Published in final edited form as:
*Nat Genet*. 2009 December ; 41(12): 1330–1334. doi:10.1038/ng.483.

# Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region

**The UK IBD Genetics Consortium** and **The Wellcome Trust Case Control Consortium 2**

## Abstract

Ulcerative colitis (UC) is a common form of inflammatory bowel disease with a complex aetiology. As part of the Wellcome Trust Case Control Consortium 2, we performed a genome-wide association scan for UC in 2361 cases and 5417 controls. Loci showing evidence of association at $P < 1 \times 10^{-5}$ were followed up by genotyping in an independent set of 2321 cases and 4818 controls. We find genome-wide significant evidence of association at three new loci, each containing at least one biologically relevant candidate gene, on chromosomes 20q13 (*HNF4A*; $P = 3.2 \times 10^{-17}$), 16q22 (*CDH1* and *CDH3*; $P = 2.8 \times 10^{-8}$) and 7q31 (*LAMB1*; $3.0 \times 10^{-8}$). Of note, *CDH1* has recently been associated with susceptibility to colorectal cancer, which is an established complication of longstanding UC. The new associations suggest that changes in the integrity of the intestinal epithelial barrier may contribute to the pathogenesis of UC.

Genetic epidemiological data clearly implicate inherited susceptibility in the pathogenesis of UC and Crohn's disease (CD), which represent the two common forms of inflammatory bowel disease (IBD) and together affect at least 1 in 250 of the Northern European population.1 Notwithstanding recent therapeutic advances, disease-related morbidity in ulcerative colitis continues to be high. Recognized complications of severe disease refractory to medical therapy include colectomy, often as an emergency, in 15-20% of patients, as well as colorectal cancer2.

Substantial progress has been made in understanding IBD pathogenesis in recent years. In genetically susceptible individuals it appears that a dysregulated mucosal immune response to commensal enteric bacteria predisposes to chronic, relapsing intestinal inflammation which is the hallmark of IBD.3 Clinical features combined with epidemiological evidence have long suggested that CD and UC are related polygenic diseases. This has recently been corroborated by the results of genetic association studies, which have highlighted both disease-specific loci and others which are shared between UC and CD. For example, while genetically-determined defects in the handling of intracellular bacteria (*NOD2* and the autophagy genes *ATG16L1* and *IRGM*) are specific to CD, multiple components in the Th17 pathway (*IL23R*, *IL12B*, *JAK2*, *STAT3*) are associated with both CD and UC.4-12

Until recently most attention had focused on CD, with genome-wide association (GWA) studies and subsequent meta-analysis yielding more than 30 confirmed CD susceptibility loci.4, 6, 7, 10-12 In addition to the longstanding known association in the MHC,13 the first

Correspondence to: JCB (barrett@sanger.ac.uk), CCAS (chris.spencer@well.ox.ac.uk).

Author contributions

JL, CL, NP, AP, EW, KP, HZ, HD, ERN, DM, KB, TE, LC were involved in establishing DNA collections, and/or assembling phenotypic data; CL, AP, EW, DM, HD, AL, CM, JeS, DPJ, CE, TA, JCM, JS, MP recruited patients; WN, CE, TA, JCM, JS, MP, CGM supervised clinical and laboratory work; WTCCC2 DNA, Genotyping, Data QC and Informatics group executed GWAS sample handling, genotyping and QC; WTCCC2 Data and Analysis group, JCB, CAA performed statistical analyses; JCB, JL, CL, CCAS, CAA, TA, PD, JS, MP, CGM contributed to writing the manuscript. WTCCC2 Management Committee conceived and oversaw the design and execution of the GWAS. WTCCC2 group memberships are specified in the full author list.


Europe PMC Funders Author Manuscripts

GWA scans in UC reported associations at *IL23R*, *IL10* and loci on chromosomes 1p36 and 12q15 which meet accepted genome-wide significance thresholds.14, 15

As part of the Wellcome Trust Case Control Consortium 2 (WTCCC2) study of 15 complex disorders and traits, we report here the results of the largest GWA scan in UC to date. All study subjects were UK residents of white, European ancestry; clinical data are presented in Table 1. Cases and controls were genotyped on the Affymetrix 6.0 array. After application of quality control filters (see Methods), we analysed GWA data from 2361 individuals with UC and 5417 controls (Figure 1). An initial analysis revealed 24 distinct loci (comprising 156 SNPs) which showed evidence of association at $P < 1\times10^{-5}$. Sixteen of these had not been previously reported, and were followed up by genotyping the most strongly associated SNP from each locus using the Sequenom iPlex platform in an independent panel of 2321 UC cases and 4818 controls. Three new loci showed evidence for association at $P < 5 \times 10^{-8}$ in the combined panel, with three further new loci showing nominal ($P < 0.05$) replication (Table 2 and Figure 2). We describe these loci below and highlight the most plausible candidate gene for each, recognizing that fine mapping and functional studies are required to define causal variants and identify the gene from which each signal arises. A list of all loci for which replication was attempted is shown in Supplementary Table 1.

