AMERICAN ACADEMY OF OTOLARYNGOLOGY-
HEAD AND NECK SURGERY FOUNDATION

ATTN: Stephanie Jones
Assistant Director, Research and Quality Improvement
1650 Diagonal Road
Alexandria, VA 22314-2857
sljones@entnet.org
PH: 703-535-3747

COVER SHEET FOR FINAL PROGRESS REPORT

Type of Grant:

_____AAFPRS Leslie Bernstein Grant
_____AAFPRS Leslie Bernstein Resident Research Grant
_____AAFPRS Leslie Bernstein Investigator Development Grant
_____AAOA Foundation/AAO-HNSF Combined Research Grant
_____AAO-HNSF Resident Research Award
_____AAO-HNSF Maureen Hannley Research Training Award
_____AAO-HNSF Percy Memorial Research Award
_____AAO-HNSF Health Services Research Grant
_____AAO-HNSF Rande H. Lazar Health Services Research Grant
__X__AHNS Pilot Grant
_____AHNS Alando J. Ballantyne Resident Research Pilot Grant
_____AHNS/AAO-HNSF Young Investigator Combined Award
_____AHNS/AAO-HNSF Surgeon Scientist Combined Award
_____AHRF Wiley H. Harrison Memorial Research Award
_____ALA/ALVRE Award
_____ANS/AAO-HNSF Herbert Silverstein Otology and Neurotology Research Award
_____ARS New Investigator Award
_____ARS Resident Research Grants
_____ASPO Research Grant
_____ASPO Daiichi Innovative Technology Grant
_____PSEF/AAO-HNSF Combined Grant
_____The Triological Career Development Award
_____XORAN Resident Research Grant

Start date: _07/01/2012_____ Stop date: _06/30/2013_______

Principal Investigator: __Daria A. Gaykalova______________________________

Institution: _Johns Hopkins University_______________________________

Title of Project: _Transcription Factor Signature of Head and Neck Squamous Cell Carcinoma_

Abstract:

Background: Head and Neck Squamous Cell Carcinoma (HNSCC) results in significant mortality and morbidity despite current therapeutic strategies. The molecular biology of HNSCC is related to abnormal transcriptional regulation. A direct comprehensive genome wide analysis of deregulation of key transcription factors (TF) in primary HNSCC has not been performed. In this study we sought to analyze the differences in transcription factor signatures in subtypes of cancer patients and the normal population.

Methods: To evaluate the TF signatures of 44 HNSCC samples and 25 healthy oral mucosa samples we used the Affymetrix GeneChip Human Exon 1.0 ST Array data and estimated transcription levels of all genes. The TF activity signature of each of 2,600 human transcription factors was characterized by the expression of its target genes, as reflected in TRANSFAC, and corrected for methylation and CNV status. The significance of each TF based on the expression levels of its targets was compared for HPV positive (HPV+) and HPV negative (HPV-) samples. The expression of the subset of target genes of NFKB, STATs and AP1 pathways was confirmed my qRT-PCR and co-activation of NFKB and STAT3 was confirmed by Immunistochemical (IHC) analysis.

Results: Of the top ranked TFs analyzed AP1, NFKB and STATs exhibited the greatest differences in TF activity in HPV+ and HPV- HNSCC tumor tissue. The changes of activity of these factors do not depend on DNA methylation or copy loss for their targets. We have confirmed coordinated activation of STAT3 and NFKB pathways in tumor samples, and showed that these pathways are the most activated in HPV-population of HNSCC patients. We have also discovered the top ten scoring genes, which expression was differentially affected by HPV infection. Expression analysis of these ten genes allows classification of HNSCC patients into three groups.

Conclusions: We have discovered that HPV+ and HPV- HNSCC differ significantly based on the level of activity of the key TF, such as AP1, STATs and NFKB. These data have implications for therapeutic targeting of tumors, as well as potential insight into biologic variability of behavior and treatment response for HPV+ and HPV- HNSCC patients.

Briefly describe progress in completing the project:

All aims and sub-aims are completed in full with minor adjustment of the project flow in agreement with the experimental set up and results.

Specific Aim 1:

a) We have assembled the validation cohort of 61 HNSCC and 28 control tissues, with clinical characteristics similar to the original discovery cohort.

b) By TSP analysis, we have discovered ten genes, which show the difference between HPV+ and HPV-samples. These genes include: CCND1, CEBPD, ICAM1, IFG1R, IL6ST, IRF1, JAG1, JAK3, NOS3 and SOCS3. Five out of ten genes belong to more than one pathway. Their expression was evaluated in the validation cohort.

c) We have performed ICH analysis on the 100 HNSCC and 13 non-cancer control samples. We have adopted total STAT3 and NFKB antibodies and performed independent scoring for total cellular staining (that reflects the protein expression) and nuclear staining (that reflects protein activation). STAT3 and NFKB demonstrated coordinated increase expression in HNSCC and especially in HPV- cancer samples.

d) The ten top-scoring genes, whose expression was evaluated in Aim 1b was used for heat map preparation and unsupervised hierarchical clustering of 61 HNSCC. The population of cancer patients was separated into three subgroups in by the differential expression of these genes.

Specific Aim 2:

a) All oropharyngeal HNSCC samples from the discovery and validation cohort have been tested by in situ hybridization (ISH) for high-risk HPV, p16 IHC staining and HPV16-specific qRT-PCR. HPV status for oropharyngeal HNSCC samples from ECOG-TMA and TCGA validation cohorts was evaluated only by HPV-ISH and p16-IHC.

b) HPV status was integrated into the heat map of the expression of ten top-scoring target genes and demonstrated that 16 out of 18 HPV+ were clustered together into the subgroup of total 25 HNSCC samples. Other 36 primarily HPV- HNSCC samples were separated into 15 and 21 sample groups.

What work was completed?

The proposed work in both aims was completed in full

What work was not completed?

The proposed work in both aims was completed in full

Were all of the funds spent? *If no, then the remaining funds will need to be return with a hard copy of the final financial report. If an AAO-HNSF grant, these can be sent to Stephanie Jones. If one of the sister society grants, contact Stephanie to obtain the name and address of the organization to whom the funds should be returned.*

Yes, the budget was spent in full, with just $8 left over.

Have the results been presented? Poster? Oral? What meeting? What publication?

The project was reported as a poster on the 8th International Conference on Head and Neck Cancer, Ontario, Canada in July 2012

The project is ready for submission as a manuscript described below

Clinical Applications, Either Immediate or Potential, of This Research:

The described different TF signatures within HNSCC population may explain the biological diversity of cancer patients in addition to HPV status. The evaluation of TF signatures may have direct clinical application: these signatures may be used to identify the key TFs and associated pathways for the development of the targeted anticancer therapy.

Utilization of the ten top-scoring genes demonstrated separation of HNSCC into three subgroups. This separation coordinated with HPV infection and discriminated HPV+ samples by overall lowered expression of target genes. HPV- patients, that in general have higher expression of these genes was classified into two groups, where one of them (25%, 15 out of 61 patients) had predominately higher gene expression for CCND1, CEBPD, ICAM1, IRF1, JAG1, JAK3 and NOS3. These data suggests that this subgroup of 25% patients would benefit from gene specific therapy against genes listed above.

Other Pertinent Information:

With collaborative efforts of investigators listed below, project utilized four independent HNSCC cohorts and a diverse set of biochemical approaches. The majority of experiments and experimental set up was performed by the project director, Daria Gaykalova, who received additional material support and help with statistical analysis from the collaborative investigators. The project was performed in the laboratory of Dr. Joseph Califano.

The project was completed with additional financial support from NIDCR/NIH Challenge Grant RC1DE020324 (Dr. Joseph Califano), NIDCR/NCI P50DE019032 Head and Neck Cancer SPORE (Dr. Joseph Califano), and NIDCR/NIH R01 DE013152 (Dr. Wayne Koch).

TITLE: Transcription Factor Signature of Head and Neck Squamous Cell Carcinoma
SHORT TITLE: NFKB and STAT3 pathways dysregulation in HNSCC

KEY WORDS: head and neck squamous cell carcinoma, HPV, Transcription Factor, Expression, high throughput

Daria A. Gaykalova1, Judi Manola2, Hiroyuki Ozawa3, Kathryn Morton1, Justin Bishop4, Michael Considine4, Rajni Sharma5, Chi Zhang1,6, Marietta Tan1, Elana Fertig4, Patrick Hennessey1, Julie Ahn1, Wayne Koch1, William Westra1,5, Zubair Khan1, Christine H Chung3, Michael F. Ochs4, Joseph A. Califano1,7

AUTHOR AFFILIATION:
1Department of Otolaryngology—Head and Neck Surgery, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA
2Dana-Farber Cancer Institute, Boston, MI
3Department of Surgery and Oncology, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA
4Division of Oncology Biostatistics, Department of Oncology, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA
5Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA
6University of Virginia, Charlottesville, VA, USA
7Milton J. Dance Head and Neck Center, Greater Baltimore Medical Center, Baltimore, Maryland, USA

GRANT SUPPORT:
The work is supported by American Head and Neck Society (AHNS) 306481 (DAG), NIDCR/NIH Challenge Grant RC1DE020324 (JAC), NIDCR/NCI P50DE019032 Head and Neck Cancer SPORE (JAC), and NIDCR/NIH R01 DE013152 (WK).

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is the fifth most common cancer worldwide, affecting an estimated 50,000 individuals in the United States and 500,000 worldwide annually (Jemal, Bray et al. 2011). HNSCC is responsible for 90% of all head and neck malignancies and is associated with tobacco use, alcohol exposure, as well as with high-risk human papilloma virus (HPV) infections (Psyrri, Boutati et al. 2011). Conventional treatment modalities, including surgery, radiotherapy and cytotoxic therapy, only modestly increase the lifespan of HNSCC patients, and five-year survival remains approximately 50% (Howlader N, Noone AM et al. 2010; Jemal, Bray et al. 2011). Both genetic and epigenetic aberrations have been shown to play a role in HNSCC development (Scully, Field et al. 2000; Ha and Califano 2006; Smith, Mydlarz et al. 2007). Current research is largely focused on genetic and epigenetic alterations resulting in the downregulation of tumor suppressor genes, such as p53, Rb and p16, and the upregulation of oncogenes, such as EGFR (Hardisson 2003). The development of prospective target-specific therapies for HNSCC has thus far been limited mostly to these genes. To date, the only targeted biological agent approved by the FDA for the treatment of HNSCC is Cetuximab, a monoclonal antibody directed against EGFR (Bonner, Harari et al. 2006; Bonner, Harari et al. 2010). The development of additional targeted therapies is urgently needed.

Interestingly, genome-wide expression array data demonstrate that many genes are differentially expressed during HNSCC carcinogenesis (Chung, Parker et al. 2004). From a biological point of view, the primary determinants in the regulation of gene expression are transcription factors (TFs). Unfortunately, direct comprehensive analysis of the dysregulation of key TFs is not feasible with existing screening techniques for several reasons: it is difficult to analyze changes in TF expression because of their low baseline expression in normal cells; in addition, the activation of most TFs requires post-translational modifications, protein cleavage, or protein translocation from cytoplasm to nucleus. Thus changes can be challenging to detect by current high throughput platforms

In this work we have developed a novel technique to define TF activity and function by expression of the target genes of a TF, acquired from the Transcription Factor Database. We have removed transcription factors with less than 5 target genes. We have also decreased the number of annotated targets for each TF. Thus, we have removed targets that were not experimentally validated, or the expression was decreased due to DNA methylation or decrease of copy number.

