Journal of Crohn's and Colitis, 2023, **17**, 277–288
https://doi.org/10.1093/ecco-jcc/jjac121
Advance access publication 16 September 2022
Original Article

An Extremes of Phenotype Approach Confirms Significant Genetic Heterogeneity in Patients with Ulcerative Colitis

Sally Mortlock,[a] Anton Lord,[b,c] Grant Montgomery,[a] Martha Zakrzewski,[b] Lisa A. Simms,[b] Krupa Krishnaprasad,[b] Katherine Hanigan,[b] James D. Doecke,[d] Alissa Walsh,[e] Ian C. Lawrance,[f] Peter A. Bampton,[g] Jane M. Andrews,[h] Gillian Mahy,[i] Susan J. Connor,[j,k] Miles P. Sparrow,[l] Sally Bell,[m] Timothy H. Florin,[n,o] Jakob Begun,[n,p,o] Richard B. Gearry,[q] Graham L. Radford-Smith[b,r,o]

[a]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia
[b]QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia
[c]Centre for Health Services Research, University of Queensland, Brisbane, QLD, Australia
[d]Australian eHealth Research Centre, CSIRO, Brisbane, QLD, Australia
[e]Department of Gastroenterology, John Radcliffe Hospital, Headington, Oxford, UK
[f]Centre of Inflammatory Bowel Diseases, Saint John of God Hospital Subiaco, University of Western Australia, WA, Australia
[g]Flinders Medical Centre, Adelaide, SA, Australia
[h]Department of Gastroenterology and Hepatology, Royal Adelaide Hospital & University of Adelaide, Adelaide, SA, Australia
[i]Department of Gastroenterology and Hepatology, Townsville University Hospital, Townsville, QLD, Australia
[j]Department of Gastroenterology and Hepatology, Liverpool Hospital, Sydney, NSW, Australia
[k]South Western Sydney Clinical School, University of New South Wales, Sydney, NSW, Australia
[l]Department of Gastroenterology, Alfred Health, Melbourne, VIC, Australia
[m]Department of Gastroenterology and Hepatology, Monash Health, Melbourne, VIC, Australia
[n]Inflammatory Bowel Diseases Group, Translational Research Institute, Brisbane, QLD, Australia
[o]Faculty of Medicine, University of Queensland, Brisbane, QLD, Australia
[p]Inflammatory Disease Biology and Therapeutics Group, Translational Research Institute, Brisbane, QLD, Australia
[q]Department of Medicine, University of Otago, Christchurch, New Zealand
[r]Department of Gastroenterology and Hepatology, Royal Brisbane and Women's Hospital, Brisbane, QLD, Australia

Corresponding author: Graham Radford-Smith, Gut Health Lab, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia. Tel: +617 3362 0499; Fax: +617 3009 0053; Email: g.radfordsmith@uq.edu.au

Abstract

Background and Aims: Ulcerative colitis [UC] is a major form of inflammatory bowel disease globally. Phenotypic heterogeneity is defined by several variables including age of onset and disease extent. The genetics of disease severity remains poorly understood. To further investigate this, we performed a genome wide association [GWA] study using an extremes of phenotype strategy.

Methods: We conducted GWA analyses in 311 patients with medically refractory UC [MRUC], 287 with non-medically refractory UC [non-MRUC] and 583 controls. Odds ratios [ORs] were calculated for known risk variants comparing MRUC and non-MRUC, and controls.

Results: MRUC–control analysis had the greatest yield of genome-wide significant single nucleotide polymorphisms [SNPs] [2018], including lead SNP = rs111838972 [OR = 1.82, $p = 6.28 \times 10^{-9}$] near MMEL1 and a locus in the human leukocyte antigen [HLA] region [lead SNP = rs144717024, OR = 12.23, $p = 1.7 \times 10^{-19}$]. ORs for the lead SNPs were significantly higher in MRUC compared to non-MRUC [$p < 9.0 \times 10^{-6}$]. No SNPs reached significance in the non-MRUC–control analysis (top SNP, rs7680780 [OR 2.70, $p = 5.56 \times 10^{-8}$). We replicate findings for rs4151651 in the Complement Factor B [CFB] gene and demonstrate significant changes in CFB gene expression in active UC. Detailed HLA analyses support the strong associations with MHC II genes, particularly *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* in MRUC.

Conclusions: Our MRUC subgroup replicates multiple known UC risk variants in contrast to non-MRUC and demonstrates significant differences in effect sizes compared to those published. Non-MRUC cases demonstrate lower ORs similar to those published. Additional risk and prognostic loci may be identified by targeted recruitment of individuals with severe disease.

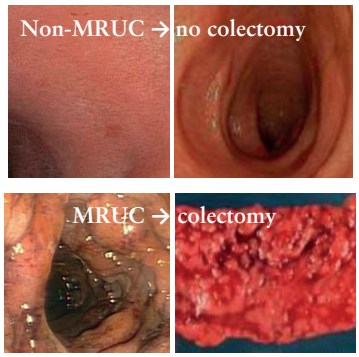
Key Words: Ulcerative colitis; genetics; disease severity

Graphical Abstract

An extremes of phenotype approach confirms significant genetic heterogeneity in patients with ulcerative colitis

Aim: To investigate the genetic contribution to UC heterogeneity

Background: unclear as to what determines disease severity in ulcerative colitis



What we know

Predictors of colectomy at diagnosis:

Age, extent, CRP or ESR, need for steroids

Predictors of acute severe UC at diagnosis:

Extent, CRP, Haemoglobin

Genetic factors? SNP rs9268877 associated with poor prognosis in Korean UC population

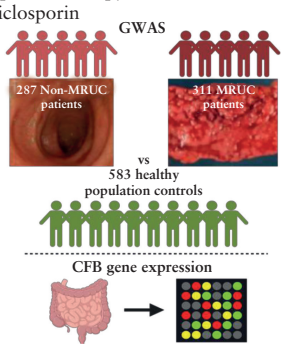
Definitions and Methods:

Non-medically refractory UC

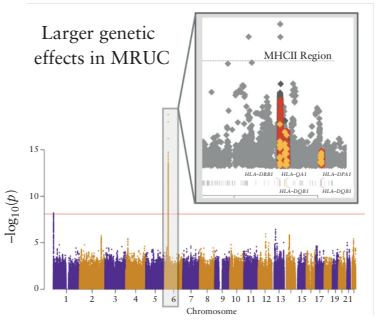
- Duration and follow up for 10+ years
- Documented colonoscopic follow up
- No biologics
- No immunomodulator > 6 months
- No steroid dependence, no surgery

Medically refractory UC requiring colectomy

- Chronic: inadequate response to steroids + immunomodulator ± biologic(s)
- Acute: documented inadequate response to inpatient therapy - IV steroid ± infliximab or Ciclosporin

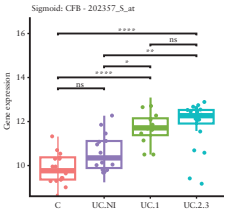


Results



MRUC: 2018 SNPS with GW significance

Non_MRUC: No SNPS with GW significance



SNP rs4151651 in Complement Factor B (CFB) replicated. CFB expression correlates with disease severity (left).

1. Introduction

Ulcerative colitis [UC] is a chronic inflammatory disorder of the large intestine and one of the major forms of inflammatory bowel disease [IBD]. IBD now has a global distribution and affects approximately 6.8 million of the world's population.[1] Intestinal inflammation in UC is typically limited to the colonic mucosa and superficial submucosa. A number of factors have been implicated in contributing to disease severity including age at onset, disease extent and genetic risk factors.[2–6] Individuals with mild disease (non-medically refractory UC [non-MRUC]) may achieve adequate disease control through lifestyle modification and limited medical therapy such as the use of 5-aminosalicylates.[7,8] Those with severe disease (medically refractory UC [MRUC]) are characterized by either severe attacks requiring hospitalization [acute severe UC; ASUC] and/or frequent disease flares that require corticosteroids, immunomodulators and biologic drugs [chronic refractory UC]. If intensive medical therapy fails to achieve sustained remission of symptoms, then surgery is regarded as a safe and effective option in achieving a reasonable quality of life.[9,10] The lifetime risk of ASUC is up to 25% and carries an additional increased risk of colectomy of up to 40% as compared to less than 15% in those individuals without a history of acute severe disease.[11] If we were able to predict disease severity at an early stage, more rapid escalation to advanced therapies may be instituted to attempt to change the natural history of the disease. Early identification of patients requiring more aggressive treatment options could assist in selecting the optimal treatment strategy on a patient-by-patient basis.

