Crohn’s Disease Loci Are Common Targets of Protozoa-Driven Selection

Rachele Cagliani,1 Uberto Pozzoli,1 Diego Forni,1 Andrea Cassinotti,2 Matteo Fumagalli,3 Matteo Giani,1 Maria Fichera,2 Marta Lombardini,2 Sandro Ardizzone,2 Rosanna Asselta,4 Roberto de Franchis,2 Stefania Riva,1 Mara Biasin,5 Giacomo P. Comi,6 Nereo Bresolin,1,6 Mario Clerici,7,8 and Manuela Sironi*1

1Bioinformatics Laboratory, Scientific Institute IRCCS E. Medea, Bosisio Parini (LC), Italy
2IBD Unit, Luigi Sacco University Hospital, Milano, Italy
3Department of Integrative Biology, University of California, Berkeley
4Dipartimento di Biologia e Genetica per le Scienze Mediche, Università degli Studi di Milano, Milano, Italy
5DISC LITA Vialba, University of Milan, Milano, Italy
6Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, Fondazione Ca’ Granda IRCCS Ospedale Maggiore Policlinico, Milano, Italy
7Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, Milano, Italy
8Fondazione Don C. Gnocchi, IRCCS, Milano, Italy
*Corresponding author: E-mail: manuela.sironi@bp.lnf.it.

Abstract

Previous studies indicated that a few risk variants for autoimmune diseases are subject to pathogen-driven selection. Nonetheless, the proportion of risk loci that has been targeted by pathogens and the type of infectious agent(s) that exerted the strongest pressure remain to be evaluated. We assessed whether different pathogens exerted a pressure on known Crohn’s disease (CD) risk variants and demonstrate that these single-nucleotide polymorphisms (SNPs) are preferential targets of protozoa-driven selection (P = 0.008). In particular, 19% of SNPs associated with CD have been subject to protozoa-driven selective pressure. Analysis of P values from genome-wide association studies (GWASs) and meta-analyses indicated that protozoan-selected SNPs display significantly stronger association with CD compared with nonselected variants. This same behavior was not observed for GWASs of other autoimmune diseases. Thus, we integrated selection signatures and meta-analysis results to prioritize five genic SNPs for replication in an Italian cohort. Three SNPs were significantly associated with CD risk, and combination with meta-analysis results yielded P values < 4 × 10−6. The bona fide risk alleles are located in ARHGEF2, an interactor of NOD2, NSF, a gene involved in autophagy, and HEBP1, encoding a possible mediator of inflammation. Pathway analysis indicated that ARHGEF2 and NSF participate in a molecular network, which also contains VAMP3 (previously associated to CD) and is centered around miR-31 (known to be disregulated in CD). Thus, we show that protozoa-driven selective pressure had a major role in shaping predisposition to CD. We next used this information for the identification of three bona fide novel susceptibility loci.

Key words: pathogen-driven selection, GWAS, Crohn’s disease, protozoa.

Introduction

Inflammatory bowel disease (IBD) is a chronic and relapsing inflammatory disease of the digestive tract comprising two main subtypes: ulcerative colitis (UC) and Crohn’s disease (CD). UC and CD differ by the intestinal localization and features of the inflammation and have a similar prevalence of 100–150 per 100,000 individuals in populations of European ancestry (reviewed in Khor et al. 2011). IBD is thought to result from a combination of genetic and environmental risk factors. In recent years, genome-wide association studies (GWASs) and meta-analyses have identified 71 and 47 susceptibility loci for CD and UC, respectively, with several variants being shared between the two diseases. Although these findings have provided valuable information into the molecular pathways underlying the pathogenesis of IBD, a large part of genetic risk factors remain to be identified, as testified by the fact that known CD and UC variants explain less than 25% of disease heritability (Khor et al. 2011). A portion of missing risk loci is likely to be accounted for by common variants with modest effect, which have been treated as false negatives in GWAS. Indeed, the heavy multiple test correction inherent in genome-wide analysis is likely to discard several true-positive associations, suggesting that leveraging GWAS results with other types of data might allow the identification of additional risk variants.

Previous analyses have indicated that polymorphisms associated with autoimmune diseases have been targets of natural selection (Fumagalli, Cagliani, et al. 2009; Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2011; Barreiro and...
Quintana-Murci 2010; Corona et al. 2010; Zhernakova et al. 2010; Abadie et al. 2011) specifically, the selective pressure exerted by infectious agents has modulated the frequency of a subset of risk alleles for autoimmune or chronic inflammatory conditions, including CD and UC (Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2011). Moreover, CD risk alleles have been shown to represent targets of recent positive selection (Corona et al. 2010). Nonetheless, the proportion of IBD risk loci that has been targeted by pathogen-driven selection and the type of infectious agent(s) that exerted the strongest pressure remain to be evaluated. Moreover, the relationship between past selection and modern susceptibility to autoimmunity is often complex and variable depending on the disease and the loci being investigated (Sironi and Clerici 2010). Also, the question remains open as to whether selection signatures can be exploited in the identification of novel susceptibility alleles for autoimmune diseases. A similar approach has been previously applied to increase power to detect malaria-resistant variants (Ayodo et al. 2007) but not to autoimmune risk loci. Also, it has been recently shown that the long-term evolutionary conservation scores at polymorphic genomic positions relates to the odds ratio of GWAS associations, with single-nucleotide polymorphisms (SNPs) at more conserved positions showing stronger effect sizes (Dudley et al. 2012). Thus, the integration of evolutionary consideration with large-scale association studies was suggested (Dudley et al. 2012).

