available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/molonc

A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients

Seon-Kyu Kim[a,1], Seon-Young Kim[a,1], Jeong-Hwan Kim[a], Seon Ae Roh[b,c], Dong-Hyung Cho[c,d], Yong Sung Kim[a,c,**], Jin Cheon Kim[b,c,*]

[a]Medical Genomics Research Centre, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea
[b]Department of Surgery, University of Ulsan College of Medicine, Seoul, Korea
[c]Department of Cancer Research, Institute of Innovative Cancer Research and Asan Institute for Life Sciences, Asan Medical Centre, Seoul, Korea
[d]Graduate School of East-West Medical Science, Kyung Hee University, Gyeonggi-do, Korea

ARTICLE INFO

Article history:
Received 13 June 2014
Accepted 16 June 2014
Available online 4 July 2014

Keywords:
Colorectal cancer
Metastasis
Markers
Prognosis
Chemotherapy

ABSTRACT

Colorectal cancer (CRC) patients frequently experience disease recurrence and distant metastasis. This study aimed to identify prognostic indicators, including individual responses to chemotherapy, in CRC patients. RNA-seq data was generated using 54 samples (normal colon, primary CRC, and liver metastases) from 18 CRC patients and genes associated with CRC aggressiveness were identified. A risk score based on these genes was developed and validated in four independent CRC patient cohorts ($n = 1063$). Diverse statistical methods were applied to validate the risk scoring system, including a generalized linear model likelihood ratio test, Kaplan–Meier curves, a log-rank test, and the Cox model. *TREM1* and *CTGF* were identified as two activated regulators associated with CRC aggressiveness. A risk score based on 19 genes regulated by *TREM1* or *CTGF* activation (TCA19) was a significant prognostic indicator. In multivariate and subset analyses based on pathological staging, TCA19 was an independent risk factor (HR = 1.894, 95% CI = 1.227–2.809, P = 0.002). Subset stratification in stage III patients revealed that TCA19 had prognostic potential and identified patients who would benefit from adjuvant chemotherapy, regardless of age. The TCA19 predictor represents a novel diagnostic tool for identifying high-risk CRC patients and possibly predicting the response to adjuvant chemotherapy.

Abbreviations: CRC, colorectal cancer; TCA, *TREM1* or *CTGF* activation; GW, genome-wide; PC, primary cancer; MC, metastatic cancer; NC, normal colon; MSI, microsatellite instability; MSI-H, high-frequency MSI; MSI-L, low-frequency MSI; MSS, microsatellite stable; DFS, disease-free survival; FPKM, fragments per kilobase of transcript per million fragments mapped; GLM, generalized linear model; MA, Mutation assessor; SIFT, sorting intolerant from tolerant; PP2, Polyphen2; ROC, receiver operating characteristic; AUC, area under curve; MMR, DNA mismatch repair; dMMR, deficient MMR; pMMR, proficient MMR; MDA114, 114 gene MD Anderson Cancer Centre prognostic predictor; OncoDX, Oncotype DX; EMT, epithelial–mesenchymal transition; CSC, cancer stem cell.

[*] *Corresponding author*. Department of Surgery, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 138-736, Korea. Tel.: +82 2 3010 3489; fax: +82 2 474 9027.

[**] *Corresponding author*. Medical Genomics Research Centre, Korea Research Institute of Bioscience and Biotechnology, 52 Eoeun-dong, Yuseong-gu, Daejeon 305-333, Korea. Tel.: +82 42 879 8110; fax: +82 42 879 8119.

E-mail addresses: yongsung@kribb.re.kr (Y.S. Kim), jckim@amc.seoul.kr, jckimmd@naver.com (J.C. Kim).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.molonc.2014.06.016
1574-7891/© 2014 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

1. Introduction

Colorectal cancer (CRC) is a leading cause of cancer death worldwide (Jemal et al., 2011). Pathological staging is the gold standard for prognosis, concurrently selecting patients for adjuvant chemotherapy (Hari et al., 2013). However, pathological staging frequently fails to accurately predict recurrence in patients undergoing curative surgery for locally advanced CRC. A large number of CRC patients relapse after complete surgical resection, and approximately half of patients with stage III disease relapse within 5 years (Carlsson et al., 1987; Midgley and Kerr, 1999). The two most common sites of recurrence in CRC are the liver and lung (Cunningham et al., 2010). Recurrence at these sites is frequently accompanied by other systemic metastasis and is often fatal. In addition, tumours with similar histopathological appearances often manifest significantly different clinical behaviour. A recent study showed that more than a half of patients with advanced CRC experience clinical responses after systemic chemotherapy including targeted regimens (Midgley et al., 2009). Although patients with stage III CRC are routinely offered chemotherapy, elderly patients (≥75 years) are often excluded from the therapy, due to frailty coping with unfavourable adverse events (Sveen et al., 2013). A guideline is thereby needed for adjuvant treatment in the elderly patients with advanced CRC. Otherwise, it is indispensable to understand the molecular heterogeneity associated with different responses to chemotherapy and to develop models that identify patients who would benefit the most from adjuvant chemotherapy.

Recent genome-wide (GW) gene expression studies in various cancers strongly indicate that tumour heterogeneity is reflected in gene expression patterns. A number of recent GW studies have successfully identified several distinct subtypes of CRC that exhibit heterogeneous biological and clinical behaviours (Cancer Genome Atlas Network, 2012; De Sousa et al., 2013; Marisa et al., 2013; Sadanandam et al., 2013), implying that there may be several prognostic signatures for CRC, each of which corresponds to different tumour behaviour. Unfortunately, few surrogate signatures have been shown to be clinically feasible, particularly as prognostic markers. Although several molecular classification criteria, such as microsatellite instability (MSI), CpG island methylation phenotype (CIMP), chromosomal instability (CIN), and *BRAF* and *KRAS* mutations, are currently used (Jass, 2007; Kang, 2011; Shen et al., 2007), such criteria clearly need to be further consolidated for efficient responses to individualized chemotherapy (Marisa et al., 2013; Oh et al., 2012). These studies strongly suggest that the characteristic molecular changes that occur during CRC tumorigenesis and progression can be used to identify molecular signatures that are able to predict patient prognosis and response to therapy.

The current study investigated putative genetic signatures associated with CRC aggressiveness. The GW identification based on RNA-seq was used to identify genes driving tumorigenesis and progression, and then we examined whether these signatures could identify patients who would benefit from adjuvant chemotherapy.

2. Materials and methods

2.1. *Patients and tissue samples*

The current study used 18 matched primary CRC (PC) samples, synchronous liver metastases (MC) that were histologically identified as adenocarcinoma, and their normal colonic epithelium (>5 cm from the tumour border) (NC), respectively. All patients were treated at the Asan Medical Centre (Seoul, Korea) between May 2011 and February 2012 (AMC cohort). Nine of these patients received curative surgery (R0 resection) for both PC and MC. All patients in the AMC cohort showed a status of microsatellite stable (MSS) except for one patient showing high-frequency MSI (MSI-H). All tumour samples were acquired following patient consent for tissue sample donation and examination. The study protocol was approved by the Institutional Review Board for Human Genetic and Genomic Research (registration no. 2009-0091), in accordance with the Declaration of Helsinki.

2.2. *RNA extraction and RNA-seq experiments*

Total RNA was isolated using the RNeasy Mini Kit (Qiagen, CA), according to the manufacturer's protocol. The quality and integrity of the RNA were confirmed by agarose gel electrophoresis and ethidium bromide staining, followed by visual examination under ultraviolet light. The sequencing library was prepared using the TruSeq RNA Sample Preparation kit v2 (Illumina, CA) according to the manufacturer's instructions. In brief, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads, fragmented, and converted into cDNA. Adapters were then ligated to the cDNA and the fragments were amplified by PCR. Sequencing was performed in paired-end reads (2 × 100 bp) using a Hiseq-2000 (Illumina).

2.3. *RNA-seq data processing*

Reference genome sequence data from *Homo sapiens* were obtained from the University of California Santa Cruz Genome Browser Gateway (assembly ID: hg19). The reference genome index was built using the Bowtie2-build component of Bowtie2 (ver. 2.0) and SAMtools (ver. 0.1.18). Tophat2 was applied to tissue samples for mapping reads to the reference genome (ver. 2.0). The statistics of the mapping activity of Tophat2 are described in Supplementary Table 1. The data set generated by RNA-seq is available in the NCBI Gene Expression Omnibus public database under the data series accession number GSE50760.

2.4. *Public gene expression data sets and study design*

A gene expression data set from the Hubrecht Institute was used as a validation data set to verify expression differences in selected genes among the NC, PC, and MC tissue groups (HI cohort, GSE14297, $n = 48$). In the HI cohort, PC and matched MC from the same patient were assayed by gene expression profiling in 18 pairs. Additionally, expression data of NC and normal liver tissue exist in the dataset (Stange et al., 2010).

