scientific reports

Check for updates

OPEN

Mining the plasma-proteome associated genes in patients with gastro-esophageal cancers for biomarker discovery

Frederick S. Vizeacoumar[1], Hongyu Guo[2], Lynn Dwernychuk[3], Adnan Zaidi[3,4], Andrew Freywald[1], Fang-Xiang Wu[2,5,6], Franco J. Vizeacoumar[3,4,8]✉ & Shahid Ahmed[3,4,7]✉

Gastro-esophageal (GE) cancers are one of the major causes of cancer-related death in the world. There is a need for novel biomarkers in the management of GE cancers, to yield predictive response to the available therapies. Our study aims to identify leading genes that are differentially regulated in patients with these cancers. We explored the expression data for those genes whose protein products can be detected in the plasma using the Cancer Genome Atlas to identify leading genes that are differentially regulated in patients with GE cancers. Our work predicted several candidates as potential biomarkers for distinct stages of GE cancers, including previously identified CST1, INHBA, STMN1, whose expression correlated with cancer recurrence, or resistance to adjuvant therapies or surgery. To define the predictive accuracy of these genes as possible biomarkers, we constructed a co-expression network and performed complex network analysis to measure the importance of the genes in terms of a ratio of closeness centrality (RCC). Furthermore, to measure the significance of these differentially regulated genes, we constructed an SVM classifier using machine learning approach and verified these genes by using receiver operator characteristic (ROC) curve as an evaluation metric. The area under the curve measure was > 0.9 for both the overexpressed and downregulated genes suggesting the potential use and reliability of these candidates as biomarkers. In summary, we identified leading differentially expressed genes in GE cancers that can be detected in the plasma proteome. These genes have potential to become diagnostic and therapeutic biomarkers for early detection of cancer, recurrence following surgery and for development of targeted treatment.

Cancers of the stomach and esophagus or gastro-esophageal (GE) cancers represent a highly aggressive disease and are one of the major causes of cancer-related death in the world. Stomach cancer is the fifth most common cancer and the third leading cause of cancer-related death worldwide. For example, in 2018, more than 1 million new cases of stomach cancer were diagnosed and about 783,000 people die from it[1]. Likewise, esophageal cancer is the seventh most common cancer and the sixth leading cause of cancer-related death. Each year more than 500,000 new cases of esophageal cancer are diagnosed and about 509,000 people die from it[1]. Despite improvements in surgical and radiation treatments and the availability of newer agents, the prognosis of patients with recurrent GE cancers remains very poor[2–4]. The need for novel strategies to improve current therapy is therefore vital in the management of GE cancers.

It is well known that cancer development and progression are triggered by altered activities and dysregulated expression of genes that control cell proliferation and differentiation[5]. Comparative assessment of genetic aberrations between cancerous and matched normal tissues as control, has facilitated identification of new biomarkers that may also serve as new therapeutics targets or predict various cancer-related outcomes. There are several

[1]Department of Pathology and Laboratory Medicine, College of Medicine, University of Saskatchewan, Saskatoon, Canada. [2]Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, Canada. [3]Cancer Research, Saskatchewan Cancer Agency, Saskatoon, Canada. [4]Division of Oncology, University of Saskatchewan, Saskatoon, Canada. [5]Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, Canada. [6]Department of Computer Science, University of Saskatchewan, Saskatoon, Canada. [7]Saskatoon Cancer Center, University of Saskatchewan, 20 Campus Drive, Saskatoon, SK S7N4H4, Canada. [8]Saskatoon Cancer Center, University of Saskatchewan, 107 Wiggins Road, Saskatoon, SK S7N5E5, Canada. ✉email: franco.vizeacoumar@usask.ca; shahid.ahmed@saskcancer.ca

known biomarkers that are associated with tumorigenesis or have prognostic and predictive values in patients with stomach and esophageal cancers[6–10]. For example, approximately 20% of stomach and gastroesophageal junction cancers are associated with the amplification of the HER2 gene that is an important therapeutic target and predicts response to trastuzumab[8]. Most biomarkers such as TP53 or CDH1, however, have limited therapeutic and predictive values and presence or absence of them does not alter treatment strategies[9]. There is a strong unmet need for novel biomarkers in the management of GE cancers to identify new therapeutic targets and to yield predictive response to the available therapies.

We conducted this study by intersecting the gene expression profiles from The Cancer Genome Atlas (TCGA) with the plasma proteome databases to identify leading genes that are differentially expressed (upregulated or downregulated) in patients with esophageal and stomach cancers[11,12]. We also applied machine-learning approaches to test our predictive accuracy. Overall, the purpose of these analyses is to identify differentially expressed novel tumor-specific genes that code for the plasma proteins and use this information to develop blood-based prognostic biomarker studies in the near future.

Methods

TCGA gene expression analyses. We obtained the level-3 HiSeq RSEM gene-normalized RNA-seq gene expression data for stomach adenocarcinoma (STAD) and esophageal carcinoma (ESCA) from the TCGA database[11]. Overall, gene expression data for 415 independent tumor samples and 35 matching normal tissue samples for STAD and for 185 ESCA cases with 11 matching normal tissue samples were available. We also downloaded the plasma proteome database from http://www.plasmaproteomedatabase.org. The database contained information on 1241 protein-coding genes, while gene expression profiles from TCGA mapped to 1232 protein-coding genes. To analyse plasma proteome genes, we used non-parametric Mann–Whitney-U test to identify genes that are expressed at significantly different levels ($p < 0.05$) in cancerous and normal tissues. The deregulated genes were grouped according to tumor stages based on the available patient data. This allowed identification of genes with expression significantly increased or decreased at multiple stages of cancer.

Computational method using gene co-expression network analysis. Gene co-expression networks were used for analyzing the importance of genes and their relationships with other genes. In a weighted gene co-expression networks (WGCN), nodes represent the gene expression profiles, edges represent the pairwise correlation between gene expressions while the edge weights represented the correlation strengths. For our study, the correlation strength of a pair of genes was measured by their similarity, which was calculated by the Pearson correlation coefficient (PCC) between their expression profiles. Specifically, for each pair of genes g_i and g_j, its strength was calculated as

$$S_{ij} = pcc(g_i, g_j) = \frac{covar(x_i, x_j)}{var(x_i)var(x_j)}$$

where x_i and x_j are expression profiles of genes g_i and g_j, respectively; $var(x_i)$ and $var(x_j)$ are the variance of x_i and x_j, respectively while $covar(x_i, x_j)$ is the covariance of x_i and x_j. Since the result of PCC has a value between − 1 and 1, we transformed the similarity measure S_{ij} into $Dissimilarity\_cor_{ij}$ and $Simmilarity\_cor_{ij}$, as follows:

$$Dissimilarity\_cor_{ij} = \frac{1 - pcc(g_i, g_j)}{2}$$

$$Similarity\_cor_{ij} = \frac{1 + pcc(g_i, g_j)}{2}$$

$Similarity\_cor_{ij}$ represents the positive correlation between the genes since the larger value indicates the stronger positive correlation between the pair of genes, while $Dissimilarity\_cor_{ij}$ represents the negative correlation between a pair of genes since the larger value indicates the stronger negative correlation between genes, which is also called the distance of a pair of genes. Both $Dissimilarity\_cor_{ij}$ and $Simmilarity\_cor_{ij}$ take on values in [0, 1], that was used for further network analysis.

We used WGCNA R package[13], that is commonly used in recent studies[14,15], to construct the weighted network for the stomach and esophageal cancer datasets. To filter out the noisy edges in WGCNs, we applied the soft thresholding scheme[13] by raising the co-expression similarity to a soft power β to shrink the lower correlations. Hence, the strengths in WGCN is represented by

$$Dissimilarity\_cor_{ij} = \left(\frac{1 - pcc(g_i, g_j)}{2}\right)^{\beta}$$

$$Similarity\_cor_{ij} = \left(\frac{1 + pcc(g_i, g_j)}{2}\right)^{\beta}$$

where $\beta \geq 1$. The criterion for the determination of the soft power β is dependent on the model fitting index of the scale-free topology[13]. The scale-free networks were constructed because they have strong ability to tolerate against errors[16]. We then removed the self-loop edges of nodes by setting the diagonal elements of the adjacency

matrix as 0. Also, we applied a hard threshold to remove weak edges between nodes by setting the threshold as 0.01.

