"Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4"

Libioulle, Cecile ; Louis, Edouard ; Hansoul, Sarah ; Sandor, Cynthia ; Farnir, Frederic ; Franchimont, Denis ; Vermeire, Severine ; Dewit, Olivier ; de Vos, Martine ; Dixon, Anna ; Demarche, Bruno ; Gut, Ivo ; Heath, Simon ; Foglio, Mario ; Liang, Liming ; Laukens, Debby ; Mni, Myriam ; Zelenika, Diana ; Van Gossum, Andre ; Rutgeerts, Paul ; Belaiche, Jacques ; Lathrop, Mark ; Georges, Michel

To identify novel susceptibility loci for Crohn disease (CD), we undertook a genome-wide association study with more than 300,000 SNPs characterized in 547 patients and 928 controls. We found three chromosome regions that provided evidence of disease association with p-values between 10(-6) and 10(-9). Two of these (IL23R on Chromosome 1 and CARD15 on Chromosome 16) correspond to genes previously reported to be associated with CD. In addition, a 250-kb region of Chromosome 5p13.1 was found to contain multiple markers with strongly suggestive evidence of disease association (including four markers with p < 10(-7)). We replicated the results for 5p13.1 by studying 1,266 additional CD patients, 559 additional controls, and 428 trios. Significant evidence of association (p < 4 x 10(-4)) was found in case/control comparisons with the replication data, while associated alleles were over-transmitted to affected offspring (p < 0.05), thus confirming that the 5p13.1 locus contributes to CD susceptibility. The CD-associated 250-kb region was saturated with 111 SNP markers. Haplotype analysis supports a complex locus architecture with multiple variants contributing to disease susceptibility. The novel 5p13.1 CD locus is contained within a 1.25-Mb gene desert. We present evidence that disease-associated alleles correlate with quantitative expression levels of the prostaglandin receptor EP4, PTGER4, the gene that resides closest to the associated region. Our results identify a major new susceptibility locus for CD, and suggest that genetic variants associated with disease risk at th...

CITE THIS VERSION


Le dépôt institutionnel DIAL est destiné au dépôt et à la diffusion de documents scientifiques émanant des membres de l'UCLouvain. Toute utilisation de ce document à des fins lucratives ou commerciales est strictement interdite. L'utilisateur s'engage à respecter les droits d'auteur liés à ce document, principalement le droit à l'intégrité de l'œuvre et le droit à la paternité. La politique complète de copyright est disponible sur la page Copyright policy

DIAL is an institutional repository for the deposit and dissemination of scientific documents from UCLouvain members. Usage of this document for profit or commercial purposes is strictly prohibited. User agrees to respect copyright about this document, mainly text integrity and source mention. Full content of copyright policy is available at Copyright policy

Available at: http://hdl.handle.net/2078.1/37616
Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4

Cécile Libioulle1, Edouard Louis2, Sarah Hansoul1, Cynthia Sandor1, Frédéric Farnir1, Denis Franchimont3, Séverine Vermeire4, Olivier Dewit5, Martine de Vos6, Anna Dixon7, Bruno Demarche8, Ivo Gut9, Simon Heath9, Mario Foglio9, Liming Liang10, Debby Laukens6, Myriam Mni1, Diana Zelenika9, André Van Gossum3, Paul Rutgeerts4, Jacques Belaiche2, Mark Lathrop9, Michel Georges1*

1 Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liége, Liége, Belgium, 2 Unit of Hepatology and Gastroenterology, Department of Clinical Sciences, GIGA-R, Faculty of Medicine, and CHU de Liége, University of Liége, Liége, Belgium, 3 Department of Gastroenterology, Erasmus Hospital, Free University of Brussels, Brussels, Belgium, 4 Department of Gastroenterology, University Hospital Leuven, Leuven, Belgium, 5 Department of Gastroenterology, Clinique Universitaire St Luc, UCL, Brussels, Belgium, 6 Department of Hepatology and Gastroenterology, Ghent University Hospital, Ghent, Belgium, 7 National Heart and Lung Institute, Imperial College London, England, 8 Unit of Bioinformatics, GIGA-R and Institut Montefiore, University of Liége, Liége, Belgium, 9 Centre National de Génotypage, Evry, France, 10 Centre for Statistical Genetics, Department of Biostatistics, Ann Arbor, Michigan, United States

