Human Molecular Genetics, 2014, Vol. 23, No. 17 ***4729–4737***
doi:10.1093/hmg/ddu177
Advance Access published on April 15, 2014

Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis

Nicola Whiffin[1], Fay J. Hosking[1], Susan M. Farrington[2], Claire Palles[3], Sara E. Dobbins[1], Lina Zgaga[2], Amy Lloyd[1], Ben Kinnersley[1], Maggie Gorman[3], Albert Tenesa[4], Peter Broderick[1], Yufei Wang[1], Ella Barclay[3], Caroline Hayward[5], Lynn Martin[3], Daniel D. Buchanan[6], Aung Ko Win[7], John Hopper[7], Mark Jenkins[7], Noralane M. Lindor[8], Polly A. Newcomb[9], Steve Gallinger[10], David Conti[11], Fred Schumacher[11], Graham Casey[11], Tao Liu[12], The Swedish Low-Risk Colorectal Cancer Study Group, Harry Campbell[13], Annika Lindblom[12], Richard S. Houlston[1,*,†], Ian P. Tomlinson[4,*,†] and Malcolm G. Dunlop[2,*,†]

[1]Molecular and Population Genetics, Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK, [2]Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital Edinburgh, Crewe Road, Edinburgh EH4 2XU, UK, [3]Wellcome Trust Centre for Human Genetics, Oxford, UK, [4]The Roslin Institute, University of Edinburgh, Easter Bush, Roslin EH25 9RG, UK, [5]Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital Edinburgh, Crewe Road, Edinburgh, EH4 2XU, UK, [6]Cancer and Population Studies Group, Queensland Institute of Medical Research, Queensland, Australia, [7]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, Victoria, Australia, [8]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ, USA, [9]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [10]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON, Canada, [11]Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA, [12]Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden and [13]Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK

Received January 7, 2014; Revised April 3, 2014; Accepted April 10, 2014

To identify common variants influencing colorectal cancer (CRC) risk, we performed a meta-analysis of five genome-wide association studies, comprising 5626 cases and 7817 controls of European descent. We conducted replication of top ranked single nucleotide polymorphisms (SNPs) in additional series totalling 14 037 cases and 15 937 controls, identifying a new CRC risk locus at 10q24.2 [rs1035209; odds ratio (OR) = 1.13, $P = 4.54 \times 10^{-11}$]. We also performed meta-analysis of our studies, with previously published data, of several recently purported CRC risk loci. We failed to find convincing evidence for a previously reported genome-wide association at rs11903757 (2q32.3). Of the three additional loci for which evidence of an association in Europeans has been previously described we failed to show an association between rs59336 (12q24.21) and CRC risk. However, for the other two SNPs, our analyses demonstrated new, formally significant associations with CRC. These are rs3217810 intronic in *CCND2* (12p13.32; OR = 1.19, $P = 2.16 \times 10^{-10}$) and rs10911251 near *LAMC1* (1q25.3; OR = 1.09, $P = 1.75 \times 10^{-8}$). Additionally, we found some evidence to support a relationship between, rs647161, rs2423297 and rs10774214 and CRC risk originally identified in East Asians in our European datasets. Our findings provide further insights into the genetic and biological basis of inherited genetic susceptibility to CRC.

*To whom correspondence should be addressed: E-mails: richard.houlston@icr.ac.uk (R.H.); iant@well.ox.ac.uk (I.T.); malcolm.dunlop@igmm.ed.ac.uk (M.D.)
†These authors should be considered to have equal status.

INTRODUCTION

Many colorectal cancers (CRC) develop in genetically susceptible individuals, most of whom are not carriers of germ-line mismatch repair or *APC* mutations (1–3). Genome-wide association studies (GWAS) conducted over recent years have demonstrated that an appreciable part of the heritable risk of CRC is attributable to common, low-risk variants. These GWAS have also provided insights into the biological basis of CRC development, highlighting the potential role of genes within the bone morphogenetic protein signalling pathway (*BMP2, BMP4, GREM1* and SMAD7) (4,5) and some candidate genes (*e.g. CDH1/CDH3*), as well as genes not previously implicated in CRC (*e.g. POLD3, TERC, CDKN1A* and *SHROOM2*) (6,7).

The statistical power of individual GWAS has been limited by the modest effect sizes of genetic variants, the need to establish stringent thresholds of statistical significance and economic constraints on the number of variants that can be followed up. Meta-analysis of existing GWAS data therefore offer the opportunity to discover additional CRC risk loci and provide further insights into disease aetiology. In this study, we conducted a discovery meta-analysis of five European GWAS datasets, followed by validation in three independent case–control series, enabling us to identify a novel susceptibility locus for CRC. We also tested a set of recently reported risk SNPs, some with statistically significant associations and others with more limited evidence of a relationship with CRC. This analysis refutes two of these associations and demonstrates two new genome-wide genetic associations with CRC risk.

RESULTS

The discovery phase comprised analysis of five non-overlapping GWAS case–control series of Northern European ancestry, which have been previously reported (Supplementary Material, Table S1): UK1, 890 familial colorectal tumour cases and 900 cancer-free controls from the COloRectal Gene Identification (CORGI) consortium (7); Scotland1, 972 early-onset CRC cases (aged <55 years at diagnosis) and 998 population controls(7); VQ58, 1794 UK cases with stage B/C CRC from the VICTOR and QUASAR2 trials, together with publicly available data from 2686 population controls from the UK 1958 Birth Cohort (8); CCFR1, 1175 familial CRC cases and 999 controls from the Colon Cancer Family Registry (CCFR) (9); and CCFR2, 795 CRC cases from CCFR and 2234 controls from the Cancer Genetic Markers of Susceptibility studies of breast and prostate cancer (10,11). Samples were genotyped using proprietary Illumina SNP arrays: UK1 on Hap550; Scotland1 on Hap300 + Hap240S; VQ58 on Hap300, Hap370, Hap660 or Hap1M; CGEMS on Hap300 + Hap240 or Hap550; and CCFR samples using Hap1M, Human1M-Duo or Omni-express arrays. Quality control of genotyping was assessed as previously described and all SNPs presented in this study passed the predetermined thresholds (6).

