# Integrated Analysis of Gene Expression Data for Colon Cancer Biomarker Discovery

Aamir Hassan, Masood UH. Zaka, Demetres Kouvatsos, Yonghong Peng
Department of Computing, University of Bradford, West Yorkshire, BD7 1DP, United Kingdom

**Abstract— The identification of molecular markers with prognostic value in colorectal cancer is a challenging task and vital for future therapeutic guidelines. Despite recent advances in the screening, diagnosis, and treatment of colorectal cancer, an estimated 608,000 people die every year from colon cancer, which is 8% of all cancer deaths. We performed two staged integrated bioinformatics analytics on gene expression data sets of three latest developed studies of colon cancer. We identified two groups of integrated signatures from the comparison of normal versus tumor and tumor versus mets patients samples. Functional analysis of 267-gene diagnostic signature shows over-represented signaling-related molecules and other significantly cancers related regulatory pathways. The metastatic 124-gene signature shows functionally involved in immune-response, lipid metabolism and PPAR signaling pathways. Kaplan-Meier estimates of 124-gene signature using independent data sets shows that higher grade/stage patients shows significantly better overall-survival (p=0.001, HR=2.61(CI 1.43-4.79)) and disease-specific survival rate (p=0.00, HR=2.41(CI 1.28-4.53)) compare to low grade patients. Further biological validation of genes identified in this study may provide vital biomarker targets for colon cancers.**

## I. INTRODUCTION

Colon Cancer (CC) is a common malignancy affecting both women and men. In 2012, it become the fourth most commonly diagnosed cancer (after prostate, breast, and lung cancer) with an estimated 103,170 new cases every year and, combined with rectal cancer, is the second most common cause of cancer deaths with 51,690 deaths (Marisa, et al., 2013; Siegel, et al., 2012). Despite recent advances in the screening, diagnosis, and treatment of colorectal cancer, an estimated 608,000 people die every year from this form of cancer, which is 8% of all cancer deaths. Pathological staging is the only prognostic classification used in clinical practice to select patients for adjuvant chemotherapy. However, pathological staging fails to predict recurrence accurately in many patients undergoing curative surgery for localized CRC. In fact, 10%-20% of patients with stage II CRC, and 30% - 40% of those with stage III CRC, develop recurrence (Zhang, et al., 2001).

Among the molecular markers that have been extensively investigated for colon cancer (CC) characterization and prognosis. DNA mismatch repair (MMR) system, is the only marker that was reproducibly found to be a significant prognostic factor in early CRC in both a meta-analysis and a prospective trial (Nannini, et al., 2009; Zhang, et al., 2001). Precise classification of tumor is critically important for cancer diagnosis and treatment. During the past decade, efforts have been made to use gene expression profiles to improve the precision of classification, with limited success (Cardoso, et al., 2007). Many studies have exploited the use of microarray technology to investigate gene expression profiles (GEPs) for the diagnosis of colon cancer in recent years, but no signature has been to be useful for clinical practice, especially for predicting prognosis (Sagynaliev, et al., 2005). It is shown that the reproducibility of GEP studies on colon cancer has not been sufficient for clinic practice, possibly because colon cancer cells are composed of distinct molecular entities that may develop through multiple pathways (Chan, et al., 2008; Shih, et al., 2005). Therefore, there may be several prognostic signatures for CRC, each corresponding to a different entity.

Indeed, GEP studies, based on integrated analysis of genetic/epigenetic data including high-throughput methylome data (Nannini, et al., 2009), have identified at least three distinct molecular subtypes of colon cancer. Therefore, colon cancer should no longer be considered as a homogeneous entity. However, the molecular classification of CC currently used, which is based on a few common DNA markers (MSI, CpG island methylator phenotype [CIMP], chromosomal instability [CIN], and BRAF and KRAS mutations) (Jass, 2007; Kang, 2011), needs to be refined, and a standard and reproducible molecular classification is still not available.

In order for identifying more robust diagnostic gene signature of colon cancer, this paper presents an investigated analysis of multiple latest competitive studies based on various stages of colon cancer. We applied tissue-based differential expression followed by supervised machine learning approach for the discovery of diagnostic/prognostic gene signatures for the earlier and outcome identification of patients with colon cancers. We identified 124-gene signature that can discriminate between the patients with good and poor outcomes, also provide evidences of functionally involved in immune response, lipid metabolism and PPAR signaling pathways.

