Human Molecular Genetics, 2006, Vol. 15, No. 21 ***3263–3271***
doi:10.1093/hmg/ddl401
Advance Access published on September 25, 2006

Search for low penetrance alleles for colorectal cancer through a scan of 1467 non-synonymous SNPs in 2575 cases and 2707 controls with validation by kin-cohort analysis of 14 704 first-degree relatives

Emily L. Webb[1,†], Matthew F. Rudd[1,†], Gabrielle S. Sellick[1], Rachid El Galta[1], Lara Bethke[1], Wendy Wood[1], Olivia Fletcher[2], Steven Penegar[1], Laura Withey[1], Mobshra Qureshi[1], Nichola Johnson[2], Ian Tomlinson[3], Richard Gray[4], Julian Peto[2,5] and Richard S. Houlston[1,*]

[1]Section of Cancer Genetics, Institute of Cancer Research, Surrey, UK, [2]The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK, [3]Molecular and Population Genetics Laboratory, Cancer Research UK, London, UK, [4]Birmingham Clinical Trials Unit, University of Birmingham, Birmingham, UK and [5]Non-Communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, London, UK

Received August 8, 2006; Revised and Accepted September 20, 2006

To identify low penetrance susceptibility alleles for colorectal cancer (CRC), we genotyped 1467 non-synonymous SNPs mapping to 871 candidate cancer genes in 2575 cases and 2707 controls. nsSNP selection was biased towards those predicted to be functionally deleterious. One SNP *AKAP9* M463I remained significantly associated with CRC risk after stringent adjustment for multiple testing. Further SNPs associated with CRC risk included several previously reported to be associated with cancer risk including *ATM* F858L [OR = 1.48; 95% confidence interval (CI): 1.06–2.07] and P1054R (OR = 1.42; 95% CI: 1.14–1.77) and *MTHFR* A222V (OR = 0.82; 95% CI: 0.69–0.97). To validate associations, we performed a kin-cohort analysis on the 14 704 first-degree relatives of cases for each SNP associated at the 5% level in the case–control analysis employing the marginal maximum likelihood method to infer genotypes of relatives. Our observations support the hypothesis that inherited predisposition to CRC is in part mediated through polymorphic variation and identify a number of SNPs defining inter-individual susceptibility. We have made data from this analysis publicly available at http://www.icr.ac.uk/research/research_sections/cancer_genetics/cancer_genetics_teams/molecular_and_population_genetics/software_and_databases/index.shtml in order to facilitate the identification of low penetrance CRC susceptibility alleles through pooled analyses.

INTRODUCTION

A recent twin study indicates that ~35% of colorectal cancer (CRC) can be ascribed to inherited susceptibility (1). Mendelian predisposition syndromes associated with mutations in known genes (*APC*, DNA mismatch repair genes, *MYH*, *SMAD4*, *ALK3* and *STK11*/*LKB1*), however, account for <5% of the overall incidence of the disease (2,3). The nature of the residual inherited susceptibility to CRC is at present undefined, but a model in which high-risk alleles account for all of the excess inherited risk seems improbable. One hypothesis about the allelic architecture of residual CRC susceptibility proposes that part of the genetic risk is caused by common, low penetrance alleles. The 'common-disease

*To whom correspondence should be addressed at: Section of Cancer Genetics, Brookes Lawley Building, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. Tel: +44 2087224175; Fax: +44 2087224359; Email: richard.houlston@icr.ac.uk
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

common-variant' hypothesis implies that testing for allelic association should be a more powerful strategy than genome-wide linkage for identifying low penetrance alleles (4).

Most association studies have focused on polymorphisms in genes involved in biologically defined mechanisms such as processing of ingested carcinogens and protection of DNA from carcinogen-induced damage (5). It is, however, likely that other as yet unrecognized genes will also influence tumour development. The spectrum of mutations in Mendelian disease genes, coupled with issues of statistical power, provides a compelling rationale for association analyses targeting non-synonymous SNPs (nsSNPs) (6).

We have sought to identify novel low penetrance susceptibility alleles for CRC by genotyping nsSNPs across 871 genes relevant to cancer biology, biasing selection of nsSNPs towards those likely to have deleterious consequences. Here, we report on the genotyping of 1467 nsSNPs in a large series of CRC cases and healthy population-based controls from the UK.

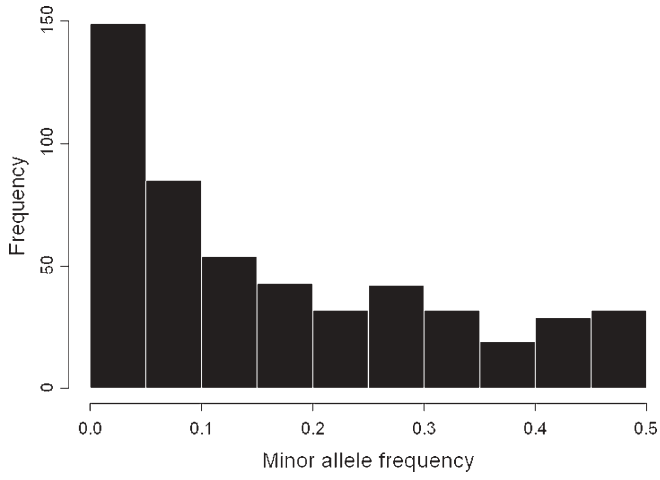

Figure 1. Distribution of MAFs of the 1047 SNPs assayed.

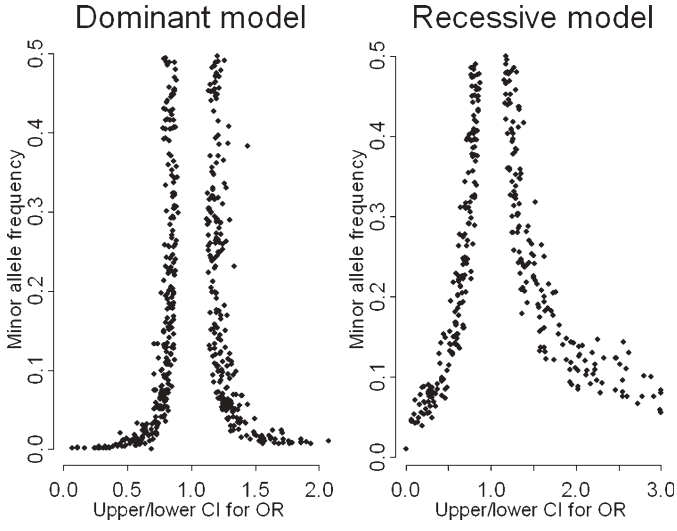

Figure 2. Upper and lower 95% CIs for risk of CRC associated with the 1041 SNPs in relation to MAF under dominant and recessive models.

