ABSTRACT

The inflammatory bowel diseases (IBD) are a group of chronic inflammatory conditions involving the gastrointestinal tract. IBD affects between 1 and 2 million Americans, and its incidence is increasing for unknown reasons. One of the predominant forms of IBD is Crohn's disease (CD). Genome-wide association studies (GWAS) have linked many single-nucleotide polymorphisms (SNPs) within genes of the immune system to CD pathogenesis, but it is unknown which of these SNPs are causative. In genetically susceptible individuals, CD can result from a disruption in the balance between pro-inflammatory and anti-inflammatory gene expression in intestinal immune cells. Gene expression is partially regulated by the binding of transcription factors to DNA to either activate or repress target genes. The ability of transcription factors to bind to DNA may be influenced by chromatin accessibility. To determine the location of accessible, or "open" chromatin regions in the genome, we used Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE). Many of these open chromatin regions were located near genes involved in immune system function. A subset of these open chromatin regions overlapped with CD-associated SNPs. I used luciferase reporter assays to determine the potential regulatory function of a CD-linked SNP found in a region of open chromatin. Understanding the function of disease-associated genetic variants will elucidate the importance of genetic susceptibility in CD pathogenesis.

INTRODUCTION

Inflammatory bowel disease (IBD)

The inflammatory bowel diseases (IBD) are a broad group of conditions that involve chronic or recurring inflammation of the gastrointestinal tract. While most cases can be managed with medication, there is no cure for IBD, making its increasing prevalence a major public health

concern. The two major forms of IBD are Crohn's disease (CD) and ulcerative colitis. While the symptoms of these two conditions are similar, they are distinct illnesses that affect different areas of the gastrointestinal tract. Crohn's disease (CD) may affect any part of the gastrointestinal tract, while ulcerative colitis (UC) is limited to the lining of the colon and rectum. In genetically susceptible individuals, IBD is caused by an inappropriate immune response to the enteric microbiota that leads to intestinal inflammation. The hyper-activation of immune cells, as demonstrated by increased production of pro-inflammatory cytokines or decreased production of anti-inflammatory cytokines, is a central event in the development of CD[Dionne,Rogler]. The increasing incidence of CD in developed nations, including the United States, suggests that environmental triggers may interact with genetic factors to contribute to the development of this disease[Dixon, Molodecky].

Genetic variation associated with Crohn's disease

Genome-wide association studies (GWAS) and subsequent meta-analyses have linked a total of 163 single-nucleotide polymorphisms (SNPs) to IBD[Jostins]. Most of these loci are associated with both Crohn's disease and ulcerative colitis and are found in genes of the immune system, specifically genes predominantly expressed in intestinal macrophages. A unique characteristic of these GWAS findings is the observation that the majority of identified CD-associated variants occur in regions of the genome that do not code for proteins[Mirza]. Non-coding regions of the genome often act as regulatory elements that influence gene expression, such as promoters that increase the transcription of proximal downstream genes or repressors that serve to downregulate gene expression. In order to influence gene expression, regulatory elements must be accessible to transcription factors. However, the default state of chromatin is a "closed" conformation characterized by tight coiling of DNA around histone proteins. In

contrast, histone-free "open" chromatin regions can be readily accessed by transcription factors. Regulatory elements are thus most often found in regions of open chromatin that may be accessed by transcription factors and subsequently influence the expression of either nearby or distant genes[Simon].

Open chromatin regions associated with Crohn's disease

Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) is a technique used to identify regions of open chromatin throughout the genome indicative of potential regulatory elements. Peaks representing open chromatin identified by FAIRE are consistently found within known regulatory regions, including transcription start sites (TSS) and enhancers[Simon]. Additional methods to identify open chromatin, including DNase-seq, have been performed on samples obtained from genotyped CD patients; these previous studies indicate that CD-associated SNPs identified by GWAS often occur in regions of open chromatin[Maurano]. We therefore hypothesize that causative CD-associated SNPs occur in gene regulatory elements and can be identified using FAIRE-seq as changes in chromatin organization, leading to more accessible chromatin that can be readily accessed by transcription factors and influence gene expression. In response to environmental cues such as the enteric microbiota, these SNPs may directly mediate chronic intestinal inflammation through the transcriptional regulation of gene products, such as pro- and anti-inflammatory cytokines.

