

32. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).

33. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409–419 (2010).

34. Cheetham, S. W., Gruhl, F., Mattick, J. S. & Dinger, M. E. Long noncoding RNAs and the genetics of cancer. *British Journal Of Cancer* **108**, 2419 (2013).

35. Tripathi, V. *et al.* The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Molecular Cell* **39**, 925–938 (2010).

36. Schmidt, L. H. *et al.* The Long Noncoding MALAT-1 RNA Indicates a Poor Prognosis in Non-small Cell Lung Cancer and Induces Migration and Tumor Growth. *Journal of Thoracic Oncology* **6**, 1984–1992 (2011).

37. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).

38. Li, D. *et al.* Long Intergenic Noncoding RNA HOTAIR Is Overexpressed and Regulates PTEN Methylation in Laryngeal Squamous Cell Carcinoma. *The American Journal of Pathology* **182**, 64–70 (2013).

39. Tang, H., Wu, Z., Zhang, J. & Su, B. Salivary lncRNA as a potential marker for oral squamous cell carcinoma diagnosis. *Mol Med Rep* **7**, 761–766 (2013).

40. Fang, Z. *et al.* Increased expression of the long non-coding RNA UCA1 in tongue squamous cell carcinomas: a possible correlation with cancer metastasis. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology* **117**, 89–95 (2014).

41. Zhang, L.-M. *et al.* Long non-coding RNA ANRIL promotes tumorgenesis through regulation of FGFR1 expression by sponging miR-125a-3p in head and neck squamous cell carcinoma. *Am J Cancer Res* **8**, 2296–2310 (2018).

42. Salyakina, D. & Tsinoremas, N. F. Non-coding RNAs profiling in head and neck cancers. *npj Genomic Med* **1**, 15004 (2016).

43. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25**, 1915–1927 (2011).

44. de la Taille, A. Progensa^{TM} PCA3 test for prostate cancer detection. *Expert Review of Molecular Diagnostics* **7**, 491–497 (2007).

45. Poser, I. *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nature Methods* **5**, 409–415 (2008).

46. Invitrogen^{TM} by life technologies^{TM}. pENTR^{TM} Directional TOPO®Cloning KitsFive-minute, directional TOPO® Cloning of blunt-end PCR products into an entry vector for the Gateway® System.

47. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* **46**, W537–W544 (2018).

48. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

49. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).

50. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

51. Guangchuang Yu. *enrichplot*. (Bioconductor, 2018). doi:10.18129/b9.bioc.enrichplot

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

53. Dickson, M. A. *et al.* Human keratinocytes that express hTERT and also bypass a p16(INK4a)-enforced mechanism that limits life span become immortal yet retain normal growth and differentiation characteristics. *Mol. Cell. Biol.* **20**, 1436–1447 (2000).

54. Balló, H. *et al.* Establishment and characterization of four cell lines derived from human head and neck squamous cell carcinomas for an autologous tumor-fibroblast in vitro model. *Anticancer Res.* **19**, 3827–3836 (1999).

55. White, J. S. *et al.* The influence of clinical and demographic risk factors on the establishment of head and neck squamous cell carcinoma cell lines. *Oral Oncol.* **43**, 701–712 (2007).

56. Momose, F. *et al.* Variant sublines with different metastatic potentials selected in nude mice from human oral squamous cell carcinomas. *J. Oral Pathol. Med.* **18**, 391–395 (1989).

57. Rangan, S. R. A new human cell line (FaDu) from a hypopharyngeal carcinoma. *Cancer* **29**, 117–121 (1972).

58. Gioanni, J. *et al.* Two new human tumor cell lines derived from squamous cell carcinomas of the tongue: establishment, characterization and response to cytotoxic treatment. *Eur J Cancer Clin Oncol* **24**, 1445–1455 (1988).

59. Rheinwald, J. G. & Beckett, M. A. Tumorigenic keratinocyte lines requiring anchorage and fibroblast support cultured from human squamous cell carcinomas. *Cancer Res.* **41**, 1657–1663 (1981).

60. Pyeon, D. *et al.* Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res.* **67**, 4605–4619 (2007).

61. Strati, K. & Lambert, P. F. HUMAN PAPILLOMAVIRUS ASSOCIATION WITH HEAD AND NECK CANCERS: UNDERSTANDING VIRUS BIOLOGY AND USING IT IN THE DEVELOPMENT OF CANCER DIAGNOSTICS. *Expert Opin Med Diagn* **2**, 11–20 (2008).

62. Božinović, K. *et al.* Genome-wide miRNA profiling reinforces the importance of miR-9 in human papillomavirus associated oral and oropharyngeal head and neck cancer. *Sci Rep* **9**, 2306 (2019).

63. Minor, J. *et al.* Methylation of microRNA-9 is a specific and sensitive biomarker for oral and oropharyngeal squamous cell carcinomas. *Oral Oncol.* **48**, 73–78 (2012).

64. Hersi, H. M., Raulf, N., Gaken, J., Folarin, N. & Tavassoli, M. MicroRNA-9 inhibits growth and invasion of head and neck cancer cells and is a predictive biomarker of response to plerixafor, an inhibitor of its target CXCR4. *Mol Oncol* **12**, 2023–2041 (2018).

65. Nowek, K., Wiemer, E. A. C. & Jongen-Lavrencic, M. The versatile nature of miR-9/9* in human cance. *Oncotarget* **9**, (2018).

