A Network-Based Methodology to Identify Subnetwork Markers for Diagnosis and Prognosis of Colorectal Cancer

Olfat Al-Harazi1, Ibrahim H. Kaya2, Achraf El Allali3 and Dilek Colak1*

1Biostatistics, Epidemiology and Scientific Computing Department, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia, 2College of Medicine, Alfaisal University, Riyadh, Saudi Arabia, 3African Genome Center, Mohammed VI Polytechnic University, Benguerir, Morocco

The development of reliable methods for identification of robust biomarkers for complex diseases is critical for disease diagnosis and prognosis efforts. Integrating multi-omics data with protein-protein interaction (PPI) networks to investigate diseases may help better understand disease characteristics at the molecular level. In this study, we developed and tested a novel network-based method to detect subnetwork markers for patients with colorectal cancer (CRC). We performed an integrated omics analysis using whole-genome gene expression profiling and copy number alterations (CNAs) datasets followed by building a gene interaction network for the significantly altered genes. We then clustered the constructed gene network into subnetworks and assigned a score for each significant subnetwork. We developed a support vector machine (SVM) classifier using these scores as feature values and tested the methodology in independent CRC transcriptomic datasets. The network analysis resulted in 15 subnetwork markers that revealed several hub genes that may play a significant role in colorectal cancer, including PTP4A3, FGFR2, PTX3, AURKA, FEN1, INHBA, and YES1. The 15-subnetwork classifier displayed over 98 percent accuracy in detecting patients with CRC. In comparison to individual gene biomarkers, subnetwork markers based on integrated multi-omics and network analyses may lead to better disease classification, diagnosis, and prognosis.

Keywords: colorectal cancer, network, biomarker, omics, subnetwork, prognostic

INTRODUCTION

Artificial intelligence (AI) and Machine learning (ML) approaches have been widely used to investigate the disease diagnosis and predict the outcome (Maciukiewicz et al., 2018; Lai et al., 2019; Eicher et al., 2020; Jamal et al., 2020; Sanchez and Mackenzie, 2020; Sinkala et al., 2020; Stafford et al., 2020; Toraih et al., 2020). The integration of multiple high-throughput omics datasets, such as messenger RNA (mRNA) expression profiles, proteomics, copy number alterations (CNAs), methylation and others, may increase the robustness and reliability in identifying disease associated biomarkers (Colak et al., 2010; Colak et al., 2013; List et al., 2014; Al-Harazi et al., 2016; Colak et al., 2016; Aldosary et al., 2020; Eicher et al., 2020). A protein–protein interaction (PPI) network can be defined as a directed or undirected network that consists of vertices as proteins or genes and edges as the interactions among them (Wiredja and Bebek, 2017; Sanchez and Mackenzie,
2020). Interactions among proteins or genes are meant to be specific and biologically meaningful (Wiredja and Bebek, 2017; Sanchez and Mackenzie, 2020). It has been reported that network-based approaches have high efficacy in identifying biomarkers for numerous complex diseases, including cancers (Wang et al., 2017; Chen et al., 2019; Liu et al., 2019; Uddin et al., 2019; Khan et al., 2020; Van et al., 2020).

Traditional statistical approaches are not suitable for detecting gene interactions, especially when interactions appear between more than two genes, or when the data are high-dimensional, meaning the data have many attributes or independent variables (McKinney et al., 2006; Lai et al., 2019). Machine learning approaches have been widely used to identify disease biomarkers (Lim et al., 2019; Moni et al., 2019; Tabl et al., 2019; Sanchez and Mackenzie, 2020). Recently, Sanchez et al. identified methylation biomarkers for leukemia by investigating PPI for differentially methylated genes (DMGs) and differentially expressed genes (DEGs) using machine learning approach (Sanchez and Mackenzie, 2020). The authors reported that the identified biomarkers are reliable and associated with cancer development and risk (Sanchez and Mackenzie, 2020). Tabl et al. proposed a hierarchical machine learning system to develop biomarkers that can support the identification of the best therapy for breast cancer patients based on their gene expression and clinical data that achieved a high classification accuracy (Tabl et al., 2019). Furthermore, Sinkala et al. applied machine learning algorithms coupled with integrative profiling of multiple data types to identify biomarkers that can differentiate between pancreatic cancer subtypes (Sinkala et al., 2020).

When a specific gene/protein is related to a particular disease or biochemical process, its associated genes/proteins may also be involved in the same disease or biochemical process (Barabasi et al., 2011). Most interaction networks can be clustered into small connected subgraphs that are called disease modules or subnetworks (Barabasi et al., 2011). A disease subnetwork or module consists of linked genes or proteins that share mutations, biological processes or expression variations which can be related to a specific disease (Al-Harazi et al., 2016). Previous reports indicated that the development of disease-related subnetwork markers is a robust approach that can provide markers with higher accuracy in disease classification in comparison to using individual genes (Al-Harazi et al., 2016; Khunlertgit and Yoon, 2016; Al-Harazi et al., 2019). Indeed, network-based analysis of gene expression profiling was performed to identify subnetworks and hub genes that are associated with different cancer, including breast cancer (Khan et al., 2020), lung cancer (Huang et al., 2015), ovarian cancer (Zhang et al., 2013), and others and have demonstrated the significance of the method in discovering genes related to development and progression of cancer (28).

