Article

Molecular Classification of Colorectal Cancer Using Gene Expression Profile of Tumor Samples

Mamoon Rashid [1], Ramesh K. Vishwakarma [1], Ahmad Deeb [2], Mohamed A. Hussein [1] and Mohammad A. Aziz [3,*]

[1] King Abdullah International Medical Research Center/King Saud bin Abdulaziz University for Health Sciences, Department of Biostatistics and Bioinformatics-MNGHA, Riyadh, 11426, Saudi Arabia; rashidma@ngha.med.sa (MR), vishwakarmara@ngha.med.sa (RKV), husseinmo2@ngha.med.sa (MAH)

[2] King Abdullah International Medical Research Center/King Saud bin Abdulaziz University for Health Sciences, Research Office-MNGHA, Riyadh, 11426, Saudi Arabia; deebah@ngha.med.sa

[3] King Abdullah International Medical Research Center/King Saud bin Abdulaziz University for Health Sciences, Colorectal Cancer Research Program-MNGHA, Riyadh, 11426, Saudi Arabia; azizmo@ngha.med.sa

* Correspondence: azizmo@ngha.med.sa; Tel.: +966-11-42-94582

Abstract: Molecular classifications of colorectal cancer (CRC) are benefitting cancer research by providing insights into subtype-specific disease prognosis and better therapeutic intervention. So far different conventional DNA markers such as microsatellite instability (MSI), CpG island methylator phenotype (CIMP), chromosomal instability (CIN), and *BRAF* and *KRAS* mutations have been used to classify CRC patients but have not shown promising prognostic values. Here, for the first time, we show classification of CRC tumors from Saudi Arabian patients based on gene expression profile (GEP). An existing method of CRC subtyping has been applied to the GEP of tumors from Saudi CRC patients. Survival analysis was carried out on predicted CRC subtypes. *In-silico* functional analyses were conducted on the gene signature used for subtype prediction. The predicted subtypes showed distinct but statistically insignificant overall survival distribution (log-rank test, p = 0.069). Comparison of predicted subtypes in Saudi CRC patients with that of the French one showed significant dissimilarity in the two populations (Chi-square test, p = 0.0091). Functional analyses of the gene signature used for subtyping suggest their association with "cancer" and "gastrointestinal diseases". Most of the signature genes were found differentially expressed in CRC tumors compared to adjacent normal tissues. Such a classification framework might help improve the treatment of colorectal cancer patients.

Keywords: Colorectal cancer; Gene expression; Molecular classification; Molecular subtyping

1. Introduction

Colorectal cancer (CRC) is the third-leading cancer type for the estimated new cancer cases and deaths in 2010 US population with 142,570 (9%) cases and 51,370 (9%) deaths respectively**[1]**. In Saudi population it is the most frequent type of cancer in male (13.9%) and third-frequent in female (10.2%) [Saudi Cancer Registry, 2013]. While the cancer mortality rate due to CRC in Saudi population is 12.5% in male and 11.1% in female [http://www.who.int/cancer/country-profiles/sau_en.pdf]. Tremendous efforts have been made to understand and characterize the disease by available molecular determinants such as microsatellite instability (MSI) **[2, 3]**, BRAF and KRAS mutation status**[4]**, CpG island methylator phenotype (CIMP) **[5]** in order to classify CRC patients for better predictable treatment outcome (i.e prognosis).

Surprisingly, the patient groups classified by these molecular markers individually or in combination showed remarkable difference in therapeutic response and patient survival. Such observations contribute to the well-known notion of CRC being a heterogeneous disease**[6, 7]**. Moreover, numerous methods to further subtype the CRC tumors/patients based on clinical, pathological, genomic, genetic and epigenetic features have been proposed in the recent past**[5, 8-13]**. In a large-scale multidimensional analysis a hypermutant group of CRC tumors has been revealed which was not fully explained by MSI status and twenty-four genes were found hypermutated providing several new therapeutic targets**[13]**. In last five years, plethora of research publications focused on the problem of CRC subtyping and most of them used the gene expression profile (GEP) of the tumor samples employing unsupervised hierarchical clustering methods**[14-19]**. These methods are independent of each other and differ in gene expression platforms (Affymetrix HGU133plus2 and Agilent gene chips), methods of clustering, and patient cohorts in training and validation sets. Unsurprisingly, these methods resulted in different number of subtypes or classes of CRC tumors with three**[16, 17]**, five**[15, 18, 19]**, and six**[14]** subtypes.

In the present study we used a genome-wide mRNA expression analysis of 48 matched normal and tumor sample pairs from Saudi CRC patients using Affymetrix exon arrays**[20]**. We applied one of the existing GEP based CRC subtyping method**[14]** on this dataset to predict the various subtypes present among the colorectal cancer patients. The predicted subtypes differ in the overall survival probabilities showing the prognostic value of the subtyping. Functional analyses concluded the biological relevance of the gene signature used for CRC subtyping. Differential gene expression analysis was done to show that most of the genes from signature list significantly differentially expressed in the CRC tumor tissues compared to the corresponding normal tissues samples.

2. Materials and Methods

2.1 Ethical approval and sample collection

The study was approved ethically by the Institutional Review Board (IRB) of King Abdullah International Medical Research Center after a review process. The CRC patients were recruited for the study and the tissue samples were collected after the informed consent signed by the patients.

The samples were collected either by biopsies or surgical resections from the forty-eight patients upon their first presentation in the clinic for CRC diagnosis. The tumor and matched normal tissue samples were collected from 48 patients totaling about 96 samples for further studies. All cases regardless of their surgical stage and histological grade were included in this study. The inclusion criteria for the tumor samples were i) confirmation of histological consistency of specimens with the colon adenocarcinoma by a board certified pathologist ii) and retaining of >60% tumor cell nuclei in the specimens. The tissue samples from each selected CRC patients that contained no tumor cells and physically adjacent (>2 cm apart) to the tumor site were designated as matched normal samples. Further, the patients have not had undergone any CRC-related therapeutic intervention prior to the time of biopsy. The patients and tumor characteristics are shown in Table 1.

