nature genetics

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

The UK IBD Genetics Consortium & the Wellcome Trust Case Control Consortium 2*

Ulcerative colitis is a common form of inflammatory bowel disease with a complex etiology. As part of the Wellcome Trust Case Control Consortium 2, we performed a genome-wide association scan for ulcerative colitis in 2,361 cases and 5,417 controls. Loci showing evidence of association at $P < 1 \times 10^{-5}$ were followed up by genotyping in an independent set of 2,321 cases and 4,818 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*; $P = 3.0 \times 10^{-8}$). Of note, *CDH1* has recently been associated with susceptibility to colorectal cancer, an established complication of longstanding ulcerative colitis. The new associations suggest that changes in the integrity of the intestinal epithelial barrier may contribute to the pathogenesis of ulcerative colitis.

Genetic epidemiological data clearly implicate inherited susceptibility in the pathogenesis of ulcerative colitis and Crohn's disease, which represent the two most common forms of inflammatory bowel disease (IBD) and together affect at least 1 in 250 individuals in 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 affected individuals, as well as colorectal cancer[2].

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 Crohn's disease and ulcerative colitis are related polygenic diseases. This has recently been corroborated by the results of genetic association studies, which have highlighted both disease-specific loci and other loci that are shared between ulcerative colitis and Crohn's disease. For example, whereas genetically determined defects in the handling of intracellular bacteria (related to the gene *NOD2*, and the autophagy genes *ATG16L1* and *IRGM*) are specific to Crohn's disease, genes encoding multiple components in the Th17 pathway (*IL23R*, *IL12B*, *JAK2* and *STAT3*) are associated with both Crohn's disease and ulcerative colitis[4–12].

Until recently, most attention had focused on Crohn's disease, with genome-wide association (GWA) studies and subsequent meta-analysis yielding more than 30 confirmed Crohn's disease-susceptibility loci[4,6,7,10–12]. In addition to the longstanding known association in the major histocompatibility complex (MHC)[13], the first GWA scans in ulcerative colitis reported associations at *IL23R*, *IL10* and loci on chromosomes 1p36 and 12q15 that 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 ulcerative colitis to date. All study subjects were UK residents of European ancestry; clinical data are presented in **Table 1**. Affected individuals (cases) and controls were genotyped on the Affymetrix 6.0 array. After application of quality control filters (see Online Methods), we analyzed GWA data from 2,361 individuals with ulcerative colitis and 5,417 controls (**Fig. 1**). An initial analysis revealed 24 distinct loci (comprising 156 SNPs) which showed evidence of association at $P < 1 \times 10^{-5}$. Sixteen of these loci 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 2,321 ulcerative colitis cases and 4,818 controls. Three new loci showed evidence for association at $P < 5 \times 10^{-8}$ in the combined panel, with three additional new loci showing nominal ($P < 0.05$) replication (**Table 2** and **Fig. 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 hot spot on chromosome 20q13 containing the 3′ untranslated region of just one gene, *HNF4A*. The SNP rs6017342 itself maps 5 kb distal to the 3′ untranslated region and is located within an expressed sequence tag DB076868, which has been detected in only a single testis cDNA library and does not

*A full list of authors and affiliations appears at the end of the paper.

Received 21 July; accepted 24 September; published online 15 November 2009; doi:10.1038/ng.483

Table 1 Clinical details of cases and controls

	GWAS	Replication cohort
Cases	2,361	2,321
Age at diagnosis[a]		
Early onset (<18 years)	5.9% (112)	6.5% (130)
Not early onset (>18 years)	94.1% (1,783)	93.5% (1,861)
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% (1,432)	88.0% (1,660)
Colorectal cancer[e]	1.00% (23)	0.87% (20)
Controls		
Total	5,417	4,818
UKBS	2,675	–
1958 Birth Cohort	2,742	1,952
PoBI	–	2,866

