Published in final edited form as:
Nat Genet. 2010 November ; 42(11): 973–977. doi:10.1038/ng.670.

Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33

Richard S Houlston[*,1], **Jeremy Cheadle**[2], **Sara E Dobbins**[1], **Albert Tenesa**[3], **Angela M Jones**[4], **Kimberley Howarth**[4], **Sarah L Spain**[4], **Peter Broderick**[1], **Enric Domingo**[4], **Susan**

[*]Corresponding authors: Richard Houlston, Institute of Cancer Research, 15, Cotswold Rd, Sutton, Surrey SM2 5NG, UK. Tel: +44-(0)-208-722-4175. Fax: +44-(0)-208-722-4359. richard.houlston@icr.ac.uk. Malcolm Dunlop, Colon Cancer Genetics Group, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, UK. Tel: +44 (0)-131-467-8454. Fax: +44 (0)-131-467-8450. malcolm.dunlop@hgu.mrc.ac.uk. Ian Tomlinson, Molecular and Population Genetics, Nuffield Dept. of Medicine, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN. Tel: +44 (0)-1865-287832. Fax: +44 (0)-1865-287501. iant@well.ox.ac.uk.
[16]A full list of members is provided in the Supplementary Information

URLs

Detailed information on the tag SNP panel can be found at http://www.illumina.com/
Haploview: http://www.broadinstitute.org/haploview/haploview
VICTOR, QUASAR2: http://www.octo-oxford.org.uk/
PLINK: http://pngu.mgh.harvard.edu/Purcell/plink/
dbSNP: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=snp
HAPMAP: http://www.hapmap.org/
KBioscience: http://kbioscience.co.uk/
STATA: http://www.stata.com/
GELCAPS: http://pfsearch.ukcrn.org.uk/StudyDetail.aspx?TopicID=1&StudyID=781
- http://www.dh.gov.uk/assetRoot/04/01/45/13/04014513.pdf
National Study of Colorectal Cancer Genetics (NSCCG): http://pfsearch.ukcrn.org.uk/StudyDetail.aspx?TopicID=1&StudyID=1269
ICR-RMH Family history and DNA resource: http://intratest.icr.ac.uk/tissueres/index.htm
1958 Birth Cohort: http://www.b58cgene.sgul.ac.uk/
WTCCC2: http://www.wtccc.org.uk/ccc2/wtccc2_studies.shtml
Genetic Power Calculator: http://pngu.mgh.harvard.edu/~purcell/gpc/
IMPUTE v2: (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html),
Eigenstrat: (http://genepath.med.harvard.edu/~reich/Software.htm),
SNAP: http://www.broadinstitute.org/mpg/snap
PolyPhen: http://genetics.bwh.harvard.edu/pph/

Author contributions

The study was designed and financial support obtained by RSH, IPMT, MGD and HC. The manuscript was drafted by IPMT, RSH and MGD. Statistical and bioinformatic analyses were conducted by SED, SLS, and AT, with contributions from IPMT, JBC and RSH.
Oxford and local collaborators: Patient recruitment and sample acquisition were undertaken by EB, MG, LM, AL, DGRE, ERM, HJWT and members of the CORGI Consortium, and by RMa, RMi, EJ and DJK. Sample preparation was performed by KH, SLS and EEMJ. Genotyping was performed and co-ordinated by LGC-C, KH, AMJ, MC, EEMJ, AW and EDo. AD and EDe supplied eQTL data.
Institute of Cancer Research and local collaborators: Patient recruitment and sample acquisition to NSCCG were undertaken by SP. Co-ordination of sample preparation and genotyping was performed by PB. Sample preparation and genotyping were performed by AMP and BO.
Colon Cancer Genetics Group, Edinburgh and local collaborators: Patient recruitment and sample acquisition were performed by SF and members of the SOCCS and COGS study teams. Sample preparation was co-ordinated by SF. Genotyping was performed and co-ordinated by SF, ET, RB and MD. JGDP performed bioinformatic analyses.
The following authors from collaborating groups conceived the local or national study, undertook assembly of case/control series in their respective regions, collected data and samples, and variously undertook genotyping and analysis: CGS, JCo, SI, TM and JCh in Cardiff; IN, ST and LAA in Finland; and PP in Cambridge. All undertook sample collection and phenotype data collection and collation in the respective centres.

Competing interests statement

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

Farrington[3], **James GD Prendergast**[3], **Alan M Pittman**[1], **Evi Theodoratou**[3], **Christopher G Smith**[2], **Bianca Olver**[1], **Axel Walther**[4], **Rebecca A Barnetson**[3], **Michael Churchman**[4], **Emma EM Jaeger**[4], **Steven Penegar**[1], **Ella Barclay**[4], **Lynn Martin**[4], **Maggie Gorman**[4], **Rachel Mager**[5], **Elaine Johnstone**[5], **Rachel Midgley**[5], **Iina Niittymäki**[6], **Sari Tuupanen**[6], **James Colley**[2], **Shelley Idziaszczyk**[2], **The COGENT Consortium**[16], **Huw JM Thomas**[7], **Anneke M Lucassen**[8], **D Gareth R Evans**[9], **Eamonn R Maher**[10], **The CORGI Consortium**[16], **The COIN Collaborative Group**[16], **The COINB Collaborative Group**[16], **Timothy Maughan**[11], **Antigone Dimas**[4,12], **Emmanouil Dermitzakis**[12], **Jean-Baptiste Cazier**[4], **Lauri A Aaltonen**[6], **Paul Pharoah**[13], **David J Kerr**[5,14], **Luis G Carvajal-Carmona**[4], **Harry Campbell**[15], **Malcolm G Dunlop**[*,3], and **Ian PM Tomlinson**[*,3]