To identify novel susceptibility loci for Crohn disease (CD), we undertook a genome-wide association study with more than 300,000 SNPs characterized in 547 patients and 928 controls. We found three chromosome regions that provided evidence of disease association with p-values between 10^-6 and 10^-9. Two of these (IL23R on Chromosome 1 and CARD15 on Chromosome 16) correspond to genes previously reported to be associated with CD. In addition, a 250-kb region on Chromosome 5p13.1 was found to contain multiple markers with strongly suggestive evidence of disease association (including four markers with p < 10^-7). We replicated the results for 5p13.1 by studying 1,266 additional CD patients, 559 additional controls, and 428 trios. Significant evidence of association (p < 4.8 x 10^-4) was found in case/control comparisons with the replication data, while associated alleles were over-transmitted to affected offspring (p < 0.05), thus confirming that the 5p13.1 locus contributes to CD susceptibility. The CD-associated 250-kb region was saturated with 111 SNP markers. Haplotype analysis supports a complex locus architecture with multiple variants contributing to disease susceptibility. The novel 5p13.1 CD locus is contained within a 1.25-Mb gene desert. We present evidence that disease-associated alleles correlate with quantitative expression levels of the prostaglandin receptor EP4, PTGER4, the gene that resides closest to the associated region. Our results identify a major new susceptibility locus for CD, and suggest that genetic variants associated with disease risk at this locus could modulate cis-acting regulatory elements of PTGER4.

Introduction

Crohn disease (CD) is a chronic relapsing inflammatory disorder of the intestinal tract, described for the first time in the 1920s [1]. Lifetime prevalence has increased to current estimates of ~0.15% in Caucasians. The precise environmental causes underlying this rise remain essentially unknown, but familial clustering and twin studies clearly identify an inherited component to predisposition. More than ten susceptibility loci have been identified by linkage and/or association studies and convincing causative mutations have been reported, particularly in CARD15 [2–3]. As known loci don't fully account for the genetic risk for CD we performed a genome-wide association scan (GWA) to contribute to the identification of additional susceptibility loci.

Results/Discussion

Genotype data from the Illumina HumanHap300 Genotyping Beadchip [4] were obtained on 547 Caucasian CD patients from Belgium and compared to genotypes for 928 healthy controls from Belgium and France. Genotype call rates were >93% for all individuals included in the study. Of the total 317,497 SNPs available, 15,046 with genotyping success rate of less than 96% or deviating from Hardy-Weinberg proportions in controls (Fisher’s exact test \( p \leq 10^{-3} \)) were eliminated.

* To whom correspondence should be addressed. E-mail: michel.georges@ulg.ac.be
from further analysis as it is known that less reliable markers generate spurious associations. For the remaining 302,451 SNPs, we compared allele frequencies between cases and controls as outlined in Methods.

Figure 1 shows the 10,000 most significant p-values obtained across the human genome. Regions on Chromosomes 1, 5, and 16 harbored clusters of markers with suggestive evidence of association at significance levels between $10^{-6}$ and $10^{-10}$. The significance of tests of association with these markers remained within this range after controlling for possible effects of population structure using a backwards stepwise regression [5]. The strongest association was found with markers of the IL23R gene on Chromosome 1, which has recently been identified as a novel CD susceptibility locus in a case-control and family-based association study of Caucasian and Jewish cohorts [6]. In our data, two markers of the IL23R gene, rs11209026 and rs11465804, gave the most significant association signals ($p < 10^{-9}$). Rs11209026 corresponds to an Arg381Gln substitution in IL23R while rs11465804 is intronic and in strong linkage disequilibrium (LD) with the former marker. A marker within the CARD15 gene on Chromosome 16, which is the first susceptibility gene to have been identified in CD [3], also showed suggestive evidence of association (rs5743289; $p < 10^{-6}$). We also examined the results of the GWA with respect to other previously reported susceptibility loci, including OCTN [7], DLG5 [8], TNFSF15 [9], and ATG16L1 [10]. None of these obtained a similar level of significance for association in our study. Genotyping our cohorts for other SNPs at these loci that are reported in the literature to be associated with CD did not improve the signals, with the exception of rs224188 corresponding to a Thr-to-Ala substitution within ATG16L1 ($p < 2 \times 10^{-4}$), thus providing confirmation of this novel susceptibility locus for the first time [10].

