Inherited DNA Repair Defects in Colorectal Cancer

Authors: Saud H. AlDubayan, MD[1,2,3,4,12], Marios Giannakis, MD, PhD[1,2,12], Nathanael D. Moore, BSc[1,2,5,6], G. Celine Han, PhD[1,2], Brendan Reardon, BSc[1,2], Tsuyoshi Hamada, MD, PhD[7,8], Xinmeng Jasmine Mu, PhD[1,2], Reiko Nishihara, PhD[9], Zhirong Qian, PhD[9], Li Liu, MD, PhD[9], Matthew B. Yurgelun, MD[1], Sapna Syngal, MD, MPH[1], Levi A. Garraway, MD, PhD[1,2,12], Shuji Ogino, MD, PhD[7,8,9,12], Charles S. Fuchs, MD, MPH[10,11,12], Eliezer M. Van Allen, MD[1,2,12,*]

Affiliations:
1 Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, 02215
2 The Broad Institute of MIT and Harvard, Cambridge, MA, 02142
3 Division of Genetics, Brigham and Women's Hospital, Boston, MA, 02115
4 Department of Medicine, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
5 Indiana University School of Medicine, Indianapolis, IN, 46202
6 Howard Hughes Medical Institute, Chevy Chase, MD, 20815
7 Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA, 02115
8 Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, MA, 02215
9 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, 02115
10 Yale Cancer Center, Yale School of Medicine, New Haven, CT, 06510
11 Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, 02115
12 Equal contribution

* Corresponding Author

Keywords: colorectal cancer genetics, germline genetics, cancer heritability, *ATM* mutations, *PALB2* mutations, homologous recombination, DNA repair deficiency

Abstract:
Colorectal cancer (CRC) heritability has been estimated to be around 30%. However, mutations in the known CRC susceptibility genes explain CRC risk in under 10% of the cases. Germline mutations in DNA-repair genes (DRGs) have recently been reported in CRC but their contribution to CRC risk is largely unknown. We evaluated the gene-level germline mutation enrichment of 40 DRGs in 680 unselected CRC individuals compared to 27728 ancestry-matched cancer-free adults. Significant findings were then examined in independent cohorts of 1661 unselected CRC cases and 1456 early-onset CRC cases. Of 680 individuals in the discovery set, 31 (4.56%) individuals harbored germline pathogenic mutations in known CRC susceptibility genes while another 33 (4.85%) individuals had DRG mutations that have not been previously associated with CRC risk. Germline pathogenic mutations in *ATM* and *PALB2* were enriched in both the discovery (OR= 2.81; P= 0.035 and OR= 4.91; P= 0.024, respectively) and validation sets (OR= 2.97; Adjusted P= 0.0013 and OR= 3.42; Adjusted P= 0.034, for *ATM* and *PALB2* respectively). Biallelic loss of *ATM* was evident in all cases with matched tumor profiling. CRC cases also had higher rates of actionable mutations in the HR pathway that can substantially increase the risk of developing cancers other than CRC. Our analysis provides evidence for *ATM* and *PALB2* as CRC risk genes, underscoring the importance of the homologous recombination pathway in CRC. In addition, we identified frequent complete homologous recombination deficiency in CRC tumors, representing a unique opportunity to explore targeted therapeutic interventions such as PARPi.

Introduction:

Colorectal cancer (CRC) [MIM: 114500] is the third most common malignancy in the US[1]. Although most CRC cases are thought to be sporadic, recent twin studies have estimated that 30% of the inter-individual variability in CRC risk is attributed to inherited genetic factors[2]. Over the past few decades, several CRC predisposition genes, including *APC* [MIM: 611731], *MLH1* [MIM: 120436], *MSH2* [MIM: 609309], *MSH6* [MIM: 600678], *PMS2* [MIM: 600259], *STK11* [MIM: 602216], *MUTYH* [MIM: 604933], *SMAD4* [MIM: 600993], *BMPR1A* [MIM: 601299], *PTEN* [MIM: 601728], *TP53* [MIM: 191170], *CHEK2* [MIM: 604373], *POLD1* [MIM: 174761] and *POLE* [MIM: 174762], have been described[3-5]. Collectively, mutations in these Mendelian CRC risk genes explain the increased risk for CRC in 5-10% of unselected cases[6-9]. The discrepancy between the proportion of CRC cases explained by these genetic risk factors and the estimated degree of heritability, known as "missing heritability", indicates that one or more undiscovered inherited risk factors contribute to CRC risk.

DNA-repair is a critical biological process that prevents permanent DNA damage and ensures genomic stability. Although defects in DNA mismatch repair and certain DNA polymerases have been implicated in CRC risk, the role of other canonical DNA repair pathways is less defined. Our group and others have reported several observational studies which showed that some CRC cases were found to have germline mutations in DNA-repair genes (DRGs), such as *ATM* [MIM: 607585], *BRCA1* [MIM: 113705], *BRCA2* [MIM: 600185], and *PALB2* [MIM: 610355], that have classically been associated with susceptibility to cancers other than CRC[6, 10, 11]. As these DRG mutations are also present in the general population at a very low frequency, it is still unclear if these DRG defects are truly associated with a higher CRC risk or merely represent incidental findings in these CRC individuals[12]. To date, there has not been a case-control study to systematically examine candidate DRGs for potential germline mutation enrichment.

Here, we build upon our previous observations to evaluate the role of gene-level DRG defects in CRC susceptibility using germline data from CRC individuals and cancer-free controls in a case-cohort study, with complementary somatic analyses of candidate genes. We hypothesized that germline mutations in DRGs previously linked to other Mendelian forms of inherited cancer predisposition account for a significant fraction of the missing CRC heritability. To investigate this hypothesis, we studied germline whole exome sequencing data in a large discovery set of CRC cases who were not preselected for early-onset disease or positive family history and subsequently validated our findings in an independent large validation set of similarly unselected CRC cases. For CRC individuals who had disruptive germline mutations in genes related to homologous recombination, we also examined somatic tumor DNA for biallelic inactivation so as to explore whether such CRCs might theoretically be treated by agents that target deficient double-strand DNA repair (e.g. PARP inhibitors).

Methods:

Study subjects

1- Discovery set:

Two independent cohorts that included 680 CRC persons were examined in the discovery phase (Figure S1). Of these, 591 CRC persons came from the population-based Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) cohorts[13]. Only cases with available self-reported ancestry information were included in this case series. CRC cases from the NHS/HPFS were not selected on the basis of their age of presentation, stage of their disease or presence of a positive family history of CRC or other cancers[13]. In addition, 89 CRC persons from the CanSeq study at Dana-Farber Cancer Institute (DFCI) were included in the discovery set[14]. The CanSeq study is a single-arm prospective study that aims to evaluate the clinical utility of using paired (tumor and normal) whole exome sequencing in the clinical care of individuals with advanced cancer without pre-selection for early age at diagnosis or high-risk family histories (hereafter referred to as "unselected cases")[15]. Both studies were approved by the Partners Human Research Committee institutional review board (NHS/HPFS: BWH IRB#2001-P-001945, CanSeq: DFCI IRB#12-078)), and informed consent was obtained from all subjects.

2- Validation set:

Germline data of 1661 subjects from two independent cohorts of unselected CRC cases, The Cancer Genome Atlas (TCGA; n = 603) and the cohort reported by Yurgelun et al. (n = 1058) were used to validate the main findings detected in the discovery phase (hereafter called "the validation set")[16, 17]. Both cohorts were not selected for early-onset disease or positive family history. Similar variant calling and pathogenicity assessment pipelines were used to evaluate germline variants in both cohorts.

3- Early-onset CRC set:

To further delineate the penetrance of DRGs with significant germline mutation enrichment in the discovery and validation sets in CRC individuals, germline mutation enrichment in 1456 early-onset (age<56) CRC cases was evaluated. These cases were part of two large CRC studies[10, 18]. In total, our study evaluated relevant germline sequencing data of 3797 CRC cases relative to cancer-free adult controls (Figure S1).

Sequencing and Bioinformatics Analysis

Germline DNA from the CRC subjects in the discovery set was obtained from whole blood or adjacent normal colon tissue that was dissected after pathology review. DNA was extracted from formalin-fixed, paraffin embedded (FFPE) blocks using commonly used practices[19]. All germline variants in the validation and early-onset CRC sets were detected form whole blood. Production pipelines of the germline variants of these cohorts are described in Table S1 and elsewhere[10, 13, 16-18]. Partial or whole gene deletions were not evaluated in this study.

Selection of DNA-repair genes and gene sets

Only genes that have been clearly associated with a Mendelian cancer-predisposition syndrome in humans were examined. A total of 14 well-known CRC risk genes, as well as 40 DRGs that have been associated with cancer phenotypes other than CRC, were evaluated (Tables S2 and S3). Some of these DRGs such as *BLM* [MIM: 210900] and *NTHL1* [MIM: 602656] have been recently linked to CRC susceptibility, however these observations have not been so far independently validated so these genes were included in the DRG set to be evaluated here. Analysis of the germline variants in *POLE* and *POLD1* was restricted to the known pathogenic missense mutations in the exonuclease domain of the protein.

Of the examined DRGs, 14 genes play an important part in the homologous recombination pathway: *ATM*, *BARD1* [MIM: 601593], *BLM, BRCA1, BRCA2, BRIP1* [MIM: 605882], *MRE11* [MIM: 600814], *NBN* [MIM: 602667], *PALB2, RAD51* [MIM: 179617], *RAD51C* [MIM: 602774], *RAD51D* [MIM: 602954], *RAD54L* [MIM: 603615], and *XRCC3* [MIM: 600675][20]. "Actionable DRGs" were defined as established cancer predisposition genes that confer a 3-fold or higher increase in the risk for cancer phenotypes other that CRC and for which enhanced screening and family genetic testing are recommended. Out of the examined DRGs, *ATM*, *BRCA1*, *BRCA2*, *BRIP1, PALB2, RAD51C,* and *RAD51D* were considered clinically actionable[21-25].

Variant Interpretation

An identical workflow for variant inclusion and pathogenicity assessment was used to evaluate the germline variants in both cases and controls (Table S1). The clinically-oriented American College of Medical Genetics and Genomics (ACMG) germline variant assessment guidelines were used to evaluate germline variants in cases and controls. Based on the available evidence, germline variants were classified into 5 categories: benign, likely benign, variants of unknown significance, likely pathogenic and pathogenic[26]. Only germline variants which had sufficient evidence of pathogenicity to be classified as pathogenic or likely pathogenic variants (hereafter collectively referred to as pathogenic mutations) were included. All variants of unknown significance (VUS) were excluded from all analyses.

Frequency of mutations in the general population:

Annotated germline variants in the examined genes in 53105 cancer-free adults from the Exome Aggregation Consortium (ExAC) (release 0.3.1 on 3/16/2016), excluding the TCGA cohort, were also evaluated using an identical workflow to the one used for cases [27]. Frequencies of germline pathogenic mutations in the genes of interest were calculated for each of the continental populations in ExAC. Gene mutation frequencies for the ExAC Non-Finnish European (n=27173) and African & African American (n=4533) cohorts were then used to calculate the predicted pathogenic gene mutation frequency in an ancestry-matched control cohort of 27728 individuals (98%; 27173 Non-Finnish Europeans (NFE), and 2%; 555 African Americans (AFR)) (Figure S2)[28]. Population-specific common variant frequencies were similar in cases and

controls decreasing the likelihood of a significant population structure (Figure S3). Ancestry information for some individuals in the validation set was not readily available. Since the majority of the cases included in these studies are expected to have European ancestry, non-Finnish European individuals from the ExAC cohort (ExAC_NFE; n= 27173) were used as a control group.

Tumor LOH analysis

MuTect was applied to identify somatic single-nucleotide variants (SNVs)[29]. Strelka was used to detect small insertions and deletions. Individual sites were reviewed with Integrated Genomics Viewer (IGV)[30]. Using filtered-based method, artifacts from DNA oxidation during sequencing were removed[31, 32]. Annotation of identified variants was performed using Oncotator[33]. Probability distributions of possible cancer cell fractions (CCFs) of mutations were calculated, based on local copy-number and the estimated sample purity, using ABSOLUTE[34].

Statistical Analysis

A logistic regression model was used to examine the clinical characteristics of CRC cases with germline pathogenic mutations. Two-sided Fisher's exact tests were used to calculate the odds ratios and confidence intervals (using "Minimum likelihood correction") for the enrichment of germline pathogenic mutations in each of the examined DRGs. In addition, Exact binomial test of proportions was used to calculate the P value for the measured enrichment of each gene in CRC cases compared with the reference population. Consistent with established statistical methods for two-stage association studies, we implemented a permissive first discovery stage analysis where genes with P values smaller than 0.05 were considered significant. These top candidate genes were then tested in a subsequent validation phase in an independent cohort, prior to performing secondary analyses, with appropriate correction for multiple testing using Bonferroni correction[35-37].

Results:

Cohort characteristics and sequencing metrics of CRC cohorts

Demographic characteristics of all 680 CRC cases from the discovery cohort are summarized in Tables 1 and S4. The average target coverage for germline WES for the discovery set was 71.69X (NHS/HPFS) and 137.11X (CanSeq). DNA-repair genes, where significant germline pathogenic mutation enrichment was seen in the discovery set, were subsequently examined in 1661 unselected CRC cases and 1456 early-onset CRC cases (methods) [10, 16]. Examined DRGs had an average coverage of 58.67X in the ExAC cohort (Figure S4 and Table S5).

Germline pathogenic mutations in known CRC risk genes

In the discovery set (n = 680), 31 (4.56%) individuals had germline CRC risk mutations. Of these, 12 (1.76%) harbored highly or moderately penetrant germline pathogenic mutations in *APC* (n=2), *CHEK2* (n=4), *MSH2* (n=1), *MSH6* (n=1), *PMS2* (n=2), and *TP53* (n=2) (Figures 1a and S5; Table S6). In addition, 19 (2.79%) individuals carried heterozygous germline pathogenic mutations in *MUTYH* (n=11, 1.62%) or the Ashkenazi founder low-penetrance variant, p.Ile1307Lys, in *APC* (n=8, 1.18%). Of 1661 unselected CRC individuals in the validation set, 93 (5.6%) individuals had at least one germline mutation in the CRC susceptibility genes (Figure 1a; Tables S7 and S8). The frequency of germline mutations in the mismatch repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) in the discovery CRC set (4 patients; 0.6%) is considerably lower than the frequency of these gene mutations in other studies[38]. This underrepresentation of Lynch syndrome patients in our discovery cohort could be attributed to the population-based nature of the NHS/HPFS cohorts as well as to the fact that these studies only enrolled cancer-free subjects, sometimes at a more advanced age for some individuals.

Germline pathogenic mutations in additional DNA-Repair genes

Next, germline variants in 40 DRGs in the discovery CRC set (n=680) were evaluated for pathogenicity. Thirty-three (4.85%) subjects had at least one germline pathogenic mutation in 21 of these DRGs (Figure 1b). Four (0.59%) individuals had 2 germline pathogenic mutations each in different DRGs (Table S9). There were no cases with germline pathogenic mutations in both sets of known CRC risk genes and the additional DRGs. Enrichment analysis of the discovery CRC set, relative to cancer-free individuals, showed significant germline pathogenic mutation enrichment in *ATM* and *PALB2* (Figure 1c; Table 2).

Germline pathogenic mutations in *ATM*

Among 680 unselected CRC individuals, five (0.74%) had mutations in *ATM*. Germline mutations in *ATM* were significantly more prevalent in the CRC discovery set than cancer-free individuals (OR= 2.81; 95% CI= 1.07-6.71; P= 0.035) (Table S9). The frequency of *ATM* germline pathogenic mutations in the CanSeq cohort was not significantly higher than that of the NHS/HPFS cohort (P= 0.5) (Figure S6). Analysis of *ATM* mutation frequency in another 1661 unselected CRC cases, from the validation set, also identified significant enrichment of *ATM*

germline pathogenic mutations (13 cases; 0.78%; OR= 2.97; 95% CI= 1.57-5.39; Adjusted P= 0.0013) (Figures 1d and 2a; Tables S10 and S11). Evaluation of an independent cohort of 1456 early-onset CRC individuals similarly showed significant enrichment of germline *ATM* mutations in these individuals (10 cases; 0.69%; OR= 2.6; 95% CI= 1.3-5.07; Adjusted P= 0.013) (Figure 2a).

Although most of the cases included in our study were of European ancestry, self-reported ancestry information, as previously shown, can be inaccurate[39]. To evaluate for spurious *ATM* mutation enrichment that could have resulted from inadequate population stratification, we next blinded the ancestry data of the CRC subjects from the validation cohort and examined *ATM* mutation enrichment relative to cancer-free controls from various continental populations in ExAC. Our analysis showed that regardless of the selected control population, rates of germline *ATM* mutations were significantly higher in the CRC validation set (n=1661) (OR= 2.4-6.5, Adjusted P< 0.05 for all pairwise comparisons; Binomial Exact with Bonferroni correction for 6 independent tests) (Figure S7).

Germline pathogenic mutations in *PALB2*

Three individuals in our discovery cohort were found to have germline *PALB2* mutations, which represented a significant enrichment, compared to cancer-free controls (0.44%; OR= 4.91; 95% CI= 1.26-16.19; P= 0.024) (Table S9). This enrichment was also evident in 1661 unselected CRC cases from the validation cohort (5 cases; 0.3%; OR= 3.42; 95% CI= 1.24-9.24; Adjusted P= 0.034) (Figure 2b and Tables S10 and S11). Interestingly, no significant enrichment of germline *PALB2* mutations was seen in 1456 early-onset CRC cases (3 cases; 0.2%; OR 2.34; 95% CI= 0.6-7.75; Adjusted P= 0.28), suggesting late-onset penetrance of *PALB2* mutations in CRC individuals.

Somatic loss of heterozygosity (LOH)

Matched tumor WES for most of the individuals with germline mutations in the discovery set (n=64) were available and examined for somatic loss of heterozygosity (LOH) (Table S12). Among the CRC risk genes, somatic inactivation of the wild-type allele was seen in *APC* (8 cases; 80%), *CHEK2* (1 case; 25%), *ERCC2* (2 case, 100%), *MSH2* (1 case; 100%), *MSH6* (1 case; 100%), *MUTYH* (2 cases; 18%), *PMS2* (2 case; 100%) and *TP53* (2 cases; 100%). Out of the examined DRGs, all individuals with germline pathogenic mutations in *ATM* (5; 100%) had evidence of somatic inactivation of the wild-type allele in the matched tumor samples (Figure S8). Somatic inactivation of the *ATM* wild-type allele, in all tumors with germline *ATM* events, provides compelling evidence for *ATM* to be etiologic for the development of CRC in these cases. No somatic LOH was detected in any of the tumors of individuals with germline *PALB2* mutations, though disruptive non-coding genetic and epigenetic events are not captured by tumor WES.

