Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease

John D. Rioux[1], Mark J. Daly[1], Mark S. Silverberg[2,3], Kerstin Lindblad[1], Hillary Steinhart[2], Zane Cohen[4], Terrye Delmonte[1], Kerry Kocher[1], Katie Miller[1], Sheila Guschwan[1], Edward J. Kulbokas[1], Sinead O'Leary[1], Ellen Winchester[1], Ken Dewar[1], Todd Green[1], Valerie Stone[1], Christine Chow[1], Albert Cohen[7], Diane Langelier[8], Gilles Lapointe[9], Daniel Gaudet[9], Janet Faith[7], Nancy Branco[7], Shelley B. Bull[6], Robin S. McLeod[4], Anne M. Griffiths[5], Alain Bitton[7], Gordon R. Greenberg[2], Eric S. Lander[1,10,*], Katherine A. Siminovitch[2,3,*] & Thomas J. Hudson[1,7,*]

These authors co-directed the project.

Linkage disequilibrium (LD) mapping provides a powerful method for fine-structure localization of rare disease genes, but has not yet been widely applied to common disease[1]. We sought to design a systematic approach for LD mapping and apply it to the localization of a gene (*IBD5*) conferring susceptibility to Crohn disease. The key issues are: (i) to detect a significant LD signal (ii) to rigorously bound the critical region and (iii) to identify the causal genetic variant within this region. We previously mapped the *IBD5* locus to a large region spanning 18 cM of chromosome 5q31 ($P<10^{-4}$). Using dense genetic maps of microsatellite markers and single-nucleotide polymorphisms (SNPs) across the entire region, we found strong evidence of LD. We bound the region to a common haplotype spanning 250 kb that shows strong association with the disease ($P<2\times10^{-7}$) and contains the cytokine gene cluster. This finding provides overwhelming evidence that a specific common haplotype of the cytokine region in 5q31 confers susceptibility to Crohn disease. However, genetic evidence alone is not sufficient to identify the causal mutation within this region, as strong LD across the region results in multiple SNPs having equivalent genetic evidence—each consistent with the expected properties of the *IBD5* locus. These results have important implications for Crohn disease in particular and LD mapping in general.

The majority of inflammatory bowel disease (IBD) patients can be classified as having either Crohn disease or ulcerative colitis. Both are idiopathic inflammatory diseases of the bowel associated with distinct clinical and pathological profiles. Our recent attempts to identify IBD susceptibility loci in 158 nuclear families (Table 1, set A) provides evidence for a gene (*IBD5*) conferring susceptibility to Crohn disease in chromosome 5q31 (ref. 2; Fig. 1).

We used a hierarchical strategy to search for a signal of LD from *IBD5*. We studied 256 father–mother–child trios, where the child had Crohn disease and at least one parent was unaffected (Table 1, set B). We genotyped 56 microsatellite polymorphisms distributed throughout the region at an average spacing of approximately 0.35 cM. We examined each allele of each marker for evidence of transmission disequilibrium using the transmission disequilibrium test (TDT)[3]. Two loci have alleles with significant TDT results: *IRF1p1* (P=0.00016) and *D5S1984* (P=0.00039; Web Table A). Notably, the two loci are adjacent to one another and within 1 cM of *D5S2497*. The observed transmission ratio of the risk allele is roughly 1.8:1. Permutation testing shows that these observations are highly significant (empirical P value=0.00091). Because the TDT tests of association are completely independent of our previously

Table 1 • Summary of pedigrees used in the SSLP and SNP genotyping

Pedigrees	Number of pedigrees/trios	Genotyping performed	Population	Age of diagnosis (average; median)	Description of pedigrees
set A	122 CD pedigrees[a]	51 SSLPs	Toronto	17.2; 18.0	50/122 linkage families having an affected offspring with age of diagnosis ≤16
set B	256 CD trios	64 SSLPs	Toronto	18.4; 16.0	124 trios from original linkage families (set A), with only one trio per family to ensure independence for LD analysis 132 new trios (not in set A)
set C	139 CD trios	301 SNPs	Toronto	16.3; 15.0 subset C1: 15.4; 14.0 subset C2: 17.2; 16.0	139 trios consisting of two subsets: subset C1: 63 trios from set B (18/63 from original linkage families) subset C2: 76 new trios (not in set A or B)
set D	88 CD trios	11 significant SNPs[b]	Quebec	23.0; 21.5	88 trios independently collected for current study

[a]As described in ref. 2. [b]Set of 11 significant SNPs as described in Table 2B. CD, Crohn disease.

