Translational Oncology

www.transonc.com

Volume 9 Number 4 August 2016 pp. 280–286 280

Colorectal Cancer with Residual Polyp of Origin: A Model of Malignant Transformation[1,2]

Brooke R. Druliner*,[3], Shahrooz Rashtak*,[3], Xiaoyang Ruan†, Taejeong Bae†, Nikolaos Vasmatzis¶, Daniel O'Brien†, Ruth Johnson‡, Donna Felmlee-Devine*, Jill Washechek-Aletto*, Nivedita Basu§, Hongfang Liu†, Thomas Smyrk#, Alexej Abyzov† and Lisa A. Boardman*

*Department of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, 55905; †Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, 55905; ‡Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, 55905; §Department of Internal Medicine, University of Michigan, Ann Arbor, MI, 48109; ¶Center for Individualized Medicine, Mayo Clinic, Rochester, MN, 55905; #Anatomic Pathology, Mayo Clinic, Rochester, MN, 55905

Abstract

The majority of colorectal cancers (CRCs) arise from adenomatous polyps. In this study, we sought to present the underrecognized CRC with the residual polyp of origin (CRC RPO +) as an entity to be utilized as a model to study colorectal carcinogenesis. We identified all subjects with biopsy-proven CRC RPO + that were evaluated over 10 years at Mayo Clinic, Rochester, MN, and compared their clinical and pathologic characteristics to CRC without remnant polyps (CRC RPO −). Overall survival and disease-free survival overlap with an equivalent hazard ratio between CRC RPO + and RPO − cases when age, stage, and grade are adjusted. The somatic genomic profile obtained by whole genome sequencing and the gene expression profiles by RNA-seq for CRC RPO + tumors were compared with that of age -and gender-matched CRC RPO − evaluated by The Cancer Genome Atlas. CRC RPO + cases were more commonly found with lower-grade, earlier-stage disease than CRC RPO −. However, within the same disease stage and grade, their clinical course is very similar to that of CRC RPO −. The mutation frequencies of commonly mutated genes in CRC are similar between CRC RPO + and RPO − cases. Likewise, gene expression patterns are indistinguishable between the RPO + and RPO − cases. We have confirmed that CRC RPO + is clinically and biologically similar to CRC RPO − and may be utilized as a model of the adenoma to carcinoma transition.

Translational Oncology (2016) 9, 280–286

Introduction

In the United States, colorectal cancer (CRC) is the second leading cause of cancer-related mortality [1]. The majority of CRCs are presumed to arise from adenomatous polyps, which are present in at least one third of individuals undergoing screening colonoscopy [2–5]. However, in the majority of CRC patients, no residual adenomatous polyp tissue is detectable in the specimen obtained at the time of surgical resection (CRC RPO –). It has not been possible, due to the ethical

Address all correspondence to Lisa Boardman, MD, Mayo Clinic College of Medicine, Gonda Building 9-411, 200 First Street SW, Rochester, MN, 55905.
E-mail: boardman.lisa@mayo.edu

[1] Disclosures: The authors declare no conflicts of interest.

[2] Funding: This work was supported by the National Cancer Institute at the National Institutes of Health (R01 CA170357) and the Clinical Core of the Mayo Clinic Center for Cell Signaling in Gastroenterology (P30DK084567). This publication was also made possible by CTSA grant number UL1 TR000135 from the National Center for Advancing Translational Sciences, a component of the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the National Institutes of Health.

[3] These authors have contributed equally.

Received 8 June 2016; Accepted 13 June 2016

1936-5233/16

http://dx.doi.org/10.1016/j.tranon.2016.06.002

concerns with leaving a premalignant lesion in a patient, to study the transition and progression of a polyp to cancer in humans, thus limiting researchers to the use of animal models, cell lines, or comparisons of polyps and cancers from different individuals. These studies have contributed valuable knowledge to the field of CRC, with the caveat that extensive extrapolation of the transformation from benign polyp to cancer within an individual was required. However, in a small portion of CRC cases, the residual polyp of origin (CRC RPO+) is still present within the resected lesion. These residual polyps, adjacent to the cancer, provide a window of opportunity to study the evolution of adenoma to carcinoma, thus expanding our understanding of malignant transformation [6]. It is not known whether there are features unique to CRC RPO+ that allow the polyp of origin to remain and not be completely consumed by the tumor or if the cancer component of the CRC RPO+ cases is similar to the CRC RPO− cancer.

In other words, in order for the CRC RPO+ model to be pertinent to the majority of CRCs in which the residual polyp of origin does not remain, it is crucial to determine if CRC RPO+ cases have the same characteristics—molecular, histological, or clinical—as CRC RPO− cases. In this study, we sought to determine if there were differences in the cancer tissue of CRC RPO+ and RPO− cases. If CRC RPO+ cases resemble CRC RPO− cases, then CRC RPO+ cases can serve as a valid model for future studies investigating the transition from the premalignant to malignant CRC state.