Germline pathogenic mutations in the homologous recombination (HR) pathway

Given the observed mutations specifically in HR genes (*ATM* and *PALB2*), we next examined the frequency of inherited mutations affecting any of HR cancer-predisposition genes (methods). Unselected CRC individuals in the discovery set had a higher rate of germline pathogenic mutations in the HR genes compared with cancer-free individuals (19 cases; 2.8%; OR= 1.77; 95% CI= 1.07-2.84; P= 0.02) (Table S9). Evaluation of the validation and early-onset CRC sets also showed that CRC cases were more likely to have inherited HR mutations (validation set: 47 cases; 2.8%; OR= 1.78; 95% CI= 1.30-2.43; P= 2.77E-04; early-onset set: 39 cases; 2.68%; OR= 1.68; 95% CI= 1.19-2.35; P= 0.002) (Figure 2c; Tables S10 and S11). This effect did not seem to be purely driven by *ATM* and *PALB2* mutations, as when excluded, there was a trend, that did not reach statistical significance, for germline disruptive events in other HR genes to be more prevalent in the CRC validation set compared with cancer-free adults (OR= 1.4; 95% CI= 0.95-2.06; P= 0.077) (Figure S9).

Clinical actionability and risk of other cancers in CRC individuals

Analysis of mutations in actionable DRGs (*ATM, BRCA1, BRCA2, BRIP1, PALB2*, *RAD51C*, and *RAD51D*) in the discovery set identified a total of 15 germline pathogenic mutations in 14 (2.1%) CRC persons. One person had two actionable mutations in *BRCA2* and *PALB2*. Compared with cancer-free individuals, actionable cancer-risk mutations were approximately twice more prevalent in CRC cases from the discovery set (OR= 1.8; 95% CI= 1.04-3.07; P= 0.04), the validation set (36 cases; 2.17%; OR= 1.88; 95% CI= 1.31-2.69; P= 5.17E-04) as well as the early-onset CRC set (32 cases; 2.2%; OR= 1.91; 95% CI= 1.32-2.75; P= 8.31E-04) (Figure 2d).

Utility of testing relevant DRGs in CRC

Collectively, CRC heritability in up to about 1.2% of unselected CRC cases may be explained by higher rates of mutations in *ATM* and *PALB2*. To examine the potential impact of performing germline testing of *ATM* and *PALB2* on diagnostic yield, we next examined the CRC-specific germline panels offered by eight of the largest commercial laboratories in the US (as of September 2017). In addition to the known CRC risk genes, our evaluation showed that germline analysis of *ATM* is only occasionally included in these panels whereas *PALB2* and other actionable DRGs are not captured by these clinical tests (Figure S10).

Clinical characteristic of mutation carriers in the discovery set

Overall, there were no significant differences in clinical characteristics between DRG mutant or non-mutant CRC cases (Table 1). Although on average, CRC individuals with high penetrance germline CRC risk mutations presented 10.5 years younger that mutation-negative individuals (P= 0.0005), CRC individuals with germline pathogenic mutations in *ATM*, *PALB2*, the HR genes or DRGs were not more likely to present earlier that mutation-negative persons. All five germline *ATM* mutation carriers presented with stage III or IV disease (compared with 46% of

mutation-negative CRC cases; P= 0.051) (Figure 3). Individuals with germline pathogenic mutations in CRC risk genes, the DRGs, *ATM* or *PALB2* were not more likely to report a first-degree family member with CRC or other cancer types (Figure S11). Interestingly, individuals carrying a high penetrance CRC risk mutations were more likely to report a positive family history of breast cancer.

Discussion:

Most of the colorectal cancer heritability is still incompletely characterized. Mutations of several cancer-predisposition DRGs that are not typically associated with CRC have been recently reported in individuals with CRC, however, the clinical significance of these results has not been firmly established. Here, we present a systematic analysis of DRG mutations in large independent CRC cohorts relative to cancer-free adults to evaluate novel observations in known CRC susceptibility genes and to identify new CRC susceptibility genes.

We found that a gene-level analysis of DRGs revealed significantly higher rates of *ATM* mutations in CRC cases compared with cancer-free controls, going beyond observational studies to implicate its role as a novel CRC susceptibility gene. *ATM* is a master regulating kinase that is activated in response to DNA damage. Heterozygous carriers of *ATM* mutations have been reported to have a higher risk of breast [MIM: 114480] and potentially pancreatic cancer [MIM: 260350][11]. A previous cohort-based study that evaluated the risk of various cancers in families of individuals with ataxia telangiectasia [MIM: 208900], which results from biallelic loss of *ATM*, showed no increased risk of CRC in the obligate carrier parents of these cases. However, a secondary analysis in that study showed that, collectively, there was an increased risk of CRC when all the heterozygous *ATM* carrier relatives were evaluated (RR=2.54, 95% CI= 1.06-6.09), though this association was not statistically significant once corrected for multiple hypothesis testing[11]. A larger subsequent study on *ATM* carriers also failed to detect any enrichment of CRC events in heterozygous *ATM* carries[40]. However, a recent GWAS that evaluated three loss-of-function *ATM* variants in several cancer phenotypes showed a higher risk for CRC in cases (OR=1.97; 95% CI= 1.20–3.23), although this study was underpowered for the CRC phenotype (corrected P=0.18; for 25 tested cancer types)[41]. Given these underpowered and contradicting observations, the most recent NCCN guidelines for genetic and familial CRC syndromes (version 2.2017; released on August 9, 2017) concluded that the evidence supporting *ATM* as a CRC-risk gene is deficient and that the risk of CRC in *ATM* mutation carriers is largely unknown[12]. This is the first association study, to our knowledge, that confirmed and independently validated *ATM* as a moderately-penetrant CRC susceptibility gene, explaining the increased risk of colorectal cancer in around 0.74% of all unselected CRC cases. Furthermore, complete loss of *ATM* as a result of acquired deleterious somatic events suggesting a critical role of *ATM* in the CRC tumorigenesis in individuals with inherited *ATM* haploinsufficiency.

In addition to *ATM*, our analysis showed validated evidence supporting germline mutations in *PALB2* as CRC-risk events. *PALB2* plays a critical role in DNA homologous recombination by recruiting *BRCA2* and *RAD51* to DNA breaks to initiate DNA repair. Germline defects in *PALB2* have been associated with breast and pancreatic cancers[25, 42]. Although germline *PALB2* mutations have been observed in several CRC cohorts, it has been so far unclear wither these events contribute to the CRC risk or they merely represent coincidental findings. So far, there has not been any study to evaluate the role of *PALB2* mutations in CRC cases, hence *PALB2* has

not been part of the recent NCCN recommendations (version 2.2017) for germline testing in CRC[12]. Our analysis showed evidence for higher-than-expected germline pathogenic *PALB2* mutation rates in around 0.44% of unselected CRC cases, though this effect was not observed in early-onset CRC cohorts. Although tumors of individuals with germline mutations in *PALB2* did not show biallelic inactivation of the gene, our analysis however was not designed to capture potential pathogenic non-coding variants or epigenetic silencing events. Although *ATM* and *PALB2* may only explain a small fraction the CRC heritability in unselected cases, this represents a 20% increase in the diagnostic yield once these two genes are included.

Both *ATM* and *PALB2* are members of homologous recombination (HR) pathway which restores the integrity of double-strand DNA breaks[43]. Inherited HR gene mutations have long been known to increase the risk of several cancers, including breast, ovarian [MIM: 167000], prostate [MIM: 176807] and pancreatic cancers[23, 44, 45]. Here, we showed evidence that germline pathogenic mutations in the HR pathway genes, in aggregate, confer a relative 60-80% increase in the baseline risk of CRC. In addition, biallelic HR gene inactivation, observed in CRCs with various germline HR gene mutations in this study (particularly *ATM* mutation carriers), suggests new venues to explore targeted therapeutic intervention in CRC cases. Breast, ovarian, and prostate cancers from individuals with germline mutations in canonical HR genes have been shown to have substantial response to poly-ADP ribose polymerase (PARP) inhibitors and platinum-based chemotherapy, compared with mutation-negative individuals[46-48]. As preclinical studies have shown substantial sensitivity of the HR and *ATM*-deficient CRC cell lines to PARPi and with clinical trials to evaluate the efficacy of PARPi in CRC underway (NCT00912743, NCT02305758, NCT01589419, NCT02921256), universal screening of CRC cases for germline HR mutations may provide very informative data that could expand treatment options for these individuals[49].

The detection of mutations in actionable DRGs has significant ramifications for the probands and their families. First, these mutations significantly increase the person's risk of developing cancers other than CRC, for several of which effective screening options are available. Furthermore, identifying such mutations in an individual represents a unique opportunity to screen other family members to identify asymptomatic at-risk individuals and implement early surveillance measures. In total, our study estimates that approximately 2.1% (95% CI= 1.1%-3.4%; Binomial Exact) of all CRC cases carry actionable mutations in genes that have not been previously associated with increased CRC risk, which is significantly higher than the combined rate of these mutations in cancer-free controls. In addition, this small but significant subset of CRC cases are, as a result of being carriers of these mutations, at a substantially higher risk of developing several cancers other than CRC. Importantly, these actionable genes are not part of the recommended germline testing for individuals with CRC[12]. Consistent with prior observations in other tumor types, our analysis also demonstrated that positive family history of CRC or other malignancies could not be used as a proxy for the presence of germline DRGs

mutations, emphasizing the potential for broader molecular testing strategies to capture these clinically actionable events[50].

Offering clinical germline molecular testing to cancer cases to evaluate for an inherited cancer-predisposition syndrome relies heavily on several factors such as the individual's age of presentation and the presence of positive family history of cancer. Intriguingly, our analysis of large CRC cohorts showed that these factors may not reliably predict the likelihood of identifying a germline cancer predisposition mutation in individuals with CRC. First, except for individuals with germline high penetrance CRC risk mutations, our study showed that CRC individuals with low-penetrance CRC risk mutations and those with germline mutations in *ATM* or *PALB2* were not more likely to present at an earlier age compared with presumed sporadic cases. In addition, our study showed that positive family history of CRC was not more commonly reported in CRC individuals who carried high-penetrance CRC risk mutations, low-penetrance CRC risk mutations or DNA repair gene mutations. This is consistent with prior similar observations in the prostate and pediatric cancer spaces[50, 51]. These findings underscore the importance of considering the possibility of carrying an inherited CRC-risk mutation in individuals with late-onset CRC as well as in those without strong family history of CRC. In addition, these observations are also relevant when evaluating the potential utility of implementing early CRC screening measures. However, larger studies are still needed to further delineate the penetrance of these germline mutations.

Our study has several limitations. First, although we performed population stratification, our cases and controls did not come from the same cohort, so enrichment of mutations secondary to non-CRC related factors cannot be completely ruled out. Also, since the raw sequencing data of the control cohort (ExAC) are not publically available, germline variants in cases and controls were not jointly called to limit potential sequencing or pipeline-related variant calling biases. We, however, mitigated this potential source of bias by using the same parameters, tools and platforms that were used to analyze the ExAC cohort. In addition, individual-level clinical information on our control group as well as the validation sets were not available which limited our ability to correct for potential confounders. However, evaluating several independent CRC cohorts makes it unlikely for a confounder to be shared across all cohorts. Finally, larger case-control studies are still necessary to confirm these clinically-relevant findings and inform future updates of clinical germline testing guidelines in CRC cases.

Broadly, our study of large CRC cohorts showed enrichment of disruptive germline pathogenic mutations in the homologous recombination pathway, suggesting its important role in CRC susceptibility and management. In addition, we presented evidence to support *ATM* and *PALB2* as new CRC susceptibility genes, explaining the missing CRC heritability in 1.2% of unselected CRC cases. We also illustrated that a relatively large proportion of all CRC cases have germline pathogenic mutations in HR genes, which may greatly impact their clinical care and inform

molecularly driven treatment strategies for individuals with mutations in these genes. Finally, since these genes are not routinely tested clinically, these results could inform revisions to CRC testing guidelines.

Acknowledgements

We thank all individuals who participated in this study. We also thank Dr. Michele Hacker for her advice on the statistical analysis of this study. We would also like to thank the participants and staff of the NHS and HPFS for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. Drs. Van Allen, Ogino, Garraway and Fuchs had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

This work was conducted with support from Harvard Catalyst, the Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award UL1 TR001102) and financial contributions from Harvard University and its affiliated academic healthcare centers. The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic healthcare centers, or the National Institutes of Health. This work was also supported by K08 CA188615-02 (E.M.V.), Damon Runyon Clinical Investigator Award (E.M.V.), KL2 TR001100 (M.G.), R01CA169141-04 (C.S.F.), R01CA118553-07 (C.S.F.), P01 CA87969; UM1 CA186107; P01 CA55075; UM1 CA167552, R35 CA197735 (S.O) and the Stand Up to Cancer Colorectal Cancer Dream Team Translational Research Grant (Grant Number SU2C-AACR-DT22-17) (M.G., C.S.F.). Stand Up to Cancer is a program of the Entertainment Industry Foundation and the research grant is administered by the American Association for Cancer Research, a scientific partner of SU2C. N.D.M. is a Howard Hughes Medical Institute Medical Research Fellow. The funding organizations were not responsible for design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Availability of data and materials

All BAM files of the CanSeq study are deposited in dbGap phs001075.v1.p1. All raw sequencing files of the NHS/HPFS study are deposited in dbGap phs000722. The TCGA data is available from the database of Genotypes and Phenotypes (dbGaP), Study Accession: phs000178.v9.p8. Raw sequencing data of the NSCCG were not available for analysis, though downstream variant data can be accessed from the "CanVar browser" (https://canvar.icr.ac.uk/).

Competing interests

Dr. Van Allen is an advisor to Genome Medical and consultant to Invitae. No other competing interests. Dr. Syngal is a consultant to Myriad Genetics.

Ethics approval and consent to participate

All individuals in the CanSeq study consented to an institutional review board-approved protocol that allows comprehensive genetic analysis of tumor and germline samples (Dana-Farber Cancer Institute #12-078). The NHS/HPFS study was approved by the Partners (IRB#2012-P000788). This study conforms to the Declaration of Helsinki.

Web Resources section:

Online Mendelian Inheritance in Man (http://www.omim.org). Exome Aggregation Consortium (http://exac.broadinstitute.org/). The Cancer Variation Resource (https://canvar.icr.ac.uk/). ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/).

Uncategorized References

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer Statistics, 2017.* CA Cancer J Clin, 2017. **67**(1): p. 7-30.
2. Mucci, L.A., et al., *Familial Risk and Heritability of Cancer Among Twins in Nordic Countries.* JAMA, 2016. **315**(1): p. 68-76.
3. Fishel, R., et al., *The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer.* Cell, 1994. **77**(1): p. 1 p following 166.
4. Bronner, C.E., et al., *Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer.* Nature, 1994. **368**(6468): p. 258-61.
5. Al-Tassan, N., et al., *Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors.* Nat Genet, 2002. **30**(2): p. 227-32.
6. Yurgelun, M.B., et al., *Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome.* Gastroenterology, 2015. **149**(3): p. 604-13 e20.
7. Lynch, H.T. and A. de la Chapelle, *Hereditary colorectal cancer.* N Engl J Med, 2003. **348**(10): p. 919-32.
8. Susswein, L.R., et al., *Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing.* Genet Med, 2016. **18**(8): p. 823-32.
9. Wang, M., et al., *Genetic testing for Lynch syndrome in the province of Ontario.* Cancer, 2016. **122**(11): p. 1672-9.
10. Pearlman, R., et al., *Prevalence and Spectrum of Germline Cancer Susceptibility Gene Mutations Among Patients With Early-Onset Colorectal Cancer.* JAMA Oncol, 2016.
11. Thompson, D., et al., *Cancer risks and mortality in heterozygous ATM mutation carriers.* J Natl Cancer Inst, 2005. **97**(11): p. 813-22.
12. *National Comprehensive Cancer Network (NCCN): Familial and Genetic High-Risk Assessment: Colorectal*. 2017. **Version 2.2017**, 96.
13. Giannakis, M., et al., *Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma.* Cell Rep, 2016.
14. Ghazani, A.A., et al., *Assigning clinical meaning to somatic and germ-line whole-exome sequencing data in a prospective cancer precision medicine study.* Genet Med, 2017.
15. Gray, S.W., et al., *Oncologists' and cancer patients' views on whole-exome sequencing and incidental findings: results from the CanSeq study.* Genet Med, 2016.
16. Yurgelun, M.B., et al., *Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer.* J Clin Oncol, 2017: p. JCO2016710012.
17. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
18. Chubb, D., et al., *Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer.* Nat Commun, 2016. **7**: p. 11883.
19. Van Allen, E.M., et al., *Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine.* Nat Med, 2014. **20**(6): p. 682-8.

20. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D457-62.
21. van Os, N.J., et al., *Health risks for ataxia-telangiectasia mutated heterozygotes: a systematic review, meta-analysis and evidence-based guideline.* Clin Genet, 2016. **90**(2): p. 105-17.
22. Goldgar, D.E., et al., *Rare variants in the ATM gene and risk of breast cancer.* Breast Cancer Res, 2011. **13**(4): p. R73.
23. Mavaddat, N., et al., *Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE.* J Natl Cancer Inst, 2013. **105**(11): p. 812-22.
24. Rafnar, T., et al., *Mutations in BRIP1 confer high risk of ovarian cancer.* Nat Genet, 2011. **43**(11): p. 1104-7.
25. Antoniou, A.C., W.D. Foulkes, and M. Tischkowitz, *Breast-cancer risk in families with mutations in PALB2.* N Engl J Med, 2014. **371**(17): p. 1651-2.
26. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* Genet Med, 2015. **17**(5): p. 405-24.
27. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
28. Barnholtz-Sloan, J.S., et al., *Ancestry estimation and correction for population stratification in molecular epidemiologic association studies.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(3): p. 471-7.
29. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
30. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
31. Saunders, C.T., et al., *Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs.* Bioinformatics, 2012. **28**(14): p. 1811-7.
32. Costello, M., et al., *Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation.* Nucleic Acids Res, 2013. **41**(6): p. e67.
33. Ramos, A.H., et al., *Oncotator: cancer variant annotation tool.* Hum Mutat, 2015. **36**(4): p. E2423-9.
34. Carter, S.L., et al., *Absolute quantification of somatic DNA alterations in human cancer.* Nat Biotechnol, 2012. **30**(5): p. 413-21.
35. Wang, H., et al., *Optimal two-stage genotyping designs for genome-wide association scans.* Genet Epidemiol, 2006. **30**(4): p. 356-68.
36. Guo, X. and R.C. Elston, *One-stage versus two-stage strategies for genome scans.* Adv Genet, 2001. **42**: p. 459-71.
37. Jeffrey C. Barrett, J.B., David Cutler, Mark Daly, Bernie Devlin, Jacob Gratten, Matthew E. Hurles, Jack A. Kosmicki, Eric S. Lander, Daniel G. MacArthur, Benjamin M. Neale, Kathryn Roeder, Peter M. Visscher, Naomi R. Wray, *New mutations, old statistical challenges.* BioRxiv, 2017.
38. Yurgelun, M.B., et al., *Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer.* J Clin Oncol, 2017. **35**(10): p. 1086-1095.

