Received: 19 May 2021 | Revised: 11 December 2021 | Accepted: 14 December 2021

DOI: 10.1111/codi.16060

ORIGINAL ARTICLE

Tumour genotypes account for survival differences in right- and left-sided colon cancers

Thomas M. Ward | Christy E. Cauley | Caitlin E. Stafford | Robert N. Goldstone | Liliana G. Bordeianou | Hiroko Kunitake | David L. Berger | Rocco Ricciardi

Section of Colon and Rectal Surgery, Division of General and Gastrointestinal Surgery, Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA

Correspondence
Rocco Ricciardi, Section of Colon and Rectal Surgery, Division of General and Gastrointestinal Surgery, Department of Surgery, Massachusetts General Hospital, 15 Parkman St, WAC-4–460, Boston, MA 02114, USA.
Email: rricciardi1@mgh.harvard.edu

Abstract

Aim: We sought to identify genetic differences between right- and left-sided colon cancers and using these differences explain lower survival in right-sided cancers.

Method: A retrospective review of patients diagnosed with colon cancer was performed using The Cancer Genome Atlas, a cancer genetics registry with patient and tumour data from 20 North American institutions. The primary outcome was 5-year overall survival. Predictors for survival were identified using directed acyclic graphs and Cox proportional hazards models.

Results: A total of 206 right- and 214 left-sided colon cancer patients with 84 recorded deaths were identified. The frequency of mutated alleles differed significantly in 12 of 25 genes between right- and left-sided tumours. Right-sided tumours had worse survival with a hazard ratio of 1.71 (95% confidence interval 1.10–2.64, $P = 0.017$). The total effect of the genetic loci on survival showed five genes had a sizeable effect on survival: DNAH5, MUC16, NEB, SMAD4, and USH2A. Lasso-penalized Cox regression selected 13 variables for the highest-performing model, which included cancer stage, positive resection margin, and mutated alleles at nine genes: MUC16, USH2A, SMAD4, SYNE1, FLG, NEB, TTN, OBSCN, and DNAH5. Post-selection inference demonstrated that mutations in MUC16 ($P = 0.01$) and DNAH5 ($P = 0.02$) were particularly predictive of 5-year overall survival.

Conclusions: Our study showed that genetic mutations may explain survival differences between tumour sites. Further studies on larger patient populations may identify other genes, which could form the foundation for more precise prognostication and treatment decisions beyond current rudimentary TNM staging.

KEYWORDS
colorectal cancer, genetics, sidedness, survival

INTRODUCTION

Despite advances in screening and therapeutics, colorectal cancer remains the second leading cause of cancer mortality worldwide [1]. Differences in survival between cancers of the right and left colon have long been reported, with the majority of studies, including multiple meta-analyses, demonstrating worse survival with right-sided tumours [2–6]. Previous studies have hypothesized a variety of mechanisms through which tumour site could cause these survival differences [7–10].

These published hypotheses group into two over-arching causal pathways through which tumour site affects survival: gene expression and lymphovascular supply. A recent study showed that right- and left-sided tumours have site-specific gene expression [11]. This differential gene expression may drive carcinogenesis and differences in tumour phenotypes; increase in methylation of mismatch repair genes in the proximal colon is one known example of this phenomenon [12]. Tumour site could also affect survival through differences in lymphovascular supply, which would manifest as site-preference towards nodal, peritoneal, and distant metastases [4].

We hypothesized that genetic signatures accounted for much of the survival differences across tumour sites. To study these differences, we used causal and predictive analyses to identify genetic loci that impact survival. Our analysis harnessed publicly available data from the Cancer Genome Atlas (TCGA).

What does this paper add to the literature?

The aetiology for lower survival in patients with right, as compared to left, colon cancers is unclear, with little investigation if differences in genetic mutations could explain this discrepancy. Our study identified genes that may explain these differences and were predictive of worse overall survival

METHOD

Study design

We conducted a retrospective study of patients with colon cancer present in TCGA to evaluate the relationship between tumour site and genetics on survival. TCGA, a cancer genomics program founded in 2006 as a joint effort between the National Cancer Institute and the National Human Genome Research Institute, collects clinical and genomic tumour data from 20 contributing North American institutions on 33 different cancers [13]. We abstracted clinical, pathological, treatment, survival, and genetic information for all patients with colon cancer from TCGA for years 2009–2020. As per Mass General Brigham guidelines, submission to the Institutional Review Board was not required to analyse this data as the project did not include any interaction or intervention with human subjects, nor did it include any access to identifiable private information.