## II. METHODS

We performed two staged integrated analysis on the expression data sets on three lately developed studies based on gene expression analysis of colon cancer for the discovery of potential gene signature. In order to extract the biological information, we further performed gene ontology enrichment analysis in order to identify the functional pathways involved in localised colon cancer as well as spread to other tissues. We searched studies involving applications of gene expression profiling on patients involving samples primary and metastatic tumor tissues.

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.85

TABLE I. SUMMARY COHORT OF STUDIES INVOLVING COLON CANCER.

| Author | Studies | Cohort | Sample Size | Platform | Dataset |
|---|---|---|---|---|---|
| Musella et al. (2013) | Time course analysis of colon cancer samples | Normal vs Tumor | N=88, T=84 | GPL6947 Illumina HumanHT-12 V3.0 expression beadchip | GSE37182 |
| Shaffer et al. (2009) | Expression data from colorectal cancer patients | Normal vs Tumor vs Mets | N=54, T=186, M=67 | Affymetrix Human Genome U133A Array | GSE41258 |
| Agesen et al. (2013) | Specific extracellular matrix remodeling signature of colon hepatic metastases | Normal vs Tumor vs Mets | N=18, T=20, M=19 | Affymetrix Human Genome U133A Array | GSE49355 |

### A. Data Collection

The three microarray expression data sets were obtained using GEOquery Bioconductor R package (Davis, 2013) from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) (Table I). The expression data sets involve samples from normal, primary tumor and metastatic tissue samples. In order to identify tissue specific mRNA signatures, we performed comparisons of among three tissues to identify specific dysregulated mRNAs.

### B. Stistical Analysis

Each of the extracted raw expression data sets were log-transformed and normalised by quantile method individually. Using R/Bioconductor, linear models for microarray data analysis were employed by forming contrast matrix comparisons for normal vs primary tumor, normal vs mets and primary tumor vs mets. Significance value (P-Value < 0.05) and log scale (logFC > 1 | logFC < -1) was used to rank the genes of interest. Corrections for multiple comparison was done using false discovery rate (FDR) method. NCBI's original genome annotation was used to obtain gene symbols for probe sets id's.

### C. Functional Analysis

We applied gene ontology enrichment analysis for the interpretation of gene signatures in order to identify potential biological processes, functional network and pathways. For this purpose, we applied Functional Annotation Tool of DAVID Bioinformatics Resources 6.7 (Huang, et al., 2009) using default settings. We used gene symbols as input gene list for each derived signature by selecting the Homo sapiens as their population background. We used (P-value < 0.05) as a cut-off value for the selection of DAVID terms.

### D. Classification Performance Evaluation

We applied supervised machine learning approach in order to study the reliability and robustness of inter-study signature. We estimated classification performance on each expression data sets by building a classifier using signature genes, as a feature vector, and their corresponding expression data from (Musella et al. (2013), Shaffer et al. (2009), Agesen et al. (2013)) using the cross-validation loop on support vector machine (SVM) (Mukherjee, et al., 1999). We used standard leave one-out cross-validation (LOOCV) to estimate the accuracy of above classifier. Hence, for every sample xn in the training set S, we train the classifier by leaving one sample (N-1) and then classifying the left out sample to predict the label of xn.

### E. Survival Analysis

We performed prognostics analysis for the 124-genes signature derived from the comparison of tumor versus mets tissues. For this purpose, we used independent data set (GSE17538) from the study conducted by Smith et al. (Smith, et al., 2010) derived from metastatic colon cancer. We tested 124-genes association with the clinical endpoints such as Overall survival (OS), Disease-specific survival (DSS) and Disease-free survival (DFS) across all the grades (grade 1, 2 and 3) by building Cox proportional hazard (PH) model. We build classifiers using genes from signature genes and their corresponding values from training set for the calculation of Wald score for each of the gene in classifier. Log-rank test P-Value were computed for both univariate and multivariate Cox model for OS, DSS and DFS. Similarly Kaplan-Meier estimates were calculated for each endpoint.