RESULTS

Data quality and genotyping success

We genotyped 1467 nsSNPs in 2575 CRC patients and 2707 controls. Of the 5282 DNA samples submitted for genotyping, 5256 samples were successfully processed, generating in excess of four million genotypes. Genotypes were obtained for 2561 of 2575 cases (99.5%) and 2695 of 2707 controls (99.6%). Samples that failed to genotype had lower sample DNA concentrations than those that genotyped successfully (*t*-test; $P < 0.001$). SNP call rates per sample for each of the 5256 DNA samples were >99.6% in cases and controls. Genotype results concurred with data from the control genotyping performed in-house using Taqman. Of the 1467 SNPs submitted for analysis, 1218 SNPs were satisfactorily genotyped (83%) with mean individual sample call rates (the percentage of samples for which a genotype was obtained for each SNP) of 99.7 and 99.8% in cases and controls, respectively. Of the 1218 SNP loci satisfactorily genotyped, 171 were fixed in all samples, leaving 1047 SNPs for which genotype data were informative. Figure 1 shows the minor allele frequency (MAF) distribution of the 1047 SNPs.

Population stratification

Of the 1047 polymorphic nsSNPs, 55 were found to violate Hardy–Weinberg equilibrium (HWE) in controls at the 5% significance level (expected number of failures, 52). After Bonferroni correction, six SNPs still violated HWE and were removed, leaving a total of 1041 for further analysis. Each of the six SNPs removed had low genotyping reliability scores. Supplementary Table S1 details genotype data for each of the 1041 nsSNPs. None of the remaining 49 SNPs that violated HWE at the nominal 5% level was associated ($P < 0.05$) with risk of CRC. The estimate of the stratification parameter of the genomic control method was close to unity ($\hat{\lambda} = 1.02$; 95% confidence interval (CI): 0.87–1.20), indicating no evidence of population stratification as a cause of false positive results. Furthermore, no evidence was found for differences in allele frequencies of SNPs between male and female controls as a source of potential confounding in subsequent analyses.

Case–control analysis

Figure 2 shows upper and lower confidence limits for the ORs under dominant and recessive models plotted against the MAF for each of the 1041 SNPs. The CIs were wider for SNPs with low MAF, but a 50% difference in risk (OR < 0.5 or >1.5) could be excluded for most SNPs. Forty-four of the 1041 SNPs showed association at the 5% level based on the χ^2_{RD} statistic (Table 1). Twenty-five of these were more strongly associated under a dominant model and 19 under a recessive model. We also carried out tests for each SNP using an additive model and found that the results correlate strongly with those from the dominant model (correlation coefficient, 0.88). Although the most significant SNPs in this study were found using the dominant model, the importance of the

