Prevalence and Penetrance of Major Genes and Polygenes for Colorectal Cancer

Running title: Prevalence & Penetrance of Major Genes & Polygenes for CRC

Aung Ko Win,[1] Mark A. Jenkins,[1] James G. Dowty,[1] Antonis C. Antoniou,[2] Andrew Lee,[2] Graham G. Giles,[1,3] Daniel D. Buchanan,[1,4] Mark Clendenning,[4] Christophe Rosty,[5] Dennis J. Ahnen,[6] Stephen N. Thibodeau,[7] Graham Casey,[8] Steven Gallinger,[9] Loïc Le Marchand,[10] Robert W. Haile,[11] John D. Potter,[12,13,14] , Yingye Zheng,[12,13] Noralane M. Lindor,[15] Polly A. Newcomb,[12,13] John L. Hopper,[1] Robert J. MacInnis.[1,3,*]

[1] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria, Australia.
[2] Centre for Cancer Genetic Epidemiology, Department of Public and Primary Care, University of Cambridge
[3] Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Victoria, Australia.
[4] Colorectal Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Parkville, Victoria, Australia.
[5] Queensland Institute of Medical Research, Brisbane, Queensland, Australia
[6] University of Colorado School of Medicine, Denver, Colorado, USA.
[7] Molecular Genetics Laboratory, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA.
[8] Department of Preventive Medicine, Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA.
[9] Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada.
[10] University of Hawaii Cancer Center, Honolulu, Hawaii, USA.
[11] Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University, California, USA.
[12] School of Public Health, University of Washington, Seattle, Washington, USA.
[13] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
[14] Centre for Public Health Research, Massey University, Wellington, New Zealand.
[15] Department of Health Science Research, Mayo Clinic Arizona, Scottsdale, Arizona, USA.

[*]**Corresponding author**
Robert J. MacInnis, PhD
Cancer Epidemiology Centre
Cancer Council Victoria
615 St Kilda Road
Melbourne VIC 3004 Australia
Phone: +61 3 9514 6248
Email: robert.macinnis@cancervic.org.au

FUNDING

This work was supported by grant UM1 CA167551 from the National Cancer Institute, National Institutes of Health (NIH) and through cooperative agreements with the following Colon Cancer Family Registry (CCFR) centers: Australasian Colorectal Cancer Family Registry (U01/U24 CA097735), Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (U01/U24 CA074800), Ontario Familial Colorectal Cancer Registry (U01/U24 CA074783), Seattle Colorectal Cancer Family Registry (U01/U24 CA074794), and USC Consortium Colorectal Cancer Family Registry (U01/U24 CA074799).

Seattle CCFR research was also supported by the Cancer Surveillance System of the Fred Hutchinson Cancer Research Center, which was funded by Control Nos. N01-CN-67009 (1996-2003) and N01-PC-35142 (2003-2010) and Contract No. HHSN2612013000121 (2010-2017) from the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute with additional support from the Fred Hutchinson Cancer Research Center.

The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute; and the Centers for Disease Control and Prevention's National Program of Cancer Registries, under agreement U58DP003862-01 awarded to the California Department of Public Health. The ideas and opinions expressed herein are those of the author(s) and endorsement by the State of California, Department of Public Health the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors is not intended nor should be inferred.

P.A. Newcomb, M.A. Jenkins, J.G. Dowty, J.L. Hopper, N.M. Lindor, R.J. MacInnis and Y. Zheng received support for this study by grant R01CA170122 from NIH. M.A. Jenkins, J.L. Hopper and G.G. Giles received further support from Centre for Research Excellence grant APP1042021 and program grant APP1074383 from National Health and Medical Research Council (NHMRC), Australia. A.K. Win is a NHMRC Early Career Fellow. M.A. Jenkins is an NHMRC Senior Research Fellow. J.L. Hopper is a NHMRC Senior Principal Research Fellow. D.D. Buchanan is a University of Melbourne Research at Melbourne Accelerator Program (R@MAP) Senior Research Fellow. A.C. Antoniou is a Cancer Research UK Senior Research Fellow (C12292/A11174).

DISCLAIMER

The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US

Government or the CFR. Authors had full responsibility for the design of the study, the collection of the data, the analysis and interpretation of the data, the decision to submit the manuscript for publication, and the writing of the manuscript.

DISCLOSURE

The authors have no conflict of interest to declare with respect to this manuscript.

ACKNOWLEDGEMENTS

The authors thank all study participants of the Colon Cancer Family Registry and staff for their contributions to this project. We also thank Associate Professor James McCaw for use of his UNIX computer cluster.

ABSTRACT

Background: While high-risk mutations in identified major susceptibility genes (DNA mismatch repair genes and *MUTYH*) account for some familial aggregation of colorectal cancer, their population prevalence and the causes of the remaining familial aggregation are not known.

Methods: We studied the families of 5,744 colorectal cancer cases (probands) recruited from population cancer registries in the USA, Canada and Australia and screened probands for mutations in mismatch repair genes and *MUTYH*. We conducted modified segregation analyses using the cancer history of first-degree relatives, conditional on the proband's age at diagnosis. We estimated the prevalence of mutations in the identified genes, the prevalence of and hazard ratio for unidentified major gene mutations, and the variance of the residual polygenic component.

Results: We estimated that 1 in 279 of the population carry mutations in mismatch repair genes (*MLH1*= 1 in 1946, *MSH2*= 1 in 2841, *MSH6*= 1 in 758, *PMS2*= 1 in 714), 1 in 45 carry mutations in *MUTYH,* and 1 in 504 carry mutations associated with an average 31-fold increased risk of colorectal cancer in unidentified major genes. The estimated polygenic variance was reduced by 30-50% after allowing for unidentified major genes and decreased from 3.3 for age <40 years to 0.5 for age ≥70 years (equivalent to sibling relative risks of 5.1 to 1.3, respectively).

Conclusion: Unidentified major genes might explain one-third to one-half of the missing heritability of colorectal cancer.

Impact: Our findings could aid gene discovery and development of better colorectal cancer risk prediction models.

Article type: Research Article

Word count: 3,987 (excluding references), 5 tables, and 1 figure

Abstract word count: 247

Keywords: colorectal cancer, risk prediction, polygenes, family history, familial aggregation

INTRODUCTION

One of the most important risk factors for colorectal cancer is having a family history of the disease. First-degree relatives of persons diagnosed with colorectal cancer are, on average, at an approximately two-fold increased risk of colorectal cancer compared with those without a family history (familial relative risk) (1). An estimated 3% to 5% of colorectal cancers are caused by high-risk mutations in the identified major colorectal cancer susceptibility genes(2): DNA mismatch repair (MMR) genes(3) and constitutional 3' end deletions of EPCAM(4, 5) implicated in Lynch syndrome; the adenomatous polyposis coli (*APC*) gene implicated in familial adenomatous polyposis(6-8); and the *MUTYH* gene implicated in colorectal polyps and subsequently cancer (*MUTYH*-associated polyposis)(9). Current estimates of MMR gene mutation carriers in the general population, inferred from the prevalence of mutations in cases and the risk of colorectal cancer for carriers, range widely from approximately 1 in 300 to 1 in 3,000 depending on differing assumptions and genes (10-16). With the availability of cost-effective sequencing technologies, improved precision in estimates of mutation prevalence would be useful for devising cost-effective genetic testing protocols.

Less than half of the excess risk of colorectal cancer associated with family history (familial aggregation) is explained by mutations in the above identified genes, and only two studies have attempted to explain the remainder of the familial aggregation (17, 18). Aaltonen *et al* could not confidently distinguish between different modes of inheritance for the hypothetical unidentified major genes (17). Jenkins *et al* estimated that 1 in 588 of the population carry major gene mutations associated with a recessively inherited risk, and these mutations would explain 15% of all colorectal cancers diagnosed before age 45 years (18).

Both these studies relied on relatively small numbers of families and did not consider the existence of both polygenic and major genes.

While much research has been conducted on the search for other major colorectal cancer susceptibility genes in addition to those described above, only a few have been confirmed (19). Genome-wide association studies have identified at least 45 independent genetic susceptibility markers (single-nucleotide polymorphisms, SNPs) that are reliably associated with small increments in the risk of developing colorectal cancer (20).

The aim of this paper was to use population-based family data to estimate: the prevalence of mutations in the identified major colorectal cancer susceptibility genes (MMR genes and *MUTYH*); the prevalence, average penetrance, and likely mode of inheritance for the unidentified major gene mutations; and the variance of the residual polygenic component before and after allowing for different major gene scenarios.

MATERIALS AND METHODS

Sample

The sample consists of nuclear families from the Colon Cancer Family Registry which has been described in detail previously(21,22). The present study used data for the first-degree relatives of the incident colorectal cancer cases (probands) who had been recruited irrespective of family history from state or regional population cancer registries in the USA (Washington, California, Arizona, Minnesota, Colorado, New Hampshire, North Carolina), Australia (Victoria) and Canada (Ontario) between 1997 and 2012. Families were excluded if the proband was known to have an *APC* mutation. Informed consent was obtained from all study participants, and the study protocol was approved by the institutional research ethics review board at each recruiting site of the Colon Cancer Family Registry.

