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

Citation for published version:
COGENT Consortium, Houlston, RS, Cheadle, J, Dobbins, SE, Tenesa, A, Jones, AM, Howarth, K, Spain, SL, Broderick, P, Domingo, E, Farrington, S, Prendergast, J, Pittman, AM, Theodoratou, E, Smith, CG, Olver, B, Walther, A, Barnetson, RA, Churchman, M, Jaeger, EEM, Penegar, S, Barclay, E, Martin, L, Gorman, M, Mager, R, Johnstone, E, Midgley, R, Niittymäki, I, Tuupanen, S, Colley, J, Idziaszczyk, S, Thomas, HJW, Lucassen, AM, Evans, DGR, Maher, ER, Maughan, T, Dimas, A, Dermitzakis, E, Cazier, J-B, Aaltonen, LA, Pharoah, P, Kerr, DJ, Carvajal-Carmona, LG, Campbell, H, Dunlop, MG & Tomlinson, IPM 2010, 'Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33', *Nature Genetics*, vol. 42, no. 11, pp. 973-979. https://doi.org/10.1038/ng.670

Digital Object Identifier (DOI):
10.1038/ng.670

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Nature Genetics

nature genetics

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 Farrington[3], James G D 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 E M 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 J W 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] & Ian P M Tomlinson[4]

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

GWAS of CRC have confirmed the hypothesis that part of the heritable risk for this disease is caused by common, low-risk variants[1]. Our previous analyses, based on two GWAS from the UK (UK1, also known as CORGI) and Scotland (Scotland1, also known as COGS), identified ten common variants 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 the individual associations and the need for stringent thresholds for establishing statistical significance inevitably constrain the power of individual GWAS to detect common variants. To augment our ability to detect additional CRC loci, we undertook a further GWAS analysis of a set of cases from the VICTOR and QUASAR2 clinical trials of adjuvant therapy in potentially curable colorectal carcinoma. These trials recruited cases from throughout the UK. The controls were from the UK population-based 1958 Birth Cohort, 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 and Hap370 SNP arrays. After filtering both the VQ data and the

[1]Section of Cancer Genetics, Institute of Cancer Research, Sutton, UK. [2]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, UK. [3]Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and Medical Research Council Human Genetics Unit, Edinburgh, UK. [4]Wellcome Trust Centre for Human Genetics, Oxford, UK. [5]Department of Clinical Pharmacology, Oxford University, Radcliffe Infirmary, Old Road Campus Research Building, Headington, Oxford, UK. [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, UK. [8]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton, UK. [9]Medical Genetics, St. Mary's Hospital, Manchester, 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, UK. [11]Department of Oncology and Palliative Care, School of Medicine, Cardiff University, Heath Park, Cardiff, UK. [12]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [13]Cancer Research UK Laboratories, Strangeways Research Laboratory Department of Oncology, University of Cambridge, Cambridge, UK. [14]Sidra Medical and Research Center, Qatar Foundation, Doha, Qatar. [15]Public Health Sciences, University of Edinburgh, Edinburgh, UK. [16]A full list of members is provided in the **Supplementary Note**. Correspondence should be addressed to R.S.H. (richard.houlston@icr.ac.uk), M.G.D. (malcolm.dunlop@hgu.mrc.ac.uk) or I.P.M.T. (iant@well.ox.ac.uk).

Received 9 April; accepted 1 September; published online 24 October 2010; doi:10.1038/ng.670

publicly available control data to remove SNPs and individuals that fell below pre-determined quality control standards (Online Methods), we examined associations between genotype and CRC status. A quantile-quantile plot showed no evidence of systematic inflation of the allelic test statistic (genomic control inflation factor (λ_{gc}) = 1.018). No individual SNP showed 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 dataset to detect associations of the magnitudes found in our previous analyses of the UK and Scottish GWAS[2]. We therefore proceeded directly to a combined analysis of UK1 (CORGI) and Scotland1 (COGS) and VQ58 (**Supplementary Table 1**). Quality control measures were standardized throughout the sample sets. We used principal components analysis (PCA) to examine whether there was evidence of distinct genetic subgroups within the three GWAS. After removal of 88 outliers and 6 duplicate samples, the Scottish and UK (UK1 (CORGI) and VQ58) samples essentially clustered together, with minor variation in the first component reflecting the known north-west to south-east cline in the UK (**Supplementary Fig. 1**).

