Human Molecular Genetics, 2014 ***1–8***
doi:10.1093/hmg/ddu398

A genome-wide association study identifies a novel locus at 6q22.1 associated with ulcerative colitis

Antonio Julià[1], Eugeni Domènech[2,3], María Chaparro[3,4], Valle García-Sánchez[5], Fernando Gomollón[3,6], Julián Panés[3,7], Míriam Mañosa[2,3], Manuel Barreiro-De Acosta[8], Ana Gutiérrez[3,9], Esther Garcia-Planella[10], Mariam Aguas[3,11], Fernando Muñoz[12], Maria Esteve[3,13], Juan L. Mendoza[14], Maribel Vera[15], Lucía Márquez[16], Raül Tortosa[1], María López-Lasanta[1], Arnald Alonso[1], Josep L. Gelpí[17,18], Andres C. García-Montero[19], Jaume Bertanpetit[20], Devin Absher[21], Richard M. Myers[21], Javier P. Gisbert[3,4,*] and Sara Marsal[1]

[1]Rheumatology Research Group, Vall d'Hebron Hospital Research Institute, Barcelona, Spain, [2]Gastroenterology and Hepatology Service, Hospital Universitari Germans Trias i Pujol, Badalona, Spain, [3]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), [4]Gastroenterology Service, Hospital Universitario de la Princesa and IP, Madrid, Spain, [5]Digestive System Service, Instituto Maimónides de Investigación Biomédica de Córdoba (IMIBIC), Hospital Universitario Reina Sofía, Universidad de Córdoba, Córdoba, Spain, [6]Digestive System Service, Hospital Clínico Universitario, Zaragoza, Spain, [7]Gastroenterology Department, Hospital Clínic de Barcelona, IDIBAPS, Barcelona, Spain, [8]Gastroenterology Service, Hospital Clínico Universitario, Santiago de Compostela, Spain, [9]Gastroenterology Service, Hospital General de Alicante, Alicante, Spain, [10]Gastroenterology Department, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain, [11]Digestive Medicine, University and Polytechnic La Fe Hospital, Valencia, Spain, [12]Gastroenterology Service, Complejo Hospitalario de León, León, Spain, [13]Gastroenterology Service, Hospital Universitari Mutua de Terrassa, Barcelona, Spain, [14]Gastroenterology Service, Hospital Clínico San Carlos, Madrid, Spain, [15]Gastroenterology Service, Hospital Universitario Puerta de Hierro, Madrid, Spain, [16]Department of Gastroenterology, IMIM. (Hospital del Mar Medical Research Institute), Pompeu Fabra University, Barcelona, Spain, [17]Life Sciences, Barcelona Supercomputing Centre, National Institute of Bioinformatics, Barcelona, Spain, [18]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain, [19]Banco Nacional de ADN Carlos III, University of Salamanca, Salamanca, Spain, [20]Nacional Genotyping Centre (CeGen), Universitat Pompeu Fabra, Barcelona, Spain and [21]HudsonAlpha Institute for Biotechnology, Alabama, USA

Received April 27, 2014; Revised July 24, 2014; Accepted July 29, 2014

The genetic analysis of ulcerative colitis (UC) has provided new insights into the etiology of this prevalent inflammatory bowel disease. However, most of the heritability of UC (>70%) has still not been characterized. To identify new risk loci for UC we have performed the first genome-wide association study (GWAS) in a Southern European population and undertaken a meta-analysis study combining the newly genotyped 825 UC patients and 1525 healthy controls from Spain with the six previously published GWAS comprising 6687 cases and 19 718 controls from Northern-European ancestry. We identified a novel locus with genome-wide significance at 6q22.1 [rs2858829, $P = 8.97 \times 10^{-9}$, odds ratio (OR) (95% confidence interval, CI) = 1.12 (1.08–1.16)] that was validated with genotype data from a replication cohort of the same Southern European ancestry consisting in 1073 cases and 1279 controls [combined $P = 7.59 \times 10^{-10}$, OR (95% CI) = 1.12 (1.08–1.16)]. Furthermore, we confirmed the association of 33 reported associations with UC and we nominally validated the GWAS results of nine new risk loci ($P < 0.05$, same direction of effect). SNP rs2858829 lies in an intergenic region and is a strong *cis*-eQTL for *FAM26F* gene, a gene that is shown to be selectively upregulated in UC colonic mucosa with active inflammation.

*To whom correspondence should be addressed at: Javier P. Gisbert, Gastroenterology Unit, Hospital Universitario de La Princesa, IP and CIBEREHD. Diego de León, 62, 28006 Madrid, Spain. Tel: +34 913093911; Fax: +34 914022299; Email: javier.p.gisbert@gmail.com

Our results provide new insight into the genetic risk background of UC, confirming that there is a genetic risk component that differentiates from Crohn's Disease, the other major form of inflammatory bowel disease.

