2013-02-07

Cumulative impact of 10 common genetic variants on colorectal cancer risk in 42,333 individuals from eight populations

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

Gut. 2013 Jun;62(6):871-81.
http://doi.org/10.1136/gutjnl-2011-300537
http://hdl.handle.net/10616/41406

Cumulative impact of 10 common genetic variants on colorectal cancer risk in 42,333 individuals from eight populations.

[1]Malcolm G Dunlop[§]*, [1]Albert Tenesa*, [1]Susan M Farrington, [1]Marion Walker, [1,2]Evropi Theodoratou, [1]James G Prendergast, [1]Rebecca A Barnetson, [3]Nicola Cartwright, [4]Roseanne Cetnarskyj, [1]David H Brewster, [3]Mary E Porteous, [5]Thibaud Kossler, [5]Paul DP Pharoah, [6, 7]Clemens Schafmayer, [7]Dieter Bröring, [8]Stefan Schreiber, [8]Stephan Buch, [8]Jochen Hampe, [9]Henry Völzke, [10]Jenny Chang-Claude, [10]Michael Hoffmeister, [10]Hermann Brenner, [11]Susanna von Holst, [11]Simone Picelli, [11]Annika Lindblom, Swedish Low-Risk Colorectal Cancer Study Group, [12]Mark A. Jenkins, [12]John L Hopper, [13]Dan Buchanan, [13]Joanne Young, [14]Christopher K Edlund, [14]David V Conti, [14]Graham Casey, [15]David Duggan, [16]Polly Newcomb, [17]Anna Abulí, [17]Xavier Bessa, [17]Montserrat Andreu, [18]Ceres Fernández-Rozadilla, [18]Angel Carracedo, [18]Clara Ruiz-Ponte, [19]Victoria Gonzalo, [19]Antoni Castells, [19]Sergi Castellví-Bel, [20]EPICOLON consortium, [21]Iina Niittymäki, [21]Sari Tuupanen, [21]Auli Karhu, [21]Lauri Aaltonen, [22,23,24]Brent W Zanke, [22,25,26]Celia MT Greenwood, [25]Jagadish Rangrej, [22]Rafal Kustra, [27]Alexandre Montpetit, [22, 23,28]Thomas J Hudson, [22, 29]Steven Gallinger, [30]Ella Barclay, [30]Lynn Martin, [30]Maggie Gorman, [30]Luis Carvajal-Carmona, [30]Sarah Spain, [30]Zoe Kemp, [30]Kimberley Howarth, [30]Enric Domingo, [30]Axel Walther, CORGI Consortium, [30]Jean-Baptiste Cazier, [31]Rachel Mager, [31]Elaine Johnstone, [31]Rachel Midgely, [31]David Kerr, [32]Steven Lubbe, [32]Peter Broderick, [32]Ian Chandler, [32]Alan Pittman, [32]Steven Penegar, COGENT consortium, [1, 2]Harry Campbell▲, [30]Ian Tomlinson▲, [32]Richard S Houlston▲.

*Joint first authors.

▲Joint authors at this position.

[§]Corresponding Author
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

Affiliations

1. Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Edinburgh EH4 2XU.
2. Public Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK
3. Southeast of Scotland Clinical Genetic Services, Western General Hospital, Crewe Rd, Edinburgh, EH4 2XU, UK
4. School of Nursing, Midwifery & Social Care, Faculty of Health, Life and Social Sciences, Napier University, Edinburgh, UK.
5. Cancer Research UK Laboratories, Department of Oncology, University of Cambridge, Cambridge CB1 8RN
6. POPGEN Biobank, University Hospital Schleswig-Holstein, Campus Kiel, Schittenhelmstrasse 12, Kiel 24105, Germany.
7. Department of General and Thoracic Surgery, University Hospital Schleswig-Holstein, Campus Kiel, Arnold-Heller-Strasse 3, Kiel 24105, Germany
8. Department of General Internal Medicine, University Hospital, Schleswig-Holstein, Campus Kiel, Arnold-Heller-Straße 3 (Haus 6, Nebengebäude Haus 5), Kiel 24105, Germany.
9. Institut fuer Community Medicine, University Hospital Greifswald, Walther-Rathenau-Strasse 48, Greifswald 17487, Germany.
10. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg (JC-C) and Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany (MH, HB).
11. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden.
12. Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, Parkville, Victoria, Australia.
13. Familial Cancer Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia.
14. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA.
15. Translational Genomics Research Institute (TGen), Phoenix, Arizona, USA.
16. Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, WA 98109–1024. USA.
17. Department of Gastroenterology, Hospital del Mar, Institut Municipal d'Investigació Mèdica (IMIM), Pompeu Fabra University, Barcelona, Catalonia, Spain.
18. Fundación Pública Galega de Medicina Xenómica (FPGMX), CIBERER, Genomic Medicine Group - University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain.
19. Department of Gastroenterology, Hospital Clínic, CIBERehd, IDIBAPS, University of Barcelona, Catalonia, Spain.
20. Gastrointestinal Oncology Group of the Spanish Gastroenterological Association.
21. Department of Medical Genetics, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland.
22. Cancer Care Ontario, 620 University Ave. Toronto Ontario M5G 1L7, Canada.
23. Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada.
24. University of Ottawa, Faculty of Medicine, Division of Hematology, 501 Smythe Rd. Ottawa, Canada K1H 8L6.
25. Genetics and Genome Biology, Hospital for Sick Children, 15-703 TMDT East, 101 College Street, Toronto, ON M5G 1L7 555 Canada.
26. University of Toronto, Department of Public Health Sciences Health Sciences Building, 155 College Street, Toronto, M5T 3M7, Canada.

27. The McGill University and Genome Quebec Innovation Centre. 700 Dr. Penfield Ave. Montreal Quebec, H3G 1A4, Canada.
28. University of Toronto, Departments of Medical Biophysics and Molecular Genetics, Toronto, ON, Canada.
29. Samuel Lunenfeld Research Institute, Mount Sinai Hospital and University of Toronto, 600 University Ave. Toronto Ontario M5G 1X5, Canada.
30. Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.
31. Department of Clinical Pharmacology, Oxford University, Radcliffe Infirmary, Oxford, OX2 6HA, UK.
32. Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK.

Abstract

Background
Stratification of colorectal cancer (CRC) risk within the population offers potential to refine surveillance guidelines and inform preventative interventions. Recent discoveries have shown that 10 common genetic susceptibility variants individually influence CRC risk. We investigated the utility of genetic risk profiling in 42,333 subjects from eight populations of European descent.

Methods
Binary logistic regression was used to assess the effects of age, gender, family history (FH) and genotypes at all 10 CRC susceptibility loci. Risk models were generated by incorporating genotypes alone (n=39,266), or in combination with gender, age and FH (n=11,324). Discriminatory performance was assessed by generating ROC curves from 10-fold internal cross-validation and an independent case-control set (n=3,067). 10-year absolute risk was estimated by modelling genotype and FH with age- and gender-specific population risk.

