nature genetics

Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease

Manuel A Rivas[1–3], Mélissa Beaudoin[4,23], Agnes Gardet[5,23], Christine Stevens[2,23], Yashoda Sharma[6], Clarence K Zhang[6], Gabrielle Boucher[4], Stephan Ripke[1,2], David Ellinghaus[7], Noel Burtt[2], Tim Fennell[2], Andrew Kirby[1,2], Anna Latiano[8], Philippe Goyette[4], Todd Green[2], Jonas Halfvarson[9], Talin Haritunians[10], Joshua M Korn[2], Finny Kuruvilla[2,11], Caroline Lagacé[4], Benjamin Neale[1,2], Ken Sin Lo[4], Phil Schumm[12], Leif Törkvist[13], National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC)[14], United Kingdom Inflammatory Bowel Disease Genetics Consortium[14], International Inflammatory Bowel Disease Genetics Consortium[14], Marla C Dubinsky[15], Steven R Brant[16,17], Mark S Silverberg[18], Richard H Duerr[19,20], David Altshuler[1,2], Stacey Gabriel[2], Guillaume Lettre[4], Andre Franke[7], Mauro D'Amato[21], Dermot P B McGovern[10,22], Judy H Cho[6], John D Rioux[4], Ramnik J Xavier[1,2,5] & Mark J Daly[1,2]

More than 1,000 susceptibility loci have been identified through genome-wide association studies (GWAS) of common variants; however, the specific genes and full allelic spectrum of causal variants underlying these findings have not yet been defined. Here we used pooled next-generation sequencing to study 56 genes from regions associated with Crohn's disease in 350 cases and 350 controls. Through follow-up genotyping of 70 rare and low-frequency protein-altering variants in nine independent case-control series (16,054 Crohn's disease cases, 12,153 ulcerative colitis cases and 17,575 healthy controls), we identified four additional independent risk factors in *NOD2*, two additional protective variants in *IL23R*, a highly significant association with a protective splice variant in *CARD9* ($P < 1 \times 10^{-16}$, odds ratio ≈ 0.29) and additional associations with coding variants in *IL18RAP*, *CUL2*, *C1orf106*, *PTPN22* and *MUC19*. We extend the results of successful GWAS by identifying new, rare and probably functional variants that could aid functional experiments and predictive models.

Crohn's disease and ulcerative colitis are classified as chronic, idiopathic inflammatory bowel diseases (IBDs) of the gastrointestinal tract with unknown etiology (MIM266600). Crohn's disease occurs in about 100–150 per 100,000 individuals of European ancestry[1]. Generally, the disease affects the ileum and colon, but it can affect any region of the gut. Ulcerative colitis has similar population prevalence, and although it has some similarities to Crohn's disease in clinical manifestation, the location of inflammation is limited to the colonic mucosa. Strong familial aggregation has been observed in twin studies of Crohn's disease and ulcerative colitis[2,3]. Recent population-based sibling risk is 26-fold greater for Crohn's disease and 9-fold greater for ulcerative colitis[2], and overall Crohn's disease and ulcerative colitis concordance rates in nonselected twin studies are 30% and 15%, respectively, among monozygotic twins compared with 4% for Crohn's disease or ulcerative colitis among dizygotic twins[3]. Like most complex diseases, Crohn's disease and ulcerative colitis result from a combination of genetic and nongenetic risk factors, and each individual factor probably has a modest effect on disease risk[4].

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK. [4]Université de Montréal and Research Centre, Montreal Heart Institute, Montreal, Quebec, Canada. [5]Gastrointestinal Unit, Center for the Study of the Inflammatory Bowel Disease and Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [6]Yale School of Medicine, New Haven, Connecticut, USA. [7]Institute of Clinical Molecular Biology, Kiel, Germany. [8]Unit of Gastroenterology, Istituto Di Ricovero e Cura a Carattere Scientifico, Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy. [9]Örebro University Hospital, Department of Medicine and School of Health and Medical Sciences, Örebro University, Örebro, Sweden. [10]The Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [11]Clarus Ventures, Cambridge, Massachusetts, USA. [12]Department of Health Studies, University of Chicago, Chicago, Illinois, USA. [13]Karolinska Institutet, Department of Clinical Science Intervention and Technology, Stockholm, Sweden. [14]A full list of consortium members is provided in the **Supplementary Note**. [15]The Pediatric Inflammatory Bowel Disease Center, Cedars-Sinai Medical Center, Los Angeles, California, USA. [16]Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA. [17]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [18]Mount Sinai Hospital Inflammatory Bowel Disease Group, University of Toronto, Toronto, Ontario, Canada. [19]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [20]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [21]Karolinska Institutet, Department of Biosciences and Nutrition, Stockholm, Sweden. [22]Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [23]These authors contributed equally to this work. Correspondence should be addressed to M.J.D. (mjdaly@atgu.mgh.harvard.edu) or M.A.R. (rivas@broadinstitute.org).

Received 5 January; accepted 31 August; published online 9 October 2011; doi:10.1038/ng.952

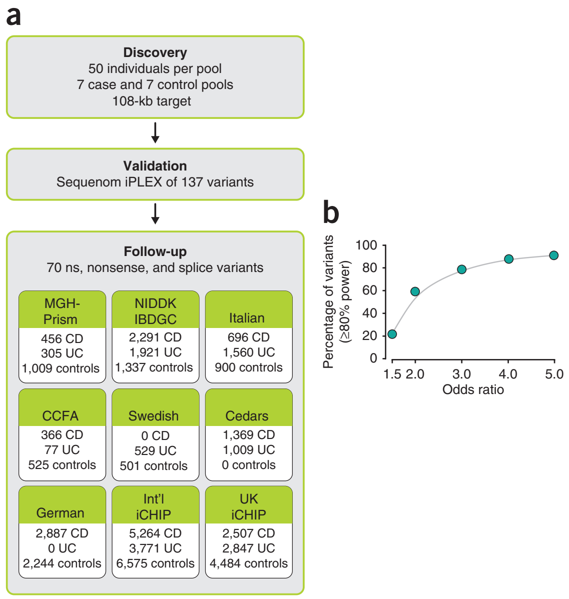


Figure 1 Overview. (**a**) Schematic of Crohn's disease rare-variant phenotype project. (**b**) Power to detect single-marker rare-variant association in follow-up sample sets. We report the results of the Crohn's disease pooled resequencing project with follow-up genotypes in 13,167 Crohn's disease cases, 12,153 ulcerative colitis cases and 15,331 healthy controls. We report that of the 70 markers successfully genotyped, 22%, 60%, 79%, 88% and 91% have at least 80% power to detect association at minor allele frequency ORs of 1.5, 2, 3, 4 and 5, respectively (see also **Supplementary Fig. 3a**,**b**), suggesting that we can address the contribution of rare and low-frequency polymorphisms in GWAS loci to IBD. OR, odds ratio.

Common immune-mediated diseases such as IBD have a genetic basis. However, until recently, identifying disease susceptibility genes has been challenging for common, polygenic disease[5,6]. With the development of HapMap and GWAS technology, the number of bona fide risk loci that have been identified and replicated for complex trait genetics in general, and for IBD in particular, has markedly increased. In Crohn's disease, individual GWAS scans and follow-up meta-analyses have identified >71 susceptibility loci and have provided insights beyond the two loci established before the GWAS era[7,8]. Similarly, in ulcerative colitis, GWAS efforts have identified 47 susceptibility loci[9,10] and, after accounting for the many alleles associated with both diseases, 99 distinct associations have been documented for IBD. Although these findings have clarified disease pathways, the common SNPs identified are generally of modest effect and explain only ~23% of the overall variance in Crohn's disease risk. Moreover, most of the associated variants do not have known or obvious function, and many implicate regions with multiple genes, limiting biological extrapolation.

