



Network-Based Predictors of Progression in Head and Neck Squamous Cell Carcinoma

Nasim Sanati^{1*}, Ovidiu D. Iancu², Guanming Wu¹, James E. Jacobs^{1,3} and Shannon K. McWeeney^{1,4}

¹ Division of Bioinformatics and Computational Biology, Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR, United States, ² Department of Behavioral Neuroscience, Oregon Health and Science University, Portland, OR, United States, ³ Division of Pediatric Hematology/Oncology, Department of Pediatrics, Oregon Health and Science University, Portland, OR, United States, ⁴ OHSU Knight Cancer Institute, Portland, OR, United States

OPEN ACCESS

Edited by:

Xiaogang Wu,
Institute for Systems Biology,
United States

Reviewed by:

Alessandro Giuliani,
Istituto Superiore di Sanità, Italy
Oksana Sorokina,
University of Edinburgh,
United Kingdom

*Correspondence:

Nasim Sanati
nasim@plenary.org

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 05 February 2018

Accepted: 07 May 2018

Published: 29 May 2018

Citation:

Sanati N, Iancu OD, Wu G, Jacobs JE
and McWeeney SK (2018)
Network-Based Predictors of
Progression in Head and Neck
Squamous Cell Carcinoma.
Front. Genet. 9:183.
doi: 10.3389/fgene.2018.00183

The heterogeneity in head and neck squamous cell carcinoma (HNSCC) has made reliable stratification extremely challenging. Behavioral risk factors such as smoking and alcohol consumption contribute to this heterogeneity. To help elucidate potential mechanisms of progression in HNSCC, we focused on elucidating patterns of gene interactions associated with tumor progression. We performed *de-novo* gene co-expression network inference utilizing 229 patient samples from The Cancer Genome Atlas (TCGA) previously annotated by Bornstein et al. (2016). Differential network analysis allowed us to contrast progressor and non-progressor cohorts. Beyond standard differential expression (DE) analysis, this approach evaluates changes in gene expression variance (differential variability DV) and changes in covariance, which we denote as differential wiring (DW). The set of affected genes was overlaid onto the co-expression network, identifying 12 modules significantly enriched in DE, DV, and/or DW genes. Additionally, we identified modules correlated with behavioral measures such as alcohol consumption and smoking status. In the module enriched for differentially wired genes, we identified network hubs including *IL10RA*, *DOK2*, *APBB1IP*, *UBASH3A*, *SASH3*, *CELF2*, *TRAF3IP3*, *GIMAP6*, *MYO1F*, *NCKAP1L*, *WAS*, *FERMT3*, *SLA*, *SELPLG*, *TNFRSF1B*, *WIPF1*, *AMICA1*, *PTPN22*; the network centrality and progression specificity of these genes suggest a potential role in tumor evolution mechanisms related to inflammation and microenvironment. The identification of this network-based gene signature could be further developed to guide progression stratification, highlighting how network approaches may help improve clinical research end points and ultimately aid in clinical utility.

Keywords: HNSCC, TCGA, RNA-Seq, progression, predictors, weighted network analysis, differentially wired, co-expression

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is the most prevalent of the mucosal head and neck cancers and represents a significant health burden in the United States with approximately 40,000 new cases and almost 8,000 deaths per year. HNSCC can arise from multiple locations, including the oral cavity, oropharynx, hypopharynx, larynx, or nasopharynx (Marur and Forastiere, 2016). Furthermore, due to its non-specific presenting symptoms, patients often go undiagnosed until the cancer has progressed beyond local involvement leading to poorer treatment outcomes. Despite current treatments, many HNSCCs will become progressive which is associated with a 40–50% 5-year survival rate (Bonner et al., 2010). There are many well-known risk factors for the development of HNSCC, including smoking or chewing tobacco products, alcohol consumption, and infection with human papilloma virus (HPV). Thus, reliable stratification of patients with HNSCC given the current tumor-node-metastasis (TNM) staging system can be quite challenging with both social and biological factors at play (Cancer Staging—National Cancer Institute¹; Patel and Shah, 2005). The ability to better predict tumor progression would be of great benefit in this patient population, allowing for the proper stratification of treatment, leading to less treatment-related morbidity in lower risk tumors as well as an increased probability of treatment success in higher risk tumors. Understanding tumor progression mechanisms is a critical step toward achieving better clinical outcomes. To this end, here we identify features predictive of tumor progression based on gene expression data.

To extract gene co-expression patterns that are predictive of tumor progression, we leverage the availability of transcriptional data coupled with clinical and behavioral measures of progression previously annotated by Bornstein et al. (2016). In their study, sample annotation with respect to progressor and non-progressor status was performed through the review and curation of The Cancer Genome Atlas (TCGA) follow-up clinical data. We hypothesize that there exist alterations in the co-expression network structure that may differentiate progressors from non-progressors. By utilizing network measures such as connectivity (pairwise gene expression correlate measures) and hubness (quantity and strength of correlation measures of a gene to all other genes), we will be able to rank the affected genes and prioritize putative predictors. These features may further aid research studies in the context of both underlying tumor progression biology and/or identifying therapeutic targets.

In order to identify mechanistically informative progressor features, we leveraged expression data to define pairwise gene relations across multiple samples as a weighted network correlation structure (Zhang and Horvath, 2005; Wu and Stein, 2012). This network model is complementary to classical approaches such as differential expression and allows for the discovery of synergies that wouldn't otherwise be evident when

assessing only single gene differences. Given the complexities of tumor progression and the different dynamics observed between clinical tumor subtypes (Sparano and Paik, 2008; Hanahan and Weinberg, 2011), it is likely that the transcriptional profiles will differ between the patient group that experiences rapid tumor progression versus the group with a less aggressive disease profile. Importantly, the genetic aberrations that drive the progressor phenotype may lie in the regulatory networks between genes rather than in abnormal expression of specific gene products. Thus, network analysis is well-suited to evaluate this overall hypothesis, and in particular to detect alterations in transcriptional coordination/co-expression that are dispersed among large sets of genes and might be undetectable at the individual gene level. Similarly, relationships between transcriptional profiles and clinical and demographic variables, such as alcohol and tobacco use, could also be possibly better detected and interpreted in a network context. Finally, network measurements can also provide an alternative procedure for prioritizing and/or ranking candidates, based on their relative network importance or connectivity.

