Novel recurrently mutated genes in African American colon cancers

Kishore Guda[a,b,c], Martina L. Veigl[b,c,1], Vinay Varadan[a,b,1], Arman Nosrati[d], Lakshmeswari Ravi[d], James Lutterbaugh[d], Lydia Beard[d], James K. V. Willson[e], W. David Sedwick[b,c,d], Zhenghe John Wang[b,f], Neil Molyneaux[f], Alexander Miron[f], Mark D. Adams[g], Robert C. Elston[b,h], Sanford D. Markowitz[b,c,d,i,2,3], and Joseph E. Willis[b,c,i,j,2]

[c]Department of Medicine, [f]Department of Genetics and Genome Sciences, [h]Department of Epidemiology and Biostatistics, [j]Department of Pathology, [a]Division of General Medical Sciences-Oncology, [d]Division of Hematology and Oncology, [b]Case Comprehensive Cancer Center, and [i]Case Medical Center, Case Western Reserve University, Cleveland, OH 44106; [e]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX 75390; and [g]J. Craig Venter Institute, La Jolla, CA 92037

Edited by Bert Vogelstein, The Johns Hopkins University, Baltimore, MD, and approved December 22, 2014 (received for review September 17, 2014)

We used whole-exome and targeted sequencing to characterize somatic mutations in 103 colorectal cancers (CRC) from African Americans, identifying 20 new genes as significantly mutated in CRC. Resequencing 129 Caucasian derived CRCs confirmed a 15-gene set as a preferential target for mutations in African American CRCs. Two predominant genes, *ephrin type A receptor 6* (*EPHA6*) and *folliculin* (*FLCN*), with mutations exclusive to African American CRCs, are by genetic and biological criteria highly likely African American CRC driver genes. These previously unsuspected differences in the mutational landscapes of CRCs arising among individuals of different ethnicities have potential to impact on broader disparities in cancer behaviors.

colon cancer | African American | next-generation sequencing | mutation | Caucasian

Colorectal cancer (CRC) is a leading cause of cancer mortality world-wide. CRC incidence and mortality rates are both increased in African Americans (AA) compared with Caucasians Americans (1–3). Although several factors likely play a role, the contribution of potential differences in tumor genetics to this disparity have yet to be fully explored (1, 3). In particular, AA CRCs were notably underrepresented in the four major published CRC sequencing studies (4–7), accounting for only two annotated AA cases of the 333 total CRCs studied (4–7). Accordingly, we initiated this study to compare the mutational landscapes of CRCs from AA individuals versus Caucasians.

Results

Using whole-exome DNA sequencing, we examined a discovery set of 31 microsatellite stable (MSS) late-stage AA CRCs (Dataset S1, Tables S1 and S2). Somatic mutations were detected using two variant calling algorithms for the point mutations and indels (8–10). One sample, 3213, identified as hypermutated, was excluded from subsequent analyses (Fig. S1 and Dataset S1, Table S1). In a second sample, 11481, library preparation from the normal tissue failed. Among the 29 informative AA CRCs, we found 2,696 protein-altering mutations in 2,156 genes (Dataset S1, Table S3). As expected for CRCs (6, 11), the mutational spectra of these 29 AA CRCs revealed C > T transitions as the predominant base substitution (Fig. S2 and Dataset S1, Table S4).

We next asked whether this discovery cohort of AA CRCs identified recurrent somatic mutations in any new genes not previously identified in CRCs (4, 6, 7). In the discovery AA CRCs, 385 genes demonstrated nonsilent mutations in at least two cancers. For 78 of these genes, we found that no protein-altering mutations had been previously reported among MSS nonhypermutator CRCs characterized in prior large-scale sequencing studies of CRCs (4, 6, 7) (Dataset S1, Table S5). We designed an Agilent SureSelect^{XT} custom-capture bait library for further resequencing of the coding and splice-junction regions of this 78-gene set (Dataset S1, Table S6). We first used this custom-capture kit to resequence the 29 paired AA CRC tumor and normal discovery samples, adding manual review of all mutation calls. Following this resequencing, 52 genes remained as demonstrating mutations in at least two individual discovery set cancers (Dataset S1, Table S7).

We next resequenced these 52 candidate mutation target genes in a second cohort of 77 MSS AA CRCs that were predominantly early stage (Dataset S1, Table S8). Three of the 77 cancers were found to be hypermutated and were excluded from subsequent analyses (Fig. S1 and Dataset S1, Table S8). Among the 74 informative AA CRC validation cases, a total of 59 nonsilent mutations were detected in 27 genes Dataset S1, Table S9). Using the statistical method detailed in our prior CRC sequencing study (6), we compared the observed to expected mutation rates (Dataset S1, Table S4) for each of the 52 candidate target genes. Twenty genes were identified as being significantly mutated [$P < 0.05$, false-discovery rate (FDR) < 0.05] in the discovery AA CRC cohort and as again being significantly mutated in the AA CRC validation cohort (Table 1 and Dataset S1, Table S10). The finding of these 20 genes as reproducibly significantly mutated above background suggests that they legitimately represent new genes targeted for recurrent mutation in colon cancer and that have been newly identified in AA CRCs.

