**Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn’s disease**

Dermot P.B. McGovern¹,²,*, Michelle R. Jones³, Kent D. Taylor², Kristin Marcianè⁴, Xiaofei Yan², Marla Dubinsky¹, Andy Ippoliti¹, Eric Vasiliaskas¹, Dror Berel¹, Carrie Derkowski¹, Deb Dutridge², International IBD Genetics Consortium, Phil Fleshner¹, David Q. Shih¹, Gil Melmed¹, Emebet Mengesha², Lily King², Sheila Pressman², Talin Haritunians², Xiuqing Guo², Stephan R. Targan¹ and Jerome I. Rotter²

¹Inflammatory Bowel and Immunobiology Research Institute, ²Medical Genetics Institute and ³Endocrinology, Diabetes & Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA, USA and ⁴Cardiovascular Health Research Unit, Department of Internal Medicine, University of Washington, Seattle, WA, USA

Received January 19, 2010; Revised June 3, 2010; Accepted June 10, 2010

Genetic variation in both innate and adaptive immune systems is associated with Crohn’s disease (CD) susceptibility, but much of the heritability to CD remains unknown. We performed a genome-wide association study (GWAS) in 896 CD cases and 3204 healthy controls all of Caucasian origin as defined by multidimensional scaling. We found supportive evidence for 21 out of 40 CD loci identified in a recent CD GWAS meta-analysis, including two loci which had only nominally achieved replication (rs4807569, 19p13; rs991804, CCL2/CCL7). In addition, we identified associations with genes involved in tight junctions/epithelial integrity (ASHL, ARPC1A), innate immunity (EXOC2), dendritic cell biology [CADM1 (IGSF4)], macrophage development (MMD2), TGF-β signaling (MAP3K7IP1) and FUT2 (a physiological trait that regulates gastrointestinal mucosal expression of blood group A and B antigens) (rs602662, P = 3.4 x 10⁻⁵). Twenty percent of Caucasians are ‘non-secretors’ who do not express ABO antigens in saliva as a result of the FUT2 W134X allele. We demonstrated replication in an independent cohort of 1174 CD cases and 357 controls between the four primary FUT2 single nucleotide polymorphisms (SNPs) and CD (rs602662, combined P-value 4.90 x 10⁻⁸) and also association with FUT2 W143X (P = 2.6 x 10⁻⁵). Further evidence of the relevance of this locus to CD pathogenesis was demonstrated by the association of the original four SNPs and CD in the recently published CD GWAS meta-analysis (rs602662, P = 0.001). These findings strongly implicate this locus in CD susceptibility and highlight the role of the mucus layer in the development of CD.

**INTRODUCTION**

Crohn’s disease (CD), one of the major forms of the inflammatory bowel diseases (IBDs), is a chronic, debilitating disease characterized by recurrent gastrointestinal (GI) inflammation, postulated to occur as a result of an abnormal immune reaction to commensal flora in genetically susceptible individuals. The role of commensal flora in precipitating chronic GI mucosal inflammation is substantiated by data from established rodent models of IBD such as the Il10⁻/⁻ mouse and the Hla-B27 transgenic rat that are disease free when housed in germ-free environments, but develop inflammation when raised under pathogen-free conditions (1,2). Furthermore, in both of these models, the bacterial load and the nature of the commensal flora can influence both the site and the degree of GI inflammation (1,3,4). In addition, in human disease, both antibiotic and probiotic therapy can be effective in modifying some of the manifestations of IBD (5,6), and our group and others have had a long-standing interest in serological responses to commensal flora and their association with CD (7).

Through utilizing genome-wide association studies (GWASs), in addition to linkage informed positional candidate gene

---

*To whom correspondence should be addressed at: Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, 8730 Alden Drive, Los Angeles, CA 90048, USA. Tel: +1 3104234100; Fax: +1 3104230146; Email: mcgovern@cshs.org

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
 approaches, there has been considerable success in identifying CD susceptibility loci in populations of Northern European origin (8–14). To date, more than 30 loci are definitively known to be associated with CD, although these loci account only for a minority of the genetic variance to CD in this population (15). A number of the CD susceptibility genes encode important components of the innate immune system genes such as NOD2/CARD15 (12,13), the Toll-like receptors (16,17) and the autophagy genes ATG16L1 (9) and IRGM (18), further emphasizing the importance of the microbial–host interaction in the development of CD. Our group and others have identified antibodies to bacterial antigens that define certain sub-groups of CD patients reinforcing the essential role that bacteria play in ‘driving’ CD (19).