Table 1: Patient and tumor characteristics of CRC cohorts.

Characteristics	**Our dataset (n=47)**	**CIT discovery dataset (n=443)**	**P-value**
Mean age (sd, range) in years	62 (13, 28-97)	67 (14, 22–97)	0.0195^^
Sex (male/female) (percent)	19/28 (40.4/59.6)	237/206 (53/47)	0.0880^
TNM stage (percent)			
I	1 (2.1)	27 (6)	<.0001^

II	7 (14.9)	198 (45)	
III	39 (83)	164 (37)	
IV	0 (0)	54 (12)	
Adjuvant chemotherapy (percent)			
Yes	26 (55.3)	161 (45)	0.0674*
No	20 (42.6)	200 (55)	
NA	1 (2.1)	1	
Tumor location			
Proximal	13 (27.7)	176 (40)	0.1060^
Distal	34 (72.3)	267 (60)	
Median follow-up (sd, range), months	36.6 (24, 0 – 69.6)	50 (39, 0–201)	
Relapse (percent)			
Yes	4 (8.5)	109 (30)	<.0001^
No	39 (83)	250 (70)	
NA	4 (8.5)	3	

^P-value was calculated based on Chi-square Test

*P-value was calculated based on Fisher's Exact Test

^^ P-value was calculated based on Two sample t-Test.

2.2 Exon Microarray

The tumor and normal tissue specimen weighed between 10-30 mg. The tissue samples were stored in RNAlater (Ambion) at 4o C for 24 hrs followed by freezing and further storage at -20 o C. RNA was extracted from these tissues using Macherey Nagel RNA extraction kit (Germany) in a single preparation. The quality and quantity of the extracted RNA was checked using Nanodrop (Thermo Fischer Scientific, USA).

Genome-wide gene expression profile of tumor and matched normal samples were obtained using GeneChipTM Human Exon 1.0 ST Arrays from Affymetrix following the manufacturer's protocol. This array is also used to study alternative splicing in human genome on a genome-wide scale. In the GeneChipTM Human Exon 1.0 ST Arrays multiple probes on different exons summarize the expression value of all transcripts for the same gene. In this study we obtained the expression value at gene level using these exon arrays. The raw signal intensity data in the form of CEL files was extracted using Expression Console Software from Affymetrix. All the data from this study was previously submitted in GEO database with the accession numbers GSE50421 and GSE77434.

2.3 Quality control and preprocessing of raw data

Before starting the downstream analysis with exon microarray data the quality control (QC) experiments was done using the "oligo" package written in R based on BioConductor [21]. The extensive QC analyses were carried out to ensure that our exon array data is of good quality.

The preprocessing process (refers to the series of complex statistical methods) comprised of different steps of microarray data analysis i) background correction ii) quantile normalization and iii) summarization of the exon probes intensities at gene level. Aforesaid steps were carried out using RMA[22-24] (Robust Multichip Average) method implemented in the "oligo" package.

2.4 Colorectal cancer subtype prediction method

We used a subtype prediction method based on GEP that classifies the CRC tumors/patients in six different subtypes[14]. This subtyping method was based on unsupervised hierarchical clustering of GEP from 443 samples of training dataset and showed that the samples clustered into six clusters or subtypes. Each subtype was characterized based on different clinicopathological, phenotypic and mutation datasets. The molecular subtypes were robust because the method adopted i) consensus clustering method using both gene and sample resampling (1000 resampling using 90% of genes and samples in each resampling) leading to the stable results, ii) large number of samples (n=443) processed with same experimental procedure to obtain subtypes, iii) classification metrics (Euclidean/Pearson) that provide same results. Moreover, the clinical and biological characteristics of the subtypes remained conserved in the large validation dataset collected across different centers in different conditions[14].

For creation of subtype prediction model five top up-regulated and five top down-regulated genes were selected from each subtype and a centroid-based predictor was built. To predict/assign a subtype to a new sample a standard distance-to-centroid approach was used[25]. This prediction approach has been implemented in the R package “citccmst”[14]. There are various steps underlying the prediction algorithm as mentioned in the manual of “citccmst” in R. Those are briefly described here for the sake of clarity.

1. Mapping the genes from our CRC tumor expression dataset to the 57 discriminating genes/probes used in centroid calculation in “citccmst” from discovery dataset[14].
2. Averaging expression measures per gene symbol both in our CRC dataset and in the citccmst discovery dataset. In any case, our CRC data and the citccmst discovery set data are reduced to discrimating probes/genes measured in both datasets.
3. Recomputing the centroids of each 6 subtypes using citccmst discovery dataset from step 2.
4. Computing distances of each CRC samples to those 6 centroids.
5. Assigning each sample to the subtype(s) based on closest distance to the centroids. If the sample is close to many centroids the sample is considered as “mixed” subtype. If the distance of a sample to closest centroid is too far to confidently assign the sample to a given subtype, the sample is considered as “outlier”. Both the mixed and outlier cases are considered as uncertain and might be removed from analysis.

Thus, in the present study the “citccmst” (http://cit.ligue-cancer.net) R package was used to predict the subtypes of colorectal cancer samples.

2.5 Survival analysis

The patient’s overall survival probabilities were analyzed using Kaplan-Meier estimator. Kaplan-Meier estimator is a non-parametric statistical test that estimates the survival function from patient’s survival data. The overall survival is defined as the time from the diagnosis or the start of treatment of CRC until the patient remains alive. The overall survival probabilities were plotted for the six predicted subtypes. The survival distribution of each molecular subtype manifests the biological significance of the subtype. The survival distributions were compared using log-rank test. The R software package “survival” and “survminer” were used for the Kaplan-Meier survival analysis and SAS procedure “Phreg” was used for cox- regression.

