A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer

Citation for published version:

Digital Object Identifier (DOI):
10.1038/srep10442

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Scientific Reports

Publisher Rights Statement:
This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer

Nada A. Al-Tassan1,*, Nicola Whiffin2,*, Fay J. Hosking2, Claire Palles2,†, Susan M. Farrington3, Sara E. Dobbins2, Rebecca Harris2, Maggie Gorman1, Albert Tenesa4,6, Brian F. Meyer2, Salma M. Wakil2, Ben Kinnersley2, Harry Campbell2, Lynn Martin3, Christopher G. Smith5, Shelley Idziaszczyk5, Ella Barclay3, Timothy S. Maughan6, Richard Kaplan3, Rachel Kerr10, David Kerr11, Daniel D. Buchanann2,3, Aung Ko Win13, John Hopper13, Mark Jenkins13, Noralane M. Lindor14, Polly A. Newcomb15, Ian P. Tomlinson3,†, Jeremy P. Cheadle5,† & Richard S. Houlston2,†

Genome-wide association studies (GWAS) of colorectal cancer (CRC) have identified 23 susceptibility loci thus far. Analyses of previously conducted GWAS indicate additional risk loci are yet to be discovered. To identify novel CRC susceptibility loci, we conducted a new GWAS and performed a meta-analysis with five published GWAS (totalling 7,577 cases and 9,979 controls of European ancestry), imputing genotypes utilising the 1000 Genomes Project. The combined analysis identified new, significant associations with CRC at 1p36.2 marked by rs72647484 (minor allele frequency [MAF] = 0.09) near \( \text{CDC42} \) and \( \text{WNT4} \) (\( P = 1.21 \times 10^{-8} \), odds ratio [OR] = 1.21 ) and at 16q24.1 marked by rs16941835 (MAF = 0.21, \( P = 5.06 \times 10^{-8} \); OR = 1.15) within the long non-coding RNA (lncRNA) \( \text{RP11-58A18.1} \) and ~500 kb from the nearest coding gene \( \text{FOXL1} \). Additionally we identified

1Department of Genetics, King Faisal Specialist Hospital and Research Center, P.O.Box 3354, Riyadh 11211, Saudi Arabia. 2Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. 3Wellcome Trust Centre for Human Genetics and NIHR Comprehensive Biomedical Research Centre, Oxford, UK. 4Colon 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. 5Institute of Cancer and Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK. 6The Roslin Institute, University of Edinburgh, Easter Bush, Roslin, EH25 9RG, UK. 7Centre for Population Health Sciences, University of Edinburgh, Edinburgh, EH8 9AG, UK. 8CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Roosevelt Drive, Oxford, OX3 7DQ, UK. 9MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London, WC2B 6NH, UK. 10Oxford Cancer Centre, Department of Oncology, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LE, UK. 11Nuffield Department of Clinical Laboratory Sciences, University of Oxford, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK. 12Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Victoria, Australia. 13Centre for Epidemiology and Biostatistics, The University of Melbourne, Victoria, Australia. 14Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ, USA. 15Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. 16Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON, Canada. 17Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA. *These authors contributed equally to this work. †These authors jointly supervised this work. Correspondence and requests for materials should be addressed to R.S.H. (email: richard.houlston@icr.ac.uk)
Twin studies indicate that heritable factors account for 35% of the variation in risk of developing colorectal cancer (CRC). However, only 5% of CRC can be attributed to the inheritance of high-penetrance mutations in the known genes. Genome-wide association studies (GWAS) conducted primarily in European and but also Asian populations have vindicated the long-held belief that part of the heritable risk of CRC is attributable to common, low-risk variants. These GWAS have provided insights into the biological basis of CRC, highlighting the role of genes within the bone morphogenetic protein signalling pathway and some candidate genes (e.g., SMAD7 and BMP2, BMP4, GREM1). Despite the success of GWAS the risk SNPs so far identified in European populations account for only 8% of the familial CRC risk (Supplementary Table 1). Together with the over-representation of association signals in GWAS strongly suggests that additional risk SNPs remain to be discovered. The statistical power of individual GWAS is limited by the modest effect sizes of the genetic variants and the requirement for a stringent threshold to establish statistical significance in order to avoid type 1 errors. Meta-analysis of GWAS data therefore offers the opportunity to identify new CRC risk loci and provide further insights into tumour biology. Furthermore, imputation of untyped variants in GWAS data using publicly available reference datasets increases the number of variants that can be tested for an association with CRC risk.