[a]Data available for 80% (GWAS) and 86% (Replication) of cases. [b]Data available for 84% (GWAS) and 81% (Replication) of cases. [c]Data available for 64% (GWAS) and 79% (Replication) of cases. [d]Data available for 72% (GWAS) and 81% (Replication) of cases. [e]Data available for 93% (GWAS) and 99% (Replication) of cases.

encode any extended 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. Because rs6017342 is located within a recombination hot spot, there are few known SNPs in strong linkage disequilibrium (LD) ($r^2 > 0.5$) with it; there are no additional SNPs in LD with rs6017342 on the Affymetrix chip used in this study or on the Illumina chips used in previous studies. As the evidence for association with ulcerative colitis rests on this single SNP, we subjected these data to careful scrutiny; genotype cluster plots for this SNP showed clear resolution of the three genotype classes (**Supplementary Fig. 1**), with 99.3% completeness of genotypes within this dataset.

Rare *HNF4A* mutations account for approximately 4% of individuals in the UK with maturity-onset diabetes of the young (MODY)[16], a monogenic form of diabetes mellitus characterized by autosomal dominant inheritance, young age of onset, pancreatic β-cell dysfunction and sensitivity to sulfonylureas. Common variants of *HNF4A* influence predisposition to type 2 diabetes (rs2144908)[17] and dyslipidemia (rs1800961)[18]. The ulcerative colitis–associated SNP, rs6017342, is not in LD with either of these two common variants, nor did it show association in our study of Crohn's disease ($P = 0.92$)[4].

HNF4A encodes the transcription factor hepatocyte nuclear factor 4α, 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 has a key role in the development of the embryonic mammalian gastrointestinal tract. Previous studies demonstrated that mice with targeted deletion of *Hnf4a* in epithelial cells of the fetal colon die perinatally. Histological analysis of colonic tissue recovered during late development (embryonic day (E) 18.5) demonstrated absent crypt formation, reduced epithelial cell proliferation and defective goblet-cell maturation[20]. In order to explore the role of Hnf4a in mouse intestinal inflammation, Ahn and colleagues circumnavigated the embryonic lethality of $Hnf4a^{-/-}$ mice by generating a conditional model of intestinal *Hnf4a* deletion[21]. These $Hnf4a^{\Delta IEpC}$ mice (*loxP*-flanked *Hnf4a* driven by the *Vil1* (villin 1) promoter) developed increased epithelial permeability and markedly more severe colitis following dextran sodium sulfate (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 ulcerative colitis[21], a finding that 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 that is bounded by recombination hot spots spans 411 kb and encodes several genes. Among the strongest candidates for ulcerative colitis susceptibility is *CDH1*, which encodes E-cadherin. This transmembrane glycoprotein is one of the main components of the adherens junction and is 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 reduced in areas of active ulcerative colitis[22].

Given the well-recognized association between ulcerative colitis and colorectal cancer[2], the observation of correlated association signals at the *CDH1* locus in both diseases is noteworthy. In particular, variants in LD ($r^2 = 0.5$) with the SNP most strongly associated with ulcerative colitis in our study were recently identified in a GWA-scan meta-analysis to be associated with colorectal cancer susceptibility[23]; conversely, we find that a perfect proxy for the most strongly associated SNP in the colorectal cancer study is also associated with ulcerative colitis ($P = 8 \times 10^{-4}$). This locus did not show association with Crohn's disease in a large international GWA meta-analysis ($P = 0.549$)[6] (**Supplementary Table 2**). However, evidence for association of *CDH1* with Crohn's disease was reported recently in the Canadian population using a candidate gene approach[24], and the presence of Crohn's disease–associated SNPs resulted in a truncated E-cadherin protein *in vitro* which accumulated in the cytoplasm and led to disorganized epithelial architecture[24].

