Published in final edited form as:

Inflamm Bowel Dis. 2013 August ; 19(9): 1799–1808. doi:10.1097/MIB.0b013e31828706a0.

A Novel Approach to Detect Cumulative Genetic Effects and Genetic Interactions in Crohn's Disease

Ming-Hsi Wang, MD, PhD^{1,2}, Claudio Fiocchi, MD^{1,2}, Richard H. Duerr, MD^{3,4}, Xiaofeng Zhu, PhD⁵, Stephan Ripke, MD⁶, Jean-Paul Achkar, MD^{1,2}, and The Wellcome Trust Case Control Consortium

¹Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, OH

²Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, OH

³Division of Gastroenterology, Hepatology, & Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA

⁴Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA

⁵Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH

⁶Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA

Abstract

Background—Genome-wide association studies (GWAS) have identified at least 71 Crohn's disease (CD) genetic risk loci but the role of gene-gene interactions is unclear. The value of genetic variants in clinical practice is not defined due to limited explained heritability.

Materials and Methods—We examined model predictability of combining the 71 CD risk alleles and genetic interactions in an ongoing inflammatory bowel disease (IBD) GWAS. The Wellcome Trust Case-Control Consortium (WTCCC) IBD GWAS was used as a replicate cohort. We used logic regression (LR), an adaptive regression methodology, to search for high order binary predictors (e.g. single nucleotide polymorphism [SNP]) interactions.

Results—The combined 71 CD SNPs had good CD risk predictability (area under the curve, AUCs of 0.75 and 0.73 in the 2 cohorts). Higher cumulative allele score predicted higher CD risk, but a relatively small difference in cumulative allele scores was observed between CD and controls (49 vs. 47, $P < 0.001$). Through LR, we identified high order genetic interactions and significantly improved the model predictability (AUC, 0.75 to 0.77, $P < 0.0001$). A genetic interaction model including *NOD2*, *ATG16L1*, *IL10/IL19*, *C13orf31* and *chr21q* loci was discovered and successfully replicated in the independent WTCCC cohort. The explained heritability of the 71 CD SNPs alone was 24% and increased to 27% after adding the genetic interactions.

Conclusions—A novel approach allowed the identification and replication of genetic interactions among *NOD2*, *ATG16L1*, *IL10/IL19*, *C13orf31* and *chr21q* loci. CD risk can be predicted by a model of 71 CD loci and improved by adding genetic interactions.

Correspondence: Jean-Paul Achkar, MD, Cleveland Clinic, Department of Gastroenterology and Hepatology, Desk A31, Cleveland, OH 44195, achkarj@ccf.org.

Disclosures for all authors: No relevant potential conflicts

Keywords

genetics; genetic interaction; cumulative genetic effect; Crohn's disease

Introduction

The pathogenesis of Crohn's disease (CD) is likely the result of an inappropriate persistent immune response in which genetic, microbial and environmental factors play important roles. Recent genome-wide association studies (GWAS) and meta-analyses have identified at least 71 CD-associated genetic risk loci (1). However, the value of currently known genetic variants in clinical practice is still unclear.

Single genetic marker analysis has limited value in predicting risk for complex diseases such as CD and studies suggest that the predictive value of combining multiple common genetic variants can be improved, especially when GWAS data are incorporated (2-4). The development of cumulative CD genetic risk scores based on GWAS results to predict disease risk could provide clinical strategies for disease screening, extent of diagnostic testing, and potentially treatment. Weersma et al proposed genetic risk profiles based on 10 CD single nucleotide polymorphisms (SNPs) identified from GWAS, but the tradeoff between sensitivity and specificity limited the discriminative accuracy of such models (5). Similarly, genetic risk profiles consisting of 22 GWAS meta-analysis CD SNPs could not explain the excessive prevalence of CD in individuals of Ashkenazi Jewish descent (6). To our knowledge, CD risk prediction using the recent 71 CD GWAS genetic loci has not been investigated.

Furthermore, even after accounting for cumulative effects of the 71 CD loci, only 23% of CD heritability can be explained (1). Using a model effect size distribution (7), only a small increase in the explained percentage of CD heritability is predicted when adding further, new CD genetic loci (1). Therefore, it is possible that identifying genetic interactions could uncover some of the missing heritability left out from genetic additive models and could provide better CD risk predictive models. In addition, such an approach could expand the understanding of the pathogenic mechanisms of CD (8).

The aims of this study are to assess the distribution and CD risk predictability of the 71 loci identified in CD GWAS meta-analyses and also to explore high order genetic interactions and associated additional contributions to model predictability. In order, to assess interaction effects, we used logic regression (LR), a novel regression method that can identify logical interactions of binary predictors (9-11). We applied LR to the 71 CD SNPs to search for logical SNP combinations that can explain the disease risk better than an additive model. We then replicated the identified genetic interactions in a larger independent population-based dataset and also assessed whether these interactions could recover any missing heritability. To our knowledge, this is the first study to evaluate model predictability of combining the 71 GWAS meta-analysis CD loci with high order genetic interactions.

Materials and Methods

GWAS Datasets

Two GWAS datasets were used for this study, the Cleveland Clinic/University of Pittsburgh (CC/UP) inflammatory bowel disease (IBD) GWAS and The Wellcome Trust Case-Control Consortium (WTCCC) IBD GWAS. The CC/UP GWAS dataset was used for the cumulative risk allele analysis and as the discovery dataset for evaluation of high order genetic interaction. The study design and data collection of this GWAS have been

previously described (12). Of note, the full GWAS has not yet been completed as the replication phase of the study is ongoing. However, we were able to pursue the current study as its main purposes were to predict CD risk using the 71 CD GWAS loci and to identify high order genetic interactions through a novel methodological approach. In brief, this GWAS consists of 692 CD cases and 1,428 controls, all of non-Jewish, European ancestry, who were genotyped using the Illumina HumanOmni-1 Quad beadchip (Illumina, San Diego, CA, USA) at the Feinstein Institute for Medical Research of the North Shore-Long Island Jewish Health System. All participants gave written informed consent. Genotype imputation was done using a 5-Mb region across the whole genome with imputation program (BEAGLE) (13). All 71 CD meta-analysis SNPs were imputed with good quality (R-square >0.90) and with Hardy-Weinberg equilibrium (HWE) $P > 10^{-05}$ in controls.

