



OPEN

# Systems-level differential gene expression analysis reveals new genetic variants of oral cancer

Syeda Zahra Abbas, Muhammad Imran Qadir &amp; Syed Aun Muhammad

Oral cancer (OC) ranked as eleventh malignancy worldwide, with the increasing incidence among young patients. Limited understanding of complications in cancer progression, its development system, and their interactions are major restrictions towards the progress of optimal and effective treatment strategies. The system-level approach has been designed to explore genetic complexity of the disease and to identify novel oral cancer related genes to detect genomic alterations at molecular level, through cDNA differential analysis. We analyzed 21 oral cancer-related cDNA datasets and listed 30 differentially expressed genes (DEGs). Among 30, we found 6 significant DEGs including CYP1A1, CYP1B1, ADCY2, C7, SERPINB5, and ANAPC13 and studied their functional role in OC. Our genomic and interactive analysis showed significant enrichment of xenobiotics metabolism, p53 signaling pathway and microRNA pathways, towards OC progression and development. We used human proteomic data for post-translational modifications to interpret disease mutations and inter-individual genetic variations. The mutational analysis revealed the sequence predicted disordered region of 14%, 12.5%, 10.5% for ADCY2, CYP1B1, and C7 respectively. The MiRNA target prediction showed functional molecular annotation including specific miRNA-targets hsa-miR-4282, hsa-miR-2052, hsa-miR-216a-3p, for CYP1B1, C7, and ADCY2 respectively associated with oral cancer. We constructed the system level network and found important gene signatures. The drug-gene interaction of OC source genes with seven FDA approved OC drugs help to design or identify new drug target or establishing novel biomedical linkages regarding disease pathophysiology. This investigation demonstrates the importance of system genetics for identifying 6 OC genes (CYP1A1, CYP1B1, ADCY2, C7, SERPINB5, and ANAPC13) as potential drugs targets. Our integrative network-based system-level approach would help to find the genetic variants of OC that can accelerate drug discovery outcomes to develop a better understanding regarding treatment strategies for many cancer types.

Oral Cancer constitutes approximately 90% among all Head and Neck Cancer (HNC) sub-types<sup>1</sup>. However, it is more prominent in urban areas of South Asia with a ratio of 15–40% among all cancer types<sup>2</sup>. In Pakistan, it ranked as 2nd most prevalent cancer-type, with increasing incidence in the past few years<sup>3,4</sup>. The complexity of genetic mechanisms in cancer has been revealed through recent investigations. Many biological systems seem to be involved in the development and progression of the cancer. But, the complications in system-interactions are limitedly understood which is a major restriction in developing effective treatments<sup>5</sup>. The gene expression studies may help to investigate the differential expression of genes in different biological states, cell cycle stages, subjects or tissues. This gene expression analysis is an important pinpoint for investigating biological processes and their functional disorders. cDNA microarrays were used to monitor and reveal the expression level for thousands of the genes which are differentially expressed in tumors simultaneously<sup>5</sup>. This technique can exploit this valuable information regarding gene expression analysis. This rapidly progressing technique provides comprehensive data for gene expression profiling of thousands of genes to the investigators in one experiment. Many studies demonstrated that this technique is useful to identify novel genes for cancer and its molecular level classification in human<sup>6,7</sup>. Thus, this novel technique may help us to identify new potential targets for drug development for optimal and effective disease therapies. It may also establish an important link between clinical medicine and gene sequences for humans<sup>8</sup>. Gene Expression Omnibus (GEO) is a publicly available database that can be used for high-throughput screening of molecular variants. It contains the microarray data including single

Institute of Molecular Biology and Biotechnology, Bahauddin Zakariya University, Multan, Pakistan. email: aunmuhammad78@yahoo.com

S. no	Geo ID	Sample count (case: control)	Platform used	Tissues
1	GSE2280	22:05	GPL96[HG-U133A] Affymetrix Human Genome U133A Array	Oral lymph node
2	GSE3524	16:04	GPL96[HG-U133A] Affymetrix Human Genome U133A Array	Oral squamous
3	GSE10063	4:30	GPL570[HG-U133_Plus_2] Affymetrix Human Genome U133	Keratinocyte cell
4	GSE13601	31:27	GPL8300 [HG_U95Av2] Affymetrix Human Genome U95 Array	Tongue
5	GSE21866	2:03	GPL201 [HG-Focus] Affymetrix Human HG-Focus Target Array	Tongue squamous
6	GSE32142	2:02	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Oral squamous
7	GSE36111	5:00	GPL571 [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array	Oral squamous
8	GSE38058	4:00	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Oral squamous
9	GSE38517	9:11	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Fibroblasts
10	GSE39376	11:17	GPL201 [HG-Focus] Affymetrix Human HG-Focus Target Array	Buccal carcinoma cell
11	GSE43862	1:03	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Oral squamous
12	GSE44458	4:00	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Tongue squamous
13	GSE49673	6:06	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Parotid adenocarcinoma
14	GSE52811	4:04	GPL8786 [miRNA-1] Affymetrix Multispecies miRNA-1 Array	Oral keratinocyte
15	GSE52915	27:00	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Tongue squamous
16	GSE57022	2:02	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Parotid adenocarcinoma
17	GSE59795	10:10	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Parotid adenocarcinoma
18	GSE70301	3:03	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133	Tongue squamous
19	GSE73171	3:03	GPL14613 [miRNA-2] Affymetrix Multispecies miRNA-2 Array	Laryngeal squamous
20	GSE75127	4:4	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	Oral squamous
21	GSE81821	5:5	GPL14613 [miRNA-2] Affymetrix Multispecies miRNA-2 Array	Metastatic tumor tissue

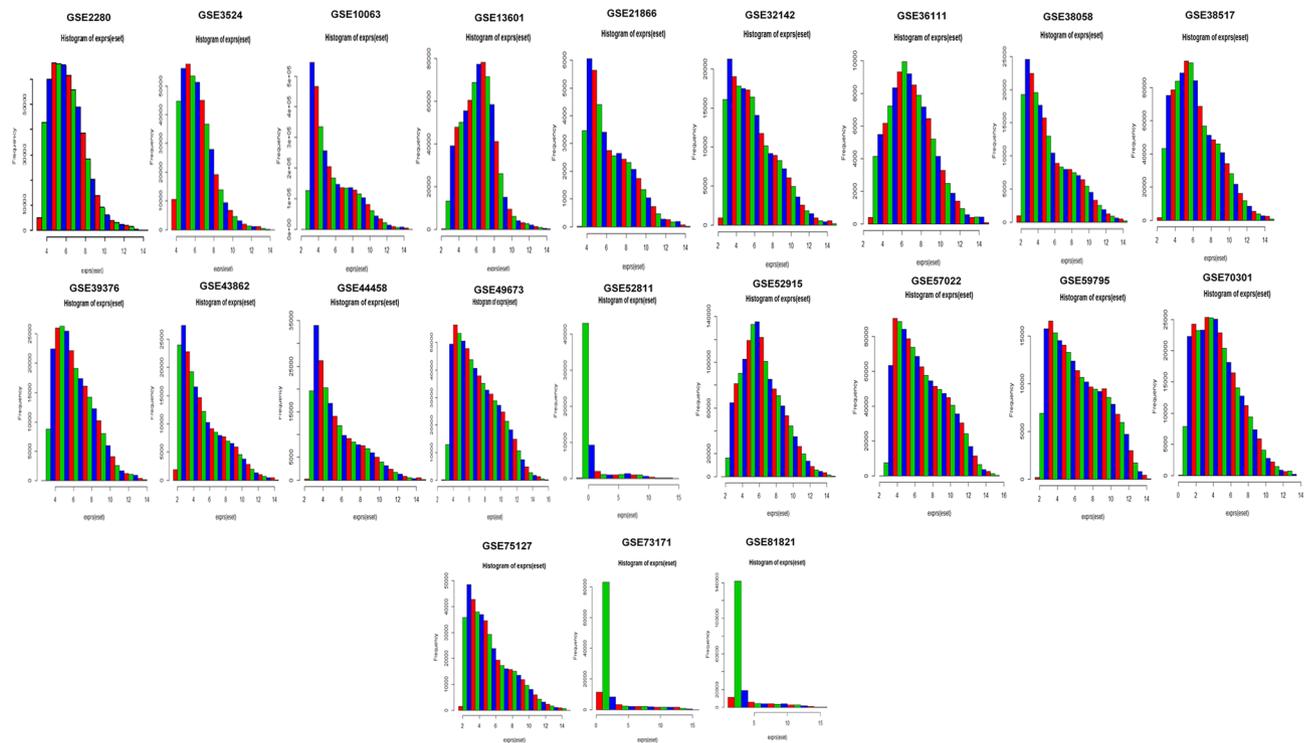
**Table 1.** List of cDNA datasets analyzed in this study.

and multiple channel-based experiments to determine abundance of genomic DNA, mRNA and protein molecules. The scientific community widely uses the gene expression database to analyze and search the molecular components for systems-level investigations. Several microarrays applications for meta-analysis are designed to investigate pathological mechanisms associated with genetic risk factors<sup>9–11</sup>. Therefore, the genome expression analysis may help to identify unusual alterations across the genome using microarray technology<sup>12</sup>. This study aims to identify genetic causes and some probable genetic variants of oral cancer that will help to modify the therapeutic strategies.

## Results

The cDNA datasets used in this study belong to various normal or cancerous oral tissues and cell lines. The list of datasets used in this study are shown in Table 1.

**Differential analysis, normalization and cross-validation.** We analyzed 21 cDNA datasets specifically associated with oral cancer. Each dataset has a different number of samples and the genes derived through mRNA expression profiling using different Affymetrix platforms for OC. The histograms representing expression after normalization indicate the density estimation of data. The shapes and ranges for arrays distributions are similar indicating the quality of the data. The array's distribution towards the right shows a high background level. The saturation of signal is specified by bulge which appears at the upper end of intensity range (Fig. 1). An automated mechanism was used to perform a comparison of biologically similar groups in pairs. We let off any subgroup without repetition from comparisons for accuracy and verification of differential analysis, and the generalized linear models' `cv.glm` method measured the error of the cross-validation prediction. The Gaussian dispersion criterion is 0.00519 which indicates the degree of confidence (Table 2). With K-folds estimation we