[1]Section of Cancer Genetics, Institute of Cancer Research, Sutton, SM2 5NG, United Kingdom [2]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, United Kingdom [3]Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Edinburgh EH4 2XU, United Kingdom [4]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN. United Kingdom [5]Department of Clinical Pharmacology, Oxford University, Radcliffe Infirmary, Old Road Campus Research Building, Headington, Oxford, OX3 7DQ, United Kingdom [6]Department of Medical Genetics, Genome-Scale Biology Research Program, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland [7]Colorectal Cancer Unit, Cancer Research UK, St Mark's Hospital, Harrow HA1 3UJ, UK [8]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton SO16 5YA, UK [9]Medical Genetics, St Mary's Hospital, Manchester M13 0JH, UK [10]Department of Medical and Molecular Genetics, School of Clinical and Experimental Medicine, University of Birmingham, Institute of Biomedical Research, and West Midlands Regional Genetics Service, Birmingham Women's Hospital, Edgbaston, Birmingham B15 2TT, UK [11]Department of Oncology and Palliative Care, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK [12]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva CH-1211, Switzerland [13]Cancer Research UK Laboratories, Strangeways Research Laboratory Department of Oncology, University of Cambridge, Cambridge CB1 8RN, United Kingdom [14]SIDRA Medical & Research Center, Qatar Foundation, PO Box 26999, Doha, Qatar [15]Public Health Sciences, University of Edinburgh, EH8 9AG, United Kingdom

Abstract

Genome-wide association (GWA) studies have thus far identified 10 loci at which common variants influence the risk of developing colorectal cancer (CRC). To enhance power to identify additional loci, we conducted a meta-analysis of three GWA studies from the UK totalling 3,334 cases and 4,628 controls, followed by multiple validation analyses, involving a total of 18,095 CRC cases and 20,197 controls. We identified new associations at 4 CRC risk loci: 1q41 (rs6691170, OR=1.06, P=9.55x10^{-10}; rs6687758, OR=1.09, P=2.27x10^{-9}); 3q26.2 (rs10936599, OR=0.93, P=3.39x10^{-8}); 12q13.13 (rs11169552, OR=0.92, P=1.89x10^{-10}; rs7136702, OR=1.06, P=4.02=x10^{-8}); and 20q13.33 (rs4925386, OR=0.93, P=1.89x10^{-10}). As well as identifying multiple new CRC risk loci this analysis provides evidence that additional CRC-associated variants of similar effect size remain to be discovered.

Genome-wide association (GWA) studies of colorectal cancer (CRC) have vindicated the hypothesis that part of the heritable risk is caused by common, low-risk variants 1. Our previous analyses, based on two GWA studies from the UK (UK1/CORGI) and Scotland (Scotland1/COGS) have identified 10 common variants that are associated with CRC risk 2. These variants map to 8q24.21 (rs6983267), 8q23.3 (rs16892766, *EIF3H*), 10p14 (rs10795668), 11q23 (rs3802842), 14q22.2 (rs4444235, *BMP4*), 15q13 (rs4779584), 16q22.1 (rs9929218, *CDH1*), 18q21.1 (rs4939827, *SMAD7*), 19q13.1 *(*rs10411210, *RHPN2)* and 20p12.3 (rs961253).

The discovered effect sizes of individual associations and the need for stringent thresholds for establishing statistical significance inevitably constrain the power of individual GWA studies to detect common variants. In order to augment our ability to detect additional CRC loci, we have undertaken a further GWA analysis of a set of cases from the VICTOR and QUASAR2 clinical trials of adjuvant therapy in potentially curable colorectal carcinoma These trials recruited patients from throughout the United Kingdom. The controls comprised a UK population-based 1958 Birth Cohort (58BC) for which genotype data are publicly available. Together, this case-control set (henceforth referred to as VQ58) comprised 1,432 cases and 2,697 controls.

The VQ58 cases were genotyped in-house using the Illumina Hap300/370 SNP arrays. After filtering of both the VQ data and the publicly-available control data to remove SNPs and individuals that fell below pre-determined quality control standards (see Methods), we examined associations between genotype and CRC status. A Q-Q plot demonstrated no evidence of systematic inflation of the allelic test statistic (λ_{gc}=1.018). No individual SNP showed a significant association with CRC under dominant, additive or recessive models at genome-wide significance (set at $P \leq 1.0 \times 10^{-7}$, based on a Bonferroni correction). This was not unexpected, given the power of the VQ58 set to detect associations of the magnitudes found in our previous analyses of the UK and Scottish GWA studies 2. We therefore directly proceeded to a combined analysis of the three GWA studies, comprising UK1/CORGI and Scotland1/COGS in addition to VQ58 (Supplementary Table 1). Quality control measures were standardised throughout the sample sets. We used principal components analysis (PCA) to examine whether there was evidence of distinct genetic sub-groups within the three GWA studies. After removal of 88 outliers and six duplicate samples, the Scottish and UK (UK1/CORGI and VQ58) samples essentially clustered together, minor variation in the first component reflecting the known North-West to South-East cline in the UK (Supplementary Figure 1).