To develop a risk score classifier for predicting CRC outcomes, data collected for the Cartes d'Identité des Tumeurs (CIT) program from the French Ligue Nationale Contre le Cancer was used as a development data set (CIT cohort, GSE39582, $n = 566$). In the CIT cohort, patients who received preoperative chemotherapy and/or radiation therapy and those with primary rectal cancer were excluded. Clinical and pathologic data were extracted from the medical records and patients were staged according to the American Joint Committee on Cancer (AJCC) staging system (Hari et al., 2013) and monitored for relapse (distant and/or locoregional recurrence) (Marisa et al., 2013). To validate our classifier, we used a gene expression data set that was generated from fresh-frozen tumour specimens retrieved from the tissue banks of the Royal Melbourne Hospital, Western Hospital, and Peter MacCallum Cancer Centre in Australia and the H. Lee Moffitt Cancer Centre in the USA (AUS cohort, GSE14333, $n = 229$). In the AUS cohort, individuals who had received preoperative chemotherapy and/or radiotherapy or samples in which tumour-derived total RNA was inadequate for microarray analysis [RNA integrity number (RIN) < 6] were excluded. Among them, 22 of 94 patients who had stage II disease and 64 of 91 patients who had stage III disease received standard adjuvant chemotherapy (either single-agent treatment with 5-fluorouracil or capecitabine, or a combination of 5-fluorouracil and oxaliplatin) or postoperative concurrent chemoradiotherapy (50.4 Gy in 28 fractions with concurrent 5-fluorouracil), according to hospital protocols. For stage II and III patients in the AUS cohort, follow-up and additional clinical data, including patient gender and TNM staging, were collected by Bio-grid Australia for Australian patients and the Moffitt Cancer Center Tumor Registry for U.S. patients (Jorissen et al., 2009). Two data sets from the Academic Medical Centre at the University of Amsterdam (GSE33113, $n = 90$) and the Institut Paoli-Calmettes in France (GSE37892, $n = 130$) were pooled to validate the classifier (UAPC cohort, $n = 220$). In the GSE33113, the patients were treated in the years 1997–2006, extensive medical records were kept of them, and long-term clinical follow-up was available for the large majority. Both paraffin-embedded and fresh frozen tissue was available from all patients, which was used to derive gene expression profiles in the GSE33113 (de Sousa et al., 2011). In the GSE37892, a series of 130 CRCs with stage II and III disease was retained and expression profiles were established on oligonucleotide microarrays. The primary end-point was disease-free survival (DFS), which was defined as the time from surgery to the first confirmed relapse. DFS was available for the CIT, AUS, and UAPC cohorts. All gene expression data sets were generated using the Affymetrix Human Genome U133 Plus 2.0 platform. The current study design and validation strategy adhered to the REMARK guideline (McShane et al., 2005), shown in Supplementary Figure 1.

2.5. *Transcriptomic profiling and significance test*

To obtain an mRNA expression landscape across tissue samples, a hierarchical clustering algorithm was applied that used the centred correlation coefficient as the measure of similarity and complete linkage clustering. For cluster analysis, the fragments per kilobase of transcript per million fragments mapped (FPKM) of each sample was used to estimate the expression level of each gene. The FPKM data were normalized by the quantile method, $\log_2$-transformed, and median-centred across genes and samples.

To estimate the significance of differences in gene expression among sample subgroups, an EdgeR package that uses a negative binomial model was used to detect differentially expressed genes from count data (Robinson et al., 2010). The gene count dispersion was estimated using a Cox-Reid profile-adjusted likelihood method. After model fitting and estimation of dispersion, differentially expressed genes were selected using a generalized linear model (GLM) likelihood ratio test that specifies probability distributions according to the mean-variance relationship. The GLM likelihood ratio test is based on the principle of fitting negative binomial GLMs with Cox-Reid dispersion estimates. Expression differences in genes were considered statistically significant if the P-value was <0.001 and the fold difference in expression between two sample groups was ≥ 2. When comparing liver metastatic cancers with primary CRCs, liver-specific genes (309 genes) were filtered out using the TiGER database (Liu et al., 2008) and differentially expressed genes were selected from among the remaining genes.

SAMtools and VarScan (ver. 2.3.4) were used to identify somatic mutations from RNA-seq data. First, the pileup command in SAMtools was used to generate pileup files from each aligned bam file. Then, VarScan's somatic command was applied to identify somatic mutations from two conditions (tumour *vs.* normal or metastatic *vs.* primary tumour). dbNSFP (ver. 2.04), an integrated database of functional prediction for non-synonymous SNPs, was used to select functionally significant non-synonymous mutations from VarScan outputs. Multiple prediction scores were considered from three prediction algorithms [Mutation assessor (MA), sorting intolerant from tolerant (SIFT), and Polyphen2 (PP2)] to determine significant variants.

2.6. *Gene set and upstream regulator analysis*

Gene set enrichment analysis was carried out to identify the most significant gene sets associated with disease processes, molecular and cellular functions, and physiological and developmental processes. The significance of over-represented gene sets was estimated by Fisher's exact test. To identify predominant upstream regulators that account for the observed gene expression changes, we performed an upstream regulator analysis, which determined the number of known targets of each regulator that were present in the data set and compared their direction of change to what is expected from the previously reported literature to predict likely relevant regulators. For each potential regulator, an overlap P-value and an activation Z-score were estimated. The overlap P-value, estimated by Fisher's exact test, measures whether there is a statistically significant overlap between the genes in a data set and the genes regulated by a regulator. The activation Z-score is used to infer likely activation states of upstream regulators based on comparison with a model that assigns random regulation direction. A positive or negative activation Z-score indicates that a potential upstream regulator was activated or inhibited, respectively. Gene set enrichment and upstream regulator analyses were performed using the Ingenuity Pathway Analysis (IPA) Tool.

2.7. *Risk score development*

To develop an easy-to-use risk score, we adopted a previously developed strategy using the Cox regression coefficient for the genes in the signature from the CIT cohort (Kim et al., 2012a; Paik et al., 2004). The risk score for each patient was calculated as the sum of each gene's score, which was derived by multiplying the expression level of a gene by its corresponding coefficient (Risk score $= \sum$ Cox coefficient of gene $G_i \times$ expression value of gene G_i). The patients were then divided into two groups (*i.e.*, high- or low-risk of recurrence) using the median cut-off of the risk score as a threshold. The coefficient and the threshold values obtained from the CIT cohort were directly applied to data from the AUS and UAPC cohorts to dichotomize the patients into high- and low-risk groups. Time-dependent receiver operating characteristic (ROC) curves of DFS using the nearest neighbour estimation method with a cut-off value of 36 months were used to identify a small number of genes that strongly retain prognostic value (Heagerty et al., 2000). Based on area under curve (AUC) values derived from ROC analysis, top genes with the highest or lowest AUC values were selected to generate an optimal signature.

2.8. *Random classifiers generation*

We randomly selected genes as many as our signature and obtained their regression coefficients with the median cut-off of risk score in the CIT cohort. Then, they were directly applied to the data from the AUS cohort to divide the patients into high-risk and low-risk groups. The prognostic values were estimated in all patients, stage II, III, or elderly patients with stage III disease. Predictive values for chemotherapy treatment were evaluated in high-risk or low-risk groups in stage III derived from a random classifier. To estimate performance of randomly generated classifiers, a bootstrap method (1000 resampling) was used to calculate the 95% confidence interval (CI) for -$\log_{10}$ (log-rank *P*-value).

2.9. *Other statistical analysis*

To estimate the significance of gene expression differences between subgroups, two-sample *t*-tests were performed for each gene. The Kaplan–Meier method was used to calculate the time to DFS, and the difference in survival between two groups was assessed using log-rank statistics. The prognostic association between the signature and potential risk factors was assessed using multivariate Cox proportional hazard regression models. A backward-forward step procedure (function *step*, R package stats) was applied to optimize the multivariate model with the most informative variables (Venables et al., 2002). To assess the strength of the correlation between outcomes predicted by the different classifiers, Cramer's V statistics and two-way contingency-table analyses were applied. Statistical analysis was carried out using R language environment software (ver. 3.0.1).

3. Results

3.1. *Baseline characteristics*

The baseline characteristics of CRC patients in the AMC and HI cohorts are shown in Supplementary Table 2. Additionally, the baseline patient characteristics of the other three cohorts for developing the prognostic classifier are described in Supplementary Table 3. Among the three classifier development cohorts, adjuvant chemotherapy data were available for the CIT and AUS cohorts. Among the 795 patients in these cohorts, 320 patients received standard adjuvant chemotherapy, while the remaining 475 patients did not receive any chemotherapy.

3.2. *Differential molecular characteristics among CRC patient samples with distinct phenotypes*

To identify a gene set significantly associated with tumorigenesis and progression of CRC, we applied various analysis methods to trio samples (NC, PC, and MC) in the AMC cohort. Hierarchical clustering analysis was initially applied to gene expression data to assess the molecular characteristics of the different sample groups. Unsupervised hierarchical clustering analysis of gene expression data yielded three major clusters: the NC, PC, and MC groups (Supplementary Figure 2). Accordingly, gene expression patterns were readily distinguishable among the PC, MC, and NC tissues.

Genes differentially expressed among the three tissue groups were next identified. When PC and MC tissues were compared, liver-specific genes (309 genes) were filtered out using the TiGER database. Genes associated with tumorigenesis and metastasis were then identified by a Venn diagram comparison between the two gene lists generated using the GLM likelihood ratio test (Figure 1A, $P < 0.001$). Gene list "A" represents genes differentially expressed between the NC and PC groups, and gene list "B" represents genes differentially expressed between the PC and MC groups. When the two gene lists were compared, three different patterns were observed: only A (2018 genes), A and B (843 genes), and only B (1003 genes) (Figure 1B). Genes in the only A category had expression patterns associated with tumorigenesis, while genes in the only B category had expression patterns associated with liver metastasis. Genes in both A and B categories were common to both tumorigenesis and liver metastasis in CRC.