The evidence that HNF4α and E-cadherin cooperate to maintain epithelial barrier integrity in the intestine is of great potential relevance. In experiments focused on the liver, *Hnf4a* knockout mice

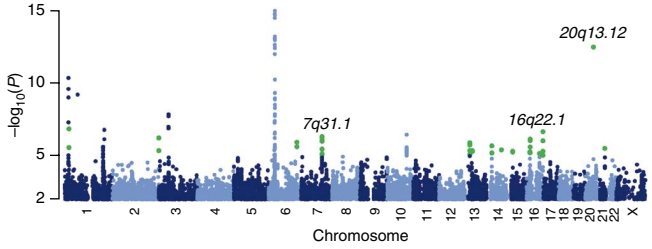

Figure 1 Plot of genome-wide association results. $-\log_{10}(P)$ values are from the 1-d.f. trend test. Alternating chromosomes are shown in shades of blue. SNPs with $P < 1 \times 10^{-5}$ which had not been previously reported are highlighted in green. The three new loci identified in this study are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

Table 2 New hits from the GWAS

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)	4.8×10^{-7}	0.005	3×10^{-8}	G	0.4136	1.11 (1.03–1.19)
rs1728785	16q22.1	66.98–67.40	*CDH1* (5)	1.8×10^{-5}	0.0004	2.8×10^{-8}	G	0.7641	1.17 (1.07–1.27)
rs6017342	20q13.12	42.49–42.52	*HNF4A* (1)	3.2×10^{-13}	7.1×10^{-6}	8.5×10^{-17}	C	0.5168	1.17 (1.09–1.26)
rs7524102[d]	1p36.12	22.54–22.61	None	1.4×10^{-7}	0.05	3.1×10^{-7}	A	0.8264	1.10 (1.00–1.21)
rs9548988	13q13.3	39.36–39.56	None	5.0×10^{-6}	0.0061	2.7×10^{-7}	T	0.4594	1.10 (1.03–1.19)

Top tier hits reach $P < 5 \times 10^{-8}$ in combined analysis; second tier hits have replication $P < 0.05$ and require further study to completely verify.
[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. [d]Replication genotyping at this locus is for SNP rs12568930, which is a proxy ($r^2 = 1$) for rs7524102.

failed to express E-cadherin[19], whereas in a human intestinal cell line, E-cadherin–dependent cell-cell contact was found to be critical in determining the amount and binding activity of nuclear HNF4α[25]. This in turn affected the expression of several genes, including *APOA4* (ref. 25), which encodes an anti-inflammatory protein known to inhibit experimental colitis[26].

The third newly confirmed ulcerative colitis susceptibility locus was a region on chromosome 7q31, previously suggested by a recent North American GWA scan[15]. 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 β1 subunit. Laminins are heterotrimers; the β1 light-chain is present in laminin-1, laminin-2 and laminin-10. Laminins are expressed in the intestinal basement membrane and play a key role in anchoring the single-layered epithelium; expression of laminins is known to be downregulated in ulcerative colitis[27]. rs886774 was not associated with Crohn's disease in the meta-analysis[6] (**Supplementary Table 2**).

Two other loci previously implicated in ulcerative colitis–related phenotypes showed strong (but not genome-wide significant) association here. These comprise a SNP previously associated with osteoporosis[28] (rs7524102 on chromosome 1p36, combined GWA and replication $P = 3.1 \times 10^{-7}$) and a SNP near (though not in LD with) a marker known to be associated with psoriasis[29] (rs9548988 on 13q.13, combined GWA and replication $P = 2.7 \times 10^{-7}$).

In addition to the newly discovered loci described above, our GWAS detected strong association at established ulcerative colitis 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 ulcerative colitis loci reported recently in genome-wide scans: the *IL10* locus[14] and the *OTUD3-PLA2G2E* locus[15], on chromosomes 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 ulcerative colitis. Moderate support was obtained for some loci previously reported to be associated with ulcerative colitis, including *ECM1*, *CARD9* (ref. 31), 1q32-*KIF21B* and 9p24-*JAK2*, but weaker support was obtained for other loci, such as *IL2-IL21* (ref. 32), *IL12B* and 12q15 (**Table 3**). Some of the ulcerative colitis loci are clearly associated with Crohn's disease, whereas others are not associated 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. We have identified three new susceptibility loci for ulcerative colitis and provide the first genetic link between ulcerative colitis and colorectal cancer. The strongest new association intervals that we detected contain, 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 ulcerative colitis 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 significance 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.