Data Collection

Information on demographics, personal characteristics, personal and family history of cancer, cancer-screening history, history of polyps, polypectomy, and other surgeries was obtained by questionnaires from all probands at baseline recruitment, which was about 1-2 years after diagnosis of their colorectal cancer, and from all participating relatives. The questionnaires are available from the Colon Cancer Family Registry website(23). We sought confirmation of all reported cancer diagnoses and ages at diagnosis for relatives using pathology reports, medical records, cancer registry reports, and death certificates, where possible. We attempted to obtain blood or buccal samples from all participants and tumor tissue from all affected participants.

Mismatch Repair (MMR) gene mutation screening

All probands had their colorectal cancers tested for MMR deficiency, defined by either tumor microsatellite instability (MSI) and/or lack of MMR protein expression by immunohistochemistry (IHC). Probands with a MMR-deficient tumor were screened for germline mutations in MMR genes. *MLH1, MSH2* and *MSH6* mutations were identified using Sanger sequencing or denaturing high performance liquid chromatography (dHPLC), followed by confirmatory DNA sequencing. Large duplication and deletion mutations including those involving *EPCAM*, which lead to *MSH2* methylation, were detected by Multiplex Ligation Dependent Probe Amplification (MLPA) according to the manufacturer's instructions (MRC Holland, Amsterdam, The Netherlands) (21,24,25). *PMS2* mutations were identified using a modified protocol from Senter *et al*(26) where exons 1-5, 9 and 11-15 were amplified in three long range PCRs followed by nested exon specific PCR/sequencing. The remaining exons (6, 7, 8 and 10) were amplified and sequenced directly from genomic DNA. Large-scale deletions in *PMS2* were detected using the P008-A1 MLPA kit according to manufacturers specifications (MRC Holland, Amsterdam, The Netherlands). Germline variants were classified for pathogenicity based on 5 class system for quantitative assessment of variant pathogenicity(27) and the application of a multifactorial likelihood model developed for MMR gene variants(28) as applied to variants catalogued within the InSiGHT database (29) where classes 4 and 5 were considered pathogenic (30). For variants not yet classified by InSiGHT, we considered a variant as pathogenic if it resulted in a stop codon, frameshift, large deletion, or if it removed a canonical splice site. The relatives of probands with a pathogenic MMR germline mutation, who provided a blood sample, underwent testing for the specific mutation identified in the proband.

***MUTYH* mutation testing**

Population-based probands were tested for 12 previously identified *MUTYH* variants: c.536A>G p.(Tyr179Cys), c.1187G>A p.(Gly396Asp), c.312C>A p.(Tyr104Ter), c.821G>A p.(Arg274Gln), c.1438G>T p.(Glu480Ter), c.1171C>T p.(Gln391Ter), c.1147delC p.(Ala385ProfsTer23), c.933+3A>C p.(Gly264TrpfsX7), c.1437_1439delGGA p.(Glu480del), c.721C>T, p.(Arg241Trp), c.1227_1228dup p.(Glu410GlyfsX43), and c.1187-2A>G p.(Leu397CysfsX89) using the MassArray MALDI-TOF Mass Spectrometry (MS) system (Sequenom, San Diego, CA) (31). To confirm the *MUTYH* mutation and identify additional mutations, screening of the entire *MUTYH* coding region, promoter, and splice site regions was performed on all samples exhibiting MS mobility shifts using denaturing high-performance liquid chromatography (Transgenomic Wave 3500HT System; Transgenomic, Omaha, NE). All MS-detected variants and WAVE mobility shifts were submitted for sequencing for mutation confirmation (ABI PRISM 3130XL Genetic Analyser). That is, if a heterozygous *MUTYH* mutation was identified, then the *MUTYH* gene was screened for any additional mutations not captured by the Sequenom genotyping screen to ensure all potential compound heterozygous carriers were identified. The relatives of probands with a pathogenic *MUTYH* germline mutation, who provided a blood sample, underwent testing for the specific variant identified in the proband. For the present study, *MUTYH* gene mutation status was recorded as monoallelic or biallelic mutation-positive or negative, with no distinction between different variants.

Statistical Methods

We used modified segregation analysis to fit a range of genetic models to the observed colorectal cancer family histories for the proband and their first-degree relatives. Individuals were assumed to be at risk of colorectal cancer from birth until the earliest of the

following: diagnosis of colorectal cancer or any other cancer (except skin cancer); first polypectomy; death; and the earlier of last known age at baseline interview or age 80 years.

The colorectal cancer incidence $\lambda_i(t,k)$ for individual i at age t in sex group k ($k = 1$ for males or 2 for females) was assumed to depend on genotype according to a parametric survival analysis model $\lambda_i(t,k)=\lambda_0(t,k)\exp(G_i+P_i(t))$, where $\lambda_0(t,k)$ is the sex-specific baseline incidence at age t . G_i is the natural logarithm of the relative risk associated with the major genotype and $P_i(t)$ is the polygenic component for age t.

The major genotype was defined by six components representing each of the genes *MLH1, MSH2, MSH6, PMS2, MUTYH* and one representing the hypothetical unidentified major genes. We fitted models in which the unidentified major genes were autosomal with a normal and a mutant allele unlinked to mutations in the MMR genes or *MUTYH*. We also fitted models in which the average relative risk for the unidentified major genes was assumed to be age dependent. We used the published age-, sex- and country-specific incidences for *MLH1* and *MSH2* mutation carriers (32), and published age- and sex-specific incidences for *MSH6, PMS2* and *MUTYH* mutation carriers (26, 33, 34).

The polygenic component for age t, $P_i(t)$, was assumed to be normally distributed with zero mean and variance $\sigma^2_p(t)$. P was approximated by the hypergeometric polygenic model (35, 36). We also fitted models where the variance of the polygenic 'modifying' component was allowed to take a different value σ^2_m for MMR gene and *MUTYH* carriers.

To compute the baseline colorectal cancer incidence $\lambda_0(t)$, we constrained the overall incidence of colorectal cancer to agree with the national age- and sex-specific incidences (1998-2002) separately for Australia, Canada and USA (37). Other cancers were ignored in this model.

We assumed that the sensitivity of the mutation testing of probands for MMR genes and *MUTYH* was 80%,(38) and we examined the effect of varying this sensitivity. For relatives, we assumed the mutation screening for the proband's mutation (i.e. predictive testing) was 100% sensitive and specific.

The genetic models were specified in terms of colorectal cancer incidence for MMR gene and *MUTYH* mutation carriers, the frequency (q_A) of the putative high risk allele ''A'' of the unidentified major genes component, the average relative risk of colorectal cancer for carriers of mutations in the unidentified major genes, and the variances of the polygenic and modifying components (σ^2_p and σ^2_m). Maximum likelihood estimation was used to estimate parameters. The estimates we present are the values that were the most likely (i.e. most consistent) with the data. Maximum likelihood is the optimal method for making such estimates, and provides confidence intervals (CIs). We adjusted for ascertainment by maximizing the likelihood of each pedigree conditioned on the colorectal cancer status of the proband and his or her age of diagnosis (but not the mutation carrier status as this information was not known at the time of recruitment).

The relative goodness of fit for nested models was tested by the likelihood ratio test. The Akaike's Information Criterion(39) [AIC=-2xlog-likelihood + 2x(no. of parameters)] was used to assess goodness of fit between non- nested models (40).

The expected versus observed number of affected relatives under each fitted model was assessed using the Pearson χ^2 goodness of fit statistic. The expected number of probands with MMR and *MUTYH* mutation carriers for families that had undergone mutation testing based on their cancer family history was computed using Bayes theorem (41). Statistical methods are described further in the Appendix.

RESULTS

A total of 5,744 families was eligible for inclusion, including 37,634 first-degree relatives of probands of whom 50% were female and 806 (2%) had been diagnosed with colorectal cancer (Table 1). Nearly two-thirds of the families were recruited from the USA (63%), with 16% and 21% of families recruited from Australia and Canada, respectively. Seventy-three percent of the probands were Caucasian whereas the rest were African American (17%), Asian (6%), Latino (1%), Native American (1%) and unknown (2%).

Approximately 7% of all probands (N=386) had been found to have a MMR-deficient colorectal tumour and therefore had been screened for germline mutations in the MMR genes, while two-thirds of all probands (N=3,796) had been tested for germline mutations in *MUTYH.* Of the probands who were screened, 136 had a MMR gene mutation (49 in *MLH1*, 39 in *MSH2*, 24 in *MSH6* and 24 in *PMS2*) and 81 had a *MUTYH* mutation (63 monoallelic and 18 biallelic) (Table 2).