39. Mersha, T.B. and T. Abebe, *Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities.* Hum Genomics, 2015. **9**: p. 1.
40. Olsen, J.H., et al., *Breast and other cancers in 1445 blood relatives of 75 Nordic patients with ataxia telangiectasia.* Br J Cancer, 2005. **93**(2): p. 260-5.
41. Helgason, H., et al., *Loss-of-function variants in ATM confer risk of gastric cancer.* Nat Genet, 2015. **47**(8): p. 906-10.
42. Zhen, D.B., et al., *BRCA1, BRCA2, PALB2, and CDKN2A mutations in familial pancreatic cancer: a PACGENE study.* Genet Med, 2015. **17**(7): p. 569-77.
43. Mladenov, E., et al., *DNA double-strand-break repair in higher eukaryotes and its role in genomic instability and cancer: Cell cycle and proliferation-dependent regulation.* Semin Cancer Biol, 2016. **37-38**: p. 51-64.
44. Ford, D., et al., *Risks of cancer in BRCA1-mutation carriers. Breast Cancer Linkage Consortium.* Lancet, 1994. **343**(8899): p. 692-5.
45. Iqbal, J., et al., *The incidence of pancreatic cancer in BRCA1 and BRCA2 mutation carriers.* Br J Cancer, 2012. **107**(12): p. 2005-9.
46. Mateo, J., et al., *DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer.* N Engl J Med, 2015. **373**(18): p. 1697-708.
47. Fong, P.C., et al., *Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers.* N Engl J Med, 2009. **361**(2): p. 123-34.
48. Tutt, A., et al., *Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial.* Lancet, 2010. **376**(9737): p. 235-44.
49. Wang, C., et al., *ATM-Deficient Colorectal Cancer Cells Are Sensitive to the PARP Inhibitor Olaparib.* Transl Oncol, 2017. **10**(2): p. 190-196.
50. Pritchard, C.C., et al., *Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer.* N Engl J Med, 2016. **375**(5): p. 443-53.
51. Zhang, J., et al., *Germline Mutations in Predisposition Genes in Pediatric Cancer.* N Engl J Med, 2015. **373**(24): p. 2336-2346.

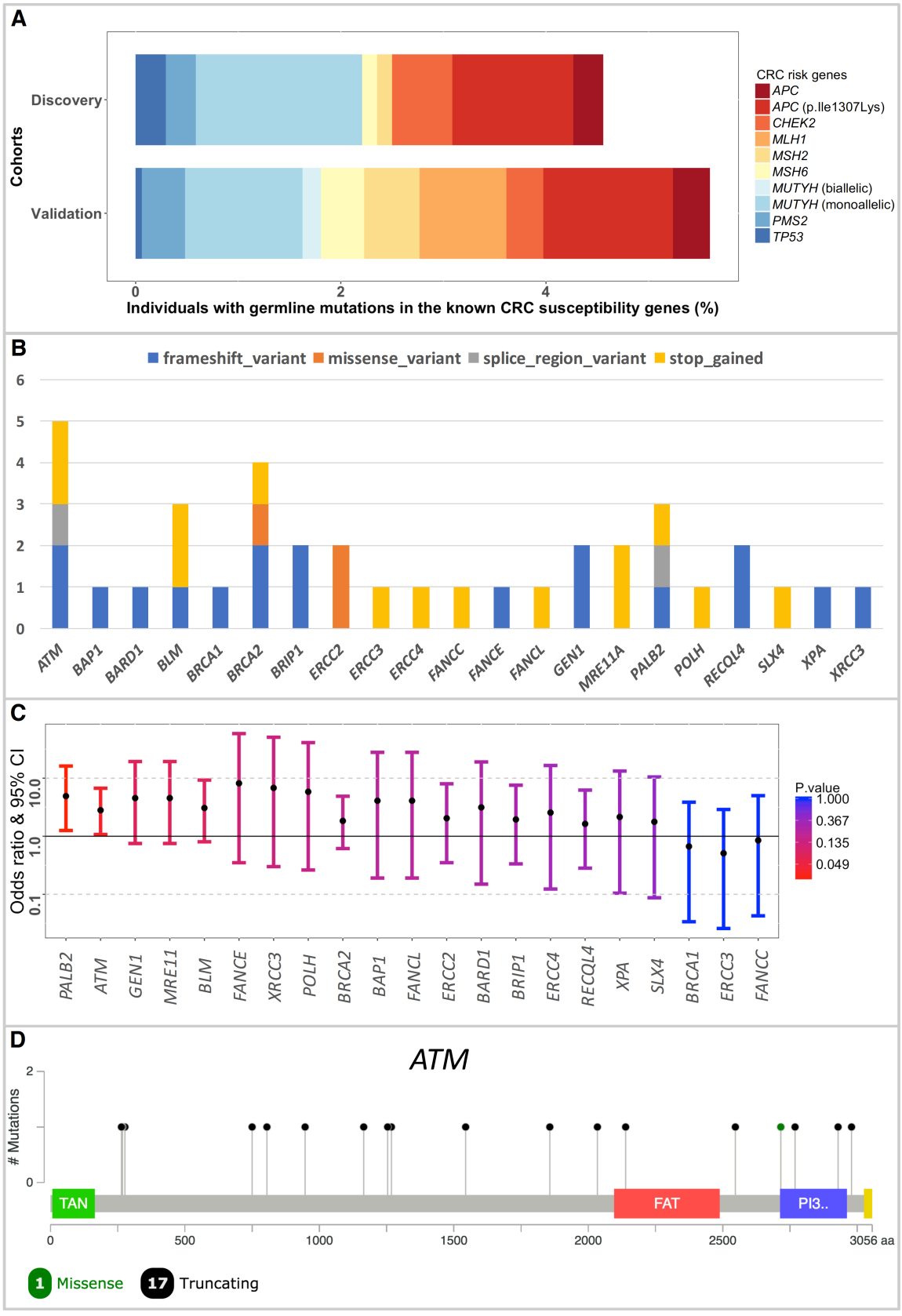
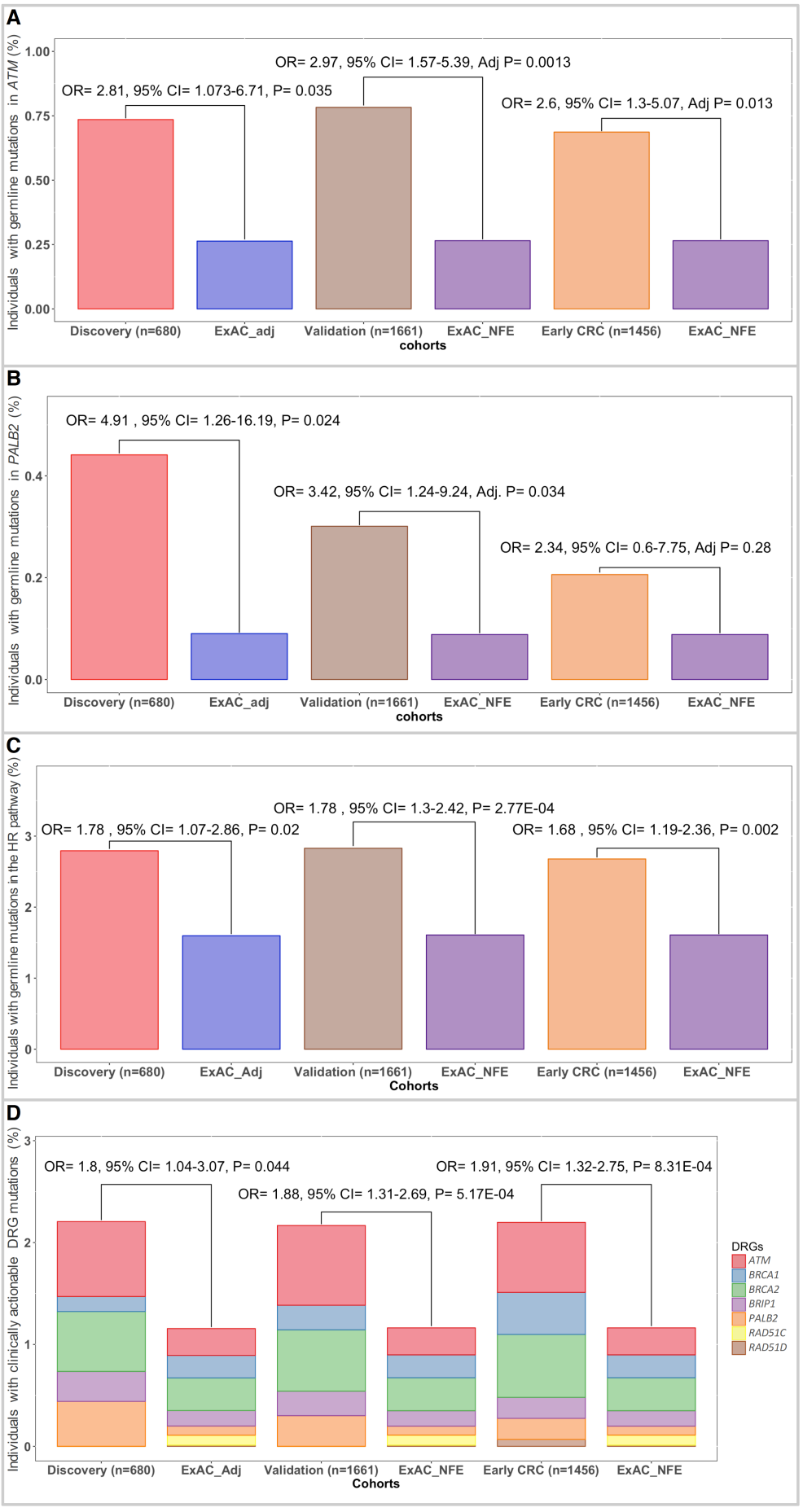


Figure 1: Germline pathogenic mutations in the known CRC predisposition genes and additional DNA repair genes. A; Proportions of cases with germline pathogenic mutations in the CRC risk genes in 680 CRC individuals in the discovery set and 1661 CRC cases in the validation set. B; Number and class of the detected germline pathogenic mutations in the DRGs in the discovery set (n=680). DRGs where no mutations were detected (n=19) are not shown here. C; Enrichment of germline pathogenic DRGs mutations in 680 CRC individuals in the discovery set. Fisher's exact test was used to calculate the ORs and 95% confidence intervals. Two-sided binomial test was used to calculate the P values. D; A total of 18 germline pathogenic *ATM* mutations were seen in the discovery and validation sets in our study. This includes seven (38.9%) nonsense mutations, six (33.3%) frameshift mutations, three (16.6%) splice-site mutations, one (5.6%) known pathogenic in-frame deletion and one (5.6%) known pathogenic missense mutation.

Figure 2: Enrichment of DRG mutations in various cohorts. A; Inherited pathogenic germline mutations in *ATM* were more commonly seen in individuals with CRC in the discovery, validation and early-onset CRC sets (n=680; n=1661, n=1456, respectively) compared with cancer-free individuals. B; Germline pathogenic mutations in *PALB2* were significantly enriched in unselected CRC cases from the discovery and validation sets. However, no significant enrichment was seen in the early-onset CRC cases. C; A secondary analysis of the homologous recombination pathway showed significant enrichment of germline HR gene mutations, as an aggregate, in all CRC cohorts. D; Individuals with CRC were also almost twice more likely to carry a clinically actionable mutation where screening recommendation do exist and which can greatly impact the clinical care offered to these individuals and their families.



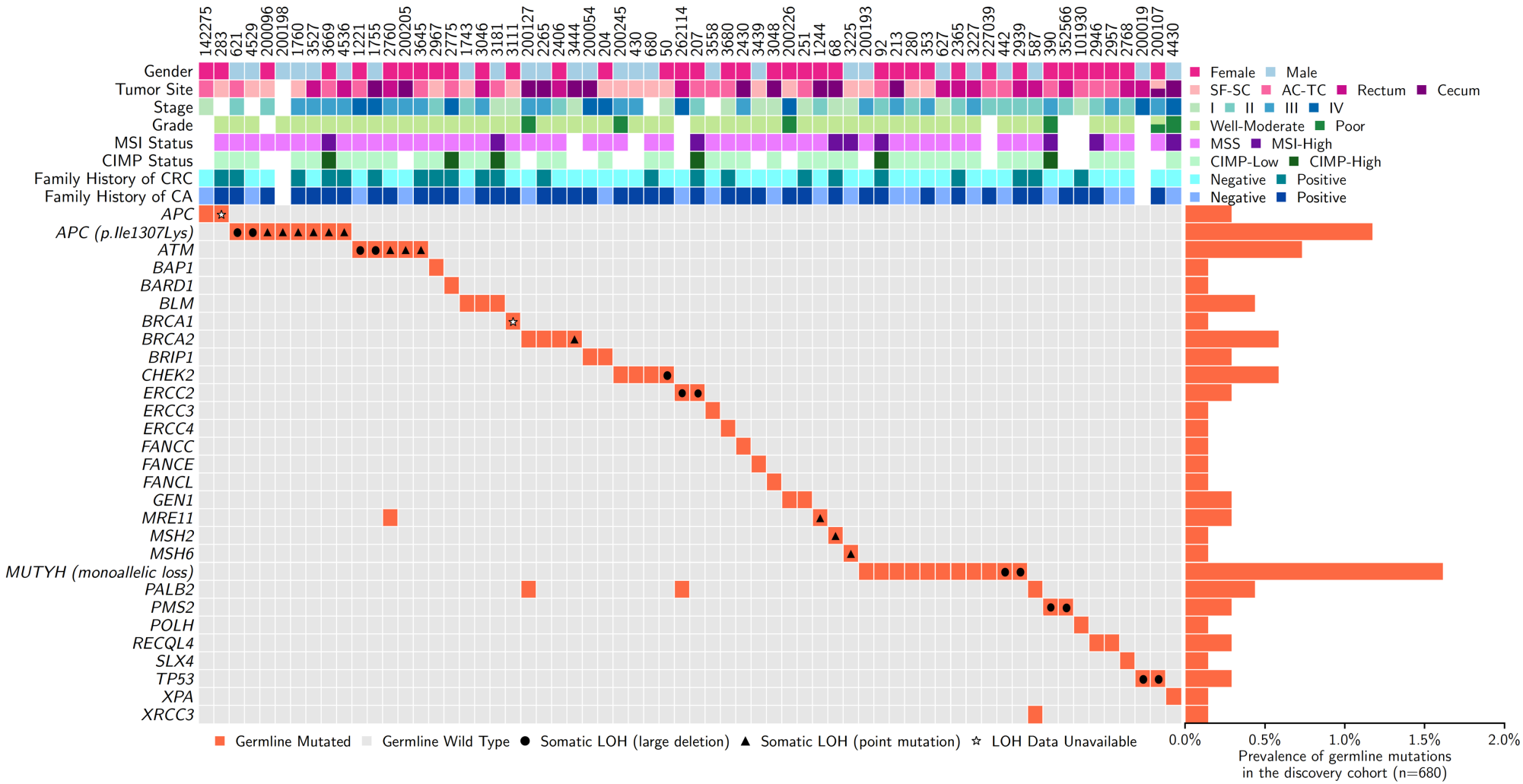


Figure 3: Clinical and molecular characteristics of all cases with germline pathogenic mutations in CRC risk genes and DRGs in our discovery set. All individuals with germline pathogenic mutations in *ATM* had somatic LOH in their tumor samples. Two of these cases had large deletions that affected the wild-type *ATM* allele while three had truncating point mutations leading to the loss of *ATM* wild-type allele as well. (AC-TC: ascending colon to transverse colon; SF-SC: splenic flexure to sigmoid colon; MSI: microsatellite instability; MSS: microsatellite stable; CIMP: CpG island methylator phenotype-specific promoters; LOH: loss of heterozygosity).

Table 1: Clinical, pathological, and molecular characteristics of 680 colorectal cancer cases who were examined in the discovery set.

Characteristic[a]	All cases (N=680)	Mutations in known CRC susceptibility genes (high penetrance)[b,c] Absent (N=616)	Present (N=12)	P^g	Mutations in known CRC susceptibility genes (low penetrance)[c,d] Absent (N=616)	Present (N=19)	P^g	Mutations in DNA repair genes[c,e] Absent (N=616)	Present (N=33)	P^g	Mutations in the homologous recombination pathway[c,f] Absent (N=616)	Present (N=19)	P^g	Mutations in *ATM*[c] Absent (N=616)	Present (N=5)	P^g	Mutations in *PALB2*[c] Absent (N=616)	Present (N=3)	P^g
Sex				0.99			0.24			0.59			0.81			0.65			0.56
Female	414 (61%)	376 (61%)	7 (58%)		376 (61%)	9 (47%)		376 (61%)	22 (67%)		376 (61%)	11 (58%)		376 (61%)	4 (80%)		376 (61%)	1 (33%)	
Male	266 (39%)	240 (39%)	5 (42%)		240 (39%)	10 (53%)		240 (39%)	11 (33%)		240 (39%)	8 (42%)		240 (39%)	1 (20%)		240 (39%)	2 (67%)	
Mean age ±SD (years)	68.8±10.3	68.9±10.2	58.4±13.8	0.0005	68.9±10.2	72.2±6.2	0.16	68.9±10.2	69.7±10.8	0.66	68.9±10.2	68.2±10.4	0.77	68.9±10.2	75.6±6.6	0.14	68.9±10.2	64.7±18.8	0.47
Missing	12	12	0		12	0		12	0		12	0		12	0		12	0	
Race/ethnicity				0.99			0.99			0.50			0.33			0.10			0.99
White	667 (98%)	604 (98%)	12 (100%)		604 (98%)	19 (100%)		604 (98%)	32 (97%)		604 (98%)	18 (95%)		604 (98%)	4 (80%)		604 (98%)	3 (100%)	
Black	13 (1.9%)	12 (2.0%)	0		12 (2.0%)	0		12 (2.0%)	1 (3.0%)		12 (2.0%)	1 (5.3%)		12 (2.0%)	1 (20%)		12 (2.0%)	0	
Ashkenazi Jewish				0.99			0.0015			0.59			0.99			0.99			0.99
No	155 (86%)	144 (88%)	2 (100%)		144 (88%)	3 (38%)		144 (88%)	6 (86%)		144 (88%)	4 (100%)		144 (88%)	1 (100%)		144 (88%)	1 (100%)	
Yes	25 (14%)	19 (12%)	0		19 (12%)	5 (62%)		19 (12%)	1 (14%)		19 (12%)	0		19 (12%)	0		19 (12%)	0	
Missing	500	453	10		453	11		453	26		453	15		453	4		453	2	
Family history of colorectal cancer in first-degree relative(s)				0.73			0.16			0.099			0.18			0.34			0.55
Absent	501 (75%)	461 (76%)	8 (73%)		461 (76%)	11 (61%)		461 (76%)	21 (64%)		461 (76%)	12 (63%)		461 (76%)	3 (60%)		461 (76%)	2 (67%)	
Present	164 (25%)	142 (24%)	3 (27%)		142 (24%)	7 (39%)		142 (24%)	12 (36%)		142 (24%)	7 (37%)		142 (24%)	2 (40%)		142 (24%)	1 (33%)	
Missing	15	13	1		13	1		13	0		13	0		13	0		13	0	
Family history of breast cancer in first-degree relative(s)				0.044			0.99			0.18			0.27			0.55			0.99
Absent	359 (81%)	329 (82%)	4 (50%)		329 (82%)	9 (90%)		329 (82%)	17 (71%)		329 (82%)	9 (69%)		329 (82%)	3 (75%)		329 (82%)	2 (100%)	
Present	85 (19%)	73 (18%)	4 (50%)		73 (18%)	1 (10%)		73 (18%)	7 (29%)		73 (18%)	4 (31%)		73 (18%)	1 (25%)		73 (18%)	0	
Missing	236	214	4		214	9		214	9		214	6		214	1		214	1	
Family history of ovarian cancer in first-degree relative(s)				0.99			0.99			0.99			0.99			0.99			0.99
Absent	425 (96%)	384 (96%)	8 (100%)		384 (96%)	10 (100%)		384 (96%)	23 (96%)		384 (96%)	13 (100%)		384 (96%)	4 (100%)		384 (96%)	2 (100%)	
Present	19 (4.3%)	18 (4.5%)	0		18 (4.5%)	0		18 (4.5%)	1 (4.2%)		18 (4.5%)	0		18 (4.5%)	0		18 (4.5%)	0	
Missing	236	214	4		214	9		214	9		214	6		214	1		214	1	
Family history of any cancer in first-degree relative(s)				0.54			0.63			0.72			0.81			0.99			0.57
Absent	270 (41%)	249 (41%)	3 (27%)		249 (41%)	6 (33%)		249 (41%)	12 (36%)		249 (41%)	7 (37%)		249 (41%)	2 (40%)		249 (41%)	2 (67%)	
Present	395 (59%)	354 (59%)	8 (73%)		354 (59%)	12 (67%)		354 (59%)	21 (64%)		354 (59%)	12 (63%)		354 (59%)	3 (60%)		354 (59%)	1 (33%)	
Missing	15	13	1		13	1		13	0		13	0		13	0		13	0	
Tumor location[h]				0.48			0.34			0.31			0.18			0.31			0.062