Colorectal cancer (CRC) is one of the most frequent cancers, with a high morbidity and mortality rate. In 2018, approximately 1.8 million new instances of CRC were diagnosed, and 881,000 deaths (Bray et al., 2018). Despite advances in screening and treatment strategies, the annual incidence and mortality rates of CRC are still increasing rapidly. Molecular studies have reported that CRC is a complex and molecularly heterogeneous disease (Hahn et al., 2016; Molinari et al., 2018; Murphy et al., 2019; Fanelli et al., 2020). Gene-expression profiling is widely used in developing prognostic and diagnostic signatures for colorectal cancer (Chen et al., 2016; Xu et al., 2017; Uddin et al., 2019; Zuo et al., 2019). However, because of the heterogeneity of CRC, minimum overlapping was observed in gene lists reported in previous studies (Cao et al., 2017).

In this study, we developed an integrated omics and network-based methodology to identify subnetwork markers for disease diagnosis and prognosis. We applied our methodology to develop subnetwork markers for CRC. We first performed integrated analysis of global gene expression and copy number data. We then constructed a PPI network for the identified DEGs using molecular interaction data from several databases, including Database of Interacting Proteins (DIP) (Salwinski et al., 2004), BioGRID (Chatr-Aryamontri et al., 2017), HPRD (Mishra et al., 2006), IntAct (Kerrien et al., 2007), BIND (Alfarano et al., 2005), and Molecular INTeraction database (MINT) (Licata et al., 2012). We calculated an activity score for each subnetwork and built a classifier using these scores as feature values. We finally validated diagnostic and prognostic potential of the identified network markers.

MATERIALS AND METHODS

Data Collection and Integrated Analysis

Whole-genome gene expression and CNA datasets for Saudi patients with colorectal cancer were gathered from the NCBI GEO (www.ncbi.nlm.nih.gov/geo). The whole-genome gene expression dataset (GSE23878) contains 35 colorectal cancer and 24 noncancerous matched samples (Uddin et al., 2011). All samples were probed using Affymetrix Human Genome U133 Plus 2.0 Array. The raw data were normalized using GC Robust Multi-array Average (GC-RMA) algorithm (Wu and Irizarry, 2004; Wu and Irizarry, 2005). The differentially expressed genes (DEGs) were identified using Analysis of Variance (ANOVA) with the adjustment of probability (p) values for multiple comparisons by false discovery rate (FDR) according to Benjamin-Hochberg step-up procedure (Benjamini and Hochberg, 1995). The DEGs were defined as those with absolute fold change (FC) > 2 and adjusted p-value < 5%.

The CNA dataset contains thirty samples (15 tumor and 15 adjacent normal samples) from Saudi patients (GSE47204) (Eldai et al., 2013). The data were generated using Affymetrix CytoScan HD arrays. The CNAs were identified as previously described in (Eldai et al., 2013) that revealed 144 genes with copy number changes (91 of which associated with CRC, that we included in our analysis). Next, we used the Venn diagram approach to identify the DEGs with copy number alterations in CRC using data from global mRNA and CNA. These genes were then used as input list or “seed genes” for building the PPI network (Martin et al., 2010).
Functional annotation and biological term enrichment analysis were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al., 2009) and Protein Analysis Through Evolutionary Relationships (PANTHER) (Mi et al., 2019). Figure 1A illustrates the framework for the integrated analysis. Statistical analyses were performed using PARTEK Genomics Suite (Partek Inc., St. Lois, MO, United States). All statistical tests were two-sided and $p$-value < 0.05 was considered statistically significant.

**Protein–Protein Interaction Network Construction and Subnetwork Identification**

We built the PPI network using BisoGenet, a Cytoscape plugin (Martin et al., 2010). BisoGenet imports the interaction data from several databases, including DIP (Salwinski et al., 2004), BioGRID (Chatr-Aryamontri et al., 2017), HPRD (Mishra et al., 2006), IntAct (Kerrien et al., 2007), BIND (Alfarano et al., 2005), and MINT (Licata et al., 2012). We input the set of seed genes (the DEGs with altered CNs) into the plugin which then builds the gene networks. The input list of genes (seed genes) are mapped to nodes, that will become initial set of network nodes (seed nodes) from which the network is expanded (Martin et al., 2010). The edges of the PPI network represent molecular interactions.