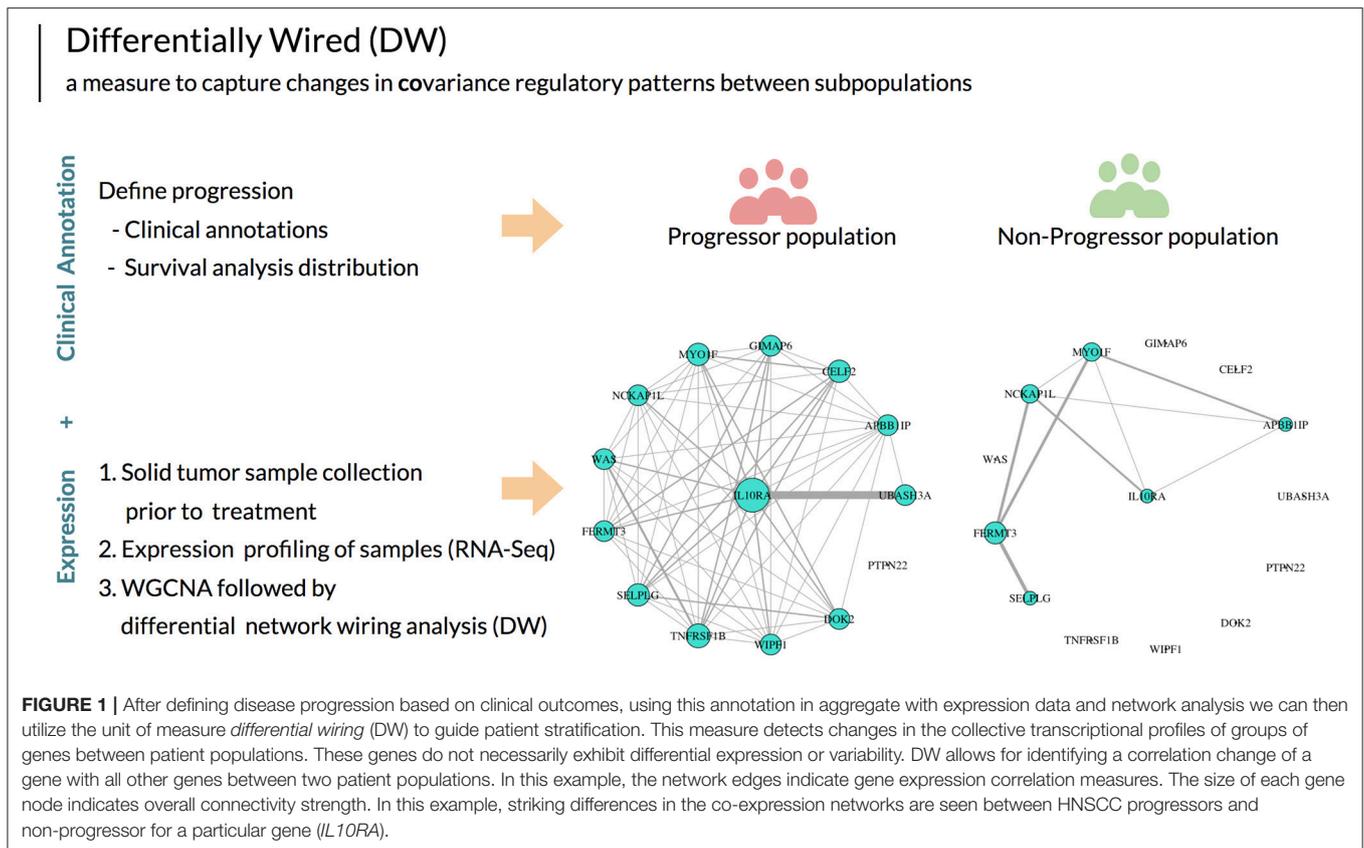
When analyzing transcriptomic data between two populations, there are a variety of methods that can be utilized based on the particular hypothesis being tested. Here, we employ three measures: differential expression (DE), differential variability (DV) and differential wiring (DW) in order to identify significant aberrations that define the progressor population. In differential expression analysis (a non-network based approach), the gene expression data over multiple patient samples is used to identify genes with statistically significant differences in transcriptional status between two populations (progressor vs. non-progressor). In contrast, differential variability analysis seeks to identify genes with significant changes in variance of expression between two populations. This analysis method has been shown to identify biologically relevant genetic aberrations which can be overlooked when performing differential expression analysis alone (Ho et al., 2008). Finally, differential wiring analysis detects changes in pairwise expression correlations between genes by integrating them over all genes in the network, and selects genes and hubs enriched for these changed correlation patterns. It is important to note that such changes often affect genes that do not necessarily exhibit differential expression or variability alone (Hudson et al., 2009; Iancu et al., 2013). An example of differentially wired (DW) between two populations is illustrated in **Figure 1**.

MATERIALS AND METHODS

Patient Clinical Demographics

Our study utilizes HNSCC data from TCGA previously annotated by Bornstein et al. (2016) (progressors and non-progressors). The data include 229 patient samples with 68 (30%) progressors and 161 (70%) non-progressors. The median last encounter days of progressor patients were considerably lower than non-progressor patients (606 vs. 4856 days). HPV status was missing in 165 (72%) patients, negative

¹Cancer Staging—National Cancer Institute. Available online at: <https://www.cancer.gov/about-cancer/diagnosis-staging/staging> (Accessed August 4, 2017).



in 48 (20%), and positive in 16 (6%) (p16 or ISH). One hundred thirty-one (57%) of the tumors occurred in the oral cavity, with 59 (25%) occurring in the larynx, and 39 (17%) in the oropharynx. The mean age in the cohort was 62 years with 138 (60%) complete documented cases of self-reported tobacco pack years smoked and 97 (42%) alcohol drinks consumed per day. 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 four alcohol drinks per day. Non-progressor patients reported slightly lower smoking and alcohol consumption estimates, with median of 40 pack years smoked and three alcoholic drinks per day.

TCGA Transcriptional Data

RNA-seq data (Level 3; Illumina HiSeq, 2000) from solid tumor tissue was retrieved from TCGA. All data alignment was mapped to genome build hg19. To capture meaningful pairwise correlate measures, we utilize normalized counts per gene as defined in Li and Dewey (2011). The normalized gene counts were then transformed via a variance stabilizing log transformation $\log_2(x + 1)$, based on WGCNA best practices (WGCNA package²: Frequently Asked Questions). This transformation

²WGCNA package: Frequently Asked Questions Available online at: <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html> (Accessed December 20, 2016).

reduces the dependence of variance on the mean and facilitates downstream network and correlational analyses. Finally, the olfactory genes were removed from the pre-processed zero/low variance gene counts (WGCNA `goodSampleGenes` function). The former have been noted to introduce noise in TCGA data across cancer types due to their locations in the chromosome (Lawrence et al., 2013) and removed from multiple previous studies (Wang et al., 2014; Araya et al., 2016).

We examined the data for extreme outliers by visualizing the samples' gene expression values as boxplot distributions and computing the inter-array correlation (IAC). The IAC is defined as the average Pearson correlation of a sample to all other samples. These procedures detected no extreme outliers as all IAC had values > 0.65 (Figure S2).

Co-expression Networks Construction

Following the WGCNA framework (Langfelder and Horvath, 2008), we constructed an adjacency network matrix by (1) computing the biweight mid-correlation between all gene pairs, (2) taking the absolute value of the resulting correlations for construction of an unsigned network (aggregating both up and down gene regulation), and then (3) raising this value to a power β chosen such that the network approaches a scale-free structure (exponential distribution of node connectivity; Figure S1).

Given that biological mechanisms of network components are best captured by the most connected genes, we restricted

the size of the network to genes that were in the top 50% with regards to connectivity. This also reduces the overall network size and decreases the computational load while preserving scale-free topology. The resulting networks contained 10,024 genes. The adjacency matrices were further transformed to a topological overlap measure (TOM) similarity matrix. This procedure integrates information not only from the direct correlation between two gene expression patterns, but also from the correlation patterns of their network neighbors (Li and Horvath, 2007). The combination of the biweight mid-correlation coupled with the use of the topological overlap matrix transformation has been shown to outperform other measures in the discovery of biologically significant co-expression modules (Song et al., 2012).

We clustered the TOM based adjacency matrices utilizing average linkage and the WGCNA dynamicTreeCut function. Identified clusters (denoted as modules) are uniquely annotated based on their size by arbitrarily chosen colors. To preclude the emergence of highly similar modules, we further refined this procedure by examining the correlations between the module eigengenes (ME; 1st principal component). Additionally, module size was restricted to 30 genes as a lower bound to preserve any downstream statistical test assumption of normality.