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. λ_{GC} values (12) for the UK1, Scotland1, VQ58, CCFR1 and CCFR2 studies were 1.02, 1.01, 1.01, 1.02 and 1.03, respectively, thereby excluding significant differential genotyping or cryptic population substructure (Supplementary Material, Fig. S1). Any ethnic outliers or individuals identified as related were excluded (Supplementary Material, Fig. S2).

Discovery of a new CRC susceptibility SNP

Using data from the above five GWAS, we derived for each directly genotyped SNP joint odds ratios (ORs) and confidence intervals (CIs) under a fixed-effects model, and the associated *P*-values. In this meta-analysis, associations for all 20 established CRC risk SNPs showed a direction of effect consistent with previously reported studies, with eight of those loci having a *P*-value of $<5.0 \times 10^{-8}$ (Supplementary Material, Table S2). We also identified 31 SNPs that showed good evidence of an association (i.e. $P < 1.5 \times 10^{-4}$) and mapped to distinct loci that had not previously been associated with CRC risk. This threshold for follow-up does not exclude the possibility that other SNPs represented genuine association signals, but was simply a pragmatic strategy for prioritizing replication.

For these 31 SNPs we conducted a replication study using a custom array-based approach. We utilized three additional case–control series comprising a total of 14 037 cases and 15 937 controls: UK NSCCG (National Study of Colorectal Cancer Genetics) replication (UK3); Edinburgh replication (Scotland3); and Swedish replication (Sweden). Genotyping was successful for 29 of the 31 SNPs (Supplementary Material, Table S3). In the combined analysis of discovery and replication phases, one SNP, rs1035209, showed an association with CRC which was genome-wide significant ($P = 4.54 \times 10^{-11}$; Fig. 1).

rs1035209 localizes to chromosome 10q24.2 (101 345 366 bp; Fig. 2A). To comprehensively analyse the 10q24.2 association, we imputed unobserved genotypes in the region in GWAS cases and controls using 1000genomes data (Phase 1 integrated version 3, March 2012). We did not find substantive evidence of stronger associations than that provided by rs1035209 (Fig. 2A). The 320 kb region of linkage disequilibrium (LD) to which rs1035209 maps encompasses five transcripts, including *ABCC2/MRP2* (multidrug resistant protein 2). We examined for relationships between rs1035209 and gender, age at diagnosis, family history of CRC and clinico-pathological features (tumour site, stage or microsatellite instability) through case-only analysis, but no significant association was found (Supplementary Material, Table S4).

Evaluation of CRC SNPs reported in Asians

Jia *et al.* (13) have recently identified three SNPs; rs647161 (5q31.1), rs2423279 (20p12.3) and rs10774214 (12p13.32, near *CCND2*) that had genome-wide significant associations with CRC in East Asian populations. The same study evaluated these SNPs in a limited number of European CRC cases and controls, including the CCFR datasets, and found evidence of associations for all SNPs, albeit at relatively modest levels of significance ($P = 0.040$, $P = 0.002$ and $P = 0.001$ respectively). To test for association of these SNPs in our data we imputed all five GWAS datasets using the 1000 Genomes project as a reference panel and performed a meta-analysis. We found similar modest levels of significance ($P = 0.02$, $P = 4.57 \times 10^{-4}$ and $P = 0.02$ respectively; Supplementary Material, Table S5) to the original study. The nominal level of association at these loci may be

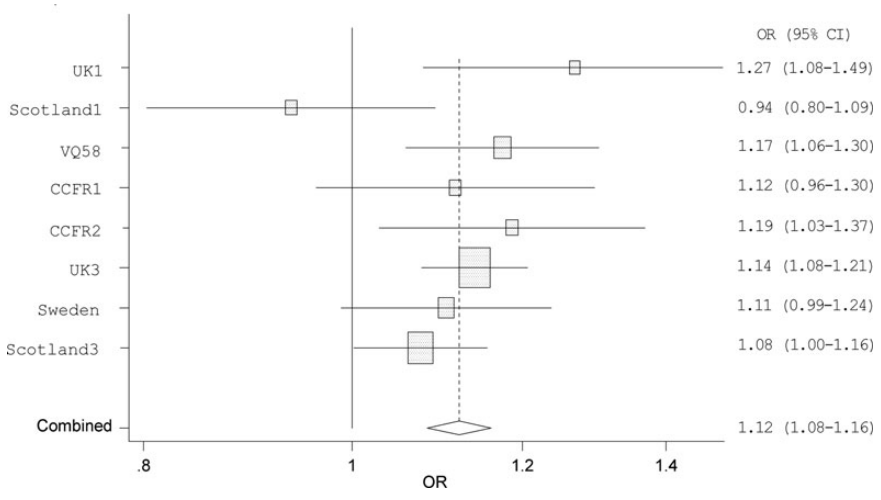

Figure 1. Forest plot of the ORs for the association between CRC and rs1035209. Studies were weighted according to the inverse of the variance of the log of the OR calculated by unconditional logistic regression. Horizontal lines: 95% CI. Box: OR point estimate; its area is proportional to the weight of the study. Diamond (and broken line): overall summary estimate, with CI given by its width. Unbroken vertical line: null value (OR = 1.0).

attributed to environmental differences or differences in risk allele frequency, specifically of rs647161, between Asians and Europeans (Supplementary Material, Table S5).

Evaluation of SNP rs11903757 previously found to be associated with CRC in europeans

Making use of imputed data, Peters *et al* (14) have recently reported a genome-wide significant association between rs11903757 at 2q32.3 and CRC risk in a combined analysis of European and Asian case–control series ($P = 3.7 \times 10^{-8}$; $P = 1.4 \times 10^{-6}$ in Europeans). Despite our overlapping datasets, with both studies including CCFR, we did not find any evidence in our samples to support a relationship between rs11903757 and CRC ($P = 0.90$). In a combined analysis of data from both our study and that of Peters *et al* (14), this SNP association was not significant at the genome-wide threshold ($P = 1.89 \times 10^{-4}$ Supplementary Material, Fig. S3).