[1]*Whitehead Institute/Massachusetts Institute of Technology, Center for Genome Research, Cambridge, Massachusetts, USA.* [2]*Department of Medicine,* [3]*Immunology and Medical Genetics,* [4]*Surgery,* [5]*Pediatrics,* [6]*Public Health Sciences, University of Toronto and the Mount Sinai Hospital Lunenfeld Research Institute and the Hospital for Sick Children, Toronto, Ontario, Canada.* [7]*Department of Gastroenterology and Montreal Genome Center, McGill University Health Center, McGill University, Montréal, Québec, Canada.* [8]*Centre Hospitalier de Sherbrooke, Sherbrooke, Quebec, Canada.* [9]*Centre Hospitalier de la Sagamie, Chicoutimi, Quebec, Canada.* [10]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to J.D.R. (e-mail: rioux@genome.wi.mit.edu), K.A.S (e-mail: ksiminovitch@mtsinai.on.ca), T.J.H. (e-mail: tjhudson@med.mcgill.ca) or E.S.L. (e-mail: lander@wi.mit.edu).*

reported linkage result, they provide strong confirmation of the presence of a susceptibility gene for Crohn disease.

Having found evidence of LD, we studied a denser collection of markers in the implicated region to confirm our results. For this purpose, we developed new microsatellite markers from the 680 kb of DNA sequence available at that time for this region[4] and genotyped the same 256 trios (Fig. 1 and Table 2A). Another significant TDT result occurs at *CAh17a* (P=0.0002), located between the two previous high points (*IRF1p1* and *D5S1984*). Analysis of two- and three-marker haplotypes increases the strength of evidence for a risk locus with nearly uninterrupted LD across the region from *GAh18a* to *CSF2p10* (Fig. 2). The region near *IRF1p1–CAh15a–CAh17a* shows the strongest evidence (nominal P values ~3×10^{-6}) and an estimated transmission ratio greater than 2:1.

We then attempted to identify the causal allele by examining all known genes within the critical region (and any additional plausible candidates just beyond this region) for allelic variants that could confer increased susceptibility to Crohn disease. The chronic inflammation in the gastrointestinal tract in individuals with Crohn disease is thought to be at least partly due to the interaction between the host immune system and enteric microflora normally present across the mucosal wall. It is therefore notable that the critical region contains the cytokine gene cluster, which includes many plausible candidate genes for inflammatory disease (Fig. 1). We studied the 11 known genes from *IL4* to *IL3* by resequencing the complete transcribed regions (5′ UTR, coding and 3′ UTR) as well as 1 kb upstream of the transcription start site. We identified a total of 16 SNPs (Web Table B). Two show significant TDT results, but neither seems to be a strong candidate for *IBD5*: a silent substitution in the coding region of *OCTN2* (P=0.004) and a missense substitution (Thr→Ile) in *OCTN1* (P<0.003). The latter seems unlikely to have a severe effect on the protein, inasmuch as isoleucine is found in the analogous position in mouse *Octn1* and in mouse and rat *Octn2* and human *OCTN2* (ref. 5). Moreover, subsequent analysis of the region turns up SNPs with stronger evidence of transmission disequilibrium with the Crohn disease phenotype and shows that these two SNPs are not unique to the risk haplotype.