66. Si, M. & Lang, J. The roles of metallothioneins in carcinogenesis. *J Hematol Oncol* **11**, 107 (2018).

67. Vogt, T. J. *et al.* Detailed analysis of adenosine A2a receptor (ADORA2A) and CD73 (5'-nucleotidase, ecto, NT5E) methylation and gene expression in head and neck squamous cell carcinoma patients. *Oncoimmunology* **7**, e1452579 (2018).

68. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **14**, 699–712 (2013).

69. Yan, X. *et al.* Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell* **28**, 529–540 (2015).

70. Piipponen, M., Heino, J., Kähäri, V.-M. & Nissinen, L. Long non-coding RNA PICSAR decreases adhesion and promotes migration of squamous carcinoma cells by downregulating α2β1 and α5β1 integrin expression. *Biol Open* **7**, (2018).

71. Wang, Y. *et al.* Long non-coding RNA AFAP1-AS1 is a novel biomarker in various cancers: a systematic review and meta-analysis based on the literature and GEO datasets. *Oncotarget* **8**, 102346–102360 (2017).

72. Kontos, C. K. & Scorilas, A. Kallikrein-related peptidases (KLKs): a gene family of novel cancer biomarkers. *Clinical Chemistry and Laboratory Medicine* **50**, (2012).

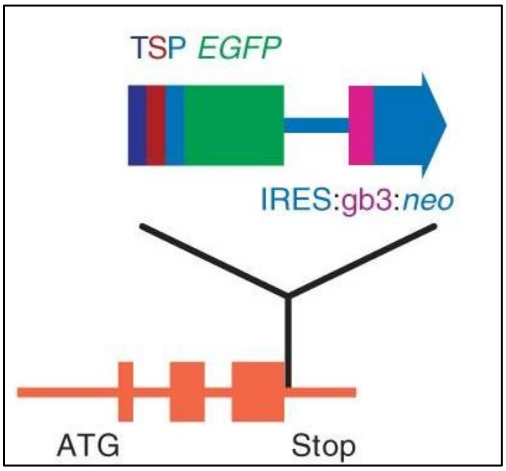
73. Jiang, R., Shi, Z., Johnson, J. J., Liu, Y. & Stack, M. S. Kallikrein-5 promotes cleavage of desmoglein-1 and loss of cell-cell cohesion in oral squamous cell carcinoma. *J. Biol. Chem.* **286**, 9127–9135 (2011).

74. Woo, S. *et al.* AKR1C1 as a Biomarker for Differentiating the Biological Effects of Combustible from Non-Combustible Tobacco Products. *Genes (Basel)* **8**, (2017).

75. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat Commun* **4**, 2513 (2013).

76. Zumsteg, Z. S. *et al.* Taselisib (GDC-0032), a Potent -Sparing Small Molecule Inhibitor of PI3K, Radiosensitizes Head and Neck Squamous Carcinomas Containing Activating PIK3CA Alterations. *Clinical Cancer Research* **22**, 2009–2019 (2016).

Supplementary material



Supplementary figure S1. Architecture of BAC construct with LAP tag (adapted from Poser et al., 2008)[3]. The C-terminal tag is inserted right after the last codon with the stop codon removed. The cassette consists of various cleavage sites for immunoprecipitation (P: PreScission cleavage site, S: S-peptide, T: TEV cleavage site) and the eGFP tag. Elements after the eGFP tag support bacterial amplification and selection (IRES: internal ribosome entry site, gb3: bacterial promoter, neo: neomycin resistance gene)

Supplementary table S2. Additional information on HNSCC patients and the respective patient-derived cell lines used in this study. M= male, p= Protein, c = coding sequence, TNM= TNM Classification of Malignant Tumors.

Cell line and reference	Age, gender	Primary site	Mutations	Risk factor exposure	Cause of death	Additional information
Cal33[58]	69, M	Tongue	P53: missense pR175H PIK3CA mutation[76]	HPV negative Non-smoker	Died 13 months after diagnosis	
HSC3[56]	63, M	Tongue	P53: insertion-frameshift c.911_912insTAAG; c.915_916insTAAG	HPV negative Non-smoker		Established from metastatic lymph nodes
SCC2[54]	58, M	Hypopharyngeal primary tumor	PTEN loss[76]	Smoker HPV-16$^+$	Died from pulmonary metastases 1 year after diagnosis	TNM: T1N3
FaDu[57]	56, M	Hypooharynx	P53: missense p.R248L CDKN2A intronic: c.151-1G>T	HPV negative Non-smoker		
SCC4[59]	55, M	Tongue	P53: missense pP151S	HPV negative Non-smoker		radiation and methotrexate treatment 16 months prior to establishment
SCC90[55]	46, M	Base of tongue recurrence	PIK3CA amplification, PTEN mutation[76]	Smoker, alcohol consumption, HPV-16$^+$	Died 4 years after diagnosis	TNM: T2N0

Supplementary table S3. Commonly DE genes in cell line and TCGA dataset. *N=167,* CL= cell line dataset.