The most significant new association was seen at rs6017342 (GWA scan $P = 3.2 \times 10^{-13}$; combined GWA and replication $P = 8.5 \times 10^{-17}$), which maps within a recombination hotspot on chromosome 20q13 containing the 3′ untranslated region (UTR) of just one gene, *HNF4A*. The SNP rs6017342 itself maps 5kb distal to the 3′UTR. Although within an expressed sequence tag DB076868, this has been detected in just a single testis cDNA library and does not encode a significant open reading frame. The region contains two small blocks of sequence that are conserved in mammals and may include regulatory sequences affecting the expression of surrounding genes. Since rs6017342 is located within a recombination hotspot, there are few known SNPs in strong linkage disequilibrium ($r^2 > 0.5$) with it; there are none on the Affymetrix chip used in this study or on the Illumina chips used in previous studies. As the evidence for this association rests on this single SNP, we subjected these data to careful scrutiny; genotype cluster plots for this SNP showed clear resolution of the 3 genotype classes (Supplementary Figure 1), with 99.3% completeness of genotypes within this dataset.

Rare *HNF4A* mutations account for approximately 4% of UK cases of maturity-onset diabetes of the young (MODY),16 a monogenic form of diabetes mellitus characterized by autosomal dominant inheritance, young age of onset, pancreatic b-cell dysfunction and sensitivity to sulphonylureas. Common variants of *HNF4A* influence predisposition to Type II diabetes (rs2144908)17 and dyslipidaemia (rs1800961).18 The UC associated SNP, rs6017342 is not in LD with either of these 2 common variants, nor did it not show association in our study of CD (P=0.92).3

*HNF4A* encodes the transcription factor hepatocyte nuclear factor 4 α which regulates the expression of multiple components within all three key compartments of the cell-cell junction, namely the adherens junction, the tight junction and the desmosome.19 Such cell-cell junctions are fundamental to epithelial organization and barrier function. HNF4α also plays a key role in the development of the embryonic mammalian gastrointestinal tract. Previous studies demonstrated that mice with targeted deletion of *HNF4α* in epithelial cells of the foetal colon die perinatally. Histological analysis of colonic tissue recovered during late development (E18.5) demonstrated absent crypt formation, reduced epithelial cell proliferation and defective goblet cell maturation.20 In order to explore the role of *HNF4α* in murine intestinal inflammation, Ahn and colleagues circumnavigated the embryonic lethality of *Hnf4α*$^{-/-}$ mice by generating a conditional model of intestinal *Hnf4α* deletion.

21 These *Hnf4a*$^{\Delta IEpC}$ mice (floxed *Hnf4a* driven by the *villin* promoter) developed increased epithelial permeability and a markedly more severe colitis following dextran sodium sulphate (DSS) challenge, than their wild-type littermates.21 The same investigators provided preliminary evidence for dysregulated *HNF4A* gene expression in the intestinal epithelium in Crohn's disease and in ulcerative colitis,21 a finding which now merits detailed re-exploration.

Significant association was also seen for a locus on chromosome 16q22, with the strongest signal at rs1728785 (GWA scan $P = 1.8 \times 10^{-5}$; combined GWA and replication $P = 2.8 \times 10^{-8}$). The interval bounded by recombination hotspots spans 411 kb and encodes several genes. Among the strongest candidates for UC susceptibility is *CDH1* which encodes E-cadherin. This transmembrane glycoprotein is one of the main components of the adherens junction and a key mediator of intercellular adhesion in the intestinal epithelium. It also plays a key role in epithelial restitution and repair following mucosal damage and expression of *CDH1* is known to be significantly reduced in areas of active UC.22

Given the well-recognised association between UC and colorectal cancer,2 the observation of correlated association signals at the *CDH1* locus in both diseases is striking. Thus variants in LD ($r^2 = 0.5$) with the most strongly UC associated SNP in our study were recently identified in a GWA scan meta-analysis to be associated with colorectal cancer susceptibility23; conversely, we find that a perfect proxy for the most associated SNP in the colorectal cancer study is also associated with UC ($P = 8 \times 10^{-4}$). This locus did not show association with CD in a large international GWA meta-analysis of CD (P = 0.549)6 (Supplementary Table 2). However, evidence for association of *CDH1* with CD was reported recently in the Canadian population using a candidate gene approach,24 and the CD associated SNPs resulted in a truncated E-cadherin protein *in vitro* which accumulated in the cytoplasm and led to disorganized epithelial architecture.24

Of great potential relevance is the evidence that *HNF4A* and E-cadherin co-operate to maintain epithelial barrier integrity in the intestine. In experiments focused on the liver, *HNF4a* knockout mice failed to express E-cadherin,19 while in the gut E-cadherin dependent cell-cell contact was found to be critical in determining the amount and binding activity of nuclear *HNF4a*. This in turn affected the expression of several genes including *ApoA-IV*,25 an anti-inflammatory protein known to inhibit experimental colitis.26

The third newly confirmed UC susceptibility locus was a region on chromosome 7q31, previously suggested by a recent North American GWA scan.14 In the current study the peak association was seen at rs886774 (GWA scan $P = 4.8 \times 10^{-7}$; combined GWA and replication $P = 3.0 \times 10^{-8}$). A strong positional candidate gene at this locus is *LAMB1*, encoding the laminin beta 1 subunit. Laminins are heterotrimers; the beta-1 light chain is present in laminins-1 -2 and -10. Laminins are expressed in the intestinal basement membrane, and play a key role in anchoring the single-layered epithelium; expression is known to be down-regulated in UC.27 rs886774 was not associated with CD in the meta-analysis.5 (Supplementary Table 2)

Two other loci previously implicated in UC-related phenotypes showed strong (but not genome-wide significant) association with UC. These comprise a SNP previously associated with osteoporosis28 (rs7524102 on chromosome 1p36, combined GWA and replication $P = 3.1 \times 10^{-7}$) and a SNP nearby (though not in LD with) a marker known to be associated with psoriasis29 (rs9548988 on 13q.13, combined GWA and replication $P = 2.7 \times 10^{-7}$).

