Enhanced Contribution of HLA in Pediatric Onset Ulcerative Colitis

Suresh Venkateswaran, PhD, Jarod Prince, BS, David J Cutler, PhD, Urko M Marigorta, PhD, David T Okou, PhD, Sampath Prahalad, MD, David Mack, MD, Brendan Boyle, MD, Thomas Walters, MD, Anne Griffiths, MD, Cary G Sauer, MD, Neal LeLeiko, MD, PhD, David Keljo, MD, PhD, James Markowitz, MD, Susan S Baker, MD, PhD, Joel Rosh, MD, Marian Pfefferkorn, MD, Melvin B Heyman, MD, Ashish Patel, MD, Anthony Otley, MD, Robert Baldassano, MD, Joshua Noe, MD, Paul Rufo, MD, Maria Oliva-Hemker, MD, Sonia Davis, PhD, Michael E Zwick, PhD, Greg Gibson, PhD, Lee A Denson, MD, Jeffrey Hyams, MD, Subra Kugathasan, MD

Enhanced Contribution of HLA in Pediatric Onset Ulcerative Colitis

Suresh Venkateswaran, PhD,[,#] Jarod Prince, BS,[*,#] David J. Cutler, PhD,[†] Urko M. Marigorta, PhD,[‡] David T. Okou, PhD,[*] Sampath Prahalad, MD,[*] David Mack, MD,[§] Brendan Boyle, MD,[¶] Thomas Walters, MD,[‖] Anne Griffiths, MD,[‖] Cary G. Sauer, MD,[*] Neal LeLeiko, MD, PhD,[**] David Keljo, MD, PhD,[††] James Markowitz, MD,[‡‡] Susan S. Baker, MD, PhD,[§§] Joel Rosh, MD,[¶¶] Marian Pfefferkorn, MD,[‖‖] Melvin B. Heyman, MD,[***] Ashish Patel, MD,[†††] Anthony Otley, MD,[‡‡‡] Robert Baldassano, MD,[§§§] Joshua Noe, MD,[¶¶¶] Paul Rufo, MD,[‖‖‖] Maria Oliva-Hemker, MD,[****] Sonia Davis, PhD,[††††] Michael E Zwick, PhD,[†] Greg Gibson, PhD,[‡] Lee A Denson, MD,[‡‡‡‡] Jeffrey Hyams, MD,[§§§§] and Subra Kugathasan, MD[*]*

Background: The genetic contributions to pediatric onset ulcerative colitis (UC), characterized by severe disease and extensive colonic involvement, are largely unknown. In adult onset UC, Genome Wide Association Study (GWAS) has identified numerous loci, most of which have a modest susceptibility risk (OR 0.84–1.14), with the exception of the human leukocyte antigen (HLA) region on Chromosome 6 (OR 3.59).

Method: To study the genetic contribution to exclusive pediatric onset UC, a GWAS was performed on 466 cases with 2099 healthy controls using UK Biobank array. SNP2HLA was used to impute classical HLA alleles and their corresponding amino acids, and the results are compared with adult onset UC.

Results: HLA explained the almost entire association signal, dominated with 191 single nucleotide polymorphisms (SNPs) ($p = 5 \text{ x } 10^{-8}$ to $5 \text{ x } 10^{-10}$). Although very small effects, established SNPs in adult onset UC loci had similar direction and magnitude in pediatric onset UC. SNP2HLA imputation identified HLA-DRB1*0103 (odds ratio [OR] = 6.941, $p = 1.92{*}10^{-13}$) as the most significant association for pediatric UC compared with adult onset UC (OR = 3.59). Further conditioning showed independent effects for HLA-DRB1*1301 (OR = 2.25, $p = 7.92{*}10^{-9}$) and another SNP rs17188113 (OR = 0.48, $p = 7.56{*}10^{-9}$). Two HLA-DRB1 causal alleles are shared with adult onset UC, while at least 2 signals are unique to pediatric UC. Subsequent stratified analyses indicated that HLA-DRB1*0103 has stronger association for extensive disease (E4: OR = 8.28, $p = 4.66\text{x}10^{-10}$) and female gender (OR = 8.85, $p = 4.82\text{x}10^{-13}$).

Conclusion: In pediatric onset UC, the HLA explains almost the entire genetic associations. In addition, the HLA association is approximately twice as strong in pediatric UC compared with adults, due to a combination of novel and shared effects. We speculate the paramount importance of antigenic stimulation either by infectious or noninfectious stimuli as a causal event in pediatric UC onset.

Key Words: GWAS, HLA-DRB1, IBD, Inflammatory Bowel Disease, Pediatric UC, Ulcerative Colitis

Received for publications 7 November 2017; Editorial Decision 14 November 2017

From the *Department of Pediatrics, Emory University School of Medicine and Children's Healthcare of Atlanta, GA; [†]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA; [‡]Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA; [§]Department of Pediatrics, Children's Hospital of Eastern Ontario IBD Centre and University of Ottawa, Ontario, Canada; [¶]Department of Gastroenterology, Nationwide Children's Hospital Columbus, OH; [‖]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, The Hospital for Sick Children, University of Toronto, Toronto, Canada; **Division of Pediatric Gastroenterology, Nutrition, and Liver Diseases, Hasbro Children's Hospital, Providence, RI; [††]Gastroenterology, Hepatology and Nutrition Department, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA; [‡‡]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Cohen Children's Medical Center of NY, New Hyde Park, NY; [§§]Department of Digestive Diseases and Nutrition Center, University at Buffalo, Buffalo, NY; [¶¶]Department of Pediatrics, Goryeb Children's Hospital, Morristown, NJ; [‖‖]Bronson Pediatric Gastroenterology, Bronson Children's Hospital, Kalamazoo, MI; ***Department of Pediatrics, University of California at San Francisco, San Francisco, CA; [†††]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX; [‡‡‡]Division of Pediatric Gastroenterology and Nutrition, Department of Pediatrics, IWK Health Centre, Dalhousie University, Halifax, Nova Scotia, Canada; §§§Department of Pediatrics, University of Pennsylvania, Philadelphia, PA; [¶¶¶]Department of Pediatric Gastroenterology, Hepatology, and Nutrition, Medical College of Wisconsin, Milwaukee, WI; [‖‖‖]Division of Gastroenterology and Nutrition, Boston Children's Hospital, Harvard Medical School, Boston, MA; ****Department of Pediatrics, John Hopkins University School of Medicine, Baltimore, MD; [††††]Collaborative Studies Coordinating Center Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC; [‡‡‡‡]Division of Pediatric Gastroenterology, Hepatology, and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; [§§§§]Division of Digestive Diseases, Hepatology, and Nutrition, Connecticut Children's Medical Center, Hartford, CT; #Equal first authorship.

Conflict of Interest: The authors have no financial relationships relevant to this article to disclose. The authors have no conflicts of interest to disclose.

Supported by NIH/NIDDK U01 DK095745, NIH/ NIDDK R01 DK087694, and CCFA 0000023837.