RefSeqID	l2fc CL	padj CL	symbol	Gene name	l2fc TCGA	padj TCGA
NM_024337	-22.869	1.47E-06	IRX1	iroquois homeobox 1	-2.665	0.00048968
NM_001129826	-22.373	1.37E-06	CSAG3	CSAG family member 3	-2.873	0.00089807
NM_001102576	-19.051	6.80E-09	CSAG1	chondrosarcoma associated gene 1	-5.706	8.89E-09
NR_027082	-11.831	1.23E-10	SFTA1P	surfactant associated 1, pseudogene	-2.656	3.59E-05
NM_001281431	-10.987	1.67E-31	KLK8	kallikrein related peptidase 8	-2.744	7.19E-07
NM_002422	-10.807	1.70E-06	MMP3	matrix metallopeptidase 3	-2.529	2.10E-05
NM_025130	-10.617	1.24E-07	HKDC1	hexokinase domain containing 1	-3.414	9.23E-08
NM_001185156	-10.531	0.00012376	IL24	interleukin 24	-2.672	5.59E-06
NM_001130014	-10.492	4.59E-06	PSG5	pregnancy specific beta-1-glycoprotein 5	-3.538	3.53E-06
NR_024089	-10.366	2.07E-05	PICSAR	P38 inhibited cutaneous squamous cell carcinoma associated lincRNA	-2.953	7.79E-05
NM_001303419	-10.339	3.02E-05	TRIML2	tripartite motif family like 2	-2.036	0.03780108
NM_012114	-10.194	0.00010175	CASP14	caspase 14	-3.682	1.03E-05
NM_000804	-10.178	0.00028908	FOLR3	folate receptor 3	-3.360	1.15E-07
NM_000802	-10.172	2.90E-06	FOLR1	folate receptor 1	-2.726	0.00036543
NM_001077491	-10.007	1.36E-29	KLK5	kallikrein related peptidase 5	-4.996	3.90E-15
NM_001271534	-10.005	0.01195643	DSCAM	DS cell adhesion molecule	-3.001	4.38E-05
NM_003880	-9.830	0.00018271	WISP3	WNT1 inducible signaling pathway protein 3	-3.259	3.80E-06
NM_032654	-9.416	3.09E-07	AFAP1-AS1	AFAP1 antisense RNA 1	-4.825	6.49E-11
NM_001207053	-9.234	3.54E-12	KLK7	kallikrein related peptidase 7	-2.632	9.97E-06
NM_000329	-9.171	2.60E-05	RPE65	RPE65, retinoid isomerohydrolase	-3.228	8.09E-05
NM_001146157	-8.830	1.60E-05	FAM25A	family with sequence similarity 25 member A	-3.534	4.68E-07
NM_012315	-8.816	9.91E-19	KLK9	kallikrein related peptidase 9	-3.033	3.04E-07
NM_001197097	-8.280	0.00032614	PRSS3	protease, serine 3	-2.571	0.00050375
NM_001008778	-7.984	0.00096199	SPDYC	speedy/RINGO cell cycle regulator family member C	-4.972	6.15E-06
NM_001796	-7.970	0.00550487	CDH8	cadherin 8	-2.101	0.00097586
NM_001008272	-7.878	0.00063642	TAGLN3	transgelin 3	-2.690	0.00012604
NM_001077500	-7.869	1.61E-24	KLK10	kallikrein related peptidase 10	-2.310	3.28E-05
NM_033197	-7.857	0.04585834	BPIFB1	BPI fold containing family B member 1	-3.357	0.00176202
NR_001447	-7.821	8.58E-06	MT1L	metallothionein 1L, pseudogene	-2.245	4.11E-05
NM_001740	-7.663	4.69E-05	CALB2	calbindin 2	-2.874	3.38E-06
NM_001785	-7.315	5.19E-11	CDA	cytidine deaminase	-3.242	7.07E-10
NM_001302813	-7.230	0.00235483	C20orf197	chromosome 20 open reading frame 197	-2.768	3.88E-05
NM_178438	-7.141	0.03300194	LCE5A	late cornified envelope 5A	-2.915	0.00163849
NM_001137556	-7.121	0.00015458	FAM25BP	protein FAM25	-3.740	3.25E-08
NM_000350	-7.100	0.0001756	ABCA4	ATP binding cassette subfamily A member 4	-2.093	0.00089545
NM_005330	-6.938	0.01128927	HBE1	hemoglobin subunit epsilon 1	-5.230	5.74E-05