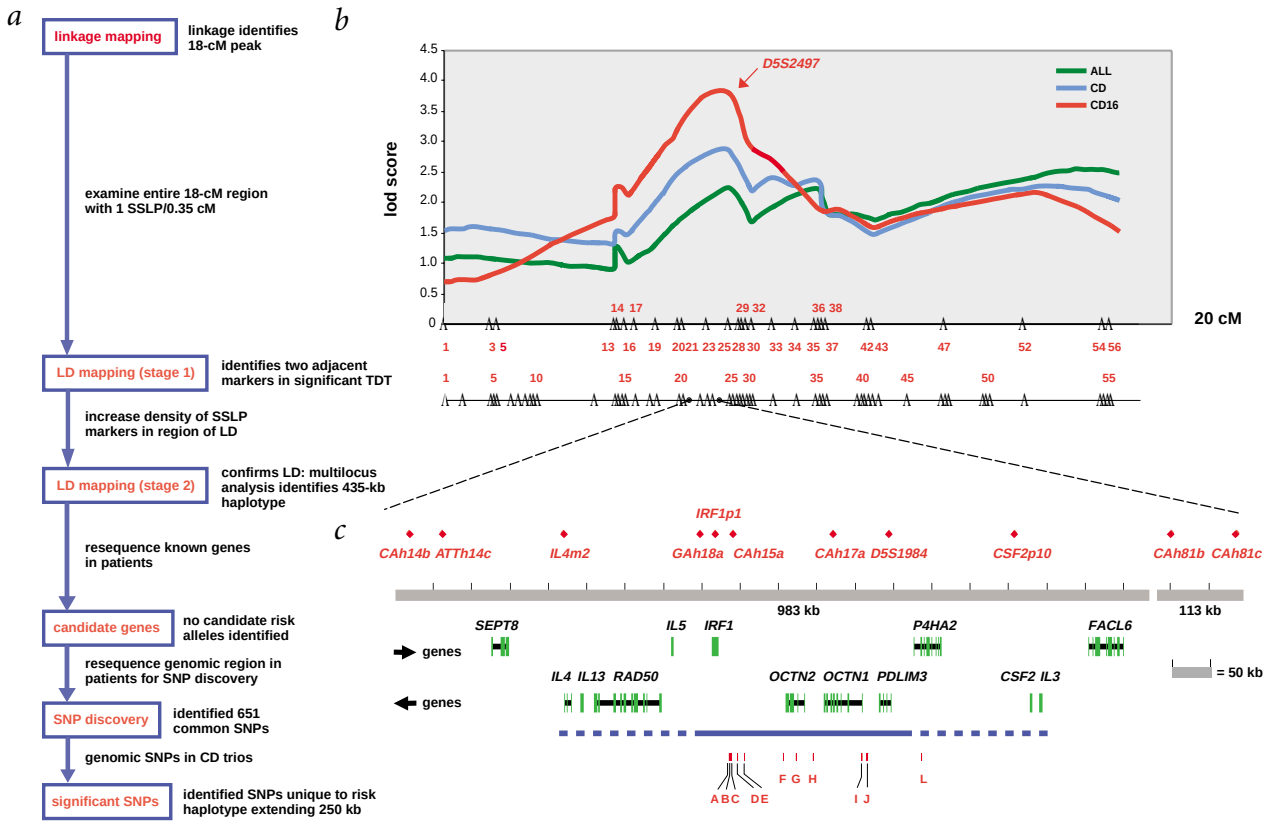

Fig. 1 Linkage and LD mapping. ***a***, Experimental scheme; ***b-c***, results and analysis. ***b***, Linkage evidence from the initial genome-wide scan for the different disease subgroups: all, all families with IBD; CD, all families with Crohn disease only; CD16, families with early-onset Crohn disease (at least one sibling with age of diagnosis ≤16)[2]. Linkage was strongest (lod score=3.9) in families with early-onset cases. *IBD5* mapped to an 18-cM region—with a maximal lod score at marker *D5S2497* and bounded by *D5S1435* and *D5S1480* (at which the lod score falls by two units, corresponding to 100-fold lower likelihood). Vertical tick marks indicate the position of the markers in the genome-wide linkage study and numbers in red refer to the marker numbers used in Table 1. Vertical tick marks on the thin horizontal line below the graph represent the position of all 56 markers used in the first stage of LD mapping in the 256 Crohn disease trios. These markers are numbered (shown in red where space available) in map order; the same numbers are used in Table 2A and Web Table A. ***c***, Expansion of the region with significant LD. The thick grey horizontal line depicts the two sequence contigs (with their lengths indicated below), with the break representing the gap between them; names and positions (indicated by red diamonds) of the microsatellite markers used in the first and second stage of LD mapping are shown above the line; and known genes are shown below the line. Exon positions and lengths are indicated by vertical green bars (drawn to scale; not all exons are distinguishable) and gene symbols are written above each gene. The thick blue line below the genes represents the region where SNP discovery was performed by resequencing DNA samples from 8 individuals (7 with Crohn disease and one CEPH DNA control). The blue line is continuous where the discovery was carried out on every base over a 285-kb contiguous ('core') region and dashed where the discovery was in noncontiguous ('proximal' and 'distal') regions. Red tick marks beneath the blue line indicate the positions of the SNPs that have alleles unique to the risk haplotype, where A=IGR2055a_1; B=IGR2060a_1; C=IGR2063b_1; D=IGR2078a_1; E=IGR2096a_1; F=IGR2198a_1; G=IGR2230a_1; H=IGR2277a_1; I=IGR3081a_1; J=IGR3096a_1; K=IGR3236a_1 (see Table 2B).

Because no obvious candidates emerged from the analysis of known genes, we undertook a comprehensive analysis of the entire critical region. We assembled a reference sequence of 983 kb and undertook systematic SNP discovery by direct resequencing of PCR products from eight individuals. This SNP discovery effort aimed for complete ascertainment in a central 'core' region of 285 kb and partial ascertainment in the surrounding 'proximal' and 'distal' regions (Fig. 1). We identified a total of 651 candidate SNPs (Web Table C).

We genotyped a large portion of these SNPs to define the risk haplotype and to search for candidates with the *IBD5* mutation. To date, 301 of the SNPs were genotyped (Web Table C) on 139 trios (Table 1, set C). Of these trios, 63 (set C1) were taken from the previously analyzed set B and 76 (set C2) were newly collected samples. Set C1 allowed the SNP haplotypes to be integrated with the previously seen microsatellite-based haplotypes; we observed extremely close correspondence (as several SNPs with alleles nearly unique to the microsatellite-defined risk haplotype show strong TDT). Set C2 provided an independent test of association and those same alleles show strong evidence of transmission distortion ($P<0.0002$).

Using this ultra-high-density SNP map, it is possible to determine the fine structure of the haplotypes in this region. These observations are interesting in their own right, as they shed light on the underlying haplotype structure in the human genome and suggest a framework for utilizing these structural properties. This framework is described in a companion paper by Daly *et al.*[6]. Specifically, the analysis shows that the region can be parsed into haplotype blocks, with each block having limited diversity (2–4 haplotypes account for 90–98% of all chromosomes), and extensive LD across the entire region. Multilocus analyses define a single common risk haplotype (frequency of 37% among untransmitted chromosomes) that shows a maximal transmission ratio of 2.5:1 (Fig. 3).