In addition to the novel loci described above, our GWAS detected strong association at established UC loci such as the MHC, *IL23R*, 3p21/*MST1* and *NKX2-3* (one tailed *P* values in the direction of the previously reported association in Table 3). We also provide robust

confirmation of two UC loci reported recently in genome-wide scans, the *IL10* locus11 and the *OTUD3/PLA2G2E* locus12 on chromosome 1q31 and 1p36 respectively. Also of interest is our finding that the *PSMG1* locus on chromosome 21, which has previously been associated with pediatric-onset IBD,30 is likely to contribute specifically to disease susceptibility in UC. Variable degrees of support were obtained for some previously reported UC loci, including *ECM1*, *CARD9*,31 *KIF21B*/chromosome 1q32, and *JAK2*/chromosome 9p24, but weaker support for other loci such as *IL2/IL21*,32 *IL12B* and 12q15 (Table 3). Some of the UC loci are clearly associated with CD, while others are not, or have not been tested (Supplementary Table 2). We also tested for epistatic interaction among all pairwise combinations of these loci (both previously described and new) but found none.

This is the first report of a new series of GWA scans undertaken by the WTCCC2 consortium. We have identified three new susceptibility loci for UC, and provide the first genetic link between UC and colorectal cancer. Each of the strongest new association intervals that we have identified contains respectively *HNF4A*, *CDH1* and *LAMB1* as the most plausible positional candidate genes, thus providing further evidence for the re-emerging concept that altered epithelial barrier function may be a key factor in UC pathogenesis.8 Indeed, this is the first time that variants within genetic loci encoding such epithelial barrier genes have shown association with IBD at stringent genome-wide significant thresholds. Fine mapping and functional studies are clearly required to investigate this connection further, but our study provides strong scientific justification for the exploration of new therapeutic targets relevant to epithelial barrier function.

## METHODS

### Subjects

**Cases—**A total of 5319 unrelated patients of white, European, non-Jewish ancestry with a diagnosis of ulcerative colitis established using standard endoscopic, radiological and histological criteria, were recruited from ten centres within the United Kingdom (Cambridge, Oxford, London, Newcastle, Sheffield, Edinburgh, Dundee, Manchester, Torbay and Exeter, Supplementary Table 4). All patients provided written consent and either a sample of blood or saliva, from which DNA was extracted according to standard protocols. Research Ethics Committee approval was obtained prior to sample collection (Cambridge, Oxford, London, Newcastle, Sheffield, Edinburgh, Dundee, Manchester, Torbay and Exeter Local Research Ethics Committees). After QC (see below), we analyzed a total of 4682 samples, which were divided between the discovery panel (2361 samples) and replication panel (2321 samples).

**Controls—**A total of 10,235 control DNA samples from 3 sources passed our QC filters (see below). 5417 samples of the WTCCC2 common control set were used for the GWA experiment. This comprised 2675 healthy blood donors recruited from the United Kingdom Blood Service (UKBS), and 2742 samples from the 1958 Birth Cohort (1958BC) obtained from EBV-transformed cell lines from individuals born in England, Wales and Scotland during one week in 1958. The 4818 samples used as controls for the replication cohort were recruited from the Wellcome Trust-funded People of the British Isles (PoBI) DNA collection, obtained from rural populations throughout the British Isles, and from a further independent set of DNA samples obtained from 1958BC. All of the control samples used were from individuals with self-reported Caucasian ethnicity.

A summary of patients and controls is shown in Table 1 and Supplementary Table 4.

**DNA sample preparation:** Genomic DNA for all cases was shipped to the Sanger Institute, Cambridge. DNA quality plus subject identity were validated using the Sequenom iPLEX

Europe PMC Funders Author Manuscripts

assay designed to genotype 4 gender SNPs and 26 SNPs present on the Affymetrix array. DNA concentrations were quantified using a PicoGreen assay (Invitrogen) and an aliquot assayed by agarose gel electrophoresis. A DNA sample was considered to pass quality control if the original DNA concentration was ≥50ng/ul, the DNA was not degraded, the gender assignment from the iPLEX assay matched that provided in the patient data manifest and genotypes were obtained for over 65% of the SNPs on the iPLEX.

## GWA Genotyping

Samples were genotyped at Affymetrix's service laboratory on the Genome-Wide Human SNP Array 6.0. For all samples passing Affymetrix's laboratory QC, raw intensities (from the .CEL files) were renormalized within collections using CelQuantileNorm (see http://outmodedbonsai.sourceforge.net/). These normalized intensities were used to call genotypes with an updated version of the Chiamo software (see www.stats.ox.ac.uk/~marchini/software/gwas/chiamo.html), adapted for Affymetrix 6.0 SNP data. The Chiamo algorithm simultaneously calls genotypes for individuals in several collections; here it was applied to 15,068 individuals from five collections genotyped as part of the WTCCC2. Chiamo generates posterior probabilities for each of the three possible genotypes plus a fourth class of outliers. Our analyses use thresholded genotypes: for each individual, if one genotype had posterior probability greater than 0.9, this was set as the genotype for that individual, otherwise the genotype was set to be missing. After applying the QC filters described below, this threshold led to a study-wide level of missing data of 0.20%.