Prognostic factors that will assist both the patient and the treating team in predicting the course of the disease have been the subject of several studies. Clinical risk factors that may assist in predicting risk of colectomy specifically at the time of diagnosis include extent of disease, age, need for systemic corticosteroids, and either C-reactive protein [CRP] or erythrocyte sedimentation rate [ESR].[3] Factors that can predict future risk of ASUC at diagnosis include disease extent, CRP and haemoglobin.[12] UC has an estimated heritability of 67% and the amount of variation captured by single nucleotide polymorphisms [SNPs] has been estimated as 33%.[13,14] The genetic basis of UC disease severity is informed by a limited number of studies that have either focused on individual genes or regions such as the major histocompatibility complex [MHC][15,16] and more recently genome wide association studies [GWAS] and Immunochip studies.[6,17–24] These have identified more than 120 independent loci associated with UC. An international study of IBD sub-phenotypes used a survival analysis to investigate markers associated with colectomy in UC. Five SNPs, all at 6p21 within the MHC, achieved genome-wide significance with the top SNP being rs4151651 (hazard ratio [HR] 1.72, 95% confidence interval [CI] 1.47–2.00).[17] This SNP is located in exon 5 of the Complement Factor B [*CFB*] gene on chromosome 6. *CFB* was also one of seven novel UC susceptibility genes identified in the first GWAS undertaken in the genetically distinct North Indian population.[25]

Monogenic mutations have been identified in specific IBD extremes of phenotype such as very early-onset disease. However, these do not explain the majority of phenotypic variance in UC. Both MRUC and non-MRUC may represent polygenic conditions, sharing variants in the same genes that determine UC in the general population, and also harbouring genes novel to these extremes. In support of this, Lee and colleagues identified a single SNP intergenic between HLA-DRA and HLA-DRB, rs9268877, that was associated with a poor UC prognosis.[18] Potential increases in statistical power afforded through analysis of extreme phenotypes has become an established approach to

investigate complex disease.[26,27] However, one of the challenges in this approach is to select criteria that accurately reflect the extreme ends of the disease under study. Hence, detailed clinical data on each UC case including evidence of long-term follow up confirming the chronic nature of the disease, colonoscopic and histological evidence of disease activity, and documentation of all treatments received will add to the quality of the study.

Given the limited treatment options currently available for MRUC, in particular ASUC, there is an ongoing need to further define the genetic contribution to disease heterogeneity, to better understand the pathogenesis of severe disease, and identify novel and effective treatment targets. In this study we used a novel UC extremes of phenotype approach, carefully selecting criteria to define individuals with either MRUC or persistent, non-MRUC with requirements for clear documentation of disease history either to colectomy for those with MRUC or a disease duration of at least 10 years in those with non-MRUC. The aims of the study are to further define the genetic differences between these subphenotypes and determine the value of a genetic risk score based upon currently available SNP data in predicting disease severity and hence UC outcome.

2. Methods

2.1. Patient samples and DNA isolation

Patients, and healthy controls, for this study were recruited from sites within the Australia and New Zealand IBD Consortium [ANZIBDC]. Briefly, consecutive patients with a diagnosis of UC based on validated criteria[28] were invited to join the ANZIBDC research programme at each participating site. Phenotype data were based upon the Montreal classification[29] together with additional detailed clinical data including smoking behaviour, medications and surgery. Predetermined criteria were used to classify patients as either MRUC or non-MRUC. Controls were recruited from the general population using the Australian electoral roll, and in proportions that reflected the age [in 2-year age bands] and sex distribution of IBD cases in the IBD database.

Non-MRUC was defined as those individuals having a minimum disease duration and follow up of 10 years, including both clinical and endoscopic assessments, during which the patient was well maintained on oral and/or rectal 5-aminosalicylate therapy with oral corticosteroids limited to one course per 12 months, and with no history of corticosteroid dependence or intravenous corticosteroids. Patients with any history of immunomodulator therapy use of greater than 6 months and/or any biologic therapy were not considered as having non-MRUC. MRUC included two sub-phenotypes defined as follows: (1) patients requiring colectomy for chronic refractory disease despite treatment with corticosteroids, an immunomodulator and/or a biologic medication; or (2) patients requiring colectomy for acute severe disease having failed to respond to intravenous corticosteroids and/or rescue therapy with either infliximab or ciclosporin. Acute severe disease was defined by the Truelove and Witts criteria for all cases.[30] An additional 41 cases of ASUC, all satisfying the Truelove and Witts criteria, and who responded to rescue therapy with either infliximab or ciclosporin with persisting response to 12 months, were included in the combined UC cohort for all case-control analyses together with the non-MRUC and MRUC [colectomy] subgroups defined above and in Table 1.

Given the published findings concerning rs4151651 within the CFB gene from both European and North Indian cohorts, we undertook gene expression analysis for *CFB* using colonic tissue biopsies collected at the study's lead site [QIMR Berghofer MRI]. Biopsies were collected by the principal investigator at the time of endoscopic examination. A total of 46 UC patients and 22 healthy controls [undergoing a screening colonoscopy and with normal endoscopic findings] were included in this analysis. Biopsies were taken from the sigmoid colon using a standard biopsy forceps technique, immediately snap frozen and stored at −80°C for RNA extraction, as previously described. Adjacent biopsies were taken from this segment for histological analysis. An inflammation score was generated for each biopsy site and each case using a validated scoring system[31] [non-inflamed, n = 14; mild, n = 12; moderate, n = 16; severe, n = 4]. RNA isolation and microarray analysis were performed as described below.[32]

Written informed consent was obtained from each patient as approved by the ethics committee of each member site. All participants were aged over 18 years at the time of recruitment to the study. A blood sample was obtained from each participant. DNA isolation and quantification were performed using well-established protocols and as previously described.

2.2. Genotyping

All genotyping was performed using Infinium technology [Illumina], specifically the OmniExpress chip containing 733 202 SNPs. Quality control [QC] was performed on genotypes using PLINK.[33,34] Call rates <0.95, SNPs with a mean GenomeStudio GenCall score <0.7, Hardy–Weinberg equilibrium $p < 10^{-6}$ and MAF <0.05 were excluded. Cryptic relatedness between individuals was identified by calculating a genomic relationship matrix in GCTA.[35] Ancestry outliers were identified using data from 1000 Genomes populations and principal components generated in GCTA. A total of 575 330 SNPs in 1222 individuals remained for imputation. Genotypes were phased using ShapeIt V2 and imputed using the 1000 Genomes Phase 3 V5 reference panel on the Michigan Imputation Server.[36] Post-imputation QC was performed in PLINK removing imputed SNPs with low MAF [<0.05] and poor imputation quality [$R^2 < 0.8$] leaving 6273 901 autosomal SNPs for analysis.

The MHC region on chromosome 6 was also imputed separately using the Multi-ethnic HLA [version 1.0 2021] reference panel[37] on the Michigan Imputation Server.[36] Post-imputation QC was performed in PLINK removing imputed SNPs with low MAF [<0.05] and poor imputation quality [$R^2 < 0.8$] leaving 120 classical *HLA* alleles, 2298 amino acids in HLA proteins, at 4-digit resolution, and 38 536 SNPs for analysis.

2.3. Data processing

2.3.1. Statistical analysis

GWAS analysis was performed for the combined UC cohort [639 cases and 583 controls], and non-MRUC [287 cases], MRUC [with colectomy, 311 cases], ASUC [148 cases] and chronic refractory UC [157 cases] separately, using logistic regression in PLINK. The first five principal components were used as covariates to account for population stratification and the genomic inflation factor was calculated [λ = 1.02]. Significant SNPs which survived the genome-wide correction [$p < 5 \times 10^{-8}$] were cross checked

Table 1. Cohort demographics. Numbers represent mean ± SD or absolute count [%] where appropriate. Percentages are calculated excluding missing data. Significance was calculated using either a Chi squared test or two-sample *t*-test as appropriate

	Control	Non-MR UC	MRUC	*p* value
Demographics				
n	583	287	311	
Female [%]	337 [57.8]	156 [54.4]	146 [47.2]	0.011
Smoking				
Ever [at diagnosis]	256 [44.4]	73 [26.9]	84 [29.7]	1.032×10^{-13}
At follow-up	–	9 [6.7]	7 [6.7]	1
Family history of IBD [%]	–	46 [25.7]	34 [31.8]	0.356
Disease features				
Age at diagnosis [mean ± SD]	–	35.6 [15.1]	32.6[14.2]	0.001
Disease duration [years ± SD]	-	20.4 [11.4]	11.9 [10.1]	$<2.2 \times 10^{-16}$
Maximum disease extent [%]				$<2.2 \times 10^{-16}$
E1	–	69 [24.0]	3 [1.0]	
E2	–	121 [42.2]	70 [22.5]	
E3	–	95 [33.1]	233 [74.9]	
Data not available	–	2 [0.7]	5 [1.6]	
Colectomy	0 [0.0]	0 [0.0]	311 [100.0]	
Colectomy date [years post diagnosis]	–	–	6.4 [7.6]	
Colectomy reason				
Chronic refractory UC	–	–	157 [50.5]	
Acute severe UC	–	–	148 [47.6]	
CRC/dysplasia with refractory disease	–	–	2 [0.6]	
Data not available[a]	–	–	4 [1.3]	
Treatment				
Anti-TNF [%]	–	0 [0.0]	106 [37.4]	
Adalimumab [%]	–	0 [0.0]	13 [5.7]	
Infliximab [%]	–	0 [0.0]	83[37.2]	
Other anti-TNF agent [%]		0 [0.0]	10 [3.2]	
Cyclosporin [%]	–	0 [0.0]	63 [23.4]	
5ASA	–	140 [96.6]	71[76.3]	4.5×10^{-6}
Oral steroids [%]	–	91 [50.6]	113 [94.1]	5.87×10^{-15}
Immunomodulator [ever]				
Yes	–	13 [4.6]	241 [79.5]	<2.2e-16