Testing the potential regulatory function of CD-associated SNPs

In the present study, I tested the potential regulatory function of a CD-associated SNP identified by GWAS: rs10896788. Our genotype and FAIRE-seq data previously obtained from uninflamed resected colon tissue from CD patients reveal that many CD-associated SNPs overlap with open chromatin regions identified by FAIRE, suggesting that these SNPs occur in regions

of open chromatin. One of these SNPs is rs1089677, which occurs at position q12.1 on chromosome 11. Luciferase reporter-based cell assays were used to study the potential regulatory function of the region containing this SNP. These assays measure the effect of a regulatory element on the transcription of a reporter gene (*LUC*) whose product catalyzes a bioluminescent reaction. In the general protocol for this assay, a SNP is inserted into the multiple cloning site of an expression vector (pGL4.23). Transformation of the plasmid containing the DNA insert into DH5α *E.coli* cells followed by heat shock allows the vector to be incorporated into the bacterial genome. Transfection of the replicated insert as well as a *Renilla* vector that constitutively transcribes luciferase into cells provides an internal control. Subsequent measurement of luminosity reveals the effect of the DNA variant on gene expression normalized to *Renilla*.

The aim of this study was to identify the effect of the rs10896788 variant on gene expression and determine its potential role in gene regulation and disease etiology. Understanding the function of disease-associated SNPs identified by GWAS is important as genetic factors are increasingly found to play a role in the development of CD and other chronic illnesses. Identifying SNPs that upregulate or downregulate the expression of coding sequences in the genome implies that when induced into a nucleosome-free state, these regulatory regions may increase the synthesis of pro-inflammatory cytokines or decrease the synthesis of anti-inflammatory cytokines. The potential for future screening efforts and targeted therapies for CD will depend on an understanding of the function of genetic variants in health and disease.

MATERIALS AND METHODS

Selection of SNP for functional study

The variant rs10896788 was chosen for functional study due to both its association with CD by GWAS and location in a nucleosome-free region in the genome, as identified by a

previous FAIRE-seq experiment performed on uninflamed resected colon tissue from CD patients recruited at UNC Hospitals. To identify the presence of this SNP in CD patients, SNP genotyping was performed at the UNC Genomics Core on uninflamed resected colon tissue obtained from these CD patients. A sample ID of 42200 was given to the patient homozygous for the disease-associated SNP (G). A sample ID of 44000 was given to the patient homozygous for the non-risk allele (A) which served as a negative control.

Table 1. Correlation between genotype and FAIRE-identified open chromatin regions.

Patient ID	Genotype	RPKM	FAIRE signal
40500	2	5.773738	48
40700	1	2.895777	65
41300	2	9.680131	84
42200	2	5.257381	47
43100	1	13.65219	63
43300	2	1.301694	88
44000	0	2.064365	No peak
45000	1	5.065185	42

Table 1. **Presence of rs10896788 correlates with more open chromatin.** Genotype 0: 0 copies of SNP of interest (rs10896788); Genotype 1: 1 copy of SNP; Genotype 2: 2 copies of SNP. RNA-seq reads per kilobase per million (RPKM) corresponds to levels of *LPXN* expression associated with each genotype. Higher values for FAIRE signal correspond to more open chromatin at the location of the SNP. Genotype, FAIRE, and RNA-seq data were collected by Adam Robinson and Chelsea Raulerson.

In Table 1, RNA-seq reads per kilobase per million (RPKM) values reflect expression values of the gene closest to this locus in hg19, *LPXN*. Patient 44000 lacked a FAIRE signal at the location of the SNP and demonstrated a lower RPKM value (2.064365), suggesting that the default state of chromatin at this locus with two non-risk alleles is closed with lower levels of *LPXN*

expression. In contrast, patient 42200 and other patients who possessed two copies of the SNP demonstrated a relatively high FAIRE signal at this locus with two copies of the SNP, and a higher RPKM value (5.257381) indicative of higher levels of *LPXN* expression. DNA from patient 42200 was chosen for use in luciferase assays over the DNA from the other patients that possessed two copies of the SNP due to the fact that more DNA was available from this patient. The difference in open chromatin between these two patients is displayed in Figure 1.

