HHLA2 and NEBL as novel prognostic biomarkers based on Rayleigh survival model in colorectal cancer

Altaf H. Khan*1, Mohammed A. Aziz2, Ghada T. Alatar2, Amal Al Ghamdi2, Mohammed A. Hussein1, and Tanveer S. Khatlani3

1Department of Biostatistics and Bioinformatics
2Colorectal Cancer Research Program
3Department of Stem Cell and Regenerative Medicine, King Abdullah International Medical Research Center, King Saud University for Health Sciences, National Guard Health Affairs, Riyadh (KSA)

March 14, 2021

Corresponding Author:
Altaf H. Khan (Senior Biostatistician)
Department of Biostatistics and Bioinformatics
King Abdullah International Medical Research Center
King Saud University of Health Sciences
Ministry of National Guard Health Affairs
Riyadh, Saudi Arabia
Email: khanahslc@gmail.com/khanal1@ngha.med.sa

Acknowledgement: Altaf H. Khan, and Mohammed Azhar Aziz contributed in writing the manuscript. Statistical Analysis was done by Altaf H. Khan, Dr. Mohammed Azhar Aziz had done the IPA (ingenuity pathway analysis) software tool to filter the genes. Ms. Ghada Al Atar and Ms. Amal AlGhandi as lab technologist were involved for microarray and qPCR experiments. Dr. Mohammed A. Hussein and Dr. Tanveer S. Khatlani reviewed the manuscript.

Potential conflicts of interest: None

Financial support: None

*Email: ahkhan25@hotmail.com
Abstract

Background
The primary focus for this study was investigation of prognostic biomarkers in colorectal cancer (CRC), since biomarkers are instrumental in clinical decision making and patient management as well as playing pivotal roles in precision medicine.

Methods
We analyzed indigenous dataset from colorectal cancer patients. Exon microarray dataset was used and 135 genes were identified as novel candidate biomarkers. 135 genes were further split into two groups: low and high gene expression values via 'maxstat' algorithm, they were analyzed using Kaplan-Meier (K-M) analysis and univariate Cox model, and a set of 33 genes were identified as statistically significant ($p<0.05$). Furthermore, using the 'VARCLUS' algorithms (a SAS software procedure) which is a useful tool for variable reduction, based on the divisive clustering techniques, a small subset of 5 genes were selected out of 33 significant genes as potential candidates to build survival models. Both parametric and semi-parametric survival models were utilized to assess whether these 5 genes could be used as prognostic biomarkers.

Results
HHLA2 ($p < 0.01$) and NEBL ($p < 0.01$) genes emerged as potential biomarkers, based on the para- and semi-parametric models such as: Rayleigh, Exponential probability distributions, and Cox models and were also validated on independent datasets, apart from validation with qPCR test as well as with the cell lines patients data in the laboratory. A rigorous statistical evaluation for model’s performance were done via Harrell’s index, Brier score, Shoenfeld residual plots as well as with comparing several predictive model plots. We also made the comparison of Akaike and Bayesian Information(AIC and BIC) criteria as well as log-likelihood estimates. The Tukey-Anscome plot and Quantile-Quantile plot as diagnostic tools were applied to validate the parametric survival models.

Conclusion
Based on predictive models HHLA2 and NEBL novel biomarkers were found as statically significant.

List of abbreviations
AIC: Akaike Information Criteria
BIC: Bayes Information Criteria
CRC: Colorectal Cancer
SAS: Statistical Analysis Software
'maxstat' and 'VARCLUS' are statistical tools.
Introduction

Precision and personalized medicine is coming of age with conspicuous results in cancer. Colorectal cancer owing to its heterogeneous nature could be a model disease for assessing the pros and cons of precision medicine [1, 2, 3, 4, 5]. Biomarkers are proposed to be important tools in practicing precision and personalized medicine [6, 7, 8]. Prognostic biomarkers are defined as measurable factors that can be used to assess the disease progression. Predictive biomarkers are used primarily for knowing the probable response of drug treatment [3, 9]. Same biomarkers can serve as prognostic and predictive as well as can also be used as potential drug target. The progression of colorectal cancer is already well defined in the form of clinical models and pathways. There are several putative biomarkers which are classified according to type, pathway, clinical utility, and method of analysis. Mutation in KRAS gene is well known and clinically being used as predictive biomarker(s) for adjuvant chemotherapy in colorectal cancer [10]. Other biomarkers like gene mutations, loss of 18q, micro-satellite instability are being studied for their clinical utility and relevance. Several other biomarkers like miRNA, circulating tumor cells and DNA had been extensively studied though yet to reach their potential application. Gene expression based biomarkers are easy to assay as well as routinely being used in clinics. There is a need to discover better biomarkers that could be biologically and clinically relevant. Gene expression databases can be explored for novel biomarker discovery. However, the importance of indigenous data from local population is underscored alongside the idea of personalized medicine. In this report we used dataset of differentially expressed genes in local colorectal cancer patient samples [11]. Statistically significant differentially expressed genes were analyzed to find putative novel biomarker candidates. These genes were tested for their potential as prognostic biomarker using parametric and semi-parametric survival models. These models identified Human Endogenous Retrovirus-H Long Terminal Repeat-Associating Protein 2 (HHLA2) and Nebulette (NEBL) genes as potential independent prognostic biomarkers in colorectal cancer. Data sets from the Broad Institute https://www.broadinstitute.org and ‘cBioPortal’ https://www.cbioportal.org/study/summary?id=coadread_tcga_pan_can_atlas_2018 were used to validate our results. These datasets will be referred as ‘external’ throughout the manuscript. Experimental validation of HHLA2 gene expression was carried out using quantitative real time PCR on matched tumor-normal samples from patients. Expression of HHLA2 gene was also analyzed in colorectal cancer cell lines. A schematic flow diagram for the work done in this paper is given below in Figure 1.

