Rare loss of function variants in candidate genes and risk of colorectal cancer

Short Running title: *NTHL1*, *BRCA2*, and *BRIP1* associated with CRC/P

Elisabeth A. Rosenthal PhD[1], Brian H. Shirts MD, PhD[2], Laura M. Amendola MS[1], Martha Horike-Pyne MPH[1], Peggy D. Robertson PhD[3], Fuki M. Hisama MD[1,4], Robin L. Bennett MS[1], Michael O. Dorschner PhD[3,5,6], Deborah A. Nickerson PhD[3], Ian B. Stanaway PhD[7], Rami Nassir PhD[8], Kathy A. Vickers[9], Chris Li[9], William M. Grady MD[10,11], Ulrike Peters PhD, MPH[9,12] and Gail P. Jarvik MD, PhD[1,3,12] on behalf of the NHLBI GO Exome Sequencing Project[11].

1. Division of Medical Genetics, School of Medicine, University of Washington, Seattle, WA
2. Department of Laboratory Medicine, School of Medicine, University of Washington, Seattle, WA
3. Department of Genome Sciences, University of Washington, Seattle, WA
4. Department of Neurology, School of Medicine, University of Washington, Seattle, WA
5. Department of Pathology, School of Medicine, University of Washington, Seattle, WA
6. Department of Psychiatry & Behavioral Sciences, School of Medicine, University of Washington, Seattle, WA
7. Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA
8. Department of Biochemistry and Molecular Medicine, University of California Davis, Davis, CA
9. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA
10. Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, D4-100, Seattle, WA 98109, USA
11. Department of Medicine, University of Washington School of Medicine, Seattle, WA 98109, USA
12. Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA
13. See supplemental information

Corresponding author:
Elisabeth A. Rosenthal
erosen@uw.edu
University of Washington Medical Center
1705 NE Pacific St, Box 357720
Seattle, WA 98195

Abstract

Purpose: Although ~25% of colorectal cancer or polyps (CRC/P) cases show familial aggregation, current germline genetic testing identifies a causal genotype in the 16 major genes associated with high penetrance CRC/P in only 20% of these cases. As there are likely other genes underlying heritable CRC/P, we evaluated the association of variation at novel loci with CRC/P.

Methods: We evaluated 158 *a priori* selected candidate genes by comparing the number of rare potentially disruptive variants (PDVs) found in 84 CRC/P cases without an identified CRC/P risk associated variant and 2440 controls. We repeated this analysis using an additional 73 CRC/P cases. We also compared the frequency of PDVs in select genes among CRC/P cases with two publicly available data sets.

Results: We found a significant enrichment of PDVs in cases versus controls: 20% of cases vs. 11.5% of controls with ≥ 1 PDV (OR=1.9, p=0.01) in the original set of cases. Among the second cohort of CRC/P cases, 18% had a PDV, significantly different from 11.5% (p=0.02). Logistic regression, adjusting for ancestry and multiple testing, indicated association between CRC/P and PDVs in *NTHL1* (p=0.0001), *BRCA2* (p=0.01) and *BRIP1* (p=0.04). However, there was no significant difference in the frequency of PDVs at each of these genes between all 157 CRC/P cases and two publicly available data sets.

Conclusion: These results suggest an increased presence of PDVs in CRC/P cases and support further investigation of the association of *NTHL1*, *BRCA2* and *BRIP1* variation with CRC/P.

Key words: Colorectal cancer/polyps; genetic testing panels; *BRIP1; NTHL1*; *BRCA2*

INTRODUCTION

Colorectal cancer (CRC) is the third most common and second most lethal cancer in the United States(Howlader et al. 2016). In 2016, CRC was estimated to be newly diagnosed in ~134,000 individuals in the United States(Howlader et al. 2016). The lifetime risk of developing CRC is ~4.5%(Howlader et al. 2016). Without CRC screening, this risk doubles for individuals with a single affected first-degree relative, and increases further with additional affected first-degree relatives (Fuchs et al. 1994, 1669-1674). At present, about 5% of all CRC cases can be attributed to a pathogenic variant in genes known to be associated with increased CRC risk(Patel and Ahnen 2012, 428-438).

There are several high penetrance Mendelian genes whose germline variants are known to be associated with CRC, which generally develops from adenomatous polyps or serrated polyps (Levin et al. 2008, 130-160). Most of these conditions have an autosomal dominant mode of inheritance. Approximately 20% of CRC and/or adenomatous polyps (CRC/P) patients who have an identified pathogenic variant are diagnosed with Lynch syndrome [MIM 120435], which is caused by pathogenic variants in the mismatch repair genes *MSH2*[MIM 609309], *MLH1*[MIM 120436], *PMS2*[MIM 600259] and *MSH6*[MIM 600678] or epigenetic silencing of *MSH2* through deletion in the 3' exons of *EPCAM* [MIM 185535](Lynch and de la Chapelle 1999, 801-818; Ligtenberg et al. 2009, 112-117). Other autosomal dominant syndromes with high risk of CRC/P include familial adenomatous polyposis, distinguished by >100 to thousands of colorectal adenomatous polyps [MIM 175100] (*APC* [MIM 611731]), Li-Fraumeni [MIM 151623] (*TP53* [MIM 191170]), juvenile polyposis [MIM 174900] (*SMAD4* [MIM 600993], *BMPR1A* [MIM 601299]), Peutz-Jeghers syndrome [MIM 175200] (*STK11*[MIM 602216]), as well as *PTEN* [MIM 601728], *AKT1* [MIM 164730], and *PIK3CA* [MIM 171834] (Nishisho et al. 1991, 665-

