bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Clusterization in head and neck squamous carcinomas based on lncRNA expression: molecular and clinical correlates

Pelayo González de Lena[1], Abel Paz-Gallardo[2], Jesús M. Paramio[1,3,4], and Ramón García-Escudero[1,3,4,5]

1. Molecular Oncology Unit, CIEMAT, Madrid, Spain.

2. CETA-CIEMAT, Trujillo, 10200 Cáceres, Spain.

3. Biomedical Research Institute I+12, University Hospital 12 de Octubre, Madrid, Spain.

4. Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Spain.

5. Institute of Oncology Research (IOR), and Oncology Institute of Southern Switzerland (IOSI), Bellinzona, Switzerland.

Correspondence to: Ramón García-Escudero, Molecular Oncology Unit, CIEMAT, Ave Complutense 40 (ed70A), 28040 Madrid, Spain. Email: ramon.garcia@ciemat.es

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background

Long non-coding RNAs (lncRNAs) have emerged as key players in a remarkably variety of biological processes and pathologic conditions, including cancer. Next-generation sequencing technologies and bioinformatics procedures predict the existence of tens of thousands of lncRNAs, from which we know the functions of only a handful of them, and very little is known in cancer types such as head and neck squamous cell carcinomas (HNSCCs).

Results

Here, we use RNA-seq expression data from The Cancer Genome Atlas (TCGA) and various statistic and software tools in order to get insight about the lncRNome in HNSCC. Based on lncRNAs expression across 426 samples, we discover five distinct tumor clusters that we compare with reported clusters based on various genomic/genetic features. Results demonstrate significant associations between lncRNA-based clustering and DNA-methylation, TP53 mutation, and human papillomavirus infection. Using "guilt by association" procedures, we infer the possible biological functions of representative lncRNAs of each cluster. Furthermore, we found that lncRNA clustering is correlated with some important clinical and pathologic features, including patient survival after treatment, tumor grade or sub-anatomical location.

Conclusions

We present a landscape of lncRNAs in HNSCC, and provide associations with important genotypic and phenotypic features that may help to understand the disease.

Keywords

lncRNA, head and neck squamous cell carcinoma, cluster, genomics

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

Head and neck cancer is the sixth leading cancer worldwide, with an estimated 600,000 new cases annually and a 50% five-year mortality rate (Globocan 2012) [1]. As more than 90% of this cancer cases are of squamous origin, they are generally referred to as head and neck squamous cell carcinoma (HNSCC). HNSCC arises in the upper aerodigestive tract, comprising the nasal cavity and paranasal sinuses, oral cavity, pharynx, larynx, and trachea. The main risk factors associated with its development are tobacco and alcohol consumption, which have a synergistic effect when combined, and also human papillomavirus (HPV) infection [2,3]. HPV is known to drive tumorigenesis through the actions of its major oncoproteins E6 and E7, which can inactivate p53 and retinoblastoma (Rb) tumor supressors, respectively, altering cell cycle regulation in infected cells. HPV-positive differ from HPV-negative HNSCCs in tumor biology and clinical characteristics, including clinical outcomes, since HPV-positive tumors have been associated with a more favorable prognosis [4]. HNSCC patients are frequently treated with surgery, together with radiotherapy and/or cisplatin-based chemotherapy. Patients with aggressive disease are treated with cetuximab, an anti-EGFR antibody. Few patients respond to this therapy, and there is not molecular stratification of the patients neither biomarkers of responsiveness [2]. Previous characterization of molecular features in HNSCC, particularly with the aid of large-scale cancer genomics initiatives such as the TCGA, has generated important insights for stratifying patients and delineating tumor subtypes [5–7]. These multiomic analyses, do not take into account the vast long non-coding transcriptome that may substantially contribute to HNSCC pathogenesis and progression.

LncRNAs are transcripts longer than 200 nucleotides that have no apparent protein coding potential [8]. They are highly diverse and actively present in many aspects of cell biology, including cellular differentiation, proliferation, DNA damage response, dosage compensation, and chromosomal imprinting. LncRNAs are categorized as exonic, intronic, intergenic, antisense, or overlapping based on their genomic location relative to a protein-coding gene. The most recent estimate of the Encyclopedia of DNA Elements (ENCODE) Project Consortium (GENCODE version 25) is that the human genome contains more than 15,000 lncRNA genes that encode almost 28,000 transcripts [9] although the total number is estimated to be much higher. Like the protein coding genes (PCGs), lncRNAs genes are regulated transcriptionally and by histone modification, and lncRNA transcripts are processed by the canonical splicing machinery [10]. In addition, lncRNAs have fewer exons than PGCs, are usually located in the nucleus, are subject to less selective pressure during evolution, and show higher tissue or cell type specificity. Given the large number of lncRNAs that are predicted to exist, it is expected that the functions in which they are involved may include many of the known (and possibly new) biological and physiological processes. Unfortunately, the possible functions or

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phenotypic effect of the vast majority of lncRNAs remains elusive in normal homeostasis or in cancer [11], and are difficult to analyze [12].

Therefore, the main objective of this report is to analyze the long non-coding transcriptome in HNSCC in order to discover new tumor clusters. Furthermore, we investigate whether lncRNA-based clusterization is useful to predict patient's clinical outcome, and is associated with important clinic-pathologic parameters. Finally, we interrogate the dataset to infer the possible biological functions of lncRNAs in HNSCC based on correlation patterns of the PCGs.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Data resources

Expression values of 12,727 lncRNA genes from 426 HNSCC primary tumor samples of the TCGA RNAseq cohort were downloaded from the TANRIC web page [13]. Methods used to extract these expression values have been described [14]. Briefly, the genomic coordinates of the human lncRNAs from the GENCODE Resource (version 19) were obtained. Thereafter, the lncRNA exons that overlapped with any known coding genes based on the gene annotations of GENCODE and RefGene were filtered out. As a result, the analysis focused on the remaining 12,727 lncRNAs. Based on the BAM files, the expression levels were quantified as RPKM, and the lncRNAs with detectable expression were defined as those with an average RPKM $\geq$ 0.3 across all samples in each cancer type, as defined in the literature. The Ensembl identifiers of the 500 lncRNA genes displaying the highest variability in HNSCC were kindly provided by Dr. H. Liang at the Research Group from the Department of Bioinformatics and Computational Biology of The University of Texas and the MD Anderson Cancer Center. We downloaded DNA methylation, CNV, miRNA, and reverse phase protein array (RPPA) clustering data, as well as HPV infection and PCG mutations for the TCGA HNSCC dataset using the cBioPortal [15], the UCSC Xena [16] and the TCGA [17] repositories.