Methods

Ethics approval and consent to participate:
The study was approved by the King Abdullah International Medical Research Center’s bioethics committee.
The matched tumor-normal samples from 41 patients will be referred as internal data were used to identify differentially expressed genes using Exon microarray (Affymetrix, Thermo Fisher Scientific, USA) and is described in [11]. All the patients were consented and the study (RC10/083) was approved by institutional review board before sample collection. Raw data was analyzed using Transcriptomics Analysis Console (Affymetrix, Thermo Fisher Scientific, USA) to obtain list of differentially expressed genes. Differentially expressed genes were analyzed using Ingenuity Pathway Analysis (IPA), Qiagen Biosciences, USA to identify novel candidate genes that qualify for biomarker identification. IPA was used with filter considering molecules associated with cancer and previously unknown as biomarker.

Exon microarray dataset was used and 135 genes were identified as novel candidate biomarkers, split into two groups: low and high gene expression values via ‘maxstat’ algorithm from an R package, see: https://cran.r-project.org/web/packages/maxstat/index.html. After split into two two groups, they were analyzed using Kaplan-Meier (K-M) survival analysis with log-rank test, and univariate Cox model. A set of 33 genes were found to be statistically significant from these analyses, as p values were less than the assumed significance level as: \( \alpha = 0.05 \) considered as statistically significant in this work for all statistical analysis.

Further, using ‘VARCLUS’ algorithms, a SAS software procedure[13], a small subset of 5 genes genes were selected from the set of 33 genes. And these 5 genes were utilized as prognostics biomarkers to build survival models based
on parametric and semi-parametric methods. To assess and evaluate predictive models, a rigorous diagnostic and performance tools were applied, namely: Harrell’s index, Brier score, Shoenfeld residual plots as well as with comparison of several predictive model plots. We also made the comparison of Akaike and Bayesian Information (AIC and BIC) criteria as well as the log-likelihood estimates. The diagnostic assessment for parametric survival models were based on Tukey-Anscombe (TA) and Quantile-Quantile plot.

A detailed outline for splitting continuous variables the ‘maxstat’ tool, feature selection the ‘VARCLUS’ algorithm, parametric, semi and non-parametric survival models, and model’s performance evaluation are described in the supplementary file. All the statistical analysis were done using SAS software [14] and R statistical packages.

**Result**

**Demographics distributions for internal and external datasets**

The demographics and clinical characteristics for internal (N=41) and external (N=102) datasets are given in supplementary Table S1.

**Filtering of genes, Kaplan-Meier and univariate Cox model**

Using IPA tool with filtering algorithms a set of 135 genes were identified as potential biomarkers, and were further split as low and high expression values groups via ‘maxstat’ tool (an R-package). The z-statistic, with their p-values based on log-rank test and K-M survival analysis as well as univariate Cox model were shown in Table S2 (see: supplementary file); and a subset of 33 genes were found to be statistically significant (p < 0.05), and K-M plots are presented in the supplementary materials. These subset of genes include: B3GNT7, BTNL3, BTNL8, C2orf88, CA1, CCDC68, CEMIP, CLDN8, COL1A2, CPM, CWH43, HHLA2, IFITM1, MT1G, NEBL, NR3C2, PDK4, PLCG1, REG1B, SELENBP1, SELENOP, SI, SLC26A3, SLC51B, SMPDL3A, TGBF1, THBS2, THRB, TMIGD1, TRPM6, TSPAN7, UGT2A3, and UGT2B17. Out of these 33 genes, 16 of them were associated with CRC disease based on the IPA analysis, and were graphically shown in Figure 2, and their functional relevance were found to be significant for survival, and makes our results more reliable and biologically explainable.