From the RNA-seq data, genes with significant somatic sequence changes between different tissue groups were also identified. For gene selection, three criteria were applied: 1) only somatic mutation was considered, 2) somatic mutation of the same gene was observed in more than two patients, and 3) somatic mutation was predicted to have a functional significance (based on cut-off values of the three scores: MA score > 1.5, SIFT score < 0.05, or PP2 score > 0.9). According to these criteria, we obtained two lists of genes that had significant sequence changes between the NC and PC groups (36 genes identified) or the PC and MC groups (57 genes identified) (Supplementary Tables 4 and 5).

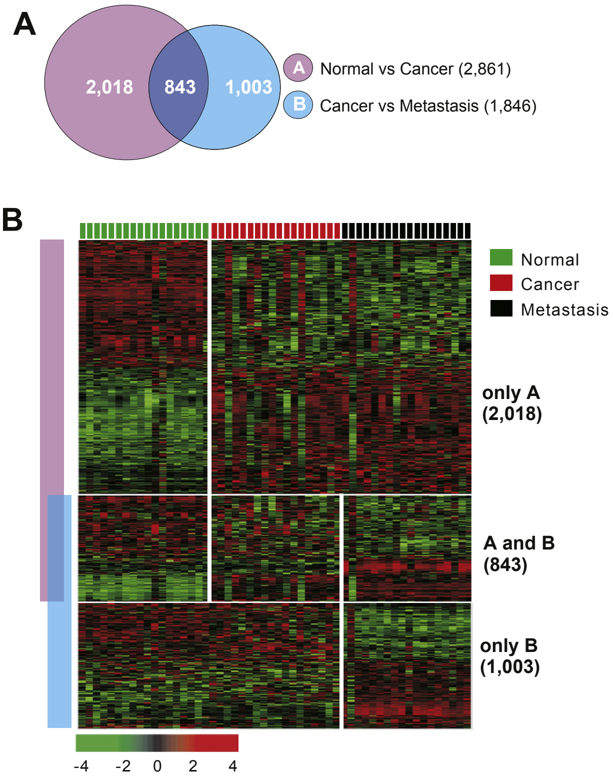

Figure 1 – Comparative analysis of differentially expressed genes among the tissue groups. (A) Venn diagram of genes selected by the GLM likelihood ratio test using EdgeR software in the distinct tissue groups. Genes in the purple circle (gene list A) represent those differentially expressed between normal-looking surrounding colon mucosae (NC) and primary colorectal cancers (PC). Genes in the blue circle (gene list B) represent those differentially expressed between PC and cancers that had metastasised to the liver (MC). Cut-off criteria of a *P*-value of less than 0.001 and a 2-fold or greater relative difference were applied to select genes whose expression were significantly different between the two groups. (B) Expression patterns of selected genes in the Venn diagram. The data are presented in matrix format, in which rows represent individual genes and columns represent each tissue. The red and green colours reflect high and low expression levels, respectively.

3.3. *Activated regulators in the tumorigenesis and liver metastasis of CRC*

A gene set enrichment test was performed to identify biological characteristics of genes associated with tumorigenesis and metastasis using the IPA software. From 843 genes in the A and B category (Figure 1A), 224 genes were eliminated that were not persistently increased or decreased between the PC group and the NC or MC groups. A tumorigenesis-associated gene set was defined by the combination of 2018 genes in the only A category (Figure 1A), 36 variant genes, and 619 genes in the A and B category (a total of 2671 genes, including two overlapping genes). Similarly, a metastasis-associated gene set was defined by the combination of 1003 genes in the only B category (Figure 1A), 57 variant genes, and 619 genes in the A and B category (a total of 1673 genes, including six overlapping genes). When the 2671 and 1673 genes were analysed using IPA, genes involved in cancer, gastrointestinal disease, cellular growth and proliferation, and cell death and survival were significantly enriched in both gene sets. Additionally, genes involved in the inflammatory response, immune cell trafficking, and inflammatory diseases were also significantly enriched (Supplementary Figure 3). These results demonstrate that the two selected gene lists reflect common biological characteristics, indicating that the pathological processes involved in tumorigenesis and liver metastasis of CRC share many biological functions.

The enriched genes included several important regulators. Among them, *TREM1* and *CTGF* were the two predominant regulators activated during CRC tumorigenesis and metastasis (Table 1 and Supplementary Figures 4 and 5). The expression level of *TREM1* was significantly higher in the PC and MC groups than in the NC group (two-sample *t*-test, $P = 8.5 \times 10^{-7}$ and 2.68×10^{-7}, respectively; Supplementary Figure 6A), indicating that activation of the *TREM1* signalling network may be a key event associated with CRC aggressiveness. Although no significant difference in the expression of *CTGF* was observed among the tissue groups ($P = 0.17$ for NC *vs.* PC and 0.27 for NC *vs.* MC; Supplementary Figure 6B), *CTGF* was strongly associated with CRC tumorigenesis and metastasis (Table 1), as the activation score for *CTGF* was estimated by differentially expressed or variant molecules directly interconnected with *CTGF* (Supplementary Figure 5).

To validate the differences in *TREM1* and *CTGF* expression among distinct tissue groups (NC, PC, and MC), the expression levels of the two genes were analysed using gene expression data from the HI cohort. The expression level of *TREM1* was significantly higher in the PC and MC groups than in the NC group (two-sample *t*-test, $P = 0.01$ and 3.5×10^{-4}, respectively), providing confidence in the validity of this molecule as a CRC marker. Furthermore, there was a significant difference in *TREM1* expression between the PC and MC groups (two-sample *t*-test, $P = 0.02$; Supplementary Figure 6C). On the other hand, the *CTGF* expression was not different among these three groups (Supplementary Figure 6D), consistent with the results of the RNA-seq data in the AMC cohort.

3.4. *Development of a risk score using TREM1 and CTGF regulatory networks and its validation in independent cohorts*

Based on our findings, the prognostic value of a set of genes regulated by *TREM1* or *CTGF* was evaluated in additional CRC cohorts. A total of 66 genes regulated by *TREM1* or *CTGF* from the RNA-seq data (Supplementary Figures 4 and 5) were used to generate a risk score classifier defined by *TREM1* or *CTGF* activation (TCA66), and this score was subsequently used as a risk assessment model for CRC prognosis in the CIT cohort. The risk score for each patient in the CIT cohort was calculated using the regression coefficient of each of the 66 genes (130 unique probes) (Supplementary Table 6). Using the median cut-off of risk score (8.410), the CRC samples in the CIT cohort were divided into two groups with either high or low TCA66 scores (Supplementary Figure 7A). The DFS rates were significantly different between the two groups in a log-rank test analysis ($P = 5.62 \times 10^{-7}$; Supplementary Figure 7B). To validate the TCA66 scoring system, the coefficient values and median cut-off derived from

Table 1 – Prediction of activated upstream regulators during tumorigenesis or metastasis.

Upstream regulator	Molecule type	Activation Z-score (normal vs. primary)	Activation Z-score (primary vs. metastasis)	P-value* (normal vs. primary)	P-value* (primary vs. metastasis)
TREM1	Other	3.57	1.71	5.98×10^{-7}	0.01
EGFR	Kinase	3.39	−0.39	8.38×10^{-4}	0.01
IL1B	Cytokine	2.92	−0.92	4.78×10^{-5}	0.002
PI3K	Complex	2.77	−0.48	0.23	0.04
IL1A	Cytokine	2.59	−0.88	0.09	0.04
TNF	Cytokine	2.54	−1.47	0.003	0.02
CTGF	Growth factor	2.2	0.88	2.34×10^{-4}	3.02×10^{-4}
EZH2	Transcription regulator	1.34	−0.43	5.88×10^{-4}	0.002
CTNNB1	Transcription regulator	0.89	−0.65	8.78×10^{-8}	0.001
CXCR4	G-protein coupled receptor	−0.15	−1.98	0.003	0.02
Hedgehog	Group	−0.2	−1.07	0.003	0.009
HIC1	Transcription regulator	−0.32	1.41	2.23×10^{-4}	0.04
SP1	Transcription regulator	−0.38	0.82	5.92×10^{-6}	3.81×10^{-7}
IL4	Cytokine	−0.74	−0.22	0.02	0.04
RUNX2	Transcription regulator	−1.05	−1.17	0.002	0.002
KIAA1524	Other	−1.39	−1.55	0.01	0.006
Oestrogen receptor	Group	−2.28	0.73	2.08×10^{-5}	0.007

*P-values were calculated by Fisher's exact test based on the number of interconnected genes with upstream regulators.
It was predicted that an upstream regulator was activated when the activated Z-score was greater than 0, while it was predicted that an upstream regulator was inhibited when the activated Z-score was less than 0. This prediction was carried out using Ingenuity Pathway Analysis software™.

the CIT cohort were directly applied to the gene expression data from the AUS cohort to dichotomize the patients into high-risk and low-risk groups. Kaplan–Meier estimation revealed significant differences in DFS between the two subgroups (log-rank test, $P = 2.14 \times 10^{-4}$; Supplementary Figure 7C).