To identify new CRC susceptibility loci, we conducted an independent primary scan of CRC using patient samples from the COIN trial and performed a genome-wide meta-analysis with five previously published GWAS. To recover untyped genotypes, thereby maximising the prospects of identifying risk variants, we imputed over 10 million SNPs in the six GWAS datasets, using data from the 1000 Genomes Project as reference (see Materials & Methods for details).

Methods

Primary GWAS. The COIN GWAS was based on 2,244 CRC cases (64% male, mean age 61 years, SD = 10) ascertained through two independent Medical Research Council clinical trials of advanced/metastatic CRC; COIN and COIN-B. Cases were genotyped using Affymetrix Axiom Arrays according to the manufacturer’s recommendations (Affymetrix, Santa Clara, CA 95051, USA), using duplicate samples and sequencing of significantly associated SNPs in a subset of samples to confirm genotyping accuracy. For all SNPs >99% concordant results were obtained. For controls, we made use of Wellcome Trust Case Control Consortium 2 (WTCCC2) Affymetrix 6.0 array data on 2,674 individuals from the UK Blood Service Control Group. Individuals were excluded with: n = 122, discordant sex information (n = 8), classed as out of bounds by Affymetrix (n = 30), duplication or cryptic relatedness (identity by descent >0.185, n = 4), evidence of non-white European ancestry using PCA in conjunction with HapMap samples (n = 130; cut-off based on the minimum and maximum values of the top two principal components of the controls; Supplementary Figure 2). The details of all sample exclusions are provided in Supplementary Figure 3. We excluded SNPs from the analysis with: call rate <95%; different missing genotype rate between cases and controls at P < 10^{-5}; MAF < 0.01; departure from Hardy–Weinberg equilibrium in controls at P < 10^{-5}. The adequacy of the case–control matching and the possibility of differential genotyping of cases and controls were assessed using quantile-quantile (Q–Q) plots of test statistics.

Published GWAS. We made use of five published and previously described GWAS (see Supplementary Methods): UK1 (CORGI) included 940 cases with colorectal neoplasia, Scotland1 (COGS) included 1,012 CRC cases and 1,012 cancer-free population controls, VQ58 comprised 1800 CRC cases and 2,690 population control genotypes from the WTCCC2 1958 birth cohort, CCFR1 comprised 1,290 familial CRC cases and 1,055 controls, CCFR2 included a further 796 cases and 2,236 controls from the Cancer Genetic Markers of Susceptibility (CGEMS) studies of breast and prostate cancer.

The VQ, UK1 and Scotland1 GWA cohorts were genotyped using Illumina Hap300, Hap240S, Hap370, Hap550 or Omni2.5M arrays. 1958BC genotyping was performed as part of the WTCCC2 study on Hap1.2M-Duo Custom arrays. The CCFR samples were genotyped using Illumina Hap1M, Hap1M-Duo or Omni-express arrays. CGEMS samples were genotyped using Illumina Hap300 and Hap240 or Hap550 arrays. After applying the same quality control as that performed for COIN and COIN-B, data on 7,577 CRC cases and 9,979 controls were available for the meta-analysis (Supplementary Figure 1).

The study was conducted in accordance with the declaration of Helsinki. Written informed consent was obtained from all subjects and the study was approved by respective ethical review boards at host institutions.
Statistical and bioinformatic analysis. Analyses were undertaken using R (v3.0.2)24 and PLINK25 software. The association between each SNP and the risk of CRC was assessed by the Cochran–Armitage trend test. ORs and associated 95% CIs were calculated by unconditional logistic regression. Phasing of GWAS SNP genotypes was performed using SHAPEIT (v2.644)26. Prediction of the untyped SNPs was carried out using IMPUTE (v2.3.0)27 based on the data from the 1000 Genomes Project (Phase 1 integrated variant set, v3.20101123)28 as reference. Imputed data were analyzed using SNPTEST (v2.4.1)29.

