

ORIGINAL ARTICLE

Role of rare variants in undetermined multiple adenomatous polyposis and early-onset colorectal cancer

Jérémy H Lefevre^{1,2,3,7}, Carolina Bonilla^{1,2,7,8}, Chrystelle Colas^{3,4}, Bruce Winney¹, Elaine Johnstone⁵, Susan Tonks¹, Tammy Day¹, Katarzyna Hutnik¹, Abdelhamid Boumertit¹, Florent Soubrier⁴, Rachel Midgley¹, David Kerr⁶, Yann Parc³ and Walter F Bodmer^{1,2}

Some 15–20% of multiple adenomatous polyposis have no genetic explanation and 20–30% of colorectal cancer (CRC) cases are thought to be due to inherited multifactorial causes. Accumulation of deleterious effects of low-frequency dominant and independently acting variants may be a partial explanation for such patients. The aim of this study was to type a selection of rare and low-frequency variants (<5%) to elucidate their role in CRC susceptibility. A total of 1181 subjects were included (866 controls; 315 cases). Cases comprised UK ($n=184$) and French ($n=131$) patients with MAP ($n=187$) or early-onset CRC ($n=128$). Seventy variants in 17 genes were examined in cases and controls. The effect of the variant effect on protein function was investigated *in silico*. Out of the 70 variants typed, 36 (51%) were tested for association. Twenty-one variants were rare (minor allele frequency (MAF) <1%). Four rare variants were found to have a significantly higher MAF in cases (EXO1-12, MLH1-1, CTNNB1-1 and BRCA2-37, $P<0.05$) than in controls. Pooling all rare variants with a MAF <0.5% showed an excess risk in cases (odds ratio = 3.2; 95% confidence interval = 1.1–9.5; $P=0.04$). Rare variants are important risk factors in CRC and, as such, should be systematically assayed alongside common variation in the search for the genetic basis of complex diseases.

Journal of Human Genetics (2012) 57, 709–716; doi:10.1038/jhg.2012.99; published online 9 August 2012

Keywords: adenomatous polyposis; APC; colorectal neoplasia; rare variant

INTRODUCTION

Familial adenomatous polyposis (FAP) and *MUTYH*-associated polyposis (MAP) are two inherited syndromes that show a high incidence of adenomatous polyps and an elevated risk of developing colorectal cancer (CRC). They account for a small fraction of CRC, <4%.¹ However, despite the fact that these two syndromes are caused by deleterious highly penetrant mutations in *APC* (GeneID 324)² and *MUTYH* (GeneID 4595),³ around 15–20% of patients with polyposis exhibit no known genetic risk factors. This is especially so for multiple polyposis patients who carry between 3 and 100 adenomatous polyps. In addition, 20–30% of CRC is thought to be due to inherited multifactorial causes.⁴ In the absence of identification of a new deleterious mutation, CRC may in part be due to the summation of the deleterious effects of a series of low-frequency dominant and independently acting variants of a variety of different genes,

each, conferring a moderate but readily detectable increase in relative risk.⁴ This ‘rare variant’ hypothesis was based upon the observation by Frayling *et al.*⁵ of the APC I1307K and E1317Q variants in patients with multiple adenomas. The I1307K variant is found in the Ashkenazi Jewish population at a frequency of ~6–7%, whereas it is absent from non-Jewish populations, and confers an increased risk of multiple adenomas and CRC.⁵ This variant implies an amino (isoleucine to lysine) substitution in a region involved in protein binding, leading to a mild dominant-negative effect. The E1317Q variant substitution may also affect the function of the APC protein presumed to translate into a slight but definitive advantage for the growth of a tumor.⁶

Following these observations, other rare variants have been tested. The candidate variants were selected because of their known involvement in sporadic or hereditary CRC or adenomas. Fearnhead *et al.*⁷

¹Department of Clinical Pharmacology, Cancer and Immunogenetics Laboratory, University of Oxford, Oxford, UK; ²Weatherall Institute of Molecular Medicine (WIMM), John Radcliffe Hospital, University of Oxford, Oxford, UK; ³Department of Digestive Surgery, Hôpital Saint-Antoine AP-HP, University Paris VI (Pierre and Marie Curie), Paris, France; ⁴Laboratory of Angiogenetics and Oncogenetics, Hôpital Pitié-Salpêtrière AP-HP, University Paris VI (Pierre and Marie Curie), Paris, France; ⁵Department of Clinical Pharmacology, Translational Oncology Group, Oxford, UK and ⁶Nuffield Department of Clinical Laboratory Sciences, John Radcliffe Hospital, University of Oxford, Oxford, UK

⁷These authors contributed equally to this work.

⁸Current address: School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK.

Correspondence: Professor WF Bodmer, Department of Digestive Surgery, Hôpital Saint-Antoine AP-HP, University Paris VI (Pierre and Marie Curie), 184, rue du Fbg Saint-Antoine, Paris 75012, France.

E-mail: walter.bodmer@hertford.ox.ac.uk

Received 26 April 2012; revised 4 July 2012; accepted 13 July 2012; published online 9 August 2012

observed a cumulative effect of 13 rare variants on five different genes in a cohort of 124 patients with adenomatous polyposis with an overall odds ratio (OR) of 2.2 ($P=0.0001$) when compared with a control set. Because of this publication, several variants in different CRC-susceptibility genes, such as *hMLH1* and *hMSH6*, have been reported to increase the risk of CRC but not cause Lynch syndrome,^{8,9} whereas *CHEK2* confers a higher CRC risk in hereditary non polyposis CRC (HNPCC)/HNPCC-related families.¹⁰

Rare variants are defined by a minor allele frequency (MAF) <1% in the general population and are unlikely to be identified by genome-wide association studies due to their low frequency and small contribution to the overall susceptibility of a disease.⁴ Only variants with a frequency >5% are detected in these large case-control association studies. Rare variants are best identified in studies with selected cases and candidate genes already known to be likely to be functionally relevant.^{1,5} Patients with early-onset CRC (before the age of 50) and multiple polyposis (3–100 polyps) with no known mutations in *APC* or *MYH* are ideal candidates to demonstrate an elevated predisposition to disease due to the accumulation of rare variants, as they are likely to involve inherited susceptibility.

The aim of this study was therefore to type a selection of rare (MAF <1%) and low-frequency variants (MAF 1–5%) in a relatively large set of patients with undetermined multiple polyposis (3–100 polyps) or early-onset CRC (diagnosed before 50 years of age) in order to elucidate the wider role of such variants in CRC susceptibility.

MATERIALS AND METHODS

A total of 315 cases and 866 controls, 1181 subjects in all, were included in this study. Collection of blood samples from cases and controls and clinicopathological information from patients were undertaken with appropriate individual informed consent and local ethical committee approvals.

