# Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease

HMG Advance Access published March 22, 2016

# Variation at 2q35 (*PNKD* and *TMBIM1*) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease

Giulia Orlando[1,‡], Philip J. Law[1,‡], Kimmo Palin[2,3,‡], Sari Tuupanen[2,3], Alexandra Gylfe[2,3,†], Ulrika Hänninen[2,3], Tatiana Cajuso[2,3], Tomas Tanskanen[2,3], Johanna Kondelin[2,3], Eevi Kaasinen[2,3], Antti-Pekka Sarin[4], Jaakko Kaprio[4,5], Johan G. Eriksson[6,6,7], Harri Rissanen[5], Paul Knekt[5], Eero Pukkala[8,9], Pekka Jousilahti[5], Veikko Salomaa[5], Samuli Ripatti[4], Aarno Palotie[4,10,11,12] , Heikki Järvinen[13], Laura Renkonen-Sinisalo[14], Anna Lepistö[14], Jan Böhm[15], Jukka-Pekka Meklin[16], Nada A. Al-Tassan[17], Claire Palles[18], Lynn Martin[18], Ella Barclay[18], Albert Tenesa[19,20], Susan Farrington[19], Maria N. Timofeeva[19], Brian F. Meyer[17], Salma M. Wakil[17], Harry Campbell[21], Christopher G. Smith[22], Shelley Idziaszczyk[22], Timothy S. Maughan[23], Richard Kaplan[24], Rachel Kerr[25], David Kerr[26], Daniel D. Buchanan[27,28], Aung Ko Win[28], John Hopper[28], Mark Jenkins[28], Noralane M. Lindor[29], Polly A. Newcomb[30], Steve Gallinger[31], David Conti[32], Fred Schumacher[32], Graham Casey[32], Jussi Taipale[2,3,33], Jeremy P. Cheadle[22], Malcolm G. Dunlop[19], Ian P. Tomlinson[18], Lauri A. Aaltonen[2,3], Richard S. Houlston[1,*]

[1] Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK

[2] Genome-Scale Biology Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland

[3] Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Helsinki, Finland

[4] Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

[5] National Institute for Health and Welfare, Helsinki, Finland

[6] Folkhälsan Research Centre, Helsinki, Finland

[7] Unit of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

[8] Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland

[9] School of Health Sciences, University of Tampere, Tampere, Finland

[10] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

[11] Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA

[12] Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

[13] Helsinki University Central Hospital, Department of Surgery, Hospital District of Helsinki and Uusimaa, Helsinki, Finland

[14] Abdominal Center, Department of Surgery, Helsinki University Hospital, Helsinki, Finland

[15] Department of Pathology, Central Finland Central Hospital, Jyväskylä, Finland

[16] Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Jyväskylä, Finland

[17] Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

[18] Wellcome Trust Centre for Human Genetics and NIHR Comprehensive Biomedical Research Centre, Oxford, UK

[19] Colon Cancer Genetics Group, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital, Edinburgh, UK

[20] The Roslin Institute, University of Edinburgh, Easter Bush, Roslin, UK

[21] Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK

[22] Institute of Cancer and Genetics, School of Medicine, Cardiff University, Cardiff, UK

[23] CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Oxford, UK

[24] MRC Clinical Trials Unit, Aviation House, London, UK

[25] Oxford Cancer Centre, Department of Oncology, University of Oxford, Churchill Hospital, Oxford, UK

[26] Nuffield Department of Clinical Laboratory Sciences, University of Oxford, John Radcliffe Hospital, Oxford, UK

[27] Colorectal Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Victoria, Australia

[28] Centre for Epidemiology and Biostatistics, The University of Melbourne, Victoria, Australia

[29] Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ, USA

[30] Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[31] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada

[32] Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

[33] SciLife Center, Department of Biosciences and Nutrition, Karolinska Institutet, SE 141 83, Sweden

[‡] These authors contributed equally to this work.

[†] Present address: Alexandra Gylfe, Human Longevity Inc., La Jolla, CA

[*] Correspondence to: Richard S Houlston, Division of Genetics and Epidemiology, The Institute of Cancer Research, London, SW7 3RP. Tel: +44 (0) 208 722 4175; Fax: +44 (0) 722 4365; E-mail: richard.houlston@icr.ac.uk

## ABSTRACT

To identify new risk loci for colorectal cancer (CRC) we conducted a meta-analysis of seven genome-wide association studies (GWAS) with independent replication, totalling 13,656 CRC cases and 21,667 controls of European ancestry. The combined analysis identified a new risk association for CRC at 2q35 marked by rs992157 ($P$ = 3.15 × $10^{-8}$, odds ratio = 1.10, 95% confidence interval = 1.06-1.13), which is intronic to *PNKD* and *TMBIM1*. Intriguingly this susceptibility SNP is in strong LD ($r^2$ = 0.90, $D'$ = 0.96) with the previously discovered GWAS SNP rs2382817 for inflammatory bowel disease (IBD). Following on from this observation we examined for pleiotropy, or shared genetic susceptibility, between CRC and the 200 established IBD risk loci, identifying an additional 11 significant associations (FDR < 0.05). Our findings provide further insight into the biological basis of inherited genetic susceptibility to CRC, and identify risk factors that may influence the development of both CRC and IBD.

# INTRODUCTION

Colorectal cancer (CRC), a leading cause of cancer-related death worldwide, has a heritable basis (1, 2). Recent genome-wide association studies (GWAS) have successfully identified a number of common single-nucleotide polymorphisms (SNPs) influencing CRC risk thereby vindicating the assertion that part of the heritable risk is polygenic (3-7). These studies have also provided insights into the biology of CRC, highlighting the importance of bone morphogenetic protein signalling pathway genes (*BMP2, BMP4, GREM1* and *SMAD7*) (4, 5), candidate genes (*CDH1*), as well as genes not previously implicated in CRC (*POLD3, TERC, CDKN1A, VIT1A* and *SHROOM2*) (6, 7). It is well established that inflammatory bowel disease (IBD), which primarily presents as Crohn's disease or ulcerative colitis, is associated with an increased CRC risk (8-11). Despite IBD being strongly heritable (12), little evidence for shared genetic susceptibility or differential effects of genetic variation on IBD and CRC risk has been reported, although the presumption is that the direction of effect will be consistent between both diseases.

A failure to uncover pleiotropy may be reflective of a lack of power of CRC GWAS conducted thus far. Indeed statistical modelling of GWAS data shows that although 19% of the heritability of CRC can be ascribed to common variation, only 10% of this is explained by currently identified risk SNPs (13). To empower the identification of new CRC susceptibility SNPs in persons of European ancestry, we conducted a genome-wide meta-analysis of a previously unreported GWAS with six published datasets in addition to independent replication totalling 13,810 cases and 21,754 controls.

We report the identification of a new CRC risk association which also impacts on IBD risk. Extending our analysis to established IBD loci, we provide evidence of shared genetic susceptibility between CRC and IBD at 11 additional loci.

## RESULTS

### Primary GWAS

In the primary scan (termed the FIN GWAS), 1,172 CRC cases ascertained through the Finnish CRC collection and Finnish Cancer Registry were analysed with control data on 8,266 individuals from the FINRISK, Health2000, Finnish Twin Cohort and Helsinki Birth Cohort Study cohorts. After applying strict quality control criteria, 283,906 autosomal SNPs were available for association with CRC risk. A quantile-quantile (Q-Q) plot of observed versus expected $\chi^2$-test statistics showed little evidence for an inflation of test statistics, thereby excluding the possibility of substantive hidden population substructure, cryptic relatedness among subjects or differential genotype calling (inflation factor $\lambda = 1.07$).