The UK1 and Scotland1 samples had been genotyped using Illumina Hap550 arrays. We therefore imputed genotype probabilities in the VICTOR and QUASAR2 samples at SNPs not present on the Hap300/370 arrays. 94,867 of 214,649 imputed SNPs passed our threshold of ≤ 5% missing genotypes and an Information Score ≥ 0.5. We then conducted a meta-analysis of the three data sets (Supplementary Table 1) using the Mantel-Haenszel method under fixed- and random-effects models. Only one SNP (rs4939827, chromosome 18q21.1), previously shown to be associated with CRC risk 3–5, achieved formal genome-wide significance for association.

At this stage, we considered whether to include in the meta-analysis data we had generated from two additional, large UK case-control sets: UK2/NSCCG (2,854 cases and 2,822 controls) and Scotland2/SOCCS (2,024 cases and 2,092 controls) (Supplementary Table 1). These additional samples had been genotyped at 55,000 SNPs with the strongest evidence of association from meta-analysis of UK1/CORGI+Scotland1/COGS GWA studies 2. If we were to include these extra data, essentially we had to weigh two factors, (i) the extra power afforded by UK2/NSCCG and Scotland2/SOCCS *versus* (ii) the probability that a true CRC SNP had not been taken forward into the top 55,000 from the UK1+Scotland1 meta-analysis, but did make it into a (smaller) set of "top" SNPs in a VQ58+UK1+Scotland1 meta-analysis. Power calculations showed that, except for rare alleles with small effects for which the power of detection was in any event low, the extra power provided by the UK2 and Scotland2 samples more than compensated for the loss of a few truly disease-associated SNPs that would not have reached the significance threshold for genotyping in UK2 and Scotland2 (Supplementary Figure 2).

We therefore undertook a meta-analysis of VQ58, UK1/CORGI, Scotland1/COGS, UK2/NSCCG and Scotland2/SOCCS (Figure 1). Seven SNPs achieved formally significant associations ($P<10^{-7}$). All these SNPs had previously been shown to be associated with CRC risk. After exclusion of SNPs in strong pairwise LD ($r^2>0.7$), we selected seven SNPs (rs11805285, rs6687758, rs6691170, rs10936599, rs7136702, rs11169552, rs4925386) with nominal associations at $P<5.0x10^{-5}$ All of these SNPs had been genotyped, rather than imputed, in VQ. The 7 SNPs underwent validation testing in 9,883 CRC cases and 10,655 controls from six independent, northern European case-control series (COIN/NBS, Helsinki, UK3/NSCCG, UK4/CORGI2BCD, Scotland3/SOCCS and Cambridge; Supplementary Table 1). This threshold for follow-up did not exclude the possibility that other SNPs represented genuine association signals, but was simply a pragmatic strategy for prioritising replication. After replication, significant associations were confirmed for six SNPs mapping to four loci: rs6687758 ($P=2.27x10^{-9}$) and rs6691170 ($P=9.55x10^{-10}$) at 1q41, rs10936599 ($P=3.39x10^{-8}$) at 3q36.2, rs7136702 ($P=4.02=x10^{-8}$) and rs11169552 ($P=1.89x10^{-10}$) at 12q13.13 and rs4925386 ($P=1.89x10^{-10}$) at 20q13.3 (Figure 2, Table 1, Supplementary Table 2). There was no significant between-study heterogeneity for these SNP associations ($P_{het}>0.05$ for all SNPs, Table 1) and no SNP showed any evidence of association with age or sex in any data set ($P>0.05$).

rs6691170 (chr1:220,112,069) and rs6687758 (chr1:220,231,571) lie 125kb from each other on chromosome 1q41 (Table 1). The region containing these two SNPs (Figure 3) is flanked by recombination hotspots close to rs3003888 (chr1:220,049,548) and rs6687797 (chr1:220,296,043). Between these sites, LD relationships are complex and blocks are not easily defined, although a minor recombination hotspot exists at chr1:220,137,516, between rs6691170 and rs66867758. rs6691170 and rs66867758 respectively lie 250kb and 125kb upstream of *DUSP10*, a dual-specificity phosphatase that inactivates p38 and SAPK/JNK. The region otherwise contains few genes, but several spliced ESTs. In the UK data sets, rs6691170 and rs6687758 were in modest pairwise LD ($r^2=0.22$; $D'=0.71$) raising the possibility that these SNPs may represent independent signals of association. We assessed this using multiple logistic regression analysis stratified by sample series in which genotypes at one SNP were assessed conditional on those at the other SNP. We found that rs6691170

was associated with an odds ratio [OR]=1.07 (P=6.15x10^{-5}) and rs6687758 with OR of 1.06 (P=1.92x10^{-4}). Individuals with the high-risk haplotype (TG) at rs6691170 and rs6687758 had a 1.15-fold increased risk of CRC compared with the low-risk haplotype (GA) (P=5.39x10^{-8}).

rs10936599 (chr3:170,974,795) is flanked by recombination hotspots at chr3:170,837,364 and chr3:171,082,143 (Figure 3). rs10936599 lies on chr3q26.2, within the myoneurin (*MYNN/OZSF*) gene which encodes a zinc finger protein of unknown function that is expressed principally in muscle. rs10936599 is also close to the actin-related protein M1 and the *TERC* telomerase loci.