Address correspondence to: Subra Kugathasan, MD, Division of Pediatric Gastroenterology, Emory University School of Medicine, 1760 Haygood Drive, W427, Atlanta, GA 30322. E-mail: skugath@emory.edu Tel: 404 727 4542; Fax: 404 727 4069

doi: 10.1093/ibd/izx084
Published online 19 March 2018

INTRODUCTION

Ulcerative Colitis (UC) is 1 of the 2 main forms of chronic inflammatory bowel disease (IBD) with a wide age range or disease onset which varies from infancy to late adulthood.[1, 2] The clinical, endoscopic, and histologic features are generally similar across the age spectrum. Numerous genome-wide association studies (GWAS) and subsequent meta-analyses in IBD have confirmed the heritability of UC by uncovering many genetic susceptibility loci.[3, 4] To date, more than 200 loci are implicated in IBD, with 29 loci being UC specific.[5] While GWAS has uncovered new and unsuspected pathways for pathogenesis, each locus only has very small to modest effect size,[5–7] with the exception of a strong association for human leukocyte antigen (HLA) loci on chromosome 6[7] that consistently shows a large effect on UC susceptibility. Despite advances in understanding the etiology of UC, 2 major questions remain unanswered: do genetic factors contribute differentially to pediatric versus adult onset, and why do some patients have limited distal disease while others have extensive disease or pancolitis? It is observed that pediatric onset UC is characterized by extensive colonic involvement (80% pancolitis in children compared to only 30–40%[8, 9] in adult onset) and younger age of onset.[10] This observation suggests that genetic effects may be higher in children and, hence, that dissection of UC genetic effects in pediatric cases may identify additional susceptibility genes and further address the unanswered questions in UC. However, 2 previous GWA studies[11, 12] in pediatric IBD where UC was included as a sub-phenotype neither identified new loci exclusive to younger onset nor shed any new information on the HLA region pertaining to pediatric onset, most likely due to lack of power.

The HLA complex contains highly polymorphic human leukocyte antigen genes and various other immune-regulatory genes.[13] HLA polymorphisms were the focus of attention in several modestly sized IBD studies involving mainly adult onset UC. These studies have indicated that multiple independent associations are likely to exist in HLA and non-HLA regions, with consistent associations observed at class-II loci, mainly HLA-DRB1 and HLADQB1[6, 7, 14]. A recent study involving HLA proteins in seropositive rheumatoid arthritis, another inflammatory disorder, was able to narrow down most of the association from 3 HLA proteins to 5 amino acids.[15] Therefore, to further define genetic associations across the HLA region and identify functional and potentially causal single nucleotide polymorphisms (SNPs) in UC, we report here the performance of a GWAS in a well-phenotyped pediatric UC cohort using a fine mapping chip with excellent coverage across the extended HLA region. We recruited an exclusively pediatric onset cohort of 768 de novo cases and used the Affymetrix UK Biobank Array, a high-density custom array designed for fine mapping. Our aim was to determine how the allele architecture of pediatric UC compares to that of adult onset UC particularly focusing on the HLA region. We also tested to see if any of the same causal variant(s) are driving the association between pediatric and adult onset UC using the Bayesian *coloc* package.[16] We report novel findings pertaining to HLA signals in pediatric UC, including a doubling of HLA effects in pediatric compared to adult onset and larger effect in female gender. Furthermore, the detected signals in HLA region can be explained by 2 classical HLA-DRB1 alleles (shared with adult onset UC) and 2 other independent signals (may be unique to the pediatric onset UC).

METHODS

Patient Population

A total of 768 children between the ages 4 and 17 (mean age 12.7, female 49%) with the diagnosis of UC were recruited into the study. Standard criteria were used to establish a diagnosis of UC.[17–19] All the patients were newly diagnosed children and adolescents with ulcerative colitis treated with consensus defined treatment regimens of mesalamine and corticosteroids (CS) after established uniform pretherapy criteria. Twenty-nine North American centers participated to recruit the study patients. Comprehensive demographic, clinical, laboratory, and serologic values were obtained along with whole genome genotypes and diagnostic ileocolonoscopy and esophagogastroduodenoscopy. Disease extent was based on visual appearance of colonic mucosa, not biopsy, and classified as proctosigmoiditis, left-sided disease (to the splenic flexure), extensive disease (to the hepatic flexure), and pancolitis (beyond the hepatic flexure). Biopsy specimens were only used for patients who met all study criteria at the time of diagnosis.

Genotyping

Whole blood from 768 newly diagnosed pediatric UC patients was stored at -80ºC until DNA was extracted using standard protocols on a Thermo Scientific King Fisher Flex system at Akesogen Inc (Norcross, GA). DNA was quantified via spectrophotometer and suspended at a concentration of 50ng/uL. Genome-wide genotyping of the DNA was done using the Affymetrix UK Biobank Axiom Array at Akesogen Inc. The UK biobank array is a high-density custom designed array with 7,348 SNPs across the HLA region (~24,000 high-quality imputed SNPs) and over 830,115 established SNPs across the genome (~4 million high-quality imputed SNPs). The UK Biobank Axiom Array covers a wide range of rare coding SNPs, pharmacogenomics markers, copy number regions, HLA, inflammation, and expression quantitative trait loci (eQTL) SNPs. It was recently used in a study to identify independent associations for prognosis relative to susceptibility in Crohn's disease.[20] In our study, the genotypes of ~27,000 disease-free ethnicity-matched healthy controls were obtained from the UK Biobank consortium.

Quality Control and Statistical Analysis

Raw genotype image files (.CEL) from cases and controls were analyzed. For each pediatric UC case, 5 closely related

UK Biobank controls (total of 3840 controls) were selected by matching their genetic distribution in principal component analysis (PCA). Cases and controls were processed as 1 batch, and genotypes were called using Axiom Analysis Suite provided by Affymetrix (Affymetrix, Inc., Santa Clara, CA). Further, the samples were tested for data completeness, and those with SNP call rates <95% were excluded. Samples that remained were tested for agreement between X/Y genotypes and reported sex and for unexpected relatedness between individuals by applying RELPAIR and GRR14[21] to ~20,000 markers in linkage equilibrium (pairwise r^2<0.1 evenly distributed across the genome). Samples with discordant gender were excluded from analyses. For pairs of individuals that appeared to be genetically related, 1 of the pair was removed from subsequent analyses. One member from all first- and second-degree relative pairs (r >= 0.25) was dropped. Overall, approximately 5% of the samples were excluded for these reasons.

Additionally, a filtering criteria was applied to 776,463 autosomal SNPs to exclude those with genotyping call rate <0.95 and Hardy-Weinberg equilibrium $P < 1 \times 10^{-7}$. Approximately 5% of the SNPs were dropped during this process.