### F. Validations

For the validation of this studying findings, we applied two approaches: first we validated our proposed signature using in-silico cross validation followed by the literature search of signature genes from previously published studies and curated cancer signature databases. Therefore, we searched gene signature database GeneSigDB (http://compbio.dfci.harvard.edu/) for the potential overlaps between the proposed signatures and previously published signatures.

## III. RESULTS

### A. Gene Expression Analysis and Microarray Data Integration

We performed differential expression analysis on each of the study by comparing expression profiles of normal, tumor and metastatic tissues samples. We employed t-test statistics among the contrast matrix comparisons for the identification of dysregulated genes.

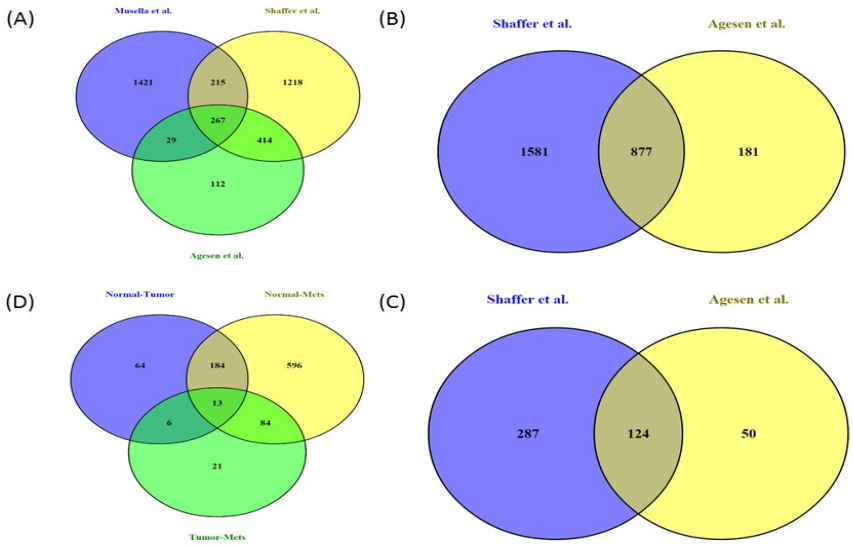


Fig. 1. Summary results of comparison between normal, tumor and mets tissues. (A) Venn diagram showing the common and unique genes of resulted gene lists after comparison between normal versus tumor samples. (B) Venn showing normal versus mets comparison (C) Venn showing results from tumor versus mets samples. (D) Venn diagram representing differentially expressed genes among three comparisons

### *B. Expression in Normal and Primary Tumor Colon Tissues*

To investigate the difference in human colon cancer, we performed differential expression using normal versus tumor samples of each data sets. We identified 2358 dysregulated (2144 up-regulated and 214 down-regulated) genes consisting of 88 normal and 84 primary tumor samples of Musella et al. study. Similarly, we observed 2696 genes (724 up-regulated and 1972 down-regulated) and 1050 genes (366 up-regulated and 684 down-regulated) from Agesen et al. and Shaffer et al. studies, respectively. We excluded all those probes set id's with no gene symbols for further analysis.

Following the identification of dysregulated gene lists from the comparison of normal versus primary tumor tissues classes of each eligible studies, we combined the resulted gene lists to form an inter-study signature gene set. In this case, we observed 267 genes were common to three gene lists of normal versus primary tumor comparison of three investigated data sets (Fig 1).

The gene ontology functional analysis of 267-genes shows over-representation of signaling-related molecules in processes and networks (Fig 2). We also identified pathways significantly (P-value < 0.05) involved in various cancers such as bladder cancer and acute myeloid leukemia. Known signaling and metabolic pathways also featured among the top ten regulatory identified pathways (Table II).