An overlapping set of 4830 controls were also genotyped on the Illumina 1.2M chip as part of a separate WTCCC2 project, and the 50,000 SNPs which are shared between that platform and the Affymetrix 6.0 (used in this study) were used to evaluate genotype accuracy. For the same QC thresholds and similar levels of missing data, discordance between Chiamo and Illuminus, which we regard as an upper bound on genotyping error rate, was 0.05857% for 1958BC and 0.07476% for UKBS.

We compared Chiamo to Birdsuite (the default Affymetrix calling algorithm applied on a plate-by-plate basis as recommended in 33) by making genotype calls at different confidence thresholds, and then plotting the fraction of calls made against concordance with the Illumina genotypes (Supplementary Figure 2). The general trend is that, when matched for the proportion of missing data, Chiamo has slightly higher concordance than Birdsuite. We are therefore confident that Chiamo is an acceptable alternative to Birdsuite.

## Replication Genotyping

In the replication stage, genotyping was carried out at the Sanger Institute using the Sequenom iPLEX Gold assay. For one locus, the most associated SNP could not be genotyped with this technology, so a perfect ($r^2 = 1$ in all HapMap populations) proxy was used instead. 19 SNPs (including 3 gender markers) were typed in a multiplex reaction; 15 passed experimental QC (one SNP with Hardy-Weinberg P value $< 1 \times 10^{-6}$ was discarded). Samples with > 20% missing genotypes (n=300) were excluded; these samples are not included in the tallies in Table 1.

## Quality Control

**Samples—**As is now standard practice for GWAS studies, we excluded sets of individuals whose genome-wide patterns of diversity are outliers compared to the bulk of those in the study, and SNPs where there is evidence that genotype calls do not provide precise estimates of genotype frequencies. Ignoring individuals and SNPs in this way throws away data gained at some expense, but because they typically violate assumptions underpinning

standard tests for association, the payback in terms of increased accuracy for these tests can be substantial.

In order to try to obtain the maximally powerful set of samples and SNPs we attempted to refine some standard QC practices. For all individuals we explicitly model the data as a mixture of “normal” and “outlier” individuals for each of ancestry, missing data and heterozygosity, and sex assignment.34 We fit each model in a Bayesian framework, and exclude individuals whose posterior probability of belonging to the outlier class was above 0.5. This approach replaces (and we believe improves upon) the traditional concept of fixed exclusion thresholds for parameters such as call rate, heterozygosity and ancestry. In total 413 case individuals and 567 control individuals were excluded from the analyses (Supplementary Table 3).

To assess relatedness amongst study individuals we compared each individual with the 100 individuals they were most closely related to (on the basis of genome-wide levels of allele sharing) and used a hidden Markov Model (HMM) to decide, at each position in their genome, whether the two individuals shared 0, 1, or 2 chromosomes identical by descent. This allows more refined assessment of the relatedness between individuals than do genome-wide sharing statistics (for example, parent-child relationships can be distinguished from siblings). We obtained a set of individuals with IBD < 5% by iteratively removing the member of each pair of putatively related individuals with more missing genotypes.

**SNPs—**For each SNP we considered a measure of the (Fisher) information carried by the genotype calls for the underlying allele frequency. Informally, this will decrease as the number of individuals with low posterior probabilities for the most likely call increases, and it can be thought of as a more refined measure of both missing data levels and minor allele frequency (Supplementary Figure 3). The measure is calculated automatically by the program SNPtest (www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html). SNPs were removed if this information measure was below 0.98, or if the estimated MAF was below 0.01% (both calculated on the combined case-control data). 14.7% of SNPs were removed by these criteria. Again, this approach appears to offer advantages over conventional SNP filters, in excluding fewer SNPs for the same level of improved data quality. Because associated SNPs are expected to be enriched in the tiny fraction of poorly performing markers on these chips, we subsequently examined 155 cluster plots for SNPs with $p < 1\times10^{-5}$, and excluded 16 from further analysis as likely genotyping errors.

Supplementary Figure 4 provides QQ plots for the post-QC comparison of our two control collections, and for association statistics based on the post-QC trend test comparing cases and the combined control set. Both visual inspection, and the inflation statistic for each ($\lambda$ = 1.037 and $\lambda$ = 1.079 respectively), suggest that the QC filtered data provides a good basis for association analyses.

## Statistical Methods

We report p-values from 1-d.f. Cochran-Armitage tests for trend as implemented in the software SNPTEST and PLINK.35 We also performed 2-d.f. genotypic tests to verify that none of our associations show significant deviation from a multiplicative model, and two marker logistic regressions to test for epistasis between associated markers. Effect size estimates are based on replication samples only, and represent per-allele increase of risk in a multiplicative model.