All seven models that incorporated a polygenic component and the hypothetical unidentified major genes provided significantly better fits than the model that included only MMR gene and *MUTYH* mutation carriers (all $P<0.001$) (Supplementary Table 1). The mixed dominant model was essentially identical to a mixed codominant model in terms of fit (likelihood ratio test, $P=0.94$), but was more parsimonious given it used less parameters. All other models were rejected when compared with the mixed codominant model (likelihood ratio test, all $P<0.001$).

When we allowed the polygenic variance to vary by age, the mixed dominant model for the unidentified major genes was the most parsimonious (i.e., had the lowest AIC) compared with all other models fitted (Table 3). Under this model, we estimated 0.19% (95%

CI, 0.04% – 1.08%) of the population carry mutations in unidentified major genes, and these are associated with on average a 31-fold (95% CI, 12 – 83) increased risk of colorectal cancer. The estimated variance of the polygenic component was 3.28 for age <40 years, 0.92 for age 40-49 years, 0.46 for age 50-59 years, 0.79 for age 60-69 years, and 0.52 for age ≥70 years. The proportion of polygenic variance after adjusting for the identified major genes explained by the unidentified major genes was 13%, 54%, 58%, 33% and 36% for ages <40, 40-49, 50-59, 60-69 and ≥70 years, respectively (Figure 1). The estimated population carrier frequency for mutations in *MLH1*, *MSH2*, *MSH6*, *PMS2,* and monoallelic and biallelic *MUTYH* are shown in Table 4.

Table 5 (A) shows the expected versus observed number of relatives of the probands, who developed colorectal cancer before age 80 years. Consistent with the AIC, the expected numbers from the mixed dominant model is closest to the observed numbers.

Table 5 (B) shows the expected and observed number of probands who are mutation carriers for each MMR gene and monoallelic and biallelic *MUTYH* mutations. The expected numbers from the mixed dominant model with an age-dependent polygenic variance were closest to the observed numbers and had the lowest χ^2 compared with other models. In general, all the models closely predicted the number of mutation carriers.

In all the fitted models above, the sensitivity of mutation testing was fixed at 0.80. When we re-fitted the models assuming the sensitivity was 0.90, the impact was negligible. Model estimates were virtually identical when the unidentified major genes were fitted as a separate locus to the MMR mutations and *MUTYH* (not shown).

Results were not materially different when we restricted analyses to Caucasian families (not shown). The relative risks for the unidentified major genes did not vary

appreciably by age in the major gene models (not shown). There was virtually no evidence of a difference between the size of the polygenic variance for non-carriers σ^2_p and the modifying variance σ^2_m for any of the models (not shown).

DISCUSSION

We have used a large population-based family data set from the Colon Cancer Family Registry, and existing penetrance estimates, to produce new estimates of the population prevalence of high-risk mutations in the identified major susceptibility genes for colorectal cancer: the DNA mismatch repair genes and *MUTYH*. We estimated that 1 in 279 (95% CI, 192 – 403) of the population carry mutations in mismatch repair genes (*MLH1* = 1 in 1946, *MSH2* = 1 in 2841, *MSH6* = 1 in 758, *PMS2* = 1 in 714), and 1 in 45 carry mutations in *MUTYH*.

Previously, researchers have inferred these carrier frequencies from the carrier frequency for cases, risk for the general population and risk for mutation carriers (Supplementary Table 2)(10-16). None, except those estimated by Song *et al*(16), were gene specific. Previous estimates of population carrier frequencies for the four MMR mutations combined (or *MLH1* and *MSH2* mutations combined) were similar to our estimates, except for those obtained by Dunlop *et al*(11). This discrepancy might be explained by different screening methods, and that knowledge about which mutations are truly pathogenic has improved substantially over time (30). For *MUTYH* mutations, a systematic review and meta-analysis estimated the population carrier frequency of monoallelic *MUTYH* mutations to be 1 in 60 and biallelic *MUTYH* mutations to be approximately 1 in 7,000, similar to our estimates(42).

We then sought to explain the residual familial aggregation of this disease. We considered a polygenic component that proposes there are multiple independent loci, and across loci and at each locus, the alleles have a multiplicative effect on risk. We also considered the existence of one or more unidentified major genes (genes for which there are mutations associated with a high risk of colorectal cancer), and allowed for different modes of disease inheritance (dominant, recessive and codominant).

We found evidence that there exist as yet unidentified major colorectal cancer susceptibility genes, and their mode of inheritance was most likely dominant (thought this does not necessarily mean that they were all dominant). It is important to note that the apparent dominant component might also reflect missed mutations in MMR genes, *MUTYH* or *APC* because the mutation screening techniques used were not 100% sensitive and not all probands had been screened. We estimated that the 1 in 504 (95% CI, 93 – 2778) of the population carry unidentified mutations associated with an average 31-fold increased risk of colorectal cancer. The estimated polygenic variance was reduced by 30-50% after allowing for these unidentified major genes, after which it decreased from 3.3 for age <40 years to 0.5 for age ≥70 years (equivalent to sibling relative risks of 5.1 to 1.3, respectively).

The term 'missing heritability' has been variously defined over the last decade to refer to the fact that not all the causes of familial aggregation, or of familial aggregation considered to be due to genetic factors, have been found (43). The latter has been addressed by assuming an all-or-nothing unmeasured liability model that makes untestable assumptions (44). For the purposes of discussion here, we assume that heritability encapsulates both genetic and non-genetic causes of familial aggregation. In this regard, it is plausible for common cancers that non-trivial heritability is due to non-genetic factors (45). In this paper, we have fitted a polygenic component to capture familial aggregation not explained by the major genes. It is based on an underlying genetic model of Fisher (1918)(46), but given are studying nuclear families it also represents non-genetic familial factors. That is, although it is labelled polygenic, it could also reflect the effect of environmental and lifestyle factors shared by first-degree relatives. Given that the polygenic variance is proportional to the log of the familial relative risk attributable to the polygenic component, the unidentified major genes might explain one-third to one-half of the missing heritability of colorectal cancer across the ages of 40 to 70 years.

The polygenic component will also capture the currently identified, and as yet unidentified, common SNPs associated with colorectal cancer risk. For example, the current 45 independent susceptibility SNPs explain 22% of familial aggregation (20). It is likely this proportion will increase as larger studies are conducted, such as the OncoArray initiative, and as more informative statistical strategies are used to devise risk-prediction SNP-based scores other than the current highly conservative paradigm of considering each SNP individually and applying stringent penalties for multiple testing. The common SNPs identified to date are not necessarily causal, and they could also be tagging rare causal variants (as was the case for *HOXB13* and prostate cancer (47)).

Our analyses suggest a role for rare variants in as yet undiscovered susceptibility genes associated with high risk. Individually they could be very rare, and difficult to discover. One recent attempt to resolve this issue was a whole exome sequencing study that identified some high-risk mutations in candidate susceptibility genes such as *POT1*, *POLE2* and *MRE11* (19). The authors concluded that the study "probably discounts the existence of further major high-penetrance susceptibility genes, which individually account for >1% of the familial risk". Therefore, if both their and our findings are correct, there is likely to be perhaps hundreds of major genes each contributing little to the missing heritability. As well as sample size, the authors recognized that restriction to exomes limited their ability to identify pathogenic mutations outside of transcribed regions, and that targeted capture is insufficiently sensitive to detected copy number variation. We, therefore, agree with the authors in their conclusion that there is a need for very large-scale sequencing studies that would benefit from including highly informative families.

Strengths of our study include a large number of families ascertained regardless of a family history, standardized questionnaires and protocols used by the Colon Cancer Family

Registry, and sophisticated statistical techniques that properly adjust for ascertainment and account for residual familial aggregation of disease (thereby avoiding bias). We also used a systematic approach for screening and testing of germline mutations in both MMR genes and *MUTYH*.

When predicting the number of relatives with colorectal cancer, we did not differentiate family history of colorectal cancer in terms of tumor location within the bowel. This approach was supported by findings from a large study in Utah, which reported similarly elevated risks of colorectal cancer associated with a family history of colorectal cancer regardless of tumor location (proximal colon, distal colon, and rectum) (48).

The response of the population-based probands approached to participate was 72% (49). MMR gene and *MUTYH* mutation carriers have both been associated with better colorectal cancer survival than non-carriers (50-52). Therefore, if probands with better prognosis are more likely to participate in the study, survivor bias could potentially lead to an overestimation of the mutation frequency. Data on participation differences by prognostic characteristics were not available to assess this.

A potential limitation of our study is inaccurate reporting of family colorectal cancer history. Of the 806 colorectal cancer diagnoses reported by first-degree relatives, 26% were confirmed by pathology reports, clinic records or cancer registries. Previous studies have found reported colorectal cancer history in first-degree relatives to be reasonably accurate (85-90% agreement)(53) so even though the colorectal cancer diagnoses in relatives were not confirmed, it is unlikely to have a great impact on our results.