**Selecting CLDN8, HHLA2, NEBL, SELENOP and THBS2 genes as predictors based on feature-selection algorithm**

In general, building predictive models we encounter with multicollinearity problem due to two or more explanatory variables in regression modeling get linearly related, which is always a challenge with omics data since genes are generally highly correlated. To overcome this multicollinearity in modeling, we utilized feature selection ‘VARCLUS’ algorithm (discussed in supplementary file). So, 5 clusters were created in such a way that genes from the same cluster are highly correlated with each other but have a low correlation with any other cluster. The selection process consisted of selecting one gene from each cluster. We selected the genes with minimum R-square ratio within its own cluster. A
Figure 2: Based on univariate Cox-model and Kaplan-Meier survival plot, a subset of 33 genes were statistically significant and out of this subset of genes 16 genes are known to be associated with colorectal cancer. IPA tool was utilized to make this graph.

gene selected this way is the best representative of the cluster. Based on this algorithm a set of five gene covariates namely: CLDN8, HHLA2, NEBL, SELENOP and THBS2 were chosen from each cluster respectively, and their K-M survival plots were given in supplementary file. The output results from this algorithm such as cluster number, name of the variable, $R^2$ with own cluster and next closest cluster, and the $1 - R^2_{ratio}$ are given in Table 1.

HHLA2 and NEBL as biomarkers based on parametric models

We utilized parametric survival models based on: Weibull, exponential, log-normal, log-logistic, and Rayleigh probability distributions to assess whether these 5 selected genes can be potentially prognostic biomarkers or not. The estimated parameters from these models namely: slopes, standard error, z-statistic and p values are given in Table 2. The CLDN8 was statistically significant with p-value less than 0.05 for log-normal and log-logistic models. The exponential, log-normal-log-logistics models showed statistical significant for the gene HHLA2 as the p-values is less than 0.05, whereas the Rayleigh model shows high significant statistically, since the p-value was less than 0.01. The NEBL gene covariate was statistically significant for Weibull and Exponential models, were highly significant based on the Rayleigh model. The SELENOP and THBS2 were not statistically significance since p-values were greater than 0.05. The output results are given in Table 2, estimated $\beta$ coefficients (slopes), standard error and p-values are marked as '*', '**', and '*' which indicates that the p-values are less than: 0.1, 0.05, and 0.01 respectively. Model’s valida-
| Cluster | Variable | Own cluster* | Next closest* | $(1 - R^2)$ Ratio |
|---------|----------|--------------|---------------|------------------|
| Cluster 1 | B3GNT7 | 0.8073 | 0.6218 | 0.5094 |
|         | CA1 | 0.8360 | 0.7355 | 0.6201 |
|         | **CLDN8** | 0.9060 | 0.6788 | **0.2926** |
|         | CWH43 | 0.7842 | 0.5616 | 0.4923 |
|         | TGFBI | 0.7798 | 0.6491 | 0.6275 |
|         | THR8 | 0.5651 | 0.2979 | 0.6193 |
|         | TMIGD1 | 0.8816 | 0.7386 | 0.4530 |
|         | TRPM6 | 0.7922 | 0.6016 | 0.5215 |
| Cluster 2 | BTNL3 | 0.7540 | 0.5785 | 0.5837 |
|         | BTNL8 | 0.2130 | 0.1401 | 0.9152 |
|         | CCDC68 | 0.6915 | 0.5189 | 0.6412 |
|         | **HHLA2** | 0.7988 | 0.5075 | **0.4085** |
|         | NR3C2 | 0.7497 | 0.6068 | 0.6367 |
|         | PLCE1 | 0.7026 | 0.5454 | 0.6542 |
|         | REG1B | 0.0015 | 0.0233 | 1.0224 |
|         | SLC26A3 | 0.7006 | 0.5337 | 0.6420 |
|         | SLC51B | 0.7585 | 0.4303 | 0.4239 |
|         | TSPAN7 | 0.4983 | 0.3890 | 0.8211 |
| Cluster 3 | IFTM1 | 0.8305 | 0.4355 | 0.3002 |
|         | **NEBL** | 0.8305 | 0.2295 | **0.2199** |
| Cluster 4 | C2orf88 | 0.8534 | 0.7389 | 0.5617 |
|         | CEMIP | 0.6997 | 0.5961 | 0.7435 |
|         | CPM | 0.7548 | 0.4932 | 0.4838 |
|         | MT1G | 0.7778 | 0.6363 | 0.6110 |
|         | PDK4 | 0.4648 | 0.1986 | 0.6678 |
|         | **SELENOP** | 0.8177 | 0.5245 | **0.3833** |
|         | SI | 0.7843 | 0.7304 | 0.8000 |
|         | SMPDL3A | 0.7784 | 0.6041 | 0.5597 |
|         | UGT2A3 | 0.5176 | 0.4272 | 0.8422 |
|         | UGT2B17 | 0.7635 | 0.5964 | 0.5858 |
| Cluster 5 | COL1A2 | 0.9025 | 0.4121 | 0.1658 |
|         | SELENBP1 | 0.6345 | 0.4092 | 0.6186 |
|         | **THBS2** | 0.9095 | 0.4178 | **0.1554** |

Table 1: Feature selection based on cluster: * indicates that $R^2$ with respect to the own cluster and next closest.
tion based Tukey-Anscombe approach have been presented in the supplementary materials.

