

Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease

Jeffrey C Barrett^{*1}, Sarah Hansoul², Dan L Nicolae³, Judy H Cho⁴, Richard H Duerr^{5,6}, John D Rioux^{7,8}, Steven R Brant^{9,10}, Mark S Silverberg¹¹, Kent D Taylor¹², M Michael Barmada⁶, Alain Bitton¹³, Themistocles Dassopoulos⁹, Lisa Wu Datta⁹, Todd Green⁸, Anne M Griffiths¹⁴, Emily O Kistner¹⁵, Michael T Murtha⁴, Miguel D Regueiro⁵, Jerome I Rotter¹², L Philip Schumm¹⁵, A Hillary Steinhart¹¹, Stephan R Targan¹², Ramnik J Xavier¹⁶, the NIDDK IBD Genetics Consortium³³, Cécile Libioulle², Cynthia Sandor², Mark Lathrop¹⁷, Jacques Belaiche¹⁸, Olivier Dewit¹⁹, Ivo Gut¹⁷, Simon Heath¹⁷, Debby Laukens²⁰, Myriam Mni², Paul Rutgeerts²¹, André Van Gossum²², Diana Zelenika¹⁷, Denis Franchimont²², Jean-Pierre Hugot²³, Martine de Vos²⁰, Severine Vermeire²¹, Edouard Louis¹⁸, the Belgian-French IBD Consortium³³, the Wellcome Trust Case Control Consortium^{33,34}, Lon R Cardon¹, Carl A Anderson¹, Hazel Drummond²⁴, Elaine Nimmo²⁴, Tariq Ahmad²⁵, Natalie J Prescott²⁶, Clive M Onnie²⁶, Sheila A Fisher²⁶, Jonathan Marchini²⁷, Jilur Ghori²⁸, Suzannah Bumpstead²⁸, Rhian Gwilliam²⁸, Mark Tremelling²⁹, Panos Deloukas²⁸, John Mansfield³⁰, Derek Jewell³¹, Jack Satsangi²⁴, Christopher G Mathew²⁶, Miles Parkes²⁹, Michel Georges² & Mark J Daly^{8,32}

Several risk factors for Crohn's disease have been identified in recent genome-wide association studies. To advance gene discovery further, we combined data from three studies on Crohn's disease (a total of 3,230 cases and 4,829 controls) and carried out replication in 3,664 independent cases with a mixture of population-based and family-based controls. The results strongly confirm 11 previously reported loci and provide genome-wide significant evidence for 21 additional loci, including the regions containing *STAT3*, *JAK2*, *ICOSLG*, *CDKAL1* and *ITLN1*. The expanded molecular understanding of the basis of this disease offers promise for informed therapeutic development.

Recent genome-wide association studies (GWAS) have identified many common variants associated with complex diseases and have rapidly expanded our knowledge of the genetic architecture of these traits. Progress in Crohn's disease (CD), a common idiopathic inflammatory bowel disease (IBD) with high heritability ($\lambda_s \sim 20\text{--}35$), has been especially notable, with recent GWAS publications increasing the number of confirmed associated loci from two to more than ten¹. The results have identified previously unknown pathogenic mechanisms of IBD and promise to advance fundamentally our understanding of CD biology. These recent discoveries highlight, for instance, the importance of autophagy and innate immunity^{2–5} as determinants of the dysregulated host–bacterial interactions implicated in disease pathogenesis. Furthermore, genetic associations have been shown to be shared between CD and other auto-inflammatory conditions—for example, *IL23R* variants⁶ are also associated with psoriasis⁷ and ankylosing spondylitis⁸, and *PTPN2* variants with type 1 diabetes^{3,5}. As in other studies of complex diseases, restricted sample sizes have resulted in early CD studies

focusing on only the strongest effects, which turn out to explain only a fraction of the heritability of the disease.

We recently published three separate GWA scans for CD in European-derived populations^{4,5,9}, the details of which are shown in **Table 1**. Motivated by the need for larger datasets to improve power to detect loci of modest effect, we carried out a genome-wide meta-analysis from our three CD scans. These analyses, together with a replication study in an equivalently sized independent panel, have enabled us to identify at genome-wide levels of significance 21 new CD susceptibility loci. This brings the number of independent loci conclusively associated with CD to more than 30 and provides unprecedented insight into both CD pathogenesis as well as the general genetic architecture of a multifactorial disease.

RESULTS

Meta-analysis of three genome-wide association scans

The combined GWAS study samples (**Table 1**) consisted of 3,230 CD cases and 4,829 controls, all of European descent. Although the

*A full list of author affiliations appears at the end of this paper.

Table 1 Samples used (post quality control) in this study

	NIDDK	BEL-FR	UK	Total
Scan cases	946	536	1,748	3,230
Scan controls	977	914	2,938	4,829
Replication cases	0	1,082	1,243	2,325
Replication controls	0	787	1,022	1,809
Replication trios	720	619	0	1,339
Nationality	US-Canadian	Belgian-French	British	
Scan platform	Illumina	Illumina	Affymetrix	
	HumanHap300	HumanHap300	GeneChip 500K	
Replication platform	Sequenom	Illumina GoldenGate	Sequenom	

individual scans did identify new risk factors, they were only well-powered to discover common alleles with odds ratios (ORs) above 1.3 (in the case of the WTCCC) or 1.5 (the smaller two scans) (Fig. 1). By contrast, the combined sample has 74% power at an OR of 1.2, allowing evaluation of the role of alleles with smaller effect sizes for the first time. As two different genotyping technologies were used in the constituent scans, we used recently developed imputation^{10,11} methods to assess association across all three studies at 635,547 SNPs contained on one or both platforms. A quantile-quantile (Q-Q) plot of the primary meta-statistic (single-SNP Z scores; Fig. 2) shows a marked excess of significant associations, well beyond what would be attributable to the modest overall distributional inflation (genomic control $\lambda < 1.16$). Despite the large sample size, the overall inflation is modest because (i) each group had separately tested for evidence of population stratification, and the meta-analysis used a test that combined the results from each study (rather than mixing the raw data and compromising the case-control matching of each study), and (ii) imputation was done on all samples ignoring case status and, thus, would not introduce artifactual differences between cases and controls¹².

We focus our attention in this study specifically on the 526 SNPs from 74 distinct genomic loci that were associated with $P < 5 \times 10^{-5}$, which is more than seven times the number of SNPs expected by chance even after correction for the modest overall inflation detected. This threshold for follow-up is not meant to imply that there are no genuine associations among SNPs with less significant association in the meta-analysis, but rather reflects a practical desire to prioritize as many true positives as possible for immediate replication. Eleven associations previously replicated and established at genome-wide significance levels (Table 2), including both 'historical' associations at *NOD2* (also known as *CARD15*)^{13,14} and 5q31 (*IBD5*)¹⁵ as well as recent replicated findings from individual GWA scans^{2-6,16} such as *IL23R*, *ATG16L1*, *IRGM*, *TNFSF15* and *PTPN2*, were among the 74 regions represented in this tail of the distribution of association statistics. However, even after removal of all SNPs in linkage disequilibrium (LD) with these 11 loci, there continued to be a substantial excess of associated alleles beyond that which would be expected by chance (Fig. 2).