Significance

Colorectal cancer is a leading cause of cancer-related deaths world-wide. African Americans exhibit the highest colon cancer incidence and mortality among all ethnic groups in the United States. Despite this finding, there is a dearth of knowledge on the genetic mechanisms underlying colon carcinogenesis in African Americans. We thus initiated this study to characterize the mutational landscapes of African American colon cancers. We identified new genes that are significantly mutated in colon cancer and that are highly preferentially targeted for mutations in colon cancers arising in African Americans as compared with Caucasians. These findings suggest differences in routes of colon carcinogenesis between the different ethnic groups and also may have implications for the ethnicity associated differences in tumor incidence and outcome.

Author contributions: K.G., S.D.M., and J.E.W. designed research; K.G., M.L.V., L.R., J.L., and L.B. performed research; K.G., V.V., A.N., J.K.V.W., N.M., A.M., M.D.A., and J.E.W. contributed new reagents/analytic tools; K.G., W.D.S., Z.J.W., R.C.E., and S.D.M. analyzed data; and K.G. and S.D.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]M.L.V. and V.V. contributed equally to this work.

[2]S.D.M. and J.E.W. contributed equally to this work.

[3]To whom correspondence should be addressed. Email: sxm10@cwru.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1417064112/-/DCSupplemental.

Table 1. Mutation rates and significance estimates of 20 candidate target genes in AA CRCs

	AA CRC discovery cohort (n = 29)*			AA CRC validation cohort (n = 74)		
Gene	Mutation rate per Mb	P value	FDR	Mutation rate per Mb	P value	FDR
TCEB3CL	81.62	0.000	0.001	47.98	0.000	0.000
MAGEB10	43.83	0.000	0.001	17.18	0.001	0.002
CPT1C	14.89	0.000	0.001	5.83	0.003	0.007
HTR1F	21.82	0.000	0.001	4.28	0.030	0.045
ANKRD36	8.34	0.000	0.001	2.18	0.027	0.045
MGAT4C	26.14	0.000	0.001	5.12	0.033	0.045
ZNF717	19.30	0.000	0.001	3.78	0.040	0.052
EPHA6	11.42	0.001	0.002	8.95	0.000	0.000
ZNF862	9.64	0.001	0.002	3.78	0.006	0.014
CP	10.44	0.001	0.002	4.09	0.008	0.016
KIAA1551	10.35	0.001	0.002	8.11	0.000	0.000
EML6	5.90	0.001	0.002	5.78	0.000	0.000
ATP8B2	7.67	0.001	0.002	3.01	0.030	0.045
JAK1	9.46	0.002	0.003	3.71	0.007	0.015
CHD5	7.46	0.007	0.009	7.30	0.000	0.000
WASH1	51.27	0.009	0.012	20.09	0.023	0.041
GPR149	14.34	0.013	0.016	16.86	0.000	0.000
CDK8	11.95	0.013	0.016	4.68	0.033	0.045
FLCN	7.25	0.026	0.028	5.68	0.000	0.001
WDR87	3.69	0.027	0.029	7.22	0.000	0.000

*Data from the unpaired AA tumor sample was excluded from the analysis.

Because these 20 genes have not been detected as mutated in previous studies of MSS nonhypermutator CRCs (4, 6, 7), we hypothesized that this gene set might be preferentially targeted for mutations in AA compared with Caucasian CRCs. To test this hypothesis, we resequenced these 20 genes in 129 predominantly late-stage MSS Caucasian CRCs (Dataset S1, Tables S11 and S12). Taken as a group, these 20 genes demonstrated a statistically significant ~twofold increase in the total number of mutations/tumor in AA CRC (83 mutations in 103 cases) compared with Caucasian CRCs (50 mutations in 129 cases; $P < 0.001$). This effect was driven by the 3.3-fold increased mutations in AA CRCs for the top 15 genes (nominal $P = 1.8 \times 10^{-8}$ from subset analysis) (Table 2 and Fig. S3). Taken individually, mutations in *ephrin type A receptor 6* (*EPHA6*) were exclusively detected in AA cases, with mutations in 5.8% of AA CRC versus 0% of Caucasian CRCs, and were significantly associated with

Table 2. Comparison of mutational frequencies of 20 candidate target genes in AA vs. Caucasian CRCs