Following identical procedures, separate networks were constructed utilizing the progressor and non-progressor samples. Each network's correlate weights were raised to the soft threshold powers of $\beta = 5$ and $\beta = 6$ for progressor and non-progressor conditions respectively (Figure S3).

Once we determined that modules were highly preserved across networks (see module preservation across networks), a consensus network was constructed utilizing the minimum adjacency values of the two networks, ensuring that high consensus co-expression values reflect high co-expression in both networks.

Consensus modules were detected utilizing the WGCNA blockwise Consensus Modules function, with the "max block size" parameter set to 10,024 to account for analysis of all genes in one block. This function takes a list of datasets as input, which in our case were the progressor and non-progressor samples. We utilized a dendrogram cut height of 0.995. Visualization of the resulting dendrograms and module color assignment is provided in Figures S4, S5.

Module Preservation Across Networks

As networks and modules were initially constructed independently in the two datasets, module overlap across networks was evaluated utilizing the WGCNA module Preservation function (Langfelder et al., 2011). This function computes module quality and preservation values. 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. In both cases, statistical significance is evaluated by bootstrapping ($N = 200$ permutations), which involves selecting random groups of genes of the same size

as the module being evaluated. The average co-expression values of random groups of genes are then compared with the co-expression values for genes in the module and the results are returned as Z scores with associated p -values.

Consensus Module Membership and Clinical Feature Relationship

For each gene, we compute measures of network connectivity and gene significance. Gene connectivity is a measure of relative gene importance within a network and/or module and is quantified utilizing two measures. Intramodular connectivity is denoted as k_{Within} and for a gene j is computed with the formula $k_{\text{Within}} = \sum_{i \neq j} a_{ij}$ where M_i denotes module i and a_{ij} denotes the adjacency between gene j and all other genes in module M_i . A distinct and complementary measure of gene connectivity is quantified by the correlation between the gene expression profile and the module eigengene; this quantity is denoted as $k_{\text{ME}} = \text{cor}(x_i, \text{ME})$. In many cases a linear relationship is expected between k_{ME} and k_{Within} ; this relationship was observed in our data (Figure S7).

Relationships between gene expression values and clinical features are denoted as gene significance (GS) and are quantified by the magnitude of the Pearson correlation coefficient $GS_i = |\text{cor}(x_i, F_{\text{clinical trait}})|$. In the present study we considered two clinical traits: tobacco use quantified as pack years, and alcohol use quantified as drinks per day. In addition to individual genes, gene modules can also be related to clinical traits; in this case the correlation is computed between the module eigengene and the clinical trait.

Differential Network Analysis

Although the consensus network contains consensus modules from both conditions, there may still be statistically significant differences between the two conditions' modules, detectable at the single gene level. Ultimately, these genes may retain informative progressor differences (Fuller et al., 2007; Iancu et al., 2013). We utilize the structure and module assignment of the consensus network to provide a system level context to the changes in the transcriptional profile between the two conditions.

At the individual gene level, we quantify three distinct and complementary changes in transcription: differentially expressed (DE), differentially variable (DV), and differentially wired (DW) genes. To detect DE genes, we used the eBayes function from the "limma" R package (limma³; Ritchie et al., 2015). For DV genes, we compared the variance of samples utilizing the var.test function in the "stats" R package (R: F Test to Compare Two Variances⁴). Differential wiring between two genes was evaluated by examining the difference in pairwise correlation between the progressor and non-progressor groups; this approach is adapted from previous approaches of quantifying network rewiring in both genomic (Gill et al., 2010) and neural imaging studies

³limma Bioconductor. Available online at: <http://bioconductor.org/packages/limma/> (Accessed September 2, 2017).

⁴R: F Test to Compare Two Variances Available online at: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/var.test.html> (Accessed September 2, 2017).

(Hosseini et al., 2012). We utilized the `r.test` function in the “psych” R package. For each gene, we recorded the number of significantly changed edges ($r.test\ p < 0.01$). Next, we evaluated whether the number of changed edges is above what can be expected by chance, given the total number of genes and the total number of changed edges. We utilized a binomial probability with the rate $r = \frac{\# \text{ changed edges}}{\# \text{ edges}}$, i.e., the number of **changed** edges divided by the total number of edges in the network. Under this model the null hypothesis is that the changed edges are randomly distributed among all the nodes and the alternative hypothesis is that some nodes will be enriched in changed edges. In the binomial distribution each gene has N trials, where N is the number of genes and also the number of edges for each gene. The probability of a gene having at least n_i changed edges is $p(n > n_i) = B(N, r)$. The set of DW genes was then taken to be the genes enriched in changed edges at $p < 0.01$.

We denote as “affected genes” the set of genes that are DE, DV, or DW. Next, we inquired whether individual modules contain more of the affected genes than what can be expected by chance given the size of the affected gene set, size of the network and the size of the module (Fisher’s exact test). Bonferroni corrections were done based on the number of modules.

Pathway Enrichment Analysis

Consensus modules’ pathway enrichment analysis was assessed utilizing ReactomeFIViz, the Reactome Cytoscape app (ReactomeFIViz—ReactomeWiki⁵; Wu et al., 2014). ReactomeFIViz’s basic data entities are known proteins and their functional interactions. The overall data network model is designed to capture known functional protein-protein interactions (PPI). This analysis was applied to the full gene set of all progressor consensus modules including those enriched in affected genes. Detailed pathway enrichment analysis tables (pathway, Binomial p -value, FDR, gene list) along with co-expression consensus modules gene lists (utilizable as input to ReactomeFIViz) can be found in Supplementary Data-2, 3, 4.

Software

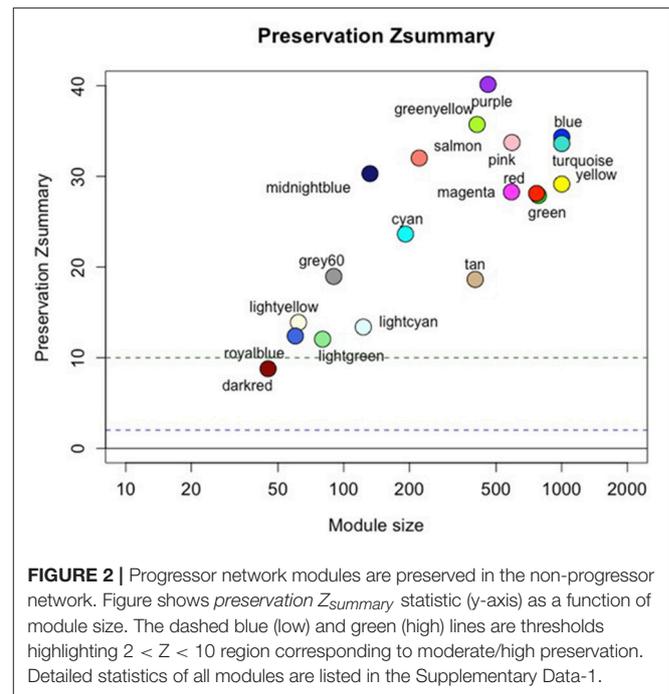
The R code for this project is open-source and can be accessed via <https://github.com/teslajoy>.

RESULTS

Progressor Network Construction and Preservation

We constructed a co-expression network based on the progressor samples, identifying 21 modules ranging in size from 45 to 1,127 genes. Modules were evaluated for both quality (Figure S4) and preservation in the non-progressor network (Figure 2—see also Supplementary Data-1). All progressor modules displayed high quality measures and also showed high preservation in the non-progressor weighted network (Figure 2; Supplementary Data-1).