Recent clinical data demonstrate that HPV+ and HPV- HNSCC are different in genetics, epigenetics, and cancer etiology; as such, they require different diagnostics and treatment. HPV affects approximately 70% of lingual and palatine tonsils of the oropharynx (Hennessey, Westra et al. 2009). The infection leads to carcinogenesis often associated with more favorable prognosis (Weinberger, Yu et al. 2006; D'Souza, Kreimer et al. 2007; Ang, Harris et al. 2010; Mydlarz, Hennessey et al. 2010; Psyrri, Boutati et al. 2011). In order to investigate the biological difference between those HNSCC subgroups and to discover new targets for gene and therapies, we have focused our attention on the transcription factors that demonstrate the difference between HPV+ and HPV- HNSCC patients. We have demonstrated that most dysregulated factors differentiate those groups involved in NFkappaB, STATs, p53, AP1 and retinoid acid signal transduction cascades.

Dysregulation of NFKB, STAT, p53 and AP1 pathways in HNSCC have previously been demonstrated by other research groups. For example, AP-1 and NFKB have been shown to be constitutively active in HNSCC cell lines that express IL-8 (Ondrey, Dong et al. 1999). STAT proteins, particularly STAT3, have been found to be frequently upregulated in many human cancers, including head and neck (Song and Grandis 2000). Conversely, multiple mutations in the *TP53* gene have been described to downregulate the p53 pathway in HNSCC (Scully, Field et al. 2000).

This work demonstrates the coordinated dysregulation of NFKB and STAT3 pathways in HNSCC. We have shown the different levels of gene expression for their targets in HPV- and HPV+ patients, as well as the difference in nuclear staining for the transcription factors. We have confirmed the correlation of HPV infection with dysregulation of NFKB and STAT3. We have also discovered the panel of genes that separated HPV+ and HPV- HNSCC patients by TF signatures.

METHODS

Tissue samples. We used three independent cohorts of HNSCC patient specimens and normal control specimens. The discovery cohort was composed of 44 primary HNSCC tissues and 25 normal mucosal samples from uvulopalatopharyngoplasty (UPPP) surgeries of non cancer affected control patients. The validation cohort was composed of 61 primary HNSCC tissues and 28 normal UPPP samples. Our study also involves Tissue Microarray (TMA) with primary cancer tissues from 100 HNSCC and normal uvula or tonsil tissues from 13 non-cancer patients from the collaborative Eastern Cooperative Oncology Group (ECOG) and Radiation Therapy Oncology Group (RTOG) (study no. ECOG E4393/RTOG 9614). All samples were obtained from the Head and Neck Tissue Bank at Johns Hopkins, acquired under the Internal Review Board-approved research protocols. TMA tissues were also approved by ECOG and RTOG protocols. All primary tissues were stored at -140°C (liquid nitrogen) until use. All cancer samples were analyzed by investigators from the Pathology Department of Johns Hopkins Hospital (WW and JB). Tumor samples were confirmed to be HNSCC and subsequently microdissected to yield at least 70% tumor purity. The clinical characteristics of the discovery and validation cohorts are listed in Tables 1 and S1. The clinical and demographic data for TMA patients was collected and managed by ECOG (Table S2). We have also used publicly available data for the TCGA HNSCC cohort that includes 279 HNSCC (including 244 HPV- and 35 HPV+ cases) and 50 non-cancer controls.

DNA preparation. Microdissected tissue samples were digested in 1% SDS (Sigma) and 50 µg/ml proteinase K (Invitrogen) solution at 48°C for 48-72 hours for removal of proteins bound to DNA. DNA was then purified by phenol-chloroform extraction and ethanol precipitation as previously described (Shao, Tan et al. 2012). DNA was resuspended in LoTE buffer (EDTA 2.5 mM and Tris-HCl 10 mM, pH 7.5), and DNA concentration was quantified using the NanoDrop ND-1000 spectrophotometer (Thermo Scientific).

RNA preparation. RNA was isolated from the microdissected tissue samples with the mirVana miRNA Isolation Kit (Ambion) per manufacturer's recommendations, and RNA concentration was quantified using the NanoDrop.

Arrays. Two micrograms of RNA and DNA from the samples of the discovery cohort were submitted to the Johns Hopkins Core Facility for quality control query and analysis by high throughput arrays. Samples were run on Affymetrix HuEx1.0 GeneChips for expression analysis (with over 1.4 million probe coverage), Illumina Infinium HumanMethylation27 BeadChips for methylation analysis (28,000 probe coverage) following bisulfite conversion, and Affymetrix Genome-wide SNP 6.0 Array (950,000 probe coverage). All arrays were run according to manufacturer protocols.

HPV analysis. We obtained pathology reports regarding the HPV status of oropharyngeal HNSCC tumors that had been tested clinically by *in situ* hybridization (ISH) for high-risk HPV and by p16 immunohistochemical staining (Singhi, Califano et al. 2012). In addition, the HPV status of all oropharyngeal HNSCC primary tissues was independently confirmed by quantitative PCR (qPCR) using HPV16 primers and probe on the 7900HT real-time PCR machine as described (Carvalho, Henrique et al. 2011). Briefly, we used specific primers and probes to amplify the E6 and E7 regions of HPV16 and normalized the data to a housekeeping gene (*β-actin*). Genomic DNA from the CaSki cell line (American Type Culture Collection, ATCC, Manassas, VA), known to have 600 copies of HPV16 per genome (6.6 pg of DNA/genome), was used in serial dilutions (50-0.005 ng) to construct a calibration curve for *β-actin* and HPV 16 *E6* and *E7* for each plate. The relative level of HPV16 DNA in each sample was determined as a mean of ratios of *E6* and *E7* amplified gene to *β-actin*, multiplied by 300, that gave number of copies per genome per tumor cell. HPV copy number ≥ 1 copy/genome/cell was regarded as HPV positive.

Reverse Transcription (RT) and quantitative Real Time PCR (qRT-PCR). One µg of RNA from the validation cohort was reverse transcribed using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Carlsbad, CA, USA). Quantitative real-time PCR was performed using gene-specific expression assays and Universal PCR Master Mix on the 7900HT real time PCR machine (all from Applied Biosystems). PCR conditions were 1 cycle: 95°C for 10 min; followed by 40 cycles: 95°C for 15 s and 60°C for 60 s. Expression of the gene of interest was quantified in triplicates relative to *GAPDH* expression using the 2-ΔΔCT method (Livak and Schmittgen 2001).

Immunohistochemistry. 100 HNSCC and 13 non-cancer formalin-fixed and paraffin-embedded samples were obtained from the Head and Neck Tissue Bank at Johns Hopkins and were used to construct a tissue microarray under ECOG approved protocols. The 5 um cuts were used for slide preparation. Immunostaining was carried out on Bond-Leica autostaining system (Leica Microsystems) using standard immunohistochemistry (IHC) protocol. IHC protocol incorporated heat-induced antigen retrieval with citrate buffer (pH 6.0) followed by peroxide-blocking step and primary antibody incubation for 15 minutes with rabbit monoclonal antibody against total NF-κB p65 (Cell Signaling, #8242, dilution 1:400) or against total Stat3 (Cell Signaling, #4904, 1:200 dilution). Reaction was developed with biotin-free Bond polymer detection system (Leica Microsystems). 3,3′-Diaminobenzidine (DAB) chromogen substrate for used for visualization of reaction. Slides were counterstained with hematoxylin, dehydrated, and cover slipped. Slides were scanned with 20x resolution. The cancer tissue or normal mucosa tissue were annotated by Aperio software. Whole cell and nuclear staining were quantified for each individual tissue and averaged for tissue triplicates.

p53 mutation analysis. Mutation status of exons 2 to 11 of the *p53* gene was evaluated using the GeneChip *p53* assay (Affymetrix) as previously described (Ahrendt, Halachmi et al. 1999; Westra, Taube et al. 2008). All mutations detected by GeneChip *p53* assay analysis were identified and confirmed by automatic (ABI BigDye cycle sequencing kit) or direct dideoxynucleotide sequencing (Ahrendt, Halachmi et al. 1999). Based on available information about the functional differences of various *p53* mutations, *p53* mutations were grouped as “disruptive” and “nondisruptive.” Disruptive mutations were defined as stop mutations, frameshift mutations, or nonconservative mutations occurring within the key DNA binding domain L2/L3. All other mutations were defined as nondisruptive mutations (Poeta, Manola et al. 2007).

Statistical analysis.

Preparation of TF target gene sets. We applied RMA analysis to an expression array dataset for samples from the discovery cohort (Carvalho and Irizarry 2010). We then annotated each TF with a list of its high probability experimentally validated targets as described in the Transcription Factor Database (TRANSFAC Professional (Matys, Fricke et al. 2003)). We have removed TFs with less than 5 targets. Overall 1325 TF target gene sets were created from 2,600 human (out of total 14,000) TF described in TRANSFAC. We simultaneously obtained promoter methylation and copy number variation (CNV) measurements on the samples using Illumina HumanMethylation27k and Affymetrix SNP6 chips respectively, with processing done in R by reading in the Illumina FinalReport files and using CRLMM respectively (ref - CRLMM). From each gene set, we removed genes that were expected to have significantly reduced expression either due to increased methylation ($\beta > 0.15$) or copy loss ($CNV < 1.2$) on a per tumor sample basis, creating tumor-specific TF gene sets. We identified these genes by integrating data from the Methylation and SNP arrays. Resulting TF target gene sets were then used to compare samples from HPV+, HPV- or non-cancer samples for each individual target gene by a conservative Wilcoxon test. Averaged p-values of all targets for each TF were used to compare different TFs**.**

Statistical significance analysis (p-values). P-values in this work were determined by the t-test (continuous variables) and the Fischer Exact test (dual variables). Wilcoxon gene set test and Wilcoxon rank sum test were used to compare group pairs with respect to marker/expression levels. Kruskal-Wallis tests were used to test for differences when there were more than two groups being compared (gene sets analysis). P-values of less than 0.05 were considered to indicate statistical significance. Markers correlation was evaluated by logistic regression, where odd ratios and 95% confidence interval (CI) were calculated.

RESULTS

Cohort assembly and array analysis. It is challenging to directly detect the changes in activity of transcription factors (TF) by high throughput platforms due to the complexity of protein activation via protein cleavage, covalent modifications and protein translocations. We have decided to investigate the expression of TF targets, as representation of TF activity. We have employed the modern high throughput platform for gene expression analysis - Affymetrix HuEx1.0 GeneChips with over 40 probes per gene to be used for 44 tumor and 25 normal samples from the discovery cohort (Table 1).

The characteristics of the HNSCC population from our discovery cohort largely reflect the demographics of head and neck cancer patients in the United States. The HNSCC patients were largely male (73%, 32 of 44) and Caucasian (91%, 40 of 44), aged 45 to 80 years (median ± SD = 58 ± 13 years). There was a history of tobacco and alcohol consumption in 61% (27 of 44) and 57% (25 of 44) of all patients, respectively, with average smoking history of 39.7 ± 30.3 pack-years. With regard to HPV status, the study population consisted of 30% (13 of 44) HPV-positive patients. Primary tumors were located in the oral cavity (23%, 10 of 44), oropharynx (39%, 17 of 44), hypopharynx (9%, 4 of 44), or larynx (30%, 13 of 44). Thirty-two of 44 patients (73%) presented with locally advanced stage IV disease. The median follow-up time of these patients was 31.4 months (range: 0.5–117.3 months). At the end of the follow-up period, 10 patients were alive with disease. During the follow-up period, 14 (32%) recurrences were detected, including 8 local recurrences. As of July 2013, a total of 21 patients (48%) have died. The cause of death was head and neck cancer in 18 out of 21 patients; the other three patients died of unrelated causes.