			AA CRC discovery + validation Cohorts (n = 103)		Caucasian CRCs (n = 129)		P value (AA CRC discovery + validation cohorts vs. Caucasian CRCs)
Gene	Mutation count (AA CRC discovery cohort, n = 29)*	Mutation count (AA CRC validation cohort, n = 74)	Mutation count	Mutation frequency	Mutation count	Mutation frequency	
EPHA6	2	4	6	5.83%	0	0.00%	0.007
FLCN	1	2	3	2.91%	0	0.00%	0.086
HTR1F	2	1	3	2.91%	0	0.00%	0.086
GPR149	1	3	4	3.88%	1	0.78%	0.123
ZNF862	2	2	4	3.88%	1	0.78%	0.123
ANKRD36	3	2	5	4.85%	2	1.55%	0.142
KIAA1551	2	3	5	4.85%	2	1.55%	0.142
EML6	2	5	7	6.80%	4	3.10%	0.158
WASH1	1	1	2	1.94%	0	0.00%	0.196
ATP8B2	2	2	4	3.88%	2	1.55%	0.243
CP	2	2	4	3.88%	2	1.55%	0.243
CPT1C	2	2	4	3.88%	2	1.55%	0.243
MAGEB10	2	2	4	3.88%	2	1.55%	0.243
CHD5	2	4	6	5.83%	4	3.10%	0.244
JAK1	2	2	4	3.88%	3	2.33%	0.377
CDK8	1	1	2	1.94%	1	0.78%	0.416
MGAT4C	2	1	3	2.91%	3	2.33%	0.547
ZNF717	2	1	3	2.91%	3	2.33%	0.547
TCEB3CL	2	3	5	4.85%	7	5.43%	0.685
WDR87	1	4	5	4.85%	11	8.53%	0.915

*Data from the unpaired AA tumor sample was excluded from the analysis.

AA ethnicity ($P = 0.007$) (Table 2). Similarly, mutations in *folliculin* (*FLCN*) and *5-hydroxytryptamine (serotonin) receptor 1F* (*HTR1F*) were also detected exclusively in AA cases (Table 2). Because of the lower mutation frequencies for *FLCN* and *HTR1F*, association with AA ethnicity was of lesser statistical significance ($P = 0.086$ for each gene) (Table 2).

Among the newly identified mutational targets, two genes, *EPHA6* and *FLCN*, were of particular interest because of their being mutated exclusively in AA CRC as well as their belonging to known oncogenic pathways. Sanger sequencing successfully reconfirmed each of the nine somatic mutations we had detected in these two genes (Fig. S4). The six somatic mutations detected in *EPHA6* occurred throughout the gene (Fig. 1). One mutation altered a canonical splice site, and four missense mutations were predicted to significantly alter protein function (Fig. 1). This genetic pattern would be consistent with *EPHA6* having activity as a CRC tumor suppressor. Supporting this argument, *EPHA6* is a member of the broader family of ephrin receptor tyrosine kinases that have been demonstrated to function as oncogenes or tumor suppressors depending on the disease context (12). For example, *EPHA3* and *EPHB6* were both identified as significantly mutated genes in our previous analysis of the colon cancer genome of a near all Caucasian MSS CRC cohort (6). Intriguingly, to our knowledge the present study is the first to implicate *EPHA6* in CRCs, with the finding of *EPHA6* somatic mutations in ~6% of the AA CRC cases suggesting a provocative ethnicity-associated difference in selection of a different EPH family member for mutational targeting (Fig. 1 and Table 2).

Although less frequent, the three mutations detected in *FLCN* were also notable in that they included two frameshift and one nonsense mutation (Fig. 1). Moreover, inactivating germ-line *FLCN* mutations are associated with Birt-Hogg-Dubé syndrome, characterized by development of medullary thyroid carcinomas, perifollicular fibromas, and renal cancers (13), a phenotype recapitulated in mice lacking a single *Flcn* allele (14). Furthermore, colon neoplasia has also been reported in some individuals with Birt-Hogg-Dubé (15). Taking these data together, we propose *EPHA6* and *FLCN* as candidate driver genes preferentially targeted for mutation in AA CRCs.

In identifying the association of the above gene mutations with AA CRC, it was key to recurate the publicly available mutational databases (4, 6, 7, 16) to identify an appropriate comparator group of nonhypermutator CRCs (details in *Materials and Methods*). For example, the Catalogue of Somatic Mutations in Cancer (COSMIC) database reports 18 *FLCN* CRC mutations; however, 17 are in hypermutator CRCs [nearly all with microsatellite instablility (MSI)], and one variant is unconfirmed as being somatic plus is in an individual of unknown ethnicity (Dataset S1, Table S13). Similarly, of 48 reported *EPHA6* mutations in CRCs, 39 are in hypermutator CRCs, 4 are silent mutations, 4 are variants not determined as somatic, and the one remaining variant is in a patient of unknown ethnicity (Dataset S1, Table S13).