Association meta-analyses only included markers with info scores >0.4, imputed call rates/SNP > 0.9 and MAFs > 0.01. The fidelity of imputation, as assessed by the concordance between imputed and sequenced SNPs, was examined in a subset of 200 UK cases. Meta-analyses were carried out using META (v2.4-1)30, under an inverse-weighted fixed-effects model using the genotype probabilities from IMPUTE, where a SNP was not directly typed. We calculated Cochran’s Q statistic to test for heterogeneity and the I² statistic to quantify the proportion of the total variation that was caused by heterogeneity – I² values ≥75% are considered characteristic of large heterogeneity31. Associations by sex, age and clinico-pathological phenotypes were examined by logistic regression in case-only analyses. The familial relative risk of CRC attributable to each variant was calculated as detailed by Pharoah et al.32 assuming the overall familial risk of CRC, as shown in epidemiological studies, is 2.33.

To explore epigenetic profiles of association signals, we used ChromHMM34. States were inferred from ENCODE Histone Modification data on the CRC cell line HCT116 (DNase, H3K4me3, H3K4me1, H3K27ac, Pol2 and CTCF)35 binarized using a multivariate Hidden Markov Model.

To examine whether any of the SNPs or their proxies (i.e. r² > 0.8 in 1000genomes CEU reference panel) annotate putative transcription factor binding/enhancer elements we used the CADD (combined annotation dependent depletion) web-server36. We assessed sequence conservation using: PhastCons (<0.3 indicative of conservation), Genomic Evolutionary Rate Profiling (GERP) (~12 to 6, with 6 being indicative of complete conservation) and CADD (>10.0 deemed to be deleterious).

Analysis of TCGA data. To examine for a relationship between SNP genotype and mRNA expression we made use of Tumor Cancer Genome Atlas (TCGA)38 RNA-seq expression and Affymetrix 6.0 SNP data (dbGaP accession number: phs000178.v7.p6) on 223 colorectal adenocarcinoma (COAD) and 75 rectal adenocarcinoma samples using a best proxy where SNPs were not represented directly. Association between normalised RNA counts per-gene and SNP genotype was quantified using the Kruskal-Wallis trend test. The frequency of somatic mutations in CRC was obtained using the CBioPortal for Cancer Genomics39,40 and TumorPortal web servers41.

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 database42. All genes within the LD block containing each tagSNP, or linked to the SNP through functional associations and known gene regulatory elements were included as a Batch query using the NCI-Nature curated data source.

Assignment of microsatellite instability (MSI), KRAS, NRAS and BRAF status in cancers. Tumour MSI status in CRCs was determined using the mononucleotide microsatellite loci BAT25 and BAT26, which are highly sensitive MSI markers. 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)43.

Tumours from the COIN study were screened for mutations in KRAS codons 12, 13, and 61 and BRAF codon 600 by pyrosequencing18. Additionally, KRAS (all three codons), BRAF (codons 594 and 600), and NRAS (codons 12 and 61) were screened for mutations by MALDI-TOF mass array (Sequenom, San Diego, CA, USA)44.

Results

In the primary scan, 2,244 advanced (stage IV) CRC cases ascertained through the Medical Research Council (MRC) trials COIN18 and COIN-B45 were analysed with control data on 2,674 individuals from the WTCCC2 UK National Blood Service Control Group. After applying strict quality control criteria (Materials and Methods), we analysed 234,675 autosomal SNPs for association with CRC risk in 1,950 cases and 2,162 controls. A Q–Q plot of observed versus expected χ²-test statistics showed little evidence for an inflation of test statistics, thereby excluding the possibility of substantive hidden population substructure, cryptic relatedness among subjects or differential genotype calling (inflation factor λ = 1.05; Supplementary Figure 1).