### URLs

Affymetrix, http://www.affymetrix.com/Auth/support/downloads/manuals/genotyping_console_manual.pdf; CelQuantileNorm, http://outmodedbonsai.sourceforge.net; Chiamo, www.stats.ox.ac.uk/~marchini/software/gwas/chiamo.html; SNPtest, www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html www.peopleofthebritishisles.org

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The principal funding for this study was provided by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project. We thank all subjects who contributed samples, and consultants and nursing staff across the UK who helped with recruitment of study subjects. We also thank Sami Bertrand, Jackie Bryant, Sarah L. Clark, Jen S. Conquer, Thomas Dibling, Stephen Gamble, Clifford Hind, Alicja Wilk, Claire R. Stribling, Sam Taylor, Julia C. Wyatt of the Wellcome Trust Sanger Institute's DNA Logistics and Genotyping Facility for technical assistance. Case collections were supported by the National Association for Colitis and Crohn's disease (NACC), the Wellcome Trust, the Medical Research Council UK, the Guy's and St Thomas' Charity, the Clinical Research Facility at the Peninsular College of Medicine and Dentistry, Exeter, the Torbay Hospital Medical Fund and the Evelyn Trust. We also acknowledge support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre awards to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London, the Cambridge University Hospitals NHS Foundation Trust in partnership with the University of Cambridge School of Clinical Medicine and the Central Manchester Foundation Trust in partnership with the University of Manchester. We acknowledge use of the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02, and thank Professor Walter Bodmer and Dr Bruce Winney for use of the People of the British Isles DNA collection which was funded by the Wellcome Trust.

## Complete List of Authors

†The complete list of authors who contributed to this study is as follows:

## : The UK IBD Genetics Consortium

Jeffrey C Barrett[1], James Lee[2], Charlie Lees[3], Natalie Prescott[4], Carl A Anderson[1,5], Anne Phillips[3], Emma Wesley[6], Kirstie Parnell[6], Hu Zhang[2], Hazel Drummond[3], Elaine R Nimmo[3], Dunecan Massey[2], Kasia Blaszczyk[4], Timothy Elliott[7], Lynn Cotterill[8], Helen Dallal[9], Alan Lobo[10], Craig Mowat[11], Jeremy Sanderson[7], Derek P Jewell[12], William

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[2]Gastroenterology Research Unit, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK
[3]Gastrointestinal Unit, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU
[4]Department of Medical and Molecular Genetics, King's College London School of Medicine, Floor 8 Tower Wing, Guy's Hospital, London SE1 9RT, UK
[5]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK
[6]Peninsula College of Medicine and Dentistry, Barrack Road, Exeter EX2 5DW, UK
[7]Dept Gastroenterology, Guy's & St Thomas' NHS Foundation Trust, St Thomas' Hospital, London SE1 9RT, UK
[8]Department of Medical Genetics, Manchester Academic Health Science Centre (MAHSC), University of Manchester and NIHR Biomedical Research Centre, Central Manchester NHS Foundation Trust, Manchester M13 0JH, UK
[9]Department of Gastroenterology, James Cook University Hospital, South Tees Hospitals NHS Trust, Marton Road, Middlesbrough TS4 3BW, UK
[10]Division of Molecular and Genetic Medicine, University of Sheffield Medical School, Royal Hallamshire Hospital, Sheffield S10 2JF, UK
[11]Department of General Internal Medicine, Ninewells Hospital and Medical School, Ninewells Avenue, Dundee DD1 9SY, UK
[12]Gastroenterology Unit, Gibson Laboratories, Radcliffe Infirmary, Woodstock Road, Oxford OX2 6HE, UK

Newman[8], Cathryn Edwards[13], Tariq Ahmad[6], John C Mansfield[14], Jack Satsangi[3], Miles Parkes[2], Christopher G Mathew[4]

## The Wellcome Trust Case Control Consortium 2

Management Committee

Peter Donnelly (Chair)[1,2], Leena Peltonen (Deputy Chair)[3], Elvira Bramon[4], Matthew Brown[5], Juan Casas[6], Aiden Corvin[7] Nicholas Craddock[8], Panos Deloukas[3], Janus Jankowski[9], Hugh Markus[10], Christopher G Mathew[11], Mark McCarthy[12], Colin Palmer[13], Robert Plomin[14], Stephen Sawcer[15], Richard C Trembath[11], Ananth Viswanathan[16], Nick Wood[17]

Data and Analysis Group

Chris C A Spencer[1], Jeffrey C Barrett[3], Celine Bellenguez[1], Daniel Davison[2], Colin Freeman[1], Amy Strange[1], Peter Donnelly[1,2]

DNA, Genotyping, Data QC and Informatics Group

Cordelia Langford[3], Sarah E Hunt[3], Sarah Edkins[3], Rhian Gwilliam[3], Hannah Blackburn[3], Suzannah J. Bumpstead[3], Serge Dronov[3], Matthew Gillman[3], Emma Gray[3], Naomi Hammond[3], Alagurevathi Jayakumar[3], Owen T McCann[3], Jennifer Liddle[3], Marc L Perez[3], Simon Potter[3], Radhi Ravindrarajah[3], Michelle Ricketts[3], 9 Matthew Waller[3], Paul Weston[3], Sara Widaa[3], Pamela Whittaker[3], Panos Deloukas[3], Leena Peltonen[3]

Publications Committee

Christopher Mathew (Chair)[11], Jenefer Blackwell[18], Matthew Brown[5], Aiden Corvin[7], Mark I McCarthy[12], Chris C A Spencer[1]