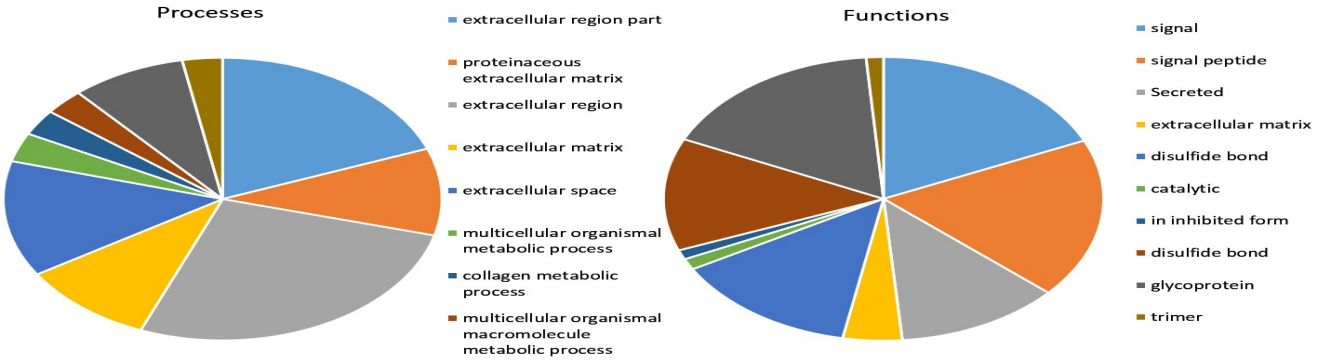


Fig. 2. Top-ten processes and molecular functions.

TABLE II. TOP 10 FUNCTIONAL REGULATORY PATHWAYS.

| Pathways | P-Value | Genes |
|---|---|---|
| Focal adhesion | 6.82E-05 | CAV1, MET, FLNC, COL5A2, COL5A1, PRKCB, MYL9, CCND1, VEGFA, COL1A2, COL1A1, COL11A1, THBS2, MYLK, SPP1 |
| ECM-receptor interaction | 8.12E-03 | COL1A2, COL1A1, COL5A2, THBS2, COL11A1, COL5A1, SPP1 |
| Bladder cancer | 1.12E-02 | CCND1, MMP9, VEGFA, MYC, MMP1 |
| Nitrogen metabolism | 1.18E-02 | CA7, CA4, CA2, CA1 |
| Complement and coagulation cascades | 1.45E-02 | C7, F12, CFB, SERPINE1, CFD, PLAU |
| Cell cycle | 1.57E-02 | CDK1, CCND1, E2F5, BUB1, MCM4, MYC, CDC25A, CDC25B |
| Vascular smooth muscle contraction | 2.99E-02 | KCNMA1, ACTG2, MYH11, KCNMB1, MYLK, PRKCB, MYL9 |
| Acute myeloid leukemia | 3.30E-02 | CCND1, LEF1, ZBTB16, RUNX1, MYC |
| Leukocyte transendothelial migration | 3.73E-02 | CLDN8, MMP9, CLDN1, CXCL12, PRKCB, THY1, MYL9 |
| Wnt signaling pathway | 3.90E-02 | WNT5A, CCND1, SFRP1, MMP7, CHP2, LEF1, MYC, PRKCB |

In the second step of bioinformatics analytics, we investigated reliability and robustness of proposed 264-gene signature using each of the expression data sets generated from different platforms (Table III). The 264-gene signature consistently achieved high classification accuracy ratios across all the data sets, classifying with 100%, 93.84% and 77.77 %, respectively.

TABLE III. LEAVE ONE OUT CROSS-VALIDATION CLASSIFICATION OF NORMAL VERSUS TUMOR SIGNATURE (267-GENES).

| Author | Expression Set | No. of Samples | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|---|
| Musella et al. (2013) | GSE37182 | N=88, T=84 | 100 | 100 | 100 |
| Shaffer et al. (2009) | GSE41258 | N=54, T=186, M=67 | 93.84 | 92.82 | 100 |
| Agesen et al. (2013) | GSE49355 | N=18, T=20, M=19 | 77.77 | 81.7 | 96.6 |

## C. Expression in Mets Tissues

Similarly, we carried out differential expression analysis for subset of Shaffer et al. data set consist of 54 normal and 67 mets samples, and identified 1328 genes were significantly dysregulated.