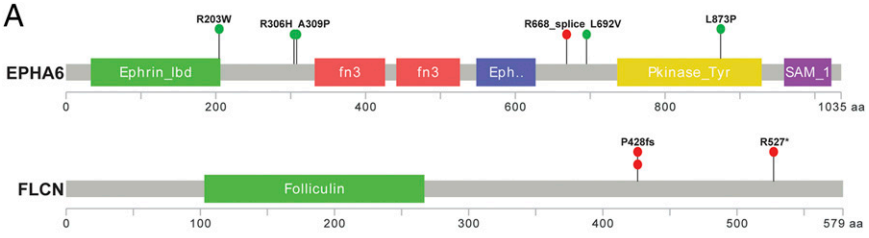
Discussion

In summary, by sequencing AA CRCs, we have identified 20 new genes as significantly mutated in CRCs. Mutations in a set of 15 of these genes appear to be strongly preferentially associated with CRCs arising in AA versus Caucasian individuals, suggesting an important difference in the mutational landscapes of CRCs arising in different ethnic groups. Moreover, mutations in *EPHA6* and *FLCN*, which are unique to AA CRCs, are highly likely to be AA CRC driver mutations, as supported by the likely inactivating nature of these mutations and by the biological pathways in which these two genes participate. One limitation of this study is that our cohort of cases was drawn exclusively from northeastern Ohio. The history of differing patterns of internal migration of African American individuals in America will make it of clear interest to compare these findings with those in African American communities from other regions of the country. Mutations in the 15 genes found preferentially targeted in AA CRC accounted for 41% of all AA CRCs (and only 15% of Caucasian CRCs) in this study. Future investigations will be warranted to examine the potential impact of mutations in this gene panel on differences in colon cancer outcome. Additionally, future investigation will be of interest to elucidate the contribution of genetic, environmental, and socioeconomic factors to these observed ethnicity associated differences in CRC mutational landscapes.

Materials and Methods

Patient Samples and DNA Extraction. Fresh-frozen tumor and matched normal specimens were collected under an Institutional Review Board approved protocol from an archive of AA and Caucasian CRC cases at the Case Medical Center. Genomic DNA from the tumor samples was extracted as previously

A



B

Hugo_ Symbol	Tumor ID	Race	Colon cancer Stage	Screen	Variant_ Class	Protein_ Change	Tumor_Mutant allele frequency	PPH2_Class
EPHA6	11843	AA	Stage IV	Discovery	Missense	R203W	0.20	Deleterious
EPHA6	15873	AA	Stage IV	Discovery	Missense	R306H	0.35	Deleterious
EPHA6	16765	AA	Stage IIIB	Validation	Missense	A309P	0.25	Deleterious
EPHA6	13129	AA	Stage IIB	Validation	Splice	R668_splice	0.47	
EPHA6	16700	AA	Stage IIB	Validation	Missense	L692V	0.21	
EPHA6	16714	AA	Stage IIB	Validation	Missense	L873P	0.10	Deleterious
FLCN	16670	AA	Stage IIA	Validation	FS ins	P428fs	0.66	
FLCN	16518	AA	Stage IIIB	Validation	FS ins	P428fs	0.46	
FLCN	11604	AA	Stage IV	Discovery	Nonsense	R527*	0.80	

Fig. 1. Somatic mutations in *EPHA6* and *FLCN*. (*A*) Somatic mutations mapped to *EPHA6* and *FLCN* protein coding regions (gray bars). Colored boxes indicated annotated protein structural domains. (*B*) Annotation of *EPHA6* and *FLCN* mutations in respective tumor samples.

described (17). DNA from all patients' tumors was tested for microsatellite instability by comparison of microsatellite alleles in tumor and matched normal DNA at microsatellite markers: BAT26, BAT40, D2S123, D5S346, and D17S250 (18). Only tumors with MSS were included in the study. The tumor stage, sex, race, MSI status, and tissue source of the DNA for the AA discovery and validation cases in the study are detailed in Dataset S1, Tables S1 and S8, respectively, and for the Caucasian CRC cases in Dataset S1, Table S11. Before use, all tissue samples were reviewed by an anatomic pathologist (J.E.W.) to confirm the colon cancer diagnosis and to confirm a cancer cell content in the sample of greater than or equal to 50%.

Whole-Exome DNA Sequencing. A total of 31 advanced-stage MSS AA CRC and matched normal sample DNAs were selected for whole-exome sequencing (WES) as part of the discovery screen (Dataset S1, Table S1). Target capture, library preparation, and deep sequencing were performed by the Oklahoma Medical Research Foundation Next Generation DNA Sequencing Core Facility (Oklahoma City, OK). Target sequence enrichments were performed using the Illumina TruSeq Exome Enrichment kit as per the manufacturer's protocols (Illumina). Briefly, sample DNAs were quantified using a picogreen fluorometric assay and 3 μg of genomic DNA were randomly sheared to an average size of 300 bp using a Covaris S2 sonicator (Covaris). Sonicated DNA was then end-repaired, A-tailed, and ligated with indexed paired-end Illumina adapters. Target capture was performed on DNA pooled from six indexed samples, following which, the captured library was PCR amplified for 10 cycles to enrich for target genomic regions. The captured libraries were precisely quantified using a quantitative PCR-based Kapa Biosystems library quantification kit (Kapa Biosystems) on a Roche Lightcycler 480 (Roche Applied Science). Deep sequencing of the capture-enriched pools was performed on an Illumina HiSEq. 2000 instrument with 100-bp, paired-end reads. For tumors, pools of six captured libraries were sequenced across three lanes on the HiSeq instrument. For normal samples, pools of six captured libraries were sequenced across two lanes on the HiSeq instrument. DNA from the normal tissue, matched to tumor sample 11481, failed at library generation phase, and accordingly was not sequenced (Dataset S1, Table S1).