With essentially perfect haplotype information (as in the present case), simulations show that the transmitted/untransmitted (T/U) ratio at the disease locus has a 99% probability of being within 0.5 of the observed peak (T/U=2.5) and a 90% probability of being within 0.25 of this value, thus delimiting a region of approximately 350 kb or 250 kb, respectively.

We tested whether the properties of the risk haplotype were adequate to explain the observed genetic properties of *IBD5*. Specifically, the best fit is a model in which one copy of the risk allele increases risk for Crohn disease by two-fold and two copies by six-fold. Simulation tests demonstrate that such a susceptibility locus could have given rise to our observed linkage data. A rarer, more penetrant, allele on a subset of this risk haplotype can be excluded, as this would result in a much higher lod score than observed. The properties of the observed risk haplotype are consistent with those inferred for the *IBD5* locus. The haplotype can thus serve as a proxy for *IBD5*.

As all of the other common haplotypes in this region are under-transmitted, the causative allele must be unique to the risk haplotype. Notably, multiple SNPs meet this genetic criterion. Of

Table 2A • LD mapping using microsatellite markers: identification of peak LD

Marker #	Marker name	Distance to next marker (kb)	LD mapping stage	Allele	TDT results T/U	χ^2	*P* value
57	*CAh14b*	43	2	–	–	–	–
58	*ATTh14c*	167	2	–	–	–	–
59	*IL4m2*	164	2	–	–	–	–
60	*GAh18a*	21	2	193	1.5	5.70	0.017
22	*IRF1p1*	24	1	156	1.9	14.25	0.00016
61	*CAh15a*	130	2	373	1.5	4.0	0.045
62	*CAh17a*	97	2	140	1.7	13.80	0.00020
23	*D5S1984*	163	1	222	1.8	12.57	0.00039
24	*CSF2p10*	N.D.	1	307	1.5	4.52	0.033
63	*CAh81b*	85	2	–	–	–	–
64	*CAh81c*		2	–	–	–	–

N.D, no data; –, no allele in transmission disequilibrium.

Table 2B • LD mapping using SNPs: summary of SNPs with significant TDT results

SNP marker name	Approximate physical position[a]	SNP type	Transmitted allele	Frequency of allele[b]	T/U[c]	χ^2	*P* value	Comment
IGR2055a_1	435.0	G/T	G	0.357	87:39	18.29	0.000019	
IGR2060a_1	437.5	C/G	C	0.351	81:34	19.21	0.000012	
IGR2063b_1	439.0	C/G	G	0.359	87:37	20.16	0.000007	Predicted to cause a missense change in a GENSCAN[c]-predicted gene
IGR2078a_1	446.5	A/G	A	0.364	48:16	16.00	0.000063	Located within the 'E3' EST, which may be a splice variant of *OCTN2* (ref. 16)
IGR2096a_1	455.5	A/C	A	0.349	75:32	17.28	0.000032	
IGR2198a_1	506.5	C/G	G	0.364	87:41	16.53	0.000048	
IGR2230a_1	522.5	C/T	T	0.415	67:28	16.01	0.000063	
IGR2277a_1	546.0	A/G	G	0.417	79:37	15.21	0.000096	
IGR3081a_1	609.0	G/T	G	0.338	79:35	16.98	0.000038	
IGR3096a_1	616.5	C/T	C	0.429	89:42	16.86	0.00004	Located within EST cluster Hs.59096
IGR3236a_1	686.5	G/T	T	0.383	79:39	13.56	0.00023	

[a]Position (kb) on the 983-kb reference sequence. [b]Frequency of allele calculated from the untransmitted parental chromosomes. [c]Ratio of the number of transmitted (T) chromosomes to untransmitted (U) chromosomes. [c]http://genes.mit.edu/GENSCAN.html.

