Title: Gene Signatures Involved in Pathways of Head and Neck Squamous Cell Carcinoma with Higher Progression Rate

By: Nasim Sanati
Date: September 2016

Code and Data availability
R code is open-source and can be accessed via https://github.com/teslajoy
Data was originally downloaded from TCGA data portal (See [3])

Biomedical Informatics Master of Science Thesis Defense Study
Biomedical Informatics and Computational Biology Track
Department of Medical Informatics & Clinical Epidemiology
Oregon Health and Science University

Abstract

Background

Head and neck squamous cell carcinoma (HNSCC) has shown to have a high progression rate in subpopulations despite current treatment methods [1-3]. Due to lack of notable symptoms most patients are diagnosed at later stages (III-IV TNM) leaving a short period of time for therapeutic decision-making and treatment. The heterogeneity in HNSCC behavioral risk factors such as smoking and alcohol consumption along with its tumors arising in a range of different anatomical locations has made reliable stratification of this cancer extremely challenging. Given this difficult cancer for patients' subpopulations with high progression rate, the goal is to control progression and administer therapy over the longest period of time and in the least invasive manner possible. To achieve this goal, with the notion of identifying targets for high progression rate mechanisms in HNSCC subpopulations, we focused on extracting measured gene signatures and driver genes involved in an aggregate of gene regulation patterns.

Methods

We analyzed 229 patients' samples from The Cancer Genome Atlas (TCGA) previously annotated by Bornstein et al [3] on their progression status. To extract gene signatures involved in solid tumor mechanisms of patients with higher rate of progression, we compared two groups of patients' tumors in 68 progressors with higher and 161 nonprogressors with lower progression rates. Leveraging expression data to define pairwise gene relations as a network correlation structure, we utilized de-novo weighted network analysis over 10,000+ genes. Both gene and exon level expression data were assessed to possibly identify interesting splicing events and consensus genes between the network levels. Association of highly organized progressor gene clusters (modules) to known HNSCC behavioral risk factors of pack years smoked and alcohol drink consumed per day were evaluated. To compare progression mechanisms, differential network analysis between the progressor and nonprogressor condition was assessed to identify progressor modules enriched in differentially expressed (DE), variable (DV), or wired (DW) genes. After identification of progressor modules enriched in DE/DV/DW genes, we characterized their biological identity based on pathway enrichment analysis.

Results

Twelve co-expression progressor consensus modules enriched in DE, DV, or DW genes were identified as putative progressor gene signatures of HNSCC. Eleven modules were enriched in DE and DV genes and showed high correlations to drink per day alcohol consumption or pack years smoking HNSCC habitual risk factors. Only one module was enriched in DW genes with putative driver genes (network hubs) of IL10RA, DOK2, APBBIP UBASH3A, SASH3 involved in inflammation and tumor microenvironment evolution mechanisms. DE/DV/DW putative progressor gene signatures showed involvement in various pathways such as cell cycle check points, abnormal mitosis and spindle bipolarity, c-myc, macrophages, immune response and T cells, receptor synapse dysregulations and associated diseases such as Alzheimer's and Parkinson, Interferon gamma signaling pathway, MAPK, Jak-Stat, P53, and more. Gene composition and characterization of each putative progressor gene signature with detailed pathway enrichment analysis are available via Supplementary Data.

Conclusion

Weighted network analysis approach gives a holistic view of tumors dynamics and allows for identification of gene signatures responsible for regulating different progression mechanisms. With the notion of identifying therapeutic targets for further clinical research, after evaluation of DW hub genes, they may be utilized as prognosis signatures to stratify patients based on high progression status. This could potentially improve clinical research end points and ultimately aid in clinical utility.

Keywords: Head and Neck Squamous Cell Carcinoma, TCGA, RNA-Seq, Gene signature, Progression, Weighted network analysis, gene co-expression, gene co-splicing

1. Background

1.1 Head and Neck Squamous Cell Carcinoma and Analysis Motivation

Overwhelming majority of head and neck carcinomas (HNC) are identified as HNSCC with mucosal malignant tissue arising from squamous cells [4]. This cancer initiates from an array of different anatomical locations with two major sites of oral cavity and oropharynx [3, 5, 6]. According to American Cancer Society, approximately 48,330 new cases of HNC of oral and oropharyngeal are expected in 2016 with an expected 9,570 cases of death events [5]. Due to its cold like symptoms, patients are diagnosed at later stages in cancer (III-IV TNM), consequently, leaving a limited time frame for treatment. HNSCC progresses despite current treatments with a 40-50% 5-year survival rate of locally advanced solid tumor (T1-2, N0) [5]. With HNSCC arising in the head and neck anatomical region, its treatment are known to be associated with clinically significant symptom burden, alterations in daily functionality, and decrease in quality of life [7]. Each year more than 550,000 people worldwide are affected by these symptoms [8].

Overtime tobacco smoking or chewing carcinogenic exposures along with other behavioral habits such as alcohol use and human papilloma virus (HPV) exchange are well-known risk factors of HNSCC [9]. The male to female ratio ranges from 2:1 to 4:1; however, this ratio is not well understood and might be indicative of variability in habitual and social risk factors associated with HNSCC [10-11]. Both biological and social disparities expose the opportunity of HNSCC tumors to be prone in receiving combination of pathway signals [12-13]. This complexity has made reliable stratification of HNSCC a challenge [14].

Given this difficult cancer, for patients' subpopulations with high progression rate, the goal is to control progression and administer therapy over the longest period of time and in the least invasive manner possible. To achieve this goal with the notion of identifying targets for further therapy research, our motive was to identify informative tumor progression mechanisms by extracting measured gene signatures.

A common analysis approach is to leverage expression data and define pairwise gene relations as a weighted network correlation structure. Weighted network analysis allows for measuring an aggregate of relational weights that assist in extraction of mechanistically informative genes signatures [15]. This network model allows for capturing interactions that can't be evaluated otherwise. Other approaches such as differential expression analysis tend to analyze single gene differences and are limited in capturing gene signatures. Analytically they are not de-novo and limited by rigorously filtering genes. These filtered genes could potentially be contributing to tumor progression mechanism but have low expression variability to be detected. Results of such studies lack reproducibility and can't be used for secondary research.

1.2 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a joint effort of the National cancer institute and the National Human Genome Research Institute in NIH with the mission of cancer prevention, diagnosis, and treatment [16]. Large quantities of HNSCC cohort samples have been collected with appropriate patient consents and standard sample normalizations. This allows the research community to extract reliable biological inferences using various data mining techniques. Two recently published TCGA study on HNSCC were completed in 2014 and 2015. The first study investigated the role of HPV in HNSCC, and the second study displayed the most comprehensive landscape of somatic genomic alterations in HNSCC by exploring variants at multi-omic levels [17-18].

2. Methods

2.1 Patients Clinical Demographics and TCGA Molecular Assay Data Types

We utilized TCGA HNSCC data previously curated in Bornstein et al. study [3]. The data includes 229 patients' samples with 68 (30%) progressors and 161 (70%) nonprogressors. The median last encounter days of progressor patients were considerably lower than nonprogressor patients (606 vs. 4856 days; Kaplan- Meyer Chisq = 39.9 on 1 degrees of freedom, p = 2.65e-10). Clinical demographics of these patients had 165 (72%) missing, 48 (20%) negative, and 16 (6%) positive HPV (p16 or ish) status with anatomical sites reported to be between three major regions of 131 (57%) oral cavity, 59 (25%) larynx, and 39 (17%) oropharynx. Mean age was 62 years old with 138 (60%) complete cases of self-reported tobacco pack years smoked and 97 (42%) alcohol drink consumed per day (Supplementary Information Fig 1). Remaining patients reported to be lifelong non-smokers and/or non-drinkers of alcohol or had no clinical documentation available on these two clinical features. Progressor patients had a median of 45 pack years and 4 alcohol drinks per day. Nonprogressor patients reported slightly lower smoking and alcohol consumption estimates, with median of 40 pack years smoked and 3 alcoholic drinks drank per day. Unfortunately the time range of alcohol consumption over years was not documented to evaluate rate or a longitudinal unit.

TCGA molecular assay data types utilized were RNA-Seq V2 (Level 3; Illumina HiSeq 2000) solid tumor tissue (01A sample-tag)[1] of genes normalized results and exons quantifications. All data alignment was mapped to genome build hg19. We extracted readily available normalized fields, normalized_counts and reads per kilo-base of exon model per million mapped reads (RPKM) per gene and exon data respectively [44-45].

$$normalized\ counts\ = \frac{raw\ count}{75th\ \ percentile\ (after\ removing\ zeros)} \times 1000$$

$$RPKM = \frac{Number\ of\ mapped\ reads\ that\ fell\ into\ a\ gene's\ exon}{Total\ number\ of\ mappable\ reads} \times \frac{10^9}{sum\ of\ the\ exons\ in\ base\ pair}$$

2.2 Study Workflow Illustration

In this study co-expression networks were constructed based on Weighted Network Correlation Analysis (WGCNA) approach [20]. To identify possible interesting splicing events and consensus genes between the network levels, we also used exon expression data and constructed co-splice networks based on Iancu et al. (2015) coSpliceEx approach [21]. Figure 1 illustrates the step-by-step workflow of our study for identification and characterization of progressor gene signatures in both co-expression and coSpliceEx networks. We examined the relation of tobacco pack years and alcohol drink consumption clinical features to progressor modules. Differential network analysis was conducted to identify progressor modules enriched in DE/DV/DW genes that were taken as potential progressor gene signatures. We characterized the identified progressor modules' biological functions by conducting pathway enrichment analysis utilizing Cytoscape ReactomeFIViz application [22].