Table 1. SNPs showing a significant association with risk of CRC

SNP ID	Gene[a]	Substitution	MAF[b]	Simulated *P*-value	Model[c]	Case–control analysis			Kin-cohort analysis		
						OR	Lower CI	Upper CI	HR	Lower CI	Upper CI
rs6964587	*AKAP9*	M463I	0.38	1.0×10^{-4}	D	1.28	1.14	1.44	0.99	0.37	1.71
rs3206824	*DKK3*	G335R	0.23	2.0×10^{-3}	D	1.20	1.07	1.33	1.01	0.42	1.43
rs17602729	*AMPD1*	Q12X	0.13	3.0×10^{-3}	D	0.81	0.71	0.92	1.35	0.89	2.46
rs3829462	*LIPC*	L356F	0.02	2.0×10^{-3}	D	0.61	0.44	0.83	0.33	0.02	1.84
rs241419	*PSMB9*	V32I	0.03	0.01	D	0.73	0.58	0.92	0.49	0.07	1.36
rs17632786	*THBS1*	N700S	0.13	0.01	D	0.83	0.73	0.95	1.07	0.43	1.70
rs2297950	*CHIT1*	G102S	0.29	0.01	D	1.17	1.05	1.30	1.06	0.42	1.38
rs11700112	*PAK7*	R335P	0.09	0.02	D	0.82	0.71	0.95	1.50	0.82	2.01
rs1950902	*MTHFD1*	R134K	0.12	0.02	D	0.86	0.76	0.96	1.09	0.72	1.63
rs11976480	*AOAH*	D28N	0.32	0.02	R	1.27	1.06	1.52	1.10	0.17	3.67
rs2302465	*BST1*	R125H	0.13	0.02	R	1.73	1.15	2.61	0.48	0.00	1.32
rs1800057	*ATM*	P1054R	0.02	0.02	D	1.34	1.05	1.70	1.86	1.13	3.08
rs2274333	*CA6*	S90G	0.31	0.02	R	0.78	0.64	0.94	1.95	0.42	3.87
rs240780	*ASCC3*	C1995S	0.42	0.02	R	1.20	1.05	1.38	0.98	0.43	2.04
rs956868	*PRKWNK1*	P1056T	0.15	0.02	R	0.63	0.44	0.90	1.30	0.00	10.80
rs5743611	*TLR1*	R80T	0.09	0.02	R	0.40	0.19	0.84	0.00	0.00	0.01
rs3747517	*IFIH1*	R843H	0.26	0.03	R	1.30	1.06	1.59	1.86	0.67	3.21
rs9438	*DHX36*	S416C	0.39	0.03	D	1.15	1.03	1.29	1.25	0.69	2.44
rs13706	*CDC6*	V441I	0.11	0.03	D	0.85	0.74	0.97	0.50	0.13	0.85
rs1800056	*ATM*	F858L	0.01	0.03	D	1.47	1.04	2.08	1.67	0.46	6.07
rs1049550	*ANXA11*	R230C	0.41	0.03	D	1.15	1.03	1.29	1.10	0.60	1.83
rs2496425	*FREM2*	F1036S	0.28	0.03	R	0.77	0.63	0.95	1.32	0.08	5.49
rs3817552	*MYBPC1*	H506Q	0.15	0.03	R	1.56	1.08	2.26	3.24	0.45	11.19
rs2295778	*HIF1AN*	P41A	0.26	0.03	D	1.14	1.03	1.27	1.16	0.84	2.80
rs17470454	*DTNBP1*	P272S	0.05	0.03	R	0.28	0.09	0.84	11.33	0.00	12.63
rs1800076	*CFTR*	R75Q	0.04	0.03	R	5.29	1.16	24.15	0.22	0.00	5.24
rs3744581	*DNAH9*	N2195S	0.22	0.03	R	0.73	0.56	0.94	0.38	0.03	5.08
rs2228615	*ICAM5*	A348T	0.39	0.03	D	0.87	0.78	0.98	0.93	0.43	2.06
rs1716	*ITGAE*	R950W	0.33	0.03	D	1.14	1.02	1.27	1.02	0.54	1.90
rs1377210	*AGXT2L1*	S185P	0.09	0.03	R	2.05	1.12	3.76	2.74	0.00	9.56
rs616114	*MEP1B*	P695L	0.41	0.04	D	0.87	0.78	0.98	1.22	0.62	2.70
rs17742683	*MPP3*	R585G	0.10	0.04	D	1.18	1.03	1.35	0.48	0.20	0.81
rs1047840	*EXO1*	E589K	0.40	0.04	R	0.84	0.72	0.97	1.55	0.57	2.94
rs3218599	*REV3L*	D1734H	0.02	0.04	D	1.32	1.02	1.71	0.33	0.00	1.03
rs17704912	*MYO18B*	G44E	0.44	0.04	D	1.20	1.03	1.40	0.81	0.16	1.29
rs1800450	*MBL2*	G54D	0.14	0.04	R	1.65	1.07	2.55	0.91	0.00	9.86
rs2295275	*TRERF1*	C590S	0.06	0.04	D	1.22	1.03	1.44	1.09	0.48	1.71
rs3826007	*BCL2A1*	G82D	0.25	0.04	R	1.30	1.04	1.62	0.60	0.01	3.36
rs17128572	*GOLGA5*	A67G	0.08	0.05	D	0.83	0.71	0.98	1.22	0.61	2.26
rs17050550	*OGG1*	A85S	0.001	0.05	D	2.38	1.03	5.47	4.68	0.00	12.55
rs16937251	*NAV2*	Q468H	0.03	0.05	R	0.13	0.00	0.88	0.32	0.00	8.09
rs4667591	*LRP2*	L4210I	0.21	0.05	R	1.35	1.04	1.76	1.25	0.24	3.58
rs17129219	*MFI2*	A559T	0.002	0.05	D	0.32	0.11	0.99	2.18	0.00	4.57
rs1801133	*MTHFR*	A222V	0.34	0.05	R	0.82	0.69	0.98	0.62	0.25	4.22

[a]NCBI Entrez Gene.
[b]MAF in controls.
[c]Most significant association under a dominant (D) or recessive (R) model.

recessive model is evidenced by the fact that 19 of the top 44 SNPs were found using the recessive model and only seven of these were associated with $P < 0.05$ using an additive model. One SNP *AKAP9* M463I was still significantly associated with CRC risk after adjustment for multiple testing, with an adjusted *P*-value of 0.035. The number of SNPs that show association at the 5% significance level (44) is less than that would be expected (52) by chance due to Type I error, albeit not significantly different ($P = 0.26$).

ATM SNPs F858L (rs1800056) and P1054R (rs1800057) are in strong linkage disequilibrium LD ($D' = 1.0$, $r^2 = 0.51$). The haplotype formed by the minor alleles of *ATM* F858L and P1054R was significantly over-represented in cases, compared with controls ($OR_D = 1.47$, 95% CI: 1.04–2.08, $P = 0.06$ after permutation testing).

Stratification by gender, family history of the disease or age at diagnosis ($\leq$60 and $>$60 years) did not alter the overall findings. Interactions between the 93 SNPs that showed some association with risk ($P < 0.1$) were examined by fitting full logistic regression models for each pair, generating 4278 models and comparing these with the main effects model. Two hundred and seven pairs of SNPs showed nominally significant interaction at the 5% level, but even the strongest interaction between *ZNF318* T1112I and *HUS1B* H130Q ($P = 3.9 \times 10^{-5}$) was non-significant after correction for multiple testing.

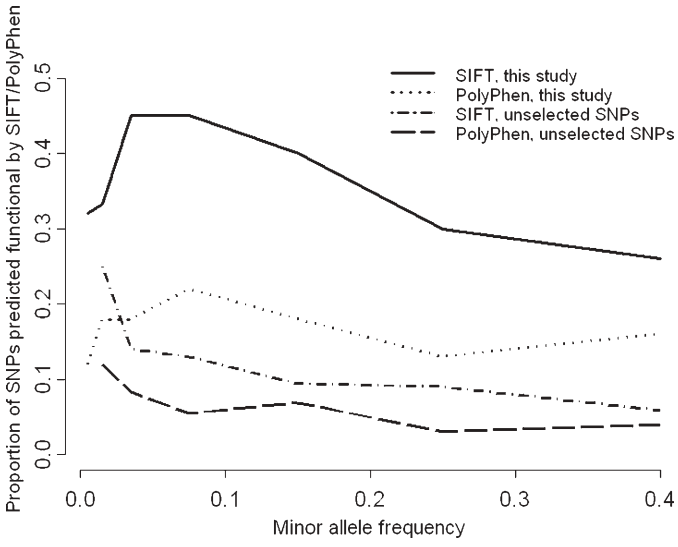

Figure 3. Relationship between MAF and predicted functionality of SNPs assayed. Values on the *y*-axis correspond to the proportion of nsSNPs predicted to be functional by SIFT (29) and PolyPhen (28) algorithms. For comparison, corresponding proportions in a series of 3009 unselected nsSNPs abstracted from the PICS database (24) are shown.