Identification of two further CRC predisposition SNPs in combined analysis of our data and published data

The study by Peters *et al.* (14) also reported non-significant, but promising, associations ($5.0 \times 10^{-7} > P > 5.0 \times 10^{-8}$) at rs3217810 (12p13.32), rs59336 (12q24.21) and rs10911251 (1q25.3). Intriguingly, rs3217810 is intronic in the *CCND2* gene close to the Asian CRC SNP rs10774214, but these two SNPs are essentially uncorrelated ($r^2 = 0.002$, $D' = 0.092$). We examined the robustness of the three promising associations in a meta-analysis of our data and the published data. This analysis showed rs3217810 and rs10911251 to be CRC risk SNPs, with both associations now attaining genome-wide significance ($P = 2.16 \times 10^{-10}$ and $P = 1.75 \times 10^{-8}$ respectively; Fig. 2B and C and Fig. 3). In contrast we did not find any evidence to support a relationship between rs59336 and CRC ($P = 0.17$; Supplementary Material, Table S5).

eQTL analysis of the three new CRC SNPs

To gain insight into the biological basis of the associations at rs1035209, rs3217810 and rs10911251, we analyzed publicly available mRNA expression data from fibroblasts, lymphoblastoid cell lines (LCLs), T cells, adipose tissue and skin cells (15,16). Additionally we interrogated The Cancer Genome Atlas (TCGA) RNA-seq expression and Affymetrix 6.0 SNP data (dbGaP accession number: phs000178.v7.p6) on 223 colonic and 75 rectal cancers using rs7075305 ($r^2 = 0.30$, $D' = 0.77$), rs3217805 ($r^2 = 0.19$, $D' = 0.87$) and rs3768617 ($r^2 = 0.90$, $D' = 1.00$) as proxies for rs1035209, rs3217810 and rs10911251, respectively. After adjustment for multiple testing, no significant associations were seen between SNP genotype and expression of genes mapping to any of the three risk loci (Supplementary Material, Tables S6 and S7).

Using HaploReg (17) and RegulomeDB (18), we examined whether any of the SNPs or their proxies (i.e. $r^2 > 0.8$ in 1000 Genomes CEU reference panel) lie at putative transcription factor binding/enhancer elements and derived GERP (Genomic Evolutionary Rate Profiling) scores to asses sequence conservation at these positions (Supplementary Material, Table S8). rs3217810 is evolutionarily conserved (GERP score 2.97) with the motif being shown to have GATA6 binding in the CRC cell line CaCo2. Intriguingly, GATA6 expression is elevated in CRC and has been linked to invasiveness (19,20). While rs10911251 is not evolutionarily conserved, the correlated SNPs rs10911205 ($r^2 = 0.81$) and rs10911211 ($r^2 = 0.83$) are conserved (GREP scores 2.95 and 3.17, respectively), and are associated with transcription factor binding, albeit weakly. In addition, rs10911205 lies within an enhancer region predicted by ChromHMM (Supplementary Material, Table S8).

Finally, we conducted pathway analysis to determine whether any genes mapping to the three newly identified regions act in pathways already over-represented in GWAS regions. All genes within the LD block containing each tagSNP, or linked to the SNP through functional experiments, were uploaded to the NCI pathway interaction database. Pathways containing three or more genes are shown in Supplementary Material,

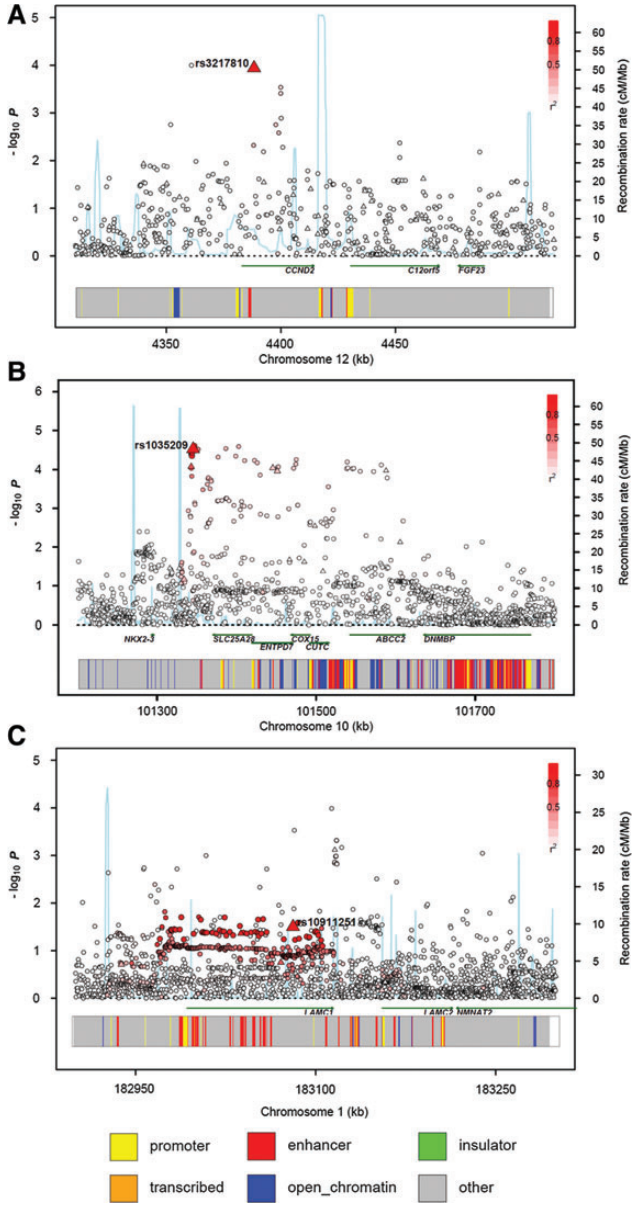

Figure 2. Regional plots of association results, recombination rates and chromatin state segmentation track for (A) 10q24.2, (B) 12p13.32 and (C) 1q25.3 susceptibility loci. Association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates for rates. $-\log_{10}$ *P*-values (*y* axis) of the SNPs are shown according to their chromosomal positions (*x* axis). The top genotyped SNP in each combined analysis is shown as a large triangle and is labelled by its rsID. Colour intensity of each symbol reflects the extent of LD with the top genotyped SNP; white ($r^2 = 0$) through to dark red ($r^2 = 1.0$) Genetic recombination rates, estimated using HapMap Utah residents of Western and Northern European ancestry (CEU) samples, 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 the region of association. Genes have been redrawn to show the relative positions; therefore, maps are not to physical scale. The lower panel shows the chromatin state segmentation track (ChromHMM).