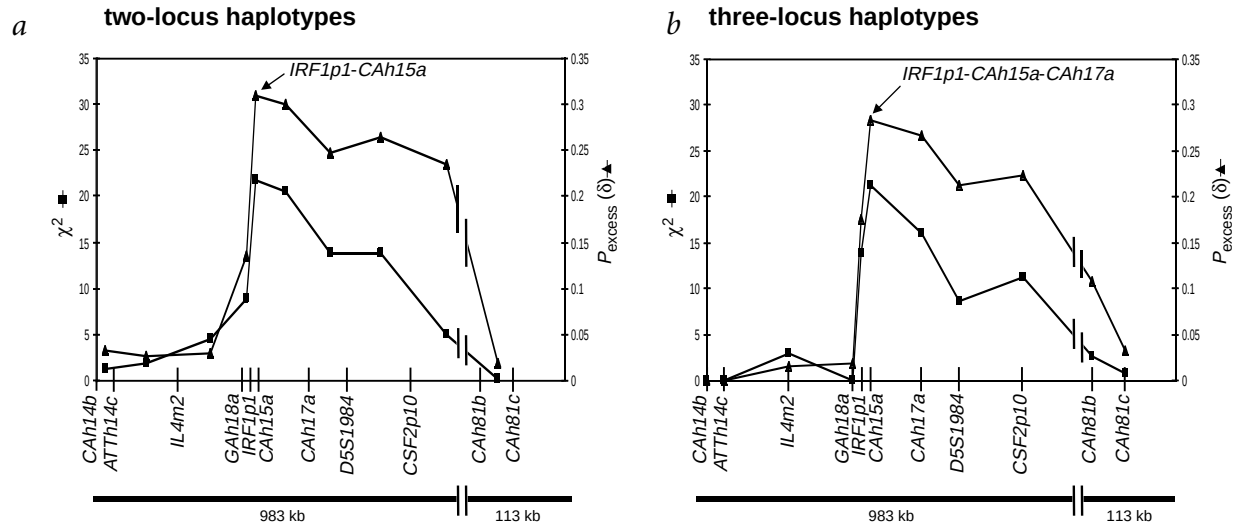

Fig. 2 Multilocus haplotype results. The curves represent the extent of association to the Crohn disease phenotype observed over the 1-cM region surrounding *IBD5* using data from the microsatellite markers described in Table 2A. The multilocus LD was measured using TDT (squares; values on left-hand *y* axis) or by P_{excess} (triangles; values on right-hand *y* axis). Tick marks along the *x* axis represent the positions of each marker. The thick black line and marker names and numbers are as described in Fig. 1. The arrow points to the peak LD in this region, observed for the haplotype formed by the *IRF1p1*, *CAh15a* and *CAh17a* markers (alleles 156, 373 and 140, respectively). ***a***, Two-locus haplotypes: results are shown for all pairs of adjacent markers where the data points are drawn at the midpoint between the two markers. ***b***, Three-locus haplotypes: results are shown for all possible combinations of adjacent markers where the data points are drawn at the position of the middle marker.

301 SNPs genotyped to date (including all SNPs in known or predicted genes), 11 had alleles that were unique to the risk haplotype. Each of these 11 SNPs shows a significant level of TDT on its own (even after conservative correction for multiple testing; Table 2B); all are essentially identical in their information content by virtue of being in nearly total LD with one another (with the allele at one SNP determining the allele at all others on nearly every phased chromosome). We are continuing to genotype the remaining 256 SNPs in the core region and expect to find approximately 12 more SNPs unique to the risk haplotype with properties identical to these first 11. The equivalent properties of these SNPs make it impossible to identify the causal SNP on the basis of genetic evidence alone.

Finally, we sought additional confirmation of our findings by examining these 11 SNPs in yet another data set derived from a different population (Table 1, set D). We observed that all of the same alleles were over-transmitted in this independent data set and that the risk haplotype had a T/U ratio of 1.75:1 (P=0.02), statistically indistinguishable from the ratio seen above. These results provide another independent replication of the findings of a risk haplotype for Crohn disease. We also confirmed that no transmission bias is seen in meioses from unaffected families (data not shown).

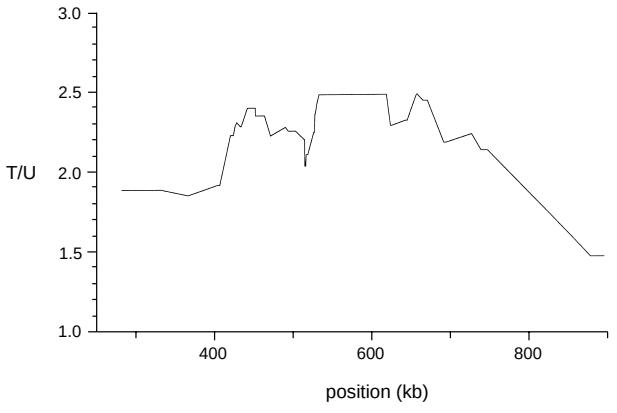

Fig. 3 Multipoint T/U plot for the *IBD5* risk haplotype. This curve represents the transmitted to untransmitted ratio (T/U) of the *IBD5* risk haplotype identified with high-density SNP genotype information for the individuals in set C (Table 1). Ancestral haplotype blocks were discovered in this region (kb positions are as per our 983-kb reference sequence) and multipoint TDT was carried out using previously described methods.

This study represents the first comprehensive application of hierarchical LD mapping involving a systematic search for LD across a linkage peak, rigorous bounding of the critical region and exhaustive ascertainment of SNPs in the critical regions. These results provide overwhelming evidence that a specific haplotype of the cytokine gene cluster on 5q31 is a risk factor for Crohn disease (Table 3).