Functionality of nsSNPs

The proportion of the SNPs successfully genotyped predicted to have functional consequences was significantly higher (χ^2 test; $P < 0.001$) than expected from a comparable random series (Fig. 3). This result reflects the fact that SNPs were chosen to have a higher prior probability of being functional. Of the 44 SNPs showing significant association, three have been previously shown to be functional: P1054R in ataxia telangiectasia mutated [*ATM* (MIM 607585)], a cell cycle checkpoint kinase required for cellular response to DNA damage; N700S in thrombospondin 1 [*THBS1* (MIM 188060)], an anti-angiogenic protein thrombospondin and A222V in 5,10-methylenetetrahydrofolate reductase [*MTHFR* (MIM 607093)], a key enzyme in folate metabolism. One SNP encodes the termination codon Q12X in adenosine monophosphate deaminase 1 [*AMPD1* (MIM 102770)], and a further 20 SNPs are predicted by at least one *in silico* algorithm to be deleterious (Table 2).

Kin-cohort analysis

We performed a kin-cohort analysis on the first-degree relatives of cases for each of the 44 significantly associated SNPs. The 2561 cases that genotyped successfully reported a total of 14 704 first-degree relatives of whom 446 (3.0%) had been diagnosed with CRC. Twenty-three of the SNPs showed an association with CRC risk in relatives in the same direction as that seen in the case–control analysis (Table 1). Many of the SNPs had relatively low MAFs (i.e. <20%), so the number of CRCs in carriers was small and the power to verify an association limited. Nonetheless, three of the 23 SNPs were significantly associated with CRC risk in the kin-cohort analysis: *ATM* P1045R (OR = 1.86; 95% CI: 1.13–3.08), *TLR1* R80T (OR = 0.06; 95% CI: 0.03–0.12) and *CDC6* V441I (OR = 0.50; 95% CI: 0.14–0.78).

DISCUSSION

We evaluated nsSNPs on the basis that each has the capacity to directly affect the function of expressed proteins, implying a higher probability of being directly causally related to susceptibility. There is good evidence that three of the nsSNPs identified (*ATM* P1054R, *THBS1* N700S and *MTHFR* A222V) directly affect the function of the expressed protein and for an additional 21 of the SNPs significantly associated with CRC risk, the substitution either produced a termination codon or was predicted to be functionally deleterious by the *in silico* algorithms PolyPhen and SIFT (i.e. classified as 'probably damaging' or 'possibly damaging' by PolyPhen and 'intolerant' by SIFT). Although predictions about the functional consequences of amino acid changes are not definitive, these algorithms have been demonstrated in benchmarking studies to successfully categorize 80% of amino acid substitutions (7). One SNP *AKAP9* M463I remained significantly associated with CRC risk after stringent adjustment for multiple testing. Although AKAP9 has been implicated in the development of thyroid cancer by virtue of generation of a fusion protein with *BRAF* (8), to our knowledge this is the first evidence implicating *AKAP9* dysfunction in CRC.

The two SNPs in *ATM* associated with risk of CRC, F858L and P1054R, are predicted to be deleterious. Heterozygosity for P1054R is associated with decreased ATM expression in tumours (9), and cell lines from breast cancer patients harbouring the linked heterozygous F858L and P1054R variants exhibit increased radiosensitivity (10). *ATM* 1054R has previously been associated with an increased risk of breast and prostate cancers (11,12). *ATM* is critical for regulation of cell cycle checkpoints, and activation of *ATM* by DNA damage leads to *ATM*-dependent phosphorylation of *CHEK2*.

Several lines of evidence support a role for inherited dysfunction in the *ATM-CHEK2* axis in predisposition to CRC. In our study, further support for this hypothesis is provided by the fact that four cases were heterozygous for *CHEK2* I157T compared with none among controls ($P = 0.052$). *CHEK2* I157T is localized in a functionally important domain of *CHEK2*, and the variant protein has been shown to be defective in its ability to bind *TP53* (13) and *BRCA1* (14). An over-representation of CRC has been documented in relatives of A–T patients (15) and both *CHEK2* I157T (16) and the 1100delC allele (17) (which was not assayed for in our analysis) have been previously reported to increase risk of CRC.

In our study, homozygosity for *MTHFR* 222V was associated with reduced CRC risk. *MTHFR* catalyses the irreversible conversion of 5,10-methyleneTHF to 5-methylTHF. 5,10-methyleneTHF is used by thymidylate synthetase in the methylation of dUMP to dTMP, the sole *de novo* source of thymidine required for DNA synthesis and repair (18). Reduced availability of dTMP results in misincorporation of uracil into DNA, repair of which may lead to double-strand DNA breaks with carcinogenic consequences (19). The observed association between *MTHFR* 222V and CRC risk

Table 2. Description and predicted functionality of SNPs showing significant association with risk of CRC