IBD, inflammatory bowel disease; UC, ulcerative colitis; TNF, tumour necrosis factor; 5ASA, 5-aminosalicylic acid.
[a]Disease severity confirmed on histology.

against known SNPs for UC, Crohn's disease [CD] and combined IBD. A post-hoc analysis was performed restricting the number of SNPs tested to only 123 independent SNPs known to associate with UC from previously published GWAS in European cohorts.[17,19–24] Results for this analysis were considered statistically significant if a *p*-value <0.05 was obtained after Bonferroni correction including all tests from the combined group, non-MRUC only, MRUC only and MRUC sub-phenotypes [acute severe and chronic refractory] [$k = 615$, critical $\alpha = 8.13 \times 10^{-5}$]. Odds ratios [ORs] for previously reported SNPs, which were significant in our dataset, were compared to investigate the consistency in effect sizes between studies and disease severity. Differences in ORs between UC subgroups and between this cohort and published ORs were assessed using the Welch modified two-sample *t*-test.

GWAS analysis was repeated for the HLA region for the combined UC cohort, and non-MRUC, MRUC and MRUC sub-phenotypes separately, using logistic regression in PLINK and including the first five principal components as covariates. A genome-wide threshold of $p < 5 \times 10^{-8}$ was used to define significant variants, and independent signals in the region were identified by conditioning on the lead SNP. Due to the extensive linkage disequilibrium [LD] in the regions, a haplotype-based [omnibus] association test was also performed using HLA-TAPAS[37] to determine the association between each amino acid position and UC, accounting for the multiple amino acid residue changes occurring at that position.

With the current sample size we had >99% power to detect significant associations [$p < 0.05$] with MAF > 0.05 and OR = 3, >85% power with MAF > 0.1 and OR = 1.5 in the combined dataset, 99% power to detect significant associations with MAF > 0.05 and OR = 3, and >75% power with MAF > 0.1 and OR = 1.5 in the MRUC or non-MRUC subgroups [estimated using the 'genpwr' R package].[38] This is

consistent with what we see in the replication results in that we were able to detect the SNPs with larger effects.

To assess if the predictive ability of SNPs differs between disease subtypes, a genetic risk score [GRS] was calculated using summary statistics obtained from Liu *et al*.[19] Summary data-based best linear unbiased prediction [sBLUP] was used to assign an effect size to each allele in the dataset based on the aforementioned summary statistics.[35] Individual GRSs were then calculated using the SNP effect estimates in PLINK. Two-sample *t*-testing was performed to test the association between GRS and disease by testing non-MRUC, MRUC and MRUC sub-phenotypes, and the combined UC cohort [n = 639] against the control cohort independently. Additional *t*-tests between GRS of non-MRUC, MRUC and MRUC sub-phenotypes were also performed. GRSs were also binned into deciles and the ORs of UC vs control were calculated using the lowest decile as a reference.

To investigate the association between the UC GRS and clinical factors, risk scores were regressed against disease extent [proctitis, n = 73; left-sided, n = 207; extensive, n = 352; total = 634] and age at diagnosis [n = 632] using the entire UC cohort. Differences in the mean disease extent and age of diagnosis between cases within the top and bottom 10% of risk scores were also tested. Associations between GRS and age were assessed as both a continuous variable and categorical [<20; 21–39; >40 years].

2.3.2. Microarray analysis

Probes representing the *CFB* genes were obtained from dbSNP at NCBI [https://www.ncbi.nlm.nih.gov/snp]. To investigate the relationship between expression of *CFB* and UC severity we tested the association between *CFB* expression in the sigmoid and clinically diagnosed UC severity subgroups. Microarray gene expression data were read into R [version 3.4.1] using the Affy package version 1.56.0.[39] Probes were pre-processed using the expresso function in which data were background corrected using the rma method, quantile normalized and summarized using the median polish method. Data were filtered according to probe variance [cut-off: 0.5] and presence in all samples. Generalized linear regression was applied to identify a relationship between *CFB* expression and UC severity. The *p*-values were adjusted using the false discovery rate [FDR]. The probe 202357s was used as a proxy for CFB expression. One-way analysis of variance with a Tukey's post-hoc comparison between groups was applied to identify differences in the *CFB* probe between healthy controls and UC severity subgroups.

3. Results

3.1. Population

A total of 1222 participants were recruited for this study, including patients with non-MRUC [n = 287] and MRUC [n = 352: n = 311, colectomy and n = 41, no colectomy], as well as a matched healthy cohort [n = 583] [Table 1]. Control participants had a significantly higher prevalence of smoking compared to both the non-MRUC and MRUC subgroups [44.4% vs 26.9% and 27.1%, respectively, $p < 0.001$]. Patients with MRUC were diagnosed younger than patients with non-MRUC [32.8 vs 35.6 years, $p < 0.01$] and had a shorter disease duration [11.5 vs 20.4 years, $p < 0.001$]. As expected, there were significant associations between disease extent and disease severity. Specifically, limited disease [E1 or E2] was reported in 190 [66.7%] of non-MRUC patients compared to 90 [26%] of those with MRUC, while extensive disease [E3] was present in 257 [74.1%] of those with MRUC disease and only 95 [33.3%] of the non-MRUC subgroup [$p < 0.001$]. In contrast to previous studies,[6] a family history of IBD was reported equally across both UC subgroups.

3.2. Identified SNPs

A GWAS using the combined UC dataset identified 1460 SNPs on chromosome 6 in the HLA region [lead SNP = rs28479879, OR = 1.97, $p = 1.63 \times 10^{-14}$] that were significantly associated with UC, reaching a conventional genome-wide significance threshold of $p < 5 \times 10^{-8}$ [Supplementary Figure 1a]. Conditioning on the lead SNP in this region identified a secondary independent risk locus in the HLA region [lead SNP = rs144717024, OR = 5.52, $p = 1.57 \times 10^{-10}$]. When considering only patients with MRUC, 2018 SNPs were significantly associated, including a locus on chromosome 1 [lead SNP = rs111838972, OR = 1.82, $p = 6.28 \times 10^{-9}$] near *MMEL1* and a locus in the HLA region on chromosome 6 [lead SNP = rs144717024, OR = 12.23, $p = 1.7 \times 10^{-19}$] [Supplementary Figure 1b]. Conditioning on the lead SNP in each of these regions identified a secondary independent risk locus in the HLA region [lead SNP = rs6916742, OR = 2.18, $p = 1.41 \times 10^{-10}$]. The risk loci also passed a more stringent Bonferroni correction [$p < 7.97 \times 10^{-9}$] accounting for the total number of SNPs tested. Similarly, rs144717024, located in the HLA region, was also the lead SNP associated with acute severe [$p = 1.92 \times 10^{-17}$, OR = 13.42] and chronic refractory [$p = 7.07 \times 10^{-13}$, OR = 9.90] UC sub-phenotypes. Upon conditioning on rs144717024, a second signal in the HLA region remained significantly associated with ASUC [rs9269233, $p = 2.38 \times 10^{-8}$, OR = 2.47]. The large effects observed for variants in the HLA region are consistent with previous reports of large effects of UC-associated haplotypes in this region.[40,41] Effect sizes observed were substantially reduced when considering non-MRUC patients only, resulting in no significant SNPs reaching genome-wide significance when compared to control participants (lead SNP, rs7680780; OR 2.70 [1.89–3.85], $p = 5.56 \times 10^{-8}$). However, the direction of these effects was consistent with MRUC. The OR for the lead SNPs, rs28479879 [lead SNP in combined], rs144717024 [lead SNP in combined and MRUC] and rs111838972 [lead SNP in MRUC], were significantly higher in MRUC compared to non-MRUC [$p < 9.0 \times 10^{-6}$].