The control population was largely female (64%, 16 of 25) and Caucasian (56%, 14 of 25), aged 18 to 65 years (29 ± 12 years). There was a history of smoking and alcohol consumption in 12% (3 of 25) and 36% (9 of 25) of all patients, respectively, with average smoking history of 29.0 ± 29.7 pack-years.

Array data annotation and preparation of TF target gene sets. The information regarding the target genes for each TF was found in TRANSFAC Professional database (Matys, Fricke et al. 2003). TRANSFAC contains information on 14,000 known TFs with 30,000 experimentally-evaluated DNA transcription factor binding sites (TFBS) on 70,000 target genes for multiple organisms. This database contains both predicted regulation through TFBS and high quality data generated by direct experiments on gene regulation determined from knock-in or knock-down of TFs. In order to utilize only high quality annotations, we retained only experimentally validated targets and retained only TFs with a minimum of 5 targets. This provided data to create sets of target genes for 1325 human TFs (out of total 2,600 human TF described in TRANSFAC). We then removed genes from each gene set that had a possibility of reduced expression regardless of TF activity. These genes were silenced by epigenetic mechanisms such as increased methylation ($\beta > 0.15$) or genetically by loss of homo- or heterozygosity (Copy Number Variation (CNV) < 1.2). These genes were identified by Methylation and SNP arrays used for the same discovery cohort of samples.

HPV positive and HPV negative patients are different by their TF signature. Methylation and CNV-corrected TF target gene sets were then analyzed, where the expression of each target gene was compared for HPV+ and HPV- HNSCC patients. The gene was considered downregulated or upregulated if its expression in two analyzed groups was significantly different by a t-test or by a conservative Wilcoxon test. The calculated p-values of target genes were averaged for each TF. While individual genes could be either upregulated or downregulated in the particular sample group, the TF activity would be considered dysregulation, assuming the presence of both upregulated and downregulated targets in its list. We then isolated the fifty TFs with the lowest averaged TF p-values for further analysis. Of note, due to the granularity of the TRANSFAC database, the majority of the fifty TFs could be assigned to a limited number of key TF pathways, such as NFKB ((RelA−p65)2, NF−kappaB1−p50, NF−kappaB1−p50:RelA-p65), AP-1 (c−Fos:c−Jun, c−Jun:c−Fos. Fra-1, ATF, NF-AT1), retinoid acid signaling (PPARalpha:RXR−alpha, RXR−alpha:PPARalpha, RXR−alpha:PPARgamma, LXR-alpha:RXR−alpha, RORalpha1, RXR−alpha1), and STATs (STAT1, STAT1alpha, STAT5, STAT6, STAT1:STAT1 and STAT3:STAT3) pathways (Table S3).

This data suggests that HPV infection plays a role in the dysregulation of several TF including NFKB and STATs. These results support the previous reports regarding dysregulation of several TF pathways in oral and cervix cancer types (Howley, Munger et al. 1989; Werness, Levine et al. 1990; Scully, Field et al. 2000; Arany, Grattendick et al. 2002; Mishra, Bharti et al. 2006; Manavi, Hudelist et al. 2007; Rampias, Sasaki et al. 2009).

To validate our discovery of TFs, where STATs and NFKB are differentially regulated in HPV+ and HPV- patients (samples?), we adopted expression array data available for the TCGA HNSCC cohort. The TCGA cohort for HNSCC includes 279 tumor samples, including 35 HPV+ and 244 HPV- samples. The expression array data (acquired by Affymetrix expression platform) for these samples are publically available as of July 2013. Using the available expression data on the TF target gene sets lists that were defined earlier for the discovery cohort,we have compared the differential expression of the target genes in HPV- and HPV+ HNSCC patients. Averaged p-values were calculated as described above for our original discovery cohort. The list of top 50 TFs separating HPV+ and HPV- samples is shown in Table S4. We have found that 62% (31 of the 50 TF were similar or identical to the top 50 TF from discovery cohort, including STATs, NFKB, AP1, p53, retinoid acid and other pathways (compare Tables S3 and S4)

HPV-relative TF signatures do not depend on DNA methylation and CNV data of TF target genes. We have originally removed the target genes from the TF gene sets, whose expression is potentially silenced by DNA methylation and chromosomal deletions. Even though our original cut-off of DNA methylation ($\beta > 0.15$) copy number (CNV < 1.2) was biologically relevant, we wanted to make sure that such cut-off setting does not discriminate against important TFs and their targets. In order to discover the most robust TFs which do not depend on DNA methylation and chromosomal deletions, we have repeated p-value calculation for additional cut-off values, such as $\beta > 0.25$, $\beta > 0.35$ and $\beta > 0.45$ for DNA methylation) and (CNV <1, CNV < 0.8, and CNV < 0). We have also used control conditions of $\beta > 1$ and CNV < 0 to avoid discrimination against any genes. The p-values for all conditions excluding non-discriminative $\beta > 1$ and CNV < 0 condition were averaged for each TF and ranked from smallest to largest. We have separated the top 27 TFs that are significantly dysregulated in HPV- and HPV+ samples per the differential gene expression of their targets. The p-values of the individual conditions for each selected TF were used to build a heat-map (Figure 2). Of note, the majority of the TFs was found in the top 50 TF lists for either discovery or TCGA cohort, with prevalence of TFs involved in STATs, NFKB, AP1 and retinoid acid pathways. We have noticed that previously observed p53 factor was not included into the final list of robustly dysregulated TFs. P53 pathway is known to be affected in HNSCC due to multiple mutation of TP53 genes and other players of p53 pathway (Poeta, Manola et al. 2007; Poeta, Manola et al. 2009) and such changes must be affected by copy number variation.

Target genes that drive the separation of HPV+ and HPV- patients. We have discovered that several TF cascades, including NFKB, STATs and AP1 were strongly affected by HPV infections, leading to differential regulation of these pathways among HPV+ and HPV- HNSCC patients. We have also demonstrated that these pathways are doubly affected by DNA methylation or DNA copy number variations. Next we wanted to depict the target genes which reflect the dysregulation of these pathways and can distinguish HPV+ and HPV- HNSCC by the gene expression. A shorter list of targets genes will also help us validate TF signatures during wet lab, and can be further adopted for clinical practice to identify patients with dysregulated pathways. Identification of dysregulated genes and pathways in HNSCC patients will also enhance the development gene specific therapies. In order to narrow down the number of most affected target genes, we have combined the DNA methylation and CNV-corrected lists of all targets for highly dysregulated NFKB, STAT1, STAT3 and AP1 factors (Table S5) and applied top scoring pair (TSP) analysis to this list. Top scoring pair (TSP) available in R script is the bioinformatics tool that allows defining pairs of genes, which can then be used for classification of patients into two groups, HPV+ and HPV- in our case (Geman, d'Avignon et al. 2004; Leek 2009). TSP is one of the most robust and easily interpretable tests to depict separation-driving genes. We have applied this technique to the expression of 72 combined target genes to the discovery cohort expression array data. We aimed for separation of HPV- and HPV- patients by 5 independent pairs of genes. TSP allowed the discovery of 10 genes among targets of mostly dysregulated STATs, NFKB and AP1 pathways (Table 2). Of note, 5 out of 10 depict genes belong to more than one dysregulated pathway. These 10 genes together could distinguish HPV+ and HPV- patients in the discovery cohort with p-value of 9.5 E-06 (Table 3). Validation of these genes on TCGA cohort demonstrated that they discriminate HPV patients with p-value of 6.7 E-08 (Table 3). The high significance of depicted genes can be explain by the fact that many of these genes can individually discriminate HPV+ samples from HPV- as well as tumor samples from the normal samples (Table 2).

Validation of the ten top scoring genes by quantitative real time PCR (qRT-PCR) technique. In order to validate our results and evaluate the discriminative ability of the top ten scoring genes, we have assembled an independent cohort of 61 HNSCC (including 43 HPV- and 18 HPV+ samples) and 28 control UPPP samples. The expression of the top ten was analyzed by qRT-PCR, as described in methods. The expression of each gene was normalized to the expression of the house-keeping *GAPDH* gene. The combination of ten genes separated HPV- and HPV+ patients with a p-value of 0.0006 (Table 3). Similar to the result for the discovery and TCGA cohorts, the majority of analyzed genes were differentially expressed in HPV+ and HPV- HNSCC samples as well as in tumor and normal samples (Table 2).

Logarithm-converted values of relative expression for the ten top genes were used for heatmap preparation in R script (Figure 3). The validation cohort was separated into three groups by unsupervised hierarchical clustering. We have found that 25 out of 61 samples were clustered into separate group that was enriched by HPV+ patients. That group contains 16 out of 18 HPV+ patients and is distinct from two other primarily HPV- groups by downregulation of CCND1, and partial downregulation of IRF1, ICAM1, IGF1R and NOS3. These data suggest that STATs, NFKB and AP1 pathways are partially donwregulated in HPV+ HNSCC patients as compared to HPV- patients. Two other groups of primarily HPV- patients can be distinguished by relative expression of several genes. Thus, one group has higher expression of CCND1, CEBPD, ICAM1, IRF1, JAG1, JAK3 and NOS3. These data suggest that dysregulation of STATs, AP1 and NFKB pathways was coordinated and that these pathway might be co-activated for the subset of HPV- patients (25%, 15 out of total 61 samples).

Co-activation of NFKB and STAT3 proteins in HNSCC. In order to investigate the coordinated dysregulation and to evaluate the direction of this dysregulation (either upregulation of downregulation) for some of TF, we performed immunohistochemical experiments on HNSCC tissue microarray (TMA). The TMA cohort includes 100 HNSCC (including 87 HPV- and 13 HPV+) and 13 control samples (Table S2).

We have performed analysis of the p53 mutation, HPV infection, as well as STAT3 and NFKB protein expression for these samples. The protein stating was scored by Aperio software, where staining was independently quantified for the whole cell (that reflects overall protein expression) or for the nuclei (that reflects protein activation and translocation to nuclei) (Table S6). We have also investigated coordinated dysregulation of NFKB and STAT3 by quantification of STAT3 and NFKB staining in nuclei. The co-staining of both markers was evaluated by linear regression, where NFKB staining values were transformed using natural log to provide a more normal distribution. Through this, we have discovered strong co-activation of NFKB and STAT3 in cancer patients (p-value 0.003) (Figure 4).

Correlation of NFKB and STAT3 staining with clinical characteristics. Quantification of protein staining demonstrated a strong increase of NFKB and STAT3 protein expression as well as activation of NFKB in cancer patients (Figure 5). HPV+ cancer patients have lowered protein expression and significant inactivation of both TFs, supporting the data that several top scoring target genes have decreased expression in HPV+ subgroup. NFKB activation was higher in patients with disruptive p53 mutations, while STAT3 activation was decreased in those patients. Such data correlate with originally found correlation of dysregulation of STAT3, NFKB and p53 pathways.

DISCUSSION

HNSCC includes two diverse populations of cancer patients: with and without HPV infection, which is responsible for 30% of HNSCC cases. HPV- patients are elder patients often with smoking history (Weinberger, Yu et al. 2006; D'Souza, Kreimer et al. 2007; Hennessey, Westra et al. 2009; Ang, Harris et al. 2010; Mydlarz, Hennessey et al. 2010; Psyrri, Boutati et al. 2011).. They are characterized by higher genetic and epigenetic alteration (Scully, Field et al. 2000; Ha and Califano 2006; Smith, Mydlarz et al. 2007). In this project we intended to elaborate on our preliminary results, that HPV- and HPV+ patients can be further discriminated by TF activity.

TF are the main drivers of gene expression variation, however the changes in TF activity is hard to detect due to the complexity of TF activation procedures. Using three high throughput platforms and TRANSFAC database we were able to annotate highly specific target genes to each known TF and evaluate the TF activity by the expression of those target genes.