Unsupervised clustering analysis

Consensus Cluster Plus (CCP) tool [18], which is implemented as an R language package from Bioconductor [19], extends the CC algorithm and is briefly described here. The algorithm begins by subsampling a proportion of items and a proportion of features from a data matrix. Each subsample is then partitioned into up to k groups by a user-specified clustering algorithm. This process is repeated for a specified number of times. Pairwise consensus values, defined as 'the proportion of clustering runs in which two items are grouped together', are calculated and stored in a consensus matrix (CM) for each k. Clustering settings used were as follows: maxK=6; number of bootstraps=1000; item subsampling proportion=0.8; feature subsampling proportion=1; cluster algorithm=pam; inner linkage type=complete; final linkage type=complete; correlation method=Euclidean.

Analysis of tumor clusters revealed by lncRNA expression.

The association between lncRNA-based sample clusters and sample clusters based on other molecular features/aberrations was done using Fisher's exact test for fourfold (2x2) tables or Chi-square test for more than fourfold comparisons. Both tests are used to determine whether there are significant differences between the expected frequencies and the observed frequencies in one or more categories. Kaplan-Meier curves were obtained using the follow-up time of the HNSCC patients of 2 end-points: recurrence and death. Statistical analyses were done with SPSS 14.0.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Guilt-by-association (GBA) analysis

Pearson correlation analysis was used to select genes with significant correlation between artificial expression vectors or templates (T1 to T5) (Fig. 2A), or between mean expression values of correlated lncRNAs and PGCs. We select both direct and inverse correlation patterns by setting thresholds either at the Pearson (r) value or at the associated p-val. Pearson's r values ranges between -1 to +1, such that 2 perfectly correlated genes display r=1, and 2 perfectly anticorrelated genes r=-1. LncRNAs surrogate selection was done setting r>0.3 with respect to the corresponding template. PGCs directly or inversely correlated with lncRNAs were selected using r>0.3, or r<-0.3 with respect to mean expression values of selected lncRNAs per cluster, respectively. Correlations as well as heatmap drawings were performed using MultiExperiment Viewer v4.9 (MeV) [20].

Gene Ontology analysis

Selected PGCs were analyzed with the Gene Functional Annotation Tool available at the DAVID website [21,22] using their official gene symbols. Gene ontology option GOTERM_BP_FAT was selected and a functional annotation chart generated. A maximum p-value of 0.05 was chosen to select only significant categories.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

HNSCC sample clusters based on lncRNA expression.

In order to discover HNSCC tumor subgroups, we selected the 500 lncRNAs (Supplementary Table 1) with the most variable expression pattern in HNSCC [14] and the Consensus Cluster Plus (CCP) software tool (see Methods). CCP analysis revealed the presence of five HNSCC clusters, which we have named clusters 1 to 5 (Fig. 1 and Supplementary Table 2). The results suggest that lncRNAs can distinguish 5 HNSCC subtypes, having significant differences in lncRNA expression.

LncRNAs clustering resembles DNA methylation and mRNA clustering.

In order to assess the correlation between lncRNAs and additional molecular features, we analyze whether lncRNA clustering resembles previously described sample clustering based on mRNA, miRNA, protein (RPPA) expression, DNA methylation and CNV. For this, we downloaded clustering data from the TCGA-HNSSC dataset [5] and compared them with lncRNAs subgroups using contingency analyses (Chi-square test, see Methods). The results showed highly significant similarities between our clusterization and DNA methylation clusters (p-val=4.6 x 10^{-34}) or mRNA-based subtypes (p-val=4.2 x 10^{-30}) (Fig. 2A). LncRNA clustering displays lower overlapping with CNV, miRNA, and RPPA subtypes, with p-values ranging from 2.8 x 10^{-15} to 3.5 x 10^{-5} (Fig. 2A). The high overlapping between lncRNA and mRNA subgroups is suggestive of similar molecular mechanisms of expression control. Also, the contingency results suggest that lncRNA expression might be strongly influenced by DNA methylation (or viceversa), an important mechanism of epigenetic transcriptional regulation.

HPV infected tumors display and specific lncRNAome.

A significant proportion of tumor samples from the reported TCGA dataset [5], having 279 samples, is infected with HPV (almost 13%). Interestingly, a deep, multiplatform analysis of the molecular features of HNSCC primary tumors from the TCGA, demonstrated strong differences between HPV-positive and HPV-negative samples [5]. Infected cancers display different mutational landscape, mainly characterized by the absence of TP53 gene mutations. Chi-square test demonstrated that lncRNA clustering is highly significantly associated with HPV (p-val=1.5 x 10^{-29}) (Fig. 2B), indicating infected tumors display a specific lncRNA landscape. Interestingly, a deeper analysis shows that almost all samples from cluster 5 (c5) are HPV positive (24 out of 29) (p-val=4.0 x 10^{-21}) (Fig. 2D).

HNSCC mutations and lncRNA clustering

In order to discover whether particular lncRNA clusters are characterized by the presence or absence of HNSCC mutations, we interrogate lncRNA-cluster and mutation associations using Chi-square test (Fig 2B-D). We selected a total number of 19 genes, found to be significantly mutated in the TCGA-HNSCC cohort: CDKN2A, FAT1, TP53, CASP8, AJUBA, PIK3CA,

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NOTCH1, KMT2D, NSD1, HLA-A, TGFBR2, HRAS, FBXW7, RB1, PIK3R1, TRAF3, NFE2L2, CUL3 and PTEN. Significant associations between these mutations and lncRNA-based clustering are shown (Fig. 2C). Cluster 1 (c1) is enriched in TP53 and NSD1 mutations, cluster 2 (c2) display frequent TP53 and KMT2D mutations, and cluster 4 (c4) frequent CASP8, NOTCH1 and CDKN2A mutations. In addition, cluster c5 is depleted of TP53 and CDKN2A mutations, as most samples are HPV-infected. Finally, cluster 4 (c4) is depleted on KMT2D mutations. Both KMT2D and NSD1 are methyltransferases involved in methylation of H3K4 or H3K36, which are important epigenetic regulators of gene expression. Most mutations found for both proteins are truncating, therefore indicating that c1 and c2 are characterized by loss-of-function of histone modifiers.

Guilt-by-association (GBA) analysis: selection of surrogate lncRNAs per cluster

In order to select candidate lncRNAs for further analysis, we decided to perform differential expression analysis using the clustering information. Our aim was to select lncRNAs surrogate of each of the 5 clusters, so they are overexpressed only in 1 cluster. We designed 5 expression vectors or templates, (t1 to t5, see Figure 3A) from which we performed Pearson correlation analysis (see Methods) onto the lncRNA genes from which expression values are available for HNSCC (n=12,727). The output are lncRNAs whose expression patterns are significantly correlated with the corresponding template ($r>0.3$ and p-val$<10^{-6}$). The number of selected lncRNAs per cluster and the heatmap showing expression patterns is shown (Fig. 3B) (Supplementary Table 3). Note that clusters c2 and c5 display the highest numbers of lncRNA genes, and c3 and c4 the lowest.

