Colon Cancer Tumor Location Defined by Gene Expression May Disagree With Anatomic Tumor Location

Emily Cannon, Steven Buechler

Abstract

Tumor location is increasingly considered in designing treatment of colon tumors. In a microarray dataset (n = 871), we showed that tumor locations predicted by gene expression may disagree with anatomic locations, and were superior to anatomic locations in specifying key molecular features of the tumors. Mesenchymal tumors (CMS4) were likely to have ambiguous location as predicted by gene expression.

Background: Cancers of the right colon have been shown to differ from left-side colon cancers in prognosis, response to epithelial growth factor receptor inhibitors, microsatellite instability and BRAF mutation status, and other molecular characteristics. Clinical application of these differences will benefit from a deeper understanding of how tumor location defines and is defined by gene expression. **Materials and Methods:** This study was carried out using Affymetrix microarray datasets (Cohort A: training set, n = 352; validation set, n = 519) and samples from The Cancer Genome Atlas Colon Adenocarcinoma database (Cohort B: n = 408), in which tumor location was reported. Gene expression patterns characteristic of tumor side were identified in a manner unbiased by statistical classification method. **Results:** In the Cohort A validation set, the anatomic locations of 75% of tumors agree with the locations predicted by gene expression (so-called genomic location), whereas 8% of tumors had genomic locations discordant with their anatomic locations, and 17% of tumors had ambiguous genomic locations. Genomic location was a better predictor of microsatellite instability, CpG island methylator phenotype status, and BRAF mutation status than anatomic location. Tumors with ambiguous genomic location were significantly ($P = 1.3 \times 10^{-7}$) more likely to have the mesenchymal consensus molecular subtype (40%) than those with a specific genomic location (18%). A genomic signature to predict genomic location was defined. **Conclusion:** Tumor location is increasingly considered in deciding treatment of a colon tumor. We showed that genomic location was superior to anatomic location as a predictor of molecular characteristics, suggesting that it may be a more accurate predictor of response.

Clinical Colorectal Cancer, Vol. 18, No. 2, 149-58

Keywords: Genomic signature, Left-side colon cancer, Molecular classification, Right-side colon cancer, Tumor sidedness

Background

Colorectal cancer is the fourth most common cancer in the United States, with heterogeneous outcomes and responses to treatment. Epidemiologic and histologic differences between left-side and right-side colon cancers have been observed, with right-side tumors occurring more frequently in women and older persons, and they are more commonly mucinous and having signet ring histology. At the molecular level, tumors that are microsatellite instable (MSI), *BRAF* mutated, CpG island methylator phenotype (CIMP)-high, or consensus molecular subtype (CMS) CMS1 are more likely right-sided,[1-4] and tumors that are chromosome instable, *APC* or *TP53* mutated, or CMS2 are more likely left-sided.[3-5] Right-sided tumors also have poorer prognosis than left-sided tumors,[6] and apparently, left and right-sided tumors follow different pathways to metastasis.[7] In fact, gene expression patterns can distinguish right-sided colon tumors from left-sided tumors,[8] and the single gene *PRAC1* can predict tumor side with 85% accuracy.[7]

Recently, retrospective analysis of data from the clinical trial Cancer and Leukemia Group B (CALGB) 80405 revealed that left-

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN

Submitted: Dec 7, 2018; Revised: Jan 22, 2019; Accepted: Feb 4, 2019; Epub: Feb 14, 2019

Address for correspondence: Steven Buechler, PhD, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 102 Crowley Hall, Notre Dame, IN 46556
E-mail contact: steve@nd.edu

1533-0028/$ - see frontmatter
https://doi.org/10.1016/j.clcc.2019.02.002

Table 1 Patient and Tumor Characteristics in the Cohort A Training and Validation Sets and Cohort B

	Cohort A Training (n = 352)[a]	Cohort A Validation (n = 519)[b]	Cohort B (n = 408)[c]
Stage			
1	45	27	64
2	137	247	160
3	116	200	116
4	0	0	58
NA	54	45	10
Microsatellite status			
MSI	14	58	8
MSS	91	311	76
NA	247	150	324
Median age, y	68	68	68
Gender			
Male	198	276	221
Female	154	243	187
Location			
Left	192	308	165
Right	160	211	243
CIMP status			
High	15	67	74
Low	46	102	89
Negative	60	170	239
NA	241	180	6
BRAF mutant			
Yes	6	41	2
No	108	311	23
NA	238	167	383
KRAS mutant			
Yes	43	150	17
No	72	232	21
NA	237	137	370

Abbreviations: CIMP = CpG island methylator phenotype; MSI = microsatellite instable; MSS = microsatellite stable; NA = not available.
[a]GSE17536 (n = 123), GSE39582 (n = 117), GSE14333 (n − 112, duplicate samples in GSE17536 were eliminated).
[b]GSE37892 (n = 120), GSE39582 (n = 399).
[c]Samples from The Cancer Genome Atlas Colon Adenocarcinoma database with tumor location reported.

side metastatic KRAS wild-type (WT) colon tumors were more likely to respond favorably to FOLFOX (leucovorin, 5-fluorouracil, and oxaliplatin) plus anti-epithelial growth factor receptor (EGFR) (cetuximab) therapy than right-side KRAS WT tumors.[9] Consequently, tumor location is now considered in selecting a treatment regimen for patients with colon cancer with unresectable synchronous liver or lung metastases in the National Comprehensive Cancer Network guidelines (https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf). A deeper understanding of the extent and nature of molecular variances by tumor location may help doctors select the most effective treatment.[2,10,11]

In this article, we developed methods that determine whether a colon tumor sample is more like a right-sided tumor or left-sided tumor with respect to gene expression. The resulting notions of genomic right-sided and genomic left-sided tumors were defined so as to be unbiased with respect to statistical methodology. We argued that there are anatomical right-sided tumors that are genomically more like left-sided tumors than right-sided tumors, and conversely, there are anatomical left-sided tumors that are genomically more similar to right-sided than other left-sided tumors. A gene expression signature to estimate genomic location will also be reported.