Controls

The controls comprised 866 individuals collected in 10 different regions across the United Kingdom as part of the People of the British Isles study¹¹ (see below) and were unselected with respect to disease status.

Cases

The UK patient group consisted of 112 individuals with 3–100 histologically proven synchronous or metachronous adenomatous polyps and 72 individuals with CRC diagnosed before 50 years of age. Sixty-three individuals with early-onset disease were obtained through the VICTOR clinical trial, a phase III double-blind placebo-controlled study of rofecoxib in Dukes stage B or C CRC patients following potentially curative therapy, whereas the remaining nine cases were recruited through the John Radcliffe and Churchill hospitals' gastrointestinal clinics. With the exception of one Black Caribbean and one Indian individual, ethnic origin was White British for all UK patients for whom information was available. Non-white individuals were excluded from further analysis. No patient fulfilled the criteria for FAP, autosomal recessive MAP or HNPCC on clinical grounds. Some of these patients had already been screened for germline mutations in the *APC* and *MYH* genes in previous studies.^{5,12}

We also collected samples from 131 French patients, including 75 with multiple adenomas and 56 with early-onset CRC, who were recruited in the Department of Digestive Surgery at the Hospital Saint-Antoine in Paris using the criteria described above. Cases were selected from those who underwent a colectomy or total colectomy for CRC or polyposis. Patients diagnosed with CRC before the age of 50 or with more than three polyps detected after 2005 were referred for a consultation with the geneticist. Immunohistochemical staining to determine loss of expression of the genes *MLH1* and *MSH2* and microsatellite status was performed for all patients with a CRC diagnosed before the age of 50. Microsatellite instability was confirmed using PCR as

already described.¹³ Sequencing of the entire *MUTYH* and *APC* genes was carried out in patients with more than three adenomatous polyps.¹⁴ Only patients with no indication of HNPCC, MAP or FAP were included in this study. No ethnic identification was available for the French patients.

All the UK and French cases had histological confirmation of adenomatous polyps, but not all of them had the precise number of polyps determined. For 24 UK and 14 French adenoma patients, only 'multiple' was recorded. Within both the UK and French patient groups, individuals with attenuated FAP may be included, as they were not purposely eliminated from the study.

Variant selection

Rare and low-frequency variants were chosen based on prior literature reports that suggested a putative association with CRC, and with features related to colorectal disease, gastric cancer or other cancers (mostly of the breast and prostate). Variants in the *BRCA* genes were specifically selected from those that were classified either as non-pathogenic or as of unknown significance (see Breast Cancer Information Core database). Common variants, such as *CDH1* rs16260, *MTHFR* rs1801133 and *TP53* rs1042522, were genotyped because it has been suggested that they are associated with several types of cancer, including CRC.^{15–19}

DNA extraction and processing

Genomic DNA was extracted from patients' peripheral venous blood using the standard techniques. The People of the British Isles control blood samples were transported at room temperature to the laboratory, where the peripheral blood lymphocytes were separated under sterile conditions within 2 days of collection. DNA was prepared from the 10-ml blood residue remaining after sterile separation using either magnetic beads (GeneCatcherTM; Invitrogen, Carlsbad, CA, USA) or spin columns (Qiagen, Valencia, CA, USA). DNA concentration was determined using Pico Green²⁰ and normalized for genotyping to 25 ng μl^{-1} . Samples from the UK cases underwent whole-genome amplification because of limited volumes and amounts of genomic DNA. We used the Repli-g Mini kit (Qiagen), which implements a multiple-displacement amplification reaction to generate up to 10 μg of DNA per 50 μl reaction from a starting amount of at least 10 ng of genomic template. Genomic DNA from French cases and UK controls was used for genotyping.

Genotyping

We examined 70 variants, in cases and controls, in the following cancer candidate genes: *APC*, *AXIN1*, *AXIN2*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *CTNNB1*, *EPHB2*, *EXO1*, *MLH1*, *MLH3*, *MSH2*, *MTHFR*, *PMS2*, *SMAD4* and *TP53*, which were selected based on their involvement in familial cancers and the presence of somatic mutations in cancer. Of these, 16 variants were genotyped in a subset of only 227 controls. The complete list of variants analyzed is given in Supplementary Table 1.

Four variants were genotyped using restriction fragment length polymorphism analysis (*CDH1*-2, *CHEK2*-1, *CTNNB1*-1 and *MSH2*-8). Variants *APC*-10 and *APC*-11 (that is, I1307K and E1317Q) were typed using allele-specific PCR.⁷ For details of primers, enzymes and fragment sizes for these see Supplementary Table 2. Primers and conditions for all other variants are available upon request. Genotyping of the remaining variants was done using the Sequenom MassArray technology, namely matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry and the iPLEX Gold assay (Sequenom Inc., San Diego, CA, USA).²¹

Statistical analysis

Hardy–Weinberg equilibrium was assessed using an exact test implemented in the program PLINK v.1.07.²² Case-control association analyses were also conducted with PLINK. Two-sided *P*-values were calculated using Fisher's exact test and those <0.05 (with no multiple comparison correction) were considered statistically significant for an initial analysis. Combined ORs were estimated using the Mantel–Haenszel test.^{23,24}

Functional *in silico* analysis

We used the web-based programs PolyPhen-2 and SNPs&GO to predict the effect of nonsynonymous variants on protein function.^{25,26} FastSNP and F-SNP were similarly used for noncoding variants.^{27,28}

RESULTS

Populations

Clinical characteristics of the patients and controls are shown in Table 1.

Variant selection

Among the 70 variants examined, 24 (34%) were monomorphic in both the UK cases and controls and therefore not useful for analysis (Supplementary Table 1). Of the remaining 46, 10 were monomorphic in cases only and 5 were monomorphic only in controls. Thirty-one variants were considered rare having a control-population MAF <1%, seven were low-frequency variants (that is, MAF between 1 and 5%) and eight were common polymorphisms (that is, MAF >5%). If we define variant class based on the combined MAF in cases and controls as recently suggested,^{29,30} only one variant (APC-17) changes categories, going from being a low-frequency variant to a rare one.

No variant was out of Hardy–Weinberg equilibrium in the control population at a Bonferroni-corrected *P*-value of ≤ 0.001 (0.05/46). Two variants were in Hardy–Weinberg disequilibrium in controls (TP53-1 and BRCA1-6, *P* < 0.05) and three in patients (EPHB2-3, EXO1-12 and BRCA1-22, *P* < 0.05) if a correction for multiple testing was not applied.