Read Mapping, Somatic Mutation Detection, and Annotation. Burrows-Wheeler Aligner (19) was used to align individual 100-bp reads from the raw FASTQ files to the human reference genome (build hg19) with default parameters. Following the conversion of aligned reads into Binary Sequence Alignment/Map format, coverage metrics of target bases were calculated using the Picard tools (samtools.sourceforge.net). The coverage statistics for the WES samples are detailed in Dataset S1, Table S2. Somatic variations (point mutations and indels) in tumor samples were detected using three algorithms specifically developed for normal-tumor paired analyses. Point mutations were detected using MuTect (8) and VarScan 2 (10), and indels were detected using VarScan 2 (10) and Genome Analysis Toolkit Somatic indel detector (9). All algorithms were set to default parameters, and with the matched normal samples serving as reference genomes. For each tumor, we took the union of all somatic variants predicted by the three algorithms for subsequent analyses. To address the possibility that some sequencing errors might still be detected in a tumor only but not in its respective matched-normal sample, thus generating a false somatic variant call, we additionally performed a comprehensive query comparing each of the presumed somatic variants identified in tumors against the entire matched-normal sample-set plus an additional 123 platform-matched germ-line samples using the SAMtools software package (20). Presumptive somatic mutations that were well detected in sequencing reads of 10% or more of germ-line normal samples sequenced on the same targeted capture platform presumptively represent recurrent sequencing or mapping errors, and were subsequently eliminated after confirmation by manual review using the Integrative Genomics Viewer (IGV) software (21). Additionally, variants identified as somatic variants by only one of the three callers were filtered-out whenever the variant allele frequency observed in respective matched normal sample was ≥10% or if the variant was found in public databases (22) (evs.gs.washington.edu/EVS). The final list of predicted somatic variants was mapped to the human transcriptome reference database (RefSeq, build hg19) using a variant annotation tool developed in-house (SLATE), which identifies variants mapping to gene-coding regions and splice-sites, including their corresponding positions and codon changes within respective transcripts.

Of the 31 colon tumors used in the discovery exome analysis, one of the tumors, 3213, was found to be hypermutated and therefore was not included in subsequent analyses (Fig. S1 and Dataset S1, Table S1). As mentioned above, in a second tumor sample, 11481, library preparation from the matched normal tissue failed. Private variants identified in this tumor were examined for the purpose of nominating genes for inclusion on an Agilent SureSelect^{XT} custom capture bait library (described below), but were otherwise excluded from all statistical analyses. The overall background mutation rate in the remaining 29 discovery exome samples was estimated using the Genome MuSiC software package (23) (Dataset S1, Table S4).

Comparative Analysis of Genes Mutated in the AA CRC Discovery Exomes with Prior Studies in CRCs. We used the following workflow to determine if AA CRCs showed recurrent somatic mutations in genes that were not previously identified in CRCs (4, 6, 7). First, we identified a panel of 385 genes showing recurrent mutations in the AA CRC discovery screen: that is, with nonsilent mutations in at least two individual cancers, including the unpaired tumor sample from the discovery screen. Next, we derived the Entrez gene identifiers for each of the 385 genes in the AA CRC discovery dataset using the mappings provided by the org.Hs.eg.db database from Bioconductor (24), and by using org.Hs.egSYMBOL2EG object. Next, we obtained a list of all genes showing nonsilent somatic mutations in MSS nonhypermutator CRCs from three prior large-scale sequencing studies in CRCs for which ethnicity data were available from most of the cases (4, 6, 7). Data from two tumors annotated as from AA cases was removed. Then, using the Entrez identifiers as anchor and org.Hs.egALIAS2EG object, we checked if any of the aliases of the 385 genes identified in the AA CRC dataset were observed in the mutational datasets from these prior studies (4, 6, 7). In parallel, we performed manual curation of respective datasets wherever necessary. Overall, 288 genes in the AA CRC dataset showed at least one nonsilent somatic mutation in prior studies of MSS nonhypermutator CRCs (4, 6, 7). For 97 of the 385 genes in the AA CRC dataset, we found no prior reports demonstrating protein-altering mutations among MSS nonhypermutator CRCs (4, 6, 7) (Dataset S1, Table S5). Further review of the mutations in 97 genes using IGV (21) resulted in the exclusion of 19 genes, with mutations being observed in regions of poor mappability and mutations that were obvious sequencing errors. In the end, 78 candidate genes in the AA CRC exome dataset that met our above inclusion criteria were selected for subsequent analyses.

