Published in final edited form as:
Gut. 2013 June ; 62(6): 871–881. doi:10.1136/gutjnl-2011-300537.

Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals

Malcolm G. Dunlop[§], **Albert Tenesa**, **Susan M. Farrington**, **Stephane Ballereau**, **David H. Brewster**, **Paul DP. Pharoah**, **Clemens Schafmayer**, **Jochen Hampe**, **Henry Völzke**, **Jenny Chang-Claude**, **Michael Hoffmeister**, **Hermann Brenner**, **Susanna von Holst**, **Simone Picelli**, **Annika Lindblom**, **Mark A. Jenkins**, **John L. Hopper**, **Graham Casey**, **David Duggan**, **Polly Newcomb**, **Anna Abulí**, **Xavier Bessa**, **Clara Ruiz-Ponte**, **Sergi Castellví-Bel**, **Iina Niittymäki**, **Sari Tuupanen**, **Auli Karhu**, **Lauri Aaltonen**, **Brent W. Zanke**, **Thomas J. Hudson**, **Steven Gallinger**, **Ella Barclay**, **Lynn Martin**, **Maggie Gorman**, **Luis Carvajal-Carmona**, **Axel Walther**, **David Kerr**, **Steven Lubbe**, **Peter Broderick**, **Ian Chandler**, **Alan Pittman**, **Steven Penegar**, **Harry Campbell**, **Ian Tomlinson**, and **Richard S. Houlston**

[1]Colon Cancer Genetics Group, MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK [2]The Roslin Institute, University of Edinburgh, Roslin, UK [3]Scottish Cancer Registry, Information Services Division, NHS National Services Scotland, Edinburgh, UK [4]Department of Oncology, Strangeways Research Laboratory, University of Cambridge, Cambridge, UK [5]POPGEN Biobank, University Hospital Schleswig-Holstein, Kiel, Germany [6]Department of General Internal Medicine, University Hospital, Schleswig-Holstein, Kiel, Germany [7]Institut fuer Community Medicine, University Hospital Greifswald, Greifswald, Germany [8]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany [9]Division of Clinical Epidemiology and Ageing Research, German Cancer Research Center (DKFZ), Heidelberg, Germany [10]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden [11]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, Parkville, Victoria, Australia [12]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA [13]Translational Genomics Research Institute (TGen), Phoenix, Arizona, USA [14]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA [15]Department of Gastroenterology, Hospital del Mar, Institut Municipal d'Investigació Mèdica (IMIM), Pompeu Fabra University, Barcelona, Catalonia, Spain [16]Fundación Pública Galega de Medicina Xenómica (FPGMX), CIBERER, Genomic Medicine Group, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain [17]Department of Gastroenterology, Hospital Clínic, CIBERehd, IDIBAPS, University of Barcelona, Catalonia, Spain [18]Department of Medical Genetics, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland [19]Cancer Care Ontario, Toronto, Ontario, Canada [20]Ontario Institute for Cancer Research, Toronto, Ontario, Canada [21]Samuel Lunenfeld Research Institute, Mount Sinai Hospital and University of Toronto, Toronto

[§]Corresponding Author/Guarantor: Malcolm Dunlop, Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Edinburgh EH4 2XU. Tel: +44-(0)131 467-8454, Fax: +44-(0)131 467-8450, Malcolm.Dunlop@hgu.mrc.ac.uk.

Conflict of interest

The authors report no conflicts of interest with respect to the work presented in this paper.

Ontario, Canada [22]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK [23]Department of Clinical Pharmacology, University of Oxford, Oxford, UK [24]Section of Cancer Genetics, Institute of Cancer Research, Sutton, UK [25]Public Health Sciences, University of Edinburgh, Edinburgh, UK

Abstract

Objective—Colorectal cancer (CRC) has a substantial heritable component. Common genetic variation has been shown to contribute to CRC risk. In a large, multi-population study, we set out to assess the feasibility of CRC risk prediction using common genetic variant data, combined with other risk factors. We built a risk prediction model and applied it to the Scottish population using available data.

Design—Nine populations of European descent were studied to develop and validate colorectal cancer risk prediction models. Binary logistic regression was used to assess the combined effect of age, gender, family history (FH) and genotypes at 10 susceptibility loci that individually only modestly influence colorectal cancer risk. Risk models were generated from case-control data incorporating genotypes alone (n=39,266), and in combination with gender, age and family history (n=11,324). Model discriminatory performance was assessed using 10-fold internal cross-validation and externally using 4,187 independent samples. 10-year absolute risk was estimated by modelling genotype and FH with age- and gender-specific population risks.

Results—Median number of risk alleles was greater in cases than controls (10 vs 9, $p<2.2\times10^{-16}$), confirmed in external validation sets (Sweden $p=1.2\times10^{-6}$, Finland $p=2\times10^{-5}$). Mean per-allele increase in risk was 9% (OR 1.09; 95% CI 1.05–1.13). Discriminative performance was poor across the risk spectrum (area under curve (AUC) for genotypes alone - 0.57; AUC for genotype/age/gender/FH - 0.59). However, modelling genotype data, FH, age and gender with Scottish population data shows the practicalities of identifying a subgroup with >5% predicted 10-year absolute risk.

Conclusion—We show that genotype data provides additional information that complements age, gender and FH as risk factors. However, individualized genetic risk prediction is not currently feasible. Nonetheless, the modelling exercise suggests public health potential, since it is possible to stratify the population into CRC risk categories, thereby informing targeted prevention and surveillance.

Introduction

Colorectal cancer (CRC) is common in Western countries, with the global annual incidence exceeding 1 million and accounting for ~9% of all cancers[1]. The variation in worldwide incidence is narrowing rapidly due to increasing exposure to “westernised” lifestyle risk factors in populations that had historically low rates. Population-based registry data indicate that CRC remains a common cause of cancer death (overall survival ~50%[2]). Screening of average risk populations using faecal occult blood testing (FOBT) has been introduced in many countries, following demonstration of mortality reductions in several large trials[3]. Furthermore, invasive screening using flexible sigmoidoscopy for a specific risk category defined only age has also been trialled with promising results[4]. Incidence reduction may also be feasible, both in the general population[56] and in genetically defined high risk groups[78]. Thus, stratifying the average risk population into risk categories offers the potential of tailoring the intensity of surveillance, or preventative approach, to the predicted level of risk. Hence, those at highest risk could be offered more frequent, or more sensitive, FOBT screening. Endoscopic surveillance by colonoscopy or flexible sigmoidoscopy might also be instigated in those at highest risk.

The heritable component of CRC variance is around 35%[9] but only ~5% of cases are attributable to highly penetrant mutations. Recent genome-wide association studies have identified a number of common genetic risk loci for CRC[10–17]. Risk associated with each locus is individually modest, but risk alleles are carried by a large proportion of the population because of the high allele frequencies[10–17]. Thus, high absolute risks, exceeding thresholds triggering clinical intervention, could be apparent in population subgroups carrying multiple risk alleles. Colonoscopic surveillance is already offered to people with a modestly elevated risk due to a personal or family history of CRC[6]. More intensive surveillance is offered to high risk individuals from Lynch Syndrome families[18–20]. Similarly, genotype data from common variants offers the possibility of partitioning risk within the *average risk* population, according to the population frequency and risk of multi-locus genotypes. More intensive surveillance could be offered to those at highest risk, whilst the remainder could remain on average risk screening protocols, as proposed for breast cancer[21].

We set out to develop and validate CRC risk prediction models and to assess model performance in profiling individual genetic risk of CRC. We developed models incorporating age, gender, family history and genotype data from 10 common genetic risk variants in over 40,000 individuals from multiple populations, mainly of north European descent. To gauge the broader future potential of genetic risk modelling, we assessed the utility in categorising risk subgroups within the population by applying the risk models to available Scottish population data on CRC risk by age and gender.