SNPs implicated by GWAS are tightly correlated with other SNPs in the region and are probably in linkage disequilibrium (LD) with the causal variant rather than causal themselves. A complete catalog of all variation is required in the search for causal variants[11,12]. However, even in denser reference data from the 1000 Genomes Project, most GWAS hits are not correlated with an obvious functional variant and therefore do not conclusively implicate a unique gene. If independently associated rare coding variation is discovered in a gene within a region implicated by GWAS, the gene harboring such variants is directly implicated. Furthermore, additional heritability can be explained and specific alleles identified for direct functional experimentation. In Crohn's disease, multiple independent associated alleles have been documented at *NOD2* and *IL23R*[13,14]. Exhaustive sequencing of genomic regions has recently become possible for the first time with the advent of next-generation sequencing (NGS) technologies. Growing collections of genome sequences through international efforts like the 1000 Genomes Project are driving the development of laboratory study designs and analytic methods for using large-scale genomic sequencing in human genetic discovery[15].

Targeted sequencing of pooled samples allows researchers to efficiently and cost-effectively capture all variation in a limited target region that has been selectively amplified in multiple DNA samples[16,17]. This approach allows efficient use of NGS technologies, which generate billions of base pairs per experimental unit yet introduce challenges in data processing and analysis, for the discovery of new variants and assess their potential association with disease. We describe here a pooled NGS study of 350 Crohn's disease cases and 350 controls across coding exons of 56 genes contained in regions of confirmed association with Crohn's disease[7], and we introduce new SNP calling methods for pooled targeted sequencing projects implemented in the software Syzygy. We further evaluated new, potentially functional rare variants identified in the survey in nine independent case-control series, confirming a role for functional, rare variants in *CARD9*, *NOD2*, *IL23R* and *IL18RAP* and identifying others in *MUC19*, *CUL2*, *PTPN22* and *C1orf106* that were significantly associated with IBD. The results lend further support to an emerging paradigm in both rare diseases (Hirschsprung's disease and Bardet-Biedl syndrome) and common phenotypes (serum lipids, QT-interval and height and type 1 diabetes) in which both common, low-penetrance and rarer, often higher-penetrance alleles exist in the same gene. They also suggest that deep sequencing of regions implicated by GWAS may be effective in increasing knowledge about the heritability of specific functional alleles in complex disease[16,18–21].

RESULTS

Discovery of new variants using pooled sequencing

We selected 350 Crohn's disease cases and 350 healthy controls of European ancestry from among samples collected by the NIDDK IBDGC with genome-wide SNP data[14,22]. We pooled samples in batches of 50 cases or 50 controls matched for European ancestry using GWAS data. One pool of 50 cases was drawn from self-reported and empirically confirmed (by GWAS data[22]) Jewish ancestry and was matched with one pool of 50 equivalently defined Jewish controls. The remaining pools of cases and controls were selected from the non-Jewish European-American samples. Samples were pooled only after two rounds of quantification and normalization to ensure that the initial DNA pool accurately reflected sample allele frequencies. For each pool, we carried out PCR amplification to capture the 107.5-kb target region, which included 645 nuclear-encoded exons (**Supplementary Tables 1** and **2**). We amplified each sample in 593 PCR reactions. The successful PCR amplicons were combined in equimolar amounts, concatenated and sheared to construct libraries. The 14 libraries were sequenced using Illumina Genome Analyzer flow cells, with one pool per lane (see Online Methods; **Fig. 1a**). High-throughput sequencing yielded large amounts of high-quality data for each pool. We captured 91% of our nuclear target regions at ≥100× coverage and achieved 1,500× median coverage per pool (corresponding to 30× per sample or 15× per individual chromosome; **Supplementary Fig. 1**).

We next aimed to identify rare and low-frequency single-nucleotide variants (SNVs) in the pooled samples. We developed a variant calling method, Syzygy, to accommodate the specific pooled study design

Table 1 Variant discovery summary

Category	High quality	Moderate quality
Variants identified	429	173
dbSNP (%)	45	24
NS/S	1.4	1.7
Ti/Tv	2.3	1.4

Using Syzygy, we detected 429 high-confidence variants (240 nonsynonymous sites, 169 synonymous sites and 20 variants within 5 bp of the nearest splice site) within our 107.5-kb targeted region with a dbSNP rate of 45%, NS/S of 1.42, and Ti/Tv of 2.3 in the pooled sequencing experiment with 350 Crohn's disease cases and 350 healthy controls.

and identify rare variants (see **Supplementary Methods**). Through empirical modeling of the sequencing error processes and filters to remove sites with strand inconsistency or clusters of variants suggestive of read misalignment, Syzygy detected 429 putatively high-confidence variants (240 nonsynonymous sites, 169 synonymous sites and 20 intronic variants within 5 bp of a splice junction) within our 107.5-kb targeted region, with 45% of the variants already included in dbSNP using dbSNP version 132, nonsynonymous/synonymous ratio (NS/S) of 1.42 and transition/transversion ratio (Ti/Tv) of 2.3 (**Table 1**). Because we designed our experiments to detect variants correctly at the limit of machine quality, we estimated the proposed set of false-positive SNPs that would need to be eliminated in subsequent genotyping. Both the proportion of variants in dbSNP and the Ti/Tv suggest a relatively high true-positive rate in this data set. Specifically, high-depth individual-level sequencing of 1,000 genes carried out by the 1000 Genomes Project (called Pilot 3) in 697 samples identified a high-quality SNP set with the same dbSNP percentage (dbSNP version 129), whereas the Ti/Tv detected here suggests a ~90% true-positive rate[23]. To confirm this, we selected a random subset of 137 high-confidence functional nonsynonymous, nonsense and putative splice-variant SNPs for Sequenom iPLEX genotyping of all samples in the sequenced pools and validated 91.2% of them (**Fig. 1a**). Using a canonical expectation of $\left(\theta \times \sum_{i=1}^{n-1} \frac{1}{i} \times Nbases\right)$, where n denotes the number of chromosomes and Nbases represents the targeted bases, or the rate observed directly in 1000 Genomes Pilot 3, we would expect to see ~470 variants across the successfully queried target. Sensitivity for singletons, however, is incomplete at the lower end of coverage in our experiment (**Supplementary Fig. 1**) and accounts for the modest deficit in our study.

A challenge in pooled genotyping or sequencing experiments is accurate recovery of allele frequencies. We were surprised to observe a strong correlation between genotype frequencies and frequencies estimated for sequence data ($r^2 \approx 0.99$) using the method in Syzygy, suggesting that accurate quantification of DNAs in the pooling steps led to experimental recovery of the pool composition. We also observed a strong correlation between the case-control test statistic estimated with the pooled data and the test statistic in the genotype data ($r^2 \approx 0.925$; **Supplementary Fig. 2**).