Replication of 21 newly identified loci

As these 74 regions included the 11 already reported as independently replicated and meeting genome-wide significance thresholds, this replication experiment effectively explored 63 putative associations in previously unreported regions with 11 positive controls (Supplementary Table 1 online). To identify the true risk factors among these 63 regions, we carried out a replication study involving 2,325 additional CD cases and 1,809 controls alongside an

independent family-based dataset of 1,339 trios of parents and their affected offspring.

Results (significance levels and ORs) for strongly replicating loci, including all positive controls, are presented in Table 2. The distribution of Z scores from the 63 putative regions shows a marked departure from the null distribution (Fig. 3), with 19 new regions showing significant replication ($P < 0.0008$; a value of 0.05/63 representing a conservative threshold expected to be exceeded only once by chance in 20 such replication experiments). SNPs on chromosome 19p13 (replication $P = 0.00347$, combined $P = 2.12 \times 10^{-9}$) and in the MHC (replication $P = 0.006$, combined $P = 5.20 \times 10^{-9}$; suspected but not previously conclusively established in CD) did not reach this conservative threshold, but so convincingly satisfy proposed thresholds for genome-wide significance ($P < 5 \times 10^{-8}$) that we propose these as the twentieth and twenty-first additional CD associated-loci defined here. A further 8 of the 42 remaining loci showed nominal replication (Table 3).

It is possible that extreme population substructure in the replication sample could give rise to such a marked excess of hits. Although unlikely, this was directly evaluated by the large family-based component of the replication study. Odds ratio estimates from the TDT analysis of the North American, French and Belgian families alone are consistent with those from the UK and Belgian case-control samples (Tables 2 and 3), with all 21 newly defined loci showing ORs in the same direction of association with the original scan in the family-based component (and nearly half showing greater OR than in the case-control arm). None of the significantly or nominally replicating loci show significant evidence for heterogeneity (across studies or between family-based and population-based arms) when corrected for the number of tests performed. This independent family-based evidence confirms that these alleles constitute true CD-associated loci.

For this newly expanded set of 32 unequivocally associated loci, we assessed whether there was evidence of significant pairwise interactions that could add further to the overall variance in liability explained by this set of loci. We carried out a case-only analysis of the 3,664 cases in the replication study and observed no interactions that withstood a correction for the number of tests performed (Supplementary Table 2 online).

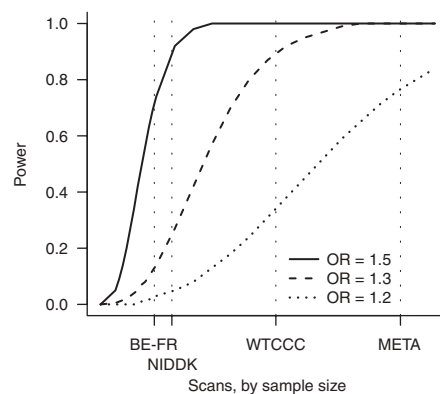


Figure 1 Power to detect a genetic effect of various sizes (odds ratio 1.2, 1.3 or 1.5) versus study sample size. Power is reported here as the probability (given a multiplicative model and risk allele frequency of 20%) of $P < 5 \times 10^{-5}$ in a scan, which was the value used to define regions for attempting replication in a larger sample set. Vertical dotted lines show the sample sizes for the three constituent scans and the meta-analysis. Relatively large effects are likely to be detected by any of these scans, whereas only the combined analysis is well powered to detect more modest effects.

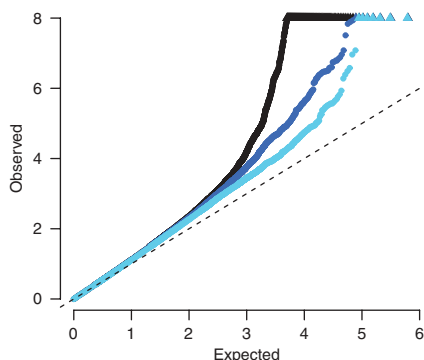


Figure 2 A quantile-quantile plot of observed $-\log_{10} P$ values versus the expectation under the null. Black points represent the complete meta-analysis, with a substantial departure from the null at the tail (values > 8 are represented along the top of the plot as triangles). Dark blue points show the distribution after removal of 11 previously published loci, demonstrating a still notable excess. Light blue points show the distribution after removal of all 40 loci that replicate at least nominally. In all the cases, the overall distribution is marginally inflated ($\lambda_{GC} < 1.16$).

Deciphering the genetic architecture of CD

The contributions of the 32 loci to disease risk were computed using a standard liability threshold model and are displayed as a histogram of individual variances (Fig. 4). The observations from this variance analysis indicate that many loci were detected for which the current study had low power and that only a minority of the variance in risk is

explained by these 32 loci, suggesting that many additional loci remain to be identified. This is reinforced by the additional eight nominal replications (Table 3), where only two or three would be expected by chance, and by the continued excess of small P values when these 40 total regions are removed (Fig. 2).

Although we recognize that fine-mapping is required to identify specific causal variants, we carried out a series of analyses to gain some general insight into the CD associations. We first queried HapMap to discover any instances where a nonsynonymous SNP (nsSNP) was correlated ($r^2 > 0.5$) to the most associated variant discovered in this study. Accepting that HapMap is not a complete catalog of nsSNPs, but