### Meta-analysis

We performed a meta-analysis of our primary scan data with six other non-overlapping GWAS of European ancestry (CCFR1, CCFR2, COIN, UK1, Scotland1, VQ58), which have been previously reported (14). To maximise the prospects of identifying novel risk variants, we imputed the data with a merged reference panel using Sequencing Initiative Suomi (for the FIN data) or UK10K (for the UK data) in addition to 1000 Genomes Project data. After quality control procedures, over 10 million variants, including over 1 million insertion-deletions, were analysed in 8,749 cases and 18,245 controls.

Associations for the 37 previously established European CRC risk SNPs showed a direction of effect consistent with previously reported studies, with 10 of these SNPs having $P < 5.0 \times 10^{-8}$ in this meta-analysis (Supplementary Table 1). Excluding these known risk SNPs, together with those correlated with $r^2 > 0.8$, from the meta-analysis two novel regions of linkage disequilibrium (LD), marked by rs992157 and rs2383207, showed the strongest association with CRC at $P < 1.0 \times 10^{-6}$ (Supplementary Table 2).

To replicate these associations, we genotyped rs992157 and rs2383207 in an additional 5,061 CRC cases and 3,509 controls, with only rs992157 showing evidence for an association with CRC ($P$ = 0.023). In the combined analysis the association was significant at the genome-wide threshold ($P$ = 3.15 x $10^{-8}$; Figure 1). There was no variation due to heterogeneity ($I^2 = 0$, $P_{het} = 0.79$). rs992157 is located at 2q35, and is intronic to two genes: paroxysmal nonkinesigenic dyskinesia (*PNKD*) on the forward strand, and transmembrane BAX inhibitor motif containing 1 (*TMBIM1*) on the reverse strand (Figure 2).

**Relationship between genotype and colorectal cancer phenotype**

Using data on microsatellite instability (MSI) status from the FIN (n = 1,146), COIN (n = 1,239) and NSCCG replication (n = 1,282) series, together with information on KRAS and BRAF mutation status in tumours in COIN, we explored the possibility that the association at rs992157 is restricted to a specific molecular subtype of CRC (Supplementary Table 3). There was no evidence of an association between these SNPs and any of the variables after adjusting for multiple testing (*i.e.* $P >$ 0.05). Additionally we observed no consistent association between age, sex or tumour site using data from the UK1, Scotland1, VQ58, COIN, and NSCCG series (Supplementary Table 3).

**Inflammatory bowel disease SNPs influence colorectal cancer**

Another association at 2q35 defined by rs2382817 has previously been shown to influence IBD risk (CRC meta $P$ = 1.02 x $10^{-5}$), which is also intronic to *PNKD* and *TMBIM1*, and is in strong LD with rs992157 ($r^2 = 0.90$, $D' = 0.96$). Paradoxically, the risk for rs2382817 in IBD is inverse to the CRC association. Given the compelling evidence for an association between IBD and CRC we sought evidence for additional shared susceptibility between the two diseases. Specifically, we examined the risk of CRC in our meta-analysis at 200 loci that have been shown in previous GWAS to affect IBD risk (15, 16) (Supplementary Table 4). A Q-Q plot of the observed CRC association *P*-values against the expected *P*-values for each of the 200 IBD risk SNPs showed significant over-dispersion ($\lambda$ = 1.33, Figure 3). This observation is compatible with a genetic relationship between CRC and IBD.

To account for multiple testing we imposed an FDR-adjusted *P*-value of 0.05 as being statistically significant. At this threshold, in addition to rs2382817, 11 IBD risk SNPs were associated with CRC risk (Table 1), of which five were positively associated with CRC risk, while the other seven displayed an inverse relationship. A number of these SNPs annotate genes with documented roles that are relevant to CRC development, such as Wnt-signalling (*WNT4*, (17)), tumour suppression (*MAPKAPK5*, *FOXO1* (18, 19)) and cellular transformation (*CDC42*, *CEBPB* (20, 21)) (Table 1). We examined for an association between the genotype of these 12 SNPs and the molecular subtype of CRC, and found no evidence of a relationship (Supplementary Table 3).

### Functional effect prediction analysis

The genomic region containing rs992157 is the site of active structure and has regulatory motifs for both enhancer and promotor function in multiple cell types (Figure 2). Moreover ChIP-seq data identifies over 122 transcription factors binding to the region, including CRC related transcription factors such as MYC, HNF4A and TCF7L2 (Supplementary Table 5). We also performed an eQTL analysis and found no significant relationship between the rs992157 genotype and *PNKD* and *TMBIM1* expression in colorectal adenocarcinoma cells (Supplementary Table 6). The risk genotype was however associated with altered gene expression in other tissues, including lymphoblastoid cells (FDR *P*-value < 0.05, Supplementary Table 6). This apparent difference in eQTLs may be reflective of the differences in epigenetic profiles at 2q35 between CRC and lymphoblastoid cells (Figure 2).

To further investigate the relationship between CRC and IBD risk we performed eQTL analysis on the 12 IBD SNPs associated with CRC risk in the colorectal adenocarcinoma data, and found two significant relationships between rs174537 and the expression of fatty acid desaturase 2 (*FADS2*, FDR *P*-value = 3.28 x $10^{-6}$) and between rs516246 and fucosyltransferase 2 (*FUT2,* FDR *P*-value = 2.08 x $10^{-17}$) (Supplementary Table 6). Additional evidence for these eQTLs was found in other tissues in the Geuvadis, Blood and GTEx databases (Supplementary Table 6). Similarly to rs992157, as reported above, rs2382817 is an eQTL for *PNKD* and *TMBIM1* in both lymphoblastoid and whole blood tissues.

Following on from this we investigated the presence of shared genetic pathways between CRC and IBD using the LENS pathway tool (22), which allows exploration of interactions between the gene products in proximity to the GWAS SNPs. Across the 594 CRC proteins and 1,574 IBD proteins, a network of 542 overlapping proteins was identified. Figure 4 shows the common network and interactions between key proteins. Of interest was the direction of association between the CRC SNPs with IBD risk. Pathways with evidence of enrichment (*i.e.* $P < 0.001$) with a consistent effect between CRC and IBD were involved in immune and inflammatory response, such as co-stimulation by the CD28 family, Fc epsilon receptor signalling, and downstream B-cell receptor signalling. In contrast, the protein networks defined by reciprocal SNPs association for CRC and IBD were enriched for interleukin and calmodulin signalling. Pathways that were enriched in both, albeit involving different proteins, included those related to the adaptive immune response, cytokine signalling and interferon signalling (Supplementary Table 7).

## DISCUSSION

In this meta-analysis we combined seven independent GWAS, and have identified a risk locus for CRC risk at 2q35 marked by rs992157. Since this SNP is intronic to both *PNKD* and *TMBIM1*, and these are the only transcripts within the region of high LD, it is a plausible that the genetic basis of the 2q35 association for CRC is through functional effects on one of these genes *a priori*. This is coupled with the fact that rs992157 localises to a genomic region with regulatory function and the eQTL data showing allele-specific *cis*-regulatory relationship between SNP genotype and *PNKD* and *TMBIM1* expression. Although speculative the long isoform of *PNKD* appears to function in a pathway to detoxify alpha-ketoaldehyde using glutathione as a cofactor (23). Since glutathione is essential for maintaining cellular redox status, reduced glutathione levels in cells through dysfunctional PNKD may lead to increasing oxidative stress levels, which have been linked to inflammation (24). TMBIM1 has been reported to have a role in regulating the level of Fas ligand (FasL) (25, 26), which mediates both apoptosis and inflammation (27). Therefore, both gene products indirectly contribute to the regulation of inflammation, a physiological process linked with the onset of IBD and CRC.