Variable definitions

Tumour pathological and genetic characteristics were abstracted from TCGA, including American Joint Commission on Cancer staging, surgical resection margin status, presence of lymphovascular invasion, and wild-type or mutated allele status of genes. The TNM status was grouped as T0-4, N0-2, and M0-1 [14]. Tumours were classified as having microsatellite instability (MSI) if their tumours had lost expression in DNA mismatch repair genes (PMS2, MLH1, MSH2, MSH6), as determined by immunohistochemistry [15]. Patients were classified as having a positive resection margin if they had either a micro (R1) or macroscopic (R2) positive pathological margin. Tumour location was classified as either right-sided (caecum to hepatic flexure) or left-sided (splenic flexure to rectosigmoid junction). Tumours of the transverse colon were excluded to avoid any inconsistency in classification with splenic flexure tumours and to minimize any effect of the different embryological origin of the proximal and distal transverse colon, as TCGA did not distinguish between proximal and distal transverse tumour location [2]. Vital status was right censored for a 5-year follow-up. Beyond mismatch repair gene expression data, TCGA included sequencing information for a large number of genes. TCGA identified 25 of these genes as being somatic recurrently mutated genes in colon cancer, which was determined through manual curation and MutSig, a technique that accounts for tumour specific background mutation rates [13,16]. Somatic recurrently mutated genes are those that are consistently mutated across different patients with the same cancer, rather than being bystander genes that happened to mutate due to the inherent aberrant gene expression regulation in cancer. We only included these loci in our analyses to avoid falsely attributing survival differences to bystander genes. We classified a loci's status as being either wild-type or mutated. We chose this binary classification as exact mutation-types were already grouped in TCGA (e.g., missense, stop gained, splice donor variant). We also wanted our analyses to provide guidance to clinicians and researchers, irrespective of the diagnostic assay used to detect mutations, as many assays can only distinguish between mutated and wild-type alleles of genes without identifying specific mutations [17].

Statistical analysis

All statistical analyses were performed using R version 4.0.3 [18]. First, differences in genetic signatures across tumour sites was compared using the Pearson's Chi-squared test with *P*-values computed by Monte Carlo simulation of one million replicates, correcting for multiple comparisons using the Holm method. All survival analyses considered 5-year overall survival (OS). Bagged tree models, which have been shown to be one of the highest performing data imputation methods, were constructed from all input variables and used to impute missing input data, which was considered missing at random [19,20]. We investigated the total effect of tumour site and genetics on survival using Cox proportional hazards models made with the survival package [21]. We used a directed acyclic graph (DAG) to design and interpret our regression models for causal inference [22]. DAGs work by designing a graphical model to represent a researcher's hypothesis on how different factors influence an outcome, in our case, patient survival. These hypothesis assumptions, often implicit and not stated when designing and publishing regression models, are then made explicit for

both researchers and those who read their works. Following the structure of the DAG, a researcher can minimize spurious associations in a statistical model by including the variables in the model necessary to prevent confounding associations, while not including extra variables, which could lead to introduction of colliders [23]. A collider is a variable that is causally influenced by two other predictor variables that do not necessarily have a causal link. For example, age and hand-eye coordination influence one's ability to play sports. Age also influences one's education, while hand-eye coordination, does not. If a statistician included sports ability (a collider), age, and hand-eye coordination, as variables in a model to predict education, a false association with hand-eye coordination would be found when one never existed. DAGs allow a modeller to avoid introducing these extra variables and minimize false associations.

Models with a large number of input variables relative to the outcome were first fit with all input variables. Variables were selected from this full fit in one step for the model's final fit if they were either part of the minimal adjustment set necessary to assess for causal effect identification or, to minimize bias, had a *P*-value under 0.50 in the full model [24]. Schoenfeld residuals were checked to confirm that the proportional hazards assumptions held for all covariables. We also analysed the effect of tumour site on survival by plotting a Kaplan-Meier curve and performing the log-rank test with the survminer package [25].

Lastly, we developed a model to identify predictive factors for survival. Due to the large number of predictors when incorporating mutated allele status for each gene, we used a lasso-penalized Cox regression. Lasso-penalized Cox regression works like classical Cox regression, except it selects, from a large number of predictor variables, the fewest number of variables needed to still explain the greatest amount of variability in the outcome, in this case, survival. This selection process eliminates variables from the model that minimally help with outcome prediction, which creates a parsimonious model and minimizes dataset overfitting. This process leaves only those variables that best predict the outcome in the model. We used the lasso penalty that minimized mean error on 10-fold cross-validation [26]. Post-selection inference to derive *p*-values and confidence intervals on the lasso-selected variables was performed using the selectiveInference package [27,28]. All data and code for analyses is available online [29].