To test the role of these rare variants, we identified all nonsynonymous, nonsense or splice-site variants that occurred in two or more copies up to a frequency of 5%, for a total of 115 variants (**Supplementary Table 3**). Excluding known GWAS-associated low-frequency coding variants at *NOD2*, *IL23R* and *LRRK2-MUC19*, we carried out follow-up genotyping for 70 of these markers in nine independent case-control series totaling 16,054 Crohn's disease cases, 12,153 ulcerative colitis cases and 17,575 healthy controls. These included (i) samples from the Prospective Registry in IBD Study at MGH (PRISM); (ii) samples assembled from throughout North America and Australia by the NIDDK IBDGC; (iii) an Italian-Dutch case-control sample; (iv) Crohn's and Colitis Foundation of America (CCFA) Repository Collection; (v) Swedish samples; (vi) Cedars samples; (vii) German samples and Immunochip genotype data provided by (viii) the International IBD Genetics Consortium and (ix) UK IBD Genetics Consortium (rare coding variants discovered in this study contributed to the Immunochip design; **Fig. 1a**). Samples i and iii–vii were genotyped for sets of markers using Sequenom iPLEX. Sample ii genotyping was done as part of a larger NIDDK IBDGC Illumina GoldenGate study. Because of design constraints and assay failures, not all markers were examined in all nine follow-up sample sets (see **Supplementary Methods** for details of follow-up genotyping). We demonstrate that the current study design is well positioned to address the overall contribution of variants in coding regions of GWAS loci to IBD (**Fig. 1b**, **Supplementary Fig. 3** and **Supplementary Methods**).

The few nonreference alleles expected for many of these variants in each substudy precludes the use of asymptotic statistics common to most association studies. Population structure is probably an even more substantial problem at low frequencies, demanding a stratified analysis retaining strict population case-control matching. Therefore, we used a mega-analysis of rare variants (MARV) that provides a permutation-based estimate of significance, constraining all permutations to be within each subgroup and thus accommodating arbitrary numbers of sample subsets of diverse population and case-control origin without power loss for single-marker and group-marker analysis (see Online Methods). Given a target set of 70 variants, we would expect <1 SNP to exceed $P < 0.01$ by chance in the follow-up analyses and define traditional experiment-wise significance to be $P = 0.0007$. As we explored both Crohn's disease and ulcerative colitis in follow-up studies, our primary analysis compared all IBD (Crohn's disease and ulcerative colitis) cases versus controls to maximize power for genes in which the same common variants have been conclusively associated with both diseases with similar effect (such as *CARD9*). For genes specifically associated with Crohn's disease only (such as *NOD2*), the ulcerative colitis group was combined with controls (for details, see ref. 10).

New protective splice variant in *CARD9*

CARD9 is associated with both Crohn's disease and ulcerative colitis risk, with a common coding variant (rs4077515 creating substitution p.S12N with both alleles of roughly equal frequency) that represents a 'typical' GWAS hit (odds ratio (OR) ≈ 1.2 in both diseases)[8,9]. In the pooled sequencing, we identified a splice-site variant in *CARD9* (**Fig. 2** and **Supplementary Fig. 4**) that altered the first base after exon 11 in six controls and zero cases, suggesting a potentially strong protective effect. Follow-up analyses confirmed a highly significant association ($P < 10^{-16}$), with the allele appearing in about 0.20% of cases and 0.64% of controls (OR ≈ 0.3; **Table 2** and **Supplementary Table 4**).

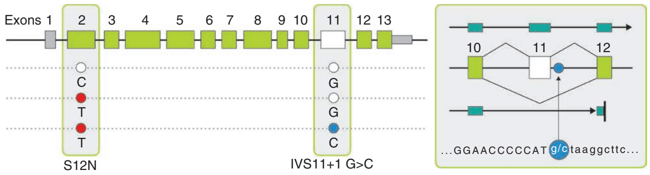


Figure 2 *CARD9* protective splice-site variant and predicted transcript. Splice-site variant IVS11+1C>G (OR = 0.29), conferring protection against Crohn's disease with predicted transcript. This hypothetical transcript has been observed in spleen, lymph-node and PBMC-derived cDNA libraries. We predict exon 11 to be skipped and the alternative transcript to include exon 9 mRNA sequence continuing to exon 12, including 21 amino acids before reaching a premature stop.

Table 2 Identification of additional rare and protective variants associated with IBD

a CD versus UC + HC

CD versus UC + HC (CD loci)	Targeted replication			International Immunochip						
	Samples			Samples					Combined	
Gene, mutation	Allele frequency			Allele frequency			Samples		OR	
chr:position[a]	CD	UC + HC	*P*	CD	UC + HC	*P*	CD	UC + HC	(L95,U95)	*P*
NOD2, p.M863V+fs1007insC	7,969	10,179	6.73×10^{-11}	6,544	16,126	2.15×10^{-7}	14,523	26,305	4.02	$<1 \times 10^{-16}$
16:49308343	0.0067	0.00157		0.0036	0.0011				(2.80,5.07)	
NOD2, p.N852S	7,962	9,590	0.00017	6,542	16,121	0.0338	14,504	25,711	2.47	2.90×10^{-5}
16:49308311	0.0046	0.0021		0.001	0.000465				(1.55,3.93)	
NOD2, p.R703C	3,090	4,100	0.00025	8,416	17,183	1.59×10^{-4}	11,506	21,283	1.51	2.33×10^{-7}
16:49303430	0.011	0.0054		0.0079	0.0052				(1.12,2.03)	
NOD2, p.S431L	7,949	9,569	0.0014	6,545	16,124	0.023	14,494	25,693	1.45	0.00025
16:49302615	0.0039	0.0019		0.0038	0.0026				(1.07,1.95)	
NOD2, p.V793M	2,227	3,252	0.0217	6,949	16,156	0.0127	9,176	19,408	1.45	0.002
16:49303700	0.0034	0.0015		0.004	0.0026				(1.07,1.95)	
NOD2, p.R311W	3,010	5,506	0.118	6,950	16,149	0.029	9,960	21,655	2.28	0.00143
16:49302254	0.0017	0.00099		0.0014	0.00073				(1.37,3.79)	
IL18RAP, p.V527L	7,920	9,561	0.0006	4,131	10,336	0.0456	12,051	19,897	3.03	2.90×10^{-4}
2:102434852	0.0036	0.0015		0.00025	0				(1.95,4.73)	
MUC19, p.V56M	2,227	3,253	0.033	4,963	11,324	0.11	7,190	14,577	4.32	0.00546
12:39226476	0.0029	0.00138		0.0003	0.00004				(1.93,9.67)	

b IBD versus HC

	Targeted replication			International Immunochip						
IBD versus HC (CD + UC loci)	Samples			Samples					Combined	
Gene, mutation	Allele frequency			Allele frequency			Samples		OR	
chr:position[a]	IBD	HC	*P*	IBD	HC	*P*	IBD	HC	(L95,U95)	*P*
CARD9, c.IVS11+1G>C	10,439	5,933	1.90×10^{-8}	16,420	10,707	3.33×10^{-16}	26,859	16,640	0.29	$<1 \times 10^{-16}$
9:138379413	0.002	0.0058		0.0024	0.0071				(0.22,0.37)	
IL23R, p.V362I	5,321	6,112	0.27	12,241	10,426	2.70×10^{-5}	17,562	16,538	0.72	1.18×10^{-5}
1:67478488	0.0131	0.0127		0.011	0.0152				(0.63,0.83)	
IL23R, p.G149R	4,629	5,305	0.064	13,789	10,707	0.0013	18,418	16,012	0.60	3.20×10^{-4}
1:67421184	0.0026	0.0045		0.0025	0.0043				(0.45,0.79)	
CUL2, c.IVS17+5A>G	5,582	1,684	0.2	16,387	10,707	0.0004	21,969	12,391	0.72	3.45×10^{-4}
10:35354137	0.0056	0.0063		0.0065	0.0092				(0.60,0.86)	
PTPN22, p.H370N	5,583	1,682	0.3	21,997	12,393	0.0046	21,997	12,393	1.60	6.20×10^{-3}
1:114182437	0.003	0.002		0.0031	0.002				(1.16, 2.24)	
C1orf106, p.Y333F	13,991	8,486	0.009	NA	NA	NA	13,991	8,486	1.44	0.009
1:199144649	0.013	0.01							(1.02, 2.06)	

[a] NCBI human genome build 36 coordinates.