669; Malkin et al. 1990, 1233-1238; Jenne et al. 1998, 38-43). Pathogenic variants that predispose to CRC/P have also been identified in *POLE* [MIM 174762] and *POLD1* [MIM 174761] (Palles et al. 2013, 136-144) and are inherited in an autosomal dominant pattern. Recessive mode of inheritance of CRC/P is attributable to pathogenic variants in *MUTYH* [MIM 604933], associated with tens to hundreds of polyps and possibly the more recently reported *NTHL1* [MIM 602656] and *MSH3* (Jones et al. 2002, 2961-2967; Weren et al. 2015, 668-671; Adam et al. 2016, 337-351).

In addition to the 5% of CRC/P cases that can be explained by a germline pathogenic variant, another ~20-25% of all CRC/P cases appear familial, but do not have a causal variant identified by current genetic testing (Lynch and de la Chapelle 1999, 801-818; Patel and Ahnen 2012, 428-438; PDQ Cancer Genetics Editorial Board 2002). Failure to detect a genetic etiology for these patients may be due to one of several reasons. First, the underlying cause may be sporadic in multiple relatives, rather than hereditary (Lichtenstein et al. 2000, 78-85), resulting in phenocopies. Second, germline genetic panel testing may detect variants of uncertain significance (VUS) which are truly pathogenic, but their pathogenicity is not yet established. Third, technical issues may result in a false negative genetic finding. Fourth, low penetrance common variants account for some portion of CRC/P (Peters et al. 2012, 217-234; Hes et al. 2014, 55-60). Finally, as pathogenic variants in other known cancer associated genes from broader panels have been found in such patients, there may be other highly penetrant pathogenic variants in genes underlying heritable CRC/P that have not been identified or validated and are not included on current CRC/P specific genetic testing panels (AlDubayan et al. 2018, 401-414).

Several thousand genes have been proposed to be involved in CRC/P risk. These include genes involved in DNA repair pathways and other cancer related biological pathways, and genes implicated by GWAS or familial linkage (Hes et al. 2014, 55-60; Gylfe et al. 2013, e1003876; Smith et al. 2013, 1026-1034). We set out to narrow this list of candidate genes to test for evidence of novel CRC/P associated genes. Identification of such genes would allow for improved diagnostic testing using evolving CRC/P germline genetic testing panels.

MATERIALS AND METHODS

Participants – Original Cohorts

Ninety-two participants with CRC/P and of European ancestry (EA) were ascertained from three sources: The Clinical Sequencing Exploratory Research consortium, New EXome Technology in Medicine study (CSER) (N=57) at the University of Washington (UW), the Northwest Institute of Genetic Medicine Family Polyps Study (PP) (N=10) at UW, and the Women's Health Initiative (WHI) (N=25). Participants from CSER and PP were ascertained through referral to the UW Genetic Medicine clinic and will be referred to jointly as UW participants. These patients were referred for clinical genetic counseling and genetic testing and tended toward early age of onset of CRC/P or a positive family history. The parent CSER study on high throughput sequencing of CRC/P patients did not enroll individuals for whom clinical usual care recommendations were to pursue a single gene test, rather than a broader gene panel (Gallego et al. 2016, 515-519). Research on these cases was approved by the biomedical IRB committee at the UW and participants granted permission for broad sharing of genomic and phenotypic data by informed consent documentation. WHI participants were ascertained through the observational WHI study(The Women's Health Initiative Study Group 1998, 61-109). Case status was defined as having a diagnosis of CRC before age 65 (UW, WHI), and/or ≥10 adenomatous polyps (lifetime total) (UW only) (Table 1). A convenience sample of 2512 EA control participants (ESP) was ascertained from the NHLBI GO Exome Sequencing Project (See web resources) (Amendola et al. 2015, 305-315). Individuals younger than 30 or with body mass index (BMI) > 50 were excluded as CRC/P rarely presents before age 30 and high BMI is associated with an increased risk for CRC/P. BMI ranged between 15 and 49.8 with a mean of 27.6. Cancer and polyp phenotypes for these ESP control individuals was unknown, therefore it is possible that some have a history of CRC/P. Demographic information for the controls is given in Table 1.

Participants – Subsequent Cohorts

A second cohort of deceased EA CRC patients was ascertained through the ColoCare study (CoCa, N=73) (Yuan et al. 2017, 1202-1210). These patients were newly diagnosed between the ages of 18 and 80, and had not had genetic testing (Table 1).

For further comparison, we used two additional, publicly available data sets. We used a cohort consisting of 7325 EA women from the FLOSSIES data set (see web resources) (Walsh et al. 2010, 12629-12633; Wang et al. 2015, 926-937; Li et al. 2008, 1100-1104). These women were over 70 years of age, had no history of a cancer diagnosis and were sequenced for genes on the BROCA panel, which includes genes known to be associated with CRC/P as well as genes associated with breast, ovarian, prostate, pancreatic, and renal cancers (Walsh et al. 2010, 12629-12633; Walsh et al. 2011, 18032-18037; Nord et al. 2011, 184; Metzker 2010, 31-46; Shirts et al. 2016, 974-981). Additionally, we used frequency summary data from the Exome Aggregation Consortium (ExAC), with the Cancer Genome Atlas (TCGA) subset removed (non-TCGA-ExAC) (see web resources). This data set consists of 53,105 individuals (27,173 EA) with whole exome data, who were ascertained at multiple sites for many different traits, excluding individuals from TCGA (Lek et al. 2016, 285-291). We used these cohorts and the cases to perform one-sided Fisher exact tests comparing the expected frequency of PDVs for some of the genes of interest.