Material and Methods

Sample Characteristics and Annotation

Demographics and clinical outcome data were reviewed and abstracted from the electronic medical record for all patients with CRC who had surgical resection at Mayo Clinic, Rochester, MN, from 2004 to 2015 and who had given authorization for research. This study was approved by our institutional review board. Patients with biopsy-proven CRC were classified into CRC RPO+ and CRC RPO−. CRC RPO+ are those cases with residual polyp contiguous with their primary cancer present on pathologic review of histologic slides from the surgical specimen, and CRC RPO− are those cases in which no residual polyp tissue was found in association with the tumor.

Mismatch repair status (MMR) was determined by review of immunohistochemistry studies for MMR proteins. MMR deficiency (d-MMR) was defined as lack of staining for the MMR proteins MLH1, MSH2, MSH6, and/or PMS2. Proficient MMR was defined as presence of staining for all four of these proteins. The frequency of d-MMR and p-MMR was compared between these two groups.

Tissue Preparation and DNA Extraction

We utilized a total of 10 CRC RPO+ cases for molecular study. Tumor specimens were harvested following surgical resection and snap frozen in liquid nitrogen. Up to three 1-cm^2 full-thickness specimens from the center and edge of the cancer were collected. In addition to cancer tissues, three 1-cm^2 normal colonic epithelium full-thickness specimens at least 8 cm from the polyp/tumor margin were harvested. Regions were identified by an expert gastrointestinal pathologist (T.J.S.) on a hematoxylin and eosin (H&E)–stained 4-μm–thick section of the frozen tissue blocks collected from surgical specimens from consented participants of the Biobank for Gastrointestinal Health Research (IRB 622-00, PI L.A. Boardman). Tumor tissue was macrodissected to enrich for tumor density (>70%). All CRC RPO+ tissues used here excluded subjects with a prior history of any malignancy, a family history of Lynch syndrome or familial adenomatous polyposis, and any other syndrome associated with hereditary CRC or inflammatory bowel disease. DNA was manually extracted utilizing PureGene chemistry (Gentra Systems, Minneapolis, MN). RNA was extracted using Qiagen MiRNeasy mini kit.

Whole Genome Sequencing of 10 CRC RPO+ Cases

For library construction, total DNA was quantified in triplicate using the Quant-iT PicoGreen DNA Assay Kit and normalized to 2-ng/μl minimum concentration. An aliquot of 100 ng for each sample was transferred into library preparation utilizing the Broad Institute–developed one-well protocol. All biochemistry occurs in a single well without the need for sample transfer (the sample is reversibly immobilized to and released from magnetic beads, allowing washes and reagent addition). The one-well protocol streamlines the process and greatly reduces sample input requirements. The product provides one library (typical median insert size of library is 330 bp) [7].

Samples were sequenced on the Illumina HiSeq X instruments producing 150-bp, paired-end reads to meet a goal of 30× mean coverage. Using the Picard Informatics Pipeline, all data from a particular sample were aggregated into a single Binary Alignment/Map format (BAM) file which included all reads, all bases from all reads, and original/vendor-assigned quality scores. A pooled Variant Call Format file using the latest version of Picard GATK software was generated and provided for each sample batch. All whole genome sequencing data analyzed in this manuscript will be uploaded to SRA through dbGaP (Accession numbers to follow).

Mutation Frequency Detection. In order to detect somatic single nucleotide variants (SNVs) between the tumor and matched normal tissue SNVs for 10 cases of CRC RPO+, we used 4 different somatic variant callers: MuTect, SomaticSniper, Strelka, and VarScan [8–11]. Those callers were run with default options for normal and tumor samples from each patient. We took common SNVs detected by at least two different callers. Variant allele frequencies for those SNVs were calculated from sample BAM files for each patient using an in-house script. To annotate them, we used Variant Effect Predictor (http://www.ensembl.org/Tools/VEP).

Tumor somatic mutation profiles for CRC RPO− were obtained from The Cancer Genome Atlas (TCGA) for 32 pMMR CRCs which were stage, site, gender, and age matched to the 10 pMMR CRC RPO+ cases described above. The somatic mutation profiles were downloaded in Mutation Annotation Format, which classifies somatic mutations into 1 of 13 categories depending on the type and sequence position of the corresponding mutation. SNVs that caused frame shift in/del, in frame in/del, missense mutation, or nonsense mutation, or involved a splice site were classified as being likely to impact a gene's function. A gene was considered as mutated when it had at least one somatic mutation in at least one of these categories [12]. The frequency of somatic mutation rates from the WGS studies on the cases of CRC RPO+ were compared with the mutation profiles reported for the matched TCGA cases presumed to be mainly CRC RPO−.

RNA Sequencing of 16 CRC RPO+ Cases

Total RNA was quantified using the Quant-iT RiboGreen RNA Assay Kit and normalized to 5 ng/μl. An aliquot of 200 ng for each sample was transferred into library preparation which was an automated variant of the Illumina TruSeq Stranded mRNA Sample Preparation Kit. This method preserves strand orientation of the RNA transcript. Oligo dT beads were used to select mRNA from the total RNA sample.

