Published in: Nature Genetics (2010), vol. 43, iss. 1, pp. 43-47.
Status: Postprint (Author's version)

Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease

Yukihide Momozawa[1], Myriam Mni[1], Kayo Nakamura[1], Wouter Coppieters[1], Sven Aimer[2], Leila Amininejad[3], Isabelle Cleynen[4], Jean-Frédéric Colombel[5], Peter de Rijk[6], Olivier Dewit[7], Yigael Finkel[8], Miquel A Gassull[9], Dirk Goossens[6], Debby Laukens[10], Marc Lémann[11], Cécile Libioulle[1], Colm O'Morain[12], Catherine Reenaers[13], Paul Rutgeerts[4], Curt Tysk[14], Diana Zelenika[15], Mark Lathrop[15], Jurgen Del-Favero[6], Jean-Pierre Hugot[16], Martine de Vos[10], Denis Franchimont[3], Severine Vermeire[4], Edouard Louis[13] & Michel Georges[1]

[1]Unit of Animal Genomics, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA-R) and Faculty of Veterinary Medicine, University of Liège (B34), Liège, Belgium. [2]Division of Gastroenterology and Hepatology, Institutionen for molekylär och klinisk medicin (IMK) Linköpings Universitet, Linköping, Sweden. [3]Department of Gastroenterology, Erasme Hospital, Université Libre de Bruxelles (ULB), Brussels, Belgium. [4]Department of Pathophysiology, Gastroenterology Section, Catholic University of Leuven, Leuven, Belgium. [5]Registre des MICI du Nord-Ouest de la France (EPIMAD), Hôpital Calmette, Lille, France. [6]Applied Molecular Genomics, Department of Molecular Genetics, Vlaams Instituut voor Biotechnologie (VIB), University of Antwerp, Antwerp, Belgium. [7]Department of Gastroenterology, Clinique Universitaire St Luc, Université Catholique de Louvain (UCL), Brussels, Belgium, [8]Department of Gastroenterology, Karolinska Children's Hospital, Stockholm, Sweden. [9]Gastroenterology Department, Hospital Universitari Germans Trias i Pujol, Badalona, Spain. [10]Department of Gastroenterology, University Hospital, Ghent University, Ghent, Belgium. [11]Department of Gastroenterology, Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Saint-Louis, Université Paris Diderot Paris-VII, Paris, France. [12]Adelaide and Meath Hospital, Dublin, Ireland. [13]Unit of Hepato-gastroenterology, GIGA-R and Faculty of Medicine, University of Liège (B34), Liège, Belgium. [14]Department of Gastroenterology, Orebro Medical Center Hospital, Orebro, Sweden. [15]Centre National de Génotypage, Evry, France. [16]INSERM U843, Hopital Robert Debré, Paris, France.

Abstract

Genome-wide association studies (GWAS) have identified dozens of risk loci for many complex disorders, including Crohn's disease[1,2]. However, common disease-associated SNPs explain at most ~20% of the genetic variance for Crohn's disease. Several factors may account for this unexplained heritability[3-5], including rare risk variants not adequately tagged thus far in GWAS[6-8]. That rare susceptibility variants indeed contribute to variation in multifactorial phenotypes has been demonstrated for colorectal cancer[9], plasma high-density lipoprotein cholesterol levels[10], blood pressure[11], type 1 diabetes[12], hypertriglyceridemia[13] and, in the case of Crohn's disease, for *NOD2* (refs. 14,15). Here we describe the use of high-throughput resequencing of DNA pools to search for rare coding variants influencing susceptibility to Crohn's disease in 63 GWAS-identified positional candidate genes. We identify low frequency coding variants conferring protection against inflammatory bowel disease in *IL23R*, but we conclude that rare coding variants in positional candidates do not make a large contribution to inherited predisposition to Crohn's disease.

An earlier meta-analysis of three GWAS identified 30 significant and 10 suggestive susceptibility loci for Crohn's disease[2]. The average confidence interval surrounding these loci was 233 kb (with a range of 20-1,140 kb), encompassing 4.1 genes (range 0-37 genes), for a total of 153 positional candidates (**Supplementary Table 1**).

We decided to sequence the open reading frame (ORF) and intron-exon boundaries of the 51 genes mapping to loci containing between one and five genes. For loci with more than six candidates, we retained 15 genes that mapped to significant networks identified when analyzing all candidates with Ingenuity Pathways (v8.5) (**Supplementary Table 2**). To these 66 genes, we added the *SLC22A4* candidate (ref. 16), as well as *PTGER4*, *ORMDL3* and *GSDMB*, on the basis of previously reported *cis* expression quantitative trait loci effects[2,17]. The list of the 70 selected genes is provided in **Supplementary Table 3**.