The constructed gene network is then clustered into subnetworks using another Cytoscape pluginclusterMaker” to cluster the network into subnetworks (Morris et al., 2011). The plugin provided the Markov Cluster Algorithm (MCL) (Van Dongen, 2001; Enright et al., 2002) that we used in our analysis. The MCL is a widely used method for analyzing complex biological networks. It uses a flow simulation to perform clustering of graphs by first building a matrix of values to be clustered that are stored in edge attributes. Then
MCL algorithm is performed iteratively. After constructing the discriminant subnetworks, we selected only the subnetworks that contained at least one gene from the seed genes and the number of nodes ≥3. We then performed functional annotation and biological term enrichment analyses of the identified subnetworks using DAVID (Huang et al., 2009) and PANTHER (Mi et al., 2019).

Scoring Subnetworks and Classification
We standardized the expression data for each gene across all samples using the z-transformation before calculating the subnetwork activity scores. We then calculated an activity score for each subnetwork as the average expression of up-regulated genes minus that of down-regulated genes in each sample. These scores were then used as feature values to build a classification model using GSE23878 dataset for training (n = 59). In order to assess the classifier’s performance, we used an independent microarray gene expression dataset for human colorectal cancer from The Cancer Genome Atlas (TCGA) database (TCGA data version 2016_01_28 for colorectal adenocarcinoma (COADREAD); https://gdac.broadinstitute.org/). The dataset contains 244 samples (222 tumor and 22 normal samples) performed on Agilent 244K Custom Gene Expression G4502A-07–3 arrays. We used level 3 preprocessed and normalized gene expression data as described in detail by the TCGA workgroup at the Broad Institute at the link above.

We tested the designed classifier based on the identified subnetwork markers by measuring its ability to differentiate patients from normal controls. The following measures were used for evaluating the performance:

\[
\begin{align*}
\text{Accuracy} &= \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{false negative} + \text{false positive} + \text{true negative}} \\
\text{Sensitivity} &= \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \\
\text{Specificity} &= \frac{\text{true negative}}{\text{false positive} + \text{true negative}}
\end{align*}
\]

Moreover, the area under the curve (AUC) with the 95% confidence interval (CI) and unsupervised principal component analyses (PCAs) are performed to further test the performance of subnetworks.

Furthermore, we compared the classification performance of the subnetwork markers with those of the previously reported CRC gene signatures as well as the DEGs identified in this study. Hence, we built several classifiers using four well-known CRC gene signatures and tested them on the same training and validation datasets. The ColoGuideEx is a gene expression classifier consisting of 13 genes designed for CRC patients at stage II (Ågesen et al., 2012). The second gene signature (ColoPrint) is 18-gene signature that is identified using whole-genome expression data and has been shown to predict high risk of recurrence in CRC patients with stage II or III (Tan and Tan, 2011). GeneFx is a 634 probe-set signature is a prognostic assay developed for patients with stage II colon cancer (Kennedy et al., 2011). The Oncotype DX, contains 12-gene signature, is also used for Stage-II CRC patients for recurrence risk prediction and guide therapy options after surgery (Clark-Langone et al., 2010).

Survival Analysis
We performed univariate and multivariate survival analyses using the Cox proportional hazard regression model on TCGA (COADREAD) to evaluate the prognostic value of the identified subnetwork markers and their relationships with overall survival of CRC patients. The multivariate Cox regression analysis was performed to examine whether the predictive ability of the subnetwork markers was independent of other clinical factors, including age, gender, pathologic stage, and lymphatic invasion.

The prognostic risk score for each patient is calculated as the sum of the product of subnetwork score with the corresponding regression coefficient in the multivariate Cox proportional hazard regression model analysis as follows:

\[
\text{Risk score} = \beta_{\text{subnetwork}_1} \times \text{SS1} + \beta_{\text{subnetwork}_2} \times \text{SS2} + \ldots + \beta_{\text{subnetwork}_n} \times \text{SSn}
\]

where SSi and \(\beta_{\text{subnetwork}_i}\) indicate the ith subnetwork score and the corresponding regression coefficient in the multivariate Cox proportional hazard analysis, respectively.

After calculating the risk scores, the median risk score is used to divide patients into high and low risk groups and Kaplan-Meier method is used to plot the survival curves. Significance between survival curves was calculated by the log-rank test. A \(p\)-value < 0.05 was considered statistically significant.

For further validation, we used an independent microarray dataset (GSE17537, n = 55) that included data from 55 CRC patients, downloaded from the NCBI GEO database and standardized using z-score transformation. We performed survival analysis using the same regression coefficients (\(\beta_s\)) that was calculated using the TCGA cohort.