UK Blood Services Controls

[13]Endoscopy Regional Training Unit, Torbay Hospital, Torbay TQ2 7AA, UK
[14]Institute of Human Genetics, Newcastle University, Newcastle upon Tyne NE1 3BZ, UK
[1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7LJ, UK
[2]Dept Statistics, University of Oxford, Oxford OX1 3TG, UK
[3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[4]Dept Psychological Medicine, King's College London Institute of Psychiatry Denmark Hill, London SE5 8AF, UK
[5]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Nuffield Orthopaedic Centre, Oxford OX3 7LD, UK
[6]Dept Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
[7]Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Eire
[8]Dept Psychological Medicine, Cardiff University School of Medicine, Heath Park, Cardiff CF14 4XN, UK
[9]Centre for Gastroenterology, Bart's and the London School of Medicine and Dentistry, London E1 2AT, UK
[10]Division of Cardiac and Vascular Sciences, Dept Clinical Neurosciences, St George's Hospital, London SW17 0RE, UK
[11]Dept Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK
[12]Oxford Centre for Diabetes, Endocrinology and Metabolism (ICDEM), Churchill Hospital, Oxford OX3 7LJ, UK
[13]Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK
[14]Social, Genetic and Developmental Psychiatry Centre, King's College London Institute of Psychiatry, Denmark Hill, London SE5 8AF, UK
[15]Dept Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK
[16]Glaucoma Research Unit, Moorfields Eye Hospital NHS Foundation Trust, London EC1V 2PD,UK
[17]Dept Molecular Neuroscience, Institute of Neurology, Queen Square, London WC1N 3BG, UK
[18]Genetics and Infection Laboratory, Cambridge Institute of Medical Research, Addenbrooke's Hospital, Cambridge CB2 0XY, UK

Antony P Attwood[3,19], Jonathan Stephens[19], Jennifer Sambrook[19], Willem H Ouwehand[3,19]

1958 Birth Cohort Controls

Wendy L McArdle[20], Susan M Ring[21], David P Strachan[22]

## Reference List

1. Rubin GP, Hungin AP, Kelly PJ, Ling J. Inflammatory bowel disease: epidemiology and management in an English general practice population. Aliment. Pharmacol. Ther. 2000; 14:1553–1559. [PubMed: 11121902]
2. Eaden JA, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut. 2001; 48:526–535. [PubMed: 11247898]
3. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. Nature. 2007; 448:427–434. [PubMed: 17653185]
4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
5. Anderson CA, et al. Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. Gastroenterology. 2009; 136:523–529. [PubMed: 19068216]
6. Barrett JC, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. 2008; 40:955–962. [PubMed: 18587394]
7. Duerr RH, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006; 314:1461–1463. [PubMed: 17068223]
8. Fisher SA, et al. Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nat. Genet. 2008; 40:710–712. [PubMed: 18438406]
9. Franke A, et al. Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. Nat. Genet. 2008; 40:713–715. [PubMed: 18438405]
10. Hampe J, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn's disease in ATG16L1. Nat. Genet. 2007; 39:207–211. [PubMed: 17200669]
11. Libioulle C, et al. Novel Crohn's disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS. Genet. 2007; 3:e58. [PubMed: 17447842]
12. Parkes M, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat. Genet. 2007; 39:830–832. [PubMed: 17554261]
13. Satsangi J, et al. Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. Lancet. 1996; 347:1212–1217. [PubMed: 8622450]
14. Franke A, et al. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. Nat. Genet. 2008; 40:1319–1323. [PubMed: 18836448]
15. Silverberg MS, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. Nat. Genet. 2009; 41:216–220. [PubMed: 19122664]
16. Yamagata K, et al. Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). Nature. 1996; 384:458–460. [PubMed: 8945471]
17. Barroso I, et al. Population-specific risk of type 2 diabetes conferred by HNF4A P2 promoter variants: a lesson for replication studies. Diabetes. 2008; 57:3161–3165. [PubMed: 18728231]

[19]Dept Haematology, University of Cambridge and National Health Service Blood and Transplant, Long Road, Cambridge CB2 2PT, UK
[20]ALSPAC DNA Bank, Dept Social Medicine, University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
[21]ALSPAC Laboratory, Dept Social Medicine, Clifton, Bristol BS8 2BN, UK
[22]Division of Community Health Sciences, St George's Hospital, London SW17 0RE, UK.