INTRODUCTION

Ulcerative colitis [UC (MIM 191390)] and Crohn's Disease [CD (MIM 266600)] are the two major forms of inflammatory bowel disease, with a combined prevalence of ~0.4% in Caucasians. While CD inflammation can take place in any part of the gastrointestinal tract, UC is characterized by a chronic inflammation that is limited to the colonic mucosa. There is now clear evidence that genetic variation is a major risk factor for the development of UC. The UC genetic risk background is complex and it is composed of variants at multiple genomic loci, most of which are also risk factors for CD (1). Furthermore, many of the susceptibility genes known to date belong to common biological pathways like the response to bacterial molecules, T- and B-cell activation, JAK-STAT signaling, as well as interleukin 12, interferon-gamma and tumor necrosis factor cytokine pathways (2). Taken together, these results underscore the role of host defense against infection as one of the main biological mechanism in the pathogenesis of inflammatory bowel diseases (IBDs).

The cumulative risk exerted by the UC risk loci identified so far is <20% (1). Therefore, the likelihood that additional, undiscovered loci contribute to UC risk is very high. Identifying the genes that are specific for UC susceptibility is an important objective since it could contribute to the development of better diagnostic tools as well as more efficient therapies (3,4). Although several loci have been found to be specific for UC risk, there is yet no evidence of a specific genetic pathway that clearly differentiates UC from CD etiology (2,5). Consequently, there is still a need to find additional genetic factors and biological mechanisms that can help explain the differential tissue localization and inflammatory characteristics of UC.

Genome-wide association studies (GWAS) are the key approach to efficiently characterize the common influential genetic variation. Most of the reported GWAS in UC, however, have been performed in Caucasian populations of marked Northern-European ancestry (6–10). In the present study we have used, for the first time, a Southern European ancestry cohort of UC patients and controls to identify new risk loci using the GWAS approach. We have conducted a meta-analysis of our GWAS results with the previous evidence from the large American-European GWAS consortium performed on populations of Northern-European ancestry. Using an independent replication case–control cohort we have then performed a validation study of the most significant findings.

RESULTS

UC risk meta-analysis and validation

We analyzed the GWAS data of newly genotyped 825 UC patients and 1525 healthy controls from Spain and combined the results with the data from the six previously published GWAS comprising 6687 cases and 19 718 controls from Northern-European ancestry. After quality control filtering, a total of 546 271 SNPs were tested for association with UC risk using 796 patients and 1493 controls from Spain.

The meta-analysis of our UC GWAS with the previous GWAS on UC based on Northern-European ancestry populations identified a genome-wide significant association of SNP rs2858829 on chromosome 6q22.1 ($P = 8.97 \times 10^{-9}$, OR (95% CI) = 1.12 (1.08–1.16)). rs2858829 is an intergenic SNP located 13.5 kb upstream from the gene encoding family with sequence similarity 26, member F (*FAM26F*) protein and 9.5 kb downstream from the gene encoding dermatan sulfate epimerase (*DSE*) protein. Figure 1 shows the meta-analysis association results for *DSE-FAM26F* locus with UC risk.

As an additional evidence, we tested the new risk locus at 6q22.1 in a novel independent cohort of 1073 cases and 1279 cohorts from Spain. We validated the association in the independent replication analysis for SNP rs2858829 at 6q22.1 [$P = 0.03$, OR (95% CI) = 1.12 (1.01–1.25)]. Finally, combining the replication data with the GWAS data the association was more significant [$P = 7.59 \times 10^{-10}$, OR (95% CI) = 1.12 (1.08–1.16)].

Study of reported UC risk loci

According to the previously described effect sizes and risk allele frequencies of each of the previously reported UC risk loci, we had >80% power to replicate ($P < 0.05$) five UC risk loci with the sample size of our study. Using our GWAS case–control cohort we confirmed the association of 33 reported UC risk loci ($P < 0.05$, same direction of effect, Supplementary Material, Table S1).

Analysis of new candidate loci

In order to identify new risk loci for UC independently from the meta-analysis, we performed a validation study of the SNPs from the GWAS on the Spain case–control cohort that showed a strong evidence for association with the disease (P-value $< 5 \times 10^{-4}$). A total of 61 SNPs from genomic regions not previously associated with UC risk were selected for validation in the independent replication cohort (Supplementary Material, Table S2). Nine of these SNPs were nominally validated ($P < 0.05$, same direction of effect as in the GWAS), and three additional SNPs showed a trend for association ($P < 0.1$, same direction of effect). Table 1 shows the association results for the group of nine SNPs that were nominally replicated in the independent case–control cohort. Combined with the GWAS results, none of the SNPs reached a genome-wide threshold of significance. However, these results are clearly enriched for nominally associated SNPs (binomial test, $P = 0.0032$) indicating that there are significantly more SNPs replicated at $P < 0.05$ than expected by chance alone.