After excluding the unqualified samples and SNPs in the preimputation process, our final dataset included 734 cases, 3651 controls, and 746,272 autosomal SNPs. To investigate population structure and identify population outliers, we used principal components analysis (PCA). Starting with the high-quality filtered SNPs, a subset of ~20,000 SNPs with moderate minor allele frequency and no linkage disequilibrium (LD) ($r^2 < 0.1$) were selected. The principal components (PCs) were inferred for all samples of each phase of the cohort using PCA analysis in plink.[22] Samples were plotted in PC space, and outliers that exceeded cut-off thresholds were dropped (Supplementary Fig. 1). After removing the outliers, a dataset of only European-ancestry samples (466 cases and 2099 controls) was selected. The observed genomic control value 1.07 indicates little evidence of population stratification after controlling for global ancestries. This suggests that there was little or no inflation leading to false positives from confounding by ancestry. All the statistical QC processing and analysis were done using plink1.09.[23]

Imputation-based Association Analysis, Meta-analysis

IMPUTE2,[24] a genotype imputation method, was applied to impute the European ancestry dataset to the 1000 Genomes Project Phase 3 integrated autosomal reference panel. The imputed SNPs with minor allele frequency (MAF) <0.05 and imputation quality scores <0.95 were excluded. Therefore, this pediatric UC GWAS analysis was narrowed down to only common SNPs. The imputation yielded ~4 million high-quality SNPs across the entire genome after excluding INDELs and copy number variants (CNVs). Genotype imputation clouds (ie, the full genotype probability values, not single point estimates) from IMPUTE2 were directly used to assess association using SNPTEST 2.5.2[25] under an additive model (-frequentist 1, -method score parameter options). An association p-value of $<5 \times 10^{-8}$ (corresponding to a genome-wide significance level of 0.05 after a Bonferroni correction for multiple testing of 1million SNPs) was considered statistically significant.

Imputing HLA Genotypes

SNP2HLA[26] was used to perform imputation of classical HLA alleles and their corresponding amino acids from SNP genotyping data. The Type 1 Diabetes Genetics Consortium (T1DGC) dataset of 5225 unrelated individuals of European ancestry is the major reference panel for SNP2HLA, which contains genotypes for 8961 HLA SNPs and amino acids covering the entire HLA region including the classical alleles HLA- A, B, C, DRB1, DQA1, DQB1, DPA1, and DPB1 at two- and four-digit resolution. The accuracy of SNP2HLA program for the T1DGC reference panel has been very well documented using a different dataset.[27] Our final GWAS dataset (before IMPUTE2 imputation) was tested against the T1DGC, and only the SNPs in the HLA region were imputed. Association was tested across the HLA by applying a logistic regression framework using plink1.09.[23]

Conditional Analysis

Given that the classical HLA-DRB1 allele had the most significant association signal in pediatric UC, we examined whether there were any other independent effects by applying the same logistic regression stated in the previous section to test the remaining markers across the HLA region conditioned on the HLA-DRB1 genotype. If any SNPs were identified as an independent association, we used their covariates to condition in subsequent analyses. Subsequent conditional analyses were performed until no marker with significant independent association at $P < 5 \times 10^{-8}$ emerged.

Colocalization of Signals in Pediatric and Adult UC

We performed colocalization to identify if the causal variants driving the association in 2 different traits (here pediatric onset and adult onset UC are considered different traits) are the same or different, using the *coloc* package.[16] Published literature[7] from adult onset UC GWAS has identified HLA-DRB1*0103 and rs6927022 as the 2 most important alleles driving the signals across the HLA class II region. Since neither rs6927022 nor any proxy SNPs with $r^2 > 0.5$ were present in the imputed SNP dataset, that site could not be evaluated for sharing. *Coloc* uses a Bayesian framework to generate posterior probabilities for 5 mutually exclusive hypotheses regarding the sharing of causal variants between 2 traits, namely H0 (no causal variant for either trait); H1 or H2 (a causal variant only for 1 or the other trait); H3 (there are 2 distinct causal variants, 1 for each trait); and H4 (a single causal variant common to both traits). We directly used the SNP-level association results and performed *coloc* analyses at each of the 6742 variants across the HLA, which are common in both traits, using 50kb windows around each target SNP. The analysis was repeated after conditioning on the pediatric associations, but because only summary p-values were available for the adult cases, the reciprocal conditional analysis could not be performed.

RESULTS

We first performed a genome-wide association study to identify any novel associations and to replicate loci reported from previous UC-centered GWAS. Additionally, the extended HLA region was imputed against the T1DGC reference panel to find HLA SNPs, classical HLA alleles, and their corresponding amino acids associated with pediatric UC. The analysis workflow is described in Fig. 1.

The Manhattan plot (Fig. 2A) shows the association signals for 3,996,700 SNPs that passed pre- and postimputation QC procedures. The p-value cut-off of $P < 5 \text{ x } 10^{-8}$ was used to identify any genome-wide significant associations. Our study identified 191 SNPs as genome-wide significant (Supplementary Table 1). The locus plot across the HLA region (Supplementary Fig. 2) exhibits not only 191 genome-wide significant SNPs, but hundreds of other associated SNPs all in linkage disequilibrium (LD) with the strong associations. The Q-Q plots with and without the HLA SNPs show significant departure from the null hypothesis driven by HLA (Fig. 2B), which is almost eliminated when the HLA region is removed (Fig. 2C).

Replication of Existing Results

We compared our GWAS results against the recently published meta-analyses in IBD.[5] After a stringent quality control process, our study produced genotypes for 183 of the 232 SNPs reported as known to be associated with CD, IBD, and/or UC in Europeans.[5] Of the 73 known UC significant loci[5] ($P < 5 \text{ x } 10^{-8}$ in adult UC) tested, 24 were replicated at $P < 0.05$ in our pediatric onset UC cohort (Supplementary Table 2). The estimated effects of 73 common risk SNPs discovered in primarily adult onset UC cohorts had very similar magnitudes and directions in our pediatric UC onset cohort (Supplementary Fig. 3) with just 8 inconsistent signs for small effect loci. Similarly, we replicated 9 out of 11 SNPs found in the adult onset UC GWAS dataset.[27] This suggests that the allelic contributions to adult and pediatric onset UC are very similar, as far as common SNPs with minor allele frequency (MAF) >5% are concerned.

Search for Associations Across the HLA Region

Because 191 SNPs in the HLA region exhibited multi-test corrected genome-wide significance (Fig. 3), the HLA region was subjected to a specific detailed analysis. The SNPs across the HLA region which passed QC were used as the backbone to impute against the T1DGC reference panel using the SNP2HLA package. This technique collapsed many of the highly redundant SNPs into haplotypes and classical HLA alleles. The final HLA dataset after imputation contained a total of the 8115 HLA alleles, SNPs, and amino acid variants which were tested for association, including 79 of 126 classical HLA alleles at two-digit resolution, 102 of 298 classical HLA alleles at four-digit resolution, and 991 of 1276 polymorphic amino acids. This analysis wrapped our initial 191 highly associated SNPs into 16 independent SNPs and HLA alleles with genome-wide significance. By far the most highly associated allele was HLA-DRB1*0103, with an odds ratio (OR) of 6.94 (95% confidence interval [CI] 6.42–7.47) and $p = 1.92*10^{-13}$. Notably, HLA-DRB1*0103 was also reported in Goyette et al.,[7] an adult onset UC study, as the most significant allele with an OR of 3.59 (CI 3.22–4.00), which is about half of the effect size we observed in our pediatric onset UC cohort. We compared our pediatric UC OR with adult UC OR for 56 classical HLA alleles, which are shown as significant with $P < 5 \text{ x } 10^{-8}$ in adult UC (Fig. 4), finding no more than 6 loci with possible differences in effect sizes and a few instances of inverted sign.