Associated PCGs to surrogate lncRNAs

Predicting the biological functions of lncRNAs is challenging. Guilt-by-association (GBA) analysis has been proposed on the basis that the function of a poorly characterized lncRNA gene can be inferred from known/predicted functions of protein coding genes which are co-expressed [23]. We leveraged our RNA-sequencing data using GBA analysis to generate hypotheses on functional significance by comparing the expression of lncRNAs to protein coding genes of known function. Therefore, we performed correlation analysis to interrogate what coding genes are directly or inversely correlated with the surrogate lncRNA genes of each cluster. Accordingly, we use 5 different expression vectors or templates, one per cluster, calculated from the mean expression value of lncRNAs selected per cluster above (Fig. 3B). The number of correlated genes per cluster and the heatmap showing their expression patterns is shown (Fig. 3C and D) for directly (overexpressed PCGs) or inversely (underexpressed PCGs) correlated PCGs. Thresholds used are $r>0.3$ or $r<-0.3$ and p-val$<1x10^{-9}$ (Supplementary Table 4).

The results shown that each lncRNA cluster have associated a significant number of coding genes, whose functions might be related with the biological roles of the correlated lncRNAs.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene Ontology analysis of lncRNA-clustering associated PGCs

In order to predict possible functions of the surrogate lncRNAs in each cluster, we perform enrichment analysis of GO Biological Processes (GOBP) on the associated PGCs, which would find enriched pathways and larger processes made up of the activities of the PGC gene products. Therefore, we analyze individually the lists of PGCs directly or inversely correlated with each lncRNA cluster (Fig. 4). Cluster c1 is characterized by the underexpression of PGCs involved in T cell, natural killer cell and myeloid leukocyte activation; these processes are overexpressed in c5, suggesting important differences in immune response between c1 and c5. Importantly, PD-L1 (CD274), an immune inhibitory receptor ligand whose inactivation using antibodies is being currently used for cancer treatment successfully, is underexpressed in c1. Cluster c2 is characterized by the depletion of PGCs involved in ubiquitin-dependent protein cleavage, including some protein component of the proteasome complex (PSM proteins). Cluster c3 display underexpression of PGCs involved in histone modification, such as deacetylases, acetyltransferases or lysine demethylases. Cluster c4 contain overexpressed PGCs involved in epidermal development, including late cornified envelope (LCE) cluster genes and small proline-rich protein (SPRR) genes. Both LCE and SPRR proteins are expressed in terminally differentiated stratified epithelia, such as skin or head and neck mucosa. Interestingly, these same genes are underexpressed in c5, suggesting distinct epithelial differentiation of c4 versus c5 carcinoma cells. Cluster c5 has many overexpressed PGCs involved in DNA replication and cell cycle processes, RNA splicing or transcription, functions which are depleted in c4. Overall, the results highlights important differences between lncRNA clusters in terms on predicted functions, somehow validating the unsupervised clustering and Pearson correlation analysis performed. Whether the surrogate lncRNAs of each cluster are also involved in these biological functions remains to be demonstrated.

Clinical and pathological features in lncRNA groups

We wanted to investigate whether our lncRNA clusters could display different clinical and pathological characteristics, possibly providing a better understanding of the patient's tumor behavior or response to current therapies. We analyze whether patient sample clusters display follow-up differences in recurrence or in death, after treatment. Therefore, we performed Kaplan-Meier survival plots for both end-points (recurrence or overall survival) and the 5 clusters. Results show significant differences in overall survival at 5 years (p-val=0.0065, log-rank test), with c5 patients having the highest probabilities of survival (Fig. 5A). In addition, patients within c1 and c2 clusters exhibit the lowest survival. No significant statistical differences were found, however, in recurrence after treatment (Supplementary Fig.1).

Also, we examined tobacco smoking history, and found more smokers in c1 and c2 than expected by chance (p-val=0.02 and 0.0002, respectively, upon Fisher's test) (Fig 5B). No differences were found in alcohol consumption between patient clusters. Interestingly, we show

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that tumor histology grade (G1, G2, G3, or G4) is associated with lncRNA clustering (p-val=6.58×10^{-10}, Chi-square test). More specifically, we found that c5 tumors have frequently poorly differentiated and undifferentiated histology (G3 and G4, respectively) compared with the other clusters, while c4 have more differentiated or moderately differentiated tumors (G1 or G2, respectively). Furthermore, sub-anatomic location of the tumors and cluster subdivision were significantly associated (p-val=5.1×10^{-41}, Chi-square test), so that c1 and c2 tumors are frequently found in larynx, c4 in tongue, and c5 in tonsil (Fig. 5B). Finally, we found that c1 patients are younger that the remaining patients (58.2 versus 61.6, p-val=0.018 after T-test), and c2 patients older (65 versus 60.2, p-val=0.001 after T-test) (Fig. 5C). In summary, lncRNA clusterization is associated with important clinico-pathologic features, providing important genomic and phenotypic relationships.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

The recent implication of lncRNAs in many biological functions has established a new scenario to better understand complex processes like cancer. Traditional genomic characterizations have produced minimal improvements in patient clinical outcome, with a mortality rate of 50%. Therefore lncRNAs may represent a promising field to discover novel diagnostic and therapeutic strategies. Here we analyzed the lncRNA expression patterns of HNSCC from 426 TCGA primary tumor samples to generate insights into the landscape of lncRNAs in HNSCC.

Interestingly, we discover 5 cluster of samples based in lncRNA expression with significant differences in molecular aberrations and clinico-pathologic features. LncRNA-clustering resembles clustering based on DNA-methylation and mRNA features. DNA-methylation is a chemical modification of genomic DNA by the addition of a methyl group (-CH3) to the cytosine or adenine DNA nucleotides. Typical DNA methylation occurs in a CpG dinucleotide context, where predominantly CpG sites are methylated in the genome. Most of the CpG clusters, known as CpG islands, occur near transcriptional start sites (TSSs) where they are predominantly un-methylated. The establishment and maintenance of methylation patterns resulting in modulation of gene expression is one of the key steps in epigenetic regulation during normal developmental programs. The significant overlapping between lncRNA and DNA-methylation clustering might be due to frequent methylation of CpG clusters closed to TSSs of lncRNAs altering their landscape. However, we cannot discard cross-regulation between DNA-methylation and specific lncRNAs, such as has been described for H19 [24] or ecCEBPA [25] recently.