We report herein the findings of a CD GWAS identifying a number of putative associations with CD. Given our group’s long-standing interest in the host–microbial interaction, we were particularly interested in the CD association with four Fucosyltransferase 2 (FUT2) single nucleotide polymorphisms (SNPs), particularly as genetic variation in FUT2 has been implicated in susceptibility to infections including Helicobacter pylori (20), Norovirus (Norwalk virus) (21–23) and progression of HIV (24). FUT2 alleles have also been associated with circulating serum vitamin B12 levels (25). FUT2 is a physiological trait that regulates the expression of the H antigen, a precursor of the blood group A and B antigens, on the GI mucosa. Approximately 20% of Caucasians are non-secretors (Se−) who do not express ABO antigens in saliva as they are homozygous for FUT2 null alleles (26). In addition to the genetic associations mentioned above, non-secretion of ABO blood group antigens into body fluids has been shown to be associated with duodenal ulceration (27), the development of oral candidiasis (28,29), rheumatic fever (30), recurrent urinary tract infection (31), cholera (32) and infection with meningococcus (33), pneumococcus (33) and haemophilus influenzae (34). These data taken together implicate FUT2 as an obvious gene of interest in IBD pathogenesis. Furthermore, our genome-wide association scan reported herein demonstrated four FUT2 SNPs with association with CD (P > 4 × 10−5), including a non-synonymous (Ser258Gly) polymorphism. The data presented herein indicate an association between the non-secretor status associated FUT2 genotype and CD.

**RESULTS**

A CD GWAS meta-analysis previously identified or confirmed association with 30 loci and demonstrated nominal association with a further 10 loci (15). We were able to confirm association (uncorrected P < 0.05 and association with the previously identified risk allele or appropriate proxy with association in the same direction) with 21 of these loci in our GWAS (Table 1) and these loci served as an internal control for our data set and as indicators of the relative power of the discovery phase of the study (see Materials and Methods). Three of these loci were from the nominally replicated list of SNPs (rs4807569, 19p13; rs991804, 17q12, CCL2, CCL7; rs917997, 2q11, IL18RAP) from the CD meta-analysis, and the data presented in Table 1 therefore provide further evidence of their relevance in CD susceptibility. The IL18RAP association has previously been confirmed by others (35). In addition to these loci, we also demonstrated the association between another locus implicated in both CD and UC namely CARD9 (35,36) [a total of five associated

