Human Molecular Genetics, 2010, Vol. 19, No. 17 **3482–3488**
doi:10.1093/hmg/ddq264
Advance Access published on July 3, 2010

Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease

Karin Fransen[1,2,†], Marijn C. Visschedijk[2,3,†], Suzanne van Sommeren[1,2], Jinyuan Y. Fu[1], Lude Franke[1], Eleonora A.M. Festen[1,2], Pieter C.F. Stokkers[4], Adriaan A. van Bodegraven[5], J. Bart A. Crusius[6], Daniel W. Hommes[7], Pieter Zanen[8], Dirk J. de Jong[9], Cisca Wijmenga[1], Cleo C. van Diemen[1] and Rinse K. Weersma[2,*]

[1]Department of Genetics and [2]Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands, [3]Department of Gastroenterology, Isala Klinieken, Zwolle, Overijssel, The Netherlands, [4]Department of Gastroenterology and Hepatology, Academic Medical Centre, Amsterdam, The Netherlands, [5]Department of Gastroenterology and Hepatology and [6]Laboratory of Immunogenetics, Department of Pathology, VU University Medical Centre, Amsterdam, The Netherlands, [7]Department of Gastroenterology and Hepatology, Leiden University Medical Centre, Leiden, The Netherlands, [8]Department of Pulmonology, University Medical Center Utrecht, Utrecht, The Netherlands and [9]Department of Gastroenterology and Hepatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Received March 22, 2010; Revised and Accepted June 23, 2010

Genome-wide association studies (GWAS) for Crohn's disease (CD) have identified loci explaining ~20% of the total genetic risk of CD. Part of the other genetic risk loci is probably partly hidden among signals discarded by the multiple testing correction needed in the analysis of GWAS data. Strategies for finding these hidden loci require large replication cohorts and are costly to perform. We adopted a strategy of selecting SNPs for follow-up that showed a correlation to gene expression [*cis*-expression quantitative trait loci (eQTLs)] since these have been shown more likely to be trait-associated. First we show that there is an overrepresentation of *cis*-eQTLs in the known CD-associated loci. Then SNPs were selected for follow-up by screening the top 500 SNP hits from a CD GWAS data set. We identified 10 *cis*-eQTL SNPs. These 10 SNPs were tested for association with CD in two independent cohorts of Dutch CD patients (1539) and healthy controls (2648). In a combined analysis, we identified two *cis*-eQTL SNPs that were associated with CD rs2298428 in *UBE2L3* ($P = 5.22 \times 10^{-5}$) and rs2927488 in *BCL3* ($P = 2.94 \times 10^{-4}$). After adding additional publicly available data from a previously reported meta-analysis, the association with rs2298428 almost reached genome-wide significance ($P = 2.40 \times 10^{-7}$) and the association with rs2927488 was corroborated ($P = 6.46 \times 10^{-4}$). We have identified *UBE2L3* and *BCL3* as likely novel risk genes for CD. *UBE2L3* is also associated with other immune-mediated diseases. These results show that eQTL-based pre-selection for follow-up is a useful approach for identifying risk loci from a moderately sized GWAS.

INTRODUCTION

Crohn's disease (CD) is a common, chronic, gastrointestinal inflammatory disorder with a prevalence of 100–200 per 100 000 in developed countries (1). The aetiology of CD is complex and is believed to originate in an aberrant immune response to the commensal intestinal bacterial flora in a genetically susceptible host (2).

*To whom correspondence should be addressed at: Department of Gastroenterology and Hepatology, University Medical Centre Groningen and University of Groningen, PO Box 30001, 9700 RB Groningen, The Netherlands. Tel: +31 503610426; Fax: +31 503619306; Email: r.k.weersma@mdl.umcg.nl
†These authors contributed equally.