The WTCCC IBD GWAS (Affymetrix GeneChip 500K arrays) was used as the replicate dataset for the interaction analysis. We downloaded the genotype data called by the CHIAMO algorithm (14) for the CD samples and the shared controls (the 1958 Birth Cohort and UK Blood Service sample) from the WTCCC website. After excluding individuals with evidence of non-European ancestry or poor call rates, 1,748 CD cases and 2,980 controls remained. We applied the following quality controls criteria to exclude SNPs: 1) HWE test P values $< 5.7 \times 10^{-07}$ in controls; 2) minor allele frequency <1% in CD cases and controls; 3) call rate <99%; and 4) plate relatedness (suggesting batch effect) provided by WTCCC. In total, 469,557 SNPs passed the quality control filters (15). Of the 71 CD meta-analysis SNPs, 24 were successfully genotyped in WTCCC dataset. The remaining 47 ungenotyped SNPs were imputed using the MaCH program (16;17) with 1000 Genome phase 1 haplotypes data as reference. We used parameters of 60 iterations of the Markov sampler and 200 states. The 1000 genomes reference panel was obtained from the University of Michigan Abecasis lab, version 20100804 (<http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-2010-08.html>). A total of 566 reference haplotypes for European ancestry served as the reference panel. Among the 47 imputed CD SNPs, 43 had good imputation quality (R-square>0.60) and four had borderline or low quality of imputation; chr19: rs281379 (R-square 0.36), chr19: rs736289 (R-square 0.33), chr19: rs12720356 (R-square 0.30), and chr9: rs4077515 (R-square 0.27).

Statistical analysis

Cumulative alleles of 71 CD GWAS loci—Risk alleles were designated as the alleles reported to be associated with CD in the GWAS meta-analysis paper (1). Odds ratios (ORs) and 95% confidence intervals (CI) were calculated to estimate effects for risk alleles and genotypes in dominant mode. These analyses were implemented in the Golden Helix SVS software suite 7 (Golden Helix, Bozeman, MT).

The cumulative allele score was calculated by summing up the number of risk alleles (in dominant mode) across the 71 CD SNPs for each study subject with complete genotype information. Prediction of CD risk by cumulative allele score was conducted through conventional logistic regression model using SAS 9.2/Genetics package PROC LOGISTIC procedures (SAS Institute, Cary, NC). The CD predicted probability was calculated from the log odds (i.e. risk=odds / (1+odds)) using the logistic regression model including all 71 CD SNPs. The correlation between the cumulative allele score and the predicted probabilities was assessed by the Spearman correlation coefficient (ρ).

Exploring high order genetic interactions using logic regression among 71 CD genes—LR (R package Logic.Reg ver. 1.4.10) was used to search for models of high order SNP combinations. All 71 CD SNPs were recoded into binary predictors in dominant mode to have consistency as we calculated the cumulative allele score. Simulated annealing, a

stochastic search algorithm with increase of model size (i.e. the complexity of SNP logical combinations), implemented in LR was used to search for models composed of logical SNP interactions (called ‘Trees’) which could explain the disease risk better than single SNP models. Through the randomization processes, a test for different model sizes can be used to determine an optimal model size and, in combination with the greedy algorithm, a best model can then be identified (9-11). To assess the significance of genetic interactions contributed from the chosen Trees (i.e. high order SNP interactions), we evaluated the CD association of each Tree after excluding the marginal effects of the SNPs from which the Tree was composed.

Model predictability—We then applied logistic regression modeling to investigate the predictive accuracy of models derived from the 71 CD SNPs with and without including the SNP interactions (i.e. Trees). Discriminative accuracy was evaluated using the area under the receiver operating characteristic (ROC) curves (AUC). The AUC can range from 0.5 (lack of discrimination) to 1.0 (perfect discrimination). AUCs were calculated for the predicted risks of the logistic regression models, the cumulative allele score, and the linear predictor values of the logistic regression models. AUCs were compared with SAS 9.2 GPLOT procedures (SAS Institute, Cary, NC). Goodness-of-fit tests between the models of 71 CD SNPs with and without genetic interactions (i.e. Trees) were conducted through the PROC LOGISTIC and ROC functions in SAS 9.2.

Assessment of heritability—The estimate of the proportion of disease variation explained by a model can be regarded as a measure of goodness of fit to the data and this can be done using the generalized pseudo R-square estimate from the logistic regression model. SAS gives the likelihood-based pseudo R-square and its rescaled estimates for the proportion of disease variation explained by the genetic predictors (18;19).

Results

Individual effects of 71 SNPs on CD risk

Table 1 shows the CD risk of each of the 71 CD SNPs in the CC/UP dataset. The association patterns of effect sizes (ORs) and association directions are similar to those in the recent CD meta-analysis paper (1), suggesting that our dataset is representative of those included in the meta-analysis study which consisted of 6,333 CD cases and 15,056 controls (1).

Cumulative allele score analysis

Figure 1 shows the ORs associated with increasing cumulative allele scores for CD subjects compared to a reference group of CD subjects with 44 risk alleles, which represented the lowest 20% of total sample in a logistic regression model.

Interestingly, the difference in average cumulative allele scores between CD subjects ($49.1 \pm \text{SD } 3.3$; range: 37 to 60) and controls ($46.9 \pm \text{SD } 3.4$; range: 36 to 56) was relatively small (Figure 2A). Not surprisingly, this absolute difference in the average cumulative allele score between CD and controls became increasingly smaller after sequentially removing the SNPs which are most strongly associated with CD risk (Fig., Supplemental Digital Content 1, <http://links.lww.com/IBD/A165>) until the top 49 CD-associated SNPs were removed. This finding highlights that the difference in the cumulative allele score between CD and controls is relatively small and is driven by more than just a few SNPs.

Figure 2B shows predicted CD risk from a logistic regression model that includes all 71 CD SNPs in relation to the cumulative allele score. The Spearman correlation coefficient was

0.80 (dominant mode), indicating good correlation between CD risk and the cumulative allele score.