Table S9. While this analysis identifies the BMP signalling pathway as expected, the analysis also reveals multiple pathways involving the LAMC1 and LAMC2 genes near rs10911251 and the regulation of nuclear beta catenin signalling pathway implicating *CCND2* to which rs3217810 maps.

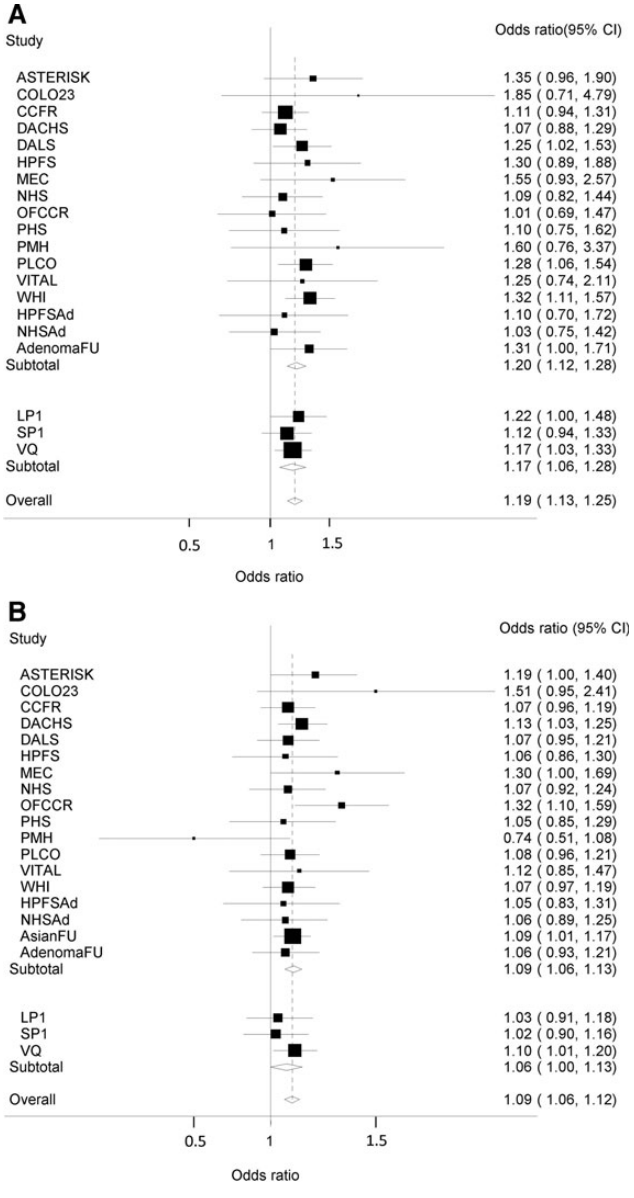

Figure 3. Forest plot of the ORs for associations between CRC and (A) rs3217810 and (B) rs10911251. Studies were weighted according to the inverse of the variance of the log of the OR calculated by unconditional logistic regression. Horizontal lines: 95% CI. Box: OR point estimate; its area is proportional to the weight of the study. Diamond (and broken line): overall summary estimate, with CI given by its width. Unbroken vertical line: null value (OR = 1.0). Summary estimates are shown for studies from Peters *et al.*, from the three GWAS not included in that study and a combined estimate.

DISCUSSION

Here, we have identified a novel CRC susceptibility SNP rs1035209 at 10q24.2. By performing a combined analyses with published data, we have also been able to show that two promising, but formally non-significant SNP associations, rs3217810 (12p13.32) and rs10911251 (1q25.3), are indeed associated with CRC risk in Europeans at genome-wide levels of significance. Additionally, we have provided evidence that a previously reported association between CRC and SNP rs11903757 does not hold in Europeans. Finally our

data lend some support to the association of three other SNPs, originally identified in East Asians, with CRC in European populations.

The relationship between genetic variation near *CCND2* (12p13.32) and risk of CRC was first reported in East Asians (13), with low levels of support in Europeans. Hence, the observation that variation at this locus also impacts on risk in Europeans at genome-wide significance thresholds provides evidence for the generalizability of this association. Additional studies are required to decipher the number of independent signals of association and the functional basis of the 12p13.32 associations in different ethnic groups. We note that rs3217810 is located within an intron of *CCND2* (Fig. 2). Cyclin D2 is a member of the D-type cyclin family which plays a critical role in cell cycle control through activation of cyclin-dependent kinases (CDK), primarily CDK4 and CDK6 (21). Additionally CCND2 has been shown to represses the transcriptional activity of *SMAD3*. While less well studied than *CCND1*, over-expression of CCND2 typifies a subset of CRC and has been reported to be an independent predictor of survival (22–24).

The 1q25.3 association implicates variation in the gene encoding the extracellular matrix protein laminin gamma 1 (*LAMC1*) which is involved in the maintenance of cell adhesion, migration and signalling (25–27). *LAMC1* is mutated in ∼8% of CRC and differential expression linked to development of metastatic CRC (http://www.cbioportal.org/public-portal/). We have previously shown that variation in *LAMA5* is a determinant of CRC risk (7) and our new observation implicates laminin genes more generally in the aetiology of CRC. Pathway analysis reveals integrin signalling pathways containing both LAMC1 and LAMA5 genes as over-represented in GWAS regions.

The susceptibility locus at 10q24.2 marked by rs1035209 encompasses a number of genes with a possible role in the development of CRC including *ABCC2/MRP2* a multi-drug resistance gene influencing responsiveness to anticancer drugs (28) which is mutated in ∼7% of CRC (http://www.cbioportal.org/public-portal/).