Methods

Study subjects

To generate the risk models, we studied a total of 44,389 subjects (24,395 CRC cases, 19,994 cancer-free controls) from seven geographically distinct populations, predominantly of European origin (Table 1). Age, gender, demographic and clinical data were collected, along with blood samples. Samples were genotyped for the 10 risk SNPs, as were the external validation case-control sets (1,563 Swedish cases and 1,504 controls; 702 Finnish cases and 418 controls). In risk model analyses, we only incorporated samples for which we generated genotypes for *all* 10 risk SNPs (ie no missing values). Table 1 shows sample numbers from each population along with the nature and origin of case and control subjects. Family history information was available for a subset (Table 1). A minority of studies used family history to select cases and/or controls.

Genotyping

DNA purification and quality control procedures are described elsewhere[1517]. Genotyping was performed using various platforms in use at each of the contributing sites. The 10 common CRC risk SNPs previously identified through genome-wide association (GWAS) studies and shown to tag independent loci were: rs6983267[10–1217], rs4779584[14], rs4939827[1315], rs3802842[1516], rs10795668[1516], rs16892766[1516], rs4444235[17], rs9929218[17], rs10411210[17], rs961253[17].

Statistical analysis

Allele frequencies for each of the 10 CRC SNPs were calculated in cases and controls for each population. The effects of SNP genotype, gender and family history were assessed using binary logistic regression. The total number of risk alleles for each population, and for all samples from the model generation set together, was then assessed and a two-sided t-test applied to compare number of risk alleles between cases and controls. In the logistic regression assessing the effect of family history, we only incorporated data from population-based studies where there was no (Scotland, DACHS), or limited (OFCR) prior selection on family history criteria (Table 2)

Risk modelling

Generation and internal validation of the risk models was based on the 39,266 subjects without any missing values, including genotype data for all 10 SNPs. The model was considered to be additive on the log risk scale. The probability that a person carrying a given number of common risk alleles develops CRC by age x was estimated using a Bayesian approach. Probability of disease by age x is expressed as P(D_x). We defined carriers as subjects with a given number of risk alleles (Z), where Z varied between 11 and 13. Thus, if Z≥13, then subjects with ≥13 alleles were classified as carriers (G=1) and subjects with <13 alleles were non-carriers (G=0). To estimate the probability that carriers of ≥Z alleles (G=1) develop CRC by age x, then

$$P(Dx|G=1)=P(G=1|Dx)*P(Dx)/P(G=1).$$

The probability that a non-carrier (G=0) develops CRC by age x is:

$$P(Dx|G=0)=P(G=0|Dx)*P(Dx)/P(G=0)$$

We assume that P(G=1|Dx) and P(G=0|Dx) are different from each other but are constant over all ages (x) and call these P(G=1|D) and P(G=0|D), respectively. This seems reasonable since each allele exerts a constant effect on risk over the observed lifetime[1517]. P(G=1|D) and P(G=0|D) can be estimated as the proportion of study cases with ≥Z or <Z risk alleles, respectively. P(G=1) and P(G=0) were estimated from control data to gauge “carrier” frequency of a given number of risk alleles in the general population. All controls were cancer-free at the time of sampling. Some control sets were enriched by selection for

absence of a CRC family history and so P(G=1) may be marginally under-estimated in the general population.

Multivariate analysis using binary logistic regression was conducted to test the effect of each covariate. Tested models included variously: genotype data for the 10 SNPs; family history status; age (continuous variable); gender. Genotype for each SNP was assumed to be additive on the log risk scale and genotypes were scored as −1, 0 or 1 in the logistic regression analysis.

Assessment of risk model performance

Risk model performance was assessed by both internal and external validation using the statistical package ROCR[22]. Internal validation comprised 10-fold cross-validation to estimate receiver operator characteristic (ROC) curves by randomly assigning study subjects and all associated data for that individual into 10 complementary datasets. One dataset at a time was used as the validation set and the remaining 9 datasets as the training set. Separate Receiver Operator Curves (ROC) were generated for models incorporating: (i) age, gender, family history and genotypes at all 10 loci for the population-based non FH-selected study populations (Table 2); (ii) 10 locus genotypes for all datasets. External validation using the two independent case-control sets (Sweden, Finland) was conducted by separately fitting the model in the analysis using all 10 SNP genotypes for 1,563 Swedish cases and 1,504 controls and 702 Finnish cases and 418 controls. Again, model performance was evaluated using ROC analysis. Probability of a subject being a case or control was determined by estimating the proportion of true and false positives at different cut-off points.

Estimating the potential public health impact by applying risk prediction models to available Scottish population data

We used Scottish population and Cancer Registry data as reference for estimating the probability of developing CRC. We consider the use of Scottish population data to be valid because: there is comprehensive population coverage and high levels of data completeness[23]; CRC incidence is broadly representative of northern European and North American populations[2]; available systematic family history data from our current and past studies. Age-specific CRC rate was calculated from 2006 cancer registration data and from age-specific estimates of the Scottish population[24]. Cumulative CRC rate for any given age was calculated separately for males and females as the sum of the age-specific rates up to that age. The cumulative probability of developing CRC in the general population by various ages, FH status and risk allele "load", is shown in Table 3 along with absolute risks in the general population.

Results

Assessment of risk prediction models

Risk allele frequencies are shown in Figure 1. Odds ratios are grouped for subjects carrying ≤4 risk alleles, and ≥14 alleles, because of very small numbers of subjects at these extremes. Figure 2 shows risk allele frequency comparisons by population as a box plot. The frequency of carriage of ≥12, ≥13 and ≥14 alleles (equating to P(G=1|D)=1−P(G=0|D)) in the combined case sets was respectively 0.205 0.091 and 0.032. Corresponding control subject allele frequencies (representative of general population) were 0.141, 0.055 and 0.017 for ≥12, ≥13 and ≥14 alleles respectively (equating to P(G=1)).

Although there was only a small difference in mean number of risk alleles in cases compared to controls (mean in cases 9.93 vs 9.39 in controls; difference - 0.53 alleles), this was highly significant (95% CI 0.57–0.49. 2-sided t-test. $p<2.2\times10^{-16}$), because of the very large sample size. Median number of alleles in cases was also highly significantly different to that in controls (10 for cases, 9 for controls, $p<2.2\times10^{-16}$ Mann-Whitney test) (Figure 2). Consistent with each locus having an independent effect on CRC risk, there was no statistically significant interactive effects between any of the 10 loci (p>0.05 for interaction, testing each locus against all others). Table 4 shows the effects of age, gender, family history and genotype for SNPs tagging each risk locus, with relative weight contributed by each variable in the logistic regression. Table 4 shows that genotype provided additional information on CRC risk that is *complementary* to family history.

The discriminative ability was assessed by ROC incorporating SNP genotypes at all 10 loci alone, or in combination with gender, age and family history data (Fig 3). The average area under the curve (AUC) for 10 iterations in the cross-validation analysis was 0.57 for the model incorporating SNP genotypes alone (39,266 subjects), and 0.59 when incorporating genotype, age, gender and family history status (11,324 subjects). Values for each of the 10 iterations of cross-validation are shown in Table 5. The relationship, and variability, between estimated risk and increasing number of risk alleles is shown in Figure 3. The association between risk and total number of alleles (the SCORE) was also highly significant in the external validation sets ($P=1.2\times10^{-6}$ for Swedish, $P=2.6\times10^{-5}$ Finnish populations). On average, each allele increased risk of CRC by 9% (OR 1.09, 95% CI 1.05–1.13) for Swedish and 3% (OR 1.03, 95% CI 1.02–1.04) for Finnish samples. Fitting Swedish and Finnish genotype data (3,067 and 1,120 subjects) generated AUC of 0.56 and 0.57 respectively (Figure 4). Thus ROC analyses shows that risk models have limited *individual* predictive performance across the observed risk spectrum and allele distributions. This is consistent with our previous estimate of an overall accuracy of prediction of the genetic component of risk of 26%, given that we previously estimated that all 10 of these SNPs explained ~6% of the excess familial risk and ~1.26% of the overall variation of liability to colorectal cancer[25]. It should be noted that these estimates do not take into account the environmental component of risk, nor age/gender effects on risk.