To select a small number of genes retaining significant prognostic power, time-dependent ROC analysis based on three-year survival was performed. Among the 66 genes in the TCA66, 19 significant genes (32 unique probes) ($P < 0.05$, Cox regression analysis) with the highest or lowest AUC scores (AUC < 0.45 or AUC > 0.55) were heuristically selected (Supplementary Table 6). A risk score classifier was generated based on the 19 genes in the CIT cohort (TCA19). When the TCA19 classifier was applied in the CIT cohort with a median cut-off (8.053), the DFS period was significantly shorter in the high-risk patient group than in the low-risk patient group (log-rank test, $P = 4.42 \times 10^{-8}$; Figure 2B). When the coefficient and median cut-off values were directly applied to the AUS cohort, the recurrence rate of the high-risk patients was significantly higher than that of the low-risk patients (log-rank test, $P = 7.52 \times 10^{-4}$; Figure 2C). Since DFS data were also available from two other cohorts (GSE33113 and GSE37892), the association between DFS and the TCA19 classifier was further assessed. For this validation, gene expression data from two cohorts (UAPC cohort, $n = 220$) were pooled and the same procedure was applied. The recurrence rate of the high-risk subgroup by TCA19 was significantly higher than those of the low-risk subgroup (log-rank test, $P = 0.005$; Figure 2D).

3.5. *The TCA19 classifier is an independent risk factor for DFS in CRC*

To estimate the independence of the TCA19 predictor, gene expression data was pooled from two validation cohorts, the AUS and UAPC cohorts ($n = 449$), in which patients were stratified by AJCC stage (Hari et al., 2013). In this analysis, stage I patients were excluded for lack of DFS events and DFS data from stage IV patients were not available. When the TCA19-based stratification was applied to stage II–III patients in the pooled cohort, the population of high-risk patients with stage III disease showed significantly worse DFS than low-risk patients ($P = 0.026$; Figure 3B), whereas there was no significant risk difference in DFS between high- and low-risk patient groups with stage II disease ($P = 0.326$; Figure 3A). This result indicates that the TCA19 classifier is a potential predictor of DFS in advanced CRC patients. In the pooled cohort, the prognostic association between the signature and other potential risk factors for DFS was also evaluated by multivariate Cox regression analysis. TCA19 was found to be an independent risk factor for DFS (HR = 1.894, 95% CI = 1.227–2.809, $P = 0.002$; Table 2) even after applying a variable selection procedure. We also carried out another multivariate analysis in the AUS cohort, and found that TCA19 still retained its statistical significance for DFS (HR = 2.24, 95% CI = 1.22–4.114, $P = 0.009$; Table 2).

Additionally, Cox proportional hazard regression model was used to analyse the interactions between TCA19 classifier and AJCC stage. In the CIT cohort, the interaction of TCA19 with stage reached a significant level ($P < 2.0 \times 10^{-16}$) with the estimated HRs for TCA19 of 1.55 in stage II (95% CI 1.103–3.119, $P = 0.019$) and 2.477 in stage III (95% CI 1.503–4.08, $P = 3.693 \times 10^{-4}$; Supplementary Figure 8A). In the cohort pooled with AUS and UAPC cohorts, a strong interaction between TCA19 and stage was also observed ($P = 1.185 \times 10^{-11}$). The HRs for TCA19 in stage II and III were 2.038 (95% CI 1.092–3.802, $P = 0.025$) and 1.778 (95% CI 1.063–2.972, $P = 0.028$), respectively (Supplementary Figure 8B). These results demonstrate that TCA19 classifier is

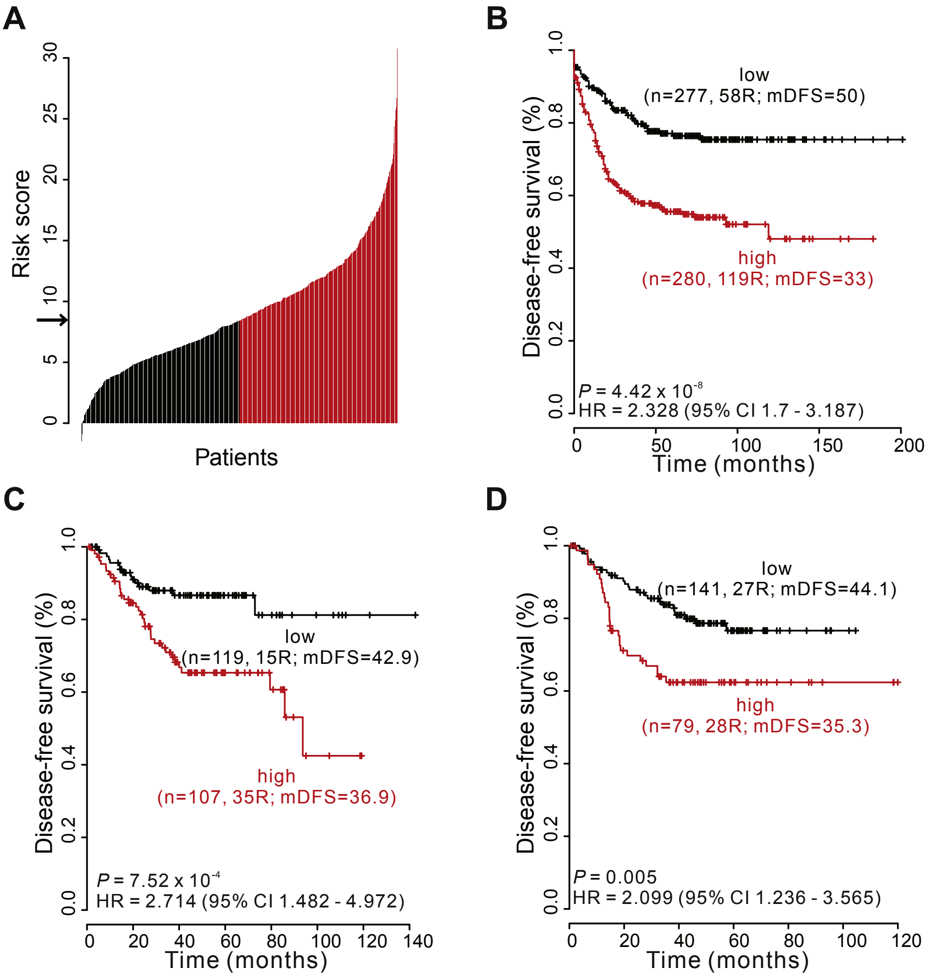

Figure 2 – CRC patient stratification with the TCA19 predictor. (A) Risk score by TCA19 in the CIT cohort. Each bar indicates the risk score for an individual patient. (B) Kaplan–Meier curves of two subgroups in the CIT cohort stratified by TCA19 risk score. (C) Kaplan–Meier curves of two subgroups in the AUS cohort stratified by the TCA19 risk score derived from the CIT cohort. (D) Kaplan–Meier curves of two subgroups in the UAPC cohort stratified by the TCA19 risk score derived from the CIT cohort. *P*-values were obtained by log-rank test. mDFS, median disease-free survival.

strongly interactive and independent with the current staging system.

We further tried to evaluate an association between the TCA19 and MSI status. Among the patient cohorts, DNA mismatch repair (MMR) status data were available in the CIT cohort [deficient MMR (dMMR) or proficient MMR (pMMR)]. Because CRCs with dMMR show a MSI-H and cancers with pMMR show low-frequency MSI (MSI-L) or MSS (Sinicrope et al., 2011), a multivariate Cox regression analysis in the CIT cohort was carried out with MMR categories. TCA19 was an independent risk factor for DFS (HR = 1.952, 95% CI = 1.407–2.708, $P = 6.155 \times 10^{-5}$) with stage and MMR status (Supplementary Table 7). Additionally, the patients were stratified according to the MMR status and prognostic value of each subgroup was estimated. We successfully identified a population of high-risk patients in both MMR conditions (log-rank test, each $P < 0.05$, respectively; Supplementary Figure 9). This finding strongly demonstrates that the TCA19 classifier is independent of the current MMR (or MSI) categories.

3.6. *The TCA19 classifier is associated with DFS after adjuvant chemotherapy*

Among the validation cohorts for the TCA19 predictor, adjuvant chemotherapy data were available for the patients from the AUS cohort. We investigated whether the TCA19 classifier could predict patients who would benefit from adjuvant chemotherapy. The analysis was performed on patients in AJCC stage III ($n = 91$) that were known to have experienced prolonged survival following adjuvant chemotherapy (Laurie et al., 1989; Moertel et al., 1990). When estimating prognostic value in stage III patients, as expected, TCA19 identified high-risk patients for DFS, consistent with the assessment including all patients (Figure 4A and B). Interestingly, when evaluating in elderly patients with CRC stage III (≥75 years, $n = 23$, 8 recurrences), TCA19 also successfully identified high-risk patients (Figure 4C), indicating that TCA19 classifier had a significant prognostic potential, even in elderly patients with advanced CRC. To assess the predictive value of the TCA19 classifier in stage III patients, CRC patients were