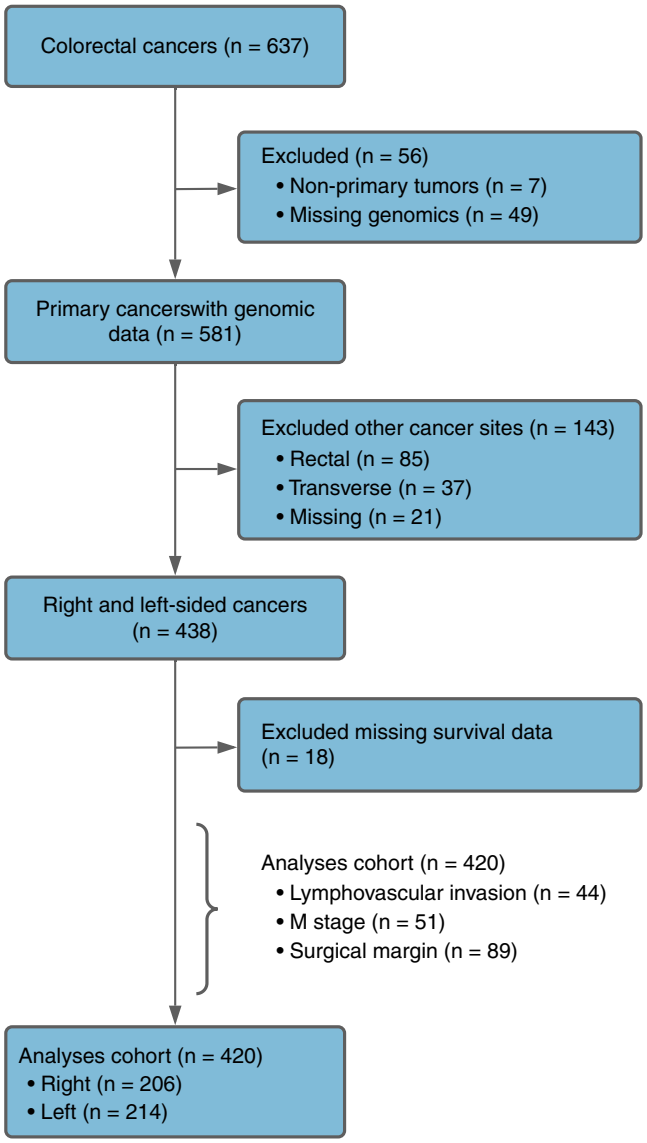

FIGURE 1 Flow diagram showing patient selection and handling of missing data for right- and left-sided colon cancers found in the Cancer Genome Atlas (as of 2020)

RESULTS

Cohort data

The Cancer Genome Atlas contained data on 637 patients with colorectal cancer. After excluding patients who had nonprimary tumours, missing genetic data, missing survival data, or who had tumours at sites other than the right or left colon, we had a final study cohort of 206 right and 214 left-sided colon cancer patients. Missing data for lymphovascular invasion ($n = 44$), M stage ($n = 51$), and surgical margin ($n = 89$) was imputed with bagged trees (Figure 1). The cohort was 47% female, with a median (interquartile range (IQR)) age of 68 (58–77) years at time of diagnosis. Most tumours were T3 (69%), N0 (56%), and M0 (83%). There were 51 and 33 deaths in the right- and left-sided colon cancer groups, respectively. The median (IQR) follow-up was 23 months (13–37; Table 1).

Causal inference

The DAG we constructed to design and interpret regression models for causal inference on tumour site and survival was predicated upon hypothesized aetiologies for right- and left-sided colon cancer survival differences. In particular, tumour site could affect survival through three pathways: TNM stage (e.g., differing embryology and therefore vascular supply leading to differences in systemic tumour spread), MSI (given right-sided gene expression differences causing

likelihood of MSI), and other genetic drivers (possibly created through differing tumour environments between sites; Figure 2) [10,13,30]. We would like to emphasize the importance the DAG had in directing our analyses. For example, one may be tempted to exclude patients with a positive resection margin (R1, R2) from the analyses. However, based on the DAG, tumour genetics could affect tumour stage, including depth of invasion, which could then increase the likelihood of a positive surgical margin and lower a patient's survival. By excluding patients with a positive resection margin, we would fail to capture the negative survival effect a gene could exert through this causal pathway.

The DAG also accounted for other factors, unobserved in TCGA, that influence survival, including predisposing patient factors (e.g., environmental exposures, heritable traits) and post-diagnosis survival factors (e.g., patient frailty, chemotherapy). These factors, while affecting survival, do not directly influence the tumour site nor its causal effect on survival. While these factors, such as chemotherapeutic regimen, would be useful in increasing precision of modeling outputs of a predictive model, they were not needed for our analyses to fully determine the true effect of the genetics on survival, based on the information found in TCGA.

TABLE 1 Patient demographic, pathological, and survival characteristics

	Tumour site	
Variable	**Right, $N = 206$[a]**	**Left, $N = 214$[a]**
Age	71 (60–79)	65 (56–74)
Sex		
Male	113 (55)	111 (52)
Female	93 (45)	103 (48)
T stage		
0	1 (0.5)	0 (0)
1	4 (1.9)	6 (2.8)
2	36 (17)	41 (19)
3	140 (68)	149 (70)
4	25 (12)	18 (8.4)
N stage		
0	128 (62)	108 (50)
1	36 (17)	67 (31)
2	42 (20)	39 (18)
M1 status	32 (16)	41 (19)
R1/R2 margin	12 (5.8)	17 (7.9)
Lymphovascular invasion	65 (32)	89 (42)
Microsatellite instability	36 (17)	9 (4.2)
Follow-up time	22 (12–36)	24 (14–37)
Deaths (within 5 years)	51 (25)	33 (15)

[a]Statistics presented: Median (IQR); *n* (%).