CD, Crohn's disease; UC, ulcerative colitis; HC, healthy controls.

We identified IVS11+1G>C to be protective against IBD with an estimated OR of 0.29 (four-fold protective effect). Five independent rare variants in *NOD2* were associated with Crohn's disease, including R311W, R703C, S431L+V793M, N852S and M863V+fs1007insC. Additional variants conferring protection against IBD were identified in *IL23R* and *CUL2*, and risk missense variants were identified in *IL18RAP*, *C1orf106*, *MUC19* and *PTPN22*.

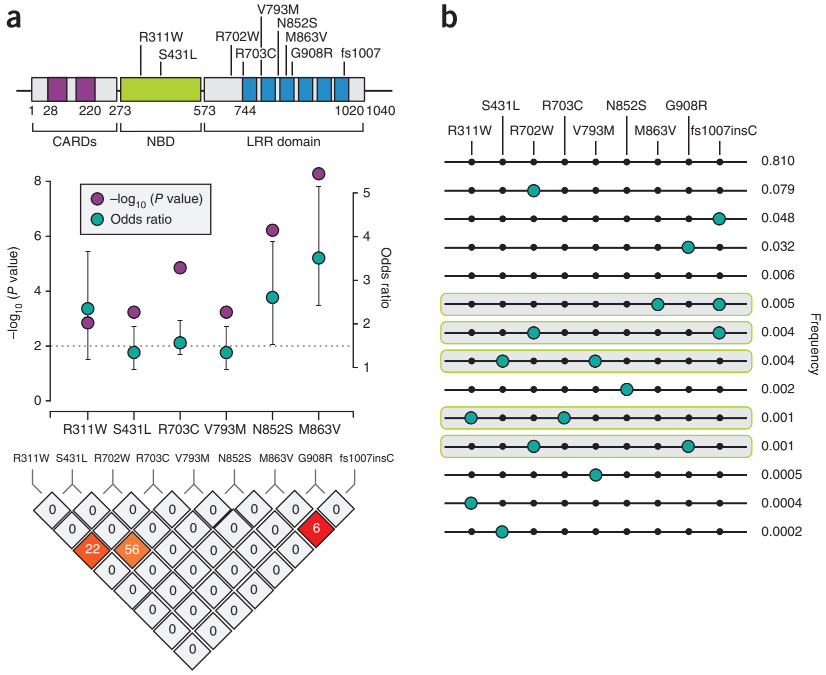
Although skipping exon 11 places translation out of frame, we predict that the resulting transcript would escape nonsense-mediated decay as premature truncation occurs close to the final splice junction in exon 12. Indeed, this hypothetical transcript (**Fig. 2** and **Supplementary Fig. 4**) has been observed in cDNA libraries derived from spleen, lymph node and peripheral blood mononuclear cells (PBMCs). Notably, this rare protective variant occurs on a haplotype carrying the risk allele at rs4077515, indicating not only that the two associations are independent but also that the splice variant completely eliminates the risk normally associated with the common haplotype. Because the Crohn's disease risk allele at rs4077515 has been associated with higher expression of *CARD9*, a consistent allelic series may exist if the splice variant is substantially lower or nonfunctional and therefore highly protective.

Rare risk variants in *NOD2*

NOD2 encodes a member of a family of human cytosolic, non–Toll/IL-1 receptor (TIR) neuronal apoptosis inhibitor protein (NACHT)-leucine-rich repeat (LRR) proteins[24] first implicated in Crohn's disease[13,25] and later discovered to be involved in Blau syndrome[26]. The three previously known causal mutations, R702W, G908R and fs1007insC, are in the LRR domain of NOD2, whereas the mutations identified in Blau syndrome are in the highly conserved NACHT nucleotide-binding domain.

We identified five distinct rare variants (R311W, S431L, R703C, N852S and M863V), and several others in LD with one of these, that are independently associated with Crohn's disease risk (**Table 2** and **Supplementary Table 4**). The S431L ($P = 0.0004$) (and the rarer V793M contained on a subset of S431L haplotypes), R703C ($P = 2.3 \times 10^{-5}$) (previously suggested in one study to be associated[27]) and N852S ($P = 1.1 \times 10^{-6}$) variants are found on distinct haplotypes that do not contain the known causal mutations R702W, G908R and fs1007insC (**Fig. 3a,b**) and are thus completely independent risk variants. R311W shares a subset of haplotypes with R703C (**Fig. 2**); however, conditional analysis and haplotype testing indicate that both alleles probably contribute independently to risk (**Supplementary Table 5**). M863V is a rarer variant that has arisen on the haplotype background of fs1007insC, and although the risk estimate of M863V+fs1007insC (OR = 4.02 (95% CI = 2.8–5.7)) is higher than the risk attributable to fs1007insC alone (OR = 3.16 (95% CI = 2.9–3.4)),

Figure 3 Identification of additional rare variants in *NOD2* associated with Crohn's disease. (**a**) Five additional risk variants were discovered in *NOD2*. $-\log_{10}P$ and minor allele ORs with 95% confidence intervals indicated with error bars and haplotype block, where D' taking values from 0–1 (white to red) represents the extent of linkage disequilibrium between markers and numbers represent r^2 between markers. (**b**) *NOD2* haplotypes observed in 700 individuals with overlapping genotype data (R311W, S431L, R702W, R703C, V793M, N852S, M863V, G908R and fs1007insC). S431L and V793M are in tight LD and we regard this as one unit (S431L V793M). R703C is at a higher frequency than R311W although they share haplotypes. Conditional analysis (**Supplementary Table 3**) demonstrates independent contributions. M863V lies on background haplotype of fs1007insC.



the low frequency of M863V makes its functionality unclear. Thus, in later calculations of variance explained, we did not count this as an additional risk factor.

Functional assessment of additional associated alleles in *NOD2*

Through assays to identify the effect of the mutations on NOD2 intracellular localization, we found that the S431L mutant and the well-studied insertion mutant (fs1007insC) did not localize to the membrane area, in contrast with N852S (**Fig. 4**). We next determined whether the NOD2 mutants S431L and N852S activated NF-κB in response to the NOD2 ligand muramyl dipeptide (MDP). HEK293T cells were transfected with the point mutants, wild-type NOD2 and the well-studied fs1007insC mutant (**Fig. 4**). Western blot analysis showed that the point mutations did not affect expression compared with the wild-type protein (**Fig. 4**). As published earlier[28,29], the fs1007insC mutant did not induce NF-κB activation after MDP stimulation. The MDP-induced NF-κB activation was also impaired in the presence of S431L and N852S (**Fig. 4**).