Estimating the proportion of the population in high risk categories

Having shown that risk prediction at the individual level is not feasible, using the Bayesian modelling approach described above, we set out to gauge the potential public health impact of applying such risk models to the general population. We estimated the proportion of people in the general population that might be included within a high risk category, sufficient to merit more intensive large bowel surveillance or intervention. Taking account of allele frequency and effect size of various risk allele combinations, we incorporated SNP genotype, family history, age, gender and Scottish population data on CRC incidence by age and gender. We estimated 10-year absolute CRC risk by age for males and females (separately) carrying >12, >13 risk or >14 alleles (Figure 5). The risk associated with a positive FH that we observed in the Scottish dataset (OR=1.75, 95% CI 1.48–2.06) was similar to, though marginally lower than, that estimated in a recent meta-analysis[26]. The frequency of control subjects reporting at least one affected first-degree relative (0.09) in the current population is very similar to that observed in a previous Scottish population-based series aged 30–70 yrs (0.094. 95% CI 5.8–14.9)[27]. We considered <5%, ≥5% and ≥10% predicted absolute risk subgroups. The 10-year absolute risk in the Scottish population approaches 5% only for males after the age of 75yrs (Figure 5, Table 3). As expected, FH+ is associated with increased risk, reaching the 5% threshold around age 70 yrs in females and 60 yrs in males. The estimated absolute risk for >12 risk alleles is very similar to that imparted by a positive family history (Figure 5a and 5b). It should be noted that genotype for the 10 common variants provides information for risk prediction *additional* to that from family history alone (Tables 3 & 4, Figure 5). In FH+ individuals with >12 alleles, the age at which the 10-year risk surpasses the 5% and 10% thresholds is substantially lower than for FH+ alone (5% - males: 52 vs 60 yrs, females 58 vs 68 yrs; 10% - males: 62yrs vs 75yrs, females: 75yrs vs >80yrs (not assessable).

Available evidence suggests benefit for advancing the age of initial FOBT screening for people with a family history[28]. Thus, offering genotyping for common variants in the general population subgroup with an affected relative (9%) could refine empiric family history guidance. Given the impact of age and gender on prior risk, it is important to take these risk factors into account when considering age at which to offer genetic testing. To gauge the practical and financial issues around population genetic testing, we extrapolate from Figure 5. If genotyping for risk SNPs was targeted to males in the Scottish population aged ≥55yrs with a positive FH then only ~59,000 tests would be necessary. This would identify the estimated 6.7% of men (~4,000) with >12 risk alleles who have >5% 10-year absolute risk of CRC, and all of whom have a 10-yr CRC risk greater than 10% from age 60yrs. Similarly, restricting genotyping to females aged ≥60yrs with a positive family history would involve ~57,000 tests and this would identify an estimated 3,800 women with >5% 10-year absolute CRC risk (Figure 5). In all, this approach would identify 7 people/ hundred of the tested population with ≥5% 10-year absolute CRC risk. It should be emphasised that this is a modelling exercise applied to population data, however robust the population data and genotyped sample sets. However, testing these models in practice will likely remain logistically and scientifically challenging for the foreseeable future. Nevertheless, this provides useful estimates to provide insight into the number of people who might be offered genetic testing for SNP markers and who might be identified to be at sufficiently high risk to merit intensive screening.

Discussion

In this study, we assessed the utility of CRC genetic risk profiling using a panel of 10 common genetic variants shown incontrovertibly to be associated with CRC susceptibility and combined this information with aqe, gender and family history information (as a proxy for genetic susceptibility factors yet to be discovered). We show clearly that genotype at common risk variants provides information over and above that of family history alone (Table 4, Figure 5). There is a small, but highly statistically significant difference

($p<2.2\times10^{-16}$), in risk allele distribution between cases and controls. The level of statistical significance was due to the very large study size, rather than the magnitude of the difference.

ROC analysis of models including genotype data alone, or in combination age, gender and family history showed very modest discriminative performance across the risk spectrum (AUC ~0.59 and 0.57 (internal validation) or 0.56 and 0.57 (external validation sets). Overall positive predictive value was between 0.51 and 0.71 for cutoff points of 0.4 and 0.7, respectively, with negative predictive values for the same cut-offs of 0.62 and 0.51, respectively. This modest level of test performance was consistent across study populations, suggesting that risk assessment algorithms based on common genetic variants are likely to have similar performance characteristics in Caucasian populations and are unlikely to be confounded by Linkage Disequilibrium (LD) structure differences.

The poor performance in *individualized* CRC risk profiling is consistent with risk prediction studies in other diseases[29–31]. Typical AUCs have range from 0.55 to 0.60 in type 2 diabetes[32–35], with slightly higher values for age-related macular degeneration (AMD), Crohn’s Disease, coronary heart disease and cardiovascular diseases[36–38]. The best predictive performances have been obtained by combining genetic, demographic, and environmental variables[39]. The great majority of true susceptibility loci are not included in these analyses because they have yet to be discovered. Improved predictive performance (AUC > 0.8) likely could be achieved by including SNPs from a much larger number of susceptibility loci[38]. Consistent with this, we previously estimated that a model with ~100 of the estimated 172 SNPs accounting for the genetic variance for CRC could provide 80% accuracy of prediction of the genetic component of risk, and explain ~17% of the phenotypic variance in the liability scale[25].

AUC generated by ROC analysis represents the probability that cases have a higher score than controls. Whilst this is important for a diagnostic test, it only gives a limited assessment of the potential value of a predictive test where the main aim is categorisation into clinically meaningful risk strata[31]. AUC does not address absolute levels of risk or whether the model stratifies correctly into high/low categories of absolute risk that are of clinical importance (such as 10-year risk of CRC). We maintain that in the context of this study, prediction of actual risk is a more important model function than sensitivity/specificity, on which ROC curve and AUC estimates are based[40].

Additional common CRC genetic risk variants identified through ongoing research efforts are likely to have effect sizes even smaller and/or allele frequencies lower than those identified to date[10–1217]. Nonetheless, predictive utility of testing for common genetic variants is likely to improve with new discoveries and individualised CRC genetic risk profiling may become feasible. The combined performance of genetic variants and other established (non-genetic) risk may vary depending on the nature of the genetic variants incorporated into the model[31]. These may have a greater impact on risk prediction if they involve novel disease pathways independent of the causal mechanisms through which the other risk factors operate[30], as is likely in this study since a number of the variants involve the TGF beta signaling pathway[25].

Application of the predictive model developed here using observed genotype and other risk factor data allows estimation of the likely effect in a population setting. It also provides some insight into the feasibility and likely outcomes of applying such a model in practice. The modeling suggests that it may be possible to identify population subgroups with substantially elevated 10-year absolute risk of CRC. The approach could identify the approximately 7% of the tested population with sufficiently high risk as to warrant additional screening, such as regular colonoscopic surveillance and/or age advancement of recruitment to population screening programs[28]. So et al[41] also recently developed a statistical framework incorporating genotype, family history and other risk factor data for prediction of breast and prostate cancer. Their findings support the notion that such modelling can stratify the population into risk categories, opening up the potential for targeted prevention and screening. Models incorporating genotype data from common variants will not identity rare high-penetrance alleles (eg those responsible for Lynch Syndrome), validated risk prediction models have been developed to identify such individuals[42].

Genotyping the population with a family history of CRC is an attractive approach. A 5% threshold of absolute 10-year CRC risk has clinical and public health validity since it exceeds the highest risk at any age in the general population and is tenfold greater than the risk for a 50-year old (Table 3) entering population-based FOBT screening programs. It should be noted that these findings are focused on identifying population subgroups with excess risk that merit *additional* screening. We have not addressed the issue of a reduction in screening for those predicted to be at lower risk. In this study, we have explored risk model performance across a range of European populations in order to reduce potential bias due to limited representativeness. Although we validated these findings in an external validation set, model performance should be tested in a large, long-term cohort study in which the genetic variants can be studied together with classical risk factors to give reassurance that model performance is not inflated due to selection, information or survival biases.