Another SNP in the 2q35 locus (rs2382817), which is in strong LD with rs992157, has previously been shown to influence IBD risk (15). In addition, contemporaneous with our analysis, a recent study (28) has also found evidence, albeit not GWAS significant, for a relationship between 2q35 variation and CRC risk ($P$ = 7.0 x $10^{-5}$), additionally finding an inverse relationship with risk of IBD. The identified SNP, rs11676348, is correlated with both rs992157 and rs2382817 (LD metrics, $r^2$ and $D'$ = 0.32, 0.65 and 0.33, 0.71 respectively). The opposing effects of the rs2382817-C allele with increased risk of CRC but decreased risk of IBD may initially appear paradoxical, given the increased risk of CRC associated with IBD. The risk of CRC in IBD increases with longer duration, extent of colitis and the degree of inflammation (11). The inflammatory response has been linked to increased oxidative stress, and this oxidative state stimulates antioxidant defences that promote the survival pathways in cancer cells, favouring tumour proliferation (29). Nonetheless, these SNPs may indicate shared pathways in which there are opposing relationships between carcinogenesis and inflammation.

Motivated by the observation that the 2q35 locus influences IBD risk, we sought additional evidence for a common genetic basis for both diseases by evaluating the CRC risk at previously established IBD loci (15, 16). While not formally significant globally, there was an over-representation of association signals for CRC defined by the IBD risk SNPs. Through this analysis we identified potential risk variants for CRC mapped to regions in the proximity of genes encoding WNT4 and CDC42, previously shown to be involved in the risk of CRC (14); MAPKAPK5, a member of the MAPK family reported to regulate MYC protein levels (18); and the transcription factor CEBPB, found to be highly expressed in samples derived from CRC patients (21). Moreover our eQTL analysis on IBD SNPs showed altered expression of *FADS2* and *FUT2* genes in CRC tissues. Both the genes have previously been reported to have a role in the development of IBD (30, 31) providing further evidence of possible shared genes. Further studies are required to delineate the genetic basis and implicate perturbation of a specific gene as the functional basis of the associations. Collectively these data are consistent with a degree of commonality in genetically-defined pathways in the development between CRC and IBD, albeit that many of the associations have opposite effects.

Considering the low prevalence of IBD in European populations (< 0.5%) (32), together with the observation that other SNPs that are strongly associated with risk of IBD were not associated with CRC, it is unlikely that sampling has biased our findings. Moreover if the association between these IBD SNPs and CRC was simply mediated by its association with IBD *per se*, we would have expected directionality of the association to be identical but this was not the case for many of the SNPs.

In summary, we have identified a new risk association for CRC which also influences IBD risk. Our association signals for CRC defined by other established IBD risk SNPs also serve to highlight the importance of shared gene pathways in the development of CRC and IBD. Deciphering the functional and biological basis of these SNPs associations has the potential to translate into a better understanding of the biological basis of how IBD transitions to CRC. Finally our analysis serves to illustrate that inter-relationships between diseases do not necessarily equate to consistent allelic architecture in risk, thus adding an extra layer of complexity to interpretation.

## MATERIALS AND METHODS

### Ethics

Collection of blood samples and clinico-pathological information from subjects was undertaken with informed consent and ethical review board approval at all sites in accordance with the tenets of the Declaration of Helsinki.

### Primary GWAS

The Finnish GWAS (FIN) was based on 1,172 CRC cases and 8,266 cancer free controls ascertained through Finnish Hospitals (33) and through the Finnish Cancer Registry. Cases were genotyped using Illumina HumanOmni 2.5M8v1 according to the manufacturer's recommendations. For controls, we made use of Illumina HumanHap 670k and 610k array data on individuals from the FINRISK (34), Health 2000 (35), Finnish Twin Cohort (36) and Helsinki Birth Cohort Studies (37). Individuals were excluded with: <90% successfully genotyped SNPs, discordant sex information, duplication or cryptic relatedness (identity by descent >0.2). We excluded SNPs from the analysis with: call rate <95%, MAF < 0.01, and departure from Hardy-Weinberg equilibrium in controls at $P < 10^{-6}$. The adequacy of the case-control matching and the possibility of differential genotyping of cases and controls were assessed using quantile-quantile (Q-Q) plots of test statistics.

### Published GWAS for meta-analysis

We made use of six previously published GWAS: UK1 (CORGI study) (7) comprised 940 cases with colorectal neoplasia and 965 controls; Scotland1 (COGS study) (7) included 1,012 CRC cases and 1,012 controls; VQ58 comprised 1,800 CRC cases from the UK-based VICTOR and QUASAR2 adjuvant chemotherapy clinical trials (38) and 2,690 population control genotypes from the Wellcome Trust Case Control Consortium 2 (WTCCC2) 1958 birth cohort (39); CCFR1 comprised 1,290 familial CRC cases and 1,055 controls from the Colon Cancer Family Registry (CCFR) (40); CCFR2 included a further 796 cases from the CCFR and 2,236 controls from the Cancer Genetic Markers of Susceptibility (CGEMS) studies of breast and prostate cancer (41, 42); and the COIN GWAS (14) was based on 2,244 CRC cases ascertained through two independent Medical Research Council clinical trials of advanced/metastatic CRC (COIN and COIN-B) (43) and controls comprised 2,162 individuals from the UK Blood Service Control Group genotyped as part of the WTCCC2 (39).

The VQ58, UK1 and Scotland1 GWAS series were genotyped using Illumina Hap300, Hap240S, Hap370, Hap550 or Omni2.5M arrays. 1958BC genotyping was performed as part of the WTCCC2 study on Hap1.2M-Duo Custom arrays. The CCFR samples were genotyped using Illumina Hap1M, Hap1M-Duo or Omni-express arrays. CGEMS samples were genotyped using Illumina Hap300 and Hap240 or Hap550 arrays. The COIN cases were genotyped using Affymetrix Axiom Arrays and the Blood Service controls were genotyped using Affymetrix 6.0 arrays. After applying the same quality control as that performed for FIN, data on 8,749 CRC cases and 18,245 controls were available for the meta-analysis.

The adequacy of the case-control matching and possibility of differential genotyping of cases and controls was assessed using Q-Q plots of test statistics. $\lambda_{GC}$ values (44) for the UK1, Scotland1, VQ58, CCFR1, CCFR2 and COIN studies were 1.02, 1.01, 1.01, 1.02, 1.03 and 1.05 respectively. Any ethnic outliers or individuals identified as related were excluded.

**Replication series**

5,061 CRC cases from the National Study of Colorectal Cancer Genetics (NSCCG) (45) were genotyped. Controls (n = 3,509) were from NSCCG and the Genetic Lung Cancer Predisposition

Study (GELCAPS) (46). None of the controls had a known history of malignancy at ascertainment. All subjects were British residents with self-reported European ethnicity and there were no obvious demographic differences between cases and controls. DNA was extracted from EDTA-venous blood samples using conventional methodologies and PicoGreen quantified (Invitrogen Corporation, Carlsbad, CA, USA). Genotyping of two SNPs was conducted using KASPar competitive allele-specific PCR chemistry (LGC, Hoddesdon, UK; primer sequences and conditions available on request). To monitor quality control, duplicate samples were included in assays, and concordance between duplicate samples was >99%.

**Imputation and meta-analysis**

Analyses were undertaken using R (v3.02) (47) and PLINK (v1.9) (48) software. Phasing of GWAS SNP genotypes was performed using SHAPEIT (v2.r644 and v2.r790 for FIN) (49). Prediction of the untyped SNPs was carried out using IMPUTE (v2.3.1) (50). The FIN dataset used a merged reference panel based on data from the 1000 Genomes Project (Phase 1 v3) (51) together with an additional population matched reference panel of 3,882 Sequencing Initiative Suomi (SISu) haplotypes. The UK samples used a merged reference panel using data from the 1000 Genomes Project and UK10K (April 2014 release). The fidelity of imputation, as assessed by the concordance between imputed and sequenced SNPs, was examined in a subset of 200 UK cases (14). The association between each SNP and the risk of CRC was assessed by a frequentist association test under an additive model, using SNPTEST (v2.5.1) (52), utilising the genotype probabilities from IMPUTE where a SNP was not directly typed. Population stratification was controlled in the FIN samples using sex and six principal components. Association meta-analyses only included markers with info scores >0.8, imputed call rates/SNP >0.9 and MAFs >0.005. Meta-analyses were carried out using META (v1.6) (53). We calculated Cochran's $Q$ statistic to test for heterogeneity and the $I^2$ statistic to quantify the proportion of the total variation that was caused by heterogeneity (54). $I^2$ values ≥75% are considered characteristic of large heterogeneity (54).