On Chromosome 5p13.1, we identified a region of
approximately 250 kb that contained six markers with \( p < 10^{-6} \) in the association test (Figure S1). This region has not previously been reported as a CD susceptibility locus. We selected ten markers from the regions of \( IL23R \) and 5p13.1 for confirmation genotyping in up to 1,266 additional Caucasian CD patients and 559 additional controls. The \( IL23R \) locus was included in the confirmation genotyping as it had not yet been reported at the time of our study [6]. The associations at these two loci were clearly replicated with \( p \)-values as low as \( 4.2 \times 10^{-15} \) at \( IL23R \) and \( 3.7 \times 10^{-4} \) at 5p13.1 (Table 1). In the combined data from the GWA and replication studies, we obtained \( p \)-values as low as \( 2.2 \times 10^{-18} \) at \( IL23R \) and \( 2.1 \times 10^{-12} \) at the 5p13.1 locus. In addition, we genotyped trios with non-affected parents for the same SNPs to perform a transmission disequilibrium test (TDT). The ten SNPs were typed on 137 trios with affected offspring included in the case-control study, while two of the 5p13.1 SNPs were typed on an additional 291 independent trios, also originating from Belgium. Significant over-transmission of the associated alleles were found at both loci, thus providing additional confirmatory evidence in support of the \( IL23R \) and 5p13.1 susceptibility loci (Table 1).

To further characterize the novel 5p13.1 locus, we genotyped a subset of 1,092 CD patients and 374 Belgian controls for 111 markers (average interval: 2.3 kb) spanning the 250-kb segment. We determined the most likely linkage phase for each individual using PHASE [11], and used the corresponding haplotype frequencies to quantify the level of

---

**Table 1. Results of Primary and Confirmatory Association Analysis and TDT for the \( IL23R \) and 5p13.1 Loci**

<table>
<thead>
<tr>
<th>Locus</th>
<th>SNP*</th>
<th>Primary Data</th>
<th>Confirmatory Data</th>
<th>Combined Data</th>
<th>TDT</th>
</tr>
</thead>
</table>
|       |      | Controlsb | Casesc | Controlsb | Casesc | Controlsb | Casesc | Controlsb | Casesc | p
|       |      |          |          |            |          |            |          |            |          |     |
|       |      | 0.915    | 0.971   | 0.934    | 0.970   | 0.922    | 0.970   | 16.4d | 137a |
|       |      | 0.98     | 3.2E-8  | 0.96     | 1.7E-5  | 0.99     | 3.5E-15 | 0.04f |
|       |      | 0.918    | 0.972   | 0.934    | 0.972   | 0.924    | 0.972   | 17.5  |
|       |      | 0.93     | 1.5E-8  | 0.64     | 4.2E-7  | 0.99     | 2.2E-18 | 0.045 |
|       |      | 0.641    | 0.712   | 0.655    | 0.722   | 0.646    | 0.719   | 76.39 |
|       |      | 0.88     | 3.0E-4  | 0.32     | 2.9E-4  | 0.87     | 2.3E-9  | 0.0003 |
|       |      | 0.291    | 0.354   | 0.31     | 0.36    | 0.30     | 0.36    | 69.44 |
|       |      | 0.91     | 0.002   | 0.75     | 0.015   | 0.73     | 2.4E-6  | 0.009 |
|       |      | 0.589    | 0.686   | 0.629    | 0.697   | 0.604    | 0.673   | 72.64 |
|       |      | 0.24     | 5.1E-7  | 0.53     | 0.067   | 0.82     | 6.6E-7  | 0.05  |
|       |      | 0.46     | 9.1E-8  | 0.45     | 0.04    | 0.95     | 1.7E-9  | 0.040 |
|       |      | 0.12     | 0.191   | 0.139    | 0.183   | 0.127    | 0.185   | 139:113 |
|       |      | 0.99     | 6.1E-7  | 0.13     | 0.37    | 0.12E-9  | 0.050   |
|       |      | 0.666    | 0.762   | 0.685    | 0.742   | 0.673    | 0.748   | 61:46 |
|       |      | 0.30     | 9.7E-8  | 0.91     | 1.8E-3  | 0.62     | 9.2E-11 | 0.073 |
|       |      | 0.585    | 0.690   | 0.607    | 0.674   | 0.593    | 0.679   | 214:177 |
|       |      | 0.13     | 4.1E-8  | 0.89     | 3.7E-4  | 0.43     | 2.1E-12 | 0.030 |
|       |      | 0.651    | 0.746   | 0.675    | 0.708   | 0.659    | 0.720   | 66:43 |
|       |      | 0.60     | 2.2E-7  | 0.99     | 0.134   | 0.71     | 6.6E-7  | 0.013 |
|       |      | 1.59     | 1.17    | 1.17     |          |          |          |