These findings have implications for current FOBT screening programs. Brenner argues that risk associated with a family history logically dictates that FH+ individuals should enter screening programs ~10 years earlier than those without[28]. Whilst there are a number of issues that need to be addressed to translate any genetic test into clinical and public health practice, the results of the modeling presented here suggest that it is possible to identify population subgroups with substantially increased CRC risk. Indeed, the risk is sufficiently high as to merit changes to screening policy for the groups in that risk category. Furthermore, amendments to criteria for age of entry to family history focused surveillance programs[643] merit evaluation. This study provides the first tangible indication that data from genome-wide studies of CRC have public health importance.

Acknowledgments

Acknowledgements and funding

Edinburgh: Work was supported by grants from Cancer Research UK (C348/A12076 and the Bobby Moore Fund), Scottish Government Chief Scientist Office (K/OPR/2/2/D333, CZB/4/94); Medical Research Council (G0000657-53203); Centre Grant from CORE as part of the Digestive Cancer Campaign (http://www.corecharity.org.uk). Cambridge: SEARCH is funded by a program grant from Cancer Research UK. TK was funded by the Foundation Dr Henri Dubois-Ferriere Dinu Lipatti. Kiel/Greifswald: Supported by the German National Genome Research Network (NGFN) through the PopGen biobank (BmBF 01GR0468) and the National Genotyping Platform. Further support through the MediGrid and Services@MediGrid projects (01AK803G, 01IG07015B). SHIP is part of the Community Medicine Research net (CMR) of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (ZZ9603), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania. Heidelberg: Support from German Research Council (Deutsche Forschungsgemeinschaft) (BR 1704/6-1, BR 1704/6-3, CH 117/1-1), and the German Federal Ministry for Education and Research (01 KH 0404). Sweden: Financial support was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and the Karolinska Institute, The Swedish Cancer Society, The Stockholm Cancer Foundation and The Swedish Research Council. Colon-CFR: Grant support NIH/NCI U01CA122839. Spain: Supported by grants from Fondo de Investigación Sanitaria/FEDER (06/1384, 06/1712, 08/0024, 08/1276), Xunta de Galicia (PGIDIT07PXIB9101209PR), Fundación de Investigación Médica Mutua Madrileña (CRP and SCB), Ministerio de Educación y Ciencia (SAF 07-64873), Asociación Española contra el Cáncer (Fundación Científica and Junta de Barcelona), Fundación Olga Torres (SCB), and Acción Transversal de Cáncer (Instituto de Salud Carlos III). CIBERER and CIBEREHD are funded by the Instituto de Salud Carlos III. SCB is supported by a contract from the Fondo de Investigación Sanitaria (CP 03-0070). We acknowledge the Santiago de Compostela and Barcelona branches of the Spanish National Genotyping Centre (CeGen) for genotyping. Finland: Grant support from Academy of Finland (Finnish Centre of Excellence Program 2006–2011), the Finnish Cancer Society, the Sigrid Juselius Foundation and the European Commission (9LSHG-CT-2004-512142). Canada: Cancer Care Ontario (host organization to the ARCTIC Genome Project) acknowledges project funding by Genome Canada through the Ontario Genomics Institute, by Génome Québec, the Ministere du Dévelopement Économique et Régional et de la Recherche du Québec and the Ontario Institute for Cancer Research. Additional funding from National Cancer Institute of Canada (NCIC) through the Cancer Risk Assessment (CaRE) Program Project Grant. The work was supported through collaboration and cooperative agreements with the Colon Cancer Family Registry and PIs, supported by the National Cancer Institute, National Institutes of Health under RFA CA-95-011, including the Ontario Registry for Studies of Familial CRC (U01 CA076783). TJH and BWZ hold Senior Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation. Oxford: The work was supported by Cancer Research UK and the Bobby Moore Fund (C1298/A8362). Additional funding provided by the European Union (CPRB LSHC-CT-2004-503465), and CORE (http://www.corecharity.org.uk). E Domingo and I Tomlinson are supported by the Oxford Biomedical Research Centre. The work was also supported by the Oxford Biomedical Research Centre and funded by the Medical Research Council. ICR, London: Grant support from Bobby Moore Cancer Cancer Research UK (C1298/A8362), CORE, and the European Commission (QLG2-CT-2001-01861). Stephen Lubbe was in receipt of a PhD studentship from Cancer Research UK and Ian Chandler a Clinical Training Fellowship from St. George's Hospital.

References

1. Parkin DM, Bray F, Ferlay J, et al. Global cancer statistics, 2002. CA Cancer J Clin. 2005; 55(2):74–108. [PubMed: 15761078]
2. Ferlay J, Parkin DM, Steliarova-Foucher E. Estimates of cancer incidence and mortality in Europe in 2008. Eur J Cancer. 46(4):765–81.
3. Towler B, Irwig L, Glasziou P, et al. A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, hemoccult. Bmj. 1998; 317(7158):559–65. [PubMed: 9721111]
4. Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. Lancet. 375(9726):1624–33.
5. Winawer SJ, Zauber AG, O'Brien MJ, et al. Randomized comparison of surveillance intervals after colonoscopic removal of newly diagnosed adenomatous polyps. The National Polyp Study Workgroup. N Engl J Med. 1993; 328(13):901–6. [PubMed: 8446136]
6. Levin B, Lieberman DA, McFarland B, et al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. Gastroenterology. 2008; 134(5):1570–95. [PubMed: 18384785]
7. Jarvinen HJ, Renkonen-Sinisalo L, Aktan-Collan K, et al. Ten years after mutation testing for Lynch syndrome: cancer incidence and outcome in mutation-positive and mutation-negative family members. J Clin Oncol. 2009; 27(28):4793–7. [PubMed: 19720893]
8. de Jong AE, Hendriks YM, Kleibeuker JH, et al. Decrease in mortality in Lynch syndrome families because of surveillance. Gastroenterology. 2006; 130(3):665–71. [PubMed: 16530507]

9. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000; 343(2):78–85. [PubMed: 10891514]

10. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet. 2007; 39(8):989–94. [PubMed: 17618283]

11. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24. 21. Nat Genet. 2007; 39(8):984–8. [PubMed: 17618284]

12. Haiman CA, Le Marchand L, Yamamato J, et al. A common genetic risk factor for colorectal and prostate cancer. Nat Genet. 2007; 39(8):954–6. [PubMed: 17618282]

13. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet. 2007; 39(11):1315–7. [PubMed: 17934461]

14. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13. 3 influence colorectal cancer risk. Nat Genet. 2008; 40(1):26–8. [PubMed: 18084292]

15. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet. 2008; 40(5):631–7. [PubMed: 18372901]

16. Tomlinson IP, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23. 3. Nat Genet. 2008; 40(5): 623–30. [PubMed: 18372905]

17. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet. 2008; 40(12):1426–35. [PubMed: 19011631]

18. Dunlop MG, Farrington SM, Carothers AD, et al. Cancer risk associated with germline DNA mismatch repair gene mutations. Hum Mol Genet. 1997; 6(1):105–10. [PubMed: 9002677]

19. Quehenberger F, Vasen HF, van Houwelingen HC. Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. J Med Genet. 2005; 42(6):491–6. [PubMed: 15937084]

20. Baglietto L, Lindor NM, Dowty JG, et al. Risks of Lynch syndrome cancers for MSH6 mutation carriers. J Natl Cancer Inst. 102(3):193–201.