Findings
There was a highly significant difference in mean and median number of risk alleles in cases compared to controls (median in cases = 10 alleles vs. 9 in controls, $p<2.2 \times 10^{-16}$). However, model discriminative performance across the risk spectrum was limited for genotypes alone (area under curve (AUC) 0.57) or incorporating genotype, age, gender, FH (AUC 0.59). Genotyping of an external case/control set validated the association between number of risk alleles and CRC risk ($p=1.2 \times 10^{-6}$). Mean per-allele increase in risk was 9% (OR 1.09; 95% CI 1.05-1.13). Modelling genotype, FH, age, gender with population risk enabled identification of a population subgroup (4 per 1000) with a predicted 10-year absolute risk of CRC greater than 5%.

Interpretation
This study demonstrates that population subgroups can be identified with a predicted absolute CRC risk sufficiently high as to merit surveillance/intervention, although individualized CRC risk profiling is not currently feasible. Nonetheless, the findings provide the first tangible evidence of public health relevance for data from genome-wide studies in CRC.

Funding
Multiple charitable and government grants.

Introduction

Colorectal cancer (CRC) is one of the most common cancers in Western countries[1]. Although there is a 25-fold variation in CRC incidence worldwide, this is rapidly narrowing due to increasing exposure to "westernised" lifestyle risk factors in countries with historically low rates. Thus, incidence projections indicate that the death toll is set to rise substantially over the next decade. Already the global annual incidence exceeds 1 million cases, accounting for ~9% of all cancer cases[1]. The lifetime risk of developing the disease is ~5% in high incidence populations such as the UK (http://info.cancerresearchuk.org/cancerstats).

Despite substantial progress made over the last twenty years, CRC remains a common cause of cancer death and suffering. Population-based registry data indicate that overall survival from CRC typically remains at only ~50%[2]. Furthermore, even when cure is attainable, there is appreciable morbidity associated with surgery and adjuvant therapies. Population-based screening has been introduced in many countries to identify prevalent disease at an early, curable stage. However, population screening modalities such as faecal occult blood (FOB) testing have only modest sensitivity. Nonetheless, early detection by screening average risk populations has been shown to shift detection to earlier stage disease, with resultant mortality reductions[3]. Incidence reduction is also achievable through colonoscopic polypectomy, both in the general population[4,5] and in genetically defined high risk groups[6,7]. Thus, identifying population subgroups with an increased CRC risk offers the potential of tailoring colorectal surveillance intensity to predicted level of risk.

Twin studies show that inherited predisposition contributes ~35% of trait variance for CRC[8]. However, only around 5% of cases are attributable to highly-penetrant mutations such as DNA mismatch repair defects responsible for Lynch Syndrome or APC mutations that cause familial adenomatous polyposis. Recent genome-wide association studies have identified 10 risk loci for CRC[9-16], thereby confirming the hypothesis that an appreciable component of the excess familial risk is a consequence of common genetic variants.

Risk associated with each of the 10 loci is individually modest, but the impact on CRC incidence is significant because of the high population frequency of risk alleles. Moreover, high absolute risks could be apparent in a subset of the population who carry multiple risk alleles. Whilst the impact of these new discoveries on the clinical management of most individuals may be limited, application of genotype data offers the possibility of identifying a population subgroup with genetic risk that exceeds a predetermined absolute risk threshold triggering clinical intervention. Indeed, it is already established clinical practice to offer colonoscopic surveillance to those with a relatively modest excess risk due to a history of CRC [5], as well as to mutation carriers from Lynch Syndrome families who have a more substantial increase in CRC risk [17-19]. This is an attractive approach because more intensive surveillance may be indicated for those at highest risk, whilst those who do not reach a predetermined threshold need only be offered screening for average risk individuals, as has been proposed for breast cancer[20].

We set out to determine the utility of 10 common genetic variants for (i) profiling the genetic risk of CRC in the population and (ii) identifying population subgroups with sufficiently high risk to merit additional, more intensive or earlier colonic surveillance. We developed and tested models using age, gender, family history and genotype data from up to 42,333 individuals from eight populations of European descent.

Methods

Study subjects

To generate the risk models, we initially 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. Blood sample collection, along with collection of age, gender demographic and clinical data from these cases and controls was undertaken with informed consent and ethical review board approval in accordance with the tenets of the Declaration of Helsinki. The 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, Seattle (Colon CFR), Germany – Heidelberg and Kiel (DACHS and POPGEN). In addition, we genotyped 3,067 subjects (1563 cases and 1504 controls) from Sweden and used data from all 10 SNPs as an external validation. Thus, a total of 47,917 subjects were included in this study and sample numbers from each population are presented in Table 1 along with the nature and source of case and control subjects. Family history information was available for a subset of study subjects. Study populations with available family history data (with limited or no selection bias on the basis of family history) are detailed in Table 2. 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.

Genotyping

DNA purification and quality control are described elsewhere[14,16]. Genotyping was performed using a variety of different platforms currently in use at each of the contributing sites. This included Illumina HumanHap550, Illumina HumanHap300 and 240S, Illumina iSelect, competitive allele-specific PCR KASPar chemistry (KBiosciences), Taqman (Applied Biosystems), single-base primer extension chemistry matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) detection (Sequenom). Primers and probes are all available on request. Appropriate genotyping quality control procedures were in place in each centre, to ensure reproducible results, including genotyping duplicate DNA samples within studies and SNP assays and direct sequencing of subsets of samples to confirm genotyping accuracy. SNPs (chromosomal locations) shown previously to be associated with CRC risk are: rs6983267 (chr 8q24)[9-11,16], rs4779584 (chr 15q23)[13], rs4939827 (chr 18q21)[12,14], rs3802842(chr 11q23)[14,15], rs10795668 (chr 10p14)[14,15], rs16892766 (chr 8q23)[14,15], rs4444235(chr 14q22)[16], rs9929218 (chr 16q22)[16], rs10411210 (chr 19q13)[16], rs961253 (chr 20p12)[16].

Statistical analysis

Allele counts were performed in cases and controls for each of the 10 CRC SNPs (as above) and allele frequencies calculated 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 the mean number of risk alleles between cases and controls.

Risk modelling

Due to missing values in each of the studies, generation and internal validation of the risk model was based on up to 39,266 samples because subjects were required to have complete data for all 10 SNPs. The probability that a person carrying a given number of common risk alleles develops CRC by a particular age was estimated using a Bayesian approach. Probability by age x is expressed as $P(D_x)$. To estimate the probability that carriers of ≥Z alleles (G=1) develop CRC by a given age, then

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

The probability that a person with < Z risk alleles develops CRC by a given age x is

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

We assume that $P(G=1|D_x)$ and $P(G=0|D_x)$ are the same for all x and call these P(G=1|D and P(G=0|D)). This is reasonable given that each allele exerts a constant effect on risk throughout measurable human lifespan [14,16]. P(G=1|D) and P(G=0|D) can be estimated from the data as the frequency of patients with ≥Z or <Z risk alleles, respectively. P(G=1) and P(G=0) were also estimated from control data to gauge the "carrier" frequency of high risk alleles in the general population. We recognise that some control sets were enriched for cancer-free status at the time of sampling and so this may marginally under-represent P(G=1) in the general population.

Multivariate analysis using binary logistic regression was conducted to test the effect of each SNP allele and each covariate. Depending on study population grouping, the model included: genotype data from 10 SNPs - rs10411210, rs9929218, rs6983267, rs4779584, rs4939827, rs3802842, rs10795668, rs16892766, rs961253, rs4444235; family history status; age; gender. Age was included as a continuous variable. An additive model was assumed for each SNP.

