
Publishers page: http://dx.doi.org/10.1002/ijc.31076

Please note: Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
Genome-wide association study and meta-analysis in Northern European populations replicate multiple colorectal cancer risk loci

Short title: GWAS and meta-analysis of CRC

Manuscript type: Short Report

Authors: Tomas Tanskanen1,2, Linda van den Berg1,2, Niko Välimäki1,2, Mervi Aavikko1,2, Eivind Ness-Jensen3,4,5,6, Kristian Hveem3,4, Yvonne Wettergren7, Elinor Bexe Lindskog7, Neeme Tõnisson6, Andres Metspalu6, Kaisa Silander6, Giulia Orlando10, Philip J. Law10, Sari Tuupanen1,2, Alexandra E. Gylfe1,2, Ulrika A. Hänninen1,2, Tatiana Cajuso1,2, Johanna Kondelin1,2, Antti P. Sarin11, Eero Pukkala12,13, Pekka Jousilahti14, Veikko Salomaa14, Samuli Ripatti11, Aarno Palotie11,15,16,17, Heikki Järvinen16, Laura Renkonen-Sinisalo18, Anna Lepistö18, Jan Böhm19, Jukka-Pekka Mecklin20, Nada A. Al-Tassan21, Claire Palles22, Lynn Martin23, Ella Barclay22, Albert Tenesa24,25, Susan Farrington24, Maria N. Timofeeva24, Brian F. Meyer21, Salma M. Waki21, Harry Campbell21, Christopher G. Smith27, Shelley Idziaszczyk27, Timothy S. Maughan26, Richard Kaplan29, Rachel Kerr20, David Kerr31, Daniel D. Buchanan32,33, Aung K. Win33, John Hopper33, Mark Jenkins33, Polly A. Newcomb34, Steve Gallinger35, David Conti36, Fred Schumacher36, Graham Casey37, Jeremy P. Cheadle27, Malcolm G. Dunlop24, Ian P. Tomlinson23, Richard S. Houlston10, Kimmo Palin1,2, and Lauri A. Aaltosen1,2.

1Department of Medical and Medical Genetics, Medicum, University of Helsinki, Helsinki, Finland.
2Genome-Scale Biology Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland.
3HUNT Research Centre, Department of Public Health, NTNU, Norwegian University of Science and Technology, Levanger, Norway.
4K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
5Department of Molecular Medicine and Surgery, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden.
6Department of Medicine, Levanger Hospital, Nord-Trengdelag Hospital Trust, Levanger, Norway.
7Department of Surgery, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Sweden.
8Estonian Genome Center, University of Tartu, Tartu, Estonia.
9National Institute for Health and Welfare, Helsinki, Finland.
10Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK.
11Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.
12Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland.
13Faculty of Social Sciences, University of Tampere, Tampere, Finland.
14National Institute for Health and Welfare, Helsinki, Finland.
15Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.
16Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA.
17Department of Neurology, Massachusetts General Hospital, Boston, MA, USA.
18Department of Surgery, Abdominal Center, Helsinki University Hospital, Helsinki, Finland.
19Department of Pathology, Central Finland Central Hospital, Jyväskylä, Finland.
20Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Jyväskylä, Finland.
21Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia.
22Wellcome Trust Centre for Human Genetics and NIHR Comprehensive Biomedical Research Centre, Oxford, UK.
23Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK.
24Colon Cancer Genetics Group, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital, Edinburgh, UK.
25The Roslin Institute, University of Edinburgh, Easter Bush, Roslin, UK.
26Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK.
27Division of Cancer and Genetics, School of Medicine, Cardiff University, Cardiff, UK.
*To whom correspondence should be addressed. Tel: +358-2941-25595; Fax: +358 2941 25610; E-mail: lauri.aaltonen@helsinki.fi.

Keywords: colorectal cancer, genetic predisposition to disease, genome-wide association study, single-nucleotide polymorphism

**Novelty & impact statements**

This study provides strong evidence for the association between rs992157 (2q35) and colorectal cancer by independent replication in 4,439 cases and 15,847 controls, as well as meta-analysis of 39,786 European-ancestry individuals. Previously published SNPs at 2q35, 6p21.2, 8q23.3, 8q24.21, 10q22.3, 10q24.2, 11q13.4, 11q23.1, 14q22.2, 15q13.3, 18q21.1, 20p12.3, and 20q13.33 were associated with colorectal cancer in the Finnish population, but new risk loci were not identified.