21. Pharoah PD, Antoniou AC, Easton DF, et al. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008; 358(26):2796–803. [PubMed: 18579814]

22. Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21(20):3940–1. [PubMed: 16096348]

23. Brewster DH, Crichton J, Harvey JC, et al. Completeness of case ascertainment in a Scottish regional cancer registry for the year 1992. Public Health. 1997; 111(5):339–43. [PubMed: 9308385]

24. Service. IaSDoSH. Scottish Cancer Statistics - Colorectal Cancer. 2006.

25. Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. Nat Rev Genet. 2009; 10(6):353–8. [PubMed: 19434079]

26. Baglietto L, Jenkins MA, Severi G, et al. Measures of familial aggregation depend on definition of family history: meta-analysis for colorectal cancer. J Clin Epidemiol. 2006; 59(2):114–24. [PubMed: 16426946]

27. Mitchell RJ, Campbell H, Farrington SM, et al. Prevalence of family history of colorectal cancer in the general population. Br J Surg. 2005; 92(9):1161–4. [PubMed: 15997443]

28. Brenner H, Hoffmeister M, Haug U. Family history and age at initiation of colorectal cancer screening. Am J Gastroenterol. 2008; 103(9):2326–31. [PubMed: 18702651]

29. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 2009; 18(18):3525–31. [PubMed: 19553258]

30. Janssens AC, van Duijn CM. Genome-based prediction of common diseases: methodological considerations for future research. Genome Med. 2009; 1(2):20. [PubMed: 19341491]

31. Wray NR, Yang J, Goddard ME, Visscher PM. The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. PLoS Genet. 6(2):e1000864.

32. Weedon MN, McCarthy MI, Hitman G, et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med. 2006; 3(10):e374. [PubMed: 17020404]

33. Vaxillaire M, Veslot J, Dina C, et al. Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. Diabetes. 2008; 57(1):244–54. [PubMed: 17977958]

34. Lango H, Palmer CN, Morris AD, et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. Diabetes. 2008; 57(11):3129–35. [PubMed: 18591388]

35. van Hoek M, Dehghan A, Witteman JC, et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes. 2008; 57(11):3122–8. [PubMed: 18694974]

36. Paynter NP, Chasman DI, Buring JE, et al. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21. 3. Ann Intern Med. 2009; 150(2):65–72. [PubMed: 19153409]

37. van der Net JB, Janssens AC, Sijbrands EJ, et al. Value of genetic profiling for the prediction of coronary heart disease. Am Heart J. 2009; 158(1):105–10. [PubMed: 19540399]

38. Wei Z, Wang K, Qu HQ, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 2009; 5(10):e1000678. [PubMed: 19816555]

39. Seddon JM, Reynolds R, Maller J, et al. Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. Invest Ophthalmol Vis Sci. 2009; 50(5):2044–53. [PubMed: 19117936]

40. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007; 115(7):928–35. [PubMed: 17309939]

41. So HC, Kwan JS, Cherny SS, et al. Risk prediction of complex diseases from family history and known susceptibility Loci, with applications for cancer screening. Am J Hum Genet. 88(5):548–65. [PubMed: 21529750]

42. Barnetson RA, Tenesa A, Farrington SM, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. N Engl J Med. 2006; 354(26):2751–63. [PubMed: 16807412]

43. Dunlop MG. Guidance on large bowel surveillance for people with two first degree relatives with colorectal cancer or one first degree relative diagnosed with colorectal cancer under 45 years. Gut. 2002; 51(Suppl 5):V17–20. [PubMed: 12221035]

44. Cotterchio M, McKeown-Eyssen G, Sutherland H, et al. Ontario familial colon cancer registry: methods and first-year response rates. Chronic Dis Can. 2000; 21(2):81–6. [PubMed: 11007659]

Summary box

What is already known about this subject?

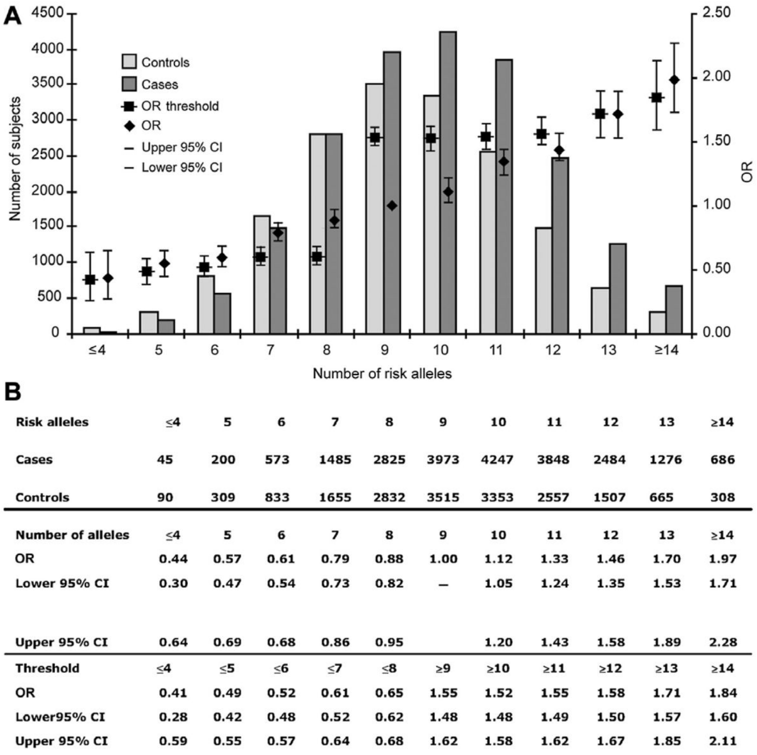
- Colorectal cancer has a heritable component to its aetiology.
- Common genetic variation contributes to risk.
- A number of common genetic variants associated with colorectal cancer risk have been identified.

What are the new findings?

- There is a highly significant difference in risk allele distribution between cases and controls.
- Genotype data from common genetic variants provides risk information over and above family history, gender and age.
- Genotype data, family history, gender and age can be incorporated into risk models.
- *Individualized* risk prediction is not yet feasible.
- A modelling exercise suggests that it is possible to identify a population subgroup with sufficiently high colorectal cancer risk to be relevant clinically and for public health strategy.

How might it impact on clinical practice in the foreseeable future?

- Applying the risk prediction model could help identify high risk groups for intensive surveillance as part of public health measures to control colorectal cancer.
- The estimated number of people who could be offered genetic testing for common genetic risk factors is both logistically and financially feasible.
- Identification of additional genetic risk factors is likely to further improve colorectal cancer risk prediction.



Risk alleles	≤4	5	6	7	8	9	10	11	12	13	≥14
Cases	45	200	573	1485	2825	3973	4247	3848	2484	1276	686
Controls	90	309	833	1655	2832	3515	3353	2557	1507	665	308

Number of alleles	≤4	5	6	7	8	9	10	11	12	13	≥14
OR	0.44	0.57	0.61	0.79	0.88	1.00	1.12	1.33	1.46	1.70	1.97
Lower 95% CI	0.30	0.47	0.54	0.73	0.82	–	1.05	1.24	1.35	1.53	1.71
Upper 95% CI	0.64	0.69	0.68	0.86	0.95		1.20	1.43	1.58	1.89	2.28

Threshold	≤4	≤5	≤6	≤7	≤8	≥9	≥10	≥11	≥12	≥13	≥14
OR	0.41	0.49	0.52	0.61	0.65	1.55	1.52	1.55	1.58	1.71	1.84
Lower95% CI	0.28	0.42	0.48	0.52	0.62	1.48	1.48	1.49	1.50	1.57	1.60
Upper 95% CI	0.59	0.55	0.57	0.64	0.68	1.62	1.58	1.62	1.67	1.85	2.11

Figure 1. Distribution of risk by allele number

Odds ratios (95% CI) for each specific number of risk alleles are shown by diamonds, using 9 alleles as the reference (A). Odds ratios (95% CI) for thresholds of risk alleles are indicated by squares (thus risk associated with carrying 10 alleles and more is compared to 9 alleles and less, and so on). Allele frequency distribution in cases and controls from all populations used in generating the models is shown in columns. Data are shown in tabular form (B) for odds ratios for number of risk alleles and partitioned by various thresholds of risk alleles.