[1] https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode

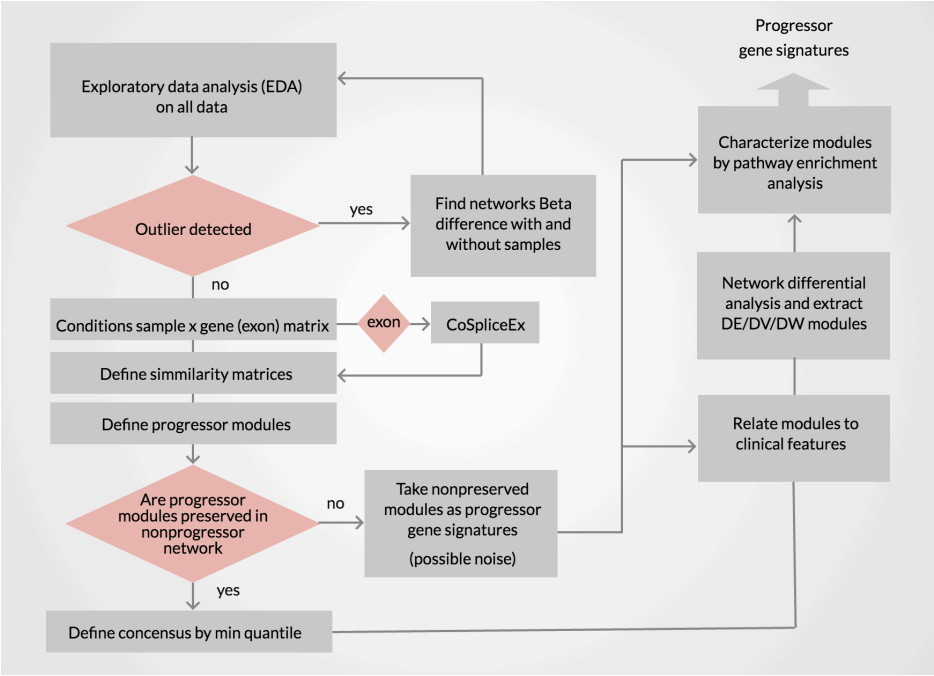


Figure 1: Step-by-step study workflow to identify and characterize putative progressor gene signatures

2.3 Co-expression and CoSpliceEx Exploratory Data Analysis and Filtration

Based on WGCNA best practices [23], normalized counts of genes were logged and shifted via log2 (x + 1) formula that is close to a variance stabilization transformation. The goal here was to create new values $y = f(x)$ such that the variability of the y values was not related to their mean and would allow for simple regression or variance based analysis. Next we removed genes with zero counts or low variance utilizing WGCNA goodSampleGenes function, and olfactory genes. The latter have been noted to introduce noise in TCGA data across cancer types due to their locations in the chromosome [24-26].

Utilizing WGCNA pickSoftThreshold function, we calculated soft threshold (β) over all gene and exon expressions and obtained an estimate of $\beta = 5$ and $\beta = 10$ for genes and exons respectively (supplementary Information Fig 2). The goal of this procedure is to reduce the influence of low and possibly noisy correlations and to bring the network towards a scale-free structure [19]. To reserve true biological identity of networks, construct de-novo networks with distinct clustering of modules, and to moderate computational time complexity we focused on extracting 10,000 to 11,000 genes. We further reduced the size of the networks by filtering the genes based on network connectivity, which is the sum of all network adjacencies (correlations raised to power β). Then we computed each gene's node connectivity measure and extracted the top 50% in genes and 70% in exons connectivity quantile. The results accounted for 10,024 genes and 66,880 exons (a subset of 10,614 genes). Between these two sets 5,706 genes overlapped.

Visualizing all samples' gene expression boxplots distribution and inter-array correlation (IAC) [27] defined as all samples' gene expression pair wise Pearson correlation histogram distribution (Supplementary Information Fig 3), no extreme outlier was detected – the IAC had values > 0.65.

However, in exon expression boxplot distributions we noted seven nonprogressor samples with low median (Supplementary Information Fig 4A) and with IAC as low as 0.3. Examining possible batch effects, neither of samples came from the same tissue source site or had the same plate id number (Supplementary Information Table 1). Exploring the difference between the scale free topology criterions, we constructed coSpliceEx on all data with and without these 7 samples. This procedure revealed extreme differences between coSpliceEx

networks scale independence and mean connectivity measures with/without the seven samples (beta estimate of 7 vs. 20+; Supplementary Information Fig 6). After removal and re-visualization of exon expression distributions we also observed a marked improvement in IAC values (Supplementary Information Fig 5B).

2.4 Co-expression Progressor and Nonprogressor Weighted Network Construction

We constructed weighted co-expression networks using the WGCNA approach [19]. The definition of modules in WGCNA was taken to represent gene signatures of HNSCC. Investigating the differences between biological mechanisms of gene signatures in progressor and nonprogressor conditions, two separate matrices for each condition were constructed with dimensions of 68 patients X 10024 genes and 161 patients X 10024 genes respectively. In order to aggregate the different types of gene regulation we defined our networks as unsigned.

Preserving the scale free topology criterion per each condition, soft threshold powers were calculated for each condition separately (Supplementary Information Fig 7). Conditions adjacency matrices were constructed by computing the absolute value of pairwise Pearson correlations between gene expression profiles (vectors of a gene in all samples) each raised to the soft threshold power of $\beta = 5$ and $\beta = 6$ for progressor and nonprogressor conditions respectively. These adjacency matrices were transformed to topological overlap measure (TOM) similarity matrices. This replaces the original adjacency matrices by a measure of interconnectedness that is based on immediate shared neighbors of each gene [20]. However, for identifying clusters a measure of dissimilarity is required, which is simply defined by subtracting one from the similarity measures ($1 - TOM$). Taken each progressor and nonprogressor dissimilarity adjacency, hierarchical clustering was evaluated based on average linkage agglomerative between Euclidean distances of clusters. Each module was identified using WGCNA dynamic tree cut function. A minimum module size was set to 30 to preserve any downstream statistical test assumption of normality distributions. This process resulted in a dendrogram and a module color band with each module assigned a unique color code based on its size. Furthermore, we analyzed the similarity of modules by evaluating the hierarchical clustering of each module's eigengene (ME; 1st principal component). Visualizing clarity of dendrogram clustering and module assignments, none of the modules required merging and all modules showed proper gene assignment in both progressor and nonprogressor networks (Supplementary Information Fig 8).

2.5 CoSpliceEx Progressor and Nonprogressor Weighted Network Construction

Progressor and nonprogressor conditions coSpliceEx weighted networks were constructed based on Iancu et al. approach [21]. This approach requires a map of exons' coordinates to its gene, thus a dictionary of exon's chromosome locus to its gene HGNC symbol was constructed. The dictionary of exon locus map to gene symbols comprised of 66,880 exons X 2 fields: chromosome locus and HGNC (10,614 total). After construction of conditions exon expression matrices of 66,880 exons X 68 progressors and 66,880 X 161 nonprogressors patients, utilizing the exon to HGNC dictionary, exons were mapped back to their genes.

Given the first step of the coSpliceEx pipeline, for each gene the Canberra distance of exons in pairwise samples was measured. This resulted in 10,614 genes lower triangle square matrices with dimensions of 68 and 161 samples for progressor and nonprogressor condition respectively. Next 10,614 genes matrices were collapsed into one 10,614 gene X 10,614 gene similarity matrix by computing the pairwise Mantel correlation. This process was done for each condition separately. The vast magnitude of data processing required large scale computing power that was achieved utilizing Oregon Health and Science University (OHSU) ExaCloud multi-processor servers.

Preserving the scale free topology criterion per each condition, soft threshold powers were calculated for each condition separately and an estimate of $\beta = 7$ was obtained for both condition (Supplementary Information Fig

9). Each condition's similarity matrix was raised to the $\beta = 7$ power and TOM transferred. Clustering was obtained similar to co-expression by taking the TOM dissimilarity measure average linkage Euclidean distance and dynamic tree cut.

2.6 Co-expression and CoSpliceEx Progressor Module Preservation Analysis in Nonprogressor Network

To determine which of the network properties of a progressor condition module changes in nonprogressor condition, we analyzed the reproducibility or preservation of progressor network modules (reference set) in nonprogressor network (test set)[28]. We used WGCNA modulePreservation function for this analysis, which requires either expression data or similarity adjacency with module color assignments of reference set (test set module assignment is not required).

Co-expression and CoSplicEx similarity was calculated by unsigned biweight midcorrelations [19-20]. This correlation measure is median based and less affected by outliers compared to mean based correlations. Module quality and preservation was validated by bootsrapping (N=200 permutations) utilizing the WGCNA modulePreservation function [28]. Module quality evaluates whether modules, as detected by the clustering procedure in the progressor network, significantly differ from random groups of genes in the same network. Module preservation evaluates whether modules detected in progressor network are different from random group of genes in the non-progressor network.

2.7 Co-expression Progressor and Nonprogressor Consensus Weighted Network Construction

A consensus network from minimum quantile of progressor and nonprogressor similarity measures was constructed. The minimum quantile allows for high conservation definition of consensus network construction by a suitable quantile. Similarity adjacency was defined by biweight midcorrelation raised to the soft threshold of $\beta = 6$, TOM transferred, and clustered based on a similar process described in section 2.4. Consensus modules were detected utilizing an automated process by using WGCNA blockwiseConsensusModules. Max block size was set to 10024 to account for analysis of all genes in one block. This function only processes expression data across different sets as a list. Both progressor and nonprogressor expression data were used in this case. Additionally, this function un-assigns genes with low intraconnectivity measure of $\mathrm{kME} = \mathrm{cor}\,(\mathrm{x_i}, \mathrm{ME})$ from modules. To reduce the degree of discordance of gene membership with module eigengenes, a high dendrogram cut height of 0.995 was used for module merging. Visualization of outcome was performed to analyze clarity of dendrogram clustering and module assignments (Supplementary Information Fig 10).

2.9 Co-expression Consensus Module Membership and Clinical Feature Relationship

Intramodular connectivity measures of kWithin ($kIM_i = \sum_{i \neq j} a_{ij}$) and $\mathrm{kME} = \mathrm{cor}\,(\mathrm{x_i}, \mathrm{ME})$ of each progressor consensus module was computed. The relationship, Pearson correlation and corresponding student t-test between the Kwithin and kME module membership measures was obtained and visualized utilizing WGCNA verboseScatterplot function. A linear relationship between kME vs. KWithin was observed and indicated proper consensus module definition as expected (Supplementary Information Fig 11).

Next we assessed the relationship of modules with clinical features of tobacco pack years smoked and alcohol drink consumption per day. A gene significance was defined based on Pearson correlations between each gene's expression profile (a gene in all samples) with pack years $GS_i = \left|cor\left(x_i, F_{pack_{years}}\right)\right|$ and drink per day $GS_i = \left|cor\left(x_i, F_{alcohol_{per_{day}}}\right)\right|$clinical features' separately. Then consensus module significance was obtained by the average gene significance measures of a module. Utilizing WGCNA plotModuleSignificance function, we visualized boxplot distribution of each module's gene significance distribution. The measure of significant

difference between consensus modules was obtained via Kruskal Wallis p-values (Fig 5). We also assessed the magnitude and sign of Pearson correlations (and corresponding student t-test) between gene significance and consensus modules eigengenes (one entity representing the overall module variability) (Fig 6).

Modules with high gene significance and kME absolute values have shown to be biologically meaningful [19]. We plotted each modules relationship between the gene significance and kME values of each progressor consensus utilizing WGCNA verboseScatterplot function (Fig 7-8; Supplementary Information Fig 12-13).