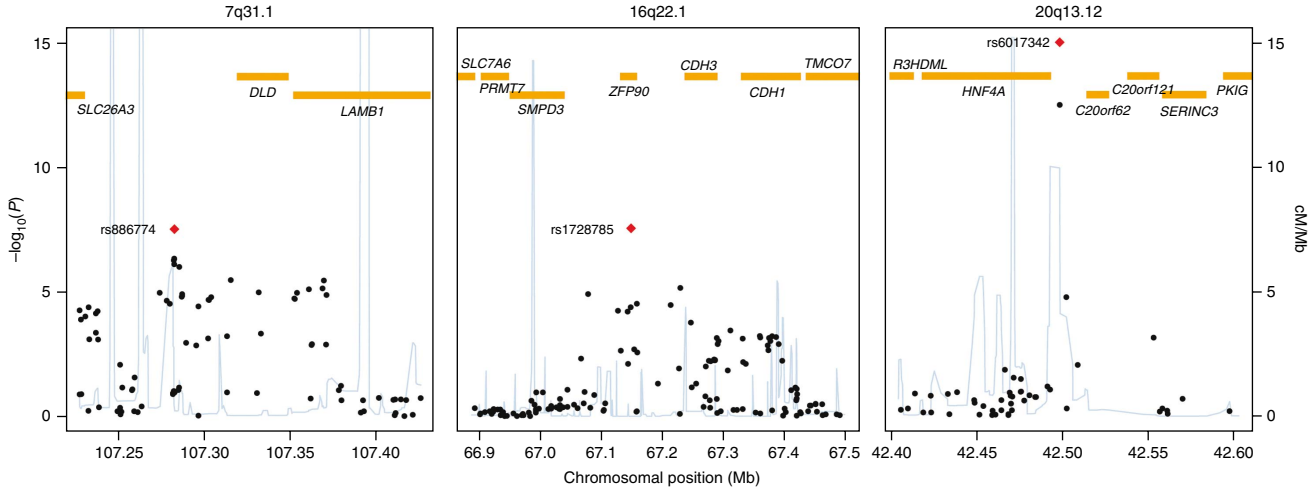

Figure 2 Regional association plots for the three newly discovered loci. $-\log_{10}(P)$ values from the 1-d.f. trend test from three new loci, along with local recombination rates estimated from HapMap data. Combined *P* values for replicated SNPs are indicated with a red diamond.

Table 3 GWAS signals from loci previously reported to be associated with ulcerative colitis

SNP	Chr.	Pos	Gene	Scan *P*	Ref.
rs6426833	1p36.13	20,044,447	*OTUD3-PLA2G2E*	2.1×10^{-11}	15
rs11209026	1p31.3	67,478,546	*IL23R*	3.0×10^{-10}	15
rs3024493	1q32.1	205,010,591	*IL10*	8.0×10^{-8}	14
rs10021288	4q27	123,224,984	*IL2-IL21*	0.0033	32
rs9268877	6p21.32	32,539,125	MHC	3.9×10^{-23}	8
rs12815372	12q15	66,765,480	*IL26*	0.00070	15
rs311497	20q13.33	61,691,693	*TNFRSF6B*	0.0018	30
rs2094871	21q22.2	39,382,729	*PSMG1*	1.6×10^{-6}	30
rs7511649	1q21.2	148,537,415	*ECM1*	0.00015	8
rs7554511	1q32.1	199,144,185	*KIF21B*	1.2×10^{-6}	5
rs12612347	2q35	218,765,583	*ARPC2*	0.024	14
rs9858542	3p21.31	49,676,987	*MST1*	7.0×10^{-9}	8
rs1368438	5q33.3	158,639,883	*IL12B*	0.0039	8
rs12529198	6p25.1	5,096,246	*LYRM4*	0.13	5
rs6908425	6p22.3	20,836,710	*CDKAL1*	0.0044	5
rs10974914	9p24.1	5,004,332	*JAK2*	1.5×10^{-5}	5
rs10781500	9q34.3	138,389,159	*CARD9*	7.0×10^{-6}	31
rs17582416	10p11.21	35,327,656	*CCNY*	0.022	5
rs10995271	10q21.2	64,108,492	none	0.32	8,9
rs6584283	10q24.2	101,280,291	*NKX2-3*	1.7×10^{-7}	8
rs916977	15q13.1	26,186,959	*HERC2*	0.26	9
rs744166	17q21.2	37,767,727	*STAT3*	0.0025	5,9
rs2542151	18p11.21	12,769,947	*PTPN2*	0.0010	9