**Validating HHLA2 and NEBL as biomarkers on external dataset**

The external dataset\(^1\) was used to validate the four gene covariates as potential biomarker predictors namely: HHLA2, NEBL, SELENOP and THBS2. For these genes, K-M survival analysis were done and based on the log-rank test, they showed statistical significance, except THBS2 gene, their survival plots are shown in supplementary file.

The parametric survival models: Weibull, exponential, log-normal, log-logistic, and Rayleigh were utilized for building the predictive models. The estimated parameters, standard errors, and p-values are presented in Table 2. HHLA2 gene showed statistical significance for Weibull, exponential and log-normal models. The HHLA2 gene covariate was highly statistically significant with p-value less than 0.01 for the Rayleigh model. The NEBL gene covariate was statistically significant for: Weibull, exponential, log-normal, and log-logistic models with p-value less than 0.05 and was highly statistically significance for p-value less than 0.01, based on the Rayleigh model. Hence, the validation for the gene covariate HHLA2 and NEBL as prognostic biomarkers were tested for internal as well as external datasets (an independent dataset for a set of different population).

**Comparison of normal versus tumor tissues for expression values of HHLA2 and NEBL**

The Wilcoxon test fort two independent samples was applied to compare the expression values of HHLA2 and NEBL genes between normal and tumor tissues, the comparison test have shown statistical significance’s for both the genes HHLA2 and NEBL with p-value less than 0.00001, and they are shown in the Violin plots of Figure 3.

**Quantitative real-time polymerase chain reaction**

We utilized SYBR green based quantitative PCR to analyze the expression levels of HHLA2 gene in matched tumor and normal samples as well as colorectal cancer cell lines. The sequence for HHLA2 primers used is as follows:

Forward primer 5’– TAC AAA GGC AGT GAC CAT TTG G – 3’
Reverse primer 5’– AAG TGT AAA TTC CTT CGT CCA GA – 3’.

HPRT1 was used as housekeeping gene. All experiments were done in triplicate using Applied Biosystems QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific, USA). RNA was extracted from frozen patient tissue samples as well as cell lines using The PureLink RNA Mini Kit (Thermo Fisher Scientific, USA). Expression Suite software was used for relative quantification analysis. Normal sample from the same patient was used as calibrator. For cell lines, CCD841 was used as calibrator.

**Cell Culture**

Colorectal cancer cell lines - CaCO\(_2\), HCT116, HCT8, HT29, RKO, SW480, SW62 and DLD1 were grown in advanced DMEM media supplemented with 10% fetal bovine serum and supplements. CCD841 cell line was used for nor-

\(^1\)CLDN8 gene was not in the external dataset
Figure 3: In the left panel the violin plot for HHLA2 gene depicts the normal vs. tumor tissues gene expression values, based on the Wilcoxon test p-value shows high statistical significance ($p < 0.00001$), moreover shape (thicker or thinner) of the violin plots tell that the thicker part implies the values in that section of the violin has higher frequency, and the thinner part implies lower frequency. Whereas, on the right panel the violin plot for NEBL gene: normal vs. tumor tissues shows the expression values, applying the Wilcoxon p-value shows high statistical significance ($p < 0.00001$), moreover shape (thicker or thinner) of the violin plots tell that the thicker part implies the values in that section of the violin has higher frequency, and the thinner part implies lower frequency,
mal colon cells. All cells were cultured in 5% $CO_2$ and 37°C in a humidified incubator.1.

The expression values for HHLA2 gene in 8 CRC cell lines were analyzed.

![Figure 4: PCR validation](image)

They were compared with the normal colon cell line. All CRC cell lines exhibited downregulation of HHLA2 gene expression. HCT116 cells show highest downregulation whereas HT29 cells were having least change in change expression, and results were presented in supplementary tables.