For assessing the difference between consensus modules, we evaluated the differential eigengene network analysis between progressors and nonprogressors consensus signed eigengene networks with tobacco pack years smoked and alcohol consumption per day clinical features. Clustering and Pearson correlation relationships of this approach are described in Langfelder (2007) study [32]. WGCNA plotEigengeneNetworks function was utilized to evaluate and visualize the summary of this analysis (Fig 9).

2.10 Co-expression Differential Network Analysis

Although consensus networks derive at consensus modules in both conditions, the effect of genes between conditions may vary and retain informative biological differences [33]. At the single gene level, we assessed module enrichment by exploring differentially expressed (DE), differentially variable (DV), and differentially wired (DW) gene enrichment in progressor modules. This assessment is called differential network analysis and was obtained using Iancu et al. 2013 approach [33]. Each test pursued to answer, over all data how many genes were DV/DE/DW between progressors and nonprogressors, how many DV/DE/DW genes were in progressor modules, and if the overlap of each progressor modules genes with DV/DE/DW genes were greater than expected by chance.

For DE test, on normalized gene expressions, we used eBayes function from limma library in R. We obtained the mean of unlogged and un-shifted normalized gene expressions in progressors and nonprogressors separately. For each gene's mean a linear model fit was obtained and p-values of empirical Bayes moderated t-test was assessed ($p < 0.01$).

Next we evaluated DV test by taking the standard deviation of progressors and nonprogressors unlogged and un-shifted gene expression values. F-statistics on each set separately was evaluated to compare the variance of samples. For this process, var.test function in stats R library was used. To reduce the family wise error rate, p-values were adjusted using Benjamini-Hochberg procedure ($p < 0.01$).

Finally we computed differentially wired test by evaluating the raw adjacency of progressors and nonrprogressors. Network weights were Pearson Correlations raised to the soft threshold powers computed in section 2.4. Using the psych R library r.test function, we tested the significance of vector correlations by number of successful edge changes ($p < 0.01$). Edge change rate was computed by taking the number of edge changes over total number of network edges. The chance of edge rate change equally likely to occur was tested by preforming a Binomial test on rate ($p < 0.01$).

For all three DE/DV/DW tests, Fishers test between all and individual progressor consensus module evaluated the significance overlap. Modules with Bonferroni corrected p-values < 0.05 were taken as enriched DE/DV/DW progressor consensus modules (Table 3).

2.11 CoSpliceEx Splicing Significance

Gene splicing significance of nonpreserved coSpliceEx modules genes were determined by Zapala et al. approach (2006) [34]. First in each progressors and nonprogressors, Manhatten distance between clinical features of pack years and alcohol per day and genes were computed [21]. This evaluates a difference vector (matrix) for each clinical feature [21]. After identification of samples with groups of exons showing the maximum splicing significance [34], each clinical feature distance/difference was Mantel correlated by each sample with the highest splicing significance. P-values were adjusted using Benjamini-Hochberg procedure. A matrix of p-values and corresponding false discovery rates (FDR) for each clinical feature is available in Supplementary Data.

2.12 Pathway Enrichment Analysis

Non-preserved and consensus modules' pathway enrichment analysis was assessed utilizing ReactomeFIViz[2] application in Cytoscape [22]. Detailed pathway enrichment analysis tables (pathway, p-value, FDR, gene list) along with coSpliceEx non-preserved and co-expression consensus modules gene lists (utilizable as input to ReactomeFIViz) can be found in Supplementary Data.

[2] http://wiki.reactome.org/index.php/ReactomeFIViz

3. Results

3.1 Progressor modules and clustering

To identify network modules conserved between progressors and nonprogressors first we studied modules in progressor network (section 2.4-5). We identified 21 modules from the progressors' co-expression network with module sizes ranging from 45 to 1127 genes. And for the progressors' coSplicEx network, we identified 19 progressor modules with sizes ranging from 34 to 3279 genes (Fig 2). Nonprogressors modules and clustering were evaluated for EDA purposes and are available via Supplementary Information Fig 8.

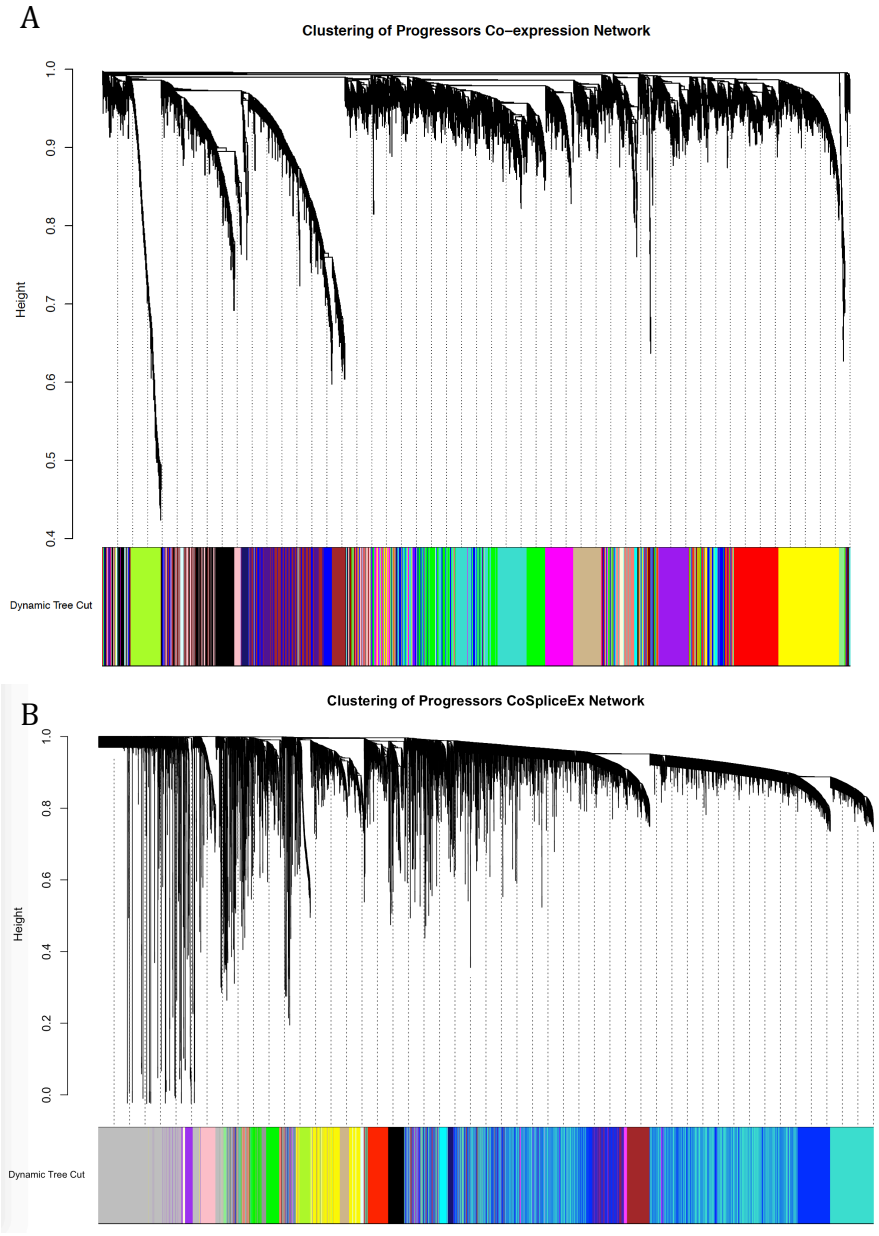


Figure 2: *Progressors Co-expression (A) and Progressors CoSpliceEx (B)* dynamic clustering and module definitions denoted by color-bands. Modules are identified based on WGCNA dynamic tree cut function (bottom-up and left-right branch distances). This function is set to distinguish clusters based on Euclidian distance average linkage dissimilarity of topological overlap interconnectedness similarity measures (TOM). The minimum module size was restricted to 30. Noncontiguous color band is due to dynamic clustering of modules.

3.2 Preservation and Reproducibility of Progressor Modules in Nonprogressor network

Then we tested the preservation of progressors' modules in nonprogressors network. Biologically we are interested in preserved modules with conserved genes. Preserved modules are made of similar genes, but their gene interactions and expression properties are possibly not the same [28].

Overall we observed high preservation between 68 progressor and 161 nonprogressor clinical conditions (Fig 3). All co-expression progressor modules reported high preservation and module quality measures in nonprogressor

weighted network (Fig 3-A; Supplementary Data). Conversely, two coSpliceEx progressor modules, purple with 149 genes and lightyellow with 34 genes revealed low preservations with all three preservation measures Zsummary, Zdensity and Zconnectivity lower than 2 and high median ranks of 20 and 19.5 (Fig 3-B; Supplementary Data). Although nonpreserved modules can be taken as progressor specific gene signatures, we hypothesize that non-preservation here could be due to utilizing readily available RPKM values. That is RPKM with large quantity of low reads has potentially cascaded noise through coSpliceEx pipeline. Given this hypothesis, heatmaps of exon expression over samples of each conditions coSpliceEx module did not report noise (Supplementary Data). Neither of the exons' genes in nonpreserved modules overlapped with co-expression consensus modules. Although pathway enrichment analysis was conducted on both nonpreserved coSpliceEx progressor modules genes, we advise thorough evaluation if the nonpreserved results are to be used in secondary research (Supplementary Data).