We have investigated the difference between HPV+ and HPV- HNSCC patients by the TF signatures. We have found that HPV- patients have strong overexpression of several pathways, including NFKB and STAT3. We have also discovered that many more pathways were differentially altered in HPV+ and HPV- patients and demonstrated that these pathways did not depend on genetic and epigenetic alterations.

We have focused our attention on STATs, NFKB and AP1 pathways. While different reported described changes in the activity of the individual factors, like STAT1, STAT3, NFKB, and p53, in HPV+ patients of HNSCC or cervix cancer. Our experiments demonstrated that HNSCC patients have coordinated dysregylation (both upregulation and downregulation) of all these pathways. We have demonstrated that HPV is the main causative of TF dysregulation and discovered the panel of the ten top-scoring genes that are the most affected by HPV infection.

Interestingly, that p53 was previously demonstrated as the most affected gene in HNSCC (Poeta, Manola et al. 2007; 2008; Poeta, Manola et al. 2009). We have shown that activity of this pathway was also affected by either DNA methylation or copy number variation of its target genes.

Utilization of ten top-scoring genes demonstrated separation of HNSCC on three subgroups. This separation coordinated with HPV infection and discriminated HPV+ samples by overall lowered expression of several target genes. Interestingly, HPV- patients, that in general had higher expression of the majority of these genes was classified into two groups, where one of them has predominately higher gene expression of CCND1, CEBPD, ICAM1, IRF1, JAG1, JAK3 and NOS3. These data suggest that this subgroup of 25% patients (15 out of 61) would benefit from gene specific therapy against genes listed above.

Overall, the described different TF signatures may explain the biological diversity of HNSCC patients in addition to HPV status. The evaluation of TF signatures may have direct clinical application: these signatures may be used to identify key TFs and associated pathways for the development of targeted anticancer therapy.

FIGURE LEGENDS

Figure 1. Experimental set up for the TF target gene sets preparation and the analysis. RMA-normalized exon expression assay data was used in the analysis. 2,600 human TF were depicted from 14,000 TF list of TRANSFAC database. The list of TF was reduced by removal of TF with less than 5 experimentally validated targets. For the remaining 1325 TFs, the target genes from the gens sets were removed if their expression was silenced by hypermethylation or copy loss. Resulting TF target gene sets were then used to compare samples from HPV+, HPV- or non-cancer samples for each individual target gene by a conservative Wilcoxon test. Averaged p-values of all targets for each TF were used to compare different TFs**.**

Figure 2. HPV-relative TF signatures do not depend on DNA methylation and DNA copy loss. We have performed p-value evaluation on each TF as described in Figure 1. Total 17 conditions of β-value and CNV cut-off were used, as a combination of $\beta > 0$, $\beta > 0.15$, $\beta > 0.25$, $\beta > 0.35$, and $\beta > 0.45$ for DNA methylation) and CNV < 1.2, CNV <1, CNV < 0.8, and CNV < 0 for DNA copy loss. P-values for 16 conditions (excluding control non-discriminative $\beta > 1$ and CNV < 0 condition were averaged for each of 1325 TF. The top-scoring 27 TFs with the lowest averaged p-value were used in heatmap preparation.

Figure 3. Separation of HNSCC by the expression of the ten top-scoring genes. The expression of the top ten target genes discovered by TSP was analyzed by qRT-PCR on 61 HNSCC and 28 UPPP samples from the validation cohort. Logarithm-converted values of GAPDH-relative expression for these genes were used for heatmap preparation in R script. Tumor samples were separated on three groups by unsupervised hierarchical clustering. The middle group of 25 HNSCC contains 16 out of total 18 HPV+ patients

Figure 4. IHC co-staining of NFKB and STAT3. TMA containing 100 HNSCC and 13 control tissues was stained by total NFKB and STAT3 staining. Representative HNSCC samples of NFKB and STAT3co-staining are shown. A. HPV+ samples. B. HPV- samples. Samples are arranged from stronger staining and better co-staining on the left to no staining on the right. Both markers co-stain in cellular nuclei of cancer tissues with p-value 0.003.

Figure 5. ICH staining quantification for different HNSCC subgroups. The level of antibody staining was quantified by the Aperio software in the whole cell (total protein expression) or in the nuclei (protein activation) for both NFKB and STAT3. The staining intensity was averaged for three tissue replicates. A. Association of protein level with cancer. B. Association of protein level with HPV infection. C. Association of protein level with p53 mutations. * represent significant difference (p-value < 0.05 or p-value < 0.025 for association with p53 mutation status).

Table 1. Clinical characteristics of HNSCC patients in the initial discovery cohort

Table 2. The top ten scoring target genes. P-values were quantified by t-test. ND - non-determined

Table 3. The comparison of the top ten target genes expression in HPV- and HPV+ patients from different cohorts. P-values were quantified by Fisher Exact Test

Table 4. NFKB and STAT3 staining correlation. P-values were calculated by Wilcoxon rank sum test

Figure S1. The top scoring pair analysis for the discovery cohort. TSP analysis was applied to the expression of 72 combined target genes (Table S5) expression array data. We aimed for separation of HPV+ and HPV- patients by 5 independent pairs of genes.

Table S1. Clinical characteristics of HNSCC patients in the validation cohort

Table S2. Clinical characteristics of HNSCC patients in the ECOG-TMA cohort

Table S3. The top 50 TF differently dysregulated in HPV+ and HPV- HNSCC patients from the discovery cohort. The p-values were calculated using a conservative Wilcoxon test on each TF target gene set comparing differential expression of each TF target in HPV+ vs HPV- tumor samples. The lowest p-value indicates a strong differential expression of TF targets in HPV+ as compared to HPV- tumor samples.

Table S4. Top 50 TF differently dysregulated in HPV+ and HPV- HNSCC patients from the TCGA cohort The p-values were calculated using a conservative Wilcoxon test on each TF target gene set comparing differential expression of each TF target in HPV+ vs HPV- tumor samples. The lowest p-value indicates strong differential expression of TF targets in HPV+ as compared to HPV- tumor samples.

Table S5. The list of target genes for STAT1, STAT3, NFKB and AP1 pathways

Table S6. Quantification of IHC staining for STAT3 and NFKB in whole cell or in nuclei. Staining correlation with HPV or p53 mutation status. Color code: Yellow represents 50-percentile, green - lowest values; red - highest value. Top 100 samples belong to tumor patients. Bottom 13 - non-cancerous patients. Samples were sorted by total staining

BIBLIOGRAPHY

(2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." Nature **455**(7216): 1061-1068.

Ahrendt, S. A., S. Halachmi, et al. (1999). "Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array." Proceedings of the National Academy of Sciences of the United States of America **96**(13): 7382-7387.

Ang, K. K., J. Harris, et al. (2010). "Human papillomavirus and survival of patients with oropharyngeal cancer." The New England journal of medicine **363**(1): 24-35.

Arany, I., K. G. Grattendick, et al. (2002). "Interleukin-10 induces transcription of the early promoter of human papillomavirus type 16 (HPV16) through the 5'-segment of the upstream regulatory region (URR)." Antiviral research **55**(2): 331-339.

Bonner, J. A., P. M. Harari, et al. (2006). "Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck." The New England journal of medicine **354**(6): 567-578.

Bonner, J. A., P. M. Harari, et al. (2010). "Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival." The lancet oncology **11**(1): 21-28.

Carvalho, A. L., R. Henrique, et al. (2011). "Detection of promoter hypermethylation in salivary rinses as a biomarker for head and neck squamous cell carcinoma surveillance." Clinical cancer research : an official journal of the American Association for Cancer Research **17**(14): 4782-4789.

Chung, C. H., J. S. Parker, et al. (2004). "Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression." Cancer cell **5**(5): 489-500.

D'Souza, G., A. R. Kreimer, et al. (2007). "Case-control study of human papillomavirus and oropharyngeal cancer." The New England journal of medicine **356**(19): 1944-1956.

Geman, D., C. d'Avignon, et al. (2004). "Classifying gene expression profiles from pairwise mRNA comparisons." Statistical applications in genetics and molecular biology **3**: Article19.

Ha, P. K. and J. A. Califano (2006). "Promoter methylation and inactivation of tumour-suppressor genes in oral squamous-cell carcinoma." The lancet oncology **7**(1): 77-82.

Hardisson, D. (2003). "Molecular pathogenesis of head and neck squamous cell carcinoma." European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies **260**(9): 502-508.

Hennessey, P. T., W. H. Westra, et al. (2009). "Human papillomavirus and head and neck squamous cell carcinoma: recent evidence and clinical implications." Journal of dental research **88**(4): 300-306.

Howlader N, Noone AM, et al. (2010). "SEER Cancer Statistics Review, 1975–2008, National Cancer Institute. Bethesda, MD." SEER **http://seer.cancer.gov/csr/1975_2008/**.

Howley, P. M., K. Munger, et al. (1989). "Molecular mechanisms of transformation by the human papillomaviruses." Princess Takamatsu symposia **20**: 199-206.

Jemal, A., F. Bray, et al. (2011). "Global cancer statistics." CA Cancer J Clin **61**(2): 69-90.

Leek, J. T. (2009). "The tspair package for finding top scoring pair classifiers in R." Bioinformatics **25**(9): 1203-1204.

Livak, K. J. and T. D. Schmittgen (2001). "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method." Methods **25**(4): 402-408.

Manavi, M., G. Hudelist, et al. (2007). "Gene profiling in Pap-cell smears of high-risk human papillomavirus-positive squamous cervical carcinoma." Gynecologic oncology **105**(2): 418-426.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res **31**(1): 374-378.

Mishra, A., A. C. Bharti, et al. (2006). "Differential expression and activation of NF-kappaB family proteins during oral carcinogenesis: Role of high risk human papillomavirus infection." International journal of cancer. Journal international du cancer **119**(12): 2840-2850.

Mydlarz, W. K., P. T. Hennessey, et al. (2010). "Advances and Perspectives in the Molecular Diagnosis of Head and Neck Cancer." Expert opinion on medical diagnostics **4**(1): 53-65.

Ondrey, F. G., G. Dong, et al. (1999). "Constitutive activation of transcription factors NF-(kappa)B, AP-1, and NF-IL6 in human head and neck squamous cell carcinoma cell lines that express pro-inflammatory and pro-angiogenic cytokines." Molecular carcinogenesis **26**(2): 119-129.

Poeta, M. L., J. Manola, et al. (2009). "The Ligamp TP53 Assay for Detection of Minimal Residual Disease in Head and Neck Squamous Cell Carcinoma Surgical Margins." Clinical cancer research : an official journal of the American Association for Cancer Research **15**(24): 7658-7665.

Poeta, M. L., J. Manola, et al. (2007). "TP53 mutations and survival in squamous-cell carcinoma of the head and neck." The New England journal of medicine **357**(25): 2552-2561.

Psyrri, A., E. Boutati, et al. (2011). "Human papillomavirus in head and neck cancers: biology, prognosis, hope of treatment, and vaccines." Anticancer Drugs **22**(7): 586-590.

Psyrri, A., E. Boutati, et al. (2011). "Human papillomavirus in head and neck cancers: biology, prognosis, hope of treatment, and vaccines." Anti-cancer drugs **22**(7): 586-590.

Rampias, T., C. Sasaki, et al. (2009). "E6 and e7 gene silencing and transformed phenotype of human papillomavirus 16-positive oropharyngeal cancer cells." Journal of the National Cancer Institute **101**(6): 412-423.

Scully, C., J. K. Field, et al. (2000). "Genetic aberrations in oral or head and neck squamous cell carcinoma 2: chromosomal aberrations." Oral oncology **36**(4): 311-327.