SNP	Substitution	Predicted functionality[a]	Gene[b]	Gene description[b]	Gene ontology[c]	MIM[d]
rs6964587	M463I	Possibly damaging[e]	*AKAP9*	A kinase (PRKA) anchor protein 9	Receptor binding, signal transduction	604001
rs3206824	G335R	Intolerant[fg]	*DKK3*	Dickkopf homologue 3	Receptor signalling	605416
rs17602729	Q12X	*Stop codon*	*AMPD1*	Adenosine monophosphate deaminase 1	Nucleotide metabolism	102770
rs3829462	L356F	Probably damaging	*LIPC*	Lipase, hepatic	Lipid catabolism	151670
rs241419	V32I	Intolerant	*PSMB9*	Proteasome subunit, beta type, 9	Proteolysis and peptidolysis	177045
rs17632786	N700S	Possibly damaging/intolerant	*THBS1*	Thrombospondin 1	Cell adhesion, cell motility	188060
rs2297950	G102S	Possibly damaging/intolerant	*CHIT1*	Chitinase 1	Carbohydrate metabolism	600031
rs11700112	R335P	Possibly damaging	*PAK7*	p21(CDKN1A)-activated kinase 7	Protein amino acid phosphorylation	608038
rs1950902	R134K		*MTHFD1*	Methylenetetrahydrofolate dehydrogenase	Folic acid and derivative biosynthesis	172460
rs11976480	D28N	Potentially damaging[e]/potentially intolerant[fg]	*AOAH*	Acyloxyacyl hydrolase	Lipid metabolism	102593
rs2302465	R125H	Possibly damaging/borderline	*BST1*	Bone marrow stromal cell antigen 1	Protein catabolism, development	600387
rs1800057	P1054R	Probably damaging[e]/intolerant[fg]	*ATM*	Ataxia telangiectasia mutated	DNA repair, cell cycle control	607585
rs2274333	S90G	Potentially damaging	*CA6*	Carbonic anhydrase VI	One-carbon compound metabolism	114780
rs240780	C1995S	Probably damaging/intolerant	*ASCC3*	Activating signal cointegrator 1 complex subunit 3	Regulation of transcription	
rs956868	P1056T	Potentially damaging	*WNK1*	WNK lysine deficient protein kinase 1	Protein phosphorylation, ion transport	605232
rs5743611	R80T	Probably damaging	*TLR1*	Toll-like receptor 1	Immune cell activation	601194
rs3747517	R843H	Possibly damaging[e]	*IFIH1*	Interferon induced with helicase C domain 1	Regulation of apoptosis	606951
rs9438	S416C	Potentially damaging	*DHX36*	DEAH box polypeptide 36	Nucleic acid binding	
rs13706	V441I	Intolerant	*CDC6*	CRC6 cell division cycle 6 homologue	Cell cycle control	602627
rs1800056	F858L	Possibly damaging[e]/potentially intolerant[fg]	*ATM*	Ataxia telangiectasia mutated	DNA repair, cell cycle control	607585
rs1049550	R230C	Intolerant	*ANXA11*	Annexin A11	Protein binding	602572
rs2496425	F1036S		*FREM2*	Fras1-related extracellular matrix protein 2	Cell signalling	608945
rs3817552	H506Q	Probably damaging/intolerant	*MYBPC1*	Myosin binding protein C, slow type	Muscle contraction	160794
rs2295778	P41A	Possibly damaging	*HIF1AN*	Hypoxia-inducible factor 1, alpha subunit inhibitor	Regulation of transcription	606615
rs17470454	P272S	Possibly damaging/intolerant	*DTNBP1*	Dystrobrevin binding protein 1	Muscle development	607145
rs1800076	R75Q	Possibly damaging/intolerant	*CFTR*	Cystic fibrosis transmembrance conductance regulator	Protein binding	602421
rs3744581	N2195S	Possibly damaging/intolerant	*DNAH9*	Dynenin, axonemal, heavy polypeptide 9	Determination of bilateral symmetry	603330
rs2228615	A348T		*ICAM5*	Intercellular adhesion molecule 5, telencephalin	Cell–cell adhesion	601852
rs1716	R950W	Intolerant	*ITGAE*	Integrin, alpha E	Receptor signalling	604682
rs1377210	S185P	Probably damaging	*AGXT2L1*	Alanine-glyoxylate aminotransferase 2-like 1		
rs616114	P695L	Possibly damaging[e]	*MEP1B*	Meprin A, beta	Proteolysis and peptidolysis	600389
rs17742683	R585G	Probably damaging[e]/intolerant[f]	*MPP3*	Membrane protein, palmitoylated 3	Signal transduction	601114
rs1047840	E589K		*EXO1*	Exonuclease 1	DNA repair	606063
rs3218599	D1734H		*REV3L*	REV3-like, catalytic subunit of DNA polymerase zeta	DNA repair, DNA replication	602776
rs133885	G44E	Possibly damaging[e]	*MYO18B*	Myosin XVIIIB	Nucleotide binding	607295
rs1800450	G54D	Probably damaging/intolerant	*MBL2*	Mannose-binding lectin (protein C) 2	Oxidative stress response	154545
rs2295275	C590S	Probably damaging[e]	*TRERF1*	Transcriptional regulating factor 1	Regulation of transcription	
rs3826007	G82D	Possibly damaging/potentially intolerant	*BCL2A1*	BCL2-related protein A1	Regulation of apoptosis	601056
rs17128572	A67G	Potentially damaging/intolerant[fg]	*GOLGA5*	Golgi autoantigen, golgin subfamily a, 5	Receptor signalling	606918
rs17050550	A85S		*OGG1*	8-oxoguanine DNA glycosylase	DNA repair	601982
rs16937251	Q468H	Possibly damaging/potentially intolerant	*NAV2*	Neuron navigator 2		607026
rs4667591	L4210I	Intolerant	*LRP2*	Low density lipoprotein-related protein 2	Cell proliferation, protein binding	600073
rs17129219	A559T	Intolerant	*MFI2*	Antigen p97 (melanoma associated)		155750
rs1801133	A222V		*MTHFR*	Methylenetetrahydrofolate reductase	Folic acid and derivative biosynthesis	607093

[a]Functional predictions based on the *in silico* SIFT (29) (intolerant) and PolyPhen (28) (probably damaging and possibly damaging) algorithms.
[b]NCBI Entrez Gene.
[c]Gene ontology.
[d]Online Mendelian inheritance in man.
[e]PolyPhen predictions based on the alignment of fewer than six sequences.
[f]SIFT predictions with divergence scores >3.25.
[g]SIFT predictions based on the alignment of fewer than six sequences.

is not without precedent (5). We have recently conducted a meta-analysis of 23 previously published case–control studies that have reported the risk of CRC associated with homozygosity for *MTHFR* 222V. In this analysis of data on over 23 000 subjects, the pooled estimate of risk was 0.82 (95% CI: 0.75–0.91; $P < 0.0001$) (unpublished data), thereby supporting the risk estimate in the current analysis. As A222V SNP generates a thermolabile enzyme, with AA homozygotes having ~30% of normal enzyme activity (20), it is plausible that this SNP could influence CRC risk because reduced *MTHFR* activity results in increased availability of 5,10-methyleneTHF for DNA synthesis.

We used the Pathway Assist programme (Stratagene, La Jolla, CA, USA), to test whether any of the 44 associated SNPs occur in genes in the same pathways. Several map to genes encoding pivotal components of the DNA damage-response and cell-signalling pathways [e.g. *EXO1* (MIM 606063) and *OGG1* (MIM 601982)] or interact with *ATM-TP53* [e.g. *HIF1AN* (MIM 606615)] and, therefore, have biological plausibility with respect to CRC risk. Moreover, 38 of the 44 SNPs associated with risk of CRC map to genes expressed in colonic tissue (21), whereas a further four, *MTHFD1* R134K, *DHX36* S416C, *CDC6* V441I and *GOLGA5* A67G, are differentially expressed in colon adenocarcinomas (GNF SymAtlas v1.2.4).