18. Kathiresan S, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat. Genet. 2009; 41:56–65. [PubMed: 19060906]
19. Battle MA, et al. Hepatocyte nuclear factor 4alpha orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver. Proc. Natl. Acad. Sci. U. S. A. 2006; 103:8419–8424. [PubMed: 16714383]
20. Garrison WD, et al. Hepatocyte nuclear factor 4alpha is essential for embryonic development of the mouse colon. Gastroenterology. 2006; 130:1207–1220. [PubMed: 16618389]
21. Ahn SH, et al. Hepatocyte nuclear factor 4alpha in the intestinal epithelial cells protects against inflammatory bowel disease. Inflamm. Bowel Dis. 2008; 14:908–920. [PubMed: 18338782]
22. Karayiannakis AJ, et al. Expression of catenins and E-cadherin during epithelial restitution in inflammatory bowel disease. J. Pathol. 1998; 185:413–418. [PubMed: 9828841]
23. Houlston RS, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet. 2008; 40:1426–1435. [PubMed: 19011631]
24. Muise AM, et al. Polymorphisms in E-cadherin (CDH1) result in a mis-localised cytoplasmic protein that is associated with Crohn's disease. Gut. 2009; 58:1121–1127. [PubMed: 19398441]
25. Peignon G, et al. E-cadherin-dependent transcriptional control of apolipoprotein A-IV gene expression in intestinal epithelial cells: a role for the hepatic nuclear factor 4. J. Biol. Chem. 2006; 281:3560–3568. [PubMed: 16338932]
26. Vowinkel T, et al. Apolipoprotein A-IV inhibits experimental colitis. J. Clin. Invest. 2004; 114:260–269. [PubMed: 15254593]
27. Schmehl K, Florian S, Jacobasch G, Salomon A, Korber J. Deficiency of epithelial basement membrane laminin in ulcerative colitis affected human colonic mucosa. Int. J. Colorectal Dis. 2000; 15:39–48. [PubMed: 10766090]
28. Styrkarsdottir U, et al. Multiple genetic loci for bone mineral density and fractures. N. Engl. J. Med. 2008; 358:2355–2365. [PubMed: 18445777]
29. Liu Y, et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet. 2008; 4:e1000041. [PubMed: 18369459]
30. Kugathasan S, et al. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. Nat. Genet. 2008; 40:1211–1215. [PubMed: 18758464]
31. Zhernakova A, et al. Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. Am. J. Hum. Genet. 2008; 82:1202–1210. [PubMed: 18439550]
32. Festen EA, et al. Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. Gut. 2009; 58:799–804. [PubMed: 19201773]
33. Korn JM, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. 2008; 40:1253–1260. [PubMed: 18776909]
34. Spencer CCA. A simple clustering approach to pre-analysis exclusion of individuals from GWAS. In preparation. 2009
35. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]

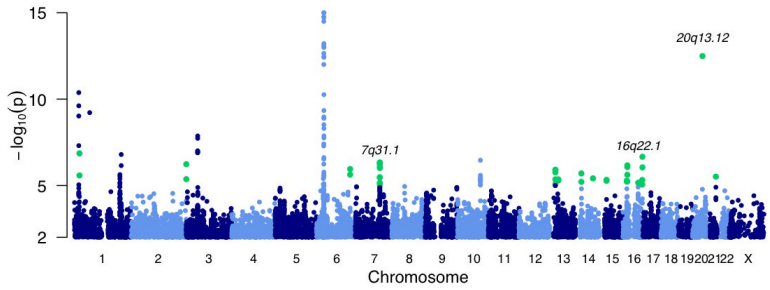


**Figure 1.**

−$\log_{10}(P)$ values from the 1 d.f. trend test. Alternating chromosomes shown in shades of blue. SNPs with $P < 1\times10\text{-}5$ which had not been previously reported are highlighted in green. The three new loci identified in this study are noted.





Europe PMC Funders Author Manuscripts

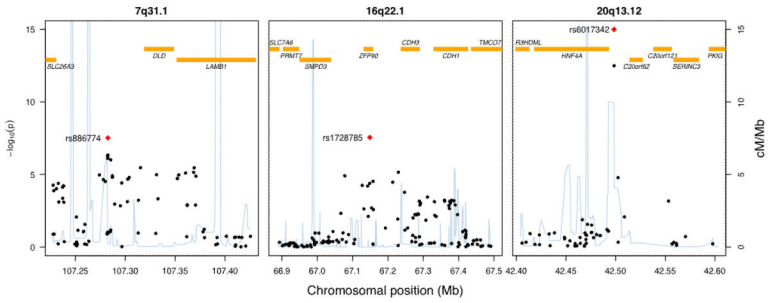


**Figure 2.**

$-\log_{10}(P)$ values from the 1 d.f. trend test from three new loci, along with local recombination rate estimated from HapMap data. Combined *P* values for replicated SNPs are indicated with a red diamond.

## Table 1

### Clinical details of cases and controls

| | GWAS | Replication cohort |
|---|---|---|
| **CASES** | 2361 | 2321 |
| | | |
| **Age at diagnosis** [a] | | |
| Early onset (<18years) | 5.9% (112) | 6.5% (130) |
| Not early onset (>18years) | 94.1% (1783) | 93.5% (1861) |
| Median | 33.4 | 35.2 |
| Mean | 36.3 | 38.2 |
| | | |
| **Disease Extent** [b] | | |
| Proctitis | 17.9% (357) | 15.1% (285) |
| Left sided | 38.8% (774) | 47.5% (897) |
| Extensive | 43.3% (864) | 37.4% (707) |
| | | |
| **Smoking at diagnosis** [c] | | |
| Ex-smoker | 36.7% (556) | 30.3% (553) |
| Current smoker | 10.8% (163) | 16.4% (300) |
| Never smoked | 52.5% (794) | 53.2% (971) |
| | | |
| **Colectomy** [d] | | |
| Yes | 15.7% (266) | 12.0% (226) |
| No | 84.3% (1432) | 88.0% (1660) |
| | | |
| **Colorectal cancer** [e] | 1.00% (23) | 0.87% (20) |
| | | |
| **CONTROLS** | | |
| Total | 5417 | 4818 |
| UKBS | 2675 | - |
| 1958 Birth Cohort | 2742 | 1952 |
| POBI | - | 2866 |

[a]Data available for 80% (GWAS) 86% (Replication),

[b]Data available for 84% (GWAS) 81% (Replication),

[c]Data available for 64% (GWAS) 79% (Replication),

[d]Data available for 72% (GWAS) 81% (Replication),

[e]Data available for 93% (GWAS) 99% (Replication).