**Discussion**

Biomarkers are proposed to be critical tools for practicing precision and personalized medicine. In this study we attempted to propose two new prognostic biomarkers for colorectal cancer. We utilized an indigenous dataset from local colorectal cancer patients and selected candidate novel biomarker genes. By using well curated databases, rigorous statistical analyses and prediction models we propose HHLA2 and NEBL genes as the novel prognostic biomarkers. Decreased expression of HHLA2 gene was found to be associated with poor survival in colorectal cancer patients. The quest for discovering biomarker has supported the growth of precision medicine. In colorectal cancer early detection is the key for better prognosis. However, the current clinical practice of using colonoscopy, gFOBT, FIT or other tests have been less adequate for technical and economic reasons. Patient compliance for invasive procedures has been poor. The clinical applications of biomarkers in CRC is pertinent for prognostic stratification, surveillance, and therapy selection [15]. Carcinoembryonic antigen (CEA) is a well-known prognostic biomarker used in the clinic for CRC management. High
levels are associated with cancer progression and can indicate recurrence after surgery. However, high CEA levels are not specific to CRC and can also be found in other malignancies and inflammatory conditions, such as inflammatory bowel disease, liver disease, and pancreatitis [16] and [17]. Microsatellite instability and BRAF mutation status are also now considered for prognostication of CRC [18]. Current and emerging biomarkers also reflect the genomic complexity of CRC, and include a wide range of aberrations such as point mutations, amplifications, fusions and hypermutator phenotypes, in addition to global gene expression subtypes [19]. Gene expression based biomarkers are relatively easier to quantify and are more accurate and sensitive than the traditional IHC based assays. To search for novel biomarkers, we utilized the ‘biomarker discovery’ tool in IPA tool which resulted in a list of 135 genes with no previously established evidence for their role in CRC prognosis. IPA based biomarker discovery has been successfully reported earlier [20] and provided a list of putative candidate novel biomarkers for CRC. There are several reasons not to include all 33 genes while building predictive model, one of the important issue is the multi-collinearity problem which is due to high correlatedness among predictors (it’s well known that gene expression values are highly correlated). Another problematic factor in model building is associated with over- or under-fitting because of high dimensional gene expression data, which have too many genes as predictors with limited numbers of samples or observations. To overcome these issues we utilized the VARCLUS algorithm [13]. Among five genes that were selected for model SELENOP and THBS2 genes are known to be involved in colorectal cancer progression. Downregulation of human SEPP1 [SELENOP] mRNA in colorectal carcinoma cells is associated with colorectal cancer in human [21] [?]. Mutant human THBS2 gene (c.1903G¿A translating to p.G635S [somatic mis-sense]) is observed with adenocarcinoma in human colon (COSMIC: observed in 2 of 24 samples) [?]. Rest of the three genes were completely novel and made good candidates for Biomarker. We attempted five different parametric models and compared them to arrive at most statistically significant tool. This approach has potential benefit of proposing and testing a biomarker with high confidence [?] and prevents high rate of false positives. Rayleigh model showed least minimum standard error. We also used the semi-parametric cox model which is widely used in biomarker discovery [33]. HHLA2 and NEBL genes were found to be significant whereas CLDN8, SELENOP and THBS2 were having higher p-value. HHLA2 gene has been found to be downregulated in CRC in a pan cancer study and that is in conformity with our results. However, results from various databases cannot conclude the favorable or unfavorable prognosis associated with HHLA2 gene expression [25] However, in renal cell carcinoma higher expression of HHLA2 have been associated with poor outcome [26]. Higher expression of HHLA2 have been reported in NSCLC based on IHC studies [?]. Two studies used IHC technique to report higher expression levels of HHLA2 in CRC [28] and [29]. These studies are limited both in technical sensitivity and the sample size as compared to other studies which concluded lower expression levels of HHLA2 in CRC and its prognostic value based on transcriptomic data from multiple databases, including cBioPortal, TCGA, Cistrome, TIMER, On-
comine, Kaplan-Meier, GeneXplain. Since CRC has not benefitted much from immunotherapy especially in non-MSI tumors (which comprise ≥85% of all CRC tumors), a new immune checkpoint molecule, HHLA2, which is widely expressed in PD-1 negative human tumors, may be a promising target. HHLA2 is being suggested as prognostic as well as immunotherapeutic target in other cancer types e.g. Intrahepatic cholangiocarcinomai[30]. Our study provide evidence to support the use of HHLA2 as prognostic biomarker. Further studies are needed to establish its role in targeted immunotherapy. Nebulette gene expression has been reported earlier to be a potential prognostic biomarker based on quantitative real time PCR experiment [31] and support our results based on more robust technique of microarray. Our results provide further evidence for use of NEBL gene expression as an independent predictor for CRC prognosis. Higher expression of NEBL was reported to be predictor for better outcome and is in conformity with our results [32]. Our independent analysis of indigenous data based on local population samples from matched tumor normal samples provide confidence in use of NEBL gene as a putative prognostic biomarker.

With the current scope and importance of personalized medicine, 2 biomarkers namely: HHLA2 and NEBL were identified as potential prognostic biomarkers. These two genes could be utilized in clinical decision making for patients with colorectal cancer. Since their significance were based on the probabilistic theory with the utilization of parametric survival models, whereas the parametric based models have better predictive power to make prediction in comparison with the semi- or non-parametric medicals. Particularly using Rayleigh survival model had shown that it has a better model fit, as it’s evident from the model’s diagnostic assessments via Tukey-Anscombe and Quantile-Quantile plot and had shown high statistical significance with 95% confidence level. The model’s assumptions were validated as well as graphically presented for the semi-parametric Cox model as well. Other evaluation for the model were based on the AIC, BIC, and log-likelihood as well as with the Harrel’s index which gives the predictive probability of 81%, were also applied to assess the performance. The statistical significance for HHLA2 and NEBL genes were further validated on the external dataset available in public domain apart from the validity were tested based on the PCR’s result; apart from comparison of normal versus tumor tissue based upon the Wilcoxon two samples test for HHLA2 and NEBL also exhibited a high statistical significance.