## Characterisation of cancer phenotype

Associations by sex, age and clinico-pathological phenotypes were examined by logistic regression. MSI status was determined using BAT25 and BAT26 markers, and samples showing ≥ 5 novel alleles when compared with normal DNA at either or both markers were assigned as MSI-H (corresponding to MSI-high) (55). Tumours were screened for *KRAS* codons 12, 13, and 61 and *BRAF* codon 600 mutations by pyrosequencing (43). Additionally, *KRAS* (all three codons) and *BRAF* (codons 594 and 600) were screened for mutations by MALDI-TOF mass array (Sequenom, San Diego, CA, USA) (56). Differences between the various sites of the tumour (colonic [ICD-9:153], rectal [ICD-9:154.1], and recto sigmoid junction [ICD9:154.0]) were also analysed.

## Functional prediction

To explore epigenetic profiles of genomic location associated with CRC, we used ENCODE histone modification data, HaploReg and RegulomeDB (57, 58) to examine whether any of the SNPs or their proxies (*i.e.* $r^2 > 0.8$ in the 1000 Genomes EUR reference panel) annotate transcription factor binding or enhancer elements. Additionally we made use of ChIP-seq data on the LoVo CRC cell line (59). We used ChromHMM to integrate DNAse, H3K4me3, H3K4me1, H3K27ac, Pol2 and CTCF states from the CRC cell line HCT116 using a multivariate Hidden Markov Model (60). ChromHMM tracks for lymphoblastoid cells were obtained from ENCODE (61). We assessed sequence conservation using: PhastCons (62) (>0.3 indicative of conservation) and Genomic Evolutionary Rate Profiling (GERP) (63) (>2 indicative of conservation). SNAP plots were created using the visPIG tool (64).

## eQTL analysis

To examine for a relationship between SNP genotype and mRNA expression in CRC we analysed Tumor Cancer Genome Atlas (TCGA) RNA-seq expression and Affymetrix 6.0 SNP data (dbGaP accession number: phs000178.v7.p6) on 416 colorectal adenocarcinoma samples (65). Association between normalised RNA counts per-gene and SNP genotype was quantified using the Kruskal-Wallis trend test. To look for a relationship between SNP genotype and expression levels in other tissues we used publicly available expression data generated from the MuTHER (66), eQTL Blood Browser (67), GTEx (68) and Geuvadis/1000 Genomes (69) resources. For the Geuvadis data, the relationship between SNPs and expression of genes located within 1 Mb was analysed using the Matrix eQTL (70) package under a linear model. When the SNPs were not directly typed, a proxy SNP was used ($r^2 \geq 0.8$). In all the datasets, eQTL results were included where FDR $P < 0.05$

**Relationship between established risk SNPs for inflammatory bowel disease and colorectal cancer**

To investigate pleiotropic (shared genetic susceptibility) between CRC and inflammatory bowel disease (IBD), we examined the 201 SNPs identified in GWAS that have been shown to affect IBD risk (15, 16). One SNP (rs71559680) is an indel that was not present in the CRC genotyping arrays or the reference panels, and was thus removed from the analysis. We obtained the lead SNPs from the IBD GWAS and extracted the *P*-values for the corresponding SNPs in our CRC meta-analysis.

**Pathway analysis**

To investigate the possibility of shared genetic susceptibility between CRC and IBD, we performed pathway analysis. First, we selected the two closest coding genes for the leading SNPs in each GWAS and then performed pathway analysis using LENS tool (22), which identifies gene product and protein-protein interactions from HPRD (71) and BioGRID (72). Enrichment of pathways was assessed using Fisher's exact test, comparing the overlap of the genes in the network with the genes in the pathway. Pathway data was obtained from REACTOME (73). Cytoscape was used to perform network analyses (74), and the Hive Plot was drawn using HiveR (academic.depauw.edu/~hanson/HiveR/HiveR.html).

## ACKNOWLEDGEMENTS

We would like to thank Amit Sud for his critical reading of the manuscript. We are grateful to all individuals who participated in the various studies. This study made use of genotyping data from the 1958 Birth Cohort, kindly made available by the Wellcome Trust Case Control Consortium 2. A full list of the investigators who contributed to the generation of the data is available at http://www.wtccc.org.uk/. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

In Finland, this work was supported by grants from the Academy of Finland [Finnish Center of Excellence Program 2012–2017, 250345], the Jane and Aatos Erkko Foundation, the Finnish Cancer Society [personal grant to K.P.], the European Research Council [ERC; 268648], the Sigrid Juselius Foundation, SYSCOL, the Nordic Information for Action eScience Center (NIASC), the Nordic Center of Excellence financed by NordForsk [project 62721, personal grant to K.P.] and State Research Funding of Kuopio University Hospital [B1401]. We acknowledge the computational resources provided by the ELIXIR node, hosted at the CSC–IT Center for Science, Finland, and funded by the Academy of Finland [grants 271642 and 263164], the Ministry of Education and Culture, Finland. V.S. was supported by the Finnish Academy [grant number 139635]. Sample collection and genotyping in the Finnish Twin Cohort has been supported by the Wellcome Trust Sanger Institute, ENGAGE – European Network for Genetic and Genomic Epidemiology, FP7-HEALTH-F4-2007 [grant agreement number 201413], the National Institute of Alcohol Abuse and Alcoholism [grants AA-12502 and AA-00145 to Richard J Rose and K02AA018755 to Danielle M Dick] and the Academy of Finland [grants 100499, 205585, 265240 and 263278 to J.K.].

At the Institute of Cancer Research, this work was supported by Cancer Research UK [C1298/A8362 - Bobby Moore Fund for Cancer Research UK]. Additional support was provided by the National Cancer Research Network and the NHS via the Biological Research Centre of the National Institute for Health Research at the Royal Marsden Hospital NHS Trust. In Edinburgh the work was supported by Programme Grant funding from Cancer Research UK [C348/A12076]. In Oxford additional funding was provided by the Oxford Comprehensive Biomedical Research Centre [to C.P. and I.P.M.T.] and the EU FP7 CHIBCHA grant [I.P.M.T.]. Core infrastructure support to the Wellcome Trust Centre for Human Genetics, Oxford was provided by grant [090532/Z/09/Z]. We are grateful to many colleagues within UK Clinical Genetics Departments (for CORGI) and to many collaborators who participated in the VICTOR and QUASAR2 trials. We also thank colleagues from the UK National Cancer Research Network (for NSCCG). Support from the European Union [FP7/207-2013, grant 258236] and FP7 collaborative project SYSCOL and COST Action in the UK is also acknowledged [BM1206]. The COIN and COIN-B trials were funded by Cancer Research UK and the Medical Research Council and were conducted with the support of the National Institute of Health Research Cancer Research Network. COIN and COIN-B translational studies were supported by the Bobby Moore Fund from Cancer Research UK, Tenovus, the Kidani Trust, Cancer Research Wales and the National Institute for Social Care and Health Research Cancer Genetics Biomedical Research Unit (2011-2014) [all awarded to J.P.C.]. N.A.A., B.F.M. and S.M.W. were funded and supported by KFSHRC.