Table 2 Convincingly (Bonferroni $P < 0.05$) replicated CD risk loci

SNP	Chr.	Critical region	P values			No. genes	Gene of interest	RAF	Risk allele	Odds ratios	
			Scan	Replication	Combined					Case Ctrl	TDI
Previously published loci											
rs11465804	1p31	67.4 ^a	1.01×10^{-35}	3.10×10^{-29}	6.66×10^{-63}	n.a.	<i>IL23R</i>	0.933	T	2.50	2.77
rs3828309	2q37	230.9 ^a	1.13×10^{-20}	7.67×10^{-14}	2.36×10^{-32}	n.a.	<i>ATG16L1</i>	0.533	G	1.28	1.30
rs3197999	3p21	48.73–49.87	2.16×10^{-7}	5.64×10^{-7}	1.15×10^{-12}	35	<i>MST1^b</i>	0.271	A	1.20	1.20
rs4613763	5p13	40.32–40.48	4.52×10^{-22}	2.79×10^{-8}	6.82×10^{-27}	0	<i>PTGER4^c</i>	0.125	C	1.32	1.28
rs2188962	5q31	131.44–131.90	4.58×10^{-9}	3.52×10^{-11}	2.32×10^{-18}	7		0.425	T	1.25	1.26
rs11747270	5q33	150.15–150.32	6.36×10^{-11}	2.57×10^{-7}	3.40×10^{-16}	3	<i>IRGM</i>	0.090	G	1.33	1.31
rs4263839	9q32	114.61–114.78	3.92×10^{-7}	6.58×10^{-5}	2.60×10^{-10}	2	<i>TNFSF15</i>	0.677	G	1.22	1.07
rs10995271	10q21	64.05–64.12	1.90×10^{-11}	1.61×10^{-10}	4.46×10^{-20}	1	<i>ZNF365</i>	0.387	C	1.25	1.53
rs11190140	10q24	101.26–101.32	1.71×10^{-10}	1.69×10^{-7}	3.06×10^{-16}	1	<i>NKX2-3</i>	0.478	T	1.20	1.28
rs2066847	16q12	49.3 ^a	n.a.	1.49×10^{-24}	2.98×10^{-24}	n.a.	<i>NOD2</i>	0.018	C	3.99	2.57
rs2542151	18p11	12.73–12.88	1.19×10^{-11}	2.41×10^{-7}	5.10×10^{-17}	1	<i>PTPN2</i>	0.152	G	1.35	1.14
Newly identified loci											
rs2476601	1p13	113.79–114.17	1.81×10^{-5}	0.000101	1.46×10^{-8}	7	<i>PTPN22</i>	0.899	G	1.31	1.17
rs2274910	1q23	157.65–157.72	3.50×10^{-7}	0.000481	1.46×10^{-9}	2	<i>ITLN1</i>	0.682	C	1.14	1.62
rs9286879	1q24	169.54–169.67	4.02×10^{-7}	0.000321	1.53×10^{-9}	0		0.243	G	1.19	1.08
rs11584383	1q32	197.60–197.77	6.82×10^{-7}	2.34×10^{-6}	1.43×10^{-11}	3		0.697	T	1.18	1.20
rs10045431	5q33	158.69–158.76	8.80×10^{-9}	3.66×10^{-6}	3.86×10^{-13}	1	<i>IL12B</i>	0.708	C	1.11	1.36
rs6908425	6p22	20.63–20.84	2.52×10^{-7}	0.000278	8.96×10^{-10}	1	<i>CDKAL1</i>	0.780	C	1.21	1.09
rs7746082	6q21	106.52–106.62	3.70×10^{-6}	7.70×10^{-6}	2.44×10^{-10}	0		0.289	C	1.17	1.19
rs2301436	6q27	167.32–167.52	3.30×10^{-7}	3.26×10^{-7}	1.04×10^{-12}	3	<i>CCR6</i>	0.463	T	1.21	1.16
rs1456893	7p12	50.03–50.11	4.92×10^{-5}	1.10×10^{-5}	4.60×10^{-9}	0		0.678	A	1.20	1.14
rs1551398	8q24	126.60–126.62	4.90×10^{-6}	0.000109	4.50×10^{-9}	0		0.619	A	1.08	1.25
rs10758669	9p24	4.94–5.26	6.80×10^{-7}	0.00043	3.46×10^{-9}	3	<i>JAK2</i>	0.348	C	1.12	1.21
rs17582416	10p11	35.30–35.60	8.48×10^{-6}	2.53×10^{-5}	1.79×10^{-9}	3		0.345	G	1.16	1.26
rs7927894	11q13	75.80–76.02	1.43×10^{-7}	0.000732	1.32×10^{-9}	1	<i>C11orf30</i>	0.386	T	1.16	1.07
rs11175593	12q12	38.61–39.31	1.33×10^{-7}	0.000165	3.08×10^{-10}	3	<i>LRRK2,MUC19</i>	0.017	T	1.54	1.44
rs3764147	13q14	43.13–43.54	1.61×10^{-7}	1.33×10^{-7}	2.08×10^{-13}	3		0.221	G	1.25	1.19
rs2872507	17q21	34.63–35.34	2.12×10^{-6}	0.000292	5.00×10^{-9}	17	<i>ORMDL3</i>	0.473	A	1.12	1.24
rs744166	17q21	37.74–37.95	5.94×10^{-6}	9.15×10^{-8}	6.82×10^{-12}	4	<i>STAT3</i>	0.565	A	1.18	1.25
rs1736135	21q21	15.73–15.76	2.06×10^{-5}	4.58×10^{-5}	7.40×10^{-9}	0		0.565	T	1.18	1.10
rs762421	21q22	44.43–44.48	1.08×10^{-5}	1.59×10^{-5}	1.41×10^{-9}	1	<i>ICOSLG</i>	0.389	G	1.13	1.21

RAF is risk allele frequency in control samples (see Supplementary Table 5 online for details). Critical region is in NCBI Build 35 coordinates, with definition as described in Methods. Risk alleles are defined relative to the +strand of the reference. P values for the TDI analysis for these loci are listed in Supplementary Table 6 online.

^aRegions where causal variants have been convincingly mapped, rendering the LD window uninformative. ^bRecently implicated through work reported in ref. 28. ^c*PTGER4* is outside the critical region, but was implicated via eQTL analysis.

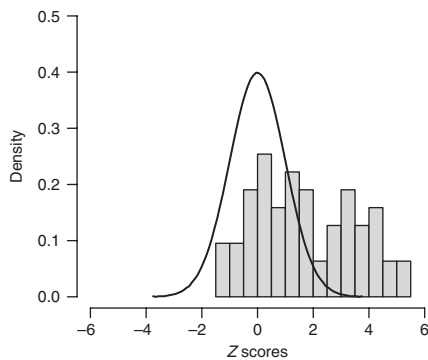


Figure 3 Distribution of observed Z scores from the 63 newly identified regions explored, along with the expected distribution under the null (a standard normal with mean 0 and variance 1). Even when the 21 regions reaching genome-wide significance are set aside, the distribution is highly skewed: four more results exceed a Z of 2 (one would be expected by chance under the null) and none showed a Z of less than -2 (same expectation under the null), suggesting that even more of the regions investigated here are likely to be found to constitute true-positive associations when additional data become available.

including four loci where fine mapping has identified coding variants, we found that just 9 of the 32 genome-wide significant associations were correlated with a known nsSNP (**Supplementary Table 3** online). To explore whether any of the associations reflect a *cis*-acting regulatory effect on a nearby gene, we evaluated genotype–expression correlation using the panel of 400 lymphoblastoid cell lines previously described¹⁷. From all genes within 250 kb of the LD-based intervals defined in **Tables 2** and **3**, we identified five correlations between expression of a nearby gene and a CD-associated variant (LOD > 2) (**Supplementary Table 4** online). This was far in excess of chance ($P \sim 0.001$) (**Supplementary Fig. 1** online) and suggests that regulatory variation also contributes to the genetic architecture identified.