The availability of detailed information on first-degree relatives of cases in our study allowed us to validate or refute associations identified in the case–control analysis by a kin-cohort analysis. Although not a formal substitute for an independent validation set, this analysis provided support for some loci identified in our case–control analysis such as *ATM* P1045R. Provided detailed family history has been ascertained from cases, the kin-cohort analysis provides an attractive method of deriving additional independent information from a conventional case–control analysis. Inferring genotypes of relatives adopting strategies such as the marginal likelihood approach obviates the need for relatives to have been genotyped. Inaccuracy in reported family histories is a theoretical limitation of this type of analysis; however, studies have shown that cancers such as CRC are generally reliably reported in first-degree relatives (22). Perhaps the major limitation of the kin-cohort approach as implemented here is that, despite 14 704 relatives being analysed, risk estimates are imprecise due to the small number of affected relatives and power will be limited for SNPs with low MAF.

Although we have only evaluated <5% of all validated nsSNPs, albeit prioritized on the basis of predicted functionality, our study provides evidence that inherited predisposition to CRC is in part mediated through low penetrance alleles. For a number of the SNPs we found to be associated with CRC, there is no precedent in the literature and our findings require confirmation in additional large data sets. However, several of our observations confirm previously reported associations, thereby strengthening inferences regarding the novel associations we identified.

Although our study is large compared with contemporaneous ones and for the majority of SNPs assayed our analysis was sufficiently powerful to exclude relative risks greater than 1.5 individually, it is acknowledged that many low penetrance alleles may confer more modest risks requiring far larger sample sets be analysed. In order to facilitate the identification of low penetrance alleles for CRC through analysis of nsSNPs, we have made data from our study publicly accessible through the following website http://www.icr.ac.uk/research/research_sections/cancer_genetics/cancer_genetics_teams/molecular_and_population_genetics/software_and_databases/index.shtml.

MATERIALS AND METHODS

Patients and control subjects

Two thousand five hundred and seventy-five patients with CRC, ascertained through an ongoing initiative at the Institute of Cancer Research/Royal Marsden Hospital NHS Trust (RMHNHST) from 1999 onwards (1474 males and 1101 females; mean age at diagnosis 59 years; SD ± 10.1), were included in the study. Cases were ascertained through either direct contact ($n = 2234$) or postal invitation ($n = 341$). CRC was defined according to the ninth revision of the International Classification of Diseases by codes 153–154 (23) and all cases had pathologically proven adenocarcinoma.

A total of 2707 healthy individuals were recruited as part of ongoing National Cancer Research Network genetic epidemiological studies (1999–2004; $n = 1075$), the Royal Marsden Hospital Trust/Institute of Cancer Research Family History and DNA Registry (1999–2004; $n = 1033$) and UK Study of Breast Cancer Genetics (1999–2004; $n = 599$) all established within the UK. Controls (836 males and 1871 females; mean age 59 years; SD ± 10.9) were the spouses or unrelated friends of patients with malignancies. None of the controls had a personal history of malignancy at the time of ascertainment. All cases and controls were British Caucasians, and there were no obvious differences in the demography of cases and controls in terms of place of residence within the UK.

Blood samples were obtained with informed consent and ethical review board approval in accordance with the tenets of the Declaration of Helsinki. Samples were collected, processed and stored under the same standardized protocol for all cases and controls. DNA was extracted from samples using conventional methodologies and quantified using PicoGreen (Invitrogen).

Selection of candidate genes and SNPs

We have previously established a publicly accessible Predicted Impact of Coding SNPs (PICS) database of potentially functional nsSNPs in genes with relevance to cancer biology (24). Briefly, candidate genes were identified by interrogating the Gene Ontology Consortium database (25), Kyoto Encyclopaedia of Genes and Genomes database (26), Iobion's Interaction Explorer PathwayAssist Programme, National Center for Biotechnology Information Entrez Gene database (27) and the CancerGene database. Both keyword and gene-pathway specific queries were performed using the following categories: catalytic activity; cellular processes, growth and death; development; enzyme regulator activity; folding, sorting and degradation; ligand–receptor interaction; nucleotide metabolism; physiological processes; regulation of

biological processes; replication and repair; signal transduction and signal transducer activity; transcription and transcription regulator activity; translation and translation regulator activity and transporter activity. A total of 9537 validated nsSNPs with MAF data were identified within 21 506 LocusLink annotated genes in NCBI dbSNP Build 123. Filtering this list and linking it to 7080 candidate cancer genes yielded 3666 validated nsSNPs with MAF ≥ 0.01 in Caucasian populations. The functional impact of each nsSNP was predicted using the *in silico* computational tools PolyPhen (28) and SIFT (version 2.1) (29). Using the PICS database and published work on resequencing of DNA repair genes (30–34), we prioritized a set of 1467 nsSNPs for the current study on the basis of SIFT and PolyPhen scores, indicative of deleterious consequences. Annotated flanking sequence information for each SNP was derived from the University of California Santa Cruz Human Genome Browser (Assembly hg17).

SNP genotyping and data manipulation

Genotyping of samples was performed using customized Illumina Sentrix Bead Arrays according to the manufacturer's protocols. DNA samples with GenCall scores <0.25 at any locus were considered 'no calls'. A DNA sample was deemed to have failed if it generated genotypes at fewer than 95% of loci. A SNP was deemed to have failed if fewer than 95% of DNA samples generated a genotype at the locus. To ensure quality of genotyping, a series of duplicate samples were genotyped and cases and controls were genotyped in the same batches. Conversion of genotype data into formats suitable for processing was performed using in-house Perl scripts (available upon request). All other statistical manipulations described were undertaken in S-Plus (version 7; Insightful, Com) or R (version 2.0.0).

Statistical methods

Population stratification. Genotypic frequencies in control subjects for each SNP were tested for departure from HWE using a χ^2 test or Fisher's exact test if an expected cell count was less than five. SNPs that violate HWE in the control population can indicate selection bias or genotyping errors and were removed from further analyses. To detect and control for possible population stratification, we employed the genomic control approach (35) using all SNPs to estimate the stratification parameter $\hat{\lambda}$ and its associated 95% CI. The possibility of gender differences as a source of population substructure in our controls was evaluated by χ^2 test or Fisher's exact test.