We have not been able to implicate a single causative mutation, because the tight LD across the region results in at least 11 SNPs having equivalent genetic information. Most of these SNPs are likely to simply be fellow travelers on the risk haplotype. None disrupts a crucial amino acid or regulatory region of a known gene. The causal SNP or SNPs may, however, affect regulation of one or more known genes, many of which are plausible candidates. Specifically, the immunoregulatory cytokines (*IL4*, *IL5*, *IL13*) are important for the T_H1/T_H2 balance, and a disturbance in this balance is believed to have a role in the pathophysiological process. In addition, individuals with Crohn disease have increased expression of *IRF1* and increased enzymatic activity of *P4HA2*[7,8]. Alternatively, the causal SNP may affect an as-yet-unidentified gene in the cytokine cluster. Biological studies will be necessary to resolve this. Interactions with other loci identified in Crohn disease and other inflammatory diseases may provide functional insight. Preliminary data collected in a subset of these individuals shows no obvious interaction between the *IBD5* haplotype and the mutations in *NOD2/CARD15* known to confer risk to Crohn disease[9,10].

Although the causal SNP has not yet been identified, the cytokine cluster can be treated as a tightly linked block analogous to the HLA region. Decades of research associate HLA haplotypes

with numerous inflammatory and autoimmune diseases, although the causal mutations elude definitive identification in most cases. The results of this study raise the possibility that haplotypes of the cytokine gene cluster may also confer risk for other inflammatory diseases. Association studies with the common haplotypes identified here should be thoroughly examined.

Methods

Families. We carried out all of the analyses in this study with father–mother–affected child (Crohn disease only) trios, where 0 or 1 of the parents was affected with Crohn disease. These trios either came from the multicase families used in our previous linkage study that identified the *IBD5* locus[2] or were identified for the purpose of the current LD study. Specifically, for the microsatellite genotyping, we genotyped 256 triads: 124 from families used in the original identification of the *IBD5* locus (only one trio per family) and 132 from newly collected families. For the SNP genotyping, we genotyped 139 trios: 18 from families used in the original identification of the *IBD5* locus and 121 from newly collected families. We identified individuals affected by Crohn disease by review of the clinical charts of all patients registered in the Mount Sinai Hospital IBD Centre patient database and from the Toronto Hospital for Sick Children IBD database. We obtained written informed consent from all participants; ethics approval for this study was granted by the University of Toronto Ethics Committee.

For confirmation of the TDT results in an independent population, we collected samples in the Canadian province of Quebec. We identified affected individuals for this collection on routine visits to their gastroenterologists in Chicoutimi (Complexe Hospitalier de la Sagamie), Montreal (McGill University Health Centre and Jewish General Hospital), Quebec City (Pavilion l'Hôtel Dieu de Quebec), and Sherbrooke (Centre Universitaire de Santé de l'Estrie). A diagnosis of Crohn disease consisted of clinical symptoms on two or more occasions and confirmation with endoscopic, radiologic or histologic evidence. A review of the patient's chart, as well as an interview with the patient, was done to complete the phenotypic data. Written informed consent was obtained from all participants and ethics approval was granted in each of the participating institutions.

Microsatellite genotyping. In the first stage of microsatellite LD mapping, a total of 56 microsatellite markers were genotyped on 256 Crohn disease triads. Information regarding primer sequence, allele size range and suggested amplification conditions for 54 of these genetic markers (excluding *IRF1p1* and *CSF2p10*) can be obtained from the Genethon (http://www.genethon.fr/), Marshfield (http://research.marshfieldclinic.org/genetics/) or Genome Database (http://www.genethon.fr) web sites. Information regarding the microsatellite markers we designed from available genomic sequence during the course of this study (*IRF1p1, CSF2p10* and the eight markers used in the second stage of LD mapping) can be obtained online (http://www.genome.wi.mit.edu/humgen/IBD5). We obtained genotypes for all of these markers by polymerase chain reaction (PCR) using fluorescently labeled primers. We amplified each locus separately and then multiplexed individual PCR products into panels by pooling on the basis of allele size range and fluorescent label. We mixed aliquots of the multiplexed samples with either TAMRA-labeled GeneScan 500 and GeneScan 2500 (PE Applied BioSystems) or rhodamine-labeled MapMarkers (Bioventures) prior to electrophoresis on ABI 377 sequencers (PE Applied BioSystems). We analyzed the genotyping gels in an automated system developed at the Whitehead Institute/MIT Center for Genome Research as previously described[11].

SNP discovery. In order to identify all SNPs in the *IBD5* critical region, we generated a 983-kb reference sequence by assembling the sequence reported by Frazer *et al.*[4] with more recent sequence (AC046165, AC023861, AC034228 and AC079320) from the Human Genome Project[12]. We then designed a tiling path of overlapping PCR products for SNP discovery by resequencing. Specifically, we targeted 285 kb of contiguous sequence in the core region delimited by markers *GAh18a* and *D5S1984* (18 kb before *GAh18a* to 27.5 kb after *D5S1984*). We also carried out SNP discovery on 100 kb of noncontiguous sequence 150 kb to the left (proximal region; from *IL4* to *GAh18a*) and on 85 kb of noncontiguous sequence 200 kb to the right (distal region; *D5S1984* to *IL3*) of the core region. In the proximal region, we designed sequencing assays primarily to cover exons of known genes and regions with significant homology (>100 bp with >80% identity) to the known mouse sequence syntenic to this region. In the distal region, we also designed sequencing assays to cover exons of known genes as well as to cover every other 500-bp segment from *D5S1984* to *IL3* (no syntenic mouse sequence available).