TABLE IV. LEAVE ONE OUT CROSS-VALIDATION CLASSIFICATION OF NORMAL VERSUS METS SIGNATURE (877-GENES).

| Author | Expression Set | No. of Samples | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|---|
| Musella et al. (2013) | GSE37182 | N=88, T=84 | 100 | 100 | 100 |
| Shaffer et al. (2009) | GSE41258 | N=54, T=186, M=67 | 93.39 | 96.82 | 100 |
| Agesen et al. (2013) | GSE49355 | N=18, T=20, M=19 | 93.33 | 93.17 | 96.6 |

Among the total identified 1328 genes, 1310 have shown over-expression whereas 18 gene were under-expressed. Likewise, analysis using subset of Musella et al. study consists of 18 normal and 19 metastatic samples have shown deregulation of 3122 genes, mostly 3098 showing over-expression (logFC > 1). We also focused our attention towards comparison of differentially expressed profiles among the tumor versus mets tissues comparison and identified 124 genes were common between the resulted gene lists. We observed number of notable genes were previously linked with metastases in colon cancer.

TABLE V. LEAVE ONE OUT CROSS-VALIDATION CLASSIFICATION OF TUMOR VERSUS METS SIGNATURE (124-GENES).

| Author | Expression Set | No. of Samples | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|---|
| Musella et al. (2013) | GSE37182 | N=88, T=84 | 100 | 100 | 100 |
| Shaffer et al. (2009) | GSE41258 | N=54, T=186, M=67 | 72.96 | 72.82 | 95.53 |
| Agesen et al. (2013) | GSE49355 | N=18, T=20, M=19 | 76.86 | 80.39 | 96.6 |

## D. 124-gene metastatic signature identified patients associated with poor outcome in independent data set

An independent human colon cancer gene expression and clinical data set was used to test the ability of 124-gene signature that discriminate between patient associated with cancer reoccurrence, overall survival and disease-specific survival. 238 patients with histopathological properties of age, gender, ethnicity, stage and grade were available for analysis. We observed, patients with higher grade (grade 3) across all the grades in independent set has significantly better OS (p=0.001, HR=2.61(CI 1.43-4.79); p=0.16, HR=1.42(CI 0.86-2.35), respectively) and DSS (p=0.00, HR=2.41(CI 1.28-4.53); p=0.25, HR=1.35(CI 0.80-2.28) compare to low grade patients.

Similarly, we determine the relative risk of reoccurrence and cancer related deaths. We observed a significant association of 124-gene signature with the risk of reoccurrence when analysed across all the tumor grades (p=0.0003, HR=1.74(CI 1.28-2.37)). We also analysed that the relative risk of reoccurrence has increase with the increase of tumor grade in patient samples (grade 3 (p=0.0005, HR=2.94(CI 1.59-5.46))) (Fig 3).

## E. The Cancer-focused Genes

We further analysed identified signatures by performing meta-analysis among inter-study signature derived from individual comparisons of normal versus tumor, normal versus mets and tumor versus mets tissues by comparing the similarities between them. We observed overlapping of 184 genes between normal versus tumor and normal versus mets gene lists. However, 64 genes of normal versus tumor comparison represents a strict tumor-specific (those genes which are not significantly dysregulated in other comparison) pool for which the functional analysis identified their targeted pathways involved: -cell cycle, acute myeloid leukemia, progesterone-mediated oocyte maturation,