Association analysis

UK cases vs controls. When comparing UK cases with controls, four rare variants were found to have a significantly higher MAF among the patients (EXO1-12, MLH1-1, CTNNB1-1 and BRCA2-37, *P* < 0.05; Table 2). Variant EXO1-12 was more frequent in individuals with cancer than in those with adenomas, as opposed to MLH1-1, CTNNB1-1 and BRCA2-37, which were only present in the multiple adenoma cases (Table 2). Results close to significance were also seen for rare variant EPHB2-3 and common variant CDH1-2 (*P* = 0.07 for both), although the CDH1-2 A allele appeared to protect against disease. When analyzing carrier frequencies instead of allele frequencies, only BRCA2-37 was significant, with an OR of 4.1 (1.2–14.3; *P* = 0.05; Table 3). Pooling together all the rare variants showed that the proportion of patients carrying rare variants was higher than the proportion of control carriers, regardless of whether the full set or a subset of controls was used (Table 3). The combined OR, obtained by merging OR 1 (effect of variants typed in the full set of controls) and OR 2 (effect of variants typed in the smaller set of controls) with the Mantel–Haenszel test, was 1.2 (95% confidence interval (CI) 0.8–1.8,

P = 0.42). This effect became much stronger when only variants with a MAF <0.5% were tested (combined OR 1.8; 95% CI 1.0–3.1; *P* = 0.05; Table 3). On the other hand, the analysis of pooled low-frequency variants showed a protective but nonsignificant effect (combined OR 0.8; 95% CI 0.5–1.1; *P* = 0.18; Supplementary Table 3). When APC-17 switched categories, from low frequency to rare variant, results changed slightly. For rare variants with MAF <1%, combined OR = 1.1 (95% CI 0.8–1.7; *P* = 0.54); for low-frequency variants, combined OR = 0.8 (95% CI 0.5–1.2; *P* = 0.24). Also, two variants (MLH3-1 and CHEK2-1) do not make the 0.5% cutoff when assessed from the frequency in the combined set of cases and controls. Taking them out of the analysis of variants with MAF <0.5% yields a combined OR of 1.6 (95% CI 0.8–2.9; *P* = 0.17).

UK multiple adenoma patients vs early-onset CRC patients. Analysis by disease group (that is, multiple adenoma vs early-onset patients) of all variants with frequencies <0.5% revealed an increase in susceptibility to disease for carriers of rare variants, especially among multiple adenoma patients (combined OR 1.9; 95% CI 1.0–3.5; *P* = 0.05; Table 4). Individually, the carrier frequency for BRCA2-37 in multiple polyp cases was significantly higher than that of controls, whereas this variant was absent among early-onset patients. MLH3-1, on the other hand, showed a significantly higher carrier frequency in the early-onset group as compared with controls, whereas BRCA2-27 was only detected in individuals with early-onset disease (Table 4). Overall, out of the 31 variants with MAF <1%, 14 were present in multiple adenoma patients only, whereas 3 (4, if counting APC-17) were present only in early-onset cases. The difference was significant whether APC-17 is included or not (*P* < 0.05), although the smaller number of early-onset CRC patients might be introducing bias.

There were significant allele-frequency differences between individuals with multiple adenomas and those with early-onset CRC in two variants in *MLH3* (one rare (MLH3-1) and one low frequency (MLH3-5)) and in a common variant in *CDH1* (CDH1-5) (Table 2). In these three instances, the allele frequency among early-onset patients was higher than among individuals with multiple adenomatous polyps.

Comparison between UK and French samples

Twenty-two of the variants genotyped in UK patients (16 rare, 4 low frequency and 2 common variants) were also examined in French subjects affected by either multiple polyps or early-onset CRC, recruited using the same set of criteria employed in the United Kingdom (Table 5). Three rare variants (MSH2-8, APC-10 and BRCA2-48) were absent from both, the UK and the French population. Eight variants that were detected in UK cases were not found in French patients (EPHB2-1, EPHB2-4, EPHB2-7, EXO1-4, CTNNB1-1, BRCA2-35, BRCA2-37 and CHEK2-1). On the other hand, no variant identified in French patients was missing from the UK sample (cases and/or controls). There was no significant difference between the French and UK cases with respect to the overall number of rare variants with MAF <1%, but UK patients show an excess of rare variants with MAF <0.5% compared with this set of French patients (*P* = 0.02; Table 6).

In silico analysis of functional effects

We investigated the putative effects of nonsynonymous variants with the programs PolyPhen-2 and SNPs&GO. Based on PolyPhen-2, 11 variants were classified as probably damaging, 6 as possibly damaging and 22 as benign, whereas according to SNP&GO, there were 18 disease variants and 21 neutral variants. Although these numbers

Table 1 Case and control sample description

	N	Mean age		Mean no. of	
		(years)	Male:female	polyps	CCR
UK multiple adenomas	112	59 ^a	68:20 ^b	11 ^b	9
UK early onset	70	42	38:31 ^c	n/a	
French multiple adenomas	75	51 ^d	44:31	26 ^e	34
French early onset	56	40	24:32	n/a	
PoBI controls	866	62	478:382 ^f	n/a	n/a

Abbreviation: PoBI, People of the British Isles; n/a, not applicable. Missing data for: ^a33, ^b24, ^c1, ^d3, ^e14, ^f6 individuals.