Next Generation Sequencing and Genotype Calling

Cases were sequenced on Roche NimbleGen SeqCap EZ v3 (UW, CoCa) and Agilent SureSelect All Exon v5 (WHI). ESP Controls were sequenced on Agilent SureSelect Human All Exon Kit v2, Nimblegen RefSeq/CCDS design, or SeqCap EZ v1, divided among two centers: BROAD (N=1122) and Nickerson UW lab (N=1390). As the target regions differed, we focused attention on the regions covered by all targets. In order to avoid an effect of different sequencing centers used for the UW/WHI cases and ESP controls, on the analysis, we only included variants that were genotyped in >90% of the participants and had depth of coverage >20 for at least 80% of the participants. The UW/WHI cases and ESP control genotypes were jointly called, simultaneously, using Genome Analysis Toolkit (McKenna et al. 2010, 1297-1303; Van der Auwera, G A et al. 2013, 33; Poplin et al. 2017). Details of the genotyping and quality controls methods are in the supplemental methods. Individual level quality control measures resulted in removal of 18 ESP controls from the analysis, leaving 2494 ESP controls. The resulting, filtered, high confidence genotype data was annotated using SeattleSeqAnnotation138 (Ng et al. 2009, 272-276; Lek et al. 2016, 285-291). Further annotation was obtained from ClinVar and HGMD (Stenson et al. 2009, 13; Landrum et al. 2016, 862). Variants were determined to be pathogenic or likely pathogenic by expert panel, using the American College of Medical Genetics and Genomics guidelines and clarifications from Amendola et al. (Richards et al. 2015, 405-424; Amendola et al. 2015, 305-315). Eight UW/WHI case subjects (8.7%) and 54 ESP controls (2.2%) were removed from analysis due to having a known pathogenic or likely pathogenic variant in any of the following 16 genes known to be associated with increased CRC/P risk:

TP53, PTEN , *MSH2, MLH1, MSH6, PMS2, EPCAM* , *APC, MUTYH, STK11, SMAD4, BMPR1A, POLE, POLD1, AKT1,* and *PIK3CA* (Supplemental Table 1). Individuals heterozygous for a pathogenic variant in *MUTYH*, which is associated with autosomal recessive disease, were also excluded, in case a second pathogenic variant at this gene was missed. We did not evaluate the *CHEK2* gene [MIM 604373] as it has low penetrance for CRC/P (Naseem et al. 2006, 388-395). A total of 84 UW/WHI cases and 2440 ESP controls remained in the initial analysis. Of these, 15 UW/WHI cases (16%) and 604 ESP controls (25%) had a VUS in a known CRC/P gene (Supplemental Table 2). European ancestry of all cases was confirmed using principal components analysis (PCA) (see Supplemental Methods).

The same methods for joint genotyping, quality control, and confirmation of EA ancestry were used on the second case cohort. As the exome sequences for the CoCa cases were collected at a much later date, their genotypes were called separately from the first set of cases and the ESP controls. All 73 CoCa cases passed quality control and none of them harbored a pathogenic variant in any of the 16 known CRC/P genes listed above. Nineteen CoCa cases had at least one VUS in the 16 known CRC/P genes (Supplemental Table 2).

Genes and Variation of Interest

We focused the enrichment analysis on 158 genes collected from multiple sources and with varying supportive evidence for a role in CRC/P (Supplemental Table 3). This evidence derives from GWAS tagging SNPs (N=12), loss of function and linkage (N=12), known somatic involvement (N=2), or involvement in DNA repair pathways according to(Smith et al. 2013, 1026-1034) (N=133). Several of these genes are *a priori* known to be associated with other cancers. It is expected that those genes that have evidence derived from linkage or are involved in other cancers are more likely candidates. However, we chose to keep a broader set of genes to allow for lower penetrance genes. We limited our analyses to potentially disruptive variation (PDV) to focus on the most relevant variants, which may increase power to detect an effect. We defined a PDV as having a minor allele frequency (MAF) < 0.005 in the ESP controls and all published populations in gnomAD (Lek et al. 2016, 285-291) (See Web Resources), and to likely cause a change in the protein product, such as a stop gain (SG), frameshift (FS), splice acceptor (SA) or splice donor (SD) change. Early terminations (SG, FS) that resulted in a termination codon within 50 base pairs of the 3' end of the penultimate exon, or occurred within the last exon, were excluded from the analysis as they are expected to result in a functional protein.

Statistical Methods

We took a step-wise approach to testing for enrichment of PDVs in the 158 tested genes among the UW/WHI cases versus the ESP controls. First, we performed a global, one-sided test across all the considered genes, to test the alternative hypothesis that the number of PDV heterozygotes in UW/WHI cases would be higher than in the ESP controls. Second, we compared the distribution of the coding changes between these cases and controls. Finally, we used logistic regression to perform one-sided burden tests for each gene, adjusting for the first 5 principal components of ancestry (PCs) (see Supplemental Methods). We used a Bonferroni correction to determine significance for single gene comparisons.