When comparing our results to the 123 previously identified SNPs associated with UC we were able to replicate eight SNPs in our dataset [Tables 2 and 3]. Of the 123 previously identified SNPs tested, 55% [n = 68] had larger effects in MRUC cases compared to non-MRUC cases [Supplementary Table 1]. We do, however, note large standard errors for OR estimates in this study due to the relatively small sample size. Overall, a large proportion of SNP effects were in the same direction as those reported previously [88% combined UC, 82% non-MRUC, 85% MRUC]. One SNP, rs7554511, on chromosome 1 was only associated with non-MRUC cases and not MRUC, or combined UC cases. rs4151651 had a statistically higher OR in MRUC compared to non-MRUC [$p = 1.08 \times 10^{-31}$]. Similarly, the ORs for three SNPs estimated in the combined UC cohort and MRUC cohort were significantly different from the published estimates [rs4151651 $p_{combined} = 8.06 \times 10^{-16}$, $p_{MRUC} = 2.56 \times 10^{-55}$; rs6667605 $p_{combined} = 2.31 \times 10^{-4}$, $p_{MRUC} = 8.70 \times 10^{-10}$; rs10761648

Table 2. Eight published SNPs associated with ulcerative colitis [UC] replicated in association analyses for combined UC cases, non-MRUC cases and MRUC cases

Previously identified SNPs							Combined cases		Non-MRUC cases	MRUC cases		
SNP	CHR	BP	Effect allele	OR	*p* value	Reference	OR [95% CI]	*p* value	OR [95% CI]	*p* value	OR [95% CI]	*p* value
rs6667605[c]	1	2502780	C	1.09	3.16E-10	19	1.39 [1.19–1.64]	5.13E-05[b]	1.14 [0.94-1.39]	1.97E-01	1.72 [1.41–2.13]	1.00E-07[b]
rs80174646	1	67708155	G	1.61	4.34E-62	19	2.04 [1.43–2.94]	1.01E-04[a]	2.17 [1.33-3.57]	1.74E-03	1.92 [1.23–3.03]	4.57E-03
rs7554511	1	200877562	C	1.18	6.83E-31	19	1.39 [1.16–1.67]	4.28E-04	1.61 [1.273-2.04]	9.56E-05[a]	1.21 [0.96–1.51]	1.02E-01
rs6556412	5	158787385	A	1.1	4.31E-12	19	1.27 [1.09–1.50]	4.66E-03	1.21 [0.98-1.50]	7.29E-02	1.35 [1.10–1.65]	4.00E-03
rs4151651[c,d]	6	31915614	A	1.72	6.05E-12	17	3.73 [2.48–5.63]	2.72E-10[b]	1.81 [1.07-3.06]	2.79E-02	6.00 [3.86–9.35]	2.03E-15[b]
rs9268853	6	32429643	T	1.41	1.35E-55	21	1.92 [1.59–2.33]	4.11E-12**	1.54 [1.23–1.92]	1.57E-04*	2.44 [1.89–3.13]	9.10E-13**
rs10761648[c]	10	64354262	T	1.16	2.99E-15	24	1.53 [1.24–1.89]	7.86E-05**	1.46 [1.13–1.89]	3.82E-03	1.57 [1.22–2.02]	5.03E-04
rs2836878	21	40465534	G	1.25	7.35E-53	19	1.43 [1.18–1.7]	2.92E-04*	1.47 [1.15–1.89]	1.81E-03	1.35 [1.06–1.72]	1.18E-02

SNP, single nucleotide polymorphism; CHR, chromosome; BP, base pair; MRUC, medically refractory UC; OR, odds ratio; CI, confidence interval.
[a]Significant association accounting for multiple testing in individual association analysis.
[b]Significant association accounting for multiple testing across all three analyses.
[c]Published OR estimates are significantly different from combined and MRUC GWAS estimates.
[d]OR estimates are significantly different between non-MRUC and MRUC GWAS.

Table 3. Eight published SNPs associated with ulcerative colitis [UC] replicated in association analyses for MRUC cases and MRUC sub-phenotypes

Previously identified SNPs							MRUC cases		Acute severe UC		Chronic refractory UC	
SNP	CHR	BP	Effect allele	OR	*p* value	Reference	OR [95% CI]	*p* value	OR [95% CI]	*p*.value	OR [95% CI]	*p* value
rs6667605[c]	1	2502780	C	1.09	3.16E-10	19	1.72 [1.41–2.13]	1.00E-07[b]	1.81 [1.39–2.35]	9.83E-06[b]	1.65 [1.28–2.14]	1.42E-04*
rs80174646	1	67708155	G	1.61	4.34E-62	19	1.92 [1.23–3.03]	4.57E-03	2.54 [1.27–5.09]	8.71E-03	1.55 [0.89–2.69]	1.23E-01
rs7554511	1	200877562	C	1.18	6.83E-31	19	1.21 [0.96–1.51]	1.02E-01	1.06 [0.80–1.41]	6.88E-01	1.34 [1–1.80]	5.15E-02
rs6556412[c,d]	5	158787385	A	1.1	4.31E-12	19	1.35 [1.10–1.65]	4.00E-03	1.62 [1.25–2.10]	3.01E-04[a]	1.13 [0.87–1.48]	3.52E-01
rs4151651[c,d]	6	31915614	A	1.72	6.05E-12	17	6 [3.86–9.35]	2.03E-15[b]	7.41 [4.48–12.24]	5.40E-15[b]	4.45 [2.58–7.68]	8.55E-08[b]
rs9268853	6	32429643	T	1.41	1.35E-55	21	2.44 [1.89–3.13]	9.10E-13[b]	2.93 [2.09–4.11]	4.05E-10[b]	2 [1.47–2.71]	8.45E-06[b]
rs10761648[c]	10	64354262	T	1.16	2.99E-15	24	1.57 [1.22–2.02]	5.03E-04	1.47 [1.08–2.02]	1.61E-02	1.61 [1.16–2.22]	4.32E-03
rs2836878	21	40465534	G	1.25	7.35E-53	19	1.35 [1.06–1.72]	1.18E-02	1.2 [0.89–1.62]	2.29E-01	1.47 [1.08–2.01]	1.40E-02

SNP, single nucleotide polymorphism; CHR, chromosome; BP, base pair; MRUC, medically refractory UC; OR, odds ratio; CI, confidence interval.
[a]Significant association accounting for multiple testing in individual association analysis.
[b]Significant association accounting for multiple testing across all three analyses.
[c]Published OR estimates are significantly different from MRUC or ASUC GWAS estimates.
[d]OR estimates are significantly different between acute severe and chronic refractory UC GWAS.

$p_{combined} = 8.57 \times 10^{-4}$, $p_{MRUC} = 1.99 \times 10^{-3}$] [Table 2; Figure 1]. In all three cases the published ORs were most similar to the non-MRUC OR estimates. In addition to differences between MRUC and non-MRUC, ORs for rs6556412 [$p = 0.01$] and rs4151651 [$p = 9.93 \times 10^{-15}$] were also significantly different between the acute severe and chronic refractory UC sub-phenotypes [Table 3].

3.3. HLA imputation analysis

A total of 4108 variants across the MHC region were significantly associated with UC using the combined dataset; the most significant variant was a SNP located in the intron of *HLA-DRB1* [DRB1_9616_32547904_intron5_Ax, OR = 0.48, $p = 1.19 \times 10^{-14}$] [Supplementary Figure 2a]. Following conditional analysis on the lead variant only one other signal was significant, led by an indel in the intron of *HLA-DRB1* [INDEL_SNPS_DRB1_6839x6840_32550680, OR = 3.84, $p = 1.42 \times 10^{-10}$] [Supplementary Figure 2b]. Seven UC-associated classical HLA alleles reached genome-wide significance, the strongest being HLA-DQA1*03 [OR = 0.47, $p = 4.83 \times 10^{-11}$] [Supplementary Table 2] although this was preceded by 1190 SNPs and amino acid variants. Similarly, 5317 variants were significantly associated with MRUC, the most significant variant was an indel located in the intron of *HLA-DRB1* [DRB1_5035x5036_32552484, OR = 10.71, $p = 2.87 \times 10^{-18}$] [Supplementary Figure 2c]. This variant remained the lead variant in both MRUC sub-phenotypes, acute severe [$p = 4.11 \times 10^{-17}$, OR = 12.34] and chronic refractory [$p = 7.42 \times 10^{-11}$, OR = 7.85]. Following conditional analysis, a second signal associated with MRUC was identified led by a SNP located between *HLA-DRB1* and *HLA-DQA1* [rs28383313, $p_{MRUC} = 2.06 \times 10^{-12}$, $OR_{MRUC} = 2.60$; $p_{acute} = 4.87 \times 10^{-10}$, $OR_{acute} = 3.34$] [Supplementary Figure 2d]. The strongest MRUC-associated HLA allele was HLA-DQA1*01 [OR = 2.14, $p = 1.83 \times 10^{-12}$] [Supplementary Table 2]; however, as in the combined analysis, this was preceded by several [$n = 1871$] SNPs and amino acid variants. No variants in this region reached genome-wide significance in the non-MRUC association analysis. Using the omnibus test we determined that position 6291 in *HLA-DRB1* was more significant than any single SNP or classical HLA allele in both the combined [$p_{omnibus} = 9.54 \times 10^{-25}$] and MRUC analysis [$p_{omnibus} = 3.81 \times 10^{-17}$]. No single position from the omnibus test exceeded the association of the intron variant [DRB1_5035x5036_32552484] for both the acute severe and chronic refractory sub-phenotypes.