Materials and Methods

Patient Cohorts With Whole-genome Analysis

Location prediction methods were developed and validated using gene expression data from multiple datasets combined to form Cohort A (Table 1). All datasets (GSE14333, GSE17536, GSE37892, GSE39582) consisted of colon cancer samples hybridized to Affymetrix hgu13plus2 arrays. Data was accessed through Colorectal Cancer Subtyping Consortium (CRCSC) (Synapse Project 2623706; https://www.synapse.org/). Duplicate samples in the GSE14333 and GSE17536 datasets were eliminated. Gene expression values

normalized using fRMA[12] were obtained from the Synapse Project 2623706 for all datasets.

In all datasets, a colon tumor was considered to be right-sided if it occurred proximally to the splenic flexure, and a colon tumor was considered left-sided if it occurred at or distally to the splenic flexure. Rectal tumors were not considered here. Throughout this study, colon tumor sidedness was defined in this way.

Clinical characteristics and the CRCSC consensus molecular subtype classifications for each sample were also obtained from Synapse Project 2623706 and are summarized in Table 1. Location annotation was available for all samples in the dataset. For classifier development, the dataset was partitioned into training (n = 352) and validation (n = 519) sets. Specifically, the Cohort A training set was defined as those samples from GSE17536, GSE14333, and the samples in GSE39582 described as the "validation dataset" in Marisa et al[13]; the Cohort A validation set consisted of GSE37892 and the remaining samples from GSE39582. The partition was made to ensure similar location distributions in each. In the Cohort A training partition, 192 samples were anatomically left (55%) and 160 were anatomically right (45%). The Cohort A validation set contained 308 anatomically left samples (59%) and 221 that were anatomically right (41%).

Cohort B consisted of samples from TCGA-COAD[3] for which tumor location was reported (Table 1). Gene expression data and patient characteristics were obtained from the National Cancer Institute Genomic Data Commons (https://gdc.cancer.gov).

Statistical Tools

All statistical analyses were performed using R (http://www.r-project.org) and Bioconductor packages (http://bioconductor.org). Different classification methods were used to classify tumor samples based on their transcriptomic profiles as either left-side or right-side. Support Vector Machines (SVMs) with multiple kernels were computed with the *e1071* package.[14] The *rpart* package was used to build the decision tree classifier,[15] and the random forest classifier was built using the *randomForest* package.[16] Throughout the analysis, significant overlap between groups or enrichment of a certain characteristic within a group was determined with a hypergeometric test using the phyper function from the *stats* package.

Consensus Molecular Subtypes

Using 6 independent classification methods and similar consensus networking analysis, the CRCSC defined 4 consensus molecular subtypes: CMS1, CMS2, CMS3, CMS4, and Mixed (non-consensus).[4] These subtypes have significant biological and clinical aspects associated with them. For all samples in this dataset, the CRCSC reports CMS classifications obtained by network analysis, the random forest classifier, and a "CMS-final" classification that agrees with the network classification when there is a consensus among the contributing classifiers, and agrees with the random forest classifier for non-consensus samples. In this article, the term "CMS subtype" refers to "CMS-final" unless otherwise specified.

Genomic Score Predicting Subtype Methods

A continuous score that predicts genomic location was defined using multistate gene methodology.[17] To apply this method, the vector of gene expression values in a microarray dataset were replaced by a so-called risk score. Descriptively, the risk score of a gene is an association of expression values to numbers 0 to 1, monotonically increasing or decreasing, so that higher values of the score correspond to increased risk of relapse, and a large majority of samples (over 80%) are greater than 0.75 or less than 0.25. Analytically, the risk score of a gene is derived from the gene expression values by combining a Gaussian mixture model with model averaging.

In the multistate genomic score methodology, after selecting a set of gene risk scores to include in the computation, a continuous score was computed as the mean of the risk score values for each sample.

A corresponding genomic score for a cohort of samples with gene expression assayed with a possibly different technology was defined as follows. First, for each gene used in computing the score, a vector of expression values of the genes was selected. If multiple vectors were possible owing to multiple array probes for the gene, we selected one with maximal variance. Subsequently, a risk score was computed for each of these vectors using the same statistical methods as employed in the original derivation. Finally, the genomic score was computed for this cohort as the mean of the risk score values for each sample.