*Chromosomal position from March 2006 assembly indicated in parentheses.

For each SNP: line 1, Allelic frequency of risk allele; line 2, number of individuals with genotype; and line 3, \( p \)-value of Hardy-Weinberg proportions (Fisher’s exact test).

For each SNP: line 1, Allelic frequency of risk allele; line 2, number of individuals with genotype; line 3, \( p \)-value of allelic association (chi-squared test); and line 4, odds ratio.

Times transmitted:times non-transmitted.

Number of genotyped trios.

\( p \)-value of segregation distortion (one-sided chi-squared test).

Results under Primary Data were obtained after re-genotyping of the initial samples using the Taqman assay conducted to verify the Illumina genotypes.

doi:10.1371/journal.pgen.0030058.t001
LD between all marker pairs. The 250 kb encompass five clearly delineated LD blocks, the central one (block III) being the largest and spanning 122 kb (Figure 2A). We first performed single-marker association analyses. The strongest effects were observed within the 122-kb block III with several SNPs yielding \( p \)-values \( < 10^{-5} \). \( p \)-values \( < 10^{-5} \) and \( 10^{-4} \) were observed in flanking blocks II and IV, respectively (Figure 2B).

We then performed haplotype analysis of the region spanned by blocks II to IV. For block III, 20 haplotypes accounted for 93\% of the observed chromosomes. These could be grouped in three clades comprising respectively six (IIA), six (IIB), and two (IIIC) haplotypes, plus a group of six haplotypes that apparently originated from various recombination events. Likewise, evaluation of block II revealed three clades (with respectively two [IIA], three [IIB], and two [IIIC] haplotypes) and two recombinant haplotypes, while block IV was characterized by two clades with two (IVa) and one (IVb) haplotype respectively. We compared the clade frequencies in cases and controls at intervals bounded by ancestral recombination events (Figure 2C). In agreement with the results of the single-marker analysis, the most significant associations were found in block III followed by IV and II. To verify whether the entire 5p13.1 effect could be attributed to block III (i.e., the effects observed for blocks II and IV would
be mere echoes of the block III effect), we performed a multi-
variante analysis as described in Methods. The clade effects of
blocks II and IV conditional on the effect of block III and vice
versa remained significant ($p_{III-II} = 0.023$, $p_{III-IV} = 0.0004$, $p_{IV-II} = 0.003$, and $p_{IV-III} = 0.026$), suggesting that multiple
variants in the region may jointly account for the observed
effect on CD. Commonly occurring recombinant haplotypes
in blocks II and III caused local drops in significance, thus
suggesting that causal variants lie outside the corresponding
subsegments (Figure 2C).

No known genes or CpG islands were found within the
region of association on 5p13.1 after examination with the
Ensembl and UCSC genome browsers. The region has an
average GC content of 38%, and an excess of interspersed
repeats given GC content (58.36% versus 42.3%), which is
mainly due to an excess of LINE1 (33.05% versus 19.6%) and
LTR elements (15.36% versus 7.70%) [12]. It contains 98
highly conserved elements [13]. It is part of a 1.25-Mb gene
desert between DAB2 (850 kb distally from the block) and
PTGER4 (270 kb proximally from the block). Interestingly,
several of the genes flanking the region have been implicated
in pathogenesis of CD, or are related to genes that have been
implicated in the disease. These include a member of the
caspase recruitment domain family (CARD6), three comple-
ment factors (C6, C7, and C9), and—most notably—the
prostaglandin receptor EP4 (PTGER4), which resides closest
to the group of disease associated markers (Figure S1).