Genome-wide association studies (GWAS) have already identified over 30 loci that convey risk for CD (3–8), representing ~20% of the total genetic risk for this disease (8). The remaining 80% of genetic risk is probably partly made up by highly prevalent loci with very modest effect sizes and by rare loci with strong effect sizes. These remaining loci are hard to identify with a GWAS, in part because of the extensive multiple testing correction needed in GWAS analyses. This multiple testing correction is necessary to exclude false-positive loci, but simultaneously it discards many true-positive risk loci. Strategies for extricating these hidden true-positive loci include: increasing the GWAS sample size, performing a meta-analysis of GWAS data sets and replicating hundreds to thousands of GWAS signals in a larger cohort. Unfortunately, all of these methods still need substantial multiple testing correction and most are expensive to perform (9).

To cut down on the size of the follow-up study for a GWAS, and thus on the costs and need for multiple testing correction, we considered selecting SNPs for follow-up on the basis of a functional effect. In this study, we focus on the effect of SNPs on human gene expression levels which have been shown to have a strong heritable component (10). By treating gene expression as a quantitative trait, it is possible to correlate gene transcription levels with SNPs (expression quantitative trait loci, eQTLs) (10). SNPs can be correlated with the expression of genes located very near the SNP itself (*cis*-eQTL) or with the expression of genes located further away, even on other chromosomes (*trans*-eQTL). In this study, the maximal distance of a *cis*-eQTL SNP to a gene is 250 kb. Since the *trans*-eQTL effects are difficult to detect due to severe multiple testing issues, we chose to study *cis*-eQTL effects.

We hypothesized that SNPs affecting gene expression are more likely to be associated with CD than SNPs without such an effect, which provides a basis for selecting SNPs for replication. *Cis*-eQTLs have already been associated with several diseases, such as celiac disease and asthma (11,12). Our hypothesis is further supported by results from a recent GWAS in celiac disease in which a *cis*-eQTL effect was seen in 20 out of 38 risk loci identified for celiac disease. Permutations showed that 50% of SNPs being *cis*-eQTLs were very unlikely to occur by chance and were not due to a bias of the genotyping platform used, nor to differences in minor allele frequency (MAF) (13). In a recent paper, Nicolae *et al*. (14) found that SNPs associated with complex traits are more likely to be eQTLs and that, by using this information, the discovery of complex disease-associated genes can be enhanced.

For this study, we first validated our hypothesis that *cis*-eQTL SNPs are overrepresented among the currently known CD-associated SNPs by comparing the amount of established CD-associated SNPs that are *cis*-eQTLs with the number of *cis*-eQTL SNPs expected by chance (Table 1) (8). Next, a set of SNPs was selected for follow-up. We did this by comparing a list of CD risk SNPs with an *cis*-eQTL SNP database and aimed to identify novel CD-associated loci by selecting *cis*-eQTL SNPs from the top 500 hits from a publicly available CD GWAS (4). This resulted in 13 putative CD-associated eQTL SNPs, 10 of them were selected and studied in two independent cohorts of Dutch CD patients and controls (Fig. 1).

Table 1. *Cis*-eQTL effect of established and SNPs selected for follow-up

SNP	Chromosome	Risk allele	Expression effect risk allele	Effected gene	eQTL *P*-value
Five out of 30 established CD-associated SNPs[a]					
rs2301436	6q27	T	−	*RNASET2*	6.52×10^{-5}
rs2872507	17q12	A	−	*GSDML*	5.20×10^{-9}
rs3197999	3p21	A	+	*UBE1L*	9.79×10^{-4}
rs2872507	17q12	A	−	*ORMDL3*	6.94×10^{-11}
rs2188962	5q31	T	−	*SLC22A5*	5.18×10^{-9}
13 SNPs selected for follow-up[b]					
rs6512121	19	G	+	*ZNF266*	5.61×10^{-19}
rs243323	16	G	+	*C16orf75*	2.07×10^{-5}
rs2298428	22	T	−	*UBE2L3*	4.03×10^{-9}
rs2066843	16	T	+	*CARD15*	7.94×10^{-4}
rs2927488	19	A	+	*BCL3*	1.11×10^{-4}
rs1156287	17	G	+	*COX11*	5.35×10^{-6}
rs9303363	17	A	+	*COX11*	8.65×10^{-6}
rs725660	19	C	+	*SYMPK*	1.24×10^{-4}
rs7142206	14	A	+	*ENTPD5*	2.00×10^{-5}
rs1005564	14	T	+	*ENTPD5*	1.28×10^{-5}
rs3118663	9	G	−	*SURF1*	5.37×10^{-6}
rs10278590	7	G	+	*RARRES2*	2.10×10^{-4}
rs359457	5	T	+	*CPEB4*	6.12×10^{-8}