Heat fragmentation and cDNA synthesis from the RNA template then followed. The resultant cDNA went through library preparation (end repair, base "A" addition, adapter ligation, and enrichment) using Broad-designed indexed adapters substituted in for multiplexing. After enrichment, the libraries were quantified with quantitative polymerase chain reaction using the KAPA Library Quantification Kit for Illumina Sequencing Platforms and then pooled equimolarly. The entire process is in 96-well format, and all pipetting is done by either Agilent Bravo or Hamilton Starlet.

Pooled libraries were normalized to 2 nM and denatured using 0.1 N NaOH prior to sequencing. Flowcell cluster amplification and sequencing were performed according to the manufacturer's protocols using either the HiSeq 2000 or HiSeq 2500. Each run was a 101-bp paired-end with an 8-base index barcode read. Data were analyzed using the Broad Picard Pipeline which includes demultiplexing and data aggregation.

Processing of RNA-Seq Data and Comparison to TCGA CRC

The paired-end RNASeq FASTQ files were then analyzed using Mayo Clinic's standard RNA-Seq application, MAPR-Seq v.2.0.0 (http://bioinformaticstools.mayo.edu/research/maprseq/). MAPR-RSeq integrates a suite of open-source bioinformatics tools along with in-house–developed methods to analyze paired-end RNA-Seq data. Read alignment was performed with Tophat [13] which uses Bowtie [14]—a fast, memory-efficient, short-sequence aligner. The reads were aligned to the transcriptome (Ensembl GTF) and to the genome (hg19) to report both existing and novel expressed regions. The BAM file produced by Tophat was processed using featureCounts [15] to summarize expression at the gene and exon levels. Reads per kilo base per million (RPKM) values were calculated from the raw gene counts produced by featureCounts and by incorporating the total number of aligned reads and the coding length of each gene. To identify possible quality control issues, RSeQC software [16] was used to detect abnormalities, such as unsymmetrical gene body coverage, high levels of read duplication, and low saturation levels of known exon junctions, within each sample.

The TCGA colorectal adenocarcinoma and rectal adenocarcinoma RNASeq expression data were obtained from the following site, https://tcga-data.nci.nih.gov/docs/publications/coadread_2012/, and their annotation file was obtained from https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/GAF/GAF_bundle/outputs/TCGA.Sept2010.09202010.gaf. RPKM values are directly impacted by gene length, so only the genes with a similar length between the TCGA annotation file and the Mayo Clinic annotation file (ftp://ftp.ensembl.org/pub/grch37/update/gtf/homo_sapiens/Homo_sapiens.GRCh37.82.gtf.gz) were used within this analysis. Specifically, genes that had a gene length within 10% of each other from both references were kept. For these similar genes identified, the RPKM values were extracted from the TCGA analyzed expression data and from the Mayo analyzed expression data. Note, the bioinformatics tools used to align and calculate expression from the TCGA analyzed samples were different from those used for the Mayo analyzed samples. All the genes that had no expression from the TCGA analyzed or the Mayo analyzed samples were removed. Mean and standard deviation values were calculated from the RPKM values across the TCGA analyzed group and for the Mayo analyzed group. The genes that had a standard deviation greater than the mean were also removed from this analysis to avoid evaluating highly variable genes, such as circadian rhythm genes. The sample specific RPKM values from the Mayo analyzed samples and the TCGA colorectal adenocarcinoma and rectal adenocarcinoma samples with similar clinical characteristics were then extracted and visualized within a heatmap. The heatmap was constructed using the Complex Heatmap function in R through the Bioconductor package [17]. All RNA-seq data analyzed in this manuscript will be uploaded to SRA through dbGaP (accession numbers to follow).

Statistical Analyses

Student's t test was used to determine statistical significance of continuous variables, and χ^2 test was used for categorical variables. Univariate and multivariate analyses were conducted to adjust for potential confounders. Overall survival (OS) and disease-free survival (DFS) were defined as the time in months from the diagnosis to death from any cause and time from the date of diagnosis to the date of recurrence or death, respectively. Survival outcomes were reported as the hazard ratio (HR) calculated in the Cox regression model and presented by Kaplan-Meier curves comparing the observed and predicted events after adjustment for sex, stage, and grade in the Cox model. A P value less than .05 was considered to be significant. Correlations were determined using Spearman's rank correlation. All the statistical analyses were performed using the SAS JMP or R software.

Results

CRC with residual polyp of origin as a model of malignant transformation

In this study, we sought to examine possible biological and clinical differences between colorectal tumors with and without the residual polyp of origin present. Because CRC with the polyp of origin remaining (CRC RPO+) is an underrecognized event, it was our goal to determine if these cases are representative of the majority of CRCs in which there is no residual polyp present (CRC RPO–) at the time of diagnosis or surgery and therefore may be a sound model for the study of colorectal carcinogenesis (Figure 1*A*). To determine whether this model is truly representative of CRC irrespective of the presence of the precursor lesion, we have studied CRC RPO+ tumors in terms of clinical course, histologic features, and genetic profiles and compared them with CRC RPO– tumors.