rs7136702 (chr12: 49,166,483) and rs11169552 (chr12:49,441,930) lie about 275kb apart, within what is essentially a large, poorly-defined haplotype block (Figure 3), composed of a set of smaller blocks, but with considerable long-range LD between markers (chr12:48,658,293-49,505,968). rs7136702 is just telomeric to the myeloproliferative oncogene binding-protein gene *LARP4* and 30kb proximal to disco-interacting protein 2B (*DIP2B*), which may have a role in determining epithelial cell fate. rs11169552 is just telomeric to *DIP2B*, and proximal to activating transcription factor 1 (*ATF1*) . *ATF1* is the 3' partner in recurrent translocations with the *EWSR1* gene (chr22q12) that contribute to the development of soft tissue clear cell sarcomas 6. rs7136702 and rs11169552 map close to a known chromosomal fragile site, but we have found that colorectal tumours rarely show somatic chromosomal breakpoints at this site 7. rs7136702 and rs11169552 are not strongly correlated (r^2=0.11, D'=0.76 in the UK samples). We therefore tested independence of these signals using conditioned logistic regression analysis as for the chromosome 1 signals. In this combined analysis, the rs11169552 signal nearly retained global significance (OR=0.91, P=4.33x10^{-7}), whereas the strength of association at rs7136702 was reduced (OR=1.06, P=4.34x10^{-4}). Individuals with the high-risk haplotype (TC) at rs7136702 and rs11169552 had a 1.14-fold increased risk of CRC compared with the low-risk haplotype (CT) (P=6.90x10^{-8}).

rs4925386 (chr20:60,354,439) is within a very small haplotype block (chr20:60,330,882-60,355,038), although it shows moderate LD with distal markers outside the block (Figure 3). rs4925386 lies within the large laminin A5 (*LAMA5*) gene which is required for the production of noggin, a secreted BMP antagonist. It is notable that other BMP pathway SNPs are likely to be involved in CRC predisposition 2. rs4935386 is in moderate/strong LD (r^2>0.5) with four non-synonymous SNPs, which lead to substitutions Ala1908Thr, Arg2226His, Asp2062Asn and Val1900Met, although all of these are predicted to be benign changes.

For both 1q41 and 12q13.12, the two signals, if independent, might have resulted from two causal variants or from a single causal variant strongly associated with disease and correlated with both SNPs in the region. For each region, we addressed the latter possibility by imputing SNPs from the HapMap2 CEU samples between the flanking recombination hotspots. We conducted logistic regression analysis of GWA studies and, UK2/NSCCG and Scotland2/SOCCS, conditioning on the genotypes at each of the two identified SNPs. Although a small number of imputed SNPs from 12q13.12 had a stronger predicted

association than the genotyped SNPs (Supplementary Figure 3), no single imputed SNP was able to account for the dual signals in either the 1q41 or 12q13.12 region.

In order to explore whether any of these novel CRC associations resulted from cis-acting regulatory elements, we examined whether any of the 6 SNPs tagged reported expression Quantitative Trait Loci (eQTLs) for nearby genes. Although four SNPs had no association with known eQTLs, rs7136702 was in moderate-strong LD (r^2=0.47-0.61, *D'*=0.80-0.84) with four SNPs (rs11169520, rs11169524, rs3742062 and rs2280503) that had been associated with *DIP2B* expression in lymphoblastoid cell lines 8. Furthermore, rs492536 was in moderate/strong LD (r^2=0.61, *D'*=0.78) with rs13043313, an eQTL for *LAMA5* expression in the liver 7.

Using a case-only design, we searched for pairwise gene-gene interactions between the 6 new CRC susceptibility SNPs and between these 6 SNPs and the 10 previously identified risk SNPs (rs6983267, rs16892766, rs10795668, rs3802842, rs4444235, rs4779584, rs9929218, rs4939827, rs10411210 and rs961253) 2. Although there was suggestive evidence of epistasis between rs6687758 and rs7136702 (P=7.70x10^{-4}), this did not meet the threshold for formal significance after adjustment for 120 comparisons (P=0.09). There was no evidence to suggest any functional relationships between genes close to these SNPs. No other evidence of gene-gene interactions was found (details not shown).

We have identified four new CRC predisposition loci, none of which maps to previously reported cancer-predisposition genes of high or low penetrance. At two of these loci, there exists the possibility that two SNPs independently predict risk. Our study illustrates other general issues that currently affect large-scale studies to identify common predisposition alleles. Allelic ORs of CRC were less than 1.10 for each of the CRC SNPs we identified. Power to detect the effects of such loci was therefore modest, the likelihood of discovery being highly sensitive to small chance differences in genotype frequencies, especially in the three GWA study data sets. Therefore, many more CRC loci of similar effect size may exist. While the new CRC risk alleles we have identified collectively account for ~1.5% of the familial CRC risk, in concert with other alleles they have potential to impact significantly on disease risk and thus have application to risk stratification at a population level. Finally, the loci we have identified are likely to provide fresh insights into the aetiological basis of CRC.

Methods

Study participants

Supplementary Table 1 provides a summary of all cases and controls in the study.

After exclusion of non-white UK cases and samples of poor quality, VQ58 comprised 1,432 CRC cases (896 males, mean age of diagnosis 62.4 years; SD ± 10.7) from the VICTOR and QUASAR2 trials. There were 2,697 population control genotypes (1,391 males,) from the Wellcome Trust Case-Control Consortium 2 (WTCCC2) 1958 birth cohort (also known as the National Child Development Study), which included all births in England, Wales and Scotland during a single week in 1958 9.

The compositions of the UK1/CORGI, Scotland1/COGS, UK2/NSCCG, Scotland2/SOCCS, UK3/NSCCG, Scotland3/SOCCS, Helsinki and Cambridge sample sets have been described previously 10 and are given in Supplementary Methods. The COIN samples were 2,151 cases (1,423 males) derived from the COIN and COIN-B clinical trials of metastatic CRC. Median age was 63 years (range 22-87). COIN cases were compared against genotypes from 2,501 population controls (1,237 males,) from the WTCCC2 National Blood Service (NBS) cohort. The UK4/CORGI2BCD samples comprised additional CRC cases and unaffected spouse/partner controls from the CORGI study collected since the UK1/CORGI samples. In all cases CRC was defined according to the ninth revision of the International Classification of Diseases (ICD) by codes 153-154 and all cases had pathologically proven disease.