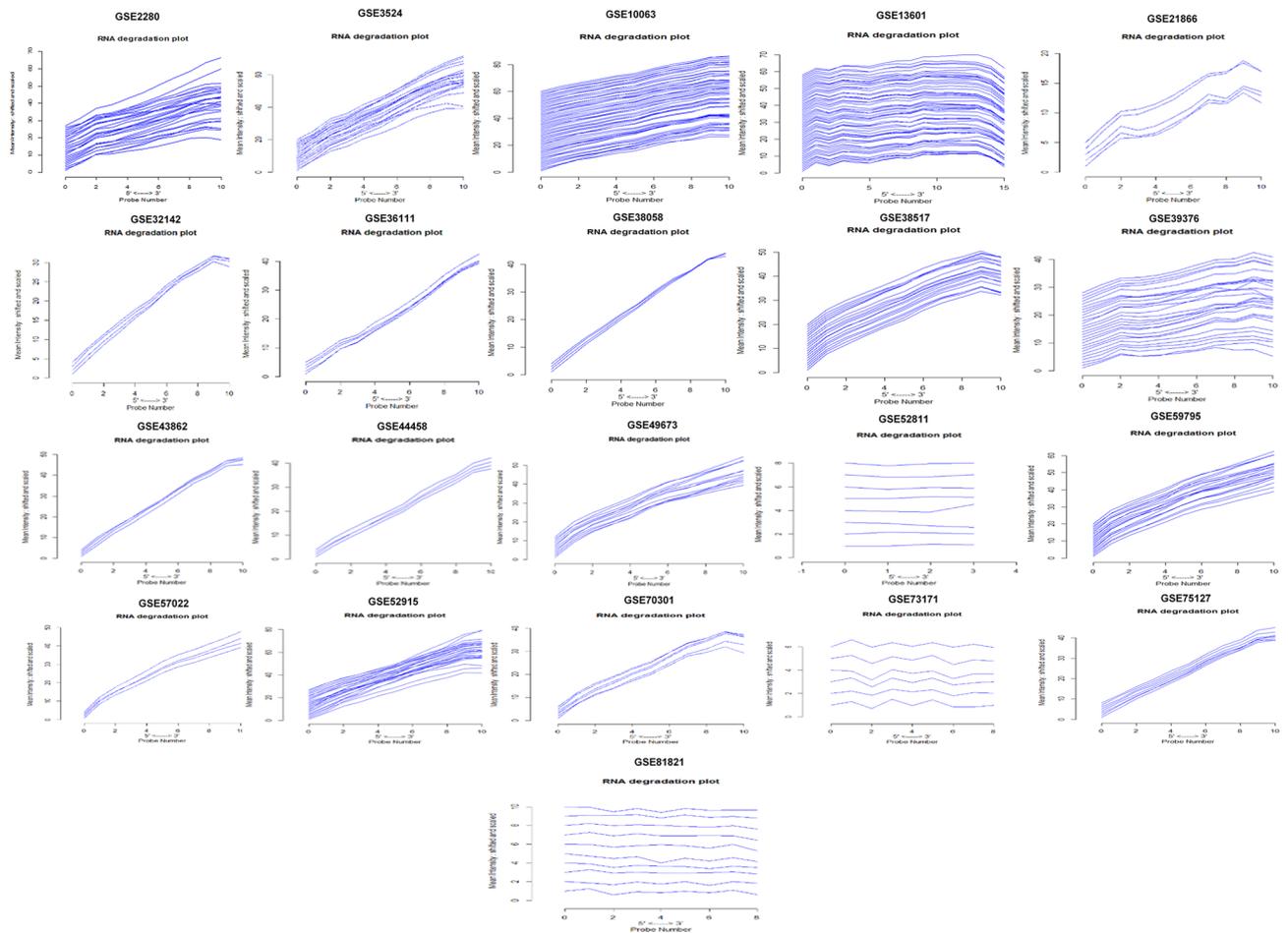


**Figure 1.** Normalization and differential analysis. Histogram (smoothed histograms) shows density estimate of the data. Typically, the distributions of the arrays have similar shapes and ranges. Arrays whose distributions are very different from the others are considered for possible problems. High levels of background shifted an array’s distribution to the right. Lack of signal diminishes its right tail. A bulge at the upper end of the intensity range indicated the signal saturation.

	Estimate	Std. error	t. value	Pr(> t )
(Intercept)	0.021148	0.003402	6.17	< 1.00E-14***
x1	0.052155	0.003825	22.009	< 1.00E-14***
x2	-0.02651	0.005117	-7.027	< 1.96E-09***
x3	0.149112	0.003105	30.007	< 1.00E-14***
x4	0.216532	0.001828	21.510	< 1.00E-14***
x5	0.048403	0.002152	31.003	< 1.00E-14***
x6	0.132542	0.002071	25.001	< 1.00E-14***
x7	-0.07733	0.001672	-29.216	< 1.00E-14***
x8	0.121504	0.002002	20.124	< 1.00E-14***
x9	0.212811	0.002825	65.278	< 1.00E-14***
x10	0.010029	0.003708	6.335	< 1.00E-14***
x11	0.022541	0.007523	22.601	< 1.00E-14***
x12	-0.02215	0.001205	-5.011	0.0089*
x13	-0.12532	0.003051	-56.055	< 1.00E-14***
x14	0.030526	0.001052	3.415	5.28E-12***
x15	-0.028691	0.001611	-20.502	< 1.00E-14***

**Table 2.** K-fold Cross-validation using “Boot” package of bioconductor software based on Gaussian dispersion modules. Deviance residuals: Min (-1.5101), 1Q (-0.0412), Median (-0.0100), 3Q (0.0132), Max (3.1977). Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1. Number of Fisher Scoring iterations: 2; \$K: [1] 10; \$delta: [1] 0.00516 = 0.00515. Null deviance: 100,813.5 on 53,225 degrees of freedom. Residual deviance: 2,817.3 on 53,209 degrees of freedom.

obtained the same delta value of 0.00515 as we used the LOOCV approach (during raw cross-validation and afterward during modified cross-validation). The substantial codes (0.1, 0.01, 0.001, and 0.05) with residuals of limited deviance suggested the consistency of the differential analysis. cDNA datasets were also analyzed for some necessary factors like RNA quality, sequence biases or RNA degradation. In genomic analysis, the use of low-quality RNA samples in the sequencing of the entire genome is inefficient. It is not clear if transcript deg-



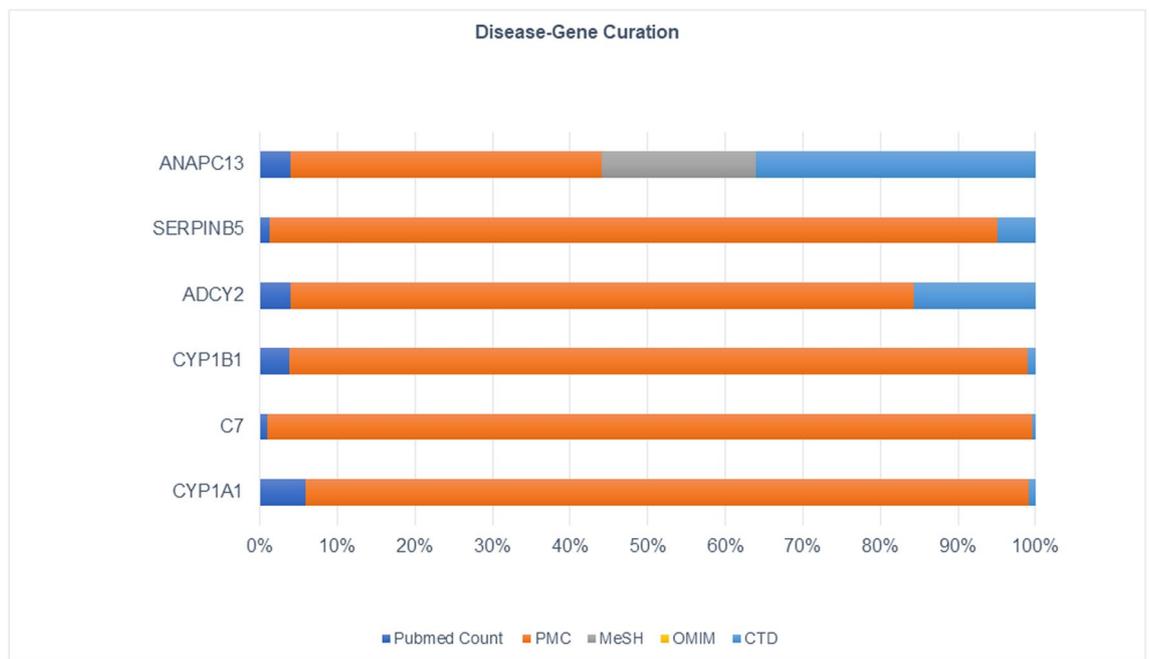
**Figure 2.** RNA degradation plot. Side-by-side plot produced by plot AffyRNAdeg representing 5'–3'-trend indicating an assessment of the severity of RNA degradation and significance level.

radation occurs reliably in low-quality RNA samples, in which case the effects of degradation can be reversed by data normalization or whether different RNA samples can be degraded at different rates, which could bias expression measurements. So, for differential expression analysis, we assessed the RNA quality. To verify the dataset reliability for identification of variation at the transcriptional level in original samples. The normalization process was used to standardize sample handling techniques and to assess optimal RNA variability threshold by using discrimination measures for statistical and algorithmic analysis. All the probe sets have their individual probes aligned at 5'-end of the target RNA molecule. The competitive binding of a particular probe to its target has been observed to depend upon a 3'/5' intensity gradient. Due to the poor quality of RNA, a reduced quantity of RNA is hybridized to the array. The low hybridization leads to a decrease in the total signal output level. But if the degree of saturation level increases the 3'/5' intensity gradient decreases. The 3'-end of the target gene contains a probe set that corresponds to the transcripts. The statistical and function summary for each batch-array is produced by 'AffyRNAdeg' to measure RNA degradation level and its significance (Fig. 2).

**Disease-gene curation for differentially expressed genes (DEGs).** From 21 datasets, we found 30 DEGs and David tool was used to retrieve their gene symbol and biological annotation. We selected the most significant ranked genes from the list of differentially expressed genes. For disease-gene curation, these genes were text mined using CTD (Comparative Toxicogenomics Database), PubMed, OMIM, MeSH, and PMC databases to filter disease-specific genes (Supplementary Table 1). We observed *CYP1B1*, *CYP1A1*, *C7*, *ADCY2*, *SERPINB5*, and *ANAPC13* are the most curated terms in the databases. These shortlisted genes were further analyzed by mapping at ( $p < 0.00005$ ) through Cancer Genetics and OMIM databases and observed their role in carcinogenesis (Fig. 3).

**Enrichment and cluster analysis of DEGs.** These genes showed enrichment substantially linked with hydroxylase, P450 pathway, steroid metabolic process, monooxygenase activity, cellular response to organic cyclic compound, and aromatase activity (Table 3). The dysregulation of these genes causes genetic heterogeneity, autosomal recessive disorder, head and neck disorders and other clinical phenotypes (Fig. 4).

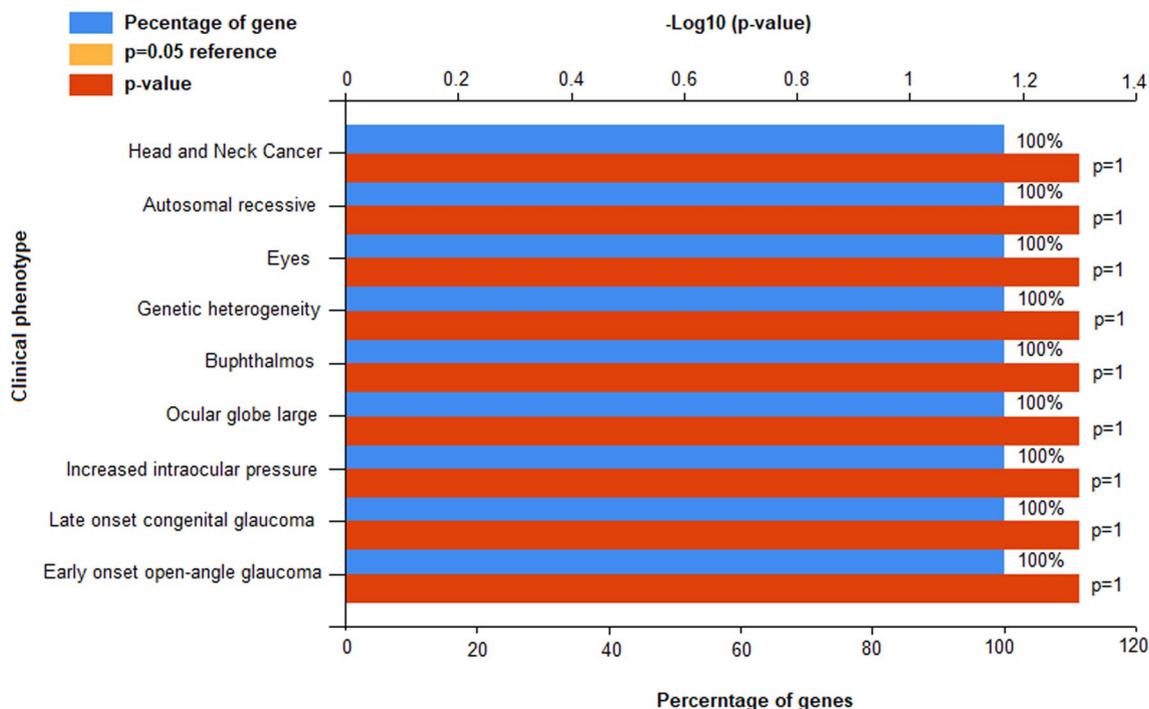
The function of the gene, its regulation, subtypes, and cellular processes play a key role in understanding its biology. The functional enrichment analysis showed that shortlisted *CYP1A1*, *CYP1B1* genes are known to



**Figure 3.** The OC-related DEGs were curated using CTD (comparative toxicogenomics database), PubMed, OMIM, and MeSH databases.