We performed a meta-analysis of our primary scan data with five non-overlapping GWAS case-control series of Northern European ancestry, which have been previously reported (Supplementary Table 2). 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. λGC values46 for the UK1, Scotland1, VQ58, CCFR1 and CCFR2 studies were 1.02, 1.01, 1.01, 1.02 and 1.03 respectively (Supplementary Figure 1). Any ethnic outliers or individuals identified as related were excluded (Supplementary Figure 2).

After quality control procedures, the six GWAS provided data on 7,577 CRC cases and 9,979 controls. To maximise the prospects of identifying novel risk variants, we imputed over 10 million variants using
1000 Genomes Project Pilot data as a reference panel. Q-Q plots for all variants post-imputation did not show evidence of substantive over-dispersion introduced by imputation (Supplementary Figure 1).

**Meta-analysis.** Associations for all 23 established European CRC risk SNPs showed a direction of effect consistent with previously reported studies, with eight of the loci having a $P$-value of $< 5.0 \times 10^{-8}$ (Supplementary Table 3; Fig. 1). Additionally six SNPs previously identified in GWAS in Asian populations as determinants of CRC risk showed evidence for an association in this meta-analysis; albeit at varying degrees of significance ($P$-values ranging from $3.64 \times 10^{-2}$ to $1.71 \times 10^{-3}$; Supplementary Table 3); thereby providing support for trans-ethnic effects.

Excluding SNPs (including those correlated with $r^2 > 0.8$) mapping to the risk loci, five variants in distinct regions of linkage disequilibrium (LD) were associated with CRC at $P < 1.0 \times 10^{-7}$ (Table 1; Fig. 1).

We assessed the fidelity of imputation in 200 UK cases by comparing imputed genotypes with those obtained by sequencing. For the three common variants (MAF $> 0.05$), rs72647484, rs16941835 and

**Figure 1.** Genome-wide $P$-values ($-\log_{10}P$, y-axis) plotted against their respective chromosomal positions (x-axis). Known regions attaining genome-wide significance (i.e. $P = 5.0 \times 10^{-8}$) are labelled with their chromosomal location. Variants in grey lie in novel regions that reach the significance threshold level ($P = 1.0 \times 10^{-7}$) required for variants to be analysed further in this study. Variants in black lie in novel regions attaining genome-wide significance.

**Table 1.** Summary statistics for variants showing an association with CRC risk at $P < 1.0 \times 10^{-7}$. For each variant shown along with meta-analysis test statistics are the $P$-values from the six individual studies and imputation Information scores. Risk alleles are given in bold. INFO, imputation Information score; $P$-het, $P$-value of heterogeneity between studies; RAF, risk allele frequency.
rs10904849 which each had imputation info scores >0.9 there was high correlation between imputed and directly typed genotype ($r^2 = 0.98, 1.00$ and 0.99, respectively). For the rare variant rs79900961 (MAF = 0.016), the correlation was poor ($r^2 = 0.60$). The call rate for the rare Indel on chromosome 5q15 (rs202110856) in the sequencing data was only 71% and both imputed heterozygotes were sequenced as homozygous reference. Therefore, only the three common variants at 1p36.12, 10p13 and 16q24.1 were subject to further analyses.

In the combined analysis of the six GWAS datasets, rs72647484, which maps to chromosome 1p36.12 (22,587,728 bps; NCBI build 37), showed the strongest evidence for association with CRC ($P = 1.21 \times 10^{-8}$; $P_{\text{het}} = 0.33$, $P_{\text{I}} = 14$%; Fig. 2a). rs72647484 maps within a 300 kb block of LD encompassing WNT4 (wingless-type mmmtv integration site family, member 4; MIM 603490) and CDC42 (cell division cycle 42, MIM 116952; Fig. 3a). The second strongest association was provided by rs16941835 ($P = 5.06 \times 10^{-8}$; $P_{\text{het}} = 0.40$, $P_{\text{I}} = 3$%; Fig. 2c) which localises to the long non-coding RNA (lncRNA) RP11-58A18.1 at chromosome 16q24.1 (86,659,720 bps; NCBI build 37) within a 65 kb region of LD (Fig. 3c). The nearest coding gene, ~500 kb away, is the transcription factor FOXL1. The third strongest association was provided by rs10904849 ($P = 7.01 \times 10^{-8}$; $P_{\text{het}} = 0.83$, $P_{\text{I}} = 0$%; Fig. 2b) which localises to chromosome 10p13 (16,997,266 bps; NCBI build 37) within intron 31 of the gene encoding cubulin (CUBN; alias intrinsic factor-cobalamin receptor [IFCR], MIM 602997; Fig. 3b).