DISCUSSION

Genome-wide association studies provide a systematic assessment of the contribution of common variation to disease pathogenesis. A limiting factor is often the size of the case-control dataset, and hence the power to detect any but the most strongly associated loci.

Meta-analysis of existing data provides an obvious potential solution. As **Figure 1** shows, our expectation was that the additional power of the combined dataset would result in the identification of a substantially larger number of readily replicating associations than were derived from any of the smaller, constituent datasets. However, the paradigm of exploring common genetic variation with similar effects across studies (in this case all of European descent) needs to be tested before its results can be accepted as valid.

On the validity of the method, our results are substantially reassuring. All 11 previously confirmed CD susceptibility loci were strongly replicated both in the meta-analysis and follow-up experiment. These include the two widely replicated findings from studies published in 2001 (refs. 13–15) as well as all of the compelling findings from individual GWAS (**Table 2**). We also identified and replicated 21 new CD susceptibility loci. Using a conservative threshold for significance (only one such region would be expected by chance in 20 such experiments), the loci with clear evidence for association in the replication panel include a very high proportion of those showing the strongest signals in the meta-analysis (**Supplementary Table 1**)—9 of 9 previously unreported regions with $P < 5 \times 10^{-7}$ in the combined scan were replicated convincingly—emphasizing the validity of the meta-analysis results. Further, all 21 of these loci exceed a conservative genome-wide level of significance ($P < 5 \times 10^{-8}$) by a substantial margin (all but two have $P < 5 \times 10^{-9}$), and equivalent strength of association was observed in the family-based subset of our replication sample.

In keeping with other regions recently identified as associated with CD, the 21 newly identified loci do not conform to any obvious pattern in terms of gene content. Thus, as shown in **Table 2**, some loci (defined by HapMap recombination hotspots flanking the set of correlated, associated variants) contain just a single gene, some contain many genes and others none. Clearly, the first category provides the most immediate clues regarding pathogenic mechanisms. These genes are discussed briefly in **Box 1**, together with a number of genes that constitute noteworthy candidates from regions with only a handful of transcripts. Included among these are compelling functional candidates such as *STAT3*, *JAK2* and *IL12B*, whereas others, such as *CDKALI* and *PTPN22*, highlight potentially intriguing contrasts between genetic susceptibility to CD and other complex disorders (**Box 1**). Consistent with previous findings from CD and other complex diseases, we did not find any strong evidence of

Table 3 Nominally (uncorrected $P < 0.05$) replicated CD risk loci

SNP	Chr	Critical region	P values			No. genes	Gene of interest	RAF	Risk allele	Odds ratios	
			Scan	Replication	Combined					CaseCtrl	TDT
rs4807569	19p13	1.05–1.15	1.16×10^{-8}	0.00347	2.12×10^{-9}	2		0.217	C	1.02	1.26
rs780094	2p23	27.30–27.77	3.82×10^{-6}	0.00381	3.14×10^{-7}	22	<i>GCKR</i>	0.397	T	1.08	1.13
rs3763313	6p21	32.44–32.79 ^a	1.45×10^{-8}	0.00602	5.20×10^{-9}	7	<i>BTNL2</i> , <i>SLC26A3</i> , <i>HLA-DRB1</i> , <i>HLA-DQA1</i>	0.188	C	1.19	1.01
rs13003464	2p16	61.09–61.14	3.44×10^{-5}	0.00565	4.60×10^{-6}	1	<i>PUS10</i>	0.376	G	1.16	1.08
rs991804	17q12	29.57–29.70	4.02×10^{-6}	0.0135	1.07×10^{-6}	4	<i>CCL2</i> , <i>CCL7</i>	0.726	C	1.10	1.08
rs12529198	6p25	5.04–5.11	7.08×10^{-7}	0.0192	6.96×10^{-7}	1	<i>LYRM4</i>	0.062	G	1.12	1.19
rs17309827	6p25	3.36–3.42	2.08×10^{-6}	0.0391	2.74×10^{-6}	1	<i>SLC22A23</i>	0.639	T	1.10	1.02
rs7758080	6q25	149.54–149.65	7.28×10^{-6}	0.044	8.78×10^{-6}	0		0.274	G	1.12	0.99
rs8098673	18q11	17.74–17.93	3.18×10^{-5}	0.0443	2.88×10^{-5}	0		0.329	C	1.05	1.09
rs917997	2q11	102.31–102.64	2.16×10^{-5}	0.0493	2.22×10^{-5}	5	<i>IL18RAP</i>	0.222	T	1.05	1.11

RAF is risk allele frequency in control samples (see **Supplementary Table 5** for details). Critical region is in NCBI Build 35 coordinates, with definition as described in Methods. Risk alleles are defined relative to the +strand of the reference.

^aSNPs with $P < 0.0001$ were observed throughout the MHC from 30.2–32.9 Mb, but only this largest signal from the region was followed up. More detailed study of the MHC will be required to identify and localize potentially independent signals from this region.

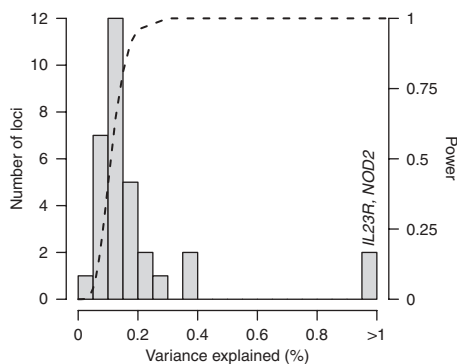


Figure 4 Histogram of percent variance explained by each of the 32 established CD risk loci. The distribution resembles the long postulated exponential distribution of effect sizes. Dashed line shows the joint power for our meta-analysis to detect ($P < 5 \times 10^{-5}$), and for our replication sample to replicate (at Bonferroni corrected P values), a 20% variant explaining a given fraction of variance. Note how quickly this curve moves from nearly zero power to detect tiny effects (less than one tenth of one percent) to nearly full power to detect larger effects (presuming they are well covered by the current generation of GWAS chips). Complete power near the origin would likely reveal a more complete exponential distribution, with many very small effects. These are likely to increase somewhat once the causal variant or variants are identified in each locus. Indeed, *NOD2* and *IL23R* are distant outliers, each explaining 1–2% of total variance, partially because multiple causal variants have already been discovered at these loci^{6,13}.

deviation from the model of multiplicative (random) effects when we tested for gene–gene interactions among the 32 confirmed associations, in spite of the fact that some of these genes seem to affect the same or overlapping pathways.