NM_001177969	-6.917	0.00393271	VIT	vitrin	-2.242	0.00042521
NM_002178	-6.848	3.88E-20	IGFBP6	insulin like growth factor binding protein 6	-2.464	8.77E-06
NM_018724	-6.845	0.01117692	IL20	interleukin 20	-2.199	0.0017731
NM_005568	-6.831	0.00948638	LHX1	LIM homeobox 1	-3.022	1.44E-05
NM_001256536	-6.685	0.02173849	PRMT8	protein arginine methyltransferase 8	-3.331	0.00035072
NM_001135639	-6.622	8.56E-06	CNGB1	cyclic nucleotide gated channel beta 1	-2.929	4.77E-08
NM_002192	-6.576	8.48E-06	INHBA	inhibin beta A subunit	-2.247	2.06E-06
NM_001080518	-6.483	0.00245604	LIPK	lipase family member K	-2.999	2.65E-06
NM_000872	-6.457	4.85E-08	HTR7	5-hydroxytryptamine receptor 7	-2.141	4.91E-06
NM_001145938	-6.340	1.68E-14	MMP1	matrix metallopeptidase 1	-2.971	1.03E-08
NM_001126063	-6.299	1.18E-06	KHDC1L	KH domain containing 1 like	-2.237	0.00209613
NM_031950	-6.284	0.04374776	FGFBP2	fibroblast growth factor binding protein 2	-4.158	3.36E-07
NM_152762	-6.192	0.02110855	TSGA10IP	testis specific 10 interacting protein	-2.182	0.01253413
NM_002427	-6.004	1.71E-13	MMP13	matrix metallopeptidase 13	-3.584	3.25E-06
NM_001272005	-6.001	0.01365226	OTOP3	otopetrin 3	-2.544	0.00854731
NM_001012964	-5.919	1.57E-07	KLK6	kallikrein related peptidase 6	-2.792	2.56E-05
NM_032556	-5.871	0.00143121	IL1F10	interleukin 1 family member 10	-3.089	3.23E-05
NM_012275	-5.642	2.56E-07	IL36RN	interleukin 36 receptor antagonist	-2.133	0.00056742
NM_002963	-5.426	0.03623282	S100A7	S100 calcium binding protein A7	-2.298	0.00210706
NM_001318189	-5.389	2.25E-05	BTBD16	BTB domain containing 16	-2.884	7.47E-05
NM_019598	-5.170	0.01818432	KLK12	kallikrein related peptidase 12	-2.784	0.00020152
NM_002994	-5.159	0.03499443	CXCL5	C-X-C motif chemokine ligand 5	-2.040	0.00181669
NM_001253908	-5.088	5.02E-17	AKR1C3	aldo-keto reductase family 1 member C3	-2.244	0.00044961
NM_004625	-5.049	1.06E-05	WNT7A	Wnt family member 7A	-2.113	0.00019856
NM_001171191	-4.977	0.00377473	GDPD2	glycerophosphodiester phosphodiesterase domain containing 2	-2.208	0.00058986
NM_001128932	-4.955	1.14E-65	CYP4F11	cytochrome P450 family 4 subfamily F member 11	-2.005	0.00435798
NM_020299	-4.947	0.00017585	AKR1B10	aldo-keto reductase family 1 member B10	-2.513	3.08E-05
NM_004959	-4.937	0.0138161	NR5A1	nuclear receptor subfamily 5 group A member 1	-4.769	1.79E-05
NM_001103160	-4.904	0.00312598	SH2D5	SH2 domain containing 5	-2.637	7.41E-08
NM_002820	-4.735	0.00017406	PTHLH	parathyroid hormone like hormone	-2.484	1.09E-07
NM_013281	-4.526	0.011774	FLRT3	fibronectin leucine rich transmembrane protein 3	-2.088	0.00170535
NM_002016	-4.426	0.00021176	FLG	filaggrin	-2.243	0.00304048
NM_001201325	-4.234	0.0014474	PDZK1	PDZ domain containing 1	-2.305	3.40E-05
NM_152611	-4.180	0.00047206	LRRN4	leucine rich repeat neuronal 4	-2.587	1.01E-05
NM_000526	-4.126	2.60E-05	KRT14	keratin 14	-2.271	2.89E-05
NM_001287758	-3.717	6.45E-05	COL4A6	collagen type IV alpha 6 chain	-2.483	1.16E-07

NM_001172439	-3.705	0.04081628	ENDOU	endonuclease, poly(U) specific	-2.260	0.00187544
NM_152763	-3.661	0.00453749	AKNAD1	AKNA domain containing 1	-2.084	0.00025828
NM_001135241	-3.563	1.59E-12	AKR1C2	aldo-keto reductase family 1 member C2	-2.439	0.00013641
NM_001145106	-3.124	1.68E-05	FIBCD1	fibrinogen C domain containing 1	-2.342	0.00096429
NM_001300845	-2.742	0.01075204	SLC35F3	solute carrier family 35 member F3	-3.109	2.66E-10
NM_152443	-2.693	0.00339528	RDH12	retinol dehydrogenase 12 (all-trans/9-cis/11-cis)	-4.014	8.53E-12
NM_001844	-2.688	0.0009779	COL2A1	collagen type II alpha 1 chain	-2.689	4.56E-06
NM_001311182	-2.680	0.02752969	KLK14	kallikrein related peptidase 14	-3.317	5.91E-07
NM_001165960	-2.624	0.03273011	ALOXE3	arachidonate lipoxygenase 3	-2.115	4.27E-05
NM_001353	-2.574	0.0107339	AKR1C1	aldo-keto reductase family 1 member C1	-2.485	0.00017684
NM_001015886	-2.557	0.02339872	HMGA2	high mobility group AT-hook 2	-2.762	2.18E-08
NM_000805	-2.281	0.04106282	GAST	gastrin	-2.810	2.96E-05
NM_001102658	-2.258	0.02390281	CT62	cancer/testis antigen 62	-2.062	0.00982914
NM_001031692	-2.208	0.03353162	LRRC17	leucine rich repeat containing 17	-2.494	1.14E-05
NM_004948	-2.152	0.02980881	DSC1	desmocollin 1	-2.988	7.14E-07
NR_024391	2.020	4.12E-07	MIR924HG	MIR924 host gene	2.275	2.30E-06
NM_001105578	2.165	1.95E-13	SYCE2	synaptonemal complex central element protein 2	3.188	2.50E-26
NM_001039780	2.350	0.00068796	CCNI2	cyclin I family member 2	2.296	3.96E-06
NM_198560	2.353	0.03572796	LHFPL4	LHFPL tetraspan subfamily member 4	5.119	8.76E-07
NM_020309	2.380	3.40E-09	SLC17A7	solute carrier family 17 member 7	2.309	1.49E-06
NM_004209	2.748	3.77E-05	SYNGR3	synaptogyrin 3	2.990	9.98E-15
NM_000077	2.915	0.00045925	CDKN2A	cyclin dependent kinase inhibitor 2A	2.353	0.00035551
NM_001039784	3.375	0.03435945	ADORA2A-AS1	ADORA2A antisense RNA 1	2.332	1.25E-20
NM_001322799	3.391	0.04082416	KCNS1	potassium voltage-gated channel modifier subfamily S member 1	3.519	1.78E-06
NM_013356	3.534	0.00177614	SLC16A8	solute carrier family 16 member 8	2.072	8.74E-06
NM_001099652	3.570	1.15E-12	GPR137C	G protein-coupled receptor 137C	2.127	3.87E-15
NM_004386	3.759	0.00014154	NCAN	neurocan	2.238	0.00014178
NM_001308165	3.823	3.68E-05	SOX30	SRY-box 30	3.675	5.81E-14
NM_015063	3.859	0.0002245	SLC8A2	solute carrier family 8 member A2	2.380	0.00042382
NM_001345843	3.881	0.00035134	C19orf57	chromosome 19 open reading frame 57	2.274	1.04E-15
NM_001145451	3.908	2.55E-05	ARHGEF33	Rho guanine nucleotide exchange factor 33	3.698	9.01E-25
NM_000092	4.163	9.31E-08	COL4A4	collagen type IV alpha 4 chain	2.281	5.27E-06
NM_001037225	4.402	0.00330947	MAJIN	membrane anchored junction protein	5.915	2.73E-20
NR_003574	4.418	0.00075759	ABCA17P	ATP binding cassette subfamily A member 17, pseudogene	4.671	4.49E-21