After extensive optimization (**Supplementary Note**), we selected a protocol involving (i) constitution of equimolar pools of genomic DNA from sets of 32 cases or controls, (ii) amplification, using Phusion Hot Start High-Fidelity DNA Polymerase (Finnzymes Oy), of the 70 targeted ORFs and intron-exon boundaries as a series of 1,045 amplicons averaging 222 bp in length (range 136-337 bp), (iii) equimolar pooling of ~300 amplicons, (iv) massive parallel pyrosequencing using the Roche FLX system[18] targeting an average sequence depth of 500 for both the Watson and Crick (W&C) strands and (v) detection of DNA sequence variants (DS Vs) using the

Amplicon Variant Analyzer (AVA) software (Roche) augmented with custom-made scripts (Online Methods).

We opted for a staged design in which all 70 candidate genes were first sequenced in 112 cases and 112 controls (stage I). This provided 98.5% and 73.3% nominal power ($P \leq 0.05$) to detect the 12% and 7% excess of rare *NOD2* (Gene ID: 64127) variants reported in references 14 and 15, respectively. The most promising genes were further evaluated in additional pools of cases and controls (stage II). To increase the impact of genetic effects other than *NOD2*, we selected 112 stage I cases that did not carry any of three known *NOD2* susceptibility variants (p.Arg702Trp, p.Gly908Arg and p.Ala1007*f*s). To avoid subtle stratification, the corresponding 112 controls underwent the same selection. All analyzed cases and controls were of European decent.

We could sequence 92.9% of the amplicons, corresponding to 63 out of the 70 genes (**Supplementary Table 3**) and a total of 108.3 kb, with coverage of at least 200 for both W&C strands in at least one case and one control pool. Simulations indicate that this coverage provided at least 83.4% power to detect singletons (that is, one variant chromosome in a DNA pool of 32 individuals), given the settings of the AVA software and the self-imposed curation filters (**Supplementary Fig. 1**). The average sequence depth (± standard deviation (s.d.)) of the retained amplicons was 1,471 ± 849 in cases and 1,420 ± 822 in controls (**Supplementary Table 3**). Analysis of the flowgrams yielded 372 DSVs (**Table 1** and **Supplementary Table 4**). Transitions accounted for 82.5% of the variants, transversions for 16.1%, dinucleotide substitutions for 0.3% and in dels for 1.1%. Synonymous variants accounted for 41.7% of the variants, missense variants for 55.9%, nonsense variants for 0.8%, in-frame in dels for 1.1% and 'boundary' (intronic, within 2 bp of an exon) (β) variants for 0.5%. DSVs with an estimated minor allele frequency (MAF) <0.05 amounted to 78.5%, whereas singletons represented 50.3% of all variants **(Fig. 1**). As expected and reflecting purifying selection on mildly deleterious variants, the frequency spectrum of non-synonymous variants was shifted toward lower frequencies when compared to synonymous variants. Non-synonymous variants represented 60% of the variants with MAF < 0.05 compared to 40% of variants with MAF > 0.05 **(Table 1** and **Fig. 1**). The high transition to transversion ratio (5.1) is thought to be due to (i) the analysis of ORFs, as transversions are more likely to be non-synonymous and are therefore selected against, (ii) idiosyncrasies of the analyzed set of genes, as their transition to transversion ratio tended to be higher than that of other ORFs in the HapMap data and (iii) the elimination of low frequency (<2.5% frequency) C>A = G>T variants **(Supplementary Table 5** and **Supplementary Note**).

Figure 1 *Frequency distribution of minor allele frequencies for different categories of variants. Synonymous variants are shown in blue. All non-synonymous variants are shown in red. Damaging non-synonymous variants (as predicted by SIFT[19]) are shown in orange.*

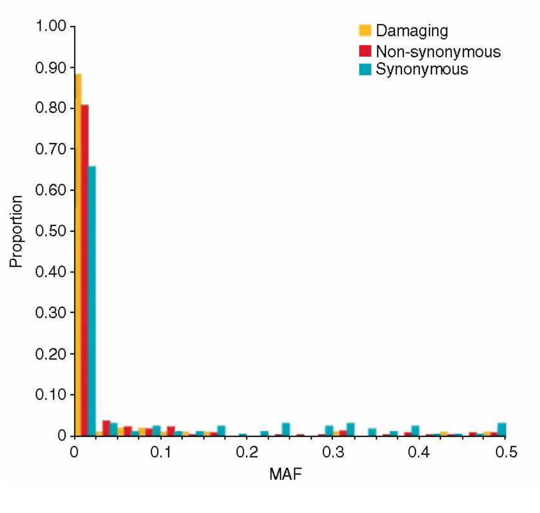

Table 1 *Type and effect of the 372 DNA sequence variants (DSVs) detected by high-throughput resequencing*