RESULTS
Identification of Overlapping Colorectal Cancer Differentially Expressed Genes
We first analyzed global mRNA expression profile from CRC \((n = 35)\) and normal samples \((n = 24)\) and identified 1,366 DEGs (up- or down-regulated) in tumor compared to normal (adjusted \(p\) value < 5% and absolute fold change >2). We obtained 91 significant genes identified in CNA regions from (Eldai et al., 2013) and performed Venn diagram approach to identify overlapping significant mRNAs that have concomitant copy number alterations (Figure 1B). The integrated omics analysis revealed 24 significant DEGs. Functional analysis using PANTHER (Mi et al., 2013) revealed that these genes are related to protein phosphorylation, locomotion, system process, cell migration, and cell motility, that are known to be
Disease-Associated Subnetwork Markers

Several reports have demonstrated that subnetwork markers are more reliable and robust than single biomarker genes and achieved higher accuracy in disease classification (Al-Harazi et al., 2016; Al-Harazi et al., 2019). Here, we constructed gene interaction network using BisoGenet (Martin et al., 2010) for the DEGs with altered CN (seed genes; shared genes in the Venn diagram in Figure 1B). The PPI network for the seed genes had 797 nodes and 9,634 edges (Figure 1C). The PPI networks are necessary to almost all cell processes, therefore investigating PPIs is essential for understanding the physiological function of human cells in normal and disease states (Al-Harazi et al., 2016). The edges of the PPI network represent molecular interactions annotated in DIP (Salwinski et al., 2004), BioGRID (Chatr-Aryamontri et al., 2017), HPRD (Mishra et al., 2006), IntAct (Kerrien et al., 2007), BIND (Alfarano et al., 2005), and MINT (Alfarano et al., 2005) databases. The associated with cancer (Yamaguchi et al., 2005; Singh et al., 2017; Stuelten et al., 2018) (Supplementary Table S1).

### Table 1: Gene ontology enrichment analysis of 15 subnetworks.

| Biological Processes                      | FE*     | p-value  | Cellular Components                      | FE*     | p-value  | Molecular Functions                     | FE*     | p-value  |
|-------------------------------------------|---------|----------|------------------------------------------|---------|----------|-----------------------------------------|---------|----------|
| Sodium ion transport                      | 16.6    | 1.8E-05  | Node of Ranvier                          | 59.6    | 1.4E-06  | Protein binding                         | 1.3     | 4.9E-08  |
| Nuclear division                          | 4.0     | 4.8E-03  | Extracellular matrix                     | 4.4     | 2.0E-05  | Growth factor binding                   | 9.1     | 6.7E-05  |
| Signaling                                 | 2.0     | 5.8E-03  | Anchoring junction                       | 3.3     | 2.0E-04  | Sodium channel activity                 | 18.5    | 8.8E-05  |
| Cell cycle                                | 3.1     | 8.5E-03  | Cytoplasm                                | 1.3     | 3.4E-04  | Voltage-gated sodium channel activity   | 27.2    | 2.5E-04  |
| Cell communication                        | 1.9     | 1.0E-02  | Main axon                                | 12.3    | 3.9E-04  | Small molecule binding                  | 2.0     | 6.5E-04  |
| Cellular response to stimulus             | 1.8     | 1.1E-02  | Cell-cell contact zone                   | 11.0    | 5.9E-04  | Cytoskeletal protein binding            | 2.7     | 1.0E-03  |
| Mitotic cell cycle                        | 3.8     | 1.2E-02  | Cell junction                            | 2.1     | 1.1E-03  | ATP binding                             | 2.2     | 2.8E-03  |
| Biological regulation                     | 1.5     | 1.4E-02  | Early endosome                           | 3.8     | 2.6E-03  | Kinase binding                          | 2.7     | 4.2E-03  |
| Regulation of cell communication           | 2.4     | 1.8E-02  | Cell-cell junction                       | 3.3     | 3.0E-03  | Receptor ligand activity                | 3.0     | 1.0E-02  |
| Cell surface receptor signaling pathway    | 2.3     | 1.9E-02  | Cytoskeleton                             | 1.9     | 3.7E-03  | Signaling receptor activator activity   | 2.9     | 1.1E-02  |

*FE, Fold Enrichment is calculated by dividing the number of genes in 15 subnetworks implicated in each GO term by the expected number.
The constructed PPI network is then clustered using the MCL algorithm of clusterMaker app in Cytoscape that revealed 174 gene-clusters (subnetworks). We selected 15 subnetworks that contained at least one seed gene (DEG) and the number of nodes $\geq 3$ (Figure 2 and Supplementary Table S2).

The functional and gene ontology enrichment analyses revealed that these subnetworks are highly enriched in biological processes that are related to sodium ion transport, nuclear division, signaling, mitotic cell-cycle, biological regulation and cell communication (Figure 2; Table 1). The enriched cellular components include extracellular matrix, anchoring junction, and cytoplasm. Protein binding, growth factor binding, sodium channel activity, and voltage-gated sodium channel activity are the significantly enriched molecular functions among the 15 subnetwork markers (Table 1).