Scully, C., J. K. Field, et al. (2000). "Genetic aberrations in oral or head and neck squamous cell carcinoma (SCCHN): 1. Carcinogen metabolism, DNA repair and cell cycle control." Oral oncology **36**(3): 256-263.

Shao, C., M. Tan, et al. (2012). "Suprabasin is hypomethylated and associated with metastasis in salivary adenoid cystic carcinoma." PLoS One **7**(11): e48582.

Singhi, A. D., J. Califano, et al. (2012). "High-risk human papillomavirus in nasopharyngeal carcinoma." Head & neck **34**(2): 213-218.

Smith, I. M., W. K. Mydlarz, et al. (2007). "DNA global hypomethylation in squamous cell head and neck cancer associated with smoking, alcohol consumption and stage." International journal of cancer. Journal international du cancer **121**(8): 1724-1728.

Song, J. I. and J. R. Grandis (2000). "STAT signaling in head and neck cancer." Oncogene **19**(21): 2489-2495.

Weinberger, P. M., Z. Yu, et al. (2006). "Molecular classification identifies a subset of human papillomavirus--associated oropharyngeal cancers with favorable prognosis." Journal of clinical oncology : official journal of the American Society of Clinical Oncology **24**(5): 736-747.

Werness, B. A., A. J. Levine, et al. (1990). "Association of human papillomavirus types 16 and 18 E6 proteins with p53." Science **248**(4951): 76-79.

Westra, W. H., J. M. Taube, et al. (2008). "Inverse relationship between human papillomavirus-16 infection and disruptive p53 gene mutations in squamous cell carcinoma of the head and neck." Clinical cancer research : an official journal of the American Association for Cancer Research **14**(2): 366-369.

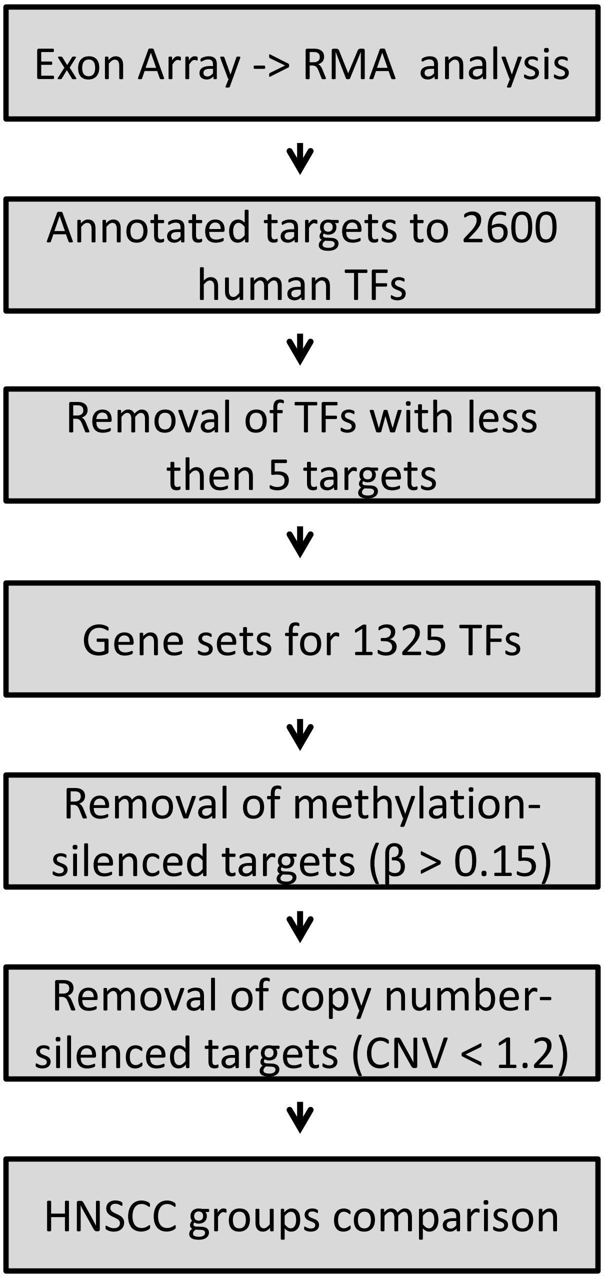

Exon Array -> RMA analysis
Annotated targets to 2600 human TFs
Removal of TFs with less then 5 targets
Gene sets for 1325 TFs
Removal of methylation-silenced targets (β > 0.15)
Removal of copy number-silenced targets (CNV < 1.2)
HNSCC groups comparison

Figure 1

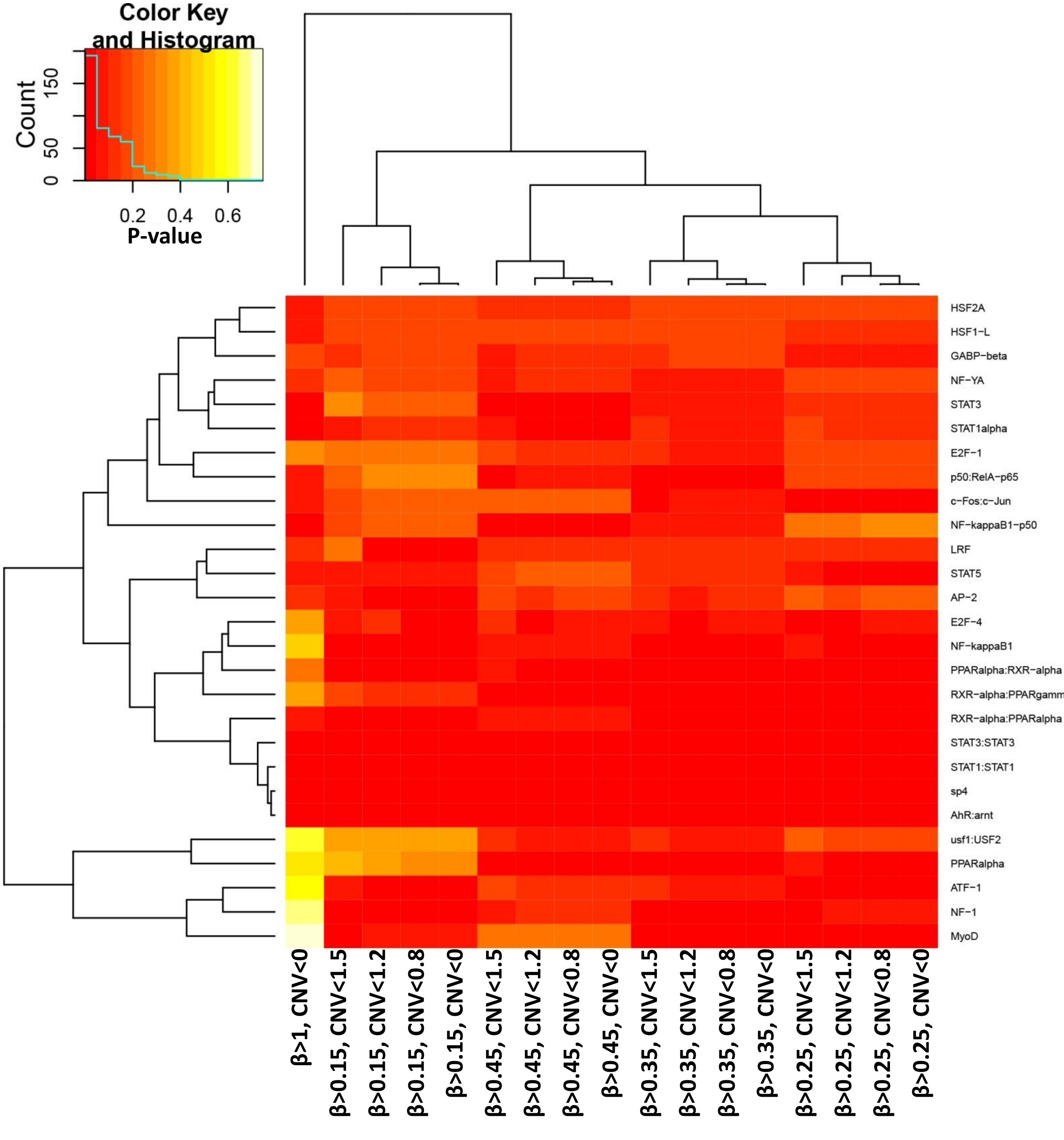

Color Key
and Histogram
Count
0
50
150
0.2
0.4
0.6
P-value
HSF2A
HSF1-L
GABP-beta
NF-YA
STAT3
STAT1alpha
E2F-1
p50:RelA-p65
c-Fos:c-Jun
NF-kappaB1-p50
LRF
STAT5
AP-2
E2F-4
NF-kappaB1
PPARalpha:RXR-alpha
RXR-alpha:PPARgamm
RXR-alpha:PPARalpha
STAT3:STAT3
STAT1:STAT1
sp4
AhR:arnt
usf1:USF2
PPARalpha
ATF-1
NF-1
MyoD
β>1, CNV<0
β>0.15, CNV<1.5
β>0.15, CNV<1.2
β>0.15, CNV<0.8
β>0.15, CNV<0
β>0.45, CNV<1.5
β>0.45, CNV<1.2
β>0.45, CNV<0.8
β>0.45, CNV<0
β>0.35, CNV<1.5
β>0.35, CNV<1.2
β>0.35, CNV<0.8
β>0.35, CNV<0
β>0.25, CNV<1.5
β>0.25, CNV<1.2
β>0.25, CNV<0.8
β>0.25, CNV<0