Together, these results indicate that the N852S alteration in the LRR domain may perturb MDP recognition without affecting NOD2 intracellular localization, similarly to the common R702W and G908R alterations[28]. This contrasts with the fs1007insC mutation, which also affects targeting of NOD2 to the membrane area. The S431L variant is in the nucleotide-binding domain of the protein and impairs both localization and MDP-induced NF-κB activation. These findings are similar to earlier studies demonstrating that critical residues in the nucleotide-binding domain region attenuate MDP-dependent NF-κB activation[29]. Further studies are needed to determine the instructive role of NOD2 mutants in coordinating autophagy, control of cellular stress signals and adaptive immune responses.

N852S and M863V variants are more common in Ashkenazi Jewish individuals

The highest reported prevalence of Crohn's disease is in subjects of Ashkenazi Jewish descent, with two to four times higher prevalence than non-Jewish populations[30]. An earlier study[31] has screened the *NOD2* gene for rare variants and identified five previously unreported changes (D113N, D357A, I363F, L550V and N852S). N852S occurs only in Ashkenazi Jewish individuals and has been proposed to predispose disease, with seven transmissions and only one nontransmission from heterozygous parents to affected offspring in an Ashkenazi Jewish family collection, concordant with the case-control observations in this study. In our study, Ashkenazi Jewish individuals had a much higher frequency of both N852S and M863V (4% and 2%, respectively, in Ashkenazi Jewish Crohn's disease cases, and 0.5% for N852S and M863V in non-Ashkenazi Jewish Crohn's disease cases). This accounts for the greater incidence of these alleles in the first replication column of **Table 2**, as NIDDK studies in particular had a specific and significant Ashkenazi Jewish collection.

We examined the haplotype carrying N852S in Ashkenazi Jewish individuals (determined given the existence of two homozygote cases) and in non-Ashkenazi Jewish individuals in the subset of samples with existing GWAS genotype data[8,9,14,22]. We found that the N852S variant in Ashkenazi individuals lies on a unique extended haplotype of several megabases (≥2 Mb to the left and right). However, the N852S variant in non-Ashkenazi Jewish individuals does not share the extended background haplotype. In Ashkenazi individuals, the average shared distance between a pair of N852S chromosomes is ≥4 Mb, whereas in non-Ashkenazi individuals, the average shared distance between a pair of N852S chromosomes is 0.5 Mb (**Supplementary Fig. 5**), suggesting that the variant is reasonably old but that a single copy was stochastically enriched in the recent Ashkenazi bottleneck ~25 generations ago.

Rare protective variants in *IL23R*

We also identified significant protective effects of substitutions G149R ($P = 3.2 \times 10^{-4}$) and V362I ($P = 1.2 \times 10^{-5}$) in *IL23R*. This confirms recent findings[32] and indicates that each of these variants have a protective effect equivalent to that of the more common R381Q substitution (**Table 2** and **Supplementary Table 4**), although they arose on different haplotype backgrounds and are not in LD with R381Q. Despite the large follow-up sample size (n = 31,747), we did not find evidence for a protective effect of the previously reported R86Q variant (P = 0.94). *IL23R* signaling is attenuated in T helper type 17 (T_H17) cells generated from healthy subjects carrying the R381Q substitution, leading to a decrease of IL-17A secretion in response to IL-23, and indicating that R381Q is associated with reduced T_H17

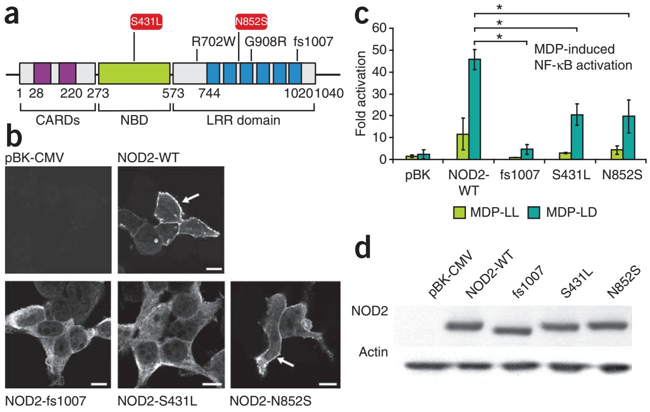


Figure 4 Functional analyses of NOD2 variants. (**a**) Schematic of NOD2 protein domains and localization. (**b**) HEK293T cells were transfected with NOD2 constructs and fixed using 4% paraformaldehyde at 24 h after transfection. Cells were then subjected to immunofluorescent staining to detect NOD2 and fluorescence was collected using a confocal microscope. Image gallery of a single confocal section. (**c**) HEK293T cells were transfected with NOD2 constructs and reporter plasmids encoding firefly luciferase cloned under a promoter containing NF-κB elements and with a plasmid encoding renilla luciferase as a transfection control. After 24 h, cells were stimulated with MDP–L-alanine–L-glutamine (LL) or MDP–L-alanine–D-glutamine (LD) (10 μg/ml) for 6 h. Transcriptional activation was quantified by ratios of firefly luciferase activity to renilla luciferase activity. Data were normalized to unstimulated condition with empty vector transfection. Statistical analyses were carried out using Student's *t*-test (**P* < 0.05). Error bars represent 95% CI for fold activation. (**d**) Cell lysates were also collected and subjected to western blot analysis to detect NOD2 and actin expression levels. Scale bars, 10 μm.

responses[33]. In addition, recent studies have highlighted a role for IL-23 in T_H17 cell lineage commitment without TGF-β. This alternate mode of T_H17 differentiation, dependent on *IL23R* expression, seems to have a greater pathogenic role than TGF-β–induced T_H17 differentiation, highlighting the value of discovering protective variants in autoimmunity[34]. Future therapies for autoimmune disease should consider the phenotypic characteristics of pathogenic T_H17 cells generated without TGF-β, and their signaling pathways as possible targets.

Additional rare risk and protective variants

Although Crohn's disease and ulcerative colitis do not share an association with the common variant rs2058660 in *IL18RAP* (minor allele frequency (MAF) = 0.23, OR ≈ 1.19 for Crohn's disease, chr2:102.17–102.67 Mb), an association of rs2058660 with celiac disease has recently been documented[35]. We identified a rare risk missense variant, V527L (MAF = 0.003), in *IL18RAP* with an estimated minor allele OR of 2.79 for Crohn's disease. In addition, a low-frequency missense variant, Y333F (MAF = 0.008), in *C1orf106* was associated with risk of both Crohn's disease and ulcerative colitis.

A common *CUL2* variant (rs12261843, MAF = 0.30, OR ≈ 1.15) is associated with both Crohn's disease and ulcerative colitis risk[8,9]. In the pooled sequencing experiment, we identified a splice-site variant in *CUL2* altering a nucleotide five bases downstream of exon 17 with an estimated OR of 0.72 in the follow-up samples (MAF = 0.007). Notably, several members of the ubiquitin proteosome are present in the autophagy interaction network, including *CUL2*, suggesting cross-talk between these processes in intracellular quality control and immunity[36].