Genetic Interactions analysis using LR

Applying LR to the 71 CD meta-analysis SNPs in the CC/UP GWAS dataset, a best model of five Trees that included 13 SNPs was identified through the LR randomization model comparison process:

Tree1: rs2076756|*NOD2* and [(rs3764147|*C13orf31* and (rs3792109^c|*ATG16L1*)) or (rs3024505|*IL10/IL19* or (rs1736020^c|*chr21q*))]

Tree2: (rs7927997^c|*C11orf30*) and [(rs13073817^c|*chr3p*) or (rs3180018^c|*MUC1/SCAMP3*)]

Tree3: (rs3792109^c|*ATG16L1*) or (rs6568421|*PRDM1* or rs8005161|*GALC/GPR65*)

Tree4: (rs7714584^c|*IRGM*) and rs2476601|*PTPN22*

Tree5: rs11209026^c|*IL23R*

(*footnote: 'c' means the complement of the SNP risk allele, i.e. not carrying the risk allele)

To test whether the LR-identified SNP interactions (i.e. Trees) added additional information beyond the SNPs from which they were composed, we examined each Tree's CD associations (except for Tree5 because it contained only one SNP) after adjusting the marginal effects of the SNPs from which the Tree was composed. For example, for Tree1, we assessed significance after adjusting for the five SNPs from which Tree1 was composed.

$$Y = \text{intercept} + \beta_1 * \text{Tree1} + \beta_2 * \text{rs2076756}|NOD2 + \beta_3 * \text{rs3764147}|C13orf31 + \beta_4 * \text{rs3792109}^c|ATG16L1 + \beta_5 * \text{rs3024505}|IL10/IL19 + \beta_6 * \text{rs1736020}^c|chr21q$$

All four tested Trees (Trees1-4) remained significant (nominal P = 0.002) after adjusting for the SNPs from which the Trees were composed (Table 2).

Analysis of model discriminative accuracy of 71 CD SNPs and their interactions in the discovery CC/UP cohort

We next assessed whether the SNPs interactions (i.e. Trees 1-5) improved CD risk prediction. The AUC was 0.75 for prediction based on the 71 CD SNPs and was significantly improved to 0.77 (P<0.0001) after adding the SNPs interactions (i.e. Tree1-5).

Replication using the WTCCC case-control dataset

Distributions of cumulative allele score between CD and controls and predicted CD risk in the replication WTCCC dataset are shown in Fig. A and B, Supplemental Digital Content 2, <http://links.lww.com/IBD/A166>. Similar to the CC/UP cohort, the average cumulative allele scores in the WTCCC were 49.1 + SD 3.3 (range: 38 to 59) for CD and 46.9 + SD 3.4 (range: 35 to 58) for controls. The Spearman correlation coefficient (rho ρ) between the cumulative allele score and the predicted probability was 0.79, indicating a good correlation between CD risk and the cumulative allele score.

Figure, Supplemental Digital Content 3, <http://links.lww.com/IBD/A167>, shows that the ORs become higher with the increasing cumulative allele scores compared with a reference group having 44 risk alleles in a logistic regression model. For example, compared to reference group (i.e. carrying 44 risk alleles), carrying 45, 47, 49, 51, and 53 risk alleles have increasing ORs of 1.46 (95% CI: 0.94-2.28, P=0.09), 2.48 (95% CI: 1.65-3.73, P<0.0001), 3.10 (95% CI: 2.09-4.59, P<0.0001), 5.21 (95% CI: 3.38-8.03, P<0.0001), and 8.51 (95% CI: 5.08-14.3, P<0.0001) respectively.

We applied the same five genetic interactions ('Tree1-5') identified in the primary CC/UP cohort in the larger independent replicate WTCCC dataset. The AUC was 0.733 for prediction based on the 71 CD SNPs with a borderline increase to 0.736 ($P=0.08$) after adding the genetic interactions (i.e. Trees).

We examined Trees1-4 (not Tree5 due to having only one SNP) in the WTCCC cohort to determine if they provided additional information beyond the SNPs from which they were composed. Only Tree1 remain significant ($P=0.0004$) after adjusting for the SNPs from which the Trees were composed (Table, Supplemental Digital Content 4, <http://links.lww.com/IBD/A168>).

Analysis of genetic heritability explained by the 71 CD SNPs and the LR-identified genetic interactions (Trees)

We estimated the likelihood-based rescaled pseudo R-square for both CC/UP and WTCCC cohorts. In the CC/UP discovery cohort, the explained heritability of the 71 CD SNPs alone in dominant mode was 24% and increased to 27% after adding the LR-identified high order genetic interactions (Trees1-5). In the replicate WTCCC cohort, only a small increase of explained heritability was observed (from 21% to 22%) by adding the same SNP interactions (Tree1-5) into the 71 CD SNP model.

Discussion

We investigated the predictive value of combining the 71 CD meta-analysis SNPs in a case-control study and replicated the findings in a larger independent population-based study. We found that the combination of the 71 CD SNPs (in dominant mode) has modest to good predictability of CD risk (AUC 0.75 in our discovery CC/UP cohort and 0.73 in the replicate WTCCC cohort). Higher cumulative allele scores predict higher CD risk. However, our study also highlights that the absolute difference of CD cumulative allele scores in CD ($49.1 \pm \text{SD } 3.3$) versus controls ($46.9 \pm \text{SD } 3.4$) is relatively small but statistically different ($P<0.001$). This small difference was mostly driven by several of the top CD-associated SNPs. We also identified high order genetic interactions among the 71 CD loci using a novel approach (LR) and found that risk prediction improved by adding genetic interactions to the 71 CD SNP model (AUC increased from 0.75 to 0.77, $P<0.0001$). High order SNP interaction Tree1, composed of five SNPs in *NOD2*, *ATG16L1*, *IL10/IL19*, *C13orf31* and *chr21q*, was successfully replicated in a large independent WTCCC cohort. To our knowledge, this is the first study investigating CD predictability using the combined 71 CD meta-analysis loci and showing improved model predictability and explained CD heritability by adding the genetic interactions identified through a novel application of LR.

There are several algorithms to search for interactions including two well-known adaptive algorithms CART (20) and MARS (21). In this study we used LR, which has been shown to outperform CART and MARS in situations in which the underlying model is particularly complicated (9-11). LR is a tool to detect interactions between binary predictors that may play important roles in complex traits. Applying LR to the 71 CD SNPs, we identified five high order logical SNP interactions (i.e., Trees1-5). Trees 1, 2, 3 and 4 were composed from two to five SNPs while Tree 5 had only one SNP (rs11209026 in *IL23R*).

Gene-gene interactions can be approached in statistical versus mechanistic ways (22). Looking for genes interacting mechanistically may be more useful than identifying statistical interactions in understanding which factors are parts of the biological mechanisms influencing the CD susceptibility. To assess if the LR-identified SNP interactions could add additional information beyond that from the individual SNPs which constituted the Tree models, we applied conventional logistic regression model to compare a model consisting of

the 70 single SNPs to a model that included both the 70 SNPs and the identified Trees. The SNP interactions remained statistically significant after excluding the marginal effects of the SNPs that made up the Trees ($P = 0.002$, Table 2). By examining the excess risk in a logistic regression model, our approach is similar to testing the mechanistic interaction to identify the underlying causal mechanisms (22). When we applied the same Trees model in the replicate WTCCC cohort, only Tree1 remained significant ($P=0.0004$). This finding further supports the evidence of genetic interaction discovered in Tree1.