**Abbreviations**

- **CI**: Confidence interval
- **CRC**: Colorectal cancer
- **GWAS**: Genome-wide association study
- **LD**: Linkage disequilibrium
- **MAF**: Minor allele frequency
- **OR**: Odds ratio
- **PCA**: Principal component analysis
- **Q-Q-plot**: Quantile-quantile plot
- **SNP**: Single-nucleotide polymorphism
Abstract

Genome-wide association studies have been successful in elucidating the genetic basis of colorectal cancer, but there remains unexplained variability in genetic risk. To identify new risk variants and to confirm reported associations, we conducted a genome-wide association study in 1,701 colorectal cancer cases and 14,082 cancer-free controls from the Finnish population. A total of 9,068,015 genetic variants were imputed and tested, and 30 promising variants were studied in additional 11,647 cases and 12,356 controls of European ancestry. The previously reported association between the single-nucleotide polymorphism rs992157 (2q35) and colorectal cancer was independently replicated (p=2.08x10^{-4}; OR, 1.14; 95% CI, 1.06-1.23), and it was genome-wide significant in combined analysis (p=1.50x10^{-9}; OR, 1.12; 95% CI, 1.08-1.16). Variants at 2q35, 6p21.2, 8q23.3, 8q24.21, 10q22.3, 10q24.2, 11q13.4, 11q23.1, 14q22.2, 15q13.3, 18q21.1, 20p12.3, and 20q13.33 were associated with colorectal cancer in the Finnish population (false discovery rate <0.1), but new risk loci were not found. These results replicate the effects of multiple loci on the risk of colorectal cancer and identify shared risk alleles between the Finnish population isolate and outbred populations.

Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and accounts for approximately 10% of global cancer incidence and mortality (http://globocan.iarc.fr/). Numerous genetic loci have been associated with CRC in genome-wide association studies (GWASs; https://www.ebi.ac.uk/gwas/), but much of its heritability remains unexplained, which limits personalized risk assessment and biological understanding of the disease.\textsuperscript{1,2} Discovery of new loci and replication of previously reported associations is thus important, and recent studies have continued to reveal novel CRC risk variants.\textsuperscript{3-7} The genetic
architecture of CRC varies between populations, and studies in isolated founder populations can offer valuable insights into disease susceptibility.\(^6\)

We conducted a GWAS of CRC in the Finnish population (the FIN cohort) using a large publicly available reference panel to impute genotypes and thus increase the odds of identifying disease-associated alleles across a wide range of allele frequencies.\(^9\) Thirty promising variants were investigated further in eleven European-ancestry studies (STHLM2, Gothenburg, HUNT, Estonia, FINRISK, COIN, UK1, Scotland1, VQ58, CCFR1, and CCFR2), adding to a total of 13,348 CRC cases and 26,438 controls.

In a recent meta-analysis of GWASs, the single-nucleotide polymorphism (SNP) rs992157 at 2q35, intronic to PANKD and TMBIM1, was found to be associated with CRC (p=3.15x10\(^{-6}\); odds ratio (OR), 1.10; 95% confidence interval (CI), 1.06-1.13).\(^6\) To replicate this finding, we genotyped and analyzed rs992157 in 4,439 CRC cases and 15,847 controls from five Northern European cohorts (STHLM2, Gothenburg, HUNT, Estonia, and a subset of the FIN cohort) that had not been previously studied for the association between rs992157 and CRC.