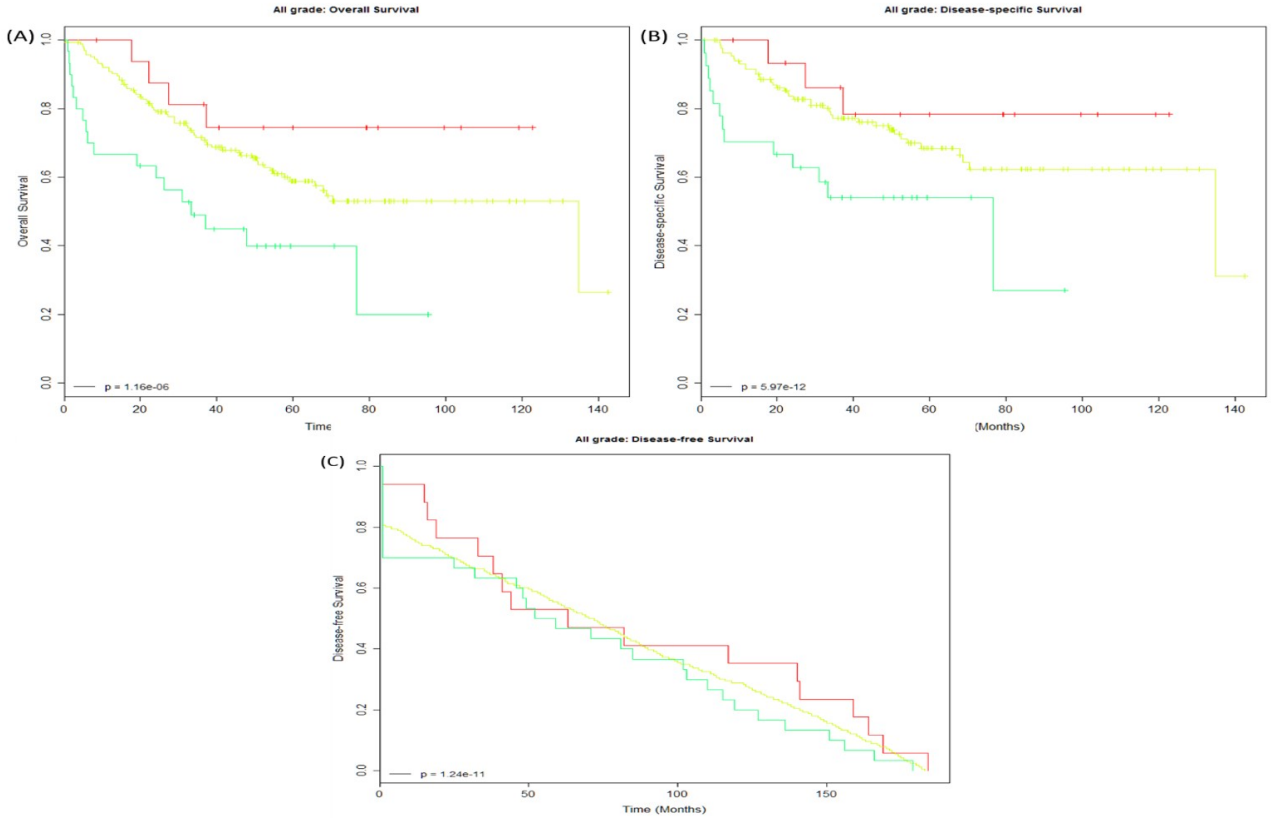


Fig. 3. The 124-gene classifier as tested in the independent data set across all grades. Kaplan–Meier estimates of overall and disease-specific survival in test set. Expression data for probes corresponding to the 124-gene recurrence classifier were used to build the Cox proportional hazard model from patient data in the Vanderbilt dataset. Plots represent survival analyses in the independent patient data set (A) Overall survival, (B) disease-specific survival analyses and (C) disease-free survival

-TGF-beta signaling pathway, role on ran in mitotic spindle regulation and G1-phase progression by myc

Similarly, in normal versus mets and tumor versus mets, a total of 701 genes were differentially expressed among met tissues. acute myeloid leukemia, progesterone- mediated oocyte maturation, TGF-beta signaling pathway, role on ran in mitotic spindle regulation and G1-phase progression by myc. We tracked the significant pathways involved: immune response, lipid metabolism and PPAR signaling pathway.

## IV. Discussion

The purpose of present study is to propose possible marker gene sets for colorectal cancer by using a two-step bioinformatical analytics. We performed meta-analysis using publically available GEO expression profiles of normal, tumor and metastatic tissues for the discovery of robust signature involving pathogenesis of colon cancer.

We identified cross-study 267-gene signature from the comparison of normal versus primary tumor samples across all the data sets that may be vital for the diagnosis for colon cancer. The functional analysis of 267-genes have revealed the involvement of cell cycle, cell-signaling and metabolic regulated pathways as reported in the previous studies (Moreno and Sanz-Pamplona, 2015; Planutis, et al., 2014). We further tested the robustness of gene signature using cross-validation, which shows excellent 90.53% overall-average accuracy-rate across all the expression data sets. We also observed the Agesen et al. expression validated with higher error-rate compare to other two data sets in the validation cohort. A close examination of the cohort present possible to explanation of these results; Agesen et al. study samples include stage IV tissues whereas Musella et al. and Shaffer et al. samples were derived from slightly earlier stages I-IV. However, we cannot rule out these variations are due to difference in sample size and/or platform differences.