2.6 Differential gene expression analysis

The genes which are significantly differentially expressed in tumor samples compared to the corresponding normal samples have been identified by the use of linear models through the R/Bioconductor software package "Limma" [26]. This package has the capability of analyzing comparisons between many genes simultaneously. It is also designed for analyzing complex experiments with variety of experimental designs. Here, the analysis was focused on identifying the genes expressed differentially in the case of colorectal cancer tissue samples and matching the list of gene signature in this differentially expressed gene set.

2.7 Functional analyses of gene signature used for subtyping

To identify the most relevant biological pathway related to the 57 gene signature, we used Ingenuity Pathway Analysis (IPA) tool (www.ingenuity.com). This web-based tool provides the statistical measure of the presence of the gene set in various biological pathway datasets. The value (–log*p-value) of 2 for e.g. explains that there is a 1% possibility that the gene set present in the pathway by random chance. It means that the score of 2 or more equates to the 99% confidence that the genes are present in the said pathway. The analysis also maps the gene set on the relevant biological gene networks and rank the networks based on a score. Moreover, it also provides the biomarker information if any of the genes in the gene set have such features to be a biomarker.

The overall analyses strategy adopted in the current study has been summarized as an illustration in Figure 1.

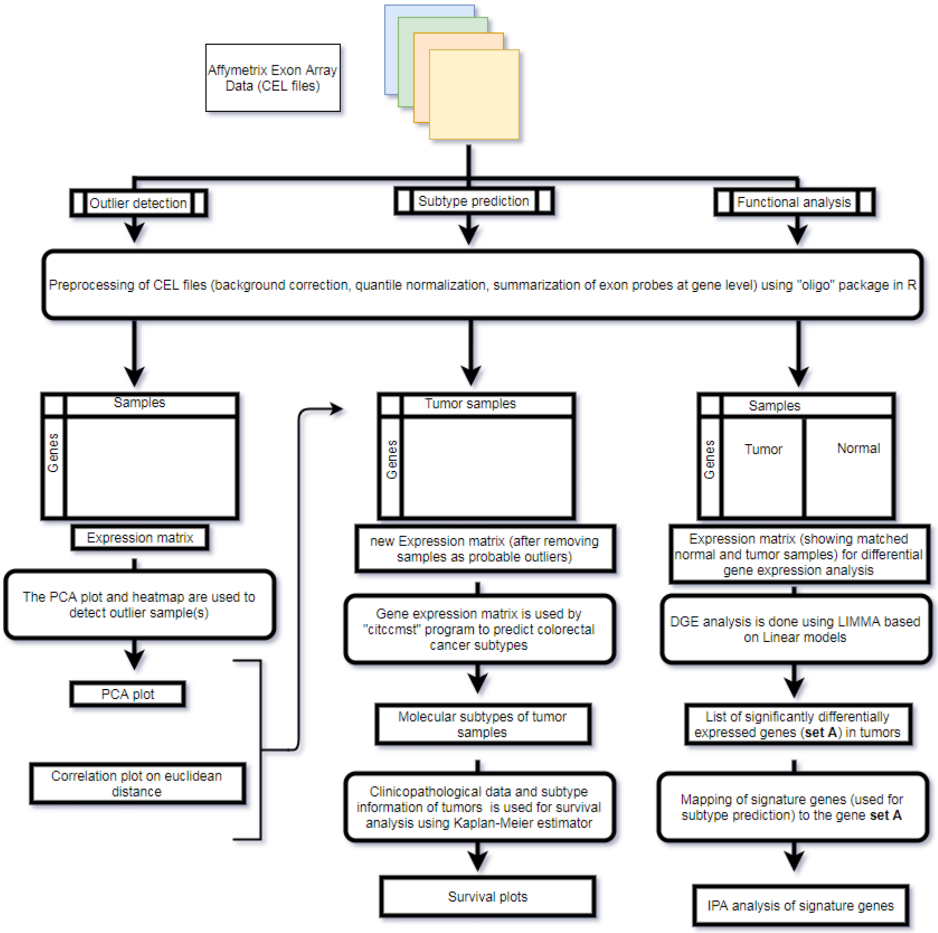
Affymetrix Exon Array Data (CEL files)
Outlier detection
Subtype prediction
Functional analysis
Preprocessing of CEL files (background correction, quantile normalization, summarization of exon probes at gene level) using "oligo" package in R
Samples
Genes
Expression matrix
The PCA plot and heatmap are used to detect outlier sample(s)
PCA plot
Correlation plot on euclidean distance
Tumor samples
Genes
new Expression matrix (after removing samples as probable outliers)
Gene expression matrix is used by "citccmst" program to predict colorectal cancer subtypes
Molecular subtypes of tumor samples
Clinicopathological data and subtype information of tumors is used for survival analysis using Kaplan-Meier estimator
Survival plots
Samples
Genes
Tumor
Normal
Expression matrix (showing matched normal and tumor samples) for differential gene expression analysis
DGE analysis is done using LIMMA based on Linear models
List of significantly differentially expressed genes (**set A**) in tumors
Mapping of signature genes (used for subtype prediction) to the gene **set A**
IPA analysis of signature genes

Figure 1: Overall analysis methodolgy adopted in the current study

3. Results

3.1 Outlier detection

We tested the CRC samples for any anomalies or outliers in the exon microarray data generation. The proximity based models such as clustering method marked two samples as potential outliers. Moreover, principal component analysis and heatmap also highlighted the same two samples as potential outliers. Those two samples (050911-01-TS and 073011-01-TS) were eliminated from the dataset for all the downstream analysis.