Here, we discuss the most relevant features found per lncRNA cluster. Cluster c1 is enriched in TP53 and NSD1 mutations, contain frequent smokers, patients are younger, tumors are mainly located in larynx, and patients display poor overall survival. Cluster c2 has mutations in TP53 and KMT2D and patients are older. As for c1, patients within c2 are frequent smokers, tumors are in larynx, and display poor overall survival. Cluster c4 has mutations in NOTCH1 and CDKN2A, tumors are located in tongue, and they are more differentiated. Cluster c5 has depletion of TP53 and CDKN2A mutations, infection with HPV, display good overall survival, contain tumors normally found in tonsil, and they are undifferentiated. TP53 mutations have been associated with poor survival in many cancer types [26]. We believe that the aggressiveness of the disease in c1 and c2 patients might be, at list partially, explained by the increased TP53 mutation frequency. The reverse would apply for c5, almost depleted on mutations, and having the best overall survival of all clusters.

In relation with the clinical parameters analyzed, the anatomical location was the most significantly associated with lncRNA clustering, with c1 and c2 in larynx, c3 and c4 in tongue, or c5 in tonsil. This finding might be explained by the already known high tissue specificity of lncRNA expression. Therefore, we consider that some of the surrogate lncRNAs in the clusters

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

might be specific for sub-anatomical locations in the head and neck region. Moreover, in the GO analysis, we found “muscle organ development” highly enriched in c4 overexpressed PGCs, a result in line with c4 carcinomas being frequent in tongue, a muscle organ. We suggest that some of the surrogate lncRNAs in c4 are related with muscle invasion of carcinomas in tongue. Similarly, c5 overexpressed coding-genes are enriched in proteins involved in “lymphocyte activation”, a process that occurs in the tonsils, which is an immune defense organ in the aerodigestive tract constituting the first line of defense against ingested or inhaled foreign pathogens. Naturally, both B- and T-cell activation occurs in the tonsils after the uptake of antigens produced by pathogens by specialized antigen capture cells called M cells. Whether the enrichment in “lymphocyte activation” found in c5 is due to tonsil-dependent functions or to a specific immune response to HPV-positive tumors remains to be ascertained.

Between the PGCs underexpressed in c1, we found programmed death-ligand 1 (PD-L1; also called CD274). PD-L1, which is expressed on many cancer and immune cells, plays an important part in blocking the 'cancer immunity cycle' by binding programmed death-1 (PD-1) and B7.1 (CD80), both of which are negative regulators of T-lymphocyte activation. Binding of PD-L1 to its receptors suppresses T-cell migration, proliferation and secretion of cytotoxic mediators, and restricts tumor cell killing. The PD-L1-PD-1 axis protects the host from overactive T-effector cells not only in cancer but also during microbial infections. Blocking PD-L1 should therefore enhance anticancer immunity and successful treatment of many patients with advanced cancer using antibodies against PD-L1 has been demonstrated. However, little is known about predictive factors of efficacy of such therapies, although some recent reports described that across multiple cancer types, responses are observed in patients with tumors expressing high levels of PD-L1, especially when PD-L1 was expressed by tumor-infiltrating immune cells. Therefore, it is tempting to speculate that c1 primary tumors would be less sensitive to anti-PD-L1 therapies than the tumors from the other clusters. In addition, some of the c1 lncRNAs might be involved in this PD-L1-PD-1 'cancer immunity cycle', and could be subject of future investigations.

Many overexpressed PGCs in c5 are related to “cell cycle” processes, suggesting that c5 carcinomas are more proliferative. This finding is in line with the oncogenic activities of HPV E7 oncoprotein, which binds and induces protein degradation of cell cycle regulators such as the retinoblastoma protein family, including pRb (RB1), p107 (RBL1) or p130 (RBL2). Degradation of these proteins, mainly pRb, allow E2F transcription factors to induce expression of genes involved in cell cycle progression, such as cyclins and cyclin-dependent kinases, or proteins involved in DNA replication, and mitotic division. Therefore, some c5 lncRNAs might also be involved in these processes, and possibly in the E7-pRb-E2F axis of cell cycle deregulation of HPV-infected tumors. Tumors from c4 overexpresses (and c5 underexpresses) PGCs involved in epidermal differentiation, such as LCE cluster SPRR genes, normally

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

involved in terminal epithelial differentiation. This result might be in line with the high frequency of differentiated carcinomas in c4 (histologic grades G1 and G2) compared with c5, whereby G3 tumors are more frequent. Therefore, some of the c4 and c5 surrogate lncRNAs could be involved in differentiation.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusions

We present the first clustering analysis of HNSCC based on lncRNA expression performed to date. The results allow selection of surrogate lncRNA genes of 5 distinct tumor groups, propose possible functions associated with them, as well as phenotypic and clinico-pathology features that may be consequence, in part, of their activities.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the TANRIC (http://ibl.mdanderson.org/tanric/_design/basic/index.html), TCGA (https://tcga-data.nci.nih.gov/docs/publications/hnsc_2014/) repositories, and webtools as cBioPortal (http://www.cbioportal.org/) and UCSC Xena (http://xena.ucsc.edu/).

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by FEDER cofounded MINECO grants SAF2012-34378 and SAF2015-66015-R (JMP), Comunidad Autónoma de Madrid grant S2010/BMD-2470 (Oncocycle Program) (JMP), ISCIII grants ISCIII-RETIC RD12/0036/0009 (JMP), PIE15/00076 (JMP), CB/16/00228 (JMP), Swiss National Science Foundation Grants 144637 and 153099 (RGE), and Swiss Cancer League Grant KLS 3243-08-2013 (RGE).