### Optimal Support Vector Machine Classification Model and Performance Comparison

We assessed the classification performance of the classifier that is designed using 15 subnetwork markers. The CRC/normal transcriptomic dataset (GSE23878) has been used as the input data for training the classifier. Expression values for each gene across all samples were normalized using z-transformation. A subnetwork activity score is then computed for each sample, as detailed in the Materials and methods section. We then designed an SVM classifier (Chang and Lin, 2011) using the 15 subnetwork scores as features to build the classification model. To evaluate the classifier’s performance, an independent microarray dataset from TCGA was used. The subnetwork markers achieved 98% accuracy, 98% sensitivity and 100% specificity, and 0.99 AUC (Table 2).

For comparison to other gene signatures, we designed classifiers for the 24-gene DEGs (DEGs with altered CN, in Figure 1B) and four well-known gene signatures for colorectal cancer, namely ColoGuideEx (Ågesen et al., 2012), ColoPrint (Tan and Tan, 2011), Genefx (Kennedy et al., 2011) and Oncotype DX (Clark-Langone et al., 2010), using the same training (GSE23878) dataset and tested each classifier’s performance on the TCGA dataset. The results demonstrated that the subnetwork markers outperformed the 24-gene DEGs and all tested gene signatures, except for the Genefx (634-probe set signature) that achieved the same performance with our subnetwork markers (Table 2). The 15 subnetwork markers and the tested

### Table 2

Disease classification results of SVM classifiers using 15 subnetwork markers and other known gene signatures.

|                  | Accuracy (%) | Sensitivity | Specificity | AUC  |
|------------------|--------------|-------------|-------------|------|
| 15 Subnetworks   | 98           | 0.98        | 1.00        | 0.99 |
| 24 DEGs*         | 97           | 0.94        | 1.00        | 0.97 |
| ColoGuideEx      | 84           | 0.83        | 1.00        | 0.91 |
| ColoPrint        | 84           | 0.82        | 1.00        | 0.91 |
| Genefx           | 98           | 0.98        | 1.00        | 0.99 |
| Oncotype DX      | 87           | 0.85        | 1.00        | 0.93 |

Abbreviations: AUC, Area Under Curve; DEG, Differentially expressed genes. "DEGs with copy number alterations identified in this study (Figure 1B). All classifiers for gene signatures, ColoGuideEx (59), ColoPrint (60), Genefx (61), and Oncotype DX (62) and 24 DEGs, are designed using GSE23878 dataset as training and TCGA dataset as validation.
gene signatures shared only two genes: \textit{INHBA} (Oncotype DX) and \textit{MACC1} (Genefx).

Furthermore, we performed unsupervised PCA to test the performance of the subnetwork markers on GSE23878 and TCGA datasets. The PCA scatter plots, in which each sphere denotes a sample in the dataset, clearly distinguished CRC patients from normal controls in both datasets (Figure 3).

**Prognostic Risk Score and Multivariate Analysis**

We performed univariate and multivariate survival analyses using the Cox proportional hazard regression model analysis. The 15 subnetwork marker-risk score for each patient in TCGA data is defined as:

\[
\text{Risk score} = \text{SS1}(-1.4) + \text{SS2}(-1.1) + \text{SS3}(-1.3) \\
+ \text{SS4}(-1.4) + \text{SS5}(-1.0) + \text{SS6}(1.2) \\
+ \text{SS7}(1.0) + \text{SS8}(-1.2) + \text{SS9}(1.1) \\
+ \text{SS10}(1.0) + \text{SS11}(-1.2) + \text{SS12}(1.0) \\
+ \text{SS13}(-1.2) + \text{SS14}(1.0) + \text{SS15}(-1.9)
\]

where \text{SSi} indicates the \text{i}th subnetwork score. The median risk score (−0.05) is used to divide the patients cohort into high and low risk groups.

The univariate Cox regression analysis revealed that three factors, the 15 subnetwork markers risk score (HR = 2.53, 95% CI = 1.29–4.99, \(p = 0.007\)), pathologic stage (HR = 3.45, 95% CI = 1.78–6.70, \(p = 0.0003\)) and lymphatic invasion (HR = 2.81, 95% CI = 1.39–5.68, \(p = 0.004\)) were significantly associated with
the CRC patients’ overall survival, but other factors did not exhibit any association with the survival (Table 3). In the multivariate analysis, the subnetwork markers showed prognostic significance for CRC overall survival risk (HR = 2.67, 95% CI = 1.35–5.30, p = 0.005). Hence, the multivariate Cox regression analysis revealed that the prognostic risk score based on the 15-subnetwork markers predicted the outcome of CRC independent of other clinical factors (Table 3). Indeed, Kaplan–Meier survival analysis displayed that patients in the high-risk group had a significantly poorer prognosis compared to low-risk group (p = 0.005) (Figure 4A). Furthermore, we used another independent dataset (GSE17537, n = 55) to perform overall survival analysis using the same risk score model with beta coefficients that were calculated on TCGA dataset, that also revealed significant survival differences between high and low risk groups (p = 0.04) (Figure 4B), confirming the prognostic significance of the 15-subnetwork markers.