Since the marker SNPs at each of the risk loci are not strongly correlated with nsSNPs, it is likely that the functional variants at each site take the form of non-coding changes influencing gene expression. Accepting the caveat that our eQTL analysis of CRC was, for rs1035209 and rs3217810, dependent upon weak proxies we found no evidence for allele-specific *cis*-regulation of neighbouring genes. It is however likely that the potential impact of common alleles on gene expression will be modest and could occur at any time before diagnosis of CRC. Moreover, expression differences may only be relevant to a subpopulation of cells that provide 'targets' for tumour mutations.

Our failure to find any evidence to support associations marked by rs11903757 and rs59336, or any proxy SNPs, was despite the CCFR GWAS datasets, which we included in our meta-analysis, previously contributing to the analysis by Peters *et al.* (14). Both our study and that of Peters relied on imputation to recover rs11903757 and rs59336 genotypes. Hence, the imputation fidelity is of paramount importance. Through sequencing, we were able to establish that imputed genotypes were accurately recovered in our study. The study of Peters did not include this quality control step, moreover, imputation used as reference HapMap data derived from 30 trios, rather than the >1000 individuals catalogued by the 1000 Genomes Project. Overall, the data suggest that the putative association with CRC at this locus may be spurious.

The power of our study to detect the major common loci conferring risks of 1.2 or greater (such as the 8q24 variant rs6983267) was high. Hence, there are unlikely to be many additional CRC SNPs with similar effects for alleles with frequencies >0.2 in populations of European ancestry. However, estimates of the missing heritability suggest that further variants of smaller effect sizes remain to be discovered. To discover these additional common CRC risk variants, there is a requirement for even larger scale GWAS meta-analyses in terms of both sample size and SNP coverage, as well as an increase in the number of SNPs taken forward to large-scale replication.

In conclusion, in this large study, we have identified a novel susceptibility locus associated with the risk of colorectal at 10q24.2 and performed combined analyses with published data to identify further associations at 12p13.32 (*CCND2*) and 1q25.3 (*LAMC1*). These findings bring to 23 the number of independent loci which influence CRC risk in Europeans.

MATERIALS AND METHODS

Ethics statement

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. Specifically: for CORGI REC 06/Q1702/99; for SOCCS REC 11/SS/0109; for NSCCG REC 02/0/97.

Subjects and datasets

UK1 (CORGI) (7) comprised 940 cases with colorectal neoplasia (47% male) ascertained through the Colorectal Tumour Gene Identification (CoRGI) consortium. All had at least one first-degree relative affected by CRC and one or more of the following phenotypes: CRC at age 75 or less; any colorectal adenoma (CRAd) at age 45 or less; ≥3 colorectal adenomas at age 75 or less; or a large (>1 cm diameter) or aggressive (villous and/or severely dysplastic) adenoma at age 75 or less. The 965 controls (45% males) were spouses or partners unaffected by cancer and without a personal family history (to second degree relative level) of colorectal neoplasia. Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic MUTYH mutation carriers were excluded. All cases and controls were of white UK ethnic origin.

Scotland1 (COGS) (7) included 1012 CRC cases (51% male; mean age at diagnosis 49.6 years, SD ± 6.1) and 1012 cancer-free population controls (51% male; mean age 51.0 years; SD ± 5.9). Cases were selected for early age at onset (age ≤55 years). Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic MUTYH mutation carriers were excluded. Control subjects were sampled from the Scottish population NHS registers, matched by age (± 5 years), gender and area of residence within Scotland.

VQ58 comprised 1800 CRC cases (1099 males, mean age of diagnosis 62.5 years; SD ± 10.9) from the VICTOR (29) and QUASAR2 (www.octo-oxford.org.uk/alltrials/trials/q2.html) trials. There were 2690 population control genotypes (1391

males) from the Wellcome Trust Case–Control Consortium 2 (WTCCC2) 1958 birth cohort (8) (also known as the National Child Development Study), which included all births in England, Wales and Scotland during a single week in 1958.

The CCFR1 data set comprised 1290 familial CRC cases and 1055 controls from the Colon Cancer Family Registry (http://epi.grants.cancer.gov/CFR/about_colon.html) (9). The cases were recently diagnosed CRC cases reported to population complete cancer registries in the USA (Puget Sound, Washington State) who were recruited by the Seattle Familial Colorectal Cancer Registry; in Canada (Ontario) who were recruited by the Ontario Familial Cancer Registry; and in Australia (Melbourne, Victoria) who were recruited by the Australasian Colorectal Cancer Family Study. Controls were population-based and for this analysis were restricted to those without a family history of colorectal cancer. The CCFR2 data set comprised a further 796 cases from the Colon Cancer Family Registry and 2236 controls from the Cancer Genetic Markers of Susceptibility studies of breast ($n = 1142$) and prostate ($n = 1094$) cancer (10,11).

UK3 (NSCCG) (7) comprised 7562 CRC cases (59% male; mean age at diagnosis 58 years, SD $\pm$ 8.4) and 7430 controls (39% male; mean age 58 years, SD $\pm$ 10.8) ascertained through NSCCG (National Study of Colorectal Cancer Genetics) post-2005 (30).

Scotland3 comprised 3969 CRC cases recruited as part of the SOCCS/COGS (7) studies and 6791 population controls. Controls comprised unrelated cancer-free participants from Generation Scotland (http://www.generationscotland.org) (31).

The Swedish study comprised 2506 CRC patients were recruited within a Swedish national study conducted by the Swedish Low-Risk Colorectal Cancer Study Group. Samples were obtained during 2004–2009 from 14 different surgical clinics in central Sweden. All CRC patients during the study period were eligible for recruitment and were invited to participate. Only those too ill or too frail to consent were excluded. Controls ($n = 1716$) comprised blood donors from Stockholm and Uppsala. Fully informed consent was obtained in accordance with the Swedish law concerning ethical approval of research on human subjects (refs: 2002:489,2003:198,2010:1213-31/4).

In all cases CRC was defined according to the ninth revision of the International Classification of Diseases (ICD) by codes 153–154.