Authors' contributions

PGdeL, APG and RGE perfomed the genomic and statistical analyses. PGdeL and RGE wrote the manuscript. PGdeL, JMP and RGE interpreted the results. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in {GLOBOCAN} 2012. Int. J. Cancer. 2015;136:E359-86.
2. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. Nat. Rev. Cancer. 2011;11:9–22.
3. Mao L, Hong WK, Papadimitrakopoulou VA. Focus on head and neck cancer. Cancer Cell. 2004;5:311–6.
4. Gillison ML, Koch WM, Capone RB, Spafford M, Westra WH, Wu L, et al. Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. J. Natl. Cancer Inst. 2000;92:709–20.
5. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576–82.
6. Ramiro I-B, Martin D, Gutkind JS. Exploiting the head and neck cancer oncogenome: widespread {PI3K-mTOR} pathway alterations and novel molecular targets. Cancer Discov. 2013;3:722–5.
7. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science (80-.). 2011;333:1157–60.
8. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
9. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res. 2014;42:D749-55.
10. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, et al. Experimental validation of the regulated expression of large numbers of non-coding {RNAs} from the mouse genome. Genome Res. 2006;16:11–9.
11. Huarte M. The emerging role of {lncRNAs} in cancer. Nat. Med. 2015;21:1253–61.
12. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, et al. {CRISPRi-based} genome-scale identification of functional long noncoding {RNA} loci in human cells. Science (80-.). 2017;355:aah7111.
13. The Atlas of ncRNAs in Cancer: TANRIC [Internet]. [cited 2016 Feb 15]. Available from: http://ibl.mdanderson.org/tanric/_design/basic/index.html
14. Li J, Han L, Roebuck P, Diao L, Liu L, Yuan Y, et al. {TANRIC:} An interactive open platform to explore the function of {lncRNAs} in cancer. Cancer Res. 2015;75:3728–37.
15. The cBioPortal for Cancer Genomics [Internet]. [cited 2016 Feb 15]. Available from: http://www.cbioportal.org/
16. UCSC Xena [Internet]. [cited 2016 Feb 15]. Available from: http://xena.ucsc.edu/
17. Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas, 2015 [Internet]. [cited 2016 Feb 15]. Available from: https://tcga-data.nci.nih.gov/docs/publications/hnsc_2014/
18. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572–3.
19. Bioconductor: Open source software for bioinformatics [Internet]. [cited 2016 Feb 20]. Available from: https://www.bioconductor.org/
20. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003;34:374–8.
21. The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 [Internet]. [cited 2016 May 15]. Available from: http://david.abcc.ncifcrf.gov/
22. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 2008;4:44–57.
23. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding {RNAs} in mammals. Nature. 2009;458:223–7.
24. Zhou J, Yang L, Zhong T, Mueller M, Men Y, Zhang N, et al. H19 {lncRNA} alters {DNA}

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

methylation genome wide by regulating S-adenosylhomocysteine hydrolase. Nat Commun. 2015;6:10221.
25. Di Ruscio A, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, Terragni J, et al. DNMT1-interacting RNAs block gene-specific DNA methylation. Nature. 2013;503:371–6.
26. Leroy B, Anderson M, Soussi T. {TP53} mutations in human cancer: database reassessment and prospects for the next decade. Hum. Mutat. 2014;35:672–88.

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure1. Unsupervised clustering of HNSCC using lncRNA expression data. A) Consensus clustering analysis of the subtypes identifies five major groups (samples, n=426). The blue and white heat map displays sample consensus. Specifications and parameters used in the analysis are described in M&M. Number of samples per cluster is shown. **B)** Consensus CDF plot for k=2 to 6. Note that best results are obtained with k=5.

Figure 2: lncRNA clusters and other molecular aberrations. **A)** LncRNA-based clustering of HNSCC samples is significantly associated with clustering based in diverse molecular features, mainly DNA methylation and expression of PCGs (mRNA). Significance values are plot upon Chi-square test computation. Dashed red line: threshold of significance (p-val<0.05). **B)** and **C)** Association with HPV infection and HNSCC mutations with lncRNA clusters. Chi-square or odds ratio values are plot upon Chi-square test **(B)** or Fisher's exact test **(C)** computation, respectively. D) Distribution of lncRNA clusters and HPV infected samples, or samples with mutations in KMT2D or NSD1. Note the enrichment of HPV infection in c5, the NSD1 mutations in c1 and the KMT2D mutations in c2 (red lines), and the depletion of KMT2D mutations in c4 (blue line). P-values are calculated with Fisher's exact test. Vertical black lines in panel D showed HPV+ samples and mutated samples for the selected genes KMT2D and NSD1, respectively.

Figure 3. Guilt-by-association (GBA) analysis. A) Schema-graph showing approach used to search for lncRNA genes as surrogates of each cluster. Briefly, Pearson correlation was computed between all lncRNAs and artificial expression vectors or templates (templates t1 to t5) whereby maximum expression in individual clusters with respect to the others was interrogated. Threshold used: r>0.3 and p-val<$1x10^{-6}$. **B)** Heatmap exhibiting selected lncRNAs upon Pearson correlation and the corresponding gene numbers per cluster. Heatmaps of PCGs directly (overexpressed PGCs) **(C)** or inversely (underexpressed PGCs) **(D)** correlated with the lncRNAs selected per cluster in panel B. Threshold used: r>0.3 and p-val<$1x10^{-9}$ for overexpressed PGCs; r<-0.3 and p-val<$1x10^{-9}$ for underexpressed PGCs.

Figure 4. GOBP terms selected by cluster. Gene Ontology analysis was done using DAVID web tool. Overexpressed or underexpressed PGCs are analyzed per cluster, and representative GO Biological Processes (GOBP) terms shown. Numbers at the end of each bar represent genes belonging to each GOBP term. Dashed red line denotes threshold for significant enrichment (p-val<0.05).

bioRxiv preprint doi: https://doi.org/10.1101/105999; this version posted February 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5. **LncRNA-based clustering and clinical parameters**. **A)** Kaplan-Meier plot of TCGA-HNSCC patients stratified by lncRNA clusters using overall survival as end-point. P-val is calculated with the log-rank test. n: number of patients with available follow-up information. **B)** Significant associations between lncRNA clusters and various clinical parameters, including tobacco, histologic grade, and sub-anatomical location. Odds ratios and significance values are calculated using Fisher's exact test. **C)** Box plots of patient age for significant differences between c1 or c2 patients- P-values are calculated using Ttest.

Supplementary Figure 1. Kaplan-Meier plot of TCGA-HNSCC patients stratified by lncRNA clusters using recurrence and end-point. P-val is calculated with the log-rank test. n: number of patients with available follow-up information.