Next, we calculated the closeness centrality measures to identify the closeness of a particular node to all other nodes in a network[17]. This closeness centrality is the inverse of the average shortest-path distance from one node to other nodes in the network. It also indicates the efficiency of one node to spread information through a network. Finally, based on the closeness centrality score of genes in networks, we defined a novel metric of gene importance in networks as the ratio of the closeness centrality (RCC) of a gene in its corresponding similarity network and its dissimilarity network, i.e.

$$RCC_N = \frac{CN_{sim}}{CN_{dis}} \quad \text{and} \quad \mathrm{RCC_P} = \frac{\mathrm{CT_{sim}}}{\mathrm{CT_{dis}}}$$

where CN_{sim} and CN_{dis} represent the closeness centrality score in the similarity network and the dissimilarity network of normal samples, while CT_{sim} and CT_{dis} represent the closeness centrality score in the similarity network and the dissimilarity network of tumor samples, respectively. We expect that the biomarkers should be significantly different in terms of log2(RCC) between the normal and tumor samples.

Machine learning approach to test the significance. To test the significance of the differentially regulated genes, we constructed a support vector machine (SVM) classifier with linear kernel. The features were based on the leading upregulated genes, the leading downregulated genes and a set of randomly selected genes for each cancer type. Accordingly, we used Receiver Operator Characteristic (ROC) curve as the evaluation metric for the classification of cancer patients and normal samples as in previous studies[18]. The ROC curve is an evaluation metric for binary classification problems, which visualizes the trade-off between true positive rate (TPR) and false positive rate (FPR). We then measured the area under the curve (AUC), as higher the AUC, the better the performance of the model in distinguishing between normal and tumor samples.

Results

Leading upregulated genes in GE cancers. We examined gene expression of 1232 protein-coding genes that were detected in plasma proteome in tumors of 185 patients with esophageal cancer and in the matching tissue of 11 subjects with no cancer. Among cancer patients, 18 (9.7%) had stage I tumors, 78 (42.2%) had stage II disease, 56 (30.3%) had stage III disease, and 9 (4.9%) had stage IV cancer. In 24 (13%) patients, cancer stage was not known. The comparison between esophageal tumors and healthy tissue showed BIRC5 (p = 2.61E−08), APOC2 (p = 3.23E−08), CENPF (p = 4.38E−08), STMN1 (p = 5.74E−08), and HNRPC (p = 8.21E−08) to be five leading genes overexpressed in esophageal cancer (Fig. 1A). The stage-based assessment of overexpressed genes showed significant overexpression of BIRC5, APOC2, CENPF, STMN1, and HNRPC across all cancer stages including early, locally advanced and metastatic esophageal tumors (Fig. 1B). The significance of expression for each gene compared between normal and tumor samples (p values), along with the number of samples in each stage are provided in Supplementary Table S1.

For stomach cancer, we evaluated 415 cases and compared them with 35 normal tissue samples. Among patients with stomach cancer, 57 (13.7%) had stage I cancer, 123 (29.6%) had stage II disease, 169 (40.7%) had stage III disease, and 41 (9.9%) had stage IV cancer. In 25 patients (6%) with stomach cancer, the disease stage was not known. Comparison between normal stomach tissue and stomach tumors showed that CST1 (p = 3.97E−21), INHBA (p = 9.22E−20), ACAN (p = 1.08E−19), HSP90AB1 (p = 2.62E−19), and HSPD1 (p = 3.91E−19) were the leading five genes that were overexpressed in stomach cancer (Fig. 1C). The stage-based assessment of overexpressed genes showed significant upregulation of CST1, INHBA, ACAN, HSP90AB1, and HSPD1 genes across all stages, including early, locally advanced and metastatic stomach cancer (Fig. 1D). The significance of expression for each gene compared between normal and tumor samples (p values), along with the number of samples in each stage are provided in Supplementary Table S2.

Stage-specific upregulation of genes in GE cancers. We next examined the pattern of gene expression based on the specific-stage of the disease in GE cancers. In addition to the five overexpressed genes reported above, the stage-based analysis showed a differential expression of following genes based on the stage of the disease. Patients with stage I esophageal cancer had significantly higher expression of CPS1 (p = 0.003), PNP (p = 0.007), SERPINB8 (p = 0.042) and EHD1 (p = 0.046). In patients with stage II disease, MSN (p = 0.003), KRT5 (p = 0.004), TNC (p = 0.007), and NAP1L4 (p = 0.018) were overexpressed compared with other stages of the disease. In patients with stage IV esophageal cancer CYCS (p = 0.014), PON3 (p = 0.14), ACPP (p = 0.047), and RPL22 (p = 0.047) were significantly upregulated compared with patients with early-stage cancer. We did not notice a stage-specific upregulated gene in stage III esophageal cancer (Fig. 2A). The significance of expression for each gene compared between normal and tumor samples (p values), are provided in Supplementary Table S3. Likewise, the stage-based analysis in patients with stomach cancer showed a differential expression of several genes at the specific stages of the disease. Thus, patients with stage I, II, III, and IV stomach cancer have significantly higher overexpression of PYGB (p = 0.043), TNF (p = 0.02), HLA-A (0.05), and EFNB2 (0.001) genes, respectively (Fig. 2B). The significance of expression for each gene compared between normal and tumor samples (p values), are provided in Supplementary Table S4.

Leading progressively upregulated genes in GE cancers. We also examined if certain gene expression pattern intensifies in parallel with the progression of the disease. This analysis showed a gradual stage-wise increasing expression pattern for ANGPT2, APOC2, CXCL5, HIST1H1E, IL17A, IL2RA, IL8, OSM, PF4V1, and