Herein, we first evaluated how different pathogens have targeted known IBD susceptibility variants and then used this information for prioritizing GWAS SNPs for replication in an independent cohort, resulting in the identification of three bona fide novel susceptibility loci.

Results

GWAS SNPs for CD Are Preferential Targets of Protozoa-Driven Selection

We have previously shown that a subset of susceptibility alleles for IBD has been a target of pathogen-driven selection (Fumagalli, Pozzoli, et al. 2009). This original observation was restricted to the analysis of SNPs located within genes encoding interleukins and their receptors and relied on the use of micropathogen diversity in distinct geographic areas as a measure of pathogen-driven selective pressure (Fumagalli, Pozzoli, et al. 2009). Subsequent works have indicated that the selective pressure exerted by distinct pathogen groups can at least partially be disentangled (Prugnolle et al. 2005; Fumagalli et al. 2010a, 2010b, 2011; Pozzoli et al. 2010). Thus, we first wished to assess whether additional susceptibility variants for IBD are natural selection targets and whether specific pathogens have shaped their allele frequency. To this aim, we estimated pathogen-driven selection for single SNPs in the Human Genome Diversity Project (HGDP)-CEPH panel (Li et al. 2008) by calculating Kendall’s rank correlation coefficient ($\tau$) between the diversity of virus, protozoan, bacterium, or helminth species transmitted in distinct geographic locations and the allele frequency for the populations living in those same areas (Prugnolle et al. 2005; Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2010a, 2010b; Pozzoli et al. 2010). Each SNP was assigned a percentile rank in the distribution of $\tau$ values calculated for all SNPs having a minor allele frequency (MAF) similar (in the 1% range) to that of the SNP being analyzed (Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2010a, 2010b; Pozzoli et al. 2010). We considered an SNP to be significantly associated with pathogen diversity if it displayed a rank higher than 0.95; this approach has the advantage of resulting in the same number of variants associated with the four pathogen groups. We next retrieved all GWAS SNPs associated to any trait or disease from the National Human Genome Research Institute (NHGRI) Catalog of Published GWASs (Hindorff et al. 2009) and collapsed SNPs in tight linkage disequilibrium ($LD, r^2 > 0.8$) into single loci. We retained only variants that have been genotyped in the HGDP-CEPH panel ($n = 2,773$); out of these, 43 and 42 SNPs were associated with CD and UC, respectively, four of them being common between the two diseases.

Having set a 0.95 cutoff for SNPs associated with pathogen diversity, we would expect 5% of CD and UC variants to be correlated with the diversity of each pathogen group. Conversely, the percentage of CD SNPs that were significantly associated with virus, bacterium, protozoan, and helminth diversity were 11.6, 4.6, 18.6, and 6.9 (table 1, supplementary table S1, Supplementary Material online), respectively, suggesting a strong enrichment for SNPs targeted by protozoa-driven selection. For UC, a percentage of SNPs close to 5% correlated with the diversity of each pathogen group (table 1, supplementary table S1, Supplementary Material online).

To assess whether the number of CD SNPs associated with protozoa diversity was significantly higher than expected, we applied a resampling approach on the 2,773 GWAS SNPs. By performing 10,000 resamplings of 43 randomly selected MAF-matched SNPs, we verified that the empirical probability of obtaining eight variants significantly associated with protozoan diversity amounted to 0.0078 (table 1), indicating that CD SNPs are preferential targets of protozoa-driven selection. A similar result was obtained when we compared CD SNPs with GWAS variants that have been identified as susceptibility loci for immune-mediated or infectious diseases/trait (supplementary table S2, Supplementary Material online) ($P = 0.0003$, table 1). Conversely, the number of CD SNPs that correlated with the diversity of virus, bacterium, and helminth species was not different from expected (table 1).