Table 2 Variants analyzed in UK cases and controls

<i>Id</i>	<i>Variant</i>	<i>dbSNP</i>	<i>Major/minor allele</i>	<i>MAF cases</i>	<i>MAF controls</i>	<i>P-value^a</i>	<i>MAF multiple adenomas</i>	<i>MAF early onset</i>	<i>P-value^b</i>	<i>PolyPhen-2/FastSNP</i>	<i>SNPs&GO</i>
MTHFR-1	A222V	rs1801133	C/T	0.338	0.336	0.95	0.337	0.342	0.94	Probably damaging	Disease
EPHB2-1	R80H	n/a	G/A	0.003	0.001	0.19	0.005	0.000	0.53	Probably damaging	Disease
EPHB2-3	I361V	rs56180036	A/G	0.007	0.001	0.07	0.010	0.000	0.37	Benign	Neutral
EPHB2-4	R568W	n/a	C/T	0.003	0.001	0.19	0.000	0.012	0.11	Probably damaging	Disease
<i>EPHB2-7</i>	M883V	n/a	A/G	0.003	0.000	0.22	0.005	0.000	0.53	Possibly damaging	Neutral
EXO1-2	E109K	n/a	G/A	0.000	0.001	0.66	0.000	0.000	n/a	Possibly damaging	Neutral
EXO1-12	D249N	rs61750993	G/A	0.020	0.007	0.03	0.018	0.024	0.74	Probably damaging	Neutral
<i>EXO1-4</i>	L410R	n/a	T/G	0.004	0.000	0.18	0.005	0.000	0.54	Probably damaging	Neutral
<i>EXO1-10</i>	G759E	rs4150001	G/A	0.003	0.009	0.38	0.000	0.013	0.11	Benign	Neutral
MLH1-1	G22A	rs41295280	G/C	0.003	0.000	0.03	0.005	0.000	0.53	Probably damaging	Neutral
CTNNB1-1	N287S	rs35288908	A/G	0.003	0.000	0.05	0.005	0.000	0.42	Benign	Disease
APC-7	L1129S	n/a	T/C	0.000	0.003	0.37	0.000	0.000	n/a	Possibly damaging	Disease
APC-11	E1317Q	rs1801166	G/C	0.011	0.007	0.46	0.014	0.007	0.59	Benign	Neutral
APC-15	G2502S	rs2229995	G/A	0.010	0.021	0.19	0.009	0.012	0.83	Benign	Neutral
APC-16	R2505Q	n/a	G/A	0.000	0.002	0.43	0.000	0.000	n/a	Probably damaging	Neutral
<i>APC-17</i>	S2621C	rs72541816	C/G	0.003	0.011	0.25	0.000	0.012	0.12	Benign	Neutral
<i>APC-20</i>	S636 C/A	n/a	C/A	0.024	0.019	0.66	0.024	0.025	0.96	n/a	n/a
PMS2-1	T511A	rs2228007	A/G	0.023	0.029	0.60	0.023	0.024	0.55	Benign	Neutral
<i>PMS2-2</i>	T597S	rs1805318	A/T	0.034	0.020	0.23	0.043	0.012	0.19	Benign	Neutral
<i>PMS2-3</i>	M622I	rs1805324	G/A	0.017	0.022	0.59	0.014	0.024	0.98	Benign	Neutral
BRCA2-7	N372H	rs144848	T/G	0.320	0.286	0.23	0.329	0.298	0.60	Benign	Disease
BRCA2-8	S384F	rs41293475	C/T	0.000	0.003	0.38	0.000	0.000	n/a	Possibly damaging	Disease
BRCA2-27	R2034C	rs1799954	C/T	0.004	0.004	0.90	0.000	0.012	0.12	Benign	Disease
<i>BRCA2-35</i>	D2665G	rs28897745	A/G	0.003	0.000	0.22	0.005	0.000	0.53	Probably damaging	Disease
BRCA2-37	V2728I	rs28897749	G/A	0.014	0.003	0.02	0.019	0.000	0.21	Benign	Disease
<i>MLH3-1</i>	V741F	rs28756990	G/T	0.014	0.004	0.16	0.005	0.036	0.04	Benign	Neutral
<i>MLH3-2</i>	M809V	rs61752722	A/G	0.003	0.009	0.37	0.005	0.000	0.53	Benign	Neutral
<i>MLH3-5</i>	S845G	rs28756992	A/G	0.014	0.033	0.09	0.005	0.036	0.04	Benign	Neutral
AXIN1-4	D495E	n/a	C/G	0.004	0.009	0.44	0.007	0.000	0.37	Benign	Neutral
AXIN1-6	R841Q	rs34015754	G/A	0.007	0.007	0.92	0.010	0.000	0.45	Probably damaging	Disease
CDH1-1	-1128	rs13335980	A/T	0.000	0.001	0.53	0.000	0.000	n/a	S8-binding site	n/a
<i>CDH1-2</i>	-284	rs16260	C/A	0.246	0.308	0.07	0.218	0.291	0.12	Affects TF-binding sites	n/a
CDH1-3	IVS1 + 6	rs3743674	T/C	0.102	0.121	0.37	0.116	0.071	0.27	Splicing site (medium-high risk)	n/a
CDH1-5	IVS4 + 10	rs33963999	G/C	0.070	0.080	0.57	0.051	0.119	0.04	Intronic enhancer (very low-low risk)	n/a
CDH1-7	A592T	rs35187787	G/A	0.003	0.007	0.45	0.005	0.000	0.53	Benign	Neutral
CDH1-8	T599S	n/a	C/G	0.000	0.001	0.65	0.000	0.000	n/a	Benign	Neutral
CDH1-10	A634V	n/a	C/T	0.000	0.001	0.52	0.000	0.000	n/a	Possibly damaging	Disease
TP53-1	R72P	rs1042522	G/C	0.255	0.248	0.80	0.270	0.220	0.38	Benign	Disease
BRCA1-6	Q356R	rs1799950	A/G	0.042	0.060	0.22	0.043	0.038	0.53	Possibly damaging	Disease
BRCA1-8	R496H	rs28897677	G/A	0.000	0.001	0.65	0.000	0.000	n/a	Benign	Disease
BRCA1-16	T826K	rs28897683	C/A	0.000	0.001	0.65	0.000	0.000	n/a	Probably damaging	Disease
BRCA1-22	E1038G	rs16941	A/G	0.338	0.342	0.90	0.346	0.317	0.64	Probably damaging	Disease
<i>BRCA1-28</i>	S1512I	rs1800744	G/T	0.003	0.004	0.82	0.005	0.000	0.86	Benign	Disease
AXIN2-1	N412S	rs115931022	A/G	0.000	0.007	0.19	0.000	0.000	n/a	Benign	Neutral
SMAD4-1	A118A	n/a	G/A	0.003	0.005	0.75	0.005	0.000	0.53	Synonymous	n/a
<i>CHEK2-1</i>	1100 delC	n/a	C/del	0.008	0.003	0.31	0.009	0.007	0.88	p.T367fsX15	n/a

Abbreviations: dbSNP, single nucleotide polymorphism database; MAF, minor allele frequency; n/a, not available; TF, transcription factor.

Number in bold: $P \leq 0.05$.

Italics: typed in a subset of 227 controls.

^aP-value for the comparison case vs control.^bP-value for the comparison multiple adenoma vs early-onset CRC.