Assessment of risk model performance

Risk model performance was assessed by rigorous internal, and external, validation. A 10-fold cross-validation approach and subsequently using a external case/control set from Sweden. Initially, 10-fold cross-validation was used to estimate receiver operator characteristic (ROC) curves. This consisted of random assignation of study subjects and all associated data for that individual into 10 complementary datasets. One dataset at a time was then used as the validation set and the remaining 9 datasets as the training set. The statistical package 'ROCR' was used for the 10-fold cross-validation[21]. Separate ROC curves 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. Next, an external validation was conducted using an independent case/control set from Sweden. The model fitted in the analysis using all 10 SNP genotypes described above was evaluated in 1,563 Swedish cases and 1,504 controls and ROC curves were generated. We estimated the probability of a subject being a case or control and estimated the proportion of true and false positives at different cut-off points.

Determination of absolute risk

In view of comprehensive population coverage and high levels of completeness[22], we used Scottish population and Cancer Registry data as reference for estimation of the probability of developing CRC in the population. We consider this to be a valid assumption because CRC incidence in Scotland is broadly representative of the northern European populations from which case and control sample sets were drawn[2]. Age-specific CRC rate was calculated from 2006 cancer registration data and from age-specific estimates of the Scottish population in that year (http://www.isdscotland.org/isd/3535.html). The cumulative CRC rate for any given age was calculated separately for males and for females as the sum of the age-specific rates up to that age. The cumulative probability was 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. The cumulative probability of developing CRC in the general population by various ages is shown in supplementary Table 3 along with absolute risks in the general population and also by FH status and by genotype.

Results

The distribution of risk allele frequency in all study populations combined is shown in Figure 1 and comparisons between populations as a box plot in Figure 2. For clarity in Figure 1, odds ratios (95% CI) are shown for subjects carrying 4 or less risk alleles as a group, and 14 or more as a group because of the very small numbers of subjects at these extremes. At the high-risk tail of the distribution curve in the combined sample sets, the frequency of carriage of $\geq$12, $\geq$13 and $\geq$14 alleles (equating to P(G=1|D)=1-P(G=0|D)) was respectively 0.205 0.091 and 0.032. Thus, approximately 20% of CRC cases carried 12 or more risk alleles. The frequency in control subjects is the most relevant to instigating preventative measures in the general population and 14.1%, 5.5% and 1.7% respectively carried $\geq$12, $\geq$13 and $\geq$14 alleles (equating to P(G=1)) in this group.

There was a highly significant difference in mean number of risk alleles between cases and controls (2-sided t-test. $p = 2.2 \times 10^{-16}$). The mean number of risk alleles in control subjects was 9.39 compared to 9.93 in cases (difference - 0.53 alleles, 95% CI 0.57-0.49) and median number of alleles in cases was highly significantly different that in controls (10 for cases, 9 for controls, $p<2.2 \times 10^{-16}$ Mann-Whitney' test). There was no evidence of statistically significant interactive effects between any of the 10 loci, compatible with each locus having an independent effect on CRC risk ($p>0.05$ for interaction, testing each locus against all others. Data not shown).

The effects of age, gender, family history and genotype for SNPs tagging each of the risk loci, along with the relative weight contributed by each variable in the logistic regression are shown in Table 3. Because of the case-control design, the expected excess CRC risk for males compared to females was abrogated. However, because the datasets were very large (n=39,266) and case-control matching was imprecise in most series (due to frequency matching), there was a highly significant effect of age on risk. Each year of advancing age had an effect on CRC risk (Table 3B). To enhance power for discovery purposes, some studies had selected familial cases and/or "super-controls" with no other affected relatives. Therefore, to assess the effect of family history in the logistic regression, we incorporated data from population-based studies where *a priori* selection on the basis of family history was absent (Scotland and DACHS), or limited (Ontario – see Table 1 legend). Odds ratios for SNP alleles at each risk locus are provided under an additive model. Family history of CRC, as measured here, imparted additional risk over and above that imparted by genotype (Table 3B) and so the FH and genotype provide complementary information on risk.

The discriminative ability (probability that cases have a higher score than controls) of incorporating SNP genotypes at all 10 loci alone (Supplementary Fig 1A), or in combination with gender, age and family history data (Supplementary Fig 1B), was assessed by ROC. 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 Supplementary Table 2. Variability in discriminative ability by number of risk alleles incorporated into the model is shown in the Supplementary Figure 2, demonstrating the relationship between increasing discriminative ability with increasing number of risk alleles. Fitting genotype data from the external validation set (3067 subjects) generated an AUC of 0.56 (Supplementary Figure 3C). The association of the total number of risk alleles (the SCORE) was highly significant ($P=1.2 \times 10^{-6}$). On average each allele increased risk of CRC by 9% (OR 1.09, 95% CI 1.05-1.13). Despite the highly significant enrichment for risk alleles in cases compared to controls ($p = 2.2 \times 10^{-16}$) and taking the results of ROC analyses together, it is clear that models incorporating genotype data, age, gender and FH

information have limited predictive performance across the observed risk spectrum and allele distributions for the individual. We estimate that the 10 common variants have an overall accuracy of prediction of genetic risk of only 26%.

Next we assessed the impact on absolute risk of incorporating SNP genotypes for the population subgroup carrying a high number of risk alleles. Taking account of population allele frequency and effect size of various risk allele combinations, we incorporated genotype, family history, age and gender with population CRC incidence data. We estimated 10-year absolute risk of CRC at various ages for males and for females carrying at least 12 or at least 13 risk alleles. The rationale for selection of carriage of these numbers of risk alleles is a pragmatic one, since the frequency in the general population of $\geq$12 and $\geq$13 alleles is 14.1% and 5.5%, and the associated odds ratios at these thresholds are OR=1.58 and OR=1.71 respectively (Figure 1). We used Scottish population data because detailed family history data was collected systematically for Scottish study subjects, who themselves were sampled from the same population used to estimate absolute risks. We consider here that CRC risks calculated from Scottish population data are broadly representative of risk within northern European and North American populations. The risk associated with a positive FH in the Scottish dataset was OR = 1.75 (95% CI 1.48-2.06), marginally lower than estimated in a recent meta-analysis of family history as a risk factor[23]. 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)[24].

Using genotype data alone, or in combination with FH, the 10-year absolute risks of developing CRC between ages 30 and 75yrs were estimated by applying the model including genotype data ($\geq$12 and >13 risk alleles) and incorporating prior probabilities calculated from 2006 Scottish population and cancer registration data. Absolute 10-year risk is highly relevant in clinical practice as it is a practical timescale in which colonscopy can be expected to influence outcome. Absolute risks by age are shown for females (Figure 4a) and for males (Figure 4b). It is possible to categorise population groups using various level of absolute risk and here we considered thresholds of 5% and 10% 10-year risk for genotyped individuals.