Another potential limitation of our study is the reliance on external estimates of colorectal cancer relative risks for carriers of MMR gene and *MUTYH* mutations. To help

mitigate this weakness, we used estimates based on the largest studies available, and all used data from the same source, the Colon Cancer Family Registry (26, 32-34). Future studies should focus on incorporating the explicit effects of other colorectal cancer susceptibility genes such as *STK11*(54) *BMPR1A*(55), *SMAD4, PTEN*(56), *POLE* and *POLD1*(57) as well as the explicit effects of identified common low risk alleles(20). In addition to colorectal cancer risk, it is known that MMR gene mutations increase the risks of other cancers such as endometrial and ovarian cancer (58). Our analyses can be extended to incorporate such information.

The polygenic variance describes the range of familial risk across a population at a given age. For example, given the estimated variances by age for the mixed dominant model, the familial relative risk was 5.1, 1.6, 1.3, 1.5 and 1.3 for ages <40, 40-49, 50-59, 60-69 and ≥70 years, respectively. Although we found no evidence that the polygenic effects differed for carriers of a MMR gene mutation compared with non-carriers, this does not imply that they are due to the same variants. Some studies have shown that the common genetic variants identified through GWAS to be associated with the risk for the general population are not relevant for MMR gene mutation carriers (59). If future studies identify specific genetic modifiers of colorectal cancer risk for MMR gene or *MUTYH* mutation carriers, it should be possible to extend the current analyses to allow for this level of complexity.

In conclusion, we have used a large population-based family study to estimate the prevalence of mutations in the identified major colorectal cancer-susceptibility genes, as well as the prevalence and relative risk of yet-to-be-discovered, high-risk susceptibility genes. This is an essential step in the development of a high quality-risk prediction model for colorectal cancer and is a major clinical and public health goal. Subsequently, screening programs can be optimized at an individual level to attain maximum benefit, however that

may be defined. This study also provides a guidepost for future new gene discovery research and will justify, and guide, the use of next-generation sequencing to find these genes. The results show that our current understanding of hereditary predisposition to colorectal cancer is incomplete and supports the existence of yet undiscovered rare but highly penetrant mutations, while also underscoring that the polygenic component is still largely unresolved.

References

1. Taylor DP, Burt RW, Williams MS, Haug PJ, Cannon-Albright LA. Population-Based Family History-Specific Risks for Colorectal Cancer: A Constellation Approach. Gastroenterology. 2010;138:877-85.

2. Rustgi AK. The genetics of hereditary colon cancer. Genes Dev. 2007;21:2525-38.

3. Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895-2015. Nat Rev Cancer. 2015;15:181-94.

4. Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, et al. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nat Genet. 2009;41:112-7.

5. Kovacs ME, Papp J, Szentirmay Z, Otto S, Olah E. Deletions removing the last exon of TACSTD1 constitute a distinct class of mutations predisposing to Lynch syndrome. Hum Mutat. 2009;30:197-203.

6. Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, et al. Identification of FAP locus genes from chromosome 5q21. Science. 1991;253:661-5.

7. Nishisho I, Nakamura Y, Miyoshi Y, Miki Y, Ando H, Horii A, et al. Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. Science. 1991;253:665-9.

8. Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, et al. Identification and characterization of the familial adenomatous polyposis coli gene. Cell. 1991;66:589-600.

9. Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, et al. Inherited variants of MYH associated with somatic G:C -->T:A mutations in colorectal tumors. Nat Genet. 2002;30:227.

10. Salovaara R, Loukola A, Kristo P, Kaariainen H, Ahtola H, Eskelinen M, et al. Population-Based Molecular Detection of Hereditary Nonpolyposis Colorectal Cancer. J Clin Oncol. 2000;18:2193-200.

11. Dunlop MG, Farrington SM, Nicholl I, Aaltonen L, Petersen G, Porteous M, et al. Population carrier frequency of hMSH2 and hMLH1 mutations. Br J Cancer. 2000;83:1643-5.

12. Terdiman JP. HNPCC: an uncommon but important diagnosis. Gastroenterology. 2001;121:1005-8.

13. de la Chapelle A. The incidence of Lynch syndrome. Fam Cancer. 2005;4:233-7.

14. Boland CR, Shike M. Report from the Jerusalem workshop on Lynch syndrome-hereditary nonpolyposis colorectal cancer. Gastroenterology. 2010;139:2197 e1-7.

15. Hampel H, de la Chapelle A. The Search for Unaffected Individuals with Lynch Syndrome: Do the Ends Justify the Means? Cancer Prev Res. 2011;4:1-5.

16. Song W, Gardner SA, Hovhannisyan H, Natalizio A, Weymouth KS, Chen W, et al. Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. Genet Med. 2016;18:850-4.

17. Aaltonen L, Johns L, JÃ¤rvinen H, Mecklin J-P, Houlston R. Explaining the Familial Colorectal Cancer Risk Associated with Mismatch Repair (MMR)-Deficient and MMR-Stable Tumors. Clin Cancer Res. 2007;13:356-61.

18. Jenkins MA, Baglietto L, Dite GS, Jolley DJ, Southey MC, Whitty J, et al. After hMSH2 and hMLH1--what next? Analysis of three-generational, population-based, early-onset colorectal cancer families. Int J Cancer. 2002;102:166-71.

19. Chubb D, Broderick P, Dobbins SE, Frampton M, Kinnersley B, Penegar S, et al. Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. Nature communications. 2016;7:11883.

20. Jenkins MA, Makalic E, Dowty JG, Schmidt DF, Dite GS, MacInnis RJ, et al. Quantifying the utility of single nucleotide polymorphisms to guide colorectal cancer screening. Future Oncol. 2016;12:503-13.

21. Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, et al. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiol Biomarkers Prev. 2007;16:2331-43.

22. Colon Cancer Family Registry. Available from: http://coloncfr.org

23. Colon Cancer Family Registry Questionnaires. Available from: http://coloncfr.org/questionnaires

24. Southey MC, Jenkins MA, Mead L, Whitty J, Trivett M, Tesoriero AA, et al. Use of molecular tumor characteristics to prioritize mismatch repair gene testing in early-onset colorectal cancer. J Clin Oncol. 2005;23:6524-32.

25. Rumilla K, Schowalter KV, Lindor NM, Thomas BC, Mensink KA, Gallinger S, et al. Frequency of deletions of EPCAM (TACSTD1) in MSH2-associated Lynch syndrome cases. J Mol Diagn. 2011;13:93-9.

26. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, et al. The Clinical Phenotype of Lynch Syndrome Due to Germ-Line PMS2 Mutations. Gastroenterology. 2008;135:419-28.

27. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat. 2008;29:1282-91.

28. Arnold S, Buchanan DD, Barker M, Jaskowski L, Walsh MD, Birney G, et al. Classifying MLH1 and MSH2 variants using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. Hum Mutat. 2009;30:757-70.

29. InSiGHT variant databases. Available from: http://insight-group.org/variants/databases/

30. Thompson BA, Spurdle AB, Plazzer JP, Greenblatt MS, Akagi K, Al-Mulla F, et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet. 2013;46:107-15.

31. Cleary SP, Cotterchio M, Jenkins MA, Kim H, Bristow R, Green R, et al. Germline MutY Human Homologue Mutations and Colorectal Cancer: A Multisite Case-Control Study. Gastroenterology. 2009;136:1251-60.

32. Dowty JG, Win AK, Buchanan DD, Lindor NM, Macrae FA, Clendenning M, et al. Cancer risks for MLH1 and MSH2 mutation carriers. Hum Mutat. 2013;34:490-7.

33. Baglietto L, Lindor NM, Dowty JG, White DM, Wagner A, Gomez Garcia EB, et al. Risks of Lynch Syndrome Cancers for MSH6 Mutation Carriers. J Natl Cancer Inst. 2010;102:193-201.

34. Win AK, Dowty JG, Cleary SP, Kim H, Buchanan DD, Young JP, et al. Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. Gastroenterology. 2014;146:1208-11.e1-5.

35. Lange K. An approximate model of polygenic inheritance. Genetics. 1997;147:1423-30.

36. Fernando RL, Stricker C, Elston RC. The finite polygenic mixed model: An alternative formulation for the mixed model of inheritance. Genetics. 1994;88:573-80.

37. Curado MP, Edwards B, Shin HR, Storm H, Ferlay J, Heanue M, et al., editors. Cancer Incidence in Five Continents, Vol. IX. Lyon, France: International Agency for Research on Cancer; 2007.

38. Palomaki GE, McClain MR, Melillo S, Hampel HL, Thibodeau SN. EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome. Genet Med. 2009;11:42-65.

39. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19:716-23.

40. Elston R. Models for discrimination between alternative modes of inheritance. In: Gianola D, Hammond F, editors. Advances in statistical methods for genetic improvement of livestock. Berlin: Springer; 1990. p. 41-55.