Cecum	129 (19%)	117 (19%)	1 (9.1%)		117 (19%)	1 (5.6%)		117 (19%)	10 (30%)		117 (19%)	7 (37%)		117 (19%)	2 (40%)		117 (19%)	1 (33%)								
Ascending-transverse colon	202 (30%)	184 (30%)	2 (18%)		184 (30%)	5 (28%)		184 (30%)	11 (33%)		184 (30%)	4 (21%)		184 (30%)	2 (40%)		184 (30%)	0								
Splenic flexure-sigmoid colon	201 (30%)	183 (30%)	6 (55%)		183 (30%)	6 (33%)		183 (30%)	6 (18%)		183 (30%)	3 (16%)		183 (30%)	0		183 (30%)	0								
Rectum	136 (20%)	122 (20%)	2 (18%)		122 (20%)	6 (33%)		122 (20%)	6 (18%)		122 (20%)	5 (26%)		122 (20%)	1 (20%)		122 (20%)	2 (67%)								
Missing	12	10	1		10	1		10	0		10	0		10	0		10	0								
Tumor differentiation				0.062			0.40			0.99			0.99			0.99			0.20							
Well to moderate	534 (90%)	483 (90%)	6 (67%)		483 (90%)	17 (100%)		483 (90%)	28 (90%)		483 (90%)	17 (94%)		483 (90%)	5 (100%)		483 (90%)	1 (50%)								
Poor	62 (10%)	56 (10%)	3 (33%)		56 (10%)	0		56 (10%)	3 (9.7%)		56 (10%)	1 (5.6%)		56 (10%)	0		56 (10%)	1 (50%)								
Missing	84	77	3		77	2		77	2		77	1		77	0		77	1								
AJCC disease stage				0.35			0.020			0.74			0.40			0.051			0.88							
I	148 (24%)	135 (24%)	4 (40%)		135 (24%)	1 (5.9%)		135 (24%)	8 (25%)		135 (24%)	3 (16%)		135 (24%)	0		135 (24%)	0								
II	188 (30%)	171 (30%)	1 (10%)		171 (30%)	8 (47%)		171 (30%)	8 (25%)		171 (30%)	4 (21%)		171 (30%)	0		171 (30%)	1 (33%)								
III	177 (28%)	157 (28%)	4 (40%)		157 (28%)	8 (47%)		157 (28%)	8 (25%)		157 (28%)	6 (32%)		157 (28%)	3 (60%)		157 (28%)	1 (33%)								
IV	110 (18%)	101 (18%)	1 (10%)		101 (18%)	0		101 (18%)	8 (25%)		101 (18%)	6 (32%)		101 (18%)	2 (40%)		101 (18%)	1 (33%)								
Missing	57	52	2		52	2		52	1		52	0		52	0		52	0								
MSI status				0.13			0.75			0.80			0.50			0.99			0.99							
MSS/MSI-low	475 (84%)	428 (84%)	5 (62%)		428 (84%)	16 (89%)		428 (84%)	26 (87%)		428 (84%)	16 (94%)		428 (84%)	5 (100%)		428 (84%)	2 (100%)								
MSI-high	92 (16%)	83 (16%)	3 (38%)		83 (16%)	2 (11%)		83 (16%)	4 (13%)		83 (16%)	1 (5.9%)		83 (16%)	0		83 (16%)	0								
Missing	113	105	4		105	1		105	3		105	2		105	0		105	1								
CIMP status				0.99			0.75			0.44			0.74			0.59			0.99							
CIMP-low/negative	382 (80%)	342 (79%)	5 (83%)		342 (79%)	13 (87%)		342 (79%)	22 (88%)		342 (79%)	12 (86%)		342 (79%)	4 (100%)		342 (79%)	1 (100%)								
CIMP-high	95 (20%)	89 (21%)	1 (17%)		89 (21%)	2 (13%)		89 (21%)	3 (12%)		89 (21%)	2 (14%)		89 (21%)	0		89 (21%)	0								
Missing	203	185	6		185	4		185	8		185	5		185	1		185	2								

a Percentage indicates the proportion of cases with a specific clinical, pathological, or molecular characteristic in all cases or in strata of germline pathogenic mutations.

b High penetrance CRC risk genes include: *APC* (excluding p.I1307K), *BMPR1A, CHEK2, MLH1, MSH2, MSH6, MUTYH* (biallelic inactivation), *PMS2, POLD1, POLE, PTEN, SMAD4, STK11, TP53*

c Individuals who had mutations in the other CRC risk genes or DNA repair genes (DRGs) were excluded.

d Low penetrance CRC risk mutations include: *APC* p.I1307K, and monoallelic inactivation of *MUTYH*

e This gene set includes 40 DNA repair genes listed in Table S2

f Homologous recombination DNA repair genes included in this analysis are: *ATM, BARD1, BLM, BRCA1, BRCA2, BRIP1 , MRE11, NBN , PALB2, RAD51 , RAD51C , RAD51D , RAD54L ,* and *XRCC3.*

g To compare characteristics between subgroups according to the germline mutation status, Fisher's exact test was used for categorical variables while unpaired t-test was used for continuous variables.

h One case who had two lesions (cecum and sigmoid colon) was excluded from the analysis.

AJCC, American Joint Committee on Cancer; CIMP, CpG island methylator phenotype-specific promoters; MSI, microsatellite instability; MSS, microsatellite stable; SD, standard deviation.

Table 2: Enrichment of germline pathogenic mutations in 680 CRC cases (discovery set) relative to 27728 ancestry-matched cancer-free adults from the ExAC cohort. Only genes with detected germline pathogenic mutations in cases are shown. (ExAC: Exome Aggregation Consortium)

Gene	Cases with mutations in the discovery cohort (n= 680)	Prevalence of cases with mutations in the discovery cohort (%)	Cases with mutations in ancestry-matched control group (n=27728)	Prevalence of mutations in the control group (%)	Enrichment of pathogenic mutations in the discovery cohort (OR; Fisher's Exact test)	95% Confidence Intervals (Fisher's Exact test)	P value (two-sided Exact Binomial test)
ATM	5	0.74%	73	0.26%	2.81	1.07-6.71	0.035
BAP1	1	0.15%	10	0.04%	4.08	0.19-27.85	0.218
BARD1	1	0.15%	13	0.05%	3.14	0.15-19.04	0.273
BLM	3	0.44%	40	0.14%	3.07	0.8-9.28	0.077
BRCA1	1	0.15%	61	0.22%	0.67	0.03-3.86	1
BRCA2	4	0.59%	89	0.32%	1.84	0.61-4.89	0.177
BRIP1	2	0.29%	42	0.15%	1.94	0.33-7.57	0.275
ERCC2	2	0.29%	40	0.14%	2.04	0.35-8.00	0.25
ERCC3	1	0.15%	80	0.29%	0.51	0.03-2.89	1
ERCC4	1	0.15%	16	0.06%	2.55	0.12-16.55	0.325
FANCC	1	0.15%	48	0.17%	0.85	0.04-5.0	1
FANCE	1	0.15%	5	0.02%	8.16	0.35-58.6	0.115
FANCL	1	0.15%	10	0.04%	4.08	0.19-27.85	0.218
GEN1	2	0.29%	18	0.06%	4.54	0.75-19.35	0.073
MRE11	2	0.29%	18	0.06%	4.54	0.75-19.35	0.073
PALB2	3	0.44%	25	0.09%	4.91	1.26-16.19	0.024
POLH	1	0.15%	7	0.03%	5.83	0.26-40.98	0.158
RECQL4	2	0.29%	50	0.18%	1.63	0.28-6.23	0.347
SLX4	1	0.15%	23	0.08%	1.77	0.09-10.49	0.431
XPA	1	0.15%	19	0.07%	2.15	0.1-13.26	0.373
XRCC3	1	0.15%	6	0.02%	6.8	0.3-50.67	0.137

Supplementary figures:

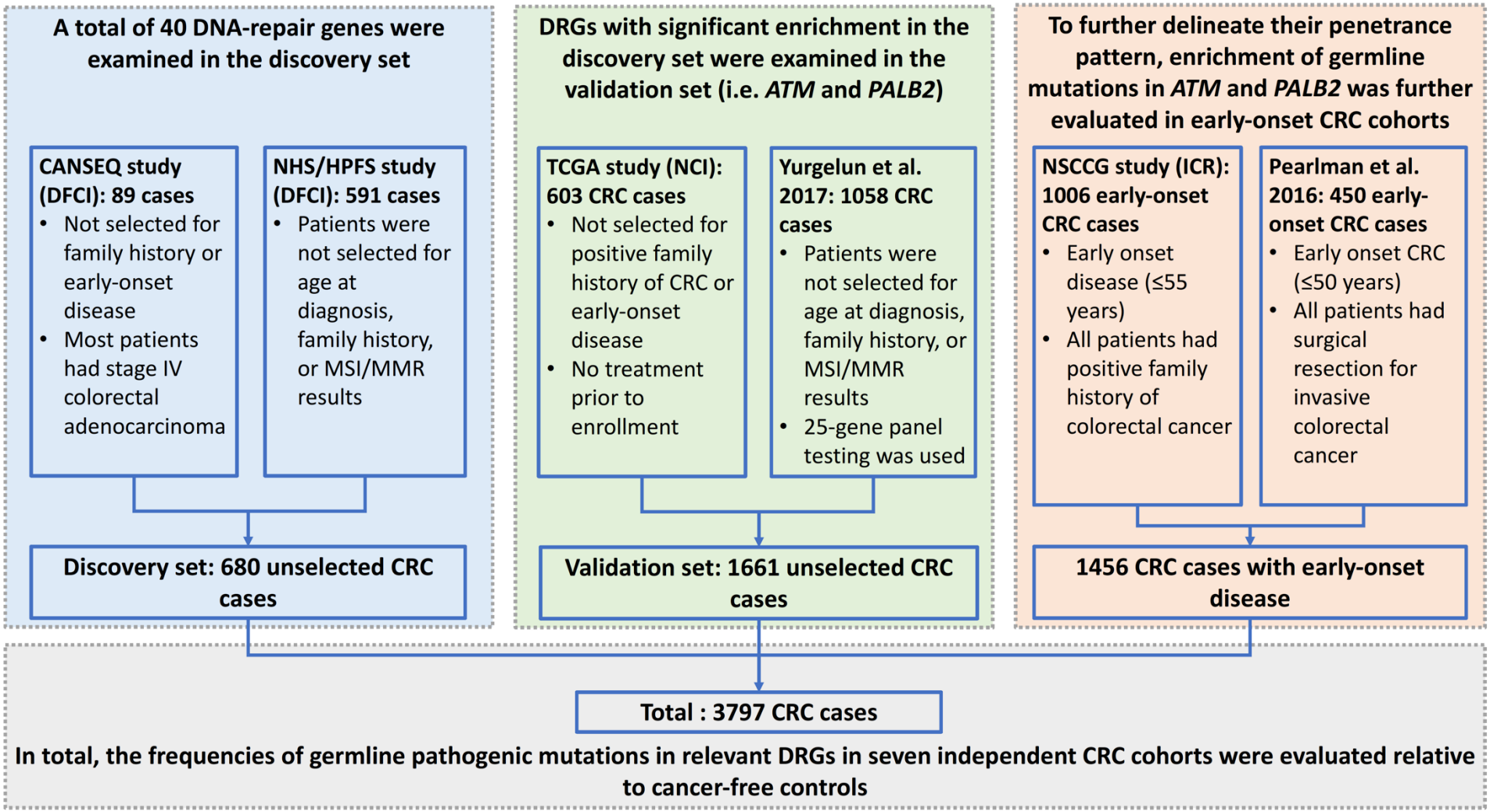


Figure S1: Various cohorts examined in the discovery and validation phases of this study. Two independent cohorts that included 680 CRC individuals were examined in the discovery phase. Of these, a total of 591 CRC cases came from the population-based Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS). In addition, 89 CRC cases from the CanSeq study at Dana-Farber Cancer Institute (DFCI) were included in the discovery set. In the validation phase, germline data of 1661 individuals from two independent CRC cohorts were evaluated. Of

those, 603 CRC individuals were included in the TCGA project. Individuals in the TCGA cohort were not selected for early-onset disease or positive family history. Germline variants of another 1058 unselected CRC cases who were recently described by Yurgelun et al. were also included in the validation set. Significant findings in the unselected CRC discovery and validation sets were also evaluated in 1456 early-onset CRC cases. In the early-onset CRC set, publically-available germline calls of 1006 early-onset (age<56) familial CRC cases, enrolled in the National Study of Colorectal Cancer Genetics (NSCCG), were examined. Raw sequencing data of the NSCCG were not available for analysis, though downstream variant data was accessed from the "CanVar browser" (https://canvar.icr.ac.uk/; accessed on December 15, 2016). The early-onset CRC set also included 450 CRC individuals who were diagnosed with CRC before the age of 50. The germline variants in these cases were recently described by Pearlman et al, 2017. Raw germline sequencing data of these cohorts were not available for examination. Only germline variants that have been reported in these studies were evaluated. (NHS: Nurses' Health Study; HPFS: Health Professional Follow Study; TCGA: The Cancer Genome Atlas; NSCCG: National Study of Colorectal Cancer Genetics; ICR: Institute of Cancer Research)

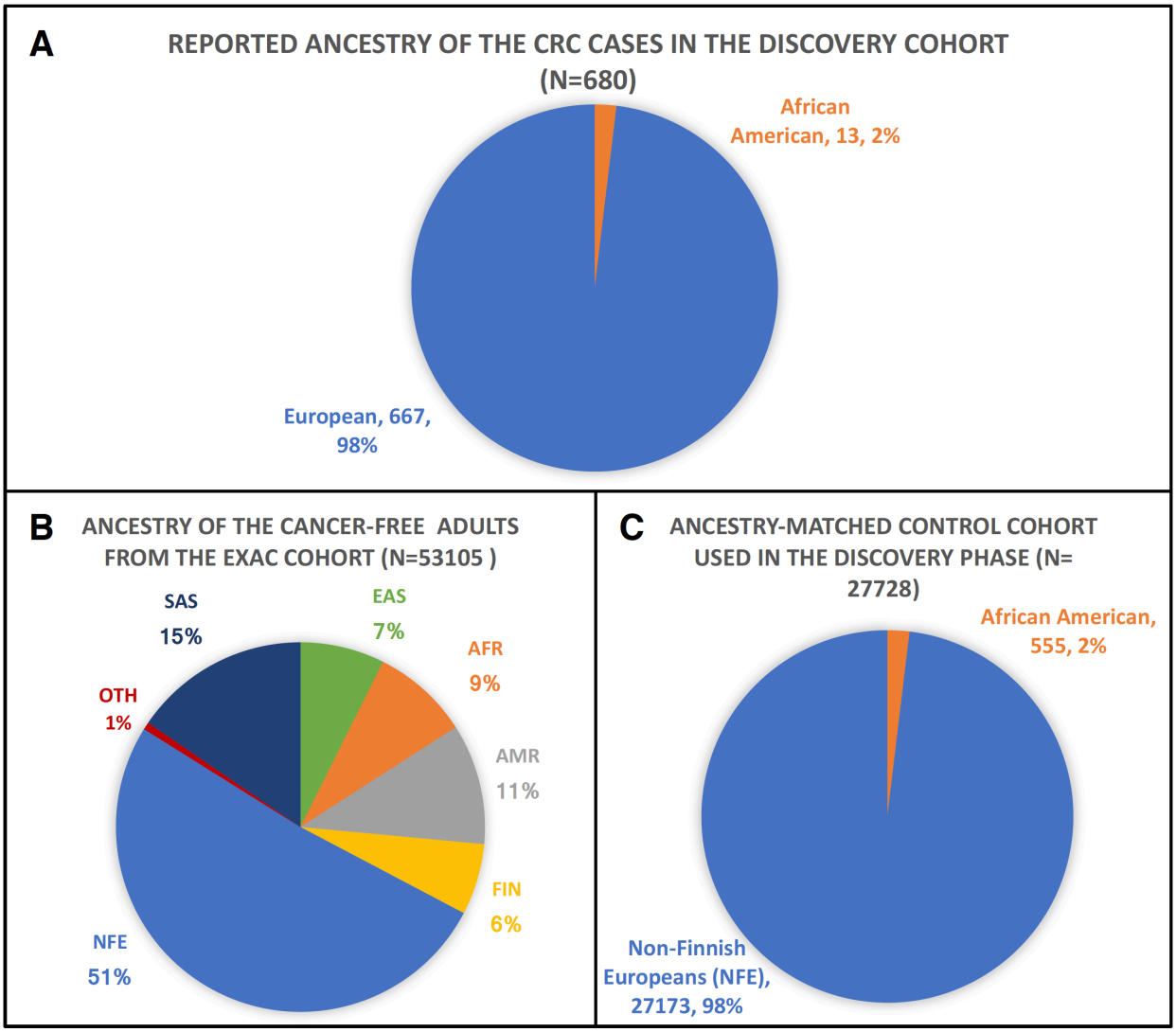


Figure S2: Proportions of cases and controls examined in the discovery phase of this study. A; most of the CRC cases in the discovery set of this study identified their ancestry as European. B&C; Rates of germline pathogenic mutations in the examined DRGs were calculated for each of the continental populations reported in the Exome Aggregation Consortium (ExAC) database (African & African American (n=4533), American (n=5608), East Asian (n=3933), Finnish (n=3307), Non-Finnish European (n=27173), South Asian (n=8204)). Based on the proportion of self-reported ancestry representation in our discovery cohort (98% European and 2% African American), ancestry-adjusted frequencies for disruptive mutations in the genes of interest were calculated as follows: Ancestry-adjusted frequency= (0.98 X gene-based frequency of germline pathogenic mutations in NFE) + (0.02 X gene-based frequency of germline pathogenic mutations in AFR). In addition to using ancestry-adjusted rates of mutations as reference values to calculate the significance of enrichment (using Binomial Exact test), we calculated the effect size of enrichment by constructing an ethnicity-matched control cohort (referred to as ExAC_Adj in this

study) that constitutes of 27728 individuals (98%; 27173 Non-Finnish Europeans (NFE), and 2%; 555 African Americans (AFR)). Expected number of germline pathogenic mutations in the ancestry-adjusted control cohort in each gene was calculated using the ancestry-adjusted frequency. (AFR: African & African American, AMR: American, EAS: East Asian, FIN: Finnish, NFE: Non-Finnish European, SAS: South Asian, OTH: Other).