The estimated Scottish population aged 35-85 years in 2006 comprised 1,310,552 males and 1,441,245 females. Using absolute risks presented in Figure 4, we estimated the number of males and females in the Scottish population where genotyping could identify those at >5% and >10% 10-year CRC risk. No males or females in the general population reach a 5% predicted 10-year absolute risk without genotype information (supplementary Table 3). The independent effect of genotype in the model in addition to family history should be noted (Figure 4a and 4b). Risk over 10 years was used for the estimates for practical reasons because it is a timescale within which a clinical surveillance intervention could realistically be expected to reduce cancer risk. The predicted 10-year risk in the general population was less than 5% for all men aged <65yrs and all women aged <70yrs, whilst only men in the age group 75-79yrs have >10% 10-year risk of CRC. Thus using actual Scottish population values, 39,211 men with $\geq$12 risk alleles of 278,091 aged 65-80 years and 30,513 women of 216,405 aged 70-80 reach the 5% risk. Only 3,836 men with $\geq$13 risk alleles of the 69,739 aged 75-79yrs reach a 10% threshold of 10-year absolute risk.

Since a positive family history is associated with an increased risk of CRC[23] and there is an evidence base for advancement of the age of recruitment to FOBT screening for FH+ individuals[25], we extrapolated the findings to the Scottish population. Thus, genotyping the ~37,000 men aged 60-80 years with a positive family history would identify ~5,200 with a >5% 10-year absolute risk of CRC. Similarly, genotyping

~31,000 women aged 65-80 years with a family history could identify ~4,400 reaching that risk threshold. Genotyping for common variants could thus be used to refine empiric guidance for people with a family history of CRC, as well as providing additional information on risk, over and above FH.

Discussion

We set out to assess the potential utility of genetic risk profiling for CRC in the general population. We used genotype information at 10 loci known to be associated with CRC susceptibility. Whilst the discriminative ability of any of the models was limited across the risk spectrum, we show that combining genotype data with gender, age and family history information (as a proxy for genetic susceptibility factors yet to be discovered) can identify individuals with a substantially raised 10-year absolute risk of CRC. We propose that the risk level is sufficiently high for those with at least 12 alleles in specific age groups that additional screening measures are warranted, such as colonoscopic surveillance and/or age advancement of recruitment to population FOBT screening programmes.

This study of 8 different populations of European origin comprised analysis of data from 42,333 subjects genotyped for all 10 genetic risk loci so far identified. In addition to the expected association with a positive family history and increasing age, there was a highly significant difference between cases and controls in the mean, and median, number of risk alleles at the 10 CRC susceptibility loci ($p<2.2 \times 10^{-16}$). We generated models to combine genotype information at each of the 10 loci associated with CRC, age, gender and family history variables, but we found limited discriminative ability to differentiate cases from controls (AUC ~0.59 and 0.57 (internal validation) or 0.56 (external validation set) from ROC curve analysis of models including or excluding FH information) and an overall positive predictive value between 0.51 and 0.71 for cut-off points of 0.4 and 0.7 respectively, The negative predictive value for the same cut-offs ranged between 0.62 and 0.51. This modest level of test performance was consistent across study populations in the internal and external validation steps. This suggests that risk assessment algorithms based on common genetic variants will have similar performance characteristics in Caucasian populations and are unlikely to be confounded by population differences, since the study populations all have very similar LD structure.

Despite the large dataset studied here, we found that common genetic variants associated with CRC which have been identified to date cannot be used alone or in combination with each other and age, gender and family history for individualized profiling across the risk spectrum in the general population. This is consistent with the recent risk prediction studies in other diseases which have been disappointing to date[26-28]. Typical AUCs have ranged from 0.55 to 0.60 in type 2 diabetes[29-32] to slightly higher values for age-related macular degeneration (AMD), Crohn's Disease, coronary heart disease and cardiovascular diseases[33-35]. The best predictive performances have been obtained by combining genetic, demographic, and environmental variables as for AMD[36]. An important issue is that the great majority of true susceptibility loci are not been included in these analyses because they have yet to be discovered. The four published prediction models for type 2 diabetes studied around 20 SNPs and this is typical of these studies. A substantially improved predictive performance (AUC > 0.8] can be achieved by including SNPs from a much larger number of susceptibility loci[35]. Consistent with this, we have estimated that a model with approximately 100 SNPs of the estimated 172 loci accounting for all the genetic variance for CRC could provide an 80% accuracy of prediction, and explain ~17% of the phenotypic variance in the liability scale[37].

Although accurate *individual* risk assessment is not currently possible, the findings presented here suggest that it is possible to partition population subgroups by absolute risk thresholds. Furthermore, we show that the number of people in this high risk subgroup is manageable, indicating that genotyping could be feasible. The use of a threshold of a 5% absolute 10-year risk of CRC has clinical and public health validity. The 5% risk exceeds the highest risk achieved throughout life in males or

females and is ten times greater than the risk of a 50-year old entering the population-based FOBT screening in the populations studied here.

ROC analysis indicates modest discriminative ability for any of the predictive models studied here. However, AUC 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 a predictive test where the main aim is categorisation into clinically meaningful risk strata[28]. AUC does not address absolute levels of risk or whether the model stratifies correctly into high/low categories of absolute risk which are of clinical importance (eg 10-year risk of CRC). Prediction of actual risk is a more important function of the model than the sensitivity/specificity (on which the ROC curve and AUC estimate are based).

It is likely that 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[28]. It would be expected that these will have a greater impact if they are involved in novel disease pathways which are independent from the causal mechanisms through which the other risk factors operate[27]– as is likely in this study in which many of the variants are involved in the TGB beta signalling pathway[37].

This study explored model performance across a range of European populations in order to reduce potential bias due to limited representativeness of study data for the settings in which the genetic testing will be applied. 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.

The findings presented here have implications for the current CRC FOB screening programmes. Brenner has argued that the "risk advancement" associated with a positive family history should logically dictate that these individuals should enter the screening programme about 10 years earlier that those with a negative family history (based on equivalent 10 year risks of CRC and assuming equal programme effectiveness in these two groups)[25].

With current intense research activity, it is likely that additional common genetic variants associated with CRC risk will be identified. Because the effect sizes are likely to be small and/or the allele frequencies may be lower than those identified to date. Nonetheless, predictive utility of testing for common genetic variants is likely to improve with new discoveries and individualized genetic risk profiling for CRC may become a reality in the foreseeable distant future. Furthermore, performance should improve further when the causal variants at these loci are discovered through fine mapping and functional studies.

Whilst there are a number of issues that need to be addressed in order to translate any genetic test into clinical and public health practice, we have shown, in principle, that it is already possible to identify a proportion of the population at substantially increased risk of CRC. These individuals have sufficiently high risk to merit individual intensive surveillance. Furthermore, amendments to criteria for age of entry to family history focussed surveillance programmes[5,38] should be considered and evaluated. This study provides tangible evidence that data from genome-wide studies of CRC have public health and clinical relevance.

Table 1: Description of study type with category of case/control subjects genotyped

*Although the majority of recruited cases were low risk, there was some enrichment of OFCR cases for FH+ since all index cases who came from high and intermediate risk families were included[39]. All OFCR controls were unselected with respect to FH.