41. Antoniou AC, Pharoah PPD, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer. 2004;91:1580-90.

42. Win AK, Hopper JL, Jenkins MA. Association between monoallelic MUTYH mutation and colorectal cancer risk: a meta-regression analysis. Fam Cancer. 2011;10:1-9.

43. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747-53.

44. Hopper JL, Mack TM. The heritability of prostate cancer-letter. Cancer Epidemiol Biomarkers Prev. 2015;24:878.

45. Hopper JL, Carlin JB. Familial Aggregation of a Disease Consequent upon Correlation between Relatives in a Risk Factor Measured on a Continuous Scale. Am J Epidemiol. 1992;136:1138-47.

46. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the royal society of Edinburgh. 1918;52:399-433.

47. MacInnis RJ, Severi G, Baglietto L, Dowty JG, Jenkins MA, Southey MC, et al. Population-based estimate of prostate cancer risk for carriers of the HOXB13 missense mutation G84E. PLoS One. 2013;8:e54727.

48. Samadder NJ, Smith KR, Mineau GP, Pimentel R, Wong J, Boucher K, et al. Familial colorectal cancer risk by subsite of primary cancer: a population-based study in Utah. Aliment Pharmacol Ther. 2015.

49. Jang JH, Cotterchio M, Gallinger S, Knight JA, Daftary D. Family history of hormonal cancers and colorectal cancer risk: a case-control study conducted in Ontario. Int J Cancer. 2009;125:918-25.

50. Nielsen M, van Steenbergen LN, Jones N, Vogt S, Vasen HFA, Morreau H, et al. Survival of MUTYH-Associated Polyposis Patients With Colorectal Cancer and Matched Control Colorectal Cancer Patients. J Natl Cancer Inst. 2010;102:1724-30.

51. Watson P, Lin KM, Rodriguez-Bigas MA, Smyrk T, Lemon S, Shashidharan M, et al. Colorectal carcinoma survival among hereditary nonpolyposis colorectal carcinoma family members. Cancer. 1998;83:259-66.

52. Sankila R, Aaltonen LA, Jarvinen HJ, Mecklin JP. Better survival rates in patients with MLH1-associated hereditary colorectal cancer. Gastroenterology. 1996;110:682-7.

53. Mai PL, Garceau AO, Graubard BI, Dunn M, McNeel TS, Gonsalves L, et al. Confirmation of family cancer history reported in a population-based survey. J Natl Cancer Inst. 2011;103:788-97.

54. Giardiello FM, Welsh SB, Hamilton SR, Offerhaus GJ, Gittelsohn AM, Booker SV, et al. Increased risk of cancer in the Peutz-Jeghers syndrome. N Engl J Med. 1987;316:1511-4.

55. Haidle JL, Howe JR. Juvenile Polyposis Syndrome. In: Pagon RA, Bird TD, Dolan CR, Stephens K, editors. Gene Reviews. Seattle, WA: University of Washington, Seattle; 1993-.

56. Mallory SB. Cowden syndrome (multiple hamartoma syndrome). Dermatol Clin. 1995;13:27-31.

57. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013;45:136-44.

58. Win AK, Young JP, Lindor NM, Tucker K, Ahnen D, Young GP, et al. Colorectal and other cancer risks for carriers and noncarriers from families with a DNA mismatch repair gene mutation: a prospective cohort study. J Clin Oncol. 2012;30:958-64.

59. Win AK, Hopper JL, Buchanan DD, Young JP, Tenesa A, Dowty JG, et al. Are the common genetic variants associated with colorectal cancer risk for DNA mismatch repair gene mutation carriers? Eur J Cancer. 2013;49:1578-87.

Table 1: Descriptive statistics of population-based families from the Colon Cancer Family Registry

	All			Australia			USA			Canada		
Relative of proband	Total No.	No. of CRC affected (%)	Mean age at CRC diagnosis (SD)	Total No.	No. of CRC affected (%)	Mean age at CRC diagnosis (SD)	Total No.	No. of CRC affected (%)	Mean age at CRC diagnosis (SD)	Total No.	No. of CRC affected (%)	Mean age at CRC diagnosis (SD)
Proband	5744	5744 (100)	52.5 (11.6)	911	911 (100)	45.8 (8.0)	3626	3626 (100)	54.7 (11.8)	1207	1207 (100)	50.7 (10.9)
Father	5737	305 (5)	61.6 (11.0)	911	68 (7)	61.3 (12.2)	3626	164 (5)	61.9 (10.8)	1200*	73 (6)	61.3 (10.5)
Mother	5737	234 (4)	61.5 (12.1)	911	48 (5)	61.7 (11.1)	3626	142 (4)	62.2 (12.4)	1200*	44 (4)	59.2 (12.0)
Sibling	15095	255 (2)	56.0 (13.3)	2228	26 (1)	47.2 (14.1)	9437	183 (2)	57.3 (12.4)	3430	46 (1)	55.6 (14.4)
Offspring	11065	12 (0.1)	40.3 (14.4)	1772	2 (0.1)	23.0 (8.5)	6884	8 (0.1)	46.9 (11.0)	2409	2 (0.1)	31.5 (16.3)

CRC, colorectal cancer; SD, standard deviation

*7 probands had no data for father and mother

Table 2: Descriptive statistics of population-based families from the Colon Cancer Family Registry by mismatch repair (MMR) gene and *MUTYH* mutation carrier status

	MMR gene mutation families (n=136)			*MUTYH* gene mutation families (n=81)			Non-carrier / unidentified carrier status families (n=5528)		
Relative of proband	Total No.	No. of CRC affected (%)	Mean age at CRC diagnosis (SD)	Total No.	No. of CRC affected (%)	Mean age at CRC diagnosis (SD)	Total No.	No. of CRC affected (%)	Mean age at CRC diagnosis (SD)
Proband	136	136 (100)	42.9 (10.5)	81	81 (100)	50.1 (12.3)	5528	5528 (100)	52.7 (11.5)
Father	136	26 (19)	49.0 (14.4)	81	8 (10)	67.8 (7.0)	5501*	271 (5)	62.7 (10.0)
Mother	136	16 (12)	51.4 (12.6)	81	0 (0)	-	5501*	218 (4)	62.3 (11.7)
Sibling	375	27 (8)	41.7 (11.5)	181	4 (2)	63.3 (9.9)	14494	224 (2)	57.6 (12.5)
Offspring	207	0 (0)	-	150	0 (0)	-	10665	12 (0.1)	40.3 (14.4)

CRC, colorectal cancer; SD, standard deviation

One proband had both a MMR gene and a monoallelic *MUTYH* germline mutation.

*7 probands had no data for father and mother

Table 3: Segregation analysis models including age-dependent polygenic variance, mismatch repair gene and *MUTYH* mutation carrier status

Model	No. Par	LL	AIC	P*	q_A (95% CI)	RR Het (95% CI)	RR Hom (95% CI)	$\sigma^2_{p\,(<40y)}$ (95% CI)	$\sigma^2_{p\,(40\text{-}49y)}$ (95% CI)	$\sigma^2_{p\,(50\text{-}59y)}$ (95% CI)	$\sigma^2_{p\,(60\text{-}69y)}$ (95% CI)	$\sigma^2_{p\,(>=70y)}$ (95% CI)	q(*MLH1*) (95% CI)	q(*MSH2*) (95% CI)	q(*MSH6*) (95% CI)	q(*PMS2*) (95% CI)	q(*MUTYH*) (95% CI)
Polygenic	10	-7218.1	14456.1	0.01	_	_	_	3.74 (1.47, 9.51)	2.02 (1.17, 3.48)	1.11 (0.64, 1.91)	1.19 (0.74, 1.90)	0.80 (0.42, 1.54)	0.000261 (0.000198, 0.000342)	0.000181 (0.000134, 0.000244)	0.000664 (0.000447, 0.000987)	0.000701 (0.000474, 0.001047)	0.01113 (0.00950, 0.01304)
Mixed Dominant	12	-7212.5	14449.0	1.0	0.000992 (0.00018, 0.00541)	31.1 (11.6, 83.4)	31.1 (11.6, 83.4)	3.28 (1.10, 9.74)	0.93 (0.26, 3.32)	0.46 (0.12, 1.81)	0.78 (0.27, 2.27)	0.52 (0.16, 1.64)	0.000257 (0.000195, 0.000338)	0.000176 (0.000130, 0.000238)	0.000660 (0.000444, 0.000982)	0.000701 (0.000471, 0.001042)	0.01113 (0.00950, 0.01304)
Mixed Recessive	12	-7216.1	14456.2	0.007	0.151 (0.057, 0.403)	1.0	10.8 (3.5, 33.4)	3.28 (1.24, 8.64)	1.50 (0.70, 3.21)	0.69 (0.27, 1.79)	0.82 (0.35, 1.94)	0.64 (0.25, 1.64)	0.000261 (0.000198, 0.000343)	0.000180 (0.000133, 0.000244)	0.000663 (0.000446, 0.000985)	0.000703 (0.000473, 0.001045)	0.01109 (0.00947, 0.01299)
Mixed Codominant	13	-7212.5	14451.0	_	0.000992 (0.00018, 0.00541)	31.1 (11.6, 83.4)	31.1 (11.6, 83.4)	3.28 (1.10, 9.74)	0.93 (0.26, 3.32)	0.46 (0.12, 1.81)	0.78 (0.27, 2.27)	0.52 (0.16, 1.64)	0.000257 (0.000195, 0.000338)	0.000176 (0.000130, 0.000238)	0.000660 (0.000444, 0.000982)	0.000701 (0.000471, 0.001042)	0.01113 (0.00950, 0.01304)

Par, number of parameters estimated in the model; LL, log-likelihood; AIC, Akaile's Information Criterion; q_A, estimated high-risk allele frequency for the unidentified major genes; q, minor allele frequency; CI, confidence interval; hom, homozygous; het, heterozygous, RR, relative risk as compared with non-carriers; σ^2_p, variance of the polygenic component; –, not applicable.