3.2 CRC subtypes using tumor samples gene expression profile

The pre-processed and normalized gene expression profile of tumor samples from CRC patients were used to classify CRC tumors into subtypes using one of the existing methods of CRC subtyping [14]. This method called "citccmst" classified the samples into six different subtypes C1, C2, C3, C4, C5, and C6 with 14, 2, 3, 3, 13, 11 (two samples were removed as outliers) number of samples in each subtype respectively. The PCA plot was also generated by the classification method to show the distribution of samples along the two-dimensional space (Figure 2A). The upper and lower panels in the figure 2A are the PCA plots showing the "CITCCMST discovery dataset" and our "input dataset" respectively. We also intended to compare the subtype prediction results using our CRC dataset with that of the discovery dataset of CITCCMST study[14]. The chi-square test suggests that these two populations (Saudi and French) of tumor samples were significantly different (p-value=0.0091) from each other in context of the proportion of different CRC molecular subtypes (Figure 2B). But, both the populations do have the same six subtypes.

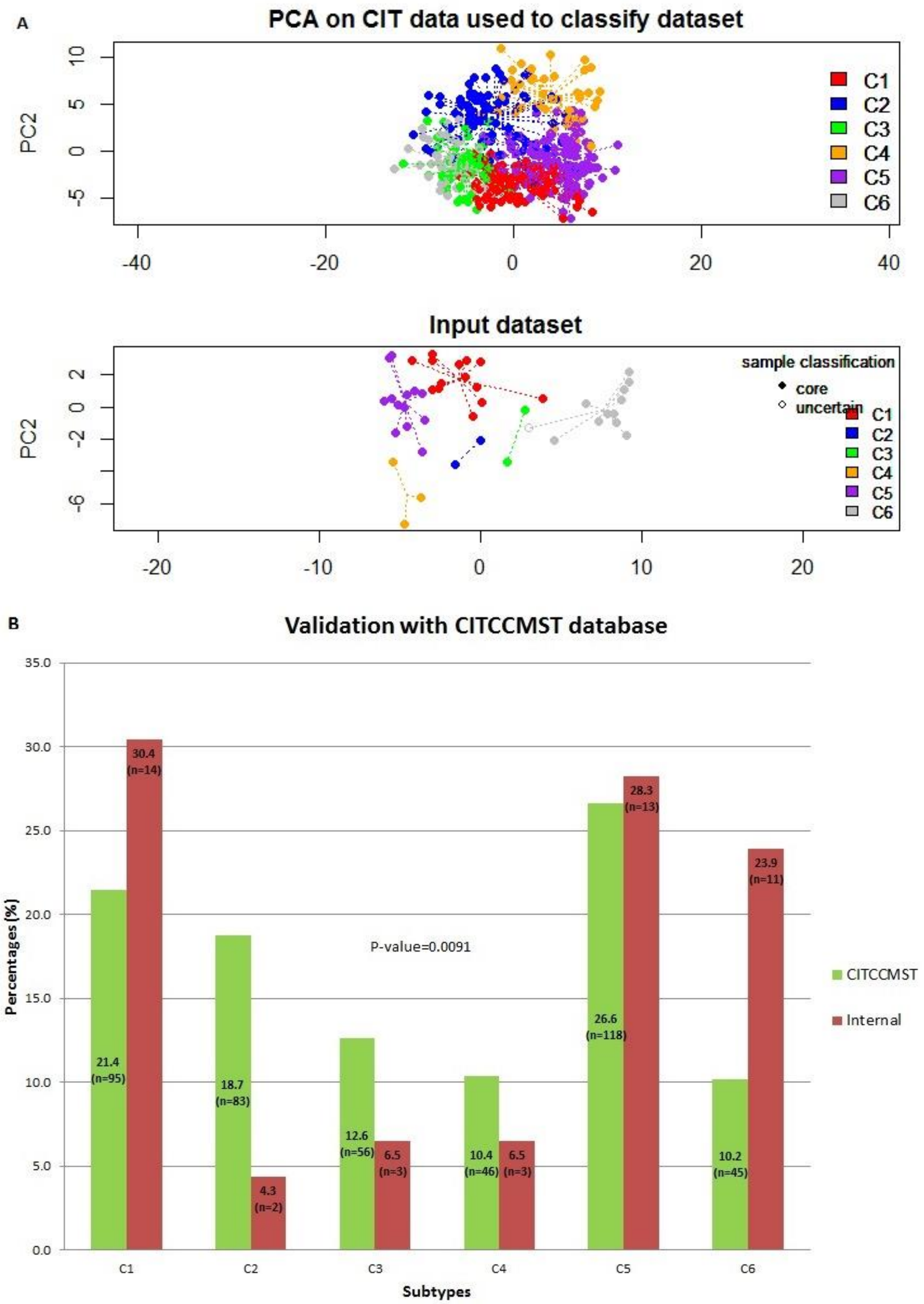

Figure 2: A) PCA plot showing the distribution of the CRC tumor samples in two dimensional space into six subtypes. The upper and lower panels in the plot display the sample distribution using "CITCCMST discovery dataset" and our "input dataset" respectively. B) Comparison of subtype

doi:10.20944/preprints201804.0016.v1

proportion from our CRC ("internal", red bar) dataset with that of the French ("CITCCMST", green bar) dataset.

3.3 Prognostic value of the predicted subtypes

The patient's survival data was analyzed to see the overall survival distributions after grouping the patients into predicted subtypes (Figure 3). The differences between survival distributions among subtypes were compared using log-rank test with an endpoint of four year overall survival. The survival probabilities among all six subtypes differ greatly to each other however not statistically significant (P-value: 0.069). This might be due to the insufficient number of subjects in each subtype. The patients with C4 and C6 subtypes showed poor outcome in overall survival (median survival time 161 and 210 days) compared to patients with C1 and C5 subtypes (median survival time 1304 and 1027 respectively). To confirm this, we recoded our classification by combining C4 and C6 into a single high-risk group, versus all other subtypes as the low-risk group. This grouping has already been reported in earlier literature [14]. From our analysis, it is found that this dichotomous classification led to a significantly different overall survival probabilities between the high-risk group and the low-risk group (P-value: 0.0151).