Baseline demographics and histological characteristics of CRC RPO+ and RPO–

From those cases identified as having undergone surgery at Mayo Clinic Rochester between 2004 till 2015, a total of 543 patients with CRC RPO+ and 4101 patients with CRC RPO– were included in the study. We found that baseline demographics including age and sex distribution were similar in both groups (Table 1). Among those RPO+ cases, 332 (61%) of the cases were found to have tubulovillous lesions, 29 (5%) villous, and 182 (34%) adenomatous not otherwise specified.

CRC RPO+ cases were more likely to be low-grade (moderately or well-differentiated) lesions (56.6% low vs 43.4% high grade) compared with RPO– lesions which were more likely to be high grade (poorly or undifferentiated) tumors (32.2% low vs 67.8% low grade), $P < .0001$ (Table 1). Immunohistochemistry studies for MMR proteins on 70 specimens of CRC RPO+ were compared with 621 CRC RPO– cases. Fifteen (21.4%) of the CRC RPO+ group and 125 (20.1%) of the RPO– group had deficient mismatch repair with absence of expression of MLH1, MSH2, MSH6, and/or PMS2, $P = .79$. Upon pathologist review of hemotoxylin and eosin–stained sections of CRC RPO+ and CRC RPO– cases, there were no qualitative distinctions overall between the tumors (Figure 1*B*).