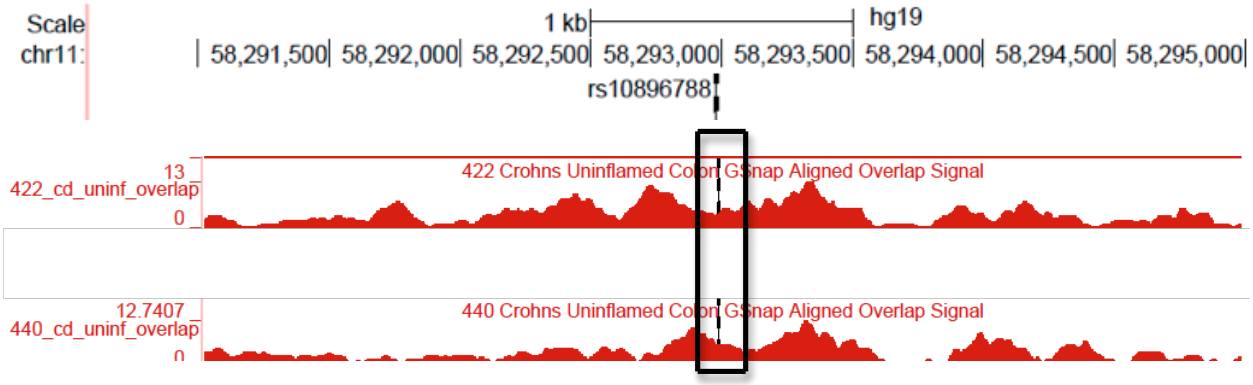

Figure 1. **FAIRE-seq demonstrates a difference in open chromatin signal.** FAIRE peaks shown in red correspond to regions of open chromatin. Patient 44000, with both non-risk alleles, lacks the FAIRE peak indicative of open chromatin at the location of rs10896788 on chromsome 11 (outlined in black). Patient 42200, with two copies of this SNP, demonstrates a FAIRE peak at this locus. FAIRE-seq data and the resulting figure were produced by Matt Weiser.

Cell culture

In preparation for the transfection, one adherent hepatocellular carcinoma cell line, human-derived HepG2, was maintained at 37^{o}C with 5% CO_2. The cells were cultured in 1X MEM-alpha (Gibco®, Life Technologies) with 1 mM sodium pyruvate supplemented with 10% FBS. The medium was discarded and replaced every 2-3 days.

Preparation of DNA sequence containing SNP for cloning

The DNA sequence (~200 base pairs) surrounding rs10896788 from both patient 42200 and patient 44000 was amplified by PCR using designed primers 30 base pairs in length that contained restriction sites for Kpn1 and Xho1. The primers were reconstituted to 100 μM in

sterile H_2O. A gradient PCR was run to select the best annealing temperature for amplification of the desired fragments. After optimizing the annealing temperature, PCR was run on the patients' DNA samples. A 1.5% agarose gel was then run using the amplified samples. The bands containing the remaining PCR product were then cut from the gel and proceeded to gel purification using the Promega Wizard SV Gel and PCR CleanUp System (Cat# A9282). The purified products were resuspended in 30 μL of sterile H_2O and the concentration of DNA remaining in each sample was read using a NanoDrop™. The amplified sequences and enhancer-specific vector (pGL4.23) were then digested in separate tubes with the restriction enzymes Kpn1 and Xho1 and incubated at 37°C for 4 hours to make the desired sequences and vector compatible for cloning. The restricted products were then purified using the Promega Wizard SV Gel and PCR CleanUp System (Cat# A9282), and the concentration of the purified products was read using a NanoDrop™. To ligate the digested insert and vector, 100 ng of vector and 65 ng of insert were incubated at room temperature for 2 hours with T4 ligase (Thermo Scientific). The ligated sample was then heated at 65°C for 10 minutes to inactivate the T4 ligase. To produce thousands of copies of the ligated vector and insert, the ligation mixture was added to MAX Efficiency® DH5α™ Competent Cells (Invitrogen 18258012). The plasmid cloning vector pUC19 (Invitrogen) was used as an internal control. The bacterial cells were incubated on ice for 20 minutes and then underwent heat shock at 42°C for 55 seconds. The cells were then incubated on ice for 5 minutes before adding 250 μL S.O.C. medium and shaking at 37°C for 1 hour. The cells containing the plasmid were then spread on agar plates and incubated at 37°C for 24 hours. The next day, 3 colonies were picked from each plate and added to 14 mL tubes containing 5 mL of LB broth with ampicillin. The tubes were shaken for 20 hours at 37°C. A QIAGEN MiniPrep Kit was then used to extract DNA from the colonies, and the plasmid concentration was