[a]Five out of 30 established CDc-associated SNPs are Cis-eQTL SNPs. The 30 CD-associated loci in the meta-analysis conducted by Barrett *et al*. were tested on their *cis*-eQTL effects in the publicly available expression database used in our study (8,15) rs2872507 is correlated with the expression of two genes *ORMDL3* and *GSDML* (26).
[b]Cis-eQTL effect of 13 identified SNPs within the top 500 of a publicly available GWAS data set from the US NIDDK Consortium.

A combined analysis was performed using the data from the discovery GWAS and both our replication cohorts (4,15). A second separate meta-analysis was then performed using the data of another publicly available database of the CD meta-analysis conducted by Barrett *et al*. (8) and both our replication cohorts.

RESULTS

CD-associated SNPs are more likely to be *cis*-eQTLs

To confirm our hypothesis that SNPs associated with CD are more likely to be eQTLs, we compared the amount of eQTL SNPs in the 30 established CD SNP with the amount expected by chance. Among the 30 top SNPs, five eQTLs were found ($P < 0.05$ corrected for FDR). We found after 100 permutations that this was higher than expected by chance ($P = 0.01$).

Allelic association analysis

Results for the allelic association analysis for replication phases 1 and 2 are depicted in Tables 2 and 3. In the first replication phase, 10 SNPs were tested in a Dutch cohort of 777 CD cases and 964 healthy controls and we observed a significant association with CD for three SNPs.

rs2298428 in *UBE2L3* [$P = 4.6 \times 10^{-4}$, odds ratio (OR) = 0.73, confidence interval (CI) 0.61–0.87], SNP rs2927488 in *BCL3* ($P = 0.011$, OR = 0.80, CI 0.68–0.95) and rs725660 in *SYMPK* ($P = 0.029$, OR = 1.16, CI 1.01–1.32).

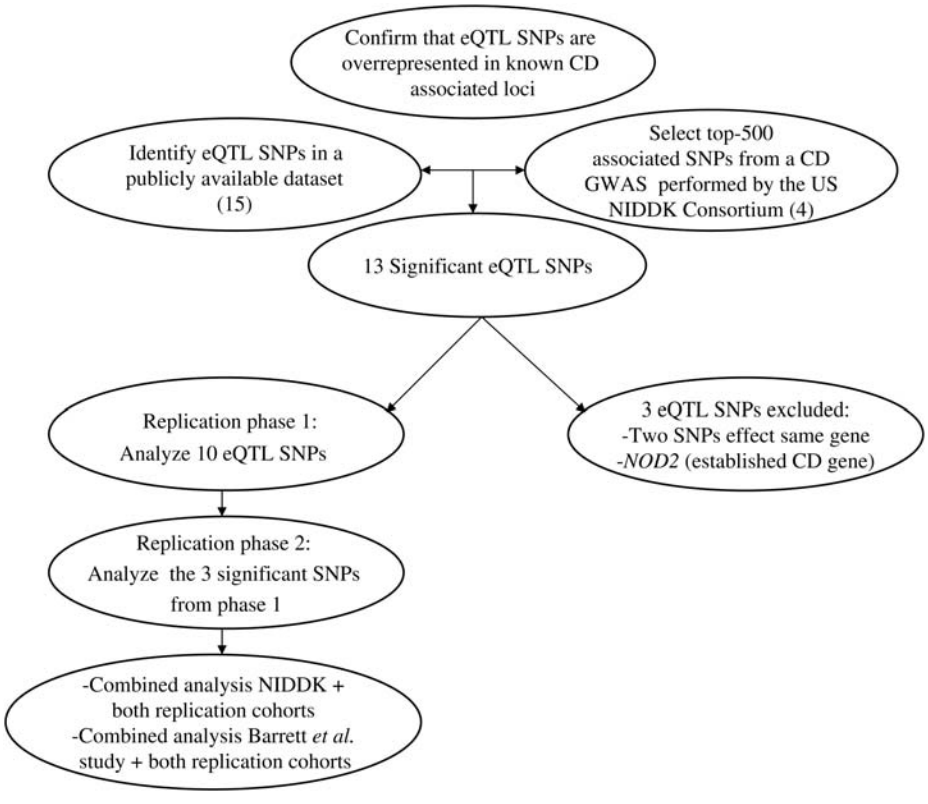


Figure 1. Study design.