Term	<i>p</i> value	Fold enrichment	FDR
hsa04913: Ovarian steroidogenesis	4.86E-04	70.5102	0.464991
GO:0,097,267 ~ omega-hydroxylase P450 pathway	2.68E-03	621.9259	2.926305
GO:0,016,712 ~ oxidoreductase activity	0.00355	450.16	2.812419
GO:0,019,373 ~ epoxigenase P450 pathway	0.005349	310.963	5.768479
GO:0,070,330 ~ aromatase activity	0.006383	250.0889	5.007033
GO:0,019,825 ~ oxygen binding	0.011091	143.6681	8.558457
GO:0,008,202 ~ steroid metabolic process	0.01274	130.1705	13.24158
IPR002401: Cytochrome P450, E-class, group I	0.0134	123.7267	9.757773
GO:0,004,497 ~ monooxygenase activity	0.013674	116.4207	10.45655
IPR017972: Cytochrome P450, conserved site	0.01553	106.6609	11.23041
IPR001128: Cytochrome P450	0.017126	96.66146	12.31951
GO:0,071,407 ~ cellular response to organic cyclic compound	0.017447	94.87006	17.71563
Monooxygenase	0.021199	77.95833	17.98805
GO:0,031,090 ~ organelle membrane	0.023645	69.82375	15.90821
Metal ion-binding site: Iron (heme axial ligand)	0.024432	67.55219	18.69242
hsa00380: Tryptophan metabolism	0.028619	57.58333	24.30038
Microsome	0.030952	53.18088	25.24757
Heme	0.031663	51.97222	25.75335
GO:0,020,037 ~ heme binding	0.032072	49.28759	23.01009
Secondary metabolites biosynthesis, transport, and catabolism	0.032577	30.69697	3.257651
GO:0,005,506 ~ iron ion binding	0.035767	44.13333	25.3362
hsa00140: Steroid hormone biosynthesis	0.041281	39.71264	33.24928
hsa00980: Metabolism of xenobiotics by cytochrome P450	0.052426	31.12613	40.32831
hsa05204: Chemical carcinogenesis	0.056578	28.79167	42.78879
hsa04914: Progesterone-mediated oocyte maturation	0.061404	26.4751	45.53397
Glycoprotein	0.075408	3.014869	51.59749
Iron	0.076281	21.17387	52.01866
hsa04114: Oocyte meiosis	0.076443	21.1315	53.34849
hsa04114: Oocyte meiosis	0.076443	21.1315	53.34849

**Table 3.** Gene Ontology and functional enrichment of differentially expressed genes.



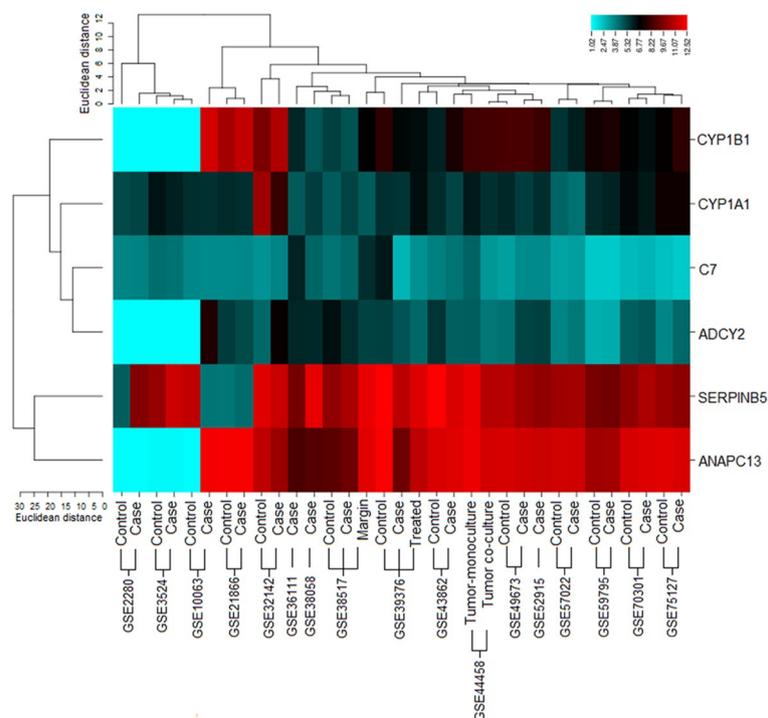
**Figure 4.** Clinical phenotypes for oral cancer related DEGs using FunRich annotation tool.

involved in xenobiotic metabolic and energy pathways. C7 is involved in immune response whereas SERPINB5 is known for protein metabolism. While ADCY2 plays an important role in cell communication and signal transduction. ANAPC13, have a potential role in class-I MHC-mediated antigens and cell-cycle progression at early tumor stages. The membrane attack complex in the extracellular region for C7, Expression of CYP1A1, SERPINB5, CYP1B1 is found in the endoplasmic reticulum. CYP1A1 is also found in nucleus and microsomes. ANAPC13 is a well-known anaphase-promoting complex. Differential expression of ADCY2, SERPINB5, CYP1B1 is also found in the cytoplasm.

Cluster analysis of selected DEGs helps us in the recognition of functional annotation and significance. The results were observed with the Euclidean distance (Fig. 5). The genetic expression of sample cells is distinguished as cases and control indicating the obvious differences between two of these groups. The analysis showed the down and up-regulated genes based on the *p* value and fold changes (Table 4).

**Mutation analysis.** ADCY2 has eight post-translational modification (PTM) sites with 234 recurrent cancer mutations at the chromosome no. 5 positive-strand encoding 1091 protein residues representing 14.02% of the predicted disordered region. The mutation visualization plot shows ADCY2 isoform ADCY2 Q474R direct network-rewiring mutation impact at the position 474, with reference amino acid residue Q and mutated amino acid residue R in the protein. The affected-site at position 472 with S amino acid residue-site enriched with a phosphorylation-type mutation affecting PTMs. Another, ADCY2 isoform ADCY2 S655R, reveals the mutation for this protein at position 655 including amino acid residue S comparison with the mutated amino acid residue R. In position 659, S-amino acid residue site enriched with phosphorylation-type mutation, this shows distal-mutation PTM impact with the affected-site. CYP1B1 showed a 10.5% predicted disordered sequence region with 100 mutations observed at chromosome no 2 on the negative strand with 543 protein residues and 3 PTM sites. CYP1B1 I87S, CYP1B1 Q479H, CYP1B1 T510I isoforms were revealed for CYP1B1 mutational enrichment, at the positions 87,479,510 respectively. The reference amino acid residue for these isoforms was I, Q, T along with mutant amino acid residues S, H, I respectively for each isoform. Similarly, the mutational analysis of C7 showed that 12.57% of the sequence predicted for disease-pathophysiology. Total 244 number of mutations were found on the positive strand of chromosome no. 5 for C7. The number of PTM sites for C7 were eight with 843 protein residues. So, 10 isoforms for C7 were found, among them C7 Q29R and C7 G41D were in distal-mutational PTM impact (Fig. 6). The reference amino acid residues for C7 Q29R, C7 G41D, and C7 T756I were Q, G, T, and mutant amino acid residues reported were I, R, and D, respectively (Supplementary Table 2).

**Protein–protein interaction analysis.** We retrieved the related nodes and edges of all oral cancer associated DEGs from the HAPPI database to construct the integrated PPI-network (Fig. 7). This interaction analysis helped us to observe the potential functional interaction among OC related DEGs and other associated genes contributing to the disease phenotype. The seeder or source OC associated DEGs CYP1A1\_HUMAN, CYP1B1\_HUMAN, ADCY2\_HUMAN, C7\_HUMAN, SERPINB5\_HUMAN, ANAPC13\_HUMAN interact with the target genes including BRAC1\_HUMAN, CO6\_HUMAN, ISG15\_HUMAN, S1PR3\_HUMAN, and other essential proteins. The network topology shows a significant relationship between seeder and target genes. The identi-



**Figure 5.** Cluster analysis of 6 oral cancer-related DEGs with Euclidean distance (Binning method). Quantile lines indicate the boundaries of the clusters in the level of the tree.

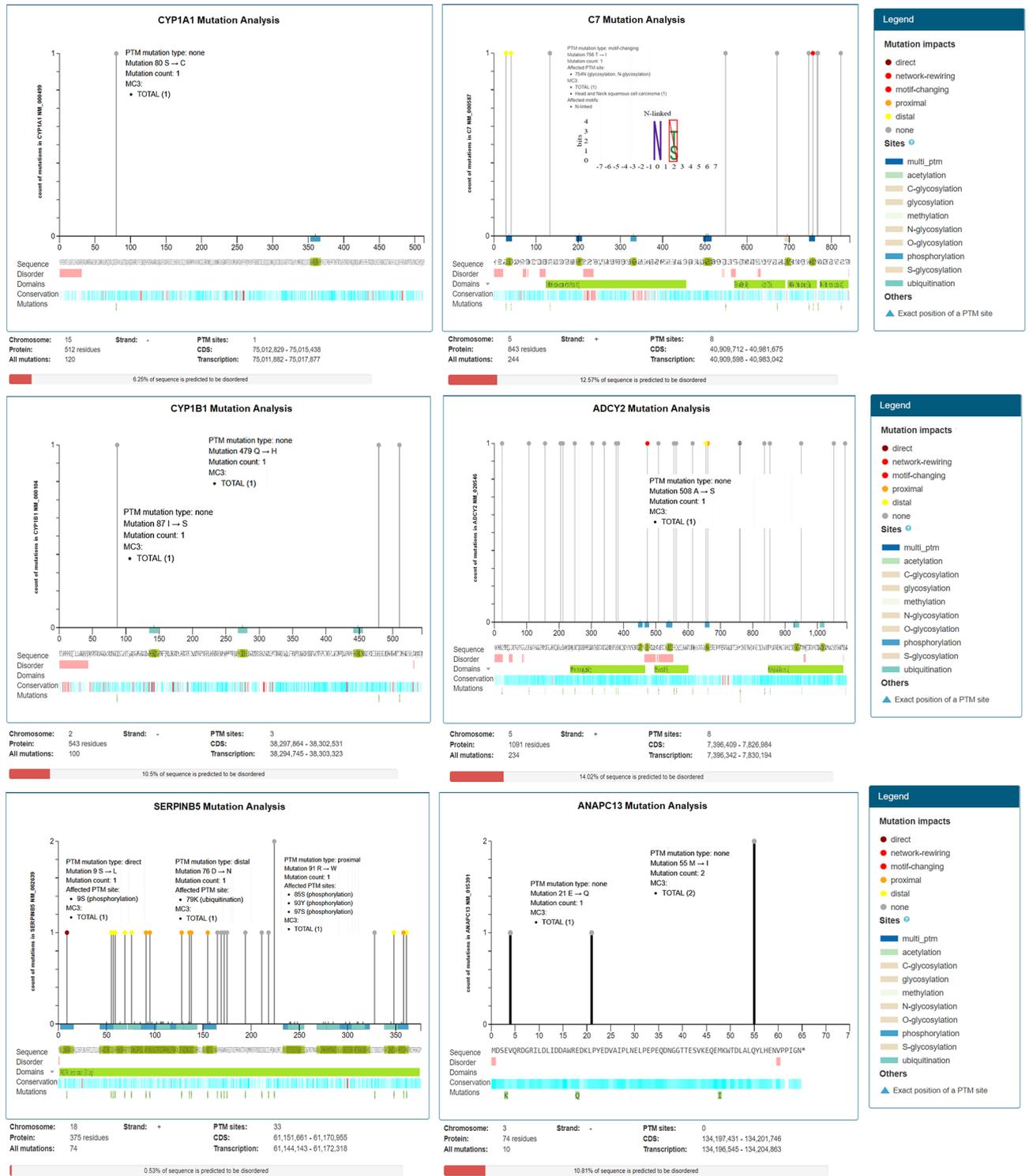
AFFYMETRIX_3PRIME_IVT_ID	Gene name	logFC	AveExpr	t	p value	adj. p-Val	B	Abberation
205749_at	<i>CYP1A1</i>	3.900696	8.398162	39.22931	2.99E-46	1.63E-41	87.06766	Up regulated
202992_at	<i>C7</i>	-1.84918	4.492108	-6.64914	3.77E-07	0.008403	4.18572	Down regulated
202437_s_at	<i>CYP1B1</i>	3.140615	8.48289	25.87402	1.89E-35	5.18E-31	67.29499	Up regulated
213217_at	<i>ADCY2</i>	-2.07208	5.693218	-7.54109	3.51E-07	0.019192	4.699974	Down regulated
204855_at	<i>SERPINB5</i>	-6.46131	9.39441	-41.2474	2.06E-26	9.05E-23	49.19926	Down regulated
209001_s_at	<i>ANAPC13</i>	-2.49883	10.52767	-31.6716	1.32E-10	7.23E-06	13.02867	Down regulated