Once high order genetic interactions are identified through LR, the next challenge relates to interpreting and proving that the statistically identified genetic interactions are of importance at a biological level. Tree1, composed of SNPs in *NOD2*, *ATG16L1*, *IL10/IL19*, *C13orf31* and a gene poor region on *chr21q*, demonstrates a good example of biological plausibility for the identified genetic interactions by highlighting mechanisms of host response to microbial organisms. *NOD2* was the first gene identified to be associated with CD (23;24) and its variants affect the leucine rich repeat domain of the nucleotide-binding oligomerization domain (Nod) microbial-recognition receptors. *ATG16L1* is part of the autophagy pathway which is a crucial element of the innate immune response to intracellular bacteria. Recent studies have provided a mechanistic link between *NOD2* and *ATG16L1* (25;26), both of which were included in the SNP interaction term Tree1. One study demonstrated that bacterial autophagy triggered by Nod proteins relied on interactions with *ATG16L1*. In the absence of other activators, Nod1 and Nod2 induced autophagy by interacting with *ATG16L1* and promoted the sequestration of bacteria inside autophagosomes (26). Similarly, a study from our group showed evidence of functional interactions between *ATG16L1* and *NOD2* in an autophagy-dependent antibacterial pathway that is altered by CD-associated variants in a cell type-specific manner (25). IL-10 is an anti-inflammatory cytokine with several lines of evidence to support its role in the pathogenesis of IBD including identification of genetic variants in the IL-10 receptor in subjects with early onset IBD (27-30). From a possible gene-gene interaction standpoint, peripheral blood mononuclear cells from patients with the *NOD2* 3020insC mutation show defective release of IL-10 after stimulation with microbial organisms (31). Furthermore, a recent study illustrated the biological relationship between *NOD2* 3020insC loss-of-function mutations and inhibited IL-10 transcription by blocking phosphorylation of the nuclear ribonucleoprotein hnRNP-A1 (32). Interleukin-19 is a member of the IL-10 family and has been shown to have a role in protecting the host from colonic bacterial infection following injury to intestinal epithelial cells (33).

A candidate gene *USP25* (ubiquitin-specific deubiquitination protease) at gene-poor region chr21q11.2 (34), located 200kb away from SNP rs1736020 in Tree1, is a member of the DUB (deubiquitinating enzymes) superfamily. The first DUB associated disease, familial cylindromatosis, disclosed the biological relationship between the nuclear factor-kappa B signaling pathway and deubiquitination enzymes (35). In addition, DUB protein has been increasingly found involved in human pathological disorders such as Parkinson's disease, prostate cancer, lung cancer and ataxia. (36;37). Another locus *C13orf31* (chromosome 13 open reading frame), which accounted for SNP rs3764147 in Tree1, has been found to be associated with leprosy (38) and IBD (39). *C13orf31* encodes laccase (multicopper oxidoreductase) domain containing 1 protein which acts as antioxidant molecule, scavenging free radicals in the bloodstream and eliminating pro-cancerous environmental estrogen bisphenol A (40). Based on all these lines of evidence we conclude that there is biological plausibility for interaction between the 5 genes/loci identified in our replicated Tree1. Follow up functional work will be necessary although we believe that the methodology we have used can help identify pathway based approaches to unravel the pathogenesis of IBD.

As previously reported, only 23% of CD heritability can be explained by the 71 CD GWAS meta-analysis SNPs (1). For case-control studies, disease heritability can be alternatively approximated by the proportion of disease variability explained in a SNP model using generalized R-square measures implemented in the logistic regression model. In our primary CC/UP GWAS, 24% of phenotypic variability was explained by the 71 CD SNPs in dominant mode. After adding the SNP interactions (i.e. Trees) into the 71 CD SNP model, the explained heritability increased from 24% to 27%. More heritability would be expected to be explained by searching more complicated interaction models within or between IBD related biological pathways (e.g. Th17/Th1 pathway, TNF α , NF- κ B signaling, the KIR-family, and the NOD2 pathway) (41). Of note, in the replicate WTCCC dataset, only a minimal increase of explained CD variability was observed (from 21% to 22%) by adding the SNPs interactions into the 71 CD SNP model, which likely relates to heterogeneity between the two populations.

The limitations of this study include the relatively small sample size in our CC/UP GWAS (692 CD and 1437 controls) which may explain why some of the 71 CD meta-analysis SNPs did not achieve genome-wide level of significance. Also, the suboptimal imputation quality in a few SNPs from the replicate WTCCC cohort may hinder the reproducibility of the selected SNP interactions model. Despite using a high imputation accuracy program like MaCH (16;17;41), several factors can still affect imputation results including the density of genotype platform in the experimental sample (the WTCCC Affymetrix 500k gene chip is considered relatively low density for a GWAS platform), level of linkage disequilibrium in different genetic regions, minor allele frequency of the marker being imputed, and degree of genetic relationship or population heterogeneity between the experimental and reference populations.

In conclusion, this is the first study investigating CD risk using the 71 CD meta-analysis loci and demonstrating increased risk with increasing CD risk allele burden. However, we also highlighted the small differences of cumulative allele scores between CD subjects ($49.1 \pm \text{S.D. } 3.3$) and healthy controls ($46.9 \pm \text{S.D. } 3.4$) which may explain why only 23% of CD heritability is attributed to the 71 CD loci. Through a novel approach, we found a high order genetic interaction which was successfully replicated in a large independent WTCCC cohort. Using the 71 CD SNPs, model predictability improved from 75% to 77% and estimated explained heritability improved from 24% to 27% after adding the genetic interactions. Potential future applications of our work will include verifying these findings in a larger population-based cohort, including environmental factors to explore gene-environment interactions, defining the biological significance of suggested interactions in cell line or animal models, and searching for genetic interactions using pathway based approaches.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the patients and the controls for participating in this study. We acknowledge the Feinstein Institute for Medical Research of the North Shore-Long Island Jewish Health System for Illumina Genotyping BeadChip processing. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

Grant Support: T32 DK083251, NIH - NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES (NIDDK)(M-H W, CF); DK068112 (J-PA), DK062420 (RHD) and DK076025 (RHD); a

Crohn's & Colitis Foundation of America Senior Research Award (RHD); and funds generously provided by Kenneth and Jennifer Rainin, Gerald and Nancy Goldberg, and Victor and Ellen Cohn.