Top tier were previously reported at genome-wide significance ($P < 5 \times 10^{-8}$); bottom tier were previously reported with weaker evidence. *P* values are one-tailed in the direction of the previously reported association.

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

The principal funding for this study was provided by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project (083948/Z/07/Z). 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 S. Bertrand, J. Bryant, S.L. Clark, J.S. Conquer, T. Dibling, J.C. Eldred, S. Gamble, C. Hind, A. Wilk, C.R. Stribling and S. Taylor of the Wellcome Trust Sanger Institute's Sample and Genotyping Facilities 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. P. Donnelly was supported in part by a Wolfson–Royal Society Merit Award. We also acknowledge support from the UK Medical Research Council (R.C.T., grant G0601387), the Special Trustees of Moorfields National Health Service (NHS) Foundation Trust, and 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, the Central Manchester NHS Foundation Trust in partnership with the University of Manchester, and Moorfields Eye Hospital in partnership with University College London Institute of Ophthalmology. 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 W. Bodmer and B. Winney for use of the People of the British Isles DNA collection, which was funded by the Wellcome Trust.

AUTHOR CONTRIBUTIONS

J.C.L., C.W.L., N.J.P., A.P., E.W., K.P., H.Z., H.D., E.R.N., D.M., K.B., T.E. and L.C. were involved in establishing DNA collections and/or assembling phenotypic data. C.W.L., A.P., E.W., D.M., H.D., A.J.L., C.M., J.D.S., D.P.J., C.E., T.A., J.C.M., J. Satsangi and M.P. recruited patients. W.G.N., C.E., T.A., J.C.M., J.D.S., M.P. and C.G.M. supervised clinical and laboratory work. The WTCCC2 DNA, genotyping, data QC and informatics group executed GWAS sample handling, genotyping and quality control. The WTCCC2 data and analysis group, J.C.B. and C.A.A. performed statistical analyses. J.C.B., J.C.L., C.W.L., N.J.P., C.C.A.S., C.A.A., T.A., P. Donnelly, J. Satsangi, M.P. and C.G.M. contributed to writing the manuscript. The WTCCC2 management committee conceived and oversaw the design and execution of the GWAS. WTCCC2 group memberships are specified in the full author list.

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

1. Rubin, G.P., Hungin, A.P., Kelly, P.J. & Ling, J. Inflammatory bowel disease: epidemiology and management in an English general practice population. *Aliment. Pharmacol. Ther.* **14**, 1553–1559 (2000).
2. Eaden, J.A., Abrams, K.R. & Mayberry, J.F. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut* **48**, 526–535 (2001).
3. Xavier, R.J. & Podolsky, D.K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**, 427–434 (2007).
4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
5. Anderson, C.A. *et al.* Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. *Gastroenterology* **136**, 523–529 (2009).
6. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
7. Duerr, R.H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
8. Fisher, S.A. *et al.* Genetic determinants of ulcerative colitis include the *ECM1* locus and five loci implicated in Crohn's disease. *Nat. Genet.* **40**, 710–712 (2008).
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.* **40**, 713–715 (2008).
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.* **39**, 207–211 (2007).
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.* **3**, e58 (2007).
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.* **39**, 830–832 (2007).