Conditioning Effect on HLA Amino Acid Variants

We also noticed the effect of HLA-DRB1*0103 conditional analysis on the associations observed at specific amino acids. Of all the amino acids tested in HLA-DRB1*0103, there are 2 strong amino acid associations mapped to position at 71. These amino acids are Glu71 (OR = 1.969 and $p = 1.24 \times 10^{-10}$) and Lys71 or Arg71 (OR = 1.724 and $p = 1.83 \times 10^{-9)}$ (Supplementary Fig. 4A). Conditioning on HLA-DRB1*0103 completely suppressed the 2 strongest amino acid associations, and none of the additional associations were shown as significant (Supplementary Fig. 4B).

Search for Additional HLA Signals

Because the association to HLA-DRB1*0103 is extremely strong (OR of 6.94), highly replicated in adults, and resides in a region with extensive LD, we next asked whether there were any additional independent signals in the HLA region. To test this, we repeated the logistic regression, treating HLA-DRB1*0103 as a co-variate and, thus, removing its effect. After this procedure, 31 SNPs remained genome-wide significant. Among the 31 SNPs, HLA-DRB1*1301 had the strongest OR of 2.25 (CI 1.97–2.5) and $p = 7.92*10^{-9}$, which is also significant in the univariate test, and thus we selected it for a second round of conditioning. Now using both HLA-DRB1*0103 and HLA-DRB1*1301 as covariates, 23 SNPs remained as genome-wide significant, with SNP rs17188113 having OR = 0.48 and $p = 7.56*10^{-9}$. The third round of conditioning included HLA-DRB1*0103, HLA-DRB1*1301, and rs17188113 as covariates, and no additional SNPs further emerged as genome-wide significant. From this series of conditioning, our

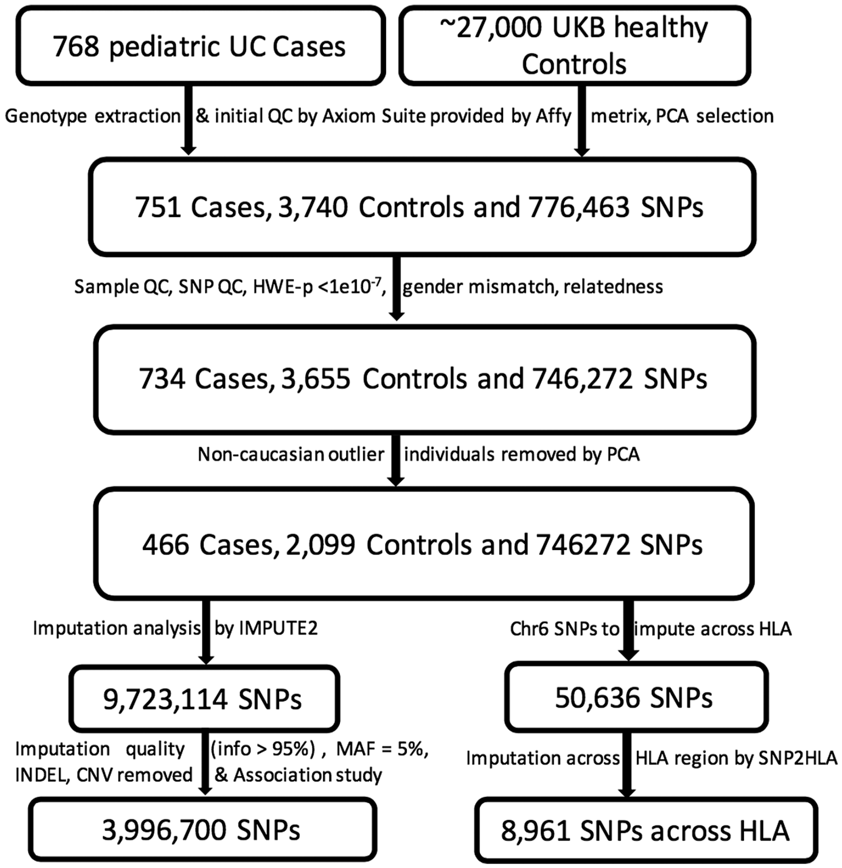
768 pediatric UC Cases
~27,000 UKB healthy Controls
Genotype extraction & initial QC by Axiom Suite provided by Affymetrix, PCA selection
751 Cases, 3,740 Controls and 776,463 SNPs
Sample QC, SNP QC, HWE-p <1e10-7, gender mismatch, relatedness
734 Cases, 3,655 Controls and 746,272 SNPs
Non-caucasian outlier individuals removed by PCA
466 Cases, 2,099 Controls and 746272 SNPs
Imputation analysis by IMPUTE2
Chr6 SNPs to impute across HLA
9,723,114 SNPs
50,636 SNPs
Imputation quality (info > 95%) , MAF = 5%, INDEL, CNV removed & Association study
Imputation across HLA region by SNP2HLA
3,996,700 SNPs
8,961 SNPs across HLA

FIGURE 1. Experiment design flowchart for the entire analysis. An independent GWAS was performed on 466 pediatric UC cases/ 2099 UK Biobank controls (GWAS genotyped on UK Biobank array). After QC and imputation based association analysis with IMPUTE2, the SNPs across the HLA region was imputed using SNP2HLA and tested for any HLA association for pediatric UC.

analysis concludes that HLA-DRB1*0103, HLA-DRB1*1301, and rs17188113 appear to capture all the independent genome-wide significant signals in the HLA region of our pediatric onset UC cohort. The univariate and conditional effects of all other SNPs across the HLA region for our entire pediatric UC dataset are given in Figs. 5A-5B and Supplementary Table 3.

HLA Signals Between Pediatric Versus Adult Onset UC

We next performed colocalization analysis,[16] testing the hypothesis that HLA associations in our pediatric dataset directly correspond to the same alleles as observed in the adult UC GWAS study.[7] The Bayesian algorithm in *coloc* generates posterior probabilities for the hypotheses that the associations at a locus are different in the 2 datasets (H3) or due to the same SNP or credible interval (H4). These are plotted in Fig. 5C, which indicates that there appear to be 4 associations, 2 each (1 shared and 1 separate) in the Class 1 (HLA- A, B and C) and class II (HLA- DP, DM, DOA, DOB, DQ, and DR) intervals. The HLA-DR locus at MB 32.3 to 32.7 shows very high posterior probabilities for a pediatric-specific haplotype, which is centered on rs17495592 in the vicinity of HLA-DRB5 (H3: blue SNPs), and a shared haplotype, which corresponds to HLA-DRB1 (H4: yellow SNPs). Conditioned on HLA-DRB1*0103, the H4 signal dominates the locus (Fig. 5D) and is likely due to HLA-DRB1*1301, although some residual secondary H4 signal is retained when we also further condition on HLA-DRB1*1301 (Fig. 5E). Alternatively, the complex LD structure at the locus may confound the analysis leading to false posterior probability estimates: there is clearly LD between rs17495592 and HLA-DRB1 and independently conditioning on rs17495592 reduces both the association signal and the posterior probabilities at the HLA-DRB1 locus (data not shown). We were unable to perform reciprocal conditioning on the adult HLA associations since only summary p-values are available, and note that rs6927022, identified as the second peak in the adult HLA,[7] was not present in our imputed genotypes.