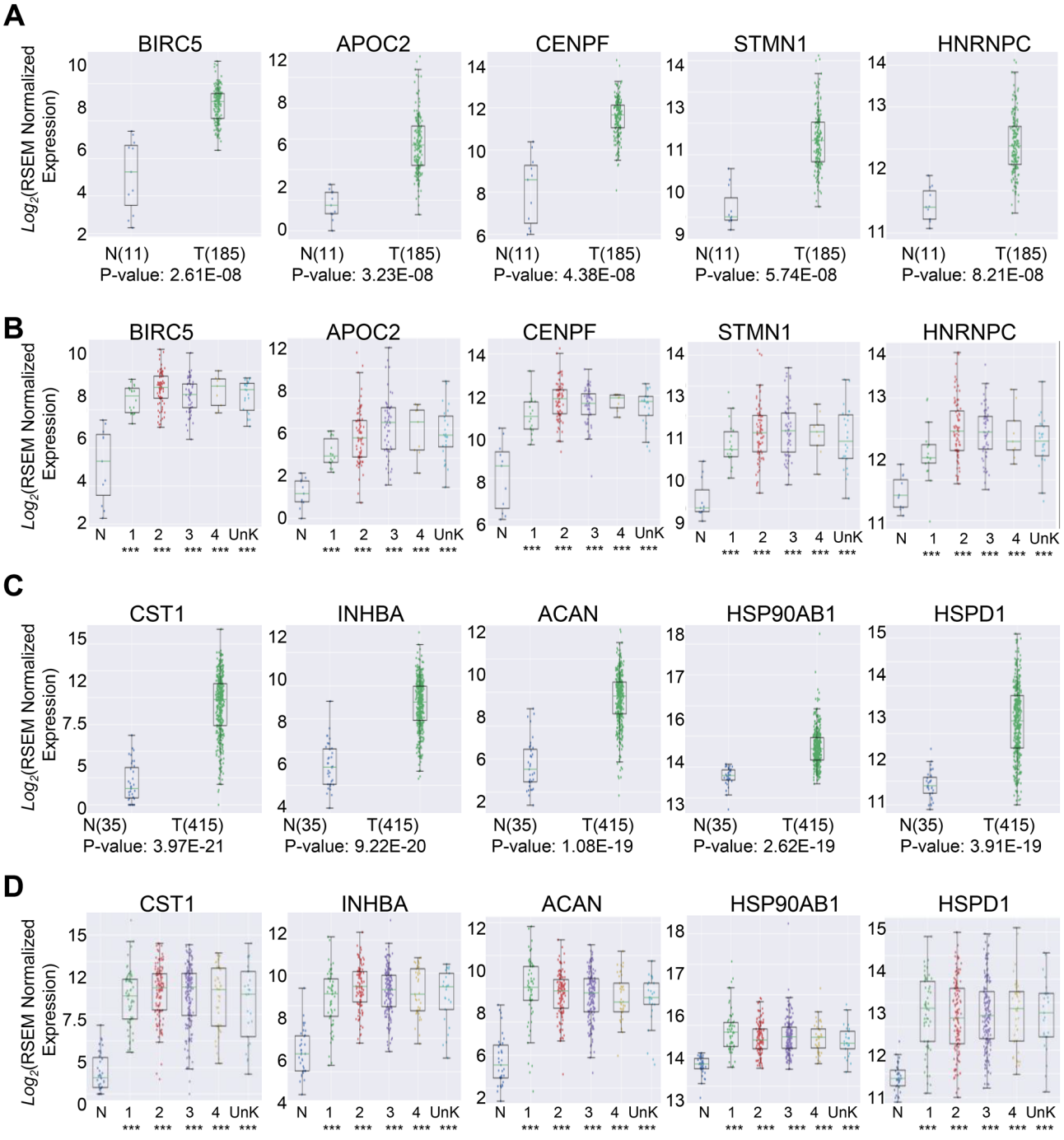

Figure 1. (**A**) The comparison between patients with esophageal cancer versus control individuals showed significant overexpression BIRC5, APOC2, CENPF, STMN1, and HNRPC in patients with esophageal cancer. (**B**) BIRC5, APOC2, CENPF, STMN1, and HNRPC significantly overexpressed across all stages in patients with esophageal cancer. (**C**) The comparison between patients with stomach cancer and healthy control showed significant overexpression of CST1, INHBA, ACAN, HSP0AB1, and HSPD1 in patients with stomach cancer. (**D**) Assessment of CST1, INHBA, ACAN, HSP90AB1, and HSPD1 in different stages of stomach cancers showed a significant overexpression of these genes across all stages in patients with stomach cancer.

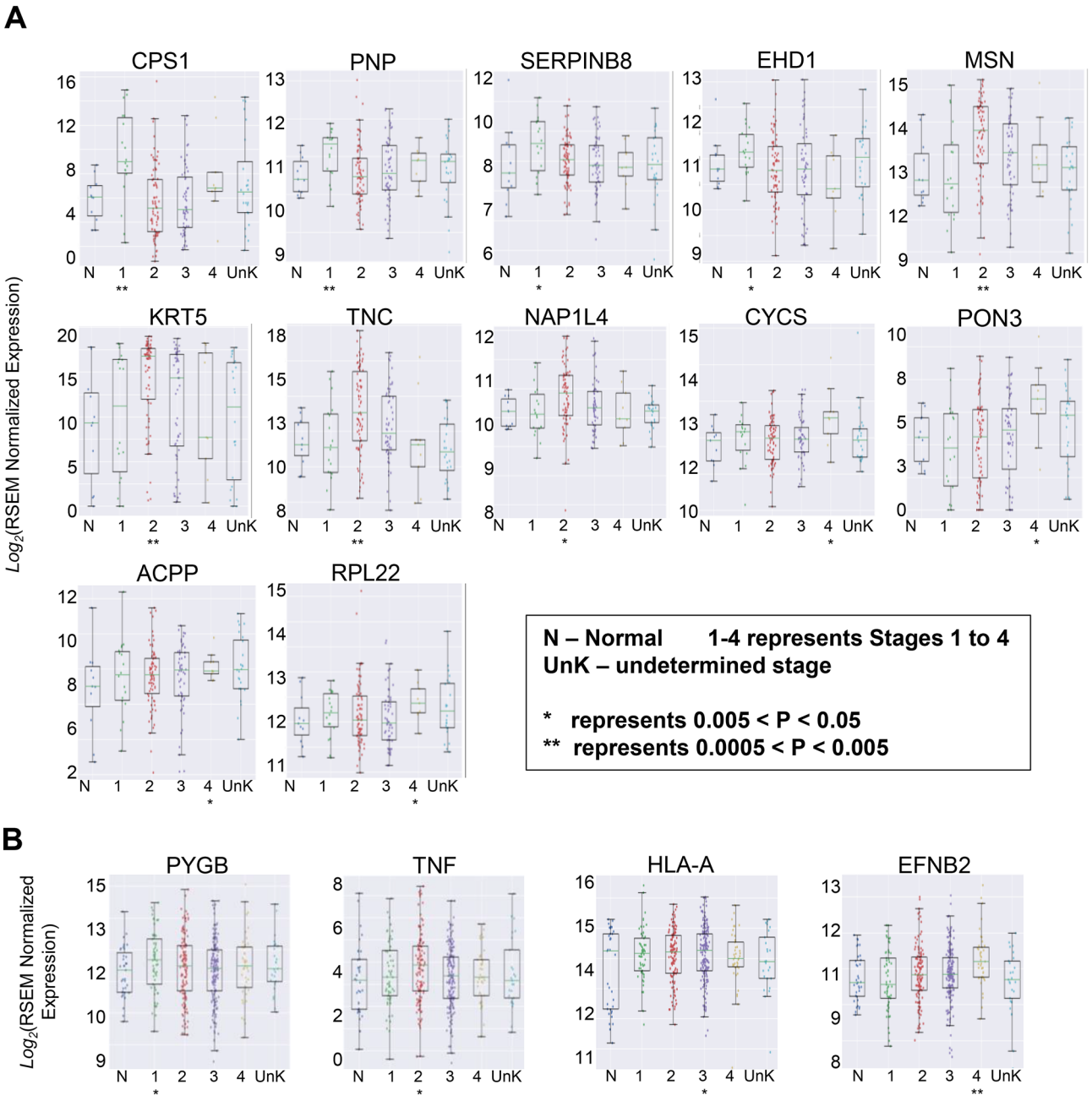

Figure 2. (**A**) Stage-based analysis of esophageal cancer showed a differential expression of various genes based on the stage of the disease. (**B**) Stage-based analysis showed a differential expression of various genes based on the stage of the disease in patients with stomach cancer.

SAA4 in esophageal cancer. Among them, a gradual increase from early-stage cancer to more advanced stages were strongest for APOC2 and IL8 genes followed by for SAA4 and OSM genes, whereas HIST1H1E, PF4V1, CXCL5, and IL2RA expression showed limited stage-wise upregulation (Fig. 3A). In the stomach cancer, the stage-wise analysis showed a gradual increasing expression pattern for AGRN, CETP, FGL1, HABP2, MDK, OSMR, RNASE2, SELE, SERPINE1, and VCAN. Among them, the upregulation from the early-stage cancer to more advanced stages were the strongest for RNASE2, SERPINE1, and CETP (Fig. 3B).

Leading downregulated genes in GE cancers. In addition to the upregulated genes, we also examined leading downregulated genes that could play a major role in pathogenesis and progression of cancer. Our analysis showed that following five genes were most significantly downregulated in esophageal cancer: C16orf89 (9.78E−08), AR (1.01E−07), CKB (1.17E−07), ADH1B (1.79E−07), and NCAM1 (2.15E−07) (Fig. 4A). We did not observe any stage-specific down-regulation for esophageal cancer. However, the stage-wise analysis showed

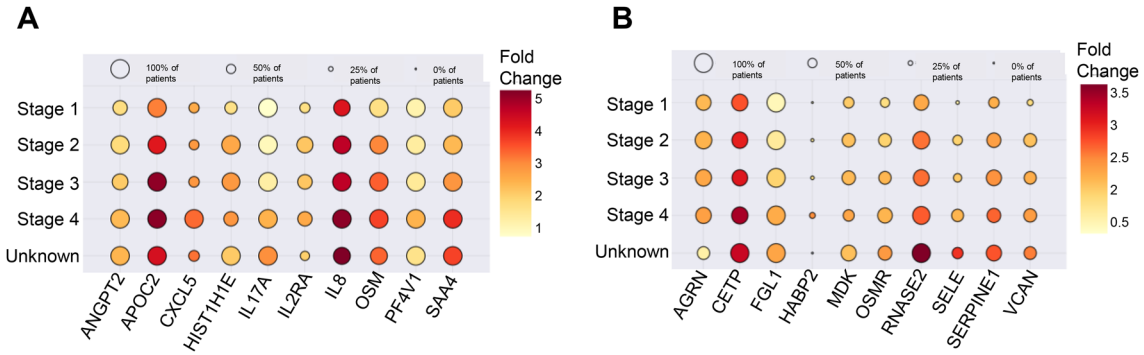

Figure 3. (**A**) Stage-wise incremental over-expressions of leading genes in patients with esophageal cancer. (**B**) Stage-wise incremental over-expressions of leading genes in patients with stomach cancer.

a gradual downregulation of the following genes from stage I to IV esophageal cancer: AHNAK, APOM, ART3, CAMK2D, MB, MEGF8, MMRN1, PROC, S100A1, and TNFRSF10C (Fig. 4B).