We further tested the alternative hypothesis that the overall frequency of PDV heterozygotes in the second case cohort (CoCa) would be higher than the frequency found in the ESP controls, above. In this situation, we are not accounting for the variation in the estimated frequency in the controls, and are assuming the estimate is representative of EA individuals, in general.

Finally, we compared the frequency of PDVs in specific genes in all cases to the observed total frequency of PDVs in EA cohorts from the FLOSSIES and non-TCGA Exac data sets using Fisher's exact test. All statistical analyses were performed using the R 3.1.0 package(R Core Team 2016). All tests are one-sided, unless otherwise indicated.

RESULTS

PDVs in UW/WHI cases vs. ESP controls

The UW/WHI cases had a statistically significant greater proportion of PDV heterozygotes in the genes tested than the ESP controls, with an odds ratio (OR) of 1.9, 95% C.I. (1.2, ∞) (one-sided Fisher's exact test p-value = 0.02). Specifically, 17 cases (20%) had a PDV in a total of 16 genes (Table 2). One case had two PDVs: one in *BRCA2* [MIM 600185] and one in *NTHL1*. The same PDV in *NTHL1*, rs150766139, was found in 2 separate cases. Two genes, *POLQ* [604419] and *RECQL* [600537], had two unique PDVs each in the cases. In contrast, only 11.5% of controls (N=282) had a PDV in a total of 84 genes (Supplemental Table 4). Seventeen controls had two PDVs each, and 49 genes had multiple PDVs in the controls (Supplemental Figure 1). The *NTHL1* PDV observed in 2 cases (2.4%), rs150766139, was found in 4 controls (0.16%). Both *RECQL* variants observed in the cases were each seen in a control (0.04%) and one *POLQ* variant observed in the cases (NM_199420.3:c.2021dupA) was seen in one control (0.04%). Sensitivity analyses removing splice variants from the analysis gave similar results (OR = 2, 95% C.I. (1.1, ∞), p = 0.02), as did reducing the MAF cutoff to 0.001 (OR=2, 95% C.I. (1.1, ∞), p=0.01).

A total of 237 unique PDVs (93 FS; 99 SG; 20 SA; 25 SD) were observed in the UW/WHI cases and ESP controls in 84 (53%) genes. Although the cases have a higher proportion of FS than the controls (Supplemental Table 5), the distribution of PDV types is not statistically significantly different between cases and controls (p > 0.5). Seven PDVs (3 FS, 3 SG, 1 SD) were observed in both cases and controls.

PDVs in CoCa cases

Thirteen CoCa cases (18%) were heterozygous for a PDV, which is statistically significantly different from the 11.5% observed in the ESP controls (t-test p=0.02, Table 3). Of these 13 PDVs (6 FS; 6 SG; 1 SA) the *POLK* frameshift, NM_016218.2:c.1243delA, was also observed in the UW/WHI cases and in seven ESP controls.

Single gene tests in the UW/WHI cases and ESP controls

There were 16 genes with PDVs in the UW/WHI cases; we performed logistic regression analyses between the UW/WHI cases and ESP controls for each. Three genes were significantly associated with case status, adjusting for multiple testing using a Bonferroni correction: *NTHL1* (p=0.0001), *BRCA2* (p = 0.01) and *BRIP1* (p = 0.04). One gene, *RECQL*, had suggestive evidence for association with case status, based on unadjusted p-value of 0.02 (Bonferroni corrected p = 0.15). In all models, the first 5 PCs, were significant with difference in deviance of 327.77 on 5 degrees of freedom (χ^2-test p < 2e-16,).

Single gene tests in all cases vs FLOSSIES and ExAC controls

We further compared the frequency of PDVs at seven genes among all cases (UW/WHI/CoCa) with that observed in the FLOSSIES and non-TCGA ExAC data (Table 4). This list of genes included *BRIP1* and genes with at least two heterozygous cases and no, or few, heterozygous ESP controls. The total frequency of PDVs was not significantly higher in the cases than the Flossies or ExAC controls, for any single gene, under the assumption that the PDVs are independent of each other within each gene.

Non-CRC/P Actionable Findings

As the list of 158 genes studied here contains genes known to be associated with other cancers, it was possible to identify individuals heterozygous for a pathogenic or likely pathogenic variant in such genes (Tables 2 and 3, Supplemental table 4). This list of subjects includes 3 UW/WHI cases (*ATM*, *BRCA2*, *BRIP1*), 2 CoCa cases (*BRCA2*, *FANCM*) and 20 ESP controls (*ATM*, *BRCA1*, *BRCA2*, *BRIP1*, *NBN*).