Table 3 • Independent statistical evidence providing support for *IBD5* locus

	Pedigrees	Analysis	*P* value*
linkage results	set A	ASP	$<10^{-4}$
LD mapping with SSLP markers	set B	TDT	$<2\times10^{-4}$
LD mapping with SNP markers	set C2	TDT	$<2\times10^{-4}$
replication of SNP results	set D	TDT	<0.05

*Combined *P* value for all three TDT samples is $<2\times10^{-7}$.

We designed PCR assays using Primer 3.0 to be ~700 bp in length, with 100-bp overlap with adjacent assays. The –21 M13 forward and the –28 M13 reverse sequences were added to each of the forward and reverse PCR primers, respectively. These PCR primers were used to amplify 50 ng of genomic DNA from six independent individuals with Crohn disease, one unaffected family member (5 of 7 Jewish; 2 of 7 non-Jewish; 4 of 7 early-onset cases; 4 of 7 with an affected parent) and one DNA sample from the Centre D'Etude du Polymorphisme Humain (CEPH) as control. We selected the seven individuals with Crohn disease to maximize the chance they carried mutations at the *IBD5* locus. Specifically, we obtained samples of Crohn disease from families showing linkage to chromosome 5q31. We also selected individuals based on whether they carried the risk haplotype at the six markers from *GAh18a* through *CSF2p10*. One individual was homozygous for the risk haplotype, three were heterozygous, and three did not appear to carry this haplotype. This set of samples reflects the diversity observed in the entire data set and was chosen to ensure the identification of the risk allele as well as SNPs in the overall sample collection. We purified the PCR products using the solid-phase reversible immobilization (SPRI) method[13] and sequenced them using the appropriate –21 M13 or –28 M13 DYEnamic Direct Cycle Sequencing kit (Amersham). All sequencing reactions were run on ABI377 automated sequencers (PE Applied BioSystems). We processed the gel files using BASS software, available on the Whitehead Institute/MIT Center for Genome Research FTP site (http://www-genome.wi.mit.edu/). We base-called sequences using the Phred program, and assembled forward and reverse reads using the Phrap program (The Phred/Phrap/Consed System Home Page, http://bozeman.mbt.washington.edu/index.html). At least two observers visually inspected all traces. A different pair of primers was designed for any assay that failed at either the PCR or sequencing step.

SNP genotyping. We carried out SNP genotyping using length-multiplexed single-base extension (LM-SBE) as previously described[14]. Briefly, we designed PCR primers as close as possible to the SNPs identified in the current study, resulting in a product of a maximum length of 150 bp. Forward primers had T7 tails at their 5′ ends and reverse primers had T3 tails at their 5′ ends. We used these T7 and T3 tails for secondary amplification. We checked primer pairs for homology to all amplicons and sorted them into pools consisting of up to 50 primer pairs. We subjected loci to two rounds of PCR amplification. In the first round, we amplified 10 ng of genomic DNA using a pool of primer pairs (0.1 μM) and 2.5 units of Amplitaq Gold (Perkin Elmer). In the second round, we amplified a 3-μl aliquot of the primary amplification product with biotinylated-T7 and biotinylated-T3 primers. We purified a 7-μl aliquot of this secondary amplification product from the unincorporated dNTPs using streptavidin-coated Dynabeads (Dynal). We then carried out a multiplex SBE reaction on the purified product using SNP-specific primers, JOE-ddATP (0.12 M), TAMRA-ddCTP (0.12 M), FAM-ddGTP (0.12 M), ROX-ddUTP (0.60 M; NEN DuPont) and Thermosequenase (0.5 U; Amersham). We removed excess ddNTPs from the SBE products using 96-well gel filtration blocks (Edge Biosystems) prior to electrophoresis on ABI 377 sequencers. We analyzed the SBE gels using a system developed at the Whitehead Institute/MIT Center for Genome Research as previously described[15].

Statistical analysis. To assess the significance of the TDT results for each marker, we carried out permutation tests using the same genotype data. For

each trio, we randomly reassigned chromosomes as transmitted or untransmitted to form a permuted data set. In 10^6 permutations of the entire data set of 56 markers, we observed a single-allele χ^2 value greater than 14.2 only 15,796 times (corresponding to an empirical *P* value of 0.016), and only 911 simulations had two markers with χ^2 value greater than 12.5 (corresponding to an empirical *P* value of 0.00091, corrected for multiple testing).