	DSV	Rare (MAF < 0.05)		Common (MAF ≥ 0.05)	
Substitution	A>C=T>G	10	3.4%	4	5.0%
	A>G=T>C	86	29.5%	23	28.8%
	A>T=T>A	10	3.4%	2	2.5%
	C>A=G>T	9	3.1%	2	2.5%
	C>G=G>C	20	6.8%	3	3.8%
	C>T=G>A	152	52.1%	46	57.5%
	AA>CG=TT>GC	1	0.3%	0	0.0%
	F-INDEL	4	1.4%	0	0.0%
Effect	Synonymous	107	36.6%	48	60.0%
	Non-synonymous	176	60.3%	32	40.0%
	Nonsense	3	1.0%	0	0.0%
	Deletion	4	1.4%	0	0.0%
	Boundary	2	0.7%	0	0.0%
	Total	292	78.5%	80	21.5%

A>C=T>G corresponds to an A to C or T to G transversion where A and T are the ancestral alleles as determined from the analysis of orthologues primate sequences. All other substitutions follow this nomenclature scheme. IF-INDEL corresponds to in-frame nsertion-deletion events. Boundary corresponds to splice site variants located within 2-bp from an exon boundary.

We evaluated our protocol in terms of sensitivity (that is, the fraction of true variants called), positive predictive value (PPV, the fraction of true variants among called variants) and accuracy in estimating allelic frequency, focusing first on common variants (MAF ≥ 0.05). Analysis of the HapMap data identified 62 *bona fide* SNPs with MAF ≥ 0.05 that were covered by the 879 retained amplicons. Five of these amplicons (8.1%) resided within 6 bp of a homopolymer track and were thus ignored (Online Methods). The remaining 57 amplicons were all detected, indicating excellent sensitivity. The 24 called SNPs with MAF ≥ 0.05 that were not genotyped in HapMap were inventoried in the dbSNP database (22 out of 24 SNPs) or confirmed by the 1000 Genomes Project data (see URLs) (the remaining two SNPs not inventoried in dbSNP), indicating excellent PPV. To evaluate the accuracy in estimating allelic frequencies, we took advantage of 31 common SNPs that had been genotyped on the same individuals as part of other projects. **Supplementary Figure 2** shows the correlation between allelic frequencies estimated from the genotyping data and the read counts. The regression coefficient was 0.975 and the correlation was 0.993.

To obtain similar estimates for rare variants, we manually (Sanger sequencing on the ABI 3730) sequenced 2,283 bp of the *NOD2* ORF in the same 112 cases and 112 controls. Sanger sequencing identified 38 variants with MAF < 0.05. Assuming faultless Sanger sequencing, the sensitivity and PPV of the massive parallel resequencing protocol were 82.4% and 97.9%, respectively. Frequency estimates from the read counts tended to underestimate the actual frequencies of the rare variants (regression coefficient = 0.822, correlation = 0.578) (**Supplementary Fig. 2**). We observed no difference in sequence depth between the amplicon by DNA pool combinations in which rare DSVs were detected and those in which no such DSVs were found (**Supplementary Fig. 3**).

Having evaluated the performance of our protocol, we searched for differences in the cumulative frequencies of rare variants (MAF < 0.05) between cases and controls. We estimated the statistical significance of the observed differences on a gene-by-gene basis using a permutation test (Online Methods). We computed the *P* values for synonymous variants, all non-synonymous plus β variants and non-synonymous variants predicted by SIFT[19] to be damaging. For each gene by DSV-type combination, we computed two *P* values corresponding, respectively, to an enrichment of rare variants in cases (risk variants) or an enrichment of rare variants in controls (protective variants). Thus, we hypothesized that disruptive variants would increase disease risk in some genes and decrease disease risk in others. After applying a Bonferroni correction, none of the 63 sequenced genes showed a significant ($P < 7.94 \times 10^{-4}$) enrichment of rare variants either in cases or in controls, whether synonymous, non-synonymous plus β or damaging (**Supplementary Table 6**). There was no evidence for a difference in the distribution of *P* values between synonymous and nonsynonymous plus β variants, whether considering variants independently or on a gene-by-gene basis (**Supplementary Note**). However, *NOD2* showed the expected enrichment of rare non-synonymous variants (excluding the well-known p.Arg702Trp, p.Gly908Arg and p.Ala1007/s variants) in cases (nominal $P = 5.94 \times 10^{-3}$; rank 3).