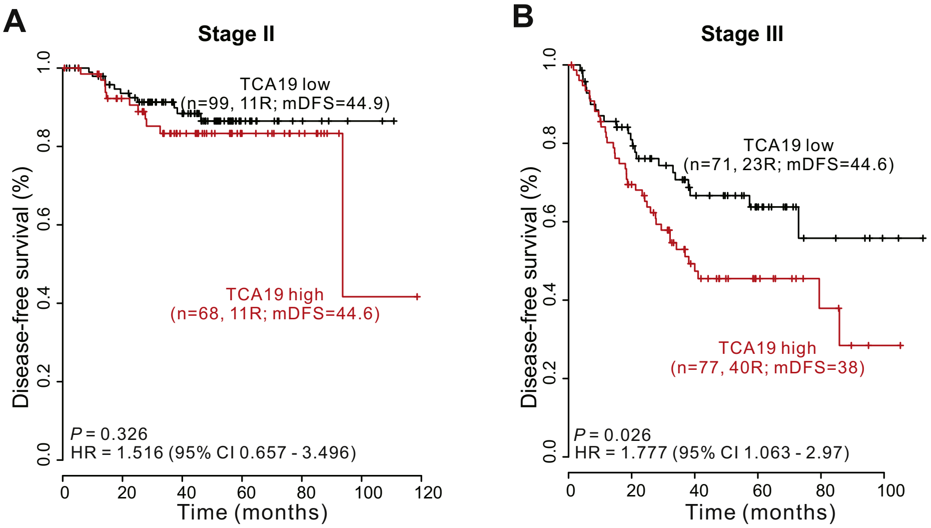

Figure 3 – Kaplan–Meier plots of disease-free survival (DFS) of patients grouped by AJCC stage in the pooled cohort with AUS and UPAC cohorts ($n = 449$). (A) TCA19 risk score-based subset analysis in stage II patients. (B) TCA19 risk score-based subset analysis in stage III patients. The TCA19 classifier was predictive in patients at stage III. *P*-values were obtained by log-rank test. mDFS, median disease-free survival.

divided into high- and low-risk subgroups based on TCA19 risk scores, and the difference in DFS was independently assessed. Adjuvant chemotherapy improved DFS in patients in the TCA19-classified high-risk patient subgroup ($P = 0.009$, Figure 4D), while no association was observed in the low-risk patient subgroup ($P = 0.704$, Figure 4E). When Cox regression model was applied, the interaction of TCA19 with adjuvant chemotherapy reached a significance level of 0.599 (Supplementary Figure 10). However, consistent with the Kaplan–Meier plot and log-rank test, the estimated HR for adjuvant chemotherapy in the high-risk group classified by TCA19 classifier was 0.363 (95% CI = 0.163–0.805; $P = 0.013$) retaining a significant predictive value, while HR for relapse for adjuvant chemotherapy in the low-risk group was 0.758 (95% CI = 0.180–3.186; $P = 0.705$). Taken together, TCA19 classifier had a significant prognostic potential in stage III CRC patients regardless of the age as well as a predictive value only in advanced CRC patients.

Another assessment for a predictive value of the TCA19 in the elderly patient with stage III were carried out. To prevent under-sampling of patients, the AUS cohort was combined with the CIT cohort, from which adjuvant chemotherapy data were also available. Elderly patients with stage III CRC (≥75 years, $n = 84$) were divided into high- and low-risk subgroups using the classifier, and the difference in DFS was independently assessed according to whether patients received chemotherapy. We have not found a statistical significance but a tendency towards benefit from adjuvant chemotherapy in the TCA19-based high-risk patients, whereas any association was not observed in the low-risk patient group (Supplementary Figure 11). Additionally, we divided CRC patients in the combined cohort into chemotherapy-treated and non-treated groups and performed multivariate analyses in these two groups, separately. TCA19 classifier remained as an independent risk factor for DFS in both chemotherapy-treated (HR = 1.851, 95% CI = 1.283–2.671, $P = 9.946 \times 10^{-4}$; Supplementary Table 8) and non-treated groups (HR = 2.287, 95% CI = 1.478–3.538, $P = 2.401 \times 10^{-4}$; Supplementary Table 8). These results demonstrated that TCA19 predictor might be independent with treatment of adjuvant chemotherapy, although patient selection bias existed in the patient cohorts used in the current study.

3.7. *Comparison of other genomic predictors with the TCA19 classifier*

Because the 114 gene MD Anderson Cancer Centre prognostic predictor (MDA114) (Oh et al., 2012) and the 7 gene Oncotype DX recurrence score (OncoDX) (Clark-Langone et al., 2010) showed robust performance in identifying patients with poor prognosis (Park et al., 2013), these two predictors were compared with the TCA19 predictor. We applied the original prediction methods and threshold values of the three classifiers (TCA19, MDA114, and OncoDX) and stratified patients in the AUS cohort according to the risk level predicted by each classifier. Kaplan–Meier plots illustrated significant differences in DFS rates between high-risk and low-risk patient groups classified by each genomic predictor (Supplementary Figure 12). When patients were stratified according to AJCC staging, all classifiers successfully identified high-risk patients with stage III disease, whereas neither of them showed significant prognostic value in classifying patients with stage II disease. Particularly in the elderly patient (≥75 years) with stage III patients, TCA19 was the only predictor identifying high-risk patients, while MDA114 or OncoDX did not show such a predictive ability (Supplementary Figure 13). Another subset analysis was performed to compare the associations between outcome predicted by the classifiers and benefit from adjuvant chemotherapy. The high-risk stage III patient subgroups classified by all genomic predictors had a potential benefit from adjuvant chemotherapy, whereas those in the low-risk subgroups did not (Supplementary Figure 14). The

Table 2 – Univariate and multivariate Cox regression analysis for prediction of DFS.

Variable	Univariate				Multivariate model 1 (AUS + UAPC)			Multivariate model 2 (AUS)		
	n	*n* Event	HR (95% CI)	*P*-value	*n*	HR (95% CI)	*P*-value	*n*	HR (95% CI)	*P*-value
Stages (I, II, or III)	446	105	3.193 (2.231–4.571)	2.226×10^{-10}	446	2.949 (2.053–4.235)	4.824×10^{-9}	226	2.784 (1.687–4.592)	6.107×10^{-5}
Gender (male or female)	446	105	0.903 (0.615–1.326)	0.604						
Age (<75 or ≥ 75)	445	105	0.757 (0.482–1.188)	0.226						
Location (distal or proximal)	226	50	1.01 (0.454–2.247)	0.981						
Chemotherapy (no or yes)	226	50	1.892 (1.085–3.301)	0.025						
TCA19 (low or high)	446	105	2.281 (1.543–3.372)	3.574×10^{-5}		1.894 (1.277–2.809)	0.002		2.24 (1.22–4.114)	0.009

The multivariate analysis described correspond to the best multivariate model obtained from a backward-forward selection procedure. Multivariate model 1 was performed in the combined cohort with the AUS and UAPC cohorts. Multivariate model 2 was carried out in the AUS cohort.
Abbreviations: DFS, disease-free survival; HR, hazards ratio; CI, confidence interval; TCA, *TREM1* and *CTGF* activation.

characteristics of each classifier are illustrated in Supplementary Table 9, showing that all classifiers were able to predict CRC outcomes but that the TCA19 predictor had a prognostic potential in elderly patients with stage III disease. Concordance among the predicted outcomes was assessed by comparing the subgroups of patients predicted by each prediction model (Cramer's V statistics; Supplementary Table 10). All correlations among predictors were statistically significant (χ^2 test, $r = 0.219$–0.472; $P < 0.001$) and the highest correlation was observed between TCA19 and OncoDX (χ^2 test, $r = 0.472$; $P = 1.209 \times 10^{-12}$).

The prognostic and predictive values of TCA19 were also validated by comparing it with randomly generated genomic predictors. Using the similar procedure for the development and validation of TCA19 classifier, we made a classifier with randomly selected 19 genes and confirmed its predictive value. In case of random classifiers, the average *P*-values by log-rank test in all subset categories didn't reach statistical significance. Compared with TCA19 classifier, the significance levels of TCA19 in all patients, stage III patients, elderly patients with stage III, and chemotherapy response in high-risk stage III were out of range of CI for significance in case of random classifiers (Supplementary Figure 15A), showing that TCA19 classifier was not generated by chance. While there were a fraction of random classifiers that outperformed TCA19 in a subset analysis, no random classifier outperformed TCA19 across all subset categories in bootstrap resampling analysis (Supplementary Figure 15B).

3.8. *Biological characteristics of the classifier and comparison with CRC subtypes*

A gene set enrichment test was performed on the TCA66 genes to explore the biological characteristics of the classifier (Supplementary Table 6). Genes involved in inflammatory disease and response, which are well-known activities of *TREM1*, were enriched. Additionally, terms involved in cellular development, cellular growth and proliferation, cell-to-cell signalling and interaction, and cell death and survival were also identified (Supplementary Figure 16). These results suggest that 66 genes including *TREM1* and *CTGF* might have a significant activity associated with cancer progression beyond inflammatory or immune response.