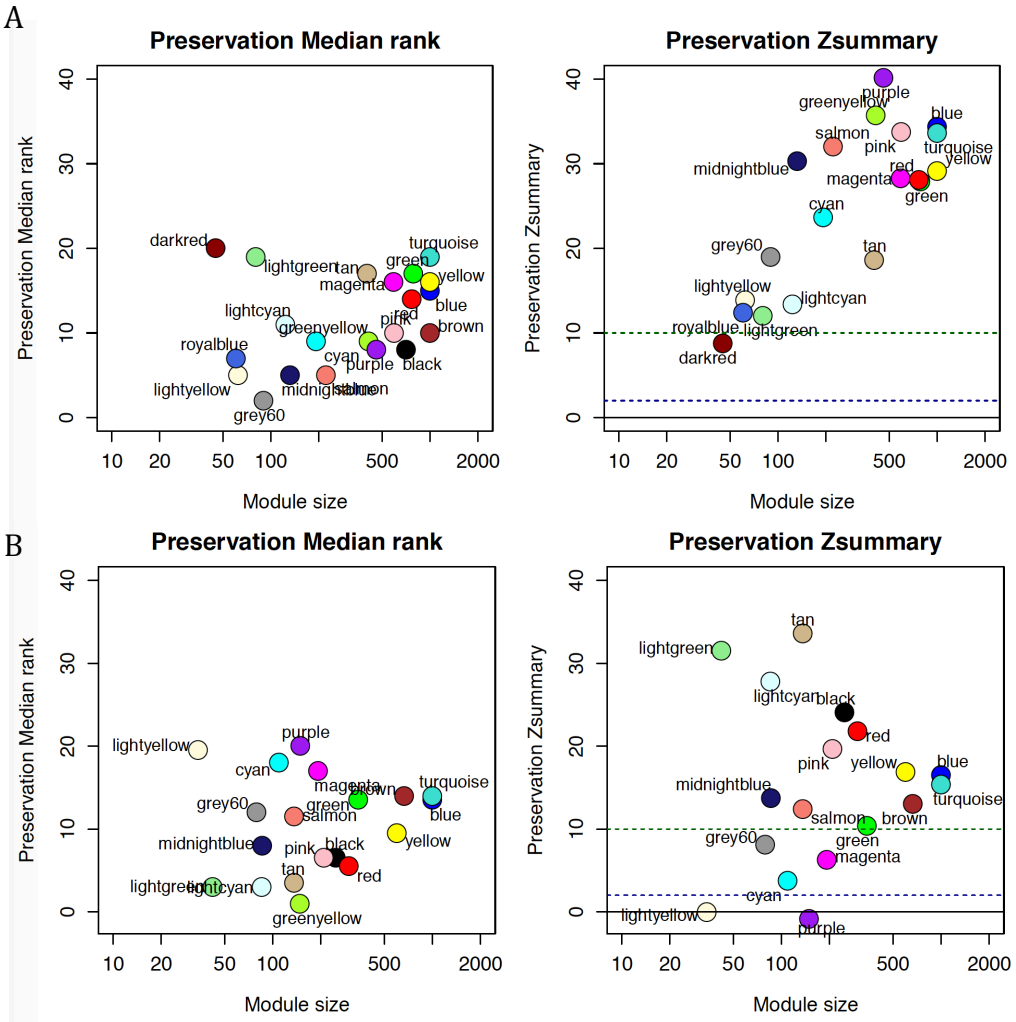


Figure 3: Co-expression (A), coSpliceEx (B) summary of preservation statistics by evaluating progressor modules (reference) in the nonprogressor network (test) over 200 permutations. Each module is represented by its color-code and name. Left figure shows the composite statistic *Preservation median rank* (y-axis) as a function of module size. This measure tends to be independent from module size with high median ranks indicating low preservation. Right figure shows *preservation* $Z_{summary}$ statistic (y-axis) as a function of module size. The dashed blue (low) and green (high) lines are thresholds highlighting $2 < Z < 10$ region. This measure is size dependent with $Z < 2$ indicating low preservation and $Z > 10$ indicating high preserved modules. All modules in Co-expression (A) show high preservation statistics summary than expected by random chance using bootsrapping validation procedures. Conversely, for co-spliceEx (B) two modules purple and lightyellow are not preserved and show low preservation statistics summary. Quality statistics of all modules were high and can be found in supplementary material.

3.3 Consensus Co-expression Network of Both Progressor and Nonprogressor Condition

After identifying highly preserved modules in both progressors and nonprogressors co-expression networks, we constructed one consensus network between the two conditions to identify consensus modules (section 2.7). Biologically consensus modules between the two conditions are made of the same genes but their interactions and expressions may be variable. Also, since each consensus module retains different samples, progressors and nonprogressors separately, their module properties are not the same (i.e. module eigengene). We identified 18 proper modules with gene numbers ranging from 71 to 1389 (Fig 4, Table 1). Overall, out of 10,024 genes, only 882 genes were unassigned to any module (Table 1). All genes showed high module membership with kME and KIM Pearson correlations > 0.9 (student t-test p-value < 0.01; Supplementary Information Fig 11).

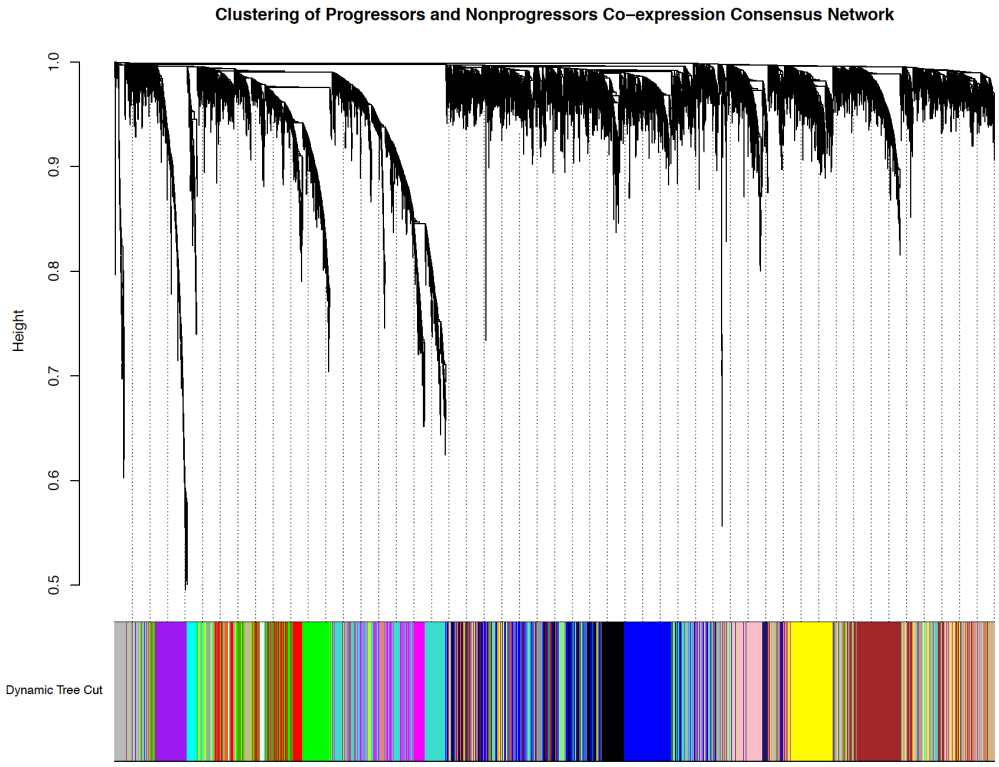


Figure 4: *RNA_SeqV2 HNSCC gene expression data of progressor and nonprogressor conditions* consensus hierarchical clustering and corresponding modules. WGCNA blockwiseConsensusModules with a max block size equal to all data (gene quantity of 10024) was utilized. Modules were identified based on the minimum quantile biweight midcorrelation TOM transformed measures between progressors and nonprogressors. The minimum module size was restricted to 30. This function un-assigns genes with low kME = cor (x_i, ME) from modules. To reduce the degree of discordance of gene membership with module eigengenes, a high dendrogram cut height of 0.995 was used for module merging. Here the hierarchical clustering is based on dynamic tree cut (bottom-up and left-right branch distances) function. Clusters are distinguished based on Euclidian distance average linkage between biweight midcorrelation TOM transferred dissimilarity measures. Each consensus module is assigned a unique color based on its size. Noncontiguous color-band is due to dynamic clustering of modules.

Table 1: Summary of auto block-wise consensus module detection with one block assigned to the total quantity of genes. Here we identified 18 proper modules with sizes ranging from 71 to 1389 genes and 882 unassigned (grey module).

Module	Black	Blue	Brown	Cyan	Green	Greenyellow	Grey60
Size	541	1357	1065	182	688	350	71
Module	Lightcyan	Lightgreen	Magneta	Midnightblue	Pink	Purple	Red
Size	120	71	447	120	533	354	634
Module	Salmon	Tan	Turquoise	Yellow			
Size	193	262	1389	765			

3.4 Relating Smoking and Alcohol Exposure to Co-expression Consensus Modules

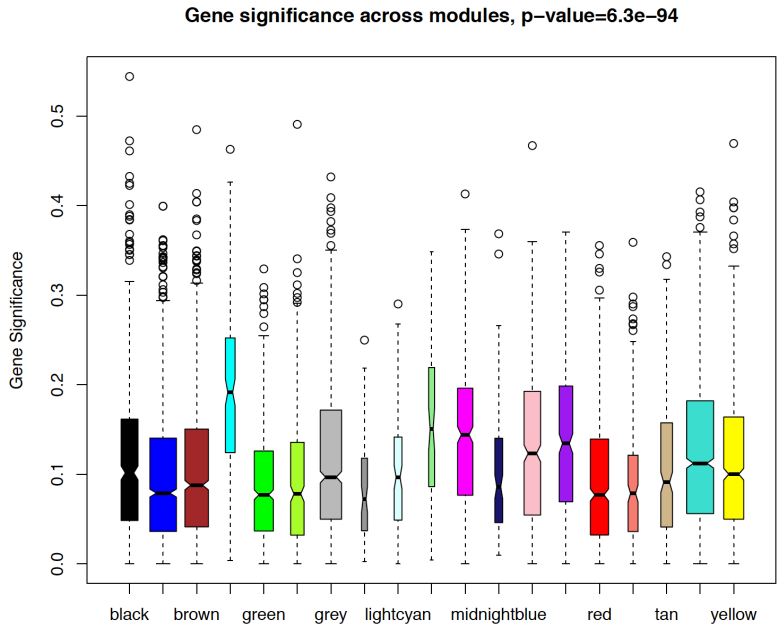
We evaluated four different steps to relate progressor consensus modules to pack years and drink per day clinical features (section 2.9). Summary of all three steps in progressor consensus module relation to clinical features are shown in Table 2. First, we assessed the relation between gene significance measures and clinical features. Gene significance measures of pack years and drink per day were significantly different between progressor consensus modules (Fig 5-A & B; Kruskal Wallis test p-value: 6.3e-94 pack years, 1.5e-251 drink per day). Cyan module showed the highest pack years module significance (Fig 5-A; mean > 0.19). Progressor consensus modules color-coded as black, brown, cyan, pink, tan, yellow showed the strongest drink per day module significance (mean > 0.25).

Heatmap of Pearson correlations between gene significance and progressor consensus module eigengenes revealed similar patterns (Fig 6). Although drink per day had less complete cases documented (42%) vs. pack years (60%), it showed an overall higher correlation measures. Progressor consensus modules color-coded as brown, yellow, black, pink, tan, and cyan show the strongest relation to drink per day (Fig 12; > 0.4). And only cyan module showed the strongest correlation to pack years (Pearson correlation 0.3 and t-test p-value 0.01; Fig 6).