Table 3 Rare variant counts in UK cases and controls

Rare variant	Cases ^a	Controls ^a	P-value
EPHB2-1	1/149	1/778	0.30
EPHB2-3	1/145	2/746	0.41
EPHB2-4	1/149	1/775	0.30
<i>EPHB2-7</i>	1/149	0/224	0.40
EXO1-2	0/150	1/751	1.00
<i>EXO1-4</i>	1/125	0/226	0.36
<i>EXO1-10</i>	1/145	4/225	0.65
EXO1-12	5/150	10/745	0.15
MLH1-1	1/147	0/740	0.17
CTNNB1-1	1/174	0/702	0.20
APC-7	0/149	4/743	1.00
APC-11	4/176	11/745	0.50
APC-16	0/150	3/729	1.00
BRCA2-8	0/147	4/748	1.00
BRCA2-27	1/142	6/740	1.00
<i>BRCA2-35</i>	1/149	0/224	0.40
<i>BRCA2-37</i>	4/148	5/744	0.05
<i>MLH3-1</i>	4/144	2/225	0.21
<i>MLH3-2</i>	1/148	4/227	0.65
AXIN1-4	1/117	14/749	0.71
AXIN1-6	2/146	11/746	1.00
CDH1-1	0/150	2/748	1.00
CDH1-7	1/147	11/747	0.70
CDH1-8	0/151	1/742	1.00
CDH1-10	0/150	2/731	1.00
BRCA1-8	0/151	1/735	1.00
BRCA1-16	0/149	1/744	1.00
<i>BRCA1-28</i>	1/149	2/225	1.00
AXIN2-1	0/128	10/737	0.37
SMAD4-1	1/148	7/739	1.00
<i>CHEK2-1</i>	3/178	1/179	0.37
<i>MAF < 1% (n = 31)</i>			
Total carriers/total noncarriers 1 ^b	24/146	108/739	
OR 1 (95% CI) ^c	1.13 (0.70, 1.81)		0.63
Total carriers/total noncarriers 2 ^b	13/146	13/217	
OR 2 (95% CI) ^c	1.49 (0.68, 3.24)		0.33
Combined OR (95% CI)	1.21 (0.80, 1.82)		0.42
<i>MAF < 0.5% (n = 23)</i>			
Total carriers/total noncarriers 1 ^b	11/149	41/740	
OR 1 (95% CI) ^c	1.33 (0.67, 2.65)		0.41
Total carriers/total noncarriers 2 ^b	11/146	5/215	
OR 2 (95% CI) ^c	3.24 (1.10, 9.52)		0.04
Combined OR (95% CI) ^c	1.77 (0.97, 3.08)		0.05

Abbreviations: CI, confidence interval; OR, odds ratio.

Number in bold: $P \leq 0.05$.

Italics: typed in a subset of 227 controls.

^aNumber of individuals with variant/total number of individuals typed.^bNumber of noncarriers corresponds to the harmonic mean of individuals without the rare variant for each variant typed.^cOR 1 was calculated pooling all variants typed in 866 controls, whereas OR 2 was calculated using a subset of 227 controls. Combined OR is the OR that results from combining OR 1 and OR 2 using a Mantel-Haenszel test.

seem fairly similar, there were several disagreements between programs with respect to the prediction of particular variants (Table 2). Among the rare variants with $MAF < 1\%$, there were 9 probably damaging, 5 possibly damaging and 14 benign, or 13 disease and 15 neutral variants. When only those variants with $MAF < 0.5\%$ were considered, the ratio of damaging (probably + possibly)/disease to

benign/neutral variants increased from ~ 50 to 60%. All of the low-frequency variants, in contrast, were predicted to be benign/neutral. In addition to the missense variants, there was one synonymous (SMAD4-1) and one deleterious coding (CHEK2-1) rare variants and five non-coding variants (one rare and one common variant in the promoter and two common intronic variants in *CDH1*, and one low-frequency variant in the 3' untranslated region of *APC*). Promoter variant CDH1-1 is predicted to eliminate a S8 transcription factor-binding site, whereas the CDH1-2 promoter variant A allele has been found to decrease transcriptional efficiency by 68% with respect to the C allele, also probably by altering transcription factor-binding sites.³¹ Using the programs FastSNP and F-SNP, a variant in *CDH1* intron 1 was determined to potentially affect a splicing site, whereas a variant in intron 4 of the same gene showed a low risk of being an intronic enhancer (Table 2).

DISCUSSION

We have examined 55 rare variants, 7 low-frequency variants and 8 polymorphisms in a sample of UK CRC and multiple adenoma cases and controls. Two of the four rare variants that were individually significantly associated with disease (that is, MLH1-1 and CTNNB1-1) had already been identified in the same set of individuals with multiple polyps, although not then found to be individually significant.⁷ In this study, we showed that these *MLH1* and *CTNNB1* variants were not present in a different and much larger UK control population, which explains the present case-control significant difference, and were also absent from a sample of early-onset CRC UK patients. Given that these two variants have not been found in our set of French patients, and that having $\sim 30\%$ fewer French cases may not in itself fully explain the UK-French apparent difference, they may represent UK founder effects, as previously suggested.⁴ However, replication of our findings in another UK sample of multiple adenoma patients, as well as functional studies, are necessary to establish their importance as CRC risk factors. The remaining two individually significant variants (EXO1-12 and BRCA2-37) have not been associated with CRC before. BRCA2-37 was classified as not clinically significant by the Breast Cancer Information Core database (in early 2010) and predicted to be benign by PolyPhen-2, yet it was recently found to be overrepresented among subjects with familial prostate cancer,³² and SNP&GO considered it to be disease-associated. As mentioned above, the fact that it was not found among French patients could indicate a restricted distribution of this variant. Variant EPHB2-3, which was detected in a Finnish individual with rectal and prostate cancer in an earlier study,³³ showed a nearly significant result. All associated variants code for nonsynonymous amino-acid changes. However, CTNNB1-1, BRCA2-37 and EPHB2-3 were predicted to be benign by PolyPhen-2, whereas MLH1-1 and EXO1-12 were considered probably damaging. SNPs&GO, on the other hand, predicted CTNNB1-1 and BRCA2-37 to be disease-associated and the remaining three variants to be neutral. Recently, SNPs&GO has been found to be more accurate than PolyPhen and other similar programs.³⁴ However, even though it identifies CTNNB1-1 and BRCA2-37 as potentially pathogenic, it misses MLH1-1 and EXO1-12. Also, APC-11, demonstrably pathogenic,⁵ was not identified as such by any of these computational methods. These discrepancies indicate that the use of *in silico* methods to evaluate the effects of nonsynonymous rare variants is not yet sufficiently reliable to be confident of their predictions. This is especially important when using them to predict which variants to focus on.

Table 4 Rare variant counts in UK multiple adenoma, early-onset CRC and control subjects for variants with MAF <0.5% in controls