In silico analysis of 6q22.1 region functional regulation

We performed a screening for significant expression quantitative trait loci (eQTL) association of rs2858829 genomic region on

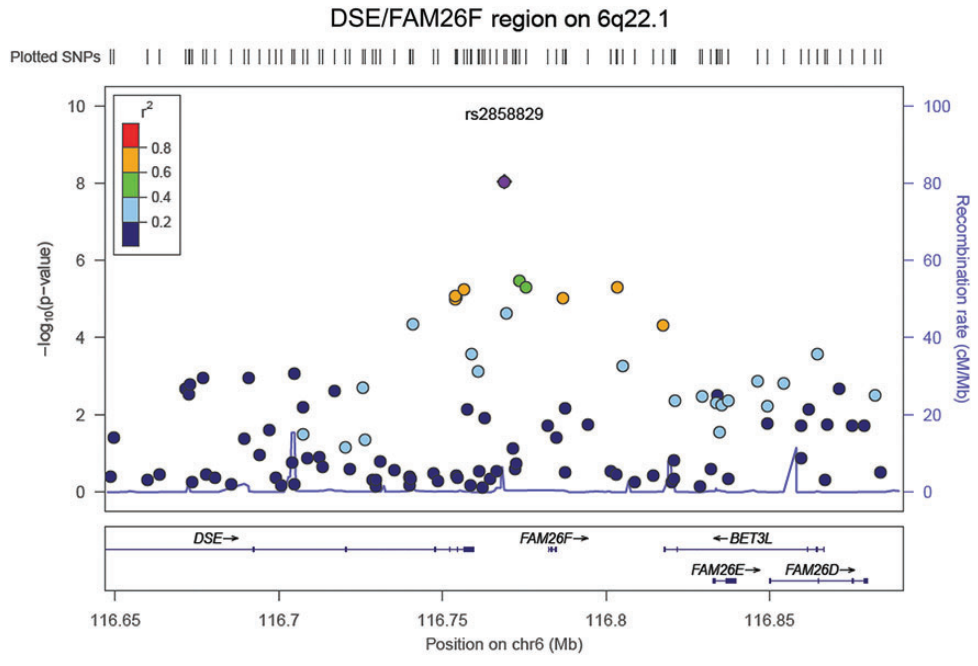

Figure 1. Meta-analysis association results for 6q22.1 locus with UC risk. Regional association plot with the significance [i.e. −log10(*P*-value), *y*-axis] of the SNPs in the associated *DSE-FAM26F* region as a function of the basepair location in chromosome 6q22.1 (*x*-axis). The most associated SNP with UC risk (i.e. rs2858829) is indicated as a diamond. The remaining SNPs are marked as circles with color coding indicating the level of LD of each SNP with rs2858829 SNP (i.e. r^2 values). The estimated recombination rates (i.e. centiMorgans/Megabase, right *y*-axis) are plotted as a continuous background line (blue). Genes in the associated 6q22.1 chromosome region and their transcription direction are also shown.

chromosome 6q22.1. In a large-scale eQTL study on the expression of human monocytes from 1490 individuals (11), we found a highly significant *cis*-regulatory evidence for rs2858829 locus with the expression of *FAM26F* gene. Although SNP rs2858829 was not directly genotyped, multiple neighboring SNPs showed a strong correlation with *FAM26F* expression. In particular, the variation at SNP rs9398434, which is at only 18 kb upstream from rs2858829 and is one of the markers showing highest linkage disequilibrium with this SNP ($r^2 = 0.82$, Supplementary Material, Fig. S1), shows a highly significant association with *FAM26F* expression ($P < 5 \times 10^{-26}$). eQTL analysis on the Hapmap reference lymphoblastoid cell lines using mRNA sequencing data (12), also shows a significant *cis*-eQTL association of this genomic region with *FAM26F* gene expression. SNP rs479454, located at 12 kb upstream from SNP rs2858829 and one of the most correlated SNPs with rs2858829 according to the 1000 Genomes Caucasian European data (Supplementary Material, Fig. S1), was also a *cis*-eQTL for *FAM26F* ($P = 0.00039$). Furthermore, recent high-throughput RNA sequencing data on 1000 Genomes Project individuals generated in the GEUVADIS international research project (13), clearly supports the association of this locus with the regulation of *FAM26F* expression. In this genomic region, the most significant eQTLs are associated with *FAM26F* expression levels (i.e. *P*-values ranging from 5e−8 to 9.6e−29). Importantly, rs2858829 itself is one of the top SNPs associated with *FAM26F* gene expression ($P = 6.1e-28$) and is in very high LD with the most significant eQTL in this region (i.e. SNP rs2637678, $P = 9.6e-29$, $r^2 = 0.95$).

Data from the ENCODE project indicated that SNP rs2858829 lies in a site that contains strong regulatory evidence (Supplementary Material, Fig. S2). We found strong evidence of histone acetylation, a pattern typical of enhancers. Also, 78 out of 125 cell types from ENCODE showed DNaseI hypersensitivity clusters overlapping this region, and data from chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments indicated a strong evidence for CCAAT/enhancer binding protein beta (CEBPB) transcription factor related activity.