For loci containing multiple genes or no genes, the picture is less well defined. The identified paucity of correlation between associated SNPs and coding variation suggests that these loci in particular may benefit from eQTL (expression quantitative trait locus) analysis. This approach seeks to correlate genotype and expression patterns and takes into account that such functional relationships need not respect the specific boundaries of LD around the association. One of our groups previously reported an eQTL effect implicating *PTGER4* at the 5p13 locus⁹. A notable outcome from our present analysis was at the established IBD5 locus¹⁵, where CD-associated SNPs were associated with decreased *SLC22A5* mRNA expression. Although a SNP had previously been proposed as regulating *SLC22A5* transcriptional activity¹⁸, these data suggest for the first time that the most disease-associated variants in the IBD5 region, including a coding variant in neighboring *SLC22A4*, are the same variants most associated with *SLC22A5* expression. Further, the most significant CD-associated eQTL reported here affects *ORMDL3* (lod score = 20) on chromosome 17, and SNPs in precisely the same region were recently shown to be strongly associated with childhood asthma¹⁹. This suggests that the same polymorphisms might underlie susceptibility to both CD and asthma, possibly by perturbing *ORMDL3* expression.

The new loci that we identified are of modest effect size, which is unsurprising given that all loci with larger impact on disease risk were discovered in the original scans (as might be expected). The small sizes of these effects explains the lack of overlap between linkage results in CD and these newly discovered loci (**Supplementary Fig. 2** online), with the possible exceptions of combined effects of multiple high-ranking associations on chromosomes 5q and 6p. Indeed, the linkage evidence that led to the discovery of the IBD5 locus was very likely boosted by the nearby effects at *IL12B* and *IRGM*. As expected, the

Box 1 Noteworthy genes within loci newly implicated in Crohn's disease pathogenesis

- ***CCR6*** (chemokine receptor 6): encoding a member of the G protein-coupled chemokine receptor family, this homing receptor is expressed by immature dendritic cells and memory T cells and is important for B-cell differentiation and tissue-specific migration of dendritic and T cells during epithelial inflammatory and immunological responses²⁹. The ligand of this receptor is macrophage inflammatory protein 3 α (MIP-3 α); both genes are expressed in granulomas of pulmonary sarcoid³⁰. Recent studies have also demonstrated that *CCR6*, *IL23R* and *RORC* are selectively expressed by IL-17 producing cells and IFN γ -producing Th17 and Th1 cells in CD³¹.

- ***IL12B***: this gene encodes the p40 subunit, which is a constituent of both heterodimeric interleukins IL-12 and IL-23 (ref. 32). Its association with CD has been previously reported⁵ but not confirmed, and it is also known to be associated with psoriasis⁷. The key role of the IL-12–IL-23 pathway in chronic intestinal inflammation is supported by the association between *IL23R* and CD³ and strong functional evidence from mouse models of colitis^{33–36}.

- ***STAT3*** (signal transducer and activator of transcription 3) and ***JAK2*** (Janus kinase 2): the JAK-STAT pathway is a focal point in signal transmission downstream of cytokine and growth factor signals from cell surface receptors to the nucleus to modify transcription of various genes, notably in hematopoietic cells. The present findings are particularly significant, given the role of both genes in IL23R signaling³⁷ and the central role of STAT3 in Th17 differentiation³⁸. However, JAK2 and STAT3 are also downstream of several other cytokines implicated in CD pathogenesis in addition to IL-23, highlighting the pathophysiologic complexity of these new associations. Further complexity is highlighted by the distinctly different roles of STAT3 in innate versus adaptive immunity in mouse colitis models: activation of STAT3 in innate immune cells enhances mucosal barrier function, whereas STAT3 activation in T cells exacerbates colitis.

- ***LRRK2*** (leucine-rich repeat kinase 2): this gene encodes a multi-domain protein expressed mainly in the cytoplasm of neurons, myeloid cells and monocytes, and mutations in *LRRK2* have been strongly associated with Parkinson's disease³⁹. A recent study³⁹ reported the induction of autophagy by mutant *LRRK2*, which is of interest given the strong associations between CD and the autophagy genes *ATG16L1* and *IRGM*^{2–5}. The same locus also contains the gene *MUC19*, which encodes a large protein with multiple serine- and threonine-rich repeats characteristic of the proteins encoded by the mucin gene family. The mucin proteins are core components of the mucus layer, which protects the intestinal epithelium from injury, and mucin deficiency leads to intestinal inflammation in mouse models of colitis⁴⁰.

- ***CDKAL1***: the protein encoded by this gene is poorly characterized, but *CDKAL1* is noteworthy for being recently confirmed as gene associated with type 2 diabetes susceptibility^{24,41–43}. In this study, we find that SNPs from the same intron of *CDKAL1* that shows association with T2D are associated with CD, but the associated alleles for the two diseases are not correlated with each other.

- ***ICOSLG*** (inducible T-cell co-stimulator ligand): this co-stimulatory molecule is expressed on intestinal (and other) epithelial cells and may have a role in their antigen presentation to and regulation of mucosal T lymphocytes⁴⁴. Upon maturation, plasmacytoid dendritic cells express *ICOSLG* and drive the generation of IL-10–producing regulatory T cells⁴⁵.

- ***PTPN2*** and ***PTPN22*** (protein tyrosine phosphatase, nonreceptor types 2 and 22): both of these genes are associated with other autoimmune and inflammatory diseases and the effect described here for *PTPN2* is similar to that previously described for type 1 diabetes (T1D)⁴⁶. However, the association of *PTPN22* with CD, although mapping to the same coding variant (R602W) that is a risk factor for T1D and rheumatoid arthritis^{47,48}, is in the opposite direction, with the T1D and rheumatoid arthritis risk allele, 602W, offering protection from CD.

- ***ITLN1*** (intelectin-1): this gene known to be expressed in human small bowel and colon, and encodes a 120-kDa homotrimeric lectin recognizing galactofuranosyl residues found in cell walls of various microorganisms but not in mammals⁴⁹. Human intelectin-1 is structurally identical to the lactoferrin receptor (LFR), expressed within the enterocyte brush border, and seems critical in membrane stabilization, preventing loss of digestive enzymes and protecting the glycolipid microdomains from pathogens⁵⁰. In addition, intelectin expression is reported in Paneth cells in both mouse and pig small intestine, further pointing to a role in innate immunity.

only gene conclusively discovered via linkage (*NOD2*) is one of two loci which stand well out from the remainder of the distribution of effect sizes (Fig. 4). The other outlier, *IL23R*, illustrates an interesting characteristic of linkage: because (unlike in *NOD2*) the most penetrant risk allele has very high frequency (93%), it is nearly invisible to linkage analysis despite the high OR; highly protective rare alleles are simply not present in multiplex affected families and thus do not influence allele sharing substantially.