⁵ReactomeFIViz—ReactomeWiki Available online at: <http://wiki.reactome.org/index.php/ReactomeFIViz> (Accessed September 2, 2017).



Consensus Network Construction

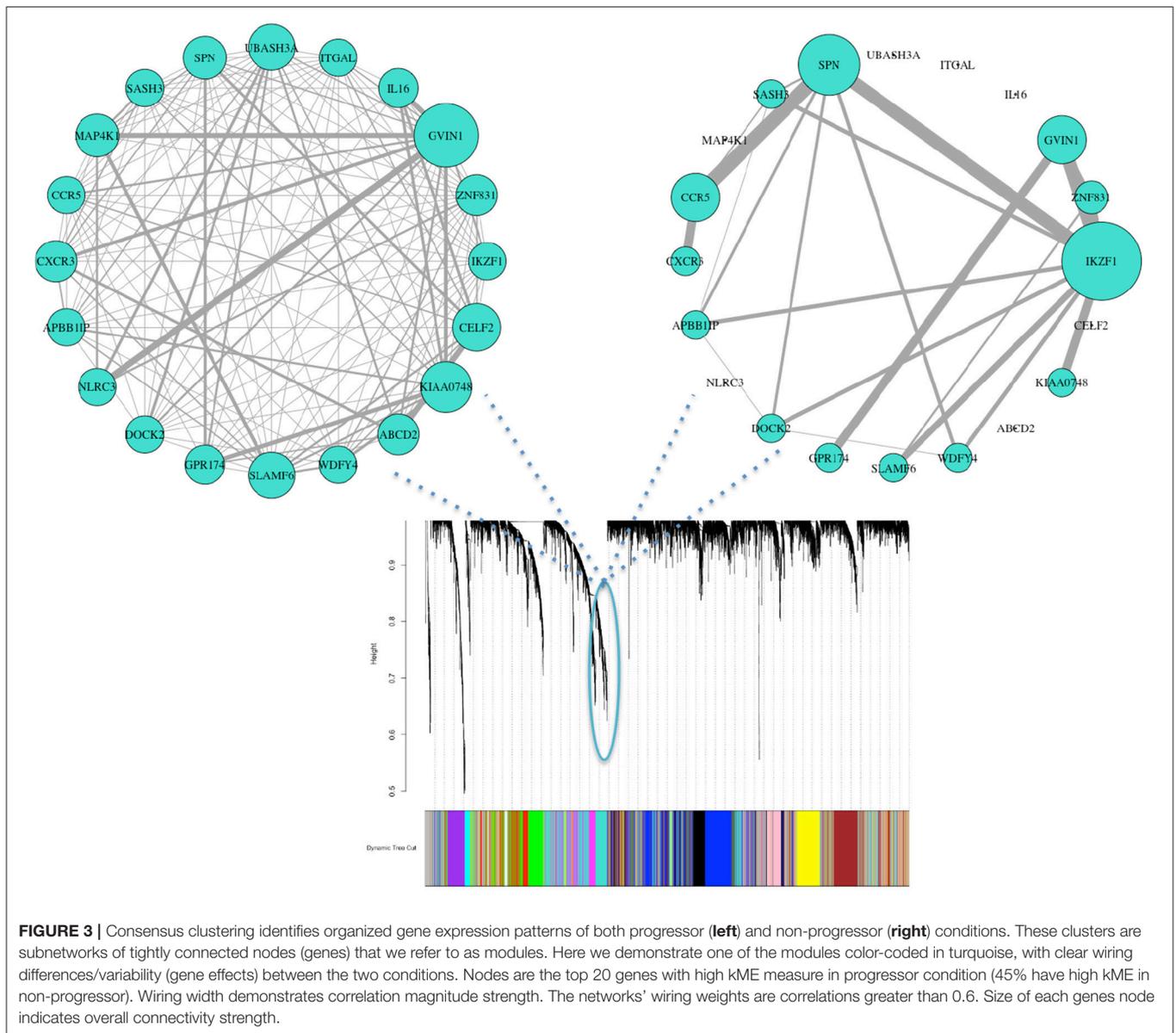
The preservation of the progressor modules in the non-progressor network indicates that overall network and module structure is preserved and no progressor-specific modules have emerged. In light of this finding, we combined the two conditions’ datasets and constructed clustering of consensus modules. Network adjacency in the consensus network was based on the minimum adjacency across the two conditions.

Biological consensus modules between the two conditions are made of similar gene structure but their interactions and expressions are variable (Figure 3). Also, since each consensus module retains different samples, progressors and non-progressors separately, their module properties are not the same (e.g., module eigengene).

We identified 18 consensus modules ranging from 71 to 1,389 genes (Figure 3, Table S1). There were 882 unclustered genes that were assigned to the “gray” pseudo module. All genes showed high module membership with kME and KIM Pearson correlations > 0.9 ($p < 0.01$; Figure S6).

Relating Smoking and Alcohol Exposure to Consensus Modules

We evaluated the relationship between gene expression and tobacco/alcohol use both at the level of individual genes and at the level of gene modules. At the individual genes, module membership (as measured by kME) showed strong correlation with gene significance for our clinical risk factors (see Figures S7–S9). At the module level, we correlated the module eigengenes to the same “pack years” and “drinks per



day” risk factors. We utilized the consensus module assignment, but eigengenes were constructed separately for the progressor and non-progressor networks. While the non-progressor module eigengenes showed only weak correlations with each risk factor, in the progressor network, seven module eigengenes had a stronger relationship with alcohol consumption; in contrast only the cyan module had a strong correlation with tobacco use (Figure 4).

Cancer Progression and Differential Network Analysis

Leveraging network properties at the individual gene level, we quantified differences in expression, variability, and correlation

(DE, DV, and DW respectively); each measures’ gene list is available via the Supplementary Data-3.

Not all modules had an equal proportion of affected genes. We identified 7,055 genes that were differentially expressed, 3,682 genes that were differentially variable, and 34 genes that were differentially wired. Overall, 12 modules were enriched in identified affected genes, with 11 having more DE or DV genes and only the turquoise module having above the expected DW genes (Table 1). All identified enriched modules’ expression profiles were evenly distributed and showed no outstanding noise/batch effect (Figure S10).

We further analyzed connectivity and hubness (potential driver events) of DW genes in the turquoise module. From the 34 DW genes, we identified 18 genes (shown in