*For all models, P value refers to the comparison with the mixed codominant model using the log-likelihood ratio test.

Table 4: Estimated population carrier frequency for each mismatch repair (MMR) gene, *MUTYH* and the unidentified major susceptibility genes based on the mixed dominant model with age-dependent polygenic component

Gene	**% (95% CI)**	**1 in (95% CI)**
Unidentified major genes	0.198 (0.036 – 1.079)	504 (93 – 2778)
MLH1	0.051 (0.039 – 0.068)	1946 (1480 – 2564)
MSH2	0.035 (0.026 – 0.048)	2841 (2101 – 3846)
MLH1 or *MSH2*	0.087 (0.065 – 0.115)	1155 (868 – 1539)
MSH6	0.132 (0.089 – 0.196)	758 (509 – 1126)
PMS2	0.140 (0.094 – 0.208)	714 (480 – 1062)
Any MMR gene	0.359 (0.248 – 0.520)	279 (192 – 403)
MUTYH monoallelic	2.214 (1.891 – 2.591)	45 (39 – 53)
MUTYH biallelic	0.012 (0.009 – 0.017)	8073 (5881 – 11080)

CI, confidence interval; MMR, mismatch repair

Table 5 (A): Expected versus observed number of colorectal cancer affected relatives

	1 parent	1 sibling	2 siblings	1 parent 1 sibling	χ^2
Observed	478	175	14	28	
Expected					
- Polygenic	466.9	189.8	9.6	21.7	5.3
- Mixed dominant	462.4	179.6	9.4	24.2	3.5
- Mixed recessive	451.9	200.1	10.8	22.4	7.0
- Mixed codominant	462.4	179.6	9.4	24.2	3.5

χ^2 value for the difference between observed and expected number of affected relatives.

Note, the lower the χ^2, the better the fit of the model.5

Table 5 (B): Expected versus observed number of mutation carriers in families that had mutation testing performed

	MLH1	*MSH2*	*MSH6*	*PMS2*	*MUTYH* biallelic	*MUTYH* monoallelic	χ^2
Number of families	3319	3319	3319	3319	3796	3796	
Observed	49	39	24	24	18	63	
Expected							
- Polygenic	49.3	43.8	24.9	24.9	18.3	66.6	0.8
- Mixed dominant	48.7	42.5	24.7	24.6	18.2	66.6	0.5
- Mixed recessive	49.4	43.9	24.7	24.7	17.9	66.3	0.8
- Mixed codominant	48.7	42.5	24.7	24.6	18.2	66.6	0.5

χ^2 value for the difference between observed and expected number of mutation carriers.

Note, the lower the χ^2, the better the fit of the model.

Figure 1. Amount of polygenic variance explained by the hypothetical unidentified major genes component (dark grey) and the polygenic component (white) for each 10-year age group.

Statistical Methods

We used modified segregation analysis to fit a range of genetic models to the observed colorectal cancer family histories for the proband and their first-degree relatives. Information on second-degree relatives and other relatives of the proband were not used for this analysis as data were incomplete. Individuals were assumed to be at risk of colorectal cancer from birth until the earliest of the following: diagnosis of colorectal cancer or any other cancer (except skin cancer), first polypectomy, death, the earlier of last known age at baseline interview or age 80 years. Individuals known to have died but with unknown age at death were censored at age 70 years. Relatives reported to have had colorectal cancer with unknown age at diagnosis were assigned an age at diagnosis as their age of death (if deceased) or their age at the time of the proband's baseline interview. Relatives with no age information were treated as lost to follow-up with age censored at zero years.

Models of Susceptibility

The model we used was an extension of the mixed model of Morton and MacLean (1974) (1). incorporating both major gene and polygenic components. In our model, the colorectal cancer incidence $\lambda_i(t,k)$ for individual i at age t in sex group k (k = 1 for males, k = 2 for females) was assumed to depend on genotype according to a parametric survival analysis model $\lambda_i(t,k) = \lambda_0(t,k) \exp(G_i + P_i(t))$, where $\lambda_0(t,k)$ is the sex-specific baseline incidence at age t . G_i is the natural logarithm of the relative risk associated with the major genotype and $P_i(t)$ is the polygenic component for age t.

The major genotype was defined by six components representing each of the genes: *MLH1, MSH2, MSH6, PMS2, MUTYH* and one representing the hypothetical unidentified major genes, here given the name *UNIDENTIFIED_MAJOR_GENES*. As the probability of having

a mutation in more than one of *MLH1, MSH2, MSH6, PMS2, MUTYH* was very small, to reduce the computation time, we coded these genes into one locus with seven alleles: *MLH1* positive, *MSH2* positive, *MSH6* positive, *PMS2* positive, *MUTYH* positive, *UNIDENTIFIED_MAJOR_GENES* positive and a normal allele. For simplicity, the hierarchical order in which mutations were assumed to be dominant over other alleles was as follows: *MLH1*, *MSH2*, *MSH6*, *PMS2*, *MUTYH*, *UNIDENTIFIED_MAJOR_GENES* and normal alleles. These assumptions are not critical to the model because carriers in multiple genes are rare. Thus, G_i takes on the following potential risk categories: *MLH1* carriers, *MSH2* carriers, *MSH6* carriers, *PMS2* carriers, *MUTYH* biallelic carriers, *MUTYH* monoallelic carriers, *UNIDENTIFIED_MAJOR_GENES* heterozygotes, *UNIDENTIFIED_MAJOR_GENES* homozygotes and non-carriers. We fitted models in which the unidentified major genes component was autosomal with a normal and a mutant allele unlinked to mutations in the MMR genes or *MUTYH*. These assumptions are not critical to the model because carriers in multiple genes are rare. This is consistent with observation, since in this data set, there were no probands observed with mutations in more than one MMR gene. We have also previously observed that having *MUTYH* monoallelic mutation does not significantly alter the risk of colorectal cancer in addition to having a MMR mutation (2, 3). We also fitted models in which the average relative risk for the unidentified major genes component was assumed to be age dependent. As the number of MMR and *MUTYH* mutation carriers in our dataset was too small to obtain reliable colorectal cancer risk estimates for mutation carriers, we used the published age-, sex- and country-specific incidence rates for *MLH1* and *MSH2* mutation carriers (4), and published age- and sex-specific incidence rates for *MSH6, PMS2* and *MUTYH* mutation carriers (2, 5, 6).

The polygenic component for age t, $P_i(t)$, was assumed to be normally distributed with zero mean and variance $\sigma^2_p(t)$. Because the standard polygenic model is not amenable to

"peeling" (7-9), we used as an approximation the hypergeometric polygenic model (HPM) (10, 11). P was approximated by $P=(R-N)\sigma_p/(N/2)^{1/2}$, where R has a binomial distribution Bin($2N$, 1/2). N, the number of loci used in the HPM, was set to 3 (see Appendices in Antoniou et al. (2001) for further details of the model(12)). We also fitted models where the variance of the polygenic 'modifying' component was allowed to take a different value σ^2_m for MMR and for *MUTYH* carriers.

To compute the baseline colorectal cancer incidence $\lambda_0(t)$, we constrained the overall incidence of colorectal cancer to agree with the national age and sex specific incidences (1998-2002) separately for Australia, Canada and USA (13). Other cancers were ignored in this model. Published incidences are reported in 5-year intervals, which can result in large variation in incidences between adjacent age intervals, so we smoothed the population incidences using locally weighted regression (LOWESS)(14) with a bandwidth of 0.2.

We assumed that the sensitivity of the mutation testing of probands for MMR genes and *MUTYH* was 80% (15), and we examined the effect of varying this sensitivity. For relatives, we assumed the mutation screening for the proband's mutation (i.e. predictive testing) was 100% sensitive and specific.