Using a liability-threshold model, we estimate that the 32 loci identified to date explain about 10% of the overall variance in disease risk, which may be as much as a fifth of the genetic risk, given previous estimates of CD heritability of approximately 50% (ref. 20). This observation is consistent with the fact that these loci collectively contribute only a factor of two to sibling relative risk (λ_s), and even this figure is dominated by the substantial contribution of *NOD2* variants. However, it should be emphasized that the full impact of the new loci cannot be determined until causal variants have been identified by directed sequencing and fine-mapping experiments. Until then, the proportion of the variance in CD risk explained must be measured from the confirmed SNPs, where association is due to LD with causal variants. As multiple causal variants might exist at each locus (ranging in frequency from rare to common), our estimates of variance explained provide only a lower bound for the true contribution of each locus.

In conjunction with results from a very similar gene discovery effort in type 2 diabetes²¹, common lessons are beginning to emerge with respect to the genetic architecture of complex traits. In each example, a substantial increase in sample size achieved through meta-analysis has led to dramatic improvements in gene discovery. In all cases, this progress has revealed an underlying architecture consistent with many individually modest effects, which conventional genetic linkage analysis, and even the largest individual genome-wide association studies, are not well powered to detect. Common variants explaining more than 1% of the genetic variance are rare, whereas well-powered studies have found dozens of variants contributing 0.1% of overall variance in liability. Perhaps surprisingly, neither we nor others have yet to document a substantial role for epistasis among these loci, and a number of associated loci are conclusively mapped to regions with no currently annotated protein-coding genes. Despite the considerable concordant success, a distinct minority of the overall heritability has been explained by these documented associations.

As our study is well powered to identify loci that explain > 0.2% of the overall variance, but the sum of such loci explains a relatively small fraction of the total, it seems likely that many loci with even more modest effect sizes remain undiscovered. Of particular note is the continued excess of associations outside of the regions studied here, as well as the nominal replication of an additional eight loci, notably greater than expected by chance. Overall, the distribution of *Z* scores in the replication experiment is clearly skewed toward replication: only 11 of the 63 *Z* scores in this replication experiment generate *Z* < 0. If only the 21 strongly confirmed loci were genuinely associated, half of the 42 remaining should end up with *Z* < 0. Indeed, that 8 of the 42 remaining tests have *Z* > 1.5 is itself a highly significant observation (*P* < 0.0001). Although modest in terms of effect size, identification of such loci is likely to still provide important insights into pathogenic mechanisms, as biological importance need not be proportional to the statistical evidence for genetic association. Closer inspection of regions showing nominal association in the replication experiment reveals that a number of transcripts in these loci are of considerable interest, including *CCL2* and *CCL7* (ref. 22), *IL18RAP*²³ and *GCKR*²⁴.

It is important to note that the GWAS arrays used for these scans did not offer complete genome coverage of common variation (additional loci may reside in poorly covered intervals) and did not address either rare SNPs or copy number variation effectively. Thus, in spite of the wealth of new susceptibility loci identified by the current study, it seems plausible that there are still more to be found; however, very large datasets are likely to be required to achieve robust statistical support for them. With respect to the present findings, there is much work to be done in resequencing and fine mapping to identify causal variants. Although we do not yet have a complete understanding of the genetic architecture of CD, dramatic progress has now been made toward this goal—and with it, the prospect of directed functional exploration of the pathways identified, insight into how risk alleles interact with environmental modifiers and the hope of new avenues for treatment.

METHODS

Subjects and GWAS. The meta-analysis was based on data from the three genome-wide scans of the NIDDK⁴, WTCCC⁵ and Belgian-French⁹ studies. Details of the numbers of cases and controls genotyped in the respective scans and of the genotyping platforms used are shown in Table 1, as are details of the case-control and family cohorts genotyped in the replication study of the meta-analysis. Details of the ascertainment and characterization of these cohorts, as well as of quality control procedures applied to the GWA datasets, were provided in the original scan and replication publications^{3–6,9}. Recruitment of study subjects was approved by local and national institutional review boards, and informed consent was obtained from all participants.

Imputation. These methods rely on observed haplotype patterns in a set of reference data (the HapMap) and the actual genotype data from each project to make predictions (along with a measure of statistical certainty) at ungenotyped SNPs. We used the program MACH¹⁰ with the NIDDK and Belgian-French data, and IMPUTE¹¹ with the WTCCC data. Comparisons between the two algorithms yielded very similar results (data not shown). We imputed the superset of polymorphic markers that passed quality control in the original scans^{4,5,9}. This set was comprised of SNPs on either the Affymetrix 500K only (*n* = 350,507), Illumina HumanHap300 version 1 only (*n* = 238,935) or both panels (*n* = 46,105) such that all association tests done were at least partially based on observed genotype data.

Test for association, effect size estimation and interactions. Using the genotype probabilities (rather than best-guess genotypes) and empirical variances for imputed markers in the case and control tallies, we summarized the standard 1 d.f. allele-based test of association as a *Z* score within each scan and combined scores across studies to produce a single meta-statistic for each SNP across all three datasets. Odds ratios were estimated separately in TDT samples and each case-control replication collection and then combined and tested for heterogeneity²⁵. Interaction tests were done using the case-only epistasis test implemented in PLINK²⁶.

Critical regions. Given that most associations contain many correlated SNPs showing signal, we demarcated independent loci by first defining the set of HapMap SNPs with $r^2 > 0.5$ to the most significantly associated SNP. We then bounded the 'critical region' by the flanking HapMap recombination hotspots that contained this set. These windows very likely contain the causal polymorphisms explaining the associations.

Replication. We defined loci to have been previously confirmed if an earlier study had both detected and replicated the association in independent samples and the association achieved $P < 5 \times 10^{-8}$ (recently proposed as an appropriate genome-wide significance level for GWAS²⁷). For replication genotyping, we selected the most significantly associated SNP from each region along with a second, correlated SNP with $P < 0.0001$ or a second assay on the opposite strand in order to have a technical backup should the first fail genotyping (Supplementary Table 1). Replication genotyping for the putatively associated loci was done using primer extension chemistry and mass spectrometric analysis (iPLEX, Sequenom) using Sequenom Genetics Services

(North American panel) and Genome Research Limited, Wellcome Trust Sanger Institute (UK panel), and using a custom-made Golden Gate assay on a Beadstation500 (Illumina), following the manufacturer's recommendations (Belgian-French panel). The more completely genotyped SNP of the two from each region was chosen to represent that regional association in analysis (if both were completely typed, the SNP that was more strongly associated in the scan was used). Samples with >10% missing data ($n = 267$ for Belgian-French data, 111 for the UK data and 8 for the North American data; these samples are not included in the tallies for **Table 1**), as well as SNPs with >10% missing data or Hardy-Weinberg P value <0.001, were excluded from this analysis.