One hypothesis is that the disease-associated region contains
cis-acting regulatory elements that control the expression
levels of the causal gene(s) located in the vicinity, and that the
causal variants modulate the activity of these elements. As a
first step to test this, we studied the effect of SNPs in the
disease-associated region on the expression levels of neigh-
boring genes. To that end we exploited a database of genome-wide
gene expression (Affymetrix HG-U133 Plus 2.0 chips; http://
www.affymetrix.com) measured in EBV-transformed lymphob-
blastoid cell lines from 378 individuals genotyped with the
Illumina HumanHap300 Genotyping Beadchip (W. Cookson,
unpublished data). Remarkably, eight of the 26 Illumina
markers spanning 264 kb coinciding precisely with the CD-
associated region yielded $p$-values $< 2 \times 10^{-5}$ for PTGER4
(Figure 2B). Three of the markers influencing PTGER4
expression are located in block III (rs16869977, rs10512739,
and rs6880934). The first two are tagging the IIIBa sub-clade
(Figure 2C), while the third one is in complete LD with it (IIIA+
IIIBa). The corresponding SNPs did not show convincing
evidence for association with CD. Two strongly associated
SNPs ($D’ = 0.84$) located respectively in block IV (rs4495224)
and V (rs7720838) showed the most significant effect on
PTGER4 expression and were also associated with CD (Table 1).
The rs4495224 A and rs7720838 T risk alleles were associated
with increased PTGER4 expression. Although these results
must be treated as preliminary, they tend to support the
hypothesis that the disease-associated polymorphisms may be
related to the expression levels of one or more genes in the
region.

CD is the most common form of inflammatory bowel
disease, the other being ulcerative colitis (UC). We genotyped
a cohort of 240 Belgian UC patients (Caucasians) for IL23R
(rs11290926), ATG16L1 (rs2241880), and the novel 5p13.1
locus (rs4613763). Consistent with published results [6,10]
we found a significant association for IL23R ($p = 1.2 \times 10^{-5}$; odds
ratio: 2.51) but not for ATG16L1 ($p = 0.78$). There was no
effect of the novel 5p13.1 locus on UC ($p = 0.54$). While
additional studies will be needed to exclude completely a role
in UC, these results suggest that the principal susceptibility
effects of the 5p13.1 locus are for CD. The restriction to CD
risk observed for ATG16L1 and the 5p13.1 locus is similar to
that found for CARD15 [3].

We herein describe the localization of a novel major
susceptibility locus for CD on 5p13.1 by GWA. The region
of strongest association coincides with a gene desert devoid
of known protein-coding genes. The observed effect may be
mediated by as-yet unknown transcripts mapping within
the region. As a matter of fact, limited numbers of spliced
and unspliced ESTs originating from the HT1080 fibrosarcoma
cell line or medulla (e.g., BG182136 and BG184600) map to
the region. An alternative explanation, however, is that the
disease-associated region contains cis-acting elements con-
trolling the expression of more distant genes. We provide
evidence in support of this hypothesis by demonstrating that
genetic variants in the CD-associated region differentially
regulate the expression levels of PTGER4, the closest known
gene located 270 kb proximally, PTGER4 is a strong candidate
gene for CD, as it is known that knock-out mice develop
severe colitis upon dextran sodium sulphate treatment,
unlike mice deficient in any of the seven other types of
prostanoid receptors. Increased susceptibility to colitis is also
observed in wild-type mice administered an EP4-selective
agonist, while EP4-selective agonists are protective [14].
We observe in particular that the CD susceptibility allele at
marker rs4495224 is associated with increased PTGER4
transcript levels in lymphoblastoid cell lines. This finding
establishes a direct link between disease susceptibility and
PTGER4 expression, although the direction of the effect
apparently contradicts the results in knock-out mice. De-
tailed studies of the effect of genetic variants in the disease-
associated region on PTGER4 expression in different tissues
and of a possible connection between PTGER4 levels and CD
susceptibility are certainly needed and work towards that goal
is in progress. The hypothesis that the 5p13.1 CD-suscepti-
bility locus operates by modulating PTGER4 expression levels
could—at least in theory—be tested by replacing the
(corresponding murine sequences with the human ortholo-
gous variants and quantitatively complement the murine
knock-out allele [15]. Our results suggest that the 5p13.1 effect
on CD could result from the combined action of multiple susceptibility variants. Extensive sequencing of the
most common haplotypes in the region of association is being
directed towards their identification.