3.4. Genetic risk score

Genome-wide risk scores were significantly increased in all UC subgroups, non-MRUC [$p = 9.60 \times 10^{-13}$], MRUC [$p = 8.03 \times 10^{-16}$], acute severe [$p = 5.94 \times 10^{-11}$] and chronic refractory [$p = 3.10 \times 10^{-9}$], compared to controls, but no difference between non-MRUC, MRUC, acute severe and chronic refractory was observed [Figure 2]. Considering all UC patients as a single group vs controls, the genome-wide risk score was also significantly higher [$p < 2.2 \times 10^{-16}$]. When separated into deciles [Figure 3], the proportion of control participants reduced from 79.8% [decile 1] to 28.7% [decile 10] as the genetic risk score increased. Conversely, the proportion of MRUC patients increased from 9.7% [decile 1] to 34.4% [decile 10] with increasing risk score. Similarly, the proportion of patients with non-MRUC increased from 10.5% [decile 1] to 32.8% [decile 10]. OR calculations between the lowest and highest deciles showed that an increased proportion of participants in the highest decile had UC [either non-MRUC or MRUC] compared to the lowest decile [OR = 9.18, 95% CI = 5.12–16.47, $Z = 7.3$, $p = 1 \times 10^{-4}$]. There was a significant positive association between UC GRS and disease extent [$p = 4.91 \times 10^{-3}$] and a significant difference [$p = 0.023$] in disease extent between cases in the top and bottom deciles. Age at diagnosis was not significantly associated with the GRS when assessed as either a continuous or a categorical variable.

Using the AVENGEME R package[42] we estimate that a training set of ~22 000 individuals would be required to achieve a clinically relevant AUC of 0.75 using 100 000 SNPs if the genetic variance explained is 33% [SNP heritability] and the proportion of SNPs having no effect on disease is 0.90 [Supplementary Table 3].

3.5. *CFB* gene expression

Regression analysis indicated an increase in *CFB* expression in sigmoid colon mucosa in the UC patients [$p = 0.002$, FDR = 0.037]. The expression of *CFB* was significantly different between the control group and mild UC and between the control group and moderate UC [Figure 4, Tukey's test, $p < 0.0001$]. In contrast, *CFB* expression in UC non-inflamed sigmoid was similar to healthy controls [Tukey's test, $p = 0.25$].

4. Discussion

GWAS using large international cohorts have identified over 200 SNPs linked to IBD that explain ~8.2% of the variance in UC risk.[19–24] These studies have been invaluable in identifying SNPs that explain disease susceptibility and hence provide important insights into disease pathogenesis. However, these SNPs do not differentiate between patients who experience particularly aggressive MRUC as opposed to those with persistent, documented, non-MRUC. Without the granularity of data to separate these sub-phenotypes, genetic influences reported in the literature to date may provide only part of the unique genetic signatures carried by each form of UC. In this study we assess two distinctly different groups of patients with UC, namely those who follow a severe course which typically requires surgery within a median of 6.4 years from diagnosis and those who have been diagnosed and followed up for at least 10 years with limited medical interventions required to control disease activity and no requirement for surgery. Previous studies indicate that these two extremes of UC phenotype account for between 25 and 40% of all UC cases.[2,3,9–11]

Our study finds that the effect sizes of known UC risk variants differ between patients with MRUC and non-MRUC. Notably, only one SNP was identified, rs7554511, which was related to non-MRUC but not MRUC in our dataset. Effect sizes reported in this study are on average 7% larger than in the published literature. This effect was even more pronounced when considering only patients with MRUC [10%]. Sub-phenotype analysis of those with MRUC demonstrated larger effect sizes in five of eight replicated SNPs for those with ASUC as compared with chronic refractory UC [Figure 1 and Table 3]. Even our non-MRUC subgroup had an effect size comparable with published effect sizes, suggesting international meta-analyses may use a mixture of patients with severity typically on the milder side of the disease spectrum.

This may relate to the recruitment process for genetic studies with many patients identified from outpatient clinics and population-based registries. The observations for non-MRUC are supported by those of Kopylov and colleagues.[43] In a North American IBD Consortium analysis of 156 index SNPs from known IBD loci in their "benign' UC cohort, none achieved the pre-defined significance threshold.

For MRUC, of note is rs4151651, a SNP in an exonic region of complement factor B [*CFB*]. This SNP had a much larger OR [6.00] in patients with MRUC compared to non-MRUC [1.81]. Further sub-phenotype analysis demonstrated an even higher OR [7.41] for those patients who had required colectomy for ASUC as compared to those undergoing surgery for chronic refractory UC [OR 4.45] [Figure 1]. CFB is a secreted protein in the alternative complement pathway and is mainly expressed by mononuclear phagocytes. The complement system plays important roles in pathogen recognition and clearance,[44] and both inflammatory and immune responses. It has also been implicated in a range of autoinflammatory disorders including IBD.[45] Recent multi-ethnic studies in IBD genetics have identified *CFB* as one of two novel UC susceptibility genes in the North Indian population, with *CFB* allelic heterogeneity demonstrated when comparing North Indian, Japanese and Dutch populations.[25,46] The driver SNP, rs537160, in the UC-associated Dutch haplotype was also replicated in this study in the combined [$p = 2.48 \times 10^{-5}$] and MRUC [$p = 2.07 \times 10^{-9}$] GWAS, and was a predicted transcription factor binding site for POLR2A and TFAP2A.[46] The over-representation of the rs4151651 and rs537160 risk alleles in patients with MRUC may be associated with abnormal CFB secretion, impaired pathogen clearance within the colonic mucosa, and/or an exaggerated and poorly controlled immune response. Our gene expression data support a potential role for *CFB* in the mucosal inflammatory response typical of UC with a stepwise increase in expression across a spectrum of disease activity from remission through to MRUC. These observations replicate and extend previous *CFB* gene expression analysis in the context of UC.[45] The study by Ostviks and colleagues identified the colonic epithelium as the major local source of this increased *CFB* expression in active UC. Functional analysis of a SNP [rs12614] in *CFB* demonstrated significantly reduced alternative complement pathway activity in UC sera from individuals homozygous or heterozygous for this variant as compared to homozygous wild-type.[25] Whilst rs12614 is not in LD with rs4151651 or rs537160, it suggests a possible role for genetic regulation of *CFB* in UC. Studies in animal models of IBD have identified potential pathogenic and protective roles for different complement pathway components in disease aetiology. Specifically, an