In addition, we carried out comparative analysis of recently reported CRC subtypes and TCA19 classifier. We first compared classifications of CRC patients between TCA19 and CRCassigner (Sadanandam et al., 2013) in the AUS cohort. When five subtypes of CRCassigner (goblet-like, enterocyte, stem-like, inflammatory, and transit-amplifying) were compared with TCA19 classifier, most patients (33 out of 38, 86.8%) in the stem-like subtype, with short time to recurrence and the greatest benefit from adjuvant chemotherapy, were classified in high-risk patients by TCA19. Among the 5 subtypes, 41.5% (17 out of 41) of the inflammatory subtype, which had the moderate benefit from chemotherapy, was also classified as TCA19 high-risk subgroup (Supplementary Figure 17A). When patients were sorted by TCA19 risk scores, most of the high-scoring CRC patients were classified in the stem-like subtype (Supplementary Figure 17B), indicating that high-risk subgroup stratified by TCA19 well reflects

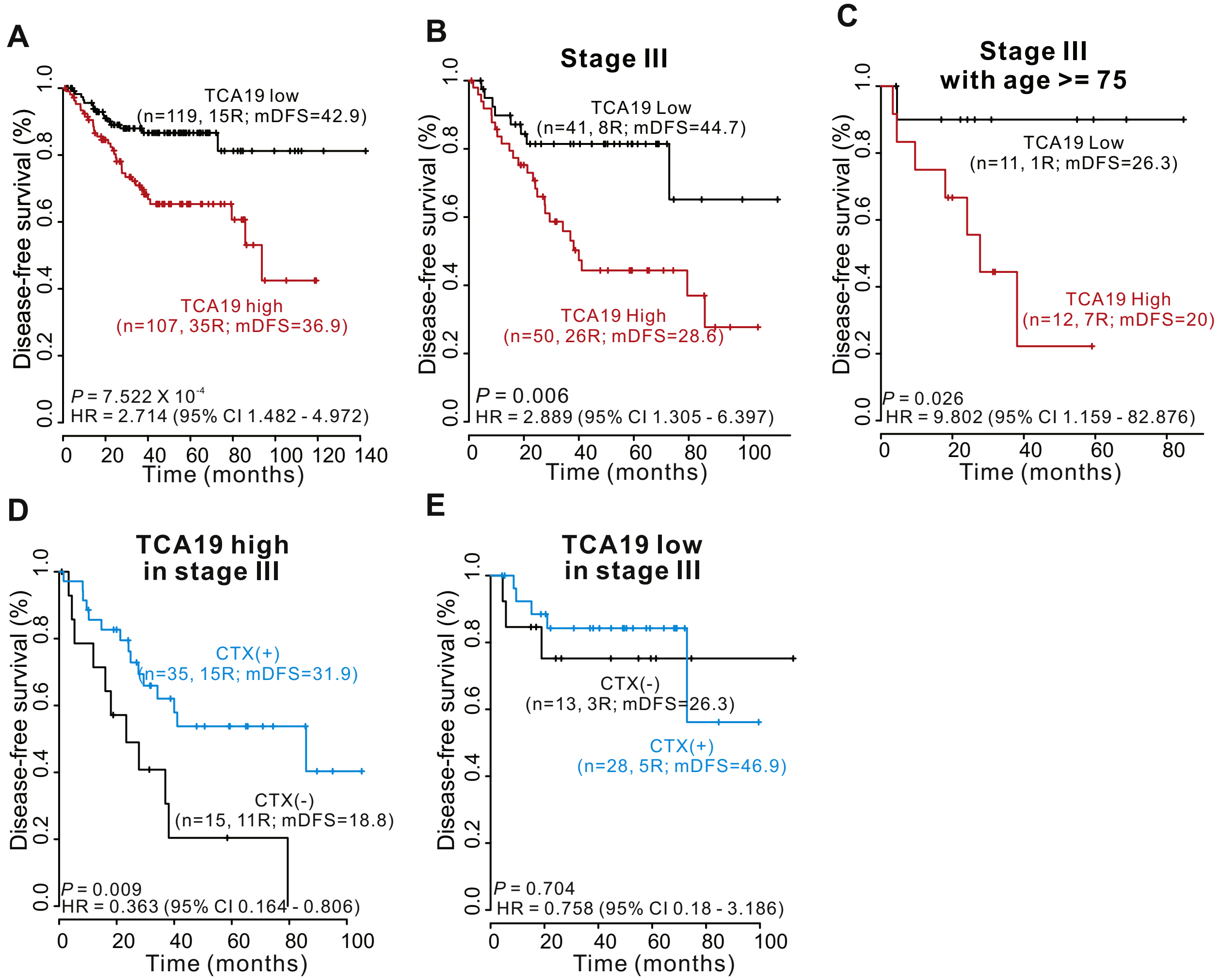

Figure 4 – Prediction of response to chemotherapy in two subgroups based on the TCA19 classifier. (A) Kaplan–Meier plots of disease-free survival (DFS) of two subgroups (TCA19-high and -low) in the AUS cohort. (B) Kaplan–Meier plots of DFS in patients with stage III disease of the AUS cohort. (C) Kaplan–Meier plots of DFS in stage III elderly patients (≥75 years) of the AUS cohort. (D) Kaplan–Meier plots of stage III patients in the TCA19 high subgroup. (E) Kaplan–Meier plots of stage III patients in the TCA19 low subgroup. The data were plotted according to whether patients received chemotherapy (CTX) or not. *P*-values were obtained by log-rank test. mDFS, median disease-free survival.

distinct biological subtypes of CRC showing poor prognosis. There were 16 common genes between CRCassigner (786 genes) and our classifier (66 genes) (Supplementary Figure 17C). Among them, *TREM1* and *CTGF* showed the highest Prediction of Microarray Analysis (PAM) score in the inflammatory and the stem-like subtypes, respectively (Supplementary Table 11) (Sadanandam et al., 2013).

We next compared classifications of CRC patients between TCA19 and 3 colon cancer subtypes (CCSs) derived by a 146-gene classifier (De Sousa et al., 2013) in the AUS cohort. When compared with 3 subtypes of CRC (CCS1, CCS2, and CCS3), a large majority of CRC patients (27 out of 36, 75%) in the CCS3 subtype, having a particularly unfavourable prognosis and showing a signature of epithelial-mesenchymal transition (EMT) and extracellular matrix remodelling, were classified in the high-risk subgroup by TCA19 (Supplementary Figure 18A). Most of the high-scoring CRC patients sorted by TCA19 were significantly associated with the CCS3 (Supplementary Figure 18B), indicating that high-risk subgroup stratified by TCA19 may be very similar with EMT and matrix remodelling signature showing poor prognosis. Interestingly, CCS3 is mostly MSS (De Sousa et al., 2013), congruent with our previous multivariate analysis including MMR status [pMMR (MSS or MSI-L) vs. dMMR (MSI-H), HR = 0.452, 95% CI = 0.238–0.858, $P = 0.015$; Supplementary Table 7]. There were only 4 overlapping genes between 146-gene classifier (146 genes) and our classifier (66 genes) (Supplementary Figure 18C).

Lastly, a comparative analysis between TCA19 and 6 CRC subtypes (C1 ~ C6) stratified by 57-gene centroid classifier (Marisa et al., 2013) in the CIT cohort was carried out. 96.6% (57 out of 59) of the CRC patients in the C4 (stem cell phenotype-like) subtype and 71.7% (43 out of 60) in the C6 (normal-like) subtype, associated with shorter relapse-free survival, were classified in the high-risk subgroup by TCA19 (Supplementary Figure 19A). When patients were sorted by TCA19 risk scores, the highest CRC patients were strongly related with C4 or C6 subtypes (Supplementary Figure 19B), demonstrating that TCA19-based high-risk subgroup may be very similar with cancer stem cell (CSC) signature consistent with our previous comparison with CRCassigner. Interestingly, no common genes were found between 57-gene centroid (57 genes) and our classifier (66 genes) (Supplementary Figure 19C).

4. Discussion

Because CRC is a heterogeneous disease with diverse clinical behaviour, it is crucial to identify distinct subtypes with different clinical outcomes and to determine who will benefit from adjuvant chemotherapy. Recently, many studies have reported distinct CRC subtypes with heterogeneous biological and clinical features (Cancer Genome Atlas Network, 2012; De Sousa et al., 2013; Marisa et al., 2013; Sadanandam et al., 2013). To develop prognostic gene signatures for CRC, a unique three-step approach was used to identify molecular changes using individually matched samples (normal colon, primary CRC, and metastatic CRC to the liver). First, unique and common genes were selected in terms of tumorigenesis and metastasis. Then, using high-throughput RNA-seq analysis, quantification profiling of mRNA expression intensity and sequence-variant profiling approaches were combined. Through these analyses, two significantly activated regulators, *TREM1* and *CTGF*, were identified. The gene sets collectively regulated by these two molecules were strongly predictive of high-risk CRC and independent of currently available clinical indicators. The survival outcomes underscore the potential value of CRC subtypes defined by *TREM1* and *CTGF* activation, in addition to the pathological prognosis features.

Based on gene set enrichment analysis, we showed that the disease events of tumorigenesis and metastasis of CRC share many similar biological functions. In addition, analysis of the two gene sets within the context of gene networks showed that *TREM1* and *CTGF* were significantly activated in both tumorigenesis and metastasis. Many downstream genes of *TREM1*, including *CASP5*, *CCL7*, *CSF2*, *CXCL3*, *IL8*, and *ITGA5*, are involved in *TREM1* signalling, of which *CSF2* and *ITGA5* are associated with cell growth and proliferation. Interestingly, *CCL7*, *CSF2*, *CXCL3*, and *ITGA5* are associated with tumour metastasis in various cancer types (Acharyya et al., 2012; Aggarwal and Sung, 2009; Liu et al., 1999; Valastyan et al., 2011), indicating that activation of the *TREM1* signalling pathway may be associated with CRC aggressiveness. A number of downstream effectors of *CTGF* (*e.g.*, *ACAN*, *FN1*, *SERPINE1*, and *TIMP1*) are involved in cell growth and proliferation, which are well-known activities of *CTGF*. Among them, *FN1*, *SERPINE1*, and *TIMP1* are associated with metastases in a number of cancers, including CRC (Binder and Mihaly, 2008; Kim et al., 2012b; Kruger et al., 1997; Malik et al., 2010; Nakagawa et al., 2004). Taken together, our findings demonstrate that the gene set defined by *TREM1* and *CTGF* activation that is overexpressed in CRC patients is strongly predictive of tumour aggressiveness.