A common missense variant (risk allele frequency = 0.90; OR = 1.31, rs2476601) in *PTPN22* is associated with Crohn's disease[7,8], type 1 diabetes[37], rheumatoid arthritis[38] and vitiligo[39]. In this instance, the direction of association differs in different diseases, with the minor allele (Trp) strongly associated with type 1 diabetes, rheumatoid arthritis and vitiligo but highly protective against Crohn's disease. Analysis of rare variants in IBD cases versus healthy controls showed a modest risk effect (P = 0.00026, minor allele OR = 1.6) for a rare (MAF = 0.003) *PTPN22* missense variant (H370N). Ongoing studies in other autoimmune diseases may help clarify the relevance of H370N and rs2476601 in different conditions.

Through examination of haplotype structure (**Supplementary Fig. 6**) and formal conditional analysis (**Supplementary Table 6**), we found that the rare variants highlighted in *IL18RAP*, *MUC19*, *C1orf106*, *PTPN22* and *CUL2* are independent of the common associated GWAS variants. Specifically, the rare variants at *IL18RAP* and *MUC19* reside on the common higher-risk background but confer independently significant risk, the rare variants at *PTPN22* and *C1orf106* occur on the common low-risk background and are therefore independent, and the rare variant at *CUL2* is protective and in weak LD with common-risk variants at that locus.

Heritability estimates of rare associated variants

We estimated the fraction of additive genetic variance explained using the liability threshold model[40,41], which assumes an additive effect at each locus and shifts the mean of a normal distribution of disease liability for each genotype class. We assumed a prevalence of Crohn's disease of 4 per 1,000 and a total narrow-sense heritability of 50% (ref. 42). We estimate that the discovered rare and low-frequency variants associated with Crohn's disease in this study contribute another 1–2% genetic variance over all populations and 2–3% genetic variance to the Ashkenazi Jewish population (**Supplementary Table 7**).

DISCUSSION

Genome-wide association has been highly successful in IBD, with 99 confirmed associations providing new insight into disease biology. However, it is the ~75% of heritability yet to be explained that fuels most debate in human genetics. NGS offers potential insights into both the biology and the heritable component explained by GWAS results by completing the allelic spectrum of functional alleles in cases and controls, including rare variation.

Using a targeted pooled approach, we carried out an efficient and cost-effective scan for rare and low-frequency polymorphisms in genes from regions identified as relevant through GWAS. After extensive follow-up genotyping, we identified highly significant variants at *CARD9*, *NOD2*, *CUL2* and *IL18RAP* that contribute to risk independently of previously defined variants at these loci, and we showed the functionality of the newly implicated *NOD2* variants. We report additional protective variants at *IL23R* and identify nominally significant variants in *MUC19*, *PTPN22* and *C1orf106* more frequently than expected by chance.

The results of this experiment are relevant to ongoing debates in human genetics. Although we found little support for the hypothesis that common-variant associations are simply an indirect LD-driven by-product of higher penetrance rare alleles, additional independently acting low-frequency alleles in genes implicated by common-variant association are documented. In the case of the *CARD9* splice variant, this newly discovered allele explains more of the overall population variance in risk than does the common S12N variant (about 0.3% and 0.2%, respectively). If these observations become commonplace through available technology, they may help make the debates about common versus rare variation biologically irrelevant. As in many quantitative traits and Mendelian disorders, we observed common alleles of modest effect and rarer alleles with more considerable impact coexisting in the same genes, with both types of variation providing insight into the same disease biology. More than simply increasing variance explained, these results will likely be of great value to functional biology. In addition to the functional confirmation of *NOD2* alleles, the identification of a new *CARD9* isoform that strongly protects against disease development provides a way to study disease biology and a model that could be mimicked therapeutically. Finally, our study suggests that additional variants should be routinely searched for by thorough sequencing of genes within significantly associated regions in GWAS in large sets of cases and appropriate controls, not simply to expand incrementally the variance explained, but to identify specific alleles that may substantially advance our understanding of the functional role of each gene.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. GenBank: *NOD2*, NP_071445.1; *IL23R*, NP_653302.2; *CARD9*, NP_434700.2; *CUL2*, NM_003591; *IL18RAP*, NP_003844.1; *PTPN22*, NP_057051.3; *C1orf106*, NP_060735.3; *MUC19*, AAP41817.1.

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

ACKNOWLEDGMENTS

We thank L. Solomon for designing figures, R. Onofrio for designing PCR primers, K. Ardlie and S. Mahan for assistance in DNA sample preparation, J. Wilkinson and L. Ambrogio for Illumina sequence project management, B. Thomson, G. Crenshaw and A. Taylor for genotyping assistance, E.S. Lander for helpful comments, the Broad Institute Sequencing Platform for sequence data production, J. Flannick and J. Maguire for assistance with pooled sequence analysis, and M. dePristo and the Broad Institute Genome Sequence Analysis Group for sequencing analysis help and useful discussions. The work was supported by US National Human Genome Research Institute sequencing grant DK83756; Funds from the Helmsley Trust to M.J.D. and R.J.X.; and US National Institutes of Health grants AI062773, DK060049, DK086502, HG005923 and DK043351 (to R.J.X.). A.G. was supported by fellowships from La Fondation pour la Recherche Medicale and the CCFA. J.D.R. holds a Canada Research Chair and is funded by grants from the US National Institutes of Allergy and Infectious Diseases (AI065687; AI067152) from the NIDDK (DK064869; DK062432), the CCFA (SRA512), La Fondation de l'Institut de Cardiologie de Montréal, the Crohn's and Colitis Foundation of Canada (CCFC), the Fonds de Recherche en Santé du Québec (17199) and the Canadian Institutes of Health Research (01038). Genotyping of the Italian samples was supported by the Italian Ministry for Health GR-2008-1144485 and unrestricted research grant of Giuliani, with case collections provided by the Italian Group for IBD. The NIDDK IBDGC is funded by the following grants: DK062431 (S.R.B.), DK062422 (J.H.C.), DK062420 (R.H.D.), DK062432 (J.D.R.), DK062423 (M.S.), DK062413 (D.P.B.M.) and DK062429 (J.H.C.). D.P.B.M. is supported by NCRR grant M01-RR00425 to the Cedars-Sinai General Research Center Genotyping Core, U01-DK062413 (IBD Genetics Research Center), P01-DK046763 (IBD Program Project Grant) and Southern California Diabetes Endocrinology Research Center grants DK063491, R21-DK84554-01 and U01 DK062413. Genotyping of the German samples was supported by the German Ministry of Education and Research through the National Genome Research Network with infrastructure support through the Deutsche Forschungsgemeinschaft cluster of excellence Inflammation at Interfaces. Genotyping of the Swedish samples was supported by the Swedish Research Council, the Bengt Ihre Foundation and the Örebro University Hospital Research Foundation.