The work of the Colon Cancer Family Registry (CCFR) was supported by from the National Cancer Institute [grant UM1 CA167551], National Institutes of Health and through cooperative agreements with the following CCFR centres: Australasian Colorectal Cancer Family Registry [U01 CA074778, U01/U24 CA097735], USC Consortium Colorectal Cancer Family Registry [U01/U24 CA074799], Mayo Clinic Cooperative Familial 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 University of Hawaii Colorectal Cancer Family Registry [U01/U24 CA074806].

The CCFR Illumina GWAS was supported by funding from the National Cancer Institute, National Institutes of Health [U01 CA122839 and R01 CA143237 to G.C.]. Seattle CCFR research was also supported by the Cancer Surveillance System of the Fred Hutchinson Cancer Research Center [Control Nos. N01-CN-67009 (1996-2003), N01-PC-35142 (2003-2010)] and the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute [Contract No. HHSN2612013000121 (2010-2017)] with additional support from the Fred Hutchinson Cancer Research Center. The collection of cancer incidence data for the State of Hawai'i used in this study was supported by the Hawai‘i Department of Health as part of the statewide cancer reporting program mandated by Hawai‘i Revised Statutes; the National Cancer Institute's Surveillance, Epidemiology and End Results Program (SEER) awarded to the University of Hawai‘i [Control Nos. N01-PC-67001 (1996-2003), N01-PC-35137 (2003-2010), Contract Nos. HHSN26120100037C (2010-2013), HHSN261201300009I (2010-current)]. The ideas and opinions expressed herein are those of the author(s) and endorsement by the State of Hawai‘i, Department of Health, the National Cancer Institute, SEER Program or their Contractors and Subcontractors is not intended nor should be inferred. 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. 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 CCFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CCFR.

## CONFLICT OF INTEREST STATEMENT

None declared.

## REFERENCES

1 Jiao, S., Peters, U., Berndt, S., Brenner, H., Butterbach, K., Caan, B.J., Carlson, C.S., Chan, A.T., Chang-Claude, J., Chanock, S. *et al.* (2014) Estimating the heritability of colorectal cancer. *Human molecular genetics*, **23**, 3898-3905.

2 Peters, U., Jiao, S., Schumacher, F.R., Hutter, C.M., Aragaki, A.K., Baron, J.A., Berndt, S.I., Bezieau, S., Brenner, H., Butterbach, K. *et al.* (2013) Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology*, **144**, 799-807 e724.

3 Whiffin, N. and Houlston, R.S. (2014) Architecture of inherited susceptibility to colorectal cancer: a voyage of discovery. *Genes*, **5**, 270-284.

4 Tomlinson, I.P., Carvajal-Carmona, L.G., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Palles, C., Broderick, P., Jaeger, E.E., Farrington, S. *et al.* (2011) Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS genetics*, **7**, e1002105.

5 Broderick, P., Carvajal-Carmona, L., Pittman, A.M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Spain, S., Sullivan, K., Fielding, S. *et al.* (2007) A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet*, **39**, 1315-1317.

6 Dunlop, M.G., Dobbins, S.E., Farrington, S.M., Jones, A.M., Palles, C., Whiffin, N., Tenesa, A., Spain, S., Broderick, P., Ooi, L.Y. *et al.* (2012) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet*, **44**, 770-776.

7 Houlston, R.S., Cheadle, J., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Spain, S.L., Broderick, P., Domingo, E., Farrington, S. *et al.* (2010) Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet*, **42**, 973-977.

8 Ullman, T.A. and Itzkowitz, S.H. (2011) Intestinal inflammation and cancer. *Gastroenterology*, **140**, 1807-1816.

9 Rizzo, A., Pallone, F., Monteleone, G. and Fantini, M.C. (2011) Intestinal inflammation and colorectal cancer: a double-edged sword? *World journal of gastroenterology : WJG*, **17**, 3092-3100.

10 O'Connor, P.M., Lapointe, T.K., Beck, P.L. and Buret, A.G. (2010) Mechanisms by which inflammation may increase intestinal cancer risk in inflammatory bowel disease. *Inflammatory bowel diseases*, **16**, 1411-1420.

11 Xie, J. and Itzkowitz, S.H. (2008) Cancer in inflammatory bowel disease. *World journal of gastroenterology : WJG*, **14**, 378-389.

12 Bengtson, M.-B., Solberg, C., Aamodt, G., Sauar, J., Jahnsen, J., Moum, B., Lygren, I. and Vatn, M.H. (2009) Familial aggregation in Crohn's disease and ulcerative colitis in a Norwegian population-based cohort followed for ten years. *Journal of Crohn's and Colitis*, **3**, 92-99.

13 Frampton, M.J., Law, P., Litchfield, K., Morris, E.J., Kerr, D., Turnbull, C., Tomlinson, I.P. and Houlston, R.S. (2015) Implications of polygenic risk for personalised colorectal cancer screening. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, in press.

14 Al-Tassan, N.A., Whiffin, N., Hosking, F.J., Palles, C., Farrington, S.M., Dobbins, S.E., Harris, R., Gorman, M., Tenesa, A., Meyer, B.F. *et al.* (2015) A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports*, **5**, 10442.

15 Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A. *et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119-124.

16 Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T. *et al.* (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*, **47**, 979-986.

17 Bernard, P., Fleming, A., Lacombe, A., Harley, V.R. and Vilain, E. (2008) Wnt4 inhibits beta-catenin/TCF signalling by redirecting beta-catenin to the cell membrane. *Biology of the cell / under the auspices of the European Cell Biology Organization*, **100**, 167-177.

18 Kress, T.R., Cannell, I.G., Brenkman, A.B., Samans, B., Gaestel, M., Roepman, P., Burgering, B.M., Bushell, M., Rosenwald, A. and Eilers, M. (2011) The MK5/PRAK kinase and Myc form a negative feedback loop that is disrupted during colorectal tumorigenesis. *Molecular cell*, **41**, 445-457.

19 Gao, F. and Wang, W. (2015) MicroRNA-96 promotes the proliferation of colorectal cancer cells and targets tumor protein p53 inducible nuclear protein 1, forkhead box protein O1 (FOXO1) and FOXO3a. *Molecular medicine reports*, **11**, 1200-1206.

20 Stengel, K. and Zheng, Y. (2011) Cdc42 in oncogenic transformation, invasion, and tumorigenesis. *Cellular signalling*, **23**, 1415-1423.

21 Birkenkamp-Demtroder, K., Mansilla, F., Sorensen, F.B., Kruhoffer, M., Cabezon, T., Christensen, L.L., Aaltonen, L.A., Verspaget, H.W. and Orntoft, T.F. (2007) Phosphoprotein Keratin 23 accumulates in MSS but not MSI colon cancers in vivo and impacts viability and proliferation in vitro. *Molecular oncology*, **1**, 181-195.

22 Handen, A. and Ganapathiraju, M.K. (2015) LENS: web-based lens for enrichment and network studies of human proteins. *BMC Med Genomics*, **8 Suppl 4**, S2.

23 Shen, Y., Lee, H.Y., Rawson, J., Ojha, S., Babbitt, P., Fu, Y.H. and Ptacek, L.J. (2011) Mutations in PNKD causing paroxysmal dyskinesia alters protein cleavage and stability. *Human molecular genetics*, **20**, 2322-2332.

24 Reuter, S., Gupta, S.C., Chaturvedi, M.M. and Aggarwal, B.B. (2010) Oxidative stress, inflammation, and cancer: how are they linked? *Free radical biology & medicine*, **49**, 1603-1616.

25 Rojas-Rivera, D. and Hetz, C. (2015) TMBIM protein family: ancestral regulators of cell death. *Oncogene*, **34**, 269-280.

26 Shukla, S., Fujita, K., Xiao, Q., Liao, Z., Garfield, S. and Srinivasula, S.M. (2011) A shear stress responsive gene product PP1201 protects against Fas-mediated apoptosis by reducing Fas expression on the cell surface. *Apoptosis*, **16**, 162-173.