Regional annotation via eQTL analysis. The effects of SNPs listed in **Tables 2** and **3** on expression of neighboring genes were studied using transcriptome data from the ~400 lymphoblastoid cell lines described previously¹⁷. SNPs that were not genotyped on this panel ($n = 14$) were replaced with a proxy with $r^2 > 0.95$ when possible ($n = 12$). Lod scores > 2 for genes (probe average) located within 250 kb of the corresponding LD windows were retrieved (see URLs section below). To evaluate the significance of the findings with the CD-associated SNPs, we compared the observed (i) number of genes yielding lod scores > 2, and (ii) sum of these lod scores, with the corresponding frequency distributions for 1,000 randomly selected sets of 31 SNPs, matched for allele frequency (± 0.02) and gene context. Window sizes determined for associated SNPs were used for the matched simulated SNPs.

URLs. mRNA by SNP Browser, <http://www.sph.umich.edu/csg/liang/asthma/>; meta-analysis test statistics and allele frequencies for all SNPs, <http://www.broad.mit.edu/~jbarret/ibd-meta/>.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We acknowledge use of DNA from the 1958 British Birth Cohort collection (R. Jones, S. Ring, W. McArdle and M. Pembrey), funded by the Medical Research Council (grant G0000934) and The Wellcome Trust (grant 068545/Z/02), and the UK Blood Services Collection of Common Controls (W. Ouwehand) funded by the Wellcome Trust. We also acknowledge the National Association for Colitis and Crohn's disease and the Wellcome Trust for supporting the case DNA collections, and support from UCB Pharma (unrestricted educational grant) and the NIHR Cambridge Biomedical Research Centre. The National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) IBD Genetics Consortium is funded by the following grants: DK62431 (S.R.B.), DK62422 (J.H.C.), DK62420 (R.H.D.), DK62432 and DK064869 (J.D.R.), DK62423 (M.S.S.), DK62413 (K.D.T.), NIH-AI06277 (R.J.X.) and DK62429 (J.H.C.). Additional support was provided by the Burroughs Wellcome Foundation (J.H.C.) and the Crohn's and Colitis Foundation of America (S.R.B., J.H.C.). We thank P. Gregersen and A. Lee (Feinstein Medical Research Institute) for their efforts and the use of control samples. This work was supported by grants from the DGTRÉ from the Walloon Region (n°315422 and CIBLES), the Communauté Française de Belgique (Biomod ARC), and the Belgian Science Policy organisation (SSTC Genefunc and Biomagnet PAI). E.L., S.Hansoul, D.F. and S.V. are fellows of the Belgian Fonds de la Recherche Scientifique (FNRS) and Fonds Wetenschappelijk Onderzoek-Vlaanderen (NFWO). C.S. is a fellow of the FRIA. We are grateful to all the clinicians, consultants and nursing staff who recruited subjects, including: J.-M. Maisin, V. Muls, J. Van Cauter, M. Van Gossum, P. Closset, P. Hayard and J.M. Ghilain (Erasmus-BBIH-IBD); P. Mainguet, F. Mokaddem, F. Fontaine, J. Deflandre and H. Demolin (Ulg collaborators); J.-F. Colombel, M. Lemann, S. Almer, C. Tysk, Y. Finkel, M. Gassul, C. O'Morain, V. Binder and J.-P. Cézard (INSERM collaborators). Sincere thanks to L. Liang for his assistance in accessing the eQTL database, and to F. Merlin for expert technical assistance. Finally, we thank all individuals who contributed samples.

Published online at <http://www.nature.com/naturegenetics/>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

- Mathew, C.G. New links to the pathogenesis of Crohn's disease provided by genome-wide association scans. *Nat. Rev. Genet.* **9**, 9–14 (2008).
- Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nat. Genet.* **39**, 207–211 (2007).

- Parkes, M. *et al.* Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39**, 830–832 (2007).
- Rioux, J.D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39**, 596–604 (2007).
- The Wellcome Trust Case-Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
- Duerr, R.H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
- Cargill, M. *et al.* A large-scale genetic association study confirms *IL12B* and leads to the identification of *IL23R* as psoriasis-risk genes. *Am. J. Hum. Genet.* **80**, 273–290 (2007).
- Burton, P.R. *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39**, 1329–1337 (2007).
- Libioulle, C. *et al.* A novel susceptibility locus for Crohn's disease identified by whole genome association maps to a gene desert on chromosome 5p13.1 and modulates the level of expression of the prostaglandin receptor EP4. *PLoS Genet.* **3**, e58 (2007).
- Li, Y. & Abecasis, G.R.S. MACH 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.* **579**, 2290 (2006).
- Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
- Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
- Hugot, J.P. *et al.* Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
- Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
- Rioux, J.D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* **29**, 223–228 (2001).
- Yamazaki, K. *et al.* Single nucleotide polymorphisms in *TNFSF15* confer susceptibility to Crohn's disease. *Hum. Mol. Genet.* **14**, 3499–3506 (2005).
- Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 (2007).
- Peltekova, V.D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat. Genet.* **36**, 471–475 (2004).
- Moffatt, M.F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
- Tysk, C., Lindberg, E., Jarnerot, G. & Floderus-Myrhed, B. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* **29**, 990–996 (1988).
- Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638–645 (2008).
- Wedemeyer, J. *et al.* Enhanced production of monocyte chemoattractant protein 3 in inflammatory bowel disease mucosa. *Gut* **44**, 629–635 (1999).
- Dinarello, C.A. Interleukin-18 and the pathogenesis of inflammatory diseases. *Semin. Nephrol.* **27**, 98–114 (2007).
- Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
- Kazeem, G.R. & Farrall, M. Integrating case-control and TDT studies. *Ann. Hum. Genet.* **69**, 329–335 (2005).
- Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genome-wide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
- Goyette, P. *et al.* Gene-centric association mapping of chromosome 3p implicates *MST1* in IBD pathogenesis. *Mucosal Immunology* **1**, 131–138 (2008).
- Salazar-Gonzalez, R.M. *et al.* CCR6-mediated dendritic cell activation of pathogen-specific T cells in Peyer's patches. *Immunity* **24**, 623–632 (2006).
- Facco, M. *et al.* Expression and role of CCR6/CCL20 chemokine axis in pulmonary sarcoidosis. *J. Leukoc. Biol.* **82**, 946–955 (2007).
- Annuziato, F. *et al.* Phenotypic and functional features of human Th17 cells. *J. Exp. Med.* **204**, 1849–1861 (2007).
- Oppmann, B. *et al.* Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity* **13**, 715–725 (2000).
- Hue, S. *et al.* Interleukin-23 drives innate and T cell-mediated intestinal inflammation. *J. Exp. Med.* **203**, 2473–2483 (2006).
- Kullberg, M.C. *et al.* IL-23 plays a key role in *Helicobacter hepaticus*-induced T cell-dependent colitis. *J. Exp. Med.* **203**, 2485–2494 (2006).
- Uhlig, H.H. *et al.* Differential activity of IL-12 and IL-23 in mucosal and systemic innate immune pathology. *Immunity* **25**, 309–318 (2006).
- Yen, D. *et al.* IL-23 is essential for T cell-mediated colitis and promotes inflammation via IL-17 and IL-6. *J. Clin. Invest.* **116**, 1310–1316 (2006).
- Parham, C. *et al.* A receptor for the heterodimeric cytokine IL-23 is composed of IL-12R β 1 and a novel cytokine receptor subunit, IL-23R. *J. Immunol.* **168**, 5699–5708 (2002).