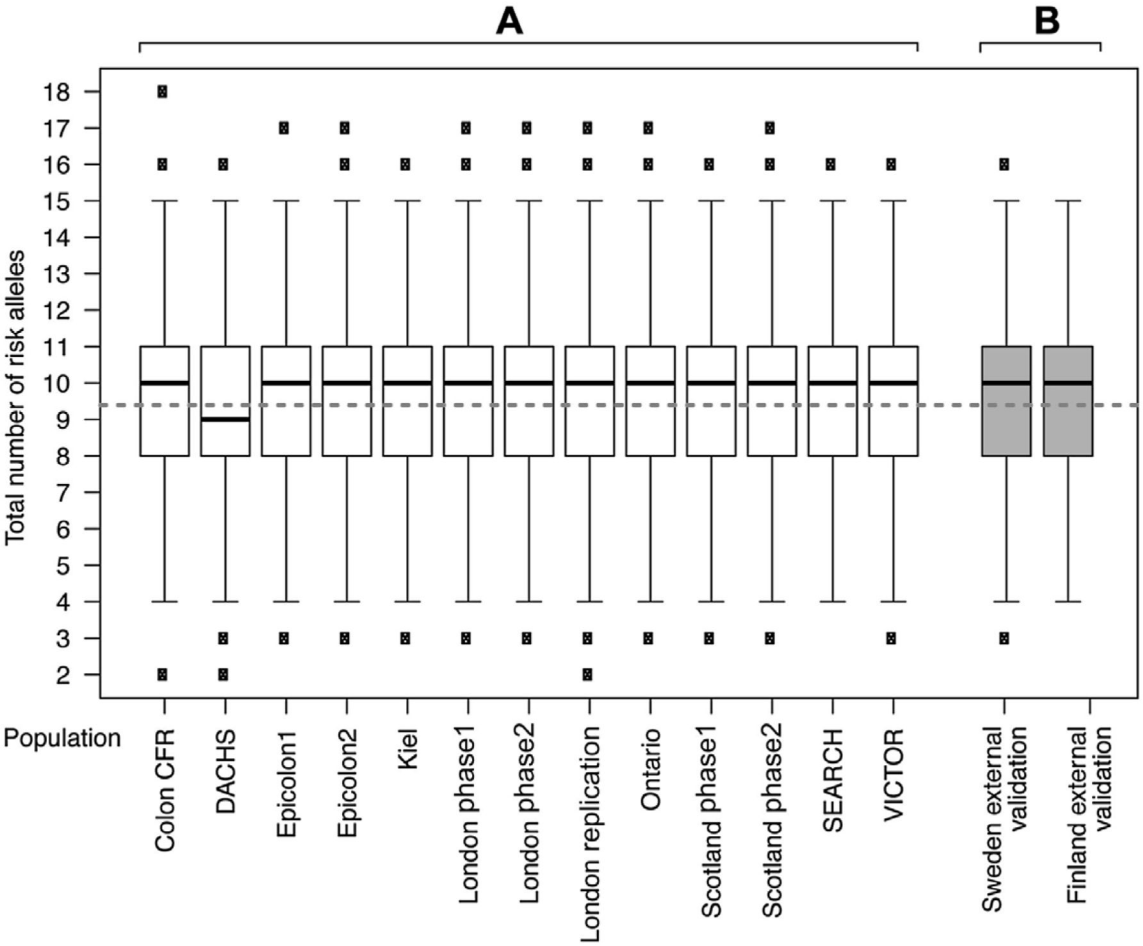
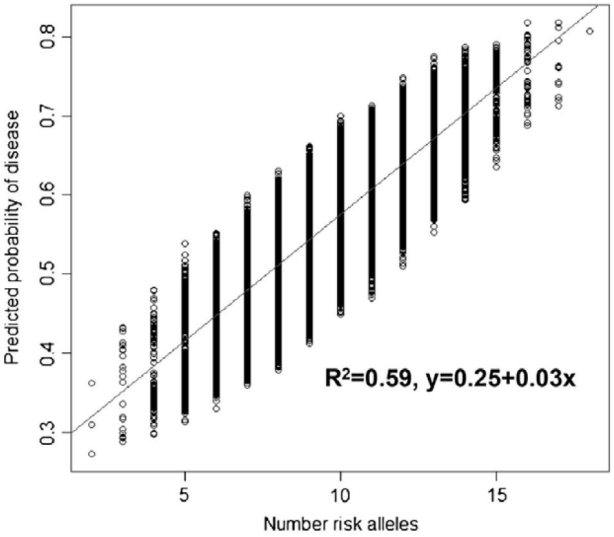


Figure 2. Box plot of risk alleles in case and control subjects by study

Box plot of number of risk alleles in case and control subjects for each study population used in the generation and internal validation of the risk models (**A**) and in the external validation sets from Sweden and Finland (**B**). Median number of risk alleles for cases and controls combined is indicated by a heavy black line. Mean number of alleles in cases by fine solid grey line and broken grey line for controls. There was a marginal difference in median number of risk alleles (9 versus 10) in DACHs compared to other populations, but the difference in mean number of alleles between cases and controls was similar to that in all other populations.



Coefficients				
	Estimate	SE	t value	Pr(>|t|)
Intercept	0.2541780	0.0013387	189.9	$<2\times10^{-16}$
Total	0.0320488	0.0001354	236.6	$<2\times10^{-16}$

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residuals	
Min	−0.137199
1Q	−0.047355
Median	0.007621
3Q	0.046141
Max	0.125501

Residual standard error: 0.05236 on 38 320 degrees of freedom (6073 observations deleted due to missing values). Multiple R-squared: 0.5937, adjusted R-squared: 0.5937. F-statistic: 5.599x104 on 1 and 38,320 DF, p Value: <2.2×10–16

Figure 3.

Variation in predicted probability of CRC (n=39,266) for a given number of risk alleles in the logistic regression model incorporating genotype data.

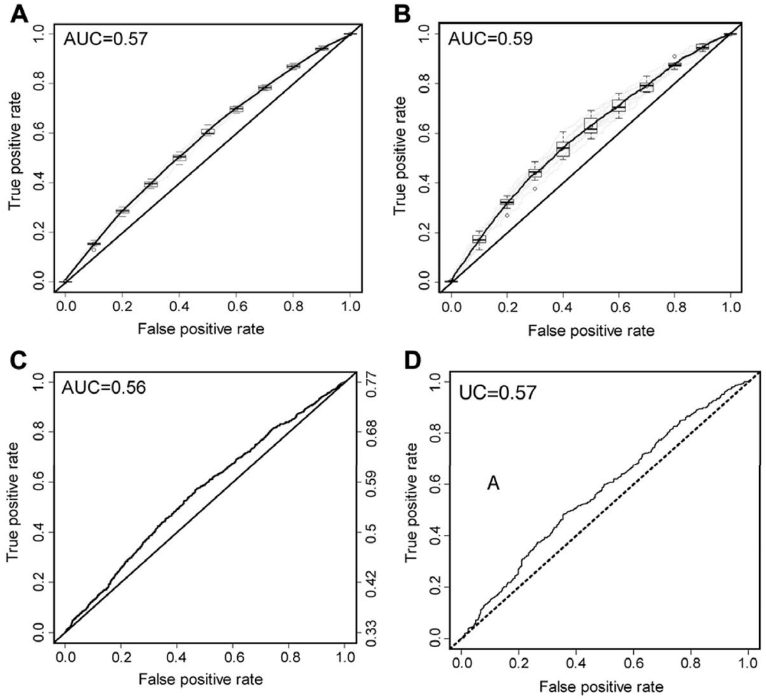


Figure 4.

ROC curves assessing the discriminative ability of the logistic regression model incorporating only genotype data for the 10 risk SNPs (A) (39,266 subjects) and of a model incorporating genotype data for the 10 SNPs along with age, FH status and gender (B) (11,324 subjects). Mean ROC is plotted and the spread of the estimates shown as a box-plot along the ROC curve is shown for A and B. External validation comprised analysis of genotype data from 3,067 Swedish subjects (C) and 1,120 Finnish subjects (D).