Correlation with pathogen diversity and, in general, approaches that exploit environmental variables to detect selection signatures have been shown to be reliable and robust to demographic effects (Hancock et al. 2008, 2010, 2011; Pozzoli et al. 2010). Nonetheless, to confirm our results using a different test for selection, we calculated population genetic differentiation (measured as $F_{ST}$ [Wright 1950]) for CD variants that were found to be or to be not targeted by protozoa-driven selection. Specifically, we used the 1000 Genomes Pilot Project data to obtain $F_{ST}$ values for the Yoruba/European, Yoruba/Asian, and European/Asian comparisons. $F_{ST}$ values for CD SNPs were compared with the distribution of $F_{ST}$ calculated for variants deriving from
approximately 1,000 randomly selected genes (see Materials and Methods). Results indicated that six of eight CD SNPs (75%) targeted by protozoa-driven selection displayed $F_{ST}$ values higher than the 95th percentile in at least one comparison (fig. 1); for CD SNPs that were not found to correlate with protozoa diversity, this proportion amounted to 31.4% (11 of 35) (fig. 1), which may still be higher than expected, in line with the observation that CD SNPs represent preferential selection targets (Corona et al. 2010), being the underlying pressure represented by protozoa or not.

Finally, we confirmed previous indications that correlation between allele frequencies and protozoa diversity is not secondary to other factors such as latitude and climatic variables (temperature and short wave radiation flux) (supplementary table S3, Supplementary Material online).

**SNPs Targeted by Protozoa-Driven Selective Pressure Display Stronger Association with CD than Nonselected Variants**

Given the results described earlier, we wished to assess whether, in single GWAS or meta-analyses, SNPs targeted by protozoa-driven selection display stronger association with CD compared with variants that have not been subject to this selective pressure. For this purpose, we retrieved association $P$ values for the WTCCC1 (Wellcome Trust Case Control Consortium 2007) and NIDDK (Rioux et al. 2007) GWAS for CD, as well as for a meta-analysis of these two studies plus a third Belgian/French cohort (Barrett et al. 2008). The WTCCC1 and NIDDK studies were performed on subjects of British and US/Canadian origin, respectively, using different platforms; thus, the studies have few SNPs in common. We compared the association $P$ values of protozoa-selected and nonselected SNPs by the use of quantile–quantile (Q-Q) plots. As shown in figure 2a and b, in both CD GWASs and Barrett’s meta-analysis, association values for protozoa-selected SNPs were stronger than for nonselected variants. This same behavior was not observed for several GWAS of other autoimmune diseases (fig. 2a). To provide confidence intervals for the observed shift in association $P$ values in the CD meta-analysis, we applied a moving block bootstrap (MBB) procedure (see Materials and Methods). This is devised to account for the nonindependence of SNPs (due to LD), which might be even more pronounced in regions that have been targeted by natural selection (Keinan et al. 2007; Keinan and Reich 2010). As shown in figure 2b, the upper edge of the confidence interval is below the diagonal identity line for most $P$ value ranges, indicating that SNPs selected by protozoa have significantly stronger association with CD compared with variants that have not been targeted by protozoa-driven selection.

The power to detect association with a trait also depends on the SNP frequency in the population being analyzed. The average MAF (over all HGDP-CEPH populations) was slightly lower for protozoa-selected compared with nonselected variants (mean MAF = 0.250 and 0.266, respectively) included in Barrett’s meta-analysis. Nonetheless, MAF calculation in subjects from Europe only (Barrett’s meta-analysis included
GWASs conducted in Europeans) revealed the opposite situation, with SNPs targeted by protozoa-driven selection (mean MAF in Europeans = 0.268) having, on average, higher MAF than nonselected SNPs (mean MAF in Europeans = 0.256; Wilcoxon Rank Sum test, \( P < 0.001 \)). Therefore, to rule out the possibility that the stronger association of SNPs selected by protozoa with CD (fig. 2b) is secondary to a frequency bias, we reanalyzed Barrett’s data taking MAF into account. In particular, we first LD pruned (with \( r^2 > 0.8 \)) the meta-analysis variants included in the HGDP-CEPH panel, and we next Q-Q plotted the association P values of protozoa-selected and nonselected variants. The same analysis was performed for 1,000 random samples of variants from the LD-pruned set (LDS) and matched for MAF (calculated in Europeans only) with protozoa-selected variants. These samples were used to obtain the 5th and 95th percentiles across P value quantiles. As shown in figure 2c, protozoa-selected SNPs show stronger association with CD compared with MAF-matched randomly selected variants.

Protozoa-Driven Selection Allows Identification of Novel Bona Fide CD Susceptibility Loci

In the CD meta-analysis described earlier, the authors set a threshold P value of \( 5 \times 10^{-5} \) for SNP follow-up, although variants with less significant association might well represent false negatives (Barrett et al. 2008). Given the results obtained, we reasoned that true-positive associations might be enriched among SNPs that display signatures of protozoa-driven selection and do not reach the meta-analysis threshold for significance. Thus, we first assessed whether protozoa-selected SNPs with nominally significant P value (i.e., \( 0.05 < P \text{ value} < 5 \times 10^{-5} \)) in Barrett’s meta-analysis would have a better chance of being identified as CD-associated in a more powerful meta-analysis of six studies (Franke et al. 2010) compared with variants that display no correlation with protozoa diversity. A total of 44,898 polymorphisms had a P value in the range of \( 0.05–5 \times 10^{-5} \) in Barrett’s meta-analysis and were analyzed by Franke et al. (2010); of these, 1,284 showed association with protozoa diversity. The fraction of variants that reached genome-wide significance (i.e., \( P < 5 \times 10^{-8} \)) in the six-study meta-analysis resulted to be 0.99% and 0.41% for protozoa-associated and nonassociated variants, respectively. Although these variants cannot be regarded as independent, a more than 2-fold enrichment is observed for protozoa-selected SNPs.