Rare variant	Multiple adenomas ^a		Controls ^a	P-value ^b	P-value ^c
	Multiple adenomas ^a	Early onset ^a			
EPHB2-1	1/107	0/42	1/778	0.23	1.00
EPHB2-3	1/103	0/42	2/746	0.32	1.00
EPHB2-4	0/107	1/42	1/775	1.00	0.10
EPHB2-7	1/107	0/42	0/224	0.32	1.00
EXO1-2	0/108	0/42	1/751	1.00	1.00
EXO1-4	1/91	0/34	0/226	0.29	1.00
MLH1-1	1/105	0/42	0/740	0.12	1.00
CTNNB1-1	1/106	0/68	0/702	0.13	1.00
APC-7	0/107	0/42	4/743	1.00	1.00
APC-16	0/108	0/42	3/729	1.00	1.00
BRCA2-8	0/105	0/42	4/748	1.00	1.00
BRCA2-27	0/101	1/41	6/740	1.00	0.05
BRCA2-35	1/107	0/42	0/224	0.32	1.00
BRCA2-37	4/106	0/42	5/744	0.02	1.00
MLH3-1	1/102	3/42	2/225	1.00	0.03
CDH1-1	0/108	0/42	2/748	1.00	1.00
CDH1-8	0/109	0/42	1/742	1.00	1.00
CDH1-10	0/109	0/41	2/731	1.00	1.00
BRCA1-8	0/109	0/42	1/744	1.00	1.00
BRCA1-16	0/107	0/42	1/734	1.00	1.00
BRCA1-28	1/107	0/42	2/225	1.00	1.00
SMAD4-1	1/106	0/42	7/739	1.00	1.00
CHEK2-1	2/111	1/67	1/179	0.56	0.47
<i>Adenomas vs controls</i>					
Total carriers/total noncarriers 1 ^d	9/106		41/740		
OR 1 (95% CI) ^e	1.53 (0.72, 3.24)			0.26	
Total carriers/total noncarriers 2 ^d	7/103		5/215		
OR 2 (95% CI) ^e	2.92 (0.91, 9.43)			0.06	
Combined OR (95% CI)	1.87 (0.98, 3.48)			0.05	
<i>Early onset vs controls</i>					
Total carriers/total noncarriers 1 ^d		2/43	41/740		
OR 1 (95% CI) ^e		0.84 (0.20, 3.59)		0.81	
Total carriers/total noncarriers 2 ^d		4/42	5/215		
OR 2 (95% CI) ^e		4.10 (1.06, 15.89)		0.03	
Combined OR (95% CI)		1.72 (0.73, 5.27)		0.25	

Abbreviations: CI, confidence interval; OR, odds ratio.

Number in bold: $P \leq 0.05$.^aNumber of individuals with variant/total number of individuals typed.^bP-value for the comparison between multiple adenoma patients and controls.^cP-value for the comparison between early-onset CRC patients and controls.^dNumber of noncarriers corresponds to the harmonic mean of individuals without the rare variant for each variant typed.^eOR 1 was calculated using the larger set of controls, whereas OR 2 was calculated using a subset of 227 controls.

The grouping of all rare variants in the association analysis (23 or 8, depending on the control set used) yielded a combined OR of ~ 1.2 , which suggested that there was no strong evidence of an effect on CRC. However, pooling all variants with a MAF <0.5%

Table 5 Variants genotyped in French patients

Id	Variant	Major/minor allele	French cases ^a	UK cases ^a	Class
			(N = 131)	(N = 182)	
EPHB2-1	R80H	G/A	0.000	0.003	Rare variant
EPHB2-4	R569W	C/T	0.000	0.003	Rare variant
EPHB2-7	M883V	A/G	0.000	0.003	Rare variant
EXO1-4	L410R	T/G	0.000	0.004	Rare variant
EXO1-10	G759E	G/A	0.015	0.003	Rare variant
MSH2-8	E808X	G/T	0.000	0.000	Rare variant
CTNNB1-1	N287S	A/G	0.000	0.003	Rare variant
APC-10	I1307K	T/A	0.000	0.000	Rare variant
APC-11	E1317Q	G/C	0.004	0.011	Rare variant
APC-17	S2621C	C/G	0.008	0.003	Low freq variant
APC-20	8636	C/A	0.051	0.024	Low freq variant
PMS2-2	T597S	A/T	0.008	0.034	Low freq variant
BRCA2-7	N372H	T/G	0.273	0.320	Polymorphism
BRCA2-35	D2665G	A/G	0.000	0.003	Rare variant
BRCA2-37	V2728I	G/A	0.000	0.014	Rare variant
BRCA2-48	P3194Q	C/A	0.000	0.000	Rare variant
MLH3-1	V741F	G/T	0.016	0.014	Rare variant
MLH3-2	M809V	A/G	0.012	0.003	Rare variant
MLH3-5	S845G	A/G	0.042	0.014	Low freq variant
CDH1-2	-284	C/A	0.315	0.246	Polymorphism
BRCA1-28	S1512I	G/A	0.004	0.003	Rare variant
CHEK2-1	1100 delC	C/del	0.000	0.008	Rare variant

^aMinor allele frequency is shown.

considerably bolstered the association, taking the OR to ~ 1.8 . Notably, even though several of the variants included in the analysis are, on the basis of the *in silico* analysis and the examination of other parameters of pathogenicity,³⁵ considered to be benign, neutral, not clinically significant or of unknown significance, there is nevertheless an elevated risk from their combined action. The conclusion is that these variants may well be pathologically relevant, but that the *in silico* approaches are not yet adequate to detect this. The low-frequency variants (MAF between 1 and 5%) do not appear to influence susceptibility to CRC, as we described earlier for *CCND1*.³⁶ It is clear that further research is needed to evaluate more fully the role of low-frequency variants in cancer.³⁷ Defining rare variants using a threshold based on the combined set of cases and controls, as compared with just the controls, only altered the classification of three variants in our study and so did not appreciably affect the results. This was to be expected because we had a substantially larger number of controls than cases.

Extensively studied common variants MTHFR A222V (rs1801133) and TP53 R72P (rs1042522) did not show significant frequency differences between cases and controls. Conversely, CDH1-284C/A (rs16260) exhibited a lower frequency of the A allele in patients than in controls, revealing weak statistical evidence of a protective effect of this polymorphism on colorectal disease (0.25 vs 0.31, $P = 0.07$). This is in agreement with previous findings on CRC where the C allele increases risk,^{15,16} whereas in gastric, prostate and breast cancer, the

Table 6 Rare variant counts in UK and French patients

Rare variant id	UK cases ^a	French cases ^a
EPHB2-1	1/149	0/130
EPHB2-4	1/149	0/124
EPHB2-7	1/149	0/129
EX01-4	1/125	0/130
EX01-10	1/145	4/130
CTNNB1-1	1/174	0/130
APC-11	4/176	1/131
BRCA2-35	1/149	0/130
BRCA2-37	4/148	0/131
MLH3-2	1/148	3/129
MLH3-1	4/144	4/127
BRCA1-28	1/149	1/129
CHEK2-1	3/178	0/131
<i>MAF</i> < 1% (n = 13)		
Total carriers/total noncarriers ^b	24/149	13/128
OR (95% CI)	1.59 (0.71, 3.02)	0.20
<i>MAF</i> < 0.5% (n = 10)		
Total carriers/total noncarriers ^b	18/148	5/128
OR (95% CI) ^e	3.14 (1.13, 8.69)	0.02

Abbreviations: CI, confidence interval; OR, odds ratio.
Number in bold; $P \leq 0.05$.

^aNumber of individuals with variant/total number of individuals typed.

^bNumber of noncarriers corresponds to the harmonic mean of individuals without the rare variant for each variant typed.

A allele tends to be the risk allele.^{38–40} However, our study is underpowered for the detection of effects from common variants.