27 O'Connell, J., Houston, A., Bennett, M.W., O'Sullivan, G.C. and Shanahan, F. (2001) Immune privilege or inflammation? Insights into the Fas ligand enigma. *Nature medicine*, **7**, 271-274.

28 Khalili, H., Gong, J., Brenner, H., Austin, T.R., Hutter, C.M., Baba, Y., Baron, J.A., Berndt, S.I., Bezieau, S., Caan, B. *et al.* (2015) Identification of a Common Variant with Potential Pleiotropic Effect on Risk of Inflammatory Bowel Disease and Colorectal Cancer. *Carcinogenesis*, in press.

29 Guina, T., Biasi, F., Calfapietra, S., Nano, M. and Poli, G. (2015) Inflammatory and redox reactions in colorectal carcinogenesis. *Annals of the New York Academy of Sciences*, **1340**, 95-103.

30 McGovern, D.P., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C. *et al.* (2010) Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Human molecular genetics*, **19**, 3468-3476.

31 Costea, I., Mack, D.R., Lemaitre, R.N., Israel, D., Marcil, V., Ahmad, A. and Amre, D.K. (2014) Interactions between the dietary polyunsaturated fatty acid ratio and genetic factors determine susceptibility to pediatric Crohn's disease. *Gastroenterology*, **146**, 929-931.

32 Molodecky, N.A., Soon, I.S., Rabi, D.M., Ghali, W.A., Ferris, M., Chernoff, G., Benchimol, E.I., Panaccione, R., Ghosh, S., Barkema, H.W. *et al.* (2012) Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. *Gastroenterology*, **142**, 46-54.e42.

33 Aaltonen, L.A., Salovaara, R., Kristo, P., Canzian, F., Hemminki, A., Peltomaki, P., Chadwick, R.B., Kaariainen, H., Eskelinen, M., Jarvinen, H. *et al.* (1998) Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N Engl J Med*, **338**, 1481-1487.

34 Vaara, S., Nieminen, M.S., Lokki, M.L., Perola, M., Pussinen, P.J., Allonen, J., Parkkonen, O. and Sinisalo, J. (2012) Cohort Profile: the Corogene study. *Int J Epidemiol*, **41**, 1265-1271.

35 Kristiansson, K., Perola, M., Tikkanen, E., Kettunen, J., Surakka, I., Havulinna, A.S., Stančáková, A., Barnes, C., Widen, E., Kajantie, E. *et al.* (2012) Genome-Wide Screen for Metabolic Syndrome Susceptibility Loci Reveals Strong Lipid Gene Contribution But No Evidence for Common Genetic Basis for Clustering of Metabolic Syndrome Traits. *Circulation: Cardiovascular Genetics*, **5**, 242-249.

36 Kettunen, J., Tukiainen, T., Sarin, A.P., Ortega-Alonso, A., Tikkanen, E., Lyytikainen, L.P., Kangas, A.J., Soininen, P., Wurtz, P., Silander, K. *et al.* (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet*, **44**, 269-276.

37 Eriksson, J.G. (2011) Early growth and coronary heart disease and type 2 diabetes: findings from the Helsinki Birth Cohort Study (HBCS). *The American journal of clinical nutrition*, **94**, 1799S-1802S.

38 Midgley, R.S., McConkey, C.C., Johnstone, E.C., Dunn, J.A., Smith, J.L., Grumett, S.A., Julier, P., Iveson, C., Yanagisawa, Y., Warren, B. *et al.* (2010) Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *J Clin Oncol*, **28**, 4575-4580.

39 Power, C. and Elliott, J. (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol*, **35**, 34-41.

40 Newcomb, P.A., Baron, J., Cotterchio, M., Gallinger, S., Grove, J., Haile, R., Hall, D., Hopper, J.L., Jass, J., Le Marchand, L. *et al.* (2007) Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev*, **16**, 2331-2343.

41 Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A. *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*, **39**, 870-874.

42 Yeager, M., Orr, N., Hayes, R.B., Jacobs, K.B., Kraft, P., Wacholder, S., Minichiello, M.J., Fearnhead, P., Yu, K., Chatterjee, N. *et al.* (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*, **39**, 645-649.

43 Maughan, T.S., Adams, R.A., Smith, C.G., Meade, A.M., Seymour, M.T., Wilson, R.H., Idziaszczyk, S., Harris, R., Fisher, D., Kenny, S.L. *et al.* (2011) Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet*, **377**, 2103-2114.

44 Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E. *et al.* (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, **37**, 1243-1246.

45 Penegar, S., Wood, W., Lubbe, S., Chandler, I., Broderick, P., Papaemmanuil, E., Sellick, G., Gray, R., Peto, J. and Houlston, R. (2007) National study of colorectal cancer genetics. *Br J Cancer*, **97**, 1305-1309.

46 Eisen, T., Matakidou, A., Houlston, R. and Consortium, G. (2008) Identification of low penetrance alleles for lung cancer: the GEnetic Lung CAncer Predisposition Study (GELCAPS). *BMC cancer*, **8**, 244.

47 R Core Team 2013. (Accessed 01/12/2014) R: A language and environment for statistical computing. . *R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (Date of access 01/12/2014);*, in press.

48 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, **81**, 559-575.

49 Delaneau, O., Marchini, J. and Zagury, J.F. (2012) A linear complexity phasing method for thousands of genomes. *Nature methods*, **9**, 179-181.

50 Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, **5**, e1000529.

51 The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.

52 Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, **39**, 906-913.

53 Liu, J.Z., Tozzi, F., Waterworth, D.M., Pillai, S.G., Muglia, P., Middleton, L., Berrettini, W., Knouff, C.W., Yuan, X., Waeber, G. *et al.* (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet*, **42**, 436-440.

54 Higgins, J.P. and Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, **21**, 1539-1558.

55 Boland, C.R., Thibodeau, S.N., Hamilton, S.R., Sidransky, D., Eshleman, J.R., Burt, R.W., Meltzer, S.J., Rodriguez-Bigas, M.A., Fodde, R., Ranzani, G.N. *et al.* (1998) A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer research*, **58**, 5248-5257.

56 Smith, C.G., Fisher, D., Claes, B., Maughan, T.S., Idziaszczyk, S., Peuteman, G., Harris, R., James, M.D., Meade, A., Jasani, B. *et al.* (2013) Somatic profiling of the epidermal growth factor receptor pathway in tumors from patients with advanced colorectal cancer treated with chemotherapy +/- cetuximab. *Clin Cancer Res*, **19**, 4104-4113.

57 Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, **22**, 1790-1797.

58 Ward, L.D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*, **40**, D930-934.

59 Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M. *et al.* (2013) Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, **154**, 801-813.

60 Jager, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H.E., Heindl, A., Whiffin, N., Carnicer, M.J., Broome, L., Dryden, N. *et al.* (2015) Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications*, **6**, 6178.

61 Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, **9**, 215-216.

62 Duret, L. and Galtier, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, **10**, 285-311.

63 Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, **6**, e1001025.

64 Scales, M., Jager, R., Migliorini, G., Houlston, R.S. and Henrion, M.Y. (2014) visPIG--a web tool for producing multi-region, multi-track, multi-scale plots of genetic data. *PloS one*, **9**, e107497.

65 (Accessed 01/12/2014) The Cancer Genome Atlas *http://cancergenome.nih.gov/*, in press.

66 Grundberg, E., Meduri, E., Sandling, J.K., Hedman, A.K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M. *et al.* (2013) Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *American journal of human genetics*, **93**, 876-890.

67 Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E. *et al.* (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*, **45**, 1238-1243.

68 The GTEx Consortium. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**, 648-660.

69 Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506-511.

70 Shabalin, A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353-1358.

71 Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. *et al.* (2009) Human Protein Reference Database--2009 update. *Nucleic acids research*, **37**, D767-772.