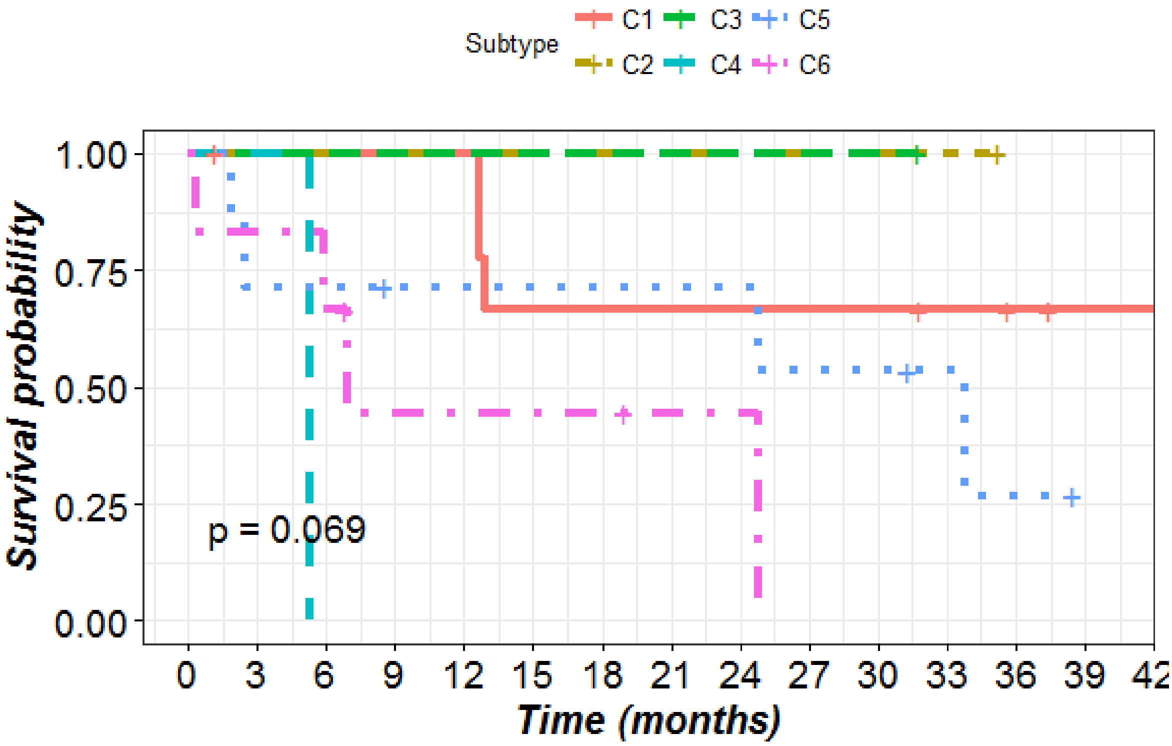

Figure 3: Survival plot showing the overall survival distribution of six predicted subtypes of CRC patients.

3.4 Cox proportional hazard analysis

We performed Cox analysis to determine the prognostic value of the predicted subtypes controlling for other known prognostic variables. Controlling for age (in 5 years), tumor size (in centimeter), gender, types of therapy and metastasis status, the effect of the predicted subtype was no longer statistically significant (hazard ratio [HR]: 3.63, 95% CI: 0.794-16.603, p=0.097). However age and metastasis status remained statistically significant (HR: 0.89, 95% CI: 0.82-.96, p=0.0152), (HR: 15.153, 95% CI: 1.74- 132.19, p=0.0048), respectively.

3.5 Differential expression of gene signature used for subtyping

The molecular subtypes predicted in this study were based on 57 genes/probes selected from a previous study for classification of colorectal cancer tumor samples. The presence of those genes in our CRC dataset prompted us to check the expression profile of the genes. The matched normal and tumor tissue samples for all the CRC patients were used for the differential gene expression (DGE) analysis. The analysis resulted in 2866 genes being significantly differentially expressed in the tumor tissues. Out of 2866 genes 1610 genes were down-regulated and 1256 genes were up-regulated. Comparison of 57 gene signature to the 2866 gene list showed that there are 22 genes (22/57= 38% genes) from gene signature which are significantly differentially expressed in our CRC dataset. The volcano plot shows the DGE of the gene signature in the CRC dataset (Figure 4).