Risk of CRC associated with nsSNPs. The most efficient test of association depends on the true mode of inheritance of alleles. Test statistics calculated by combining the heterozygotes with the rare or common homozygotes and comparing these frequencies in cases and controls are most powerful under dominant and recessive modes of inheritance, respectively. We based our analyses on the statistic χ^2_{RD}, the higher of two χ^2 statistics obtained from dominant and recessive tests. This test statistic is not quite as powerful as if the most efficient test were used, but when the mode of action is not known, this loss of power is offset by the reduction in multiple testing. As the dominant test is strongly correlated with co-dominant statistics, it was deemed that the additional burden of multiple testing was not warranted. The distribution of χ^2_{RD} is non-standard, so significance levels were obtained using a Monte Carlo simulation approach implemented in the programming language C (source code available on request). Risks associated with the minor allele of each SNP were subsequently estimated by dominant or recessive odds ratios (ORs), dependent on the maximal mode of inheritance, using unconditional logistic regression; associated 95% CIs were calculated in each case. Where it was not possible to calculate ORs and their CIs by asymptotic methods, an exact approach was implemented using LogXact software (Cytel Inc., Cambridge, MA, USA). The test statistic χ^2_{RD} was also computed for each SNP within subgroups based on the family history of CRC and age at diagnosis, together with ORs and their associated 95% CIs. We present 95% CIs to give an indication of the precision of the estimates; however, these should not be used to infer global significance. Under certain conditions, a two-stage process incorporating estimates of pairwise interactions between significant SNPs can yield greater power to detect association (36). To investigate epistatic interactions, each pair of SNPs that displayed a significant association at the 10% level was evaluated by fitting a saturated logistic regression model and the log likelihood ratio statistic for comparison with the main effects model computed. This was compared against a χ^2 distribution with one degree of freedom. Statistics were then adjusted for multiple testing using a Bonferroni correction.

Correction for multiple testing in association studies using a simple Bonferroni correction is conservative due to the assumption of independence between tests. We therefore adopted an empirical Monte Carlo simulation approach (37) based on 10 000 permutations, thus allowing for correlations due to LD throughout the genome. At each iteration, case and control labels were permuted at random and the maximum test statistic $\max(\chi^2_{RD})$ determined. The significance level for each SNP was estimated as the proportion of permutation samples with $\max(\chi^2_{RD})$ larger than the observed value.

To assess the level of LD between SNPs, we calculated the pairwise LD measure D' between consecutive pairs of markers throughout the genome using the expectation–maximization algorithm to estimate two-locus haplotype frequencies. This information was used to investigate the relationship between haplotypes and disease status. Haplotypes were reconstructed using a Markov chain Monte Carlo method and their frequencies in case and control samples compared by permutation testing, using the programme PHASE (38,39).

Kin-cohort analysis. We also performed a kin-cohort analysis on the (un-genotyped) first-degree relatives of the CRC patients for each SNP significant at the 5% level. Data on first-degree relatives of probands were collected by questionnaire. For all relatives, follow-up ceased at diagnosis of CRC, death, age 80 years or completion of the interview or questionnaire by the index case, whichever was first. When age at diagnosis had not been accurately documented, current age or age at death was used as a proxy for age at diagnosis. In the kin-cohort analysis for each SNP, CRC risk in first-degree

relatives of carriers and non-carriers, proband genotypes and the allele frequency were used to infer the genotypes of relatives and hence to estimate the risk in carriers. Age-specific cumulative CRC distributions in first-degree relatives were estimated using a marginal likelihood approach (40), which has the advantage that it is robust in the presence of residual correlation between family members.

For each SNP, carrier status of the proband was defined by the most probable mode of inheritance for the SNP determined from the case–control study, i.e. for dominantly inherited SNPs, the carrier group consisted of all carriers (homozygote or heterozygote) of the mutant allele, whereas for recessively inherited SNPs, susceptibles were the homozygote carriers of the mutant allele. The marginal maximum likelihood method was implemented using an adaptation of functions created by Nilanjan Chatterjee (National Cancer Institute, USA) in MatLab v. 7.0.1 (MathWorks, Natick, MA, USA), which maximize the marginal likelihood using an expectation–maximization algorithm. Cumulative survival estimates are generated for both carriers and non-carriers. Bootstrap estimates for the hazard ratios were then computed, using 1000 resamples of the data. These estimates were used to generate 95% CIs for the hazard ratios by taking the 25th and 975th order statistics.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We gratefully acknowledge the participation of all patients and control individuals. The authors are indebted to Richard Coleman, Christina Fleischmann, Nicholas Hearle, Rosalind Mutch, Elaine Ryder-Mills, Hayley Spendlove and Remben Talaban for sample collection, identification and preparation. Finally, we would like to thank Nilanjan Chatterjee for assistance with implementing his functions for marginal likelihood computations. *Grant support*: funding for this work was undertaken with support from Cancer Research UK, CORE, the National Cancer Research Network, the European Union (CCPRB), Leukaemia Research, Arbib Foundation, Breakthrough Breast Cancer and HEAL. The funding sources had no role in the study design, collection and analysis of data, nor in the preparation of the manuscript and the decision to submit the paper for publication.

Conflict of Interest statement. None declared.