Results

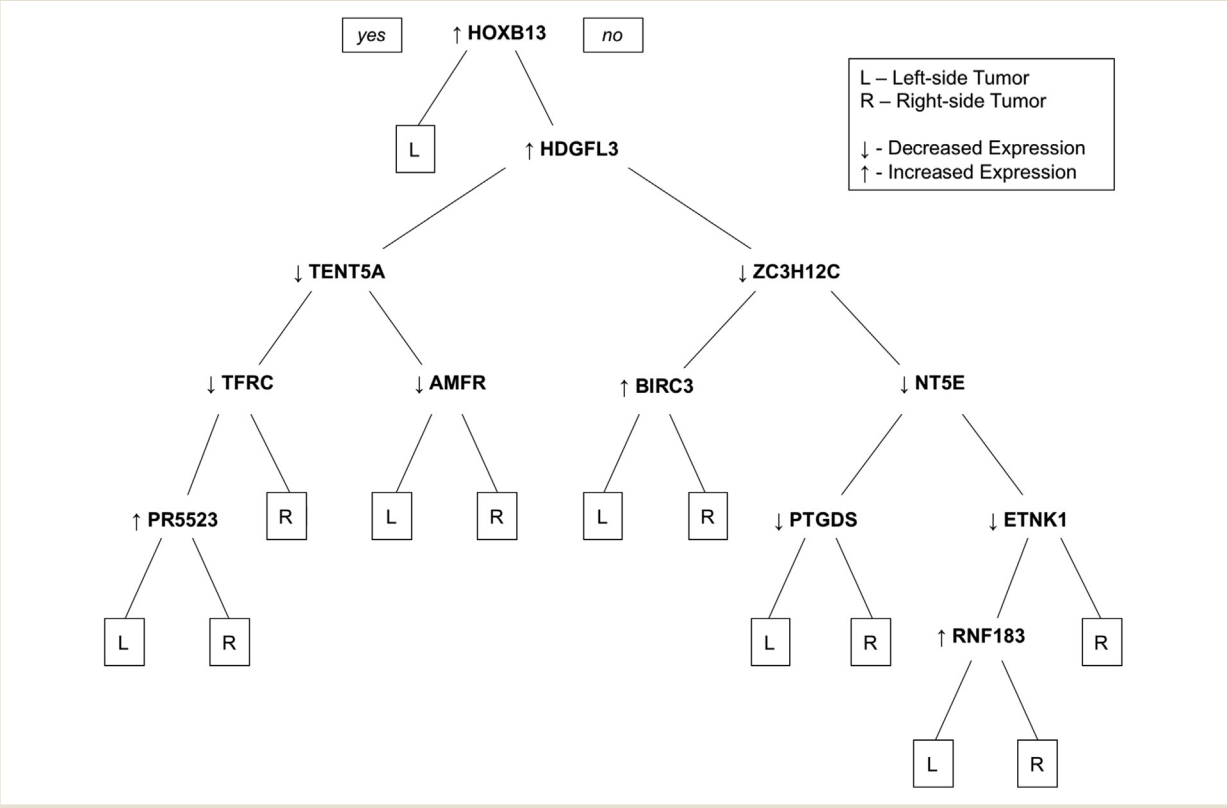
Scheme for Unbiased Prediction of Tumor Location by Gene Expression

To identify genomic features characteristic of anatomic left-side or right-side tumors, we define predictors of anatomic location based on gene expression. Statistical classification methods were used to predict anatomic location using gene expression data in the Cohort A training set. To avoid biasing the prediction to a particular classification method, multiple predictors were developed using different classification methods, and sets of samples predicted to be left-sided or right-sided by a consensus of predictors were identified. Samples that were predicted to be left-sided, respectively right-sided, by a consensus of predictors, were called genomic left-sided tumors, respectively, genomic right-sided tumors. Following the definition of these subtypes, the degrees of agreement with anatomic location and other molecular tumor features were studied.

Independent Classifiers of Location

Independent classifiers were built and validated using SVM with linear, polynomial, and radial kernels, a decision tree, and a random forest. Each method was developed independently using the Cohort A training set with anatomic location as the response variable of the classifier. To begin, the set of probes in the microarray data was restricted to those with coefficient of variation > 0.7 (n = 2081). These features were used for all classifiers. For an SVM classification with linear kernel, the cost parameter of the model was set at the value that minimized error in 10-fold cross validation of the training dataset (cost = 10,000). SVM classifiers with radial and polynomial kernels were defined similarly. Using the 2081 features described above as variables, a decision tree was built to predict anatomic location. Using cross validation, the full decision tree had the lowest cross validation error within 1 standard deviation of the smallest tree; hence, the final decision tree was unpruned. This tree used a total of 12 genomic features (Figure 1). A random forest classifier

Figure 1 The Final Decision Tree for Tumor Side Computed With Gene Expression in the Cohort A Training Set. Up Arrow Indicates Increased Expression; Down Arrow Indicates Decreased Expression



Abbreviations: L = left-side tumor; R = right-side tumor.

was also built using the same set of variables as the other classifiers. The number of trees (10,001) was chosen to achieve the most accurate classification in the training dataset, while maintaining a reasonable computation time. An odd number of trees was needed to avoid random ties.

Samples predicted to be left-sided or right-sided by each of these 5 classifiers in the Cohort A validation set were computed and compared with anatomic location (Table 2). Each classification method resulted in a group of discordant samples for which the anatomic location was different from the predicted genomic location (Table 3). There was a high degree of overlap between the location predicted by different classifiers, even for the samples whose predicted location disagrees with anatomic location, suggesting that discordance is a feature of the tumor genome and not simply owing to statistical error.

Consensus Genomic Left and Genomic Right Subtypes

Each of the 5 classifiers defines a set of samples in Cohort A that it predicts to be right-sided or left-sided tumors. We next identified the samples for which there was a consensus prediction among the classifiers. Specifically, for each sample we computed the probability that the multiple classifiers predicted the sample to be left-sided or right-sided using a hypergeometric test. A sample was termed genomic left-sided, respectively, genomic right-sided if said probability of the hypergeometric test for prediction as left-sided, respectively, right-sided, was less than .05. Eighty-three percent of samples were classified as genomic left-sided or genomic right-sided. There was significant concordance between genomic location and anatomic location; however, there was discordance, giving rise to 6 distinct subtypes (with the percentage of samples in the Cohort A validation set): anatomic right – genomic right (aR-gR, 28%), anatomic right – genomic left (aR-gL, 5%), anatomic left – genomic right (aL-gR, 4%), anatomic left – genomic left (aL-gL, 46%), anatomic left – genomic noncensensus (aL-gNC, 9%), and anatomic right – genomic noncensensus (aR-gNC, 8%). Because the genomic nonconsensus samples are divided equally by anatomic location, we subsequently combine them into a single group

Table 2 Training and Validation Error Associated With Each Independent Classifier

	Training Error, %	Validation Error, %
SVM – Linear	0.00	15.99
SVM – Radial	0.00	21.39
SVM – Polynomial	0.00	15.06
Decision Tree	6.79	21.34
Random Forest	25.60	26.78

Abbreviation: SVM = support vector machine.