13. Satsangi, J. *et al.* Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. *Lancet* **347**, 1212–1217 (1996).
14. Franke, A. *et al.* Sequence variants in *IL10, ARPC2* and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.* **40**, 1319–1323 (2008).
15. Silverberg, M.S. *et al.* Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* **41**, 216–220 (2009).
16. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-4α gene in maturity-onset diabetes of the young (MODY1). *Nature* **384**, 458–460 (1996).
17. Barroso, I. *et al.* Population-specific risk of type 2 diabetes conferred by *HNF4A* P2 promoter variants: a lesson for replication studies. *Diabetes* **57**, 3161–3165 (2008).
18. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
19. Battle, M.A. *et al.* Hepatocyte nuclear factor 4α orchestrates expression of cell adhesion proteins during the epithelial transformation of the developing liver. *Proc. Natl. Acad. Sci. USA* **103**, 8419–8424 (2006).
20. Garrison, W.D. *et al.* Hepatocyte nuclear factor 4α is essential for embryonic development of the mouse colon. *Gastroenterology* **130**, 1207–1220 (2006).
21. Ahn, S.H. *et al.* Hepatocyte nuclear factor 4α in the intestinal epithelial cells protects against inflammatory bowel disease. *Inflamm. Bowel Dis.* **14**, 908–920 (2008).
22. Karayiannakis, A.J. *et al.* Expression of catenins and E-cadherin during epithelial restitution in inflammatory bowel disease. *J. Pathol.* **185**, 413–418 (1998).
23. 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).
24. Muise, A.M. *et al.* Polymorphisms in E-cadherin (*CDH1*) result in a mis-localised cytoplasmic protein that is associated with Crohn's disease. *Gut* **58**, 1121–1127 (2009).
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.* **281**, 3560–3568 (2006).
26. Vowinkel, T. *et al.* Apolipoprotein A-IV inhibits experimental colitis. *J. Clin. Invest.* **114**, 260–269 (2004).
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.* **15**, 39–48 (2000).
28. Styrkarsdottir, U. *et al.* Multiple genetic loci for bone mineral density and fractures. *N. Engl. J. Med.* **358**, 2355–2365 (2008).
29. Liu, Y. *et al.* A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet.* **4**, e1000041 (2008).
30. Kugathasan, S. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* **40**, 1211–1215 (2008).
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.* **82**, 1202–1210 (2008).
32. Festen, E.A. *et al.* Genetic variants in the region harbouring *IL2/IL21* associated with ulcerative colitis. *Gut* **58**, 799–804 (2009).

The UK IBD Genetics Consortium

Jeffrey C Barrett[1], James C Lee[2], Charles W Lees[3], Natalie J 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 J Lobo[10], Craig Mowat[11], Jeremy D Sanderson[7], Derek P Jewell[12], William G 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)[5,15], Leena Peltonen (Deputy Chair)[1], Jenefer M Blackwell[16], Elvira Bramon[17], Matthew A Brown[18], Juan P Casas[19], Aiden Corvin[20], Nicholas Craddock[21], Panos Deloukas[1], Audrey Duncanson[22], Janusz Jankowski[23], Hugh S Markus[24], Christopher G Mathew[4], Mark I McCarthy[25], Colin N A Palmer[26], Robert Plomin[27], Anna Rautanen[5], Stephen J Sawcer[28], Nilesh Samani[29], Richard C Trembath[4], Ananth C Viswanathan[30,31] & Nicholas Wood[32]

Data and Analysis Group: Chris C A Spencer[5], Jeffrey C Barrett[1], Céline Bellenguez[5], Daniel Davison[15], Colin Freeman[5], Amy Strange[5] & Peter Donnelly[5,15]