We also found strong positive correlations between gene significance and kME absolute values between progressor consensus module and the same clinical feature noted (Fig 7; Supplementary Information Fig 12-13). In summary, purple, pink, and cyan modules had a strong positive relation between pack years gene significance and progressor modules kME absolute values (Pearson correlation > 0.25; student t-test p-value > 0.001; Fig 7-A). Modules yellow, purple, cyan, and brown had a strong positive relation between drink per day gene significance and progressor modules kME absolute values (Pearson correlation > 0.3; student t-test p-value > 0.001; Fig 7-B).

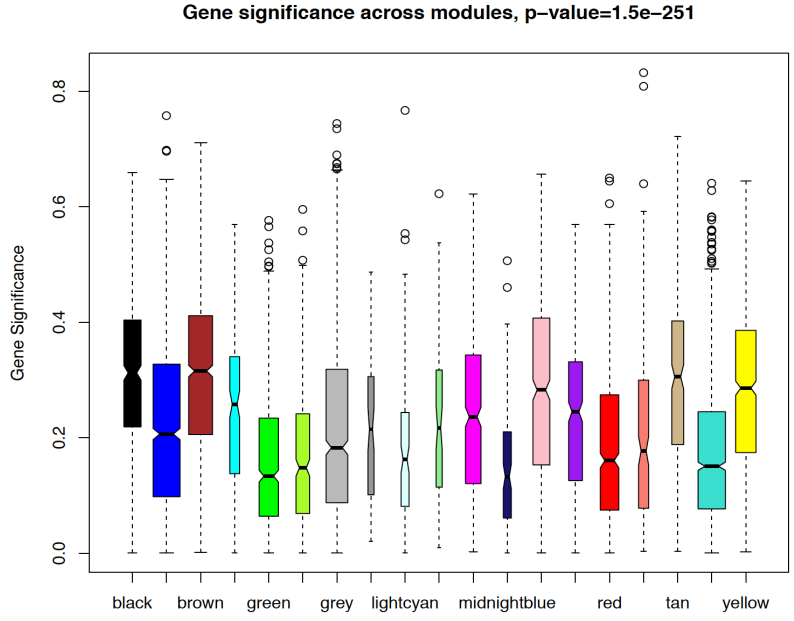
Differential eigengene network analysis revealed strong correlations results with drinks per day and pack years smoked clinical features. We found brown and yellow progressor eigengene modules showing the strongest relationship with alcohol drinks consumed per day. Also, cyan module showed strongest relationship with tobacco pack years smoking habits (Fig 8). For nonprogressors, turquoise eigengene module showed strong relationship with both clinical features of smoking and alcohol (Fig 8).

A



black	0.1177
blue	0.0966
brown	0.106
cyan	0.1924
green	0.0885
greenyellow	0.0958
grey	0.1169
grey60	0.0813
lightcyan	0.104
lightgreen	0.1562
magenta	0.1422
midnightblue	0.1021
pink	0.1302
purple	0.1377
red	0.0924
salmon	0.0915
tan	0.1056
turquoise	0.1249
yellow	0.1142

B



black	0.3065
blue	0.2235
brown	0.3081
cyan	0.2497
green	0.1604
greenyellow	0.1675
grey	0.2149
grey60	0.2144
lightcyan	0.1791
lightgreen	0.2308
magenta	0.2398
midnightblue	0.145
pink	0.2837
purple	0.2295
red	0.1859
salmon	0.2069
tan	0.2975
turquoise	0.1714
yellow	0.2789

Figure 5: A) Boxplot distribution of pack years gene significance with each color-coded by the corresponding consensus module color (black, blue, brown, cyan, green, greenyellow, grey, grey60, lightcyan, lightgreen, magenta, midnightblue, pink, purple, red, salmon, tan, turquoise, yellow) and table of module significance measures. The module significance here is defined as the average gene significance of the genes within a module. Gene significance of each gene is defined by the absolute Pearson correlations with tobacco pack years smoking clinical feature $GS_i = |cor(x_i, F_{pack-years})|$. The $GS = (GS_1, GS_2, \ldots, GS_n)$ measures are significantly different between modules (Kruskal Wallis test p-value: 6.3e-94). Only cyan module showed the highest pack years module significance (mean > 0.19). B) Boxplot distribution of drink per day gene significance and table of module significance mean measures. Gene significance of each gene is defined by the absolute Pearson correlations with drink per day alcohol consumption clinical feature $GS_i = |cor(x_i, F_{drink\_per\_day})|$. The $GS = (GS_1, GS_2, \ldots, GS_n)$ measures are significantly different between modules (Kruskal Wallis test p-value: 1.5e-251). Modules black, brown, cyan, pink, tan, yellow showed the highest module significance (mean > 0.25)

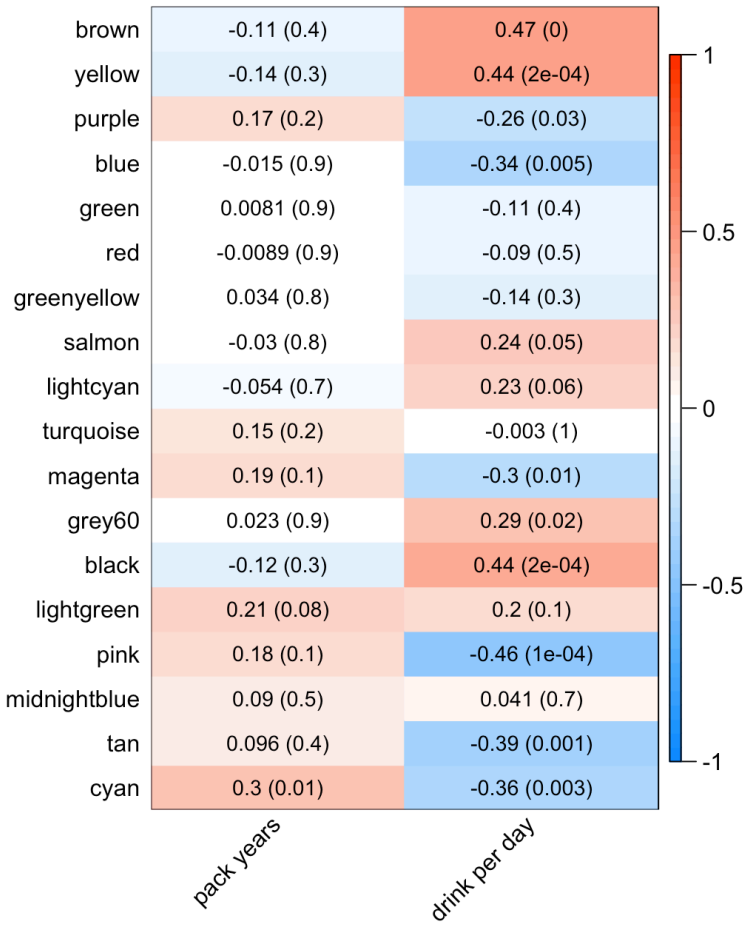
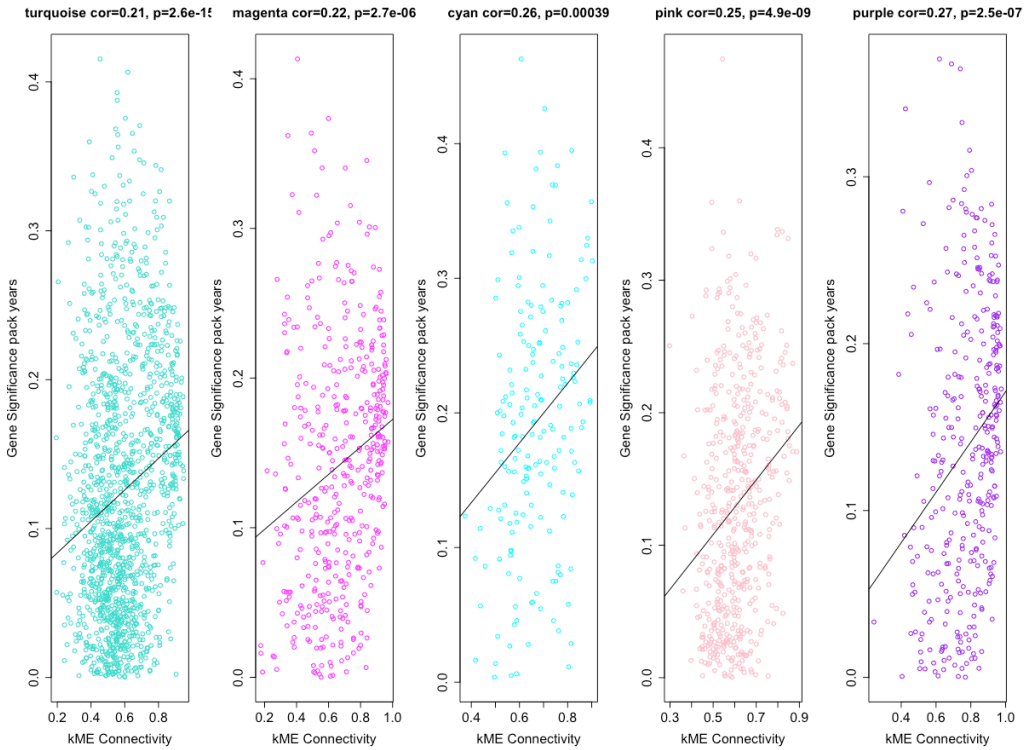


Figure 6: Heatmap plot of clinical feature Pearson correlations (-1:1 shown by color legend) with co-expression consensus module eigengene and corresponding student t-test p-values. Modules brown, yellow, and black, pink, tan, and cyan show the highest positive correlation with alcohol consumption per day. The cyan module shows the highest positive correlation with tobacco pack years smoked.