Table 3 Genomic and Anatomic Location Discordance by Side and Classifier

	Anatomic Left		Anatomic Right	
	Samples, n	Discordance, %	Samples, n	Discordance, %
Random Forest				
Left	429	14.20	160	43.13
Right	71		211	
Decision Tree				
Left	440	12.00	74	19.95
Right	60		297	
SVM — Linear				
Left	476	4.80	58	15.63
Right	24		313	
SVM — Radial				
Left	475	5.00	84	22.64
Right	25		287	
SVM — Polynomial				
Left	476	4.80	64	15.88
Right	24		339	

Abbreviation: SVM = support vector machine.

(gNC). Classifications for each sample in Cohort A are found in Supplemental Table 1 (in the online version).

Molecular and Clinical Features of Genomic Location

Several molecular and clinical features of colon cancer, notably, MSI, CIMP, BRAF mutations, KRAS mutations, and gender are differentially distributed by anatomic location. Figure 2 displays the distribution of these features among the anatomic and genomic subtypes. χ^2 tests were performed to determine the association of these characteristics within the genomic subtypes; both training and validation sample cohorts were used for this analysis. MSI (reported in 471 samples) is more common in right-sided tumors than left-sided tumors.[1] We found that in anatomic left tumors, MSI is significantly more common in genomic right-side tumors than others ($P = 1.4 \times 10^{-13}$), and in anatomic right tumors, MSI is significantly less common in genomic left tumors than others ($P = .02$). This suggests that the difference in distribution of MSI by tumor location is owing to genomic traits. Similarly, CIMP high status was significantly ($P = 1.1 \times 10^{-14}$) more common in aL-gR than other anatomic left tumors. BRAF mutants have previously been shown to be enriched in right-side colon cancer.[3] In this dataset, BRAF mutations were only found in genomic right-side tumors. Interestingly, KRAS-mutated tumors, which were significantly more common in anatomic right-side tumors than left-side tumors ($P = 9.4 \times 10^{-6}$), were not differentially distributed by genomic location within the anatomic right-sided or anatomic left-sided tumors. Finally, right-side tumors are reportedly more common in females than in males[2] and were found so in Cohort A ($P = .006$). Among the anatomic left-side tumors, the genomic right-side tumors were not more likely to be female ($P = .16$), and among the anatomic right-side tumors, the genomic left-side tumors were not less likely to be female ($P = .76$).

Genomic Location and Consensus Molecular Subtypes

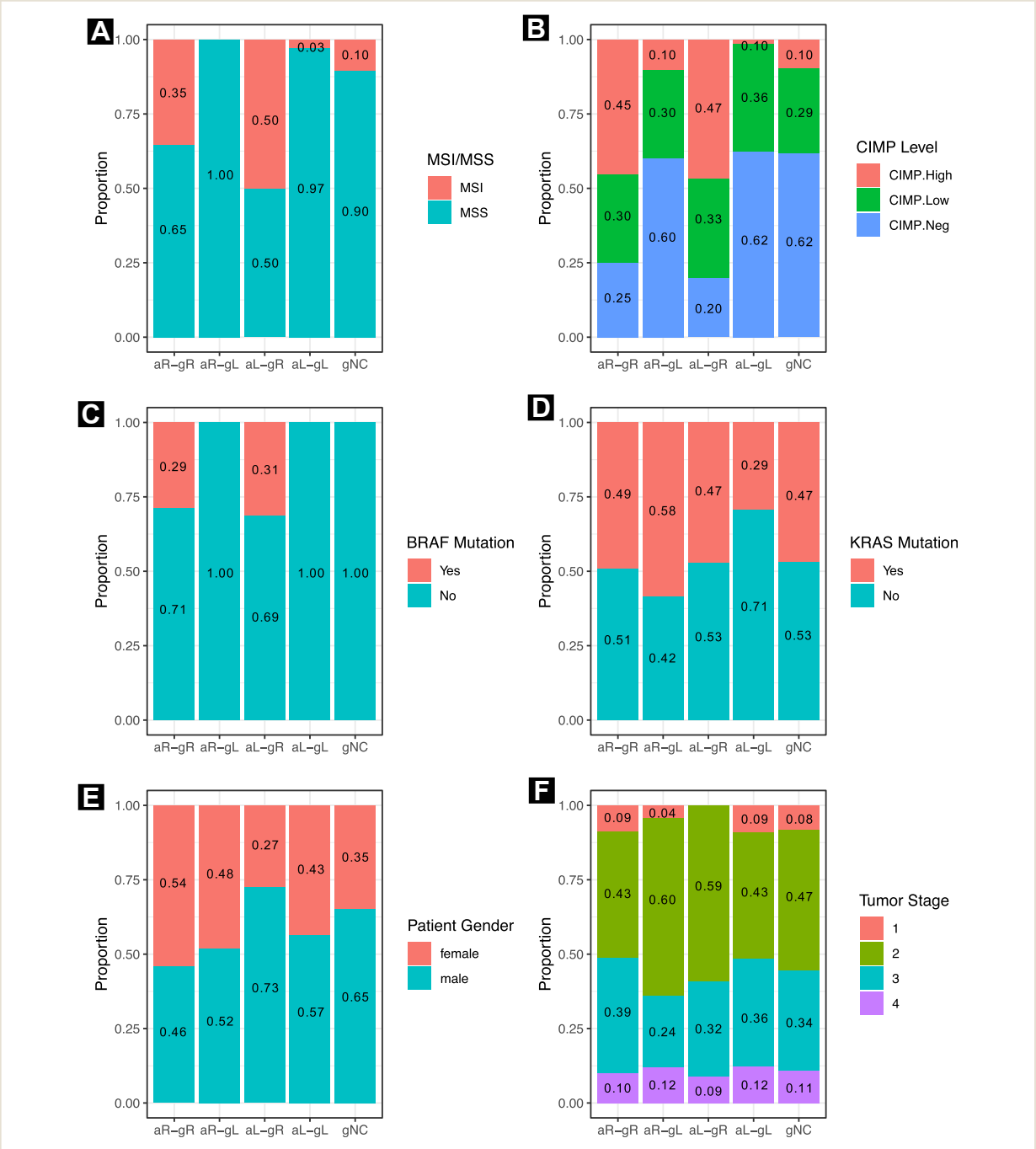
The distribution of consensus molecular subtypes with respect to anatomic and genomic location was analyzed (Figure 3). In Cohort A, and previously reported in other cohorts,[2] CMS1 and CMS3 were significantly more common in anatomic right-side tumors, CMS2 is more common in anatomic left-side tumors, and CMS4 is not biased to a location. We also found that, among the anatomic left-side tumors, CMS1 was significantly more common in the genomic right-side than the genomic left-side tumors ($P = 1.7 \times 10^{-15}$). Similarly, CMS2 is significantly biased to aR-gL among the anatomic right-side tumors ($P = 8.8 \times 10^{-5}$), whereas CMS3 is not significantly biased to aL-gR among the anatomic left-side tumors. Mixed subtype tumors were not biased with respect to location.