We next explored the possibility of integrating selection signatures with meta-analysis results to identify novel susceptibility variants for CD. To this aim, we exploited the three-study meta-analysis for variant selection, so that the association results could be combined with the partially independent six-study meta-analysis. Thus, we extracted the 1,284 protozoa-selected SNPs showing a meta-analysis P value higher than \( 5 \times 10^{-5} \) and lower than 0.05 and ranked them according to their meta-analysis P value (supplementary table S4, Supplementary Material online). After discarding a variant located in \( \text{PUS10} \) (previously associated with CD) and one located in \( \text{REL} \) (which is close to \( \text{PUS10} \)), the top five genic SNPs (supplementary table S4, Supplementary Material online) were selected for replication in an Italian cohort of 809 CD cases and 632 controls. Association analysis indicated that three of the five SNPs we tested were significantly associated with the risk to develop CD (table 2, supplementary table S5, Supplementary Material online). Under the null hypothesis with five tested SNP, no variant would be expected to yield a significant association with \( P < 0.05 \).
FIG. 2. Quantile–quantile plots of GWAS and meta-analyses for protozoa-selected or nonselected SNPs. (a) Q-Q plots are shown for the two CD GWAS (WTCCC1 and NIDDK) and for other autoimmune diseases with available GWAS data (SLE, systemic lupus erythematosus; MS, multiple sclerosis; AS, ankylosing spondylitis; ATD, autoimmune thyroid disease; RA, rheumatoid arthritis; T1D, type 1 diabetes). Q-Q plots are performed by comparing the $-\log_{10}(P\text{value})$ for SNPs that have been selected (x axis) or have not been selected (y axis) by the selective pressure exerted by protozoa. The identity diagonal line is also shown (hatched). (b) Q-Q plot for the CD meta-analysis with 90% confidence intervals (gray area) calculated by bootstrap resampling of block of 40 SNPs (see Materials and Methods). As above, the $-\log_{10}(P\text{values})$ are compared for SNPs that have been selected (x axis) or have not been selected (y axis) by protozoa. (c) Q-Q plot for the LD-pruned CD meta-analysis variants targeted (x axis) or not targeted (y axis) by protozoa-driven selection (black solid line). Probability intervals (gray area delimiting the 5th and 95th percentiles) were calculated by drawing 1,000 random samples of MAF-matched SNPs (MAF calculated in Europeans only; see Materials and Methods).
suggesting that these represent true positive associations. Combination of association \( P \) values with Franke’s meta-analysis results indicated that the three variants display \( P \) values \(< 4 \times 10^{-6} \) (table 2), whereas no association is observed for the two remaining SNPs. The three significant SNPs are located in \( ARHGEF2 \) (Rho/Rac guanine nucleotide exchange factor 2), \( NSF \) (N-ethylmaleimide-sensitive factor), and \( HEBP1 \) (heme-binding protein 1).

To verify whether the identified genes functionally interact, we applied unsupervised Ingenuity Pathway Analysis (IPA). The three genes carrying variants associated with CD were used as input, and IPA identified two networks. One network (fig. 3) contains \( ARHGEF2 \) and \( NSF \) plus VAMP3, a gene previously associated with CD in a recent meta-analysis of GWAS studies (Franke et al. 2010). The second network contains \( HEBP1 \) only among the input genes (not shown).

**Discussion**

In recent years, the increasing availability of human genetic variation data allowed large-scale analysis of genomic regions subject to the selective pressure exerted by the environment. The identification of polymorphisms that strongly correlate in frequency with environmental variables has turned out as a powerful strategy to identify signatures of local adaptation to climate, dietary regimes, and infectious agents (Prugnolle et al. 2005; Hancock et al. 2008, 2010, 2011; Fumagalli, Cagliani, et al. 2009; Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2010a, 2010b, 2011; Pozzoli et al. 2010). In particular, these approaches were suggested to be particularly useful for detecting selection on standing variation, which is otherwise difficult to identify using more common tests of natural selection (Pennings and Hermisson 2006).

Among the environmental factors investigated to date, pathogens have exerted the strongest pressure on the human genome (Fumagalli et al. 2011). These analyses have a clear relevance from an evolutionary point of view and help explain why specific disorders such as IBD and celiac disease (CeD) (Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2011; Barreiro and Quintana-Murci 2010; Zhernakova et al. 2010; Abadie et al. 2011) are so prevalent in humans, despite the fitness reduction they cause.