Sample preparation and genotyping

DNA was extracted from samples using conventional methods and quantified using PicoGreen (Invitrogen). The VQ, UK1 and Scotland1 GWA cohorts were genotyped using Illumina Hap300, Hap240S, Hap370 or Hap550 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. Genotyping quality and validity was very high in GWA datasets, as demonstrated by genotyping duplicate samples on orthogonal platforms which revealed 99.9% genotype concordance for all samples tested (data not shown). The replication samples were typed using custom 32 SNP genotyping plates Openarray (TaqMan® OpenArray®). The SNP content was designed based on SNP associations from analysis of GWA data, including imputed data. After technical assay failures, the final replication array comprised 29 SNPs (Supplementary Material, Table S3). Additional genotyping was conducted using competitive allele-specific PCR KASPar chemistry (LCG, Hertfordshire, UK). All primers, probes and conditions used are available on request. Genotypes for the Scottish control replication dataset were SNPs genotyped on the Illumina OmniExpressExome array (only SNPs that genotyped were included, imputed data were not used). Genotyping quality control was tested using duplicate DNA samples within studies and SNP assays, together with direct sequencing of subsets of samples to confirm genotyping accuracy. For all SNPs, >99% concordant results were obtained.

Quality control and sample exclusion

We excluded SNPs from analysis if they failed one or more of the following thresholds: GenCall scores <0.25; overall call rates <95%; minor allele frequency (MAF) < 0.01; departure from Hardy–Weinberg equilibrium (HWE) in controls at $P < 10^{-4}$ or in cases at $P < 10^{-6}$; outlying in terms of signal intensity or X:Y ratio; discordance between duplicate samples; and, for SNPs with evidence of association, poor clustering on inspection of X:Y plots. We excluded samples from analysis if they failed one or more of the following thresholds: duplication or cryptic relatedness to estimated identity by descent (IBD) >6.25%; overall successfully genotyped SNPs < 95%; mismatch between predicted and reported gender; outliers in a plot of heterozygosity versus missingness; and evidence of non-white European ancestry by principal component analysis (PCA)-based analysis in comparison with HapMap samples (http://hapmap.ncbi.nlm.nih.gov). Details of all sample exclusions are provided in Supplementary Material, Figure S4.

To identify individuals who might have non-northern European ancestry, we merged our case and control data from all sample sets with the 60 European (CEU), 60 Nigerian (YRI), 90 Japanese (JPT) and 90 Han Chinese (CHB) individuals from the International HapMap Project. For each pair of individuals, we calculated genome-wide identity-by-state distances based on markers shared between HapMap2 and our SNP panel, and used these as dissimilarity measures upon which to perform PCA. PCA was performed in R. The first two principal components for each individual were plotted and any individual not present in the main CEU cluster (i.e. >5% of the PC distance from HapMap CEU cluster centroid) was excluded from subsequent analyses.

We have previously shown the adequacy of the case–control matching and possibility of differential genotyping of cases and controls using Q–Q plots of test statistics. The inflation factor λ_{GC} was calculated before and after imputation by dividing the mean of the lower 90% of the test statistics by the mean of the lower 90% of the expected values from a χ^2 distribution with 1 d.f. Only SNPs with minor allele frequency >0.05, imputation INFO > 0.4, P-heterogeneity > 0.01 and P-HWE > 0.01 were considered. In addition, calculations were made using the median values of the observed and expected χ^2 distribution but no significant differences were seen. Deviation of the genotype frequencies in the controls from those expected under HWE was assessed by χ^2 test (1 d.f.), or Fisher's exact test where an expected cell count was <5.

Statistical and bioinformatic analysis

Main analyses were undertaken using R (v2.6), Stata v.11 (College Station, TX, US) and PLINK (v1.06) software (32). The association between each SNP and risk of CRC was assessed by the Cochran–Armitage trend test. ORs and associated 95% CIs were calculated by unconditional logistic regression. Meta-analysis was conducted using standard methods (33). GWAS meta-analysis data is available from EGA (accession number EGAS00001000759). Cochran's Q statistic to test for heterogeneity (33) and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated (34). I^2 values $\geq$75% are considered characteristic of large heterogeneity (34–36). Associations by sex, age and clinic-pathological phenotypes were examined by logistic regression in case-only analyses.

Prediction of the untyped SNPs was carried out using IMPUTEv2, based on 1000genomes Phase 1 integrated version 3 reference data. Imputed data were analyzed using SNPTESTv2.3.0 to account for uncertainties in SNP prediction and meta-analysis was performed using METAv1.4 with an information score threshold of 0.4. To filter poorly imputed SNPs, as previously recommended, we excluded variants having overall information scores from SNPTESTv2.3.0 of $<$0.4. LD metrics were calculated from 1000genomes pilot release I data and viewed using SNAP. Where SNPs had not been catalogued, LD metrics were calculated using in house Perl scripts using the 1000genomes Phase 1 data. Regional association plots of LD metrics were then plotted using SNAP. LD blocks were defined on the basis of HapMap recombination rate and were viewed using the Haploview software (v4.2).

Accuracy of imputation was assessed using sequence data of 200 CRC cases. The genotypes of the SNPs which passed QC in the UK1 dataset were extracted and used for imputation of each sample. The concordance between imputed SNPs and those obtained from sequencing was calculated for all individuals and for those individuals that are heterozygous or homozygous for the rare allele in the sequencing data. rs3217810, rs10911251, rs59336 and rs11903757 showed strong correlation between imputation and sequencing with r^2 of 1.00, 0.99, 0.95 and 1.00 over all SNPs, respectively (Supplementary Material, Table S10).

To explore epigenetic profiles of association signals, we used ChromHMM (37). States were inferred from ENCODE Histone Modification data on the CRC cell line Hct116 (DNAse, H3K4me3, H3K4me1, H3K27ac, Pol2 and CTCF) binarized using a multivariate Hidden Markov Model.