DISCUSSION

Overall, both the UW/WHI and CoCa CRC/P case cohorts carried a significantly higher fraction of PDVs than controls in the 158 genes hypothesized to be associated with CRC/P. Three genes (*NTHL1*, *BRCA2,* and *BRIP1)* had variants associated with case status, adjusting for ancestry principal components and multiple testing. Support for an association between *NTHL1* and CRC/P via a recessive mode of inheritance was reported by others during the course of this study (Weren et al. 2015, 668-671; Helgason et al. 2015, 906-910). Perhaps incidentally, a known *BRCA2* pathogenic variant was identified in one of the two cases in our study with a PDV in *NTHL1*. Overall, our study supports the association of *NTHL1* with CRC/P, but our design does not address mode of inheritance. Interestingly, both *BRCA2* and *BRIP1* are associated with increased breast and/or ovarian cancer risk (Rafnar et al. 2011, 1104-1107; Ford and others 1998, 676-689; Antoniou et al. 2003, 1117-1130), and we found pathogenic variants for these genes in one female with CRC/P (age at last evalutaion 49 years) without a personal history of either cancer, as well as in two males with CRC/P (ages at last evaluation 59 and 66 years). This suggests that these genes may be associated with a cancer syndrome that includes CRC/P, possibly at lower penetrance. In addition, we found a PDV in each of *ATM* and *FANCM*, two genes known to be associated with autosomal dominant inheritance of breast cancer in two other CRC/P cases without a personal history of breast cancer: one male (age at last evaluation 75 years) and one female (age at last evaluation 46 years), respectively (Swift et al. 1987, 1289-1294; Kiiski et al. 2014, 15172-15177; Peterlongo et al. 2015, 5345-5355). These findings of an association with CRC/P for genes associated with breast and/or ovarian cancer are in alignment with that of AlDubayan et. al.. That study included a larger sample size and focused on a smaller set of genes known *a priori* to be associated with other heritable cancers and involved in DNA repair. In contrast, due to the small sample size of cases in our study, there is limited power to detect a significant difference in the frequency of PDVs for any single gene when comparing with the FLOSSIES or not-TCGA ExAC control data sets.

In addition to the small number of cases, our study is limited by reduced power. First, several cases and controls that remained in the analysis had a VUS in a gene known to be associated with CRC/P; a subset of these variants may be pathogenic. Second, exome sequencing can detect point mutations and small indels, but cannot detect genetic rearrangements or interference from pseudoegenes. Therefore, our analysis may have not detected some pathogenic variants. Third, the ESP control participants, for whom we did not have phenotypes, and the non-TCGA ExAC data set likely included individuals with CRC/P or other cancers which lowered our power to detect a difference in the frequency of PDVs between cases and controls. To address this issue, we removed ESP control individuals with pathogenic variants in known CRC/P genes. However, we were unable to clean the ExAC data beyond removing the TCGA individuals. Fourth, the single gene tests may also be limited by lower penetrance, lowering the power of the tests. One future strategy to increase power to test these genes is to perform joint linkage and association family testing, adjusting for the presence of the proband, where family members are available.

This work demonstrates support for considering additional CRC/P associated genes be included in CRC/P gene mutation panel tests on a research basis. The authors caution that these genes should be considered as research genes, in the context of CRC/P, and should not be used to direct clinical care for CRC/P at this time. Given the expectation of high locus heterogeneity with many genes having an effect on CRC/P risk, larger studies will be required to determine which genes account for the observed excess of PDVs in CRC/P cases vs. controls identified here.

Web Resources

esp.gs.washington.edu/drupal/dbGaP_Releases

https://whi.color.com/

ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/subsets

https://software.broadinstitute.org/gatk/documentation/article.php?id=3893

http://gnomad.broadinstitute.org/

Supplementary information is available at the Human Gentics website.

Acknowledgements

The authors thank Karynne Patterson for help obtaining the ESP raw data. The authors thank Niha Zubair and Tabitha Harrison for help obtaining the WHI raw data. This work was supported by the NHGRI and NCI Clinical Sequencing Exploratory Research program, UO1 HG006507 and HG007307. The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, HHSN268201600004C, HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf. Whole exome sequencing of WHI was funded via X01-HG006196. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO) and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). The ColoCare study is funded by grants 2P30CA015704-40; RO1CA194663, RO1CA189184, UO1152756, Seattle Tumor Translational Research Program, and the Cottrell Family Fund. Additional acknowledgements are in the Supplemental Information.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Tables

Table 1: Demographic data for cases and controls

Cases	N	N Male (%)	N Family History (%)	Age range (mean)	N CRC (%)	N P (%)
UW	67	39 (58)	44 (66)	18,64 (44)	39 (58)	46 (69)
WHI	25	0 (0)	25 (100)	55,64 (61)	25 (100)	1 (4)
Total	92	39 (42)	69 (75)	18,64 (48)	64 (70)	47 (51)
CoCa	73	41 (56)	UN	20,80 (57)	73	UN
ESP Controls						
Nickerson	1390	681 (49)	UN	30,92 (58)	UN	UN
BROAD	1122	413 (37)	UN	30,92 (56)	UN	UN
Total	2512	1094 (44)	UN	30,92 (57)	UN	UN

Age is minimum age of diagnosis for CRC/P (UW, WHI, CoCa) or adenomatous polyps (UW) in cases and age of ascertainment for controls. N=count; CRC=colorectal cancer; P=any adenomatous polyps; UN=Unknown; UW=University of Washington samples (CSER and PP); WHI=Women's Health Initiative samples; CoCa=ColoCare.

Table 2: Observed PDVs in suspected CRC/P genes in UW/WHI cases. Func=function. FS=frameshift. SG=stop gain. SA=splice acceptor. SD=splice donor. CRC=colorectal cancer, P=Polyps; Number of adenomas. Age=age at diagnosis. NC = number of ESP controls with the same variant. Variants in bold are known to be pathogenic for a different cancer.