DISCUSSION

Integration of biological information, especially from biological networks is considered an important step for achieving more robust, stable and interpretable biomarker signatures. Previous studies have indicated that integrating gene expression with CN data may lead to key cancer driver genes that are involved in tumor initiation and progression (Colak et al., 2010; Colak et al., 2013; Ohshima et al., 2017). Our integrated omics with the network-based analysis revealed potential subnetwork markers for CRC that may play an important role in tumorigenesis. The PTTPA3 gene, the hub gene in Subnetwork 1 (Figure 2) was previously identified as a metastasis biomarker for the colorectal cancer (Guzińska-Ustymowicz et al., 2011). A recent study indicated that frameshift mutation in ANK3 (hub gene in Subnetwork 4) in colon cancer (Yeon et al., 2018). PTX3 (hub gene in Subnetwork 5) is involved in immune system process and has been shown to be prognostic marker for CRC (Liu et al., 2018). In another study, FGFR2 (hub gene in Subnetwork 7) is reported to promote the PD-L1 expression via the JAK/STAT3 signaling in colorectal tumors and associated with disease progression and poor survival (Carter et al., 2017; Hu et al., 2019). Network analysis also indicated other cancer-associated genes, such as ITG3B4 and FGF10. ITG3B4 is considered to be a therapeutic target and prognosis marker for colon cancer (Li et al., 2019). The high expression of FGF10 is found to be correlated with the size of the CRC tumors, indicating its critical role in the prognosis and survival of colorectal cancer patients (Farajihaye Qazvini et al., 2019). Similarly, AURKA (Jacobsen et al., 2018), BAG3 (Li et al., 2018), NUBP1 (Liu et al., 2017), and ANLN (Wang et al., 2016) are all found to be dysregulated in colorectal cancer and involved in cancer progression and invasion. The subnetworks revealed genes that are previously reported as CRC-associated as well as several yet undeciphered genes that may contribute colorectal cancer, such as DNAAF5, RASL10B, DUSP19, and TTC27.

In conclusion, our results demonstrated that the subnetwork markers based on integrated omics (genomics and transcriptomics datasets) are robust as disease biomarkers and may lead to better disease diagnosis and prognosis compared to single genes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: The Cancer Genome Atlas (TCGA) and the NCBI Gene Expression Omnibus.
AUTHOR CONTRIBUTIONS

DC conception, design, and supervision. DC, OA, and IHK collected, analysed, interpreted the data, and drafted the manuscript. AE participated in interpretation and revising the manuscript. All authors read and approved the manuscript.

FUNDING

This study is funded by the Research Grant (RAC#2110006 to DC). We would like to extend our thanks and appreciation to our individual sponsor who generously donated to this research. The funder had no role in the study design and collection, analysis, and interpretation of the results.

REFERENCES

Ágesen, T. H., Sveen, A., Merok, M. A., Lind, G. E., Nesbakken, A., Skothem, R. I., et al. (2012). ColoGuideEx: a Robust Gene Classifier Specific for Stage II Colorectal Cancer Prognosis. Gut 61 (11), 1560–1567. doi:10.1136/gutjnl-2011-301179

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., et al. (2020). The Heterogeneous Clinical and Pathological Landscapes of Young and Old. Metabolites 10 (5), 202. doi:10.3390/metabo10050202

Chen, H., Sun, X., Ge, W., Qian, Y., Bai, R., and Zheng, S. (2016). A Seven-Gene Signature Predicts Overall Survival of Patients with Colorectal Cancer. Oncotarget 7 (56), 95054–95065. doi:10.18632/oncotarget.10982

Colak, D., Chishii, M. A., Al-Bakheet, A.-B., Al-Qahatani, S., Shoukri, M. M., Goyns, M. H., et al. (2010). Integrative and Comparative Genomics Analysis of Early Hepatocellular Carcinoma Differentiated from Liver Regeneration in Young and Old. Mol. Cancer 9, 146. doi:10.1186/1476-4598-9-146

Dai, G. P., Wang, L. P., Wen, Y. Q., Ren, X. Q., and Zuo, S. G. (2019). Identification of Biomarkers Associated with Diagnosis and Prognosis of Colorectal Cancer Patients Based on Integrated Bioinformatics Analysis. Gene 692, 119–125. doi:10.1016/j.gene.2019.01.001

Clark-Langone, K. M., Sangli, C., Krishnakumar, J., and Watson, D. (2010). Translating Tumor Biology into Personalized Treatment Planning: Analytical Performance Characteristics of the Oncotype DXColon Cancer Assay. BMC Cancer 10, 691. doi:10.1186/1471-2407-10-691

Colak, D., Al-Ali, A., Kaya, N., Muiya, N. P., Al-Harazi, O., Shinwari, Z., et al. (2016). Integrated Left Ventricular Global Transcriptome and Proteome Profiling in Human End-Stage Dilated Cardiomyopathy. PLoS One 11 (10), e0162669. doi:10.1371/journal.pone.0162669