**Table 4.** Expression profiling of cDNA microarray datasets.

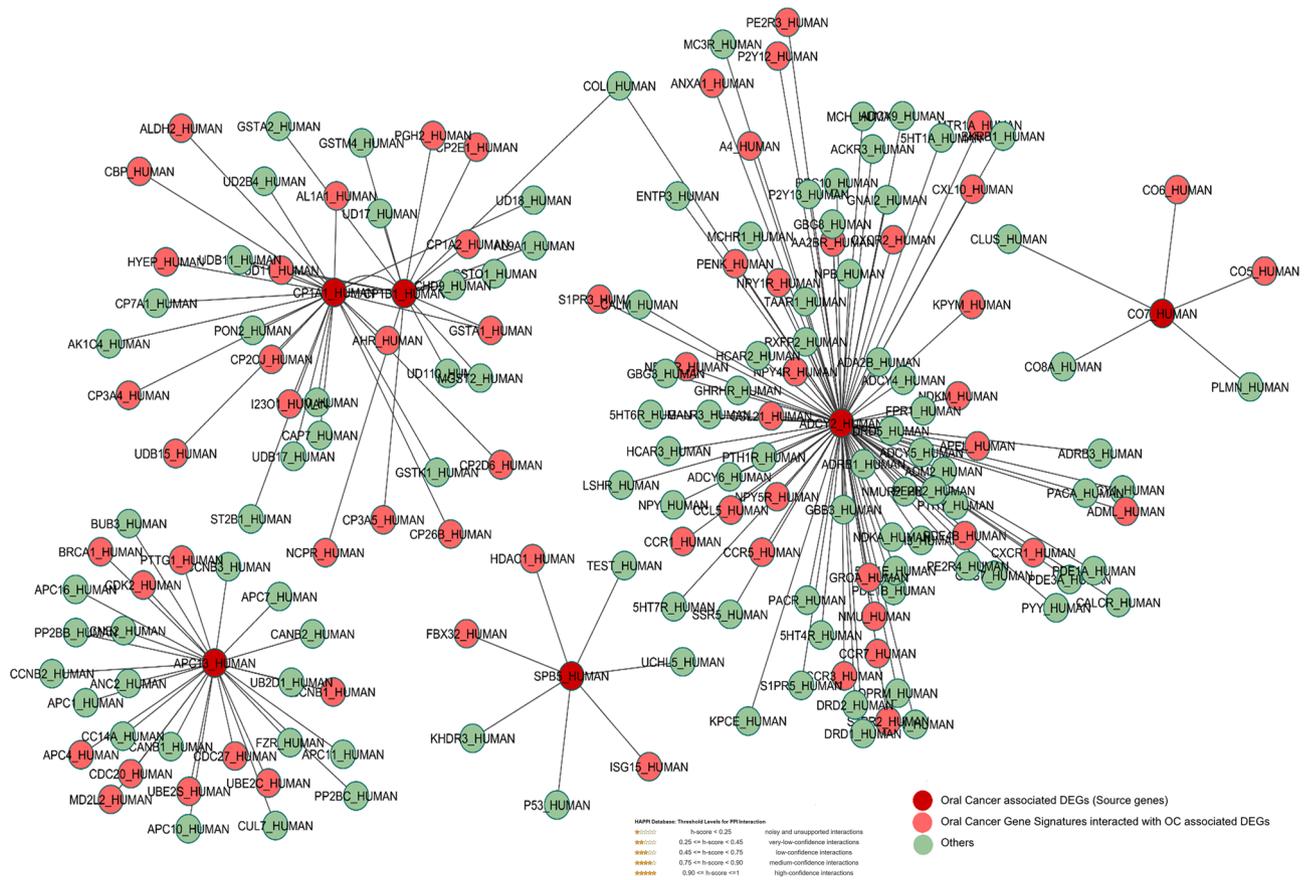
fied proteins showed a significant association with disease development. The target proteins GST2\_HUMAN and GSTK1\_HUMAN play important role in phase-II carcinogen metabolism and interacted with source genes CYP1A1\_HUMAN and CYP1B1\_HUMAN. The target proteins including CO7\_HUMAN, CLUS\_HUMAN, CO6\_HUMAN and CO5\_HUMAN involved in laryngeal carcinoma, oral cavity squamous cell carcinoma and oropharyngeal carcinoma, respectively. Oral cancer ADCY2\_HUMAN linked with KNG1\_HUMAN as potential biomarkers. SPB5\_HUMAN is interacted with P53\_HUMAN, which is a potential tumor suppressor gene. We have observed that OC related genes APC13\_HUMAN is interacting with CDK2\_HUMAN, a prognostic indicator of oral cancer.

**Pathway modelling of OC associated DEGs.** Integrated pathways were modeled to observe the possible role of DEGs in pathophysiological mechanisms. Ras, p53, MEK, SOS, Rb, Bax, PTEN, and Raf are important interacting genes associated with the pathophysiological mechanism of oral cancer. We found that p53 signaling, microRNA signaling, salivary secretion, human papillomavirus, cell cycle, alcoholism, and xenobiotic metabolism-related pathways are interconnected in the progression and development of the disease (Fig. 8).

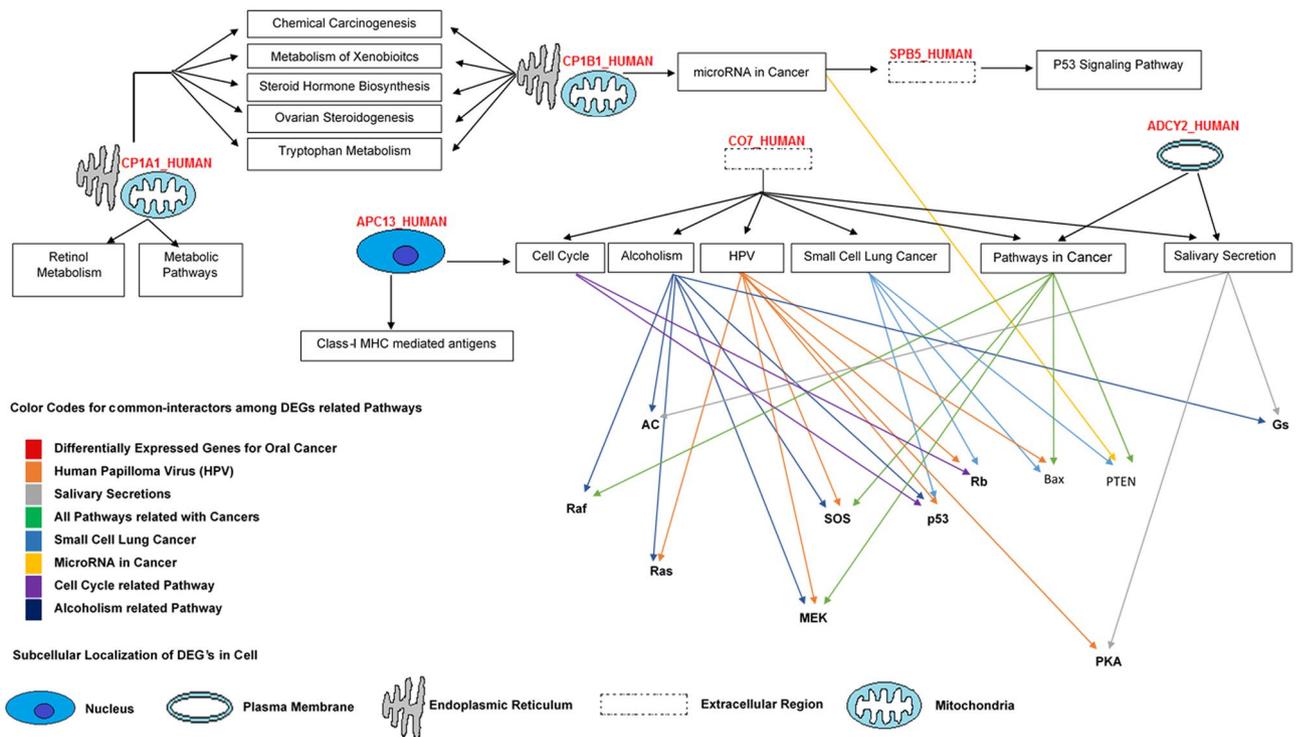
**Toxicogenomic analysis.** The toxicogenomic analysis enabled us to explore chemical genotype-phenotype exposomic information that may lead to disease progression. OC associated DEGs were curated in terms of their activity and expression with different environmental chemicals. The data revealed the activity and expression of DEGs, which either increase or decrease the expression, increase or decrease towards gene activity at different cellular events. It may also affect the cotreatment expression leading to disease occurrence. It was revealed that the same chemical exposure may show different reactivity for different genes. In this case, benzo(a)pyrene increases the expression of CYP1A1 but it affects the reactivity of ADCY2. Methylcholanthrene, albenazole,



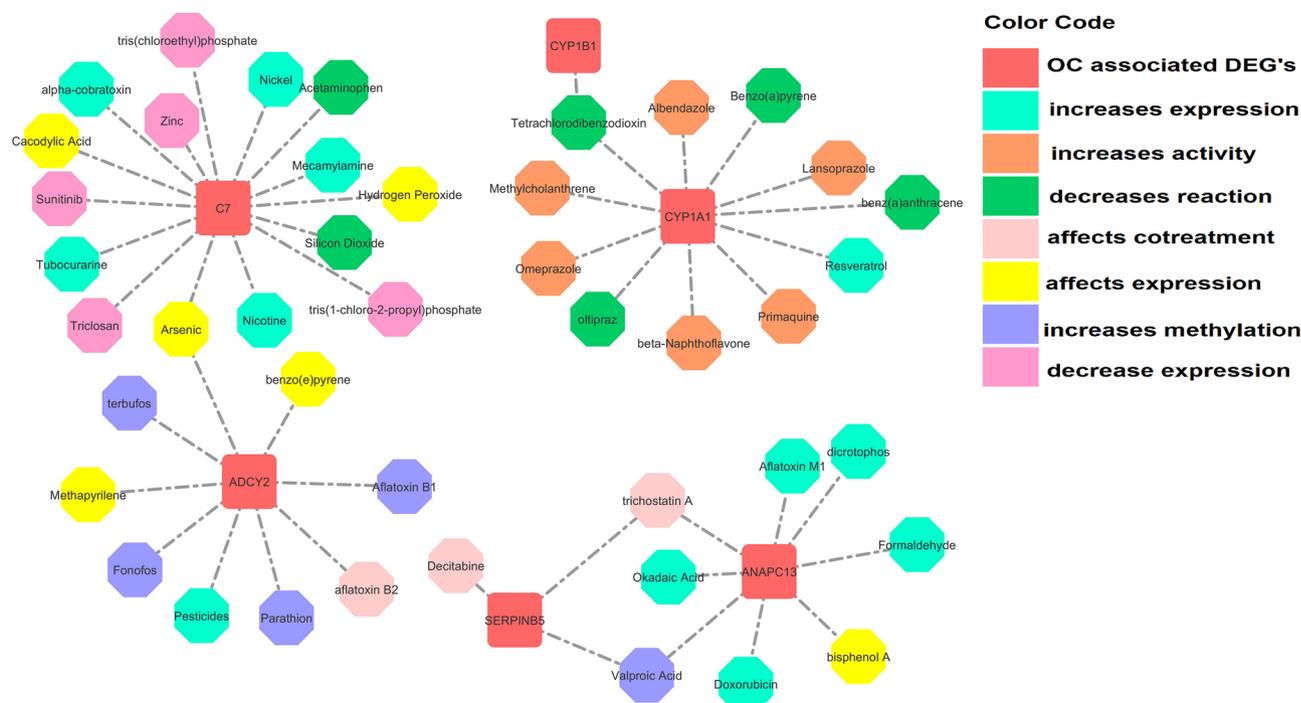
**Figure 6.** Analysis and exploration of mutations affecting post-translational modification (PTM) sites in human genes/proteins using online ActiveDriverDB database. Needle plots indicate the PTM site mutations in our genes/proteins. The graph shows the outcomes based on the specific type of PTM, cancer type, and mutation subset (presented in legend color codes). Height (y-axis) represents the number of occurrences of the mutation while Horizontal position (x-axis) indicates the position of protein's amino acid sequence. Pinhead color signifies the mutation impact and X-axis coloring shows the type of PTM associated with the mutation location. Mutational Impacts: Rewiring: mutation-induced gains and losses of kinase-bound sequence motifs (predicted by MIMP software); Distal: mutation affects an amino acid located 4–7 amino acids away from the PTM site; Direct: mutation affects the post transcriptionally modified amino acid; Proximal: mutation affects an amino acid located 1–3 amino acids away from PTM site; Sites: Amino acid sites/ regions enriched for mutations affecting post-translational modifications (PTMs).