Analysis of *FAM26F* expression in microarray studies

Given that the regulatory evidence for rs2858829 genomic region on 6q22.1 was clearly strong for *FAM26F* gene, we analyzed the expression of this gene in publicly available microarray studies involving UC. A total of 14 microarray studies on UC were found to be published in the Gene Expression Omnibus (GEO) database. After discarding studies on transformed cell lines ($n = 1$), with low sample sizes ($n = 2$) and studies using microarray platforms that lack probes to measure *FAM26F* ($n = 5$), 6 microarray studies were finally available to study *FAM26F* functional activity in UC (14–18). Four of these studies were performed in human samples and the remaining two were performed on colonic samples of the dextran sodium sulfate (DSS) colitis mouse model. In all microarray studies involving colon tissue samples (three studies in human samples and two in mouse samples), *FAM26F* was found to be consistently and significantly overexpressed in UC samples (Supplementary Material, Table S3). The most significant differential expression was found when comparing *FAM26F* mRNA levels in colonic mucosa from UC patients with active inflammation compared with normal mucosa from controls (4.7 log2 fold-change, $P = 1.88 \times 10^{-5}$) as well as paired colon biopsies of involved and

Table 1. GWAS candidate loci for UC risk replicated in the independent validation cohort

Chr	SNP	bp	Locus	Minor allele	OR (GWAS)	CI (GWAS)	*P*-value (GWAS)	MAF	OR (replication)	CI (replication)	*P*-value (replication)	*P*-value (combined)
2	rs11679592	2406488	*MYT1L* (flanking 3′)	C	1.28	(1.13–1.45)	6.40×10^{-5}	0.48	1.16	(1.03–1.31)	0.0097	9.40×10^{-6}
2	rs8179646	157598678	*GPD2* (flanking 3′)	T	1.31	(1.15–1.49)	3.70×10^{-5}	0.38	1.17	(1.03–1.34)	0.0094	5.50×10^{-6}
3	rs358803	55318709	*CACNA2D3-WNT5A* intergenic	G	1.27	(1.11–1.45)	3.30×10^{-4}	0.35	1.13	(0.99–1.29)	0.034	1.41×10^{-4}
4	rs11133504	58380322	*Chr4q12* (intergenic)	A	0.71	(0.60–0.83)	2.48×10^{-5}	0.14	0.75	(0.63–0.88)	0.00024	1.17×10^{-7}
6	rs1924466	17772764	*KIF13A* (intron)	A	1.28	(1.13–1.44)	9.80×10^{-5}	0.5	1.16	(1.02–1.31)	0.01	1.50×10^{-5}
7	rs4725479	152745810	*ACTR3B* (flanking 3′)	T	0.75	(0.66–0.85)	4.90×10^{-6}	0.34	0.84	(0.74–0.96)	0.0056	5.10×10^{-7}
10	rs4747437	22030023	*MLLT10* (intron)	T	1.27	(1.12–1.44)	1.74×10^{-4}	0.41	1.17	(1.03–1.32)	0.0091	2.28×10^{-5}
10	rs4246949	125517080	*CPXM2* (intron)	T	0.76	(0.66–0.87)	1.06×10^{-4}	0.23	0.82	(0.71–0.95)	0.004	6.70×10^{-6}
15	rs4776239	54967175	*UNC13C* (flanking 3′)	A	1.4	(1.2–1.67)	6.80×10^{-5}	0.19	1.2	(1.02–1.42)	0.017	1.70×10^{-5}

Results for the nine GWAS candidate genes for UC risk that are replicated in the independent validation cohort ($P < 0.05$). All loci show the same direction of effect as originally identified in the GWAS.
Chr: chromosome; bp: basepair; OR: odds ratio; CI: confidence interval for the odds ratio; MAF: minor allele frequency in the GWAS control cohort.

non-involved mucosa from UC patients with active disease (4.2 log2 fold-change, $P = 3.73 \times 10^{-5}$) (14).

In order to identify the biological pathway associated with *FAM26F*, we determined the genes showing the highest correlation to its expression in colon samples from the human and mouse UC studies (14–18). Using the Gene Ontology (GO) annotation of this set of genes, we performed an enrichment analysis study. The gene ontology showing the strongest statistical significance was 'immune response' (GO:0006955, $P = 6.5 \times 10^{-56}$) and was found in the study with human UC samples from different colonic localizations (15). Consequently, the ontologies associated with immune cell activation showed a highly significant overrepresentation in all UC microarray studies [i.e. GO:0006955—immune response, GO:0001775—cell activation and GO:0002684—positive regulation of immune system process, false discovery rate (FDR) corrected $P < 0.05$]. More specifically, we found a highly significant enrichment of ontologies associated with lymphocyte activation (GO:0045619—regulation of lymphocyte differentiation, GO:0051249—regulation of lymphocyte activation and GO:0050863—regulation of T-cell activation, FDR corrected $P < 0.05$) (Supplementary Material, Table S4).

Analyzing the intersection between *FAM26F* gene expression networks in human and mouse studies in UC colonic mucosa, we found 28 common genes (Supplementary Material, Table S5). The 'immune response' ontology was the most significantly overrepresented biological process, being present in 20 of the 28 common human–mouse genes ($P < 5 \times 10^{-10}$). Among these genes are genes that encode for chemokines (*CXCL10* and *CXCL9*), Tumor Necrosis Factor pathway associated proteins (*CD40* and *TNFSF13B*) as well as genes associated with immune cell signaling (*SLAMF8*, *PTPRC*, *LYN*, *LCP1*, *LAIR1*, *GIMAP4*, *CIITA*, *CD84* and *CD300A*).