Most notably, the proportion of CMS4 samples in the gNC subgroup (40%) was significantly higher than in the samples with a clearly defined genomic location (18%) ($P = 1.3 \times 10^{-7}$). Apparently, CMS4 tumors, described as mesenchymal or stem-like subtype, are not only evenly distributed between the anatomic left and right tumors, but do not have a clear genomic location.

Score Classifier to Predict Genomic Location

The multiple statistical models used to define genomic right-side and genomic left-side tumors make it cumbersome to compute these subtypes in other cohorts. For this reason, we defined a genomic score that predicts genomic location as follows. To begin selection of genes to include in the classifier, the probes in the Affymetrix hgu133plus2 array were restricted to those with a non-missing ENTREZ identifier and having inter-quartile range > 0.6 in the Cohort A training set (n = 17,887). Subsequently, the expression values for these probes in Cohort A were replaced by gene risk scores (see Methods). Using a binary indicator for genomic location, these gene risk scores were ranked by the level of significance with which

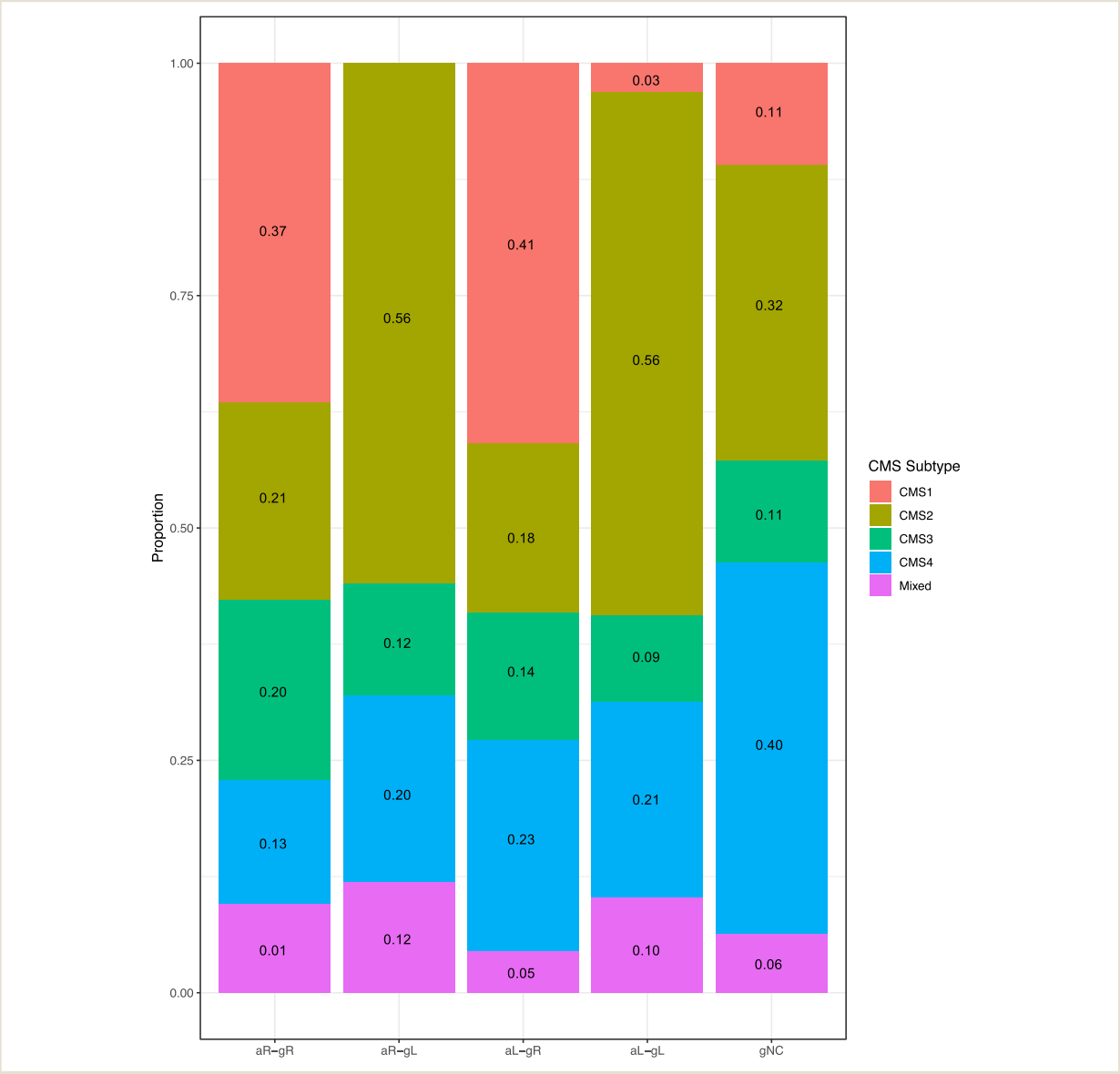
Figure 2 The Proportions of Samples in Each of the Categories Defined by Anatomic and Genomic Location for Multiple Clinicopathologic Features Were Plotted. Genomic Location Is a Significant Predictor of Within the Anatomic Left-side Tumors for MSI Status ($P = 1.4 \times 10^{-13}$) (A), CIMP Status ($P = 1.1 \times 10^{-14}$) (B), and BRAF Mutation Status ($P = 6.7 \times 10^{-19}$) (C). The Differences in Distribution With Respect to Genomic Location and Anatomic Location Are Not Statistically Significant for KRAS Mutation Status (D), Patient Gender (E), and Tumor Stage (F)