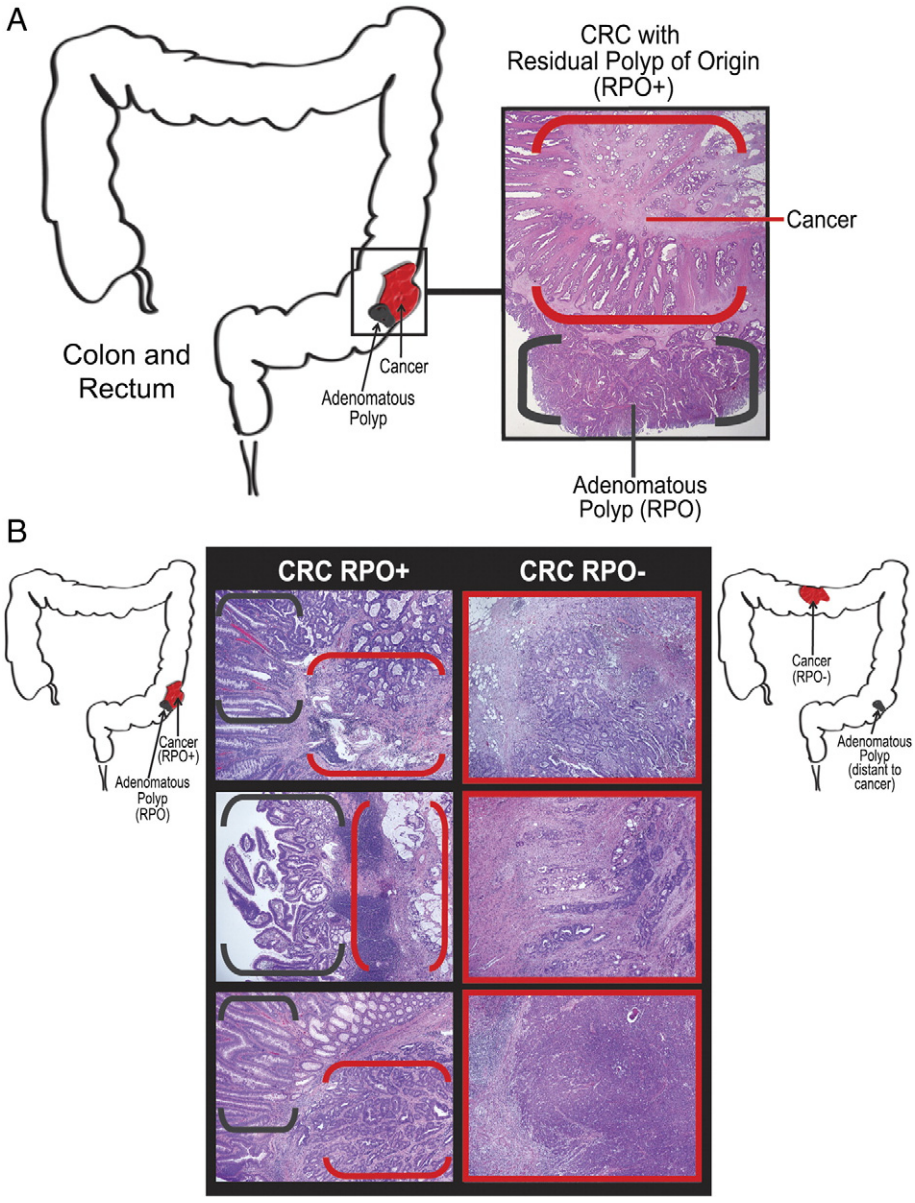

Figure 1. A model of malignant transformation: CRC with residual polyp of origin. (A) CRC with the residual polyp of origin (CRC RPO +) is represented schematically here and by H&E straining. The anatomical location in the colon of the polyp and cancer in the diagram serves only as an exemplar case as tumor location has no impact on the likelihood of finding a CRC RPO + case. The H&E stain is a cross-section of the residual polyp of origin adjacent to the cancer tissue. (B) H&E tissue sections from three CRC RPO + cases (left) and five CRC RPO − cases (right). For the CRC RPO+ cases, the cancer tissue is shown in red brackets, and the polyp tissue is shown in black brackets. All cases are poorly differentiated tumors, and the cases are also matched by stage. An additional schematic is shown to illustrate a typical CRC RPO− case, where the cancer is found at a distant location to the presence of the polyp.

Clinical features and survival outcomes of CRC RPO+ and RPO–

The average tumor size reported by Surveillance, Epidemiology, and End Results statistics was significantly larger (47.3 mm) in CRC RPO – cases compared with that in RPO + cases (35.3 mm), $P < .0001$. There was no significant difference in the distribution of the anatomical location of the tumor between the two groups. The majority (65%) of CRC RPO + cases were stage 1 compared with 20.6% of the CRC RPO – cases, $P < .0001$. CRC RPO – cases were diagnosed at a higher stage than RPO + cases (25.0% CRC RPO – vs 12.3% CRC RPO + for stage II, 28.7% vs 13.6% for stage III, and 25.6% vs 9.0% for stage IV cancers, respectively, $P < .0001$ across all stages).

Over 11 years of follow-up, there were a total of 1528 (33%) deaths within the 2 groups combined, out of which 116 (21%) events occurred among CRC RPO + patients and 1412 (38%) among CRC RPO – patients. CRC RPO + cases have similar OS and DFS performance as CRC RPO – cases after adjustment for age, sex, stage, and grade (OS HR = 1.18, 95% confidence interval [CI] 0.96-1.44, $P = .11$; DFS HR = 1.12, 95% CI 0.91-1.38, $P = .27$). The predicted OS (Figure 2*A*) and DFS (Figure 2*B*) curves after adjusting for sex, stage, and grade overlapped between the two groups.

Genomic analysis of CRC RPO+ and RPO–

The Cancer Genome Atlas Network published a report in 2012 that identified 17 somatic recurrently mutated genes in nonhypermutated (pMMR) CRC [12]. Among those 17 genes and in the nonhypermutated cancers, the 8 most frequently mutated genes were *APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, TCF7L2,* and *NRAS*. We compared the

Table 1. Sample Characteristics.

Parameter	CRC RPO+ (%)	CRC RPO− (%)	*P* Value
Number of Patients	543 (11.7)	4101 (88.3)	
Age			.82
Mean (SD)	64.6	65.7	
Median	66	67	
SD	13.6	14.3	
Range	(16-98)	(19-97)	
Gender			.15
Female	216 (39.8)	1762 (43)	
Male	327 (60.2)	2339 (57)	
Tumor Location			.17
Right colon	206 (37.9)	1610 (39.3)	
Left colon	129 (23.8)	1033 (25.2)	
Rectosigmoid	198 (36.5)	1338 (32.6)	
Not specified	10 (1.8)	120 (2.9)	
Tumor Stage			<.0001
I	353 (65.0)	845 (20.6)	
II	67 (12.3)	1026 (25.0)	
III	74 (13.6)	1177 (28.7)	
IV	49 (9.0)	1053 (25.7)	
Number of examined lymph nodes			.0001
Mean (SD)	18.2	21.0	
Median	16	19	
SD	0.7	0.2	
Range	(0-90)	(0-90)	
Number of positive lymph nodes			<.0001
Mean (SD)	0.7	1.8	
Median	0	0	
SD	0.17	0.06	
Range	(0-19)	(0-48)	
Grade			<.0001
Well differentiated	25 (4.6)	108 (2.6)	
Moderate differentiated	262 (48.2)	1136 (27.7)	
Poor differentiated	218 (40.2)	237 (57.7)	
Undifferentiated	2 (0.4)	259 (6.3)	
Unknown	36 (6.6)	233 (5.7)	
Tumor Size (EOD)			<.0001
Mean (SD)	35.3	47.2	
Median	16.25	42	
SD	1.6	0.5	
Range	(1-198)	(1-600)	
MMR Status			.79
pMMR	55 (78.6)	496 (79.9)	
dMMR	15 (20.3)	125 (20.1)	