The UK1 and Scotland1 samples had previously been genotyped using Illumina Hap550 arrays. We therefore imputed genotype probabilities in the VICTOR and QUASAR2 samples at SNPs not present on the Illumina Hap300 and Hap370 arrays. Of the SNPs we imputed, 94,867 out of 214,649 passed our threshold of ≤5% missing genotypes and an information score of ≥0.5. We then conducted a meta-analysis of the three datasets (**Supplementary Table 1**) using the Mantel-Haenszel method under both fixed- and random-effects models. Only one SNP (rs4939827 on chromosome (chr.) 18q21.1), which was previously shown to be associated with CRC risk[3–5], achieved genome-wide significance for association.

At this stage, we considered whether to include data we had generated from two additional large UK case-control sets in our meta-analysis: UK2 (NSCCG) (comprising 2,854 cases and 2,822 controls) and Scotland2 (SOCCS) (comprising 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 a meta-analysis of the UK1 (CORGI) and Scotland1 (COGS) GWAS[2]. If we were to include these extra data, essentially we had to weigh two factors: (i) the extra power afforded by including UK2 (NSCCG) and Scotland2 (SOCCS) data compared to (ii) the probability that a true CRC SNP had not been taken forward into the top 55,000 SNPs from the UK1 and Scotland1 meta-analysis but did make it into a smaller set of top SNPs in a VQ58, UK1 and 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 true disease-associated SNPs that would not have reached the significance threshold for genotyping in UK2 and Scotland2 (**Supplementary Fig. 2**).

We therefore undertook a meta-analysis of VQ58, UK1 (CORGI), Scotland1 (COGS), UK2 (NSCCG) and Scotland2 (SOCCS) (**Fig. 1**). Seven SNPs achieved significant associations ($P < 10^{-7}$) in this analysis. All these SNPs had previously been shown to be associated with CRC risk. After exclusion of SNPs in strong pairwise linkage disequilibrium (LD) ($r^2 > 0.7$), we selected seven SNPs (rs11805285, rs6687758, rs6691170, rs10936599, rs7136702, rs11169552 and rs4925386) with nominal associations at $P < 5.0 \times 10^{-5}$. All of these SNPs had been genotyped, rather than imputed, in the VQ dataset. These seven 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 prioritizing replication. After further genotyping, significant associations were confirmed for six SNPs mapping to four loci: rs6687758 ($P = 2.27 \times 10^{-9}$) and rs6691170 ($P = 9.55 \times 10^{-10}$) at 1q41; rs10936599 ($P = 3.39 \times 10^{-8}$) at 3q36.2, rs7136702 ($P = 4.02 \times 10^{-8}$) and rs11169552 ($P = 1.89 \times 10^{-10}$) at 12q13.13; and rs4925386 ($P = 1.89 \times 10^{-10}$) at 20q13.3 (**Fig. 2**, **Table 1** and **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 dataset ($P > 0.05$).