<table>
<thead>
<tr>
<th>Known CD SNP</th>
<th>Chr.#</th>
<th>Position</th>
<th>Minor allele</th>
<th>OR (95%CI)</th>
<th>P-value</th>
<th>Proxy marker for known CD SNP (D′)</th>
<th>Gene(s) of interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1465804</td>
<td>1q31</td>
<td>67 475 114</td>
<td>G</td>
<td>0.53 (0.41–0.68)</td>
<td>1.02 × 10−6</td>
<td>Same SNP</td>
<td>IL23R</td>
</tr>
<tr>
<td>rs9268789</td>
<td>1q24</td>
<td>171 128 857</td>
<td>G</td>
<td>1.21 (1.07–1.37)</td>
<td>0.002</td>
<td>Same SNP</td>
<td>Gene desert</td>
</tr>
<tr>
<td>rs1154833</td>
<td>1q22</td>
<td>199 202 489</td>
<td>C</td>
<td>0.88 (0.78–0.99)</td>
<td>0.047</td>
<td>rs3767498 (0.96)</td>
<td>CACNA1S, KIF21B</td>
</tr>
<tr>
<td>rs917997</td>
<td>2q11</td>
<td>102 437 000</td>
<td>T</td>
<td>1.19 (1.05–1.35)</td>
<td>0.001</td>
<td>rs3105714 (1.0)</td>
<td>IL18RAP</td>
</tr>
<tr>
<td>rs2241880</td>
<td>2q37</td>
<td>233 848 107</td>
<td>A</td>
<td>0.76 (0.68–0.85)</td>
<td>3.29 × 10−6</td>
<td>Same SNP</td>
<td>ATG16L1</td>
</tr>
<tr>
<td>rs3197999</td>
<td>3p21</td>
<td>49 696 536</td>
<td>A</td>
<td>1.25 (1.11–1.40)</td>
<td>1.8 × 10−4</td>
<td>rs6425 (1.0)</td>
<td>MST1</td>
</tr>
<tr>
<td>rs4613763</td>
<td>5p13</td>
<td>40 428 485</td>
<td>C</td>
<td>1.31 (1.11–1.54)</td>
<td>0.001</td>
<td>Same SNP</td>
<td>Gene desert (PTGER4)*</td>
</tr>
<tr>
<td>rs2188962</td>
<td>5q31</td>
<td>131 798 704</td>
<td>T</td>
<td>1.36 (1.21–1.52)</td>
<td>1.37 × 10−7</td>
<td>Same SNP</td>
<td>IBDS locus</td>
</tr>
<tr>
<td>rs10045341</td>
<td>5q35</td>
<td>158 747 111</td>
<td>A</td>
<td>0.69 (0.61–0.79)</td>
<td>7.12 × 10−8</td>
<td>Same SNP</td>
<td>IL12B</td>
</tr>
<tr>
<td>rs6908425</td>
<td>6p22</td>
<td>20 836 710</td>
<td>T</td>
<td>0.81 (0.70–0.93)</td>
<td>0.004</td>
<td>Same SNP</td>
<td>CDKAL1</td>
</tr>
<tr>
<td>rs2301436</td>
<td>6q27</td>
<td>167 357 978</td>
<td>T</td>
<td>1.37 (1.22–1.53)</td>
<td>5.92 × 10−8</td>
<td>Same SNP</td>
<td>CCR6, FGFR10P, RNASE2</td>
</tr>
<tr>
<td>rs10758669</td>
<td>9p24</td>
<td>4 971 602</td>
<td>C</td>
<td>1.15 (1.03–1.29)</td>
<td>0.016</td>
<td>Same SNP</td>
<td>JAK2</td>
</tr>
<tr>
<td>rs11190140</td>
<td>10q24</td>
<td>101 281 583</td>
<td>G</td>
<td>0.79 (0.66–0.71)</td>
<td>5.3 × 10−5</td>
<td>Same SNP</td>
<td>NRR2−3</td>
</tr>
<tr>
<td>rs7927894</td>
<td>11q13</td>
<td>75 978 964</td>
<td>T</td>
<td>1.20 (1.07–1.34)</td>
<td>0.0012</td>
<td>rs7130588 (1.0)</td>
<td>C11orf30</td>
</tr>
<tr>
<td>rs2066847</td>
<td>16q12</td>
<td>49 321 279</td>
<td>T</td>
<td>1.62 (1.42–1.86)</td>
<td>2.15 × 10−5</td>
<td>rs5743289 (26)</td>
<td>NOD2b</td>
</tr>
<tr>
<td>rs991804</td>
<td>19q12</td>
<td>29 611 383</td>
<td>A</td>
<td>0.78 (0.68–0.89)</td>
<td>0.002</td>
<td>Same SNP</td>
<td>CCL2, CCL7</td>
</tr>
<tr>
<td>rs2872507</td>
<td>19q21</td>
<td>35 294 289</td>
<td>A</td>
<td>1.17 (1.04–1.31)</td>
<td>0.006</td>
<td>rs907092 (1.0)</td>
<td>ORM1D3</td>
</tr>
<tr>
<td>rs744166</td>
<td>17q21</td>
<td>37 767 727</td>
<td>T</td>
<td>0.86 (0.76–0.96)</td>
<td>0.010</td>
<td>Same SNP</td>
<td>STAF3</td>
</tr>
<tr>
<td>rs2542152</td>
<td>18p11</td>
<td>12 769 947</td>
<td>G</td>
<td>1.22 (1.05–1.43)</td>
<td>0.009</td>
<td>rs1893217 (1.0)</td>
<td>PTPN2</td>
</tr>
<tr>
<td>rs4807569</td>
<td>19p13</td>
<td>1 074 378</td>
<td>C</td>
<td>1.25 (1.09–1.43)</td>
<td>0.001</td>
<td>rs2024092 (1.0)</td>
<td>STK11, SNBO2, 729119</td>
</tr>
<tr>
<td>rs1736135</td>
<td>21q11</td>
<td>15 727 091</td>
<td>C</td>
<td>0.84 (0.75–0.94)</td>
<td>0.003</td>
<td>rs1736148 (1.0)</td>
<td>—</td>
</tr>
</tbody>
</table>

Chr. #, chromosome number. Chromosomal position according to db129 (Build 36). P-value is from the current GWAS.