Targeted Resequencing of Candidate AA CRC Genes. A custom Agilent SureSelect^{XT} bait library (Agilent Technologies) was designed for targeted capture and resequencing of a panel of 78 candidate genes that were selected based on: (*i*) no prior evidence of their being mutational targets in microsatellite stable nonhypermutator colon cancers (4, 6, 7), as described above; and (*ii*) each of these 78 candidate AA CRC genes showing nonsilent somatic mutations in two or more individual cancers from the exome dataset. Sixteen of these 78 genes were nominated by noticing a private nonsilent variant in the unpaired 11481 tumor sample plus an additional somatic mutation in the 29 discovery samples (Dataset S1, Table S5). Although used to help nominate these 16 candidates, data from this unpaired tumor was excluded from all further statistical analysis. In addition to the set of 78 candidate genes, a panel of 135 genes was included as controls on the custom-capture bait library to facilitate the identification of hypermutated cancers in analysis of the validation tumor cohort (Dataset S1, Table S6).

Custom capture and resequencing were performed as follows. Briefly, 3 μg of sample DNAs were randomly sheared to obtain a target peak size of 150–200 bp using a Covaris S2 sonicator. Sonicated DNA was then end-repaired, A-tailed, adapter-ligated, amplified, and quality-assessed using a TapeStation instrument (Agilent Technologies). Next, 750 ng of the library was individually captured using the Agilent custom-capture bait library, and PCR amplification of the captured libraries was carried out for 14 cycles during which unique indexed tags were added to respective sample libraries. The captured libraries were precisely quantified using a quantitative PCR-based Agilent library quantification kit. Samples libraries were pooled in batches of 24–48, and each pooled library was deep-sequenced on a single lane of Illumina HiSEq. 2000 instrument to obtain 100-bp paired-end reads. Estimation of coverage statistics in samples used for targeted resequencing showed an average of 95% of bases covered at 30X depth, with 99% of samples showing at least 90% of the bases covered at 30X. Read mapping, somatic mutation detection, and annotation for targeted resequencing analysis was carried out as detailed above.

Resequencing analysis of the 78 candidate AA CRC genes in the discovery exome samples resulted in 52 genes confirmed as being mutated in at least two cancers (Dataset S1, Table S7). The remaining 26 genes included 19 genes with a mutation confirmed in only one of the discovery cancers, and an additional 7 genes for which none of the mutations observed in the WES validated. These 26 genes were excluded from subsequent analyses. The 52 genes, with confirmed mutations in at least two cancers, were accordingly selected for subsequent resequencing in an independent cohort of 77 fresh-frozen, predominantly early-stage MSS AA CRC and matched normal DNA

samples, using the Agilent custom-capture bait library described above (Dataset S1, Tables S8 and S9).

Mutational Significance Estimates for the 52 Candidate AA CRC Genes in the Discovery and Validation Cohorts. To identify which among the 52 candidate AA CRC genes were mutated more frequently than expected based on the overall background mutation rate in AA CRCs, we followed a statistical framework similar to our previously published study detailing mutational profiles of colon cancer (6). First, the overall background mutation rates at six different nucleotide contexts, including AT transitions, AT transversions, CG transitions, CP transversions, CpG transitions, and CpG transversions were calculated using the Genome MuSiC suite (23) in the exome samples (Dataset S1, Table S4). Next, for each of the 52 candidate genes, we counted the number of mutations in each of the above six nucleotide contexts plus indels in the discovery exome dataset, and calculated the probability of the observed number of mutations in a particular category using an exact binomial distribution. The total probability of a gene exhibiting the observed number of mutations in all of the seven categories was then calculated to be the product of the seven context-specific probabilities. To correct these probabilities for multiple comparisons, we used the algorithm described by Benjamini and Hochberg (25). Genes with a P value < 0.05 and FDR < 0.05 were considered to be significantly mutated. Mutational significance for the candidate genes in the African American validation cohort were similarly calculated using context-specific background mutation rate values derived from the exome dataset (Dataset S1, Table S4). Overall, 20 genes were found to be significantly mutated in both the discovery and validation African American cohorts (Table 1 and Dataset S1, Table S10). We note that using a selection criteria requiring that genes be significantly mutated in both the discovery and the validation cohorts may overlook some genes that could validate on further testing. However, we selected the above algorithm in the interest of minimizing any false discovery.