Gene	Chr.	Pos.	rsID	Func.	HGVS-name	Phenotype (Age)	NC
ATM	11	108216600	876658716	SG	**NM_000051.3:c.8549T>A**	P25(64)	0
BRCA2	13	32932022	80359679	FS	**NM_000059.3:c.7762delA**	^{A}P10 (34)	0
BRIP1	17	59853848	NA	FS	**NM_032043.2:c.2010dupT**	CRC/P>10 (43)	0
CCDC18	1	93680404	NA	SG	NM_206886.3:c.1600C>T	CRC (56)	0
DCLRE1A	10	115601320	NA	SA	NM_001271816.1:n.3381G>A	CRC (33)	0
ERCC2	19	45855877	NA	SG	NM_000400.3:c.1933C>T	P>14 (50)	0
^{B}*NTHL1*	16	2096239	150766139	SG	NM_002528.5:c.268C>T	^{A}P10 (34); CRC (45)	4
NUDT7	16	77759403	200408443	SG	NM_001105663.2:c.111T>A	CRC (40)	10
PNKP	19	50365057	NA	FS	NM_007254.3:c.1253_1269dupGGGTCGCCATCGACAAC	P28 (30)	3
POLK	5	74882863	NA	FS	NM_016218.2:c.1243delA	P13 (58)	7
POLQ	3	121186441	NA	SG	NM_199420.3:c.6892C>T	CRC/P10 (42)	0
POLQ	3	121217455	NA	FS	NM_199420.3:c.2021dupA	P20 (44)	1
RAD50	5	131895029	NA	FS	NM_005732.3:c.186delA	CRC/P3 (31)	0
RECQL	12	21624504	199925437	SG	NM_002907.3:c.1525A>T	P10 (58)	1

RECQL	12	21636367	376839517	SG	NM_002907.3:c.643C>T	CRC/P3 (38)	1
TDG	12	104374735	NA	FS	NM_003211.4:c.478dup	CRC (40)	0
UACA	15	70957000	377649125	SD	NM_001008224.1:c.4074+1G>A	P100 (25)	1
WRN	8	30921820	NA	FS	NM_000553.4:c.229dupG	P23 (40)	0

[A]: Observed in presence of another PDV in same individual

[B]: Observed in two cases

Table 3: Observed PDVs in suspected CRC/P genes in CoCa cases. Func=function. FS=frameshift. SG=stop gain. SA=splice acceptor. SD=splice donor. Age=age at diagnosis of colorectal cancer. NC = number of ESP controls with the same variant. Variants in bold are known to be pathogenic for a different cancer.

Gene	Chr.	Pos.	rsID	Func.	HGVS-name	Age	NC
ALKBH3	11	43905557	1.45E+08	SG	NM_139178.3:c.208C>T	36	5
BRCA2	13	32914766	11571658	FS	**NM_000059.3:c.6275_6276delTT**	65	0
CCDC18	1	93646368	NA	FS	XM_005270815.1:c.282_283delCT, XM_005270816.1:c.282_283delCT	65	0
ERCC3	2	1.28E+08	rs774261851	FS	NM_000122.1:c.1757_1758delAG	48	0
ERCC3	2	1.28E+08	NA	SG	NM_000122.1:c.1300G>T	76	0
FANCM	14	45667921	1.45E+08	SG	NM_020937.2:c.5791C>T	44	5
LAMA5	20	60899562	NA	SG	NM_005560.4:c.5578C>T	76	0
MSH4	1	76345740	rs751781089	FS	NM_002440.3:c.1686delA	51	1
NEIL1	15	75641315	rs528340029	FS	NM_001256552.1:c.330_331insAGGC, NM_024608.3:c.72_73insAGGC	53	9
POLH	6	43555090	NA	SG	NM_006502.2:c.354C>G	50	0
POLK	5	74882863	rs773201725	FS	NM_016218.2:c.1243delA	69	1
PRKDC	8	48826626	NA	SA	NM_001081640.1:c.2618-2A>G	60	0
RAD50	5	1.32E+08	rs750586158	SG	NM_005732.3:c.3598C>T	72	0

Table 4: Comparison of total frequency of PDV alleles between all cases with the FLOSSIES and non-TCGA ExAC data sets. Q=total frequency of PDVs. N=number of individuals with data at each gene for the cases and Flossies data sets, and the minimum number of individuals for each gene in non-TCGA ExAC. P=unadjusted p-value

Gene	Qcases	QFlossies	NFlossies	PFlossies	Qexac	Nexac	Pexac
BRCA2	0.006	0.001	7325	**0.08**	0.002	18084	0.12
BRIP1	0.003	0.0008	7325	0.22	0.0007	15371	0.21
CCDC18	0.006	NA	NA	NA	0.002	17597	0.14
ERCC3	0.006	NA	NA	NA	0.001	19517	**0.05**
NTHL1	0.006	NA	NA	NA	0.003	16316	0.21
RAD50	0.006	NA	NA	NA	0.002	23607	0.19
RECQL	0.006	0.002	3646	0.1	0.002	23648	0.12

Bibliography

References

Adam, R., I. Spier, B. Zhao, M. Kloth, J. Marquez, I. Hinrichsen, J. Kirfel, A. Tafazzoli, S. Horpaopan, S. Uhlhaas, et al. 2016. *Exome sequencing identifies biallelic MSH3 germline mutations as a recessive subtype of colorectal adenomatous polyposis*. Vol. 99. United States: American Society of Human Genetics. Published by Elsevier Inc.

AlDubayan, S. H., M. Giannakis, N. D. Moore, G. C. Han, B. Reardon, T. Hamada, X. J. Mu, R. Nishihara, Z. Qian, L. Liu, et al. 2018. *Inherited DNA-repair defects in colorectal cancer*. Vol. 102. United States: American Society of Human Genetics. Published by Elsevier Inc.