We therefore decided to pursue sequencing (stage II) of the top 20% of genes (that is, 12 genes) in 288-928 additional cases and 288-1,216 additional controls, depending on intermediate results. The procedure we used was identical to that in stage I, that is, high-throughput resequencing of pooled amplicons obtained from DNA pools of cases or controls (32 individuals per pool and up to 29 case and 38 control pools per gene). We appended amplicons with DNA pool-specific tags, which allowed simultaneous sequencing of multiple DNA pools. The average sequence depth in stage II was 988 ± 512 (range 411-13,506) in cases and 1,019 ± 415 (range 405-10,414) in controls (**Supplementary Fig. 1** and **Supplementary Table 3**). We detected two new common and 112 new rare variants (**Supplementary Table 7**). We observed no difference in sequence depth between the amplicon by DNA pool combinations in which rare DSVs were detected and those in which no such DSVs were found (**Supplementary Fig. 3**).

We tested for differences in the cumulative frequencies of rare variants in cases and controls using the same permutation test as above except that here we only tested the significance of enrichment with the same polarity as in stage I, meaning enrichment either for rare risk variants in cases *(FGFR1OP*, *GSDMB*, *IKZF3*, *IL1RL1*, *NOD2*, *SLC9A4* or *TNFSF8*) or rare protective variants in controls (*CCL8*, *CDKAL1*, *ENOX1*, *IL23R* or *SLC22A5*). In the stage II analysis (**Table 2**), one gene yielded a suggestive association (*FGFR1OP*, Gene ID: 11116; nominal $P = 0.040$, Bonferroni-corrected $P = 0.386$), and one gene yielded a significant association (*IL23R*, Gene ID: 149233; nominal $P = 2.67 \times 10^{-3}$, Bonferroni-corrected $P = 0.0314$).

Table 2 *Statistical significance of the difference in cumulative frequency of rare risk or protective variants between cases and controls*

	Stage I					Stage II						
	Cases		Controls			Cases		Controls				
Gene	Frequency	Number	Frequency	Number	*P*	Frequency	Number	Frequency	Number	*P*	Risk or protective	Comment
CCL8	0.0232	112	0.0405	112	0.036	0.0096	320	0.0102	320	0.500	Protective	
CDKAL1	0.0096	112	0.0507	112	0.062	0.0467	384	0.0675	384	0.208	Protective	
ENOX1	0.0116	112	0.0232	112	0.059	0.0048	320	0.0017	320	0.671	Protective	
FGFR1OP	0.0437	112	0.0000	112	4.78×10^{-3}	0.0459	320	0.0240	320	0.040	Risk	
GSDMB	0.0539	112	0.0068	112	1.74×10^{-3}	0.0206	704	0.0318	384	0.977	Risk	
IKZF3	0.0514	112	0.0205	112	0.043	0.0282	288	0.0253	288	0.354	Risk	
IL1RL1	0.0290	112	0.0083	112	0.021	0.0012	288	0.0034	288	0.888	Risk	
IL23R	0.0052	112	0.0370	112	0.071	0.0110	896	0.0210	1,216	2.67×10^{-3}	Protective	$^-$p.Arg381Gln
	0.0205	112	0.1809	112	6.00×10^{-5}	0.0408	896	0.0849	1,216	$<1.00 \times 10^{-6}$	Protective	$^+$ p.Arg381Gln
NOD2	0.1192	112	0.0580	112	5.94×10^{-3}	0.0587	928	0.0542	992	0.157	Risk	$^-$p.Gly908Arg
						0.1081	928	0.0802	992	8.91×10^{-3}	Risk	$^+$ p.Gly908Arg
SLC22A5	0.0087	112	0.0224	112	0.058	0.0064	320	0.0093	320	0.419	Protective	
SLC9A4	0.0502	112	0.0099	112	0.017	0.0433	702	0.0388	384	0.318	Risk	
TNFSF8	0.0307	112	0.0078	112	0.089	0.0000	288	0.0000	288	1.000	Risk	

For *IL23R* and *N0D2*, data are presented with and without consideration of the previously reported low frequency p.Arg381Gln (*IL23R*) and p.Gly908Arg (*N0D2*) variants, when appropriate.