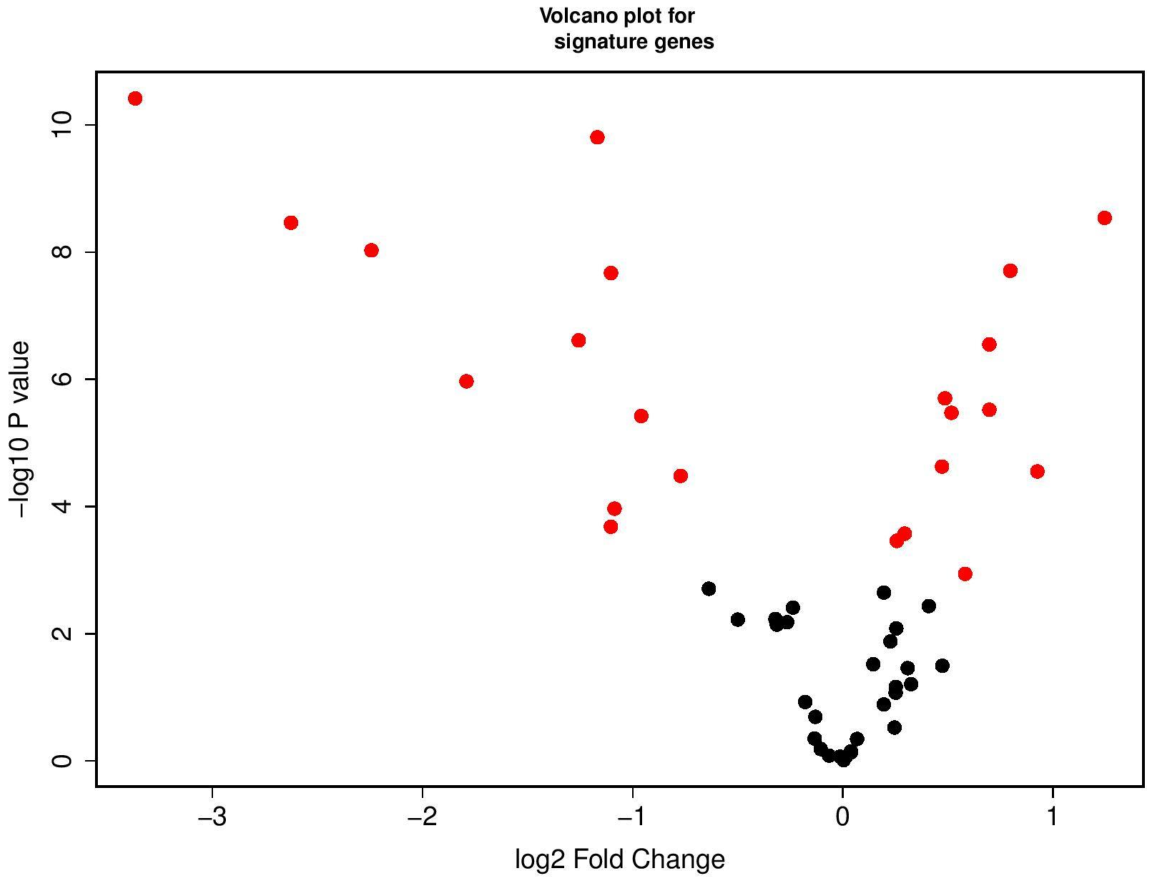

Figure 4: Differential gene expression analysis of 57 genes in our CRC tumor samples compared to the matched normal samples. Red solid circles represent 22 out of 57 genes found differentially expressed in the CRC tumor dataset.

3.6 IPA core analysis of gene signature

3.6.1 Gene signature used for classification is functionally relevant as indicated by their association with cancer

Fifty-seven gene signature was subjected to ingenuity pathway core analyses to analyze its functional relevance (Figure 5A). The most statistically significant function associated with these genes was cancer followed by gastrointestinal disease, hereditary disorder and metabolic disease. 54/57 genes were associated with cancer while 48/57 genes were found to be associated with gastrointestinal diseases. This gene signature had only 4 genes that were found to be associated with colorectal adenoma (CA1, CA2, HSD11B2 and BEST2) but 44 genes were associated with gastrointestinal neoplasia (Table S1).

3.6.2 Top network involving gene signature molecules is significantly associated with cancer

We carried out network analysis of the 57 genes used for classification (Figure 5B). Eleven of these genes were part of the network which has top score of nineteen. Only three out of these 11 genes were found to be differentially expressed in our CRC tumor samples compared to the matched normal tissue samples. This network was functionally associated with cancer, hematological disease and immunological disease. Two miRNAs were also part of this network (miR-101 and miR-101-3p), which provide tools to modulate the function of the genes. Further, we checked the differential expression of some of the genes in the network and found CA1 to be significantly down regulated.

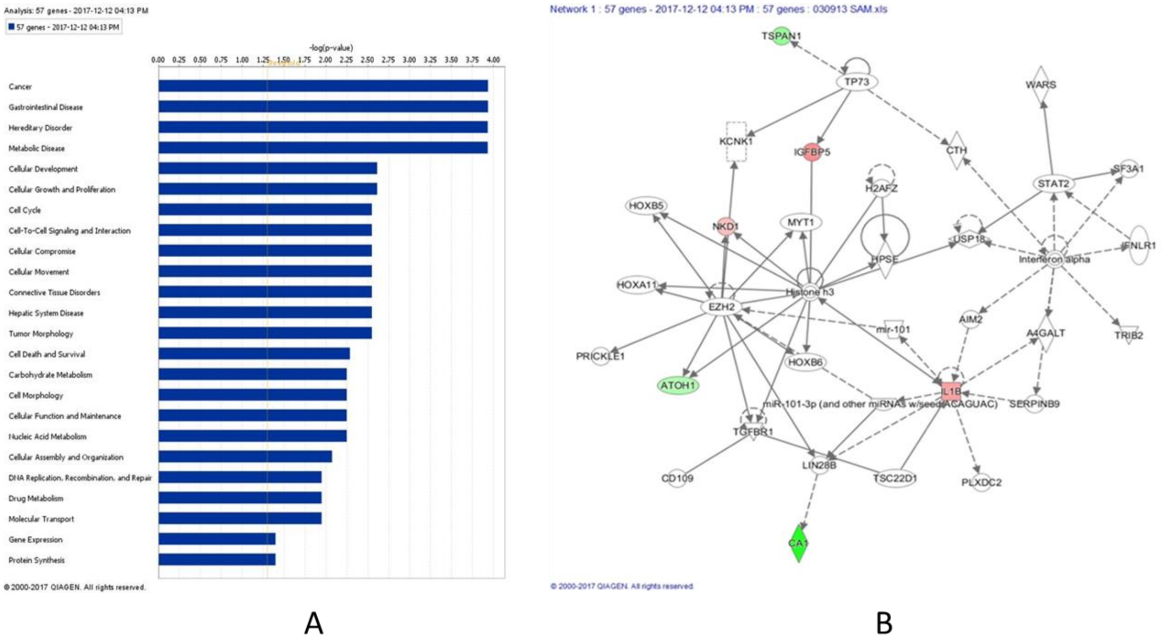

Figure 5: A) Ingenuity pathway analysis of 57 gene signature showing "cancer" as the most significant function associated with these genes. B) Top scoring network containing 11 out of 57 genes indicating the associated with cancer, hematological disease and immunological disease.