In the second replication phase, we performed a follow-up analysis of these three SNPs in an independent cohort of 762 cases and 1648 controls. In this second cohort, we did not find any association for these SNPs ($P = 0.70, 0.68, 0.50$).

Combined analysis

The risk-increasing effect could be confirmed in a combined analysis including the original CD NIDDK GWAS data set and both our replication cohorts for two SNPs (*UBE2L3* $P = 5.22 \times 10^{-5}$, *BCL3* $P = 2.79 \times 10^{-4}$). The risk-increasing effect could not be confirmed for *SYMPK* with a *P*-value of 0.25. In a second combined analysis containing data of the CD meta-analysis by Barrett *et al*. (8) and both our replication cohorts, the risk-increasing effect could be confirmed for both SNPs (*UBE2L3* $P = 2.40 \times 10^{-7}$ and *BCL3* $P = 6.46 \times 10^{-4}$). For *SYMPK* the risk-increasing effect was not significant ($P = 0.06$) (Table 3). The meta-analysis performed by Barrett *et al*. contains the data of the GWAS used in the first combined analysis; to prevent overlap, this GWAS was excluded from the second combined analysis.

Risk alleles and expression

The eQTL SNP alleles associated with increased risk for CD had diverse effects on the expression of their correlated genes in a publicly available expression data set (15). For *UBE2L3*, the gene most strongly associated with CD, the minor allele that conferred risk was correlated with a higher expression of *UBE2L3* ($P = 4.21 \times 10^{-9}$) (Fig. 2A). In contrast, the risk variant of the *BCL3*-associated eQTL SNP was correlated with the lower expression of *BCL3* $P = 5.0 \times 10^{-5}$ (Fig. 2B).

DISCUSSION

We have identified two novel potential risk genes for CD: *UBE2L3* and *BCL3*. The SNPs that correlated with the expression of these genes were among the top 500 SNPs in the original GWAS but were not followed up (4,15). The association was strengthened in a combined analysis with two independent Dutch replication cohorts, although this could not be confirmed in all replication cohorts. By adding extracted data from a publicly available meta-analysis, the association of *UBE2L3* with CD is even further strengthened and almost reaching genome-wide significance ($P = 2.40 \times 10^{-7}$), whereas the association of *BCL3* was corroborated. In addition, we have shown that prioritizing eQTL SNPs from the top nominally associated SNPs of a GWAS for follow-up is a potentially promising strategy for identifying novel risk loci. This hypothesis is supported by the fact that *NOD2*, an established CD risk allele, is among the selected *cis*-eQTL SNPs in the top 500, although not in the top regions that were selected for follow-up in the original US NIDDK GWAS.

UBE2L3, the most significantly associated gene, encodes a protein involved in ubiquitination. This is the process in which abnormal or short-lived proteins are modified with ubiquitin to mark them for degradation. The protein encoded by

Table 2. Characteristics of cases and controls and genotyping techniques

	CD GWAS		Replication cohort I		Replication cohort II	
	Cases	Controls	Cases	Controls	Cases	Controls
Total number of samples	946	977	777	964	762	1648
Nationality	US-Canadian	US-Canadian	Dutch	Dutch	Dutch	Dutch
Platform	Illumina HumanHap300	Illumina HumanHap300	ABI Taqman	Illumina Quad670	ABI Taqman	ABI Taqman

GWAS, genome-wide association study; UMCG, University Medical Centre Groningen; AMC, Academic Medical Centre Amsterdam; UMCU, University Medical Centre Utrecht; VUMC, VU University Medical Centre; ABI, Applied Biosystems; NA, not available.