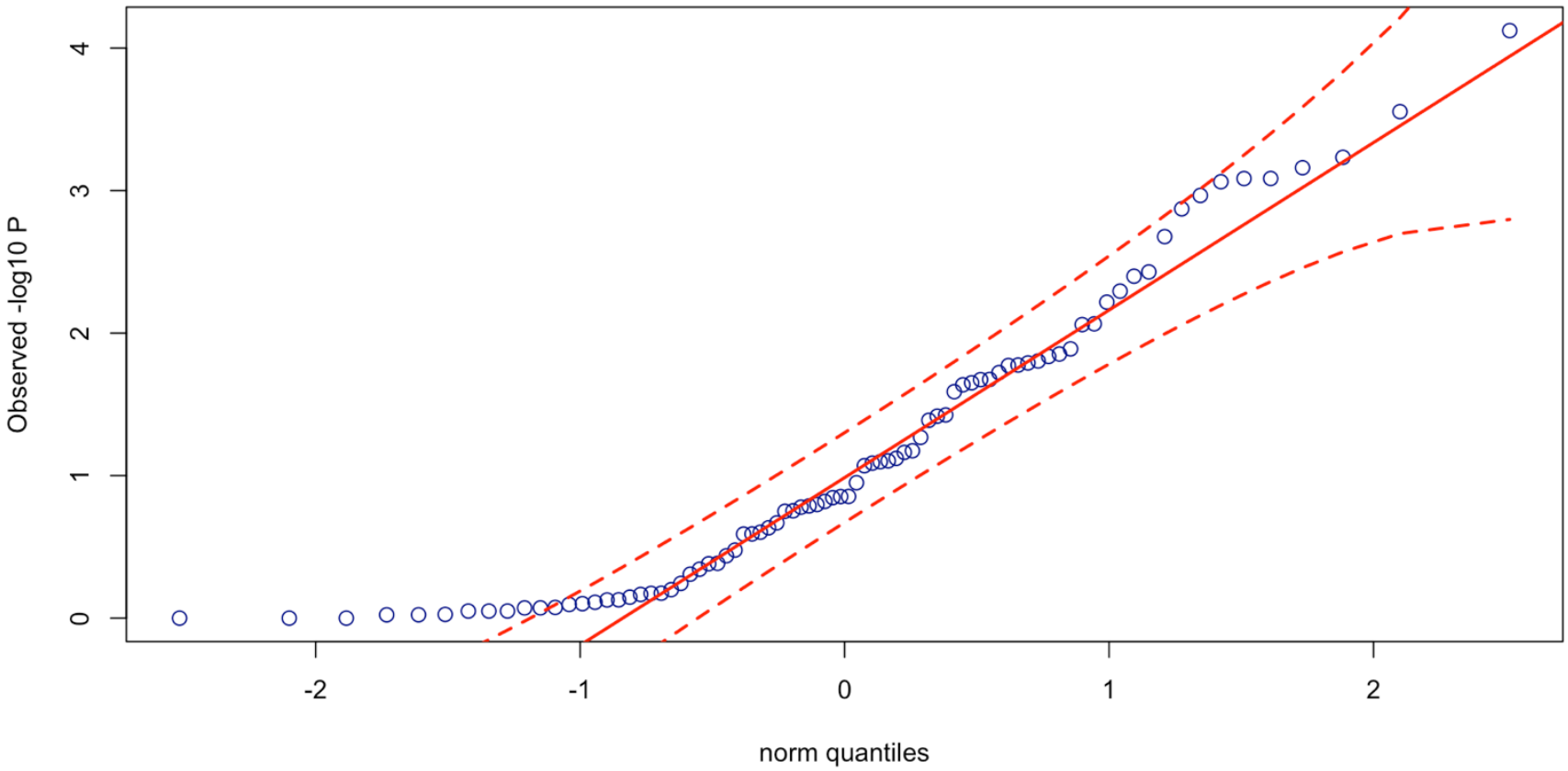


Figure S3: Quantile-quantile plot of the P value of common SNPs in the examined DRGs in the discovery CRC cases compared with the control group (ExAC). No significant deviation from the expected distribution was seen.

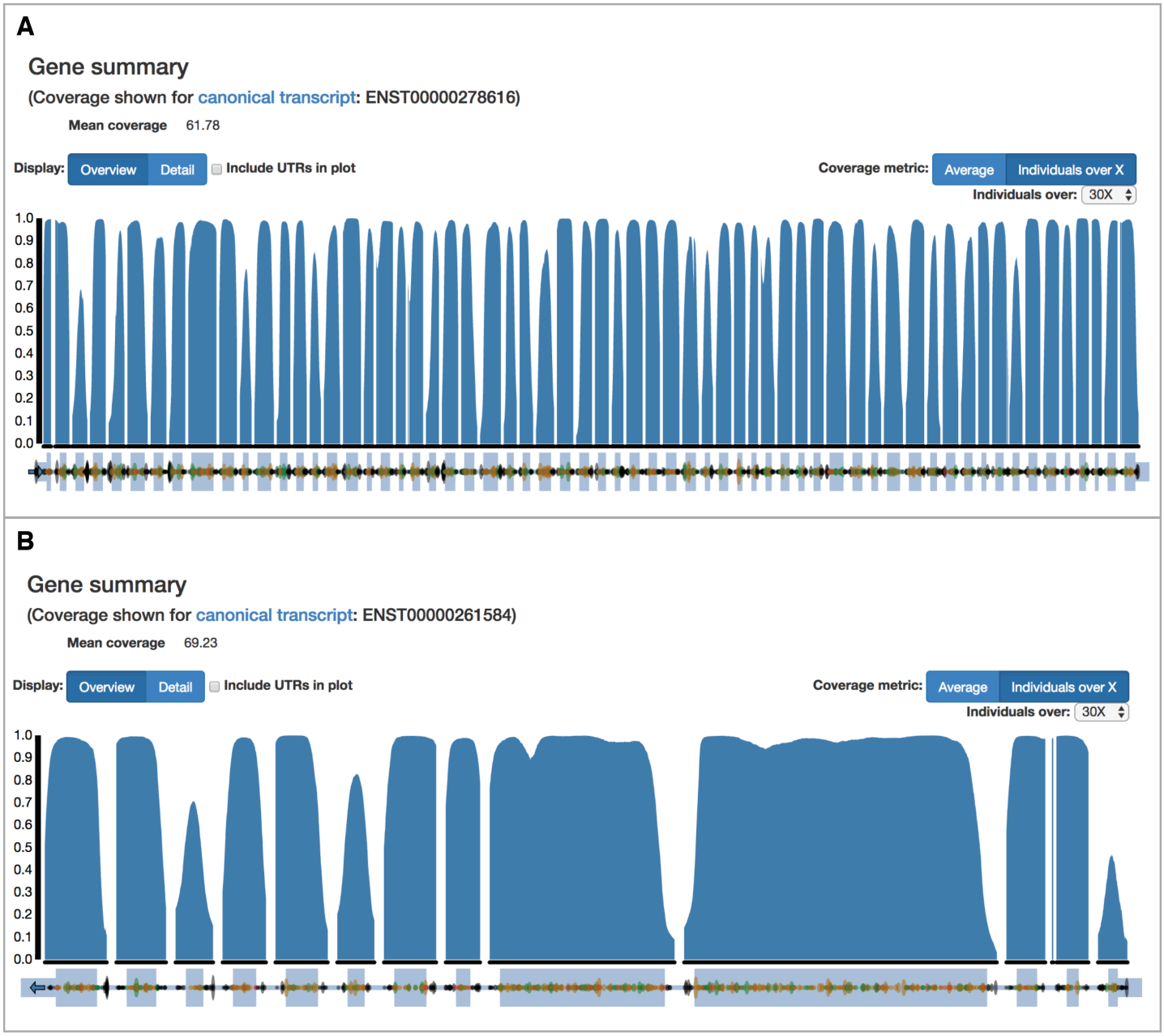


Figure S4: Sequencing coverage of (A) *ATM* and (B) *PALB2* genes in the ExAC cohort, showing the proportion of individuals who had at least 30X coverage for the coding exons.

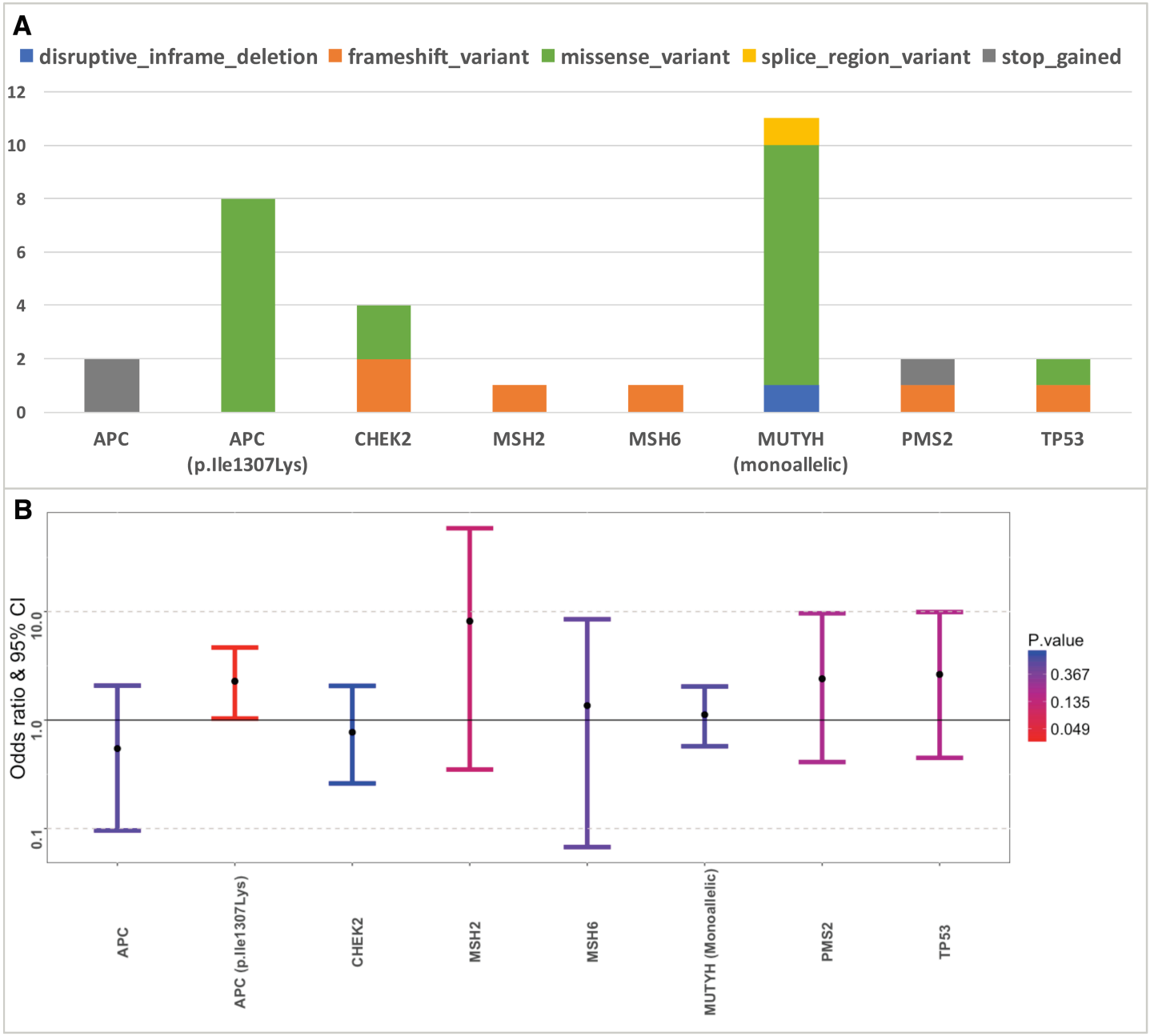


Figure S5: Pathogenic germline mutations in the CRC risk genes in the discovery cohort (n=680). A; Number and impact of detected germline mutations in the examined CRC risk genes. B; Enrichment of germline mutations in the CRC risk genes in the discovery cohort (n=680).

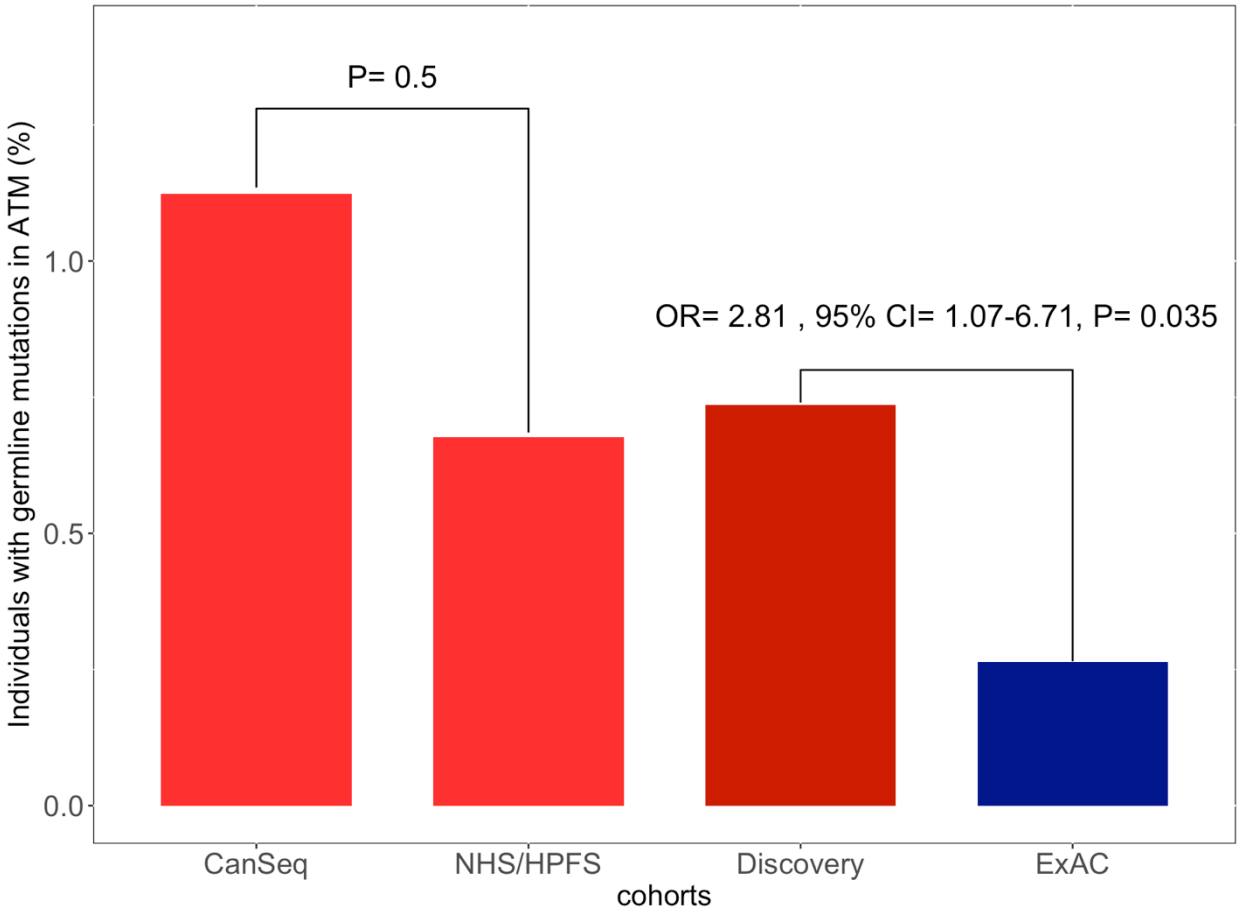


Figure S6: Enrichment of germline pathogenic mutations in *ATM* in each cohort of the discovery set. Our analysis showed that both NHS/HPFS and Canseq cohorts were enriched for *ATM* mutations. There was no statistically significant difference in the frequency of these disruptive events in the Canseq cohort compared with NHS/HPFS (P = 0.5). (NHS: Nurses' Health Study; HPFS: Health Professional Follow up Study; CanSeq: Cancer Sequencing study)

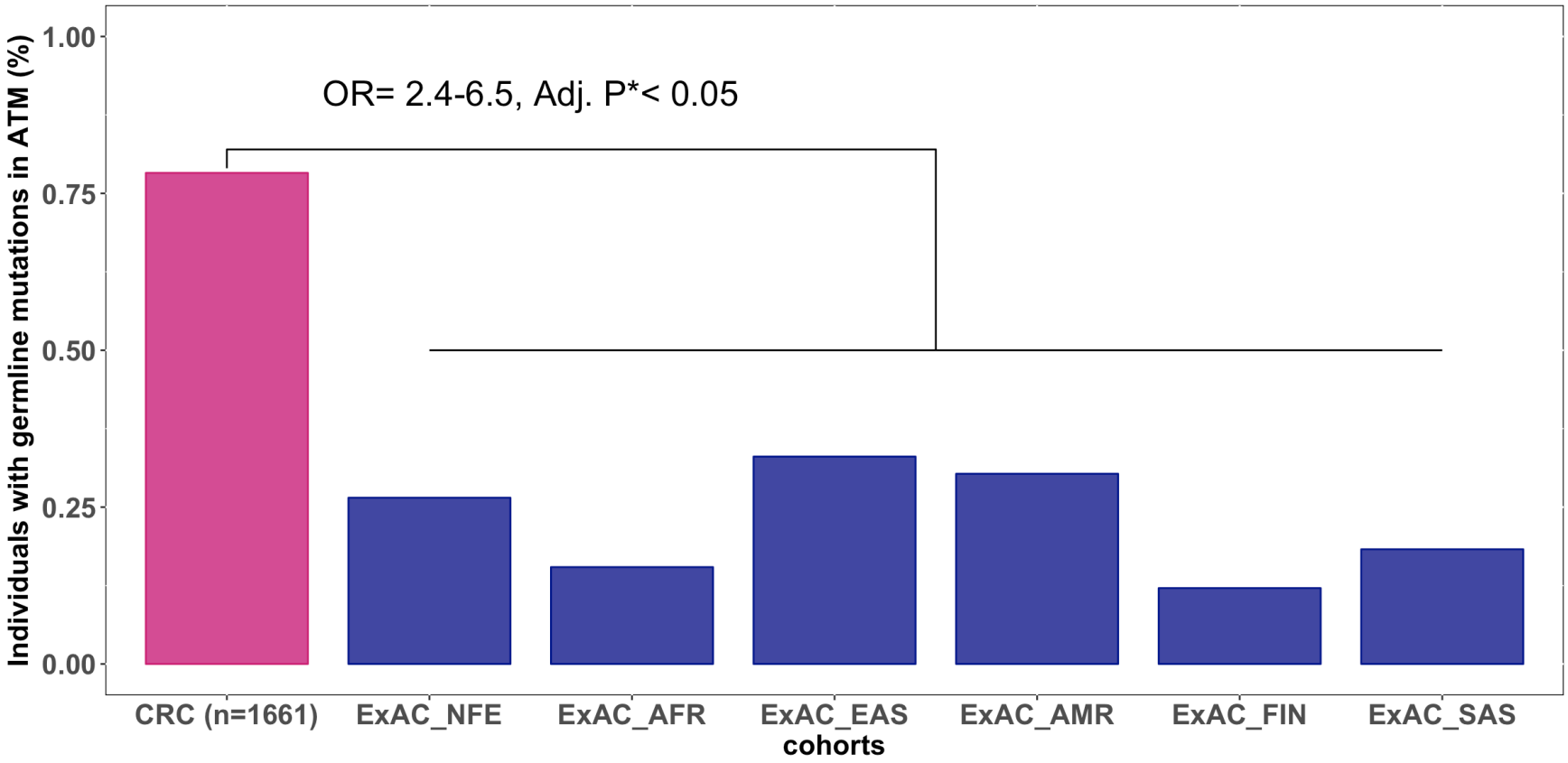


Figure S7: Enrichment of germline *ATM* mutations in the validation set (n= 1661) compared with the various major populations in the ExAC cohort (n=53105; TCGA data excluded; AFR: African & African American, AMR: American, EAS: East Asian, FIN: Finnish, NFE: Non-Finnish European, SAS: South Asian).