**Bioinformatic analysis of risk variants.** To gain insight into the biological basis of the associations we analysed publicly available RNA-seq expression and SNP data from TCGA on 223 colon and 75 rectal cancers using rs10904850 and rs7247753 as proxies for rs10904849 ($r^2 = 0.97$; $D^\prime = 1.00$) and rs72647484 ($r^2 = 0.64$; $D^\prime = 0.89$) 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 Table 4).

We examined whether any of the SNPs or their proxies (i.e. $r^2 > 0.8$ in 1000 Genomes CEU reference panel) lie in putative transcription factor binding/enhancer elements and derived GERP and PhastCons scores to assess sequence conservation at these positions (Supplementary Table 5). rs16941835 maps to a regulatory feature with histone modification suggestive of an enhancer element. rs10904852, in LD with rs10904849 ($r^2 = 0.95$, $D^\prime = 1.00$) is conserved (GERP and PhastCons scores of 1.20 and 0.47 respectively) with CADD score of 115.3. A moderate CADD score (8.21) was associated with rs72647484 (22,590,125 bps) which is strong LD with rs72647489 ($r^2 = 0.93$, $D^\prime = 1.00$). Six proxy SNPs in LD with rs16941835 showed some evidence of transcription factor binding (Supplementary Table 5). We made use of TCGA data to examine the frequency of somatic mutation of CDC42, WNT4, FOXL1 or CUBN in CRC. None of these genes showed evidence of significant somatic mutation. Next, we conducted pathway analysis to determine whether any genes mapping to the three newly identified regions act in pathways already over-represented in GWAS. Pathways containing three or more genes are shown in Supplementary Table 6. While this analysis identifies the BMP-signalling pathway as expected, we conducted pathway analysis to determine whether any genes mapping to the three newly identified regions act in pathways already over-represented in GWAS. Pathways containing three or more genes are shown in Supplementary Table 6. While this analysis identifies the BMP-signalling pathway as expected, no catalogued pathways were discernible involving genes mapping to any of the newly identified regions.

It is increasingly recognized that some genetic variants can have pleiotropic effects, influencing the risk of more than one cancer type. To explore the possibility that rs72647484, rs10904849 or rs16941835 affects the risk of other malignancies, we examined the association with lung cancer, acute lymphoblastic leukaemia, multiple myeloma, glioma and meningioma using data from previously reported GWAS. However, for these cancers, there was no evidence of rs72647484, rs10904849 or rs16941835 (or correlated SNP $r^2 \geq 0.8$) being associated with tumour risk (i.e. $P > 0.05$).

Finally, the relationship between clinico-pathological variables (sex, age at diagnosis, family history of CRC, tumour stage or microsatellite instability (MSI), KRAS-mutant status and BRAF-mutant status) and genotype at rs72647484, rs10904849 and rs16941835 was assessed by case-only logistic regression (Supplementary Table 7). There was evidence of a relationship between rs72647484 and KRAS-mutant status ($P = 0.03$) with the T risk allele associated with KRAS-mutant CRC; however this finding was not significant after accounting for multiple testing. None of the other SNPs showed any association with any of the clinico-pathological variables examined (i.e. $P > 0.05$).