References

1. Franke A, McGovern DPB, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet.* 2010; 42:1118–1125. [PubMed: 21102463]
2. Liu S, Song Y. Building Genetic Scores to Predict Risk of Complex Diseases in Humans: Is It Possible? *Diabetes.* 2010; 59:2729–2731. [PubMed: 20980472]
3. Yang Q, Khoury MJ, Botto L, et al. Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *Am J Hum Genet.* 2003; 72:636–649. [PubMed: 12592605]
4. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research.* 2007; 17:1520–1528. [PubMed: 17785532]
5. Weersma RK, Stokkers PCF, Cleynen I, et al. Confirmation of Multiple Crohn's Disease Susceptibility Loci in a Large Dutch-Belgian Cohort. *Am J Gastroenterol.* 2009; 104:630–638. [PubMed: 19174780]
6. Peter I, Mitchell A, Ozelius L, et al. Evaluation of 22 genetic variants with Crohn's Disease risk in the Ashkenazi Jewish population: a case-control study. *BMC Medical Genetics.* 2011; 12:63. [PubMed: 21548950]
7. Park JH, Gail MH, Weinberg CR, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences.* 2011; 108:18026–18031.
8. Zuk, O.; Hechter, E.; Sunyaev, SR., et al. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences;* 2012; p. 1193-1198.
9. Ruczinski I, Kooperberg C, LeBlanc L. Logic Regression. *Journal of Computational and Graphical Statistics.* 2003; 12:475–511.
10. Ruczinski I, Kooperberg C, LeBlanc L. Exploring interactions in high-dimensional genomic data: an overview of Logic Regression, with applications. *Journal of Multivariate Analysis.* 2004; 90:178–195.
11. Schwender H, Ruczinski I. Logic regression and its extensions. *Adv Genet.* 2010; 72:25–45. [PubMed: 21029847]
12. Achkar JP, Klei L, Bakker PIW, et al. Amino acid position 11 of HLA-DR[beta]1 is a major determinant of chromosome 6p association with ulcerative colitis. *Genes Immun.* 2012; 13:245–252. [PubMed: 22170232]
13. Browning BL, Browning SR. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. 2009:210–223.
14. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007; 39:906–913. [PubMed: 17572673]
15. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447:661–678. [PubMed: 17554300]
16. Li Y, Willer C, Sanna S, et al. Genotype imputation. *Annu Rev Genomics Hum Genet.* 2009; 10:387–406. [PubMed: 19715440]
17. Li Y, Willer CJ, Ding J, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2010; 34:816–834. [PubMed: 21058334]
18. Makowsky R, Pajewski NM, Klimentidis YC, et al. Beyond missing heritability: prediction of complex traits. *PLoS Genet.* 2011; 7:e1002051. [PubMed: 21552331]
19. Stokes, ME.; Davis, C.; Koch, G. Categorical data analysis using the SAS system. SAS Institute Inc.; 2000.
20. Breiman, L.; Friedman, JH.; Stone, C., et al. Classification and regression trees. Monterey, CA: Wadsworth & Brooks; 1984.
21. Friedman JH, Roosen CB. An introduction to multivariate adaptive regression splines. *Stat Methods Med Res.* 1995; 4:197–217. [PubMed: 8548103]

22. Berzuini, C.; Dawid, P.; Zhang, H., et al. Analysis of interaction for identifying causal mechanisms. In: Berzuini, C.; Dawid, P.; Bernardinelli, L., editors. *Causality: Statistical Perspectives and Applications*. Wiley; 2012. p. 192-207.
23. Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*. 2001; 411:599–603. [PubMed: 11385576]
24. Ogura Y, Bonen DK, Inohara N, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*. 2001; 411:603–606. [PubMed: 11385577]
25. Homer CR, Richmond AL, Rebert NA, et al. ATG16L1 and NOD2 Interact in an Autophagy-Dependent Antibacterial Pathway Implicated in Crohn's Disease Pathogenesis. *Gastroenterol*. 2010; 139:1630–1641.
26. Travassos LH, Carneiro LAM, Ramjeet M, et al. Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nat Immunol*. 2010; 11:55–62. [PubMed: 19898471]
27. Li MC, He SH. IL-10 and its related cytokines for treatment of inflammatory bowel disease. *World J Gastroenterol*. 2004; 10:620–625. [PubMed: 14991925]
28. Kuhn R, Lohler J, Rennick D, et al. Interleukin-10-deficient mice develop chronic enterocolitis. *Cell*. 1993; 75:263–274. [PubMed: 8402911]
29. Kotlarz D, Beier R, Murugan D, et al. Loss of interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and therapy. *Gastroenterology*. 2012; 143:347–355. [PubMed: 22549091]
30. Glocker EO, Kotlarz D, Boztug K, et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N Engl J Med*. 2009; 361:2033–2045. [PubMed: 19890111]
31. Netea MG, Kullberg BJ, de Jong DJ, et al. NOD2 mediates anti-inflammatory signals induced by TLR2 ligands: implications for Crohn's disease. *Eur J Immunol*. 2004; 34:2052–2059. [PubMed: 15214053]
32. Noguchi E, Homma Y, Kang X, et al. A Crohn's disease-associated NOD2 mutation suppresses transcription of human IL10 by inhibiting activity of the nuclear ribonucleoprotein hnRNP-A1. *Nat Immunol*. 2009; 10:471–479. [PubMed: 19349988]
33. Azuma YT, Matsuo Y, Kuwamura M, et al. Interleukin-19 protects mice from innate-mediated colonic inflammation. *Inflamm Bowel Dis*. 2010; 16:1017–1028. [PubMed: 19834971]
34. Valero R, Marfany G, Gonzalez-Angulo O, et al. USP25, a novel gene encoding a deubiquitinating enzyme, is located in the gene-poor region 21q11.2. *Genomics*. 1999; 62:395–405. [PubMed: 10644437]
35. Kovalenko A, Chable-Bessia C, Cantarella G, et al. The tumour suppressor CYLD negatively regulates NF-kappaB signalling by deubiquitination. *Nature*. 2003; 424:801–805. [PubMed: 12917691]
36. Leroy E, Boyer R, Auburger G, et al. The ubiquitin pathway in Parkinson's disease. *Nature*. 1998; 395:451–452. [PubMed: 9774100]
37. Graner E, Tang D, Rossi S, et al. The isopeptidase USP2a regulates the stability of fatty acid synthase in prostate cancer. *Cancer Cell*. 2004; 5:253–261. [PubMed: 15050917]
38. Zhang FR, Huang W, Chen SM, et al. Genomewide association study of leprosy. *N Engl J Med*. 2009; 361:2609–2618. [PubMed: 20018961]
39. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*. 2008; 40:955–962. [PubMed: 18587394]
40. Keri RA, Ho SM, Hunt PA, et al. An evaluation of evidence for the carcinogenic activity of bisphenol A. *Reprod Toxicol*. 2007; 24:240–252. [PubMed: 17706921]
41. Cho JH. IIBDGC. Abstract #868. Immunochip-based analysis of a large IBD case-control cohort identifies 50 novel loci, refining definitions of disease pathways. DDW. 2012