Figure 1

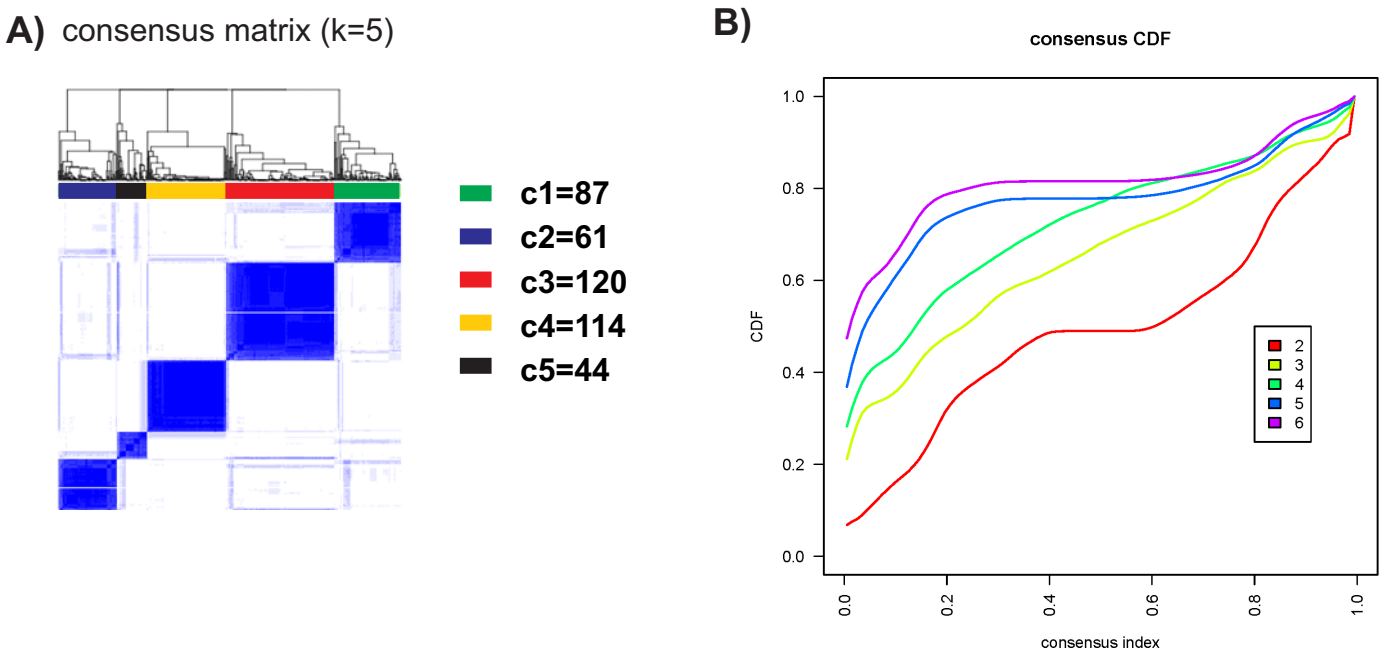

A) consensus matrix (k=5)
c1=87
c2=61
c3=120
c4=114
c5=44
B)
consensus CDF
CDF
consensus index
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6

Figure 2

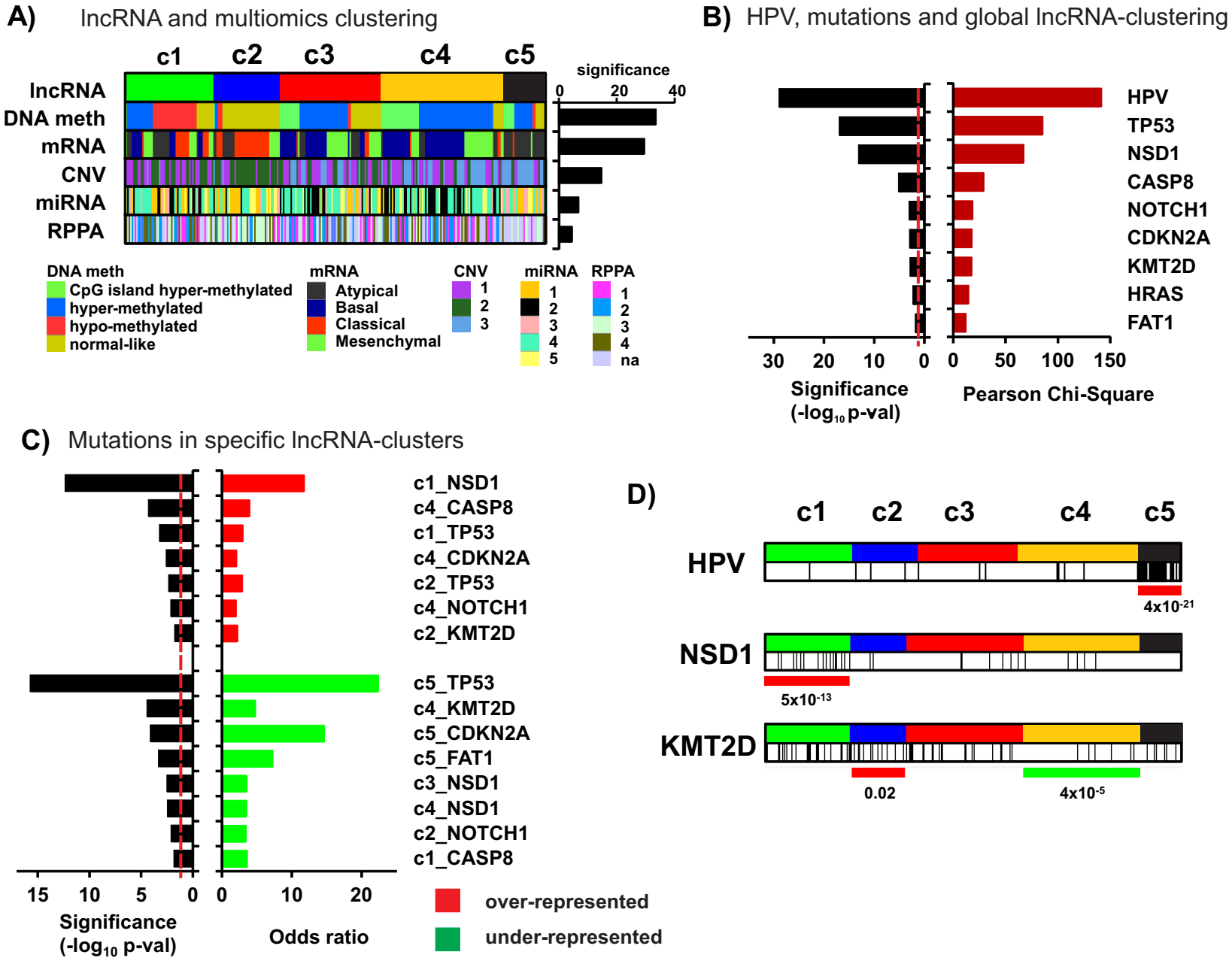

A) lncRNA and multiomics clustering
c1
c2
c3
c4
c5
lncRNA
DNA meth
mRNA
CNV
miRNA
RPPA
significance
0
20
40
DNA meth
CpG island hyper-methylated
hyper-methylated
hypo-methylated
normal-like
mRNA
Atypical
Basal
Classical
Mesenchymal
CNV
1
2
3
miRNA
1
2
3
4
5
RPPA
1
2
3
4
na
B) HPV, mutations and global lncRNA-clustering
HPV
TP53
NSD1
CASP8
NOTCH1
CDKN2A
KMT2D
HRAS
FAT1
30
20
10
0
0
50
100
150
Significance
(-log10 p-val)
Pearson Chi-Square
C) Mutations in specific lncRNA-clusters
c1_NSD1
c4_CASP8
c1_TP53
c4_CDKN2A
c2_TP53
c4_NOTCH1
c2_KMT2D
c5_TP53
c4_KMT2D
c5_CDKN2A
c5_FAT1
c3_NSD1
c4_NSD1
c2_NOTCH1
c1_CASP8
15
10
5
0
0
10
20
Significance
(-log10 p-val)
Odds ratio
over-represented
under-represented
D)
c1
c2
c3
c4
c5
HPV
4x10-21
NSD1
5x10-13
KMT2D
0.02
4x10-5