In patients with stomach cancer following five genes were most significantly downregulated: GPX3 (1.65E−19), CLEC3B (5.70E−19), CFD (5.68E−18), GSN (4.5 IE−17), and CCL14 (1.12E−16) (Fig. 4C). The stage-based analysis showed that C1R (2.34E−04), A2M (3.15E−04), LTBP1 (3.15E−04), SERPING1 (4.23E−04), and BASP1 (4.88E−04) were significantly downregulated only in patients with stage I (early-stage) stomach cancer whereas SERPINB6 (2.0E−03), G0S2 (1.80E−02), HSD17B10 (2.50E−02), ECM1 (2.60E−02), and PRDX1 (2.80E−02) were significantly downregulated in patients with stage III (locally advanced) stomach cancer (Fig. 4D). The significance of expression for each gene compared between normal and tumor samples (p values), are provided in Supplementary Table S5. The stage-wise analysis also revealed a progressive downregulation of the following genes from stage I to IV stomach cancer: ACAA1, AZGP1, BLVRB, CYB5A, EPHX2, FBP1, REG1A, SECTM1, SELENBP1, and TPPP3 (Fig. 4E,F). The significance of expression for each gene compared between normal and tumor samples (p values), are provided in Supplementary Table S6.

Measuring robustness and significance of the identified biomarkers. We next sought to evaluate the predictive accuracy of the leading candidates as potential biomarkers in GE cancers. Towards this, we calculated the similarity and dissimilarity matrices. These matrices were used to construct the co-expression network with weighted co-expression network analysis (WGCNA)[13]. Using complex network analysis, we calculated the centrality of the genes in the constructed network and verified the importance of biomarkers. The log RCC calculated for the differentially expressed genes between normal and tumor samples are shown in Fig. 5A,B for esophageal and stomach cancers, respectively. To statistically confirm our conclusion, we applied the paired samples Wilcoxon signed-rank test to $\log_2$(RCC) of the identified biomarkers between normal and tumor samples for both cancer types and found them to be significant ($p < 0.05$). Based on these results, we are confident on our predictive accuracy of these genes as potential biomarkers.

Since the sample sizes of the esophageal and stomach cancer datasets were small, a fivefold cross validation was adopted to evaluate the performances of SVM models on each dataset. The corresponding mean ROC curves of 50 executions of fivefold cross validation are illustrated in Fig. 5C,D. For the esophageal cancer dataset, the feature sets include the up-regulated genes (Fig. 1A), the down-regulated genes (Fig. 4A) and five randomly selected genes. Similarly, for the stomach cancer dataset, the feature sets include the up-regulated genes (Fig. 1C), the down-regulated genes (Fig. 4C) and five randomly selected genes as the third set. The model based on the up-regulated gene group has the AUC of 0.9941 with standard deviation of 0.0031 for esophageal cancer and the AUC of 0.9924 with standard deviation of 0.0038 for stomach cancer. Likewise, the AUC of the downregulated genes are 0.9788 (standard deviation 0.0265) and 0.9770 (standard deviation 0.0114) for esophageal and stomach cancers respectively. Meanwhile, the classifier model using the random selected genes has the lowest AUC score 0.9280 with standard deviation of 0.1137 for esophageal cancer and 0.5603 with standard deviation of 0.1664 for stomach cancer. This suggests that the features based on the differentially expressed genes are significant at identifying patients from normal samples compared with randomly selected genes. Specifically, the AUC scores of both the esophageal and stomach cancer using the up-regulated genes are greater than 0.99, which illustrates our proposed biomarkers have strong capability at differentiating the class of cancer patient samples from normal samples.

Discussion

Our investigation identified leading genes that are upregulated or downregulated in patients with GE cancers. We specifically focussed on those genes whose protein products can be detected in the plasma, as measured in the plasma proteome database. Thus, our investigation has a direct translational impact. The abnormal gene expression plays a pivotal role in tumor development and progression[5]. We noted that compared to normal tissue, BIRC5, CENPF, STMN1, APOC2, and HNRPC were the five most significantly upregulated genes in esophageal cancer. Furthermore, these genes were also overexpressed in stomach cancer.

The baculoviral IAP repeat containing 5 (BIRC5) gene, also known as survivin, is a member of the inhibitor of apoptosis (IAP) family, where it encodes regulatory proteins that prevent apoptotic cell death. Survivin localizes to the mitotic spindle and participates in regulating mitosis. In addition to GE cancers, it is highly expressed in