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

Genotyping

DNA was extracted from samples using conventional methods and quantified using PicoGreen (Invitrogen). The VQ, UK1 and Scotland 1 GWA cohorts were genotyped using Illumina Hap300, Hap370, Hap240S or Hap550 arrays. 1958BC and NBS genotyping was performed as part of the WTCCC2 study In UK2/NSCCG and Scotland2/SOCCS, genotyping was conducted using custom Illumina Infinium arrays according to the manufacturer's protocols. To ensure quality of genotyping, a series of duplicate samples was genotyped, resulting in 99.9% concordant calls.

Other genotyping was conducted using competitive allele-specific PCR KASPar chemistry (KBiosciences Ltd, Hertfordshire, UK). All primers, probes and conditions used are available on request. 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, >99.9% concordant results were obtained.

Quality control

We excluded SNPs from analysis if they failed one or more of the following thresholds: GenCall scores <0.25; overall call rates <95%; 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 individuals 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 PCA-based analysis in comparison with HapMap samples. In addition, PCA was used to exclude individuals or groups distinct from the main cluster using the first three principal components, initially based on separate analysis of VQ58, UK1 and Scotland1 (and also NBS), and subsequently on combined analysis of all three data sets (Supplementary Figure 1). To identify individuals who might have non-northern European

ancestry, we merged our case and control data with the 60 European (CEU), 60 Nigerian (YRI), and 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 principal components analysis. The first two principal components for each individual were plotted and any individual not present in the main CEU cluster (that is, >5% of the PC distance from HapMap CEU cluster centroid) was excluded from subsequent analyses.

The adequacy of the case-control matching and possibility of differential genotyping of cases and controls was formally evaluated using Q-Q plots of test statistics. The inflation factor λ was calculated 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. 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.

Association between SNP genotype and disease status was primarily assessed in PLINK v1.07 using allelic and Cochran-Armitage tests (both with 1df) or by Fisher's exact test where an expected cell count was <5. Genotypic (2df), dominant (1df) and recessive (1df) tests were also performed. The risks associated with each SNP were estimated by allelic, heterozygous and homozygous odds ratios (ORs) using unconditional logistic regression, and associated 95% confidence intervals (CIs) were calculated.

Joint analysis of data generated from multiple phases was conducted using standard methods for combining raw data based on the Mantel-Haenszel method in STATA and PLINK. The reported meta-analysis statistics were derived from analysis of allele frequencies, and joint ORs and 95% CIs were calculated assuming fixed- and random-effects models. Tests of the significance of the pooled effect sizes were calculated using a standard normal distribution. Cochran's Q statistic to test for heterogeneity 11 and the I^2 statistic 12 to quantify the proportion of the total variation due to heterogeneity were calculated. Large heterogeneity is typically defined as $I^2 \geq 75\%$. Where significant heterogeneity was identified, results from the random effects model were reported. Alongside, we also performed meta-analysis based on allele dosage (0, 1, 2) and incorporated age and sex as co-variates. Although age and sex are associated with colorectal cancer risk, they were not associated with SNP genotype and did not materially affect the significance of any of the 6 reported associations (details not shown).

We used Haploview software v4.2 to infer the LD structure of the genome in the regions containing loci associated with disease risk. The combined effects of pairs of loci identified as associated with CRC risk were investigated by multiple logistic regression analysis in PLINK to test for independent effects of each SNP and stratifying by sample series. Evidence for interactive effects between SNPs (epistasis) was assessed by likelihood ratio test assuming an allelic model in PLINK. The ORs for increasing numbers of deleterious alleles were estimated by counting two for a homozygote and one for a heterozygote at each of the 16 risk SNPs, and a trend test was performed on the resulting data.

The sibling relative risk attributable to a given SNP was calculated using the formula

$$\lambda^{*}=\frac{p(pr_2+qr_1)^2+q(pr_1+q)^2}{(p^2r_2+2pqr_1+q^2)^2}$$

where p is the population frequency of the minor allele, q=1-p, and r_1 and r_2 are the relative risks (estimated as OR) for heterozygotes and rare homozygotes, relative to common homozygotes 13. Assuming a multiplicative interaction the proportion of the familial risk attributable to a SNP was calculated as $\log(\lambda^*)/\log(\lambda_0)$, where λ_0 is the overall familial relative risk estimated from epidemiological studies of CRC, assumed to be 2.2 14. UK2/NSCCG2 samples were used for this estimation.

Imputation from HapMap2 build 36 was performed using the IMPUTE2 program (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html), incorporating as a reference panel for the VQ58 Hap300-panel genotypes the Hap550-typed UK controls from the UK1/CORGI study. SNPs were included in the analysis if there were ≤5% missing genotypes and an Information Score ≥0.5. SNPtest was used to perform association meta-analysis. Principal components analysis was performed using Eigenstrat/SmartPCA using CEU, YRI and HCB HapMap samples as reference.