3.7 Biomarker analysis of gene signature

We carried out biomarker analysis of the 57 gene signature to assess the potential of these genes as biomarkers for diagnosis, efficacy, disease progression and prognosis. Six of these genes were found to be candidate biomarkers that could be detected in Human blood, Plasma/serum, Urine, blood platelets, cytotoxic and effector T cells and large intestine. Out of these six genes five (83 %) were found to be differentially expressed in our CRC tumor samples compared to the matched normal tissue samples. CA2 and HPSE were the genes which were targets for many drugs (Table 2).

Table 2: Identification of biomarkers in 57 genes signature.

Symbol	**Entrez Gene Name**	**Location**	**Family**	**Drug(s)**	**Human**	**Blood**	**Plasma/Serum**	**Urine**	**Blood platelets**	**Cytotoxic T cells**	**Effector T cells**	**Large Intestine**	**Biomarker Application (s)**
ASCL2	achaete-scute family bHLH transcription factor 2	Nucleus	transcription regulator	NA	x								diagnosis

CA2	carbonic anhydrase 2	Cytoplasm	enzyme	ethoxyzolamide, dichlorphenamide, phentermine/topiramate, brimonidine/brinzolamide, methazolamide, diazoxide, hydrochlorothiazide, acetazolamide, trichloromethiazide, dorzolamide, chlorothiazide, dorzolamide/timolol, brinzolamide, chlorothiazide/reserpine, quinethazone, chlorthalidone, benzthiazide, sulfacetamide, topiramate	x	x	x	x		x	x	x	diagnosis,
CHGA	chromogranin A	Cytoplasm	other	NA	x	x	x	x				x	diagnosis,efficacy,
HPSE	heparanase	Plasma Membrane	enzyme	2-O,3-O-desulfated heparin, PG 545, SST 0001, heparanase inhibitor PI-88	x	x			x			x	efficacy
PLAGL2	PLAG1 like zinc finger 2	Nucleus	transcription regulator	NA	x	x				x	x	x	disease progression

TIMP3	TIMP metallopeptidase inhibitor 3	Extracellular Space	other	NA	x	x			x	x	x	x	diagnosis,prognosis

4. Discussion

Colorectal cancer is a very heterogeneous disease among the patients and hence it is difficult to classify it in a clinically relevant manner. There have been several attempts to capture this heterogeneity by proposing different classification schemes that have evolved along with better understanding of molecular details pertaining to CRC. The latest scheme of classification which is considered to be the most comprehensive till date employed an amalgamation of classification scheme from six groups [27]. All six classification schemes were based on gene expression profile from different populations and platforms. In the present study, we aimed to enrich the classification efforts by employing one of the six classification schemes for subtyping CRC patient samples from Saudi Arabia. We also analyzed the biological relevance of the genes used for classification and found their association with important biological functions and disease along with pathways and networks.

Though the number of samples used by 'citccmst' for classification (n=443) was much higher than our dataset (n=48), this particular classification scheme was able to capture all six subtypes in our sample. This was expected given that the least subtype group (C6) in the citccmst' dataset represents about 10.2% which suggests that in our dataset one might expect to observe 4.8 subjects on average. Our results suggests that the distribution of the subtypes across our dataset and citccmst' CRC tumor samples are significantly different (Chi-square test, p=0.0091). One explanation of these findings is that the patterns of the genes involved in the subtyping differ across populations.. Another explanation might be that the distribution of the subtypes might reflects the clinical heterogeneity between our population and the original citccmst' dataset. This is apparent by the fact that patients in our dataset are younger, and tend to have less of stage IV compared to CITCCMST. The latter is more plausible given the fact that the different subtypes reflects the underlying moleculrate state of the cancer as described by Marisital et. al[14]. This is an important feature of a subtyping scheme especially in the context of personalized medicine where one might need a method by which clinicians could capture the entire molecular state of that specific patient or a cohort of patients. To confirm the sensitivity of this classification approach to the underlying state of the population of interest more studies need to be carried in different populations with different clinical presentations and characteristics.

The prognostic value of the identified subtypes is evident by the survival pattern of the patients belonging to specific subgroups. Though our dataset is limited by the number of patients in each subgroup, the pattern of survival probability is similar with subgroups C4 and C6 exhibiting the worst outcome whereas C2 and C3 show the best prognosis. Since there is no survival analysis available for the validation datasets used by Marisa et.al., [14] our data provides validation of the survival pattern associated with the predicted subgroups identified using the 57 genes signature. Our data suggest that patient within subtypes C4 and C6 have poor outcome which could be ascribed to the associated molecular characteristics as discussed earlier. An interesting observation in our analysis is that we could not establish a statistically significant effect of the subtyping in the presence of other known prognostic variables such as age, gender, and metastasis status. Our results is not consistent with Maridsa et. al. findings where in their analysis it appears that the subtyping does offer prognostic value beyond the other prognostic variables that they have added in their model. This observation could be very will likely due to our limited sample size. A study with a larger pool of patient from different populations might be important to validate the additional value of subtyping beyond currently known prognostic factors.