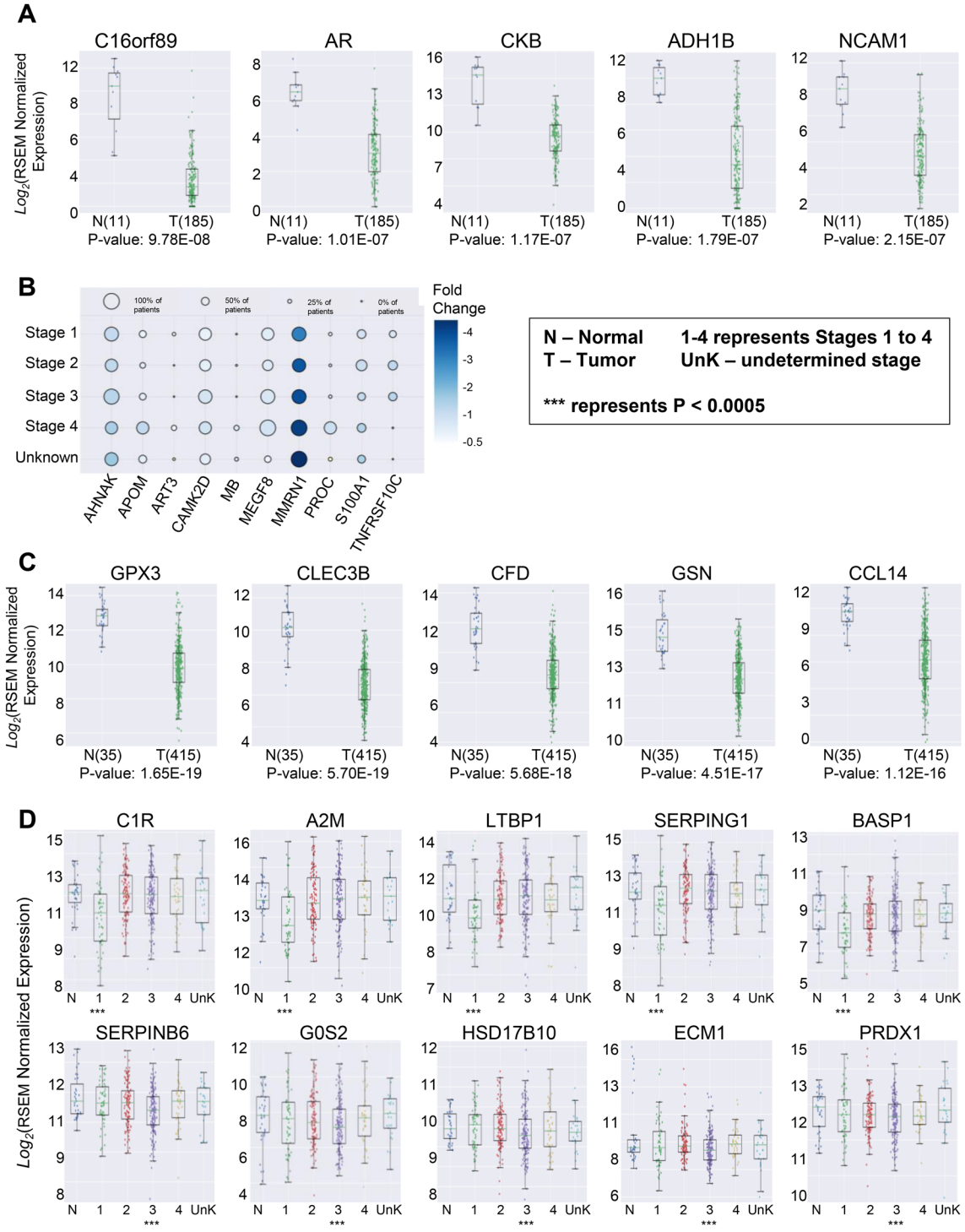

Figure 4. (**A**) Comparison between healthy control and patients with esophageal cancer showed that C16orf89, AR, CKB, ADH1B, and NCAM1 were the top five genes that were significantly down regulated in patients with esophageal cancer. (**B**) Stage-wise incremental downregulation of leading genes in patients with esophageal cancer. (**C**) Comparison between healthy control and patients with stomach cancer showed that: GPX3, CLEC3B, CFD, GSN, and CCL14 were the top five genes that were significantly down regulated in patients with stomach cancer. (**D**) The stage-based analysis showed a differential downregulation of various genes based on the stage of the disease. (**E**) Stage-wise incremental downregulation of leading genes in patients with stomach cancer. (**F**) Stage-wise downregulation of leading genes in patients with stomach cancer.

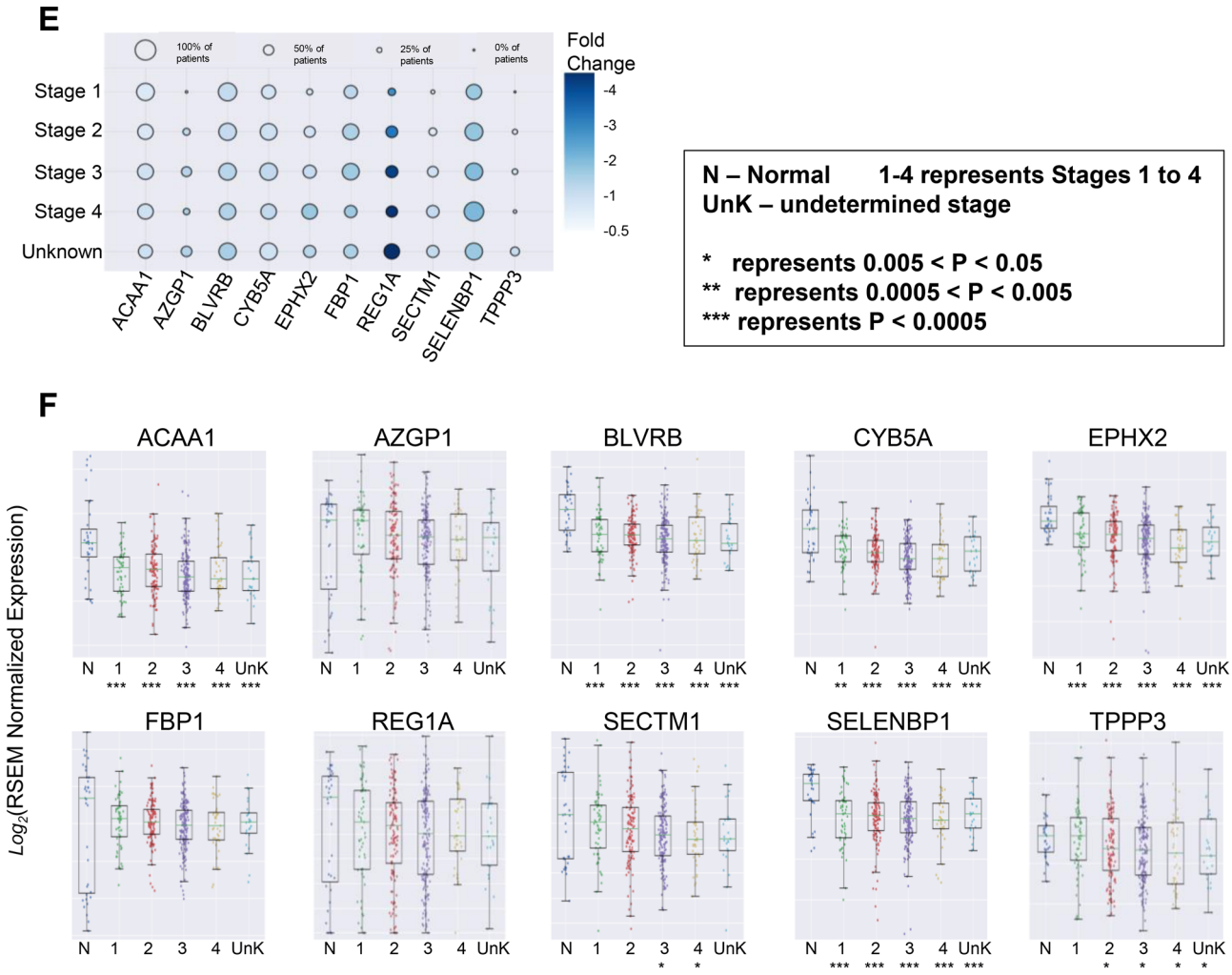

Figure 4. (continued)

various other malignancies and is associated with poor outcomes including a shorter survival period[19–21]. CENPF gene encodes centromere protein F that associates with the centromere–kinetochore complex. CENP-F protein is thought to be a cell cycle regulated protein that may play a role in chromosome segregation during mitosis. Interestingly, there is evidence that CENPF expression is associated with inferior outcomes in patients with esophageal cancer and patients with lower CENPF expression had a better survival rate compared with those with higher CENPF expression[22]. The CENPF gene is also amplified in other solid tumors including hepatocellular and breast cancers and correlates with patients' outcomes[23–25]. As cancer cells undergo active division, perhaps the up regulation of genes like BIRC5, CENPF could be a direct consequence of active mitosis.

STMN1 belongs to the stathmin family of genes and encodes a cytosoplasmic phosphoprotein stathmin 1. The encoded protein belongs to the family of microtubule-destabilizing proteins that control the assembly and disassembly of the mitotic spindle and thereby, regulate mitosis. Similar to esophageal cancer, STMN1 is highly expressed in various cancers, including leukemia, breast, prostate and lung cancer, and is a promising target for cancer therapy[26,27]. There is some evidence that it may have a prognostic significance in the early-stage gastric cancer. For example, a study that evaluated STMN1 role in both operable and advanced gastric cancers showed that in the operable cohort, STMN1 expression correlated with cancer recurrence, and resistance to adjuvant therapies[28].

In contrast to the relatively known roles of BIRC5, CENPF, and STMMN1 in malignancies, functions of APOC2 and HNRNPC genes in cancer cell are less well defined. The APOC2 gene encodes a lipid-binding protein that belongs to the apolipoprotein family and is a component of the very low-density lipoprotein. This protein activates the enzyme lipoprotein lipase, which hydrolyzes triglycerides. APOC2 mutations could cause hyperlipoproteinemia type IB, characterized by hypertriglyceridemia, xanthomas, and early atherosclerosis[29,30]. HNRNPC gene encodes a protein that belongs to the subfamily of ubiquitously expressed heterogeneous nuclear ribonucleoproteins (hnRNPs). The hnRNPs are RNA binding proteins and are associated with pre-mRNAs in the nucleus. These proteins are involved in pre-mRNA processing and other aspects of mRNA metabolism and

transport along with cell proliferation and differentiation[31]. However, functions of hnRNPs in tumorigenesis and cancer progression in solid and hematological malignancies are not well understood[32].