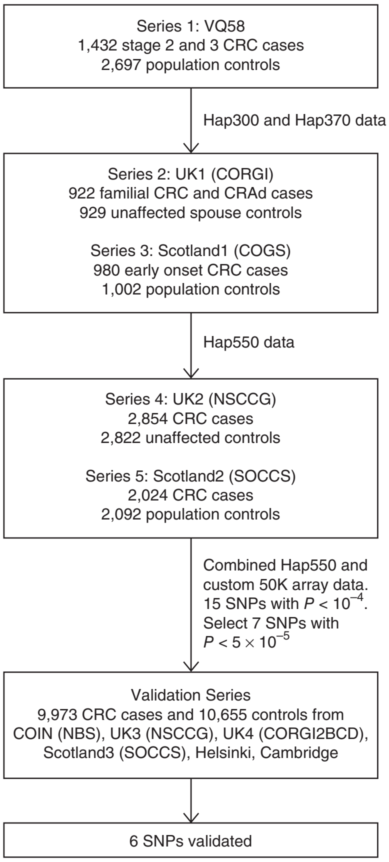
Series 1: VQ58
1,432 stage 2 and 3 CRC cases
2,697 population controls
Hap300 and Hap370 data
Series 2: UK1 (CORGI)
922 familial CRC and CRAd cases
929 unaffected spouse controls
Series 3: Scotland1 (COGS)
980 early onset CRC cases
1,002 population controls
Hap550 data
Series 4: UK2 (NSCCG)
2,854 CRC cases
2,822 unaffected controls
Series 5: Scotland2 (SOCCS)
2,024 CRC cases
2,092 population controls
Combined Hap550 and custom 50K array data.
15 SNPs with P < 10^-4.
Select 7 SNPs with P < 5 × 10^-5
Validation Series
9,973 CRC cases and 10,655 controls from COIN (NBS), UK3 (NSCCG), UK4 (CORGI2BCD), Scotland3 (SOCCS), Helsinki, Cambridge
6 SNPs validated

Figure 1 Overall study design.

rs6691170 (on chr. 1 at location 220,112,069) and rs6687758 (on chr. 1 at 220,231,571) lie 125 kb from each other on chromosome 1q41 (**Table 1**). The region containing these two SNPs (**Fig. 3**) is flanked by recombination hotspots close to rs3003888 (on chr. 1 at 220,049,548) and rs6687797 (on chr. 1 at 220,296,043). Between these sites, LD relationships are complex and LD blocks are not easily defined, although a minor recombination hotspot exists on chr. 1 at 220,137,516 between rs6691170 and rs66867758. rs6691170 and rs66867758, respectively, lie 250 kb and 125 kb upstream of *DUSP10*, which encodes a dual-specificity phosphatase that inactivates p38 and SAPK (JNK). The region otherwise contains few genes but does contain several spliced ESTs. In the UK datasets, rs6691170 and rs6687758 were in modest pairwise LD ($r^2 = 0.22$ and $D' = 0.71$), 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