		Cases	Controls	Study type	Source of controls	Case or control selection by FH	FH data available for case *and* controls
Cambridge (SEARCH)	Male	1277	949	Population-based	Population - frequency matched age/gender	NO	NO
	Female	941	1313				
	Total	2218	2262				
Ontario (OFCR)	Male	514	673	Population-based	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	Population - frequency matched age/gender	YES. FH-ve super-controls	YES
	Female	442	300				
	Total	905	515				
Heidelberg (DACHS)	Male	789	719	Population-based	Frequency matched by age/gender/county of residence	NO	YES
	Female	582	760				
	Total	1371	1479				
Epicolon1	Male	649	249	Population-based	Frequency matched age/gender	YES. FH-ve super-controls	YES
	Female	447	196				
	Total	1096	445				
Epicolon2	Male	573	320	Population-based	Frequency matched age/gender	YES. FH-ve super-controls	YES
	Female	339	229				
	Total	912	549				
Kiel/Greifswald	Male	1089	1059	Population-based	Population - frequency matched age/gender	YES. FH-ve super-controls	YES
	Female	1080	1086				
	Total	2169	2145				
London (CORGI)	Male	275	419	Population-based Clinical genetics centres	Cancer-free spouses of cases	YES. FH+ve enriched cases	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	Matched age/gender, area of residence	NO	YES
	Female	482	488				

	Total	980	1002				
Scotland (SOCCS)	Male	1222	1230	Population-based	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
Total study subjects	**Total**	**26,172**	**21,745**	**(47,917)**			

Table 2: Population-based studies with family history information.

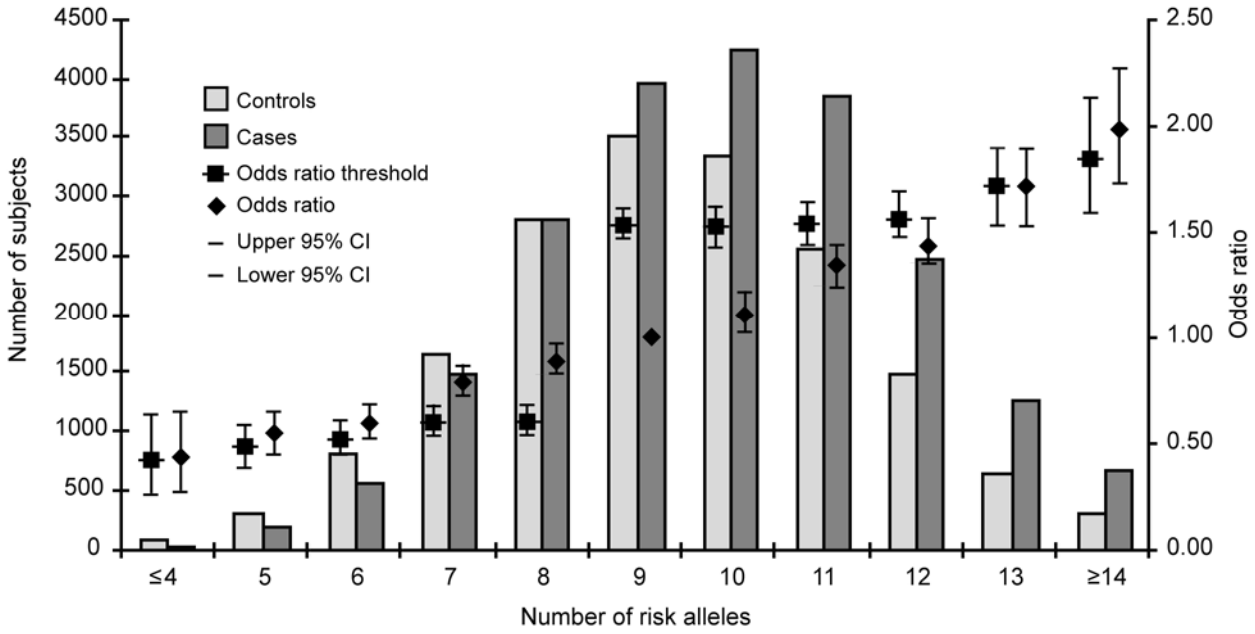
	FH	Controls	Cases
Ontario (OFCR)	No	1039	879
	Yes	155	310
DACHS	No	1313	1187
	Yes	163	180
Scotland COGS	No	936	861
	Yes	66	119
Scotland SOCCS	No	1881	1709
	Yes	211	315
Total	No	5169	4636
	Yes	595	924

Table 3.
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).

	Estimate	SE	Pr(>|z|)	OR	Upper 95%CI	Lower 95%CI
A. Study populations with SNP genotype data for all ten risk loci (n=39,266)						
rs10411210	0.12	0.02	2.07×10^{-6}	1.13	1.18	1.07
rs9929218	0.11	0.02	1.60×10^{-11}	1.11	1.15	1.08
rs6983267	0.17	0.01	$< 2 \times 10^{-16}$	1.19	1.22	1.15
rs4779584	0.13	0.02	6.86×10^{-14}	1.14	1.18	1.10
rs4939827	0.19	0.01	$< 2 \times 10^{-16}$	1.21	1.25	1.18
rs3802842	0.13	0.02	$< 2 \times 10^{-16}$	1.14	1.18	1.11
rs10795668	0.11	0.02	6.53×10^{-13}	1.12	1.15	1.09
rs16892766	0.20	0.03	3.32×10^{-15}	1.23	1.29	1.16
rs961253	0.10	0.02	4.68×10^{-12}	1.11	1.14	1.08
rs4444235	0.09	0.01	2.77×10^{-9}	1.09	1.12	1.06
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	$< 2 \times 10^{-16}$	1.66	1.87	1.48
rs10411210	0.16	0.05	1.26×10^{-3}	1.17	1.29	1.06
rs9929218	0.13	0.03	3.16×10^{-5}	1.14	1.21	1.07
rs6983267	0.16	0.03	3.77×10^{-8}	1.17	1.24	1.11
rs4779584	0.15	0.03	2.35×10^{-5}	1.16	1.24	1.08
rs4939827	0.18	0.03	4.39×10^{-10}	1.19	1.26	1.13
rs3802842	0.19	0.03	1.89×10^{-10}	1.21	1.29	1.14
rs10795668	0.06	0.03	0.057	1.06	1.12	1.00
rs16892766	0.24	0.05	7.41×10^{-7}	1.27	1.40	1.16
rs961253	0.15	0.03	6.35×10^{-7}	1.16	1.23	1.09
rs4444235	0.09	0.03	2.34×10^{-3}	1.09	1.15	1.03

Figure 1 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.

A



B

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**
Odds ratio	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**
Odds ratio	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 2: 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 (**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.