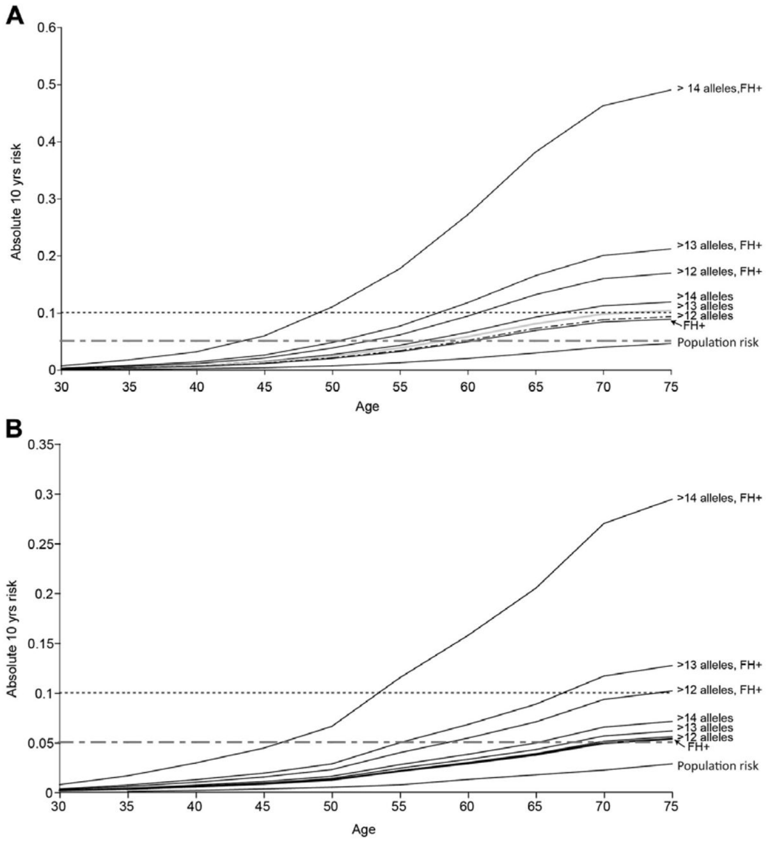


Figure 5. Estimated absolute 10-year CRC risk

10-year absolute risk for cancer-free males (A) and females (B) within the general population carrying >12,>13, >14 risk alleles using 2006 Scottish population estimates (1,310,552 males, 1,441,245 females aged ≥35yrs) using a Bayesian risk modelling approach. The rationale for assessing risk associated with carriage of various numbers of alleles is based on population frequency of that number of alleles and the associated risk (see Figure 1). 10 years is taken as the predicted risk period because it is reasonable to expect colonoscopy to influence CRC stage, mortality and/or incidence over that timescale. Cumulative probability is estimated from 1−exp(−cumulative rate) and the absolute risk in the next 10 years obtained by subtraction of the estimated cumulative risk up to the current age from the estimated cumulative risk for 10 years older than the current age. Risk is shown for males and females in each age group in the average risk population, FH+ subgroups, and by genotype groups (note scale difference in plotting male and female risks).

Table 1

Samples sets used to generate the models comprised UK- (COGS and SOCCS studies), UK – (CORGI and NSCCG studies), UK – VICTOR study; UK - East Anglia (SEARCH); Canada - Ontario (ARCTIC); Spain (EPICOLON1 and EPICOLON2); Melbourne and Seattle (Colon CFR), Germany – Heidelberg and Kiel (DACHS and POPGEN). Recruitment of cases and controls was undertaken with informed consent and ethical review board approval in accordance with the tenets of the Declaration of Helsinki.

Population	Gender	Cases	Controls	Source of Cases*	Source of controls	FH selection in cases or controls	FH data available for case *and* controls
Cambridge (SEARCH)	Male	1277	949	Population-based	Healthy Individuals Sampled From Same Population - frequency matched age/gender	NO	NO
	Female	941	1313				
	Total	2218	2262				
Ontario (OFCR)	Male	514	673	Population-based	Healthy Individuals Sampled From Same Population - frequency matched age/gender	Some case selection *	YES
	Female	676	524				
	Total	1191	1197				
Colon CFR (excludes Ontario Subjects)	Male	463	215	Population-based	Healthy Individuals Sampled From Same Population - frequency matched age/gender	Marginal enrichment for FH+ cases	YES
	Female	442	300				
	Total	905	515				
Heidelberg (DACHS)	Male	789	719	Population-based	Healthy Individuals Sampled From Same Population - Frequency matched by age/gender/ county of residence	NO	YES
	Female	582	760				
	Total	1371	1479				
Epicolon1	Male	649	249	Population-based	Healthy Individuals Sampled From Same Population - Frequency matched age/gender	Controls FH-ve	YES
	Female	447	196				
	Total	1096	445				
Epicolon2	Male	573	320	Population-based	Healthy Individuals Sampled From Same Population - Frequency matched age/gender	Controls FH-ve	YES
	Female	339	229				
	Total	912	549				
Kiel/Greifswald	Male	1089	1059	Population-based	Healthy Individuals Sampled From Same Population - frequency matched age/gender	Controls FH-ve	YES
	Female	1080	1086				
	Total	2169	2145				

Population	Gender	Cases	Controls	Source of Cases*	Source of controls	FH selection in cases or controls	FH data available for case *and* controls
London (CORGI)	Male	275	419	Clinical Genetics Centres across UK	Cancer-free spouses of cases	Controls FH-ves	YES
	Female	335	507				
	Total	610	926				
London (NSCCG)	Male	1159	1094	Population-based Oncology clinics	Cancer-free spouses/friends of cases	NO	NO
	Female	1636	1605				
	Total	2795	2699				
London (NSCCG)	Male	4560	1246	Population-based Oncology clinics	Cancer-free spouses/friends of cases	NO	NO
	Female	2363	2103				
	Total	6925	3352				
Scotland (COGS)	Male	498	514	Population-based Age ≤55yrs	Healthy Individuals Sampled From Same Population -Matched age/gender, area of residence	NO	YES
	Female	482	488				
	Total	980	1002				
Scotland (SOCCS)	Male	1222	1230	Population-based	Healthy Individuals Sampled From Same Population -Matched age/gender, area of residence	NO	YES
	Female	802	862				
	Total	2024	2092				
VICTOR	Male	764	628	Cases recruited to RCT	WTCCC 1958 Birth Cohort and cancer-free spouse controls, and European Cell Culture Collection random human control DNA samples.	NO	NO
	Female	438	706				
	Total	1202	1334				
Total subjects used for model generation and internal validation	Male	13832	9315				
	Female	10563	10679				
	Total	24395	19994				
External validation – Sweden	Total	1,777	1,751	Population-based	Cancer-free blood donor and spouse controls	NO	NO
External validation – Finland	Total	702	418	Population-based	Cancer-free blood donor controls		

Population	Gender	Cases	Controls	Source of Cases*	Source of controls	FH selection in cases or controls	FH data available for case *and* controls
Total study subjects	**Total**	**26,874**	**22,163**	**(49,037)**			

*Population-based refers to systematic collections of all cases diagnosed, identified through hospitals after diagnosis of colorectal cancer. There is minimal case selection based on family history or tumour stage. Cases ascertained through Clinical Genetics Centres are enriched for familial cases. Cases ascertained through oncology clinics are enriched for more invasive disease because referrals to such clinics are based on Tumour Staging. However, there is no enrichment for familial cancer. Patients recruited through the VICTOR study were enriched for poorer stage by nature of being eligible for the trial of adjuvant chemotherapy and rofecoxib.

Table 2

Study populations with available family history data and where there was no (or limited) selection bias on the basis of family history. Family history of CRC was considered as a categorical variable, dependent on the presence or absence of at least one first degree relative affected by CRC at any age at the time of recruitment to the respective study.

	FH	Controls	Cases
Ontario (OFCR) *	No	1039	879
	Yes	155	310
DACHS	No	1313	1187
	Yes	163	180
Scotland COGS	No	936	861
	Yes	66	119
Scotland SOCCS	No	1881	1709
	Yes	211	315
Total	No	5169	4636
	Yes	595	924

*There was a marginal over-representation FH+ in the OFCR series because index cases from high and intermediate risk families from the OFCR registers were over-sampled[44]. However, the majority of OFCR cases were recruited from average risk families and all OFCR controls were unselected with respect to FH.