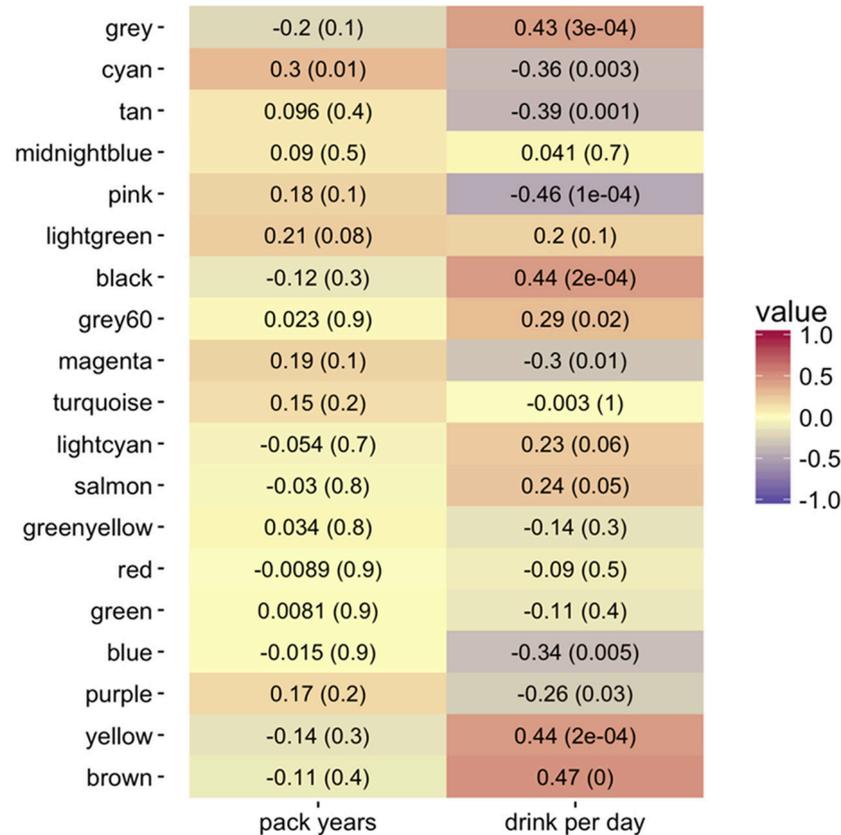


FIGURE 4 | Heatmap of Pearson correlations (−1:1 shown by color legend) for alcohol (drinks per day) and tobacco use (pack years) with the co-expression consensus module progressor eigengenes. The corresponding p -values are in parentheses. The brown, yellow, black, pink, tan, and cyan modules show the highest positive correlation with drinks per day. The cyan module shows the highest positive correlation with tobacco pack years smoked.

Table 2) that were both hubs ($kME > 0.8$) and enriched in changed edges ($p < 0.01$; Binomial). The visual network structure and wiring differences of these 18 genes between HNSCC progressor and non-progressor conditions can be seen in **Figure 5**. This result confirms our hypothesis on the existence of alterations in the co-expression network structure that differentiate progressor from non-progressor populations.

Pathway Enrichment Analysis of Enriched Modules

We performed a pathway enrichment analysis of the modules enriched in DE/DV/DW affected genes. To capture known pathways, we included all genes of enriched modules.

Pathway enrichment analysis of the progressor module enriched in DW genes showed involvement in inflammation, T cells, B cells, natural killer cells, adhesion, and the Jak-Stat pathway. This result adds to the evidence in the Bornstein et al. study (2016) and reemphasizes the potential importance of cancer associated inflammation pathways

and tumor microenvironment evolutionary mechanisms in tumors with higher progression rates. We also found the pathway association between the identified 18 differentially wired hub genes has not been discovered and remains unknown.

Modules enriched in DE or DV genes showed involvement in multiple hallmarks of cancer and tumor mechanism dynamics such as cell cycle check points, abnormal mitosis, spindle bipolarity, receptor synapse deregulation and important cellular pathways such as *c-myc*, MAPK, Jak-Stat, and P53. There were also noted associations between the enriched pathways and diseases such as Alzheimer's and Parkinson's. This characterized information could potentially further our research on tumor pathway completion studies.

We also assessed the overlap of enriched modules in pathways with $FDR < 0.05$. This was to investigate the co-dependency or overlap of biological mechanisms despite the measure of affected genes (**Table 3**). We found 149 pathways that overlapped between modules and 16 pathways that involved DV or DW enriched modules. The affected genes

TABLE 1 | Summary of identified consensus modules enriched in affected genes via differential network analysis between progressor and non-progressor conditions.

DW	DV	DE
Turquoise (14)	Black (242)	Black (513)
	Purple (212)	Blue (1298)
	Pink (236)	Cyan (148)
	Light cyan (71)	Tan (215)
	Salmon (95)	Gray60 (71)
		Yellow (658)
		Light green (69)

The number of affected genes found in each module is noted in parentheses.

were identified to be unique to each associated pathway, but overlapping between pathways. Notably, pathways associated with DW measure show high involvement of inflammation and tumor microenvironment evolution mechanisms.

DISCUSSION

Previously annotated and curated TCGA HNSCC data (Bornstein et al., 2016) provides the opportunity to study samples from both progressors and non-progressors from a network perspective. Weighted networks analysis can provide a holistic view on disease dynamics, but also enables us to reduce the complexity into organized and measurable relations. We were able to reduce the expression data from over 10,000 genes to 18 mechanistic modules. Within these modules we further focused on gene hubs with high correlation with alcohol/tobacco use, or with changing network profiles. Finally, through pathway enrichment analysis, we provide a context for the activity of these genes and relate them to biologically relevant pathways through the human protein Interactome.

Among the modules enriched in DE/DV/DW genes we identified aberrations in molecular pathways important for cell cycle check points, mitosis and spindle bipolarity, macrophage activity, immune response and T cells, interferon gamma signaling pathway, as well as c-myc, MAPK, Jak-Stat, and P53 pathways (see Supplementary Materials for a complete list of affected pathways). More specifically, pathway enrichment analysis of the module enriched in DW genes highlighted involvement of inflammation and tumor microenvironment evolution mechanisms. These results are an important extension to previous finding of Bornstein et al. (2016). We identified *IL10RA*, *DOK2*, *APBB1IP*, *UBASH3A*, *SASH3*, *CELF2*, *TRAF3IP3*, *GIMAP6*, *MYO1F*, *NCKAP1L*, *WAS*, *FERMT3*, *SLA*, *SELPLG*, *TNFRSF1B*, *WIPF1*, *AMICA1*, *PTPN22* as differentially wired hub genes (putative network driver genes). This putative signature represents an early step toward developing stratification techniques based on tumor

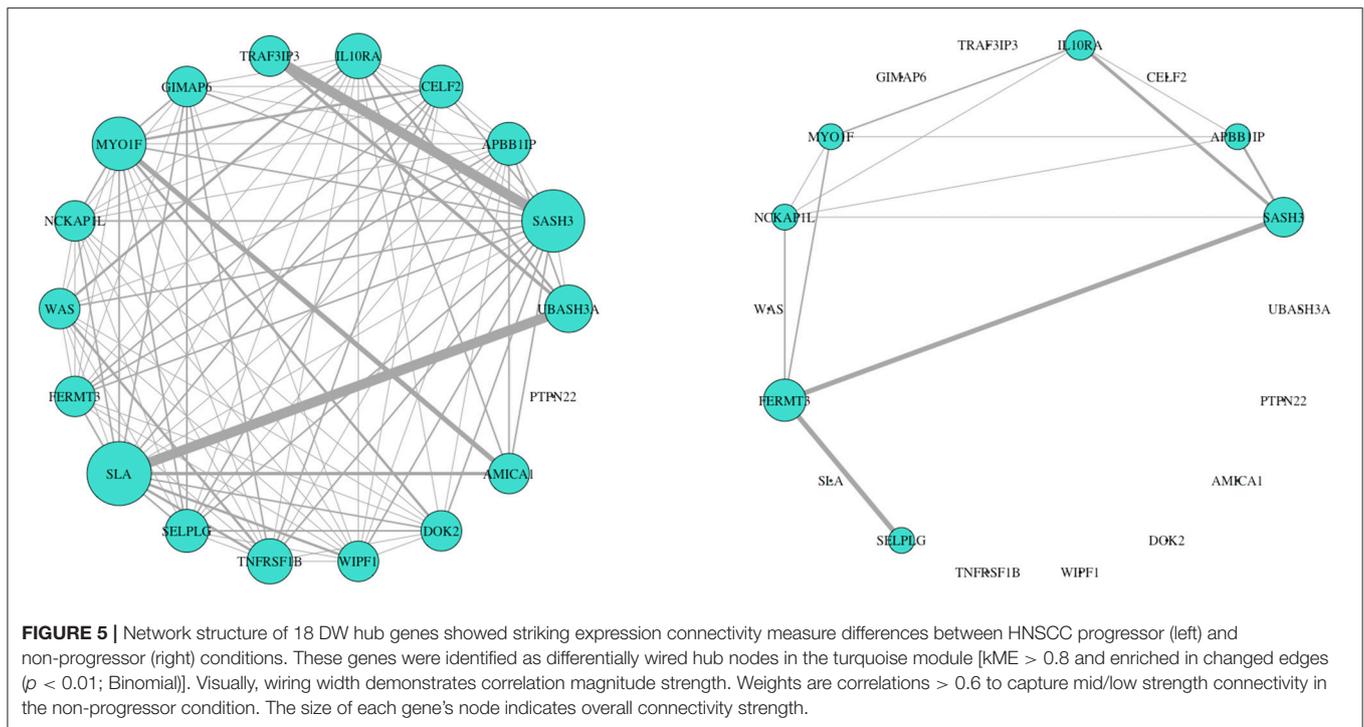
TABLE 2 | Summary measure of 18 genes identified as differentially wired hub genes in the turquoise module [kME > 0.8 and enriched in changed edges ($p < 0.01$; Binomial)].