**Discussion**

We have provided evidence supporting the existence of new susceptibility loci for CRC at 1p36.12, 10p13 and 16q24.1. The 1p36.12 association implicates WNT4 and/or CDC42 as possible determinates of CRC risk. WNT4 is part of a family of structurally related genes that encode cysteine-rich secreted glycoproteins that act as extracellular signalling factors. WNT4, WNT14, and WNT16 may play redundant roles in signalling through the CTNNB1-mediated canonical Wnt-pathway which is known to play a central role in colorectal tumorigenesis. Additionally, WNT4 signalling appears to play a pivotal role during organogenesis, acting as an autoinducer of mesenchyme-to-epithelial transition. Inactivating germline mutations in WNT4 cause Mullerian aplasia and hyperandrogenism (MIM 158330) and are responsible for the autosomal recessive SERKAL syndrome (Sex Reversal and Kidney, Adrenal, and Lung dysgenesis; MIM 611812). A priori dysfunction of either WNT4 or CDC42 could be the biological basis for the 1p36.12 association. Cdc42 is a Ras-related GTP-binding protein with roles in establishment of cell polarity, regulation of cell morphology, motility, and cell cycle progression in mammalian cells, and malignant transformation. Notably, Cdc42 regulates the actin cytoskeleton through activation of WASP.
proteins and cell polarity through GSK3-beta and APC. Rho-GTPase signalling has a documented role in the development of CRC\textsuperscript{54}. Activation of Rho GTPase Cdc42 promotes adhesion and invasion in CRC\textsuperscript{55} and targeting Cdc42 with AZA197 suppresses primary colon cancer growth and prolongs survival in a xenograft model through down regulation of PAK1\textsuperscript{56}.

Since rs10904849 is intronic to CUBN and the region of LD does not encompass any other genes or transcripts, there is a high likelihood that the functional basis of the 10p13 association is mediated

Figure 2. Forest plot of the odds ratios for the association between rs72647484, rs16941835, rs10904849 and CRC. Studies were weighted according to the inverse of the variance of the log of the OR calculated by unconditional logistic regression. Horizontal lines: 95% confidence intervals (95% CI). Box: OR point estimate; its area is proportional to the weight of the study. Diamond (and broken line): overall summary estimate, with confidence interval given by its width. Unbroken vertical line: null value (OR = 1.0).
Figure 3. Regional plot of association results and recombination rates for the (a) 1p36.12, (b) 10p13 and (c) 16q24.1 risk loci. Association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates within the loci at 1p36.12 (a), 10p13 (b) and 16q24 (c). For each plot, $-\log_{10} P$ values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The top imputed SNP in each combined analysis is shown as a large triangle and is labelled by its rsID. The colour intensity of each symbol reflects the extent of LD with the top SNP: white ($r^2 = 0$) through to dark red ($r^2 = 1.0$), with $r^2$ estimated from the 1000 Genomes Phase 1 data. Genetic recombination rates (cM/Mb), are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to each region of association. The lower panel shows the chromatin state segmentation track (ChromHMM).
through CUBN. Cubilin is the intestinal receptor for the endocytosis of intrinsic factor-vitamin B12 and a receptor in epithelial apoA-I/HDL metabolism\(^5\). Additionally cubilin is an important co-receptor in the endocytic pathway for retrieval of 25(OH)D3-DBP complexes by megalin-mediated endocytosis in the kidney\(^6\). Germline mutations in CUBN cause recessive megaloblastic anemia-1 (MGA1; MIM 261100). It is conceivable that common genetic variance in CUBN, while being insufficient to cause a "MGA type phenotype" would have physiological effects by virtue of long term effect on the cellular bioavailability of B12. Although it is entirely speculative, as epidemiological studies have yet to convincingly establish levels of B12 as a risk factor for CRC\(^9\),\(^10\), its role in DNA biosynthesis makes genetically determined variation in B12 availability a plausible candidate for a role in the development of CRC.

LncRNAs are regulators of transcription and are increasingly recognised as playing a role in cancer biology. While there is currently no evidence to implicate the RP11-58A18.1 lncRNA in CRC, lncRNAs CCAT1 and CCAT2 probably do play such roles\(^61\),\(^62\), and it is entirely plausible that the impact of variation at 16q24.1 on risk is mediated through similar long range effects.

One of the reasons for the failure to identify these CRC-loci previously is that, in addition to the issue of study power, they were not optimally tagged by SNPs featured on many commercial arrays. The power of our study to detect the major common loci conferring risks of 1.2 or greater (such as the 18q24 variant) was high. Hence, it is very unlikely there are additional CRC SNPs with similar effects for alleles with frequencies >0.2 in populations of European ancestry.