Amendola, L. M., M. O. Dorschner, P. D. Robertson, J. S. Salama, R. Hart, B. H. Shirts, M. L. Murray, M. J. Tokita, C. J. Gallego, D. S. Kim, et al. 2015. *Actionable exomic incidental findings in 6503 participants: Challenges of variant classification*. Vol. 25. United States: Amendola et al.; Published by Cold Spring Harbor Laboratory Press.

Antoniou, A., P. D. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, N. Loman, H. Olsson, O. Johannsson, A. Borg, et al. 2003. *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies*. Vol. 72. United States: .

Ford, D., and others. 1998. *Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families*. Vol. 62.

Fuchs, Charles S., Edward L. Giovannucci, Graham A. Colditz, David J. Hunter, Frank E. Speizer, and Walter C. Willett. 1994. *A prospective study of family history and the risk of colorectal cancer*. Vol. 331Massachusetts Medical Society, http://dx.doi.org.offcampus.lib.washington.edu/10.1056/NEJM199412223312501.

Gallego, C. J., M. L. Perez, A. Burt, L. M. Amendola, B. H. Shirts, C. C. Pritchard, F. M. Hisama, R. L. Bennett, D. L. Veenstra, and G. P. Jarvik. 2016. *Next generation sequencing in the clinic: A patterns of care study in a retrospective cohort of subjects referred to a genetic medicine clinic for suspected lynch syndrome*. Vol. 25. United States: .

Gylfe, A. E., R. Katainen, J. Kondelin, T. Tanskanen, T. Cajuso, U. Hanninen, J. Taipale, M. Taipale, L. Renkonen-Sinisalo, H. Jarvinen, et al. 2013. *Eleven candidate susceptibility genes for common familial colorectal cancer*. Vol. 9. United States: .

Helgason, H., T. Rafnar, H. S. Olafsdottir, J. G. Jonasson, A. Sigurdsson, S. N. Stacey, A. Jonasdottir, L. Tryggvadottir, K. Alexiusdottir, A. Haraldsson, et al. 2015. *Loss-of-function variants in ATM confer risk of gastric cancer*. Vol. 47. United States: .

Hes, F. J., D. Ruano, M. Nieuwenhuis, C. M. Tops, M. Schrumpf, M. Nielsen, P. E. Huijts, J. T. Wijnen, A. Wagner, E. B. Gomez Garcia, et al. 2014. *Colorectal cancer risk variants on*

11q23 and 15q13 are associated with unexplained adenomatous polyposis. Vol. 51. England: .

Howlader, N., A. M. Noone, M. Krapcho, D. Miller, K. Bishop, S. F. Altekruse, C. L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, et al. 2016. *SEER cancer statistics review, 1975-2014*. Bethesda, MD: National Cancer Institute.

Jenne, D. E., H. Reimann, J. Nezu, W. Friedel, S. Loff, R. Jeschke, O. Muller, W. Back, and M. Zimmer. 1998. *Peutz-jeghers syndrome is caused by mutations in a novel serine threonine kinase*. Vol. 18. UNITED STATES: .

Jones, S., P. Emmerson, J. Maynard, J. M. Best, S. Jordan, G. T. Williams, J. R. Sampson, and J. P. Cheadle. 2002. *Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations*. Vol. 11. England: .

Kiiski, J. I., L. M. Pelttari, S. Khan, E. S. Freysteinsdottir, I. Reynisdottir, S. N. Hart, H. Shimelis, S. Vilske, A. Kallioniemi, J. Schleutker, et al. 2014. *Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer*. Vol. 111. United States: .

Landrum, M. J., J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, et al. 2016. *ClinVar: Public archive of interpretations of clinically relevant variants*. Vol. 44. England: . This work is written by (a) US Government employee(s) and is in the public domain in the US.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, et al. 2016. *Analysis of protein-coding genetic variation in 60,706 humans*. Vol. 536. England: .

Levin, B., D. A. Lieberman, B. McFarland, R. A. Smith, D. Brooks, K. S. Andrews, C. Dash, F. M. Giardiello, S. Glick, T. R. Levin, et al. 2008. *Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the american cancer society, the US multi-society task force on colorectal cancer, and the american college of radiology*. Vol. 58. United States: .

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. 2008. *Worldwide human relationships inferred from genome-wide patterns of variation*. Vol. 319. United States: .

Lichtenstein, P., N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki. 2000. *Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from sweden, denmark, and finland*. Vol. 343. United States: .

Ligtenberg, M. J., R. P. Kuiper, T. L. Chan, M. Goossens, K. M. Hebeda, M. Voorendt, T. Y. Lee, D. Bodmer, E. Hoenselaar, S. J. Hendriks-Cornelissen, et al. 2009. *Heritable somatic* *methylation and inactivation of MSH2 in families with lynch syndrome due to deletion of the 3' exons of TACSTD1*. Vol. 41. United States: .

Lynch, H. T., and A. de la Chapelle. 1999. *Genetic susceptibility to non-polyposis colorectal cancer*. Vol. 36. England: .

Malkin, D., F. P. Li, L. C. Strong, J. F. Fraumeni, C. E. Nelson, D. H. Kim, J. Kassel, M. A. Gryka, F. Z. Bischoff, M. A. Tainsky, and al et. 1990. *Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms*. Vol. 250, http://science.sciencemag.org/content/250/4985/1233.abstract.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. 2010. *The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data*. Vol. 20.

Metzker, M. L. 2010. *Sequencing technologies - the next generation*. Vol. 11. England: .