Table 3 *Low frequency IL23R variants protecting against Crohn's disease (enriched in controls) detected by two rounds of high-throughput sequencing*

	Sequencing					Individual genotyping					
Variant	Frequency (CD)	*N* (CD)	Frequency (controls)	*N* (controls)	*P*	Frequency (CD)	*N* (CD)	Frequency (controls)	*N* (controls)	*P*	OR (95% Cl)
1 p.Arg86Gln	0.0007	912	0.0029	1,200	0.077	0.0005	1,005	0.0027	1,273	0.069	0.180 (0.000-1.125)
2 p.Gly149Arg	0.0007	848	0.0042	944	0.029	0.0010	1,011	0.0043	1,281	0.031	0.230 (0.000-0.879)
3 p.Val362Ile	0.0076	880	0.0141	1,264	0.026	0.0119	922	0.0216	1,276	0.010	0.548 (0.000-0.851)
1+2+3[a]	0.0090		0.0211		5.78×10^{-4}	0.0255	903	0.0552	1,251	4.22×10^{-4}	0.448 (0.000-0.683)
4 p.Arg381Gln	0.0291	880	0.0669	1,264	8.49×10^{-9}	0.0240	959	0.0693	1,269	6.53×10^{-13}	0.330 (0.000-0.439)
1+2+3+4[a]	0.0381		0.0881		$<1.00 \times 10^{-6}$	0.0749	854	0.1840	1,234	2.22×10^{-13}	0.360 (0.000-0.463)

[a] Frequency for individual genotyping is the proportion of individuals carrying at least one of the low frequency variants. CD, Crohn's disease

Closer examination of *FGFR1OP* showed that three non-synonymous variants (p.Thr184Ile, p.Lys251Asn and p.Ser281Pro), located within 2,436 bp of each other, segregated identically across DNA pools, strongly suggesting that they were in complete linkage disequilibrium. When considering these variants as a single event, nominal *P* values dropped to 0.124 in stage I and to 0.081 in stage II. Hence, *FGFR1OP* was not considered for further analysis.

The *IL23R* signal was entirely due to three variants (p.Arg86Gln, p.Gly149Arg and p.Val362Ile), with a cumulative frequency of 0.0052 in cases compared to 0.0370 in controls in stage I and 0.0088 in cases compared to 0.0193 in controls in stage II (**Table 3** and **Supplementary Table 8**). The observation of an enrichment in controls of their presumably protective *IL23R* variants was consistent with the protective effect of p.Arg381Gln that lead to the discovery of *IL23R* in a previous GWAS[20]. p.Arg381Gln was enriched in our controls as expected ($P = 8.49 \times 10^{-9}$).

The fact that these *IL23R* variants are low frequency rather than very rare DSVs[4] allowed targeted genotyping in independent samples. We developed TaqMan assays for p.Arg86Gln, p.Gly149Arg and p.Val362Ile, in addition to p.Arg381Gln. We first genotyped the sequenced individuals, which confirmed enrichment of the p.Arg86Gln, p.Gly149Arg and p.Val362Ile variants in the controls **(Table 3** and **Supplementary Fig. 2**). We then genotyped an additional 1,565 individuals with Crohn's disease, 2,000 controls and 3,101 familial samples (740 affected and 2,361 non-affected individuals) (stage III). All analyzed individuals were of European decent and most of them were previously used in GWAS replications. p.Gly149Arg ($P = 0.022$) and p.Val362Ile ($P = 1.51 \times 10^{-3}$) were significantly enriched in controls in the replication cohort, and we observed a similar trend, albeit not formally significant, for the rarer p. Arg86Gln variant ($P = 0.057$) (**Table 4**).

p.Arg381Gln confers protection against ulcerative colitis[20] as well. We thus genotyped a cohort of 1,251 ulcerative colitis cases of European decent for the same four *IL23R* variants. Both p.Arg381Gln ($P = 9.03 \times 10^{-9}$) and p.Val362Ile ($P = 8.31 \times 10^{-3}$) were significantly depleted in ulcerative colitis cases, whereas we observed the expected trend for p.Glyl49Arg ($P = 0.087$) but not for p.Arg86Gln ($P = 0.613$) (**Table 5**).

Table 4 *Confirmation by individual genotyping (case-control plus transmission disequilibrium test) of an enrichment in controls of low frequency IL23R variants protecting against Crohn's disease*

	Stage III - individual genotyping									
	Case-control						TDT			
Variant	Frequency (CD)	*N* (CD)	Frequency (controls)	*N* (controls)	*P*	OR (95% Cl)	T/U	*P*	Combined	Stage I+II+III
1 p.Arg86Gln	0.0013	1,873	0.0033	1,951	0.057	0.400 (0.000-1.031)	NA	NA	0.057	0.010
2 p.Gly149Arg	0.0013	1,942	0.0038	1,959	0.022	0.335 (0.000-0.843)	NA	NA	0.022	1.46×10^{-3}
3 p.Val362Ile	0.0116	1,331	0.0204	1,817	4.53×10^{-3}	0.567 (0.000-0.821)	7/15	0.067	1.51×10^{-3}	2.89×10^{-4}
1+2+3[a]	0.0314	1,114	0.0536	1,772	2.89×10^{-3}	0.573 (0.000-0.809)	7/15	0.067	1.14×10^{-3}	4.00×10^{-6}
4 p.Arg381Gln	0.0300	1,266	0.0786	1,946	3.37×10^{-17}	0.363 (0.000-0.452)	12/42	3.10×10^{-5}	$<1.00 \times 10^{-6}$	$<1.00 \times 10^{-6}$
1+2+3+4[a]	0.0880	1,114	0.1986	1,772	1.28×10^{-16}	0.389 (0.000-0.478)	19/57	8.00×10^{-6}	$<1.00 \times 10^{-6}$	$<1.00 \times 10^{-6}$