* P value was adjusted for 6 independent tests using Bonferroni correction

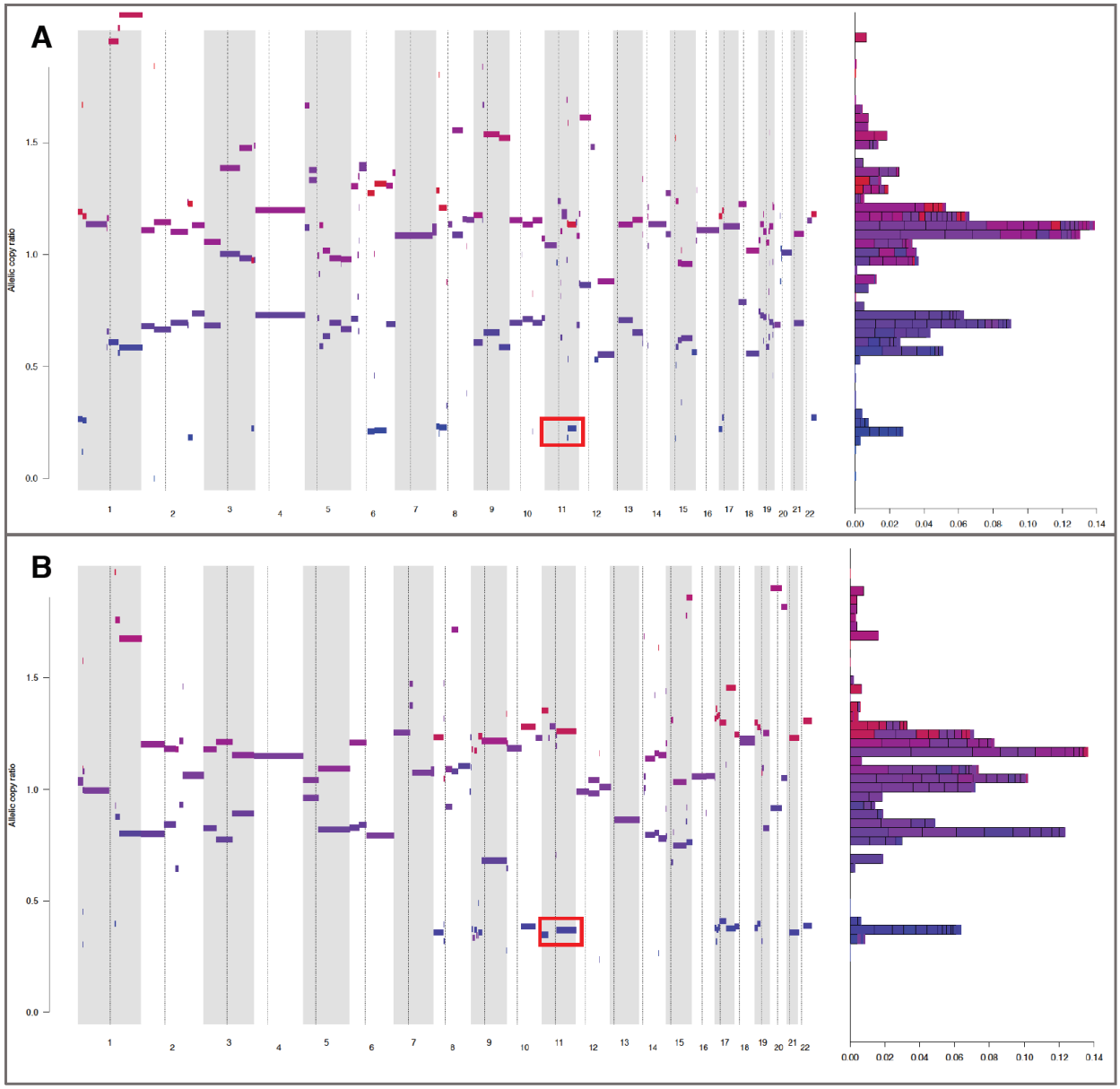


Figure S8: Evaluation of the tumors of cases with germline *ATM* mutations showed LOH of the *ATM* wild-type allele. Two individuals (top: 1221; bottom: 1755) had large deletions involving the cytogenetic region,11q22, which encompasses the *ATM* gene (highlighted).

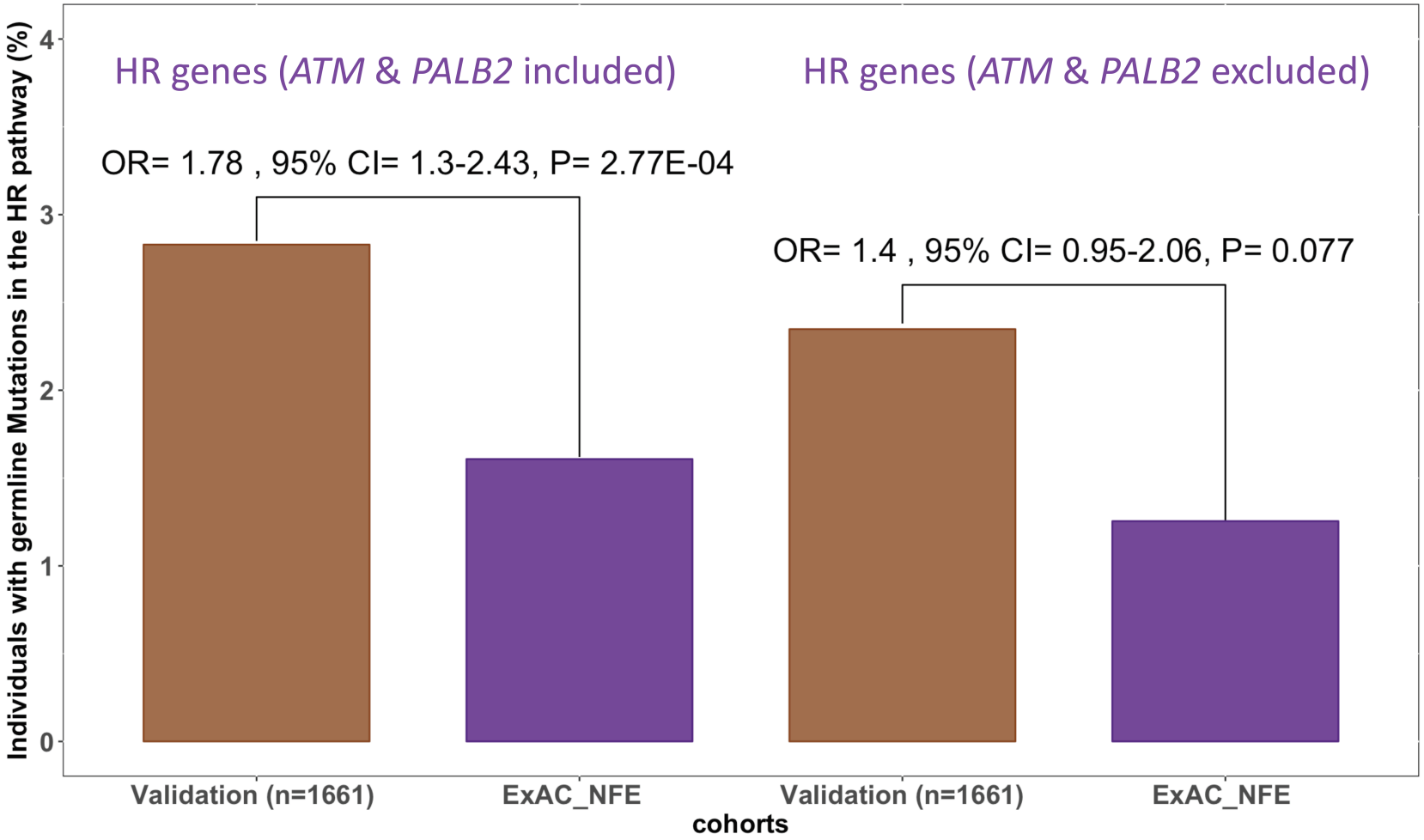


Figure S9: Enrichment of germline pathogenic mutations in the homologous recombination pathway in the CRC validation set. (ExAC: Exome Aggregation Consortium; NFE: Non-Finnish European)

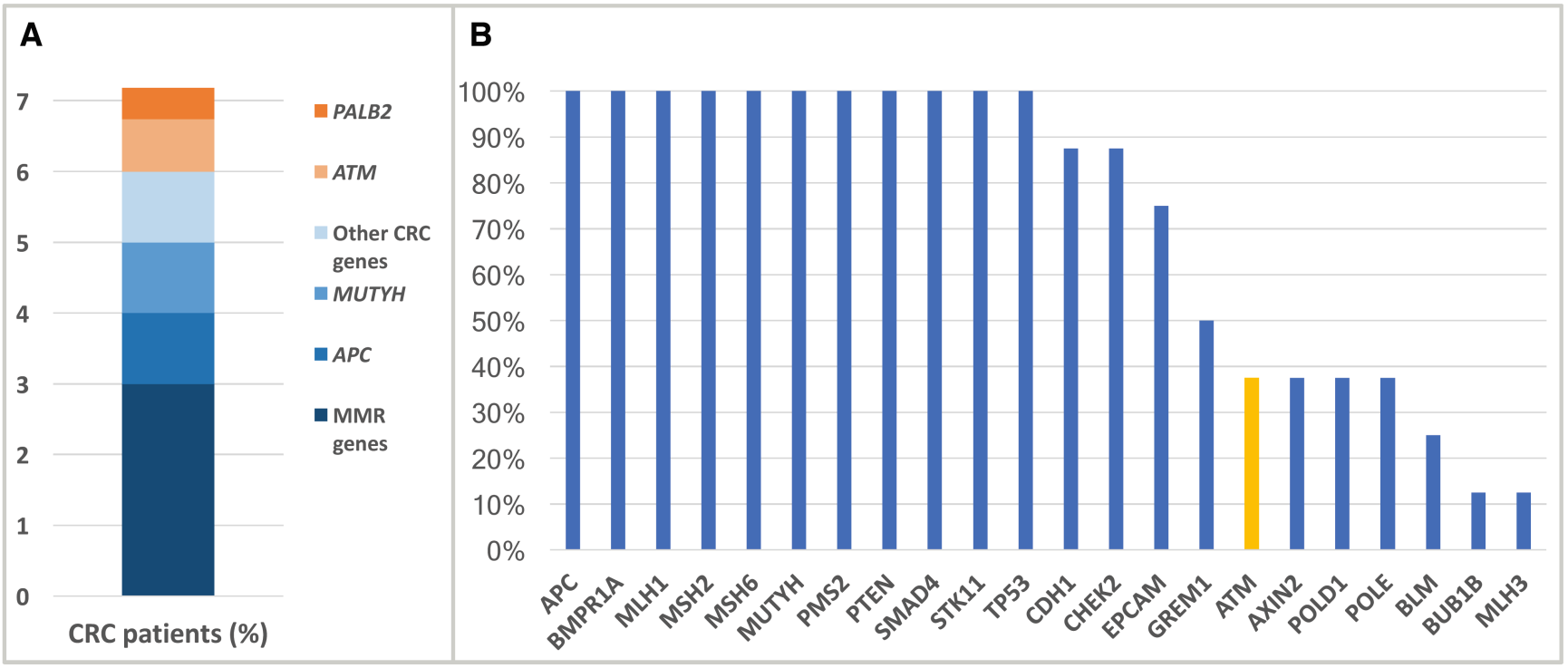


Figure S10: Diagnostic yield of germline testing in unselected CRC cases. A; Although *ATM* and *PALB2* may only explain the CRC heritability in ~1.2% of unselected CRC cases, this represents a potential 20% increase in the current diagnostic yield. B; Genes typically included in the CRC-specific germline testing panels offered by 8 of the largest commercial laboratories in the US (as of August 2017). As shown, *ATM* is only occasionally included in these panels whereas *PALB2* and other highly actionable DRGs are not captured by these clinical tests.

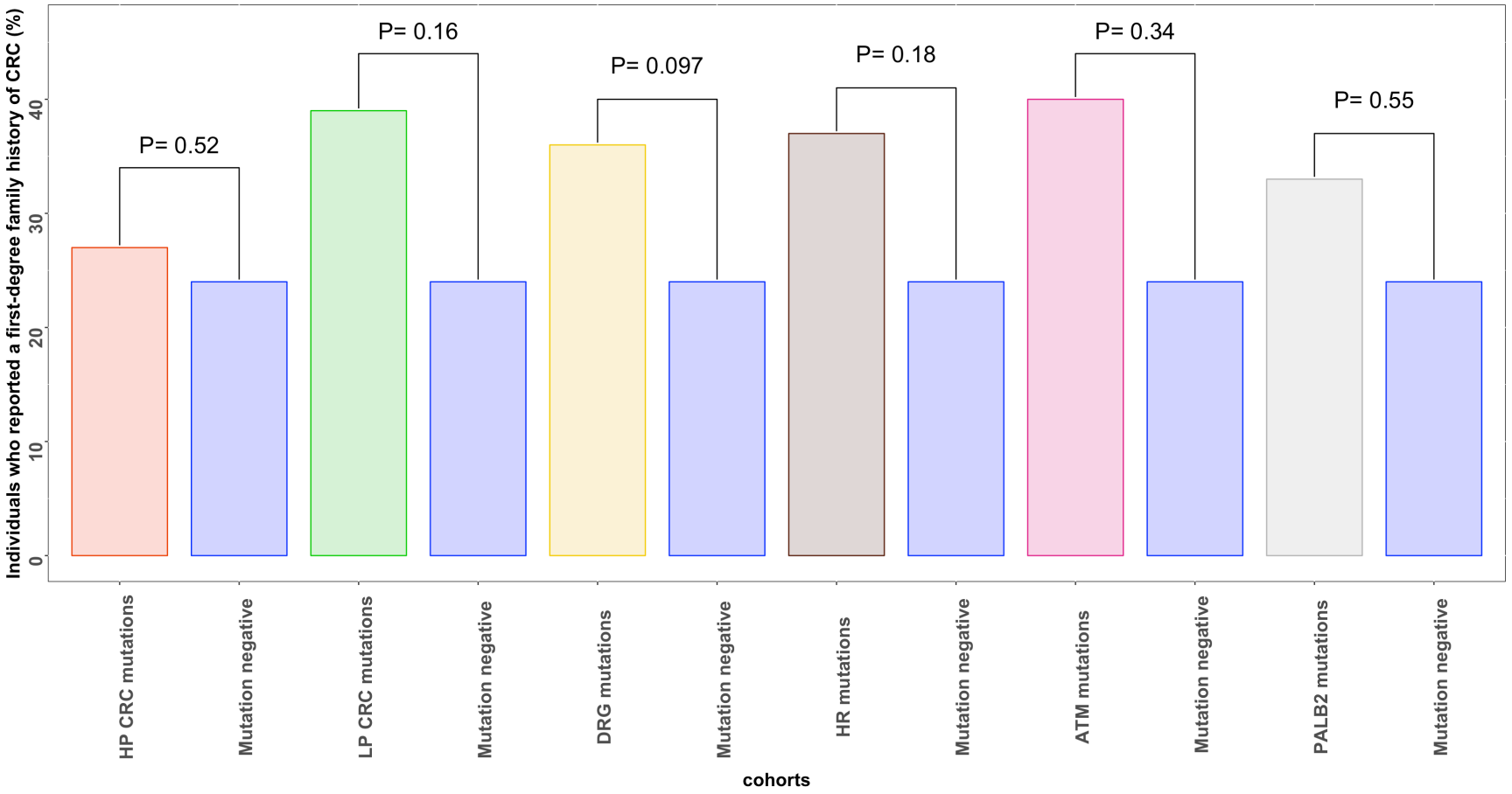


Figure S11: proportions of CRC Individuals who reported positive family history of CRC in one or more first-degree relatives. Individuals with germline pathogenic mutations in the CRC risk genes, DRGs, HR, *ATM* or *PALB2* were not more likely to have a positive family of CRC. Genes contained in each set are listed in Tables 1, S2, and S3.

Supplementary tables:

Table S1: The germline analysis workflows for the examined CRC cohorts in our study.