DW Genes	kME	Changed edge enrichment p -value
<i>UBASH3A</i>	0.9378	2.4596e-04
<i>SASH3</i>	0.9325	1.8283e-06
<i>APBB1IP</i>	0.9278	2.4596e-04
<i>CELF2</i>	0.9205	1.7880e-40
<i>IL10RA</i>	0.9054	2.4596e-04
<i>TRAF3IP3</i>	0.9005	1.8283e-06
<i>GIMAP6</i>	0.8985	1.5076e-18
<i>MYO1F</i>	0.8969	2.4596e-04
<i>NCKAP1L</i>	0.8965	2.4596e-04
<i>WAS</i>	0.8943	2.4596e-04
<i>FERMT3</i>	0.8919	2.4596e-04
<i>SLA</i>	0.8846	2.4596e-04
<i>SELPLG</i>	0.8813	2.4596e-04
<i>TNFRSF1B</i>	0.8745	1.0199e-08
<i>WIPF1</i>	0.8677	1.8283e-06
<i>DOK2</i>	0.8419	2.3464e-43
<i>AMICA1</i>	0.8368	8.6198e-35
<i>PTPN22</i>	0.8226	2.4596e-04

The genes are sorted by the magnitude of kME.

genomics that could potentially identify patients with a high likelihood of tumor progression early on. Interestingly, it has been shown that gene expression networks show characteristic changes in correlation during transition from normal to a disease state (Censi et al., 2011) as well as during normal cellular differentiation (Mojtahedi et al., 2016). Given these measurable changes in network properties and the ability of differential wiring analysis to provide a global descriptor for the network, it is feasible that—with more samples—a gene expression network threshold could be defined for the progressor phenotype. Utilized as a surveillance tool, this could provide clinicians with an early warning for patients at risk of developing progressive disease. Furthermore, if the progressor-specific molecular pathways that we have identified are validated in cell and animal models, our findings could aid in the development of targeted therapy for HNSCC.

We also assessed the overlap of significantly enriched modules in pathways (Table 3). We found 149 pathways that overlapped between modules and 16 that overlapped modules specifically enriched in DV/DW genes. The affected genes were identified to be unique to each associated pathway, but overlapping between pathways. Given the clinical phenotype of interest, progression, pathways associated with DW measure show high involvement of inflammation and tumor microenvironment evolution mechanisms. The identification of Ras signaling as a pathway that overlaps with multiple modules enriched



in DV/DW genes serves as a validation of this type of network analysis in tumor genomics. This well-known cellular signaling pathway is involved in critical functions such as cell growth, migration, adhesion, cell survival, and cell differentiation (Rajalingam et al., 2007; Fernández-Medarde and Santos, 2011). Aberrations in the Ras pathway are some of the most frequent findings in human cancers and have been previously described in some HNSCC samples (Hoa et al., 2002; Rothenberg and Ellisen, 2012). Interestingly, in colorectal cancer, the EGFR antibody, cetuximab, showed benefit in clinical trials, but the benefits did not hold for patients with Ras mutations. Thus, the development of targeted agents for the Ras pathway is an active area of research (Bahrami et al., 2017; Simanshu et al., 2017). Also of interest is the identification of the VEGF signaling pathway. This family of growth factors is involved in angiogenesis and its members have been shown to be involved in a variety of human malignancies (Olsson et al., 2006; Stacker and Achen, 2013). Furthermore, previous studies have linked VEGF signaling to HNSCC (Tong et al., 2008; Lucas et al., 2010). As therapeutic options are currently limited in HNSCC, our finding adds strength to the evidence that VEGF signaling may be a potentially targetable pathway in HNSCC.

The current focus of therapeutic advancement in cancer calls for a precision/personalized approach based on detectable molecular abnormalities of a patient's tumor that can then be exploited in a targeted manner. However, many questions still remain open as to what the best methods of molecular analysis

are, who is likely to benefit, and why. Our results suggest that a stratification procedure could benefit from inclusion of transcriptional hub genes related to disease progression. Additionally, DW measure assessment over temporal data could potentially be informative regarding targeted therapy failure or drug outcome predictions between subpopulations which has not been clinically evaluated (e.g., currently we have only assessed EGFR and cetuximab). If this measure is captured longitudinally, it has the potential to reveal novel discoveries on disease state and tumor evolution mechanisms.

Beyond progression status, we also evaluated gene expression correlations with alcohol and tobacco use. 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 modules (Figure 4; Figure S9). We hypothesize that this is potentially due to patients' stronger recall on the quantity they drink per day vs. packs of cigarettes they smoke through a year. This finding underlies the importance of advancing the quality of measured clinical data for improving research results.

Overall we have shown that the use of *de-novo* weighted network inference in the context of biological pathways provides new insights for both mechanistic and prognostically relevant information in HNSCC. Follow-up studies can incorporate other clinical phenotypes such as measurements derived from tumor imaging and could ultimately lead to a greater understanding of these tumors.

TABLE 3 | Summary of pathway enrichment analysis showing pathways (FDR < 0.05) that overlapped between modules and are enriched in genes showing DV/DW.