Naseem, H., J. Boylan, D. Speake, K. Leask, A. Shenton, F. Lalloo, J. Hill, D. Trump, and D. G. Evans. 2006. *Inherited association of breast and colorectal cancer: Limited role of CHEK2 compared with high-penetrance genes*. Vol. 70. Denmark: .

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure. 2009. *Targeted capture and massively parallel sequencing of 12 human exomes*. Vol. 461.

Nishisho, I., Y. Nakamura, Y. Miyoshi, Y. Miki, H. Ando, A. Horii, K. Koyama, J. Utsunomiya, S. Baba, and P. Hedge. 1991. *Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients*. Vol. 253, http://science.sciencemag.org/content/253/5020/665.abstract.

Nord, A. S., M. Lee, M. C. King, and T. Walsh. 2011. *Accurate and exact CNV identification from targeted high-throughput sequence data*. Vol. 12. England: .

Palles, C., J. B. Cazier, K. M. Howarth, E. Domingo, A. M. Jones, P. Broderick, Z. Kemp, S. L. Spain, E. Guarino, I. Salguero, et al. 2013. *Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas*. Vol. 45.

Patel, S. G., and D. J. Ahnen. 2012. *Familial colon cancer syndromes: An update of a rapidly evolving field*. Vol. 14. United States: .

PDQ Cancer Genetics Editorial Board. 2002. *Genetics of colorectal cancer (PDQ(R)): Health professional version*. PDQ cancer information summaries. Bethesda (MD): .

Peterlongo, P., I. Catucci, M. Colombo, L. Caleca, E. Mucaki, M. Bogliolo, M. Marin, F. Damiola, L. Bernard, V. Pensotti, et al. 2015. *FANCM c.5791C>T nonsense mutation* *(rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor*. Vol. 24. England: . Published by Oxford University Press.

Peters, U., C. M. Hutter, L. Hsu, F. R. Schumacher, D. V. Conti, C. S. Carlson, C. K. Edlund, R. W. Haile, S. Gallinger, B. W. Zanke, et al. 2012. *Meta-analysis of new genome-wide association studies of colorectal cancer risk*. Vol. 131. Germany: .

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Van der Auwera, Geraldine A, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, et al. 2017. *Scaling accurate genetic variant discovery to tens of thousands of samples*, http://biorxiv.org/content/early/2017/11/14/201178.1.abstract.

R Core Team. 2016. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rafnar, T., D. F. Gudbjartsson, P. Sulem, A. Jonasdottir, A. Sigurdsson, A. Jonasdottir, S. Besenbacher, P. Lundin, S. N. Stacey, J. Gudmundsson, et al. 2011. *Mutations in BRIP1 confer high risk of ovarian cancer*. Vol. 43. United States: .

Richards, S., N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, and ACMG Laboratory Quality Assurance Committee. 2015. *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*. Vol. 17. United States: .

Shirts, B. H., S. Casadei, A. L. Jacobson, M. K. Lee, S. Gulsuner, R. L. Bennett, M. Miller, S. A. Hall, H. Hampel, F. M. Hisama, et al. 2016. *Improving performance of multigene panels for genomic analysis of cancer predisposition*. Vol. 18. United States: .

Smith, C. G., M. Naven, R. Harris, J. Colley, H. West, N. Li, Y. Liu, R. Adams, T. S. Maughan, L. Nichols, et al. 2013. *Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer*. Vol. 34. United States: WILEY PERIODICALS, INC.

Stenson, P. D., M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper. 2009. *The human gene mutation database: 2008 update*. Vol. 1. England: .

Swift, M., P. J. Reitnauer, D. Morrell, and C. L. Chase. 1987. *Breast and other cancers in families with ataxia-telangiectasia*. Vol. 316. United States: .

The Women's Health Initiative Study Group. 1998. *Design of the women's health initiative clinical trial and observational study. the women's health initiative study group*. Vol. 19. United States: .

Van der Auwera, G A, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, et al. 2013. *From FastQ data to high* *confidence variant calls: The genome analysis toolkit best practices pipeline*. Vol. 43. United States: .

Walsh, T., S. Casadei, M. K. Lee, C. C. Pennil, A. S. Nord, A. M. Thornton, W. Roeb, K. J. Agnew, S. M. Stray, A. Wickramanayake, et al. 2011. *Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing*. Vol. 108. United States: .

Walsh, T., M. K. Lee, S. Casadei, A. M. Thornton, S. M. Stray, C. Pennil, A. S. Nord, J. B. Mandell, E. M. Swisher, and M. C. King. 2010. *Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing*. Vol. 107. United States: .

Wang, C., X. Zhan, L. Liang, G. R. Abecasis, and X. Lin. 2015. *Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation*. Vol. 96. United States: The American Society of Human Genetics. Published by Elsevier Inc.

Weren, R. D., M. J. Ligtenberg, C. M. Kets, R. M. de Voer, E. T. Verwiel, L. Spruijt, W. A. van Zelst-Stams, M. C. Jongmans, C. Gilissen, J. Y. Hehir-Kwa, et al. 2015. *A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer*. Vol. 47. United States: .

Yuan, Z., K. Baker, M. W. Redman, L. Wang, S. V. Adams, M. Yu, B. Dickinson, K. Makar, N. Ulrich, J. Bohm, et al. 2017. *Dynamic plasma microRNAs are biomarkers for prognosis and early detection of recurrence in colorectal cancer*. Vol. 117. England: .