**Materials and methods**

This study was conducted in accordance with the Declaration of Helsinki and approved by the Finnish National Supervisory Authority for Welfare and Health, National Institute for Health and Welfare (THL/151/5.05.00/2017), and the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS/408/13/03/03/09). We derived 1,627 cases with colorectal adenocarcinoma from the ongoing Finnish CRC collection and genotyped normal tissues (colorectal tissue or blood) with Illumina (San Diego, CA) HumanOmni2.5-8 SNP arrays.\(^{10,11}\) Illumina HumanCoreExome SNP array data for additional 91 CRC patients and 14,187 Finnish cancer-free controls were obtained from the National FINRISK Study (https://www.thl.fi/fi/web/thlfi-en/research-and-expertwork/population-studies/the-national-
finrisk-study). Data on diagnosed cancers in the FINRISK study participants were collected from the Finnish Cancer Registry. PLINK v.1.90b3i (www.cog-genomics.org/plink/1.9/) was used for quality control. A total of 122 samples (17 genotyped with the HumanOmi2.5-8 array and 105 genotyped with HumanCoreExome array) were excluded on the basis of close relatedness (identity-by-descent coefficient >0.2), duplication, discordant sex information, or low genotyping rate. The FIN cohort consisted of the remaining 1,701 CRC cases and 14,082 cancer-free controls. By design, the HumanOmi2.5-8 SNP array contained 2,315,673 autosomal sites, 273,074 of which overlapped with the HumanCoreExome SNP array (https://support.illumina.com/downloads.html). Exclusion criteria for SNPs were genotyping rate <95%, excess homozygosity (frequency of rare homozygotes exceeding the frequency of heterozygotes, or any rare homozygote with minor allele frequency (MAF) <2%), deviation from the Hardy-Weinberg equilibrium (p<1x10^{-8}), differential missingness between genotyping batches (p<1x10^{-8}), differential patterns of linkage disequilibrium (LD) in cases versus controls, and LD-based strand inconsistency. After quality control, 214,705 SNPs were pre-phased with SHAPEIT v2 (r790), and genotypes were imputed with a publicly available reference panel (https://imputation.sanger.ac.uk/; http://www.haplotype-reference-consortium.org/). Variants with low allele frequency (<0.4%) or low IMPUTE2 info score (<0.4) were excluded prior to association analysis. In stage 1, disease associations were tested with a linear mixed model (BOLT-LMM-inf; https://data.broadinstitute.org/alkesgroup/BOLT-LMM/), adjusting for log-transformed age and sex. A linear mixed model was used because it can control for population structure and cryptic relatedness in genetic association studies of both quantitative and binary traits. The age covariate was defined as age at CRC diagnosis in cases and age at right censoring (end of follow-up or death) in controls. An additive genetic model was assumed. The genomic inflation factor was estimated by dividing the observed median of the BOLT-LMM-inf test statistic by the median of the chi-squared distribution with one degree of freedom.
In stage 2, the MassARRAY System by Agena Bioscience (San Diego, CA) was utilized at the Institute for Molecular Medicine Finland (FIMM) to genotype single-nucleotide variants in Nordic cohorts (STHLM2, 544 cases/541 controls; Gothenburg, 1,903 cases/258 controls; HUNT, 1,168 cases/1,147 controls; Estonia, 257 cases/259 controls; and FINRISK, 198 cases/172 controls), as well as 1,038 individuals from the FIN cohort who had also been genotyped with SNP arrays (925 with the HumanOmnI2.5-8 array and 113 with the HumanCoreExome array). The STHLM2 cohort consisted of men who had been referred to prostate-specific antigen screening in Stockholm County, Sweden between 2010 and 2012; DNA samples were provided by the Karolinska Institute Biobank (http://ki.se/forskning/ki-biobank). The Gothenburg cohort was formed from CRC patients who had been operated at the Sahlgrenska University Hospital, Gothenburg, Sweden; DNA samples from cases and controls were provided by the Sahlgrenska Biobank (https://www.gothiaforum.com/sab). DNA samples from the HUNT cohort were provided by the Norwegian Nord-Trøndelag Health Study (HUNT) and Biobank (https://www.ntnu.edu/hunt). The Estonia cohort was derived from the sample collections of the Estonian Genome Center (www.geenivaramu.ee/en). The FINRISK cohort consisted of participants of the National FINRISK Study (198 CRC cases and 172 cancer-free controls) who had not been included in the FIN cohort due to unavailable SNP array data; DNA samples were provided by the THL Biobank, Finland (https://www.thl.fi/fi/web/thlfi-en/topics/information-packages/thl-biobank).

When possible, cancer-free controls were matched to CRC cases on year of birth and sex.

To assess imputation accuracy, squared Pearson correlation coefficients ($r^2$) between IMPUTE2 genotype dosage and MassARRAY genotype were calculated.