TREM1 encodes a transmembrane receptor belonging to the Ig superfamily that is expressed on myeloid cells. The protein encoded by *TREM1* is known to be involved in activation, differentiation, and phosphorylation, as well as the inflammatory response in various cells, including myeloid, lymphoma, and leukaemia cell lines. Although few associations between *TREM1* and cancer have been investigated (Huang et al., 2008), up-regulation of *TREM1* in the intestinal mucosa was reportedly associated with active ulcerative colitis in humans (Schenk et al., 2007), suggesting that *TREM1* may be a possible mediator of CRC.

The protein encoded by *CTGF* is a mitogen that is secreted by vascular endothelial cells. *CTGF* plays a role in proliferation, growth, and activation in many cancer cell types. Many previous studies support a significant correlation between *CTGF* expression and various cancer types (Cascione et al., 2013; Nakagawa et al., 2004; Yang et al., 2012). In addition, a significant association between CTGF protein expression and initiation or development of CRC was reported (Ladwa et al., 2011). Interestingly, up-regulation of *CTGF* in stromal fibroblast cells from liver metastases was reportedly correlated with metastatic CRC (Nakagawa et al., 2004), underscoring our conclusion that changes in *CTGF* might reflect aggressive clinical behaviour in CRC.

Due to the clinical heterogeneity of CRC, it is difficult to identify patients who will benefit the most from adjuvant chemotherapy. Although the benefits of adjuvant chemotherapy have been well established for patients with AJCC stage III cancer (Laurie et al., 1989; Moertel et al., 1990), clinical relevance is limited to patients that are not routinely offered adjuvant chemotherapy, in particular elderly patients (Sveen et al., 2013). In the present study, subset analysis of patients with stage III demonstrated that the TCA19 risk score could identify a high-risk patient subgroup, particularly high-risk stage III elderly patients, who would get benefits from adjuvant chemotherapy. In patients with stage III disease, chemotherapy was significantly associated with improved outcome for patients in the high-risk subgroup predicted by TCA19, whereas its benefit was not significant for patients in the low-risk group. As the first-line chemotherapy was less effective in tumours with TCA19 under-expression than those with over-expression, other regimens including targeted biologics would be recommended in the former group of patients. When TCA19 signature was evaluated in subgroup containing stage III elderly patients, a tendency towards benefit from adjuvant chemotherapy in TCA19-based high-risk patients was observed. Although elderly patients with stage III disease have often been excluded from standard chemotherapy, patients older than 75 years reportedly had a similar survival benefit from fluorouracil-based treatment as younger patients (Amorena et al., 2009; Verbeek, 2008). Therefore, traditional chemotherapy appears to benefit elderly stage III patients who were identified as a high-risk group by TCA19, while alternative chemotherapy using 2nd- or 3rd-line may be required for those in the low-risk group.

To exhibit translational and clinical relevance of the TCA19 genomic predictor, various biological characteristics of the classifier were explored. CSCs have aggressive characteristics including increased invasion, metastatic ability and poor patient prognosis (Findlay et al., 2014). EMT is a unique process initially characterized in embryonic development in which cells lose epithelial features and gain mesenchymal properties

(Thiery and Sleeman, 2006). The association between EMT and CSC properties across multiple organ systems including CRC underscore the crucial linkage of EMT with aggressive cancer biology as well as with the CSCs (Fan et al., 2012; Findlay et al., 2014; Mani et al., 2008). The dual processes of EMT and CSCs enhances the aggressiveness of tumour cells and allows the cells to escape to distant sites more conducive for survival (Findlay et al., 2014). In the current study, TCA19, consisting of genes significantly associated with CRC aggressiveness, was a classifier to predict high-risk CRC patients who showed clear characteristics of stem-cell or EMT signature. About 15% of CRCs develop for defective function of the MMR system. Patients with dMMR CRCs have reduced rates of disease recurrence and delayed time to recurrence compared with pMMR CRCs (Sinicrope et al., 2011), which is consistent with the multivariate regression analysis with MMR status and TCA19 in the present study. Because distant recurrences were reduced by adjuvant treatment in dMMR stage III tumours (Sinicrope et al., 2011) and the TCA19 classifier was an independent risk factor with MMR condition, TCA19 might have a predictive value for the high-risk dMMR patients.

Although we illustrated the clinical potential of our findings, we also note limitations of our study. First, the RNA-seq data in the AMC cohort did not contain normal liver tissue samples to accurately eliminate liver-specific genes. Second, the number of patients who received adjuvant chemotherapy in the AUS cohort (87 out of 229 patients) was not enough to rigorously determine chemo-responsiveness of the classifier. The predictive value of TCA19 is also limited without comparison with groups using various regimens as well as with a real control group of surgery only. Third, our analyses provided only indirect evidences for the roles of *TREM1* and *CTGF* in colorectal tumorigenesis, so the activity of *TREM1* and *CTGF* needs to be validated by biological assays.

In conclusion, based on RNA-seq transcriptome results of CRCs including various stages and aggressiveness, we identified high-risk prognostic subgroups defined by *TREM1* and *CTGF* activation and a possible association between the TCA19 classifier and the response to adjuvant chemotherapy. Since the TCA19 may categorize the CRC patients like CSC or EMT subtypes showing poor prognosis, a close follow-up and potent chemotherapy would be needed regardless of curative surgery in patients whose tumours showed higher expression of TCA19. Although our data demonstrate that the TCA19 classifier is effective as a potential prognostic marker, its usefulness as a predictive marker for response to adjuvant chemotherapy should be further evaluated using larger cohorts with respective chemotherapy regimen and a real control patient group of surgery only.

Acknowledgements

This study was supported by grants from the Asan Institute for Life Sciences (2014-69); the Korea Research Foundation, the Ministry of Science, ICT and Future Planning (2013R1A2A1A03070986); and the Korea Health 21 R&D Project of the Ministry of Health and Welfare, Republic of Korea (HI06C0868 and HI13C1750).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.molonc.2014.06.016.

REFERENCES

Acharyya, S., Oskarsson, T., Vanharanta, S., Malladi, S., Kim, J., Morris, P.G., Manova-Todorova, K., Leversha, M., Hogg, N., Seshan, V.E., Norton, L., Brogi, E., Massague, J., 2012. A CXCL1 paracrine network links cancer chemoresistance and metastasis. Cell 150, 165–178.

Aggarwal, B.B., Sung, B., 2009. Pharmacological basis for the role of curcumin in chronic diseases: an age-old spice with modern targets. Trends Pharmacol. Sci. 30, 85–94.

Amorena, M., Visciano, P., Giacomelli, A., Marinelli, E., Sabatini, A.G., Medrzycki, P., Oddo, L.P., De Pace, F.M., Belligoli, P., Di Serafino, G., Saccares, S., Formato, G., Langella, V., Perugini, M., 2009. Monitoring of levels of polycyclic aromatic hydrocarbons in bees caught from beekeeping: remark 1. Vet. Res. Commun. 33 (Suppl 1), 165–167.

Binder, B.R., Mihaly, J., 2008. The plasminogen activator inhibitor "paradox" in cancer. Immunol. Lett. 118, 116–124.

Cancer Genome Atlas Network, 2012. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337.

Carlsson, U., Lasson, A., Ekelund, G., 1987. Recurrence rates after curative surgery for rectal carcinoma, with special reference to their accuracy. Dis. Colon Rectum 30, 431–434.

Cascione, L., Gasparini, P., Lovat, F., Carasi, S., Pulvirenti, A., Ferro, A., Alder, H., He, G., Vecchione, A., Croce, C.M., Shapiro, C.L., Huebner, K., 2013. Integrated microRNA and mRNA signatures associated with survival in triple negative breast cancer. PLoS One 8, e55910.

Clark-Langone, K.M., Sangli, C., Krishnakumar, J., Watson, D., 2010. Translating tumor biology into personalized treatment planning: analytical performance characteristics of the Oncotype DX colon cancer assay. BMC Cancer 10, 691.

Cunningham, D., Atkin, W., Lenz, H.J., Lynch, H.T., Minsky, B., Nordlinger, B., Starling, N., 2010. Colorectal cancer. Lancet 375, 1030–1047.

de Sousa, E.M.F., Colak, S., Buikhuisen, J., Koster, J., Cameron, K., de Jong, J.H., Tuynman, J.B., Prasetyanti, P.R., Fessler, E., van den Bergh, S.P., Rodermond, H., Dekker, E., van der Loos, C.M., Pals, S.T., van de Vijver, M.J., Versteeg, R., Richel, D.J., Vermeulen, L., Medema, J.P., 2011. Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. Cell Stem Cell 9, 476–485.

De Sousa, E.M.F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L.P., de Jong, J.H., de Boer, O.J., van Leersum, R., Bijlsma, M.F., Rodermond, H., van der Heijden, M., van Noesel, C.J., Tuynman, J.B., Dekker, E., Markowetz, F., Medema, J.P., Vermeulen, L., 2013. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat. Med. 19, 614–618.