Our analysis showed that compared with normal stomach tissue, CST1, INHBA, ACAN, HSP90AB1, and HSPD1 were the leading five genes that were overexpressed in stomach cancer. These genes were also upregulated in patients with esophageal cancer. The CST1 gene encodes a secretory peptide called Cystatin SN, which is a cysteine proteinase inhibitor. Cysteine proteases are involved in tissue remodeling during development, and they support the migration of cancer cells. CST1 itself is known to promote proliferation, clone formation, and metastasis in breast cancer cells and high CST1 expression is negatively correlated with breast cancer survival[33]. CST1 has also been considered as a potential tumor marker in various epithelial malignancies[33,34]. Of note, a study involving patients with esophageal squamous cell carcinoma whose tumors express high levels of Cystatin SN showed favorable survival compared with those patients with low Cystatin SN expression[35]. Inhibin-βA (INHBA), a ligand belonging to the transforming growth factor-β superfamily, is associated with cell proliferation in cancer. INHBA is overexpressed in various types of cancers including esophageal and stomach tumors[36,37]. Overexpression of the INHBA gene is considered a useful independent predictor of outcomes in patients with gastric cancer after the curative surgery. High INHBA gene expression has shown to be associated with significantly poorer 5-year overall survival compared with low expression cases in patients with stomach cancer[37]. HSPD1 and HSP90AB1 belong to heath shock protein (HSP) group and encodechaperonin family proteins[38]. HSPD1 encodes a mitochondrial protein, which is important for assembly of imported proteins in the mitochondria and may function as a signaling molecule in the immune system. HSP90AB1 is thought to play a role in gastric apoptosis and inflammation. HSPs control a wide variety of signaling and cellular responses and have been classified into several subfamilies such as the HSP60s, HSP70s, HSP90s, and HSP100s[39]. HSP expression often correlates with patient prognosis in various malignancies[40–42] For example, HSP60 has been identified as an independent prognostic factor for both overall survival and recurrence-free survival in patients with early-stage stomach cancer[42]. The ACAN gene is a member of the aggrecan/versican proteoglycan family. The encoded protein is an integral part of the extracellular matrix in cartilaginous tissue, and it withstands compression in cartilage. Mutations in this gene may be involved in skeletal dysplasia and spinal degeneration, however, its role in cancer is not well understood.

With respect to downregulated genes, C16orf89, AR, CKB, ADH1B, and NCAM1 were the leading downregulated genes in patients with esophageal cancer. C16orf89 is predominantly expressed in the thyroid gland and is involved in the development and function of the thyroid[43]. Its role in tumorigenesis and progression has not been elucidated yet. The androgen-receptor (AR) gene encodes AR. Once AR binds its hormone ligand testosterone, it translocates into the nucleus, and stimulates transcription of androgen responsive genes[44]. In vitro evidence suggests a significant influence of sex hormones upon cancer growth[44,45]. For example, AR pathway plays an important role in the development of prostate cancer and various other epithelial malignancies including bladder, kidney, lung, breast, liver and ovary[45]. However, AR role in GE cancers development and progression is not known[46]. CKB or creatinine kinase B gene encodes a cytoplasmic enzyme that is involved in energy homeostasis. Its dysregulation could promote cancer invasiveness and progression[47]. Similar to AR, its disease modulating effect in GE cancers is unknown. ADHIB encodes alcohol dehydrogenase 1B enzyme. Evidence suggests that genetic polymorphisms of this enzyme has been associated with the increased risk of the aerodigestive cancer triggered by alcohol consumption[48]. The NCAM1 gene encodes a cell adhesion protein, a member of the immunoglobulin superfamily that is involved in both cell to cell and cell to matrix interactions. Its downregulation has been linked to cancer progression and development of metastases in gastrointestinal and other malignancies[49].

In patients with stomach cancers, GPX3, CLEC3B, CFD, GSN, and CCL14 were the leading five genes that were most significantly downregulated. GPX3 encodes glutathione peroxidase that belongs to a family of selenocysteine-containing redox enzymes that play important roles in cell signaling and immune modulation[50]. Consistent with our observation of its downregulation, promoter hypermethylation and downregulation of GPX3 in melanoma, stomach, head and neck, cervical and lung cancers suggest that GPX3 serves as a tumor suppressor in these cancers[50,51]. C-Type Lectin Domain Family 3 Member B (CLEC3B) is a member of the C-type lectin superfamily that encodes tetranectin. Dysregulation of CLEC3B has been reported in various epithelial cancers including stomach cancer[52,53]. Chen and others using TCGA database also noted downregulation of CLEC3B in stomach cancer. However, when they evaluated 328 patients with early-stage stomach cancer, high intratumoral tetranectin level was significantly associated with tumor invasion, lymph node metastasis, advanced TNM stage, and a shorter overall survival[53]. CFD or complement factor D encodes a serine protease that catalyze breakdown of factor B a rate limiting step of alternative pathway of complement activation. Impaired balance of complement activation could promote inflammation and tumorigenesis resulting in malignant cells proliferation, migration, invasiveness and metastasis[54,55]**.** GSN or Gelsolin gene encodes a protein that is involved in assembly and disassembly of actin filaments. Gelsolin has been attributed in prostate tumorigenesis and malignant transformation[56,57]. The C–C type chemokine 14 gene is known to induce targeted cell migration and is thought to play a role in carcinogenesis and metastasis of certain malignancies including breast cancer[58,59].

Aside from the leading upregulated and downregulated genes in patients with GE cancers, we also noted a stage-wise upregulation of several genes, such as APOC2, IL8, RNASE2, SERPINE1, and CETP, and stage-wise downregulation of other genes that play important role in and survival, including AHNAK, MEGF8, MMRN1, PROC, REG1A, SECTM1, TNFRSF10C, and TPPP3. The stage-related expression of these genes suggests their potential role in the disease progression and utility as monitoring markers or therapeutic targets. The cholesteryl ester transfer protein (CETP) for example maintains cholesterol homeostasis and has been identified as a potential target for estrogen positive breast cancer[60]. Conversely, AHNAK can act as a tumour suppressor gene and mediates the negative regulation of cell growth[61].

Furthermore, we also evaluated the predictive accuracy and the significance of the genes we identified. In recent years, deep neural networks have achieved enormous successes for such applications[62–64]. However, deep