somatic mutation frequency of these genes between 10 cases of pMMR CRC RPO+ and 33 tumor samples matched for age, gender, and stage downloaded from the TCGA. The most commonly mutated genes in our sample set of CRC RPO+ were *APC, TP53, and KRAS,* which were mutated in 6 of the 10 cases (60%). When compared with the TCGA data, the expected frequency of mutated genes in our CRC RPO+ cases mirrored the mutation profile of those TCGA CRC cases that likely did not have RPO (Figure 3*A*). The mutations included missense and inactivating variants. The TCGA reported that *ACVR2A, APC*, TGFBR2, *MSH3, MSH6, SLC9A9,* and *TCF7L2* were frequently mutated in the hypermutated tumors, so we obtained sequencing data on an additional two hypermutated (dMMR) CRC RPO+ cases. When we compared the mutation frequencies of genes identified by the TCGA as commonly mutated in hypermutated CRC tumors, we found that the majority of those genes were also mutated in our two CRC RPO+ dMMR tumors.

Gene expression analysis of CRC RPO+ and RPO−

We also compared gene expression patterns between CRC RPO+ cases that were subjected to RNA-seq with data from TCGA CRC tumors, matched by site, stage, gender, and age [12]. Expression values (RPKM) were plotted in a heatmap for 16 CRC RPO+ cases and 32 tumor samples from the TCGA (Figure 3*B*). Overall, there are virtually no differences between the CRC RPO+ and TCGA groups ($r = 0.88$, $P < 2.2e\text{-}16$). TCGA samples cluster with the RPO+ samples, and there are minor clusters even within a major branch of the TCGA group. It is important to note that any differences seen may be attributed to differences in sample prep, sequencing technology, and bioinformatics between the CRC RPO+ and TCGA cases.

Discussion

Adenomatous polyps are the presumed precursor lesion of CRC, and reports on the decline in CRC incidence from large population-based epidemiologic studies being attributable to the increased frequency of colonoscopy and subsequent polypectomy support that polyps are likely to be the precursors to CRC [18,19]. The multistep model of colorectal carcinogenesis has been identified by combining molecular data on polyps and CRCs from different individuals. This is due to the fact that the polyp from which the cancer was presumably derived often does not

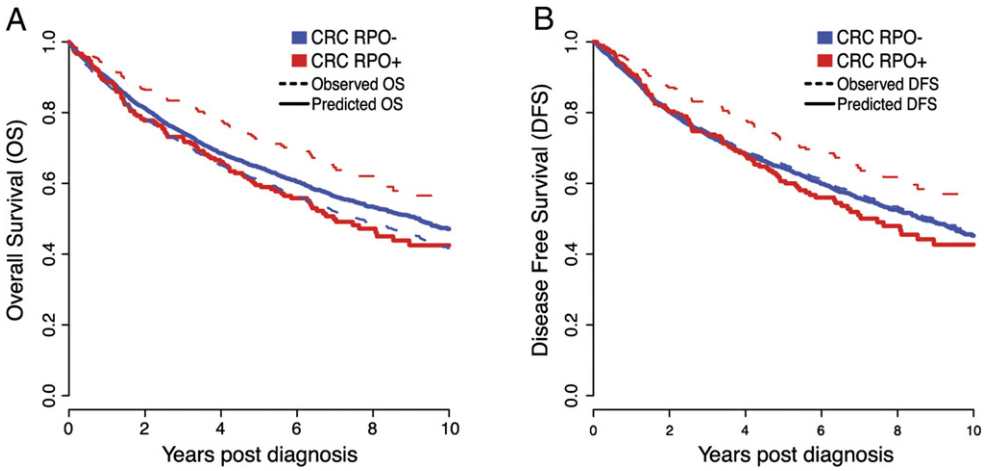

Figure 2. Survival analysis between CRC RPO + and RPO − tumors. Kaplan-Meier curves comparing (A) OS and (B) DFS for CRC RPO + and CRC RPO −. Solid lines are representative of unadjusted observations, and dashed lines are representative of predicted outcomes after adjustment for disease stage and grade.