The analysis by disease group showed that, even though the collection of rare variants in each set of patients carries a higher risk of disease, our findings are mostly driven by the effects on individuals with multiple adenomas. BRCA2-37, the rare variant with the strongest effect in this study, was, for example, found only in patients with multiple adenomas. Although the sample size for the early-onset group was limited, our results clearly suggest that the genetic influence on CRC may mostly be seen in individuals with multiple adenomas, as compared with early-onset cases. This parallels to what is found in the clear-cut familial cases of inherited CRC. The extra layer of activity needed to go from polyp to cancer leads to an additional amount of variation that may be 'less genetically determined' and so obscure the underlying genetic susceptibility due to the multiple adenomas. There were, however, no significant differences in carrier frequencies between the two groups of patients, despite the fact that over half of the variants with $MAF < 1\%$ were found only in the multiple adenoma group. This, again, is probably due to the relatively smaller size of the early-onset group of patients. However, the allelic frequencies of two missense variants (one rare and one low frequency) in *MLH3* and one intronic common variant in *CDH1* differed significantly between multiple adenoma and early-onset CRC cases, with the latter exhibiting higher frequencies of these variants. This suggests that different sets of rare variants are quite likely to be involved in different pathologies, but to detect their effect would require larger numbers of patients than we were able to study.

The association *P*-values reported in this study have not been corrected for multiple hypotheses testing. Taking into account the number of variants analyzed, a Bonferroni correction would take the significance threshold to 0.001. Nonetheless, we believe that

because there is an *a priori* case for each candidate variant to be potentially functional, such a correction would be unsuitably stringent. The lack of French controls precluded a similar association study from being carried out with French samples as was done for the UK samples. Using UK controls would be inappropriate because of population stratification within Europe, especially for analysis of rare variants as they are likely to be population specific. The presence of such founder effects is appreciably suggested by the fact that the variants *MLH1-1* and *CTNNB1-1*, which are very clearly associated with multiple adenomas in UK cases, were not found in the French multiple adenoma cases. Further analysis of such differences requires larger numbers of French cases and appropriately selected French controls. Moreover, larger cohorts, such as the EPICOLON consortium experience,⁴¹ are needed to confirm these preliminary results and meta-analysis should be performed to ensure the pathogenic effects of variants described in the present work.

In summary, because rare variants appear to be associated with higher ORs than common variants, a relatively small study like ours can uncover the effects of candidate variants with low population frequencies on complex diseases such as, in this case, CRC. We have also shown that variants with frequencies $< 0.5\%$ appear to have the biggest effects regardless of the *in silico* prediction of their function. The role of the individual variant BRCA2-37 (V2728I) on the development of multiple adenomatous polyps deserves further examination. In general, the multiple adenoma phenotype seems to be more susceptible to genetic influence than early-onset CRC, but a larger early-onset patient sample would be necessary to confirm this finding. We have found some differences between UK and French patients in terms of the distribution of rare variants that justify closer inspection as population stratification within Europe can lead to spurious association results.

To conclude, we have confirmed that rare variants are important risk factors in CRC and as such, should be systematically assayed alongside common variation in the search for the genetic basis of complex diseases, taking great care to match cases with appropriate controls.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank all individuals who contributed samples for this study. We are also grateful to Shazad Asraf, Oliver Sieber and Haitao Wang for making samples available. We thank the patients who participated in VICTOR, without whose help, the study would not have been possible. The trial was enabled by an educational study grant from Merck & Co. Inc. who also supplied the study drug. VICTOR was endorsed by the UK National Cancer Research Institute (NCRI) and Cancer Research Campaign Clinical Trials Awards and Advisory Committee (CTAAC) allowing support from the UK National Cancer Research Network. The University of Birmingham was the center for randomization and for statistics initially before these moved to the Universities of Oxford and Warwick, respectively. We also thank VICTOR trial staff: A Stanley, J Stokes, F Duchesne, K Reed, S Pendlebury, R Duvvuri, I Kennedy, G Davis, J Birtwistle, L Blair, A Bange, K Bell, L Dunham, R Deacon, A Mellor, B Mikolajczyk, M Davoudianfar and A Ferch, and VICTOR laboratory support staff: K Elliot, M Prez and H Wang, Clinical Coordinator: D Rea, The VICTOR Trial Advisory Group and the VICTOR Data and Safety Monitoring Committee and all nurses and medical staff who participated in this trial. CB was Cancer Research UK Julia Bodmer Fellow. This work was made possible by a Cancer Research UK program grant, a Wellcome Trust grant and a Serroussi Foundation grant to WFB.

URLs:

- People of the British Isles study (PoBI): www.peopleofthebritishisles.org
- Breast Cancer Information Core (BIC) dataset: <http://research.nhgri.nih.gov/bic/index.shtml>
- PolyPhen v.2: <http://genetics.bwh.harvard.edu/pph2>
- PLINK: <http://pngu.mgh.harvard.edu/~purcell/plink/>
- SNPs&GO: <http://snps-and-go.biocomp.unibo.it/snps-and-go/>
- FastSNP: http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp
- F-SNP: <http://compbio.cs.queensu.ca/F-SNP>

Author contributions: CB, BW and WB designed research; JHL, CB, BW, EJ and ST performed research; CC, TD, KH, AB, WFB, RM, DK and YP collected samples; JHL and CB analyzed data; and JHL, CB and WFB wrote the paper.