**Figure 7.** Protein–protein interaction of OC genes. Red nodes represent DEGs interacting with Pink nodes (target genes/gene signatures). High-confidence interactions of HAPPI database were selected in this network (the five stars are equivalent to high score (0.90–1).



**Figure 8.** Pathway modeling for genome signaling and metabolic reconstruction revealed the pathological mechanism of oral cancer using KEGG and Wiki pathway databases.



**Figure 9.** Toxicogenomic analysis of differentially expressed genes carried out by a comparative toxicogenomic database (CTD) helps to study the chemical-genome to phenome relationships.

primaquine increases CYP1A1 activity. Resveratrol and tetrachlorodibenzodioxin were found to affect the binding, decrease the reaction, and increase the CYP1A1 and CYP1B1 activity. Similarly, acetaminophen was found to increase the expression of C7, whereas alpha-cobra toxin may account for decrease reaction. Arsenic may affect C7 expression by increasing its abundance. While nickel was found to decrease C7 expression. However, decitabine was found to affect the cotreatment of SERPINB5 and decrease the gene reactivity with trichostatin A. It was observed that fonofos, methapyrilene, parathion increases and affects the ADCY2 reactivity at different cellular events while valproic acid, okadaic acid, doxorubicin, and bisphenol A decreases the ANAPC13 expression (Fig. 9).

**De novo prediction of regulatory motifs.** Oral cancer-associated DEGs were used for de novo analysis to predict the regulatory motifs. The transcriptional factors include ARNT, AHR, CEBPA, CTCF, HNF1B, ELK4, TCF3, and NR2E3. The conservation cutoff is 0.40 with a matrix-score threshold of 85% were set as default parameters. The parameter settings standardize to analyze oPOSSUM-tool showed how the transcriptional factor controls its related targets (Table 5).

**MiRNAs targets for oral cancer-specific genes.** miRDB database helped us to predict the miRNA targets based on the algorithms. We found the reliable score (> 80) of oral cancer-specific miRNAs. Mainly miRNAs including hsa-miR-4786-5p, hsa-miR-4282, hsa-miR-2052, hsa-miR-216a-3p, hsa-miR-3148, and hsa-let-7f-1-3p targets were predicted for CYP1A1, CYP1B1, C7, ADCY2, SERPINB5 and ANAPC13 genes respectively. The functional dysregulation of these genes may lead to disease progression. The predicted target scores, total miRNA hits, seed-binding locations, and 3'-UTR length of DEGs were analyzed (Table 6).

**Drug-gene interaction analysis.** The toxicogenomic approach was used to investigate the drug-genes interaction to explore available treatments. The genes that interact with anticancer drugs are docetaxel, hydroxyurea, bleomycin, daunorubicin, lansoprazole, doxorubicin, liothyronine sodium, risperidone using DGIdb database. We identified sixteen proteins as potential alternative drug targets including CXCL1, FBXO32, PTTG1, CCNB1, ADCY2, NMU, ANAPC13 and others (Fig. 10). The dysregulation of these proteins may affect the normal expression level and could be a potential part of therapeutic strategies.

## Discussion

This study focused on the genetic expression and functional enrichment of genetic variants of oral cancer. The six most significantly OC associated genes (*CYP1A1*, *CYP1B1*, *SERPINB5*, *ANAPC13*, *ADCY2*, *C7*) found through a differential analysis were considered as seeder or source genes. This analysis provides us a list of new genes aberrantly expressed in oral cancer including *SERPINB5*, *ANAPC13*, *ADCY2*. We have investigated differential expression between cases and controls of cDNA datasets at cellular level in oral tissues and found the possible association of these genes in oral cancer. We can get more information about the mechanism of human genetic

TF	Class	Family	Tax group	IC	Target gene hits	Target gene non-hits	Back-ground gene hits	Back-ground gene non-hits	Target TFBS hits	Target TFBS nucleotide rate	Back-ground TFBS hits	Back-ground TFBS nucleotide rate	Z-score	Fisher score
ESR1	Zinc-coordinating	Hormone-nuclear Receptor	Vertebrates	13.563	1	5	345	24,407	1	0.00319	355	0.000197	16.419	2.514
Arnt::Ahr	Zipper-Type	Helix-Loop-Helix	Vertebrates	9.532	5	1	13,597	11,155	23	0.022	66,086	0.011	8.273	1.815
CEBPA	Zipper-Type	Leucine Zipper	Vertebrates	8.712	5	1	12,828	11,924	18	0.0258	55,606	0.0139	8.01	2.06
CTCF	Zinc-coordinating	Beta Alpha-zinc finger	Vertebrates	17.205	2	4	3,236	21,516	2	0.00606	3,982	0.0021	6.703	1.718
HNF1B	Helix-Turn-Helix	Homeo	Vertebrates	16.821	3	3	4,107	20,645	3	0.00574	6,910	0.0023	5.543	2.787
ELK4	Winged Helix-Turn-Helix	Ets	Vertebrates	14.123	2	4	6,034	18,718	4	0.00574	9,342	0.00234	5.452	0.796
TAL1::TCF3	Zipper-Type	Helix-Loop-Helix	Vertebrates	14.07	2	4	7,491	17,261	5	0.00957	14,681	0.0049	5.21	0.535
Nr2e3	Zinc-coordinating	Hormone-nuclear Receptor	Vertebrates	12.028	4	2	6,354	18,398	5	0.00558	12,287	0.00239	5.045	3.187
Lhx3	Helix-Turn-Helix	Homeo	Vertebrates	16.354	2	4	5,806	18,946	4	0.00829	12,758	0.00461	4.212	0.846
Nkx2-5	Helix-Turn-Helix	Homeo	Vertebrates	8.27	6	0	16,973	7,779	43	0.048	197,210	0.0384	3.939	2.263
Arnt	Zipper-Type	Helix-Loop-Helix	Vertebrates	10.992	4	2	7,052	17,700	5	0.00478	13,904	0.00232	3.928	2.827
NFE2L2	Zipper-Type	Leucine Zipper	Vertebrates	14.394	2	4	5,635	19,117	3	0.00526	8,970	0.00274	3.696	0.886
ESR2	Zinc-coordinating	Hormone-nuclear Receptor	Vertebrates	13.618	1	5	2,219	22,533	1	0.00287	2,670	0.00134	3.155	0.842
Tal1::Gata1	Zipper-Type	Helix-Loop-Helix	Vertebrates	11.297	2	4	6,212	18,540	3	0.00861	11,200	0.0056	3.107	0.758
Evi1	Zinc-coordinating	Beta-BetaAlpha-zinc finger	Vertebrates	17.909	1	5	1931	22,821	1	0.00223	2,457	0.000956	3.067	0.952
PLAG1	Zinc-coordinating	Beta-BetaAlpha-zinc finger	Vertebrates	19.352	1	5	1971	22,781	1	0.00223	2,462	0.000958	3.059	0.936
MZF1_5-13	Zinc-coordinating	Beta-BetaAlpha-zinc finger	Vertebrates	9.4	5	1	13,425	11,327	17	0.0271	77,649	0.0216	2.97	1.868
STAT1	Ig-fold	Stat	Vertebrates	13.119	2	4	6,394	18,358	3	0.00717	10,990	0.00458	2.949	0.722
Mycn	Zipper-Type	Helix-Loop-Helix	Vertebrates	11.104	3	3	8,332	16,420	5	0.00797	18,928	0.00526	2.883	1.12
Prrx2	Helix-Turn-Helix	Homeo	Vertebrates	9.063	4	2	15,063	9,689	27	0.0215	123,270	0.0171	2.636	0.576
FOXO1	Winged Helix-Turn-Helix	Forkhead	Vertebrates	11.926	6	0	13,087	11,665	14	0.0179	62,921	0.014	2.556	3.823
MEF2A	Other Alpha-Helix	MADS	Vertebrates	15.709	2	4	7,006	17,746	4	0.00638	15,631	0.00434	2.354	0.612

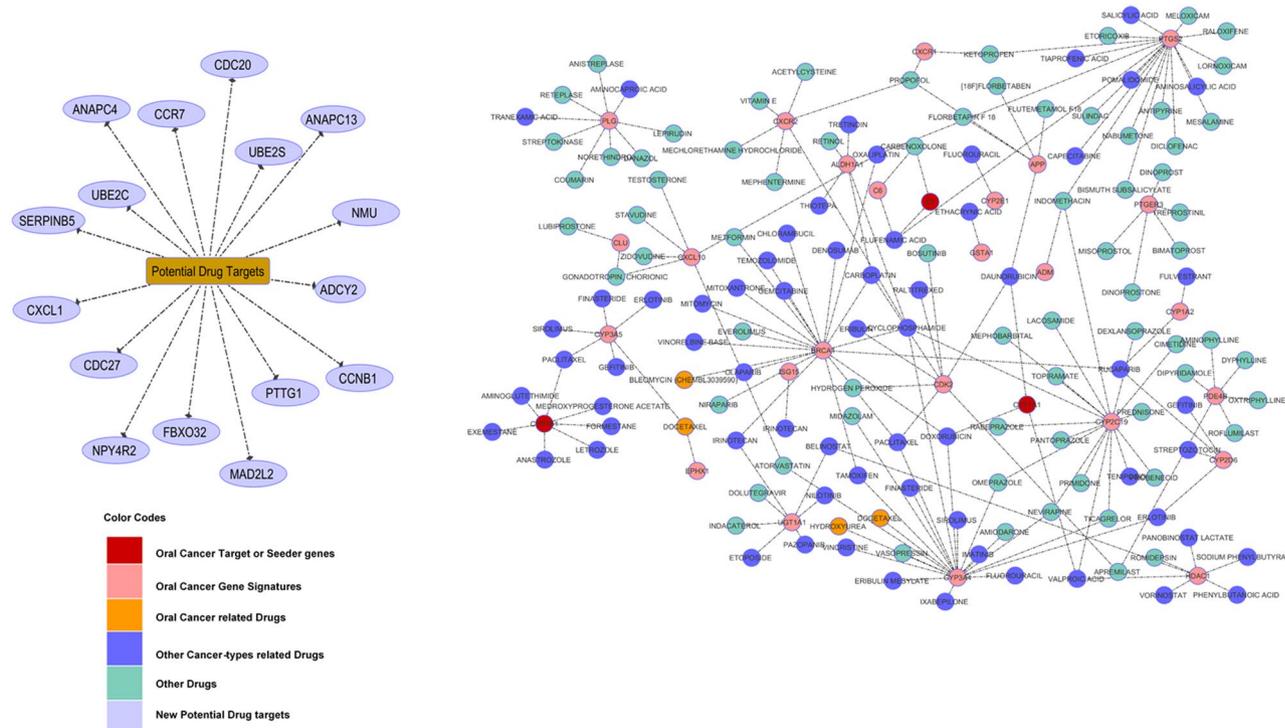
Continued

TF	Class	Family	Tax group	IC	Target gene hits	Target gene non-hits	Back-ground gene hits	Back-ground gene non-hits	Target TFBS hits	Target TFBS nucleotide rate	Back-ground TFBS hits	Back-ground TFBS nucleotide rate	Z-score	Fisher score
Gfi	Zinc-coordinating	Beta-BetaAl-pha-zinc finger	Vertebrates	9.47	4	2	13,731	11,021	14	0.0223	65,844	0.0183	2.332	0.796
FOXO3	Winged Helix-Turn-Helix	Forkhead	Vertebrates	11.734	5	1	13,611	11,141	15	0.0191	69,410	0.0154	2.328	1.811

**Table 5.** Over-representation of oral cancer-related DEGs using oPOSSUM with 80% matrix score.

Serial no	Gene symbol	Gene description	*Target score	microRNA name	Total hits	miRNA sequence	Seed location	3'-UTR length
1	CYP1A1	cytochrome P450 family 1 subfamily A member 1	90	hsa-miR-4786-5p	35	UGAGACCAGGAC UGGAUGCACC	735	966
2	CYP1B1	cytochrome P450 family 1 subfamily B member 1	98	hsa-miR-4282	141	UAAAAUUUGCAUCCA GGA	19, 704, 1,092, 1,235, 1,246, 1753, 2028, 2,137	3,125
3	C7	complement C7	97	hsa-miR-2052	84	UGUUUUGAUAAAC AGUAAUGU	1,197, 1,203, 1,311	3,073
4	ADCY2	adenylate cyclase 2	96	hsa-miR-216a-3p	142	UCACAGUGGUCU CUGGGAUUU	190, 1575, 1641, 2,853	3,210
5	SERPINB5	serpin family B member 5	98	hsa-miR-3148	64	UGGAAAAAACUG GUGUGUGCUU	169, 196, 1,323	1,363
6	ANAPC13	anaphase promoting complex subunit 13	89	hsa-let-7f-1-3p	61	CUAUACAAUCUAUUG CCUUCCC	699	903

**Table 6.** Prediction of gene specific MiRNA-targets associated with oral cancer. \*Highly reliable score—> 80, least reliable score—< 50.