Similarly, at the HLA class I locus (position at 30.5-30.6 MB), an independent strong nonshared signal (H3) was seen due to rs17188113. This signal is unambiguously driven by pediatric onset UC because rs17188113 was seen with significant association in our dataset but not in the adult UC. In addition, there

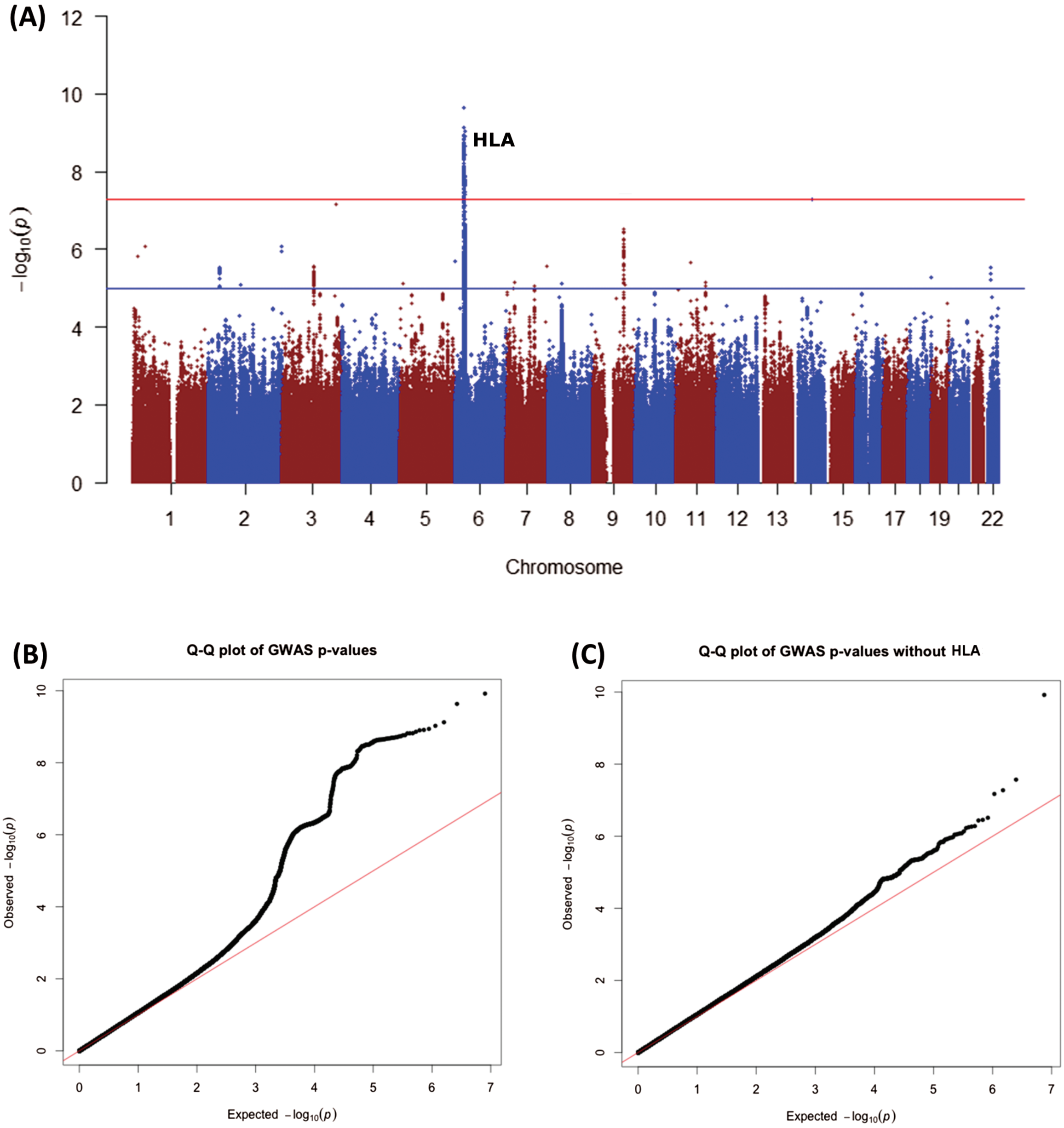

FIGURE 2. (A) Manhattan plot of SNP association p-values result for Pediatric UC cases vs UKB controls for 3,996,700 SNPs qualified after pre- and postimputation QC procedures. All SNPs are plotted according to their position on each chromosome on *x*-axis, against their association on *y*-axis. The red and blue lines indicate the genome-wide significance ($p \leq 5x10^{-8}$) and the suggestive significance threshold ($p \leq 1x10^{-5}$), respectively. Genome-wide significant signals are labeled with corresponding gene names. The inset QQ plots shows the observed (*y*-axis) against the expected (*x*-axis) distribution of p-values under the null hypothesis with (B) and without HLA region (C).

appears to be some evidence for a shared H4 association further conditioned on this SNP, at MB 30.7 (Fig. 5F), although it fails to drop below $P < 0.005$ in association modeling.

Search for Gender and Sub-phenotype Associations

A recent Juvenile Idiopathic Arthritis (JIA) study reported that the SNP rs2476601 association in the PTPN22 gene is restricted to females and is not observed in males.[28] We, therefore, attempted to generate further evidences for male (239 cases) or female (227 cases) specific/consistent association of classical HLA alleles in our pediatric UC cohort. Somewhat surprisingly, our major allele HLA-DRB1*0103 seems to have a significantly stronger effect in females ($p = 4.82x10^{-13}$ and OR = 8.85 [CI 8.2–9.5]) than in males ($p = 1.53 x 10^{-5}$ and OR 4.84 [CI 4.1–5.5]). Conditioning on HLA-DRB1*0103 in males and females seemed to reverse this pattern, although less strikingly