measured using a NanoDrop™. A diagnostic restriction test was then performed to ensure that the plasmids contained the insert containing the SNP. Plasmid DNA was digested with Kpn1 and Xho1 and incubated at 37°C for 2 hours. A 1% agarose gel was then run to test for inserts in each restricted sample. Samples that produced a band at the expected location of the insert (~200 base pairs) were sequenced at the UNC Genome Analysis Facility. The sequences were analyzed using Sequencher® to ensure that the SNP was correct and that the insert did not contain any mutations.

Transfection and luciferase reporter assays

To test the effect of rs10896788 on luciferase production in HepG2 cells, the SNP was cloned in the forward orientation into a minimal promoter-containing firefly luciferase vector pGL4.23 (Promega, Madison, WI). The plasmids for both alleles (G and A) were transfected in duplicate (~700 ng/well) into the HepG2 cell line. Approximately 70,000 cells per well were transfected with both luciferase constructs and the *Renilla*-luciferase control vector using Lipofectamine® 2000 in a 24-cell plate. Cells were assayed 48 hours after transfection using the Dual Luciferase Assay (Promega). The plates were then read using a luminometer with Luciferase Assay Reagent and Stop and Glo Buffer in separate injectors. A two-sided *t*-test was used to compare luciferase activity and normalize the values to *Renilla* luciferase activity.

RESULTS

Optimal annealing temperature determined for PCR

A gradient PCR was run on DNA that was not obtained from patients 42200 and 44000 at six different annealing temperatures, using both the forward and reverse primers at each temperature. A 1% agarose gel was then run to identify at which temperature the greatest concentration of DNA appeared on the imaged gel (Fig. 2). Based on the fact that the brightest

bands for both the forward and reverse orientations appeared at 62°C, this was the temperature chosen for PCR later run on DNA from patients 42200 and 44000.

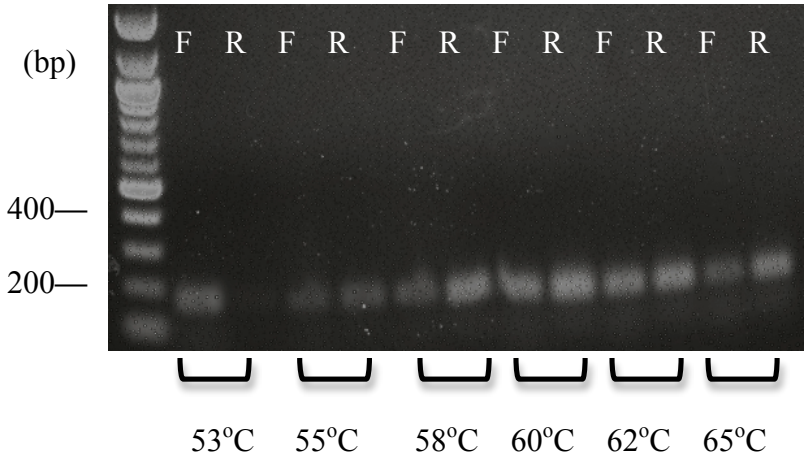

Figure 2. **Nucleic acid gel obtained for gradient PCR performed on junk DNA.** Forward and reverse primer products are represented by F and R, respectively. A 100 base-pair ladder was used to confirm the size of the PCR products (~200 base pairs).

Potential enhancer-specific activity of rs10896788

Each plasmid was transfected into HepG2 cells in duplicate with three separate plasmids for each allele: the *Renilla* plasmid, the plasmid containing the rs10896788 SNP (positive control), and the plasmid containing the non-risk allele (negative control). Relative *LUC* expression normalized to the empty vector (*Renilla* control) was read using a luminometer. GraphPad Prism® analysis indicated that the plasmid containing the risk allele (G) increased *LUC* transcription compared to the control allele (A) (Fig. 3). However, it is inconclusive from this analysis whether rs10896788 increases overall *LUC* transcription in HepG2 cells.