Colak, D., Chishii, M. A., Al-Bakheet, A.-B., Al-Qahatani, S., Shoukri, M. M., Goyns, M. H., et al. (2010). Integrative and Comparative Genomics Analysis of Early Hepatocellular Carcinoma Differentiated from Liver Regeneration in Young and Old. Mol. Cancer 9, 146. doi:10.1186/1476-4598-9-146

Colak, D., Chishii, M. A., Al-Bakheet, A.-B., Al-Qahatani, S., Shoukri, M. M., Goyns, M. H., et al. (2010). Integrative and Comparative Genomics Analysis of Early Hepatocellular Carcinoma Differentiated from Liver Regeneration in Young and Old. Mol. Cancer 9, 146. doi:10.1186/1476-4598-9-146

Colak, D., Nofal, A., Albakheet, A., Nirmal, M., Jadidi, A., El Allali, A., et al. (2013). Novel Genes Associated with Colorectal Cancer Are Identified with MetaGenomic Analysis. PLoS One 8 (5), e63204. doi:10.1371/journal.pone.0063204

Dai, G. P., Wang, L. P., Wen, Y. Q., Ren, X. Q., and Zuo, S. G. (2020). Identification of Key Genes for Predicting Colorectal Cancer Prognosis by Integrated Bioinformatics Analysis: Oncol. Lett. 19 (1), 388–398. doi:10.3892/ol.2019.11068

Chen, H., Sun, X., Ge, W., Qian, Y., Bai, R., and Zheng, S. (2016). A Seven-Gene Signature Predicts Overall Survival of Patients with Colorectal Cancer. Oncotarget 7 (56), 95054–95065. doi:10.18632/oncotarget.10982

Colak, D., Chishii, M. A., Al-Bakheet, A.-B., Al-Qahatani, S., Shoukri, M. M., Goyns, M. H., et al. (2010). Integrative and Comparative Genomics Analysis of Early Hepatocellular Carcinoma Differentiated from Liver Regeneration in Young and Old. Mol. Cancer 9, 146. doi:10.1186/1476-4598-9-146

Colak, D., Nofal, A., Albakheet, A., Nirmal, M., Jadidi, A., El Allali, A., et al. (2013). Novel Genes Associated with Colorectal Cancer Are Identified with MetaGenomic Analysis. PLoS One 8 (5), e63204. doi:10.1371/journal.pone.0063204

Eicher, T., Kinnebrew, G., Patt, A., Spencer, K., Ying, K. M., et al. (2020). Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources. Metabolites 10 (5), 202. doi:10.3390/metabo10050202

Eldai, H., Periyasamy, S., Al Qarni, S., Al Rodaaiyan, M., Muhammed Mustafa, S., Deeb, A., et al. (2013). Novel Genes Associated with Colorectal Cancer Are Revealed by High Resolution Cytogenetic Analysis in a Patient Specific Manner. PLoS One 8 (10), e76251. doi:10.1371/journal.pone.0076251

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An Efficient Algorithm for Large-Scale Detection of Protein Families. Nucleic Acids Res. 30 (7), 1575–1584. doi:10.1093/nar/gkf170

Fanelli, G. N., Dal Pozzo, C. A., Depetris, I., Schirripa, M., Brignola, S., Biason, P., et al. (2020). The Heterogeneous Clinical and Pathological Landscapes of Metastatic Breast-Mutated Colorectal Cancer. Cancer Cell Int 20 (1), 30. doi:10.1186/s12935-020-1117-2