NR_003927	4.437	0.02342401	UOX	urate oxidase (pseudogene)	3.088	0.00172094
NM_001080448	4.511	0.00093939	EPHA6	EPH receptor A6	2.812	0.00126095
NM_006593	4.512	0.00998127	TBR1	T-box, brain 1	3.857	0.00022751
NM_001039905	4.592	5.31E-09	C15orf56	chromosome 15 open reading frame 56	2.001	0.00040101
NM_001168474	4.679	8.05E-06	TAF7L	TATA-box binding protein associated factor 7 like	5.135	1.87E-35
NM_001127608	4.783	3.49E-17	FAM189A2	family with sequence similarity 189 member A2	2.007	8.98E-05
NM_014258	4.865	1.42E-12	SYCP2	synaptonemal complex protein 2	5.122	9.19E-37
NR_015411	4.874	6.57E-05	MIR9-3HG	MIR9-3 host gene	3.392	6.23E-14
NM_000835	4.937	7.13E-05	GRIN2C	glutamate ionotropic receptor NMDA type subunit 2C	3.184	6.81E-31
NM_001308245	5.185	0.00013099	BTNL9	butyrophilin like 9	4.253	6.34E-16
NM_033176	5.226	9.55E-07	NKX2-4	NK2 homeobox 4	7.999	2.84E-05
NM_003026	5.374	4.16E-07	SH3GL2	SH3 domain containing GRB2 like 2, endophilin A1	3.137	3.72E-06
NM_182583	5.476	0.00759281	FAM182A	family with sequence similarity 182 member A	2.465	4.35E-06
NM_001039548	5.526	1.10E-05	KLHL35	kelch like family member 35	3.344	4.82E-20
NM_001009565	5.627	4.72E-06	CDKL4	cyclin dependent kinase like 4	2.382	0.00028106
NM_144688	5.662	1.04E-05	CCDC155	coiled-coil domain containing 155	5.426	7.98E-11
NM_001168465	5.676	4.18E-19	MAP7D2	MAP7 domain containing 2	2.837	4.06E-05
NM_015982	5.802	1.93E-07	YBX2	Y-box binding protein 2	3.254	5.04E-11
NM_001100411	5.831	1.17E-09	FAM184A	family with sequence similarity 184 member A	2.067	7.00E-06
NM_016327	5.923	5.06E-07	UPB1	beta-ureidopropionase 1	4.517	7.20E-36
NM_020991	6.097	9.95E-08	CSH2	chorionic somatomammotropin hormone 2	3.194	0.02378686
NM_001098475	6.128	0.00011401	TDRD10	tudor domain containing 10	4.410	2.52E-25
NM_001008783	6.148	1.36E-06	SLC35D3	solute carrier family 35 member D3	2.434	0.00235994
NM_001089	6.173	5.55E-14	ABCA3	ATP binding cassette subfamily A member 3	2.309	4.56E-06
NM_001321525	6.186	2.08E-13	GPAT2	glycerol-3-phosphate acyltransferase 2, mitochondrial	2.805	1.54E-09
NM_016170	6.197	3.92E-27	TLX2	T-cell leukemia homeobox 2	3.056	1.59E-06
NM_001145720	6.511	0.00051536	ZBTB8B	zinc finger and BTB domain containing 8B	2.052	6.70E-08
NM_001002838	6.516	2.34E-12	WNK3	WNK lysine deficient protein kinase 3	2.039	2.83E-14
NM_000625	6.523	0.00126733	NOS2	nitric oxide synthase 2	2.778	8.47E-06
NM_001098834	6.660	2.97E-30	GBX1	gastrulation brain homeobox 1	6.099	3.90E-15
NM_005519	6.697	7.03E-05	HMX2	H6 family homeobox 2	4.345	3.85E-06
NM_001330438	6.739	1.34E-15	DDX25	DEAD-box helicase 25	3.979	6.96E-08
NR_002947	6.745	2.09E-06	TCAM1P	testicular cell adhesion molecule 1, pseudogene	3.450	7.07E-42
NM_001291501	6.787	8.59E-36	SMC1B	structural maintenance of chromosomes 1B	6.848	5.05E-45
NM_001271507	7.025	9.76E-23	CCDC177	coiled-coil domain containing 177	3.113	5.14E-06