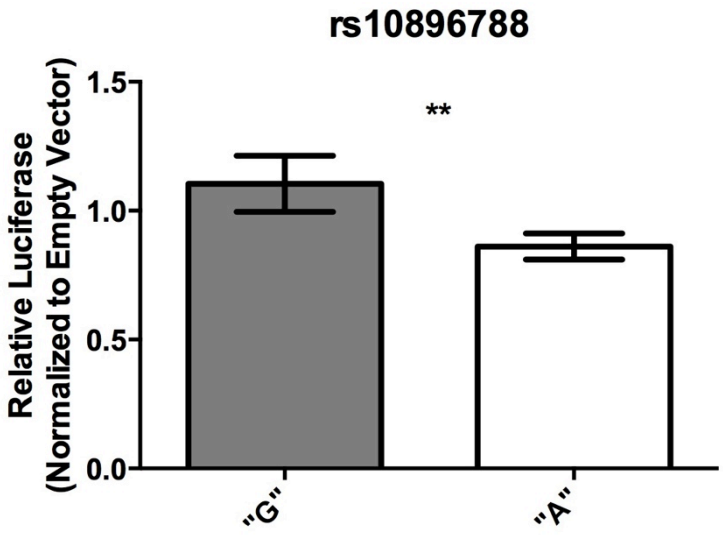

Figure 3. **rs10896788 alleles exhibit differential transcriptional activity in pGL4.23.** Enhancer activity in the forward orientation with respect to *LUC* was studied in HepG2 cells with the CD-associated allele (G) and the non-risk allele (A). Statistically significant allele-specific enhancer activity was observed. The risk allele G displays greater transcriptional activity in the forward orientation than the non-risk allele A with respect to *LUC* in a minimal promoter vector (pgL4.23). Error bars represent SD of 5 independent clones for each allele. Results are expressed as a fold change normalized to the empty vector control. P value was calculated using a two-sided *t*-test. $P = 1.9 \times 10^{-3}$.

Evidence that rs10896788 does not function as a promoter

The cloning protocol was repeated using a basic vector not containing a promoter region (pGL4.10) instead of the minimal promoter vector designed to confirm enhancer-specific activity (pGL4.23). Each plasmid was again transfected into HepG2 cells in duplicate with three separate plasmids for each allele: the *Renilla* plasmid, the plasmid containing the risk allele (G), and the plasmid containing the non-risk allele (A). Relative *LUC* expression normalized to the empty vector (*Renilla* control) was read using a luminometer. GraphPad Prism® analysis indicated that the plasmid containing the risk allele (G) demonstrated similar *LUC* transcription compared to the control allele (A). However, these results were not statistically significant and the error bars displayed significant overlap (Fig.4).

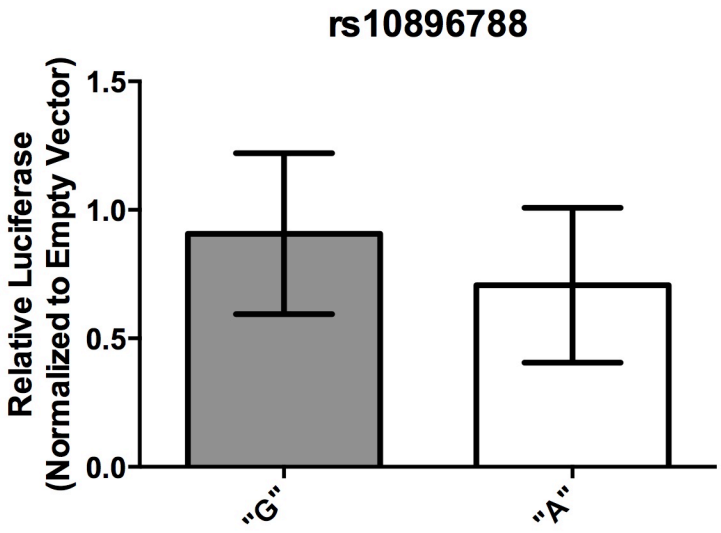

Figure 4. **rs10896788 alleles exhibit similar transcriptional activity in pGL4.10.** Promoter activity in the forward orientation with respect to *LUC* was studied in HepG2 cells with the CD-associated allele (G) and the non-risk allele (A). No significant difference in transcriptional activity was observed with respect to *LUC* in a basic vector lacking a promoter (pgL4.10). Error bars represent SD of 5 independent clones for each allele. Results are expressed as a fold change normalized to the empty vector control. $P >> 0.05$.