**Figure 10.** The drug-gene network was constructed between the FDA approved drugs and target genes. A broken line indicates the interaction between known drugs while solid line represents the novel drug targets. Anticancer drugs were retrieved from drug B+ ank.

disorder through microarray studies. The expression profiling of these important genes shows obvious differences between cases and controls. Some DEGs were found upregulated while other downregulated in this analysis. These genes are abnormally expressed to affect physiological functions including cellular signaling, replication, mitotic division, and programmed cell death. We have observed that our differentially expressed genes are associated with the cancer pathways including biological oxidations, metabolism, adenylate cyclase-activating pathway, xenobiotics, G alpha signaling events, transcriptional targets of TAp63 isoforms, p53, and IFN-gamma pathway revealed the biological significance of these genes specifically for oral cancer progression. CDC20\_HUMAN, HDAC1\_HUMAN, CXL10\_HUMAN interacting with source genes are potential drug targets<sup>13–15</sup>. The inherent mutations are reported for genes that encode drug-metabolizing enzymes. Such somatic gene mutations are induced chemically that play a vital role in cell differentiation and growth<sup>16</sup>. The sequencing investigations not only characterizes the genomics but also revealed thousands of SNVs (single nucleotide variants), the alterations in copy number along with many types of genetic variations. Such genomic to phenomic association identification, their molecular-level mechanisms, disease-related variants along with cancer-derived mutations are the current challenges in the biomedical research<sup>17,18</sup>. Deciphering inter-individual genetic variation is the latest trend in personal genomic era investigations. The interpretation of genomic to proteomic information may integrate the impact of mutations on cellular system-level investigations in the future with a higher magnitude<sup>19</sup>. Human proteomic data analysis uses PTMs to interpret disease mutations and inter-individual genetic variations. PTMs being important regulators of protein function and signaling pathways facilitate the missense mutational analysis investigations<sup>19</sup>.

We have observed that *CYP1A1* belongs to potential and well-preserved phase-I xenobiotic metabolizing gene family which is involved in the activation of procarcinogens. The *CYP1A1* enzyme is highly associated with increased risk of tumors in the oral cavity, bronchial and laryngeal regions in smokers<sup>4</sup>. Similarly, the association of *CYP1B1* has found in many cancer types<sup>20</sup>. Many investigations are reported about the substantial link of HNC with *CYP1A1* and *CYP1B1*<sup>16,21</sup>.

*SERPINB5* belongs to serpin encoding serine protease which plays a vital role in tumor metastasis<sup>22,23</sup>. It is a tumor suppressor in epithelial cells and can suppress cancerous cell invasion and their metastasizing in surrounding tissues<sup>24</sup>. The paradoxical expression of *SERPINB5* has been observed in various types of tumor<sup>25–28</sup>. We have seen a highly significant association between *SERPINB5* expression and oral carcinoma<sup>29</sup>. *ADCY2* is a membrane-associated enzyme which converts adenosine-5'-triphosphate (ATP) into 3',5'-adenosine monophosphate (cyclic AMP/cAMP) and pyrophosphate<sup>30</sup> involving in the regulation of cAMP synthesis<sup>31</sup>. This gene catalyzes the signaling molecule cAMP through G-protein beta as well as gamma subunit signaling<sup>32,33</sup>. Therefore, changes in expression patterns of the gene are mediated through down-streaming of the signaling cascades muscarinic acetylcholine receptors which increases IL6 production. The high regulation of the gene is observed in G-proteins, calcium, calmodulin, pyrophosphate, and post-translational modifications. The signaling pathways include RET signaling, Oocyte meiosis, calcium, and chemokine signaling pathways<sup>34–36</sup>. Aberrant methylation of *ADCY2* is observed in colorectal, prostate cancer<sup>35–37</sup> and urinary bladder cancer<sup>38–40</sup>. It has been studied that *ANAPC13*<sup>41</sup> is a large-sized ubiquitin ligase that controls the cell cycle progression<sup>42</sup> and involved at early steps of malignancy in tumor cells<sup>43</sup>. Similarly, *C7* (complement component-7), the terminal component for complement cascade and as a cytolytic effector for complement system, lyses transformed malignant cells<sup>44–51</sup>. The integration of chemical- gene interaction revealed different environmental chemical exposure to disease progression<sup>52,53</sup>. This analysis helped to reveal the mechanism of action between the chemical and the related gene products and their effect on human disease influence by environmental exposure<sup>53–56</sup>.

The PPI network predicted the important association of these genes with disease. These genes have a potential role in xenobiotic metabolism, tumor progression, suppression, cell cycle, HPV (human papilloma virus), alcoholism, and microRNA signaling pathway<sup>30,40,48,49,51,57–60</sup>. The transcriptomic analysis showed expressive transcription factors like JUND, FOXO, STAT1. We found the role of these genes in metabolism of xenobiotics, p53 signaling, salivary secretion, class-I MHC mediated antigens and microRNA cancer pathways. The miRNAs regulate post-transcriptional and translational events and expressional dysregulation in these molecules leads towards the progression of many diseases<sup>61–63</sup>. Therefore, the reliable miRNA target prediction is crucial for the functional annotation of miRNAs<sup>64,65</sup>.

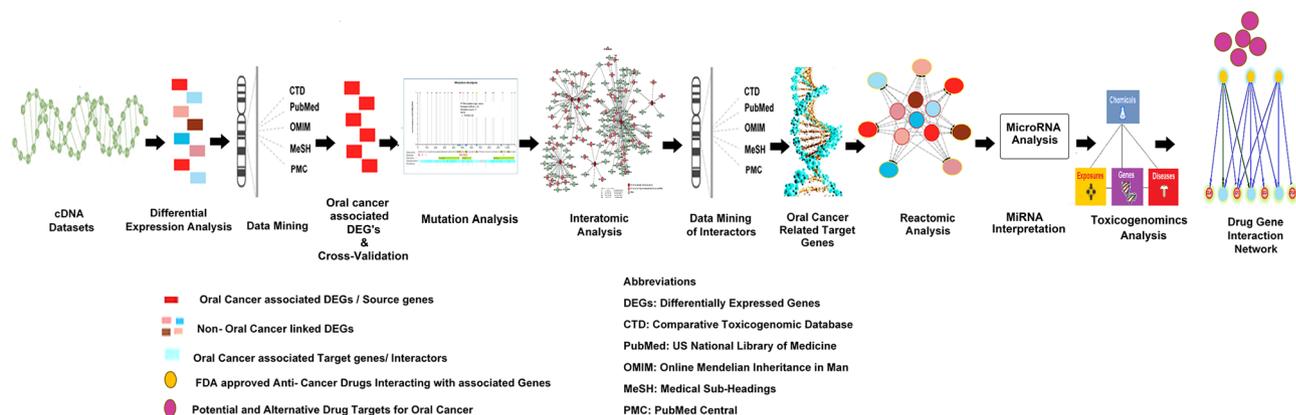
Recent reports proved that the drug-gene network enables us to understand not only the disease pathophysiology but also important in drug designing or new drug target identification or establishing novel biomedical linkages. More importantly, this network proposed many testable assumptions with the potential of great success, though the real achievement can only be justified by experimental studies.

## Conclusion

Our simulation-based systems-level hypothesis is comprehensive and effective to sort out the disease-specific genetic variants from cDNA datasets repositories. Therefore, this approach will support understanding the genetic basis of complex phenotypes including cellular replication, protein signaling, mitotic division, and programmed cell death. Based on genomic to phenomic investigations, we have found new genes including *ADCY2*, *SERPINB5*, and *ANAPC13* linked with oral cancer that could be potential diagnostic or drug targets. These source genes are clearly interacting with other essential genes affecting cell cycle and apoptosis causing carcinogenesis. These findings can provide a valuable framework for developing new therapeutic strategies against oral cancer.

## Methods

**Accession to cDNA datasets.** 5.2 We downloaded cDNA datasets related to oral cancer from the Gene Expression Omnibus database (GEO) NCBI. The comprehensive framework has been illustrated in Fig. 11 using tools, online servers, and software (Table 7).



**Figure 11.** The steps have been integrated in basic framework of our study.

Databases/software/tools	Accessibility	Utility	References
STRING database version 11	<a href="https://string-db.org/">https://string-db.org/</a>	For known and predicted protein/COGs interaction	74
National Center for Biotechnology Information (NCBI)	<a href="https://ncbi.nlm.nih.gov">https://ncbi.nlm.nih.gov</a>	Biomedical and genomic information source	–
Cytoscape version 3.6.0	<a href="https://www.cytoscape.org/">https://www.cytoscape.org/</a>	For network analysis and visualization	31
DAVID Bioinformatics tool 6.8	<a href="https://david.abcc.ncifcrf.gov">https://david.abcc.ncifcrf.gov</a>	Gene ontology/ Functional Annotation tool	75
Uniprot	<a href="https://www.uniprot.org/">https://www.uniprot.org/</a>	Resource of protein functional information	76
Kyoto Encyclopedia of Genes and Genomes (KEGG)	<a href="https://www.genome.jp/">https://www.genome.jp/</a>	Pathways analysis and comparison	77
FunRich version 3	<a href="http://www.funrich.org/">www.funrich.org/</a>	Enrichment analysis	24
R version 3.3.3	<a href="https://www.r-project.org/">https://www.r-project.org/</a>	Statistical computing/data mining	14–16
Opossum Version 3.0	<a href="https://opossum.cisreg.ca/oPOSSUM3/">https://opossum.cisreg.ca/oPOSSUM3/</a>	Single site analysis	78
Wiki-Pathways	<a href="https://www.wikipathways.org/index.php/WikiPathways">https://www.wikipathways.org/index.php/WikiPathways</a>	Pathways analysis	79
Path-Visio 3.3.0	<a href="https://www.pathvisio.org/">https://www.pathvisio.org/</a>	pathway analysis and drawing software	34
Comparative Toxicogenomics Database CTD	<a href="https://ctdbase.org/">https://ctdbase.org/</a>	gene–disease relationships	80
CIMminer	<a href="https://discover.nci.nih.gov/cimminer/home.do">https://discover.nci.nih.gov/cimminer/home.do</a>	Cluster analysis	25–26
HAPPI version 2,0	<a href="https://discovery.informatics.uab.edu/HAPPI/">https://discovery.informatics.uab.edu/HAPPI/</a>	Protein–protein interaction	30

**Table 7.** Databases and tools used during this meta-analysis.

**Normalization and differential analysis.** These datasets were analyzed in identifiable format to easily access pheno-data files and missing values were imputed<sup>66</sup>. R software version 3.3.3 and Bioconductor packages were used in computational analysis. The normalization and quality control analysis was performed to preprocess the information available by ArrayQuality<sup>67–69</sup>. The background and normalization were aligned by using Robust Multi-Array Analysis (RMA) to detect the PM (perfect matches) and the MM (mismatches) to impute the values for statistical analysis<sup>70</sup>. RMA is the widely used preprocessing algorithm applied for background correction to remove local artifacts<sup>67</sup>.

$$PM_{ijk} = BG_{ijk} + S_{ijk}$$

where  $PM$  indicates a perfect match, Background by  $BG$  and non-specific binding ( $S$ );  $ijk$  is the signal for probe  $j$  of probe set  $k$  on array  $i$ .

$$BG(PM_{ijk}) = E[S_{ijk}|PM_{ijk}] > 0$$

$$S_{ijk} \sim \text{Exp}(\lambda_{ijk}) \quad BG_{ijk} \sim N(\beta_i, \sigma^2).$$

The perfect match involves the combined signals of background ( $BG$ ) and expression ( $E$ ). The "ArrayQualityMetrics" software was used to evaluate the quality of dataset that is normalized to each gene's median expression level<sup>67,68,71</sup>. The gene–gene covariance matrix of each data set was calculated across all arrays while ignoring the missing values. The transformation formula is:

$$X_{norm} = F2^{-1}(F1(x))$$

where  $F1$  and  $F2$  represents distribution functions of the actual and reference chips.