NM_001101419	7.090	3.41E-10	ZNF541	zinc finger protein 541	5.038	1.45E-18
NM_001163560	7.393	2.81E-58	MEIOB	meiosis specific with OB domains	3.033	0.00080196
NM_174978	7.399	1.53E-14	C14orf39	chromosome 14 open reading frame 39	2.608	0.0001546
NM_138370	7.646	2.79E-06	PKDCC	protein kinase domain containing, cytoplasmic	2.637	1.06E-05
NM_001297764	7.767	3.02E-07	USH1C	USH1 protein network component harmonin	3.302	0.00073197
NM_013435	8.015	5.63E-06	RAX	retina and anterior neural fold homeobox	3.233	0.00080607
NM_001007563	8.084	9.47E-07	IGFBPL1	insulin like growth factor binding protein like 1	2.193	0.00212843
NM_001079533	8.201	7.48E-09	CPEB1	cytoplasmic polyadenylation element binding protein 1	2.783	3.94E-11
NM_152513	8.446	1.13E-46	MEI1	meiotic double-stranded break formation protein 1	2.926	1.41E-20
NM_031909	8.768	1.42E-06	C1QTNF4	C1q and TNF related 4	3.565	3.51E-10
NM_001271862	8.811	9.07E-38	PNLDC1	PARN like, ribonuclease domain containing 1	3.484	0.00011116
NM_015720	8.811	2.02E-61	PODXL2	podocalyxin like 2	2.468	6.19E-07
NM_001330375	9.078	2.57E-34	HLF	HLF, PAR bZIP transcription factor	2.147	8.53E-05
NM_015253	9.152	9.11E-17	WSCD1	WSC domain containing 1	2.242	1.58E-06
NM_153046	9.357	1.35E-50	TDRD9	tudor domain containing 9	3.561	2.32E-11
NM_001345928	9.433	0.0033834	SHCBP1L	SHC binding and spindle associated 1 like	6.778	1.17E-14
NM_001012415	9.556	7.74E-22	SOHLH1	spermatogenesis and oogenesis specific basic helix-loop-helix 1	6.175	1.42E-05
NM_001145640	9.661	6.95E-20	ZFR2	zinc finger RNA binding protein 2	5.664	2.07E-101
NM_001168647	9.862	1.01E-24	CNKSR2	connector enhancer of kinase suppressor of Ras 2	2.157	0.00137417
NM_000162	9.959	7.43E-40	GCK	glucokinase	3.537	9.03E-10
NM_021076	10.301	7.15E-225	NEFH	neurofilament heavy	4.869	3.72E-18
NM_001008537	10.344	8.25E-46	NEXMIF	neurite extension and migration factor	3.104	1.18E-05
NM_001244008	10.604	1.40E-40	KIF1A	kinesin family member 1A	2.707	0.00503971
NM_032727	10.731	1.15E-76	INA	internexin neuronal intermediate filament protein alpha	2.709	0.00198722
NM_022897	11.262	6.47E-36	RANBP17	RAN binding protein 17	2.204	2.28E-07
NM_001131034	13.813	1.15E-21	RNF212	ring finger protein 212	2.450	3.88E-05

Supplementary table S4. List of overrepresented GO Molecular function terms in commonly DE genes. GeneRatio refers to the number of genes from the dataset that were associated with the respective GO term. In total, n=135 genes out of n=167 genes from the list were contained in the database for GO term analysis (denominator of GeneRatio). The total number of genes annotated to a respective term divided by the total number of genes in the GO Molecular Function database is given in the BgRatio.

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	GeneSymbol
GO:0008236	serine-type peptidase activity	13/135	273/17632	1.80E-07	3.69E-05	3.38E-05	KLK14, ENDOU, KLK12, KLK6, MMP13, MMP1, KLK10, PRSS3, KLK9, KLK7, KLK5, MMP3, KLK8
GO:0017171	serine hydrolase activity	13/135	277/17632	2.13E-07	3.69E-05	3.38E-05	KLK14, ENDOU, KLK12, KLK6, MMP13, MMP1, KLK10, PRSS3, KLK9, KLK7, KLK5, MMP3, KLK8
GO:0004252	serine-type endopeptidase activity	12/135	250/17632	5.02E-07	5.81E-05	5.32E-05	KLK14, KLK12, KLK6, MMP13, MMP1, KLK10, PRSS3, KLK9, KLK7, KLK5, MMP3, KLK8
GO:0008106	alcohol dehydrogenase (NADP+) activity	5/135	23/17632	7.36E-07	6.38E-05	5.85E-05	AKR1C1, RDH12, AKR1C2, AKR1B10, AKR1C3
GO:0004033	aldo-keto reductase (NADP) activity	5/135	29/17632	2.50E-06	0.00017367	0.0001591	AKR1C1, RDH12, AKR1C2, AKR1B10, AKR1C3
GO:0004032	alditol:NADP+ 1-oxidoreductase activity	4/135	14/17632	3.10E-06	0.00017929	0.00016425	AKR1C1, AKR1C2, AKR1B10, AKR1C3
GO:0048018	receptor ligand activity	14/135	478/17632	1.83E-05	0.00079264	0.00072616	C1QTNF4, CSH2, GAST, ENDOU, FLRT3, PTHLH, WNT7A, CXCL5, IL36RN, IL1F10, INHBA, IL20, WISP3, IL24
GO:0043177	organic acid binding	8/135	204/17632	0.00017672	0.00681348	0.00624198	NOS2, NCAN, AKR1C1, AKR1C2, CYP4F11, HBE1, FOLR1, FOLR3
GO:0005125	cytokine activity	8/135	219/17632	0.00028606	0.00992636	0.00909377	C1QTNF4, WNT7A, CXCL5, IL36RN, IL1F10, INHBA, IL20, IL24
GO:0019838	growth factor binding	6/135	137/17632	0.0006524	0.02016026	0.01846928	IGFBPL1, COL2A1, IL36RN,