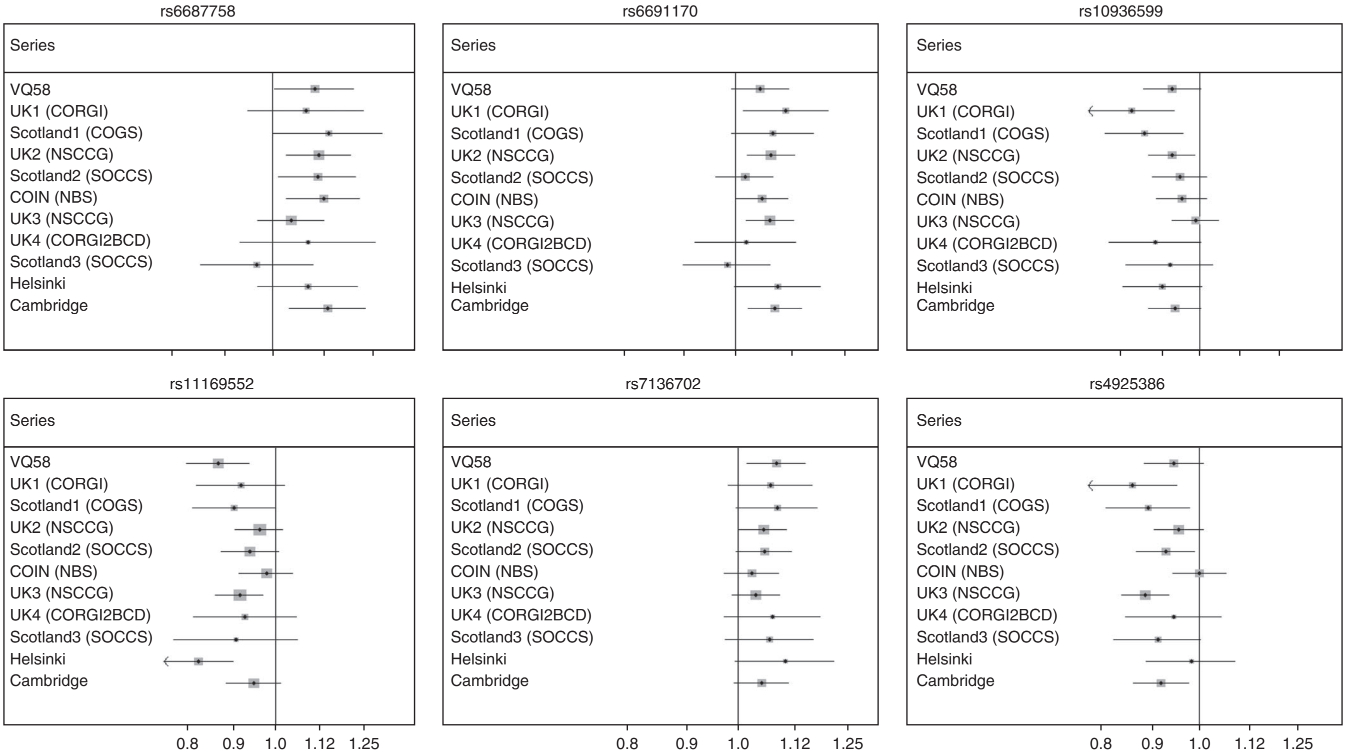

Figure 2 Forest plots of effect size and direction for the six SNPs associated with CRC. Boxes denote allelic OR point estimates with their areas being proportional to the inverse variance weight of the estimate. Horizontal lines represent 95% CIs. The diamond (and broken line) represents the summary OR computed under a fixed effects model, with the 95% CI indicated by its width. The unbroken vertical line is at the null value (OR = 1.0).

those at the other SNP. We found that rs6691170 had an OR of 1.07 (P = 6.15 × 10^{-5}), and rs6687758 had an OR of 1.06 (P = 1.92 × 10^{-4}). Individuals with the high-risk haplotype (TG) at rs6691170 and rs6687758 had a 1.15-fold increased risk of CRC compared to those individuals with the low-risk haplotype (GA) (P = 5.39 × 10^{-8}).

rs10936599 (on chr. 3 at 170,974,795) is flanked by recombination hotspots at chr. 3 position 170,837,364 and chr. 3 position 171,082,143 (**Fig. 3**). rs10936599 lies at 3q26.2 within *MYNN* (the myoneurin 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 locus.

rs7136702 (on chr. 12 at 49,166,483) and rs11169552 (on chr. 12 at 49,441,930) lie about 275 kb apart within what is essentially a large, poorly-defined haplotype block (**Fig. 3**) composed of a set of smaller blocks but with considerable long-range LD between markers (chr. 12 spanning 48,658,293–49,505,968). rs7136702 is just telomeric to the myeloproliferative oncogene binding-protein gene *LARP4* and 30 kb proximal to *DIP2* (encoding disco-interacting protein 2B), which may have a role in determining epithelial cell fate. rs11169552 is just telomeric to *DIP2B* and proximal to *ATF1* (encoding activating transcription factor 1). *ATF1* is the 3′ partner in the recurrent translocations with *EWSR1* (22q12) 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 found that colorectal tumors rarely show somatic chromosomal breakpoints at this site[7]. rs7136702 and rs11169552 are not strongly correlated (r^2 = 0.11 and D' = 0.76 in the UK samples). We therefore tested independence of these signals using conditioned logistic regression analysis just as we did for the chromosome 1 signals. In this combined analysis, the rs11169552 signal nearly retained global significance (OR = 0.91, P = 4.33 × 10^{-7}), whereas the strength of association at rs7136702 was reduced (OR = 1.06, P = 4.34 × 10^{-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 = 6.90 × 10^{-8}).