Genetics

Based on the DAG, no adjustment was necessary to estimate the total effect of tumour site on the frequency of mutated alleles, therefore we performed a Chi-squared test, which showed that the frequency differed significantly (Holm corrected q-value < 0.05) from expected in 12 of 25 genes. These genes included CSMD3 ($P = 0.001$), FAT3 ($P = 0.002$), FAT4 ($P = 0.002$), MUC16 ($P < 0.001$), NEB ($P = 0.001$), OBSCN ($P = 0.001$), PCLO ($P < 0.001$), PI3K ($P < 0.001$), SYNE1 ($P = 0.002$), TP53 ($P < 0.001$), USH2A ($P < 0.001$), and ZFH ($P < 0.001$; Table 2).

Overall five-year survival

Investigation of the total effect of tumour site on survival, through the previously described three pathways, was, following the DAG, possible through a univariable approach. 1-, 3-, and 5-year estimated overall survival (95% CI) was 93% (90%–97%), 83% (76%–90%), and 65% (53%–79%) for left-sided tumours and 88% (83%–92%), 73%

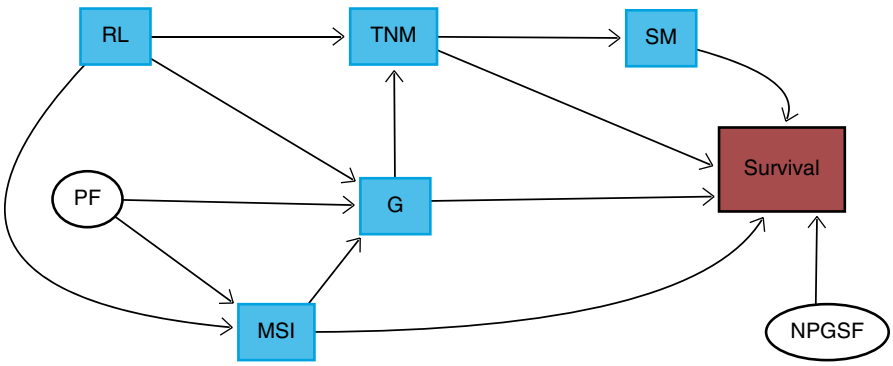

FIGURE 2 Directed acyclic graph that guided model adjustments necessary to estimate causal effects on the outcome, 5-year overall survival (survival). G, tumour genetics; MSI, microsatellite instability; NPGSF, nonpathological/genetic survival factors (e.g., frailty); PF, patient factors (e.g., exposures, inherited traits); RL, tumour site (right or left); SM, surgical margin; TNM, tumour stage. White colour, unobserved variable; blue colour, ancestor of outcome; maroon, outcome

TABLE 2 Mutations at genetic loci across right- and left-sided colon tumours

Gene	Tumour site			
	Right, $N = 206$[a]	Left, $N = 214$[a]	*P*-value[b]	q-value[c]
APC				
Wild-type	110 (53)	87 (41)	0.011	0.11
Mutated	96 (47)	127 (59)		
CSMD1				
Wild-type	160 (78)	189 (88)	0.004	0.055
Mutated	46 (22)	25 (12)		
CSMD3				
Wild-type	162 (79)	193 (90)	0.001	0.022
Mutated	44 (21)	21 (9.8)		
DNAH11				
Wild-type	168 (82)	191 (89)	0.027	0.20
Mutated	38 (18)	23 (11)		
DNAH5				
Wild-type	168 (82)	191 (89)	0.027	0.20
Mutated	38 (18)	23 (11)		
FAT3				
Wild-type	157 (76)	188 (88)	0.002	0.030
Mutated	49 (24)	26 (12)		
FAT4				
Wild-type	149 (72)	182 (85)	0.002	0.029
Mutated	57 (28)	32 (15)		
FBXW7				
Wild-type	179 (87)	196 (92)	0.16	0.31
Mutated	27 (13)	18 (8.4)		
FLG				
Wild-type	167 (81)	187 (87)	0.082	0.25
Mutated	39 (19)	27 (13)		
KRAS				
Wild-type	116 (56)	150 (70)	0.005	0.055
Mutated	90 (44)	64 (30)		
LRP				
Wild-type	166 (81)	191 (89)	0.014	0.13
Mutated	40 (19)	23 (11)		
MUC16				
Wild-type	158 (77)	193 (90)	<0.001	0.006
Mutated	48 (23)	21 (9.8)		
NEB				
Wild-type	165 (80)	195 (91)	0.001	0.024
Mutated	41 (20)	19 (8.9)		
OBSCN				
Wild-type	151 (73)	184 (86)	0.001	0.025
Mutated	55 (27)	30 (14)		
PCLO				

(Continues)

TABLE 2 (Continued)