Comparison of Mutational Frequencies of the 20 Candidate Genes in AA vs. Caucasian CRCs. The 20 significantly mutated genes in AA CRCs (Table 2) were resequenced in a panel of 129 fresh-frozen, predominantly late-stage MSS Caucasian CRC and matched normal DNA samples, using the Agilent custom-capture bait library described above (Dataset S1, Tables S6, S11, and S12). We then evaluated if the overall mutational frequency of these 20 genes, as a group, is significantly higher in the AA (discovery + validation cohorts, $n = 103$) versus Caucasian ($n = 129$) CRCs by using a paired Student's t test. Taken as a group, these 20 genes averaged a statistically significant ~twofold increase in the total number of mutations/tumor in AA CRCs compared with Caucasian CRCs ($P < 0.001$). After finding the statistical significance for differences in these 20 genes taken as a group between AA and Caucasian CRCs, subgroup analysis was done by comparing for each individual gene the difference in the mutational frequencies between AA and Caucasian CRCs using a Fisher's exact test. *EPHA6* showed the most significant nominal P value of 0.007. Further subgroup analysis was done by ranking the 20 genes from most to least significant nominal P values, and consecutively performing a paired Student's t test on groups of the top 2, 3, 4 ... 20 genes. The nominal P value was most significant for the group consisting of the top 15 genes ($P = 1.8 \times 10^{-8}$) (Fig. S3).

The above statistical results are not critically dependent on use of a paired Student's t test. For example, analysis using the less powerful Wilcoxon signed-rank and Sign tests also identify the set of 20 candidate genes as having significantly more mutations in AA vs. Caucasian CRCs ($P < 0.002$ and < 0.001, respectively).

In the above analysis, presented in Table 2, genes with two mutations in the same patient were counted only once. These included one instance each of: *CHD5* (AA validation), *KIAA1551* (AA validation), *WDR87* (AA validation), *TCEB3CL* (Caucasian validation), *WDR87* (Caucasian validation) (Dataset S1, Tables S9 and S12). Counting these additional mutations would not have changed the above conclusions.

Of note, comparison of mutational frequencies of well-known CRC driver genes (*APC*, *TP53*, *KRAS*, *PIK3CA*, *SMAD4*, *BRAF*, *FBXW7*), as a group, showed no significant differences between AA versus Caucasian CRCs, although *KRAS* and *FBXW7* individually showed a significantly higher mutation frequency in AA than Caucasian CRCs (Dataset S1, Table S14). Increased *KRAS* mutations in AA CRCs has also been previously reported by others (1, 3).

Identification of Hypermutator Tumors. All tumors sequenced in the above analyses were first screened for MSI, with only MSS CRC cases included for sequencing. Moreover, following sequencing, MSS hypermutator tumors were also identified and excluded from analysis. In the whole-exome data, MSS hypermutator tumors were identified using the criterion from the published The Cancer Genome Atlas (TCGA) study (4), assigning samples with a mutation rate >12 per 10^6 base pairs (corresponding to total mutation count >237) as hypermutators (4). One discovery sample met this definition and was excluded (sample 3213, total mutations = 587, total nonsilent mutations = 447) (Fig. S1). In targeted capture data, cases with mutation counts outside the body of the mutation distribution were classified as hypermutators. Three AA validation cases met this definition and were excluded (Fig. S1). Of note, no Caucasian cases were excluded from the analysis using these criteria. Thus, the exclusion of one AA discovery and three AA validation cases is highly conservative with respect to our final conclusions.

Curation of *EPHA6* and *FLCN* Mutations in COSMIC Database. We obtained a list of all mutations reported for *EPHA6* and *FLCN* genes in cecum, colon, and rectal cancers within the COSMIC database (16), which includes mutational data from all four published large-scale sequencing studies in CRCs (4–7) (accession date 07/20/2014) (Dataset S1, Table S13). For each tumor sample, we calculated the total number of unique mutations. Tumor samples derived from TCGA datasets within COSMIC, that were not in published literature, were annotated for race, MSI status, and tumor stage, where available, using the TCGA data-portal (tcga-data.nci.nih.gov/tcga, accession date 06/05/2014). Tumor samples with MSI status designated as MSI-H were automatically assigned hypermutator status. Moreover, among the tumors designated as either MSS, MSI-L, or unknown, we observed total mutation counts ranging from 106 to 12,202, suggesting that several of these samples were also hypermutators. Accordingly, to identify MSS hypermutator tumors we adapted the criterion from the published TCGA study, assigning samples with a mutation rate >12 per 10^6 (corresponding to total mutation count >237) as hypermutators (4). The median mutation count among these MSS hypermutator tumors was 1,204 in the COSMIC dataset. The condensed summary of our analyses as provided in Dataset S1, Table S13 shows that the vast majority of reported *EPHA6* and *FLCN* mutations (56 of 66) were derived from hypermutator cases, with the remaining variants not characterized as being somatic versus germ line (5 of 66) or being silent mutations (4 of 66). Overall, we found one MSS nonhypermutator tumor showing a confirmed somatic missense mutation in *EPHA6*; however, the ethnicity of this patient was unknown (Dataset S1, Table S13).