UBE2L3 ubiquitinates, among others, the NF-κB precursor p105. The risk allele of the *UBE2L3* eQTL SNP correlates with a higher expression of the *UBE2L3* gene. Theoretically, overexpression of *UBE2L3* could lead to a quicker degradation of the NF-κB precursor and thus to a lower production of NF-κB and consequently a diminished innate immune response. A similar effect is seen for the CD risk variants of *NOD2*, the strongest CD risk locus. The CD-associated *NOD2* variants also lead to an inadequate innate immune response because of a lack of the NF-κB precursor (16). Moreover, the protein encoded by *UBE2L3* has been shown *in vitro* to be involved in natural killer cell cytotoxic function, which is an important part of the innate immune response (17). SNPs in *UBE2L3* have also been found to be associated with celiac disease, rheumatoid arthritis and systemic lupus erythematosus (13,18,19), three immune-related diseases known to share risk loci with CD. Our study suggests that *UBE2L3* is yet another shared risk locus (20).

BCL3, the second likely novel CD risk gene, plays a role in mediating bacteria-induced colitis. Impaired Bcl3 expression in dendritic cells from $Il10^{-/-}$ mice leads to an increased expression of IL23 in reaction to bacterial lipopolysaccharides. *BCL3* also diminishes the inflammatory response induced by bacterial lipopolysaccharides in macrophages (21). The risk variant associated with CD in our study is correlated with a low expression of *BCL3*. This could point to an increased adaptive immune response in CD patients mediated by the increased expression of IL23. Indeed, IL23 appears to play an important role in the aberrant immune response that underlies CD (22).

Although the association for *UBE2L3* was strengthened in the combined analysis, we could not confirm it in the individual second replication phase. This might have several explanations; the first is possible lack of power. The more recently associated SNPs have lower ORs than the already established associations, so in order to detect new associations, the power of the studies needs to increase. As replication cohorts get exhausted, implementation is difficult. Secondly, there might be true heterogeneity in the populations we genotyped. For example *NOD2*, the most established risk allele for CD, cannot be confirmed in all populations (23). In favour of the association is that *P*-values become more significant after performing a combined analysis.

Our results show that selecting SNPs with an eQTL effect for replication is a potentially useful strategy for identifying novel CD risk genes. One disadvantage is that it will only detect risk loci which effect gene expression, whereas not all consistently replicated disease susceptibility loci have such eQTL effects. Therefore, selecting loci for follow-up on additional criteria (i.e. other functional effects) could further improve the yield of this follow-up strategy.

Newly identified CD risk loci can only improve our understanding of the disease mechanism if the effect of the risk-causing variant is known. This method of prioritizing eQTLs for replication not only improves the chances of finding relevant associations, but also provides a lead to functional studies. Since the eQTL SNP variants correlate with the expression of nearby genes, we would expect to see a difference in the expression of these genes in relevant tissues taken from patients and healthy controls. After measuring the expression of such genes in tissues relevant to the disease, assessing the functional effects of the differences in model systems might increase better understanding of CD pathogenesis.

We might have missed associated SNPs because we used gene expression data of celiac patients and HapMap data for finding *cis*-eQTL SNPs. It would be relevant to confirm the eQTL effect of SNPs on the expression level of *UBE2L3* and *BCL3* in blood or colonic mucosal biopsies of CD patients. Since CD is characterized by an aberrant immune response, causal variants are probably in the immune cells, e.g. blood.

In summary, we have identified two novel potential risk genes for CD, *UBE2L3* and *BCL3*, by prioritizing *cis*-eQTL SNPs for follow-up from the top 500 SNPs of a CD GWAS. *UBE2L3* is shared between several immune-related diseases (21), but both loci fit with the proposed role of aberrant immune responses in CD pathogenesis. This strategy for following up GWAS data provides both an effective and cost-efficient way of finding new risk loci and leads for functional studies.