Fan, F., Samuel, S., Evans, K.W., Lu, J., Xia, L., Zhou, Y., Sceusi, E., Tozzi, F., Ye, X.C., Mani, S.A., Ellis, L.M., 2012. Overexpression of snail induces epithelial-mesenchymal transition and a cancer stem cell-like phenotype in human colorectal cancer cells. Cancer Med. 1, 5–16.

Findlay, V.J., Wang, C., Watson, D.K., Camp, E.R., 2014. Epithelial-to-mesenchymal transition and the cancer stem cell phenotype: insights from cancer biology with therapeutic implications for colorectal cancer. Cancer Gene Ther 21, 181–187.

Hari, D.M., Leung, A.M., Lee, J.H., Sim, M.S., Vuong, B., Chiu, C.G., Bilchik, A.J., 2013. AJCC Cancer Staging Manual 7(th) Edition criteria for Colon Cancer: do the complex modifications improve prognostic assessment? J. Am. Coll. Surg. 217, 181–190.

Heagerty, P.J., Lumley, T., Pepe, M.S., 2000. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56, 337–344.

Huang, L.Y., Shi, H.Z., Liang, Q.L., Wu, Y.B., Qin, X.J., Chen, Y.Q., 2008. Expression of soluble triggering receptor expression on myeloid cells-1 in pleural effusion. Chin. Med. J. 121, 1656–1661.

Jass, J.R., 2007. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 50, 113–130.

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D., 2011. Global cancer statistics. CA: Cancer J. Clin. 61, 69–90.

Jorissen, R.N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L.A., Arango, D., Kruhoffer, M., Orntoft, T.F., Andersen, C.L., Gruidl, M., Kamath, V.P., Eschrich, S., Yeatman, T.J., Sieber, O.M., 2009. Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. Clin. Cancer Res.: An Official Journal of the American Association for Cancer Research 15, 7642–7651.

Kang, G.H., 2011. Four molecular subtypes of colorectal cancer and their precursor lesions. Arch. Pathol. Lab. Med. 135, 698–703.

Kim, S.M., Leem, S.H., Chu, I.S., Park, Y.Y., Kim, S.C., Kim, S.B., Park, E.S., Lim, J.Y., Heo, J., Kim, Y.J., Kim, D.G., Kaseb, A., Park, Y.N., Wang, X.W., Thorgeirsson, S.S., Lee, J.S., 2012a. Sixty-five gene-based risk score classifier predicts overall survival in hepatocellular carcinoma. Hepatology 55, 1443–1452.

Kim, Y.S., Kim, S.H., Kang, J.G., Ko, J.H., 2012b. Expression level and glycan dynamics determine the net effects of TIMP-1 on cancer progression. BMB Reports 45, 623–628.

Kruger, A., Fata, J.E., Khokha, R., 1997. Altered tumor growth and metastasis of a T-cell lymphoma in Timp-1 transgenic mice. Blood 90, 1993–2000.

Ladwa, R., Pringle, H., Kumar, R., West, K., 2011. Expression of CTGF and Cyr61 in colorectal cancer. J. Clin. Pathol. 64, 58–64.

Laurie, J.A., Moertel, C.G., Fleming, T.R., Wieand, H.S., Leigh, J.E., Rubin, J., McCormack, G.W., Gerstner, J.B., Krook, J.E., Malliard, J., et al., 1989. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. J. Clin. Oncol. 7, 1447–1456.

Liu, X., Yu, X., Zack, D.J., Zhu, H., Qian, J., 2008. TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics 9, 271.

Liu, Y., Thor, A., Shtivelman, E., Cao, Y., Tu, G., Heath, T.D., Debs, R.J., 1999. Systemic gene delivery expands the repertoire of effective antiangiogenic agents. J. Biol. Chem. 274, 13338–13344.

Malik, G., Knowles, L.M., Dhir, R., Xu, S., Yang, S., Ruoslahti, E., Pilch, J., 2010. Plasma fibronectin promotes lung metastasis by contributions to fibrin clots and tumor cell invasion. Cancer Res. 70, 4327–4334.

Mani, S.A., Guo, W., Liao, M.J., Eaton, E.N., Ayyanan, A., Zhou, A.Y., Brooks, M., Reinhard, F., Zhang, C.C., Shipitsin, M., Campbell, L.L., Polyak, K., Brisken, C., Yang, J., Weinberg, R.A., 2008. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell 133, 704–715.

Marisa, L., de Reynies, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Flejou, J.F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P., Boige, V., 2013. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 10, e1001453.

McShane, L.M., Altman, D.G., Sauerbrei, W., Taube, S.E., Gion, M., Clark, G.M.Statistics Subcommittee of the, N.C.I.E.W.G.o.C.D, 2005. REporting recommendations for tumour MARKer prognostic studies (REMARK). Eur. J. Cancer 41, 1690–1696.

Midgley, R., Kerr, D., 1999. Colorectal cancer. Lancet 353, 391–399.

Midgley, R.S., Yanagisawa, Y., Kerr, D.J., 2009. Evolution of nonsurgical therapy for colorectal cancer. Nature clinical practice. Gastroenterol. Hepatol. 6, 108–120.

Moertel, C.G., Fleming, T.R., Macdonald, J.S., Haller, D.G., Laurie, J.A., Goodman, P.J., Ungerleider, J.S., Emerson, W.A., Tormey, D.C., Glick, J.H., et al., 1990. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. N. Engl. J. Med. 322, 352–358.

Nakagawa, H., Liyanarachchi, S., Davuluri, R.V., Auer, H., Martin Jr., E.W., de la Chapelle, A., Frankel, W.L., 2004. Role of cancer-associated stromal fibroblasts in metastatic colon cancer to the liver and their expression profiles. Oncogene 23, 7366–7377.

Oh, S.C., Park, Y.Y., Park, E.S., Lim, J.Y., Kim, S.M., Kim, S.B., Kim, J., Kim, S.C., Chu, I.S., Smith, J.J., Beauchamp, R.D., Yeatman, T.J., Kopetz, S., Lee, J.S., 2012. Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. Gut 61, 1291–1298.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., Hiller, W., Fisher, E.R., Wickerham, D.L., Bryant, J., Wolmark, N., 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N. Engl. J. Med. 351, 2817–2826.

Park, Y.Y., Lee, S.S., Lim, J.Y., Kim, S.C., Kim, S.B., Sohn, B.H., Chu, I.S., Oh, S.C., Park, E.S., Jeong, W., Kim, S.S., Kopetz, S., Lee, J.S., 2013. Comparison of prognostic genomic predictors in colorectal cancer. PLoS One 8, e60778.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Sadanandam, A., Lyssiotis, C.A., Homicsko, K., Collisson, E.A., Gibb, W.J., Wullschleger, S., Ostos, L.C., Lannon, W.A., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A.B., Wiedenmann, B., Cantley, L.C., Gray, J.W., Hanahan, D., 2013. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat. Med. 19, 619–625.

Schenk, M., Bouchon, A., Seibold, F., Mueller, C., 2007. TREM-1–expressing intestinal macrophages crucially amplify chronic inflammation in experimental colitis and inflammatory bowel diseases. J. Clin. Invest. 117, 3097–3106.

Shen, L., Toyota, M., Kondo, Y., Lin, E., Zhang, L., Guo, Y., Hernandez, N.S., Chen, X., Ahmed, S., Konishi, K., Hamilton, S.R., Issa, J.P., 2007. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc. Natl. Acad. Sci. U. S. A. 104, 18654–18659.

Sinicrope, F.A., Foster, N.R., Thibodeau, S.N., Marsoni, S., Monges, G., Labianca, R., Kim, G.P., Yothers, G., Allegra, C., Moore, M.J., Gallinger, S., Sargent, D.J., 2011. DNA mismatch repair status and colon cancer recurrence and survival in clinical trials of 5-fluorouracil-based adjuvant therapy. J. Natl. Cancer. Inst. 103, 863–875.

Stange, D.E., Engel, F., Longerich, T., Koo, B.K., Koch, M., Delhomme, N., Aigner, M., Toedt, G., Schirmacher, P., Lichter, P., Weitz, J., Radlwimmer, B., 2010. Expression of an

ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15.5 gain. Gut 59, 1236–1244.

Sveen, A., Nesbakken, A., Agesen, T.H., Guren, M.G., Tveit, K.M., Skotheim, R.I., Lothe, R.A., 2013. Anticipating the clinical use of prognostic gene expression-based tests for colon cancer stage II and III: is Godot finally arriving? Clin. Cancer Res. 19, 6669–6677.

Thiery, J.P., Sleeman, J.P., 2006. Complex networks orchestrate epithelial-mesenchymal transitions. Nat. Rev. Mol. Cell Biol. 7, 131–142.

Valastyan, S., Chang, A., Benaich, N., Reinhardt, F., Weinberg, R.A., 2011. Activation of miR-31 function in already-established metastases elicits metastatic regression. Genes Dev. 25, 646–659.

Venables, W.N., Ripley, B.D., Venables, W.N., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York.

Verbeek, J., 2008. Moose Consort Strobe and Miame Stard Remark or how can we improve the quality of reporting studies. Scand. J. Work Environ. Health 34, 165–167.

Yang, M.H., Lin, B.R., Chang, C.H., Chen, S.T., Lin, S.K., Kuo, M.Y., Jeng, Y.M., Kuo, M.L., Chang, C.C., 2012. Connective tissue growth factor modulates oral squamous cell carcinoma invasion by activating a miR-504/FOXP1 signalling. Oncogene 31, 2401–2411.