Genome co-ordinates were taken from the NCBI build 36/hg18 (dbSNP b126).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Cancer Research UK provided principal funding for this study individually to IPMT, MGD, RSH, PP and JCh. Additional funding was provided by the Oxford Comprehensive Biomedical Research Centre (to EDo and IPMT and, the EU FP7 CHIBCHA grant (to LGC-C and IPMT). Core infrastructure support to the Wellcome Trust Centre for Human Genetics, Oxford was provided by grant 075491/Z/04. 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 would also like to thank colleagues from the UK National Cancer Research Network (for NSCCG). Additional funding (to MGD) was provided by the Medical Research Council (G0000657-53203), CORE and Scottish Executive Chief Scientist's Office (K/OPR/2/2/D333, CZB/4/449). We (Edinburgh) gratefully acknowledge the work of the COGS and SOCCS administrative teams; Roseanne Cetnarskyj and the research nurse teams, all who recruited to the studies; the Wellcome Trust Clinical Research Facility for sample preparation and to all clinicians and pathologists throughout Scotland at collaborating centres. ET was funded by a Cancer Research UK Fellowship (C31250/A10107). The study used the biological and data resource of Generation Scotland. COIN and COIN-B were funded by the UK Medical Research Council. COIN sample analysis (JCh) was also funded by Cancer Research Wales, Tenovus & Wales Gene Park. For Helsinki, the work was supported by grants from Academy of Finland (Finnish Centre of Excellence Program 2006-2011), the Finnish Cancer Society, and the Sigrid Juselius Foundation. For Cambridge, we thank the SEARCH study team and all the participants in the study. PP is a Cancer Research UK Senior Clinical Research Fellow. This study made use of genotyping data on the 1958 Birth Cohort and NBS samples, 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 from http://www.wtccc.org.uk/. Finally, we would like to thank all individuals who participated in the study.

References

1. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant…or not? Hum Mol Genet. 2002; 11:2417–23. [PubMed: 12351577]

2. Houlston RS, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet. 2008; 40:1426–35. [PubMed: 19011631]
3. Haiman CA, et al. A common genetic risk factor for colorectal and prostate cancer. Nat Genet. 2007; 39:954–6. [PubMed: 17618282]
4. Tomlinson I, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. 2007; 39:984–8. [PubMed: 17618284]
5. Zanke BW, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet. 2007; 39:989–94. [PubMed: 17618283]
6. Zucman J, et al. EWS and ATF-1 gene fusion induced by t(12;22) translocation in malignant melanoma of soft parts. Nat Genet. 1993; 4:341–5. [PubMed: 8401579]
7. Jones AM, et al. Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. Oncogene. 2005; 24:118–29. [PubMed: 15531920]
8. Dixon AL, et al. A genome-wide association study of global gene expression. Nat Genet. 2007; 39:1202–7. [PubMed: 17873877]
9. Power C, Jefferis BJ, Manor O, Hertzman C. The influence of birth weight and socioeconomic position on cognitive development: Does the early home and learning environment modify their effects? J Pediatr. 2006; 148:54–61. [PubMed: 16423598]
10. Tomlinson IP, et al. COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer. Br J Cancer. 2009; 102:447–54. [PubMed: 19920828]
11. Petitti DB. Coronary heart disease and estrogen replacement therapy. Can compliance bias explain the results of observational studies? Ann Epidemiol. 1994; 4:115–8. [PubMed: 8205277]
12. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21:1539–58. [PubMed: 12111919]
13. Houlston RS, Ford D. Genetics of coeliac disease. QJM. 1996; 89:737–43. [PubMed: 8944229]
14. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol. 2001; 96:2992–3003. [PubMed: 11693338]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