72 Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic acids research*, **41**, D816-823.

73 Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic acids research*, **42**, D472-477.

74 Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431-432.

## FIGURE LEGENDS

**Figure 1: Forest plot of the odds ratios for the association between rs992157 and colorectal cancer.** Studies were weighted according to the inverse of the variance of the log of the OR. *Horizontal lines:* 95% confidence intervals *(95% CI). Box:* OR point estimate; its area is proportional to the weight of the study. *Diamond:* overall summary estimate, with confidence interval given by its width. *Vertical line:* null value (OR = 1.0).

**Figure 2: Regional plot of association results and recombination rates for the 2q35 locus.** In the panel, $-\log_{10} P$ values (*y* axis) of the SNPs are shown according to their chromosomal positions (*x* axis). The top SNP is shown as a large triangle and is labelled by its rsID. The colour intensity of each symbol reflects the extent of LD with the top SNP: white ($r^2$ = 0) through to dark red ($r^2$ = 1.0), with $r^2$ estimated from the 1000 Genomes Phase 1 data. Genetic recombination rates (cM/Mb), are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to each region of association. The lower panel shows the chromatin state segmentation track (ChromHMM) in HCT116 CRC and GM12878 lymphoblastoid cell lines.

**Figure 3: Quantile-Quantile (Q-Q) plot of observed and expected colorectal cancer association *P*-values for 200 inflammatory bowel disease risk SNPs (15, 16)**.

**Figure 4: Hive Plot of common protein-protein interactions between colorectal cancer and inflammatory bowel disease defined by risk SNPs**. Each arc represents an interaction between two proteins, and the distance from the centre of the plot corresponds to a greater number of protein-protein interactions (higher degree of the node). The left arm represents proteins that were only identified using the CRC SNPs, the right arm represents proteins that were only identified

using the IBD SNPs, and the central arm represents the common proteins, highlighting the previously associated tag genes.

1

## TABLES

**Table 1**: Table of the IBD SNPs with FDR corrected *P*-value < 0.05 in the CRC GWAS.

| rsid | Chr | Position | Tag genes | CRC risk allele | IBD risk allele | CRC RAF | CRC p-value | CRC FDR corrected | CRC OR | CRC 95% CI |
|---|---|---|---|---|---|---|---|---|---|---|
| rs12568930 | 1 | 22,702,231 | WNT4, CDC42 | T | T | 0.85 | 6.58x10$^{-05}$ | 3.29x10$^{-03}$ | 1.12 | (1.06; 1.18) |
| rs7554511 | 1 | 200,877,562 | GPR25, C1orf106 | A | C | 0.29 | 6.95x10$^{-04}$ | 0.02 | 1.08 | (1.03; 1.13) |
| rs7608910 | 2 | 61,204,856 | PUS10, REL | A | G | 0.63 | 7.28x10$^{-04}$ | 0.02 | 1.07 | (1.03; 1.12) |
| rs17229285 | 2 | 199,523,122 | PLCL1, SATB2 | C | C | 0.49 | 2.46x10$^{-03}$ | 0.04 | 1.06 | (1.02; 1.1) |
| rs2382817 | 2 | 219,151,218 | TMBIM1, PNKD | C | A | 0.62 | 1.02x10$^{-05}$ | 1.02x10$^{-03}$ | 1.09 | (1.05; 1.14) |
| rs4722672 | 7 | 27,231,762 | HOXA13, HOXA11 | C | C | 0.20 | 2.46x10$^{-03}$ | 0.04 | 1.08 | (1.03; 1.13) |
| rs174537 | 11 | 61,552,680 | MYRF, TMEM258 | G | T | 0.67 | 2.63x10$^{-03}$ | 0.04 | 1.06 | (1.02; 1.11) |
| rs653178 | 12 | 112,007,756 | ATXN2, MAPKAPK5 | T | C | 0.54 | 2.23x10$^{-05}$ | 1.49x10$^{-03}$ | 1.09 | (1.05; 1.13) |
| rs17085007 | 13 | 27,531,267 | GPR12, UPS12 | C | C | 0.19 | 5.81x10$^{-04}$ | 0.02 | 1.09 | (1.04; 1.15) |
| rs941823 | 13 | 41,013,977 | MRPS31, FOXO1 | T | C | 0.27 | 2.47x10$^{-03}$ | 0.04 | 1.07 | (1.02; 1.12) |
| rs516246 | 19 | 49,206,172 | FUT2, MAMSTR | T | T | 0.54 | 4.71x10$^{-04}$ | 0.02 | 1.07 | (1.03; 1.11) |
| rs913678 | 20 | 48,955,424 | CEBPB, PTPN1 | C | T | 0.34 | 7.30x10$^{-06}$ | 1.02x10$^{-03}$ | 1.10 | (1.05; 1.14) |

## ABBREVIATIONS

CRC: colorectal cancer

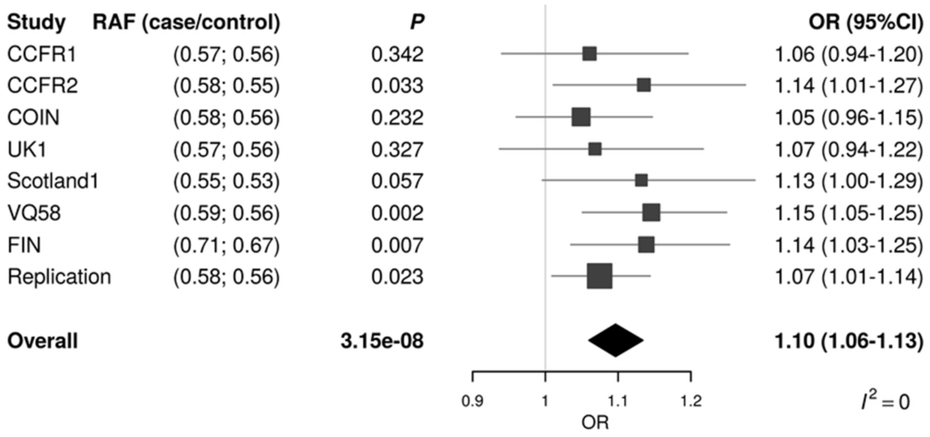
GWAS: genome-wide association studies

LD: linkage disequilibrium

IBD: inflammatory bowel disease

MSI: microsatellite instability

SNP: single-nucleotide polymorphism



| Study | RAF (case/control) | *P* | OR (95%CI) |
|---|---|---|---|
| CCFR1 | (0.57; 0.56) | 0.342 | 1.06 (0.94-1.20) |
| CCFR2 | (0.58; 0.55) | 0.033 | 1.14 (1.01-1.27) |
| COIN | (0.58; 0.56) | 0.232 | 1.05 (0.96-1.15) |
| UK1 | (0.57; 0.56) | 0.327 | 1.07 (0.94-1.22) |
| Scotland1 | (0.55; 0.53) | 0.057 | 1.13 (1.00-1.29) |
| VQ58 | (0.59; 0.56) | 0.002 | 1.15 (1.05-1.25) |
| FIN | (0.71; 0.67) | 0.007 | 1.14 (1.03-1.25) |
| Replication | (0.58; 0.56) | 0.023 | 1.07 (1.01-1.14) |
| **Overall** | | **3.15e-08** | **1.10 (1.06-1.13)** |

Figure 1 Forest plot of the odds ratios for the association between rs992157 and CRC. Studies were weighted according to the inverse of the variance of the log of the OR. Horizontal lines: 95% confidence intervals (95% CI). Box: OR point estimate; its area is proportional to the weight of the study. Diamond: overall summary estimate, with confidence interval given by its width. Vertical line: null value (OR = 1.0).
76x36mm (300 x 300 DPI)