Gene	Tumour site		*P*-value[b]	q-value[c]
	Right, *N* = 206[a]	Left, *N* = 214[a]		
Wild-type	164 (80)	197 (92)	<0.001	0.007
Mutated	42 (20)	17 (7.9)		
PI3K				
Wild-type	149 (72)	186 (87)	<0.001	0.007
Mutated	57 (28)	28 (13)		
RYR				
Wild-type	158 (77)	183 (86)	0.025	0.20
Mutated	48 (23)	31 (14)		
SMAD4				
Wild-type	181 (88)	202 (94)	0.025	0.20
Mutated	25 (12)	12 (5.6)		
SPTA1				
Wild-type	175 (85)	197 (92)	0.031	0.20
Mutated	31 (15)	17 (7.9)		
SYNE1				
Wild-type	143 (69)	177 (83)	0.002	0.029
Mutated	63 (31)	37 (17)		
TP53				
Wild-type	129 (63)	96 (45)	<0.001	0.007
Mutated	77 (37)	118 (55)		
TTN				
Wild-type	93 (45)	126 (59)	90.006	0.070
Mutated	113 (55)	88 (41)		
UNC13C				
Wild-type	183 (89)	195 (91)	0.52	0.52
Mutated	23 (11)	19 (8.9)		
USH2A				
Wild-type	161 (78)	194 (91)	<0.001	0.010
Mutated	45 (22)	20 (9.3)		
ZFH				
Wild-type	156 (76)	190 (89)	<0.001	0.011
Mutated	50 (24)	24 (11)		

[a]Statistics presented: *n* (%).

[b]Statistical tests performed: Pearson's Chi-squared test with simulated *p*-value (based on 1e+06 replicates).

[c]Holm correction for multiple testing.

(66%–81%), and 59% (49%–72%) for right-sided tumours (Figure 3). Right-sided tumours had significantly worse survival on Cox proportional hazards analysis of 5-year OS, with a hazard ratio (HR) of 1.71 (95% confidence interval (CI) 1.10–2.64, $P = 0.017$).

We then looked at the total effect of the genetic loci measured in TCGA on survival, which, following the DAG, required adjusting for MSI and tumour site. This model showed five genes had a sizeable effect on survival (HR, 95% CI), including DNAH5 (0.39, 0.17–0.89), MUC16 (1.60, 0.91–2.83), NEB (2.00, 1.07–3.76), SMAD4 (2.24, 1.18–4.25), and USH2A (HR 1.65, 0.89–3.05) (Table 3).

Prediction

Lastly, lasso-penalized regression of 5-year OS was then performed to generate a predictive survival model that incorporated all pathological and genomic information present in TCGA. This included TNM stage, surgical resection margin, presence of lymphovascular invasion, MSI, and wild-type or mutated allele status of 25 genes. The highest performing model on cross-validation selected for 13 variables, including TNM stage, R0 status, and nine genetic loci: MUC16 (HR 2.06), USH2A (HR 1.80), SMAD4 (HR 1.61), SYNE1 (HR

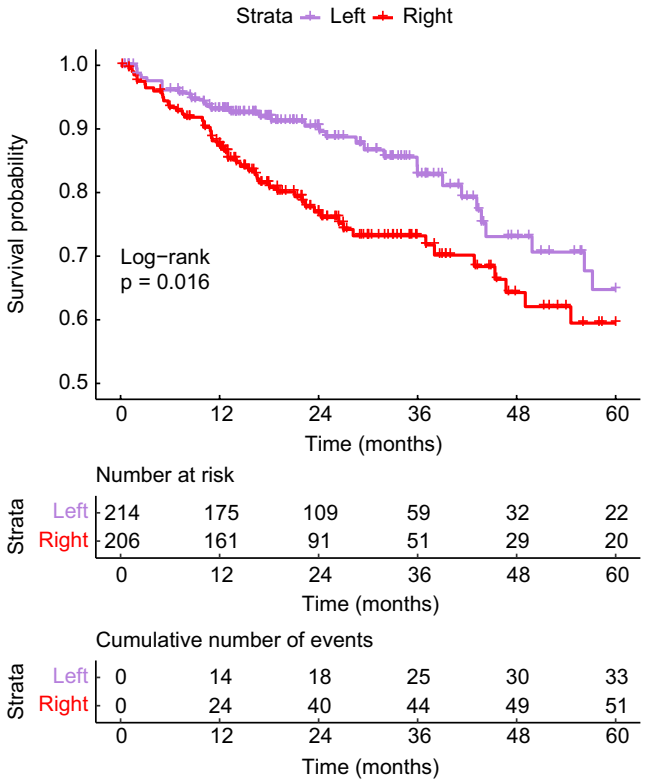

FIGURE 3 Kaplan Meier curves of overall survival in patients with right- (red) and left-sided (lavender) colon tumours. *P*-value calculated with the log-rank test

TABLE 3 Total effect of the genetic loci on 5-year overall survival

Variable	HR	95% CI	*P*-value
APC mutation	0.78	0.49, 1.24	0.30
DNAH11 mutation	0.66	0.33, 1.30	0.23
DNAH5 mutation	0.39	0.17, 0.89	0.026
FLG mutation	1.53	0.82, 2.87	0.18
MUC16 mutation	1.60	0.91, 2.83	0.10
NEB mutation	2.00	1.07, 3.76	0.031
OBSCN mutation	0.66	0.34, 1.27	0.22
PCLO mutation	0.73	0.35, 1.51	0.39
SMAD4 mutation	2.24	1.18, 4.25	0.014
SYNE1 mutation	1.19	0.69, 2.06	0.52
TP53 mutation	1.17	0.73, 1.86	0.51
TTN mutation	1.24	0.76, 2.03	0.40
USH2A mutation	1.65	0.89, 3.05	0.11
ZFH mutation	0.64	0.32, 1.26	0.19

Note: Genes in bold type preferentially mutate at different tumour sites.
Abbreviations: CI, confidence interval; HR, hazard ratio.