We made use of HaploReg (17) and RegulomeDB (18) to examine whether any of the SNPs or their proxies (i.e. $r^2 >$ 0.8 in 1000genomes CEU reference panel) annotate putative transcription factor binding/enhancer elements. We assessed sequence conservation using Genomic Evolutionary Rate Profiling (GERP). GERP scores (-12 to 6, with 6 being indicative of complete conservation) reflect the proportion of substitutions at that site rejected by selection compared with observed substitutions expected under a neutral evolutionary model, based on sequence alignment of 34 mammalian species (38). We also analyzed expression data generated from (i) fibroblasts, lymphoblastoid cell lines (LCLs) and T cells derived from the umbilical cords of 75 Geneva GenCord individuals (15); (ii) 166 adipose, 156 LCL and 160 skin samples derived from a subset of healthy female twins of the MuTHER resource (16) using Sentrix Human-6 Expression BeadChips (Illumina) (39,40).

Relationship between SNP genotype and mRNA expression

To examine for a relationship between SNP genotype and mRNA expression we made use of Tumor Cancer Genome Atlas (TCGA) RNA-seq expression and Affymetrix 6.0 SNP data (dbGaP accession number: phs000178.v7.p6) on 223 colorectal adenocarcinoma (COAD) and 75 rectal adenocarcinoma tumour samples using a best proxy where SNPs were not represented directly. Association between normalized RNA counts per-gene and SNP genotype was quantified using the Kruskal–Wallis trend test.

Pathway analysis

To determine whether any genes mapping to the three newly identified regions act in pathways already over-represented in GWAS regions we utilized the NCI pathway interaction database (http://pid.nci.nih.gov/index.shtml). All genes within the LD block containing each tagSNP, or linked to the SNP through functional experiments (MYC) were submitted as a Batch query using the NCI-Nature curated data source.

Assignment of microsatellite instability (MSI) in colorectal cancers

Tumour MSI status in CRCs was determined using the mononucleotide microsatellite loci BAT25 and BAT26, which are highly sensitive MSI markers. Briefly, 10 μm sections were cut from formalin-fixed paraffin-embedded CRC tumours, lightly stained with toluidine blue and regions containing at least 60% tumour microdissected. Tumour DNA was extracted using the QIAamp DNA Mini kit (Qiagen, Crawley, UK) according to the manufacturer's instructions and genotyped for the BAT25 and BAT26 loci using ^{32}P-labelled oligonucleotide primers. Samples showing more than or equal to five novel alleles, when compared with normal DNA, at either or both markers were assigned as MSI-H (corresponding to MSI-high) (41).

URLS

The R suite can be found at http://www.r-project.org.
Detailed information on the tag SNP panel can be found at http://www.illumina.com.
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP.
HapMap: http://www.hapmap.org.
1000Genomes: http://www.1000genomes.org.
SNAP http://www.broadinstitute.org/mpg/snap.
IMPUTE: https://mathgen.stats.ox.ac.uk/impute/impute.html.
SNPTEST: http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html.
Cancer Genome Atlas project: http://cancergenome.nih.gov.
The ENCODE Project: ENCyclopedia Of DNA Elements: http://www.genome.gov.
HaploReg: http://www.broadinstitute.org/mammals/haploreg/haploreg.php.
RegulomeDB: http://regulome.stanford.edu/.

AUTHORS' CONTRIBUTIONS

The study was designed and financial support was obtained by R.S.H., I.P.M.T. and M.G.D. The manuscript was drafted by R.S.H., N.W., I.P.M.T and M.G.D. All authors had access to data, analysis and had opportunity to contribute to drafting the manuscript. Statistical and bioinformatic analyses were conducted by N.W. and F.H., with contributions from S.D., Y.W. and B.K. ICR and local collaborators: coordination of sample preparation and genotyping was performed by P.B. Sample preparation and genotyping were performed by A.L. and N.W. Oxford and local collaborators: subject recruitment and sample acquisition were done by E.B., M.G., L.M. and members of the CORGI Consortium. Sample preparation and genotyping were performed by C.P. Colon Cancer Genetics Group, Edinburgh and local collaborators: subject recruitment and sample acquisition were performed by S.M.F., C.H., H.C., I.D. and M.G.D., as well as members of SOCCS and COGS recruitment teams. Sample preparation was coordinated by S.M.F.. Genotyping was performed and coordinated by S.M.F., C.H. and M.G.D. Data curation and analysis was conducted by L.Z., S.M.F. and A.T. Recruitment sample preps, wet lab expression analysis and genotyping was performed by C.H. and S.M.F. M.T. L.Z. performed the bioinformatic analyses.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

Cancer Research UK provided principal funding for this study to R.S.H., I.P.M.T. and M.G.D. In addition, this work was supported by the European Union (FP7/207-2013) under grant 258236, FP7 collaborative project SYSCOL and COST Action BM1206.

At the Institute of Cancer Research, the work was supported by a Programme Grant from 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. N.W. and B.K. were in receipt of PhD studentships from the ICR. B.K. additionally receives funding from the Sir John Fisher Foundation.

In Edinburgh the work was supported by Programme Grant funding from Cancer Research UK (C348/A12076). We are most grateful to all participants of the SOCCS/COGS studies and Generation Scotland, without whom our research could not be conducted. We acknowledge the excellent technical support from CCGG technicians and the COGS/SOCCS operational team: study coordinators, recruitment nurses, data entry and collation. Generation Scotland is a collaboration between the University Medical Schools and NHS in Aberdeen, Dundee, Edinburgh and Glasgow. The Generation Scotland Family Health Study (http://www.generationscotland.org) was funded by a grant from the Scottish Government Health Department, Chief Scientist Office, CZD/16/6. Genotyping was funded by the UK Medical Research Council. We are grateful to GPs and Scottish School of Primary Care for help with recruitment to GS:SFHS. We acknowledge the excellent work of the whole GS team: academic researchers, clinic staff, laboratory technicians, clerical workers, statisticians and research managers. We acknowledge the expert support on sample preparation and genotyping by the Genetics Core of the Edinburgh University Wellcome Trust Clinical Research Facility.

In Oxford additional funding was provided by the Oxford Comprehensive Biomedical Research Centre (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).

The Swedish sample and data resource was funded by the Swedish Cancer Society, the Swedish Scientific Research Council and the Stockholm Cancer Foundation. We acknowledge the contribution to recruitment and data collection of the Swedish Low-Risk Colorectal Cancer Study Group.