Abbreviations used in this paper

CD Crohn's disease

NOD2	Nucleotide-binding oligomerization domain containing-2
ATG16L1	Autophagy-related protein 16-1
IL-10	interleukin-10
IL-19	interleukin-19
IL23R	interleukin (IL) 23 receptor
IBD	inflammatory bowel disease
GWAS	genome wide association study
SNP	single nucleotide polymorphism
HWE	Hardy-Weinberg Equilibrium
WTCCC	Wellcome Trust Case-Control Consortium

Figure 1. Odds Ratios (ORs) and 95% C.I. for CD according to the number of risk alleles carried in CC/UP cohort

Figure 2A. Histograms of cumulative allele scores in CD vs. controls in CC/UP cohort

Figure 2B. Correlation between the predicted CD risks and the cumulative allele score in CC/UP cohort

Table 1
Basic characteristics and individual effects of 71 CD meta-analysis SNPs in previous CD GWAS meta-analysis study and in our CC/UP GWAS

dbSNP RS ID	chr	position (b.p.) Build 37.3	Genes of interest	Risk allele (meta)	P meta	OR meta	Risk allele (CC/UP) case/control	P (CC/UP)	OR (CC/UP)
rs2476601	1p13	114377568	<i>PTPN22</i>	G/A - 0.907	4.47×10^{-9}	1.26 (1.17-1.37)	0.95/0.92	1.10E-03	1.55 (1.19-2.03)
rs11209026	1p31	67705958	<i>IL23R</i>	G/A - 0.932	1.00×10^{-64}	2.66 (2.36-3.00)	0.99/0.93	9.56E-18	5.43 (3.38-8.73)
rs2797685	1p36	7879063	<i>VAMP3</i>	A/G - 0.190	7.10×10^{-9}	1.05 (1.01-1.10)	0.20/0.19	4.03E-01	1.07 (0.91-1.26)
rs3180018	1q22	155230131	<i>SCAMP3, MUC1</i>	A/G - 0.250	2.30×10^{-13}	1.13 (1.06-1.19)	0.27/0.23	9.20E-03	1.22 (1.05-1.41)
rs4656940	1q23	160830268	<i>ITLN1/CD244</i>	A/G - 0.801	6.17×10^{-7}	1.15 (1.09-1.21)	0.82/0.79	5.55E-02	1.17 (1.00-1.38)
rs7517810	1q24	172853460	<i>TNFSF18, TNFSF4, FASLG</i>	T/C - 0.246	1.51×10^{-15}	1.22 (1.16-1.28)	0.26/0.25	4.08E-01	1.06 (0.92-1.23)
rs1998598	1q31	197727642	<i>DENND1B</i>	G/A - 0.302	8.70×10^{-9}	1.04 (1.00-1.09)	0.30/0.30	9.15E-01	0.99 (0.86-1.14)
rs3024505	1q32	206939904	<i>IL10, IL19</i>	T/C - 0.157	1.60×10^{-14}	1.12 (1.07-1.17)	0.20/0.16	1.66E-03	1.31 (1.11-1.55)
rs7554511	1q32	200877562	<i>KIF21B/C1orf106</i>	C/A - 0.726	1.58×10^{-7}	1.14 (1.08-1.19)	0.77/0.75	1.68E-01	1.12 (0.96-1.30)
rs10181042	2p16	61224259	<i>RELCC2orf74</i>	T/C - 0.420	6.61×10^{-9}	1.14 (1.09-1.19)	0.45/0.42	5.99E-02	1.13 (1.00-1.29)
rs10495903	2p21	43806918	<i>THADA</i>	T/C - 0.129	1.60×10^{-14}	1.14 (1.09-1.20)	0.15/0.13	2.81E-02	1.23 (1.02-1.48)
rs13428812	2p23	25492467	<i>DNM3A</i>	G/A - 0.326	8.50×10^{-10}	1.06 (1.03-1.10)	0.33/0.31	1.30E-01	1.11 (0.97-1.28)
rs780093	2p23	27742603	<i>GCKR</i>	T/C - 0.418	4.70×10^{-11}	1.15 (1.10-1.21)	0.47/0.43	1.63E-02	1.17 (1.03-1.33)
rs2058660	2q12	103054449	<i>IL18RAP</i>	G/A - 0.231	1.58×10^{-12}	1.19 (1.14-1.26)	0.27/0.24	5.49E-02	1.15 (1.00-1.34)
rs6738825	2q33	198896895	<i>PLCL1</i>	A/G - 0.473	3.50×10^{-9}	1.06 (1.02-1.11)	0.52/0.47	8.01E-03	1.19 (1.05-1.35)
rs3792109	2q37	234184417	<i>ATG16L1</i>	A/G - 0.529	6.76×10^{-41}	1.34 (1.29-1.40)	0.61/0.51	1.80E-10	1.53 (1.34-1.74)
rs7423615	2q37	231116874	<i>SP140</i>	T/C - 0.187	3.10×10^{-13}	1.12 (1.07-1.18)	0.22/0.19	3.38E-02	1.19 (1.01-1.39)
rs3197999	3p21	49721532	<i>MST1, GPX1, BSN</i>	A/G - 0.297	6.17×10^{-17}	1.22 (1.16-1.27)	0.35/0.29	3.15E-04	1.29 (1.12-1.48)
rs13073817	3p24	18706858		A/G - 0.322	6.70×10^{-9}	1.08 (1.03-1.13)	0.36/0.34	2.29E-01	1.09 (0.95-1.24)
rs11742570	5p13	40410584	<i>PTGER4</i>	C/T - 0.606	7.08×10^{-36}	1.33 (1.27-1.39)	0.69/0.62	7.90E-06	1.36 (1.19-1.56)
rs7702331	5q13	72551134	<i>TMEM174</i>	A/G - 0.600	5.90×10^{-12}	1.12 (1.07-1.17)	0.64/0.61	1.21E-01	1.11 (0.97-1.27)
rs2549794	5q15	96244549	<i>ERAP2/LRAP</i>	C/T - 0.409	1.10×10^{-10}	1.05 (1.02-1.09)	0.42/0.41	3.36E-01	1.07 (0.94-1.21)
rs11167764	5q31	141479065	<i>NDFIPI</i>	C/A - 0.796	2.00×10^{-9}	1.06 (1.02-1.11)	0.79/0.80	5.40E-01	0.95 (0.81-1.11)
rs12521868	5q31	131784393	<i>SLC22A4/A5, IRF1, CSF2, IL3</i>	T/G - 0.422	1.41×10^{-20}	1.23 (1.18-1.28)	0.47/0.42	8.51E-04	1.25 (1.10-1.42)
rs6556412	5q33	158787385	<i>IL12B</i>	A/G - 0.332	5.37×10^{-14}	1.18 (1.13-1.24)	0.37/0.33	8.69E-03	1.20 (1.05-1.37)