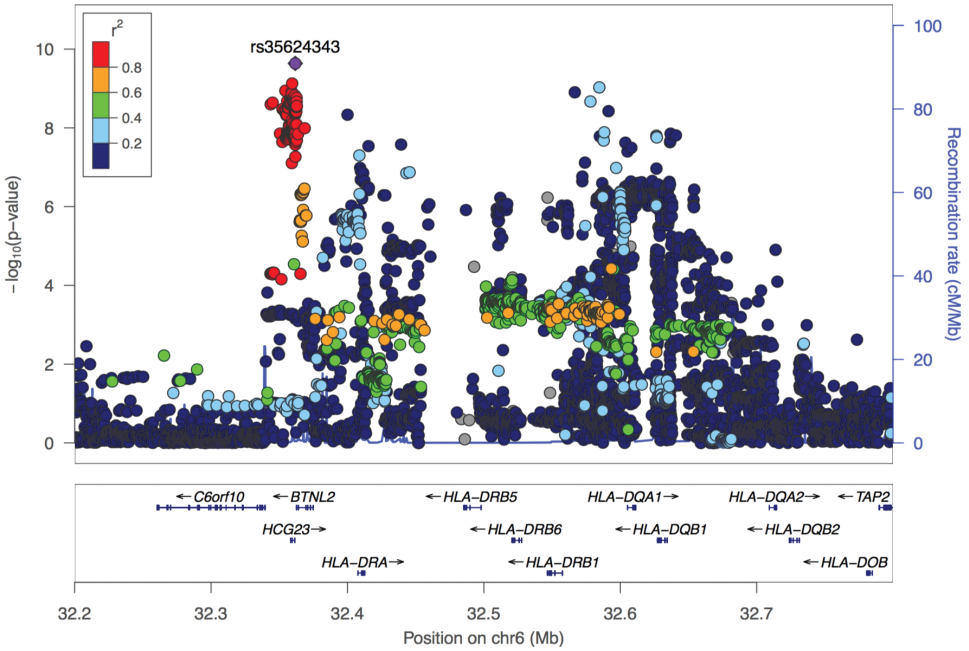

FIGURE 3. LocusZoom plots of SNPs around HLA-II positions (32.2MB to 32.8MB) on chromosome6 against –log10 p-value for their genetic associations from IMPUTE2. The top SNP is highlighted in purple. The surrounding SNPs, shown within 500kb of the top SNP are color coded to reflect their linkage disequilibrium in r² with the top SNP. Estimated recombination rates are plotted in pale blue to reflect local LD structure on secondary *y*-axis.

with males having an OR of 2.70 (CI 2.35–3.04) and females 1.6911 (CI 1.29–3.0) for HLA-DRB1*1301. Lastly, conditioning on both HLA-DRB1*0103 and HLA-DRB1*1301 in males and females showed that ORs of 0.44 (CI 0.34–0.71) and 0.53 (CI 0.19–0.86), respectively for rs17188113. Overall, it appears that HLA-DRB1*0103 has a significantly stronger apparent effect in females, HLA-DRB1*1301 seems to be slightly more associated in males, and rs17188113 has an equal effect in both. The conditioning effect of all other SNPs across the HLA region for males and females are given in Supplementary Tables 4 and 5 and Supplementary Figures 5 and 6, respectively. Finally, we performed an association analysis for sub-phenotypes such as "extensive disease" (E4), because previous reports have indicated the higher effect of HLA on pancolitis phenotype in adult UC. The OR has increased to 8.28 (CI 7.6–9.0) with E4 pediatric UC, and the OR increased to 9.4 (CI 8.6–10.3) when the analysis was restricted to females with E4 (Table 1). The conditioning effect of the remaining SNPs across the HLA region for E4 in pediatric UC and females with E4 are given in Supplementary Tables 6 and 7, respectively. We also tested if

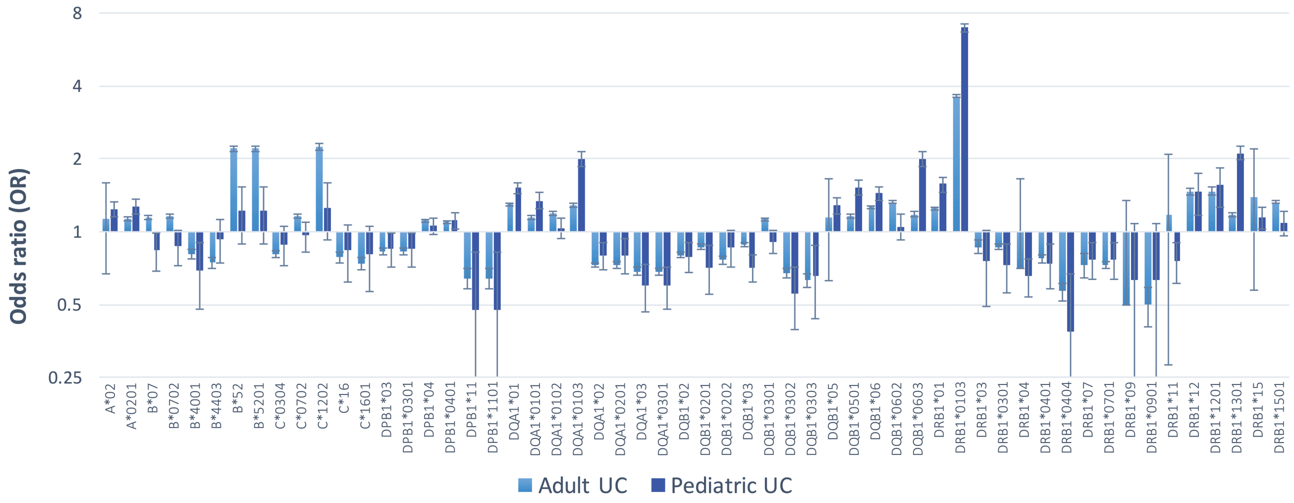

FIGURE 4. Comparison of pediatric and adult UC odds ratios for 56 classical HLA alleles which are considered as significant with $P < 5 \times 10^{-8}$ in adult UC cohort. OR values for pediatric and adult UC are red and blue, respectively.