*PTGER4 is outside the region but has been implicated by eQTL analysis.

No proxy found for specific NOD2 SNP cited in Barrett et al.—alternative SNP used in view of widespread replication of NOD2/CD association.
SNPs $\leq 1 \times 10^{-4}$, including a non-synonymous N12S SNP ($P = 1.4 \times 10^{-6}$).

We did not demonstrate association ($P < 0.05$) with CD and the other 19 loci identified in the GWAS meta-analysis, including 5q33 (IRGM) (no proxy), 9q32 (TNFSF15) (no proxy), 10p11 [proxy SNP (D’ 1.0) $P = 0.07$], 10q21 (ZNFX2) (no proxy), 12q12 (SLC2A13, LRRK2) (no proxy), 13q14 (rs3764147) (no proxy), 1p13 (PTPN22) ($P = 0.43$), 6q21 (PRDM1) (no proxy), 8q24 (no proxy), 1q23 (ITLN1, CD24) (no proxy), 6p25 (LYM4) (no proxy), 2p16 (PUS10) (no proxy), 6p25 (SLC2A243) ($P = 0.051$), 6p25 ($P = 0.97$), 2p23 (GKR) ($P = 0.25$), 7p12 ($P = 0.40$), 21q22 (ICOSLG) (no proxy), 18q11 (rs8098673) ($P = 0.07$) and 6p21 (BTNL2, DRA, DQA, DRB) [proxy SNP (D’ 0.93) $P = 0.15$].

In addition, we identified associations between CD and a number of putative loci (see Supplementary Material, Table S1). These include genes involved in tight junctions/epithelial integrity [ASHL on 1q22 ($4.3 \times 10^{-5}$)], Wnt and JNK signaling [RHOU on 1q42 ($4.2 \times 10^{-5}$)], substance P signaling [TACR3 on 4q25 ($5.4 \times 10^{-5}$)], innate immunity [EXOC2 on 6p25 ($5.6 \times 10^{-5}$)] (37), dendritic cell biology [CADM1 (IGSF4) on 11q23 ($2.3 \times 10^{-5}$)], macrophage development [MMD2 on 7p22 ($2.8 \times 10^{-5}$)], asthma susceptibility [DENND1B on 1q31 ($4.7 \times 10^{-5}$)] (38,39), integrin regulation [ACER2 on 9p22 ($6.5 \times 10^{-5}$)], and TGF-β signaling [MAP3K7IP1 on 22q13 ($6.0 \times 10^{-5}$)]. We also identified two CD-associated loci specifically involved in the host–microbial interaction, namely SPG20 on 13q13 ($3.92 \times 10^{-5}$) (endosomal trafficking) and FUT2 on 19p13 ($3.44 \times 10^{-5}$). An expression study previously utilized to implicate PTGER4 in CD pathogenesis (14,40) indicated a strong cis effect in three of the putative loci that we identified, including FDPS (effect of rs11264359A LOD score 6.6, $P = 3.3 \times 10^{-8}$), SPG20 (effect of rs912927A LOD score 7.2, $P = 7.6 \times 10^{-8}$) and SMHIT (effect of rs8089066C LOD score 4.2, $P = 1.1 \times 10^{-5}$) (Supplementary Material, Table S1). The overall GWA results are summarized in the form of a Manhattan plot (Supplementary Material, Fig. S3).

Independent replication of FUT2

From our candidate associations, we chose FUT2 as the leading gene for independent replication, given our group’s interest in the host–microbial interaction in CD pathogenesis and FUT2’s known association with a number of infective processes and GI diseases. Furthermore, FUT2 is located under a previously identified peak of linkage for CD on chromosome 19 (41) and there were a total of four SNPs (one of which is non-synonymous) with strong association to CD in our GWAS (see Table 2 and Fig. 1). In addition to these four SNPs (rs504963: 3’ UTR; rs676388: 3’ UTR; rs485186: synonymous exon 2 SNP; rs602662: S258G) from the GWAS, we also genotyped rs492602 (synonymous exon 2) and rs601338 (W143X), the common null allele in Caucasians associated with the ABO non-secretory phenotype, in the independent confirmatory cohort. We were able to replicate the initial association with the four SNPs from the discovery cohort, as well as demonstrate association with the additional two SNPs, including the allele for non-secretor status. Further analysis of the FUT2 association seen in the replication