A



B

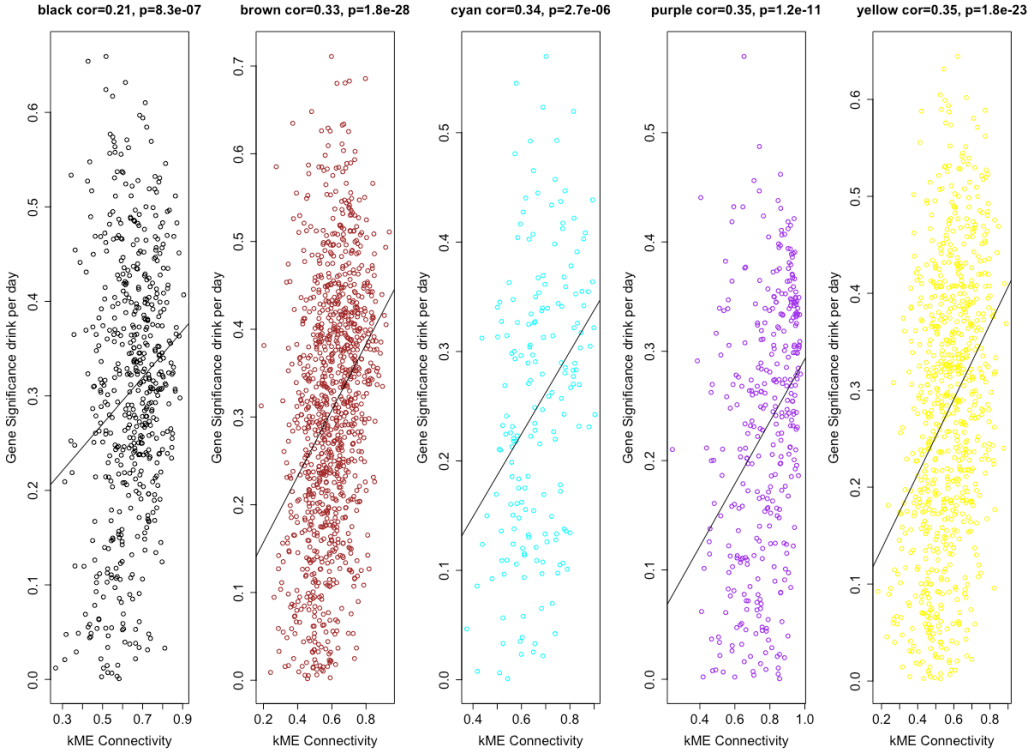


Figure 7: A) Scatterplot of top five strongest consensus module's gene significance$_{\text{pack-years}}$ (y-axis) vs. absolute value of kME intramodule connectivity (x-axis). Regression lines and corresponding Pearson correlations and p-values (student t-test) are also demonstrated in each plot. Plotted dots represent a gene that is color-coded by its corresponding module. Purple, pink, cyan, modules have the highest absolute value correlations and significant student t-test p-values between gene significance$_{\text{pack-year}}$ display and absolute value kME measures. B) Scatterplot of top five strongest consensus module's gene significance$_{\text{drink-per-day}}$ (y-axis) vs. kME intramodule connectivity (x-axis Regression lines and corresponding Pearson correlations and p-values (student t-test) are also displayed in each plot. Yellow, purple, cyan, brown modules showed the highest correlations and significant student t-test p-values between gene significance$_{\text{drink-per-day}}$ and absolute value kME measures.

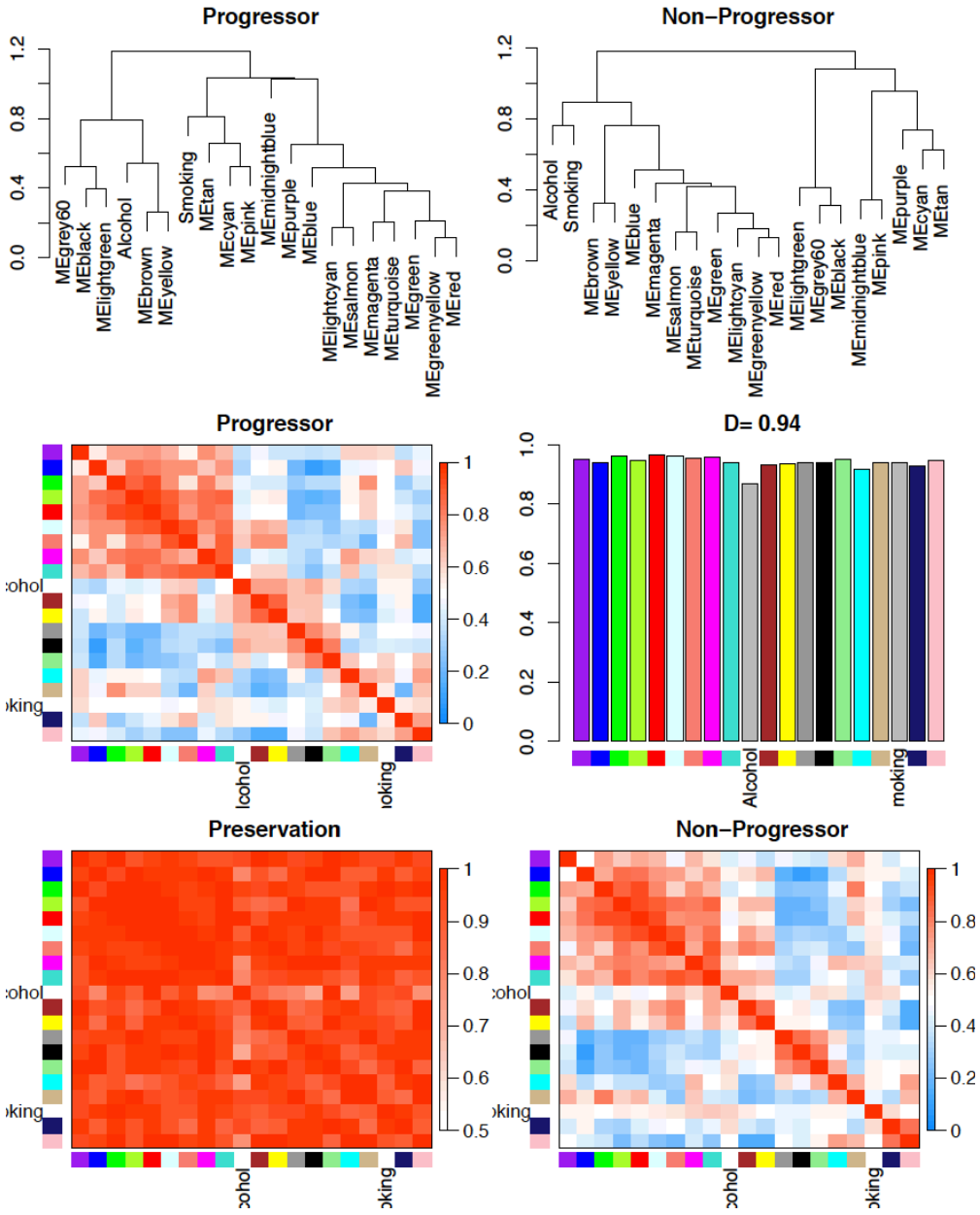


Figure 8: Visualization of the differential eigengene network analysis between TCGA HNSCC progressors and nonprogressors consensus eigengene networks and their relationship among tobacco pack years smoked and alcohol consumption per day clinical feature. This visualization aids in identifying similarities and differences across sets. Clustering dendrograms of consensus module eigengenes and features are defined based on Euclidian distance average linkage dissimilarity of Pearson correlation dissimilarity measures ($diss\left(A_{E_{IJ}}\right) = \frac{1-cor(E_I,E_J)}{2}$). Diagonal heatmaps represent consensus eigengene network of conditions. Each row and column represents clinical feature and modules eigengene labeled and color-coded accordingly. The preservation measure heatmap shows preservation between clinical feature and consensus eigengene modules with red indicating highest preservation (0.5-1 shown by color legend). Barplot shows eigengene modules preservation with height of bar (y-axis) representing preservation measure. $D\left(preservation^{(1,2)}\right) = 1 - \frac{\sum_i \sum_{j \neq i} |a_{ij}^1 - a_{ij}^2|}{n(n-1)}$ is an aggregate measure of eigengene network preservation between conditions and here shows a high overall preservation (D = 0.94). Clinical features alcohol drinks per day (Alcohol) and tobacco pack years smoked (Smoking) show strong relation with progressor condition eigengene modules. Brown and yellow modules with alcohol drinks per day and cyan module with tobacco pack years smoked. Turquoise module in both alcohol drinks per day and smoking shows a strong relationship with nonprogressor eigengene modules.

Table 2: Summary of consensus modules identified to show the *strongest* relation with clinical features (alcohol drink per day and pack years) from four methods: differential eigengene network analysis (DENA), module significance (MS), gene significance (GS) and ME Pearson correlations, and gene significance (GS) vs. kME analysis.

Condition	Progressor		Nonprogressor	
Analysis/ Clinical feature	Drink per day	Pack years	Drink per day	Pack years
DENA	Brown, yellow	Cyan	Turquoise	Turquoise
MS	Brown, black, pink, tan, yellow, cyan	Cyan		
GS and ME	Brown, yellow, and black, pink, tan, and cyan	Cyan		
GS vs. kME	Yellow, purple, cyan, brown	Purple, pink cyan		

3.5 Differential Network Analysis of Co-expression Consensus Modules

Leveraging network properties, to identify progressor modules that are enriched in genes holding informative progression mechanisms compared to nonprogressors we computed differential network analysis (section 2.10). Comparing single genes based on expression intensity and range, we identified 11 progressor consensus modules enriched in DE/DV genes shown in Table 3. Most of all DE/DV modules had an overlap with consensus modules identified to have the strongest relation with pack years and drink per day clinical features (Table 2). However we know that genes don't act in isolation and their different interactions drive different disease states [29-30]. Comparing network correlation structure information of a gene between progressors and nonprogressors, we identified genes with higher differential wiring rates (section 2.10). Only one progressor consensus module color-coded as turquoise was enriched in DW genes (Table 3). Identified DE/DV/DW genes lists are available via Supplementary Data.

Table 3: Summary of consensus network differential analysis between progressors and nonprogressors genes. DE, DV, and DW are indicative of 12 progressor modules enriched in genes that are differentially expressed, differentially variable, and differentially wired.

DE	Black	Cyan	Blue	Tan	Yellow	Grey60	Lightgreen
DV	Purple	Black	Pink	Lightcyan	Salmon		
DW	Turquoise						

To ensure proper biological findings of each DE/DV/DW progressor consensus module, we visualized the heatmap of scaled gene expressions over progressor samples with corresponding eigengene values (Fig 9; Supplementary Data). All modules were evenly distributed and showed no outstanding noise/batcheffect. Conclusively, we took these modules as putative progressor gene signatures of HNSCC.