Cohort	NHS/HPFS study	CANSEQ study	TCGA study	Yurgelun et al. 2017	NSCCG study	Pearlman et al. 2016
Number of cases	591	89	603	1058	1006	450
Sequenced tissue	Adjacent normal tissue	Blood	Blood or adjacent normal tissue	Blood	Blood	Blood and
Bioinformatics analysis	Germline DNA from the CRC patients in the NHS/HPFS cohort was obtained from adjacent normal colon tissue that was dissected after pathology review. DNA was extracted from formalin-fixed, paraffin embedded (FFPE) blocks using the QIAGEN QIAamp DNA FFPE Tissue Kit. Whole-exome capture libraries were constructed from tumor and normal DNA after sample shearing, end repair, phosphorylation, and ligation to barcoded sequencing adaptors. DNA reads were then captured using SureSelect v.2 Exome bait (Agilent Technologies) and then sequenced on Illumina HiSeq 2000.	Whole blood, from the CRC patients in the CanSeq study, was used for germline DNA extraction. Whole-exome capture libraries were constructed from tumor and normal DNA after sample shearing, end repair, phosphorylation, and ligation to barcoded sequencing adaptors. DNA was then subjected to solution-phase hybrid capture using Agilent baits. The samples were multiplexed and sequenced using Illumina HiSeq technology as previously described.	All sequence data for TCGA cohort were aligned to the GRCh37 reference genome. Where available, pre-aligned data were acquired from the NCI GDC Legacy Archive. An additional 104 samples only available via the NCI GDC Data Portal (pre-aligned to the GRCh38 reference genome) were manually realigned to the GRCh37 reference genome. To perform realignment, the GATK CleanSam and RevertSam tools were first applied to revert previous alignment data and split samples by read group. Subsequently, BWA mem was used to realign each sample (per read group) to the GRCh37 reference genome, after which read groups belonging to a single sample were merged using MergeSamFiles and the Genome Analysis ToolKit Best Practices for performing quality control in aligned sequence data were followed. Production pipelines of the raw sequencing data of the TCGA cohort has been previously described .	The analysis pipeline for this cohort has been previously described (J Clin Oncol. 2017 Apr 1;35(10):1086-1095).	The analysis pipeline for this cohort has been previously described (Br J Cancer. 2007 Nov 5; 97(9): 1305–1309).	The analysis pipeline for this cohort has been previously described (JAMA Oncol. 2017 Apr 1;3(4):464-471).
Variant discovery and functional annotation	Germline whole exome sequencing data were used to perform variant calling of single nucleotide variants (SNVs) and small deletions/duplications (indels) across all samples in each cohort. Genome Analysis Toolkit (GATK) HaplotypeCaller pipeline was used according to the recommended GATK best practices. GATK Variant Quality Score Recalibration (VQSR) was used to filter variants. The SNP VQSR model was trained using HapMap3.3 and 1KG Omni 2.5 SNP sites and a 99.5% sensitivity threshold was applied to filter variants. In addition, Mills et. al. 1KG gold standard and Axiom Exome Plus sites were used for insertions/deletion sites and a 95% sensitivity threshold, similar to that used for the ExAC cohort, was used to call indel variants in the discovery cohort patients. A more stringent filter (VQSR90) was applied to filter germline indel calls on the TCGA cohort to significantly minimize the risk of false positive calls secondary to sequencing artifacts. Variant annotation was performed using SnpEff, version 4.1, on GRCh37. SnpEff was used to determine Ensemble Gene ID and gene symbol, and Ensemble Transcript ID for each functional consequence of the variant. Only variants impacting the canonical transcript of the gene were included.			The analysis pipeline for this cohort has been previously described (J Clin Oncol. 2017 Apr 1;35(10):1086-1095).	The analysis pipeline for this cohort has been previously described (Br J Cancer. 2007 Nov 5; 97(9): 1305–1309).	The analysis pipeline for this cohort has been previously described (JAMA Oncol. 2017 Apr 1;3(4):464-471).
Variant Interpretation	An identical workflow for variant inclusion and pathogenicity assessment was used to evaluate the germline variants in both cases and controls. The analysis of germline variants focused on variants identified among the examined 54 genes (14 established CRC genes and 40 additional DRGs). Pathogenicity of the detected variants was determined according to the most recent guidelines published jointly by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP). Germline variants were evaluated against the published literature and publicly available databases such as ClinVar and variant-specific databases. Population minor allele frequencies were extracted from publicly available databases such as the Exome Aggregation Consortium (ExAC) and the 1000 genomes project. Only pathogenic and likely pathogenic variants referred to as pathogenic mutations) with sufficient evidence of pathogenicity were included. Variants of unknown significance (VUS) were excluded from all analyses. In cases and controls, all coding non-synonymous variants (such as missense, nonsense, inframe deletions, inframe insertions, frameshift insertions and deletions as well as splice site variants) were evaluated. Large alterations in the genes of interest were not examined as access to extra DNA to perform MLPA, or other testing modalities for copy number alterations, was not available.					

Table S2: DNA repair genes that were evaluated in this study.

Gene	HGNC Approved Name	Cytogenetic region	Cancer Predisposition Syndrome
ATM	ATM serine/threonine kinase	11q22-q23	Ataxia Telangiectasia
ATR	ATR serine/threonine kinase	3q23	Other Cancer Predisposition
BAP1	BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase)	3p21.1	Melanocytic Tumor syndrome, Familial Uveal Melanoma
BARD1	BRCA1 associated RING domain 1	2q35	Other Cancer Predisposition
BLM	Bloom syndrome, RecQ helicase-like	15q26.1	Bloom Syndrome
BRCA1	breast cancer 1, early onset	17q21.31	Hereditary Breast and Ovarian Cancer
BRCA2	breast cancer 2, early onset	13q12-q13	Hereditary Breast and Ovarian Cancer
BRIP1	BRCA1 interacting protein C-terminal helicase 1	17q22.2	Other Cancer Predisposition
DDB2	damage-specific DNA binding protein 2, 48kDa	11p12-p11	Xeroderma Pigmentosa
ERCC2	excision repair cross-complementation group 2	19q13.3	Xeroderma Pigmentosa
ERCC3	excision repair cross-complementation group 3	2q21	Xeroderma Pigmentosa
ERCC4	excision repair cross-complementation group 4	16p13.3	Xeroderma Pigmentosa
ERCC5	excision repair cross-complementation group 5	13q22-q34	Xeroderma Pigmentosa
FANCA	Fanconi anemia, complementation group A	16q24.3	Fanconi Anemia
FANCB	Fanconi anemia, complementation group B	Xp22.2	Fanconi Anemia
FANCC	Fanconi anemia, complementation group C	9q22.3	Fanconi Anemia
FANCD2	Fanconi anemia, complementation group D2	3p25.3	Fanconi Anemia
FANCE	Fanconi anemia, complementation group E	6p22-p21	Fanconi Anemia
FANCF	Fanconi anemia, complementation group F	11p15	Fanconi Anemia
FANCG	Fanconi anemia, complementation group G	9p13	Fanconi Anemia
FANCI	Fanconi anemia, complementation group I	15q26.1	Fanconi Anemia
FANCL	Fanconi anemia, complementation group L	2p16.1	Fanconi Anemia
FANCM	Fanconi anemia, complementation group M	14q21.3	Fanconi Anemia
GEN1	Holliday junction 5' flap endonuclease	2p24.2	Other Cancer Predisposition
MRE11	MRE11 homolog, double strand break repair nuclease	11q21	Ataxia-Telangiectasia-Like Disorder
NBN	nibrin	8q21-q24	Nijmegen Breakage Syndrome
NTHL1	nth like DNA glycosylase 1	16p13.3	Familial adenomatous polyposis 3
PALB2	partner and localizer of BRCA2	16p12.1	Fanconi Anemia
PCNA	proliferating cell nuclear antigen	20p12.3	Ataxia-telangiectasia-like disorder
RAD51	RAD51 recombinase	15q15.1	Breast cancer
RAD51C	RAD51 paralog C	17q25.1	Ovarian cancer
RAD51D	RAD51 paralog D	17q11	Ovarian cancer
RAD54L	RAD54 like	1p34.1	Breast cancer
RECQL4	RecQ protein-like 4	8q24.3	Rothmund Thomson Syndrome
SLX4	SLX4 structure-specific endonuclease subunit	16p13.3	Fanconi anemia
UBE2T	ubiquitin conjugating enzyme E2 T	1q32.1	Fanconi anemia
WRN	Werner syndrome, RecQ helicase-like	8p12	Werner Syndrome
XPA	xeroderma pigmentosum, complementation group A	9q22.3	Xeroderma Pigmentosa
XPC	xeroderma pigmentosum, complementation group C	3p25.1	Xeroderma Pigmentosa
XRCC3	X-ray repair cross complementing 3	14q32.3	Breast cancer

Table S3: Established CRC risk genes that were evaluated in this study.

Gene	HGNC Approved Name	Cytogenetic region	Cancer Predisposition Syndrome
APC	adenomatous polyposis coli	5q21-q22	Familial Adenomatous Polyposis
BMPR1A	bone morphogenetic protein receptor, type IA	10q22.3	Hereditary Mixed Polyposis Syndrome
CHEK2	checkpoint kinase 2	22q12.1	Hereditary Breast
MLH1	mutL homolog 1	3p22.3	Lynch Syndrome / CMMRD
MSH2	mutS homolog 2	2p21	Lynch Syndrome / CMMRD
MSH6	mutS homolog 6	2p16	Lynch Syndrome / CMMRD
MUTYH	mutY homolog	1p34.1	Colorectal cancer
PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	7p22.1	Lynch Syndrome / CMMRD
POLD1	polymerase (DNA directed), delta 1, catalytic subunit	19q13.3	Colorectal cancer
POLE	polymerase (DNA directed), epsilon, catalytic subunit	12q24.3	Colorectal cancer
PTEN	phosphatase and tensin homolog	10q23	Cowden syndrome
SMAD4	SMAD family member 4	18q21.1	Juvenile Polyposis
STK11	serine/threonine kinase 11	19p13.3	Peutz Jeghers syndrome
TP53	tumor protein p53	17p13.1	Li Fraumeni Syndrome

Table S5: Depth of sequencing of the examined DRGs in the ExAC cohort.

Gene	Average depth of coverage (reads)
ATM	61.78
ATR	58.70
BAP1	56.83
BARD1	58.65
BLM	61.70
BRCA1	66.14
BRCA2	59.19
BRIP1	64.51
DDB2	66.05
ERCC2	50.78
ERCC3	65.46
ERCC4	62.51
ERCC5	58.01
FANCA	52.07
FANCB	60.29
FANCC	48.46
FANCD2	65.88
FANCE	62.03
FANCF	79.42
FANCG	74.03
FANCI	69.40
FANCL	53.48
FANCM	58.54
GEN1	56.12
MRE11	55.41
NBN	60.07
NTHL1	52.13
PALB2	69.23
PCNA	62.64
RAD51	64.25
RAD51C	58.95
RAD51D	50.01
RAD54L	64.53
RECQL4	32.49
SLX4	71.76
UBE2T	67.68
WRN	57.92
XPA	40.62
XPC	45.37
XRCC3	23.71
Average	58.67

Table S6: Germline mutations in the well-known CRC risk genes in the CRC discovery set (n=680).

Case ID	chrom	gene	start	end	ref	alt	impact	codon_change	amino_acid _change	AF _EXAC (%)	genotype
283	chr5	*APC*	112175418	112175419	T	G	stop_gained	c.4128T>G	p.Tyr1376*	0.000	Heterozygous
142275	chr5	*APC*	112162890	112162891	C	T	stop_gained	c.1495C>T	p.Arg499*	0.000	Heterozygous
200096	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
200198	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
1760	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
3527	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
3669	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
4529	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
4536	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
621	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
200245	chr22	*CHEK2*	29091855	29091857	AG	A	frameshift_variant	c.1229delC	p.Thr410Metfs*15	0.177	Heterozygous
50	chr22	*CHEK2*	29091855	29091857	AG	A	frameshift_variant	c.1229delC	p.Thr410Metfs*15	0.177	Heterozygous
430	chr22	*CHEK2*	29090053	29090054	G	A	missense_variant	c.1556C>T	p.Thr519Met	0.038	Heterozygous
680	chr22	*CHEK2*	29090053	29090054	G	A	missense_variant	c.1556C>T	p.Thr519Met	0.038	Heterozygous
68	chr2	*MSH2*	47707897	47707898	T	TA	frameshift_variant	c.2523dupA	p.Glu842Argfs*4	0.000	Heterozygous
3225	chr2	*MSH6*	48033743	48033748	AAAGC	A	frameshift_variant	c.3959_3962delCAAG	p.Ala1320Glufs*6	0.001	Heterozygous
213	chr1	*MUTYH*	45798474	45798475	T	C	missense_variant	c.536A>G	p.Tyr179Cys	0.162	Heterozygous
227039	chr1	*MUTYH*	45797227	45797228	C	T	missense_variant	c.1187G>A	p.Gly396Asp	0.278	Heterozygous
2365	chr1	*MUTYH*	45798474	45798475	T	C	missense_variant	c.536A>G	p.Tyr179Cys	0.162	Heterozygous
280	chr1	*MUTYH*	45797227	45797228	C	T	missense_variant	c.1187G>A	p.Gly396Asp	0.278	Heterozygous
2939	chr1	*MUTYH*	45797227	45797228	C	T	missense_variant	c.1187G>A	p.Gly396Asp	0.278	Heterozygous
3227	chr1	*MUTYH*	45797227	45797228	C	T	missense_variant	c.1187G>A	p.Gly396Asp	0.278	Heterozygous
353	chr1	*MUTYH*	45797227	45797228	C	T	missense_variant	c.1187G>A	p.Gly396Asp	0.278	Heterozygous
442	chr1	*MUTYH*	45797834	45797835	T	G	splice_region_variant	c.933+3A>C		0.007	Heterozygous
627	chr1	*MUTYH*	45797227	45797228	C	T	missense_variant	c.1187G>A	p.Gly396Asp	0.278	Heterozygous
92	chr1	*MUTYH*	45797227	45797228	C	T	missense_variant	c.1187G>A	p.Gly396Asp	0.278	Heterozygous
200193	chr1	*MUTYH*	45796889	45796893	TTCC	T	disruptive_inframe_ deletion	c.1437_1439delGGA	p.Glu480del	0.012	Heterozygous
352566	chr7	*PMS2*	6026563	6026564	A	AT	frameshift_variant	c.1831dupA	p.Ile611fs	0.001	Heterozygous
390	chr7	*PMS2*	6026708	6026709	G	A	stop_gained	c.1687C>T	p.Arg563*	0.002	Heterozygous
200019	chr17	*TP53*	7577598	7577600	CA	C	frameshift_variant	c.681delT	p.Asp228Thrfs*19	0.000	Heterozygous
200107	chr17	*TP53*	7577537	7577538	C	T	missense_variant	c.743G>A	p.Arg248Gln	0.006	Heterozygous

Table S7: Germline mutations in the well-known CRC risk genes in the TCGA cohort (n=603).

Case ID	chrom	gene	start	end	ref	alt	impact	codon_change	amino_acid_change	AF_EXAC (%)	genotype
TCGA_CRC_18	chr5	*APC*	112102092	112102093	T	A	stop_gained	c.206T>A	p.Leu69*	0.001	Heterozygous
TCGA_CRC_01	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
TCGA_CRC_04	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
TCGA_CRC_11	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
TCGA_CRC_16	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
TCGA_CRC_17	chr5	*APC*	112175210	112175211	T	A	missense_variant	c.3920T>A	p.Ile1307Lys	0.169	Heterozygous
TCGA_CRC_31	chr22	*CHEK2*	29091855	29091857	AG	A	frameshift_variant	c.1229delC	p.Thr410fs	0.177	Heterozygous
TCGA_CRC_32	chr22	*CHEK2*	29091855	29091857	AG	A	frameshift_variant	c.1229delC	p.Thr410fs	0.177	Heterozygous
TCGA_CRC_33	chr22	*CHEK2*	29091855	29091857	AG	A	frameshift_variant	c.1229delC	p.Thr410fs	0.177	Heterozygous
TCGA_CRC_35	chr22	*CHEK2*	29091206	29091207	G	A	missense_variant	c.1412C>T	p.Ser471Phe	0.030	Heterozygous
TCGA_CRC_12	chr3	*MLH1*	37053588	37053589	C	T	stop_gained	c.676C>T	p.Arg226*	0.001	Heterozygous
TCGA_CRC_03	chr2	*MSH2*	47703537	47703538	C	T	stop_gained	c.2038C>T	p.Arg680*	0.001	Heterozygous
TCGA_CRC_09	chr2	*MSH2*	47657023	47657024	T	TC	frameshift_variant	c.1221dupC	p.Tyr408fs	0.001	Heterozygous
TCGA_CRC_07	chr2	*MSH6*	48025862	48025864	AC	A	frameshift_variant	c.742delC	p.Arg248fs	0.001	Heterozygous
TCGA_CRC_01	chr1	*MUTYH*	45798465	45798466	C	T	missense_variant	c.545G>A	p.Arg182His	0.002	Heterozygous

Table S8: Germline mutations in the well-known CRC risk genes in the Yurgelun et al. 2017 cohort (n=1058).

gene	codon change	amino acid change	gene (2)	codon change (2)	amino acid change (2)
APC	c.1495C>T	p.R499X			
APC	c.3183_3187del	p.Q1062X			
APC	c.1213C>T	p.R405X			
APC	c.70C>T	p.R24X			
APC	c.937_938del	p.E313Nfs*13			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T>A	p.I1307K			
APC (p.Ile1307Lys)	c.3920T.A	p.I1307K	*BRCA1*	c.68_69del	p.E23Vfs*17
APC (p.Ile1307Lys)	c.3920T.A	p.I1307K	*BRCA1*	c.68_69del	p.E23Vfs*17
CHEK2	c.1100del	p.T367Mfs*15			
CHEK2	exons 8	9 deletion			
MLH1	c.2070_2071insTT	p.I691Lfs*ext			
MLH1	c.1411_1414del	p.K471Dfs*19			
MLH1	c.55A>T	p.I19F			
MLH1	c.230G>A	p.C77Y			
MLH1	c.1852_1854del	p.K618del			
MLH1	c.1667G>A	p.S556N			
MLH1	c.5C>A	p.S2X			
MLH1	c.350C>T	p.T117M			
MLH1	c.678	1G>A			
MLH1	c.2195_2198dup	p.H733Qfs*14			
MLH1	whole gene deletion				
MLH1	exons 16	19 deletion			
MLH1	exons 16	19 deletion	*BRCA2*	c.3199del	p.T1067Kfs*10
MSH2	c.1906G>C	p.A636P	*APC (p.Ile1307Lys)*	c.3920T.A	p.I1307K
MSH2	c.2074G>T	p.G692W			
MSH2	c.2082dup	p.V695Cfs*4			
MSH2	c.1906G>C	p.A636P			
MSH2	exons 9	12 deletion			
MSH2	exons 1	6 deletion			
MSH2	exon 8 duplication				
MSH6	c.3939_3957dup	p.A1320Sfs*5			
MSH6	c.10C>T	p.Q4X			
MSH6	c.3939_3957dup	p.A1320Sfs*5			
MSH6	c.1519dup	p.R507Kfs*9			
MSH6	c.1519dup	p.R507Kfs*9			
MSH6	whole gene deletion				
MUTYH (Biallelic loss)	c.494A>G	p.Y165C	*MUTYH*	c.1145G.A	p.G382D
MUTYH (Biallelic loss)	c.1145G>A	p.G382D	*MUTYH*	c.1145G.A	p.G382D
MUTYH (Biallelic loss)	c.1145G>A	p.G382D	*MUTYH*	c.283C.T	p.R95W

MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.891+3A>C				
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.494A>G	p.Y165C			
MUTYH (monoallelic loss)	c.494A>G	p.Y165C			
MUTYH (monoallelic loss)	c.494A>G	p.Y165C			
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.494A>G	p.Y165C			
MUTYH (monoallelic loss)	c.1282	1G>T			
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.494A>G	p.Y165C			
MUTYH (monoallelic loss)	c.892	2A>G			
MUTYH (monoallelic loss)	c.1145G>A	p.G382D			
MUTYH (monoallelic loss)	c.503G>A	p.R168H			
MUTYH (monoallelic loss)	c.494A.G	Y165C	*BRCA2*	c.1796_1800del	p.S599X
PMS2	c.2174+1G>A				
PMS2	c.2117del	p.K706Sfs*19			
PMS2	c.765C>G	p.Y255X			
PMS2	c.1067del	p.K356Rfs*4			
PMS2	c.736_741delins11	p.P246Cfs*3			
PMS2	exon 13 deletion				
PMS2	exons 6	15 deletion			
TP53	c.681del	p.D228Tfs*19			

Table S9: Germline mutations in the examined DNA-repair genes in the CRC discovery set (n=680).