In this study, we have only considered SNPs showing evidence of an association with a stipulated P-value threshold of <1 × 10\(^{−7}\). There exist, however, many variants with P-values just above this threshold which may also warrant investigation in a further study (Fig. 1). Hence further efforts to expand the scale of GWAS meta-analyses, in terms of both sample size and SNP coverage, and to increase the number of SNPs taken forward to large-scale replication, may identify additional variants for CRC.

In conclusion, we have provided evidence for 3 new susceptibility loci for CRC. Our data also provide further evidence for the value of meta-analysis and the value of imputation as a means of enhancing the detection of novel risk loci thereby extending the utility of GWAS data.

References

We thank the patients and their families who participated in COIN and COIN-B, and gave

Acknowledgements
The COIN and COIN-B trials were funded by Cancer Research UK and the Medical Research Council and were conducted with the support of the National Institute of Health Research Cancer Research Network. COIN and COIN-B translational studies were supported by the Bobby Moore Fund from Cancer Research UK, Tenovus, the Kidani Trust, Cancer Research Wales and the National Institute for Social Care and Health Research Cancer Genetics Biomedical Research Unit (2011–2014) (all awarded to J.P.C.). We thank the patients and their families who participated in COIN and COIN-B, and gave
their consent for this research, and the investigators and pathologists throughout the UK who submitted samples for assessment. N.A.A., B.F.M. and S.M.W. were funded and supported by KFSHRC. At the Institute of Cancer Research, the work was supported by Cancer Research UK (C1298/A8362 - Bobby Moore Fund for Cancer Research UK). Additional support was provided by the National Cancer Research Network and the NHS via the Biological Research Centre of the National Institute for Health Research at the Royal Marsden 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). 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), Support from the European Union (FP7/207-2013) under grant 238236 and FP7 collaborative project SYSCOL and COST Action BM1206 in the UK is also acknowledged. The work of the Colon Cancer Family Registry CFR was supported by grant UM1 CA167551 from the National Cancer Institute, National Institutes of Health and through cooperative agreements with members of the Colon CFR and Principal Investigators. Collaborating centers include the Australasian Colorectal Cancer Family Registry (U01/U24 CA097735), the USC Colorectal Cancer Family Registry (U01/U24 CA074799), Mayo Clinic Cooperative Familial Registry for Colon Cancer Studies (U01/U24 CA074800), Ontario Registry for Studies of Familial Colorectal Cancer (U01/U24 CA074783), the Seattle Colorectal Cancer Family Registry (U01/U24 CA074794), and the University of Hawaii Colorectal Cancer Family Registry (U01/U24 CA074806). The Colon CFR GWAS was supported by funding from the National Cancer Institute, National Institutes of Health (U01 CA122839 and R01 CA143237 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/. 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 which constitute the TCGA research network can be found at http://cancergenome.nih.gov/. 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 J.P.C. The manuscript was drafted by R.S.H. with input from I.P.M.T., M.G.D. and J.P.C. All authors had access to data, analysis and had opportunity to contribute to drafting the manuscript. J.P.C. initiated and directed the GWAS of COIN and COIN-B. S.I. was responsible for COIN blood DNA extractions and quantification, R.H. for aliquoting and manifest preparation for Axiom genotyping and C.G.S. for somatic profiling of COIN tumours. N.A.A. and B.F.M. coordinated, and S.M.W. performed, the genotyping of COIN and COIN-B samples on the Axiom platform. T.S.M. was CI/co-CI of COIN and COIN-B, and R.K. provided access to linked clinico-pathological data. Cleaning of the COIN/COIN-B genotyping data and all statistical and bioinformatic analyses were conducted by N.W. and F.H., with contributions from S.D. and B.K., under the supervision of R.S.H. ICR - 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., R.K., D.K., 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., 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 and analysis was performed and coordinated by S.M.F., C.H., M.G.D. and A.T. For the colon CFR datasets - D.D.B., A.K.W., J.H., M.J., F.S., G.C., S.G., N.L., P.N. and D.C. performed sample ascertainment and analysis.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep
Competing financial interests: The authors declare no competing financial interests.


This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/