Figure 2

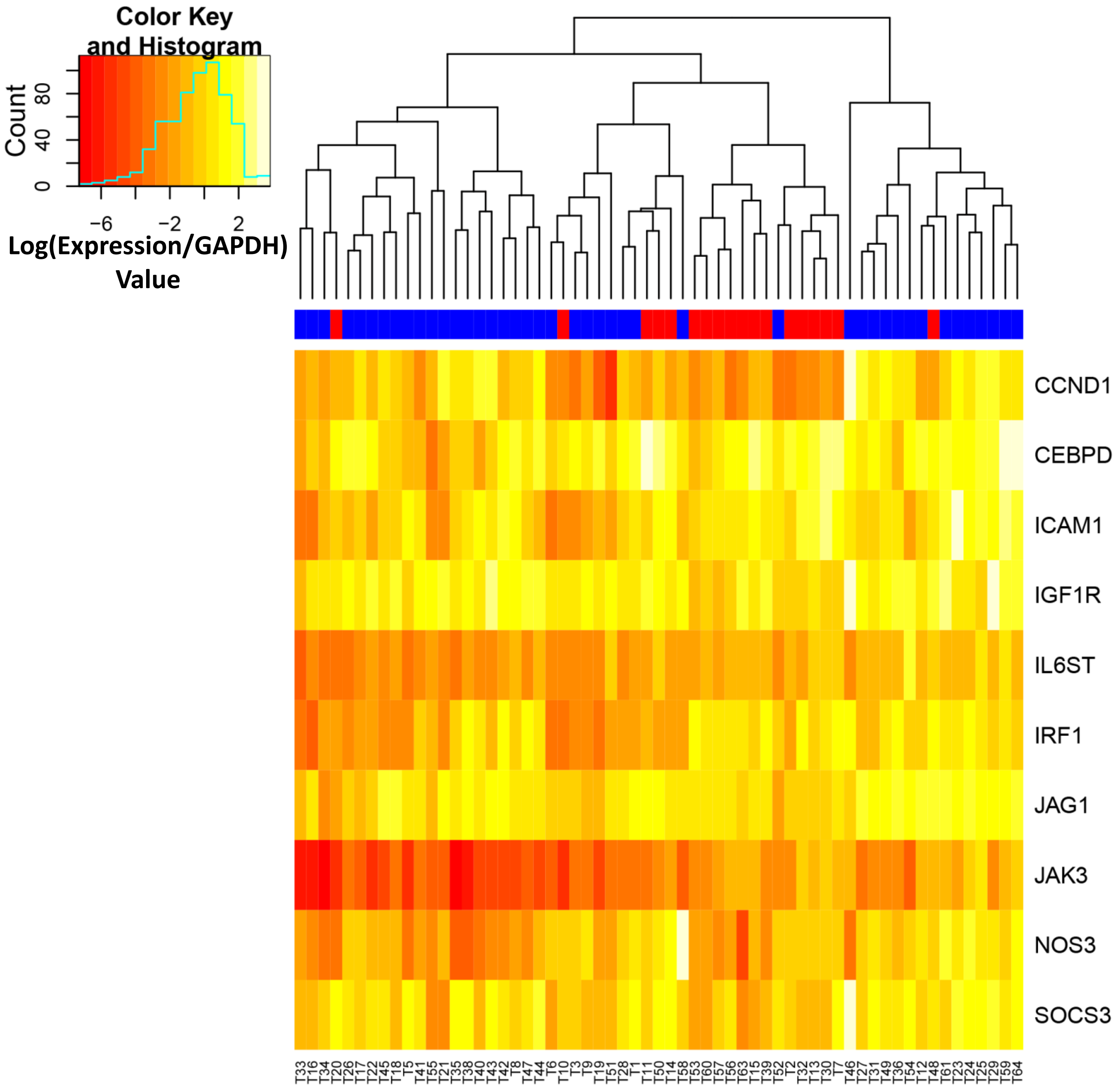

Color Key
and Histogram
Count
0
40
80
−6
−2
2
Log(Expression/GAPDH)
Value
CCND1
CEBPD
ICAM1
IGF1R
IL6ST
IRF1
JAG1
JAK3
NOS3
SOCS3
T33
T16
T34
T20
T26
T17
T22
T45
T18
T5
T41
T55
T21
T35
T38
T40
T43
T42
T8
T47
T44
T6
T10
T3
T9
T19
T51
T28
T1
T11
T50
T14
T58
T53
T60
T57
T56
T63
T15
T39
T52
T2
T32
T13
T30
T7
T46
T27
T31
T49
T36
T54
T12
T48
T61
T23
T24
T25
T29
T59
T64

Figure 3

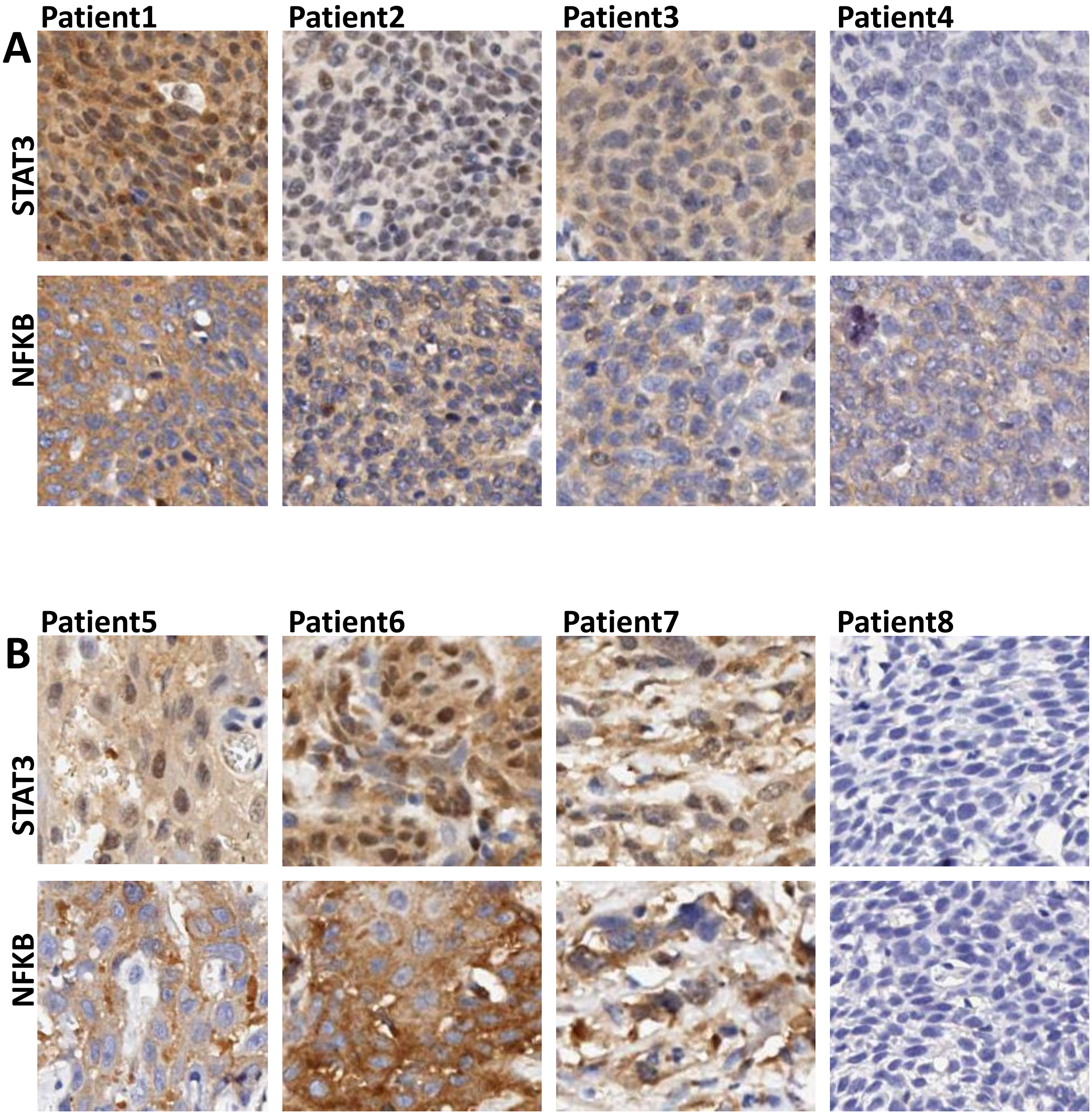

Patient1
Patient2
Patient3
Patient4
A
STAT3
NFKB
Patient5
Patient6
Patient7
Patient8
B
STAT3
NFKB

Figure 4

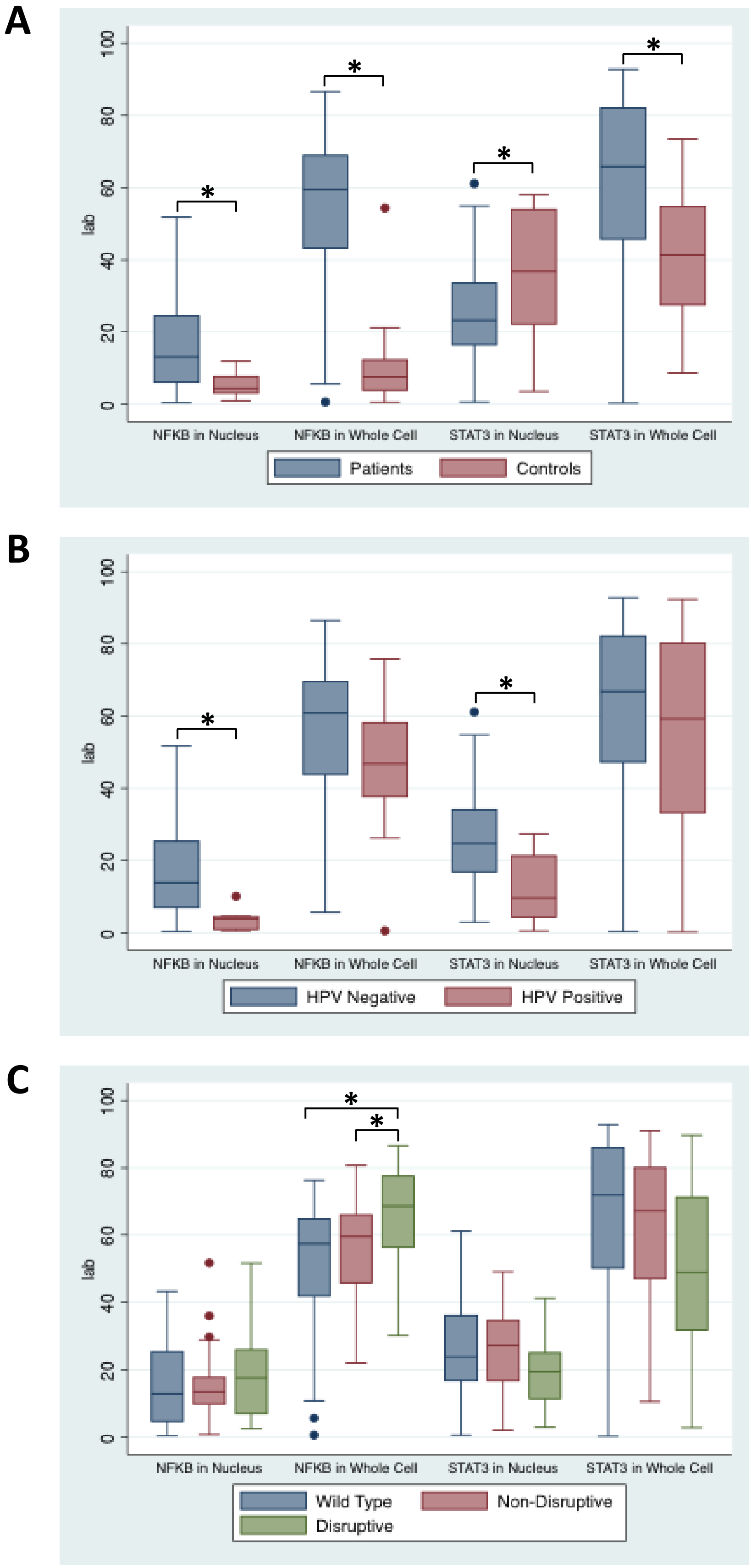
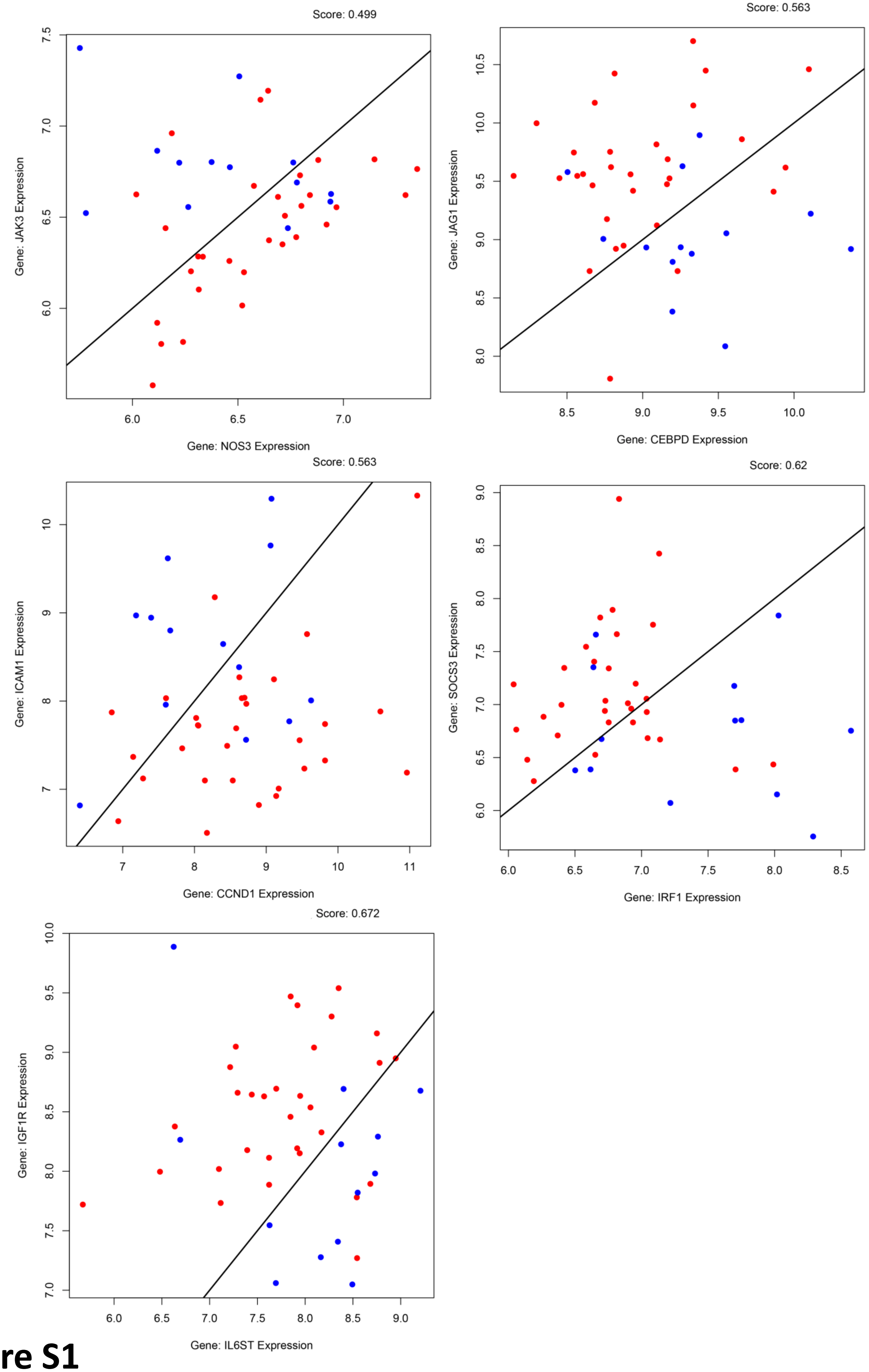