**Conclusion**

Higher expression of HHLA2 and NEBL genes could predict better overall survival in colorectal cancer. This study provides substantial evidence based on indigenous patient data and validated in other datasets. Rigorous experimental validation is presented to support this evidence, apart from predictive statistical survival models.

**Scopes and limitation of this study**
• The HHLA2 and NEBL genes can be utilized as potential prognostic biomarkers in the clinical decision making for patients with CRC.

• The Rayleigh survival model had shown as the most suitable predictive model from the family of parametric survival model and could be used as a suitable and reliable predictive model since parametric models have better statistical power. This is based upon the diagnostics and the validation tools.

• The comparison between the normal and tumor tissues had shown statistically significant for the HHLA2 and NEBL genes.

• The statistical significance for both the genes HHLA2 and NEBL based on parametric, semi- and non-parametric statistical survival models for the internal as well as an external data as an independent dataset on different set of population. The gene HHLA2 was also validated with the help of PCR generated data for a subset of subjects from the internal dataset.

• This study report is based on the retrospective data which may have some bias due to statistical sampling procedure, as the subjects were of mostly older age population.

• The total subjects for the internal dataset had only 41 subjects with event of died subjects were 13, and 28 subjects censored (alive or status unknown). The external dataset was having only 100 subjects and events (dead) were 8 and remaining subjects were either alive or their status were unknown, since very few events were observed in this external dataset though it had suffice one of the requirement for a good model fit to have minimum 5 events per coariate [38]. Another independent dataset TCGA dataset with 469 observations with events (93 dead) and 376 subject were censored at the time of last follow-up, and had shown statistical significance with parametric as well as semi- and non-parametric survival model for the gene covariate HHLA2 but not for the gene NEBL.

• The PCR laboratory test for the gene NEBL had not been done due to logistic as well as limited resources.

• We had also explored to validate on on few other ‘GEO-series’ independent datasets, but the genes HHLA2 and NEBL, did not show statistical significance. This shows the complexity and heterogeneity of subject population as well as poses challenges to build predictive models based on the probabilistic based approach.

• Since the genes HHLA2 and NEBL had shown statistical significant for HHLA2 and NEBL for survival models for internal as well as the external dataset, apart from the laboratory validation for the gene HHLA2 based on the PCR laboratory analysis.
• We are also intending to further explore both these genes in vitro as well as animal based models to confirm whether both these genes could be utilized as prognostic biomarkers, though it’s solely based upon the pending grant from the research center.

References

[1] Dienstmann, R. et al. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer 17, 268, doi:10.1038/nrc.2017.24 (2017)

[2] Turanli, B. et al. A Network-Based Cancer Drug Discovery: From Integrated Multi-Omics Approaches to Precision Medicine. Curr Pharm Des, doi:10.2174/1381612824666181106095959 (2018)

[3] Aziz, M. A., Yousef, Z., Saleh, A. M., Mohammad, S. Al Knawy, B. Towards personalized medicine of colorectal cancer. Crit Rev Oncol Hematol 118, 70-78, doi:10.1016/j.critrevonc.2017.08.007 (2017)

[4] Burke, W. Psaty, B. M. Personalized medicine in the era of genomics. JAMA 298, 1682-1684, doi:10.1001/jama.298.14.1682 (2007)

[5] Tursz, T., Andre, F., Lazar, V., Lacroix, L. Soria, J. C. Implications of personalized medicine—perspective from a cancer center. Nat Rev Clin Oncol 8, 177-183, doi:10.1038/nrclinonc.2010.222 (2011)

[6] Goel, G. Molecular characterization and biomarker identification in colorectal cancer: Toward realization of the precision medicine dream. Cancer Manag Res 10, 5895-5908, doi:10.2147/CMAR.S162967 (2018)

[7] Patil, H. et al. Chasing the personalized medicine dream through biomarker validation in colorectal cancer. Drug Discov Today 22, 111-119, doi:10.1016/j.drudis.2016.09.022 (2017)

[8] Ross, J. S., Torres-Mora, J., Wagle, N., Jennings, T. A. Jones, D. M. Biomarker-based prediction of response to therapy for colorectal cancer: current perspective. Am J Clin Pathol 134, 478-490, doi:10.1309/AJCP2Y8KTDPOAORH (2010).