Previous works have shown that, albeit not independent, the diversity of different pathogen groups (i.e., protozoa, bacteria, helminths, and viruses) in distinct geographic areas can be regarded as a reliable measure of the selective pressure exerted by these agents and that their effects can be disentangled (Prugnolle et al. 2005; Fumagalli, Cagliani, et al. 2009; Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2010a, 2010b, 2011; Pozzoli et al. 2010). We used this same approach to verify whether known variants associated with CD and UC have been targeted by pathogen-driven selective pressure more often than expected; our data indicate that although UC susceptibility loci are not preferential selection targets, approximately 19% of SNPs associated with CD have been subject to a selective pressure exerted by protozoa. By applying a resampling approach, this proportion was compared with that of all GWAS SNPs or with variants that have been associated with immune-mediated/infectious diseases.

### Table 2. Association Analysis of Five Selected SNPs

<table>
<thead>
<tr>
<th>SNP*</th>
<th>Gene</th>
<th>Risk</th>
<th>Allele</th>
<th>Barrett’s Meta-Analysis, ( P ) Value</th>
<th>Italian Cohort (809 Cases, 632 Controls)</th>
<th>Franco’s Meta-Analysis, ( P ) Value</th>
<th>Combined ( P ) Valued</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2364403*</td>
<td>ARHGEF2</td>
<td>A</td>
<td>0.00011</td>
<td>0.19</td>
<td>0.0126</td>
<td>1.28 (1.05–1.55)</td>
<td>3 ( \times 10^{-6} )</td>
</tr>
<tr>
<td>rs199533*</td>
<td>NSF</td>
<td>G</td>
<td>0.00035</td>
<td>0.78</td>
<td>0.0005</td>
<td>1.36 (1.14–1.63)</td>
<td>0.00038</td>
</tr>
<tr>
<td>rs3782567*</td>
<td>HEBP1</td>
<td>T</td>
<td>0.00041</td>
<td>0.48</td>
<td>0.0139</td>
<td>1.21 (1.04–1.40)</td>
<td>1.3 ( \times 10^{-5} )</td>
</tr>
<tr>
<td>rs9635320</td>
<td>ARID3B</td>
<td>G</td>
<td>0.00041</td>
<td>0.14</td>
<td>0.0559</td>
<td>1.24 (0.99–1.54)</td>
<td>0.25</td>
</tr>
<tr>
<td>rs1011312</td>
<td>TSP12</td>
<td>T</td>
<td>0.00056</td>
<td>0.085</td>
<td>0.5603</td>
<td>1.1 (0.80–1.51)</td>
<td>0.17</td>
</tr>
</tbody>
</table>

---

*SNPs are denoted with an asterisk when the risk allele (if any) positively correlates with protozoa diversity.

**P** value from a three-study meta-analysis (Barrett et al. 2008).

**P** value from a six-study meta-analysis (Franke et al. 2010).

**P** values were combined using Fisher’s method (Italian cohort and Franke’s meta-analysis; Franke et al. 2010).
only. This procedure is devised to assess the excess of selected SNPs for CD against variants that share the same biases (i.e., being tag SNPs, and most often identified in European populations) and that have an effect on human phenotypes. In both comparisons, the proportion of CD variants selected by protozoa was statistically significant. Similarly, analysis of GWAS data and a meta-analysis for CD indicated that variants selected by protozoa have stronger association P values compared with variants that have not been targeted by these infectious agents. This was not observed for all other autoimmune diseases we analyzed, including UC. A recent large-scale analysis (Fumagalli et al. 2011) identified genes involved in CeD, UC, type 1 diabetes, and multiple sclerosis as being common targets of pathogen-driven selection. Nonetheless, a different approach was used in that study, as the authors did not disentangle the effect of different pathogens, and performed gene-wise analyses. Herein, we applied SNP-wise tests, and the data we present are in agreement with a previous study (Fumagalli, Pozzoli, et al. 2009) showing that a few variants located in genes encoding interleukins/interleukin receptors and associated with IBD and/or CeD have been targets of pathogen-driven selection. Unfortunately, no complete GWAS data for CeD is publicly available, so we could not verify whether pathogen-selected SNPs show stronger association with this disease compared with nonselected variants; however, analysis of GWAS-significant SNPs for CeD revealed no significant enrichment of polymorphisms that correlate with the diversity of any pathogen group (not shown).