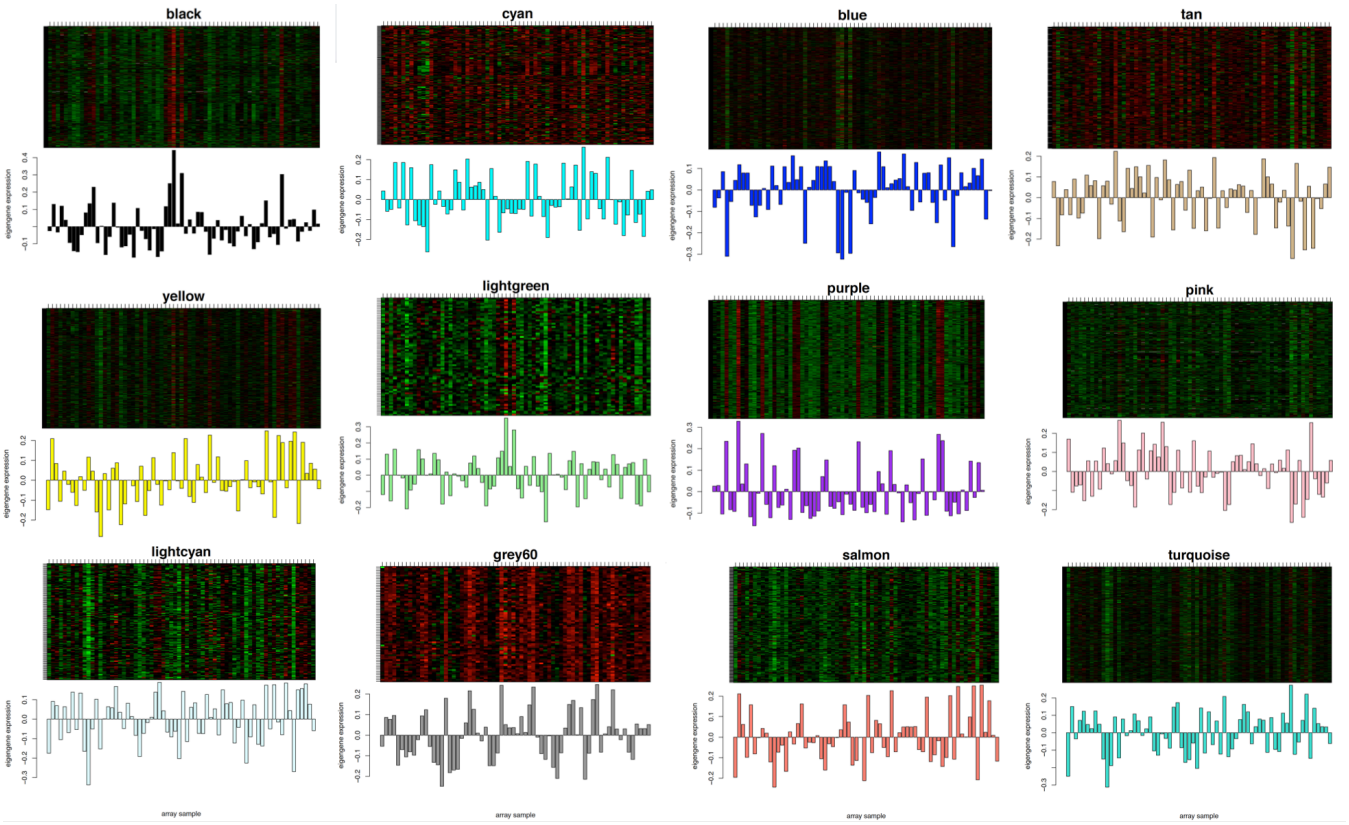


Figure 9: DE, DV, and DW consensus modules of progressor condition scaled ($scale = \frac{x-mean(x)}{sd(x)}$)[3] gene X samples heatmaps and corresponding samples eigengene values barplot (color-coded by module membership). Rows correspond to genes of consensus modules and columns correspond to progressor condition patients' samples. Heatmap colors red indicates high and green indicates low-scaled expression values. All consensus modules heatmaps of both conditions can be found in supplementary Data.

Top 20 genes with the highest kME values of these modules were extracted and are available in Supplementary Data. Genes with high kME (>0.8) and GS (>0.2) in each clinical feature of these modules were also extracted (Supplementary Data). These genes are the network modules' hubs and known to be potential driver genes of the biological events in modules/gene signatures [19-20].

3.6 Pathway Enrichment Analysis of DE/DV/DW Co-expression Consensus and Nonpreserved CoSpliceEx Modules

We assessed pathway enrichment analysis on all genes in DE/DV/DW progressor consensus modules (Supplementary Data). We also assessed pathway enrichment on genes with kME (>0.8) and GS (>0.2) of pack years and drink per day clinical features (Supplementary Data). Portable text files of gene lists to Cytoscape ReactomeFIViz are available via Supplementary Data. Most DE/DV progressor consensus modules pathway enrichment analysis showed involvement with regulating internal tumors mechanism to change the state of cancer. DW pathway enrichment analysis showed involvement of genes with inflammation and tumor microenvironment evolutionary mechanisms. We also obtained National Center for Biotechnology Information (NCBI) known biological roles of DW genes with high kME measures of > 0.9 (hubs). The biological mechanism of these genes revealed similar biological roles with DW progressor consensus modules pathway enrichment results

[3] An essential unit of measure in WGCNA is eigengene (ME) that is closely defined as the 1st principal component. Principal components rely on scale dependency of variables and normalization is required ($scale = \frac{x-mean(x)}{sd(x)}$).

(Table 4). The latter result extends onto Bornstein et al. study [3] and reemphasizes the potential importance of inflammation pathways and tumor microenvironment evolution in tumors with higher progression rate.

Table 4: One progressor consensus module color-coded as turquoise was denoted as enriched in DW genes. This table demonstrates NCBI known functions of turquoise module genes with kME > 0.9 (hubs) that are also DW between progressor and nonprogressors.

Node	NCBI Known function
IL10RA	Mediates immunosuppressive signaling of interleukin 10
DOK2	Provides docking platform for the assembly of signaling complexes in proliferation
APBBIIP	Controls adhesion (scar tissue) and cell migration
UBASH3A	Negative regulation of T-cell signaling and involved in apoptosis
SASH3	Cell signaling adaptor protein in lymphocytes

Although all twelve DE/DV/DW modules pathway enrichment analysis results are biologically rich and informative, we demonstrate two. Following are pathway enrichment analysis results of the yellow module with high correlation to drink per day and cyan module with high correlation to both clinical features.

3.6.1 DE enriched Yellow Co-expression Consensus module with high kME and Drink per day GS

Yellow module pathway enrichment analysis included only 3 genes SGOL1, KNTC1, and BUB1B out of 20 (Table 5). These genes show involvement in cell division cycle roles. Noteworthy, the gene KIF15 with the highest kME value of 0.89 (hub) wasn't captured here, but is associated with spindle bipolarity. Asymmetric bipolarity in cells is known to cause abnormal mitosis and proliferation in cancer [41].

Table 5: Yellow module pathway enrichment analysis of genes with kME > 0.8 and drink per day GS > 0.2

Pathway	Ratio of protein in gene set	Number of protein in gene set	Protein from network	P-value	FDR	Nodes
Mitotic Prometaphase (R)	0.0101	99	3	0.0000	9.40E-06	SGOL1, KNTC1, BUB1B
Mitotic Metaphase and Anaphase (R)	0.0165	161	3	0.0000	1.80E-05	SGOL1, KNTC1, BUB1B
PLK1 signaling events (N)	0.0045	44	2	0.0001	1.83E-04	SGOL1, BUB1B
Aurora B signaling (N)	0.0041	40	1	0.0122	0.0243	SGOL1
APC/C-mediated degradation of cell cycle proteins (R)	0.0082	80	1	0.0244	0.0243	BUB1B
Oocyte meiosis (K)	0.0116	113	1	0.0343	0.0343	SGOL1
Cell Cycle Checkpoints (R)	0.0119	116	1	0.0352	0.0352	BUB1B
Cell cycle (K)	0.0127	124	1	0.0376	0.0376	BUB1B
HTLV-I infection (K)	0.0267	260	1	0.0778	0.0778	BUB1B

3.6.2 DE enriched Cyan Co-expression Consensus module with high kME and Pack years GS

Cyan module showed a high correlation to both pack years and drink per day. Both pathway enrichment analyses over genes with high pack years and drink per day GS measures revealed similar results. Here we show genes with high pack years GS. The pathway enrichment analysis involved 9 out of 29 genes with STAT1 activating all other hub genes of this module (Fig 10). STAT1 is a member of the STAT protein family. STAT is known to be involved in many of the hallmarks of cancer such as apoptosis, proliferation and tumor suppression pathways. For example MAPK, Jak-Stat, and P53. Also, UBE2L6 is notable for activating STAT1 (Fig 10). This gene is known to be associated with Parkinson's disease, a type of synapse degenerative disease.

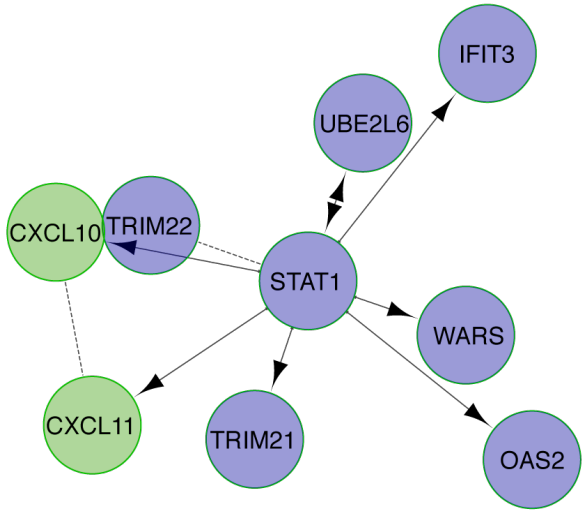


Figure 10: Cyan module genes with kME > 0.8 and pack years GS > 0.2 pathway enrichment analysis network.

4. Discussion and Conclusion

Previously annotated and curated TCGA HNSCC data gave us the opportunity to study tumors with higher progression rate [3]. The collaborative effort of TCGA and accessibility of its open data with proper sample collection and patient consents has enabled researches to drive and resolve theories with an adequate power. OHSU Knight Cancer Institute annotations on this cohorts' progression status enabled us to assess regulatory mechanisms between two subtypes of HNSCC tumors [3].

Given this cohorts' annotated and curated clinical and genomic information, we were able to identify and characterize twelve putative gene signatures involved in HNSCC tumors with high progression rate. Methodologically, this was possible by leveraging network properties and expression data to define pairwise gene relations as a network correlation structure. In different states of a disease, genes are known to co-regulate with different groups of other genes and this can be seen as differential wiring or correlation magnitude differences [20-30]. These interaction patterns or relations in a network model can't be captured from other approaches. Weighted networks analysis gives a holistic view on disease dynamics, but also enables us to reduce the complexity into organized and measurable relations. We were able to reduce 10,000+ genes down to 10+ mechanistic modules with measured hubs. We then were able to further characterize them. This was done by relating them to known clinical risk factors of HNSCC and conducting pathway enrichment analysis to reveal their biological identity.