A
100
80
60
40
20
0
lab
*
*
*
*
NFKB in Nucleus
NFKB in Whole Cell
STAT3 in Nucleus
STAT3 in Whole Cell
Patients
Controls
B
100
80
60
40
20
0
lab
*
*
NFKB in Nucleus
NFKB in Whole Cell
STAT3 in Nucleus
STAT3 in Whole Cell
HPV Negative
HPV Positive
C
100
80
60
40
20
0
lab
*
*
NFKB in Nucleus
NFKB in Whole Cell
STAT3 in Nucleus
STAT3 in Whole Cell
Wild Type
Non-Disruptive
Disruptive

Figure 5



Score: 0.499
Gene: JAK3 Expression
7.5
7.0
6.5
6.0
6.0
6.5
7.0
Gene: NOS3 Expression
Score: 0.563
Gene: JAG1 Expression
10.5
10.0
9.5
9.0
8.5
8.0
8.5
9.0
9.5
10.0
Gene: CEBPD Expression
Score: 0.563
Gene: ICAM1 Expression
10
9
8
7
7
8
9
10
11
Gene: CCND1 Expression
Score: 0.62
Gene: SOCS3 Expression
9.0
8.5
8.0
7.5
7.0
6.5
6.0
6.0
6.5
7.0
7.5
8.0
8.5
Gene: IRF1 Expression
Score: 0.672
Gene: IGF1R Expression
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Gene: IL6ST Expression

Figure S1

Table 1. Clinical characteristics of HNSCC patients in the initial discovery cohort

	HNSCC (n = 44)	Normal samples (n = 25)
	n (%)	n (%)
Median age (range)	58±13 (45-80)	29±12 (18-65)
Male	32 (73%)	9 (36%)
Female	12 (27%)	16 (64%)
Race		
Caucasian	40 (91%)	14 (56%)
African American	3 (7%)	11 (44%)
Others	1 (2%)	
Smoking status		
Pack-years (range)	39.7±30.3 (4-125)	29.0±29.7 (8-50)
Smokers	27 (61%)	3 (12%)
Non-smokers	12 (28%)	22 (88%)
Unknown	5 (11%)	
Drinking status		
Drink	25 (57%)	9 (36%)
Do not drink	12 (27%)	16 (64%)
Unknown	7 (16%)	
HPV16 positive	13 (30%)	
Tumor site		
Oral cavity	10 (23%)	
Oropharynx	17 (38%)	
Larynx	13 (30%)	
Hypopharynx	4 (9%)	
TNM stage		
I	5 (11%)	
II	2 (5%)	
III	5 (11%)	
IV	32 (73%)	
Disease status		
No evidence of disease	22 (50%)	
Alive with disease	1 (2%)	
Dead of disease	18 (41%)	
Dead of unrelated causes	3 (7%)	

Table 2. The top ten scoring target genes

		pValue. HPV+ vs HPV-			pValue. Tumor vs Normal		
Gene ID	Pathway	Discovery Cohort	TCGA Cohort	Validation Cohort	Discovery Cohort	TCGA Cohort	Validation Cohort
CCND1	STAT1	0.0856	2.71E-13	0.003662	0.075238	ND	1.52E-05
CEBPD	STAT1/STAT3/NFKB	0.0173	0.0019	0.006042	1.74E-08	ND	1.34E-06
ICAM1	AP1/STAT1/STAT3/NFKB	0.0007	0.0083	4.78E-12	0.391872	ND	2.90E-14
IFG1R	NFKB	ND	ND	0.949207	ND	ND	0.009576
IL6ST	STAT1/STAT3	0.1206	0.0009	2.21E-07	0.013416	ND	4.28E-07
IRF1	STAT1/STAT3/NFKB	0.0004	0.0025	0.05451	3.01E-06	ND	0.113348
JAG1	NFKB	0.0038	0.0003	0.867499	1.10E-08	ND	0.018084
JAK3	STAT3	0.0416	3.62E-09	1.08E-17	0.136659	ND	1.16E-19
NOS3	STAT3	0.2154	0.0694	1.79E-38	0.003192	ND	1.53E-27
SOCS3	STAT1/STAT3	0.0427	0.0718	0.512127	2.19E-05	ND	0.001857

P-values were quantified by t-test. ND - non-determined

Table 3. The comparison of the top ten target genes expression in HPV- and HPV+ patients from different cohorts.

	Discovery Cohort	TCGA	Validation Cohort
p value	9.50E-06	6.70E-08	0.0006
odds ratio [95% CI]	44 [5.0 - 2185]	8.9 [3.6 26]	9.6 [2.2 60]

P-values were quantified by Fisher Exact Test

Table 4. NFKB and STAT3 staining correlation

	Spearman correlation	p-value
In nucleus of tumor samples	0.3	0.003
In whole cell of tumor samples	0.19	0.06
In nucleus of control samples	0.29	0.33
In whole cell of control samples	0.29	0.07

P-values were calculated by Wilcoxon rank sum test

Table S1. Clinical characteristics of HNSCC patients in the validation cohort

	HNSCC (n = 61)	Normal samples (n = 28)
	n (%)	n (%)
Median age (range)	60±10 (35-87)	33±11 (18-57)
Male	48 (79%)	14 (50%)
Female	13 (21%)	14 (50%)
Race		
Caucasian	53 (87%)	13 (46%)
African American	7 (11%)	13 (46%)
Others	1 (2%)	2 (8%)
Smoking status		
Pack-years (range)	52.7±165.0 (5-1095)	20.6±63.1 (3-274)
Smokers	47 (77%)	10 (36%)
Non-smokers	12 (20%)	18 (64%)
Unknown	2 (3%)	
Drinking status		
Drink	39 (64%)	3 (11%)
Do not drink	19 (31%)	25 (89%)
Unknown	3 (5%)	
HPV16 positive	18 (30%)	
Tumor site		
Oral cavity	14 (23%)	
Oropharynx	30 (49%)	
Larynx	15 (25%)	
Hypopharynx	2 (3%)	
TNM stage		
I	3 (5%)	
II	6 (10%)	
III	5 (8%)	
IV	29 (48%)	
Unknown	18 (29%)	
Disease status		
No evidence of disease	21 (35%)	
Alive with disease	11 (18%)	
Dead of disease	7 (11%)	
Dead of unrelated causes	7 (11%)	
Unknown	15 (25%)	

Table S2. Clinical characteristics of HNSCC patients in the ECOG-TMA cohort

	HNSCC (n = 100)	Control (n = 13
	n (%)	n (%)
Male	73 (73%)	
Female	27 (27%)	
Race		
Caucasian	73 (73%)	
African American	24 (24%)	
Other	3 (3%)	
Smoking status		
Smokers	78 (78%)	
Non-smokers	18 (18%)	
Unknown	4 (4%)	
HPV16 positive	10 (10%)	
Tumor site		
Oral cavity	43 (43%)	
Oropharynx	25 (25%)	
Larynx	20 (20%)	
Hypopharynx	11 (11%)	
Salivary Gland	1 (1%)	
p53 mutation status		
Wild type	46 (46%)	
Non-disruptive	29 (29%)	
Disruptive	15 (15%)	
Unknown	10 (10%)	

Table S3. Top 50 TF differently dysregulated in HPV+ and HPV- HNSCC patients from the discovery cohort

	TF	pValue
1	AhR:arnt	0.001048352
2	HNF-4alpha1	0.002866295
3	LEF1-isoform1	0.008492089
4	STAT1:STAT1	0.009404568
5	Sp1-isoform1	0.011978578
6	NF-kappaB1	0.014236153
7	PPARalpha:RXR-alpha	0.018835914
8	ATF-1	0.026191031
9	AP-2	0.031000472
10	STAT3:STAT3	0.03166898
11	Elk1-isoform1	0.034849856
12	LRF	0.035148693
13	NF-1	0.039408411
14	STAT6	0.039778518
15	RXR-alpha:PPARalpha	0.040351852
16	E12	0.048328688
17	FOXO3a	0.051936645
18	STAT5	0.056376692
19	MyoD	0.05894446
20	TBP	0.059987462
21	c-Myc	0.060184546
22	CREB	0.063842797
23	Nrf2	0.065482991
24	HSF1-L	0.074826318
25	MITF	0.085353634
26	HSF2A	0.089854075
27	E2F-4	0.109071313
28	HNF-4	0.112508574
29	RXR-alpha:PPARgamma	0.11545721
30	STAT1	0.123446702
31	RORalpha1	0.129812833
32	LXR-alpha:RXR-alpha	0.129870013
33	Fra-1	0.134245053
34	NF-YA	0.144235603
35	(RelA-p65)2	0.153160281
36	STAT1alpha	0.154080055
37	ctcf	0.158620894
38	ATF	0.173694793
39	FOXO1A	0.174851014
40	Clock:BMAL1	0.175402849
41	HIF-1alpha:arnt	0.181393146
42	GABP-alpha:GABP-beta1	0.190710431
43	RXR-alpha	0.193104629
44	GABP-beta	0.195373129
45	HES-1	0.198611536
46	NF-kappaB1-p50:RelA-p65	0.20627903
47	c-Fos:c-Jun	0.210316376
48	PEA3	0.210447901
49	NF-AT1	0.223356953
50	p53	0.234342437

The p-values were calculated using a conservative Wilcoxon test on each TF target gene set comparing differential expression of each TF target in HPV+ vs HPV- tumor samples. The lowest p-value indicates strong differential expression of TF targets in HPV+ as compared to HPV- tumor samples.