Europe PMC Funders Author Manuscripts



**Table 2**

**New hits from the GWAS**

Top tier reaches $5\times10^{-8}$ in combined analysis, second tier hits have replication $P < 0.05$ and require further study to completely verify.

| SNP | Chr | LD region(Mb)[a] | Gene of interest (#)[b] | $P_{scan}$ | $P_{repl}$ | $P_{comb}$ | Risk allele | RAF[c] | OR (95% CI) |
|---|---|---|---|---|---|---|---|---|---|
| rs886774 | 7q31.1 | 107.25-107.39 | *LAMB1(2)* | $4.8 \times 10^{-7}$ | 0.005 | $3 \times 10^{-8}$ | G | 0.4136 | 1.11 (1.03-1.19) |
| rs1728785 | 16q22.1 | 66.98-67.40 | *CDH1(5)* | $1.8 \times 10^{-5}$ | 0.0004 | $2.8 \times 10^{-8}$ | G | 0.7641 | 1.17 (1.07-1.27) |
| rs6017342 | 20q13.12 | 42.49-42.52 | *HNF4A(7)* | $3.2 \times 10^{-13}$ | $7.1 \times 10^{-6}$ | $8.5 \times 10^{-17}$ | C | 0.5168 | 1.17 (1.09-1.26) |
| rs7524102 * | 1p36.12 | 22.54-22.61 | (0) | $1.4 \times 10^{-7}$ | 0.05 | $3.1 \times 10^{-7}$ | A | 0.8264 | 1.10 (1.00-1.21) |
| rs9548988 | 13q13.3 | 39.36-39.56 | (0) | $5.0 \times 10^{-6}$ | 0.0061 | $2.7 \times 10^{-7}$ | T | 0.4594 | 1.10 (1.03-1.19) |

* replication genotyping at this locus is for SNP rs12568930, which is an $r^2 = 1$ proxy for rs7524102.

[a] LD region of 0.2 cM centered on focal SNP, in NCBI Build 36 coordinates.

[b] Number of genes in LD region.

[c] Risk allele frequency.

Europe PMC Funders Author Manuscripts





**Table 3**

**GWAS signals from previously reported UC loci**

Top tier were previously reported at genome-wide significance ($5\times10^{-8}$), bottom tier were previously reported with weaker evidence. *P* values are one-tailed in the direction of the previously reported association.

| SNP | Chrom | Pos | Gene | Scan *P* | Ref |
|---|---|---|---|---|---|
| rs6426833 | 1p36.13 | 20044447 | *OTUD3/PLA2GE* | $2.1 \times 10^{-11}$ | 15 |
| rs11209026 | 1p31.3 | 67478546 | *IL23R* | $3.0 \times 10^{-10}$ | 15 |
| rs3024493 | 1q32.1 | 205010591 | *IL10* | $8.0 \times 10^{-8}$ | 14 |
| rs10021288 | 4q27 | 123224984 | *IL2/21* | 0.0033 | 32 |
| rs9268877 | 6p21.32 | 32539125 | *MHC* | $3.9 \times 10^{-23}$ | 8 |
| rs12815372 | 12q15 | 66765480 | *IL26* | 0.00070 | 15 |
| rs311497 | 20q13.33 | 61691693 | *TNFRSF6B* | 0.0018 | 30 |
| rs2094871 | 21q22.2 | 39382729 | *PSMG1* | $1.6 \times 10^{-6}$ | 30 |
| rs7511649 | 1q21.2 | 148537415 | *ECM1* | 0.00015 | 8 |
| rs7554511 | 1q32.1 | 199144185 | *KIF21B* | $1.2 \times 10^{-6}$ | 5 |
| rs12612347 | 2q35 | 218765583 | *ARPC2* | 0.024 | 14 |
| rs9858542 | 3p21.31 | 49676987 | *MST1* | $7.0 \times 10^{-9}$ | 8 |
| rs1368438 | 5q33.3 | 158639883 | *IL12B* | 0.0039 | 8 |
| rs12529198 | 6p25.1 | 5096246 | *LYRM4* | 0.13 | 5 |
| rs6908425 | 6p22.3 | 20836710 | *CDKAL1* | 0.0044 | 5 |
| rs10974914 | 9p24.1 | 5004332 | *JAK2* | $1.5 \times 10^{-5}$ | 5 |
| rs10781500 | 9q34.3 | 138389159 | *CARD9* | $7.0 \times 10^{-6}$ | 31 |
| rs17582416 | 10p11.21 | 35327656 | *CCNY* | 0.022 | 5 |
| rs10995271 | 10q21.2 | 64108492 | none | 0.32 | 8, 9 |
| rs6584283 | 10q24.2 | 101280291 | *NKX2-3* | $1.7 \times 10^{-7}$ | 8 |
| rs916977 | 15q13.1 | 26186959 | *HERC2* | 0.26 | 9 |
| rs744166 | 17q21.2 | 37767727 | *STAT3* | 0.0025 | 5, 9 |
| rs2542151 | 18p11.21 | 12769947 | *PTPN2* | 0.0010 | 9 |