To get a description of intensities, we used the RMA-algorithm to measure averages between probes in a sample set. During this analysis RNA quality was evaluated in samples. The RNA degradation analysis was performed by using AffyRNAdeg, summary AffyRNAdeg, and plot AffyRNAdeg packages<sup>72</sup>. The DEGs were observed using the LIMMA package, that process the information based on modified statistics which is proportional to sample variance offsets. The LIMMA package measured the duplicate spots and quality weights. The statistical analysis was performed to categorize the genes based on the significant cutoffs values logFC greater than 1, FDR less than 0.05, AEL  $\geq$  40% and  $p$  value  $\leq$  0.05<sup>73</sup>.

**K-fold cross-validation.** We used K-Fold Cross-validation and Bootstrap test to estimate accuracy in the differential analysis<sup>74</sup> and this approach has the advantage that all the samples in the dataset can ultimately be used for both training and research. This technique is usually easier to calculate estimated average error and has been used to validate the shortlisted differentially expressed genes using the 'Boot' package of Bioconductor. Boots trapping is used effectively in molecular analysis to correct biases<sup>75</sup>. In such cases, we applied the generalized linear Gaussian models and used the 'cv.glm' method to test the k-fold cross-validation. It estimates the true error as the average rate error:

$$E = 1/K \sum_{i=K}^K E_i$$

Leave-one-out-cross-validation (LOOCV) continued to trail the Gaussian rule. The LOOCV approach is instinctively termed as the test set is left out and the rest of the data is used as the training-set<sup>75</sup>. We used  $N - 1$  subsets for training and the rest for testing. Increasing the number of folds would make the bias of the true error rate estimator low and valid<sup>75,76</sup>.

The true error is assessed as the average error rate on test cases:

$$E = 1/N \sum_{i=K}^N E_i$$

**Disease-gene curation of differentially expressed genes (DEGs).** The text mining is important in biomedical research to extract useful information<sup>77</sup>. This analysis is designed to identify the most significant DEGs, all 30 genes from 21 datasets were curated from the DAVID database to retrieve their gene symbol, gene name, Uniprot\_ID. These genes were curated using Comparative Toxicogenomics Database (CTD), Online Mendelian Inheritance in Man (OMIM), PubMed and MeSH databases to observe their role in oral cancer. This screening further shortlisted the significant DEGs<sup>78</sup>.

**Enrichment and cluster analysis.** The biological functions of the genes help us to understand the cellular level signaling network. We performed enrichment analysis using the DAVID tool<sup>78–80</sup>. FunRich tool was used to observe the biological functions of oral cancer-related DEGs at molecular level<sup>81</sup>. The list of DEGs were analyzed for their  $p$ - and FDR values<sup>79</sup>. For cluster analysis, gene expression values of cases and controls of each dataset were studied to observe genetic variations and expression profiling using One Matrix CIMminer tool<sup>82,83</sup>.

**Mutation analysis.** Mutations resulting from cancer and the inherited-disease process can be understood to decode the genetic variation by associations of genotype-phenotype. The human genome contains thousands of SNVs (single nucleotide variants) and many are known for the progression of the disease. Approximately 21% of amino acid substitutions are known to be associated with disease-progression in correspondence with missense single nucleotide variants located at PTM protein sites (post-translation modifications). The chemical modification of the amino acid thus basically extends the functionality of the associated protein<sup>19</sup>.

Mutation of differentially expressed genes were analyzed using online ActiveDriverDB database<sup>19</sup>. The needle plot mutations analysis provides a visual overview of the position, frequency, and functional significance of all identified mutations in our DEGs. PTM sites with all mutations and the predicted disordered region of protein sequences were observed. Placing the pins corresponds to the position along the sequence of the genes and protein, whereas the related mutation effect and PTM are explained in the figure legend.

**Protein–protein interaction.** The biological functions are mainly carried out by protein–protein interactions<sup>30,84</sup>. The interaction of proteins reveals that each protein interacts with one or more genes related to their molecular functions<sup>85</sup>. The biological networks indicate altered activity in normal or disease conditions. This gene-network aims to identify potentially OC associated gene signatures whose dysfunction directly contributes to disease phenotype are functionally associated. The gene signatures related to each source protein was measured. Human Annotated and Predicted Protein Interaction (HAPPI) and String databases were used to analyze gene–gene/protein–protein interactions of microarray dataset DEGs<sup>86</sup>. This database annotates and mine comprehensive physical as well as genetic mapping and includes experimentally validated data to simulate biological networks. We have mentioned the threshold for PPI network from HAPPI database. We used high-confidence interactions in our network (the five stars are equivalent to high score (0.90–1)). The role and association of these source and target genes in oral cancer were evaluated from Cancer Genetics Web, National Cancer and OMIM database. The molecular networks were visualized by Cytoscape software (version 3.6.0)<sup>87</sup>. The

Cytoscape Network Analyzer calculates topological properties of networks. The degree of annotation between the gene and disease is categorized by nodes in the network.

**Pathway analysis of oral cancer linked genes.** Reactomic analysis enable us to explore all metabolic networks of DEGs regarding their molecular mechanism. We analyzed these pathways to inter-connect DEGs to show the pathological mechanism of oral cancer. The KEGG and Wiki pathways databases were used to map target genes<sup>88,89</sup>. PathVisio tool was used to reconstruct the pathway model for understanding system-level analysis<sup>90</sup>.

**Toxicogenomic analysis.** The toxicogenomic analysis is carried out by a comparative toxicogenomic database (CTD) to retrieve exposome data. The exposome data helps investigate chemical-genome to phenome relationships to interpret the functional pathway cellular signaling-mechanism towards disease progression influenced by environmental exposures. It provides information regarding chemical-gene/protein and disease interactions which may reveal the particular gene-activity or expression regarding gene-disease connections. The curation of environment-disease exposure helps to analyze the available toxicogenomic information<sup>55</sup>.

**Prediction of regulatory motifs.** Cancer has a complex mechanism that can be explored by understanding the biological functions at transcription and post-transcriptional level. oPOSSUM version 3.0 was used to analyze promoter region target motifs like transcription factor binding sites (TFBS) or the overexpression of target matrices<sup>91,92</sup>. This information helps to understand the functional role of gene targets and eventually gene ontology<sup>93</sup>.

**Prediction of oral cancer-associated miRNA targets.** Numerous genes are involved in the biological signaling cascade. These cascades are influenced through small noncoding RNAs as post-transcriptional regulators, known as microRNAs (miRNAs). The function and expression of miRNA play a significant role in understanding gene etiology<sup>5,93</sup>. miRNA target prediction helps to explore the functional and molecular annotation of disease-specific DEGs<sup>5,94</sup>. Therefore, oral cancer associated DEGs miRNA targets were predicted by miRDB, an online database for functional microRNA target prediction. The target prediction data involves specie specific 3'-UTR sequences, 3'-UTR region length, miRNA seed binding-sites, miRNA-candidate target pairs along with target prediction scores, miRNA-target sequences, and other important description<sup>95,96</sup>. The MiRNA target predictive score is ranked and >80 was considered as a reliable score<sup>95,96</sup>.

**Drug-gene interaction analysis.** In our study, the drug-gene network analysis was performed to correlate our shortlisted DEGs with FDA approved commercially available anti-cancer drugs. CTD database was used to investigate the relationship between the chemical and disease at default parameters. In this analysis, DEGs were directly linked to anticancer drugs. All drugs, used in this interaction, were verified through the Drug Bank database to check approval status by the FDA.

## Data availability

All the other data that support the findings of this study are available from the corresponding author upon request.

Received: 11 December 2019; Accepted: 20 July 2020

Published online: 04 September 2020

## References

- Lozano, R. *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **380**, 2095–2128. [https://doi.org/10.1016/S0140-6736\(12\)61728-0](https://doi.org/10.1016/S0140-6736(12)61728-0) (2010).
- Khawaja, M. R. *et al.* Chewing of betel, areca and tobacco: Perceptions and knowledge regarding their role in head and neck cancers in an urban squatter settlement in Pakistan. *Asian Pac. J. Cancer Prev.* **7**, 95–100 (2006).
- Bhurgri, Y., Bhurgri, A. & Usman, A. Epidemiological review of head and neck cancers in Karachi. *Asian Pac. J.* **7**, 195–200 (2006).
- Masood, N., Kayani, M. A., Malik, F. A., Baig, R. M. & Faryal, R. Genetic variations in carcinogen metabolizing genes associated with oral cancer in pakistani population. *Asia. Pac. J. Cancer Prev.* **12**, 491–495 (2011).
- Muhammad, S. A. *et al.* Cellular signaling pathways in insulin resistance-systems biology analyses of microarray dataset reveals new drug target gene signatures of type 2 diabetes mellitus. *Front. Physiol.* **8**, 1–15 (2017).
- Golub, T. R. *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**(80), 531–527 (1999).
- Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 1–9 (2000).
- Okabe, H. *et al.* Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray. *Cancer Res.* **61**, 2129–2137 (2001).
- Hong, F. & Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24**, 374–382 (2008).
- Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–349 (2005).
- Wheeler, E. & Barroso, I. Genome-wide association studies and type 2 diabetes. *Brief. Funct. Genomics* **10**, 52–60 (2011).
- Siddiqui, A. S. *et al.* Sequence biases in large scale gene expression profiling data. *Nucleic Acids Res.* **34**, e83 (2006).
- Wagner, J. M., Hackanson, B., Lübbert, M. & Jung, M. Histone deacetylase (HDAC) inhibitors in recent clinical trials for cancer therapy. *Clin. Epigenet.* <https://doi.org/10.1007/s13148-010-0012-4> (2010).
- Wang, Z. *et al.* Cdc20: A potential novel therapeutic target for cancer treatment. *Curr. Pharma. Des.* **19**, 3210–3214 (2013).
- Muller, A. *et al.* Chemokine receptors in head and neck cancer: Association with metastatic spread and regulation during chemotherapy. *Int. J. Cancer* <https://doi.org/10.1002/ijc.21514> (2014).