To enable standard meta-analysis, data from the FIN cohort were reanalyzed by unconditional logistic regression under an additive genetic model, adjusting for sex, log-transformed age, and ten principal components (SNPTEST v.2.5.2). In the MassARRAY-genotyped Nordic cohorts, unconditional logistic regression was applied using R v.3.3.3, provided that at least ten minor alleles were observed. Details of the previously published
GWASs (COIN, UK1, Scotland1, VQ58, CCFR1, and CCFR2) can be found in Reference 15. Genomic control was applied by multiplying the standard errors of regression coefficients by the square root of the inflation factor of the respective study. PLINK v.1.90b3i was used for LD-based SNP pruning and principal component analysis (PCA). PCA was performed using 13,012 LD-pruned SNPs with allele frequency >5% and IMPUTE2 info score >0.9. R v.3.3.3 was used for meta-analysis. Estimated log ORs and standard errors were combined to obtain summary p-values, ORs, and 95% CIs under inverse-variance weighted random-effects and fixed-effect models (function “rma.uni” in the metafor package v.1.9-9). All reported p-values are two-sided. The type I error rate (α) was 0.05, corresponding to a genome-wide significance threshold of 5x10^{-8}. The Benjamini-Hochberg method was used to adjust for false discovery rate.

Results

In stage 1, we used a linear mixed model (BOLT-LMM-inf)\textsuperscript{13} to test 9,068,015 single-nucleotide variants for association with CRC in the FIN cohort, which comprised 1,701 Finnish CRC cases and 14,082 population-matched, cancer-free controls. The median of the BOLT-LMM-inf test statistic was 0.512, corresponding to an inflation factor of 1.12, which was used for genomic control. A quantile-quantile (Q-Q) plot is shown in Supplementary Figure 1, PCA plots in Supplementary Figures 2 and 3, and a Manhattan plot in Supplementary Figure 4. A low-frequency variant at 12q14.3 (rs73121704; MAF, 0.860\%) displayed the smallest p-value in stage 1 (p=4.07x10^{-9}). Among the highest-ranking SNPs were the CRC-associated variants rs10505477 (p=5.29x10^{-8}), rs6589219 (p=4.34x10^{-7}; r\textsuperscript{2} with rs3802842, 0.942 in 1,000 Genomes Phase 3 European populations), and rs6983267 (p=1.38x10^{-6})\textsuperscript{16-18} Thirty-eight previously published CRC risk SNPs were tested for association with CRC in the FIN cohort, and 14 of the 38 SNPs showed associations with false discovery rate <0.1. Directions of effects were consistent with earlier publications for each of the 14 SNPs, which were located at 11q23.1 (rs3802842, q=1.77x10^{-5}), 8q24.21
(rs6983267, q= 1.77x10^{-5}; rs7014346, q=1.77x10^{-5}), 20p12.3 (rs961253, q=6.92x10^{-5}),
15q13.3 (rs4779584, q=1.29x10^{-3}), 10q22.3 (rs704017, q=1.91x10^{-3}), 18q21.1 (rs4939827, 
q=7.96x10^{-3}), 2q35 (rs992157, q=7.96x10^{-3}), 8q23.3 (rs16892766, q=0.0113), 14q22.2 
(rs4444235, q=0.0231), 6p21.2 (rs1321311, q=0.0231), 20q13.33 (rs4925386, q=0.0501), 
10q24.2 (rs1035209, q=0.0536), and 11q13.4 (rs3824999, q=0.0604). Stage 1 results and 
LocusZoom (http://locuszoom.org/) plots are shown in Supplementary Tables 1 and 2 and in 
Supplementary Figures 35 to 102, respectively.

From 20 loci that were ranked highest in stage 1, we selected 40 variants for MassARRAY 
genotyping in five Nordic cohorts (STHL2, Gothenburg, HUNT, Estonia, and FINRISK; 
stage 2). Two variants were selected from each locus. rs992157 (2q35) was also selected 
for stage 2 because it had been recently reported as a CRC risk factor. We were unable to 
design genotyping assays for seven variants because of sequence context, and four variants 
failed genotyping. Consequently, 30 variants representing 20 loci were successfully 
genotyped in a total of 4,070 Nordic CRC cases and 2,377 controls. The MAF of 
6:73457627G>C was low in all five Nordic cohorts, ranging from 0.000923 to 0.00954 (allele 
count, 2-7). To evaluate imputation accuracy, 1,038 individuals from the FIN cohort were 
directly genotyped with the MassARRAY platform. Squared Pearson correlation coefficients 
(r^2) between IMPUTE2 genotype dosage and MassARRAY genotype for the 30 variants 
ranged from 0.816 to 1.00 (median, 0.978).