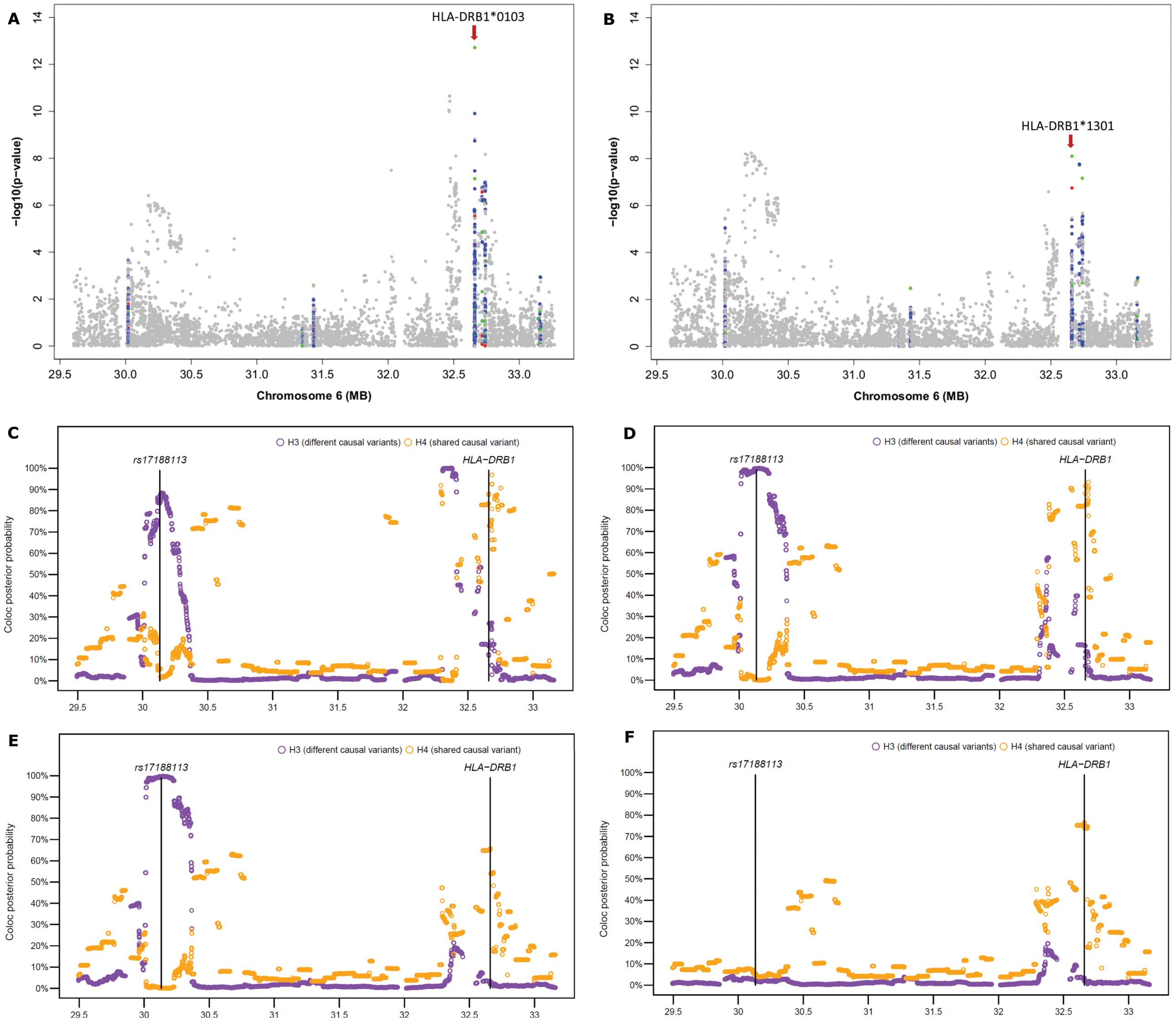

FIGURE 5. Association tests across the HLA region in our pediatric onset UC dataset (466 Cases and 2099 Controls). (A) The major genetic determinants of UC risk map to HLA-DRB1 (OR = 6.94; p = $1.92*10^{-13}$). Un-conditional association analysis results for 79 two-digit (red), 102 four-digit HLA alleles (green), 741 single–amino acid variants (blue) and 5,568 other SNPs across the HLA region. (B) Subsequent conditional analyses on HLA-DRB1*0103 controlling all other high LD alleles revealed an independent association to HLA-DRB1*1301 (OR = 2.25; p = $7.92*10^{-9}$). Results obtained for this dataset is given in Supplementary Table 3. (C-F) Tests of colocalization of causal variants across the HLA region between pediatric and adult onset UC showing posterior probabilities computed by *coloc* at each SNP for H3 (distinct nonshared causal variants, blue circles) and H4 (shared causal variant, yellow points) (C) The *coloc* comparison between pediatric versus adult onset UC on univariate p-values from the pediatric and adult studies. (D) The *coloc* comparison between pediatric onset UC (after first conditioning on HLA-DRB1*0103) with adult onset UC univariate results. (E) The *coloc* comparison between pediatric onset UC (after additionally conditioning on HLA-DRB1*1301) with adult onset UC univariate results. (F) The *coloc* comparison between pediatric onset UC (after thirdly conditioning additionally on rs17188113) with adult onset UC univariate results.

the younger age of onset (dichotomized to under 12 versus over 12 years) among the pediatric UC and did not see increasing association with younger onset.

DISCUSSION

Ulcerative colitis is characterized by chronic mucosal inflammation limited to the colon with histo-pathological features that are strikingly similar, regardless of age of onset.[29] However, pediatric onset UC (onset of less than 17 years of age) differs from adult IBD in that 75–80% of children present with pancolonic distribution, compared with about 30–40% in adult UC.[9, 30] This difference, in part, may explain why pediatric onset UC is also characterized by a high need for corticosteroid therapy, frequent medically refractory disease, and a high rate of treatment failure, which leads to higher risk for colectomy compared to adult UC.[31] Surprisingly, neither disease

TABLE 1: Association of HLA-DRB1*0103 in Different Phenotypes of Our Pediatric UC Cohort

Phenotypes	Odd ratio (95% CI)	p
Pediatric UC - entire dataset	6.94 (6.4–7.5)	1.92E-13
Males only	4.84 (4.1–5.6)	1.53E-05
Females only	8.85 (8.2–9.5)	4.82E-13
Extensive (E4) disease only	8.28 (7.6–9.0)	4.66E-10
Females + E4	9.46 (8.6–10.3)	1.18E-07

distribution nor the age of onset differences in UC have been explained by the advances attributed to GWAS studies, despite the inclusion of pediatric UC as a sub-phenotype.[32] Common susceptibility SNPs in UC show similar allele architectures in pediatric and adult onset UC including the dominance of the HLA region. In this study, we have performed a GWAS narrowed to pediatric UC cases only, with the objective to identify any novel associations, to refine existing associations, and to better explore the role of HLA in early onset UC.

Comparing our results against the recently published meta-analyses in IBD,[5] we were able to replicate 24 of the 73 known UC SNPs at $P < 0.05$. Low power in a sample of 466 cases is the most likely explanation for the lack of replication of the remaining 49 SNPs, since the estimated effects of all the Liu et. al.[5] had very similar magnitudes and directions in our pediatric cohort.

The most intriguing results from our study are the findings pertaining to the HLA region in pediatric onset UC. Not only did the locus plot across the HLA region indicate 191 genome-wide significant SNPs, but it also showed hundreds of more weakly associated SNPs all in LD with the strong associations. The Q-Q plots with and without HLA showed that the genetic signal was almost entirely eliminated when the HLA region was removed, confirming that genetic susceptibility of pediatric onset UC is predominately explained and driven by HLA SNPs. This observation has been observed in numerous studies before.[8, 33] Because the HLA region merited specific detailed analysis to determine if functional and causal SNPs can be identified, we proceeded with imputing our dataset against the T1DGC reference panel using the SNP2HLA package. This process collapsed many of the highly redundant SNPs into haplotypes and classical HLA alleles. The most associated allele in our study was HLA-DRB1*0103, with an odds ratio of nearly 7 (OR = 6.941; p = 1.92×10^{-13}). The same allele was reported in an adult onset UC study as the most significant variant with an odds ratio of only 3.59, which is about half of what we observed in our pediatric onset UC. The *coloc* analysis suggests that the signal at HLA-DRB1*0103 may be due to different main effect alleles in the pediatric and adult samples, possibly with a secondary shared effect in high LD. A shared association at nearby HLA-DRB1*1301 also appears to have a considerably larger effect in the pediatric cases. Conditional analysis also revealed a third pediatric-specific signal centered on rs17188113, and together, these 3 alleles appear to explain most of the signals within the HLA region in our dataset.

Finally, we have performed association analysis for sub-phenotypes such as gender and extensive disease, as much as previous reports indicated the higher effect of HLA on pancolitis phenotype in adult UC. Although pediatric UC is not a gender biased disorder,[34] our finding of higher HLA risk to more of the female gender warrants further investigation. The HLA risk in our study also increased with “extensive disease” (E4) in pediatric UC, and increased further when the analysis was restricted to females with extensive disease. Because conditional analysis used here is based generally on the assumption that single primary association signals allow conditioning on a relatively homogeneous set of cases and controls that exhibit this association, additional layers of complexity can exist. Nevertheless, our data clearly suggests that in contrast to a single locus model, the strong HLA associations with younger onset and gender results from complex, multi-locus effects that span the entire HLA region. Population studies reveal HLA class I and class II gene polymorphisms associated with almost all the common chronic autoimmune diseases, notably spondylarthropathies, rheumatoid arthritis, multiple sclerosis, and type I diabetes. Despite no known reliably measurable auto-antibodies in UC, the stronger association of HLA in early onset UC suggests a role for auto-immunity in pediatric onset UC.