[9] Oldenhuis, C. N., Oosting, S. F., Gietema, J. A. de Vries, E. G. Prognostic versus predictive value of biomarkers in oncology. Eur J Cancer 44, 946-953, doi:10.1016/j.ejca.2008.03.006 (2008)
[10] Allegra, C. J. et al. American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. J Clin Oncol 27, 2091-2096, doi:10.1200/JCO.2009.21.9170 (2009).

[11] Aziz, M. A. et al. Integrated exon level expression analysis of driver genes explain their role in colorectal cancer. PLoS One 9, e110134, doi:10.1371/journal.pone.0110134 (2014) Eldai, H. et al. Novel genes associated with colorectal cancer are revealed by high resolution cytogenetic analysis in a patient specific manner. PLoS One 8, e76251 (2013).

[12] Eldai, H. et al. Novel genes associated with colorectal cancer are revealed by high resolution cytogenetic analysis in a patient specific manner. PLoS One 8, e76251 (2013).

[13] VARCLUS procedure: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_varclus_overview.htm&docsetVersion=14.3&locale=en

[14] The SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, USA.

[15] Alves Martins BA, de Bulhoes GF, Cavalcanti IN, Martins MM, de Oliveira PG, Martins AMA. Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research. Front Oncol. 2019;9:1284.

[16] Gonzalez-Pons M, Cruz-Correa M. Colorectal Cancer Biomarkers: Where Are We Now? Biomed Res Int. 2015;2015:149014.

[17] Campos-da-Paz M, Dorea JG, Galdino AS, Lacava ZGM, de Fatima Menezes Almeida Santos M. Carcinoembryonic Antigen (CEA) and Hepatic Metastasis in Colorectal Cancer: Update on Biomarker for Clinical and Biotechnological Approaches. Recent Pat Biotechnol. 2018;12(4):269-79.

[18] Bhalla A, Zulfiqar M, Bluth MH. Molecular Diagnostics in Colorectal Carcinoma: Advances and Applications for 2018. Clin Lab Med. 2018;38(2):311-42.

[19] Sveen A, Kopetz S, Lothe RA. Biomarker-guided therapy for colorectal cancer: strength in complexity. Nat Rev Clin Oncol. 2020;17(1):11-32.

[20] Yang Y, Pospisil P, Iyer LK, Adelstein SJ, Kassis AI. Integrative genomic data mining for discovery of potential blood-borne biomarkers for early diagnosis of cancer. PLoS One. 2008;3(11):e3661.
[21] Pesson M, Volant A, Uguen A, Trillet K, De La Grange P, Aubry M, et al. A gene expression and pre-mRNA splicing signature that marks the adenoma-adenocarcinoma progression in colorectal cancer. PLoS One. 2014;9(2):e87761.

[22] Short SP, Williams CS. Selenoproteins in Tumorigenesis and Cancer Progression. Adv Cancer Res. 2017;136:49-83.

[23] Lim B, Mun J, Kim JH, Kim CW, Roh SA, Cho DH, et al. Genome-wide mutation profiles of colorectal tumors and associated liver metastases at the exome and transcriptome levels. Oncotarget. 2015;6(26):22179-90.

[24] Vicini P, Gastonguay MR, Foster DM. Model-based approaches to biomarker discovery and evaluation: a multidisciplinary integrated review. Crit Rev Biomed Eng. 2002;30(4-6):379-418.

[25] Wang B, Ran Z, Liu M, Ou Y. Prognostic Significance of Potential Immune Checkpoint Member HHLA2 in Human Tumors: A Comprehensive Analysis. Front Immunol. 2019;10:1573.

[26] Chen L, Zhu D, Feng J, Zhou Y, Wang Q, Feng H, et al. Overexpression of HHLA2 in human clear cell renal cell carcinoma is significantly associated with poor survival of the patients. Cancer Cell Int. 2019;19:101.

[27] Cheng H, Janakiram M, Borczuk A, Lin J, Qiu W, Liu H, et al. HHLA2, a New Immune Checkpoint Member of the B7 Family, Is Widely Expressed in Human Lung Cancer and Associated with EGFR Mutational Status. Clin Cancer Res. 2017;23(3):825-32.

[28] Zhu Z, Dong W. Overexpression of HHLA2, a member of the B7 family, is associated with worse survival in human colorectal carcinoma. Onco Targets Ther. 2018;11:1563-70.

[29] Janakiram M, Chinai JM, Fineberg S, Fiser A, Montagna C, Medavarapu R, et al. Expression, Clinical Significance, and Receptor Identification of the Newest B7 Family Member HHLA2 Protein. Clin Cancer Res. 2015;21(10):2359-66.

[30] Jing CY, Fu YP, Yi Y, Zhang MX, Zheng SS, Huang JL, et al. HHLA2 in intrahepatic cholangiocarcinoma: an immune checkpoint with prognostic significance and wider expression compared with PD-L1. J Immunother Cancer. 2019;7(1):77.