In stage 3, we obtained summary statistics from previously published GWASs that 
comprised 7,577 CRC cases and 9,979 controls of European ancestry. Summary-level 
data were available for 27 of the 30 variants that were genotyped in stage 2 (data for 
rs150509351, rs186867472, and 6:73457627G>C were missing).

To increase statistical power, datasets from stages 1 to 3 were combined (Figure 1), totaling 
13,348 CRC cases and 26,438 controls. The FIN cohort was reanalyzed by logistic
regression to obtain log ORs and corresponding standard errors; the inflation factor was 1.11. The post-imputation inflation factors for the COIN, UK1, Scotland1, VQ58, CCFR1 and CCFR2 studies were 1.10, 1.03, 1.04, 1.04, 1.03, and 1.08, respectively. Genomic control was applied for each of these studies. Inflation factors for the STHLM2, Gothenburg, HUNT, or Estonia studies were not estimated because of the small number of genotyped markers. Fixed-effect meta-analysis was performed, but because of possible study heterogeneity, we considered the random-effects model (Supplementary Table 3). Under the random-effects model, rs10505477 (8q24.21), rs6983267 (8q24.21), and rs992157 (2q35) were genome-wide significant (for rs10505477, p=7.63x10^{-14}, \phi_{het}=0.144, \hat{I}^2=34.4\%; for rs6983267, p=7.45x10^{-13}, \phi_{het}=0.0985, \hat{I}^2=37.7\%; for rs992157, p=1.50x10^{-9}, \phi_{het}=0.777, \hat{I}^2=0\%), and rs6589219 (11q23.1) displayed suggestive evidence of association (p=9.14x10^{-6}, \phi_{het}=0.153, \hat{I}^2=36.5\%). Combined effect size estimates and directions of effects for these four SNPs were consistent with prior studies.\(^{6,16-18}\)

Next, we studied rs992157 (2q35) in a replication dataset comprising 4,439 CRC cases and 15,847 controls (STHLM2, Gothenburg, HUNT, Estonia, and a subset of the FIN cohort) who had not been previously studied for the association between rs992157 and CRC (Figure 2). In the FIN cohort, rs992157 had been directly genotyped with SNP arrays in both cases and controls, and the other Nordic cohorts were genotyped with the MassARRAY platform. Logistic regression models were fit within each cohort. In the independent subset of the FIN cohort (567 CRC cases and 13,642 cancer-free controls), the inflation factor was 1.11, and genomic control was applied accordingly. Estimated log ORs were combined under random-effects and fixed-effect models, the results of which were highly similar without notable study heterogeneity (\phi_{het}=0.462, \hat{I}^2=0\%). Applying Bonferroni correction for the 30 variants that were genotyped in the MassARRAY experiment (\alpha=0.05/30=0.00167), rs992157 was significantly associated with CRC with an OR of 1.14 (95% CI, 1.06-1.23; p=2.08x10^{-4}). Consistent with prior results, the alternative allele (A) conferred a higher risk of CRC than the
reference allele (G). For rs992157, $r^2$ between IMPUTE2 genotype dosage and MassARRAY genotype was 1.00 in the FIN cohort.

**Discussion**

The identification of CRC susceptibility alleles and quantification of their effects is biologically and clinically meaningful. The genome-wide statistical analysis of tag SNPs has highlighted new genes and regulatory mechanisms in the pathogenesis of CRC while concurrently allowing more accurate estimation of the personalized risk of colorectal neoplasms.20,21 We conducted a GWAS of CRC in the Finnish population (stage 1), genotyped 30 promising variants in five Nordic cohorts (stage 2), and analyzed corresponding summary statistics from previously published GWASs (stage 3). A total of 39,786 individuals (13,348 CRC cases and 26,438 controls) were analyzed in stages 1 to 3. New genotype data generated in this study were used to analyze the recently reported effect of rs992157 (2q35) on CRC risk.