- 1 Fearnhead, N. S., Wilding, J. L. & Bodmer, W. F. Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. *Br. Med. Bull.* **64**, 27–43 (2002).
- 2 Bodmer, W. F., Bailey, C. J., Bodmer, J., Bussey, H. J., Ellis, A., Gorman, P. et al. Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature* **328**, 614–616 (1987).
- 3 Al-Tassan, N., Chmiel, N. H., Maynard, J., Fleming, N., Livingston, A. L., Williams, G. T. et al. Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nat. Genet.* **30**, 227–232 (2002).
- 4 Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**, 695–701 (2008).
- 5 Frayling, I. M., Beck, N. E., Ilyas, M., Dove-Edwin, I., Goodman, P., Pack, K. et al. The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proc. Natl Acad. Sci. USA* **95**, 10722–10727 (1998).
- 6 Bodmer, W. Familial adenomatous polyposis (FAP) and its gene, APC. *Cytogenet. Cell Genet.* **86**, 99–104 (1999).
- 7 Fearnhead, N. S., Wilding, J. L., Winney, B., Tonks, S., Bartlett, S., Bicknell, D. C. et al. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl Acad. Sci. USA* **101**, 15992–15997 (2004).
- 8 Allan, J. M., Shorto, J., Adlard, J., Bury, J., Coggins, R., George, R. et al. MLH1 -93G>A promoter polymorphism and risk of mismatch repair deficient colorectal cancer. *Int. J. Cancer* **123**, 2456–2459 (2008).
- 9 Tulupova, E., Kumar, R., Hanova, M., Slysokva, J., Pardini, B., Polakova, V. et al. Do polymorphisms and haplotypes of mismatch repair genes modulate risk of sporadic colorectal cancer? *Mutat. Res.* **648**, 40–45 (2008).
- 10 Wasielewski, M., Vasen, H., Wijnen, J., Hooning, M., Dooijes, D., Tops, C. et al. CHEK2 1100delC is a susceptibility allele for HNPCC-related colorectal cancer. *Clin. Cancer Res.* **14**, 4989–4994 (2008).
- 11 Winney, B., Boumertit, A., Day, T., Davison, D., Echeta, C., Evseeva, I. et al. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur. J. Hum. Genet.* **20**, 203–210 (2012).
- 12 Lamlum, H., Al Tassan, N., Jaeger, E., Frayling, I., Sieber, O., Reza, F. B. et al. Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. *Hum. Mol. Genet.* **9**, 2215–2221 (2000).
- 13 Canard, G., Lefevre, J. H., Colas, C., Coulet, F., Srcek, M., Lascols, O. et al. Screening for lynch syndrome in colorectal cancer: are we doing enough? *Ann. Surg. Oncol.* **19**, 809–816 (2012).
- 14 Lefevre, J. H., Rodrigue, C. M., Mourra, N., Bennis, M., Flejou, J. F., Parc, R. et al. Implication of MYH in colorectal polyposis. *Ann. Surg.* **244**, 874–880 (2006).
- 15 Grunhage, F., Jungck, M., Lamberti, C., Berg, C., Becker, U., Schulte-Witte, H. et al. Association of familial colorectal cancer with variants in the E-cadherin (CDH1) and cyclin D1 (CCND1) genes. *Int. J. Colorectal Dis.* **23**, 147–154 (2008).
- 16 Pittman, A. M., Twiss, P., Broderick, P., Lubbe, S., Chandler, I., Penegar, S. et al. The CDH1-160C>A polymorphism is a risk factor for colorectal cancer. *Int. J. Cancer* **125**, 1622–1625 (2009).
- 17 Levine, A. J., Figueiredo, J. C., Lee, W., Poynter, J. N., Conti, D., Duggan, D. J. et al. Genetic variability in the MTHFR gene and colorectal cancer risk using the colorectal cancer family registry. *Cancer Epidemiol. Biomarkers Prev.* **19**, 89–100 (2010).
- 18 Le Marchand, L., Wilkens, L. R., Kolonel, L. N. & Henderson, B. E. The MTHFR C677T polymorphism and colorectal cancer: the multiethnic cohort study. *Cancer Epidemiol. Biomarkers Prev.* **14**, 1198–1203 (2005).
- 19 Polakova, V., Pardini, B., Naccarati, A., Landi, S., Slysokva, J., Novotny, J. et al. Genotype and haplotype analysis of cell cycle genes in sporadic colorectal cancer in the Czech Republic. *Hum. Mutat.* **30**, 661–668 (2009).
- 20 Ahn, S. J., Costa, J. & Emanuel, J. R. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Res.* **24**, 2623–2625 (1996).
- 21 Sauer, S. & Gut, I. G. Genotyping single-nucleotide polymorphisms by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **782**, 73–87 (2002).
- 22 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 23 Sveigaard, A., Jersild, C., Nielsen, L. S. & Bodmer, W. F. HL-A antigens and disease. Statistical and genetical considerations. *Tissue Antigen* **4**, 95–105 (1974).
- 24 Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl Cancer Inst.* **22**, 719–748 (1959).
- 25 Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
- 26 Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30**, 1237–1244 (2009).
- 27 Yuan, H. Y., Chiou, J. J., Tseng, W. H., Liu, C. H., Liu, C. K., Lin, Y. J. et al. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.* **34**, W635–W641 (2006).
- 28 Lee, P. H. & Shatky, H. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.* **36**, D820–D824 (2008).
- 29 Lemire, M. Defining rare variants by their frequencies in controls may increase type I error. *Nat. Genet.* **43**, 391–392 (2011).
- 30 Pearson, R. D. Bias due to selection of rare variants using frequency in controls. *Nat. Genet.* **43**, 392–393 (2011).
- 31 Li, L. C., Chui, R. M., Sasaki, M., Nakajima, K., Perinchery, G., Au, H. C. et al. A single nucleotide polymorphism in the E-cadherin gene promoter alters transcriptional activities. *Cancer Res.* **60**, 873–876 (2000).
- 32 Luedeke, M., Linnert, C. M., Hofer, M. D., Surowy, H. M., Rinckleb, A. E., Hoegel, J. et al. Predisposition for TMPRSS2-ERG fusion in prostate cancer by variants in DNA repair genes. *Cancer Epidemiol. Biomarkers Prev.* **18**, 3030–3035 (2009).
- 33 Kokko, A., Laiho, P., Lehtonen, R., Korja, S., Carvajal-Carmona, L. G., Järvinen, H. et al. EPHB2 germline variants in patients with colorectal cancer or hyperplastic polyposis. *BMC Cancer* **6**, 145 (2006).
- 34 Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32**, 358–368 (2011).
- 35 Spurdle, A. B. Clinical relevance of rare germline sequence variants in cancer genes: evolution and application of classification models. *Curr. Opin. Genet. Dev.* **20**, 315–323 (2010).
- 36 Bonilla, C., Lefevre, J. H., Winney, B., Johnstone, E., Tonks, S., Colas, C. et al. Cyclin D1 rare variants in UK multiple adenoma and early-onset colorectal cancer patients. *J. Hum. Genet.* **56**, 58–63 (2011).
- 37 Zhu, Q., Ge, D., Maia, J. M., Zhu, M., Petrovski, S., Dickson, S. P. et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.* **88**, 458–468 (2011).
- 38 Humar, B., Graziano, F., Cascinu, S., Catalano, V., Ruzzo, A. M., Magnani, M. et al. Association of CDH1 haplotypes with susceptibility to sporadic diffuse gastric cancer. *Oncogene* **21**, 8192–8195 (2002).
- 39 Qiu, L. X., Li, R. T., Zhang, J. B., Zhong, W. Z., Bai, J. L., Liu, B. R. et al. The E-cadherin (CDH1)-160 C/A polymorphism and prostate cancer risk: a meta-analysis. *Eur. J. Hum. Genet.* **17**, 244–249 (2009).
- 40 Beeghly-Fadiel, A., Shu, X. O., Long, J., Li, C., Cai, Q., Cai, H. et al. Genetic polymorphisms in the MMP-7 gene and breast cancer survival. *Int. J. Cancer* **124**, 208–214 (2009).
- 41 Castellvi-Bel, S., Ruiz-Ponte, C., Fernandez-Rozadilla, C., Abuli, A., Munoz, J., Bessa, X. et al. Seeking genetic susceptibility variants for colorectal cancer: the EPICOLON consortium experience. *Mutagenesis* **27**, 153–159 (2012).

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)