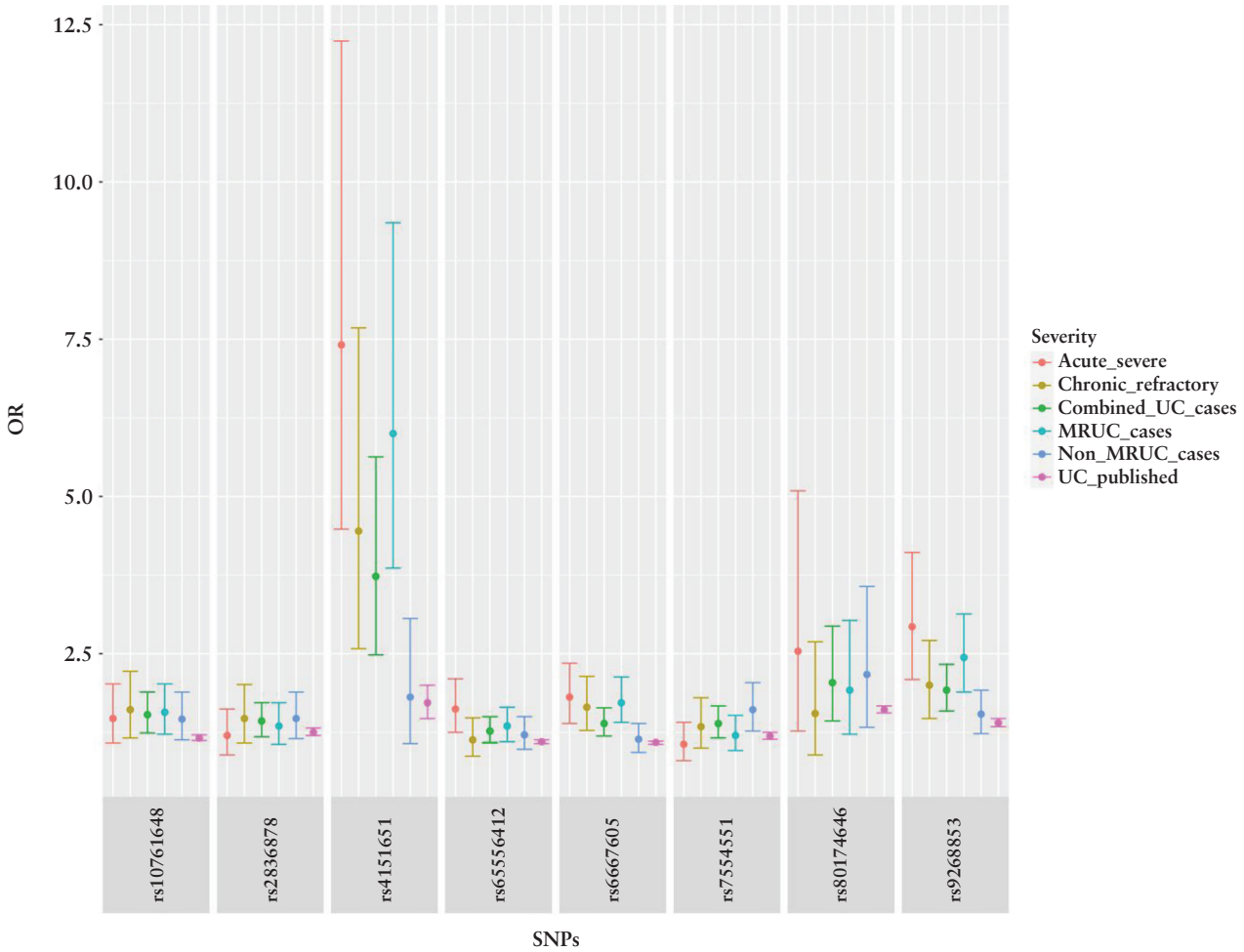


Figure 1. Odds ratios with 95% confidence intervals for eight published SNPs associated with ulcerative colitis [UC] and replicated in association analyses for combined UC cases, non-medically refractory UC [non-MRUC] only, MRUC only, acute severe UC only and chronic refractory UC only.

alternative pathway knockout ameliorated the early effects of a dextran sodium sulphate-induced colitis,[47] and subsequent work demonstrated therapeutic potential for CR2-fH, a targeted inhibitor of the alternative pathway.[48] There has also been interest in the development of agents that can block complement pathway components such as C5a or its receptor. The far stronger association with MRUC, in particular ASUC, in this study supports genetic heterogeneity within UC and the need to further explore the genetic regulation of complement in mucosal immune responses and how this is influenced by local environmental factors such as the intestinal microbiome.

In our study, people in the highest decile of the genetic risk score are nine times more likely to have UC compared to those in the lowest decile of genetic risk. We also found a significant association between disease extent and the genome-wide GRS developed on all UC types calculated in this study. However, the GRS was unable to separate non-MRUC from MRUC in our cohort. This limitation to the GRS based upon currently available data probably reflects the milder disease course of many UC participants in GWAS to date and the clinical data available to define extreme phenotypes. There may be a lack of access to patients who have undergone surgery for MRUC given that their follow up is often with the surgical service at their local hospital, and that they remain a minority within the total recruited UC population. As such, independent larger and well-defined subgroups would be required to further develop robust indicators of disease course.

To date, two publications have explored genetic nuances between patients with non-MRUC and MRUC.[6,18] Descriptors in these studies have included 'medically-refractory versus non-medically refractory' or 'poor prognosis versus good prognosis'. Our study used a stricter definition for those responsive to limited medical therapy; specifically, all patients in this subgroup had not undergone colectomy within 10 years of diagnosis, had not experienced an episode of severe colitis requiring hospital admission and/or intravenous corticosteroids, nor required immunosuppression therapy for greater than 6 months. These extremes of phenotype criteria are similar to those used by Lee and colleagues in their analysis of a Korean UC cohort, and probably result in more distinct non-MRUC and MRUC subgroups.[18] This study of UC identified one SNP that was associated with the MRUC subgroup and which reached genome-wide significance. This SNP, rs9268877, located in the HLA region, was not associated with overall UC disease susceptibility. The lead SNP in our combined analysis, rs28479879, is in LD with rs9268877 [$r^2 = 0.70$] and was the strongest association signal for UC reported in a transethnic analysis by Degenhardt *et al.*[49] In our cohort, rs9268877 reached genome-wide significance in both the combined [OR = 1.87, CI: 1.58–2.23; $p = 7.21 \times 10^{-13}$] and MRUC analysis [OR = 2.22, CI: 1.79–2.76; $p = 3.02 \times 10^{-13}$]. The effect size was greater in MRUC compared to non-MRUC [OR = 1.57, CI: 1.27–1.93; $p = 2.18 \times 10^{-5}$].

Our HLA imputation analysis supported and underscored the differences between our MRUC and non-MRUC subgroups. Specifically, over 5000 HLA variants achieved genome-wide significance in those with MRUC while none was found for the non-MRUC subgroup. Otherwise, association of variants and classical HLA alleles within *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* with combined UC

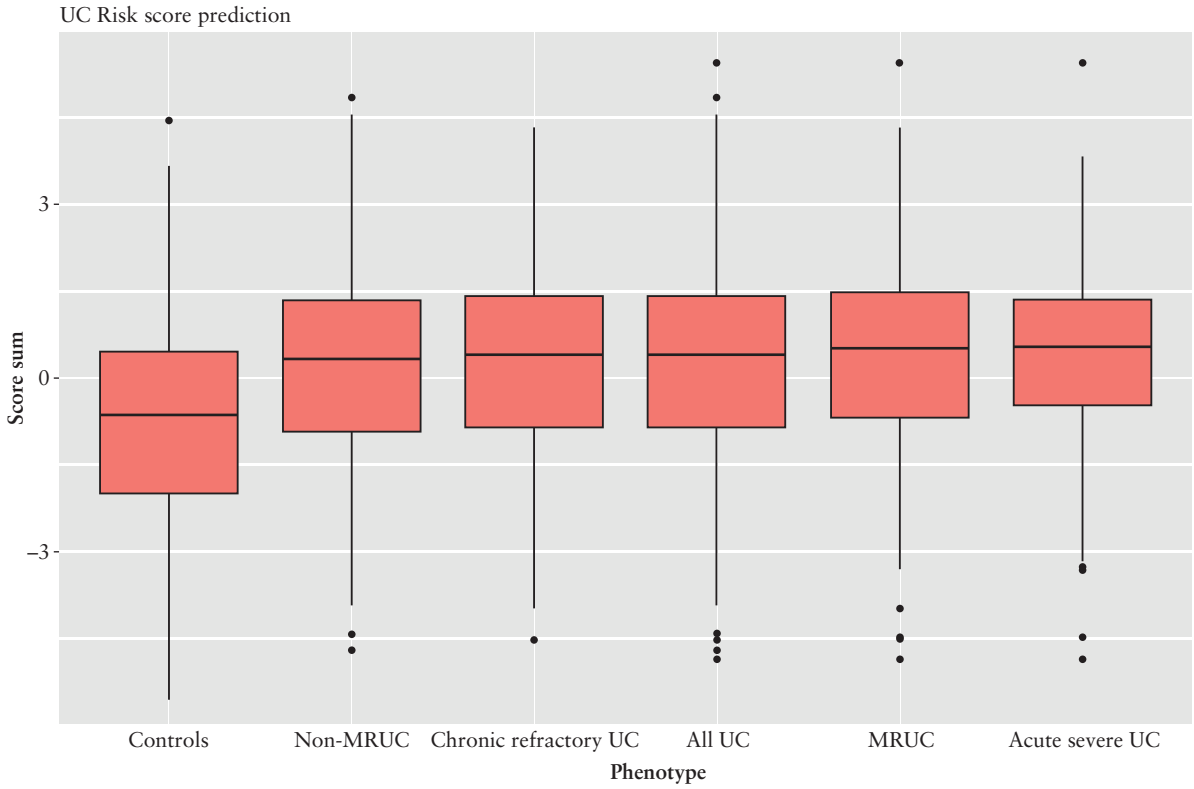


Figure 2. Distribution of ulcerative colitis [UC] genetic risk scores for controls, all UC patients [All UC], non-MRUC patients only, MRUC patients only, chronic refractory UC patients only and acute severe UC patients only. Sub-phenotypes are ordered from lowest mean risk score [left] to highest mean risk score [right].