Case ID	chrom	gene	start	end	ref	alt	impact	codon_change	amino_acid_change	AF_EXAC (%)	genotype
200205	chr11	*ATM*	108213986	108213987	G	A	stop_gained	c.8307G>A	p.Trp2769*	0.001	Heterozygous
1221	chr11	*ATM*	108155006	108155008	AG	A	frameshift_variant	c.3802delG	p.Val1268fs*	0.003	Heterozygous
1755	chr11	*ATM*	108141873	108141874	G	T	splice_region_variant	c.2921+1G>T		0.000	Heterozygous
2760	chr11	*ATM*	108190743	108190746	CAG	C	frameshift_variant	c.6415_6416delGA	p.Glu2139Ilefs*6	0.000	Heterozygous
3645	chr11	*ATM*	108115680	108115681	G	T	stop_gained	c.829G>T	p.Glu277*	0.000	Heterozygous
2967	chr3	*BAP1*	52439281	52439282	G	GC	frameshift_variant	c.959dupG	p.Cys320fs	0.000	Heterozygous
2775	chr2	*BARD1*	215595201	215595204	CAT	C	frameshift_variant	c.1932_1933delAT	p.Cys645fs	0.000	Heterozygous
1743	chr15	*BLM*	91306245	91306246	C	T	stop_gained	c.1933C>T	p.Gln645*	0.004	Heterozygous
3046	chr15	*BLM*	91306245	91306246	C	T	stop_gained	c.1933C>T	p.Gln645*	0.004	Heterozygous
3181	chr15	*BLM*	91310195	91310196	C	CAAAT	frameshift_variant	c.2250_2251insAAAT	p.Leu751fs	0.000	Heterozygous
3111	chr17	*BRCA1*	41209078	41209079	T	TG	frameshift_variant	c.5329dupC	p.Gln1777fs	0.016	Heterozygous
200127	chr13	*BRCA2*	32914173	32914174	C	G	stop_gained	c.5682C>G	p.Tyr1894*	0.000	Heterozygous
2265	chr13	*BRCA2*	32936731	32936732	G	C	missense_variant	c.7878G>C	p.Trp2626Cys	0.002	Heterozygous
2406	chr13	*BRCA2*	32912963	32912968	TGAAA	T	frameshift_variant	c.4478_4481delAAAG	p.Glu1493Valfs*10	0.000	Heterozygous
3444	chr13	*BRCA2*	32890598	32890600	TG	T	frameshift_variant	c.3delG	p.Met1fs	0.000	Heterozygous
200054	chr17	*BRIP1*	59761412	59761417	CTTTG	C	frameshift_variant	c.2990_2993delCAAA	p.Thr997Argfs*61	0.002	Heterozygous
204	chr17	*BRIP1*	59885904	59885906	GA	G	frameshift_variant	c.840delT	p.His281Ilefs*8	0.000	Heterozygous
207	chr19	ERCC2	*45856058*	45856059	C	G	missense_variant	c.1847G>C	p.Arg616Pro	0.013	Heterozygous
262114	chr19	ERCC2	*45856058*	45856059	C	G	missense_variant	c.1847G>C	p.Arg616Pro	0.013	Heterozygous
3558	chr2	*ERCC3*	128050331	128050332	G	A	stop_gained	c.325C>T	p.Arg109*	0.048	Heterozygous
3680	chr16	*ERCC4*	14014079	14014080	C	T	stop_gained	c.58C>T	p.Arg20*	0.000	Heterozygous
2430	chr9	*FANCC*	97864023	97864024	G	A	stop_gained	c.1642C>T	p.Arg548*	0.002	Heterozygous
3439	chr6	*FANCE*	35423605	35423607	GA	G	frameshift_variant	c.334delA	p.Ser112Valfs*14	0.000	Heterozygous
3048	chr2	*FANCL*	58388743	58388744	A	C	stop_gained	c.948T>G	p.Tyr316*	0.000	Heterozygous
200226	chr2	*GEN1*	17962406	17962411	TAAAG	T	frameshift_variant	c.1933_1936delAAAG	p.Lys645Cysfs*29	0.007	Heterozygous
251	chr2	*GEN1*	17942842	17942845	AAG	A	frameshift_variant	c.347_348delAG	p.Glu116Valfs*20	0.001	Heterozygous
2760	chr11	*MRE11*	94180441	94180442	G	A	stop_gained	c.1735C>T	p.Arg579*	0.003	Heterozygous
1244	chr11	*MRE11*	94200986	94200987	G	A	stop_gained	c.1099C>T	p.Arg367*	0.005	Heterozygous
200127	chr16	*PALB2*	23640534	23640535	G	T	stop_gained	c.2576C>A	p.Ser859*	0.000	Heterozygous
262114	chr16	*PALB2*	23649451	23649452	T	A	splice_region_variant	c.49-2A>T		0.000	Heterozygous
587	chr16	*PALB2*	23647355	23647358	ATC	A	frameshift_variant	c.509_510delGA	p.Arg170Ilefs*14	0.006	Heterozygous
101930	chr6	*POLH*	43555063	43555064	G	T	stop_gained	c.328G>T	p.Glu110*	0.000	Heterozygous
2946	chr8	*RECQL4*	145741630	145741632	GC	G	frameshift_variant	c.871delG	p.Ala291Leufs*2	0.000	Heterozygous
2957	chr8	*RECQL4*	145738490	145738493	CAT	C	frameshift_variant	c.2492_2493delAT	p.His831Argfs*52	0.007	Heterozygous
2768	chr16	*SLX4*	3641254	3641255	G	C	stop_gained	c.2384C>G	p.Ser795*	0.000	Heterozygous
4430	chr9	*XPA*	100459470	100459471	G	GC	frameshift_variant	c.103dupG	p.Ala35Glyfs*27	0.000	Heterozygous
587	chr14	*XRCC3*	104165866	104165869	GAC	G	frameshift_variant	c.606_607delGT	p.Arg204Glyfs*18	0.001	Heterozygous

Table S10: Germline mutations in *ATM*, *PALB2* and other HR genes in the TCGA cohort (n=603).

Case ID	chrom	gene	start	end	ref	alt	impact	codon_change	aa_change	AF_EXAC (%)	genotype
TCGA_CRC_02	chr11	*ATM*	108186741	108186742	C	T	stop_gained	c.6100C>T	p.Arg2034*	0.000	Heterozygous
TCGA_CRC_14	chr11	*ATM*	108205831	108205832	T	C	missense_variant	c.8147T>C	p.Val2716Ala	0.004	Heterozygous
TCGA_CRC_05	chr11	*ATM*	108224607	108224608	G	A	splice_donor_variant	c.8786+1G>A		0.002	Heterozygous
TCGA_CRC_22	chr2	*BARD1*	215610565	215610566	G	A	stop_gained	c.1690C>T	p.Gln564*	0.005	Heterozygous
TCGA_CRC_26	chr15	*BLM*	91293264	91293267	ACT	A	frameshift_variant	c.772_773delCT	p.Leu258Glufs*7	0.004	Heterozygous
TCGA_CRC_27	chr15	*BLM*	91303903	91303904	C	G	stop_gained	c.1301C>G	p.Ser434*	0.001	Heterozygous
TCGA_CRC_28	chr15	*BLM*	91304244	91304245	C	T	stop_gained	c.1642C>T	p.Gln548*	0.018	Heterozygous
TCGA_CRC_29	chr15	*BLM*	91304244	91304245	C	T	stop_gained	c.1642C>T	p.Gln548*	0.018	Heterozygous
TCGA_CRC_30	chr15	*BLM*	91306245	91306246	C	T	stop_gained	c.1933C>T	p.Gln645*	0.004	Heterozygous
TCGA_CRC_08	chr17	*BRCA1*	41245089	41245091	TG	T	frameshift_variant	c.2457delC	p.Asp821fs	0.000	Heterozygous
TCGA_CRC_15	chr13	*BRCA2*	32914436	32914438	GT	G	frameshift_variant	c.5946delT	p.Ser1982fs	0.026	Heterozygous
TCGA_CRC_34	chr13	*BRCA2*	32936731	32936732	G	C	missense_variant	c.7878G>C	p.Trp2626Cys	0.002	Heterozygous
TCGA_CRC_10	chr17	*BRIP1*	59793411	59793412	G	A	stop_gained	c.2392C>T	p.Arg798*	0.015	Heterozygous
TCGA_CRC_23	chr8	*NBN*	90983440	90983446	ATTTGT	A	frameshift_variant	c.657_661delACAAA	p.Lys219fs	0.019	Heterozygous
TCGA_CRC_24	chr8	*NBN*	90983440	90983446	ATTTGT	A	frameshift_variant	c.657_661delACAAA	p.Lys219fs	0.019	Heterozygous
TCGA_CRC_25	chr8	*NBN*	90983440	90983446	ATTTGT	A	frameshift_variant	c.657_661delACAAA	p.Lys219fs	0.019	Heterozygous
TCGA_CRC_20	chr16	*PALB2*	23647107	23647108	T	TA	frameshift_variant	c.758dupT	p.Ser254fs	0.003	Heterozygous
TCGA_CRC_06	chr16	*PALB2*	23647355	23647358	ATC	A	frameshift_variant	c.509_510delGA	p.Arg170fs	0.006	Heterozygous
TCGA_CRC_13	chr16	*PALB2*	23647355	23647358	ATC	A	frameshift_variant	c.509_510delGA	p.Arg170fs	0.006	Heterozygous

Table S11: Germline mutations in *ATM*, *PALB2* and other HR genes in the Yurgelun et al. 2017 cohort (n=1058).

gene	codon change	amino acid change	gene (2)	codon change (2)	amino acid change (2)
BRCA1	c.68_69del	p.E23Vfs*17	*APC (p.Ile1307Lys)*	c.3920T.A	p.I1307K
BRCA1	c.68_69del	p.E23Vfs*17	*APC (p.Ile1307Lys)*	c.3920T.A	p.I1307K
BRCA2	c.3199del	p.T1067Kfs*10	*MLH1*	exons 16-19 deletion	
BRCA2	c.1796_1800del	p.S599X	*MUTYH (monoallelic loss)*	c.494A.G	Y165C
ATM	c.8934_8935del	p.E2979Afs*9			
ATM	c.7638_7646del	p.R2547_S2549del			
ATM	c.4632_4635del	p.Y1544X			
ATM	c.3760del	p.V1254Ffs*2			
ATM	c.802C>T	p.Q268X			
ATM	c.790del	p.Y264Ifs*12			
ATM	c.5570C>A	p.S1857X			
ATM	c.2413C>T	p.R805X			
ATM	c.2250G>A	p.K750K			
ATM	c.3480_3492dup	p.S1165Gfs*5			
BRCA1	c.5095C>T	p.R1699W			
BRCA2	c.7602del	p.C2535Vfs*16			
BRCA2	c.5946del	p.S1982Rfs*22			
BRCA2	c.4477G>T	p.E1493X			
BRCA2	c.3847_3848del	p.V1283Kfs*2			
BRCA2	c.8537_8538del	p.E2846Gfs*22			
BRCA2	c.8537_8538del	p.E2846Gfs*22			
BRIP1	c.2990_2993del	p.T997Rfs*61			
BRIP1	c.2379+1G>T				
BRIP1	c.1970del	p.G657Vfs*31			
NBN	c.657_661del	p.K219Nfs*16			
NBN	c.1142del	p.P381Qfs*23			
PALB2	c.2711G>A	p.W904X			
PALB2	c.751C>T	p.Q251X			

Table S12: Somatic inactivating mutations presumably affecting the wild-type allele of genes where germline mutations were detected.

Case ID	gene	germline mutation			Somatic LOH evaluation		
		impact	codon_change	aa_change	Large Deletion	Point mutation	LOH Call
283	*APC*	stop_gained	c.4128T>G	p.Tyr1376*	unknown	unknown	unknown
142275	*APC*	stop_gained	c.1495C>T	p.Arg499*	No	No	No
621	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	Yes	Yes (1 frameshift insertion)	Yes
1760	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	No	Yes (1 nonsense mutation)	Yes
3527	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	No	Yes (1 splice site)	Yes
3669	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	No	Yes (2 frameshift insertion)	Yes
4529	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	Yes	Yes (1 nonsense mutation)	Yes
4536	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	No	Yes (1 nonsense mutation)	Yes
200096	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	No	Yes (1 nonsense mutation, 1 frameshift insertion)	Yes
200198	*APC (p.Ile1307Lys)*	missense_variant	c.3920T>A	p.Ile1307Lys	No	Yes (1 nonsense mutation, frameshift deletion)	Yes
1221	*ATM*	frameshift_variant	c.3802delG	p.Val1268fs	Yes	No	Yes
1755	*ATM*	splice_region_variant	c.2921+1G>T		Yes	No	Yes
2760	*ATM*	frameshift_variant	c.6415_6416delGA	p.Glu2139Ilefs*6	No	Yes (1 nonsense mutation)	Yes
3645	*ATM*	stop_gained	c.829G>T	p.Glu277*	No	Yes (1 nonsense mutation)	Yes
200205	*ATM*	stop_gained	c.8307G>A	p.Trp2769*	No	Yes (1 splice mutation)	Yes
2967	*BAP1*	frameshift_variant	c.959dupG	p.Cys320fs	No	No	No
2775	*BARD1*	frameshift_variant	c.1932_1933delAT	p.Cys645fs	No	No	No
1743	*BLM*	stop_gained	c.1933C>T	p.Gln645*	No	No	No
3046	*BLM*	stop_gained	c.1933C>T	p.Gln645*	No	No	No
3181	*BLM*	frameshift_variant	c.2250_2251insAAAT	p.Leu751fs	No	No	No
3111	*BRCA1*	frameshift_variant	c.5329dupC	p.Gln1777fs	unknown	unknown	unknown
2265	*BRCA2*	missense_variant	c.7878G>C	p.Trp2626Cys	No	No	No
2406	*BRCA2*	frameshift_variant	c.4478_4481delAAAG	p.Glu1493Valfs*10	No	No	No
3444	*BRCA2*	frameshift_variant	c.3delG	p.Met1fs	No	Yes (1 missense mutation)	Yes
200127	*BRCA2*	stop_gained	c.5682C>G	p.Tyr1894*	No	No	No
204	*BRIP1*	frameshift_variant	c.840delT	p.His281Ilefs*8	No	No	No
200054	*BRIP1*	frameshift_variant	c.2990_2993delCAAA	p.Thr997Argfs*61	No	No	No
200245	*CHEK2*	frameshift_variant	c.1229delC	p.Thr410Metfs*15	unknown	unknown	unknown
50	*CHEK2*	frameshift_variant	c.1229delC	p.Thr410Metfs*15	Yes	No	Yes
430	*CHEK2*	missense_variant	c.1556C>T	p.Thr519Met	No	No	No
680	*CHEK2*	missense_variant	c.1556C>T	p.Thr519Met	No	No	No
207	*ERCC2*	missense_variant	c.1847G>C	p.Arg616Pro	No	Yes (1 frameshift deletion)	Yes
262114	*ERCC2*	missense_variant	c.1847G>C	p.Arg616Pro	Yes	No	Yes
3558	*ERCC3*	stop_gained	c.325C>T	p.Arg109*	No	No	No
3680	*ERCC4*	stop_gained	c.58C>T	p.Arg20*	No	No	No
2430	*FANCC*	stop_gained	c.1642C>T	p.Arg548*	No	No	No
3439	*FANCE*	frameshift_variant	c.334delA	p.Ser112Valfs*14	No	No	No
3048	*FANCL*	stop_gained	c.948T>G	p.Tyr316*	No	No	No
251	*GEN1*	frameshift_variant	c.347_348delAG	p.Glu116Valfs*20	No	No	No
200226	*GEN1*	frameshift_variant	c.1933_1936delAAAG	p.Lys645Cysfs*29	No	No	No
1244	*MRE11*	stop_gained	c.1099C>T	p.Arg367*	No	Yes (1 missense mutation)	Yes
2760	*MRE11*	stop_gained	c.1735C>T	p.Arg579*	No	No	No
68	*MSH2*	frameshift_variant	c.2523dupA	p.Glu842Argfs*4	No	Yes (1 nonsense mutation)	Yes
3225	*MSH6*	frameshift_variant	c.3959_3962delCAAG	p.Ala1320Glufs*6	No	Yes (1 frameshift insertion)	Yes
92	*MUTYH (monoallelic loss)*	missense_variant	c.1187G>A	p.Gly396Asp	No	No	No
213	*MUTYH (monoallelic loss)*	missense_variant	c.536A>G	p.Tyr179Cys	No	No	No
280	*MUTYH (monoallelic loss)*	missense_variant	c.1187G>A	p.Gly396Asp	No	No	No
353	*MUTYH (monoallelic loss)*	missense_variant	c.1187G>A	p.Gly396Asp	No	No	No
442	*MUTYH (monoallelic loss)*	splice_region_variant	c.933+3A>C		Yes	No	Yes
627	*MUTYH (monoallelic loss)*	missense_variant	c.1187G>A	p.Gly396Asp	No	No	No
2365	*MUTYH (monoallelic loss)*	missense_variant	c.536A>G	p.Tyr179Cys	No	No	No
2939	*MUTYH*	missense_variant	c.1187G>A	p.Gly396Asp	Yes	No	Yes

	(monoallelic loss)						
3227	*MUTYH (monoallelic loss)*	missense_variant	c.1187G>A	p.Gly396Asp	No	No	No
200193	*MUTYH (monoallelic loss)*	disruptive_inframe_deletion	c.1437_1439delGGA	p.Glu480del	No	No	No
227039	*MUTYH (monoallelic loss)*	missense_variant	c.1187G>A	p.Gly396Asp	No	No	No
587	*PALB2*	frameshift_variant	c.509_510delGA	p.Arg170Ilefs*14	No	No	No
200127	*PALB2*	stop_gained	c.2576C>A	p.Ser859*	No	No	No
262114	*PALB2*	splice_region_variant	c.49-2A>T		No	No	No
352566	*PMS2*	frameshift_variant	c.1831dupA	p.Ile611fs	Yes	No	Yes
390	*PMS2*	stop_gained	c.1687C>T	p.Arg563*	Yes	No	Yes
101930	*POLH*	stop_gained	c.328G>T	p.Glu110*	No	No	No
2946	*RECQL4*	frameshift_variant	c.871delG	p.Ala291Leufs*2	No	No	No
2957	*RECQL4*	frameshift_variant	c.2492_2493delAT	p.His831Argfs*52	No	No	No
2768	*SLX4*	stop_gained	c.2384C>G	p.Ser795*	No	No	No
200019	*TP53*	frameshift_variant	c.681delT	p.Asp228Thrfs*19	Yes	Yes (1 splice site)	Yes
200107	*TP53*	missense_variant	c.743G>A	p.Arg248Gln	Yes	No	Yes
4430	*XPA*	frameshift_variant	c.103dupG	p.Ala35Glyfs*27	No	No	No
587	*XRCC3*	frameshift_variant	c.606_607delGT	p.Arg204Glyfs*18	No	No	No