16. Ko, Y. *et al.* Association of CYP1B1 codon 432 mutant allele in head and neck squamous cell cancer is reflected by somatic mutations of p53 in tumor tissue. *Cancer Res.* **61**, 4398–4404 (2001).
17. Gonzalez-perez, A., Mustonen, V., Reva, B. & Ritchie, G. R. S. *NIH Public Access.* **10**, 723–729 (2014).
18. Macarthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
19. Krassowski, M. *et al.* *ActiveDriverDB: Human disease mutations and genome variation in post-translational modification sites of proteins.* **46**, 901–910 (2018).
20. Shimada, T. *et al.* Metabolic activation of polycyclic aromatic hydrocarbons and other procarcinogens by cytochromes p450 1a1 and p450 1b1 allelic variants and other human cytochromes p450 in *Salmonella typhimurium* NM2009. *Drug Metab. Dispos.* **29**, 1176–1182 (2001).
21. Boccia, S. *et al.* Polymorphisms in metabolic genes, their combination and interaction with tobacco smoke and alcohol consumption and risk of gastric cancer: A case–control study in an Italian population. *BMC Cancer* **7**, 3–10 (2007).
22. Luo, J. *et al.* Nuclear cytokine-activated IKK a controls prostate cancer metastasis by repressing. *Maspin* <https://doi.org/10.1038/nature05656> (2007).
23. Khalkhali-ellis, Z. *Maspin. New Front.* **12**, 7279–7284 (2006).
24. Zou, Z. *et al.* Maspin, a serpin with tumor-suppressing activity in human mammary epithelial cells. *Science* **80**(263), 526–529 (1994).
25. Sood, A. K. *et al.* The paradoxical expression of Maspin in ovarian carcinoma. *Clin Cancer Res* **1**(8), 2924–2932 (2015).
26. Ahtens, S. M. *et al.* Expression of the p53 and Maspin protein in primary prostate cancer: Correlation with clinical features. *Int. J. Cancer* **342**, 337–342 (2001).
27. Ohike, N., Maass, N., Mundhenke, C. & Biallek, M. Clinicopathological significance and molecular regulation of maspin expression in ductal adenocarcinoma of the pancreas. *Cancer Lett.* **199**, 193–200 (2003).
28. Yu, M., Zheng, H. & Tsuneyama, K. Paradoxical expression of Maspin in gastric carcinomas: Correlation with carcinogenesis and progression. *Hum. Pathol.* **38**, 1248–1255. <https://doi.org/10.1016/j.humpath.2006.11.025> (2007).
29. Xia, W. *et al.* High tumoral Maspin expression is associated with improved survival of patients with oral squamous cell carcinoma. *Oncogene* **19**, 2398–2403 (2000).
30. Li, Y. X. *et al.* Molecular cloning, sequence identification, and gene expression analysis of bovine ADCY2 gene. *Mol. Biol. Rep.* **41**, 3561–3568 (2014).
31. Ding, Q. *et al.* Raf kinase activation of adenylyl cyclases: Isoform-selective regulation. *Mol. Pharmacol.* **66**, 921–928 (2004).
32. Drozdov, I. *et al.* Gene network inference and biochemical assessment delineates GPCR pathways and CREB targets in small intestinal neuroendocrine neoplasia. *PLoS ONE* **6**, e22457 (2011).
33. Iyengar, R. Molecular and functional diversity of mammalian GS – stimulated adenylyl cyclases. *The FASEB J.* **7**, 768–775 (1993).
34. Plaza-menacho, I., Mologni, L. & Mcdonald, N. Q. Mechanisms of RET signaling in cancer: Current and future implications for targeted therapy. *Cell Signal.* **26**, 1743–1752 (2014).
35. Liang, B., Li, C. & Zhao, J. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis. *Med. Oncol.* **33**, 1–8 (2016).
36. Liu, J. *et al.* Aberrantly methylated—differentially expressed genes and pathways in colorectal cancer. *Cancer Cell Int.* <https://doi.org/10.1186/s12935-017-0444-4> (2017).
37. Sun, Y., Jia, X., Hou, L. & Liu, X. Screening of differently expressed miRNA and mRNA in prostate cancer by integrated analysis of transcription data. *Urology* <https://doi.org/10.1016/j.urology.2016.04.041> (2016).
38. Dolićanin, Z., Lj, J. V. & Katić, V. Biomarkers for detection treatment decision and prognosis ontogeneses telomerase tumor suppressor genes genomic studies in bladder cancer. *Facta Universitatis* **14**, 1–5 (2007).
39. Hardin, M. *et al.* Severe chronic obstructive pulmonary disease IN Poland. *Am. J. Respir. Cell Mol. Biol.* **47**(2), 203–208. <https://doi.org/10.1165/rcmb.2012-0011OC> (2012).
40. Yu, S. J. *et al.* SPARCL1, Shp2, MSH2, E-cadherin, p53, ADCY-2 and MAPK are prognosis-related in colorectal cancer. *World J. Gastroenterol.* **17**, 2028–2036 (2011).
41. Yoon, H. *et al.* Proteomics analysis identifies new components of the fission and budding yeast anaphase-promoting complexes. **12**, 2048–2054 (2002).
42. Peters, J. *The anaphase-promoting complex Proteolysis in mitosis and beyond.* **9**, 931–943 (2002).
43. Aparecida, C. *et al.* Translational oncology down-regulation of ANAPC13 and CLTCL1: Early events in the progression of pre-invasive ductal carcinoma of the breast 1, 2. *Transl. Oncol.* **5**, 113–123 (2012).
44. Ricklin, D., George, H., Kun, Y. & John D. L. Complement – a key system for immune surveillance and homeostasis. *Nat. Immunol.* **11**, 785–797 (2010).
45. Serna, M., Giles, J. L., Morgan, B. P. & Bubeck, D. Structural basis of complement membrane attack complex formation. *Nat. Commun.* **7**, 10587 (2016).
46. Podack, E. R. Molecular composition of the tubular structure of the membrane attack complex of complement. *J. Biol. Chem.* **259**, 8641–8647 (1984).
47. Botto, M. *et al.* Complement in human diseases: Lessons from complement deficiencies. *Mol. Immunol.* **46**, 2774–2783 (2009).
48. Gelderman, K. A., Tomlinson, S., Ross, G. D. & Gorter, A. Complement function in mAb-mediated cancer immunotherapy. *Trends Immunol.* **25**, 158–164 (2004).
49. Ying, L. *et al.* Complement component 7 (C7), a potential tumor suppressor, is correlated with tumor progression and prognosis. *Oncotarget* **7**, 86536–86546 (2016).
50. Shen, Y. B. *et al.* A 425 T > C polymorphism in complement C7 association with resistance to *Aeromonas hydrophila* in grass carp. *Genet Mol. Res.* **1**, 1–9 (2016).
51. Ying, L. *et al.* Complement component 7 (C7), a potential tumor suppressor, is correlated with tumor progression and prognosis. *Oncotarget* **7**, 86536 (2016).
52. Mortensen, H. M. & Euling, S. Y. Integrating mechanistic and polymorphism data to characterize human genetic susceptibility for environmental chemical risk assessment in the 21st century. *Toxicol. Appl. Pharmacol.* **271**(3), 395–404. <https://doi.org/10.1016/j.taap.2011.01.015> (2011).
53. Mattingly, C. J. *et al.* The comparative toxicogenomics database: A cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.* **92**, 587–595 (2006).
54. Davis, A. P. *et al.* The comparative toxicogenomics database: Update 2011. *Nucleic Acids Res.* **92**, 1067–1072 (2011).
55. Davis, A. P. *et al.* The comparative toxicogenomics database: Update 2019. *Nucleic Acids Res.* **92**, 948–954 (2019).
56. Davis, A. P. *et al.* Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical—gene—disease networks. *Nucleic Acids Res.* **37**, 786–792 (2009).
57. Lynam-Lennon, N., Maher, S. G. & Reynolds, J. V. The roles of microRNA in cancer and apoptosis. *Biol. Rev.* **84**, 55–71 (2009).
58. Mamidi, S., Höne, S. & Kirschfink, M. The complement system in cancer: Ambivalence between tumour destruction and promotion. *Immunobiology* **222**, 45–54 (2017).
59. Suryawanshi, S. *et al.* Complement pathway is frequently altered in endometriosis and endometriosis-associated ovarian cancer. *Clin. Cancer Res.* **20**, 6163–6174 (2014).
60. Pio, R. *Tumor Microenvironment and Cellular Stress* (Springer, Berlin, 2014), 772.

61. Thierfelder, N., Fang, Z., Khanin, R., Rajewsky, N. & Selbach, M. Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
62. Lim, L. P., Lau, N. C., Garrett-engele, P. & Grimson, A. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **292**, 288–292 (2005).
63. Shin, C. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008).
64. Wong, N. & Wang, X. miRDB: An online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.* **43**, 146–152 (2015).
65. Wang, X. & Wang, X. miRDB: A microRNA target prediction and functional annotation database with a wiki interface miRDB: A microRNA target prediction and functional annotation database with a wiki interface. *RNA* **14**, 1012–1017. <https://doi.org/10.1261/rna.965408.was> (2008).
66. Tibshirani, R. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520 (2001).
67. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. Gene expression omnibus a comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185–193 (2003).
68. Fujita, A., Sato, J. R., de Oliveira Rodrigues, L., Ferreira, C. E. & Sogayar, M. C. Evaluating different methods of microarray data normalization. *BMC Bioinform.* **7**, 1–11 (2006).
69. Obenchain, V. *et al.* Variant annotation: A bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**, 2076–2078 (2014).
70. Yoon, D., Yi, S. G., Kim, J. H. & Park, T. Two-stage normalization using background intensities in cDNA microarray data. *BMC Bioinform.* **5**, 1–12 (2004).
71. Obenchain, V. *et al.* Variant annotation: A bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**, 2076–2078 (2014).
72. Affymetrix® Microarray Suite User's Guide. Version 5.0. (2000, 2001).
73. Jin, Y. & Da, W. Screening of key genes in gastric cancer with DNA microarray analysis (retraction of vol 18, 37, 2013). *Eur. J. Med. Res.* **20**, 30 (2015).
74. Geisser, S. The predictive sample reuse method with application. *J. Am. Stat. Assoc.* **70**, 320–328 (1975).
75. Picard, R. R. & Cook, R. D. Cross-validation of regression models. *J. Am. Stat. Assoc.* **79**, 575–583 (1984).
76. MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **8**, 827–838 (2010).
77. Clematide, S. & Rinaldi, F. Ranking relations between diseases, drugs and genes for a curation task. *J. Biomed. Semant.* **3**, S5 (2012).
78. Nam, D. & Kim, S. Y. Gene-set approach for expression pattern analysis. *Brief. Bioinform.* **9**, 189–197 (2008).
79. Muhammad, S. A. *et al.* MicroRNA expression profiling of human respiratory epithelium affected by invasive candida infection. *PLoS ONE* **10**, 1–14 (2015).
80. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
81. Pathan, M. *et al.* FunRich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics* **15**, 2597–2601 (2015).
82. Weinstein, J. N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **80**(275), 343–349 (1997).
83. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **24**, 236–244 (2000).
84. Muhammad, S. A. *et al.* Prioritizing drug targets in Clostridium botulinum with a computational systems biology approach. *Genomics* **104**, 24–35 (2014).
85. Rachlin, J., Cohen, D. D., Cantor, C. & Kasif, S. Biological context networks: A mosaic view of the interactome. *Mol. Syst. Biol.* **2**, 1–12 (2006).
86. Chen, J. Y., Mamidipalli, S. R. & Huan, T. HAPPI: An online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics* **10**, 1–11 (2009).
87. Cline, M. S. *et al.* Integration of biological networks and gene expression data using cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
88. Bergholdt, R. *et al.* Identification of novel type 1 diabetes candidate genes by integrating genome-wide association data, protein–protein interactions, and human pancreatic islet gene expression. *Diabetes* **61**, 954–962 (2012).
89. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).
90. Kutmon, M. *et al.* PathVisio 3: An extendable pathway analysis toolbox. *PLoS Comput. Biol.* **11**, 1–14 (2015).
91. Ho Sui, S. J. *et al.* oPOSSUM: Identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.* **33**, 3154–3164 (2005).
92. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32**, 199–203 (2004).
93. Alshalalfa, M. & Alhaji, R. Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures. *BMC Bioinform.* **14**, 1–13 (2013).
94. Ambros, V. R. The function of animal MicroRNAs. *Nature* <https://doi.org/10.1038/nature02871> (2019).
95. Liu, W. & Wang, X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol.* **20**, 1–10 (2019).
96. Chen, Y. & Wang, X. miRDB: An online database for prediction of functional microRNA targets. *Nucleic Acids Res.* **48**, 127–131 (2020).

## Author contributions

SZA: Collected materials & interpreted the data; MIQ: Designed and write the manuscript; SAM: Directed the project.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-71346-7>.

**Correspondence** and requests for materials should be addressed to S.A.M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020