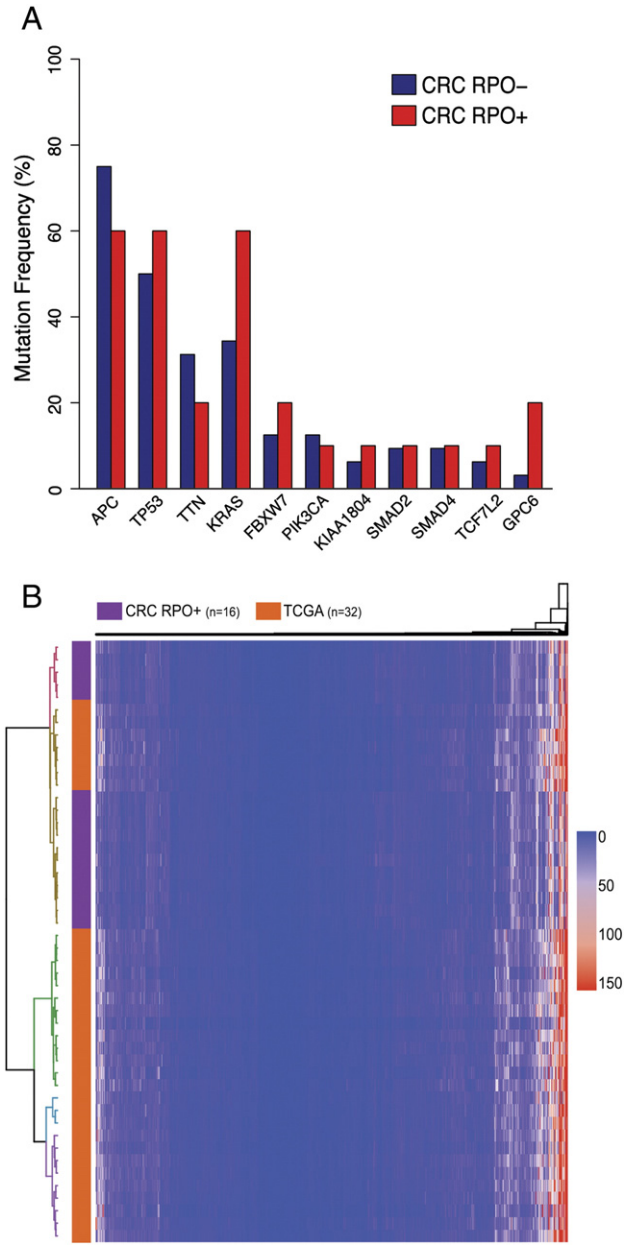

Figure 3. Mutation frequency and gene expression signatures are comparable between CRC RPO + and RPO − tumors. (A) Somatic mutation frequency of 11 genes found to be commonly mutated in CRC. We compared the mutation frequencies of these genes from 10 CRC RPO + cases and 32 CRC RPO − cases from the TCGA. TCGA samples were matched by age and stage to our 10 CRC RPO + cases. These mutation frequencies were not significantly different between CRC RPO + and RPO − cases. (B) CRC RPO + and TCGA CRC tumors are virtually indistinguishable from each other on the basis of their gene expression patterns. Plotted in this heatmap are the RPKM values for expression of genes following RNA-seq. Hierarchical clustering with $k = 5$ was performed for genes and both CRC RPO + and TCGA tumor cases. The left axis is the clustering for the CRC RPO + and TCGA tumors, and the top axis is the clustering for the genes. For the case clustering, each cluster is colored for clarity. Red represents genes with highest expression, and blue is the lowest.

remain at the time of detection, or mention of the polyp component from a surgical cancer specimen may be absent in the pathology report because the cancer diagnosis has the more significant impact on the clinical care for the patient. We propose CRC RPO + as a model in which to study malignant transformation, but the features of the cancer with polyp remaining have not been well characterized. Through our examination of clinical behavior, histologic, genetic, and transcriptional features, we have shown the CRC RPO + cancers are indistinguishable from CRC RPO −. Based on these results, studies that examine the adenoma to carcinoma transition can directly compare CRC tumor tissues, regardless of RPO status.

We have found that CRC RPO + cases are more likely to be smaller in size and earlier stage than their RPO − counterparts. It is important to note that the stage-adjusted survival outcomes of these patients are almost identical to those with CRC RPO −. This observation can be explained by the fact that, over time, as the tumor grows and metastasizes to regional and distant locations, it will outgrow and consume the polyp of the origin, making it less likely to be detected when diagnosed at the higher stage. This theory is supported by the fact that RPO − tumors tend to be larger in size when compared with RPO + ones.

We discovered that CRCs RPO − more commonly present as poorly differentiated tumors. This observation suggests that high-grade tumors grow faster and are more likely to obliterate the precursor lesion before resection. The other possibility would be that CRCs RPO − follow a different pathway in the process of cancer development that does not require origination from a premalignant neoplasm. Although CRCs RPO − tend to present as a high-grade lesion, there is no tendency for deficient MMR status in either group.

CRC RPO + resembles CRC in general regardless of RPO status in terms of somatic mutation frequency rate of genes reported as commonly altered in colorectal carcinogenesis. Given the relatively small sample size in our cohort of 10 CRC RPO + patients who underwent whole genome sequencing, our study could be underpowered to pick up small differences in the frequency of the mutated genes between the two groups, but given the similarity between the groups, this is not likely. It is feasible that some of the TCGA CRC cases were RPO +, but the odds that they represented the majority of these subjects are quite low, as frequency of these types of cases is approximately 10% as shown in this study. It is also possible that the frequency of RPO + lesions is underestimated secondary to the histologic undercalling, as pathologists may not report coexistence of the premalignant polyp with cancer. Although less likely, the frequency of the polyp as the lesion of origin for CRC potentially can be also overestimated if the polyp and the cancer both arise independently from the same anatomical location.