The association between rs992157 and CRC was independently replicated (p=2.08x10^{-4}), and its effect size was approximately 1.1 (OR, 1.14; 95% CI, 1.06-1.23). In the combined analysis of 13,348 CRC cases and 26,438 controls, the p-value and OR for rs992157 were 1.50x10^{-9} and 1.12 (95% CI, 1.08-1.16), respectively, with no indication of study heterogeneity ($p_{het}$=0.777, $I^2$=0%). In addition to CRC, rs992157 has shown pleiotropic effects on adult human height and inflammatory bowel disease.6,22

In stage 1, we found evidence supporting multiple previously published SNPs as risk factors for CRC in the Finnish population with false discovery rate <0.1. The corresponding chromosomal regions and nearby genes were 2q35 (PNKD and TMBIM1), 6p21.2 (TRNAI25), 8q23.3 (LINC00536 and EIF3H), 8q24.21 (CCAT2 and LOC101930033), 10q22.3 (ZMIZ1-AS1), 10q24.2 (NKX2-3 and SLC25A28), 11q13.4 (POLD3), 11q23.1
(COLCA1 and COLCA2), 14q22.2 (RPS3AP46 and MIR5580), 15q13.3 (SCG5 and GREM1), 18q21.1 (SMAD7), 20p12.3 (FGFR3P3 and CASC20), and 20q13.33 (LAMA5).

We did not find Finnish population-specific CRC risk variants, which may reflect limitations in replicating them in other populations, their rarity, or small contributions to inherited risk. A low-frequency variant at 12q14.3 (rs73121704; MAF, 0.860%) displayed a notable association in stage 1 (p=4.07x10⁻⁹), but the finding was not supported by meta-analysis (random-effects p=0.466, fixed-effect p=0.122). Bias due to genotype imputation or population stratification remains a concern, and further data is needed.

A limitation of the study is that the number of variants selected for stages 2 and 3 was relatively small, and disease-associated variants may have been omitted from further investigation because of low rank in the primary analysis. It is also difficult to assess whether there was residual confounding due to population stratification or different genotyping platforms. For rs992157, r² between IMPUTE2 genotype dosage and MassARRAY genotype was 1.00, making technical bias unlikely. Genomic control was applied for all primary GWASs to avoid type I error.

In conclusion, we replicated the association between rs992157 (2q35) and CRC in Northern European studies and found it to be genome-wide significant in a meta-analysis of twelve European-ancestry studies. SNPs at 2q35, 6p21.2, 8q23.3, 8q24.21, 10q22.3, 10q24.2, 11q13.4, 11q23.1, 14q22.2, 15q13.3, 18q21.1, 20p12.3, and 20q13.33 were associated with CRC in the Finnish population, which validates findings from previous studies and reveals shared genetic architecture of CRC between the Finnish population isolate and outbred populations.

Acknowledgments
We are thankful to Sini Nieminen, Sirpa Soisalo, Marjo Rajalaakso, Inga-Lill Svedberg, Iina Vuoristo, Alison Olikainen, and Heikki Metsola for their technical support. The study was supported by grants from Academy of Finland (Finnish Center of Excellence Program 2012-2017, Project No. 1250345), Cancer Society of Finland, Sigrid Juselius Foundation, Jane and Aatos Erkko Foundation, SYSCOL (an EU FP7 Collaborative Project No. 258236), Nordic Information for Action eScience Center (NIASC), Nordic Center of Excellence financed by NordForsk (Project No. 62721), NordForsk Colorectal Cancer Pilot Project (07 BM 11/424), Cancer Research UK (C348/A18927 for Edinburgh Colon Cancer Genetics Group (CCGG)), and UK Medical Research Council (MR/K018647/1 for Edinburgh CCGG), Estonian RC (IUT20-60 and PUT736), and European Regional Development Fund (Project No. 2014-2020.4.01.15-0012). Niko Välimäki received grant No. 287665 from the Academy of Finland. The Colon Cancer Family Registry (CCFR) was supported by grant UM1 CA167551 from the National Cancer Institute, USA, and through cooperative agreements with the following centres: Australasian Colorectal Cancer Family Registry (U01 CA074778 and 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), University of Hawaii Colorectal Cancer Family Registry (U01/U24 CA074806), and USC Consortium Colorectal Cancer Family Registry U01/U24 CA074799). The CCFR GWASs were supported by grants U01 CA122839, R01 CA143237, and U19 CA148107 from National Cancer Institute, USA. We acknowledge Karolinska Institute Biobank, Sahlgrenska Biobank, HUNT Biobank, THL Biobank, and Estonian Genome Center for providing DNA samples from Nordic CRC cases and controls. The Nord-Trøndelag Health Study (HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, NTNU, Norwegian University of Science and Technology), Nord-Trøndelag County Council, Central Norway Health Authority, and Norwegian Institute of Public Health.
Conflict of interest statement

We have no conflicts of interest to declare.