<table>
<thead>
<tr>
<th>SNP</th>
<th>GWAS</th>
<th>P-value for association in GWAS</th>
<th>Replication</th>
<th>P-value for association in replication</th>
<th>Combined P-value in replication</th>
<th>H. W. Hardy–Weinberg analysis using logistic regression in the replication cohort</th>
<th>Combined P-value calculated using logistic regression in the replication cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs492602 Exon 2</td>
<td>G</td>
<td>0.003</td>
<td>0.407</td>
<td>27.464</td>
<td>0.017</td>
<td>16.4863</td>
<td>0.99</td>
</tr>
<tr>
<td>rs492602 Exon 2</td>
<td>A</td>
<td>0.005</td>
<td>0.412</td>
<td>27.740</td>
<td>0.999</td>
<td>16.4863</td>
<td>0.99</td>
</tr>
<tr>
<td>rs492602 Exon 2</td>
<td>G</td>
<td>0.005</td>
<td>0.413</td>
<td>27.740</td>
<td>0.999</td>
<td>16.4863</td>
<td>0.99</td>
</tr>
<tr>
<td>rs492602 Exon 2</td>
<td>A</td>
<td>0.005</td>
<td>0.413</td>
<td>27.740</td>
<td>0.999</td>
<td>16.4863</td>
<td>0.99</td>
</tr>
<tr>
<td>rs492602 Exon 3</td>
<td>G</td>
<td>0.005</td>
<td>0.413</td>
<td>27.740</td>
<td>0.999</td>
<td>16.4863</td>
<td>0.99</td>
</tr>
<tr>
<td>rs492602 Exon 3</td>
<td>A</td>
<td>0.005</td>
<td>0.413</td>
<td>27.740</td>
<td>0.999</td>
<td>16.4863</td>
<td>0.99</td>
</tr>
</tbody>
</table>
cohort reveals deviation from the Hardy–Weinberg equilibrium (HWE) in the cases ($P < 0.05$) but not the controls ($P > 0.84$) (Table 2). The raw genotype frequencies suggest an excess of homozygotes (A allele in rs601338) in cases [325 out of 1174 cases (27.7%)] compared with controls [58 out of 357 (16.2%)] (Table 2). There is no excess of heterozygotes in CD compared with controls and would also be in keeping with the proposed hypothesis that the association seen at this locus is ‘driven’ by an association between non-secretor status and $FUT2$. In addition, we demonstrated no association between $FUT2$ and ulcerative colitis (UC), and there was no evidence of deviation from the HWE in the UC cases (data not shown).

Further evidence for the association between this locus and CD susceptibility is provided by the CD meta-analysis published by Barrett et al. (15) in which all four of the $FUT2$ SNPs highlighted in the GWAS presented here are also associated with CD (Table 2). We combined the $FUT2$ association signals from the study index GWAS and the replication cohort in a meta-analysis comprising in total of our two studies consisting of a total of 2270 CD cases and 4337 controls, achieving the stringent criteria for genome-wide statistical significance ($P > 5 \times 10^{-8}$) for all four SNPs (Table 2 and Fig. 1).

The six SNPs included in the replication study are in strong linkage disequilibrium (Supplementary Material, Fig. S4).

**DISCUSSION**

In this study, we have confirmed the association with a number of known CD loci, provided further evidence for association to CD with two other loci previously only nominally associated with disease (19p13 and 17q12), and identified a number of candidate loci. The region on 19p13 contains $SBNO2$ and
GPX4 (glutathione peroxidase 4). Little is known about SBN02, while GPX4 is known to protect cells against oxidative damage and may have a regulatory role in leukotriene biosynthesis (42). The I7q12 locus is located in a cytokine gene cluster, containing CCL2, CCL8, CCL11 and CCL7 genes. These genes encode Cys–Cys cytokine genes that are involved in immunoregulatory and inflammatory processes and are therefore attractive candidate genes for CD susceptibility. This locus has previously also been implicated in susceptibility to asthma (43), mycobacterial infection (44) as well as with HIV progression (45).