rs4925386 (on chr. 20 at 60,354,439) is within a very small haplotype block (on chr. 20 spanning 60,330,882–60,355,038), although

Table 1 Summary results for six SNPs associated with colorectal cancer

		Discovery		Replication		Overall	
SNP	Locus	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
rs6691170	1q41	1.06 (1.03–1.09)	3.05 × 10^{-5}	1.06 (1.03–1.09)	6.48 × 10^{-6}	1.06 (1.03–1.09)	9.55 × 10^{-10}
rs6687758	1q41	1.10 (1.06–1.15)	2.73 × 10^{-6}	1.08 (1.04–1.12)	1.57 × 10^{-4}	1.09 (1.06–1.12)	2.27 × 10^{-9}
rs10936599	3q26.2	0.91 (0.88–0.95)	2.03 × 10^{-6}	0.95 (0.91–0.98)	1.87 × 10^{-3}	0.93 (0.91–0.96)	3.39 × 10^{-8}
rs7136702	12q13.13	1.06 (1.03–1.09)	1.19 × 10^{-5}	1.05 (1.02–1.08)	6.50 × 10^{-4}	1.06 (1.04–1.08)	4.02 × 10^{-8}
rs11169552	12q13.3	0.92 (0.89–0.96)	1.24 × 10^{-5}	0.93 (0.90–0.96)	3.66 × 10^{-6}	0.92 (0.90–0.95)	1.89 × 10^{-10}
rs4925386	20q13.33	0.93 (0.90–0.96)	6.80 × 10^{-6}	0.93 (0.91–0.96)	6.48 × 10^{-6}	0.93 (0.91–0.95)	1.89 × 10^{-10}

Odds ratios (95% CIs) and P values from the allelic test are shown for the discovery phase, the replication phase and overall for each of the six SNPs associated with risk of CRC. Further details are provided in **Supplementary Table 2**.