Sanger Sequencing. The primers for amplifying mutant positions in *EPHA6* and *FLCN* genes are listed in Dataset S1, Table S15. The PCR conditions included 95 °C for 4 min, 38 cycles of 95 °C for 45 s, 62.3 °C for 30 s, and 72 °C for 45 s. Each reaction was carried out in a 50-μL reaction volume using 2.5 U of Fast-TAQ DNA polymerase (Roche Applied Science) with 25–50 ng of template DNA. The PCR products were purified and were either directly sequenced using universal M13 forward and reverse primers or sequenced following subcloning the PCR products in bacteria. Analysis of Sanger sequencing data were performed using Mutation Surveyor software package (SoftGenetics).

Mutational Spectrum Analysis in the WES Dataset. Mutational spectrum was evaluated in the 29 AA colon cancer discovery exome sequence dataset using the SomaticSignatures package in Bioconductor (bioconductor.org/packages/release/bioc/html/SomaticSignatures.html, github.com/julian-gehring/SomaticSignatures). Briefly, all somatic mutations identified in each of the tumors were grouped into the six possible classes of nucleotide substitutions (C > A, C > G, C > T, T > A, T > C, T > G) and further stratified across the 16 dinucleotide sequence contexts based on the identity of the reference nucleotides on the 5′ and 3′ end of the mutation. These counts were normalized to the total number of mutations in the cohort resulting in the frequency of the individual somatic alterations within their respective sequence contexts.

ACKNOWLEDGMENTS. We thank Courtney Montgomery, Graham Wiley, and Simone Edelheit for performing massively parallel DNA sequencing; Debora Poruban for assisting with DNA library generation for targeted resequencing studies; Jill Barnholtz-Sloan for helping with the curation of the COSMIC database; and the honorable Louis Stokes for his inspiration and encouragement of this project. This research was supported by Public Health Service Awards Case GI SPORE P50 CA150964 (to S.D.M.) and KO8 CA148980 (to K.G.); Career Development Program of Case GI SPORE Awards P50 CA150964 (to K.G. and V.V.), R21 CA149349 (to J.E.W.), and P30 CA043703; and by gifts from the Marguerite Wilson Foundation (S.D.M.), the Leonard and Joan Horvitz Foundation (S.D.M), and the Richard Horvitz and Erica Hartman-Horvitz Foundation (S.D.M.).

1. Gupta S, et al. (2014) Challenges and possible solutions to colorectal cancer screening for the underserved. *J Natl Cancer Inst* 106(4):dju032.
2. Jemal A, et al. (2008) Cancer statistics, 2008. *CA Cancer J Clin* 58(2):71–96.
3. Sylvester BE, et al. (2012) Molecular analysis of colorectal tumors within a diverse patient cohort at a single institution. *Clin Cancer Res* 18(2):350–359.
4. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330–337.
5. Seshagiri S, et al. (2012) Recurrent R-spondin fusions in colon cancer. *Nature* 488(7413): 660–664.
6. Sjöblom T, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314(5797):268–274.
7. Wood LD, et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318(5853):1108–1113.
8. Cibulskis K, et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31(3):213–219.
9. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
10. Koboldt DC, et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22(3):568–576.
11. Alexandrov LB, et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013) Signatures of mutational processes in human cancer. *Nature* 500(7463):415–421.
12. Boyd AW, Bartlett PF, Lackmann M (2014) Therapeutic targeting of EPH receptors and their ligands. *Nat Rev Drug Discov* 13(1):39–62.
13. Nickerson ML, et al. (2002) Mutations in a novel gene lead to kidney tumors, lung wall defects, and benign tumors of the hair follicle in patients with the Birt-Hogg-Dubé syndrome. *Cancer Cell* 2(2):157–164.
14. Hartman TR, et al. (2009) The role of the Birt-Hogg-Dubé protein in mTOR activation and renal tumorigenesis. *Oncogene* 28(13):1594–1604.
15. Khoo SK, et al. (2002) Clinical and genetic studies of Birt-Hogg-Dubé syndrome. *J Med Genet* 39(12):906–912.
16. Forbes SA, et al. (2011) COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 39(Database issue):D945–D950.
17. Fecteau RE, Lutterbaugh J, Markowitz SD, Willis J, Guda K (2014) GNAS mutations identify a set of right-sided, RAS mutant, villous colon cancers. *PLoS ONE* 9(1):e87966.
18. Umar A, et al. (2004) Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst* 96(4): 261–268.
19. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
20. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
21. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* 14(2):178–192.
22. Abecasis GR, et al.; 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
23. Dees ND, et al. (2012) MuSiC: Identifying mutational significance in cancer genomes. *Genome Res* 22(8):1589–1598.
24. Carlson M (2013) org.Hs.eg.db: Genome Wide Annotation for Human, R Package Version 2.14.0. Available at www.bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html. Accessed May 19, 2014.
25. Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. *Stat Med* 9(7):811–818.