							FGFBP2, IGFBP6, WISP3
GO:0031406	carboxylic acid binding	7/135	192/17632	0.00069718	0.02016026	0.01846928	NOS2, NCAN, AKR1C1, AKR1C2, CYP4F11, FOLR1, FOLR3
GO:0004497	monooxygenase activity	5/135	98/17632	0.00093839	0.02486506	0.02277946	NOS2, AKR1C1, AKR1C2, CYP4F11, AKR1C3
GO:0016628	oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor	3/135	26/17632	0.0010032	0.02486506	0.02277946	AKR1C1, AKR1C2, AKR1C3
GO:0005520	insulin-like growth factor binding	3/135	28/17632	0.00124999	0.0289164	0.02649099	IGFBPL1, IGFBP6, WISP3
GO:0016229	steroid dehydrogenase activity	3/135	33/17632	0.00202451	0.04390654	0.0402238	AKR1C1, AKR1C2, AKR1C3
GO:0016616	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	5/135	121/17632	0.00238879	0.04848955	0.0444224	AKR1C1, RDH12, AKR1C2, AKR1B10, AKR1C3
GO:0005536	glucose binding	2/135	10/17632	0.00251531	0.04848955	0.0444224	GCK, HKDC1

Appendix

A1. Commercially available products used in this study

Product name	Manufacturer	Art. number
Cell culture		
Dulbecco's Modification of Eagle's Medium (DMEM)	Sigma®	D5546-1L
Ham's F-12 Nutrient Mix	Gibco™	11765070
Fetal Bovine Serum (FBS)	Sigma®	F0926
Penicillin-Streptomycin, 100x	Corning	MT30002CI
L-Glutamine, 200 mM	Gibco™	25030081
Trypsin/EDTA 0.25%	Gibco™	25200072
Dulbecco's Phosphate Buffered Saline (PBS)	Sigma®	D8662
G418 (Geneticin), 100 mg/ml	Gold Biotech	G-418-1
Puromycin, 1mg/ml	Sigma®	P8833
Keratinocyte-SFM (1x) kit including supplements: human EGF 1-53 and BPE	Gibco™	17005042
Western Blot		
ECL™ Prime Western Blotting Detection Reagent	GE Healthcare, UK	RPN2232
Trans-Blot® Turbo™ RTA Transfer Kit, PVDF	BIO-RAD	170-4273
Pierce™ BCA Protein Assay Kit	Thermo Scientific™	23225
MagicMark™ XP Western Protein Standard	Invitrogen™	LC5602
Precision Plus Protein™ Dual Color Standards	BIO-RAD	1610374
Restore™ Western Blot Stripping Buffer	Thermo Scientific™	21059
Phosphate Buffered Saline (PBS) 10x, pH 7.4	Quality Biological	119-069-101
Cell fixation		
DAPI solution 1 mg/ml	Invitrogen™	D1306
Shandon™ Immu-Mount™	Thermo Scientific™	9990402
Cloning		
pENTR™/D-TOPO® Cloning Kit, with One Shot™ TOP10 Chemically Competent *E. coli*	Invitrogen™	K240020
Gateway® LR Clonase™ II Enzyme Mix	Invitrogen™	11791-020
QIAprep® Spin miniprep Kit	QIAGEN, Germany	27106
HiSpeed® Plasmid Midi Kit	QIAGEN, Germany	12643
QIAquick® Gel Extraction Kit	QIAGEN, Germany	28706
NotI	New England Biolabs®	R0189L
NEBuffer™ 3.1	New England Biolabs®	B7203S
S.O.C Medium	Invitrogen™	15544034
DH5α chemically competent *E.coli* cells	own production	-
PCR		
Phusion® High-Fidelity DNA Polymerase	New England Biolabs®	M0530S
Phusion® HF buffer 5x	New England Biolabs®	M0530S
Phusion® GC buffer 5x	New England Biolabs®	M0530S
One*Taq*® DNA Polymerase	New England Biolabs®	M0480S
One*Taq*® GC buffer 5x	New England Biolabs®	M0480S
Gel Loading Dye, Purple (6X)	New England Biolabs®	B7024S
HyperLadder™ 1kb	Bioline	BIO-33053

RNA extraction and library preparation

RNeasy®mini Kit	QIAGEN, Germany	74160
NEBNext® UltraTM II Directional RNA Library Prep Kit for Illumina®	New England Biolabs®	E7765S
qPCR		
High-Capacity cDNA Reverse Transcription Kit	Applied Biosystems™	4368814
TURBO™ DNase (2 U/µL)	Invitrogen™	AM2238
FastStart Universal SYBR® Green Master (ROX)	Roche Diagnostics	04913914001