We identified a number of putative loci associated with CD in our population. However, our most intriguing result is the evidence of association with FUT2. We present independent confirmation for association between FUT2 and CD in a distinct replication cohort, and also see association with these four SNPs in the meta-analysis published by Barrett et al. (15). The combined analysis at this locus of both our index study and the replication cohort demonstrate association attaining stringent criteria of genome-wide statistical significance (P < 5.0 × 10⁻⁸). We were particularly interested in this gene given our interest in the host–bacterial interface and the previously documented associations between this gene and infective processes. Furthermore, the association identified herein potentially extends our knowledge regarding the scope of the host–microbial interaction in CD as previously identified genetic associations with CD have highlighted the role of innate (12,13,16,17) and adaptive immune systems (46,47). The data presented here extend this interaction to the mucus layer of the GI tract. FUT2 encodes the secretor type α (1,2) fucosyltransferase (also known as the Se enzyme) that is responsible for regulating the secretion of the ABO antigens in both the digestive mucosa and secretory glands. Approximately 20% of individuals are non-secretors who fail to express ABO antigens in both the GI tract and saliva as a result of being homozygous for non-secretor alleles (26). The prevalence of the non-secretor status (Se-) is similar between populations (48), although the point mutations that lead to Se- differ. The dominant non-secretor polymorphism in Caucasians is the Trp143Ter (W143X) (26) and our detailed analysis lead us to conclude that this polymorphism is the most likely causative SNP at this locus.

Pathogens utilize host cell surface molecules, including oligosaccharides (synthesized by glycosyltransferases), for invasion. It is likely that the high prevalence of non-secretor phenotypes in the population occurs as the absence of particular carbohydrate molecules in the mucosa may have conferred some historical protection to infection as demonstrated with non-secretor status and protection from Helicobacter pylori infection and GI ulceration (20,27). Lactobacilli, a known commensal bacteria, bind to the precursor glycolipid GA1, implying a role of the GI mucosal glycolipid profile in the adherence of commensal and ‘beneficial’ bacteria, in addition to pathogenic organisms (49). Furthermore, Lactobacilli can also displace pathogens such as Clostridium from mucus (50) and also inhibit the Shigella–host interaction (51). Commensal bacteria probably induce glycolipid expression, as the fucosylglycolipid GA1 is found in the small bowel of conventionally bred mice but not in germ-free mice (52). Furthermore, GA1 expression is induced by administration of microbes, and FUT2 transcripts in the ileum were induced in germ-free mice 48 h after administration of feces from conventionally bred mice (53). Fut2-null mice do not express the fucosylglycolipid GA1 in the cecum and colon, whereas normal mice do (52). In the mammalian gut, blocking the CRK and JNK pathways inhibits the ability of bacterial colonization to induce fucosyltransferase activity and FUT2 mRNA expression, both of which are hallmarks of the adult mammalian colon (54). Commensal bacteria and probiotics may exert their protective effects through preventing adherence or even displacing pathogenic bacteria, thus emphasizing the potential role of FUT2 and non-secretor status on GI bacterial profile (55). Se- individuals may thus have a disrupted immunogenic/homeostatic equilibrium that makes them more susceptible to the development of chronic mucosal inflammation. Furthermore, changes in the microflora of IBD patients have been well documented (56). In addition, Fut2 null mice display an increased susceptibility to experimental yeast vaginitis and cervical mucins containing Fut2 are partly protected from induced vaginal candidiasis (57). Mucin 2 (muc-2), the predominant secreted mucin in the colon, plays a key barrier role in intercepting and excluding bacteria from the mucosal cell surface, thereby reducing host susceptibility to colitis and inflammation-associated neoplasia (58–60). Recent genetic studies have clarified the importance of glycan structures of muc-2 protein in these biologic roles (61,62). Both the core 1- and core 3-derived α-glycans of mucin core proteins are terminally fucosylated, which serves as a binding structure for bacterial interception (63). Accordingly, the present findings with FUT2 may represent human genetic evidence linking IBD susceptibility to the functional state of intestinal mucin.