Figure 3

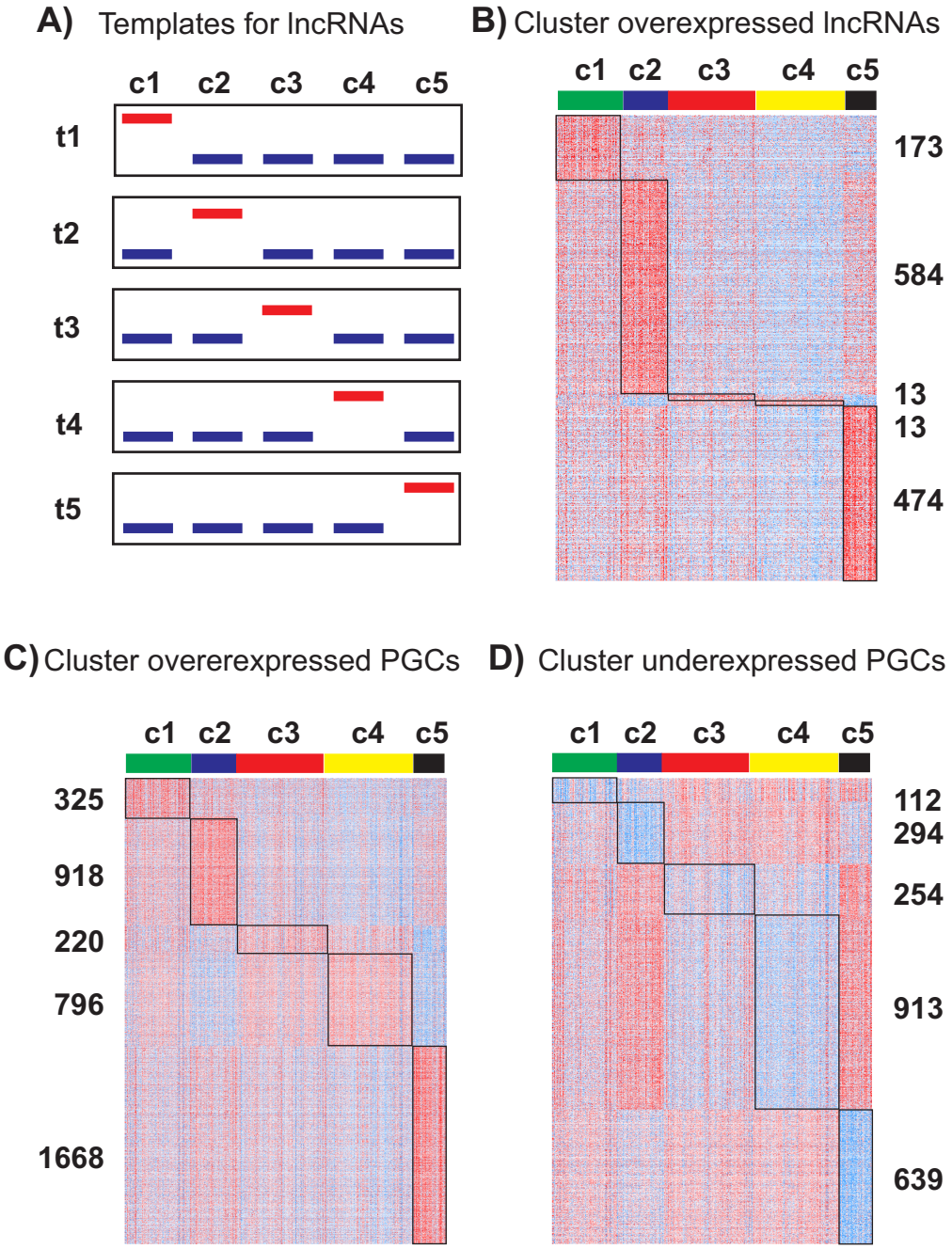

A) Templates for lncRNAs
c1 c2 c3 c4 c5
t1
t2
t3
t4
t5
B) Cluster overexpressed lncRNAs
c1 c2 c3 c4 c5
173
584
13
13
474
C) Cluster overerexpressed PGCs
c1 c2 c3 c4 c5
325
918
220
796
1668
D) Cluster underexpressed PGCs
c1 c2 c3 c4 c5
112
294
254
913
639