1.55), FLG (HR 1.51), NEB (HR 1.48), TTN (HR 1.35), OBSCN (HR 0.49), and DNAH5 (HR 0.37). Post-selection inference demonstrated that mutations in MUC16 ($P = 0.01$) and DNAH5 ($P = 0.02$) were significantly associated with five-year OS after adjustment with all other covariables (Table 4).

TABLE 4 Post selection inference on the variables selected in the highest-performing Lasso-penalized Cox regression of 5-year overall survival

Variable	HR	95% CI	*P*-value
T stage	2.06	1.14, 3.27	0.027
N stage	1.75	1.27, 2.12	0.001
Metastases	1.77	1.15, 3.91	0.025
R1/R2 margin	1.74	0.74, 7.27	0.084
DNAH5 mutation	0.37	0.07, 0.67	0.024
FLG mutation	1.51	0.19, 5.42	0.34
MUC16 mutation	2.06	1.51, 8.14	0.011
NEB mutation	1.48	0.15, 4.78	0.32
OBSCN mutation	0.49	0.13, 2.39	0.20
SMAD4 mutation	1.61	0.58, 5.62	0.12
SYNE1 mutation	1.55	0.45, 5.62	0.20
TTN mutation	1.35	0.01, 2.14	0.68
USH2A mutation	1.80	0.80, 5.86	0.079

Note: Genes in bold type preferentially mutate at different tumour sites.
Abbreviations: CI, confidence interval; HR, hazard ratio.

DISCUSSION AND CONCLUSION

This study demonstrated that the prevalence of mutations at loci genotyped by TCGA varies across right- and left-sided colon cancers and affects survival outcome. We confirmed that patients with right-sided colon cancers have worse survival than patients with left-sided cancers. We hypothesized three causal pathways through which tumour site impacts observed survival differences, including underlying genetic differences between the tumour sites. The use of DAGs allowed us to minimize spurious associations and conclusions by only including variables in our model necessary to avoid confounding pathways while not including excessive variables that could introduce colliders. We revealed that there were differences in mutation frequency in 12 out of 25 genes compared to expected distributions for tumour sites. We then identified the presence of five mutated genes which portended changes in survival.

Survival differences were seen in tumours with mutations in DNAH5, MUC16, NEB, SMAD4, and USH2A in the causal analyses for the total effect of genes on survival. These genes were also selected for by lasso regularization in the predictive model. DNAH5, which encodes for dynein, has previously been reported to be commonly dysregulated in colon cancer, though little investigative work has investigated mechanisms on how it could affect survival [31]. MUC16, produces the CA-125 serum tumour marker for ovarian epithelial cancer. MUC16, unlike BRAF and MSI, has not previously been associated with mucinous status [32]. Prior

works have shown that it protects tumours from the innate immune system's anticancer response by inhibiting natural killer cell recognition. Unfortunately, targeted MUC16 immunotherapy in ovarian cancer patients has not been shown to affect overall survival [33–35]. NEB, nebulin, is a protein that regulates actin filaments, with members of its family being implicated in migration of cancer cells in breast, ovarian, and liver cancer [36]. SMAD4, a tumour suppressor gene, mediates the TGF-β signalling pathway and is commonly mutated in advanced colon cancers [37]. The last gene, USH2A, is likely involved in regulating the actin cytoskeleton and has been found to be frequently mutated in triple-negative breast cancers [38].

MUC16, NEB, and USH2A, while often implicated in other cancers, have not been described as drivers of colon cancer. Interestingly, we showed that all three of these were mutated more frequently in the right colon. These genes could be potential targets of basic science research to help explain the differences in tumour biology that leads to worsened survival and to the investigate why they are more commonly mutated on the right, be it differing baseline gene expression, exposure to bile acids, microbiome differences, or embryological origins [7–9]. However, since TCGA only provided data on a limited number of genes, we do caution that these three genes being mutated may, rather than directly driving mortality, instead indicate a higher patient-specific mutation rate, with other genes driving the tumour's aggressive nature [16].

We also created a predictive model that could be used to help prognosticate colon cancer patients beyond current rudimentary TNM staging, which included mutations at nine genetic loci. Tumour genotypes may help clinicians with prognostication and treatment decisions. We have already seen genome-informed treatment decisions for some systemic therapeutics [39–41]. Other cases, such as the worse efficacy of cetuximab in right-sided cancers, leads one to ask if a more specific genetic marker exists that could help one discriminate between patients who will respond best to a certain systemic therapy beyond tumour site [42]. For example, mutations in SMAD4, one of the selected loci, has been shown to increase 5-flurouracil resistance of colon cancer cells [43]. An additional possible use of genetic markers would be to discriminate between those patients who would do well with shorter courses of adjuvant chemotherapy [44].