In order to quantify the extent of LD in the *IBD5* region, we examined three-marker haplotypes using TDT and P_{excess} (δ). P_{excess} represents the strength of LD and is calculated by $(P_{affected}–P_{normal})/(1–P_{normal})$. In our study, the $P_{affected}$ is calculated from the frequency of the haplotype among the transmitted parental chromosomes and P_{normal} is the frequency among untransmitted parental chromosomes.

To define the confidence intervals around the TDT results, we simulated 100,000 data sets with genomic regions containing a risk haplotype of the nature we observed. We allowed recombination to break down LD on both sides of the risk locus and examined the entire region to find the location of the maximum evidence of LD as measured by T/U. We defined confidence intervals around this maximum point and asked how often the true locus was contained within.

To examine whether the properties of the risk haplotype were adequate to explain the observed genetic properties of *IBD5*, we carried out simulation experiments. Specifically, the risk haplotype has the following properties: (i) its frequency among untransmitted chromosomes is 37% (ii) the transmission ratio from heterozygous parents to progeny with Crohn disease is 2.5:1 and (iii) the proportion of homozygotes to heterozygotes among affected individuals is 1:1. From these characteristics, one can infer the genetic properties of a Crohn disease locus carried on such a risk haplotype. Specifically, the best fit is a model in which one copy of the risk allele increases risk for Crohn disease by 2-fold and two copies by 6-fold. We then tested whether such a Crohn disease–susceptibility locus could have given rise to our observed linkage data (lod score=3.0 in 122 families with Crohn disease only). Simulation tests showed that such a locus would yield a lod score greater than 3.0 with a probability of approximately 7%, and a lod score greater than 2.0 with a probability of 20%.

In addition to the identification of known genes, we also examined this genomic region for the presence of unidentified genes by using GENSCAN gene-prediction software and searched for expressed sequence tag (EST) clusters using BLAST alignment. Eight predicted genes and four EST clusters were identified that did not overlap with the previously known genes; only one of the predicted genes had overlap with an EST cluster. From the comparison of the known mouse and human sequence, 47 conserved, noncoding sequences were reported in this region[15].

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

The authors would like to thank the following individuals for their invaluable participation in the patient sample collection: G. Wild, A. Watier, P. Pare, L. Dion, M.-L. Bernier, J. Rivard, J. Letourneau, J. Delisle, L. Lapierre and D. Lafrance. The authors would also like to thank L. Gaffney for her help in the preparation of this manuscript. This work was supported by the Crohn's and Colitis Foundation of Canada, the McGill Inflammatory Bowel Disease Network, and the Canadian Genetic Diseases Network, and by research grants from Bristol-Myers Squibb, Millennium Pharmaceutical, Affymetrix and Ellipsis Biotherapeutics Corporation. K.A.S. is a Senior Scientist of the Canadian Institutes for Health Research, M.S.S. is a fellow of the Canadian Association of Gastroenterology and the Canadian Institutes of Health Research, S.B.B. is a Scholar of the National Health Research Development Programme , A.B. is a Research Scholar of the Fonds de la Recherche en Santé du Quebec and T.J.H. is recipient of a Clinician Scientist Award from the Canadian Institutes for Health Research.

Received 9 May; accepted 29 August 2001.

1. Horikawa, Y. *et al.* Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175 (2000).
2. Rioux, J.D. *et al.* Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *Am. J. Hum. Genet.* **66**, 1863–1870 (2000).
3. Spielman, R.S. & Ewens, W. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* **59**, 983–989 (1996).
4. Frazer, K.A. *et al.* Computational and biological analysis of 680 kb of DNA sequence from the human 5q31 cytokine gene cluster region. *Genome Res.* **7**, 495–512 (1997).
5. Burckhardt, G. & Wolff, N.A. Structure of renal organic anion and cation transporters. *Am. J. Physiol. Renal. Physiol.* **278**, F853–F866 (2000).
6. Daly, M.J., Rioux, J.D., Schaffner, S., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
7. Clavell, M. *et al.* Detection of interferon regulatory factor-1 in lamina propria mononuclear cells in Crohn's disease. *J. Pediatr. Gastroenterol. Nutr.* **30**, 43–47 (2000).
8. Farthing, M.F., Dick, A.P., Heslop, G. & Levene, C.I. Prolyl hydroxylase activity in serum and rectal mucosa in inflammatory bowel disease. *Gut* **19**, 743–747 (1978).
9. Hugot J-P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
10. Ogura Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
11. Rioux, J.D. *et al.* Familial eosinophilia maps to the cytokine gene cluster on human chromosomal region 5q31-q33. *Am. J. Hum. Genet.* **63**, 1086–1094 (1998).
12. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. Hawkins, T.L., O'Connor-Morin, T., Roy, A. & Santillan, C. DNA purification and isolation using a solid-phase. *Nucleic Acids Res.* **22**, 4543–4544 (1994).
14. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24**, 381–386 (2000).
15. Zhu, Y. *et al.* Genomic interval engineering of mice identifies a novel modulator of triglyceride production. *Proc. Natl Acad. Sci. USA* **97**, 1137–1142 (2000).