Materials and Methods

Genotyping. Genotyping for the whole genome scan was performed on a Illumina HumanHap300 Genotyping Beadchip (http://
www.illumina.com) [4]. Genotyping of individual SNPs was performed on an ABI7900HT Sequence Detection System using TaqMan MGB
probes from Pre-designed SNP Genotyping or Custom TaqMan SNP Genotyping assays (Applied Biosystems, http://www. appliedbiosystems.com).

Association analyses. Association analyses were conducted using
Fisher’s exact test (whole genome scan) or chi-squared tests of
independence (confirmation analysis). We applied the logistic
regression method of Setakis et al. [5] to test for the possible effect
of population structure on the most significant association results.
The 110 control markers included in the logistic regression had 100%
genotype success rate with minor allele frequency $>30\%$, and no two

Whole Genome Association for Crohn Disease
markers were within 20 Mb of one another. To test for an effect of block I on the effect of an adjacent block II, we compared the proportion of I haplotype clades nested within a given II clade (for instance, proportion of IA, IB and IC within IIA) between cases and controls by chi-squared. Chi-squared values (and degrees of freedom) were summed across II clades to yield an overall (I|II) test statistic.

Expression database. The database genome-wide expression analysis data was provided by W. Cookson (Imperial College, London, United Kingdom). Briefly, expression data were generated from RNA extracted from EBV-transformed cells from 378 genotyped offspring in nuclear families. Annotations for individual transcripts on the Affymetrix arrays were extracted from the Affymetrix NetAffx database (http://www.affymetrix.com). Data from the gene expression experiment were normalized together using the RMA (Robust Multi-Average Array) package [16,17] to remove any technical or spurious background variation. An inverse normalization transformation step was also applied to each trait to avoid any outliers. A variance components model was used to estimate heritability of each trait using Merlin-regress (RandomSample option) [18, 19]. For PTGER4, we obtained a mean quantitative expression value of ~0.017 and a variance of 0.722 while the heritability estimate for PTGER4 estimated using the sibship data was 0.84. Association analysis was applied with Merlin (FASTASSOC option). We estimated an additive effect for SNPs and tested its significance using a score test that adjusts for family structure and takes into account uncertainty in the inference of missing genotypes.

Supporting Information

Figure S1. Log(p) Values for Illumina SNPs Located in a ~2-Mb Region Centered around the 122-krb Block III (Marked by Solid Vertical Lines) Described in Detail in Figure 2

Found at doi:10.1371/journal.pgen.0030058.s001 (418 KB PPT).

References


Acknowledgments

We are grateful to Véronique Dhennin, Dimitri Pirottin, and Stéphanie Glineur for their assistance, and to all the clinicians that took part in patient recruitment. (Erasmu collaborators: Jean-Marc Maisin, Vinciane Mub, Jean Van Cahter, Marc Van Goossen, Philippe Closset, Pierre Hayard, and Jean Michel Ghilain; University of Liège collaborators: Paul Mainguet, Faddy Mokaddem, Fernand Fontaine, Jacques Deflandre, and Hubert Demolin). Sincere thanks to W. Cookson for providing us access to the genome-wide expression data prior to publication.

S. Hansoul and C. Sandor contributed equally to this work.


Funding. This work was supported by grants from the Région Wallonne, the Communauté Française de Belgique (Game and Biomod actions de recherche concertées), the Belgian Science Policy organization (Geneunc and Biomagnet pôles d’attraction interuniversitaires), and the University of Liège. E. Louis, S. Hansoul, and D. Franchimont are fellows of the Fonds National de la Recherche Scientifique. S. Vermeire is a fellow of the Nationaal Fonds voor Wetenschappelijk Onderzoek and C. Sandor is a fellow of the Fonds pour la formation à la Recherche dans l’Industrie et dans l’Agriculture.

Competing interests. The authors have declared that no competing interests exist.