dbSNP RS ID	chr	position (b.p.) Build 37.3	Genes of interest	Risk allele (meta)	P meta	OR meta	Risk allele (CC,UP) case/control	P (CC,UP)	OR (CC,UP)
rs7714584	5q33	150270420	<i>IRGM</i>	G/A - 0.088	7.76×10 ⁻¹⁹	1.37 (1.28-1.47)	0.12/0.08	1.36E-04	1.51 (1.23-1.87)
rs359457	5q35	173279842	<i>CPEB4</i>	T/C - 0.571	2.50×10 ⁻¹²	1.08 (1.04-1.12)	0.61/0.56	3.93E-03	1.21 (1.06-1.38)
rs1799964	6p21	31542308	<i>LTA,HLA-DQA2,TNFI,LTB</i>	C/T - 0.209	3.98×10 ⁻¹¹	1.19 (1.13-1.25)	0.24/0.21	1.12E-02	1.22 (1.05-1.42)
rs6908425	6p22	20728731	<i>CDKALI</i>	C/T - 0.784	1.41×10 ⁻⁸	1.17 (1.11-1.23)	0.79/0.77	2.72E-01	1.09 (0.93-1.28)
rs17309827	6p25	3433318		T/G - 0.639	6.70×10 ⁻⁹	1.10 (1.05-1.16)	0.64/0.64	7.34E-01	1.02 (0.90-1.17)
rs1847472	6q15	90973159	<i>BACH2</i>	G/T - 0.658	5.10×10 ⁻⁹	1.07 (1.03-1.11)	0.69/0.66	4.03E-02	1.16 (1.01-1.33)
rs6568421	6q21	106435025	<i>PRDMI</i>	G/A - 0.301	4.37×10 ⁻⁸	1.13 (1.07-1.18)	0.36/0.28	3.25E-07	1.43 (1.25-1.64)
rs212388	6q25	159490436	<i>TAGAP</i>	G/A - 0.393	2.30×10 ⁻¹¹	1.10 (1.05-1.14)	0.42/0.39	1.16E-01	1.11 (0.98-1.27)
rs415890	6q27	167406633	<i>CCR6</i>	C/G - 0.522	2.51×10 ⁻¹²	1.17 (1.12-1.22)	0.56/0.53	8.80E-02	1.12 (0.98-1.27)
rs1456896	7p12	50304461	<i>IKZF1, ZPBP, FIGNLI</i>	T/C - 0.69	1.20×10 ⁻⁸	1.14 (1.09-1.20)	0.71/0.70	7.47E-01	1.02 (0.89-1.18)
rs4871611	8q24	126537570		A/G - 0.609	1.51×10 ⁻¹²	1.17 (1.12-1.23)	0.65/0.62	4.55E-02	1.15 (1.00-1.31)
rs6651252	8q24	129567181		T/C - 0.865	3.90×10 ⁻¹⁸	1.23 (1.17-1.30)	0.88/0.87	2.79E-01	1.12 (0.92-1.36)
rs10758669	9p24	4981602	<i>JAK2</i>	C/A - 0.349	1.00×10 ⁻¹³	1.18 (1.13-1.23)	0.39/0.35	7.89E-03	1.20 (1.05-1.37)
rs3810936	9q32	117552885	<i>TNFSF15, 8</i>	C/T - 0.682	1.00×10 ⁻¹⁵	1.21 (1.15-1.27)	0.70/0.69	4.15E-01	1.06 (0.92-1.22)
rs40777515	9q34	139266496	<i>CARD9, SNAPC4</i>	T/C - 0.411	1.30×10 ⁻³⁶	1.18 (1.13-1.22)	0.45/0.39	4.21E-04	1.26 (1.11-1.44)
rs12242110	10p11	35535695	<i>CREM</i>	G/A - 0.315	1.10×10 ⁻⁹	1.15 (1.10-1.20)	0.37/0.32	5.94E-04	1.27 (1.11-1.45)
rs12722489	10p15	6102012	<i>IL2RA</i>	C/T - 0.852	2.90×10 ⁻⁹	1.11 (1.05-1.16)	0.88/0.85	1.36E-02	1.27 (1.05-1.53)
rs10761659	10q21	64445564	<i>ZNF365</i>	G/A - 0.538	4.37×10 ⁻²²	1.23 (1.18-1.29)	0.57/0.56	4.68E-01	1.05 (0.92-1.19)
rs1819658	10q21	59913151	<i>UBE2D1</i>	C/T - 0.774	9.10×10 ⁻¹⁷	1.19 (1.13-1.25)	0.81/0.78	5.44E-02	1.17 (1.00-1.37)
rs1250550	10q22	81060317	<i>ZMIZ1</i>	G/T - 0.669	1.10×10 ⁻³⁰	1.19 (1.15-1.23)	0.69/0.66	2.38E-02	1.17 (1.02-1.35)
rs4409764	10q24	101284237	<i>NKX2-3</i>	T/G - 0.492	2.29×10 ⁻²⁰	1.22 (1.17-1.27)	0.52/0.48	2.20E-02	1.16 (1.02-1.32)
rs102275	11q12	61557803	<i>FADS1</i>	C/T - 0.341	2.30×10 ⁻¹¹	1.08 (1.04-1.12)	0.37/0.34	1.60E-01	1.10 (0.96-1.26)
rs694739	11q13	64097233	<i>PRDX5, ESRRRA</i>	A/G - 0.626	6.00×10 ⁻¹⁰	1.10 (1.05-1.16)	0.66/0.63	8.82E-02	1.13 (0.98-1.29)
rs7927997	11q13	76301375	<i>C11orf50</i>	T/C - 0.389	5.62×10 ⁻¹³	1.17 (1.12-1.22)	0.45/0.37	1.06E-06	1.38 (1.22-1.58)
rs11564258	12q12	40792300	<i>LRRK2,MUC9</i>	A/G - 0.025	6.17×10 ⁻²¹	1.74 (1.55-1.95)	0.04/0.03	1.27E-01	1.33 (0.93-1.90)
rs2062305	13q14	43052880	<i>TNFSF11</i>	G/A - 0.346	4.90×10 ⁻¹⁰	1.10 (1.05-1.15)	0.49/0.48	5.56E-01	1.04 (0.91-1.18)
rs3764147	13q14	44457925	<i>C13orf31</i>	G/A - 0.245	1.41×10 ⁻¹⁰	1.17 (1.12-1.23)	0.28/0.24	6.48E-03	1.23 (1.06-1.42)
rs4902642	14q24	69210199	<i>ZFP36L1</i>	G/A - 0.584	1.60×10 ⁻¹⁰	1.07 (1.11-1.04)	0.61/0.58	1.35E-01	1.11 (0.97-1.26)