Upon examining the values read for the *Renilla* vector by the luminometer, it was discovered that the empty vector yielded higher values for *LUC* expression than the *Renilla* vector. This finding is inconsistent with the fact that the empty vector should yield similar *LUC* expression values as the *Renilla* vector, as neither contain a regulatory element. Based on these considerations, the analysis was performed again by normalizing all luciferase readings from pGL4.10 to the average of the empty vector values. This analysis indicated that there was no significant difference in transcriptional activity between the risk allele (G) and non-risk allele (A), but the results were not statistically significant with widely overlapping error bars (Fig. 5).

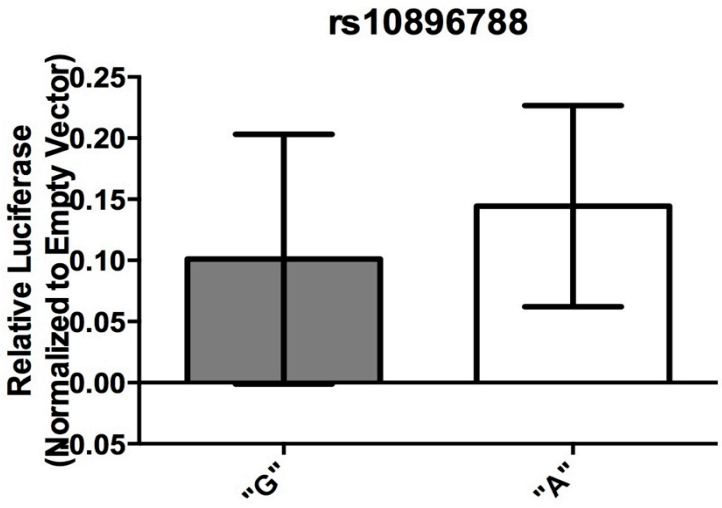

Figure 5. **rs10896788 alleles exhibit similar transcriptional activity in pGL4.10.** Promoter activity in the forward orientation with respect to *LUC* was studied in HepG2 cells with the CD-associated allele (G) and the non-risk allele (A). No significant difference in transcriptional activity was observed with respect to a basic vector with no promoter (pgL4.10). Error bars represent SD of 5 independent clones for each allele. Results are expressed as a fold change normalized to the empty vector control. $P >> 0.05$.

DISCUSSION

Here we show that the region containing the CD-associated SNP rs10896788 has a potential role as a regulatory element that may increase gene expression in liver cells. In the HepG2 cell line, this variant displayed an increase in gene expression compared to the non-risk allele when cloned into a minimal promoter vector (pGL4.23). We also show that this variant displayed no significant difference in transcriptional activity compared to the non-risk allele when cloned into a promoter-specific vector. These results suggest that the presence of the G allele at this locus may be associated with a change in chromatin accessibility, allowing this region to act as a regulatory element; specifically, it may act as an enhancer.

Our hypothesis at the beginning of this study was that CD-associated SNPs found in regions of open chromatin may act as regulatory elements that contribute to CD pathogenesis by affecting the production of pro- or anti-inflammatory cytokines, contributing to the imbalance in

pro-and anti-inflammatory gene expression characteristic of the chronic inflammation found in CD. Confirming the fact that rs10896788 occurs in a region of open chromatin using FAIRE-seq, while the non-risk allele was associated with closed chromatin, was the first indication that this variant is associated with a change in chromatin organization. Our suggestion that this variant may affect gene expression was also supported by our observations using luciferase assays.