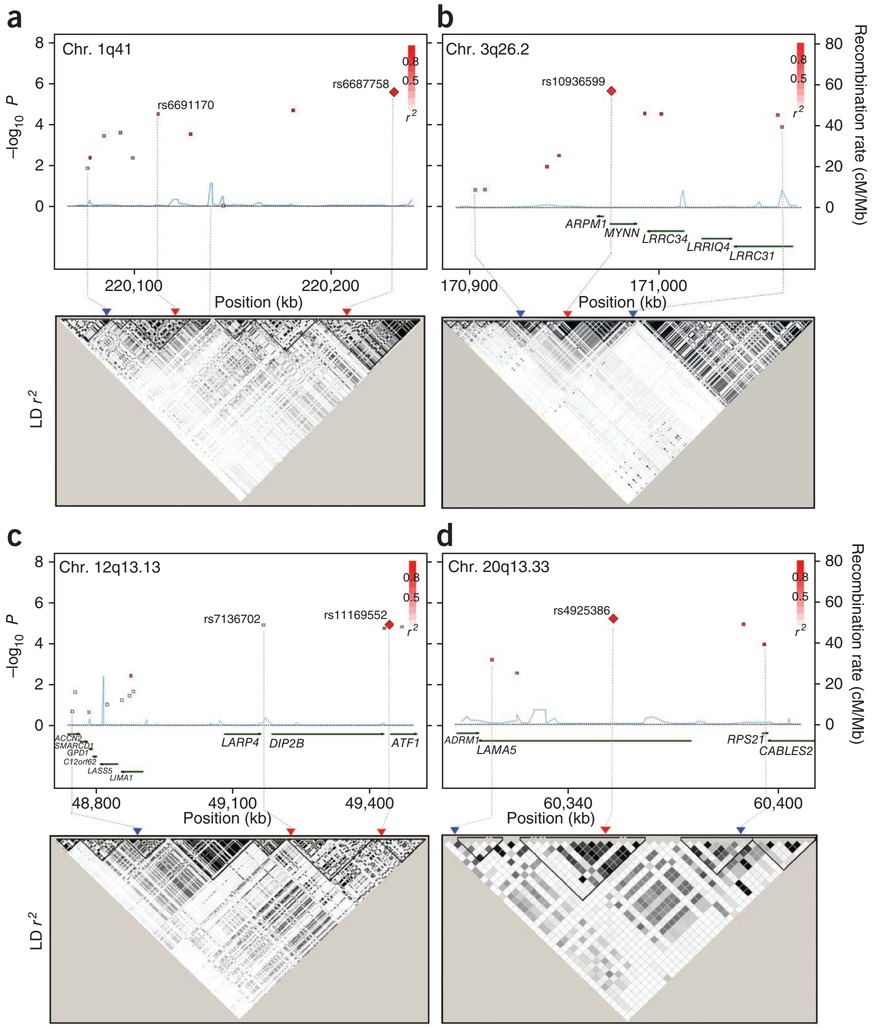

Figure 3 Regional plots. (**a**–**d**) Maps of the 1q41 (**a**), 3q26.2 (**b**), 12q13.13 (**c**) and 20q13.33 (**d**) regions showing evidence of association with CRC and local LD structure. In the association plot, each point represents a SNP genotyped at this locus. For each SNP at the position (kb) shown on the *x* axis, $-\log_{10} P$ from the allelic association test is indicated on the *y* axis. The 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 the VQ58, UK1, Scotland1, UK2 and Scotland2 cohorts, which resulted in relatively few SNPs being shown for each region but which 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 color intensity of each SNP represents the strength of LD according to the standard Haploview scheme for r^2 (with black indicating values >0.90 through shades of gray to white, which indicate a value of 0.0). Physical positions are based on NCBI build 36 of the human genome.

it shows moderate LD with distal markers outside the block (**Fig. 3**). rs4925386 lies within *LAMA5* (encoding large laminin A5), 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 to strong LD (r^2 > 0.5) with four non-synonymous *LAMA5* SNPs which lead to the substitutions p.Ala1908Thr, p.Arg2226His, p.Asp2062Asn and p.Val1900Met, although all of these alterations are predicted to be benign.

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 European CEU samples between the flanking recombination hotspots. We conducted logistic regression analysis of the GWAS and UK2 (NSCCG) and Scotland2 (SOCCS) datasets, 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 Fig. 3**), no single imputed SNP was able to account for the dual signals in either the 1q41 or 12q13.12 region.

To explore whether any of these newly discovered CRC associations resulted from *cis*-acting regulatory elements, we examined whether any of the six SNPs tagged reported expression quantitative trait loci (eQTLs) for nearby genes. Although four SNPs had no association with known eQTLs, rs7136702 was in moderate to strong LD ($r^2 = 0.47$ to $r^2 = 0.61$ and $D' = 0.80$ to $D' = 0.84$) with four SNPs (rs11169520, rs11169524, rs3742062 and rs2280503) that have previously been associated with *DIP2B* expression in lymphoblastoid cell lines[8]. Furthermore, rs492536 was in moderate to 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 six new CRC susceptibility SNPs and also between these six SNPs and the ten 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 ($P = 7.70 \times 10^{-4}$), this evidence did not meet the threshold for significance after adjustment for 120 comparisons ($P = 4.2 \times 10^{-4}$). There was no evidence to suggest any functional relationships between genes close to these SNPs. No other evidence of gene-gene interactions was found (data not shown).

We identified four new CRC risk 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 were less than 1.10 for each of the CRC SNPs we identified. Power to detect the effects of such loci was therefore modest, with the likelihood of discovery being highly sensitive to small chance differences in genotype frequencies, especially in the three GWAS datasets. Therefore, many more CRC loci of similar effect size may exist. Although 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 the potential to substantially influence disease risk and thus have an application to risk stratification at a population level. Finally, the loci we identified are likely to provide fresh insights into the etiological basis of CRC.