Abbreviations: aL-gL = anatomic left – genomic left; aL-gR = anatomic left – genomic right; aR-gL = anatomic right – genomic left; aR-gR = anatomic right – genomic right; CIMP = CpG island methylator phenotype; gNC = genomic noncensensus; MSI = microsatellite instable; MSS = microsatellite stable.

they predict the genomic left samples using a linear model in the Cohort A training set. Candidate scores were computed (see Methods) using the highest ranked 2, 5, 7, 10, or 15 risk scores and assessed as predictors of the genomic location using the receiver operating characteristic method. We evaluated each score S as a predictor of the genomic left samples, and 1-S as a predictor of the genomic right samples. The highest area under the curve (AUC) values for both left (AUC = 0.95) and right (AUC = 0.94) were

Figure 3 The Proportions of Samples in Each of the Categories Defined by Anatomic and Genomic Location for the Consensus Molecular Subtypes Were Plotted. The Proportion of CMS4 Samples in gNC (40%) Is Significantly ($P = 1.3 \times 10^{-7}$) Greater Than in Other Tumors (18%). Genomic Location Is Also Superior to Anatomic Location in Predicting CMS1 (Right-side Dominant) and CMS2 (Left-side Dominant)



Abbreviations: aL-gL = anatomic left − genomic left; aL-gR = anatomic left − genomic right; aR-gL = anatomic right − genomic left; aR-gR = anatomic right − genomic right; CMS = consensus molecular subtype; gNC = genomic noncensensus.

obtained with the 7-probe score (Table 4), hereafter called the genomic location score. In the Cohort A validation set, the genomic location score predicted genomic location with AUC = 0.90 for genomic left-side tumors (Figure 4A) and AUC = 0.91 for genomic right-side tumors (Figure 4B).

Owing to the high degree of overlap between genomic location and anatomic location, the genomic location score is also predictive of anatomic location (AUC = 0.90, in the Cohort A validation set). The ability of genomic location score to predict genomic location goes beyond prediction of anatomic location as follows. In the Cohort A validation set, the genomic location score predicted genomic left-side tumors within the anatomic right-side tumors (Figure 4C) (AUC = 0.86) and genomic right-side tumors within the anatomic left-side tumors (Figure 4D) (AUC = 0.89). Score values for each sample in Cohort A are found in Supplemental Table 1 (in the online version).

To show that the utility of the genomic location score is not limited to Cohort A, it was computed for samples in Cohort B using gene expression values for the 7-panel genes (Table 4) assessed by RNA-seq (see Methods). The genomic location score predicted