REFERENCES

1. Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. and Hemminki, K. (2000) Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.*, **343**, 78–85.
2. Bonaiti-Pellie, C. (1999) Genetic risk factors in colorectal cancer. *Eur. J. Cancer Prev.*, **8**(Suppl. 1), S27–S32.
3. Fleischmann, C., Peto, J., Cheadle, J., Shah, B., Sampson, J. and Houlston, R.S. (2004) Comprehensive analysis of the contribution of germline MYH variation to early-onset colorectal cancer. *Int. J. Cancer*, **109**, 554–558.
4. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
5. Houlston, R.S. and Tomlinson, I.P. (2001) Polymorphisms and colorectal tumor risk. *Gastroenterology*, **121**, 282–301.
6. Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33**(suppl.), 228–237.
7. Xi, T., Jones, I.M. and Mohrenweiser, H.W. (2004) Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics*, **83**, 970–979.
8. Lee, J.H., Lee, E.S., Kim, Y.S., Won, N.H. and Chae, Y.S. (2006) BRAF mutation and AKAP9 expression in sporadic papillary thyroid carcinomas. *Pathology*, **38**, 201–204.
9. Stankovic, T., Weber, P., Stewart, G., Bedenham, T., Murray, J., Byrd, P.J., Moss, P.A. and Taylor, A.M. (1999) Inactivation of ataxia telangiectasia mutated gene in B-cell chronic lymphocytic leukaemia. *Lancet*, **353**, 26–29.
10. Gutierrez-Enriquez, S., Fernet, M., Dork, T., Bremer, M., Lauge, A., Stoppa-Lyonnet, D., Moullan, N., Angele, S. and Hall, J. (2004) Functional consequences of ATM sequence variants for chromosomal radiosensitivity. *Genes Chromosomes Cancer*, **40**, 109–119.
11. Larson, G.P., Zhang, G., Ding, S., Foldenauer, K., Udar, N., Gatti, R.A., Neuberg, D., Lunetta, K.L., Ruckdeschel, J.C., Longmate, J. *et al*. (1997) An allelic variant at the ATM locus is implicated in breast cancer susceptibility. *Genet. Test.*, **1**, 165–170.
12. Angele, S., Falconer, A., Edwards, S.M., Dork, T., Bremer, M., Moullan, N., Chapot, B., Muir, K., Houlston, R., Norman, A.R. *et al*. (2004) ATM polymorphisms as risk factors for prostate cancer development. *Br. J. Cancer*, **91**, 783–787.
13. Falck, J., Lukas, C., Protopopova, M., Lukas, J., Selivanova, G. and Bartek, J. (2001) Functional impact of concomitant versus alternative defects in the Chk2-p53 tumour suppressor pathway. *Oncogene*, **20**, 5503–5510.
14. Li, J., Williams, B.L., Haire, L.F., Goldberg, M., Wilker, E., Durocher, D., Yaffe, M.B., Jackson, S.P. and Smerdon, S.J. (2002) Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. *Mol. Cell*, **9**, 1045–1054.
15. Thompson, D., Duedal, S., Kirner, J., McGuffog, L., Last, J., Reiman, A., Byrd, P., Taylor, M. and Easton, D.F. (2005) Cancer risks and mortality in heterozygous ATM mutation carriers. *J. Natl Cancer Inst.*, **97**, 813–822.
16. Cybulski, C., Gorski, B., Huzarski, T., Masojc, B., Mierzejewski, M., Debniak, T., Teodorczyk, U., Byrski, T., Gronwald, J., Matyjasik, J. *et al*. (2004) CHEK2 is a multiorgan cancer susceptibility gene. *Am. J. Hum. Genet.*, **75**, 1131–1135.
17. Meijers-Heijboer, H., Wijnen, J., Vasen, H., Wasielewski, M., Wagner, A., Hollestelle, A., Elstrodt, F., van den Bos, R., de Snoo, A., Fat, G.T. *et al*. (2003) The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype. *Am. J. Hum. Genet.*, **72**, 1308–1314.
18. Kim, Y.I. (1999) Folate and carcinogenesis: evidence, mechanisms, and implications. *J. Nutr. Biochem.*, **10**, 66–88.
19. Blount, B.C., Mack, M.M., Wehr, C.M., MacGregor, J.T., Hiatt, R.A., Wang, G., Wickramasinghe, S.N., Everson, R.B. and Ames, B.N. (1997) Folate deficiency causes uracil misincorporation into human DNA and chromosome breakage: implications for cancer and neuronal damage. *Proc. Natl Acad. Sci. USA*, **94**, 3290–3295.
20. Jacques, P.F., Bostom, A.G., Williams, R.R., Ellison, R.C., Eckfeldt, J.H., Rosenberg, I.H., Selhub, J. and Rozen, R. (1996) Relation between folate status, a common mutation in methylenetetrahydrofolate reductase, and plasma homocysteine concentrations. *Circulation*, **93**, 7–9.
21. Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. *et al*. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
22. Kerber, R.A. and Slattery, M.L. (1997) Comparison of self-reported and database-linked family history of cancer data in a case–control study. *Am. J. Epidemiol.*, **146**, 244–248.
23. Diseases, I.S.C.o. (1977) *1975 Revision*, WHO, Geneva.
24. Rudd, M.F., Williams, R.D., Webb, E.L., Schmidt, S., Sellick, G.S. and Houlston, R.S. (2005) The predicted impact of coding single nucleotide polymorphisms database. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 2598–2604.

25. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
26. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
27. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
28. Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
29. Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
30. Ford, B.N., Ruttan, C.C., Kyle, V.L., Brackley, M.E. and Glickman, B.W. (2000) Identification of single nucleotide polymorphisms in human DNA repair genes. *Carcinogenesis*, **21**, 1977–1981.
31. Mohrenweiser, H.W., Xi, T., Vazquez-Matias, J. and Jones, I.M. (2002) Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 1054–1064.
32. Kuschel, B., Auranen, A., McBride, S., Novik, K.L., Antoniou, A., Lipscombe, J.M., Day, N.E., Easton, D.F., Ponder, B.A., Pharoah, P.D. *et al*. (2002) Variants in DNA double-strand break repair genes and breast cancer susceptibility. *Hum. Mol. Genet.*, **11**, 1399–1407.
33. Fearnhead, N.S., Wilding, J.L., Winney, B., Tonks, S., Bartlett, S., Bicknell, D.C., Tomlinson, I.P., Mortensen, N.J. and Bodmer, W.F. (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl Acad. Sci. USA*, **101**, 15992–15997.
34. Savas, S., Kim, D.Y., Ahmad, M.F., Shariff, M. and Ozcelik, H. (2004) Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. *Cancer Epidemiol. Biomarkers Prev.*, **13**, 801–807.
35. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
36. Marchini, J., Donnelly, P. and Cardon, L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.
37. Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.
38. Stephens, M. and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.
39. Stephens, M., Smith, N.J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.
40. Chatterjee, N. and Wacholder, S. (2001) A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics*, **57**, 245–252.