MATERIALS AND METHODS

CD-associated SNPs are more likely to be eQTLs

We first assessed the 30 SNPs that recently have been reported to be associated with CD (8). We used two genetical genomics data sets in a meta-analysis setting, as reported by Heap *et al.* (15). These data sets comprise 109 celiac disease samples and 90 HapMap CEU samples. As the 109 celiac disease samples had been genotyped using Illumina HumanHap300 arrays, we attempted to impute all HapMap SNPs using Impute v2 and HapMap CEU release 23a. For 29 of the 30 SNPs, genotype data were eventually available, each having an MAF of at least 0.05, a call rate of at least 95% and exact HWE $P > 0.0001$. We investigated the 12 013 expression probes that were present in both genetical genomics data sets.

Table 3. Results for allelic association analysis of the NIDDK GWAS, the replication phases, the Barrett *et al.* meta-analysis and both combined analyses

Gene	Marker	Minor allele	Major allele	Expression effect minor allele	Risk effect minor allele	*P*-value GWAS	*P*-value first replication	OR first replication	*P*-value second replication	OR second replication	*P*-value Barrett *et al.* data set	Combined NIDDK and replication phases	Combined Barrett *et al.* and replication phases
UBE2L3	rs2298428	A	G	+	+	$7.71\times\mathbf{10^{-4}}$	$\mathbf{4.60\times10^{-4}}$	0.73	0.70	0.97	$\mathbf{2.04\times10^{-6}}$	$\mathbf{5.22\times10^{-5}}$	$\mathbf{2.39\times10^{-7}}$
BCL3	rs2927488	A	G	−	+	$5.94\times\mathbf{10^{-4}}$	**0.0107**	0.80	0.68	1.03	**0.003**	$\mathbf{2.79\times10^{-4}}$	$\mathbf{6.46\times10^{-4}}$
SYMPK	rs725660	A	C	+	+	$4.84\times\mathbf{10^{-4}}$	**0.029**	1.16	0.50	0.96	**0.008**	0.25	0.06
SURF1	rs3118663	A	G	−	+	$2.90\times\mathbf{10^{-4}}$	0.231						
COX11	rs1156287	G	A	−	−	$5.41\times\mathbf{10^{-5}}$	0.12						
C16orf75	rs243323	G	A	−	−	$3.35\times\mathbf{10^{-4}}$	0.287						
RARRES2	rs10278590	C	A	−	−	$3.46\times\mathbf{10^{-5}}$	0.665						
ZNF266	rs6512121	C	A	−	−	$6.41\times\mathbf{10^{-4}}$	0.5						
ENTPD5	rs1005564	A	G	−	+	$2.39\times\mathbf{10^{-4}}$	0.732						
CPEB4	rs359457	G	A	+	−	$6.20\times\mathbf{10^{-4}}$	0.586						

Values in bold are statistically significant. *UBE2L3*, ubiquitin-conjugating enzyme *E2L3*; *BCL3*, B-cell CLL/lymphoma 3; *SYMPK*, symplekin; *SURF1*, surfeit 1; *COX11*, COX11 homolog, cytochrome *c* oxidase assembly protein; *C16orf75*, chromosome 16 open-reading frame 75; *RARRES2*, retinoic acid receptor responder (tazarotene induced) 2; ZNF266, zinc finger protein 266; *ENTPD5*, ectonucleoside triphosphate diphosphohydrolase 5; *CPEB4*, cytoplasmic polyadenylation element-binding protein 4.

Figure 2. (**A**) eQTL effect of rs2298428 on the expression of *UBE2L3*. On the *X*-axis, the three different genotypes for SNP rs2298428 are displayed and on the *Y*-axis the level of expression for *UBE2L3*. Each dot represents the expression level of *UBE2L3* for one individual; the individuals are grouped per genotype. The level of expression of gene *UBE2L3* is correlated to the different genotypes. Data for this analysis were obtained from publicly available expression data from patients with celiac disease and HapMap. (**B**) eQTL effect of rs2927488 on the expression of *BCL3*. On the *X*-axis, the three different genotypes for SNP rs2927488 are displayed and on the *Y*-axis the level of expression for *BCL3*. Each dot represents the expression level of *BCL3* for one individual; the individuals are grouped per genotype. The level of expression of gene *BCL3* is correlated to the different genotypes. Data for this analysis were obtained from publicly available expression data from patients with celiac disease and HapMap.