Pathway	Modules	Measure	FDR	Genes
Chemokine signaling pathway(K)	Turquoise, salmon	DW, DV	0, 0.002	<i>GNG2, PIK3CG, NCF1, GNGT2, LYN, WAS, CCL25, CCL22, CCL21, ARRB1, CX3CL1, CCL5, DOCK2, VAV1, CXCL13, XCL2, RAF1, JAK2, JAK3, ADCY6, CCR8, CCR7, CCR6, CCR5, CCR4, CX3CR1, CCR2, XCR1, CXCL9, STAT3, ITK, CXCR3, PLCB3, CXCR4, CXCR6, PLCB2, PRKCB, RASGRP2, PIK3R5, PIK3R1, BCAR1, CXCR5, PIK3R3, GNG7, CCL19</i>
Platelet activation(K)	Turquoise, pink	DW, DV	5e-04, 0.038	<i>BTK, ACTG1, PIK3CG, PPP1CA, LYN, APBB1IP, P2RY12, MYL12B, P2RX1, LCP2, ADCY6, PLA2G4C, ITPR1, GP5, PLCB3, PLCB2, RASGRP1, RASGRP2, PIK3R5, PIK3R1, MAPK3, GNAI3, GNAI1, PLA2G4F, PLA2G4E, PLA2G4D</i>
B cell receptor signaling pathway(K)	Turquoise, salmon	DW, DV	6e-04, 0	<i>BTK, PIK3CG, NFKBIE, RAC3, LYN, CARD11, VAV1, INPP5D, NFATC2, RAF1, PTPN6, PRKCB, PIK3R5, PIK3R1, CD22, PIK3R3, CR2, CD19, CD79B, CD79A</i>
BCR signaling pathway(N)	Turquoise, salmon	DW, DV	8.00E-04, 8.00E-04	<i>BTK, LYN, CARD11, RASA1, INPP5D, RAF1, MAP4K1, PTPN6, MAP3K1, POU2F2, PIK3R1, PTPRC, SH3BP5, CD22, CD19, CD79B, CD79A</i>
Fc gamma R-mediated phagocytosis(K)	Turquoise, pink	DW, DV	0.0018, 0.038	<i>PIK3CG, NCF1, SPHK2, LYN, WAS, DOCK2, VAV1, LAT, PIP5K1B, INPP5D, RAF1, PRKCB, PIK3R5, PIK3R1, PTPRC, PLD2, MAPK3, PLA2G4F, PLA2G4E, PLA2G4D</i>
GPCR downstream signaling(R)	turquoise, light cyan	DW, DV	0.011, 0.0438	<i>GNG2, PIK3CG, SSTR3, GAST, RGS1, NMUR1, ARHGEF4, ARHGFE6, GPR65, CCL25, CCL21, P2RY12, P2RY13, P2RY10, P2RY14, PDE3B, PDE4D, GPR55, FGD2, DRD1, CYSLTR1, ADORA2A, CCL5, ADRA2A, FGD3, VAV1, CXCL13, XCL2, CSF2RB, RHOB, CSF2RA, GPR18, JAK2, JAK3, ADCY6, S1PR4, GPR132, CCR8, CCR7, CCR6, CCR5, CCR4, CCR2, ABR, CCKAR, OXER1, XCR1, CXCL9, NMB, ITPR1, VIPR2, PTHLH, IL3RA, INSL3, CXCR3, ARHGAP4, PLCB3, CXCR4, CXCR6, PLCB2, RAMP3, GHRL, RASGRP1, RASGRP2, IL2RG, PIK3R5, PIK3R1, IL2RB, GNA14, TACR1, HTR2B, MCF2L</i>
Oxytocin signaling pathway(K)	Turquoise, purple, pink	DW, DV	0.0159, 0, 0.0024	<i>ACTG1, PIK3CG, PPP1CA, KCNJ2, CACNA2D4, CACNA1F, NFATC2, RAF1, ADCY6, CAMK4, PLA2G4C, ITPR1, PLCB3, PLCB2, PRKCB, PPP1R12B, PIK3R5, PIK3R1, MEF2C, ADCY2, MYLK3, MYLK2, RYR1, CACNB1, CACNA1S, PRKAG3, CACNG6, CACNG1, CAMK2B, PRKAA2, CAMK2A, PRKAB1, MAPK3, GNAI3, GNAI1, ELK1, CDKN1A, PLA2G4F, PLA2G4E, PLA2G4D, CALM1</i>
Ras signaling pathway(K)	Turquoise, pink	DW, DV	0.0183, 0.0206	<i>GNG2, PIK3CG, RAC3, ZAP70, GNGT2, RASA1, RASSF5, RGL1, MET, LAT, RRAS2, RAF1, EFNA3, PLA2G4C, FOXO4, STK4, RASAL3, FASLG, PRKCB, RASGRP1, RASGRP2, PIK3R5, PIK3R1, GAB1, PLD2, MAPK3, PLA2G2F, ELK1, PLA2G4F, PLA2G4E, PLA2G4D, CALM1</i>
Gastrin-CREB signaling pathway via PKC and MAPK(R)	Turquoise, pink, light cyan	DW, DV	0.026, 0.0129, 0.0127	<i>GNG2, GAST, NMUR1, GPR65, P2RY10, CYSLTR1, XCL2, RAF1, JAK2, GPR132, CCKAR, XCR1, NMB, ITPR1, PLCB3, PLCB2, GHRL, RASGRP1, RASGRP2, PIK3R1, GNA15, LTB4R, LPAR5, MAPK3, BDKRB1, BDKRB2, LTB4R2, GNRHR, ANXA1, GNA14, TACR1, HTR2B</i>
Neurotransmitter Receptor Binding And Downstream Transmission In The Post-synaptic Cell(R)	Turquoise, pink, purple	DW, DV	0.0295, 0.0049	<i>GNG2, KCNJ2, GNGT2, GRIK5, RAF1, ADCY6, CHRNA7, CHRNA6, CAMK4, KCNJ10, AKAP5, PLCB3, PLCB2, PRKCB, MDM2, ADCY2, CHRNA1, ACTN2, CHRND, CHRNG, CAMK2B, CAMK2A</i>
VEGF signaling pathway(K)	Turquoise, pink	DW, DV	0.0334, 0.0451	<i>PIK3CG, RAC3, SPHK2, NFATC2, RAF1, PLA2G4C, PRKCB, PIK3R5, PIK3R1, MAPK3, PLA2G4F, PLA2G4E, PLA2G4D</i>
Gastric acid secretion(K)	Turquoise, purple	DW, DV	0.0354, 2e-04	<i>GAST, KCNJ2, ADCY6, KCNJ10, ATP1A3, ATP1A4, ITPR1, PLCB3, PLCB2, PRKCB, ADCY2, ATP1B4, MYLK3, MYLK2, CAMK2B, CAMK2A, ATP1A2</i>
Central carbon metabolism in cancer(K)	turquoise, pink	DW, DV	0.0494, 0.0176	<i>PIK3CG, TP53, SLC7A5, SLC2A1, MET, FLT3, RAF1, PIK3R5, PIK3R1, PGAM1, PGAM4, HIF1A, MAPK3, HK2</i>
Calcium signaling pathway(K)	purple, pink, light cyan	DV	0, 0.0451, 5e-04	<i>SLC8A3, ADCY2, TNNC2, TNNC1, ATP2B2, MYLK3, MYLK2, RYR1, CACNA1S, ATP2A1, PHKG1, CAMK2B, CAMK2A, PLN, GNA15, ITPKA, BDKRB1, BDKRB2, PPIF, LTB4R2, CALM1, GNA14, TACR1, HTR2B, CACNA1D</i>
Glutamatergic synapse(K)	pink, salmon	DV	0.0024, 0.0437	<i>PLD2, MAPK3, GNAI3, GNAI1, HOMER1, HOMER2, PLA2G4F, PLA2G4E, PLA2G4D, HOMER3, GNG7, GRM7</i>
Serotonergic synapse(K)	pink, light cyan	DV	0.0122, 0.0438	<i>MAPK3, GNAI3, GNAI1, PLA2G4F, PLA2G4E, PLA2G4D, ALOX12B, HTR2B, CACNA1D</i>

In order to highlight findings from the network analysis, DE modules have been excluded. Genes of the associated pathway are unique within each pathway, but overlap between pathways (Supplementary Data-5).

AUTHOR CONTRIBUTIONS

NS conceived of the study and developed or extended code, performed analysis and wrote the majority of the manuscript. GW, OI, and SM provided input into the design and analyses. JJ provided clinical perspective. GW, OI, SM, and JJ provided guidance on presentation of data and contributed to editing of the manuscript.