URLs. Detailed information on the tag SNP panel, http://www.illumina.com/; Haploview, http://www.broadinstitute.org/haploview/haploview; VICTOR and 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; National Study of Colorectal Cancer Genetics (NSCCG), http://pfsearch.ukcrn.org.uk/StudyDetail.aspx?TopicID=1&StudyID=1269; 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/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

Cancer Research UK provided principal funding for this study individually to I.P.M.T., M.G.D., R.S.H., P.P. and J. Cheadle. Additional funding was provided by the Oxford Comprehensive Biomedical Research Centre (to E. Domingo and I.P.M.T.) and the EU FP7 CHIBCHA grant (to L.G.C.-C. and I.P.M.T.). 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 also thank colleagues from the UK National Cancer Research Network (for NSCCG). Additional funding (to M.G.D.) 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; R. Cetnarskyj and the research nurse teams, who all recruited subjects to the studies; the Wellcome Trust Clinical Research Facility for sample preparation; and to all clinicians and pathologists throughout Scotland at the collaborating centres. E.T. was funded by a Cancer Research UK Fellowship (C31250/A10107). The study used the biological and data resources of Generation Scotland. COIN and COINB were funded by the UK Medical Research Council. COIN sample analysis (J. Cheadle) was also funded by Cancer Research Wales, Tenovus and Wales Gene Park. For the Helsinki study, 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 the Cambridge study, we thank the SEARCH study team and all the participants in the study. P.P. is a Cancer Research UK Senior Clinical Research Fellow. This study made use of genotyping data from 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 at http://www.wtccc.org.uk/. Finally, we would like to thank all individuals who participated in the study.

AUTHOR CONTRIBUTIONS

The study was designed and financial support was obtained by R.S.H., I.P.M.T., M.G.D. and H.C. The manuscript was drafted by I.P.M.T., R.S.H. and M.G.D. Statistical and bioinformatic analyses were conducted by S.E.D., S.L.S. and A.T., with contributions from I.P.M.T., J.-B.C. and R.S.H.

Oxford and local collaborators: subject recruitment and sample acquisition were done by E.B., M.G., L.M., A.M.L., D.G.R.E., E.R.M., H.J.W.T. and members of the CORGI Consortium, and by R. Mager, R. Midgley, E.J. and D.J.K. Sample preparation was performed by K.H., S.L.S. and E.E.M.J. Genotyping was performed and coordinated by L.G.C.-C., K.H., A.M.J., M.C., E.E.M.J., A.W. and E. Domingo. A.D. and E. Dermitzakis supplied eQTL data.

Institute of Cancer Research and local collaborators: subject recruitment and sample acquisition to NSCCG were undertaken by S.P. The coordination of sample preparation and genotyping was performed by P.B. Sample preparation and genotyping were performed by A.M.P. and B.O.

Colon Cancer Genetics Group, Edinburgh and local collaborators: subject recruitment and sample acquisition were performed by S.F. and members of the SOCCS and COGS study teams. Sample preparation was coordinated by S.F. Genotyping was performed and coordinated by S.F., E.T., R.A.B. and M.G.D. J.G.D.P. performed the 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: C.G.S., J. Colley, S.I., T.M. and J. Cheadle in Cardiff; I.N., S.T. and L.A.A. in Finland; and P.P. in Cambridge. All other authors undertook sample collection and phenotype data collection and collation in the respective centers.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Pritchard, J.K. & Cox, N.J. The allelic architecture of human disease genes: common disease-common variant…or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
2. Houlston, R.S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–1435 (2008).
3. Haiman, C.A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nat. Genet.* **39**, 954–956 (2007).
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.* **39**, 984–988 (2007).
5. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
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.* **4**, 341–345 (1993).
7. Jones, A.M. *et al.* Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. *Oncogene* **24**, 118–129 (2005).
8. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 (2007).

ONLINE METHODS

Study participants. A summary of all cases and controls in the study is provided in **Supplementary Table 1**. After exclusion of self-reported non-white UK cases and samples of poor quality, VQ58 comprised 1,432 CRC cases (896 males with a mean age of diagnosis of 62.4 years ± 10.7 (standard deviation)) from the VICTOR and QUASAR2 trials. There were 2,697 population control genotypes (1,391 male) from the Wellcome Trust Case Control Consortium 2 (WTCCC2) 1958 birth cohort[9] (also known as the National Child Development Study), which included all births in England, Wales and Scotland during a single week in 1958.