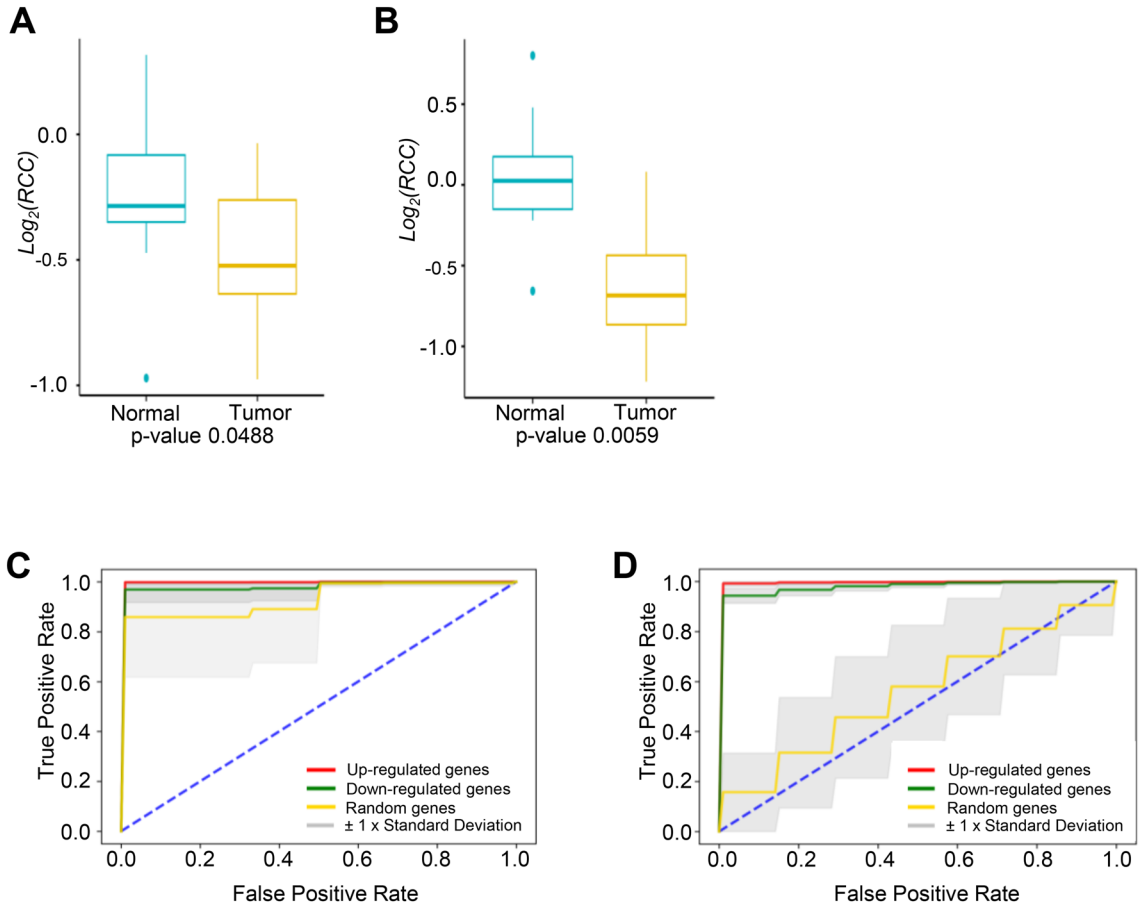

Figure 5. (**A**) The log ratio of closeness centrality (RCC) between normal and tumor samples for the proposed potential biomarkers in Esophageal cancer. (**B**) The log ratio of closeness centrality (RCC) between normal and tumor samples for the proposed potential biomarkers in Stomach cancer. (**C**) Comparison of ROC curves with features of the leading up-regulated genes, the leading down-regulated genes and the random selected genes for classification of tumor and normal samples in esophageal cancer. The dashed blue line in the diagonal presents the ROC curve of a random predictor, which has an AUC of 0.5 and can be used as the baseline to validate the effectiveness of our models. The mean AUC scores of the SVM models based on the up-regulated genes and the down-regulated genes are 0.9941 (standard deviation: 0.0031) and 0.9788 (standard deviation: 0.0265), respectively. The mean AUC of the comparison group, which uses the random selected genes, has the lowest score 0.9280 (standard deviation: 0.1137). (**D**) Comparison of ROC curves with features of the leading up-regulated genes, the leading down-regulated genes and the random selected genes for classification of tumor and normal samples in stomach. The dashed blue line in the diagonal presents the ROC curve of a random predictor, which has an AUC of 0.5 and can be used as the baseline to validate the effectiveness of our models. The mean AUC scores of the SVM models based on the up-regulated genes and the down-regulated genes are 0.9924 (standard deviation: 0.0038) and 0.9770 (standard deviation: 0.0114), respectively. The mean AUC of the comparison group, which uses the random selected genes, has the lowest score 0.5603 (standard deviation: 0.1664).

learning networks was not adopted in this study. This is primarily because, deep learning algorithms are non-linear and normally has millions of parameters[65]. Since the aim of our study is to identify key biomarkers for GE cancers, the explanation of model is crucial to evaluate the significance of genes. Contrarily, the classical machine learning models, such as linear models, provide a direct relationship between features and their prediction which makes it relatively straightforward to reason the decision mechanism of the model. Also, to avoid overfitting, more data are needed for the training of deep learning models. Moreover, the variations of the training data are necessary to construct a robust model. As the dataset sizes are not large in this study, the deep learning model will result in overfitting, if we apply deep learning to these datasets.

While our work provides a significant amount of novel information regarding the behavior of cancer-related molecules in GE cancers, it does not assess the level of gene expression based on the molecular classification of stomach and esophageal cancers[7]. Furthermore, we did not have information on histopathology of these cancer types and therefore, were not able to segregate the data based on histopathology. Finally, while we examine the up/down regulated genes, solely from the perspective of their differential expression, it will be interesting to investigate these candidates in cohorts of immunodeficient patients as it will provide additional knowledge on

how these candidates may promote adaptive alterations of host gut- and tissue-based microbiome[66]. In summary, the present study identified leading upregulated and downregulated genes in GE cancers. Since expression of the upregulated genes was minimal in both stomach and esophageal normal tissues, these genes have a strong potential to become diagnostic and therapeutic biomarkers for screening and early detection of cancer, recurrence following surgery and for anti-cancer therapies. Future studies will be required for validating diagnostic, therapeutic and prognostic importance of these genes. Our group plans to prospectively evaluate prognostic and predictive values of selected genes in a cohort of patients with metastatic gastroesophageal cancer who are treated with combination chemotherapy.

Received: 4 December 2020; Accepted: 15 March 2021
Published online: 07 April 2021

References

1. Bray, F. *et al.* Global Cancer Statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
2. Bang, Y. J. *et al.* Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial. *Lancet* **376**, 687 (2010).
3. Bouché, O. *et al.* Randomized multicenter phase II trial of a biweekly regimen of fluorouracil and leucovorin (LV5FU2), LV5FU2 plus cisplatin, or LV5FU2 plus irinotecan in patients with previously untreated metastatic gastric cancer: A Federation Francophone de Cancerologie Digestive Group Study—FFCD 9803. *J. Clin. Oncol.* **22**, 4319 (2004).
4. Ajani, J. A. *et al.* Clinical benefit with docetaxel plus fluorouracil and cisplatin compared with cisplatin and fluorouracil in a phase III trial of advanced gastric or gastroesophageal cancer adenocarcinoma: The V-325 Study Group. *J. Clin. Oncol.* **25**, 3205 (2007).
5. Li, M., Sun, Q. & Wang, X. Transcriptional landscape of human cancers. *Oncotarget* **8**, 34534–34551 (2017).
6. Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. *Nature* **12**, 169–175 (2017).
7. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
8. Chua, T. C. & Merret, N. D. Clinicopathologic factors associated with HER2-positive gastric cancer and its impact on survival outcomes—a systematic review. *Int. J. Cancer* **130**, 2845–2856 (2012).
9. Katona, B. W. & Rustgi, A. K. Gastric cancer genomics: Advances and future directions. *Cell Mol. Gastroenterol. Hepatol.* **3**, 211–217 (2017).
10. Liu, X. & Meltzer, S. J. Gastric cancer in the era of precision medicine. *Cell Mol. Gastroenterol. Hepatol.* **3**, 348–358 (2017).
11. National Cancer Institute GDC Data Portal. https://portal.gdc.cancer.gov/. Accessed 24 Apr 2019.
12. Nanjappa, V. *et al.* Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic Acids Res.* **42**(1), D959–D965 (2013).
13. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559. https://doi.org/10.1186/1471-2105-9-559 (2008).
14. Chen, J. *et al.* Genetic regulatory subnetworks and key regulating genes in rat hippocampus perturbed by prenatal malnutrition: Implications for major brain disorders. *Aging (Albany NY).* **12**(9), 8434–8458 (2020).
15. Li, H. *et al.* Co-expression network analysis identified hub genes critical to triglyceride and free fatty acid metabolism as key regulators of age-related vascular dysfunction in mice. *Aging (Albany NY).* **11**(18), 7620–7638 (2019).
16. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17. https://doi.org/10.2202/1544-6115.1128 (2005) (**(Epub 2005 Aug 12)**).
17. Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, 581–603. https://doi.org/10.1007/BF02289527 (1966).
18. Yu, H. *et al.* LEPR hypomethylation is significantly associated with gastric cancer in males. *Exp. Mol. Pathol.* **116**, 104493 (2020).
19. Zhang, S. *et al.* Prognostic role of survivin in patients with glioma. *Medicine* **97**, e0571 (2018).
20. Gu, Y. *et al.* Autophagy-related prognostic signature for breast cancer. *Mol. Carcinog.* **55**, 292–299 (2016).
21. Shang, X. *et al.* Downregulation of BIRC5 inhibits the migration and invasion of esophageal cancer cells by interacting with the PI3K/Akt signaling pathway. *Oncol. Lett.* **16**, 3373–3379 (2018).
22. Mi, Y. J. *et al.* Prognostic relevance and therapeutic implications of centromere protein F expression in patients with esophageal squamous cell carcinoma. *Dis. Esophagus.* **26**, 636–643 (2013).
23. Kim, H. E. *et al.* Frequent amplification of CENPF, GMNN and CDK13 genes in hepatocellular carcinomas. *PLoS One* **7**, e43223 (2012).
24. O'Brien, S. L. *et al.* CENP-F expression is associated with poor prognosis and chromosomal instability in patients with primary breast cancer. *Int. J. Cancer* **120**, 1434–1443 (2007).
25. Alli, E., Yang, J. M. & Hait, W. N. Silencing of stathmin induces tumor-suppressor function in breast cancer cell lines harboring mutant p53. *Oncogene* **26**, 1003–1012 (2007).
26. Bao, P. *et al.* High STMN1 expression is associated with cancer progression and chemo-resistance in lung squamous cell carcinoma. *Ann. Surg. Oncol.* **24**, 4017–4024 (2017).
27. Ghosh, R. *et al.* Increased expression and differential phosphorylation of stathmin may promote prostate cancer progression. *Prostate* **67**, 1038–1052 (2007).
28. Bai, T. *et al.* High STMN1 level is associated with chemo-resistance and poor prognosis in gastric cancerpatients. *Brit. J. Cancer* **116**, 1177–1185 (2017).
29. Surendran, R. P. *et al.* Mutations in LPL, APOC2, APOA5, GPIHBP1 and LMF1 in patients with severe hypertriglyceridaemia. *J. Intern. Med.* **272**, 185–196 (2012).
30. Wolska, A. *et al.* Apolipoprotein C-II: New findings related to genetics, biochemistry, and role in triglyceride metabolism. *Atherosclerosis* **267**, 49–60 (2017).
31. Hossain, M. N. *et al.* Downregulation of hnRNP C1/C2 by siRNA sensitizes HeLa cells to various stresses. *Mol. Cell. Biochem.* **296**, 151–157 (2007).
32. Park, Y. M. *et al.* Heterogeneous nuclear ribonucleoprotein C1/C2 controls the metastatic potential of glioblastoma by regulating PDCD4. *Mol. Cell Biol.* **32**, 4237–4244 (2012).
33. Dai, D. *et al.* elevated expression of CST1 promotes breast cancer progression and predicts a poor prognosis. *J. Mol. Med. (Berl).* **95**, 873–886 (2017).
34. Jiang, J. *et al.* Identification of cystatin SN as a novel biomarker for pancreatic cancer. *Tumour Biol.* **36**, 3903–3910 (2015).
35. Chen, Y.-F. *et al.* Overexpression of Cystatin SN positively affects survival of patients with surgically resected esophageal squamous cell carcinoma. *BMC Surg.* **13**, 15 (2013).