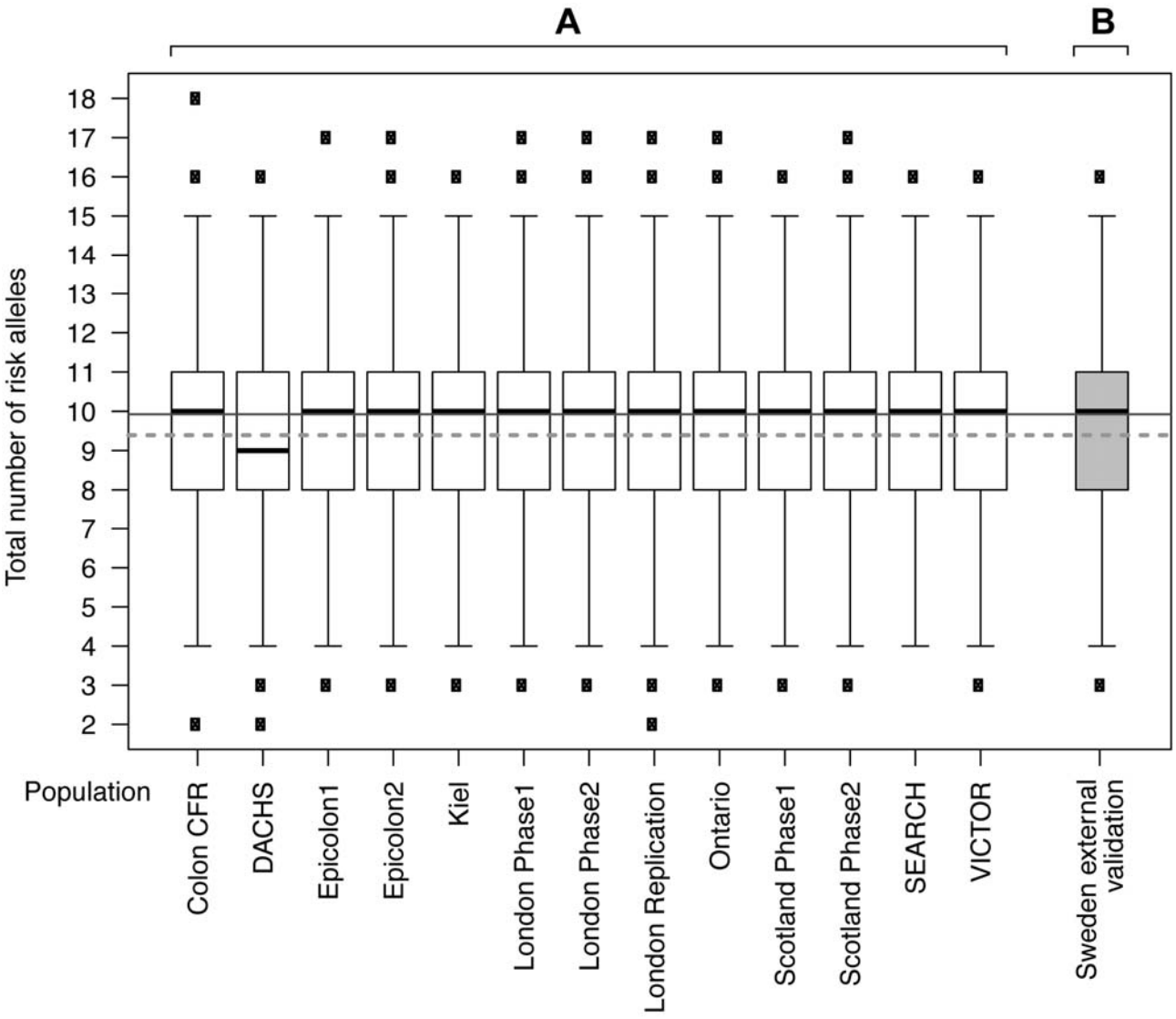
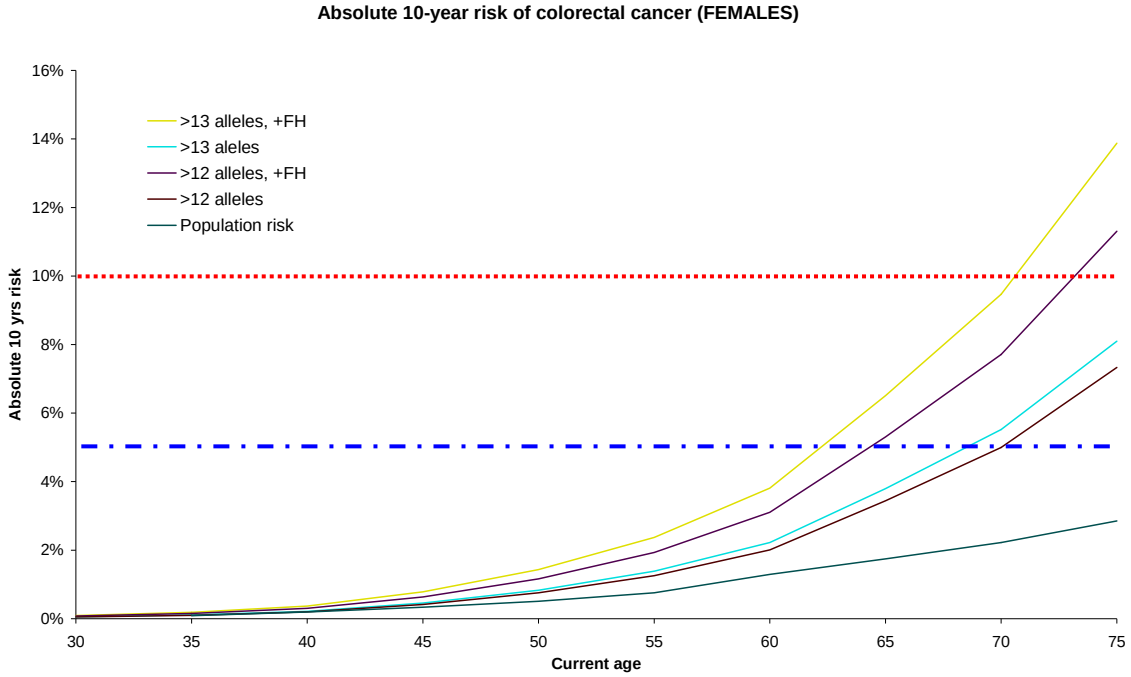
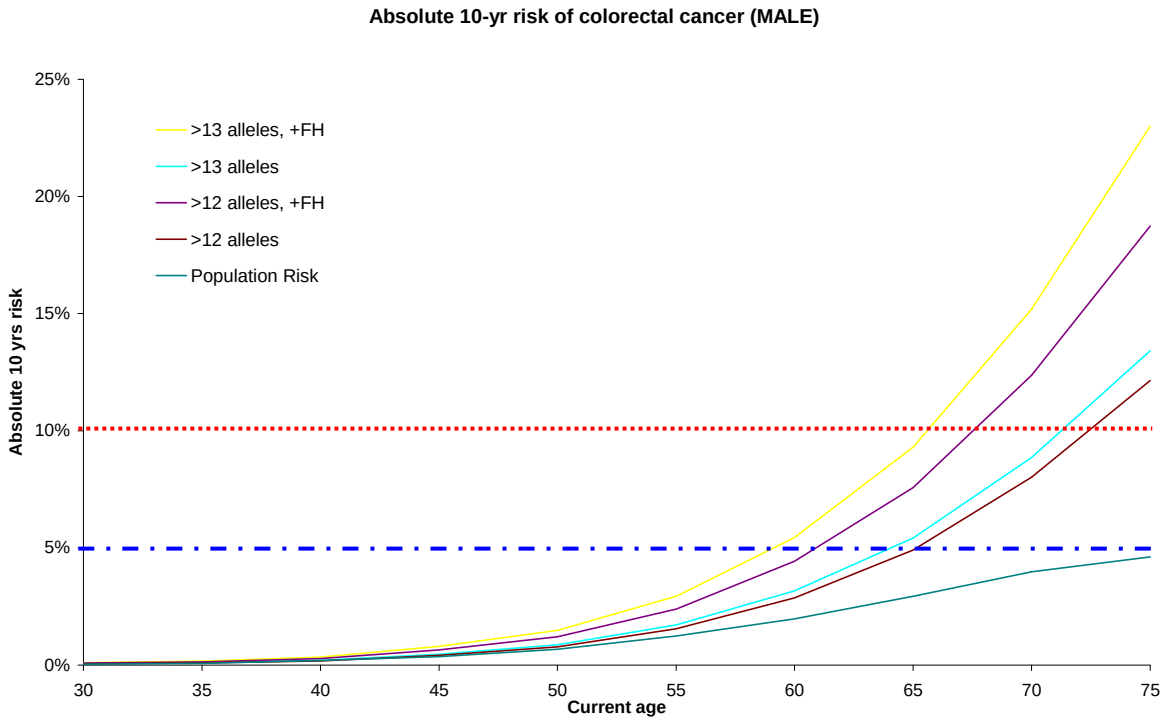