CD is thought to occur in genetically susceptible individuals as a consequence of disregulated immune recognition of commensal gut flora, which is mainly accounted for by bacterial species. Therefore, protozoa have generally attracted little interest in relation to this disease. Nonetheless, immune response pathways often allow protection against a wide variety of pathogens. For example, mice deficient in Ifngm1, a known CD gene, are highly susceptible to infection with both Listeria monocytogenes and Toxoplasma gondii (Collazo et al. 2001). Interestingly, recent evidences have indicated T. gondii as a possible environmental trigger/cause of autoimmune diseases, including IBD (Lidar et al. 2009; Shapira et al. 2012). This parasite might have a particular interest in relation to CD, as infection of certain mouse strains with Toxoplasma triggers an inflammatory pathology resembling CD (Egan et al. 2011, 2012). This observation has also led to the proposal that T. gondii infection might be used as a novel experimental model of CD (Denkers 2010; Egan et al. 2012). Our data indicate that variants in eight genes identified in GWASs for CD have been targeted by protozoa-driven selection. Among these, NOD2 and TYK2 are necessary for clearance of T. gondii in mouse models (Shaw et al. 2003, 2009), whereas IL23, the ligand of IL23R is essential for the triggering of intestinal inflammation by the parasite (Munoz et al. 2009). These observations do not necessarily imply that T. gondii is the selective agent responsible for shaping allele frequencies at these loci. Indeed, protozoan-transmitted diseases, and malaria in particular, have been shown to represent an exceptionally strong selective pressure for humans (Kwiatkowski 2005; Pozzoli et al. 2010), and different pathogenetic protozoa species, including Trypanosoma, Leishmania, and Giardia have likely represented (and still represent) major health issues throughout human history.

Although GWAS and meta-analyses have unveiled several susceptibility loci for CD, the majority of risk alleles for the disease remain to be identified (Khor et al. 2011). A portion of these are thought to be accounted for by small effect variants that do not reach statistical significance in genome-wide analyses (Altshuler and Daly 2007; McCarthy et al. 2008). As previously indicated (McCarthy et al. 2008), loci with weak effects are of little value in predictive testing but may be extremely useful in providing insight into disease pathogenesis. Thus, the integration of additional information with genome-wide data has been advocated as a means to disease loci discovery (McCarthy et al. 2008). Natural selection acts through time; thus, if the selective pressure is sufficiently ancient, even a small fitness advantage can leave a signature on the allele frequency spectrum. This information can be exploited to identify variants with a functional effect (or in LD with a functional polymorphism). Therefore, based on the observations that SNPs targeted by protozoa-driven selection are over-represented among GWAS SNPs and have stronger association P values with CD, we prioritized five variants from a meta-analysis for replication in an independent CD case/control cohort. Three of these variants, located in HEBP1, NSF, and ARHGEF2, showed significant association with the disease, representing a strong enrichment over the expected based on a null hypothesis. HEBP1 encodes an intracellular tetra- and/or-binding protein with possible involvement in heme homeostasis (Taketani et al. 1998). Processing of HEPB1 by cathepsin D (CTSD) originates a 21-amino acid peptide (F2L) that acts as a specific endogenous agonist for formyl peptide receptor 3 (FPR3) (Migeotte et al. 2005; Devosse et al. 2011). FPR3 is expressed by monocytes, dendritic cells, and by macrophage subpopulations (including those residing in the colon) (Devosse et al. 2009), and F2L binding promotes calcium mobilization and chemotaxis in these cells. Whether specific stimuli induce HEBP1 proteolysis remains to be investigated, but it is worth noting that CTSD is upregulated in macrophages isolated from IBD patients compared with those derived from controls (Hausmann et al. 2004); also, inhibition of CTSD in an experimental model of colitis improves disease symptoms (Menzel et al. 2006), suggesting a possible role for the HEPB1-CTSD-FPR3 axis in sustaining mucosal inflammation in IBD.

NSF is a protein involved in vesicle sorting and autophagy (Behrends et al. 2010). This latter process is central in the clearance of intracellular bacteria, as well as in the pathogenesis of CD (Khor et al. 2011). In line with this view, NSF is necessary for the recruitment of VAMP3 (fig. 2) at sites of Salmonella infection and for the maturation of bacterium-containing vacuoles (Coppolino et al. 2001). Notably, the NSF variant we associated with CD has previously been shown to confer risk for Parkinson’s disease (Simon-Sanchez et al. 2009); thus, NSF joins LRRK2 (Zimprich et al. 2004) as a risk gene shared between the two diseases. This is in line with the recent observation that pleiotropy is widespread for complex disease risk loci (Sivakumaran et al. 2011) and with the
description of higher frequency of NOD2 variants in sporadic Parkinson’s disease (Bialecka et al. 2007).