Importantly, interrogating the GeneNetwork database, the most significant prediction functions for *FAM26F* showed a strong overlap with the functions found in colon-restricted studies (Supplementary Material, Table S6).

DISCUSSION

In this study, we present the first GWAS in UC using a Southern European population. We have performed a meta-analysis with the data from the six previous GWAS performed in Caucasian populations of Northern-European ancestry which represents the combined evidence of 28 755 individuals. In this meta-analysis we have found the new UC risk SNP rs2858829 at 6q22.1 which was also confirmed in an independent replication cohort. We have found strong evidence implicating this genomic region in the *cis*-regulation of *FAM26F,* a gene that is significantly overexpressed in inflamed colonic mucosa from UC patients. Finally, we have also performed a validation study of the most significant loci in our GWAS representing new candidate loci for UC. Although none reaches a genome-wide level of significance, we have found a significant enrichment for nominally replicated loci.

The associated marker rs2858829 is an intergenic SNP that is located in a site with strong genetic regulatory evidence. Among these enriched regulatory elements there is a significant evidence for CEBPB transcription factor binding overlapping the SNP

region. Of relevance, variation at *CEBPB* locus has been also found to be a common risk factor for inflammatory bowel diseases (2). Consequently, future experimental studies aimed at characterizing the implication of *DSE-FAM26F* region in UC should prioritize the confirmation of this mechanistic hypothesis.

SNP rs2858829 lies close to the proximal promoter of *FAM26F* gene. We have found strong evidence from eQTL studies that variation at this particular genomic region is associated with the regulation of the expression levels of *FAM26F*. To date, there is no known biological function for the protein encoded by this gene. Using functional data from functional studies in UC, however, we have found a strong association of *FAM26F* mRNA levels with the presence of the disease. Of relevance, when comparing the gene expression of *FAM26F* in involved mucosa from UC patients with the gene expression of non-involved mucosa of the same patients (14), we found a clear upregulation of the expression levels of the gene. Together, this evidence suggests that *FAM26F* is relevant in the inflammatory process that takes place in UC colonic mucosa and, consequently, the regulation of this gene could be important to the development of the disease.

The analysis of the *FAM26F* gene expression networks in colon tissue of human UC and mouse colitis models strongly supports the association of this gene with the immune response activity. Furthermore, analysis of global coexpression patterns in multiple tissues and traits also supports the implication of this gene with the regulation of the immune response activity. In human colon biopsies in particular (14–16), we found a highly significant overrepresentation of genes involved in human leukocyte antigen (*HLA*) mediated activation of lymphocytes. Importantly, genetic variation at the *HLA* region is the strongest genetic risk factor for UC (5) and is also the genetic region that shows a more distinct genetic effect compared with CD despite being also associated to its etiology (2). In our previous GWAS (19) SNP rs2858829 showed no association with CD risk (data not shown) and, to date, there is no evidence from other GWAS involving 6q22.1 locus with CD risk. Therefore, this evidence suggests that 6q22.1 locus is specific for UC risk, and it participates in the *HLA*-mediated activation of lymphocytes in UC colonic mucosa. Functional studies to determine the precise functional role of FAM26F in UC related inflammation are warranted.

We found highly suggestive evidence supporting the association of nine additional risk loci for UC. While these loci do not reach the genome-wide level of significance combining the GWAS and replication datasets, there are far more SNPs replicated at $P < 0.05$ than expected by chance alone. Consequently, these loci constitute strong candidate loci for UC risk. In particular, *4q12* intergenic SNP rs11133504 showed a nominal level of significance in the Northern-European ancestry GWAS (Supplementary Material, Table S7) and, therefore, could constitute another common risk factor for UC. Additional studies in independent case–control cohorts will be needed to identify the true risk variants from this set of candidate loci.

In conclusion, in the present GWAS we have found a new risk locus for UC at 6q22.1 and identified nine new candidate risk loci for UC. Furthermore, we have found substantial evidence that the genomic region at 6q22.1 is associated with the *cis*-regulation of *FAM26F* gene expression. Although no functionality has yet been described for *FAM26F*, we have found a significant overexpression of this gene in UC inflamed mucosa and a strong association to HLA-mediated immune cell activation. The results of this GWAS contribute to the identification of the genetic basis that is specific for UC risk and differential from CD.

MATERIALS AND METHODS

GWAS samples

To identify new loci associated with UC risk using the GWAS approach, we recruited 827 UC patients and 1558 controls from the Spanish population. The whole-blood sample collection and DNA extraction process was performed by the Immune-Mediated Inflammatory Disease Consortium (IMIDC) (19). In particular, UC samples were obtained from the Gastroenterology departments of 15 different Spanish University Hospitals (Supplementary Material, Table S8). All patients fulfilled the currently accepted diagnostic criteria for UC [mean $\pm$ standard deviation (SD) of years of follow-up since diagnosis: 13.5 ± 8.7] and were >18 years old (20). For all patients, the four grandparents were Caucasian and born in Spain.