[a]Frequency is the proportion of individuals carrying at least one of the low frequency variants. TDT, transmission disequilibrium test; T, transmitted minor allele count; U, untransmitted minor allele count, NA, not analyzed.

Table 5 *Rare IL23R variants protecting against ulcerative colitis (enriched in controls) analyzed by individual genotyping (case-control)*

Variant	Frequency (UC)	*N* (UC)	Frequency (controls)	*N* (controls)	P	OR (95% Cl)
1 p.Arg86Gln	0.0034	1,176	0.0033	1,951	0.613	1.021 (0.000-2.321)
2 p.Gly149Arg	0.0016	1,232	0.0038	1,959	0.087	0.423 (0.000-1.146)
3 p.Val362Ile	0.0119	1,172	0.0204	1,817	8.31×10^{-3}	0.582 (0.000-0.854)
1+2+3[a]	0.0357	1,065	0.0536	1,772	0.017	0.653 (0.000-0.653)
4 p.Arg381Gln	0.0430	1,185	0.0786	1,946	9.03×10^{-9}	0.527 (0.000-0.642)
1+2+3+4[a]	0.1155	1,065	0.1986	1,772	3.06×10^{-9}	0.527 (0.000-0.637)

[a]Frequency is the proportion of individuals carrying at least one of the low frequency variants. UC, ulcerative colitis

We herein describe the systematic search for rare coding variants influencing inherited predisposition to Crohn's disease in 63 positional candidates identified by GWAS. We report three new low frequency *IL23R* variants protecting against Crohn's disease: p.Arg86Gln, p.Gly149Arg and p.Val362Ile. The three same variants were found to be protective in an independent study, strengthening our claims (M. Rivas & M. Daly, personal communication). We present preliminary evidence that p.Gly149Arg and p. Val362Ile act protectively against ulcerative colitis as well, as would be expected from the equivalent effect of p.Arg381Gln.

As was the case for the previously described p.Arg381Gln variant, the newly described p.Arg86Gln, p.Gly149Arg and p.Val362Ile variants are assumed to be hypomorphs that dampen IL23R signaling. p.Gly149Arg and p.Arg381Gln affect extremely conserved residues in the extracellular and intracellular domain of the receptor, respectively, and are predicted by SIFT[19] to be damaging. p.Arg86Gln and p.Val362Ile, on the contrary, affect poorly conserved residues and are predicted by SIFT[19] (using sequence information only) to be 'tolerated' and by PolyPhen[21] (using both sequence and structural information) to be 'benign'. Moreover, the reference *IL23R* sequences of some mammals carry the glutamine and isoleucine residues associated with inflammatory bowel disease in humans. Although we cannot exclude the possibility that p.Arg86Gln and p.Val362Ile are enriched in controls because of their association with causative variants lying outside the coding region, we consider it more parsimonious that they affect IL23R signaling directly. Of note, relative protection conferred by the 'damaging' p.Gly149Arg and p.Arg381Gln variants (2.98 and 2.75) tended to be higher than that conferred by the 'tolerated' p.Arg86Gln and p.Val362Ile variants (2.50 and 1.76), and we observed the same tendency for ulcerative colitis.

Relative protection against Crohn's disease conferred by the newly detected low frequency variants was -2.4, on average. Although apossible overestimation (due to winner's curse), this value appears considerably larger than the -1.2 relative risk conferred by most common variants detected in GWAS and supports an increase in effect size with decreasing frequency[6]. However, the newly detected variants jointly explain only ~0.44% of the variance of the underlying liability, compared with ~1.23% explained by the more common p.Arg381Gln and rs7517847 variants[19] (**Supplementary Note**). Haplotype analysis indicated that the protection conferred by p.Arg86Gln, p.Gly149Arg and p.Val362Ile is largely independent of the more common p.Arg381Gln and rs7517847 variants (**Supplementary Table 9**). Thus, we provide no evidence for 'synthetic association'[22] at the *IL23R* locus (**Supplementary Note**).