DNA, Genotyping, Data QC and Informatics Group: Cordelia Langford[1], Sarah E Hunt[1], Sarah Edkins[1], Rhian Gwilliam[1], Hannah Blackburn[1], Suzannah J Bumpstead[1], Serge Dronov[1], Matthew Gillman[1], Emma Gray[1], Naomi Hammond[1], Alagurevathi Jayakumar[1], Owen T McCann[1], Jennifer Liddle[1], Marc L Perez[1], Simon C Potter[1], Radhi Ravindrarajah[1], Michelle Ricketts[1], Matthew Waller[1], Paul Weston[1], Sara Widaa[1], Pamela Whittaker[1], Panos Deloukas[1] & Leena Peltonen[1]

Publications Committee: Christopher G Mathew (Chair)[4], Jenefer M Blackwell[16], Matthew A Brown[18], Aiden Corvin[20], Mark I McCarthy[25] & Chris C A Spencer[5]

UK Blood Services Controls: Antony P Attwood[1,33], Jonathan Stephens[33], Jennifer Sambrook[33] & Willem H Ouwehand[1,33]

1958 Birth Cohort Controls: Wendy L McArdle[34], Susan M Ring[35] & David P Strachan[36]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [2]Gastroenterology Research Unit, Addenbrooke's Hospital, Hills Road, Cambridge, UK. [3]Gastrointestinal Unit, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, UK. [4]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London, UK. [5]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, UK. [6]Peninsula College of Medicine and Dentistry, Barrack Road, Exeter, UK. [7]Department of Gastroenterology, Guy's & St. Thomas' NHS Foundation Trust, St Thomas' Hospital, London, 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, UK. [9]Department of Gastroenterology, James Cook University Hospital, South Tees Hospitals NHS Trust, Middlesbrough, UK. [10]Division of Molecular and Genetic Medicine, University of Sheffield Medical School, Royal Hallamshire Hospital, Sheffield, UK. [11]Department of General Internal Medicine, Ninewells Hospital and Medical School, Ninewells Avenue, Dundee, UK. [12]Gastroenterology Unit, Gibson Laboratories, Radcliffe Infirmary, Woodstock Road, Oxford, UK. [13]Endoscopy Regional Training Unit, Torbay Hospital, Torbay, UK. [14]Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, UK. [15]Department of Statistics, University of Oxford, Oxford, UK. [16]Genetics and Infection Laboratory, Cambridge Institute of Medical Research, Addenbrooke's Hospital, Cambridge, UK. [17]Division of Psychological Medicine and Psychiatry, Biomedical Research Centre for Mental Health at the Institute of Psychiatry, King's College London and The South London and Maudsley NHS Foundation Trust, Denmark Hill, London, UK. [18]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia. [19]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [20]Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Eire. [21]Department of Psychological Medicine, Cardiff University School of Medicine, Heath Park, Cardiff, UK. [22]Molecular and Physiological Sciences, The Wellcome Trust, London, UK. [23]Centre for Gastroenterology, Bart's and the London School of Medicine and Dentistry, London, UK. [24]Clinical Neurosciences, St. George's University of London, London, UK. [25]Oxford Centre for Diabetes, Endocrinology and Metabolism (ICDEM), Churchill Hospital, Oxford, UK. [26]Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee, UK. [27]Social, Genetic and Developmental Psychiatry Centre, King's College London Institute of Psychiatry, Denmark Hill, London, UK. [28]University of Cambridge Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge, UK. [29]Department of Cardiovascular Science, University of Leicester, Glenfield Hospital, Leicester, UK. [30]Glaucoma Research Unit, Moorfields Eye Hospital NHS Foundation Trust, London, UK. [31]Department of Genetics, University College London Institute of Ophthalmology, London, UK. [32]Department of Molecular Neuroscience, Institute of Neurology, Queen Square, London, UK. [33]Department of Haematology, University of Cambridge and NHS Blood and Transplant, Long Road, Cambridge, UK. [34]Avon Longitudinal Study of Parents and Children (ALSPAC) DNA Bank, Department of Social Medicine, University of Bristol, Bristol, UK. [35]ALSPAC Laboratory, Department of Social Medicine, Clifton, Bristol, UK. [36]Division of Community Health Sciences, St. George's Hospital, London, UK. Correspondence should be addressed to J.C.B. (barrett@sanger.ac.uk) or C.C.A.S. (chris.spencer@well.ox.ac.uk).