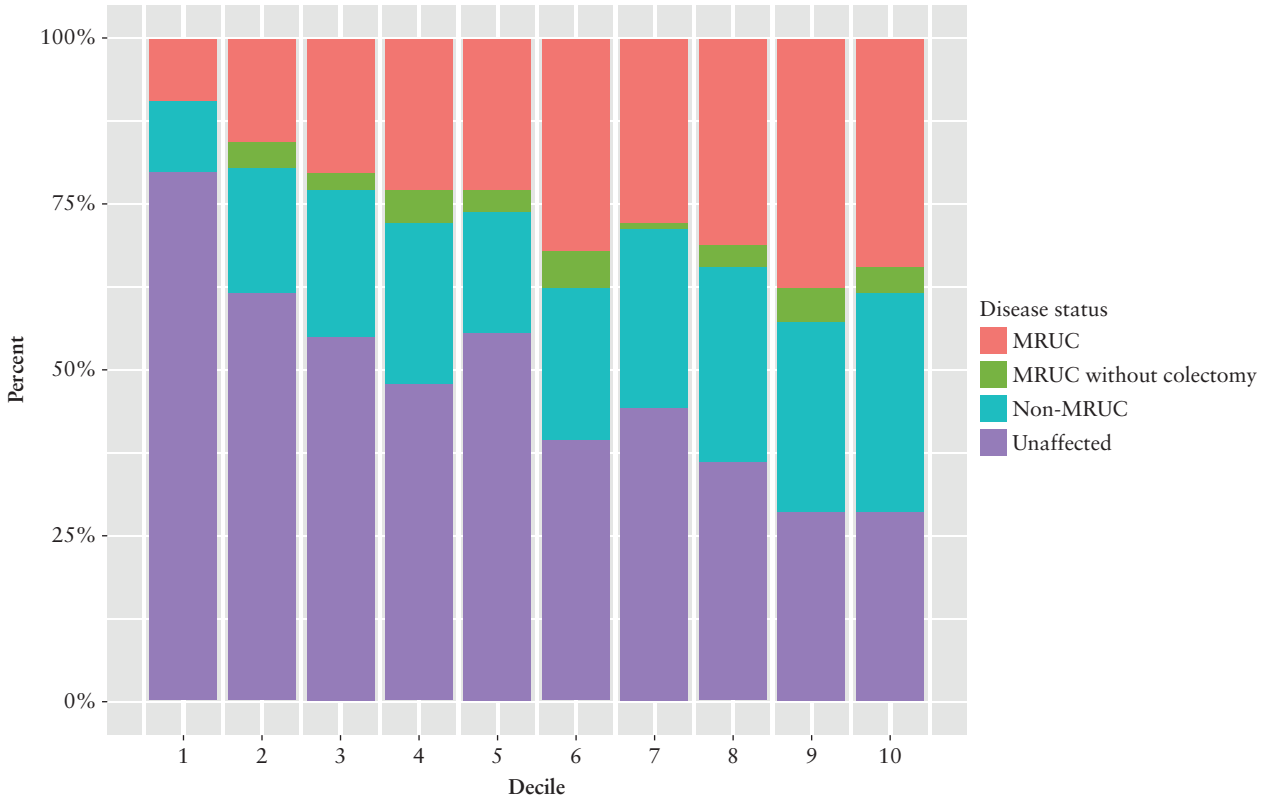


Figure 3. Patients divided into deciles according to ulcerative colitis [UC] genetic risk score and the proportion of patients with medically refractory UC [MRUC] with colectomy [red], MRUC without colectomy [green], non-MRUC [blue] and unaffected [purple].

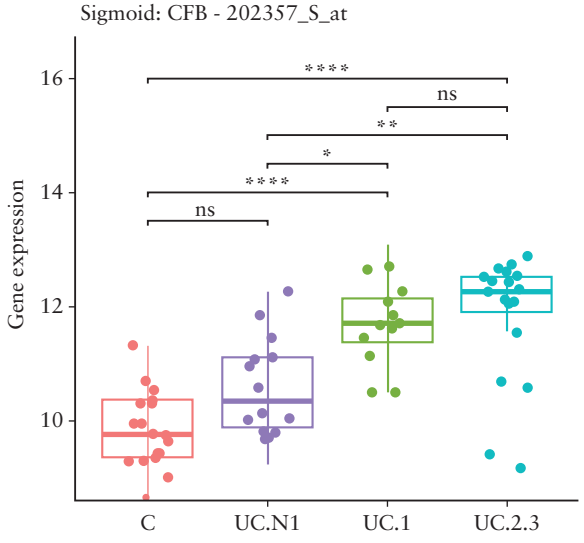


Figure 4. Microarray gene expression levels for *CFB* using probe 202357_s_at, for controls [C], non-inflamed ulcerative colitis [UC.NI], mild UC [UC.1] and moderate to severe UC [UC.2.3].

and MRUC, and the larger effects on MRUC, were consistent with previous studies.[49,50] The associations identified are evidence of an effect of these genes in UC, but given the strong LD across this region, and correlation between signals, identification of potential causal alleles would require larger studies in ancestrally diverse populations to utilize LD differences[49] and subsequent functional studies. With respect to results for HLA-DRB1*0103 and those reported by previous studies,[16,49,50] this allele was imputed as part of our additional analyses but not included in our association analysis as it was filtered out with a frequency of <5%. Looking at previous studies in the context of sub-phenotypes, Roussomoustakaki and colleagues[16] included only UC patients who had undergone proctocolectomy for either MRUC [n = 97] or colonic neoplasia [n = 10]. They identified strong associations for all of extensive disease, and/or need for surgery, and/or presence of specific extra-intestinal manifestations [EIMs], compared with their controls. It was not clear whether the association with need for surgery was independent of EIMs. The study did not include a non-MRUC subgroup. Goyette and colleagues[50] performed high-density SNP mapping of the MHC in a very large population of IBD patients and controls. Sub-phenotype analysis of disease severity, specifically well-defined MRUC and non-MRUC, was not undertaken and hence the relative contributions of these sub-phenotypes to the top signal [HLA-DRB1*0103] was not included in their analyses. These points also apply to the more recent multi-ethnic analysis by Degenhardt and colleagues.[49] In the large genotype–phenotype study published by Cleynen *et al.*,[17] the top association with colectomy in UC was rs4151651 as alluded to above, while Haritunians and colleagues identified rs17207986 as the top HLA association in their MRUC–control analysis.[6] Neither study discussed HLA-DRB1*0103 in the context of their colectomy or MRUC subgroups, respectively.

Our findings have important clinical implications. The clinical criteria used to define non-MRUC and MRUC have been successful in identifying significant genetic differences between these two extremes of phenotype in a modest sample size. This supports the concept of related but distinct polygenic disorders and hence the potential to uncover novel treatment targets for MRUC. The findings

also question whether these extreme phenotypes share a common pathogenesis, which has implications for approaches to treatment.

Further studies are needed to develop a clinically useful risk score that combines clinical and genetic variables at diagnosis that can stratify patients by disease course. Successful stratification can provide more informed treatment decisions at diagnosis and hence more personalized care. Those identified as having the potential for severe disease may require early referral to a specialized IBD service while those stratified as mild may be successfully managed by lifestyle modification guided by their general practitioner.

The strengths of our study include the *a priori* prescriptive case definitions for non-MRUC and MRUC, the recruitment of population controls from the same population, the detailed clinical metadata ascertained for all cases and the extensive analysis undertaken within subgroups of MRUC. Our clinical and genetic findings are predominantly consistent with previous published data while highlighting the genetic heterogeneity within the sub-phenotype of UC. Limitations relate to statistical power across the study and within subgroups.

5. Conclusion

MRUC and non-MRUC show distinct genetic signatures characterized by differences in effect sizes of risk variants. Genetic heterogeneity between sub-phenotypes can make the development of a diagnostic genetic risk score difficult. While the direction of effects is relatively consistent, the influence of genetics on non-MRUC is noticeably reduced with no statistically significant hits at the genome-wide significance level in our dataset. Combining non-MRUC and MRUC patients into a single cohort for GWAS increases genetic heterogeneity, probably reducing the ability of the GRS to distinguish between clinically relevant sub-phenotypes. We identified *CFB* as an important candidate for UC susceptibility within a Caucasian population and highlighted its potential role in influencing UC disease activity. Future studies should consider the severity of disease when trying to elucidate genetic nuances of UC.

Funding

This work was supported by a grant from the National Health and Medical Research Council of Australia [APP1028569] awarded to Graham Radford-Smith [Chief Investigator] and colleagues. G.R.S. is supported by the QIMR Berghofer MRI and the Royal Brisbane and Women's Hospital Research Foundation.

Conflict of Interest

The authors have no conflicts of interest to declare.