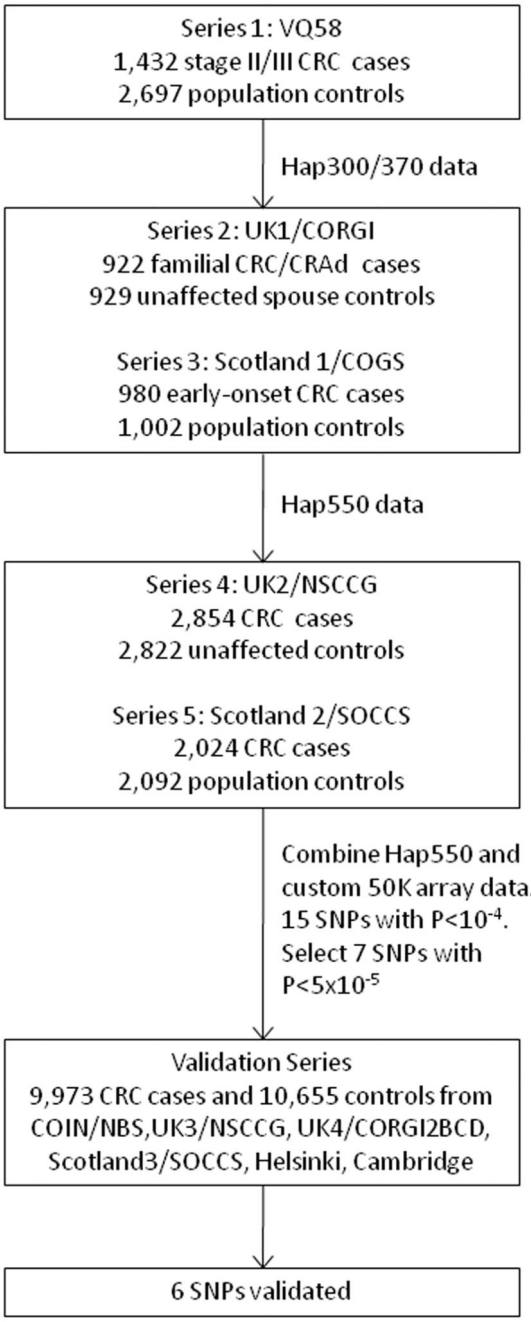

Figure 1. Overall study design

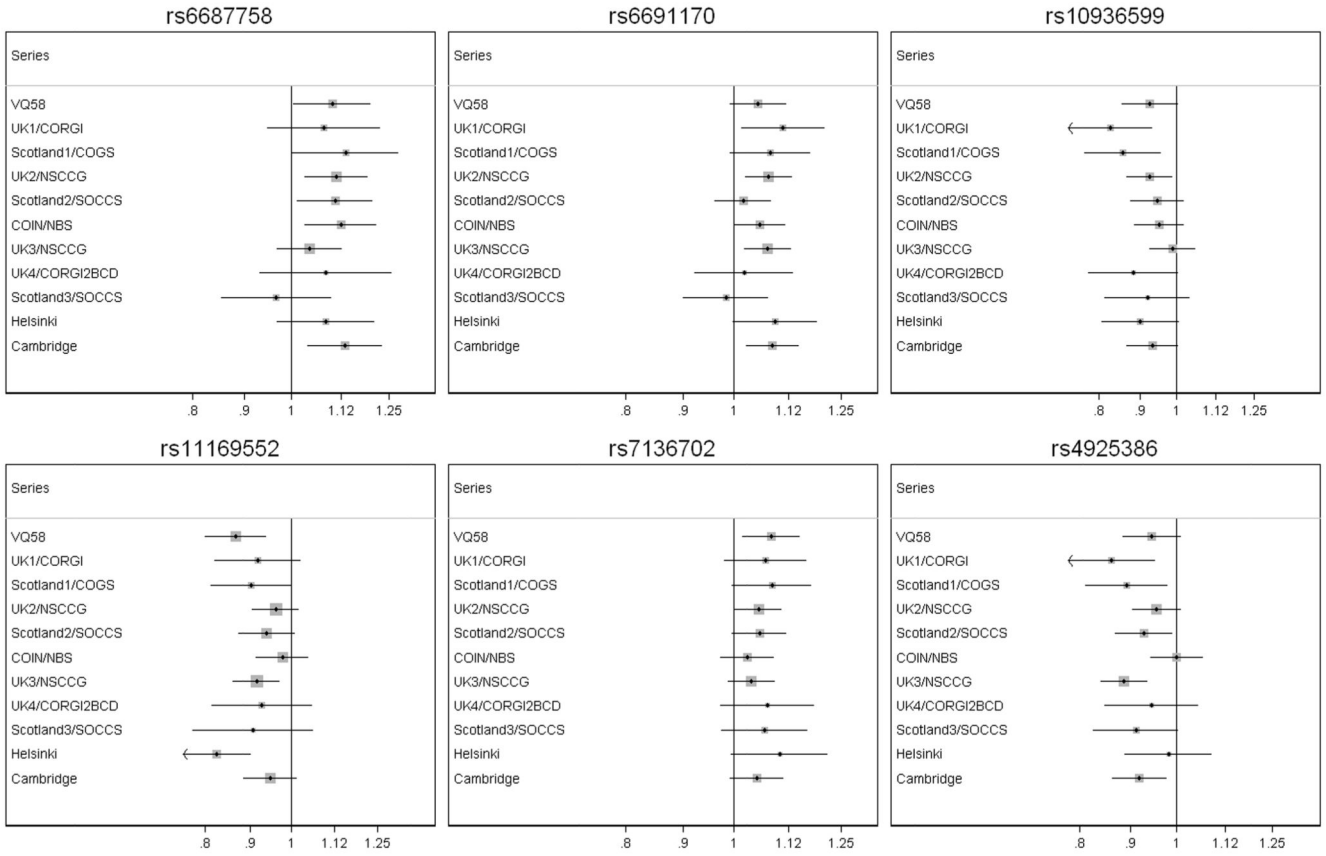

Figure 2. Forest plots of effect size and direction for the six SNPs associated with CRC.

Boxes denote allelic OR point estimates, their areas being proportional to the inverse variance weight of the estimate. Horizontal lines represent 95% confidence intervals. The diamond (and broken line) represents the summary OR computed under a fixed effects model, with 95% confidence interval given by its width. The unbroken vertical line is at the null value (OR=1.0).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