Control individuals were recruited from blood bank donors attending at 13 Hospitals from different regions in Spain in collaboration with the Spanish National DNA Bank. All the controls included in this study were screened for the presence of UC or any autoimmune disorder, as well as for a family history of autoimmune disorders in first-degree relatives. Individuals with an autoimmune disease or a family history of autoimmune disease were excluded. To increase the hypernormality in this cohort, only control individuals who were >30 years old and who fulfilled the previous criteria were included in the study. In total, 1558 controls, 40% of whom were females, were genotyped. Of note, 98% of the control individuals were ≥ 40 years old at the time of recruitment (mean age $\pm$ SD 49.6 ± 7 years). For all controls, the four grandparents had also to be Caucasian and born in Spain. Table 2 summarizes the main epidemiological variables of the new Southern European GWAS cohort.

Replication sample

A replication sample of 1073 UC patients and 1279 healthy controls from Spain was collected. 773 of the replication cohort patients (72%) were collected by the IMIDC using the same selection criteria as in the GWAS stage. The remaining 300 UC patients (28%) were obtained from the ENEIDA project collection of the Spanish Working Group in Crohn's Disease and Ulcerative Colitis (GETECCU) (19). The former group of patients were also collected using the same selection criteria as in the GWAS phase, except that the origin of the four grandparents was not available. All control individuals were selected from the Spanish DNA Bank repository using the same criteria as in the discovery phase.

Informed consent

Informed consent was obtained from all participants, and protocols were reviewed and approved by local institutional review

Table 2. GWAS and replication cohort epidemiological and clinical statistics

	GWAS UC	GWAS controls	Replication UC	Replication controls
Female %	45.3%	40.0%	46.9%	48.2%
Age (mean years ± SD)	48.8 ± 14.13	50.7 ± 12.7	48.25 ± 14.7	43.3 ± 15.5
Familial CD %	1.5%	0%	2.8%	0%
Familial UC %	5.6%	0%	6.2%	0%
Smoking at diagnosis				
Yes %	26.5%		24.8%	
Ex-smoker %	20.2%	NA	21.3%	NA
Age at diagnosis (mean years, IQR)	33 (24–43)	NA	34 (26–45)	NA

The table shows the main epidemiological features of the discovery (GWAS) case–control cohort and the independent validation cohort. Familial CD/UC: presence of one or more first or second order relatives with CD or UC. NA: not applicable. IQR: interquartile range, SD: standard deviation. CD: Crohn's disease. UC: ulcerative colitis.

boards. This study was conducted in accordance with the Declaration of Helsinki principles.

Genotyping procedures

The genome-wide scan was performed using Illumina Quad610 Beadchips (Illumina, San Diego, CA, USA) on 827 UC patients and 1558 controls. The Quad610 arrays genotype >550 000 SNPs and have ∼60 000 probes specific for copy number variant (CNV) detection. GWAS genotyping was performed at the Centro Nacional de Genotipado (CeGen, Spain). After excluding mitochondrial, X and Y chromosome SNPs, a total of 600 470 markers were considered for GWAS analysis, of which 17 879 were exclusively CNV Probes. The SNP genotype calling was performed using Illumina GenomeStudio software v2010.1 (Illumina), and CNV calls were performed using CNstream software (21). Only samples that had a >95% genotype completion rate were considered for analysis (99% of samples), and only SNPs that had a >95% call rate (99.4% of SNPs) and a minor allele frequency (MAF) >0.05 (90.5% of SNPs) and that showed Hardy–Weinberg Equilibrium (99.5% of SNPs, $P > 0.0001$ in controls) were considered for association analysis. Also, SNPs showing a differential missingness rate between cases and controls (0.1% SNPs significantly different at $P < 1 \times 10^{-7}$) were excluded from any further analysis. From this set of quality control-filtered SNPs ($n =$ 533 476) inferred the main axis of variation using the principal components approach implemented in EIGENSTRAT software (22). Using the 10 first principal components, $n = 39$ outlying samples were excluded. Supplementary Material, Figure S3A shows the case and control distribution according to the first and second principal components after excluding the outliers. Supplementary Material, Figure S3B also shows the projection of our GWAS cohort onto the eigenvectors calculated from the Hapmap reference populations. The genomic inflation factor was close to 1 ($\lambda = 1.04$), and consequently, no adjustment was performed over the GWAS association statistics.

Replication genotyping was performed at the HudsonAlpha Institute for Biotechnology (Huntsville, Alabama, USA) using the Illumina GoldenGate assay (Illumina) on 1073 UC patients and 1279 controls. Similar quality control measures were applied, including genotyping call rate >90%, sample completion rate >90%, HWD P-value of control group $P > 0.001$. Five percent samples were genotyped in duplicate giving an estimated 1% genotyping error rate and no marker was excluded for deviation from Hardy–Weinberg Equilibrium.

Statistical analyses

GWAS meta-analysis

Using the set of QC-filtered GWAS data, we performed a case–control association analysis using an allelic χ^2 test of association using PLINK software version 1.07 (23). First, we evaluated the association of previously reported UC risk loci in our cohort. We calculated the statistical power of our GWAS cohort to replicate at $\alpha = 0.05$ the reported UC risk loci association using the Genetic Power Calculator software (24). For this objective we used the minor allele frequencies and effect sizes reported in the Northern-European population UC meta-analysis (2). We performed the allelic χ^2 test for association for each of the 133 reported UC risk SNPs. In those cases were the SNP was not directly genotyped, we imputed the SNP genotype using the MACH imputation software and the 1000 Genomes Caucasian European reference dataset (25).