Table 3

Scottish population data and incidence rates, along with cumulative probably of developing colorectal cancer by age and gender. The Bayesian risk model was used to estimate the 10-year absolute risk associated with carriage >12, >13 or >14 risk alleles in those with a family history of colorectal cancer and irrespective of FH. Estimated Scottish population over age 35yrs in 2006 comprised 1,310,552 males and 1,441,245 females (http://www.isdscotland.org/isd/3535.html)

Age at risk estimation	**0–29**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**
Males											
Scottish population											
Colorectal cancer registrations	3	6	8	22	53	72	158	214	276	345	330
Number at risk in age group	653725	153686	185147	194867	183306	164736	169377	135028	113650	94702	69739
Cumulative probability of CRC	0.0001	0.0003	0.0005	0.0011	0.0026	0.0047	0.0094	0.0172	0.0290	0.0466	0.0689
Male 10-year absolute risk											
Average risk population	0.0003	0.0004	0.0008	0.0020	0.0036	0.0068	0.0125	0.0197	0.0294	0.0398	0.0462
≥12 risk alleles	0.0007	0.0013	0.0033	0.0060	0.0113	0.0209	0.0336	0.0515	0.0723	0.0877	0.0929
≥13 risk alleles	0.0007	0.0014	0.0037	0.0066	0.0125	0.0231	0.0371	0.0569	0.0799	0.0969	0.1026
≥14 risk alleles	0.0009	0.0016	0.0042	0.0076	0.0145	0.0267	0.0429	0.0658	0.0924	0.1120	0.1187
FH+ population	0.0006	0.0012	0.0032	0.0057	0.0108	0.0199	0.0321	0.0491	0.0690	0.0837	0.0887
FH+, ≥12 risk alleles	0.0012	0.0023	0.0060	0.0109	0.0206	0.0380	0.0612	0.0937	0.1316	0.1596	0.1692
FH+, ≥13 risk alleles	0.0015	0.0029	0.0075	0.0136	0.0258	0.0476	0.0766	0.1173	0.1648	0.1998	0.2118
FH+, ≥14 risk alleles	0.0036	0.0068	0.0174	0.0315	0.0596	0.1102	0.1771	0.2714	0.3812	0.4622	0.4898
Females											
Scottish population											
Colorectal cancer registrations	5	2	12	25	51	71	106	132	222	218	270
Number at risk in age group	632555	163497	199628	210261	194619	170649	175422	144542	129719	117673	98732
Cumulative probability of CRC	0.0002	0.0003	0.0006	0.0012	0.0025	0.0046	0.0076	0.0121	0.0205	0.0295	0.0427
Female 10-year absolute risk											
Average risk population	0.0003	0.0004	0.0009	0.0019	0.0034	0.0051	0.0075	0.0129	0.0174	0.0222	0.0285
≥12 risk alleles	0.0006	0.0015	0.0031	0.0056	0.0084	0.0126	0.0219	0.0299	0.0389	0.0512	0.0559
≥13 risk alleles	0.0007	0.0016	0.0035	0.0062	0.0093	0.0139	0.0241	0.0330	0.0430	0.0566	0.0617
≥14 risk alleles	0.0008	0.0019	0.0040	0.0071	0.0108	0.0160	0.0279	0.0382	0.0497	0.0654	0.0714

Age at risk estimation	**0–29**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**
FH+ population	0.0006	0.0014	0.0030	0.0053	0.0080	0.0120	0.0209	0.0285	0.0371	0.0489	0.0533
FH+, ≥12 risk alleles	0.0011	0.0027	0.0057	0.0102	0.0153	0.0229	0.0398	0.0544	0.0708	0.0933	0.1017
FH+, ≥13 risk alleles	0.0014	0.0034	0.0071	0.0127	0.0192	0.0286	0.0498	0.0681	0.0886	0.1167	0.1273
FH+, ≥14 risk alleles	0.0031	0.0078	0.0165	0.0294	0.0444	0.0662	0.1152	0.1575	0.2050	0.2700	0.2944

Table 4

A. Results of logistic regression to assess the effect of genotype at each of the 10 risk loci for samples with genotypes at all 10 SNPs. B. Effect of genotype, age, gender and family history for study populations where study design did not involve case or control selection on the basis of FH criteria (see Table 2). The anticipated effects of gender and age were blunted due to the case-control design for each study population but because the datasets were very large and case-control matching was imprecise in most series (due to frequency matching), a significant effect on CRC risk was observed for age.

SNP	Location	Estimate	SE	Pr(>|z|)	OR	Lower 95%CI	Upper 95%CI
A. Study populations with SNP genotype data for all ten risk loci (n=39,266)							
rs10411210	19q13	0.12	0.02	$2.07{\times}10^{-6}$	1.13	1.07	1.18
rs9929218	16q22	0.11	0.02	$1.60{\times}10^{-11}$	1.11	1.08	1.15
rs6983267	8q24	0.17	0.01	$< 2 \times10^{-16}$	1.19	1.15	1.22
rs4779584	15q23	0.13	0.02	6.86×10^{-14}	1.14	1.10	1.18
rs4939827	18q21	0.19	0.01	$< 2 \times10^{-16}$	1.21	1.18	1.25
rs3802842	11q23	0.13	0.02	$< 2 \times10^{-16}$	1.14	1.11	1.18
rs10795668	10p14	0.11	0.02	6.53×10^{-13}	1.12	1.09	1.15
rs16892766	8q23	0.20	0.03	3.32×10^{-15}	1.23	1.16	1.29
rs961253	20p12	0.10	0.02	4.68×10^{-12}	1.11	1.08	1.14
rs4444235	14q22	0.09	0.01	2.77×10^{-9}	1.09	1.06	1.12
B. Study populations with genotypes for all ten risk loci. Not selected for FH (n=11,324)							
Age		−0.01	0.00	4.16×10^{-5}	0.99	1.00	0.99
Gender M>F		0.00	0.04	0.97	1.00	1.08	0.93
FH		0.51	0.06	$< 2 \times10^{-16}$	1.66	1.87	1.48
rs10411210	19q13	0.16	0.05	1.26×10^{-3}	1.17	1.29	1.06
rs9929218	16q22	0.13	0.03	3.16×10^{-5}	1.14	1.21	1.07
rs6983267	8q24	0.16	0.03	3.77×10^{-8}	1.17	1.24	1.11
rs4779584	15q23	0.15	0.03	2.35×10^{-5}	1.16	1.24	1.08
rs4939827	18q21	0.18	0.03	4.39×10^{-10}	1.19	1.26	1.13
rs3802842	11q23	0.19	0.03	1.89×10^{-10}	1.21	1.29	1.14
rs10795668	10p14	0.06	0.03	0.057	1.06	1.12	1.00
rs16892766	8q23	0.24	0.05	7.41×10^{-7}	1.27	1.40	1.16
rs961253	20p12	0.15	0.03	6.35×10^{-7}	1.16	1.23	1.09

SNP	Location	Estimate	SE	Pr(>|z|)	OR	Lower 95%CI	Upper 95%CI
rs4444235	14q22	0.09	0.03	2.34×10^{-3}	1.09	1.15	1.03

Table 5

Results of 10 successive iterations of validation of the logistic regression model in those subjects with age, sex and genotype data who were not selected in any way by FH criteria and in subjects all subjects with genotype data at every SNP.

Iteration	Age, sex, FH, 10 genotypes (11,324 subjects)	10 genotypes alone (39,266 subjects)
	AUC	
1	0.61	0.57
2	0.59	0.57
3	0.60	0.58
4	0.61	0.58
5	0.62	0.59
6	0.59	0.57
7	0.56	0.57
8	0.60	0.58
9	0.58	0.57
10	0.58	0.57
Mean	**0.59**	**0.57**