A2. Buffers and solutions used in this study

Name	Contents
Blocking buffer	5% non-fat dry milk powder in PBST
10x SDS-PAGE running buffer	250 mM Tris 1.92 M Glycine 1% SDS
RIPA buffer	10 mM Tris 150 mM NaCl 0.5 mM EDTA 1% Triton X-100 0.1% SDS 1% deoxycholate
PBST	1x PBS 0.1% Tween-20
SDS lysis buffer	4% SDS in dH$_2$O
LB medium	25% (w/v) LB Broth, Miller in milliQ H$_2$O (optional: 50 µg/ml Kanamycin)
3x SDS-PAGE loading buffer	150 mM Tris-HCl (pH 6.8) 300 mM DTT 6% SDS 0.3% bromophenol blue 30% glycerol
Tissue digestion buffer	100 mM NaCl 10 mM Tris, pH 8 25 mM EDTA, pH 8 0.5% SDS 0.2 mg/ml ProteinaseK (added right before use)
Lysis buffer – HNSCC cell lines	20 mM Tris-HCl, pH 7.4 150 mM NaCl 5 mM MgCl$_2$ 1 mM DTT 100 µg/ml cyclohexamide 1% Triton X-100 25U/ml Turbo DNAse

A3. List of primers

Name			**5' – 3'**
Sam68		forward	CACCATGCAGCGCCGGGACGAT
		reverse	TTAATAACGTCCATATGGATGCTCTCTGTATGCTCCC
DND1		forward	CACCATGCAGTCCAAGCGGGAT
		reverse	TCACTGTTTAACCATGGTACCTGCCT
HNRNPA1		forward	CACCATGTCTAAGTCAGAGTCT
		reverse	TTAAAATCTTCTGCCACT
RBFOX2		forward	CACCATGGAGAAAAAGAAAATG
		reverse	TTAGTAGGGGGCAAATCG
TRIM32		forward	CACCATGGCTGCAGCAGCAGCT
		reverse	CTATGGGGTGGAATATCTTCTC
TRIM71		forward	CACCATGGCTTCGTTCCCCGAG
		reverse	TTAGAAGACGAGGATTCGATTGTTGCC
genomic PCR primers			
FMR1		forward	CCAGTGAAGGTAGTCGGCTG
HSP90AA1		forward	ACCAGTGCTGCTGTAACTGA
METTL3		forward	ACCCTTGGAAACCAACTGGA
eGFP		forward	ATGGTGAGCAAGGGCGAGGAGC
		reverse	GTCTTGTAGTTGCCGTCGTC
qPCR primers			
Name	**RefSeq ID**		**5' – 3'**
IGFBPL1		forward	TCACGTGGAGAAAGGTCACG
		reverse	CCTTTCGCAGGGGGTTGAT
TCAM1P		forward	TGGTTTCCACTGCCCTTTTCT
		reverse	AGTCCTGTCCCTGGACTACA
MIR9-3HG		forward	AGAGCTCTCAGTAGGGCCTC
		reverse	GCCCCACAGCCAATTTGAAG
ABCA17P		forward	CCACTTTCTGGGGTGTTTGG
		reverse	TTGCTCTCGTTGGTCTTCGC
ADORA2A-AS1		forward	GCCCTGTGAAAGGACAAGCC
		reverse	CCAGGAGTGACTTCCTCTCCA
KLK5		forward	TGGGGGTCACAGAGCATGT
		reverse	ATCCATTGATGATGCGGCTG
AFAP1-AS1		forward	ATGGGGTAACTCAAAAAGCCTG
		reverse	TGGTTCATACCAGCCCTGTC
SFTA1P		forward	TATACAGCATTCCAGGTGGGC
		reverse	TGGTGAATGCCTTTCCCTTGT
IL36RN		forward	AGCTTCACCTTCTACCGGC
		reverse	GGCATTCCAGCCACCATTCT
WNT7A		forward	CTTGCACAACAACGAGGCAG
		reverse	TTGTCCTTGAGCACGTAGCC
GAPDH		forward	CAACTACATGGTTTACATGTTC
		reverse	GCCAGTGACTCCACGAC

A4. List of plasmids

Insert of interest	Plasmid of origin	Addgene ID
DND1	Inhouse plasmid (Dr. Wayne Miles)	-
TRIM32	peGFP-N1_hTRIM32	#69541
TRIM71	pMXS-hs-3xHA-LIN-41	#52717
Sam68	pcDNA3 HA Sam68 WT (mouse)	#17690
RBFOX2	peGFP rbFOX2	#63086
HNRNPA1	pET9d-hnRNP-A1	#23026
Entry vector	pENTR/D-TOPO	-
Destination vector	pEZY-hPGK-eGFP (own production)	-

A5. Antibodies and dilutions. All antibodies were diluted in blocking solution.

Product name	Manufacturer	Art. number	Lot	Dilution
GFP (D5.1) XP® Rabbit mAb	Cell Signaling	#2956S	4	1:1000
β-Actin (8H10D10) Mouse mAb	Cell Signaling	#3700S	17	1:4000
p16 mouse Ab	Gift from J.W. Rocco's lab	(own production)	-	1:100
ECL™ Rabbit IgG, HRP-linked whole Ab (donkey)	GE Healthcare	NA934V	16897770	1:1000
ECL™ Mouse IgG, HRP-linked whole Ab (sheep)	GE Healthcare	NA931V		1:2000