AUTHOR CONTRIBUTIONS

M.A.R. and M.J.D. conceived and designed the study. Functional characterization of *NOD2* mutants was coordinated and designed by A.G. and R.J.X. Study subject recruitment and phenotyping was supervised by R.H.D., M.C.D., D.P.B.M., M.D., R.J.X., J.H.C., J.D.R., M.C.D., M.D., A.F., D.E., M.S.S., and A.L. Sequenom assay designs were developed by P.G., T.H., J.H., L.T., and A.K. NIDDK IBDGC BeadXpress typing was coordinated and supervised by Y.S. and J.H.C. The pooled sequencing protocol was designed and established at the Broad Institute by N.B., M.A.R., C.S., D.A., M.J.D. and S.G. NIDDK IBDGC, UK IBDGC and IIBDGC contributed sample collection and Immunochip genotype data for replication. The project was managed by M.A.R., G.L., M.S., J.D.R., J.H.C., R.J.X., D.P.B.M., R.H.D., S.R.B. and M.J.D. C.S. and M.B. carried out pooling. C.S., Y.S., P.G., C.L., D.E. and M.B. carried out genotyping. M.A.R. and M.J.D. designed and carried out the statistical and computational analyses, with assistance from K.S.L., G.B., B.N., J.M.K., T.G., S.R., F.K., T.F., P.S. and C.K.Z. S.R. assisted with quality control, principal-component analysis and analysis of Immunochip data. Syzygy was developed by M.A.R. and M.J.D. M.J.D. supervised all aspects of the study. The manuscript was written by M.A.R., J.D.R., R.J.X. and M.J.D.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Loftus, E.V. Jr. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology* **126**, 1504–1517 (2004).
2. Bengtson, M.B. *et al.* Familial aggregation in Crohn's disease and ulcerative colitis in a Norwegian population-based cohort followed for ten years. *J. Crohns Colitis* **3**, 92–99 (2009).
3. Brant, S.R. Update on the heritability of inflammatory bowel disease: the importance of twin studies. *Inflamm. Bowel Dis.* **17**, 1–5 (2011).
4. Rioux, J.D. & Abbas, A.K. Paths to understanding the genetic basis of autoimmune disease. *Nature* **435**, 584–589 (2005).
5. Nadeau, J.H. Single nucleotide polymorphisms: tackling complexity. *Nature* **420**, 517–518 (2002).
6. Plenge, R. & Rioux, J.D. Identifying susceptibility genes for immunological disorders: patterns, power, and proof. *Immunol. Rev.* **210**, 40–51 (2006).
7. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
8. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
9. McGovern, D.P. *et al.* Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.* **42**, 332–337 (2010).
10. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
11. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nat. Genet.* **39**, 813–815 (2007).
12. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
13. Hugot, J.P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
14. Duerr, R.H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
15. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
16. Nejentsev, S. *et al.* Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
17. Calvo, S.E. *et al.* High-throughput, pooled sequencing identifies mutations in *NUBPL* and *FOXRED1* in human complex I deficiency. *Nat. Genet.* **42**, 851–858 (2010).
18. Zaghloul, N.A. *et al.* Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. *Proc. Natl. Acad. Sci. USA* **107**, 10602–10607 (2010).
19. Emison, E.S. *et al.* Differential contributions of rare and common coding and noncoding *Ret* mutations to multifactorial Hirschsprung disease liability. *Am. J. Hum. Genet.* **87**, 60–74 (2010).
20. Cohen, J.C., Boerwinkle, E., Mosley, T.H.J. & Hobbs, H.H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
21. Cohen, J.C. *et al.* Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl. Acad. Sci. USA* **103**, 1810–1815 (2006).
22. 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).
23. Marth, G.T. *et al.* The functional spectrum of low-frequency coding variation. *Genome Biol.* **12**, R84 (2011).
24. Chamaillard, M. *et al.* Gene-environment interaction modulated by allelic heterogeneity in inflammatory diseases. *Proc. Natl. Acad. Sci. USA* **100**, 3455–3460 (2003).
25. Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
26. Miceli-Richard, C. *et al.* *CARD15* mutations in Blau syndrome. *Nat. Genet.* **29**, 19–20 (2001).
27. King, K. *et al.* Mutation, selection, and evolution of the Crohn disease susceptibility gene CARD15. *Hum. Mutat.* **27**, 44–54 (2006).

28. Barnich, N., Aguirre, J.E., Reinecker, H.C., Xavier, R.J. & Podolsky, D.K. Membrane recruitment of NOD2 in intenstinal epithelial cells is essential for nuclear factor-κB activation in muramyl dipeptide recognition. *J. Cell Biol.* **170**, 21–26 (2005).
29. Tanabe, T. *et al.* Regulatory regions and critical residues of NOD2 involved in muramyl dipeptide recognition. *EMBO J.* **23**, 1587–1597 (2004).
30. Roth, M.P. *et al.* Geographic origins of Jewish patients with inflammatory bowel disease. *Gastroenterology* **97**, 900–904 (1989).
31. Tukel, T. *et al.* Crohn disease: frequency and nature of *CARD15* mutations in Ashkenazi and Sephardi/Oriental Jewish families. *Am. J. Hum. Genet.* **74**, 623–636 (2004).
32. Momozawa, Y. *et al.* Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. *Nat. Genet.* **43**, 43–47 (2011).
33. Di Meglio, P. *et al.* The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced T_H17 effector response in humans. *PLoS ONE* **6**, e17160 (2011).
34. Ghoreschi, K. *et al.* Generation of pathogenic T(H)17 cells in the absence of TGF-β signalling. *Nature* **467**, 967–971 (2010).
35. Festen, E.A. *et al.* A meta-analysis of genome-wide association scans identifies *IL18RAP, PTPN2, TAGAP*, and *PUS10* as shared risk loci for Crohn's disease and celiac disease. *PLoS Genet.* **7**, e1001283 (2011).
36. Behrends, C., Sowa, M.E., Gygi, S.P. & Harper, J.W. Network organize of the human autophagy system. *Nature* **466**, 68–76 (2010).
37. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
38. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
39. Jin, Y. *et al.* Variant of *TYR* and autoimmunity susceptibility loci in generalized vitiligo. *N. Engl. J. Med.* **362**, 1686–1697 (2010).
40. Pearson, K. Mathematical contributions to the theory of evolution VIII: On the inheritance of characters not capable of exact quantitative measurement. *Phil. Trans. R. Soc. Lond. A* **195**, 79–150 (1900).
41. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
42. Ahmad, T., Satsangi, J., McGovern, D., Bunce, M. & Jewell, D.P. The genetics of inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **15**, 731–748 (2001).

ONLINE METHODS

DNA preparation and pooling. We selected Crohn's disease cases and controls from the NIDDK IBDGC, with priority given to samples with adequate amounts of DNA and those with GWAS data available. Samples from the NIDDK IBDGC underwent rigorous clinical phenotyping and control matching for genetic studies. DNA purification methods were also carried out on these samples. The case-control samples selected have already been stringently matched in previous GWAS studies. The baseline concentration of genomic DNA was quantified by Quant-iT PicoGreen dsDNA reagent and detected on the Thermo Scientific Varioskan Flash. All DNAs were normalized to 20 ng/μl and quantification was repeated to assess accuracy of the normalization step. The quantification and normalization was repeated again to ensure that all samples fell within the desired concentration range. The normalization steps were done with robotic automation using the Packard Multiprobe II HT EX. After each individual sample was normalized to 10 ng/μl, groups of 50 individuals were pooled together using a Multiprobe or Packard Robotic to total 14 pools (700 people).