Although FUT2 is a strong candidate gene for CD susceptibility given its tissue expression, its influence on the GI bacterial profile and the mode of inheritance we have observed at this locus, the associations identified herein may reflect association with other genetic variants at this locus in linkage disequilibrium with the described FUT2 SNPs (see Fig. 1). We therefore explored the LD pattern at this locus using the latest version of HapMap (64) and identified that LD (defined as D' > 0.80) extends into neighboring genes, including interesting candidate genes that are also potentially involved in the host–bacterial interaction, such as FUT1 [alpha-1-2-fucosyltransferase 1—FUT, genetic variation in pigs is associated with alterations in Escherichia coli adherence (65)] and RASIP1 (RAS interacting protein 1—an RAS effector localized to the Golgi membranes), as well as DBP (D-site of albumin promoter-binding protein) and FGF21 [fibroblast growth factor 21—involved in insulin sensitivity, adipocyte function and growth hormone signaling (66,67)]. While we believe FUT2 is the most attractive candidate gene at this locus, and we have demonstrated the association with a variant with a known consequence on gene expression, further work will be needed to fully map this locus. We have also identified some candidate loci for further investigation, including genes involved in tight junctions, Substance P signaling, macrophage development, dendritic cell function and NK T-cell function. Further work on these and other loci listed in Supplementary Material, Table S1 will be necessary.

In summary, the data presented here provide strong evidence that non-secretor status increases CD susceptibility. The non-secretor variants from other ethnic groups have...
been well documented, and studies of these variants within the relevant IBD populations will help elucidate the exact role of *FUT2* in CD susceptibility. Studies on the effect of *FUT2* on clinical and serological phenotype, and in particular its role on the microbiome of non-secretor individuals, may help investigators understand further the role of commensal bacteria in CD susceptibility, and also further determine those CD patients who might most benefit from probiotic- or antibiotic-based therapies for prevention and treatment of CD.

**MATERIALS AND METHODS**

**Study subjects**

*Discovery cohort.* The discovery cohort used in the GWAS comprised 1096 CD subjects and 3970 healthy population controls. Cases were recruited from the Cedars-Sinai IBD and Pediatric IBD Centers and were diagnosed with CD according to standard clinical, radiological, endoscopic and histological criteria. This population consists of a pediatric cohort (39% of the sample, with a mean age of onset of 12.8 years) and adult CD subjects (61% of the sample, with a mean age of onset 37.7 years). Controls were obtained from the Cardiovascular Health Study (CHS), a population-based longitudinal study of risk factors for cardiovascular disease and stroke in adults 65 years of age or older, recruited at four field centers (68). A total of 5201 predominantly Caucasian individuals were recruited in 1989–1990 from random samples of Medicare eligibility lists, followed by an additional 687 African-Americans recruited in 1992–1993 (total n = 5888).

*Replication cohort.* The replication cohort used in Taqman genotyping of the *FUT2* locus consisted of 1174 CD cases and 357 healthy controls. All subjects in the replication cohort were of Northern European origin and independent of the cohort in the GWAS. Cases were recruited at the Cedars-Sinai IBD and Pediatric IBD Centers and diagnosed with CD according to the same criteria as those included in the discovery cohort. As with the discovery cohort, this replication sample consisted of both pediatric (9.4% of the total sample, mean age of onset of 11.2 years) and adult cases (90.6% of the total sample, mean age of onset of 31.1 years). Controls were recruited through the Cedars-Sinai IBD center as unrelated acquaintances and spouses of cases (who were not included in the current analysis set) with no personal or family history of IBD or autoimmune disease. All cases and controls provided informed consent prior to study participation and following approval of participating centers’ institutional review boards. All cases and controls were independent of the cases and controls included in the published CD meta-analysis.

**Genotyping**

*Discovery GWAS.* All genotyping was performed at the Medical Genetics Institute at Cedars-Sinai Medical Center using Illumina Infinium whole-genome genotyping technology, following the manufacturer’s protocol (Illumina, San Diego, CA, USA) (69,70). All cases were genotyped with the Illumina Human 610Quad platform. Controls were genotyped with the Illumina 370Duo platform. Samples with genotyping rates >98% were retained in the analysis. In addition, case and control cohorts were both investigated using Identity-By-Descent using the --genome command in PLINK. Identity-By-Descent is estimated within PLINK utilizing Identity-By-State distance clustering which estimates relatedness within the sample as $\text{Pi hat} = \frac{\text{P(IBD} = 2\text{)}}{\text{P(IBD} = 1\text{)}} + 0.5 \times \text{P(IBD} = 1\text{)}$ and identifies cryptic relatedness. Pairwise subjects comparisons with Pi hat scores >0.5 had one subject from the pair being analyzed for relatedness removed from the downstream analysis. Eleven CD subjects and 259 controls were removed based on cryptic relatedness. SNPs were excluded based on the following criteria: test of HWE ($P < 0.001$); SNP failure rate >3%; minor allele frequency <5%; and SNPs not found in dbSNP Build 129. SNPs were also examined in order to exclude case/control disparity in missingness, and SNPs with a missingness $P$-value <0.001 were excluded from the results [PLINK (71)]. A total of 304 825 SNPs passed our QC criteria were available in all data sets and included in the logistic regression association analysis.