This study has limitations related to the retrospective nature of case selection. In addition, the tumour samples derive from TCGA's twenty contributing North American sites and thus may not fully represent the general population. Limited information on chemotherapeutic regimens limited the precision of estimates on the effect of mutations on survival. Additionally, the data is limited, with 84 deaths in 420 patients. Beyond sample size alone, the power of our analyses, which examined 5-year OS, was limited by the 2-year median follow-up. Despite the power limitations, though, we still were able to capture a number of substantial effects for different genes on 5-year overall survival, rather than needing to use less clinically useful metrics to increase our power such as disease-free survival. Our analyses also only captured a limited amount of epigenetic phenomenon, such as MSI, which we know exerts a strong effect on survival in colorectal cancer [45]. As previously discussed, of the genes reported by TCGA, we only included the 25 genes TCGA denoted as somatic recurrently mutated genes in colon cancer, so other genetic drivers of patient survival may exist. Lastly, mutations at the 25 loci could just reflect higher patient-specific mutation rate. This is less of a concern since many of the genes we found have evidence for their direct link with cancer, and, as previously mentioned, TCGA minimized including bystander genes [13,16]. Given colon cancer's heterogeneous nature, it will be important to validate the study's findings as a foundation for future work on datasets with a more diverse and larger patient population.

In conclusion, our study showed that genetic mutations, with prevalence differences in right- and left-sided colon cancers, may lead to differences in overall survival. Further studies on larger patient populations may identify other genes that impact survival differences, which could form the foundation for more precise prognostication and treatment decisions beyond current rudimentary TNM staging.

FUNDING INFORMATION

The authors have no relevant financial or non-financial interests to disclose.

CONFLICT OF INTEREST

Dr Ward receives research support from the Olympus Corporation. Ms Stafford and Drs Cauley, Goldstone, Bordeianou, Kunitake, Berger, and Ricciardi have no conflicts of interest nor financial ties to disclose.

AUTHOR CONTRIBUTIONS

All manuscript authors: made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data, drafted the work or revised it critically for important intellectual content, approved the version to be published, agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ETHICAL STATEMENT

The authors declare they have no financial interests.

DATA AVAILABILITY STATEMENT

The data and analyses that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.4773064