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 described in the **Supplementary Note**. The COIN samples comprised 2,151 cases (1,423 males) derived from the COIN and COIN-B clinical trials of metastatic CRC with a median age of 63 years (range, 22–87 years). 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 or partner controls from the CORGI study collected in the time since collection of 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 cases 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 GWAS cohorts were genotyped using Illumina Hap300, Hap370, Hap240S or Hap550 arrays. The genotyping of the 1958 Birth Cohort and NBS cohort was performed as part of the WTCCC2 study. In the UK2 (NSCCG) and Scotland2 (SOCCS) samples, genotyping was conducted using custom Illumina Infinium arrays according to the manufacturer's protocols. To ensure the quality of the genotyping, a series of duplicate samples was genotyped, which resulted in 99.9% concordant calls.

Other genotyping was conducted using competitive allele-specific PCR KASPar chemistry (KBiosciences Ltd). 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 score <0.25; overall call rate <95%; minor allele frequency <0.01; departure from Hardy-Weinberg equilibrium 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 the 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-northern 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 the VQ58, UK1 and Scotland1 cohorts (as well as the NBS cohort) and subsequently based on combined analysis of all three datasets (**Supplementary Fig. 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 PCA. The first two principal components for each individual were plotted, and any individual not present in the main CEU cluster (that is, having >5% of the principal component distance from the HapMap CEU cluster centroid) was excluded from subsequent analyses.

The adequacy of the case-control matching and the possibility of differential genotyping of cases and controls was formally evaluated using quantile-quantile 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 degree of freedom (d.f.). Deviation of the genotype frequencies in the controls from those expected under Hardy-Weinberg equilibrium was assessed by a χ^2 test with 1 d.f. or a Fisher's exact test where an expected cell count was greater than five.

Association between SNP genotype and disease status was primarily assessed in PLINK v1.07 using allelic and Cochran-Armitage tests (both with 1 d.f.) or by a Fisher's exact test where an expected cell count was greater than five. Genotypic (2 d.f.), dominant (1 d.f.) and recessive (1 d.f.) tests were also performed. The risks associated with each SNP were estimated by allelic, heterozygous and homozygous ORs using unconditional logistic regression, and associated 95% 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 an 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. The 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. We also performed a meta-analysis based on allele dosage (0, 1 or 2) and incorporated age and sex as covariates. 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 six reported associations (data 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 a 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 following formula:

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

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, which was assumed to be 2.2 (ref. 14). The UK2 (NSCCG2) samples were used for this estimation.

Imputation from HapMap2 build 36 was performed using the IMPUTE2 program (see URLs), incorporating the Hap550-typed UK controls from the UK1 (CORGI) study as a reference panel for the VQ58 Hap300 panel genotypes. SNPs were included in the analysis if there were ≤5% missing genotypes and an information score ≥0.5. SNPtest was used to perform the association meta-analysis. PCA was performed using Eigenstrat using CEU, YRI and HCB HapMap samples as references.

Genome coordinates were taken from the NCBI build 36/hg18 (dbSNP b126).

9. Power, C., Jefferis, B.J., 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.* **148**, 54–61 (2006).

10. Tomlinson, I.P. *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* **102**, 447–454 (2010).
11. Petitti, D.B. Coronary heart disease and estrogen replacement therapy. Can compliance bias explain the results of observational studies? *Ann. Epidemiol.* **4**, 115–118 (1994).
12. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).
13. Houlston, R.S. & Ford, D. Genetics of coeliac disease. *QJM* **89**, 737–743 (1996).
14. Johns, L.E. & Houlston, R.S. A systematic review and meta-analysis of familial colorectal cancer risk. *Am. J. Gastroenterol.* **96**, 2992–3003 (2001).