The work of the Colon Cancer Family Registry CFR was supported by the National Cancer Institute, National Institutes of Health under RFA #CA-95-011 and through cooperative agreements with members of the Colon CFR and Principal Investigators. Collaborating centres include the Australasian Colorectal Cancer Family Registry (U01 CA097735), the USC Familial Colorectal Neoplasia Collaborative Group (U01 CA074799), Mayo Clinic Cooperative Familial Registry for Colon Cancer Studies (U01 CA074800), Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783) and the Seattle Colorectal Cancer Family Registry (U01 CA074794). The Colon CFR GWAS was supported by funding from the National Cancer Institute, National Institutes of Health (U01CA122839 to GC).

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/. Finally, we would like to thank all individuals who participated in the study. 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 who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Conflict of Interest statement. The authors declare no competing financial interests.

REFERENCES

1. Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. and Hemminki, K. (2000) Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.*, **343**, 78–85.
2. Aaltonen, L., Johns, L., Jarvinen, H., Mecklin, J.P. and Houlston, R. (2007) Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clin. Cancer Res.*, **13**, 356–361.
3. Lubbe, S.J., Webb, E.L., Chandler, I.P. and Houlston, R.S. (2009) Implications of familial colorectal cancer risk profiles and microsatellite instability status. *J. Clin. Oncol.*, **27**, 2238–2244.
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 Genet.*, **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. Power, C. and Elliott, J. (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.*, **35**, 34–41.
9. 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.
10. 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.
11. 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.
12. 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.
13. Jia, W.H., Zhang, B., Matsuo, K., Shin, A., Xiang, Y.B., Jee, S.H., Kim, D.H., Ren, Z., Cai, Q., Long, J. *et al.* (2013) Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat. Genet.*, **45**, 191–196.
14. 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.
15. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.
16. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K. *et al.* (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.*, **7**, e1002003.
17. 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–D934.
18. 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 Res.*, **22**, 1790–1797.
19. Belaguli, N.S., Aftab, M., Rigi, M., Zhang, M., Albo, D. and Berger, D.H. (2010) GATA6 promotes colon cancer cell invasion by regulating urokinase plasminogen activator gene expression. *Neoplasia*, **12**, 856–865.
20. Shureiqi, I., Zuo, X., Broaddus, R., Wu, Y., Guan, B., Morris, J.S. and Lippman, S.M. (2007) The transcription factor GATA-6 is overexpressed in vivo and contributes to silencing 15-LOX-1 *in vitro* in human colon cancer. *FASEB J.*, **21**, 743–753.
21. Matsushime, H., Quelle, D.E., Shurtleff, S.A., Shibuya, M., Sherr, C.J. and Kato, J.Y. (1994) D-type cyclin-dependent kinase activity in mammalian cells. *Mol. Cell. Biol.*, **14**, 2066–2076.
22. Musgrove, E.A., Caldon, C.E., Barraclough, J., Stone, A. and Sutherland, R.L. (2011) Cyclin D as a therapeutic target in cancer. *Nat. Rev. Cancer*, **11**, 558–572.
23. Mermelshtein, A., Gerson, A., Walfisch, S., Delgado, B., Shechter-Maor, G., Delgado, J., Fich, A. and Gheber, L. (2005) Expression of D-type cyclins in colon cancer and in cell lines from colon carcinomas. *Br. J. Cancer*, **93**, 338–345.
24. Sarkar, R., Hunter, I.A., Rajaganeshan, R., Perry, S.L., Guillou, P. and Jayne, D.G. (2010) Expression of cyclin D2 is an independent predictor of the development of hepatic metastasis in colorectal cancer. *Colorectal Dis.*, **12**, 316–323.
25. Turck, N., Gross, I., Gendry, P., Stutzmann, J., Freund, J.N., Kedinger, M., Simon-Assmann, P. and Launay, J.F. (2005) Laminin isoforms: biological roles and effects on the intracellular distribution of nuclear proteins in intestinal epithelial cells. *Exp. Cell Res.*, **303**, 494–503.
26. Pouliot, N., Saunders, N.A. and Kaur, P. (2002) Laminin 10/11: an alternative adhesive ligand for epidermal keratinocytes with a functional role in promoting proliferation and migration. *Exp. Dermatol.*, **11**, 387–397.
27. Patarroyo, M., Tryggvason, K. and Virtanen, I. (2002) Laminin isoforms in tumor invasion, angiogenesis and metastasis. *Semin. Cancer Biol.*, **12**, 197–207.
28. Taniguchi, K., Wada, M., Kohno, K., Nakamura, T., Kawabe, T., Kawakami, M., Kagotani, K., Okumura, K., Akiyama, S. and Kuwano, M. (1996) A human canalicular multispecific organic anion transporter (cMOAT) gene is overexpressed in cisplatin-resistant human cancer cell lines with decreased drug accumulation. *Cancer Res.*, **56**, 4124–4129.
29. 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.
30. 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.
31. Smith, B.H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S.M., Deary, I.J., Macintyre, D.J., Campbell, H., McGilchrist, M. *et al.* (2013) Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int. J. Epidemiol.*, **42**, 689–700.
32. 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. *Am. J. Hum. Genet.*, **81**, 559–575.
33. Petitti, DB. (1994) *Meta-Analysis Decision Analysis and Cost-Effectiveness Analysis*. Oxford University Press, New York.
34. Higgins, J.P. and Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Stat. Med.*, **21**, 1539–1558.
35. Ioannidis, J.P., Ntzani, E.E. and Trikalinos, T.A. (2004) 'Racial' differences in genetic effects for complex diseases. *Nat. Genet.*, **36**, 1312–1318.
36. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
37. Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
38. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
39. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
40. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.
41. 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 Res.*, **58**, 5248–5257.
42. Tomlinson, I.P., Dunlop, M., Campbell, H., Zanke, B., Gallinger, S., Hudson, T., Koessler, T., Pharoah, P.D., Niittymaki, I., Tuupanen, S. *et al.* (2010) COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer. *Br. J. Cancer*, **102**, 447–454.