In summary, we performed a genome-wide association and fine mapping study in a narrow phenotype of pediatric UC patients using 2 well characterized cohorts. Although underpowered to detect small effects of disease associations, this is the first such study using exclusively pediatric onset UC cases. The HLA explained the almost entire association signal, dominated with 191 SNPs in high LD. The first being that the HLA-DRB1*0103 contribution is almost twice as high in pediatric UC when compared to adult onset UC. Several other HLA alleles for UC also have a higher OR in children than in adults. The *coloc* results further suggest that primary associations in both the class I and class II regions involve pediatric-specific effects, but that the 2 HLA DRB1 alleles are shared since they associate in both pediatric and adult cohorts. The larger effect size of HLA DRB1*0103 in children may in part be due to some signal from the pediatric-specific peak near the locus, but is mainly intrinsic to the haplotype. Based on these findings we propose that pediatric UC is similar to an extreme form of adult onset UC, with a major contribution from the HLA. We observed that the effect of HLA is larger in females when compared to males in pediatric UC. We are intrigued by this discovery because never has a gender biased genetic risk been seen in UC in either pediatric onset or adult onset UC. The HLA being the dominant association in pediatric UC onset further fueling our speculation that antigenic stimulation either infectious or non-infectious as a precipitating event for such an early onset. The most interesting findings from this study are (1) 2 classical HLA-DRB1 alleles are shared with adult onset UC and, (2) other 2 independent signals are specific to pediatric onset UC. Further studies in pediatric UC with larger sample sizes from more ethnically diverse cohorts combining microbiome studies should be dedicated to better dissect the HLA region, and in turn, those findings may help in individualized medical approaches, pharmacogenomics, and risk stratification in pediatric UC.

SUPPLEMENTARY DATA

Supplementary data are available at *Inflammatory Bowel Diseases* online.

REFERENCES

1. Bequet E, Sarter H, Fumery M, et al. Incidence and Phenotype at Diagnosis of Very-early-onset Compared with Later-onset Paediatric Inflammatory Bowel Disease: A Population-based Study [1988-2011]. *J Crohns Colitis.* 2016;11:519–526.
2. Ordás I, Eckmann L, Talamini M, et al. Ulcerative colitis. *Lancet*. 2012;380:1606–1619.
3. Brant SR, Okou DT, Simpson CL, et al. Genome-wide association study identifies african-specific susceptibility loci in african americans with inflammatory bowel disease. *Gastroenterology*. 2017;152:206–217.e2.
4. 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;22:2311–2316.
5. 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.
6. 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.
7. 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.
8. Kolho KL, Paakkanen R, Lepistö A, et al. Novel associations between major histocompatibility complex and pediatric-onset inflammatory bowel disease. *J Pediatr Gastroenterol Nutr*. 2016;62:567–572.
9. Nambu R, Hagiwara S, Kubota M, et al. Difference between early onset and late-onset pediatric ulcerative colitis. *Pediatr Int.* 2016;58:862–6.
10. Rinawi F, Assa A, Hartman C, et al. Long-term Extent Change of Pediatric-Onset Ulcerative Colitis. *J Clin Gastroenterol.* 2017; doi:10.1097/MCG.0000000000000741.
11. Cutler DJ, Zwick ME, Okou DT, et al. Dissecting Allele Architecture of Early Onset IBD Using High-Density Genotyping. *PLoS One* 2015;10:e0128074.
12. Andreoletti G, Ashton JJ, Coelho T, et al. Exome analysis of patients with concurrent pediatric inflammatory bowel disease and autoimmune disease. *Inflamm Bowel Dis*. 2015;21:1229–1236.
13. Ahmad T, Marshall SE, Jewell D. Genetics of inflammatory bowel disease: the role of the HLA complex. *World J Gastroenterol*. 2006;12:3628–3635.
14. Westerlind H, Mellander MR, Bresso F, et al. Dense genotyping of immune-related loci identifies HLA variants associated with increased risk of collagenous colitis. *Gut*. 2017;66:421–428.
15. Raychaudhuri S, Sandor C, Stahl EA, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet*. 2012;44:291–296.
16. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *Plos Genet*. 2014;10:e1004383.
17. Dotson JL, Crandall WV, Zhang P, et al. Feasibility and validity of the pediatric ulcerative colitis activity index in routine clinical practice. *J Pediatr Gastroenterol Nutr*. 2015;60:200–204.
18. de Bie CI, Buderus S, Sandhu BK, et al. Diagnostic workup of paediatric patients with inflammatory bowel disease in Europe: results of a 5-year audit of the EUROKIDS registry. *J Pediatr Gastroenterol Nutr.* 2012;54:374–80.
19. Dignass A, Eliakim R, Magro F, et al. Second european evidence-based consensus on the diagnosis and management of ulcerative colitis part 1: definitions and diagnosis. *J Crohns Colitis*. 2012;6:965–990.
20. Lee JC, Biasci D, Roberts R, et al. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat Genet.* 2017;49:262–268.
21. Abecasis GR, Cherny SS, Cookson WO, et al. GRR: graphical representation of relationship errors. *Bioinformatics*. 2001;17:742–743.
22. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904–909.
23. 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;81:559–575.
24. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *Plos Genet*. 2009;5:e1000529.
25. Marchini J, Cutler D, Patterson N, et al. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet.* 2006;78:437–50.
26. Jia X, Han B, Onengut-Gumuscu S, et al. Imputing amino acid polymorphisms in human leukocyte antigens. *Plos One*. 2013;8:e64683.
27. Silverberg MS, Cho JH, Rioux JD, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet*. 2009;41:216–220.
28. Chiaroni-Clarke RC, Li YR, Munro JE, et al. The association of PTPN22 rs2476601 with juvenile idiopathic arthritis is specific to females. *Genes Immun*. 2015;16:495–498.
29. Ungaro R, Mehandru S, Allen PB, et al. Ulcerative colitis. *Lancet* 2017;389:1756–1770.
30. Kolho KL. Assessment of disease activity in pediatric ulcerative colitis. *Expert Rev Gastroenterol Hepatol* 2016;10:1127–34.
31. Romano C, Syed S, Valenti S, et al. Management of acute severe colitis in children with ulcerative colitis in the biologics era. *Pediatrics* 2016;137:e20151184.
32. Ostrowski J, Paziewska A, Lazowska I, et al. Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the polish population. *Sci Rep*. 2016;6:39831.
33. 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.
34. Malaty HM, Abraham BP, Mehta S, et al. The natural history of ulcerative colitis in a pediatric population: a follow-up population-based cohort study. *Clin Exp Gastroenterol*. 2013;6:77–83.