Table 4 Genes Used to Compute Genomic Location Score

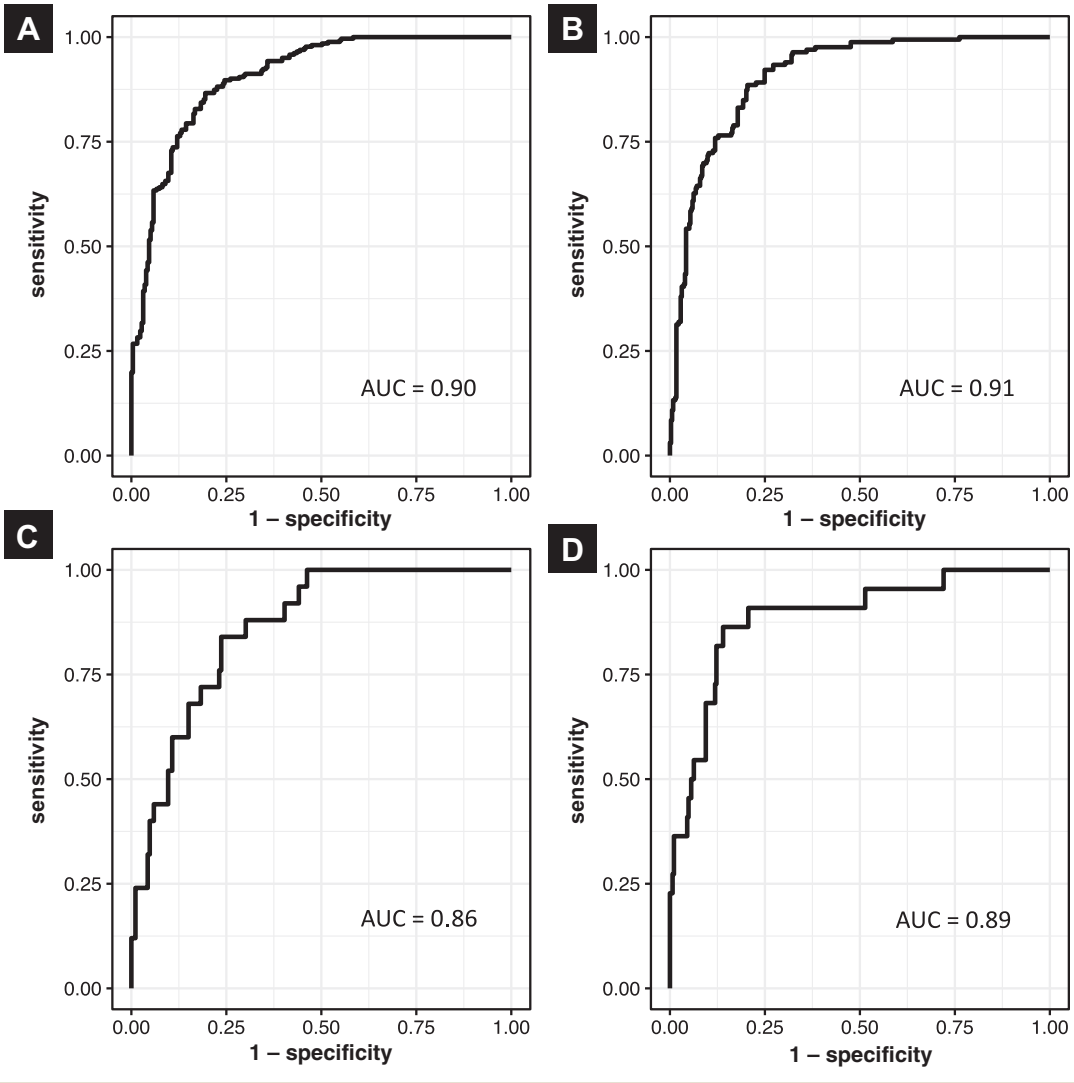
Feature	ENTREZID	SYMBOL	Side of Elevation
230784_at	84366	PRAC1	Left
206858_s_at	3223	HOXC6	Right
209844_at	10481	HOXB13	Left
228260_at	1993	ELAVL2	Left
228821_at	84620	ST6GAL2	Left
217665_at	126017	ZNF813	Left
236300_at	5139	PDE3A	Left

anatomic location in Cohort B with AUC = 0.92. Thus, the relationship between genomic location score and anatomic location was comparable in Cohort B and Cohort A validation set.

Discussion

Tumor location is increasingly a consideration in planning the treatment of colon cancer.[18] Because anti-tumor drugs act at the molecular level, there is need for greater knowledge of the molecular differences between left- and right-side colon cancer. In this study, we analyzed the degree to which gene expression determines the side of a tumor. Using a methodology that reduces bias owing to statistical methodology, we showed that gene expression predicted tumor location with 81% accuracy; however, 6% of tumors were

Figure 4 Receiver Operator Characteristic Curves Were Plotted in the Cohort A Validation Set for the 7-Probe Genomic Location Score as a Predictor of Genomic Left-side Tumors (A), Genomic Right-side Tumors (B), Genomic Left-side Tumors Within the Anatomic Right-side Tumors (C), and Genomic Right-side Tumors Within the Anatomic Left-side Tumors (D)



Abbreviation: AUC = area under the curve.

consistently predicted to have the opposite side from anatomic location, and 13% of tumors had no clear predicted location.

The differences between left-side and right-side tumors are often attributed to the fact that the right colon is formed from the midgut and the left colon is formed from the hindgut in the embryo.[3] Indeed, 2 of the probes in the 7-probe genomic location score are HOX family genes, which are expressed differentially across the embryo.[19] However, the hindgut and midgut join at a nonspecific location in the transverse colon about one-third of the way from the splenic flexure.[2] This could account for genomic left tumors that are anatomically right, but not genomic right tumors that are anatomically left, which we discovered. We also discovered that genomic non-consensus tumors are almost equally distributed between the right and left colon. Apparently, colon embryology is a factor in tumor genomic location, but not the whole story.

Although the derivation of genomic location by network analysis may be cumbersome to reproduce in multiple cohorts, the genomic location score can be derived and applied in a straightforward manner. The genomic location score could be used in statistical models of treatment response in comparison with tumor location to assess the relative predictive significance of genomic location and anatomic location. A study of this nature, beyond the scope of this project, would be to test the relative significance of anatomic or genomic location to predict response to anti-EGFR therapy for patients with KRAS WT metastatic colorectal cancer.

Differences in the prevalence of CMSs between tumor sides were reflected in genomic location as well as anatomic location. For example, CMS1 tumors were more prevalent in anatomic right-side than left-side colon cancer.[4] In this study, tumors that were genomic right and anatomic left were more likely to be CMS1 than those that were genomic left and anatomic left ($P = 1.7 \times 10^{-15}$) (Figure 3). However, the fact that significant percentages of both right-side and left-side tumors (with respect to both anatomic and genomic locations) were CMS2 and CMS4, shows a degree of independence between location and CMS. Although a consensus molecular subtype is considered molecularly homogeneous to some degree, genomic location is one sign of heterogeneity within some CMS subtypes.