References


Figure Legends

**Figure 1.** Study scheme. Sources of genetic markers are shown on the left, analytic stages in the center, and sources of samples on the right.

**Figure 2.** Study cohorts, sample sizes, and estimated odds ratios for rs992157. The vertical line corresponds to the null hypothesis (odds ratio=1). The horizontal lines and square brackets indicate 95% confidence intervals. Areas of the boxes are proportional to the weight of the study. Diamonds represent combined estimates. FE, fixed-effect. RE, random-effects.
Figure 1. Study scheme. Sources of genetic markers are shown on the left, analytic stages in the center, and sources of samples on the right.

146x129mm (300 x 300 DPI)
<table>
<thead>
<tr>
<th>Replication cohorts</th>
<th>Cases</th>
<th>Controls</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIN (independent subset)</td>
<td>567</td>
<td>13,642</td>
<td>1.17 [1.02, 1.35]</td>
</tr>
<tr>
<td>STHLM2</td>
<td>544</td>
<td>541</td>
<td>1.29 [1.09, 1.53]</td>
</tr>
<tr>
<td>Gothenburg</td>
<td>1,903</td>
<td>258</td>
<td>1.07 [0.89, 1.28]</td>
</tr>
<tr>
<td>HUNT</td>
<td>1,168</td>
<td>1,147</td>
<td>1.11 [0.99, 1.25]</td>
</tr>
<tr>
<td>Estonia</td>
<td>257</td>
<td>259</td>
<td>1.03 [0.80, 1.31]</td>
</tr>
<tr>
<td>Total (replication)</td>
<td>4,439</td>
<td>15,847</td>
<td></td>
</tr>
<tr>
<td>FE Model (replication)</td>
<td></td>
<td></td>
<td>1.14 [1.06, 1.23]</td>
</tr>
<tr>
<td>RE Model (replication; p=2.08e−04)</td>
<td></td>
<td></td>
<td>1.14 [1.06, 1.23]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other cohorts</th>
<th>Cases</th>
<th>Controls</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINRISK</td>
<td>198</td>
<td>172</td>
<td>1.20 [0.86, 1.68]</td>
</tr>
<tr>
<td>FIN</td>
<td>1,701</td>
<td>14,082</td>
<td>1.15 [1.06, 1.26]</td>
</tr>
<tr>
<td>COIN</td>
<td>1,950</td>
<td>2,162</td>
<td>1.06 [0.96, 1.16]</td>
</tr>
<tr>
<td>UK1</td>
<td>890</td>
<td>900</td>
<td>1.07 [0.93, 1.22]</td>
</tr>
<tr>
<td>Scotland1</td>
<td>973</td>
<td>998</td>
<td>1.13 [0.99, 1.29]</td>
</tr>
<tr>
<td>VQ50</td>
<td>1,794</td>
<td>2,000</td>
<td>1.15 [1.05, 1.25]</td>
</tr>
<tr>
<td>CCFR1</td>
<td>1,175</td>
<td>999</td>
<td>1.06 [0.94, 1.20]</td>
</tr>
<tr>
<td>CCFR2</td>
<td>795</td>
<td>2,234</td>
<td>1.14 [1.01, 1.28]</td>
</tr>
<tr>
<td>Total (combined)</td>
<td>13,348</td>
<td>26,438</td>
<td></td>
</tr>
<tr>
<td>FE Model (combined)</td>
<td></td>
<td></td>
<td>1.12 [1.08, 1.16]</td>
</tr>
<tr>
<td>RE Model (combined; p=1.50e−09)</td>
<td></td>
<td></td>
<td>1.12 [1.08, 1.16]</td>
</tr>
</tbody>
</table>

Odds ratio (log scale)

0.8  1  1.2  1.5

Figure 2. Study cohorts, sample sizes, and estimated odds ratios for rs992157. The vertical line corresponds to the null hypothesis (odds ratio=1). The horizontal lines and square brackets indicate 95% confidence intervals. Areas of the boxes are proportional to the weight of the study. Diamonds represent combined estimates. FE, fixed-effect. RE, random-effects.

142x121mm (300 x 300 DPI)