Figure 4

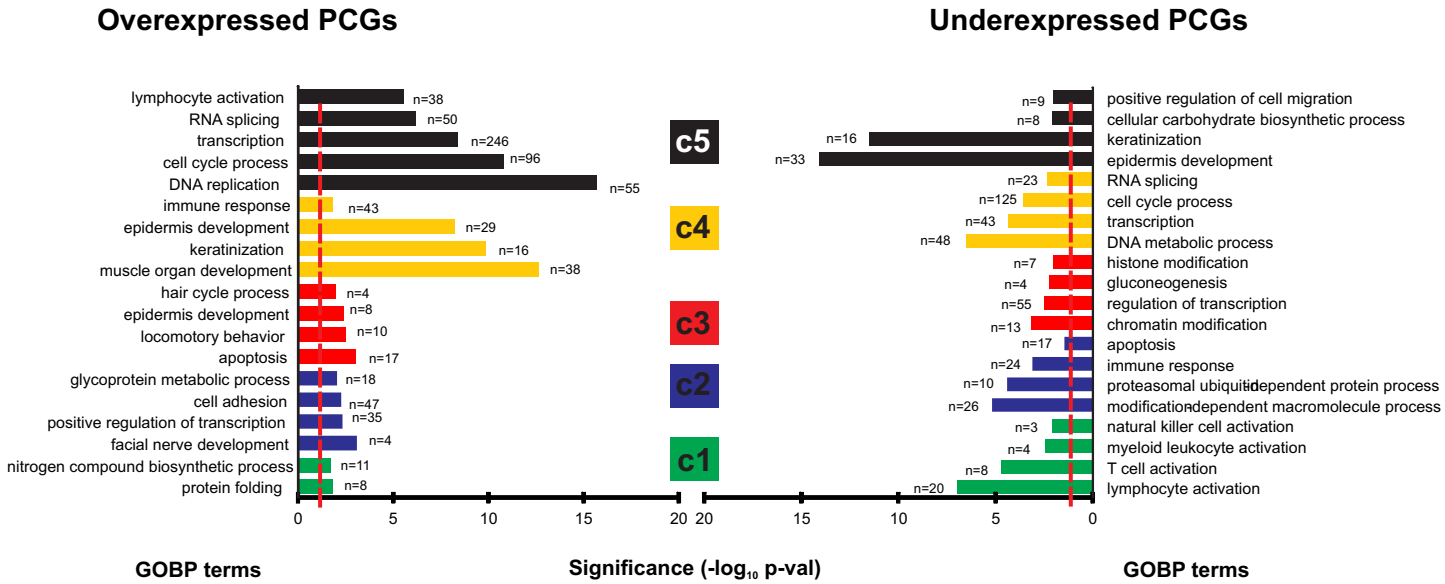
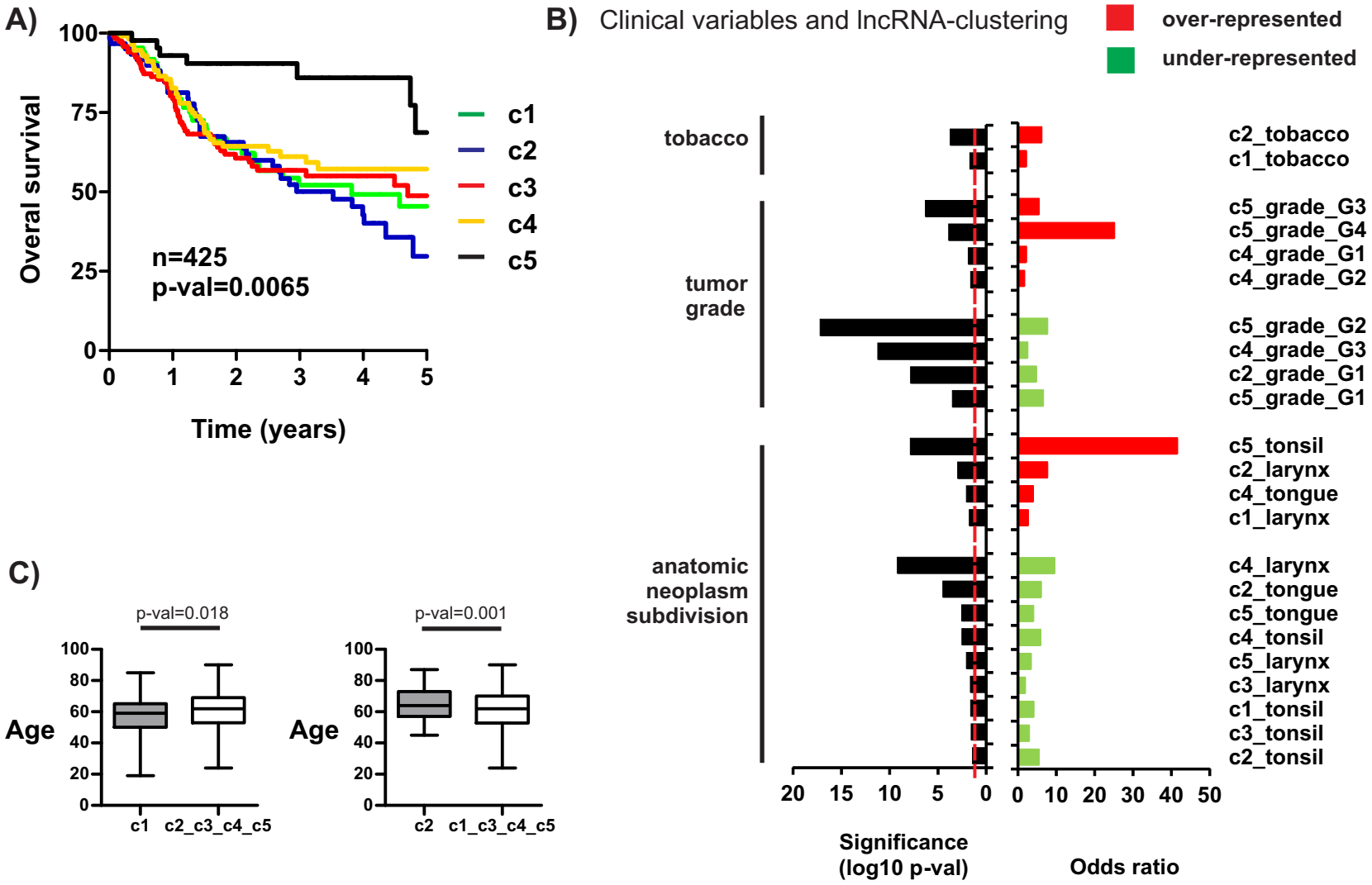

Overexpressed PCGs
Underexpressed PCGs
lymphocyte activation
RNA splicing
transcription
cell cycle process
DNA replication
immune response
epidermis development
keratinization
muscle organ development
hair cycle process
epidermis development
locomotory behavior
apoptosis
glycoprotein metabolic process
cell adhesion
positive regulation of transcription
facial nerve development
nitrogen compound biosynthetic process
protein folding
n=38
n=50
n=246
n=96
n=55
n=43
n=29
n=16
n=38
n=4
n=8
n=10
n=17
n=18
n=47
n=35
n=4
n=11
n=8
c5
c4
c3
c2
c1
n=9
n=8
n=16
n=33
n=23
n=125
n=43
n=48
n=7
n=4
n=55
n=13
n=17
n=24
n=10
n=26
n=3
n=4
n=8
n=20
positive regulation of cell migration
cellular carbohydrate biosynthetic process
keratinization
epidermis development
RNA splicing
cell cycle process
transcription
DNA metabolic process
histone modification
gluconeogenesis
regulation of transcription
chromatin modification
apoptosis
immune response
proteasomal ubiquitin-dependent protein process
modification-dependent macromolecule process
natural killer cell activation
myeloid leukocyte activation
T cell activation
lymphocyte activation
0
5
10
15
20
20
15
10
5
0
GOBP terms
Significance (-log10 p-val)
GOBP terms

Figure 5


A)
Overal survival
100
75
50
25
0
0
1
2
3
4
5
Time (years)
n=425
p-val=0.0065
c1
c2
c3
c4
c5
B) Clinical variables and lncRNA-clustering
over-represented
under-represented
tobacco
tumor grade
anatomic neoplasm subdivision
c2_tobacco
c1_tobacco
c5_grade_G3
c5_grade_G4
c4_grade_G1
c4_grade_G2
c5_grade_G2
c4_grade_G3
c2_grade_G1
c5_grade_G1
c5_tonsil
c2_larynx
c4_tongue
c1_larynx
c4_larynx
c2_tongue
c5_tongue
c4_tonsil
c5_larynx
c3_larynx
c1_tonsil
c3_tonsil
c2_tonsil
20 15 10 5 0
Significance (log10 p-val)
0 10 20 30 40 50
Odds ratio
C)
p-val=0.018
Age
100
80
60
40
20
0
c1
c2_c3_c4_c5
p-val=0.001
Age
100
80
60
40
20
0
c2
c1_c3_c4_c5