We performed a meta-analysis combining our GWAS results with the results from six previous GWAS on 6687 UC patients and 19 718 controls from Northern-European ancestry available at the International IBD Genetics Consortium (IIBDGC, http://www.ibdgenetics.org/) (2). For this objective we used the METAL meta-analysis software tool (26), which combined the association statistics from the different studies to generate a combined P-value. In this approach, the significance values are weighted according to the sample size of each study.

In order to validate the risk locus identified in the meta-analysis we performed an allelic χ^2 test of association in the independent replication sample consisting of 1073 cases and 1279 controls from Spain. In order to combine the resulting association result with the previous GWAS results we performed a meta-analysis using the METAL software.

Candidate loci replication

To identify additional UC risk loci, we selected those SNPs from our GWAS with a significance $P < 5 \times 10^{-4}$ and representing loci not previously associated with UC. These markers were genotyped and analyzed in the independent validation cohort of cases and controls using the allelic χ^2 test of association.

In silico functional study of 6q22.1 locus

eQTL screening

We used the eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl) developed at the Pritchard Lab (University of Chicago, USA) to identify the presence of regulatory evidence at the 6q22.1 locus. This software tool summarizes, for a specific genomic location, the results obtained from multiple eQTL studies involving different human tissues and cell types. Also, we screened for gene expression regulatory evidence in the associated *DSE-FAM26F* region using the recent eQTL results generated from the Genetic European Variation in Health and Disease project (GEUVADIS, www.geuvadis.org) (13). In this international large-scale mRNA sequencing project, the transcriptome from lymphoblastoid cell lines from 1000 Genomes Project individuals has been sequenced and deeply analyzed to identify relevant functional effects in the genome.

Gene expression analysis from microarray studies

In order to characterize the potential functionality of *FAM26F* we first looked for all relevant microarray studies in UC available at the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/). Using the search term 'ulcerative colitis', a total of 14 different and GEO datasets were identified: three performed using the DSS colitis mouse model and 11 performed using human UC samples. From this initial set of studies, we excluded those based on transformed cell lines ($n = 1$, GDS1478), studies with only one individual per group ($n = 2$, GDS4366 and GDS560) as well as studies with microarray platforms lacking probes to measure the expression of *FAM26F* gene ($n = 5$, GDS1615, GDS2642, GDS2014, GDS559 and GDS1330). Consequently, we had a final set of six microarray studies from which we were able to study the gene expression pattern of *FAM26F* in relation to UC pathology (GEO datasets: GDS4367, GDS3859, GDS4270, GDS4365, GDS3268 and GDS3119) (14–18, 27). From all the microarray datasets, we obtained the normalized log2 gene expression data for the probes capturing *FAM26F* gene expression. In those cases were the study design had a meaningful grouping, we performed a statistical test for differential expression (*t*-test, $P < 0.05$ for significance). Supplementary Material, Table S9 summarizes the main traits of the six microarray studies in UC used in this study.

In order to identify the biological pathway where *FAM26F* is involved, we selected from each microarray study the group of genes showing the strongest statistical correlation with *FAM26F* expression (i.e. Pearson's product moment correlation test). This statistical network of highly correlated genes was subsequently used to determine the enrichment of specific biological features. For this objective, we used the Gene Ontology (GO) database annotation of human and mouse genes (28). Using these gene annotations we tested for the overrepresentation of each GO using the Fisher's exact test method as described previously (29). All statistical analyses were performed using the *R* statistical software and the Bioconductor set of libraries for genomic analysis (30).

In a second microarray-based functional characterization approach we used the GeneNetwork online tool (www.GeneNetwork.nl). In this method, co-regulation information obtained from a compendium of ~80 000 microarray expression arrays and, therefore, provides a more global (and non-colon-specific) prediction of the functionality of *FAM26F*.

ENCODE Annotation

In order to identify useful fine mapping and mechanistic information, we screened for overlap of the *DSE-FAM26F* region harboring the associated SNP rs2858829 using the Encyclopedia of DNA Elements (ENCODE) database (31).

DATA AVAILABILITY

The complete GWAS data (SNP, chromosome, basepair, odds ratio and *P*-value) are available for download at http://www.urr.cat/data/.

SUPPLEMENTARY DATA

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the use of the published results from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) meta-analysis (results available at: http://www.ibdgenetics.org/). We thank Genoma España Foundation, the Barcelona Supercomputing Centre (BSC), Instituto Nacional de Bioinformática (INB), the Centro Nacional de Genotipado (CeGEN), the Banco Nacional de ADN (USAL) and the Hudson Alpha Institute for Biotechnology for their support.

Conflict of Interest statement: None declared.

FUNDING

This study was supported by of the Ministry of Economy and Competitiveness, Spain (grant numbers PSE-010000-2006-6, IPT-010000-2010-36).