CRC RPO + is an entity that has the potential to be utilized as a unique model representative of the mutagenic path of malignant transformation. This model is reminiscent of dysplastic nevi in a malignant melanoma lesion, which has been presented as an ideal model for studying nevus transformation to nevus-derived melanomas [20,21]. Through our careful collection of CRC RPO + cases, we have shown that they are indistinguishable from CRC RPO −, and based on that result, we conclude that CRC regardless of RPO status can be compared directly in future studies.

Conclusions

Overall, in this study, we have shown that cancer tissues of CRC RPO + cases are clinically and biologically similar to the CRC RPO − cases despite the tendency to be more commonly detected at an earlier disease stage and lower grade. This model provides the opportunity for future studies to discover the underlying mechanisms involved in the carcinogenesis through interrogation of genetic, transcriptional, and epigenetic events in this process.

This study also confirms for existing studies that the cancer component of CRC RPO+ cases is not a unique malignancy and is comparable to CRC in general. Because CRC RPO+ recapitulates the clinical and biological profile of CRC RPO−, we propose that CRC RPO+ will provide the opportunity to advance the understanding of neoplastic transformation mechanisms and will result in better recognition of high-risk precursor lesions leading ultimately to developing preventative measures and targeted therapies.

References

[1] Surveillance, E, and End Results (SEER) Program (www.seer.cancer.gov) research data (1973-2013). (National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission).

[2] Citarda F, Tomaselli G, Capocaccia R, Barcherini S, Crespi M, and Italian Multicentre Study, G (2001). Efficacy in standard clinical practice of colonoscopic polypectomy in reducing colorectal cancer incidence. *Gut* **48**, 812–815.

[3] Fearon ER and Vogelstein B (1990). A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767.

[4] Heitman SJ, Ronksley PE, Hilsden RJ, Manns BJ, Rostom A, and Hemmelgarn BR (2009). Prevalence of adenomas and colorectal cancer in average risk individuals: a systematic review and meta-analysis. *Clin Gastroenterol Hepatol* **7**, 1272–1278. http://dx.doi.org/10.1016/j.cgh.2009.05.032.

[5] Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, and Ederer F (1993). Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N Engl J Med* **328**, 1365–1371. http://dx.doi.org/10.1056/NEJM199305133281901.

[6] Kim TM, An CH, Rhee JK, Jung SH, Lee SH, Baek IP, Kim MS, Lee SH, and Chung YJ (2015). Clonal origins and parallel evolution of regionally synchronous colorectal adenoma and carcinoma. *Oncotarget* **6**, 27725–27735. http://dx.doi.org/10.18632/oncotarget.4834.

[7] Fisher S, et al (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* **12**, R1. http://dx.doi.org/10.1186/gb-2011-12-1-r1.

[8] Cibulskis K, et al (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213–219. http://dx.doi.org/10.1038/nbt.2514.

[9] Koboldt DC, et al (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568–576. http://dx.doi.org/10.1101/gr.129684.111.

[10] Larson DE, et al (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317. http://dx.doi.org/10.1093/bioinformatics/btr665.

[11] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, and Cheetham RK (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817. http://dx.doi.org/10.1093/bioinformatics/bts271.

[12] Cancer Genome Atlas N (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337. http://dx.doi.org/10.1038/nature11252.

[13] Trapnell C, Pachter L, and Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111. http://dx.doi.org/10.1093/bioinformatics/btp120.

[14] Langmead B, Trapnell C, Pop M, and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25.

[15] Liao Y, Smyth GK, and Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930. http://dx.doi.org/10.1093/bioinformatics/btt656.

[16] Wang L, Wang S, and Li W (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185. http://dx.doi.org/10.1093/bioinformatics/bts356.

[17] Team R. C. (2015). R: a language and environment for statistical computing. https://www.R-project.org/; 2015.

[18] Corley DA, Levin TR, and Doubeni CA (2014). Adenoma detection rate and risk of colorectal cancer and death. *N Engl J Med* **370**, 2541. http://dx.doi.org/10.1056/NEJMc1405329.

[19] Loberg M, Kalager M, Holme O, Hoff G, Adami HO, and Bretthauer M (2014). Long-term colorectal-cancer mortality after adenoma removal. *N Engl J Med* **371**, 799–807. http://dx.doi.org/10.1056/NEJMoa1315870.

[20] Duffy K and Grossman D (2012). The dysplastic nevus: from historical perspective to management in the modern era: part II. Molecular aspects and clinical management. *J Am Acad Dermatol* **67**, 19 e11-12. http://dx.doi.org/10.1016/j.jaad.2012.03.013 [quiz 31–12].

[21] Duffy K and Grossman D (2012). The dysplastic nevus: from historical perspective to management in the modern era: part I. Historical, histologic, and clinical aspects. *J Am Acad Dermatol* **67**, 1 e1-16. http://dx.doi.org/10.1016/j.jaad.2012.02.047 [quiz 17–18].