Utilizing network differential analysis, we compared single genes between progressor and nonprogressor and identified eleven progressor consensus modules enriched in DE/DV genes. These modules showed high

correlations to drink per day or pack years HNSCC habitual risk factors. We know that genes don't act in isolation and their combined interaction regulates different disease states. To capture interaction differences we compared differential wiring rates (DW) between progressor and nonprogressor conditions. Only one progressor consensus module was enriched in DW genes. These twelve DE/DV/DW putative progressor gene signatures showed involvement in various pathways such as cell cycle check points, abnormal mitosis and spindle bipolarity, c-myc, macrophages, immune response and T cells, receptor synapse dysregulations and associated diseases such as Alzheimer's disease and Parkinson's disease, Interferon gamma signaling pathway, MAPK, Jak-Stat, P53, and more. Gene composition and characterization of each putative progressor gene signature with detailed pathway enrichment analysis are available via Supplementary Data.

IL10RA, DOK2, APBBIP UBASH3A, SASH3 genes were identified as DW and putative driver genes (network hubs). The known NCBI biological function of these genes and the pathway enrichment analysis of DW module showed involvement in inflammation and tumor microenvironment evolution mechanisms. This result is a direct extension to Bornstein et al study findings [3]. After further evaluation of DW putative gene signature and driver genes, they can possibly be used in prognosis and targeted therapy research. With this information we potentially have a better chance of disrupting progression rate mechanisms in the least invasive manner possible.

Today with NCI-match precision medicine clinical trials, patients are stratified based on their tumor molecular abnormalities vs. randomly being placed in case and control sub-groups [42]. This is an enormous improvement, but many questions still remain open, e.g. what are best methods, who is likely to benefit, and why. In this growing body of clinical research, if we stratify patients based on gene signatures involved in different progression rates (i.e. evaluated DW module and hubs), we possibly have a better chance of improving the end points that inform us whether a drug is working or not and potentially observe clinical utility [40]. Patient stratification for prognosis based on different progression rate may be applied using OncotypeDX approach [43].

While we looked at both gene and exon RNA level expressions to gain a biologically more complete insight on tumor progression, we found two nonpreserved coSpliceEx modules. We utilized readily available RPKM as a normalized/scaled unit for exon expression data. RPKMs may not be capturing accurate splicing events and we hypothesis that this has cascaded noise through coSpliceEx pipeline. It has been noted that RPKM's are not a well-defined unit for expression analysis [35-36]. In future studies we will use a more robust normalized/scaled unit for exon expression.

Although data annotation of alcohol consumption per day was less complete than pack years, we found stronger associations of alcohol habits with co-expression consensus and coSpliceEx nonpreserved modules (Supplementary Information Fig 1). We hypothesis that the self report estimates captured in a clinical setting for drink per day are closer to the patients' biology. This is potentially due to patient's stronger recall on the quantity they drink per day vs. packs of cigarettes they smoke through a year. Given this direct impact of self-reported clinical data on our research study (secondary research), advancing the quality of measured clinical data has the potential to improve secondary research results [37-38].

Overall we showed the use of de-novo weighted network inference in the context of biological pathways suggests the initiation of new insights for both mechanistic and prognostically relevant information. For Future directions, other phenotypic features such as CT scan image processing together with weighted network approaches has the potential to reveal a more holistic view of the tumors progression dynamics. Expanding this notion over longitudinal data may be utilized as predictive prognosis in clinical research.

Supplementary

Supplementary Information pdf file
Supplementary Data: folder of co-expression & coSpliceEx excel, pdf, and txt files

Reference

1. Leemans C, Braakhuis B, Brakenhoff R. The molecular biology of head and neck cancer. Nature Reviews Cancer. 2010;11(1):9-22.
2. B. Burtness and E. A. Golemis. Overview: The Pathobiology of Head and Neck Cancer. Molecular Determinants of Head and Neck Cancer. 2014;23-53.
3. Bornstein S, Schmidt M, Choonoo G, Levin T, Gray J, Thomas C et al. IL-10 and integrin signaling pathways are associated with head and neck cancer progression. BMC Genomics. 2016; 17(1).
4. Worsham M. Identifying the risk factors for late-stage head and neck cancer. Expert Review of Anticancer Therapy. 2011;11(9):1321-1325.
5. Head and Neck Cancer [Internet]. National Cancer Institute. 2016 [cited August 2016]. Available from: http://www.cancer.gov/types/head-and-neck
6. Cancer.org. What are the key statistics about oral cavity and oropharyngeal cancers? [Internet]. 2016 [cited February 2016]. Available from: http://www.cancer.org/cancer/oralcavityandoropharyngealcancer/detailedguide/oral-cavity-and-oropharyngeal-cancer-key-statistics
7. Murphy B, Ridner S, Wells N, Dietrich M. Quality of life research in head and neck cancer: A review of the current state of the science. Critical Reviews in Oncology/Hematology. 2007;62(3):251-267.
8. Who.int. [Internet]. 2016 [cited August 2016]. Available from: http://www.who.int/selection_medicines/committees/expert/20/applications/HeadNeck.pdf?ua=1
9. Lawrence M et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-218.
10. Burtness B, Golemis E. Molecular determinants of head and neck cancer. New York: Springer; 2014.
11. Roberts JC, Li G, Reitzel LR, Wei Q, Sturgis EM. No evidence of sex-related survival disparities among head and neck cancer patients receiving similar multidisciplinary care: a matched-pair analysis. Clin Cancer Res. 2010;16:5019–27
12. Curry J, Sprandio J, Cognetti D, Luginbuhl A, Bar-ad V, Pribitkin E et al. Tumor Microenvironment in Head and Neck Squamous Cell Carcinoma. Seminars in Oncology. 2014;41(2):217-234.
13. Balkwill F, Capasso M, Hagemann T. The tumor microenvironment at a glance. Journal of Cell Science. 2012;125(23):5591-5596
14. Patel S, Shah J. TNM Staging of Cancers of the Head and Neck: Striving for Uniformity Among Diversity. CA: A Cancer Journal for Clinicians. 2005; 55(4):242-258.
15. Wu G, Stein L. A network module-based method for identifying cancer prognostic signatures. Genome Biol. 2012;13(12):R112.
16. About TCGA [Internet]. The Cancer Genome Atlas - National Cancer Institute. 2016 [cited August 2016]. Available from: http://cancergenome.nih.gov/abouttcga
17. Parfenov M, Pedamallu C, Gehlenborg N, Freeman S, Danilova L, Bristow C et al. Characterization of HPV and host genome interactions in primary head and neck cancers. Proceedings of the National Academy of Sciences. 2014;111(43):15544-15549.
18. Lawrence M, Sougnez C, Lichtenstein L, Cibulskis K, Lander E, Gabriel S et al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576-582.
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9(1): 559.
20. Horvath S. Weighted network analysis. New York: Springer; 2011.
21. Iancu O, Colville A, Oberbeck D, Darakjian P, McWeeney S, Hitzemann R. Cosplicing network analysis of mammalian brain RNA-Seq data utilizing WGCNA and Mantel correlations. Front Genet. 2015;

22. Wu G, Dawson E, Duong A, Haw R, Stein L. ReactomeFIViz: the Reactome FI Cytoscape app for pathway and network-based data analysis. F1000Research. 2014.
23. WGCNA package: Frequently Asked Questions [Internet]. Labs.genetics.ucla.edu. 2016 [cited August 2016]. Available from: https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html
24. Lawrence M et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-218.
25. Araya C, Cenik C, Reuter J, Kiss G, Pande V, Snyder M et al. Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nature Genetics. 2015; 48(2): 117-125.
26. Wang K et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nature Genetics. 2014; 46(6): 573-582.
27. Identification and Removal of Outlier Samples (Illumina) [Internet]. Labs.genetics.ucla.edu. 2016 [cited August 2016]. Available from: https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/HumanBrainTranscriptome/
28. Langfelder P, Luo R, Oldham M, Horvath S. Is My Network Module Preserved and Reproducible? PLoS Comput Biol. 2011; 7(1): e1001057.
29. Iancu O, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S. Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics. 2012;28(12):1592-1597.
30. Iancu O, Darakjian P, Malmanger B, Walter N, McWeeney S, Hitzemann R. Gene networks and haloperidol-induced catalepsy. Genes, Brain and Behavior. 2012;11(1):29-37.
31. Iancu O, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S. Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics. 2012; 28(12):1592-1597.
32. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology. 2007;1(1):54.
33. Iancu O, Oberbeck D, Darakjian P, Kawane S, Erk J, McWeeney S et al. Differential Network Analysis Reveals Genetic Effects on Catalepsy Modules. PLoS ONE. 2013;8(3):e58951.
34. Zapala M, Schork N. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences. 2006;103(51):19430-19435.
35. Wagner G, Kin K, Lynch V. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131(4):281-285.
36. RPKM measure is inconsistent among samples - Next Genetics [Internet]. Blog.nextgenetics.net. 2012 [cited August 2016]. Available from: http://blog.nextgenetics.net/?e=51
37. Weiskopf N, Weng C. 3x3 DQA: Dynamic, evidence-based guidelines to enable electronic health record data quality assessment and reporting for retrospective research (Version 1.0). 2014.
38. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013; 46(5):830-6.
39. Hanahan D, Weinberg R. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-674.
40. Simon R, Roychowdhury S. Implementing personalized cancer genomics in clinical trials. Nature Reviews Drug Discovery. 2013;12(5):358-369.
41. Gómez-López S, Lerner R, Petritsch C. Asymmetric cell division of stem and progenitor cells during homeostasis and cancer. Cell Mol Life Sci. 2013;71(4):575-597.
42. NCI-Molecular Analysis for Therapy Choice (NCI-MATCH) Trial [Internet]. National Cancer Institute. 2016 [cited August 2016]. Available from: http://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match
43. Sparano JPaik S. Development of the 21-Gene Assay and Its Application in Clinical Practice and Clinical Trials. Journal of Clinical Oncology. 2008;26(5):721-728.
44. Li, Bo and Colin N Dewey. RSEM: Accurate Transcript Quantification From RNA-Seq Data With Or Without A Reference Genome. *BMC Bioinformatics* 12.1 (2011): 323.
45. Mortazavi Ali et al. Mapping And Quantifying Mammalian Transcriptomes By RNA-Seq. *Nature Methods* 5.7 (2008): 621-628. Web.