*Replication of the FUT2 locus.* The six SNPs tested in the replication cohort were genotyped using the TaqMan™ Minor Groove Binder chemistry utilizing Assays on Demand according to the manufacturer’s instructions (Applied Biosystems, Foster City, CA, USA). Genotyping concordance among duplicate samples was 100% and the genotyping rate was 96.0% across all SNPs (94.8–97.6% for individual SNPs). All SNPs were in HWE in the controls.

**Statistical analysis**

*GWAS analysis.* Population structure was detected using multidimensional scaling [PLINK (71)]. In total, 10 principal components (PCs) were calculated and plotted for graphical representation of population substructure within the cohort. Self-reported ethnicity data were used to confirm the identification of ethnicity based on cluster plots. Subjects lying above 0.025 on the y-axis (PC 1) of the population structure plot (Supplementary Material, Fig. S1) were identified as subjects of African-American ancestry. To reduce false-positive discovery due to population substructure and the predominantly Caucasian make-up of the cases, the African-American subjects were excluded from downstream analysis. In total, 896 CD and 3204 control subjects were carried forward for association testing with the CD phenotype using a logistic regression in PLINK using all 10 PCs as covariates in the model (71). We calculated that the cohorts from this study give us power ranging from 0.36 [mean allele frequency (MAF) = 0.1] to 0.73 (MAF = 0.4) assuming an alpha of 0.05 and an effect size of 1.15 (approximate effect size seen in the 40 meta-analysis loci) to reproduce the associations demonstrated by Barrett *et al.* (15). Genomic control was calculated using the --adjust function within PLINK simultaneously with a logistic regression with the appropriate quality control filters and revealed a genomic inflation ($\lambda_{GC}$) of 1.09. The Q–Q plot is shown in Supplementary Material, Figure S2.

*Confirmation study.* The association of the FUT2 SNPs with CD was tested in an independent confirmation cohort. While
a logistic regression model is typically used for testing association between a single SNP and disease status, testing deviation from Hardy–Weinberg proportion among cases has been proposed as another approach for genetic association studies (72–73). Combining information from HWE tests in association studies has been proposed and demonstrated to be effective in improving study power while type I error can be effectively controlled (74–76). We utilized the mean-based tail-strength (TS) measure for association as proposed by Taylor and Tibshirani (75) and the extended median-based measure (TSM) for association as proposed by Wang and Shete (76) to combine the evidence for association from both resources to improve our study power. The program CSIG which can be used to implement both TS and TSM was used to perform the confirmation association analysis. An additive genetic model was assumed in the association analysis.

Meta-analysis. To improve study power and combine the evidence for association from our GWAS, we performed a meta-analysis for the four overlapping SNPs genotyped in both the initial GWAS and replication genotyping of the FUT2 locus. Given that a consistent directionality between our GWAS sample and the confirmation sample was observed for each of the four SNPs, we utilized Fisher’s method for combining P-values in our meta-analysis, consisting of a total of 2270 CD cases and 4337 controls.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

Conflict of Interest statement. None declared.

FUNDING

This study was supported in part by NCRR grant M01-RR00425 to the Cedars-Sinai General Research Center Genotyping core; NIH/NIDDK grant P01-DK046763; the Diabetes Endocrinology Research Center grant, DK 063491; Cedars-Sinai Medical Center Inflammatory Bowel Disease Research Funds; The Feintech Family Chair in IBD (S.R.T.); The Abe and Claire Levine Chair in Pediatric IBD (M.D.) and The Cedars-Sinai Board of Governors’ Chair in Medical Genetics (J.I.R.). Additional funding through grants DK76984 (M.D.), DK062413 (D.B.P.M.) and DK084554 (M.D. and D.P.B.M.). CHS research reported in this article was supported by contract numbers N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133; grant numbers U01 HL080295 and R01 HL087652 from the NHLBI, with additional contribution from the National Institute of Neurological Disorders and Stroke. A full list of principal CHS investigators and institutions can be found at http://www.chs-nhlinstitute.org/pi.htm.

REFERENCES