ACKNOWLEDGMENTS

The results published here are based upon data generated by the TCGA Research Network. This work was supported by the National Library of Medicine Informatics Training

Grant T15LM007088 (JJ), the National Cancer Institute 1R01CA192405 (SM), the National Human Genome Research Institute 2U41HG003751 (GW), and National Center for Advancing Translational Sciences 5UL1TR000128 (GW, SM). We thank both reviewers for their constructive feedback which improved the paper tremendously. We also wish to acknowledge the contribution of Reviewer 1 with regard to the suggestion of the extension of this work for individual patient connectivity analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00183/full#supplementary-material>

REFERENCES

- Araya, C. L., Cenik, C., Reuter, J. A., Kiss, G., Pande, V. S., Snyder, M. P., et al. (2016). Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nat. Genet.* 48, 117–125. doi: 10.1038/ng.3471
- Bahrami, A., Hassanian, S. M., ShahidSales, S., Farjami, Z., Hasanzadeh, M., Anvari, K., et al. (2017). Targeting RAS signaling pathway as a potential therapeutic target in the treatment of colorectal cancer. *J. Cell. Physiol.* 233, 2058–2066. doi: 10.1002/jcp.25890
- Bonner, J. A., Harari, P. M., Giralt, J., Cohen, R. B., Jones, C. U., Sur, R. K., et al. (2010). Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival. *Lancet Oncol.* 11, 21–28. doi: 10.1016/S1470-2045(09)70311-0
- Bornstein, S., Schmidt, M., Choonoo, G., Levin, T., Gray, J., Thomas, C. R., et al. (2016). IL-10 and integrin signaling pathways are associated with head and neck cancer progression. *BMC Genomics* 17:38. doi: 10.1186/s12864-015-2359-6
- Censi, F., Giuliani, A., Bartolini, P., and Calcagnini, G. (2011). A Multiscale graph theoretical approach to gene regulation networks: a case study in atrial fibrillation. *IEEE Trans. Biomed. Eng.* 58, 2943–2946. doi: 10.1109/TBME.2011.2150747
- Fernández-Medarde, A., and Santos, E. (2011). Ras in cancer and developmental diseases. *Genes Cancer* 2, 344–358. doi: 10.1177/1947601911411084
- Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm. Genome* 18, 463–472. doi: 10.1007/s00335-007-9043-3
- Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics* 11:95. doi: 10.1186/1471-2105-11-95
- Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013
- Ho, J. W., Stefani, M., dos Remedios, C. G., and Charleston, M. A. (2008). Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics* 24, i390–i398. doi: 10.1093/bioinformatics/btn142
- Hoa, M., Davis, S. L., Ames, S. J., and Spanjaard, R. A. (2002). Amplification of wild-type k-ras promotes growth of head and neck squamous cell carcinoma. *Cancer Res.* 62, 7154–7156.
- Hosseini, S. M., Hoefft, F., and Kesler, S. R. (2012). GAT: a graph-theoretical analysis toolbox for analyzing between-group differences in large-scale structural and functional brain networks. *PLoS ONE* 7:e40709. doi: 10.1371/journal.pone.0040709
- Hudson, N. J., Reverter, A., and Dalrymple, B. P. (2009). A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput. Biol.* 5:e1000382. doi: 10.1371/journal.pcbi.1000382
- Grant T15LM007088 (JJ), the National Cancer Institute 1R01CA192405 (SM), the National Human Genome Research Institute 2U41HG003751 (GW), and National Center for Advancing Translational Sciences 5UL1TR000128 (GW, SM). We thank both reviewers for their constructive feedback which improved the paper tremendously. We also wish to acknowledge the contribution of Reviewer 1 with regard to the suggestion of the extension of this work for individual patient connectivity analysis.
- Iancu, O. D., Oberbeck, D., Darakjian, P., Kawane, S., Erk, J., McWeeney, S., et al. (2013). Differential network analysis reveals genetic effects on catalepsy modules. *PLoS ONE* 8:e58951. doi: 10.1371/journal.pone.0058951
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559
- Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Comput. Biol.* 7:e1001057. doi: 10.1371/journal.pcbi.1001057
- Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213
- Li, A., and Horvath, S. (2007). Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics Oxf. Engl.* 23, 222–231. doi: 10.1093/bioinformatics/btl581
- Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323
- Lucas J. T. Jr, Salimath, B. P., Slomiany, M. G., and Rosenzweig, S. A. (2010). Regulation of invasive behavior by vascular endothelial growth factor is HEF1-dependent. *Oncogene* 29: onc2010185. doi: 10.1038/onc.2010.185
- Marur, S., and Forastiere, A. A. (2016). Head and neck squamous cell carcinoma: update on epidemiology, diagnosis, and treatment. *Mayo Clin. Proc.* 91, 386–396. doi: 10.1016/j.mayocp.2015.12.017
- Mojtahedi, M., Skupin, A., Zhou, J., Casta-o, I. G., Leong-Quong, R. Y. Y., Chang, H., et al. (2016). Cell fate decision as high-dimensional critical state transition. *PLoS Biol.* 14:e2000640. doi: 10.1371/journal.pbio.2000640
- Olsson, A.-K., Dimberg, A., Kreuger, J., and Claesson-Welsh, L. (2006). VEGF receptor signalling? in control of vascular function. *Nat. Rev. Mol. Cell Biol.* 7:nrm1911. doi: 10.1038/nrm1911
- Patel, S. G., and Shah, J. P. (2005). TNM staging of cancers of the head and neck: striving for uniformity among diversity. *CA Cancer J. Clin.* 55, 242–264. doi: 10.3322/canjclin.55.4.242
- Rajalingam, K., Schreck, R., Rapp, U. R., and Albert, Š. (2007). Ras oncogenes and their downstream targets. *Biochim. Biophys. Acta BBA Mol. Cell Res.* 1773, 1177–1195. doi: 10.1016/j.bbamcr.2007.01.012
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007
- Rothenberg, S. M., and Ellisen, L. W. (2012). The molecular pathogenesis of head and neck squamous cell carcinoma. *J. Clin. Invest.* 122, 1951–1957. doi: 10.1172/JCI59889
- Simanshu, D. K., Nissley, D. V., and McCormick, F. (2017). RAS proteins and their regulators in human disease. *Cell* 170, 17–33. doi: 10.1016/j.cell.2017.06.009

- Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 13:328. doi: 10.1186/1471-2105-13-328
- Sparano, J. A., and Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 26, 721–728. doi: 10.1200/JCO.2007.15.1068
- Stacker, S. A., and Achen, M. G. (2013). The VEGF signaling pathway in cancer: the road ahead. *Chin. J. Cancer* 32, 297–302. doi: 10.5732/cjc.012.10319
- Tong, M., Lloyd, B., Pei, P., and Mallery, S. R. (2008). Human head and neck squamous cell carcinoma cells are both targets and effectors for the angiogenic cytokine, VEGF. *J. Cell. Biochem.* 105, 1202–1210. doi: 10.1002/jcb.21920
- Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H. N., Shi, S. T., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* 46, 573–582. doi: 10.1038/ng.2983
- Wu, G., Dawson, E., Duong, A., Haw, R., and Stein, L. (2014). ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Research* 3:146. doi: 10.12688/f1000research.4431
- Wu, G., and Stein, L. (2012). A network module-based method for identifying cancer prognostic signatures. *Genome Biol.* 13:R112. doi: 10.1186/gb-2012-13-12-r112
- Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17. doi: 10.2202/1544-6115.1128

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Sanati, Iancu, Wu, Jacobs and McWeeney. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.