REFERENCES

1. Khor, B., Gardet, A. and Xavier, R.J. (2011) Genetics and pathogenesis of inflammatory bowel disease. *Nature*, **474**, 307–317.
2. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A. *et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119–124.
3. Okada, Y., Yamazaki, K., Umeno, J., Takahashi, A., Kumasaka, N., Ashikawa, K., Aoi, T., Takazoe, M., Matsui, T., Hirano, A. *et al.* (2011) HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology*, **141**, 864–871 e861–865.
4. Sartor, R.B. (2006) Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nat. Clin. Pract. Gastroenterol. Hepatol.*, **3**, 390–407.
5. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A. *et al.* (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.*, **43**, 246–252.
6. McGovern, D.P., Gardet, A., Torkvist, L., Goyette, P., Essers, J., Taylor, K.D., Neale, B.M., Ong, R.T., Lagace, C., Li, C. *et al.* (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.*, **42**, 332–337.
7. Franke, A., Balschun, T., Sina, C., Ellinghaus, D., Hasler, R., Mayr, G., Albrecht, M., Wittig, M., Buchert, E., Nikolaus, S. *et al.* (2010)

Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). *Nat. Genet.*, **42**, 292–294.
8. Kugathasan, S., Baldassano, R.N., Bradfield, J.P., Sleiman, P.M., Imielinski, M., Guthery, S.L., Cucchiara, S., Kim, C.E., Frackelton, E.C., Annaiah, K. *et al.* (2008) Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.*, **40**, 1211–1215.
9. Imielinski, M., Baldassano, R.N., Griffiths, A., Russell, R.K., Annese, V., Dubinsky, M., Kugathasan, S., Bradfield, J.P., Walters, T.D., Sleiman, P. *et al.* (2009) Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.*, **41**, 1335–1340.
10. Silverberg, M.S., Cho, J.H., Rioux, J.D., McGovern, D.P., Wu, J., Annese, V., Achkar, J.P., Goyette, P., Scott, R., Xu, W. *et al.* (2009) Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.*, **41**, 216–220.
11. Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H. *et al.* (2010) Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS One*, **5**, e10693.
12. Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E.T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.
13. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
14. Planell, N., Lozano, J.J., Mora-Buch, R., Masamunt, M.C., Jimeno, M., Ordas, I., Esteller, M., Ricart, E., Pique, J.M., Panes, J. *et al.* (2013) Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. *Gut*, **62**, 967–976.
15. Noble, C.L., Abbas, A.R., Cornelius, J., Lees, C.W., Ho, G.T., Toy, K., Modrusan, Z., Pal, N., Zhong, F., Chalasani, S. *et al.* (2008) Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut*, **57**, 1398–1405.
16. Olsen, J., Gerds, T.A., Seidelin, J.B., Csillag, C., Bjerrum, J.T., Troelsen, J.T. and Nielsen, O.H. (2009) Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflamm. Bowel. Dis.*, **15**, 1032–1038.
17. Tang, A., Li, N., Li, X., Yang, H., Wang, W., Zhang, L., Li, G., Xiong, W., Ma, J. and Shen, S. (2012) Dynamic activation of the key pathways: linking colitis to colorectal cancer in a mouse model. *Carcinogenesis*, **33**, 1375–1383.
18. Fang, K., Bruce, M., Pattillo, C.B., Zhang, S., Stone, R. 2nd, Clifford, J. and Kevil, C.G. (2011) Temporal genomewide expression profiling of DSS colitis reveals novel inflammatory and angiogenesis genes similar to ulcerative colitis. *Physiol. Genomics*, **43**, 43–56.
19. Julia, A., Domenech, E., Ricart, E., Tortosa, R., Garcia-Sanchez, V., Gisbert, J.P., Nos Mateu, P., Gutierrez, A., Gomollon, F., Mendoza, J.L. *et al.* (2012) A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300. *Gut*, **62**, 1440–1445.
20. Van Assche, G., Dignass, A., Panes, J., Beaugerie, L., Karagiannis, J., Allez, M., Ochsenkuhn, T., Orchard, T., Rogler, G., Louis, E. *et al.* (2010) The second European evidence-based Consensus on the diagnosis and management of Crohn's disease: definitions and diagnosis. *J. Crohns Colitis*, **4**, 7–27.
21. Alonso, A., Julia, A., Tortosa, R., Canaleta, C., Canete, J.D., Ballina, J., Balsa, A., Tornero, J. and Marsal, S. (2010) CNstream: a method for the identification and genotyping of copy number polymorphisms using Illumina microarrays. *BMC Bioinformatics*, **11**, 264.
22. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
23. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
24. Purcell, S., Cherny, S.S. and Sham, P.C. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, **19**, 149–150.
25. Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.
26. Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.
27. Kabakchiev, B., Turner, D., Hyams, J., Mack, D., Leleiko, N., Crandall, W., Markowitz, J., Otley, A.R., Xu, W., Hu, P. *et al.* (2010) Gene expression changes associated with resistance to intravenous corticosteroid therapy in children with severe ulcerative colitis. *PLoS One*, **5**, e13085.
28. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32 Database issue**, D258–D261.
29. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
30. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
31. 2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.