Figure 4
Absolute 10-year risk of CRC for cancer-free females (A) and males (B) within the general population carrying ≥12 or ≥13 risk alleles. (Note scale difference in absolute risk between female and male graphs).

A



B



Supplementary Table 1

Odds ratio (95% CI) estimated by logistic regression by number of risk alleles with reference to 9 alleles as the median number of risk alleles in the study populations.

	Estimate	SE	z value	Pr(>|z|)	OR	Upper_95% CI	Lower_95% CI
(Intercept)	0.12	0.02	5.29	1.23×10^{-7}	1.13	1.18	1.08
≤4	-0.82	0.18	-4.43	9.34×10^{-6}	0.44	0.63	0.31
5	-0.56	0.09	-5.95	2.64×10^{-9}	0.57	0.69	0.48
6	-0.50	0.06	-8.42	$<2 \times 10^{-16}$	0.61	0.68	0.54
7	-0.23	0.04	-5.42	5.93×10^{-8}	0.79	0.86	0.73
8	-0.12	0.04	-3.54	3.94×10^{-4}	0.88	0.95	0.82
10	0.11	0.03	3.48	4.99×10^{-4}	1.12	1.19	1.05
11	0.29	0.03	8.31	$<2 \times 10^{-16}$	1.33	1.42	1.24
12	0.38	0.04	9.43	$<2 \times 10^{-16}$	1.46	1.58	1.35
13	0.53	0.05	9.96	$<2 \times 10^{-16}$	1.70	1.88	1.53
≥14	0.68	0.07	9.37	$<2 \times 10^{-16}$	1.97	2.27	1.71

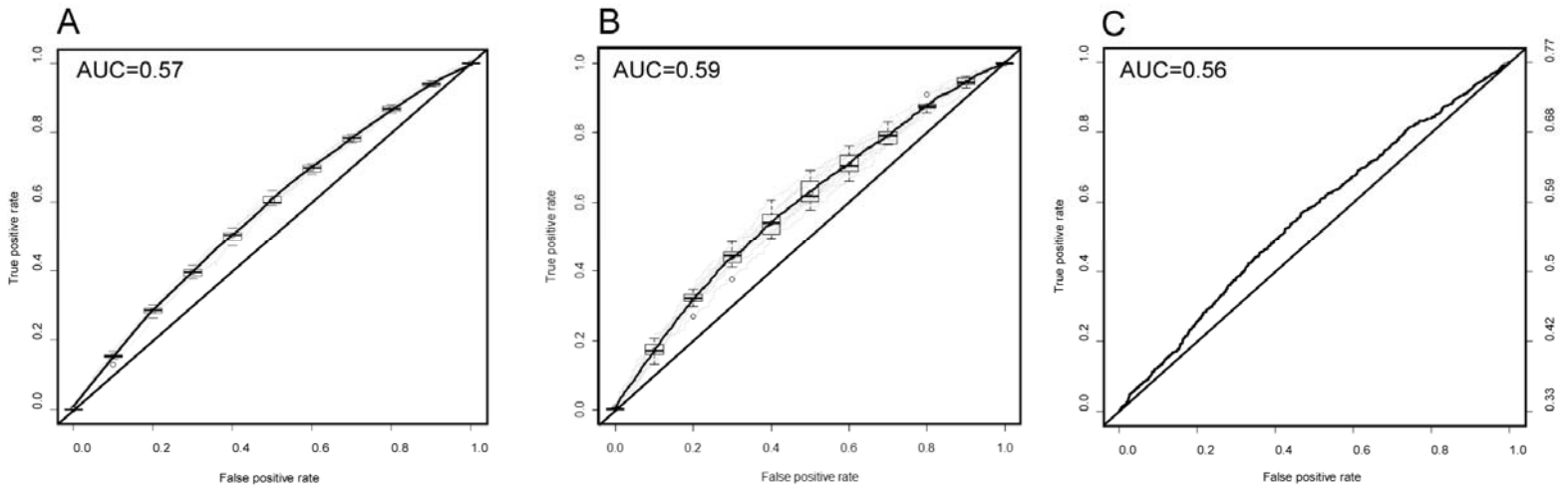
Supplementary Table 2.
Results of 10 successive iterations of validation in those subjects with age, sex and genotype data who were not selected in any way by FH criteria and in subjects all subjects with genotype data at every SNP.

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**

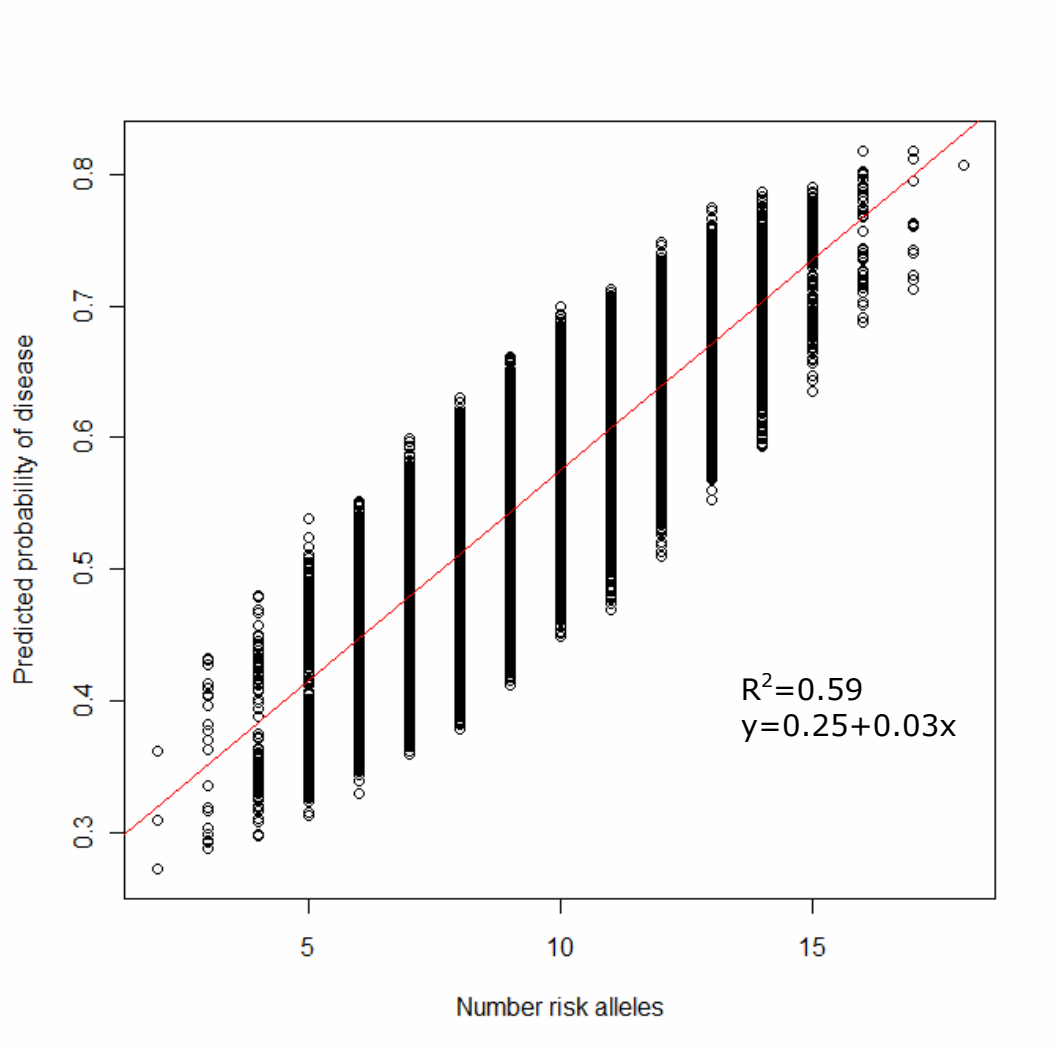
Supplementary Table 3
Scottish population data and incidence rates, along with cumulative probably of developing CRC by age and gender. 10-year risk associated with carriage of 12 or more and 13 or more alleles in those with a family history of CRC and irrespective of FH.