Estimation

The genetic models were specified in terms of colorectal cancer incidence for MMR gene and *MUTYH* mutation carriers, the frequency (q_A) of the putative high risk allele ''A'' of the unidentified major genes, the average relative risk of colorectal cancer for carriers of mutations in the unidentified major genes, the number of polygenic loci assumed in the HPM (N), and the variances of the polygenic and modifying components (σ^2_p and σ^2_m). Maximum likelihood estimation was used to estimate the parameters. The estimates we

present are the values that were the most likely (i.e. most consistent) with the data. Maximum likelihood is the optimal method for making such estimates, and provides confidence intervals (CIs). We adjusted for ascertainment by maximizing the likelihood of each pedigree conditioned on the colorectal cancer status of the proband and his or her age of diagnosis (but not mutation carrier status as this information was not known at the time of recruitment). The variances of the parameter estimates were obtained by inverting the observed information matrix. To allow for the restricted ranges of the parameter values and to provide estimates likely to be more nearly normally distributed, we used transformed values of the parameters as a basis for calculating CIs: the high-risk allele frequencies q_A were transformed using the logit function $\log[p/(1-p)]$, while for relative risk and σ^2_p we used the log transformation.

Goodness of Fit

The relative goodness of fit for nested models was tested by the likelihood ratio test. The Akaike's Information Criterion(16) [AIC=-2 log-likelihood + 2x(no. of parameters)] was used to assess goodness of fit between non-nested models (1).

Estimation of Familial Relative Risk

The familial relative risk of the polygenic component was estimated using $\exp(\sigma^2/2)$ (17).

Expected versus Observed Number of Affected Relatives

The goodness of fit of the models was assessed by comparing the observed number of affected relatives with that expected under each model. For each model, we computed the predicted number of probands with one parent with colorectal cancer, one sibling with colorectal cancer, two siblings with colorectal cancer and one parent and one sibling with colorectal cancer. The probability of observing each of these events for a family was

computed from the ratio of likelihoods for two pedigrees: the likelihood for a family in which the predicted event occurred and the likelihood of the affected proband (12). The expected versus observed number of affected relatives under each fitted model was assessed using the Pearson χ^2 goodness of fit statistic $\Sigma_j[(O_j-E_j)^2/E_j]$ where O_j is the observed number of probands or families with a particular characteristic and E_j is the corresponding predicted numbers under each model.

Expected versus Observed number of Mutation Carriers

The expected number of probands with MMR and *MUTYH* mutation carriers for families that had undergone mutation testing based on their cancer family history was computed using Bayes theorem (18). For example, P(*MLH1* carrier | family history) = P(*MLH1* carrier, family history) / ΣP(mutation carrier status, family history) = $L_1/(L_0+L_1+L_2+L_3+L_4+L_5)$, where L_i is the likelihood of observing the family with the proband carrying mutation i (=0, 1, 2, 3, 4, 5 for mutation negative, *MLH1*, *MSH2*, *MSH6*, *PMS2* and *MUTYH*, respectively). These probabilities were then added over all families to compute the total number of probands expected to carry a mutation for each gene (12).

Statistical Computation

Segregation analyses were performed using an optimised version(19) of the pedigree analysis software MENDEL version 3.2 (20). Incidence smoothing and other descriptive statistics were calculated using STATA 13.0 (21).

References

1. Elston R. Models for discrimination between alternative modes of inheritance. In: Gianola D, Hammond F, editors. Advances in statistical methods for genetic improvement of livestock. Berlin: Springer; 1990. p. 41-55.

2. Win AK, Dowty JG, Cleary SP, Kim H, Buchanan DD, Young JP, et al. Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. Gastroenterology. 2014;146:1208-11.e1-5.

3. Win AK, Cleary SP, Dowty JG, Baron JA, Young JP, Buchanan DD, et al. Cancer risks for monoallelic MUTYH mutation carriers with a family history of colorectal cancer. Int J Cancer. 2011;129:2256-62.

4. Dowty JG, Win AK, Buchanan DD, Lindor NM, Macrae FA, Clendenning M, et al. Cancer risks for MLH1 and MSH2 mutation carriers. Hum Mutat. 2013;34:490-7.

5. Baglietto L, Lindor NM, Dowty JG, White DM, Wagner A, Gomez Garcia EB, et al. Risks of Lynch Syndrome Cancers for MSH6 Mutation Carriers. J Natl Cancer Inst. 2010;102:193-201.

6. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, et al. The Clinical Phenotype of Lynch Syndrome Due to Germ-Line PMS2 Mutations. Gastroenterology. 2008;135:419-28.

7. Cannings C, Thompson E, Skolnick M. Probability functions on complex pedigrees. Adv Appl Prob. 1978;10:26-61.

8. Hasstedt S. A mixed-model likelihood approximation on large pedigrees. Comput Biomed Res. 1982;15:295-307.

9. Lalouel J. Probability calculations in pedigrees under complex models of inheritance. Hum Hered. 1980;30:320-3.

10. Lange K. An approximate model of polygenic inheritance. Genetics. 1997;147:1423-30.

11. Fernando RL, Stricker C, Elston RC. The finite polygenic mixed model: An alternative formulation for the mixed model of inheritance. Genetics. 1994;88:573-80.

12. Antoniou AC, Pharoah PDP, McMullan G, Day NE, Ponder BAJ, Easton D. Evidence for further breast cancer susceptibility genes in addition to *BRCA1* and *BRCA2* in a population-based study. Genet Epidemiol. 2001;21:1-18.

13. Curado MP, Edwards B, Shin HR, Storm H, Ferlay J, Heanue M, et al., editors. Cancer Incidence in Five Continents, Vol. IX. Lyon, France: International Agency for Research on Cancer; 2007.

14. Royston P. The use of cusums and other techniques in modelling continuous covariates in logistic regression. Stat Med. 1992;11:1115-29.

15. Palomaki GE, McClain MR, Melillo S, Hampel HL, Thibodeau SN. EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome. Genet Med. 2009;11:42-65.

16. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19:716-23.

17. Pharoah PDP, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BAJ. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. 2002;31:33-6.

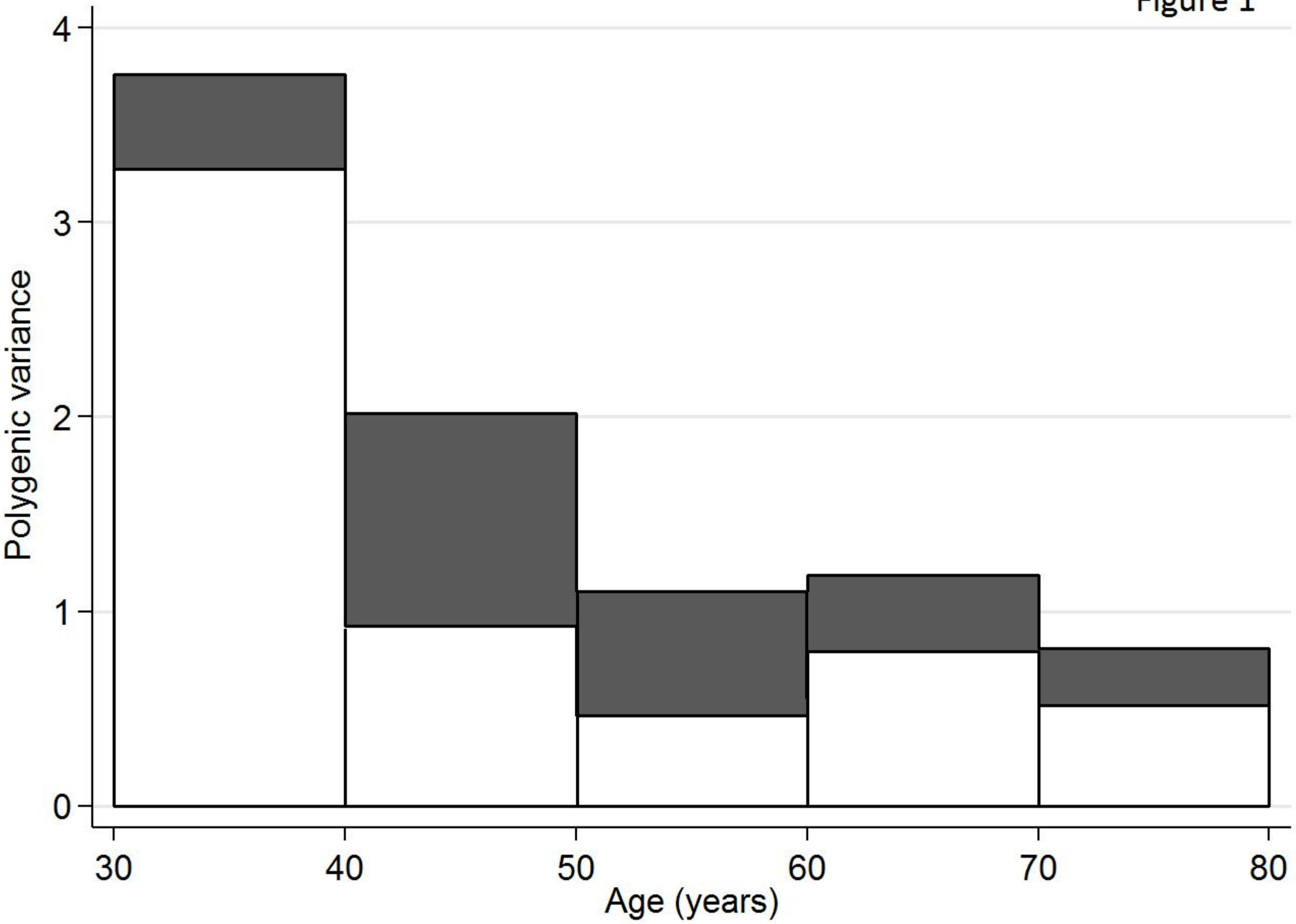
18. Antoniou AC, Pharoah PPD, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer. 2004;91:1580-90.