ONLINE METHODS

Subjects. *Cases*: A total of 5,319 unrelated individuals of European ancestry with a diagnosis of ulcerative colitis established using standard endoscopic, radiological and histological criteria were recruited from ten centers within the UK (Cambridge, Oxford, London, Newcastle, Sheffield, Edinburgh, Dundee, Manchester, Torbay and Exeter) and defined as cases. All study participants 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 before sample collection (Cambridge, Oxford, London, Newcastle, Sheffield, Edinburgh, Dundee, Manchester, Torbay and Exeter Local Research Ethics Committees). After applying quality control measures (see below and **Supplementary Table 3**), we analyzed a total of 4,682 samples, which were divided between the discovery panel (2,361 samples) and replication panel (2,321 samples).

Controls: A total of 10,235 population control DNA samples from three sources passed our quality control filters (see below). 5,417 samples of the WTCCC2 common control set were used for the GWA experiment. This comprised 2,675 healthy blood donors recruited from the United Kingdom Blood Service (UKBS) and 2,742 samples from the 1958 Birth Cohort (1958BC) obtained from Epstein Barr virus–transformed cell lines from individuals born in England, Wales and Scotland during 1 week in 1958. The 4,818 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 obtained from individuals with self-reported British ancestry.

A summary of cases 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 and subject identity were validated using the Sequenom iPLEX 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 standards if the original DNA concentration was ≥50 ng/μl, 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 quality control, raw intensities (from the .CEL files) were renormalized within collections using CelQuantileNorm. These normalized intensities were used to call genotypes with an updated version of the Chiamo software 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 used thresholds for assigning 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 quality control filters described below, this threshold led to a study-wide level of missing data of 0.20%.

An overlapping set of 4,830 controls was 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 quality control 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 ref. 33) by making genotype calls at different confidence thresholds and then plotting the fraction of calls made against concordance with the Illumina genotypes (**Supplementary Fig. 2**). The general trend was 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 strongly 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 quality control measures (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 we excluded SNPs for which 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 the data typically violate assumptions underpinning standard tests for association, the payback in terms of increased accuracy for these tests can be substantial.

To try to obtain the maximally powerful set of samples and SNPs, we attempted to refine some standard quality control practices. For all individuals, we explicitly modeled the data as a mixture of 'normal' and 'outlier' individuals for each of ancestry, missing data and heterozygosity, and sex assignment (unpublished). 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 among 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 allowed more refined assessment of the relatedness between individuals than genome-wide sharing statistics (for example, parent-child relationships can be distinguished from those of siblings). We obtained a set of individuals with identity by descent < 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 Fig. 3**). The measure is calculated automatically by the program SNPtest. SNPs were removed if this information measure was below 0.98 or if the estimated minor allele frequency 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 $P < 1 \times 10^{-5}$, and we excluded 16 from further analysis as likely genotyping errors.

Supplementary Figure 4 provides quantile-quantile plots for the post–quality control comparison of our two control collections and for association statistics based on the post-quality control trend test comparing cases and the combined control set. Both visual inspection and the inflation statistics for each ($\lambda = 1.037$ and $\lambda = 1.079$, respectively) suggest that the quality control 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[34]. We also performed 2-d.f. genotypic tests to verify that none of our associations show significant deviation from a multiplicative model, and we performed

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, http://www.stats.ox.ac.uk/~marchini/software/gwas/chiamo.html; SNPtest, http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html; PoBI http://www.peopleofthebritishisles.org.

33. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).