Finally, ARHGEF2 (also known as GEF-H1) encodes a nucleotide exchange factor for Rac and Rho small GTPases (Ren et al. 1998). The gene is involved in multiple cellular processes, including pathogen recognition by NOD1 and NOD2. Specifically, by signaling downstream of NOD2, GEF-H1 mediates the activation of RIP2, a process that is abolished in the presence of the 3020insC variant of NOD2 associated with CD (Zhao et al. 2012). Similarly, ARHGEF2 participates in sensing muramyl dipeptides through NOD1 and, by activating RHOA (fig. 2), induces the NF-kB signaling pathway (Fukazawa et al. 2008). The interaction between RHOA and ARHGEF2 at tight junctions is also central in regulating paracellular permeability in human colonic epithelia (Samarin et al. 2007), and enteropathogenic Escherichia coli strains have been shown to hijack GEF-H1 activity to facilitate infection (Matsuzawa et al. 2004). Interestingly, an increased activation of RHOA has been described in the inflamed intestinal mucosa of CD patients and in rats with experimentally induced colitis (Segain et al. 2003).

Thus, the three genes we have identified have biological functions consistent with their involvement in the pathogenesis of CD, and two of them participate in a molecular pathway that contains additional loci previously associated with the disease (VAMP3 [Franke et al. 2010]) or disregulated in CD patients (RHOA) (Segain et al. 2003). The network also contains CD28, a costimulatory molecule essential for T-cell differentiation and survival, and is centered around miR-31. Notably, this microRNA shows higher expression in both unaffected and inflamed colonic tissue from IB patients compared with controls, with miR-31 abundance being higher in inflamed versus uninflamed IBD tissue (Dalal and Kwon 2010). Similarly, increased expression of miR-31 was detected in Il10−/− mice, an animal model of Th1-mediated IB and preceded the ensuing of colonic pathology (Schaefer et al. 2011).

It is worth noting that we observed both risk and protective alleles for CD to correlate with protozoa diversity; thus, as anticipated elsewhere (Sironi and Clerici 2010), the vision whereby alleles that predispose to autoimmunity are maintained in human populations due to their being protective against infections is oversimplistic. Possibly, some risk alleles have increased in frequency as a consequence of the selection operated by protozoa, whereas others might have reached higher representation in areas where the selective pressure exerted by these parasites has been relaxed. This is also in line with the observation that both risk and protective alleles for autoimmune diseases increased in frequency due to selection (Barreiro and Quintana-Murci 2010).

**Materials and Methods**

**GWAS Data Retrieval and Analysis**

A list of GWAS SNPs was obtained from the NHGRI Catalog of Published GWASs (Hindorff et al. 2009) (as of 19 October 2011). LD between SNP pairs was calculated using SNAP (SNP Annotation and Proxy Search; http://www.broadinstitute.org/mpg/snap/ldsearchpw.php, last accessed October 2, 2012) using data for Europeans (CEU); SNPs showing r² ≥ 0.8 were collapsed in a single locus by randomly selecting one of the two polymorphisms.

Resampling experiments for CD and UC SNPs were performed using SNP sets matched for MAF; specifically for each CD or UC SNP, one GWAS variant with a MAF differing less than 1% was sampled (i.e., we divided MAFs into 50 bins).

Association P values for the WTCCC GWAS for CD (Wellcome Trust Case Control Consortium 2007) were retrieved from the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/, last accessed October 2, 2012), which is hosted by the EBI, under accessions EGAS00000000006. Meta-analysis data were derived from previous works (Barrett et al. 2008; Franke et al. 2008) and are available at http://www.broad.mit.edu/~jcbarret/ibd-meta/ (last accessed October 2, 2012) and http://www.ibdgenetics.org/downloads.html (last accessed October 2, 2012). GWAS data for the NIDDK study of CD (Rioux et al. 2007) were retrieved from the dbGAP website (http://www.ncbi.nlm.nih.gov/gap/, last accessed October 2, 2012). GWAS data for other autoimmune diseases were retrieved either from the EGA (WTCCC1 studies) or from dbGAP.

For each GWAS (or meta-analysis) panel, a subset (S) of SNPs genotyped in HGDP-CEPH was obtained. The S subset was then divided into protozoan-selected (PS, n = 17,115) and protozoan nonselected (NPS, n = 304,710) SNPs. PS cumulative frequencies (CFs) of association P values were obtained and the corresponding quantiles in the NPS distribution calculated and Q-Q plotted.

Confidence intervals for the meta-analysis Q-Q plot were calculated using an MBB resampling procedure (Keinan et al. 2007; Keinan and Reich 2010). This is suitable to correct for the nonindependence among loci, also taking into account the possibility of an increased level of LD near the positively selected sites. Chromosome blocks were sampled following a previously described procedure (Fumagalli et al. 2011), and SNPs belonging to the subset S were extracted, so as to obtain a sample subset (SS). Each SS was Q-Q plotted as described for the subset S, using the same CFs. For each Q-Q plot point, we calculated the distance from the identity line. After repeating the procedure for 1,000 samples, we obtained a distribution of distances at each CF. We then used the 5th and 95th percentiles of these distributions to plot confidence intervals around the real Q-Q plot data points.