For the mets tissues analysis, we identified two gene sets of deregulated genes from the comparison of normal, tumor and mets tissues. Further analysis of 124-genes deregulated among tumor versus mets tissues have shown involvement in key regulatory pathways such as complement cascade, formation of fibrin clot, extracellular matrix organization, collagen degradation and lipoprotein metabolism. Survival analysis of 124-gene signature using independent data sets have separated patients with high grades from lower grades when analysed for overall survival rate and disease-specific survival. The ability of 124-gene signature to discriminate between patient outcomes may be useful in patient prognosis, but further biological validation will be required. The prognostics results

also shows the positive correlation between the risk of reoccurrence and disease related deaths with the increase of tumor grade.

We also compared the similarities between the results of three signatures. The deregulated 64 genes specific to normal versus primary tumor comparison have also shown significant linkages to cancer related pathways. The other group of genes (701), strictly related to mets tissues have also shown significantly involvement in pathways previously observed in colon cancer.

In conclusion, this study shows the importance of integrated techniques of individually conducted gene expression studies and provide further insights in understanding of colon cancer data for clinical purposes. The cross-validation analysis of gene signature shows samples scarcity and different platform used for generation of expression remains challenging area. This study have also shown valuable knowledge and future direction for the treatment of colon cancers but a more robust approach using multiple biological stage data may answer question related to molecular heterogeneity.

## CONFLICT OF INTEREST

None declare

## REFERENCES


[1] Cardoso, J., et al. (2007) Expression and genomic profiling of colorectal cancer, Biochimica et biophysica acta, 1775, 103-137.

[2] Chan, S.K., et al. (2008) Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers, Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 17, 543-552.

[3] Davis, S. (2013) GEOquery R package: Get data from NCBI Gene Expression omnibus (GEO). Sean Davis, sdavis2@mail.nih.gov, pp. The NCBI Gene Expression Omnibus (GEO) is a public repository of microarray data. Given the rich and varied nature of this resource, it is only natural to want to apply BioConductor tools to these data. GEOquery is the bridge between GEO and BioConductor.

[4] 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, 44-57.

[5] Jass, J.R. (2007) Classification of colorectal cancer based on correlation of clinical, morphological and molecular features, Histopathology, 50, 113-130.

[6] Kang, G.H. (2011) Four molecular subtypes of colorectal cancer and their precursor lesions, Archives of pathology & laboratory medicine, 135, 698-703.

[7] Marisa, L., et al. (2013) Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value, PLoS medicine, 10, e1001453.

[8] Moreno, V. and Sanz-Pamplona, R. (2015) Altered pathways and colorectal cancer prognosis, BMC Med, 13, 76.

[9] Mukherjee, S., et al. (1999) Support vector machine classification of microarray data.

[10] Nannini, M., et al. (2009) Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives, Cancer treatment reviews, 35, 201-209.

[11] Planutis, K., Planutiene, M. and Holcombe, R.F. (2014) A novel signaling pathway regulates colon cancer angiogenesis through Norrin, Sci. Rep., 4.

[12] Sagynaliev, E., et al. (2005) Web-based data warehouse on gene expression in human colorectal cancer, Proteomics, 5, 3066-3078.

[13] Shih, W., Chetty, R. and Tsao, M.S. (2005) Expression profiling by microarrays in colorectal cancer (Review), Oncology reports, 13, 517-524.

[14] Siegel, R., Naishadham, D. and Jemal, A. (2012) Cancer statistics, 2012, CA: a cancer journal for clinicians, 62, 10-29.

[15] Smith, J.J., et al. (2010) Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer, Gastroenterology, 138, 958-968.

[16] Zhang, H., et al. (2001) Recursive partitioning for tumor classification with gene expression microarray data, Proceedings of the National Academy of Sciences of the United States of America, 98, 6730-6735.