Although preliminary, these results are significant in that they propose an alternative explanation for CD etiology that incorporates the role of genetic variation in the development and varying levels of severity of this chronic disease. When SNPs occur in protein-coding regions of the genome, their contribution to disease can be readily observed as changes in protein structure or function. For example, sickle cell anemia results from a mutation in a single gene that changes the conformation of hemoglobin. The fact that the majority of SNPs associated with CD occur in noncoding regions of the genome suggests that gene regulation and chromatin accessibility contribute to disease pathogenesis in a manner distinct from many other chronic illnesses[Natoli]. If it is found that certain genetic variants associated with CD have a regulatory function that serves to increase or decrease cytokine production, more targeted immunotherapies and genetic screening programs may be developed that can serve to personalize CD treatment.

While our finding that the CD-associated allele at this locus increases *LUC* transcription relative to the non-risk allele is statistically significant, there are several alternative interpretations of these results that should be addressed. It cannot be concluded that rs10896788 increases overall luciferase production, as the relative luciferase values were similar between the empty vector and positive control as determined using a two-sided *t*-test. To further investigate whether rs10896788 acts as an enhancer element, the reverse insertion will need to be tested in a similar luciferase assay, as only the forward insertion was tested here. Another consideration is

the fact that in the human genome, this locus occurs upstream of the *LPXN* gene. This gene encodes leupaxin, a protein preferentially expressed in hematopoietic cells. It cannot be concluded that rs10896788 increases *LPXN* transcription, as only tested *LUC* transcription was tested here. Further, RNA-seq reads per kilobase per million indicated relatively low levels of *LPXN* expression in intestinal tissue, even in the patient with both copies of the SNP (Table 1). In addition, our study contained several limitations due to the difficulty of replicating a sequence found in human cells into a luciferase reporter assay. Primarily, the HepG2 cell line used in these experiments are liver cells, while macrophages are the disease-relevant cell type in CD.

To address these limitations and further elucidate the role of this potential variant in CD pathogenesis, future directions will involve transfection of this variant into the disease-relevant cell type, specifically a macrophage cell line. The effect of rs10896788 on transcriptional activity when transfected into a macrophage cell line (such as J774) will confirm whether this variant demonstrates greater *LUC* transcription compared to the non-risk allele. Additional dual-luciferase reporter assays will be performed in J774 cells transfected with plasmids containing different CD-associated SNPs to investigate which SNPs are causal with respect to CD and which SNPs are marker CD variants.

ACKNOWLEDGEMENTS

I gratefully acknowledge Dr. Shehzad Sheikh, Department of Medicine and Genetics, UNC-Chapel Hill; Dr. Kelly Hogan, Department of Biology, UNC-Chapel Hill; and Gregory Gipson, Adam Robinson, and Eric Lee, Center for Gastrointestinal Biology and Disease, UNC-Chapel Hill, for their helpful advice and guidance. I appreciate the helpful support of Maren Cannon, Chelsea Raulerson, Dr. Jeremy Simon, and Matthew Weiser, Department of Genetics, UNC-Chapel Hill. I acknowledge funding obtained from the Summer Undergraduate Research Fellowship from the Office for Undergraduate Research at UNC-Chapel Hill.

LITERATURE CITED

1. Dionne S, D'Agata DD, Hiscott J, Vanounou T, Seidman EG. Colonic explant production of IL-1 and its receptor antagonist is imbalanced in inflammatory bowel disease (IBD). *Clin*

Exp Immunol 1998 Jun; 112(3):435-442.

2. Rogler G, Andus T. Cytokines in inflammatory bowel disease. *World Journal of Surgery*, 1998 April; 22(4):382-389.

3. Dixon LJ, Kabi A, Nickerson KP, McDonald C. Combinatorial effects of diet and genetics on inflammatory bowel disease pathogenesis. *Inflamm Bowel Dis* 2015 Jan 9. [Epub ahead of print].

4. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, Kaplan GG. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*, 2012 Jan;142(1):46-54.

5. Jostins, L., et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119-124 (2012).

6. Mirza AH, Kaur S, Brorsson CA, Pociot F. Effects of GWAS-associated genetic variants on lncRNAs within IBD and T1D candidate loci. *PLoS One* 2014 Aug 21. DOI: 10.1371/journal.pone.0105723.

7. Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* 2012 Jan 19;7(2):256-67.

8. Maurano, M.T., et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012;337:1190-1195.

9. Natoli, G. Specialized chromatin patterns in the control of inflammatory gene expression. *Current topics in microbiology and immunology* 349, 61-72 (2011).