dbSNP RS ID	chr	position (b.p.) Build 37.3	Genes of interest	Risk allele (meta)	P meta	OR meta	Risk allele (CC,UP) case/control	P (CC,UP)	OR (CC,UP)
rs8005161	14q35	88472595	<i>GALC, GPR65</i>	T/C - 0.119	4.20×10 ⁻¹⁸	1.23 (1.16-1.31)	0.10/0.08	2.16E-02	1.30 (1.04-1.61)
rs17293632	15q22	67442596	<i>SMAD3</i>	T/C - 0.233	2.70×10 ⁻¹⁹	1.12 (1.07-1.16)	0.27/0.24	5.83E-02	1.16 (1.00-1.34)
rs151181	16p11	28490517	<i>APOB48R, IL27, SULT1A1/A2</i>	G/A - 0.386	1.50×10 ⁻¹¹	1.07 (1.03-1.12)	0.42/0.40	2.06E-01	1.09 (0.96-1.24)
rs2076756	16q12	50756881	<i>NOD2</i>	G/A - 0.26	3.98×10 ⁻⁶⁹	1.53 (1.46-1.60)	0.36/0.25	5.98E-13	1.67 (1.45-1.92)
rs3091315	17q12	32593665	<i>CCL2, CCL7</i>	A/G - 0.723	1.70×10 ⁻¹³	1.20 (1.14-1.26)	0.76/0.73	5.63E-02	1.16 (1.00-1.34)
rs11871801	17q21	40570772	<i>STAT3, MLX</i>	A/C - 0.756	2.51×10 ⁻⁸	1.15 (1.10-1.21)	0.76/0.76	7.01E-01	0.97 (0.84-1.13)
rs2872507	17q21	38040763	<i>GSMDL, ZBP2, IKZF3</i>	A/G - 0.458	1.51×10 ⁻⁹	1.14 (1.09-1.19)	0.47/0.44	5.25E-02	1.14 (1.00-1.29)
rs1893217	18p11	12809340	<i>PTPN2</i>	G/A - 0.153	1.29×10 ⁻¹⁴	1.25 (1.18-1.32)	0.18/0.15	2.65E-02	1.22 (1.03-1.44)
rs12720356	19p13	10469975	<i>TYK2, ICAM1, ICAM3</i>	G/T - 0.084	1.40×10 ⁻¹²	1.12 (1.06-1.19)	0.09/0.08	4.50E-01	1.09 (0.87-1.36)
rs740495	19p13	1124835	<i>GPX4, SBNO2</i>	G/A - 0.247	8.13×10 ⁻¹²	1.16 (1.10-1.21)	0.30/0.28	2.76E-01	1.08 (0.94-1.25)
rs281379	19q13	49214274	<i>FUT2, RASIP1</i>	A/G - 0.487	7.40×10 ⁻¹²	1.07 (1.04-1.11)	0.54/0.49	3.97E-03	1.21 (1.06-1.38)
rs736289	19q13	33757062		T/C - 0.612	8.70×10 ⁻⁹	1.06 (1.02-1.11)	0.64/0.60	6.43E-03	1.20 (1.05-1.37)
rs4809330	20q13	62349586	<i>RTEL1</i>	G/A - 0.709	2.70×10 ⁻¹⁵	1.12 (1.06-1.18)	0.71/0.71	8.57E-01	0.99 (0.86-1.14)
rs1736020	21q21	16812552		C/A - 0.579	9.33×10 ⁻¹²	1.16 (1.11-1.21)	0.63/0.58	2.96E-03	1.22 (1.07-1.39)
rs2838519	21q22	45615023	<i>ICOSLG</i>	G/A - 0.391	2.09×10 ⁻¹⁴	1.18 (1.13-1.23)	0.41/0.39	2.16E-01	1.09 (0.95-1.24)
rs181359	22q11	21928641	<i>YD1C</i>	T/C - 0.203	4.80×10 ⁻¹⁶	1.10 (1.06-1.15)	0.21/0.19	1.17E-01	1.14 (0.97-1.34)
rs713875	22q12	30592487	<i>MTMR3</i>	C/G - 0.471	7.30×10 ⁻¹²	1.08 (1.04-1.13)	0.50/0.49	6.01E-01	1.04 (0.91-1.18)
rs2413583	22q13	39659773	<i>MAP3K7IP1</i>	C/T - 0.830	1.10×10 ⁻²⁶	1.23 (1.17-1.29)	0.86/0.83	2.07E-02	1.24 (1.03-1.48)

Table 2
Assessment of the significance of SNP interactions (Trees) by adjusting for the marginal effects of SNPs from which the Trees were composed in CC/UP cohort

	Before control for SNPs		After control for SNPs	
	OR (95%CI)	P value	OR (95%CI)	P value
Tree 1	2.49 (2.03-3.04)	<0.0001	2.59 (1.80-3.73)	<0.0001
Tree 2	0.46 (0.37-0.58)	<0.0001	0.44 (0.28-0.69)	0.0003
Tree 3	2.24 (1.77-2.83)	<0.0001	1.83 (1.26-2.67)	0.001
Tree 4	0.43 (0.31-0.61)	<0.0001	0.30 (0.14-0.65)	0.002