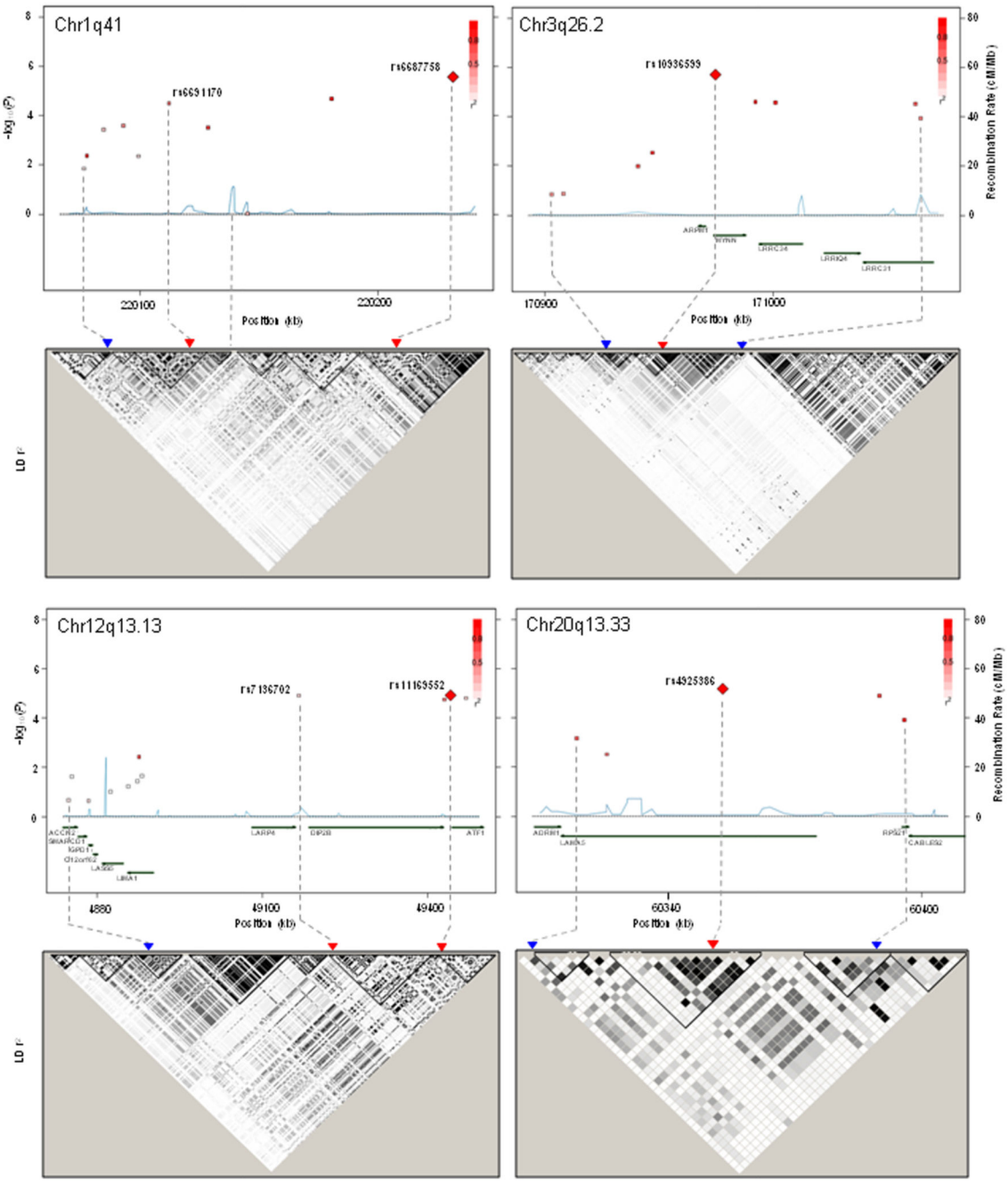

Figure 3. Maps of the (a) 1q41, (b) 3q26.2, (c) 12q13.13 and (d) 20q13.33 regions, showing evidence of association with CRC and local LD structure.

In the association plot, each point represents a SNP genotyped at this locus. For each SNP at the position (kb) shown on the x-axis, $-\log_{10}P$ from the allelic association test is indicated on the y-axis. Recombination rate is shown in blue. The SNP with the strongest association in each region is shown as a red diamond. Data were derived from the combined analysis of VQ58, UK1, Scotland1, UK2 and Scotland2; this resulted in relatively few SNPs being shown for each region, but illustrates the rationale for the selection of SNPs for genotyping in the validation sample sets. In the LD plots (lower), derived from HapMap CEU

individuals in Haploview. the colour intensity of each SNP represents the strength of LD according to the standard Haploview scheme for r^2 (black >0.90 through shades of grey to white 0.0). Note that *DUSP10* is not shown for chromosome 1q41 but maps to 219,941,389-219,982,084; similarly, *TERC* is not shown for chromosome 3q26.2, but lies at 170,965,092-170,965,542. Physical positions are based on NCBI build 36 of the human genome.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Summary of results for six SNPs associated with colorectal cancer.

Odds ratios (95% confidence intervals) and P values from the allelic test are shown for the Discovery phase, RFeplication Phase and Overall for each of the 6 SNPs associated with risk of CRC. Further details are provided in Supplementary Table 2.

	Discovery phase	Replication phase	Overall
rs6691170	OR=1.06 (1.03-1.09), P=3.05x10^{-5}	OR=1.06 (1.03-1.09), P=6.48x10^{-6}	OR=1.06 (1.03-1.09), P=9.55x10^{-10}
rs6687758	OR=1.10 (1.06-1.15), P=2.73x10^{-6}	OR=1.08 (1.04-1.12), P=1.57x10^{-4}	OR=1.09 (1.06-1.12), P=2.27x10^{-9}
rs10936599	OR=0.91 (0.88-0.95), P=2.03x10^{-6}	OR=0.95 (0.91-0.98), P=1.87x10^{-3}	OR=0.93 (0.91-0.96), P=3.39x10^{-8}
rs7136702	OR=1.06 (1.03-1.09), P=1.19x10^{-5}	OR=1.05 (1.02-1.08), P=6.50x10^{-4}	OR=1.06 (1.04-1.08), P=4.02x10^{-8}
rs11169552	OR=0.92 (0.89-0.96), P=1.24x10^{-5}	OR=0.93 (0.90-0.96), P=3.66x10^{-6}	OR=0.92 (0.90-0.95), P=1.89x10^{-10}
rs4925386	OR=0.93 (0.90-0.96), P=6.80x10^{-6}	OR=0.93 (0.91-0.96), P=6.48x10^{-6}	OR=0.93 (0.91-0.95), P=1.89x10^{-10}