A second analysis was performed on the S set to account for MAF differences in European populations. We first LD pruned the meta-analysis variants included in the HGDP-CEPH panel by calculating LD using the 1000 Genomes Pilot Project data (phased variants) and retaining only variants with r² < 0.8 (for each SNP pair with LD > 0.8 one of the two variants was randomly discarded). For each SNP in the LDS, we next calculated MAF averaged over HGDP-CEPH populations living in Europe (Euro-MAF). These were binned in 50 quantiles. As above, the LDS subset was then divided into PS (LDPS, n = 14,205) and NPS (LDNPS, n = 250,697) SNPs. LDPS and LDNPS association P values were Q-Q plotted. The same procedure was performed for 1,000 random samples of LDS variants matched in Euro-MAF.
with LDS (i.e., for each LDPS, one SNP was randomly selected from the same Euro-MAF bin). These samples were used to calculate the probability intervals that the association P value distribution for 14,205 randomly selected SNPs deviates from the distribution of the remaining LDS SNPs.

Identification of Variants Targeted by Pathogen-Driven Selection

The approach used to identify variants selected by different pathogen species has been extensively described elsewhere (Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2010a, 2010b; Pozzoli et al. 2010). Briefly, it is based on calculating Kendall’s correlation coefficient (τ) between allele frequencies of HGDP-CEPH SNPs (Li et al. 2008) and pathogen diversity in the countries where populations included in the Panel live. To account for demographic events, each SNP is then assigned a percentile rank in the distribution of τ values calculated for all SNPs having a MAF similar (in the 1% range) to that of the SNP being analyzed (Fumagalli, Pozzoli, et al. 2009; Fumagalli et al. 2010a, 2010b; Pozzoli et al. 2010). MAF was calculated as the average over all populations in the panel. The annual minimum and maximum temperature, as well as short wave radiation flux, were retrieved from the NCEP/NCAR database (http://www.ngdc.noaa.gov/ecosys/cdroms/ged_iia/datasets/a04/, last accessed October 2, 2012) (Legates and Willmott Average, regredded data set) using the geographic coordinates reported by HGDP-CEPH (http://www.cephb.fr/en/hgdp/table.php, last accessed October 2, 2012) for each population.

1000 Genome Data Analysis

Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website (http://www.1000genomes.org/, last accessed October 2, 2012) (1000 Genomes Project Consortium et al. 2010). SNP genotypes were organized in a MySQL database. A set of programs was developed to retrieve genotypes from the database and to analyze them according to selected regions/populations. These programs were developed in C++ using the GeCo++ (Cereda et al. 2011) and the libsequence (Thornton 2003) libraries. For each pair of populations, \( F_{ST} \) was calculated for CD SNPs and for variants located within 1,000 randomly selected gene regions. These latter were used to construct an empirical distribution. Because \( F_{ST} \) values are not independent from allele frequencies, we binned variants based on their MAF (50 quantile classes) and calculated the 95th and 99th percentiles for each MAF class.

Patients and Controls

CD cases (459 males and 350 females) and controls (371 males and 261 females) were recruited at the IBD Unit of the Luigi Sacco University Hospital in Milano, a third-level centre for the management of IBD patients. The diagnosis of CD was based on international published criteria, according to clinical, endoscopic, histological, and/or radiological data (Lennard-Jones 1989). A detailed clinical history, as well as laboratory and instrumental diagnostic data, was collected. All patients and controls were Italian of European ancestry. All subjects gave informed consent according to protocols approved by the local Ethic Committee.

Genotyping and Statistical Analysis

Genotyping of all SNPs was performed by TaqMan probe assays (Applied Biosystems, Foster City, CA) using the allelic discrimination real-time polymerase chain reaction method.

All polymorphisms complied to Hardy–Weinberg equilibrium. Association analysis was performed by means of logistic regressions with sex as a covariate using PLINK (Purcell et al. 2007).

Network Construction

Biological network analysis was performed with IPA software using an unsupervised analysis (www.ingenuity.com, last accessed October 2, 2012). IPA builds networks by querying the Ingenuity Pathways Knowledge Base for interactions between the identified genes and all other gene objects stored in the knowledge base; it then generates networks with a maximum node size of 35 genes/proteins. We used the three genes significantly associated with CD as the input set and the Ingenuity Knowledge Base (genes and endogenous chemicals) as the reference set (as of April 2012). We considered both direct and indirect relationships (confidence = experimentally observed or high). Only relationships observed in human, mouse, or rat were considered.

Supplementary Material

Supplementary tables S1–S5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the Wellcome Trust Case Control Consortium for allowing access to genotype data. They are also grateful to Grazia Giusti and Lina Raucci for technical support. This work was supported by the Broad Medical Research Program of The Broad Foundation (grant IBD-0294), by 2010 Ricerca Corrente (Italian Ministry of Health), by Fondazione CARIPLO, and by an EMBO Long-Term Post-doctoral Fellowship (ALTF 229-2011) to M.F.

References


Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678.