Here, we found that 40% of the samples without a consensus genomic location were CMS4, in comparison to 18% for those with a genomic location. That CMS4 tumors appear to lose genomic features specific to a location is worthy of further analysis in light of the "stem-like" traits reported for CMS4 tumors.[4] Combined, these results show that features of CMS subtypes may vary with genomic location, extending the study by Lee et al[2] on the integration of tumor location and CMSs.

One limitation of this study is the lack of information on tumor histology in the study population. For example, it has been reported that right-side tumors are more likely to arise from serrated polyps.[2] Conceivably, tumor histology could also contribute to genomic location.

Conclusion

In this study, we analyzed the degree to which the side of a colon tumor is recapitulated by tumor gene expression. We found that gene expression predicted tumor location with 81% accuracy; however, 19% of tumors had an ambiguous or discordant predicted location. Tumors with an ambiguous gene expression-predicted tumor location were significantly more likely to have a mesenchymal phenotype (CMS4) than tumors with a definitive predicted location. Tumors whose predicted location was discordant with anatomic location shared many features with their genomically predicted locations. This work suggests that, in clinical studies of disparities in treatment response, a predominately molecular phenomenon, tumors with an ambiguous genomic location or a genomic location discordant with anatomic location, may deserve separate analysis.

Clinical Practice Points

- Subset analysis of multiple clinical trials have shown that patients with KRAS WT metastatic colon cancer with left-side disease benefit from the addition of anti-EGFR therapy to chemotherapy, whereas those with right-side disease receive no significant benefit from this addition.
- Results herein suggest that genomic sidedness, as represented by the genomic location score, may be considered for stratifying patients by sensitivity to anti-EGFR therapy. Disparities in treatment response owing to tumor side are less clear in stage I to III disease; however, colon cancers of all stages show differences in distributions owing to tumor location for multiple clinically-relevant molecular features (eg, MSI, BRAF mutation, and CIMP statuses) and consensus molecular subtypes.
- In this study, the molecular features intrinsic to left-side or right-side colon cancer were analyzed using whole-genome transcription analysis. We discovered a set of left-side tumors (4%) with gene expression signature more characteristic of right-side tumors, and similarly for 4% of right-side colon cancers. Moreover, 17% of tumors have ambiguous location with respect to gene expression and tend towards a mesenchymal phenotype. We also defined a score based on expression of 7 genes that predicts so-called genomic location.
- This work suggests that, in clinical studies of disparities in treatment response, a predominately molecular phenomenon, tumors with an ambiguous genomic location or a genomic location discordant with anatomic location, may deserve separate analysis.

Disclosure

The authors have stated that they have no conflicts of interest.

Supplemental Data

Supplemental table accompanying this article can be found in the online version at https://doi.org/10.1016/j.clcc.2019.02.002.

References

1. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A* 1999; 96:8681-6.
2. Lee MS, Menter DG, Kopetz S. Right versus left colon cancer biology: integrating the consensus molecular subtypes. *J Natl Compr Canc Netw* 2017; 15:411-9.
3. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; 487:330-7.
4. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015; 21:1350-6.
5. Missiaglia E, Jacobs B, D'Ario G, et al. Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. *Ann Oncol* 2014; 25:1995-2001.
6. Petrelli F, Tomasello G, Borgonovo K, et al. Prognostic survival associated with left-sided vs right-sided colon cancer: a systematic review and meta-analysis. *JAMA Oncol* 2016; 3:211.

7. Bauer KM, Hummon AB, Buechler S. Right-side and left-side colon cancer follow different pathways to relapse. *Mol Carcinog* 2011; 51:411-21.
8. Glebov OK, Rodriguez LM, Nakahara K, et al. Distinguishing right from left colon by the pattern of gene expression. *Cancer Epidemiol Biomarkers Prev* 2003; 12:755-62.
9. Venook AP, Niedzwiecki D, Innocenti F, et al. Impact of primary (1o) tumor location on overall survival (OS) and progression-free survival (PFS) in patients (pts) with metastatic colorectal cancer (mCRC): analysis of CALGB/SWOG 80405 (Alliance). *J Clin Oncol* 2016; 34(15 Suppl):3504.
10. Demurtas L, Puzzoni M, Giampieri R, et al. The role of primary tumour sidedness, EGFR gene copy number and EGFR promoter methylation in RAS/BRAF wild-type colorectal cancer patients receiving irinotecan/cetuximab. *Br J Cancer* 2017; 117:315-21.
11. Ulivi P, Scarpi E, Chiadini E, et al. Right- vs. left-sided metastatic colorectal cancer: differences in tumor biology and bevacizumab efficacy. *Int J Mol Sci* 2017; 18:1240.
12. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostatistics* 2010; 11:242-53.
13. Marisa L, de Reyniès A, Duval A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* 2013; 10:e1001453.
14. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-0. Available at: https://CRAN.R-project.org/package=e1071. Accessed: July 28, 2018.
15. Therneau T, Atkinson B. rpart: recursive partitioning and regression trees. R package version 4.1-13. Available at: https://CRAN.R-project.org/package=rpart. Accessed: February 23, 2018.
16. Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002; 2:18-22.
17. Buechler SA, Gökmen-Polar Y, Badve SS. EarlyR: a robust gene expression signature for predicting outcomes of estrogen receptor-positive breast cancer. *Clin Breast Cancer* 2019; 19:17-26.e8.
18. Arnold D, Lueza B, Douillard J-Y, et al. Prognostic and predictive value of primary tumour side in patients with RAS wild-type metastatic colorectal cancer treated with chemotherapy and EGFR directed antibodies in six randomized trials. *Ann Oncol* 2017; 28:1713-29.
19. Zacchetti G, Duboule D, Zakany J. Hox gene function in vertebrate gut morphogenesis: the case of the caecum. *Development* 2007; 134:3967-73.