Target selection and design. Candidate exonic targets from top published, confirmed GWAS loci and a sample of other highly significant regions of interest were uploaded for design using human genome build 17 and an in-house database, which houses PRIMER3 software. Amplicons encompassing each target region (coding exons only) were designed using Illumina parameters including a minimum amplicon length of 150 bp and maximum amplicon length of 600 bp with no buffer sequence added. Additionally, NotI tails were added to the primer pairs to provide a recognition site for downstream concatenation and shearing step. Amplicons were validated by running PCR product on agarose gels to assess clarity of single bands. Amplicons with two-thirds clear bands were considered validated. Pfu enzyme, used in Illumina sequencing protocol for PCR, was used in the characterization process. In total, 593 primer pairs passed and covered 95% of the 108-kb target. PCRs contained 20 ng of pooled genomic DNA, 1× HotStar buffer, 0.8 mM dNTPs, 2.5 mM $MgCl_2$, 0.2 units of HotStar Enzyme (Qiagen) and 0.25 μM forward and reverse primers in a 6- or 10-μl reaction volume. PCR cycling parameters were one cycle of 95 °C for 15 min; 35 cycles of 95 °C for 20 s, 60 °C for 30 s and 72 °C for 1 min; followed by one cycle of 72 °C for 3 min. Each PCR product was then treated to similar steps used for pooling DNA individuals. The quantification, normalization and pooling process ensured that equimolar PCR product went into library construction for equal representation of all targets. PCR yield was assessed by the same quantification system and the lowest product yield was then used for normalization across PCR plates. Secondary confirmation was ascertained by testing one column of PCR product per plate on 2% agarose E-gel versus a 1-kb DNA ladder to visualize PCR product size. The 593 PCR products were then combined using the Packard Multiprobe II HT EX, leading to an amplified target product per sample pool for sequencing.

Sequencing. The PCR products for each pooled sample were concatenated using NotI adaptors and sheared into fragments as described[43]. Libraries were constructed by a modified Illumina single-end library protocol, with 225–275 bp gel size selection and PCR enrichment using 14 cycles of PCR, and then were single-end sequenced with 76 cycles on an Illumina Genome Analyzer. Each sample pool was sequenced using a single lane of an Illumina GAII analyzer flow cell. Reads of 76 bp, 36 bp and 52 bp were aligned to the genome using MAQ algorithm[44] within the Picard analysis pipeline, and further processed using SAMtools software[45] and custom scripts.

Genotyping. We assayed 137 high-confidence SNVs in two phases of genotyping using Sequenom MassARRAY iPLEX GOLD chemistry50. The first phase comprised 72 SNVs and the second phase comprised 65 SNVs on 350 NIDDK Crohn's disease cases and 350 NIDDK controls for validation purposes. In each phase of genotyping, oligonucleotides were synthesized and quality control using mass spectrometry was carried out at Integrated DNA Technologies. All SNVs were genotyped in multiplexed pools of 25–36 assays, designed by AssayDesigner v.3.1 software, starting with 10 ng DNA per pool. About 7 nl of reaction was loaded onto each position of a 384-well SpectroCHIP preloaded with 7 nl matrix (3-hydroxypicolinic acid). SpectroCHIPs were analyzed in automated mode by a MassArray MALDI-TOF Compact System 2 with a solid-phase laser mass spectrometer (Bruker Daltonics). We obtained high-quality data (>95% genotype call rate, Hardy-Weinberg Equilibrium (HWE) $P > 0.001$) in all samples that had at least one SNV. Variants were called by real-time SpectroCaller algorithm, analyzed by SpectroTyper v.4.0 software and manually reviewed for rare variants. Additional Sequenom genotyping was carried out for nine SNVs in 2,887 Crohn's disease cases and 2,244 healthy controls from the German PopGen Biobank collection. German patients were recruited either at the Department of General Internal Medicine of the University of Kiel, the Charité University Hospital of Berlin, through local outpatient services, or nationwide with the support of the German Crohn and Colitis Foundation. German healthy control individuals were obtained from the PopGen Biobank[46].

Beadexpress data generated by the NIDDK IBDGC on 5,549 NIDDK samples aided in validation and follow-up of associated variants. Genotyping of IIBDGC samples was done with the Illumina Immunochip, in which SNVs discovered in this experiment were included. Independent Crohn's disease and ulcerative colitis cases, along with unaffected population controls, were genotyped at five genotyping centers (see **Supplementary Methods** for details on quality control steps).

Cells, antibodies and plasmids. HEK293T cells were obtained from American Type Culture Collection (ATCC) and maintained according to the instructions of ATCC. Antibody to β-actin was obtained from Santa Cruz. Antibody to NOD2 (clone NOD-15) was obtained from BioLegend. Human wild-type *NOD2* cDNA was cloned in pBK-CMV vector (Stratagene) for expression of untagged NOD2. Mutated constructs were made using QuikChange site-directed mutagenesis kit (Stratagene). Inserts were fully sequenced to confirm that only the desired mutations were present.

Immunostaining. HEK293T cells were seeded on polylysine-coated slides and transfected with *NOD2* constructs using lipofectamine 2000. The next day, cells were fixed with 4% paraformaldehyde (10 min) and permeabilized with 0.1% Triton X-100 in PBS (10 min). After washing with PBS, the sections were incubated 15 min in PBS containing 1% BSA. The sections were then incubated with antibody to NOD2 (1:200) for 1 h, washed using PBS, incubated with dylight 488 conjugated donkey anti-mouse Ig antibody (Jackson ImmunoResearch) for 1 h, washed using PBS and incubated with PBS containing 100 μg/ml of DABCO (Sigma) as antifading reagent before mounting in Glycergel medium (Dako). Fluorescence signals were captured using a laser confocal microscope (model Radiance 2000 Bio-Rad).

Luciferase reporter assays. HEK293T cells were co-transfected with 0.025 ng renilla luciferase plasmid, 2.5 ng Ig-pIV firefly luciferase reporter and 5 ng *NOD2* plasmids using lipofectamine 2000 (Invitrogen). After 24 h of transfection, cells were stimulated with MDP-LL or MDP-LD (10 μg/ml) for 6 h. Luciferase activities were measured using the Dual Luciferase reporter assay system (Promega) in a BD Moonlight 3010 luminometer (BD Biosciences) and normalized to the internal transfection control of renilla luciferase activity.

Variant discovery software. We used methods in Syzygy to analyze pooled sequencing data. The software enables investigators to carry out SNP calling on pooled data, estimate allele frequencies of discovered variants, apply single-marker association tests in a pooled setting, carry out group-wise testing of rare and low-frequency variants, carry out power evaluation and quality control summary, and annotate variants discovered in regions from primary sequencing data in Sequence Alignment/Map format. Thus, researchers can prioritize variants and regions for follow-up experiments and dissection of genetic architecture in target regions of interest.

Mega-analysis of rare variants. A goal of the project was to combine data from different groups and subpopulations in which samples were carefully matched. We propose the following approach, called MARV, to analyze rare variants.

Step 1. Let our random variable X = number of nonreference alleles observed across all collections genotyped.

Step 2. The affected or unaffected status is permuted among the individuals within each subgroup, and step 1 is repeated k times to sample $x_1^*, \ldots, x_k^*$ under the null hypothesis.

Step 3. The average ($\hat{\mu}$) and sample s.d. ($\hat{\sigma}$) of $x_1^*, \ldots, x_k^*$ are calculated, and the standardized score is $Z = \dfrac{X - \hat{\mu}}{\hat{\sigma}}$.

Under the null hypothesis, Z has an approximately standard normal distribution (see **Supplementary Fig. 7**). Thus, a P value for the association test can be obtained by comparing Z to the quantiles of the standard normal. Alternatively, a P value can be obtained by using a standard permutation test, where the P value is found by $(k_0 + 1) / (k + 1)$, and k_0 is the number of the k permutations that are at least as extreme as x.

Software availability. Syzygy, http://www.broadinstitute.org/software/syzygy/; MARV, http://www.broadinstitute.org/ftp/pub/mpg/syzygy/MARV.R.

43. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
44. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet.* **9**, 55–61 (2006).