19. Lee AJ, Cunningham AP, Kuchenbaecker KB, Mavaddat N, Easton DF, Antoniou AC. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Br J Cancer. 2014;110:535-45.

20. Lange K, Weeks D, Boehnke M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol. 1988;5:471-2.

21. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP; 2013.

Figure 1



Supplementary Table 1: Results of segregation analyses incorporating mismatch repair gene and *MUTYH* mutation carrier status

	No. Par	LL	AIC	*P**	q_A (95% CI)	RR Het (95% CI)	RR Hom (95% CI)	σ^2_p (95% CI)	q(*MLH1*) (95% CI)	q(*MSH2*) (95% CI)	q(*MSH6*) (95% CI)	q(*PMS2*) (95% CI)	q(*MUTYH*) (95% CI)
Base	5	-7296.0	14602.0	$3.8x10^{-33}$	_	_	_	_	0.000313 (0.000240, 0.000407)	0.000234 (0.000177, 0.000309)	0.000701 (0.000472, 0.001042)	0.000753 (0.000508, 0.001118)	0.01135 (0.00969, 0.01330)
Dominant	7	-7229.0	14472.0	$6.0x10^{-6}$	0.00202 (0.00087,0.00468)	28.8 (17.6,47.1)	28.8 (17.6,47.1)	_	0.000275 (0.000209, 0.000361)	0.000191 (0.000141, 0.000258)	0.000668 (0.000450, 0.000993)	0.000712 (0.000479, 0.001059)	0.01118 (0.00954, 0.01310)
Recessive	7	-7238.1	14490.2	$6.5x10^{-10}$	0.1804 (0.11,0.296)	1.0	19.0 (11.0,32.7)	_	0.000284 (0.000216, 0.000372)	0.000203 (0.000151, 0.000272)	0.000667 (0.000449, 0.000992)	0.000711 (0.000479, 0.001057)	0. 01113 (0.00950, 0.01304)
Codominant	8	-7227.2	14470.3	$6.2x10^{-6}$	0.007024 (0.0032,0.0152)	14.0 (9.0,22.0)	830.5 (261.5,2637.1)	_	0.000272 (0.000206, 0.000357)	0.000189 (0.000140, 0.000256)	0.000667 (0.000449, 0.000991)	0.000711 (0.000479, 0.001057)	0. 01114 (0.00951, 0.01305)
Polygenic	6	-7223.6	14459.2	0.004	_	_	_	1.32 (1.08,1.62)	0.000272 (0.000208, 0.000357)	0.000191 (0.000142, 0.000257)	0.000667 (0.000449, 0.000992)	0.000705 (0.000474, 0.001047)	0. 01119 (0.00955, 0.01311)
Mixed Dominant	8	-7217.0	14449.9	0.94	0.00063 (0.00010,0.00398)	40.5 (13.2,124.1)	40.5 (13.2,124.1)	0.87 (0.53,1.41)	0.000263 (0.000199, 0.000346)	0.000181 (0.000133, 0.000245)	0.000662 (0.000445, 0.000984)	0.000701 (0.000471, 0.001041)	0. 01116 (0.00953, 0.01307)
Mixed Recessive	8	-7221.2	14458.3	0.004	0.116 (0.046,0.290)	1.0	14.7 (4.8,45.0)	1.04 (0.71,1.52)	0.000270 (0.000206, 0.000354)	0.000189 (0.000140, 0.000254)	0.000664 (0.000447, 0.000987)	0.000702 (0.000473, 0.001044)	0. 01112 (0.00949, 0.01303)
Mixed Codominant	9	-7216.9	14451.9	_	0.00062 (0.00009,0.00412)	40.8 (12.8,129.6)	19.6 (0,∞)	0.87 (0.53,1.41)	0.000262 (0.000199, 0.000345)	0.000179 (0.000132, 0.000244)	0.000662 (0.000445, 0.000984)	0.000701 (0.000472, 0.001042)	0. 01115 (0.00952, 0.01306)

Par, number of parameters estimated in the model; LL, log-likelihood; AIC, Akaile's Information Criterion; q_A, estimated high-risk allele frequency for the unidentified major genes; q, minor allele frequency; CI, confidence interval; hom, homozygous; het, heterozygous, RR, relative risk as compared with non-carriers; σ^2_p, variance of the polygenic component; –, not applicable.

*For all models, P value refers to the comparison with the mixed codominant model using the log-likelihood ratio test.

Supplementary Table 2. Estimated population carrier frequency of a mismatch repair gene or *MUTYH* mutation from previous studies and current study

Author	**Population**	**Gene**	**Estimate of population carrier frequency (95% CI)**	**Calculation of carrier frequency based on these assumptions.**
Salovaara et al. (2000) (1)	Finland	*MLH1*, *MSH2*	1 in 740	2.7% carrier frequency in CRC × 5% lifetime risk of CRC = 0.135%
Dunlop (2000) (2)	Scotland (15-74 years)	*MLH1*, *MSH2*	1 in 3139 (1247 - 7626)	2.66% carrier frequency in CRC × 0.17% population prevalence of CRC ÷ 14.6% prevalence of CRC in carriers = 0.031%
Terdiman (2001) (3)	USA	*MLH1*, *MSH2*	1 in 800 - 1 in 1600	1-2% carrier frequency in CRC × 5% lifestyle risk of CRC ÷ 80% lifetime risk for carriers = 0.0625% to 0.125%
de la Chapelle (2005) (4)	Literature review	*MLH1*, *MSH2*	1 in 660 - 1 in 2000	1-3% carrier frequency in CRC × 5% lifetime risk of CRC = 0.05% to 0.15%
Boland and Shike (2010) (5)	USA	*MLH1*, *MSH2*, *MSH6*, *PMS2*	1 in 300	2.8% carrier frequency in CRC × 6% lifetime risk of CRC ÷ 50% lifetime risk for carriers = 0.33%
Hampel and de la Chapelle (2011) (6)	USA	*MLH1*, *MSH2*, *MSH6*, *PMS2*	1 in 370	2.8% carrier frequency in CRC × 5% lifetime risk of CRC ÷ 50% lifetime risk for carriers = 0.28%
Win et al. (2011) (7)	Literature review	*MUTYH*	mono *MUTYH* 1 in 60 bi *MUTYH* 1 in 7320	243 monoallelic carriers ÷ 14639 controls 2 biallelic carriers ÷ 14639 controls

CRC, colorectal cancer; CI, confidence interval

References

1. Salovaara R, Loukola A, Kristo P, Kaariainen H, Ahtola H, Eskelinen M, et al. Population-Based Molecular Detection of Hereditary Nonpolyposis Colorectal Cancer. J Clin Oncol. 2000;18:2193-200.
2. Dunlop MG, Farrington SM, Nicholl I, Aaltonen L, Petersen G, Porteous M, et al. Population carrier frequency of hMSH2 and hMLH1 mutations. Br J Cancer. 2000;83:1643-5.
3. Terdiman JP. HNPCC: an uncommon but important diagnosis. Gastroenterology. 2001;121:1005-8.
4. de la Chapelle A. The incidence of Lynch syndrome. Fam Cancer. 2005;4:233-7.
5. Boland CR, Shike M. Report from the Jerusalem workshop on Lynch syndrome-hereditary nonpolyposis colorectal cancer. Gastroenterology. 2010;139:2197 e1-7.
6. Hampel H, de la Chapelle A. The Search for Unaffected Individuals with Lynch Syndrome: Do the Ends Justify the Means? Cancer Prev Res. 2011;4:1-5.
7. Win AK, Hopper JL, Jenkins MA. Association between monoallelic MUTYH mutation and colorectal cancer risk: a meta-regression analysis. Fam Cancer. 2011;10:1-9.