[31] Hosseini SM, Mahjoubi F, Majidzadeh T, Khaje-Hosseini F, Haghpanah M. Nebulette Expression Is Associated with Lymph Node Metastasis in Patients with Colorectal Cancer. Middle East J Dig Dis. 2018;10(3):174-9.
[32] Qiu X, Feng JR, Wang F, Chen PF, Chen XX, Zhou R, et al. Profiles of differentially expressed genes and overexpression of NEBL indicates a positive prognosis in patients with colorectal cancer. Mol Med Rep. 2018;17(2):3028-34.

[33] Zaixiang Tang, Yueping Shen, Xinyan Zhang, Nengjun Yi, The spike-and-slab lasso Cox model for survival prediction and associated genes detection, Bioinformatics, Volume 33, Issue 18, 15 September 2017, Pages 2799–2807, https://doi.org/10.1093/bioinformatics/btx300

[34] Frank E. Harrell Jr, Robert M. Cali, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. Journal of the American Medical Association, 247 (18):2543-2546, 1982.

[35] Terry Therneau, Patricia Grambsch, and Thomas Fleming, ’Martingale based residuals for survival plots’, Technical Report Series 840, April 1988.

[36] Terry M Therneau, Survival Analysis,’survival’ package https://cran.r-project.org/web/packages/survival/index.html

[37] Alboukadel Kassambara, Claire Gourzones Dmitriev, Surinder Sahota, Thierry Rene, Jerome Moreaux, Hartmut Goldschmidt, Angelos Constantinou, Philippe Pasero, Dirk Hose, Bernard Klein. A DNA repair pathway score predicts survival in human multiple myeloma: the potential for therapeutic strategy. Onco-target, Vol 5, No.9. February 24, 2014.

[38] Eric Vittinghoff, Stephen C. Shiboski, David V. Glidden, Charles E. McCulloch, Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, Springer 2005.

**Ethics approval and consent to participate**
The study was approved by the Ethics Committee of King Abdullah International Medical Research Center for internal review board.

**Consent for publication**
'Not applicable'

**Availability of supporting data**
The data could be obtained from corresponding author upon request.

**Competing interests**
Authors declare no conflict of interest.

**Funding**
'Not applicable'

**Authors’ contributions**
Altaf Khan contributed in writing/editing and statistical analyses, Mohammed Azhar Aiz contributing in writing and IPA analysis. Ghada Alatar and Amal
AlGhamdi had contributed as laboratory related work. Other authors had reviewed the manuscript.

Acknowledgments
Authors acknowledging and extending their thanks to the open-source software the LATEX and the 'R Development Core Team' from the 'R Foundation for Statistical Computing, Vienna, Austria’, for free availability of useful utilities and nice packages. Without these tools the work done in this work was not possible.
Table 2: The parametric models: Weibull, exponential, log-normal, log-logistic, and Rayleigh based probability distribution are given. The estimated parameters, standard errors and statistically significant parameters are marked with ‘*, **, and ***’ for $\alpha$ significance values as: 0.1, 0.05 and 0.01.

| Dependent variable: overall survival in days | Weibull | Exponential | Log-normal | Log-logistic | Rayleigh |
|---------------------------------------------|---------|-------------|------------|--------------|----------|
| CLDN8                                       | −0.951  | −0.735*     | −1.356**   | −1.262**     | −0.359   |
| (0.584)                                     | (0.424) | (0.628)     | (0.618)    | (0.242)      |          |
| HHLA2                                       | 0.912*  | 0.732***    | 1.027**    | 1.042**      | 0.439*** |
| (0.474)                                     | (0.334) | (0.508)     | (0.510)    | (0.169)      |          |
| NEBL                                        | 1.242** | 1.110***    | 1.062*     | 0.981*       | 0.878*** |
| (0.509)                                     | (0.396) | (0.560)     | (0.503)    | (0.274)      |          |
| SELENOP                                     | 0.586   | 0.450       | 0.830      | 0.746        | 0.243    |
| (0.418)                                     | (0.297) | (0.509)     | (0.483)    | (0.158)      |          |
| THBS2                                       | 0.207   | 0.160       | 0.027      | 0.128        | 0.066    |
| (0.449)                                     | (0.334) | (0.390)     | (0.381)    | (0.154)      |          |
| Constant                                    | −12.516 | −9.694      | −10.649    | −10.768      | −4.897   |
| (9.407)                                     | (6.979) | (9.410)     | (8.495)    | (4.102)      |          |

Observations | 41 | 41 | 41 | 41 | 41
Log Likelihood | −109.406 | −110.221 | −107.867 | −108.204 | −122.674
$\chi^2$ (df = 5) | 14.819** | 17.600*** | 16.787*** | 16.718*** | 23.027***

Note: *p<0.1; **p<0.05; ***p<0.01