Although not significant when accounting for multiple testing, we obtained evidence suggesting an enrichment of rare non-synonymous *NOD2* risk variants in cases in stage I, supporting the results presented in previous reports[14,15]. We did not confirm this enrichment in stage II despite the sequencing of 928 cases and 992 controls. This discrepancy may be related to the selection of stage I samples not carrying the previously described p.Arg702Trp, p.Gly908Arg or p.Ala1007*f*s *NOD2* susceptibility variants, which were consequently enriched in stage II samples. Considering the stage I and II samples jointly, however, indicates that the excess *NOD2* mutation load in Crohn's disease cases is likely to be lower than previously assumed[14,15] and more in line with recent estimates from a similarly conducted North American scan for rare variants influencing Crohn's disease (M. Rivas & M. Daly, personal communication).

Our findings are highly reminiscent of those of a previous study[12], in which researchers resequenced the ORF and regulatory regions of ten candidate genes for type 1 diabetes in 480 cases and 480 controls and reported four low frequency protective variants in *IFIH1*. These modest success rates contrast with the findings of another previous study[13], in which researchers reported an enrichment of rare variants associated with hypertriglyceridemia (defined as fasting plasma triglyceride concentrations above the ninety-fifth percentile) in all four resequenced (438 cases and 327 controls) candidate genes from GWAS (*APOA5*, *GCKR*, *LPL* and *APOB*). Simulation studies indicate that this discrepancy is more likely to result from a difference in the genomic architecture of the studied traits rather than from methodological idiosyncrasies (high-throughput sequencing of DNA pools in two stages versus Sanger sequencing of individual samples in one stage) (**Supplementary Table 10** and **Supplementary Note**).

This study confirms the enrichment of low frequency variants (either in cases or controls) in at least some genes underlying inherited predisposition to complex diseases. Our results support an increase in effect size with decreasing variant frequency. However, because of their frequency, rare variants explain less of the heritability than their common counterparts. Achieving adequate power to reliably detect low frequency variants will require resequencing of cohorts larger than those used in this study. This will become increasingly feasible as sequencing technology continues to improve. The demonstration of an enrichment of rare or low frequency variants in candidate genes could then become an effective way to demonstrate the causality of candidate genes from GWAS.

URLs. 1000 Genomes Project, http://www.1000genomes.org/.

METHODS

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

Accessions codes. Accession codes are deposited in NCBI's RefSeq under the following accession codes: Human NOD2 protein, NP_ 071445.1; human IL23R protein, NP_653302.2; and human FGFR1OP protein, NP_919410.1.

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

ACKNOWLEDGMENTS

This work was funded by the Walloon Region (IPSEQ, Crohn & CIBLES projects), by the Politique Scientifique Fédérale (IAP GENFUNC), by the Fonds National de la Recherche Scientifique (FNRS) and by the Communauté Française de Belgique (ARC BIOMOD). E.L. is a senior research associate from the FNRS. K.N. benefitted from a post-doctoral fellowship from the University of Liège. We are grateful to A. Kvasz, F. Farnir and D. Baurain for their help with bioinformatic analyses. We thank the personnel of the Flanders Institute for Biotechnology (VIB)

Genetic Service Facility (http://www.vibgeneticservicefacility.be) and the GIGA-R genotranscriptomic platform for their contribution to sequencing. We are grateful to N. Wray and P. Visscher for fruitful discussions.

AUTHOR CONTRIBUTIONS

Y.M., M.M., K.N., L.A., D.G. and D.Z. performed experiments. Y.M., W.C., P.d.R. and M.G. analyzed data. M. Lathrop and J.D.-F. supervised experiments. S.A., L.A., J.-F.C., O.D., Y.F., M.A.G., M. Lémann, C.O., C.R., P.R., C.T., J.-P.H., M.d.V., D.F., S.V. and E.L. examined cases and collected samples. I.C., D.L. and C.L. prepared and organized samples. Y.M. and M.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* 322, 881-888 (2008).

2. 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).

3. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565-569 (2010).

4. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* 461, 747-753 (2009).

5. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* 462, 868-874 (2009).

6. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124-137 (2001).

7. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 40, 695-701 (2008).

8. Kruglyak, L. The road to genome-wide association studies. *Nat. Rev. Genet.* 9, 314-318 (2008).

9. Fearnhead, N.S. *et al.* Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl. Acad. Sci. USA* 101, 15992-15997 (2004).

10. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305, 869-872 (2004).

11. Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.* 40, 592-599 (2008).

12. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, l.A. Rare variants of *IFIH1,* a gene implicated in antiviral responses, protect

against type 1 diabetes. *Science* 324, 387-389 (2009).

13. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* 42, 684-687 (2010).

14. Hugot, l.P. *et al.* Association of *N0D2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411, 599-603 (2001).

15. Lesage, S. *et al. CARD15/NOD2* mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* 70, 845-857 (2002).

16. Peltekova, V.D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat. Genet.* 36, 471-475 (2004).

17. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of *PTGER4. PLoS Genet.* 3, e58 (2007).

18. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376-380 (2005).

19. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073-1081 (2009).

20. Duerr, R.H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* 314, 1461-1463 (2006).

21. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894-3900 (2002).

22. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 8, e1000294 (2010).

23. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* 132, 365-386 (2000).

ONLINE METHODS

High-throughput pyrosequencing on Roche FLX. Genomic DNA concentrations were determined by Quant-iT PicoGreen dsDNA Reagent and Kits (Invitrogen) for the constitution of equimolar pools of 32 cases or controls (except one pool of 48 in stage I). Primer pairs for PCR were selected using Primer 3 (ref. 23), avoiding known SNP positions. Amplicon-specific PCR reactions were set up in 30 µl volumes containg 6 µl of 5× Phusion HF buffer, 200 µM of each dNTP, 0.5 µM of each primer and 0.6 U of Phusion High-Fidelity DNA Polymerase (Finnzymes Oy). Cycling conditions were 98 °C for 2 min, 32 cycles at 98 °C for 10 s, 60 °C for 30 s and 72 °C for 15 s, followed by 72 °C for 10 min on a GeneAmp PCR System 2700 thermal cycler (Applied Biosystems). PCR products were purified using MultiScreen PCR$_{\mu 96}$ Filter Plates (Millipore) and quantified with the Quant-iT PicoGreen dsDNA reagent and kit. Up to 300 amplicons were combined in equimolar ratios. Pooled amplicons were concentrated using the Montage PCR Filter Units (Millipore) and purified using the AMPure kit (Agencourt Biosciences). The final concentration and length distribution were measured using the Experion DNA 1K Analysis kit (Bio-Rad) on the Experion Automated Electrophoresis Station (Bio-Rad). High-throughput pyrosequencing was carried out using both primer A and B on a Roche 454 Genome Sequencer FLX instrument following the recommendations of the manufacturer[18].

DSV detection. Image and data were processed with the Genome Sequencer FLX System Software Package (Roche). DSVs were extracted from sff files using the AVA software. AVA reports DSVs if they are observed at least four times and represent ≥ 0.5% of the reads. From the AVA-generated list, we eliminated DSVs based on the following criteria: (i) not observed on both W8;C strands, (ii) having flanking DSVs within 2 bp on both sides, (iii) in or within 6 bp from a homopolymer (≥5×) track and (iv) corresponding to C> A or G>T substitutions with frequency <0.025 (**Supplementary Note**).

Testing for a differential load of rare variants in cases and controls from resequencing data. The excess load of rare variants in cases (risk variants) or controls (protective variants) was tested on a gene-by-gene basis, and within each gene by DSV variant type (synonymous, non-synonymous plus β, or damaging). Rare variants were defined as DSVs with MAF < 0.05. The results were essentially unaffected by the threshold frequency used to define rare variants (MAF < 0.02-0.05) (data not shown). DSV read counts (number of reads with the DSV per total number of reads for that amplicon) were converted to the closest chromosome counts (>0) (number of chromosomes with the DSV per total number of chromosomes in the pool = 64) and these were summed over DNA pools separately for cases and controls. The *P* value of the difference in DSV chromosome counts between cases and controls was then computed using two one-tailed Fisher's exact tests, one testing an excess in cases (risk) and the other in controls (protective). For a given gene, we then multiplied hypothesis-specific (risk and protective) *P* values across rare variants to generate two gene-specific summary *P* values. The statistical significance of these summary statistics was estimated by permutation testing. For each rare DSV, case versus control status of the mutant chromosomes were assigned randomly, but accounting for the possibility that the number of successfully sequenced chromosomes (that is, DNA pools) might differ between cases and controls. The same two gene-specific summary *P* values were generated for 1,000,000 permutations, and the significance of the *P* values obtained with the real data was estimated as the proportion of permutations with a lower, hypothesis-specific, summary *P* value.

Case-control and familial association test based on individual genotypes.SNPs were tested on individual DNA using custom TaqMan assays (Applied Biosystems). The statistical significance of the difference in DSV frequency between cases and controls was estimated using a one-sided Fisher's exact test. The familial cohort was used to evaluate the significance of the distorted segregation (transmission disequilibrium test (TDT)) of the DSVs p.Arg381Gln and p.Val362Ile from heterozygous parents to affected offspring using a custommade script. As no heterozygous parents were available in the familial cohort for p.Arg86Gln and p.Gly149Arg, one affected individual per family was added to the case cohort in the case-control analysis for the analysis of these variants. We combined test statistics across resequencing, case-control and TDT experiments using a permutation test akin to the one described above.