ORCID

Thomas M. Ward https://orcid.org/0000-0003-1965-8657

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of

incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Bufill JA. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. Ann Intern Med. 1990;113(10):779.
3. Wray CM, Ziogas A, Hinojosa MW, Le H, Stamos MJ, Zell JA. Tumor subsite location within the colon is prognostic for survival after colon cancer diagnosis. Dis Colon Rectum. 2009;52(8):1359–66.
4. Benedix F, Kube R, Meyer F, Schmidt U, Gastinger I, Lippert H, et al. Comparison of 17,641 patients with right- and left-sided colon cancer: differences in epidemiology, perioperative course, histology, and survival. Dis Colon Rectum. 2010;53(1):57–64.
5. Petrelli F, Tomasello G, Borgonovo K, Ghidini M, Turati L, Dallera P, et al. Prognostic survival associated with left-sided vs right-sided colon cancer: a systematic review and meta-analysis. JAMA Oncol. 2017;3(2):211–9.
6. Yahagi M, Okabayashi K, Hasegawa H, Tsuruta M, Kitagawa Y. The worse prognosis of right-sided compared with left-sided colon cancers: a systematic review and meta-analysis. J Gastrointest Surg. 2016;20(3):648–55.
7. Missiaglia E, Jacobs B, D'Ario G, Di Narzo AF, Soneson C, Budinska E, et al. Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. Ann Oncol. 2014;25(10):1995–2001.
8. Bernstein C, Holubec H, Bhattacharyya AK, Nguyen H, Payne CM, Zaitlin B, et al. Carcinogenicity of deoxycholate, a secondary bile acid. Arch Toxicol. 2011;85(8):863–71.
9. Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ, et al. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut. 2017;66(4):633–43.
10. Lee MS, Menter DG, Kopetz S. Right versus left colon cancer biology: integrating the consensus molecular subtypes. J Natl Compr Canc Netw. 2017;15(3):411–9.
11. Jiang Y, Yan X, Liu K, Shi Y, Wang C, Hu J, et al. Discovering the molecular differences between right- and left-sided colon cancer using machine learning methods. BMC Cancer. 2020;20(1):1012.
12. Menigatti M, Truninger K, Gebbers J-O, Marbet U, Marra G, Schär P. Normal colorectal mucosa exhibits sex- and segment-specific susceptibility to DNA methylation at the hMLH1 and MGMT promoters. Oncogene. 2009;28(6):899–909.
13. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330–7.
14. Amin MB, Edge S, Greene F, Byrd DR, Brookland RK, Washington MK, et al. AJCC Cancer Staging Manual [Internet], 8th ed. Springer International Publishing; 2017 [cited 2021 Feb 13]. Available from: https://www.springer.com/gp/book/9783319406176
15. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology. 2010;138(6):2073–87. e3.
16. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
17. Loree JM, Kopetz S, Raghav KPS. Current companion diagnostics in advanced colorectal cancer; getting a bigger and better piece of the pie. J Gastrointest Oncol. 2017;8(1):199–212.
18. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/
19. Kuhn M, Johnson K. Applied Predictive Modeling. Springer; 2013. p. 600.
20. Jordanov I, Petrov N, Petrozziello A. Classifiers accuracy improvement based on missing data imputation. J Artif Intell Soft Comput Res. 2017;8(1):31–48.
21. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox model. Springer; 2000.
22. Textor J, Zander Bvd, Gilthorpe MS, Liśkiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int J Epidemiol. 2016;45(6):1887–94.
23. Rohrer JM. Thinking clearly about correlations and causation: graphical causal models for observational data. Adv Methods Pract Psychol Sci. 2018;1(1):27–42.
24. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19(8):1059–79.
25. Kassambara A, Kosinski M, Biecek P. survminer: Drawing survival curves using "ggplot2"; 2020. Available from: https://CRAN.R-project.org/package=survminer. Accessed 1 Feb 2022.
26. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5):1–13.
27. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. Ann Stat. 2016;44(3):907–27.
28. Taylor J, Tibshirani R. Post-selection inference for l1-penalized likelihood models. Can J Stat. 2018;46(1):41–61.
29. Ward TM. tmward, tcga. Tumor genotypes account for survival differences in right and left-sided colon cancers analyses code [Internet]. Zenodo; 2021. https://doi.org/10.5281/zenodo.4773064. Accessed 1 Feb 2022.
30. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa J-PJ. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA. 1999;96(15):8681–6.
31. Xiao WH, Qu XL, Li XM, Sun YL, Zhao HX, Wang S, et al. Identification of commonly dysregulated genes in colorectal cancer by integrating analysis of RNA-Seq data and qRT-PCR validation. Cancer Gene Ther. 2015;22(5):278–84.
32. Reynolds IS, Furney SJ, Kay EW, McNamara DA, Prehn JHM, Burke JP. Meta-analysis of the molecular associations of mucinous colorectal cancer. Br J Surg. 2019;106(6):682–91.
33. Yin BWT, Dnistrian A, Lloyd KO. Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. Int J Cancer. 2002;98(5):737–40.
34. Gubbels JA, Felder M, Horibata S, Belisle JA, Kapur A, Holden H, et al. MUC16 provides immune protection by inhibiting synapse formation between NK and ovarian tumor cells. Mol Cancer. 2010;9(1):11.
35. Felder M, Kapur A, Gonzalez-Bosquet J, Horibata S, Heintz J, Albrecht R, et al. MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. Mol Cancer. 2014;13(1):129.
36. Pappas CT, Bliss KT, Zieseniss A, Gregorio CC. The Nebulin family: an actin support group. Trends Cell Biol. 2011;21(1):29–37.
37. Zhang B, Halder SK, Kashikar ND, Cho Y, Datta A, Gorden DL, et al. Antimetastatic role of Smad4 signaling in colorectal cancer. Gastroenterology. 2010;138(3):969–980. e3.
38. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486(7403):395–9.
39. van Dijk E, Biesma HD, Cordes M, Smeets D, Neerincx M, Das S, et al. Loss of chromosome 18q11.2-q12.1 is predictive for survival in patients with metastatic colorectal cancer treated with bevacizumab. J Clin Oncol. 2018;36(20):2052–60.
40. Pietrantonio F, Petrelli F, Coinu A, Di Bartolomeo M, Borgonovo K, Maggi C, et al. Predictive role of BRAF mutations in patients with advanced colorectal cancer receiving cetuximab and panitumumab: a meta-analysis. Eur J Cancer. 2015;51(5):587–94.
41. Schrock AB, Ouyang C, Sandhu J, Sokol E, Jin D, Ross JS, et al. Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer. Ann Oncol. 2019;30(7):1096–103.

42. Brulé SY, Jonker DJ, Karapetis CS, O'Callaghan CJ, Moore MJ, Wong R, et al. Location of colon cancer (right-sided versus left-sided) as a prognostic factor and a predictor of benefit from cetuximab in NCIC CO.17. Eur J Cancer. 2015;51(11):1405–14.
43. Papageorgis P, Cheng K, Ozturk S, Gong Y, Lambert AW, Abdolmaleky HM, et al. Smad4 inactivation promotes malignancy and drug resistance of colon cancer. Cancer Res. 2011;71(3):998–1008.
44. André T, Meyerhardt J, Iveson T, Sobrero A, Yoshino T, Souglakos I, et al. Effect of duration of adjuvant chemotherapy for patients with stage III colon cancer (IDEA collaboration): final results from a prospective, pooled analysis of six randomised, phase 3 trials. Lancet Oncol. 2020;21(12):1620–9.
45. Okugawa Y, Grady WM, Goel A. Epigenetic alterations in colorectal cancer: emerging biomarkers. Gastroenterology. 2015;149(5):1204–25. e12.

How to cite this article: Ward TM, Cauley CE, Stafford CE, Goldstone RN, Bordeianou LG, Kunitake H, et al. Tumour genotypes account for survival differences in right- and left-sided colon cancers. Colorectal Dis. 2022;00:1–10. https://doi.org/10.1111/codi.16060