Table S4. Top 50 TF differently dysregulated in HPV+ and HPV- HNSCC patients from the TCGA cohort

	TF	p-Value
1	STAT1	1.97E-08
2	HNF3A	7.37E-08
3	Smad4	4.72E-06
4	E2F	6.59E-06
5	Smad3	6.68E-06
6	MAZ	9.42E-06
7	GABP-beta	3.44E-05
8	NR1B1:RXR-alpha	4.71E-05
9	c-Fos:c-Jun	4.98E-05
10	NF-kappaB1-p50:RelA-p65	6.46E-05
11	E2F:DP	7.55E-05
12	HNF-1alpha	0.000288889
13	MyoD	0.000412333
14	GATA-1	0.000526017
15	Elf-1	0.000543379
16	p53	0.000660128
17	SREBP-1c	0.000661377
18	STAT1:STAT3	0.00068438
19	Sp1	0.000786779
20	NF-kappaB1	0.000909947
21	STAT5	0.001033089
22	POU2F1	0.001126797
23	RXR-alpha:PPARgamma	0.001478215
24	STAT1alpha	0.001823834
25	sp4	0.001911328
26	NF-AT1	0.002081297
27	TCF-4	0.002492685
28	Pax-5	0.003634786
29	HES-1	0.00364241
30	p63alpha	0.004063488
31	STAT5A	0.004183604
32	NF-Y	0.004312941
33	E12	0.004523778
34	Clock:BMAL1	0.004709103
35	(STAT1)2	0.004935939
36	STAT5B	0.0052125
37	IRF-3	0.00526602
38	c-Ets-1	0.006466447
39	AP-2alpha	0.007082614
40	p53-isoform1	0.011030381
41	STAT3:STAT3	0.011733879
42	NF-1A	0.012539552
43	p50	0.012716994
44	usf1	0.012965927
45	STAT6	0.015361393
46	IRF-8	0.016724151
47	Sp1:Sp3	0.01708546
48	C/EBPalpha	0.017388899
49	p73alpha	0.019113828
50	Elk1-isoform1	0.019307869

The p-values were calculated using a conservative Wilcoxon test on each TF target gene set comparing differential expression of each TF target in HPV+ vs HPV- tumor samples. The lowest p-value indicates strong differential expression of TF targets in HPV+ as compared to HPV- tumor samples.

Table S5. The list of target genes for STAT1, STAT3, NFKB and AP1 pathways

Gene ID	Pathway
CAPNS1	AP1
ACAT1	STAT1
BACE1	STAT1
BAX	NFKB
CCND1	STAT1
CD40	STAT1
CDKN1A	STAT1/STAT3
CDKN1B	STAT3
CEBPD	STAT1/STAT3/NFKB
CISH	STAT3
CLGN	AP1/STAT3
CXCL5	NFKB
DUSP1	STAT3
ESR1	AP1/STAT3
FAAH	STAT3
FAS	NFKB
FASN	STAT1
FCER2	STAT1
FCGRT	STAT1/NFKB
FOS	STAT1/STAT3
G1P2	STAT1
GADD45B	STAT3
HAMP	STAT1
HBEGF	AP1
HMOX1	STAT3
ICAM1	AP1/STAT1/STAT3/NFKB
IFNB1	NFKB
IGF1R	NFKB
IL6ST	STAT1/STAT3
IRF1	STAT1/STAT3/NFKB
IRF8	STAT1/NFKB
IVL	AP1/STAT1
JAG1	NFKB
JAK3	STAT3
JUNB	NFKB
KLF4	STAT1
LPL	STAT1
LY6E	STAT1
MCL1	STAT3
MMP2	STAT3
MMP9	NFKB
MTHFR	NFKB
MUC1	STAT1
MYC	STAT1/STAT3/NFKB
MYD88	STAT1/STAT3
NFKBIA	NFKB
NOS2A	STAT1

NOS3	STAT3
NR4A2	NFKB
OPRD1	NFKB
PEMT	STAT1
PIM1	STAT1/STAT3
POMC	STAT3
PPARG	STAT1
PSMB9	STAT1
PTGS2	AP1/NFKB
RELB	NFKB
SDC4	NFKB
SERPINB9	NFKB
SLC9A3	STAT3
SNCG	AP1
SOCS3	STAT1/STAT3
SOD2	NFKB
SST	AP1
STAT3	STAT3
TF	AP1
TNF	NFKB
TRH	STAT1/STAT3
TWIST1	STAT3
VEGF	STAT3
VIM	STAT3
WARS2	STAT1

Table S6. Quantification of IHC staining for STAT3 and NFKB in whole cell or in nuclei. Staining correlation with HPV or p53 mutation status

#	Nuclear NFKB staining	Cellular NFKB staining	Nuclear STAT3 staining	Cellular STAT3 staining	0 = HPV-; 1 = HPV+	p53 mutation status
1	0.55	0.51	0.52	0.22	0	wt
2	4.81	10.78	2.99	0.31	0	wt
3	0.70	26.20	1.94	12.20	1	missense
4	8.96	64.62	2.88	2.73	0	missense
5	6.74	32.08	16.89	24.92	0	wt
6	2.44	44.81	4.27	33.26	1	wt
7	2.42	30.19	20.92	31.81	0	10 bp deletion
8	6.90	62.36	8.21	10.49	0	missense
9	16.56	45.81	8.06	18.68	0	missense
10	15.08	59.52	5.78	10.52	0	missense
11	13.08	57.61	3.83	21.20	0	wt
12	3.76	48.80	4.58	39.32	1	missense
13	4.22	20.56	19.85	53.25	0	ND
14	13.04	49.71	9.73	28.52	0	wt
15	3.86	32.62	16.95	48.63	0	wt
16	4.63	34.23	16.13	48.56	0	wt
17	2.34	28.68	28.99	47.08	0	missense
18	5.84	34.91	17.45	52.30	0	wt
19	7.00	39.88	19.51	44.34	0	missense
20	11.84	59.42	8.84	32.24	0	ND
21	6.45	19.49	23.23	66.77	0	ND
22	16.41	41.93	25.91	38.84	0	wt
23	12.52	63.04	15.73	32.52	0	missense
24	4.38	58.12	8.71	53.54	1	wt
25	4.76	43.90	11.88	64.35	0	missense
26	6.45	37.42	29.12	54.34	0	wt
27	11.76	31.97	27.23	57.98	0	missense
28	0.32	5.60	36.87	87.59	0	wt
29	25.92	80.50	11.36	13.11	0	non sense
30	5.31	34.55	24.78	67.25	0	missense
31	13.38	22.08	40.81	55.63	0	missense
32	15.56	47.36	16.80	52.95	0	wt
33	12.19	61.35	16.76	44.19	0	missense
34	4.01	66.52	16.63	50.98	0	missense
35	11.60	46.01	19.44	62.47	0	missense
36	21.46	51.66	24.21	42.58	0	missense
37	6.33	52.60	18.80	62.19	0	wt
38	14.26	36.62	19.59	69.59	0	ND
39	25.29	59.97	16.50	41.81	0	wt
40	10.08	44.22	10.64	79.81	1	missense
41	27.60	66.17	13.94	37.28	0	ND
42	11.81	42.37	24.72	68.06	0	wt
43	13.35	64.83	18.76	50.17	0	wt
44	3.88	37.71	21.44	87.55	1	wt
45	17.74	58.73	27.21	47.22	0	missense
46	11.13	49.44	16.70	74.23	0	wt
47	0.98	49.15	21.30	80.15	1	wt
48	9.48	32.39	33.95	79.20	0	wt
49	24.78	51.10	24.63	56.05	0	missense
50	28.84	61.85	23.10	44.23	0	missense
51	4.58	75.79	14.80	64.91	1	wt
52	4.76	36.65	33.66	85.17	0	wt
53	6.86	36.27	25.66	92.77	0	wt
54	15.10	59.52	21.11	66.89	0	wt
55	43.32	76.30	8.96	34.39	0	wt

56	23.70	56.47	24.96	61.21	0	missense
57	24.27	82.53	10.65	48.91	0	3 bp deletion
58	9.83	72.37	15.54	68.86	0	missense
59	3.08	69.00	24.62	71.66	0	ND
60	24.46	71.96	25.52	47.64	0	non sense
61	11.12	72.21	28.03	58.87	0	silente
62	38.29	83.58	11.93	37.27	0	ND
63	22.74	35.86	34.61	83.04	0	missense
64	4.02	61.74	22.79	89.57	1	wt
65	37.69	70.20	18.90	51.87	0	wt
66	4.96	70.23	30.28	74.25	0	missense
67	17.60	72.79	19.83	72.09	0	missense
68	20.49	70.14	22.17	69.55	0	wt
69	2.43	51.53	40.52	88.09	0	wt
70	26.40	68.88	30.13	57.60	0	wt
71	13.58	48.45	36.04	88.66	0	wt
72	4.38	64.13	27.35	92.34	1	wt
73	7.75	67.93	28.75	85.27	0	missense
74	14.69	61.86	34.07	80.08	0	missense
75	16.23	49.88	42.70	82.03	0	missense
76	19.65	62.89	30.73	77.90	0	wt
77	13.49	63.38	33.31	86.46	0	missense
78	20.30	54.46	44.79	77.17	0	wt
79	11.42	63.18	46.23	78.19	0	missense
80	12.53	76.54	20.69	89.71	0	missense
81	11.78	69.74	44.60	75.91	0	missense
82	35.01	79.99	20.98	67.22	0	ND
83	28.92	59.52	39.35	82.15	0	wt
84	1.00	71.86	46.13	91.12	0	missense
85	51.64	86.55	16.42	60.03	0	missense
86	29.78	66.05	37.67	82.56	0	missense
87	17.07	60.35	61.12	78.76	0	wt
88	38.98	72.46	32.62	74.33	0	wt
89	27.50	69.60	38.69	83.18	0	silente
90	17.54	76.31	39.80	85.75	0	ND
91	12.65	61.33	54.85	92.73	0	wt
92	30.70	57.15	46.28	88.19	0	wt
93	49.14	77.60	25.40	71.18	0	non sense
94	31.17	68.17	40.44	84.45	0	wt
95	35.11	74.70	28.67	87.71	0	wt
96	30.06	70.09	38.01	88.13	0	wt
97	35.55	68.65	41.25	82.67	0	non sense
98	51.73	80.67	29.97	72.21	0	missense
99	33.82	63.08	51.83	85.96	0	wt
100	35.95	63.44	49.06	88.70	0	missense
101	0.93	0.48	22.11	8.57	ND	tonsil control
102	3.04	2.29	3.46	27.56	ND	uvula control
103	4.33	3.83	18.92	21.11	ND	uvula control
104	5.20	5.60	29.64	23.63	ND	tonsil control
105	1.60	7.57	36.76	37.40	ND	uvula control
106	2.08	0.62	38.60	50.62	ND	tonsil control
107	3.72	4.29	32.89	54.27	ND	uvula control
108	7.66	9.43	44.52	54.85	ND	uvula control
109	11.52	54.29	18.32	33.39	ND	tonsil control
110	8.05	17.77	57.25	41.27	ND	uvula control
111	11.83	21.12	53.85	54.72	ND	uvula control
112	5.84	11.36	56.40	69.89	ND	tonsil control
113	4.30	12.19	57.99	73.37	ND	tonsil control

Color code: Yellow represents 50-percentile, green - lowest values; red - highest value. Top 100 samples belong to tumor patients. Bottom 13 - non-cancerous patients. Samples were sorted by total staining