Farajihaye Qazvini, F., Samadi, N., Safiari, M., Emami-Razavi, A. N., and Shirkoohi, R. (2019). Fibroblast Growth Factor-10 and Epithelial-
Mesenchymal Transition in Colorectal Cancer. EXCLI J. 18, 530–539. doi:10.17179/excli2018-1784
Guzikta-Ustymowicz, K., Przybyczyn, A., Kemenia, A., and Ustymowicz, M. (2011). Immunohistochemical Assessment of PRL-3 (PTP4A3) Expression in Tumor Buds, Invasion Front, central Region of Tumor and Metastases of Colorectal Cancer. Adv. Med. Sci. 56 (1), 39–43. doi:10.2478/v10039-011-0015-1
Hahn, M. M., de Voer, R. M., Hoogerbrugge, N., Ligtheng, M. J. L., Kuiper, R. P., and van Kessel, A. G. (2016). The Genetic Heterogeneity of Colorectal Cancer Predisposition - Guidelines for Gene Discovery. Cell Oncol. 39 (6), 491–510. doi:10.3233/ccc-150268
Hu, L. P., Zhang, X. X., Jiang, H., Tao, L. Y., Li, Q., Zhu, L. L., et al. (2019). Targeting Purinergic Receptor P2Y2 Prevents the Growth of Pancreatic Ductal Adenocarcinoma by Inhibiting Cancer Cell Glycolysis. Clin. Cancer Res. 25 (4), 1318–1330. doi:10.1158/1078-0432.CCR-18-2297
Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. Nat. Protoc. 4 (1), 44–57. doi:10.1038/nprot.2008.211
Huang, H.-H., Liang, Y., and Liu, X. Y. (2015). Network-Based Logistic Classification with an Enhanced1,2S dover Reveals Biomarker and Subnetwork Signatures for Diagnosing Lung Cancer. BioMed. Res. Int. 2015, 1–7. doi:10.1155/2015/713993
Jacobsen, A., Bosch, L. J. W., Martens-de Kemp, S. R., Carvalho, B., Sillars-Hardebol, A. H., Dobson, R. J., et al. (2018). Aurora Kinase A (AURKA) Interaction with Wnt and Ras-MAPK Signalling Pathways in Colorectal Cancer. Sci. Rep. 8 (1), 7522. doi:10.1038/s41598-018-24982-z
Jamal, S., Khubai, M., Gangwar, R., Grover, S., Grover, A., and Hasnain, S. E. (2020). Artificial Intelligence and Machine Learning Based Prediction of Resistant and Susceptible Mutations in Mycobacterium tuberculosis. Sci. Rep. 10 (1), 5487. doi:10.1038/s41598-020-62368-2
Kennedy, R. D., Bylesjo, M., Kerr, P., Davison, T., Black, J. M., Kay, E. W., et al. (2018). GWAS-based Machine Learning Approach to Predict Duloxetine Response in Major Depressive Disorder. J. Psychi atr. Res. 99, 62–68. doi:10.1016/j.jpsychires.2017.12.009
Martin, A., Ochagavia, M. E., Rabasa, L. C., Miranda, J., Fernandez-de-Cossio, J., and Bringas, R. (2010). Bioisogenet: A New Tool for Gene Network Building, Visualization and Analysis. BMC Bioinformatics 11, 91. doi:10.1186/1471-2105-11-91
McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine Learning for Detecting Gene-Gene Interactions. Appl. Bioinformatics 5 (2), 77–88. doi:10.1016/S0082-2942(06)00502-00002
Mi, H., Muruganujan, A., Casasgrande, J. T., and Thomas, P. D. (2013). Large-scale Gene Function Analysis with the PANTHER Classification System. Nat. Protoc. 8 (8), 1551–1566. doi:10.1038/nprot.2013.092
Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). PANTHER Version 14: More Genomes, a New PANTHER GO-Slim and Improvements in Enrichment Analysis Tools. Nucleic Acids Res. 47 (D1), D419–D426. doi:10.1093/nar/gky1038
Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, B., Bala, P., et al. (2006). Human Protein Reference Database–2006 Update. Nucleic Acids Res. 34 (Database issue), D411–D414. doi:10.1093/nar/gkj141
Molinar, C., Marisi, G., Passardi, A., Matteucci, L., De Maio, G., and Ulivi, P. (2018). Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine? Jms 19 (12), 3733. doi:10.3390/jms19123733
Moni, M. A., Islam, M. B., Rahman, M. R., Ashraf-AL-Mahfuz, M., Awal, M. A., Islam, S. M. S., et al. (2019). Network-based Computational Approach to Identify Delineating Common Cell Pathways Influencing Type 2 Diabetes and Diseases of Bone and Joints. IEEE Access 8, 1486–1497. doi:10.1109/ACCESS.2019.2962991
Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., et al. (2011). clusterMaker: a Multi-Algorithm Clustering Plugin for Cytoscape. BMC Bioinformatics 12, 436. doi:10.1186/1471-2105-12-436
Murphy, N., Ward, H. A., Jenab, M., Rothwell, J. A., Boutron-Ruault, M.-C., Carbonnel, F., et al. (2019). Heterogeneity of Colorectal Cancer Risk Factors by Analysing Subsites in 10 European Countries: A Multinational Cohort Study. Clin. Gastroenterol. Hepatol. 17 (7), 1323–1331. e6. doi:10.1016/j.cgh.2018.07.030
Nair, J., Ghatge, M., Kakkar, V. V., and Shanker, J. (2014). Network Analysis of Inflammatory Genes and Their Transcriptional Regulators in Coronary Artery Disease. PloS one 9 (4), e94328. doi:10.1371/journal.pone.0094328
Ohshima, K., Hatakeyama, K., Nagashima, T., Watanabe, Y., Kanto, K., Doi, Y., et al. (2017). Integrated Analysis of Gene Expression and Copy Number Identified Potential Cancer Driver Genes with Amplification-dependent Overexpression in 1,454 Solid Tumors. Sci. Rep. 7 (1), 641. doi:10.1038/s41598-017-00219-3
Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 Update. Nucleic Acids Res. 32 (Database issue), 449D–451D. doi:10.1093/nar/gkh086