Acknowledgments

We wish to acknowledge all the participants who took part in the study and the clinicians, clinical nurses, administrative staff and research nurses who assisted in the study. The graphical abstract was created with BioRender.com

Author Contributions

GRS and GM obtained financial support and designed the study. GRS coordinated participant recruitment, clinical data collection, assisted with first draft, any additional support, and wrote all subsequent manuscript drafts. GM supervised genotyping and analysis approach. SM was lead analyst, assisted by AL, JD and MZ. LS coordinated all sample processing with sites and supervised laboratory work at the lead site. KK and KH assisted in collection of all samples and data/data verification from participating sites. KK and KH completed all ethics requirements. AW, IL, PAB, JMA, GM, SC, MS, SB, TF, JB and RG coordinated participant recruitment at each site including suitability, consent, data and sample acquisition.

Data Availability

The data used and analysed for this study are available from the corresponding author upon reasonable request. Consent from participants and ethics approval for uploading of data to a publicly available database was not obtained at the time of this study.

Supplementary Data

Supplementary data are available at *ECCO-JCC* online.

References

1. Alatab S, Sepanlou SG, Ikuta K, *et al*. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol* 2020;**5**:17–30.
2. Quezada SM, Cross RK. Association of age at diagnosis and ulcerative colitis phenotype. *Dig Dis Sci* 2012;**57**:2402–7.
3. Solberg IC, Høivik ML, Cvancarova M, Moum B. Risk matrix model for prediction of colectomy in a population-based study of ulcerative colitis patients (the IBSEN study). *Scand J Gastroenterol* 2015;**50**:1456–62.
4. Reinisch W, Reinink AR, Higgins PDR. Factors associated with poor outcomes in adults with newly diagnosed ulcerative colitis. *Clin Gastroenterol Hepatol* 2015;**13**:635–42.
5. Lee H-S, Cleynen I. Molecular profiling of inflammatory bowel disease: is it ready for use in clinical decision-making? *Cells* 2019;**8**:535.
6. Haritunians T, Taylor KD, Targan SR, *et al*. Genetic predictors of medically refractory ulcerative colitis. *Inflamm Bowel Dis* 2010;**16**:1830–40.
7. Rubin DT, Ananthakrishnan AN, Siegel CA, Sauer BG, Long MD. ACG clinical guideline: ulcerative colitis in adults. *Am J Gastroenterol* 2019;**114**:384–413.
8. Chicco F, Magrì S, Cingolani A, *et al*. Multidimensional impact of Mediterranean diet on IBD patients. *Inflamm Bowel Dis* 2020;**27**:1–9.
9. Harbord M, Eliakim R, Bettenworth D, *et al*. Third European evidence-based consensus on diagnosis and management of ulcerative colitis. part 2: current management. *J Crohns Colitis* 2017;**11**:769–84.
10. Magro F, Gionchetti P, Eliakim R, *et al*. Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part 1: definitions, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and ileo-anal pouch disorders. *J Crohns Colitis* 2017;**11**:649–70.
11. Dinesen LC, Walsh AJ, Protic MN, *et al*. The pattern and outcome of acute severe colitis. *J Crohns Colitis* 2010;**4**:431–7.

12. Cesarini M, Collins GS, Rönnblom A, *et al*. Predicting the individual risk of acute severe colitis at diagnosis. *J Crohns Colitis* 2016;**11**:335–41.
13. Gordon H, Trier Moller F, Andersen V, Harbord M. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. *Inflamm Bowel Dis* 2015;**21**:1428–34.
14. Chen G-B, Lee SH, Brion M-JA, *et al*. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum Mol Genet* 2014;**23**:4710–20.
15. De La Concha EG, Fernandez–Arquero M, Lopez–Nava G, *et al*. Susceptibility to severe ulcerative colitis is associated with polymorphism in the central MHC gene IKBL. *Gastroenterology* 2000;**119**:1491–5.
16. Roussomoustakaki M, Satsangi J, Welsh K, *et al*. Genetic markers may predict disease behavior in patients with ulcerative colitis. *Gastroenterology* 1997;**112**:1845–53.
17. Cleynen I, Boucher G, Jostins L, *et al*. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* 2016;**387**:156–67.
18. Lee H-S, Yang S-K, Hong M, *et al*. An intergenic variant rs9268877 between HLA-DRA and HLA-DRB contributes to the clinical course and long-term outcome of ulcerative colitis. *J Crohns Colitis* 2018;**12**:1113–21.
19. Liu JZ, van Sommeren S, Huang H, *et al*. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 2015;**47**:979–86.
20. McGovern DPB, Gardet A, Törkvist L, *et al*. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat Genet* 2010;**42**:332–7.
21. Anderson CA, Boucher G, Lees CW, *et al*. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* 2011;**43**:246–52.
22. de Lange KM, Moutsianas L, Lee JC, *et al*. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* 2017;**49**:256–61.
23. Jostins L, Ripke S, Weersma RK, *et al*. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;**491**:119–24.
24. Ellinghaus D, Jostins L, Spain SL, *et al*. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet* 2016;**48**:510–8.
25. Juyal G, Negi S, Sood A, *et al*. Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. *Gut* 2015;**64**:571–9.
26. Duncan EL, Danoy P, Kemp JP, *et al*. Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. *PLoS Genet* 2011;**7**:e1001372.
27. Amanat S, Requena T, Lopez-Escamez JA. A systematic review of extreme phenotype strategies to search for rare variants in genetic studies of complex disorders. *Genes* 2020;**11**:987.
28. Lennard-Jones JE. Classification of inflammatory bowel disease. *Scand J Gastroenterol* 1989;**24**:2–6.
29. Silverberg MS, Satsangi J, Ahmad T, *et al*. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can J Gastroenterol Hepatol* 2005;**19**:5A–36A.
30. Truelove SC. Cortisone in ulcerative colitis. Final report on a therapeutic trial. *Br Med J* 1955;**2**:104–8.
31. Geboes K, Riddell R, Öst A, Jensfelt B, Persson T, Löfberg R. A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. *Gut* 2000;**47**:404–9.
32. Tye H, Yu C-H, Simms LA, *et al*. NLRP1 restricts butyrate producing commensals to exacerbate inflammatory bowel disease. *Nat Commun* 2018;**9**:3728.
33. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 2015;**4**:7.
34. Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
35. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82.
36. Das S, Forer L, Schönherr S, *et al*. Next-generation genotype imputation service and methods. *Nat Genet* 2016;**48**:1284–7.
37. Luo Y, Kanai M, Raychaudhuri S, *et al*. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat Genet* 2021;**53**:1504–16.
38. Moore CM, Jacobson SA, Fingerlin TE. Calculations for genetic association studies in the presence of genetic model misspecification. *Hum Hered* 2019;**84**:256–71.
39. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004;**20**:307–15.
40. Okada Y, Yamazaki K, Umeno J, *et al*. HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology* 2011;**141**:864–71.e1.
41. Venkateswaran S, Prince J, Cutler DJ, *et al*. Enhanced contribution of HLA in pediatric onset ulcerative colitis. *Inflamm Bowel Dis* 2018;**24**:829–38.
42. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013;**9**:e1003348.
43. Kopylov U, Boucher G, Waterman M, *et al*. Genetic predictors of benign course of ulcerative colitis—a North American Inflammatory Bowel Disease Genetics Consortium Study. *Inflamm Bowel Dis* 2016;**22**:2311–6.
44. Reis ES, Mastellos DC, Hajishengallis G, Lambris JD. New insights into the immune functions of complement. *Nat Rev Immunol* 2019;**19**:503–16.
45. Østvik AE, Granlund A, Gustafsson BI, *et al*. Mucosal toll-like receptor 3-dependent synthesis of complement factor B and systemic complement activation in inflammatory bowel disease. *Inflamm Bowel Dis* 2014;**20**:995–1003.
46. Gupta A, Juyal G, Sood A, *et al*. A cross-ethnic survey of CFB and SLC44A4, Indian ulcerative colitis GWAS hits, underscores their potential role in disease susceptibility. *Eur J Hum Genet* 2017;**25**:111–22.
47. Schepp-Berglind J, Atkinson C, Elvington M, Qiao F, Mannon P, Tomlinson S. Complement-dependent injury and protection in a murine model of acute dextran sulfate sodium-induced colitis. *J Immunol* 2012;**188**:6309–18.
48. Elvington M, Schepp-Berglind J, Tomlinson S. Regulation of the alternative pathway of complement modulates injury and immunity in a chronic model of dextran sulphate sodium-induced colitis. *Clin Exp Immunol* 2015;**179**:500–8.
49. Degenhardt F, Mayr G, Wendorff M, *et al*. Transethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals not only shared but also ethnicity-specific disease associations. *Hum Mol Genet* 2021;**30**:356–69.
50. Goyette P, Boucher G, Mallon D, *et al*. High density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat Genet* 2015;**47**:172–9.