Further, we analyzed the biological relevance of the 57 genes signature in terms of the associated disease and networks. As expected, the most significantly associated disease was cancer followed by gastrointestinal disease. However, only 4 genes matched to the genes associated with colorectal adenoma. Of these CA1 gene was significantly down-regulated in our patient cohort which confirms previous results in TCGA data set[28]. CA1 has also been used in the gene classifier that is associated with cellular phenotype[18] and using single cell approach[29]. Usually classification gene signatures with functionally relevant genes are helpful in explaining the biology of the colorectal cancer subtypes. As we have reported earlier 28/30 genes used for classification were associated with colorectal cancer. However these genes were used to classify tumor and normal samples[30]. We further analyzed the differential expression of the 57 genes between our normal and matched cases and found some of them to be significantly differentially expressed. We constructed network of genes in the classification signature based on their association. The most statistically significant network had 11 of the 57 genes. Of these IGFBP5, IL1B and NKD1 were found to be up regulated while CA1 and TSPAN1 were down regulated in our patient cohort. Out of these 11 genes, 8 genes were not differentially expressed in our CRC tumor samples. This might reflect the underlying difference in gene expression program in Saudi CRC patients. In the biomarker analysis using IPA the six (out of 57) genes are found to be as potential biomarkers. And to our surprise 5 of these six genes were found to be differentially expressed in our CRC tumor samples. It proves the usability of these five genes as potential biomarkers in Saudi CRC patients. Moreover, each of the 17 genes (22 - 5) which were shown to be differentially expressed but not reported as biomarkers in the IPA analysis in Saudi CRC tumor samples is a target for further investigation to be used as a potential biomarker in Saudi population.

We also checked the overlap of statistically significant differentially expressed genes across the predicted subtypes. There was variable number of genes in each subtype that were differentially expressed with respect to the rest of the subtypes. Most of the genes in each subtype were common with one or more subtypes. But some of the genes are unique in each subgroup except for C3. These unique genes provide an opportunity for suggesting subtype specific targets which may have utility as biomarkers.

5. Limitations

One obvious limitation of our study is the small sample size and therefore larger cohort of Saudi colorectal cancer patients might be needed to confirm our observations. Our analysis did not include classical features such as CIMP, MSI, and MMR status of the patients. This was because of low availability of the patients' samples.

Supplementary Materials: The following are available online, Table S1: List of 44 genes associated with gastrointestinal neoplasia.

Acknowledgments: Data used in the study was generated from grant (RC/1083) awarded to MAA from KAIMRC.

Author Contributions: MR, MH, and MA conceived the project. MR, MH, and MA designed the experiments. MR, RV, and MA performed the experiments and analyzed the data. AD contributed the clinicopathological data of the colorectal cancer patients. MR and MA wrote the paper. All the co-authors read and approved the content of the manuscript.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Jemal, A., et al., *Cancer statistics, 2010.* CA Cancer J Clin, 2010. **60**(5): p. 277-300.
2. Tejpar, S., et al., *Microsatellite instability, prognosis and drug sensitivity of stage II and III colorectal cancer: more complexity to the puzzle.* J Natl Cancer Inst, 2011. **103**(11): p. 841-4.
3. Sinicrope, F.A. and D.J. Sargent, *Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications.* Clin Cancer Res, 2012. **18**(6): p. 1506-12.
4. Vecchione, L., et al., *EGFR-targeted therapy.* Exp Cell Res, 2011. **317**(19): p. 2765-71.
5. Hinoue, T., et al., *Genome-scale analysis of aberrant DNA methylation in colorectal cancer.* Genome Res, 2012. **22**(2): p. 271-82.
6. Martini, M., et al., *Targeted therapies: how personal should we go?* Nat Rev Clin Oncol, 2011. **9**(2): p. 87-97.
7. Popovici, V., et al., *Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer.* J Clin Oncol, 2012. **30**(12): p. 1288-95.
8. Jass, J.R., *Classification of colorectal cancer based on correlation of clinical, morphological and molecular features.* Histopathology, 2007. **50**(1): p. 113-30.
9. Shen, L., et al., *Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer.* Proc Natl Acad Sci U S A, 2007. **104**(47): p. 18654-9.
10. Ogino, S. and A. Goel, *Molecular classification and correlates in colorectal cancer.* J Mol Diagn, 2008. **10**(1): p. 13-27.
11. Furlan, D., et al., *Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas.* Mod Pathol, 2011. **24**(1): p. 126-37.
12. Loboda, A., et al., *EMT is the dominant program in human colon cancer.* BMC Med Genomics, 2011. **4**: p. 9.
13. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
14. Marisa, L., et al., *Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value.* PLOS Medicine, 2013. **10**(5): p. e1001453.
15. Budinska, E., et al., *Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer.* J Pathol, 2013. **231**(1): p. 63-76.
16. Roepman, P., et al., *Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition.* Int J Cancer, 2014. **134**(3): p. 552-62.
17. De Sousa, E.M.F., et al., *Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions.* Nat Med, 2013. **19**(5): p. 614-8.
18. Sadanandam, A., et al., *A colorectal cancer classification system that associates cellular phenotype and responses to therapy.* Nat Med, 2013. **19**(5): p. 619-25.

19. Schlicker, A., et al., *Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines.* BMC Med Genomics, 2012. **5**: p. 66.
20. Aziz, M.A., et al., *Integrated exon level expression analysis of driver genes explain their role in colorectal cancer.* PLoS One, 2014. **9**(10): p. e110134.
21. Carvalho, B.S. and R.A. Irizarry, *A framework for oligonucleotide microarray preprocessing.* Bioinformatics, 2010. **26**(19): p. 2363-7.
22. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
23. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.
24. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
25. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.
26. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
27. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.
28. Liu, H.Y. and C.J. Zhang, *Identification of differentially expressed genes and their upstream regulators in colorectal cancer.* Cancer Gene Ther, 2017. **24**(6): p. 244-250.
29. Dalerba, P., et al., *Single-cell dissection of transcriptional heterogeneity in human colon tumors.* Nat Biotechnol, 2011. **29**(12): p. 1120-7.
30. Gabere, M.N., M.A. Hussein, and M.A. Aziz, *Filtered selection coupled with support vector machines generate a functionally relevant prediction model for colorectal cancer.* Onco Targets Ther, 2016. **9**: p. 3313-25.