Age group	**0-29**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**
Scottish population											
Males in age group	653725	153686	185147	194867	183306	164736	169377	135028	113650	94702	69739
Population 10-yr risk - male	0.000	0.000	0.001	0.002	0.004	0.007	0.012	0.020	0.029	0.040	0.046
Cumulative probability of CRC-male		0.01%		0.05%		0.26%		0.94%	1.72%	2.90%	4.66%
Females in age group	632555	163497	199628	210261	194619	170649	175422	144542	129719	117673	98732
Population 10-yr risk - female		0.000	0.001	0.002	0.003	0.005	0.008	0.013	0.017	0.022	0.029
Cumulative probability of CRC-female		0.02%		0.06%		0.25%		0.76%	1.21%	2.05%	2.95%
10-yr risks FH+											
Male >=12 alleles, 10-yr risk	0.000	0.001	0.001	0.003	0.006	0.012	0.024	0.044	0.076	0.124	0.187
Female >=12 alleles, 10-yr risk	0.001	0.001	0.002	0.003	0.006	0.012	0.019	0.031	0.053	0.077	0.113
Male >=13 alleles, 10-yr risk	0.000	0.001	0.002	0.003	0.008	0.015	0.029	0.054	0.093	0.152	0.230
Female >=13 alleles, 10-yr risk	0.001	0.001	0.002	0.004	0.008	0.014	0.024	0.038	0.065	0.095	0.139
10-yr risks irrespective of FH status											
Male >=12 alleles, 10-yr risk	0.000	0.001	0.001	0.002	0.004	0.008	0.016	0.029	0.049	0.080	0.121
Female >=12 alleles, 10-yr risk	0.000	0.000	0.001	0.002	0.004	0.008	0.013	0.020	0.034	0.050	0.073
Male >=13 alleles, 10-yr risk	0.000	0.001	0.001	0.002	0.005	0.009	0.017	0.032	0.054	0.089	0.134
Female >=13 alleles, 10-yr risk	0.000	0.001	0.001	0.002	0.005	0.008	0.014	0.022	0.038	0.055	0.081

Supplementary Figure 1
ROC curves assessing the discriminative ability of a 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 subjects (C).



Supplementary Figure 2
Variation in predicted probability of CRC (n=39,266) for a given number of risk alleles in the model incorporating genotype data.



Coefficients

	Estimate	SE	t value	Pr(>|t|)
Intercept	0.2541780	0.0013387	189.9	$<2x10^{-16}$
Total	0.0320488	0.0001354	236.6	$<2x10^{-16}$

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: $5.599x10^4$ on 1 and 38,320 DF, p-value: $< 2.2x10^{-16}$

Acknowledgements

Edinburgh: Work was supported by grants from Cancer Research UK (C348/A8896, C348/A6361 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). ET is funded by a Cancer Research UK Fellowship (C31250/A10107). We acknowledge the Wellcome Trust Clinical Research Facility in Edinburgh for sample preparation and some genotyping. Cambridge: We thank the SEARCH study team and all the participants in the study. 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. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating institutions or investigators in the Colon CFR, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government or the Colon CFR. 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. E Domingo and I Tomlinson are supported by the Oxford

Biomedical Research Centre. The study made use of genotyping data on the 1958 Birth Cohort. Genotyping data on controls was generated and generously supplied to us by Panagiotis Deloukas of the Wellcome Trust Sanger Institute. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We are grateful to all participating centres, all laboratory members involved in sample preparation. We are grateful to colleagues at the UK National Cancer Research Network and also those at UK Clinical Genetics Centres and the UK National Cancer Research Network. We are grateful to colleagues at OCTO, and all centres that participated in the VICTOR trial. The work was 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. We are grateful to colleagues at the UK National Cancer Research Network.

Bibliography

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005;**55**(2)**:**74-108.
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, Kewenter J, Weller D, Silagy C. A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, hemoccult. *Bmj* 1998;**317**(7158)**:**559-65.
4. 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.
5. 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.
6. Jarvinen HJ, Renkonen-Sinisalo L, Aktan-Collan K, Peltomaki P, Aaltonen LA, Mecklin JP. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med* 2008;**358**(26)**:**2796-803.

21. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;**21**(20)**:**3940-1.
22. Brewster DH, Crichton J, Harvey JC, Dawson G. Completeness of case ascertainment in a Scottish regional cancer registry for the year 1992. *Public Health* 1997;**111**(5)**:**339-43.
23. 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.
24. Mitchell RJ, Campbell H, Farrington SM, Brewster DH, Porteous ME, Dunlop MG. Prevalence of family history of colorectal cancer in the general population. *Br J Surg* 2005;**92**(9)**:**1161-4.
25. Brenner H, Hoffmeister M, Haug U. Family history and age at initiation of colorectal cancer screening. *Am J Gastroenterol* 2008;**103**(9)**:**2326-31.
26. 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.
27. Janssens AC, van Duijn CM. Genome-based prediction of common diseases: methodological considerations for future research. *Genome Med* 2009;**1**(2)**:**20.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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.
33. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann Intern Med* 2009;**150**(2)**:**65-72.
34. van der Net JB, Janssens AC, Sijbrands EJ, Steyerberg EW. Value of genetic profiling for the prediction of coronary heart disease. *Am Heart J* 2009;**158**(1)**:**105-10.
35. 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.
36. Seddon JM, Reynolds R, Maller J, Fagerness JA, Daly MJ, Rosner B. 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.
37. 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.
38. 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.
39. 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.