36. Seder, C. W. *et al.* INHBA overexpression promotes cell proliferation and may be epigenetically regulated in esophageal adenocarcinoma. *J. Thorac. Oncol.* **4**, 455–462 (2009).
37. Oshima, T. *et al.* Relation of INHBA gene expression to outcomes in gastric cancer after curative surgery. *Anticancer Res.* **34**, 2303–2310 (2014).
38. Bross, P. & Paula, F.-G. Disease-associated mutations in the *HSPD1* gene encoding the large subunit of the mitochondrial HSP60/HSP10 chaperonin complex. *Front Mol. Biosci.* **3**, 49 (2016).
39. Quintana, F. J. & Cohen, I. R. The HSP60 immune system network. *Trends Immunol.* **32**, 89–95 (2011).
40. Calderwood, S. K., Khaleque, M. A., Sawyer, D. B. & Ciocca, D. R. Heat shock proteins in cancer: Chaperones of tumorigenesis. *Trends Biochem. Sci.* **31**, 164–172 (2006).
41. Campanella, C. *et al.* Heat shock protein 60 levels in tissue and circulating exosomes in human large bowel cancer before and after ablative surgery. *Cancer* **121**(18), 3230–3239 (2015).
42. Li, X. S. *et al.* Heat shock protein 60 overexpression is associated with the progression and prognosis in gastric cancer. *PLoS One* **9**, e107507 (2014).
43. Afink, G. B. *et al.* Initial characterization of C16orf89, a novel thyroid-specific gene. *Thyroid* **20**, 811–821 (2010).
44. Brooke, G. N. & Bevan, C. L. The role of androgen receptor mutations in prostate cancer progression. *Curr. Genom.* **10**, 18–25 (2009).
45. Antonarakis, E. S. AR signaling in human malignancies: Prostate cancer and beyond. *Cancers (Basel)* **10**, E22 (2018).
46. Sukocheva, O. A. *et al.* Androgens and esophageal cancer: What do we know?. *World J. Gastroenterol.* **21**, 6146–6156 (2015).
47. Mello, A. A. *et al.* Deregulated expression of SRC, LYN and CKB kinases by DNA methylation and its potential role in gastric cancer invasiveness and metastasis. *PLoS One* **10**, e0140492 (2015).
48. Guo, H., Zhang, G. & Mai, R. Alcohol dehydrogenase-1B Arg47His polymorphism and upper aerodigestive tract cancer risk: A meta-analysis including 24,252 subjects. *Alcohol Clin. Exp. Res.* **36**, 272–278 (2012).
49. Crnic, I. *et al.* Loss of neural cell adhesion molecule induces tumor metastasis by up-regulating lymphangiogenesis. *Cancer Res.* **64**, 8630–8638 (2004).
50. Barrett, C. W. *et al.* Tumor suppressor function of the plasma glutathione peroxidase Gpx3 in colitis-associated carcinoma. *Cancer Res.* **73**, 1245–1255 (2013).
51. Zhao, H. *et al.* Silencing GPX3 expression promotes tumor metastasis in human thyroid cancer. *Curr. Protein Pept. Sci.* **16**, 316–321 (2015).
52. Liu, J. *et al.* CLEC3B is downregulated and inhibits proliferation in clear cell renal cell carcinoma. *Oncol. Rep.* **40**, 2023–2035 (2018).
53. Chen, H. *et al.* High intratumoral expression of tetranectin associates with poor prognosis of patients with gastric cancer after gastrectomy. *J. Cancer* **8**, 3623–3630 (2017).
54. Afshar-Kharghan, V. The role of the complement system in cancer. *J. Clin. Invest.* **127**, 780–789 (2017).
55. Reis, E. S. *et al.* Complement in cancer: Untangling an intricate relationship. *Nat. Rev. Immunol.* **18**, 5–18 (2018).
56. Chen, C. *et al.* Secreted gelsolin desensitizes and induces apoptosis of infiltrated lymphocytes in prostate cancer. *Oncotarget* **8**, 77152–77167 (2017).
57. Kim, J. C. *et al.* Opposite functions of GSN and OAS2 on colorectal cancer metastasis, mediating perineural and lymphovascular invasion, respectively. *PLoS One* **13**(8), e0202856 (2018).
58. Li, Q. *et al.* Binding of the JmjC demethylase JARID1B to LSD1/NuRD suppresses angiogenesis and metastasis in breast cancer cells by repressing chemokine CCL14. *Cancer Res.* **71**, 6899–6908 (2011).
59. Zhang, X. *et al.* TMEM88, CCL14 and CLEC3B as prognostic biomarkers for prognosis and palindromia of human hepatocellular carcinoma. *Tumour Biol.* **39**, 1–9 (2017).
60. Esau, L. *et al.* Identification of *CETP* as a molecular target for estrogen positive breast cancer cell death by cholesterol depleting agents. *Genes Cancer* **7**, 309–322 (2016).
61. Lee, I. H. *et al.* Ahnak functions as a tumor suppressor via modulation of TGFbeta/Smad signaling pathway. *Oncogene* **33**, 4675–4684 (2014).
62. Liu, M. *et al.* A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuroimage* **208**, 116459 (2020).
63. Liu, M., Cheng, D., Wang, K. & Wang, Y. Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics* **20**, 295–308 (2018).
64. Chen, J., Yang, L., Zhang, Y., Alber, M. & Chen, D. Z. Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation. *Adv. Neural Inf. Process. Syst.* **20**, 3036–3044 (2016).
65. Samek, W. *et al. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Vol 11700* (Springer, 2019).
66. Zheng, S. *et al.* Immunodeficiency promotes adaptive alterations of host gut microbiome: An observational metagenomic study in mice. *Front Microbiol.* **10**, 2415. https://doi.org/10.3389/fmicb.2019.02415.eCollection2019 (2019).

Acknowledgements

The study was supported by a research Grant provided by the Saskatchewan Cancer Agency (Operating Grant) to FJV and SA. F.S.V. is supported by funds from the College of Medicine, University of Saskatchewan. We thank Bjorn Haave for his assistance in the early stages of this analysis.

Author contributions

Conceptualization: F.S.V, A.F., F.J.V. and S.A.; methodology: F.S.V. and H.G.; writing—review and editing: F.S.V., H.G., L.D., A.Z., A.F., F.W., F.J.V. and S.A.; funding acquisition: F.J.V. and S.A.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-87037-w.

Correspondence and requests for materials should be addressed to F.J.V. or S.A.

Reprints and permissions information is available at www.nature.com/reprints.

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