38. Mathur, A.N. *et al.* Stat3 and Stat4 direct development of IL-17-secreting Th cells. *J. Immunol.* **178**, 4901–4907 (2007).
39. Plowey, E.D., Cherra, S.J., III, Liu, Y.J. & Chu, C.T. Role of autophagy in G2019S-LRRK2-associated neurite shortening in differentiated SH-SY5Y cells. *J. Neurochem.* **105**, 1048–1056 (2008).
40. Van der Sluis, M. *et al.* Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology* **131**, 117–129 (2006).
41. Steinhilber, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39**, 770–775 (2007).
42. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
43. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
44. Nakazawa, A. *et al.* The expression and function of costimulatory molecules B7H and B7-H1 on colonic epithelial cells. *Gastroenterology* **126**, 1347–1357 (2004).
45. Ito, T. *et al.* Plasmacytoid dendritic cells prime IL-10-producing T regulatory cells by inducible costimulator ligand. *J. Exp. Med.* **204**, 105–115 (2007).
46. Bottini, N. *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type 1 diabetes. *Nat. Genet.* **36**, 337–338 (2004).
47. Criswell, L.A. *et al.* Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *Am. J. Hum. Genet.* **76**, 561–571 (2005).
48. Rieck, M. *et al.* Genetic variation in *PTPN22* corresponds to altered function of T and B lymphocytes. *J. Immunol.* **179**, 4704–4710 (2007).
49. Tsuji, S. *et al.* Human intelectin is a novel soluble lectin that recognizes galactofuranose in carbohydrate chains of bacterial cell wall. *J. Biol. Chem.* **276**, 23456–23463 (2001).
50. Wrackmeyer, U., Hansen, G.H., Seya, T. & Danielsen, E.M. Intelectin: a novel lipid raft-associated protein in the enterocyte brush border. *Biochemistry* **45**, 9188–9197 (2006).

¹Bioinformatics and Statistical Genetics, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. ²Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Belgium. ³University of Chicago, Department of Medicine, 5801 South Ellis, Chicago, Illinois 60637, USA. ⁴Yale University, Departments of Medicine and Genetics, Division of Gastroenterology, Inflammatory Bowel Disease (IBD) Center, 300 Cedar Street, New Haven, Connecticut 06519, USA. ⁵University of Pittsburgh, School of Medicine, Department of Medicine, Division of Gastroenterology, Hepatology and Nutrition, University of Pittsburgh Medical Center (UPMC) Presbyterian, 200 Lothrop Street, Pittsburgh, Pennsylvania 15213, USA. ⁶University of Pittsburgh, Graduate School of Public Health, Department of Human Genetics, 130 Desoto Street, Pittsburgh, Pennsylvania 15261, USA. ⁷Université de Montréal and the Montreal Heart Institute, Research Center, 5000 rue Belanger, Montreal, Quebec H1T 1C8, Canada. ⁸The Broad Institute of Massachusetts Institute of Technology and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ⁹Johns Hopkins University, Department of Medicine, Harvey M. and Lyn P. Meyerhoff Inflammatory Bowel Disease Center, 1503 East Jefferson Street, Baltimore, Maryland 21231, USA. ¹⁰Johns Hopkins University, Bloomberg School of Public Health, Department of Epidemiology, 615 E. Wolfe Street, Baltimore, Maryland 21205, USA. ¹¹Mount Sinai Hospital IBD Centre, University of Toronto, 441-600 University Avenue, Toronto, Ontario M5G 1X5, Canada. ¹²Medical Genetics Institute and Inflammatory Bowel Disease (IBD) Center, Cedars-Sinai Medical Center, 8700 W. Beverly Blvd., Los Angeles, California 90048, USA. ¹³Department of Medicine, Royal Victoria Hospital, McGill University, Montreal, Quebec, H3A 1A1, Canada. ¹⁴The Hospital for Sick Children, University of Toronto, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada. ¹⁵University of Chicago, Department of Health Studies, 5841 S. Maryland Avenue, Chicago, Illinois 60637, USA. ¹⁶Gastrointestinal Unit and Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, 185 Cambridge Street, Boston, Massachusetts 02114, USA. ¹⁷Centre National de Génotypage, Evry, France. ¹⁸Unit of Hepatology and Gastroenterology, Department of Clinical Sciences, GIGA-R, Faculty of Medicine and CHU de Liège, University of Liège, Belgium. ¹⁹Department of Gastroenterology, Clinique universitaire St Luc, UCL, Brussels, Belgium. ²⁰Department of Hepatology and Gastroenterology, Ghent University Hospital, Belgium. ²¹Department of Gastroenterology, University Hospital Leuven, Belgium. ²²Department of Gastroenterology, Erasmus Hospital, Free University of Brussels, Belgium. ²³INSERM; Université Paris Diderot; Assistance Publique Hôpitaux de Paris; Hôpital Robert Debré, Paris, France. ²⁴Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. ²⁵Peninsula Medical School, Barrack Road, Exeter, EX2 5DW, UK. ²⁶Department of Medical and Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Tower, Guy's Hospital, London, SE1 9RT, UK. ²⁷Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. ²⁸The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ²⁹IBD research group, Addenbrooke's Hospital, University of Cambridge, Cambridge CB2 2QQ, UK. ³⁰Department of Gastroenterology and Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK. ³¹Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford, OX2 6HE, UK. ³²Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, 185 Cambridge Street, Boston, Massachusetts 02114, USA. ³³This study is a joint effort of the Wellcome Trust Case Control Consortium, the NIDDK IBD Genetics Consortium and the French-Belgian IBD Consortium. ³⁴A full list of authors is provided in the **Supplementary Note** online. Correspondence should be addressed to M.J.D. (mj Daly@chgr.mgh.harvard.edu).