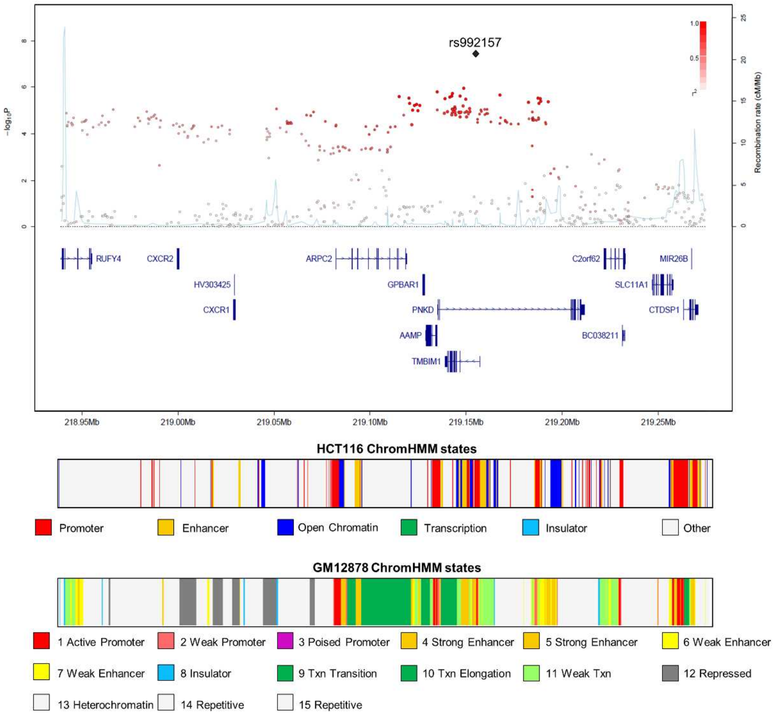


Figure 2 Regional plot of association results and recombination rates for the 2q35 locus. In the panel, −log10 P values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The top SNP is shown as a large triangle and is labelled by its rsID. The colour intensity of each symbol reflects the extent of LD with the top SNP: white (r2 = 0) through to dark red (r2 = 1.0), with r2 estimated from the 1000 Genomes Phase 1 data. Genetic recombination rates (cM/Mb), are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to each region of association. The lower panel shows the chromatin state segmentation track (ChromHMM) in HCT116 CRC and GM12878 lymphoblastoid cell lines.
254x190mm (300 x 300 DPI)

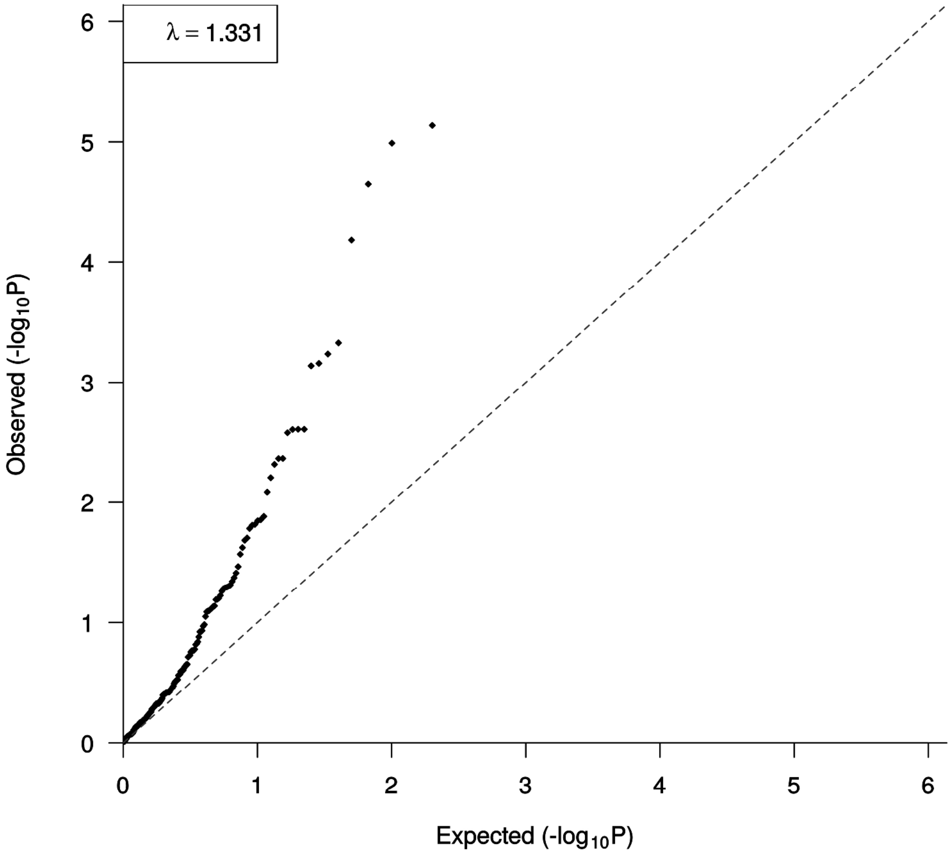


Figure 3 Quantile-Quantile (Q-Q) plot of observed and expected colorectal cancer association P-values for 200 inflammatory bowel disease risk SNPs

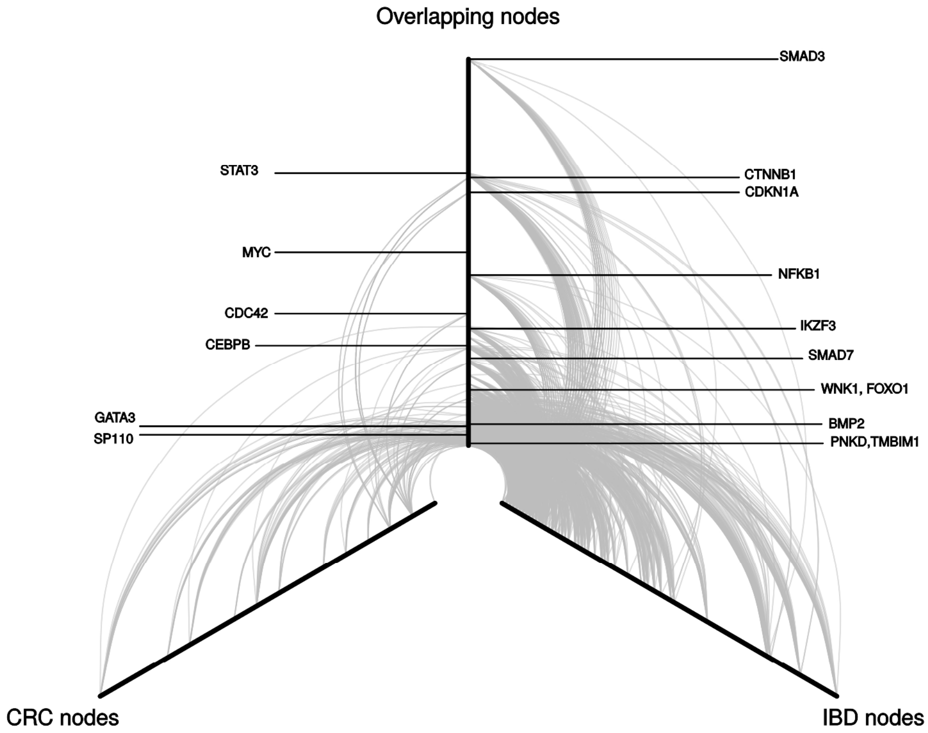


Figure 4 Hive Plot of common protein-protein interactions between colorectal cancer and inflammatory bowel disease defined by risk SNPs. Each arc represents an interaction between two proteins, and the distance from the centre of the plot corresponds to a greater number of protein-protein interactions (higher degree of the node). The left arm represents proteins that were only identified using the CRC SNPs, the right arm represents proteins that were only identified