We conducted a *cis*-eQTL analysis (SNP–probe distance $<$250 kb, 1000 permutations) and identified five significant *cis*-eQTLs (FDR controlled at 0.05) (Supplementary Material, Fig. S1).

We subsequently assessed whether the five *cis*-eQTLs we had detected were higher than expected by chance. For each of the 29 included SNPs, we determined the MAF and assessed how many probes mapped within 250 kb distance. We then selected a random set of 29 SNPs, but ensured that

each randomly selected SNP had an MAF and number of probes in its vicinity that matched the original SNP. We subsequently assessed how many significant *cis*-eQTLs could be identified in this permuted set of SNPs (using identical settings as in the original *cis*-eQTL analysis). We ran 100 permutations and observed that none of the permutations identified at least five *cis*-eQTLs for the random set of matched SNPs (four *cis*-eQTLs were found at most, occurring in nine out of 100 permutations). This indicates that the top 30 CD SNPs are significantly enriched for *cis*-eQTLs ($P < 0.01$).

SNP selection

Based on these results, we reasoned that if a high-ranking SNP, but not reaching genome-wide significance, affects gene expression in *cis*, it is more likely to be a true disease association. We decided to investigate the top 500 SNPs of a publicly available GWAS performed by the US NIDDK Consortium (http://www.ncbi.nlm.nih.gov/gap) (4). Four hundred and ninety-eight of these 500 SNPs had been genotyped or imputed in our genetical genomics data sets. Four hundred and ninety-four SNPs out of 498 SNPs passed QC (having an MAF of at least 0.05, a call rate of at least 95% and an exact HWE $P > 0.0001$). Using identical eQTL analysis settings, we identified 13 significant *cis*-eQTLs. One of these SNPs correlated with the expression of *NOD2*. Since *NOD2* is an established CD risk gene, it was not included in our independent replication study. For two genes, *COX11* and *ENTPD5*, we had more than one eQTL SNP in our database, so we selected the SNP with the strongest eQTL effect for replication because this is more likely to be a causative variant. In total, we analysed the 10 remaining SNPs for replication in an initial cohort. The three SNPs that were significantly associated with CD ($P < 0.05$) were replicated in an independent second cohort (Fig. 1).

Subjects

Our initial analysis, in which we selected SNPs for follow-up, was done in a GWAS data set from a US-Canadian cohort of 946 CD patients and 977 healthy controls (4). The first replication analysis of the selected SNPs was then performed in a Dutch cohort of 777 CD patients and 964 healthy controls (Replication cohort I). The CD patients for this replication were collected by the University Medical Centre Groningen ($n = 322$) and by the Academic Medical Centre in Amsterdam ($n = 455$) (24). The 964 healthy controls were blood donors recruited from donor centres in Utrecht and Amsterdam (Table 2) (25).

The SNPs that were found to be associated with CD in Replication cohort I ($P < 0.05$) were genotyped in a second cohort (Replication cohort II) of 762 Dutch CD patients and 1684 Dutch controls. The CD patients for the second cohort were collected by the University Medical Centre Leiden ($n = 287$), the VU University Medical Centre in Amsterdam ($n = 317$) and the Radboud University in Nijmegen ($n = 158$). The healthy controls were blood donors recruited from the donor centre in Groningen ($n = 720$) and healthy controls participating in a chronic obstructive pulmonary disease GWAS ($n = 964$).

Barrett *et al*. (8) performed a meta-analysis based on three GWAS performed by the US NIDDK consortium, The UK Wellcome Trust Case Control Consortium and a Belgian-French collaboration. This analysis contained a total of 3230 cases and 4829 controls. The results were used in a second combined analysis.

Recruitment of participants was approved by the institutional review boards of each of the hospitals, and informed consent was obtained from all participants.

Genotyping

Genotyping of all CD cases from both replication cohorts was performed using TaqMan technology (Applied Biosystems, Foster City, CA, USA). SNP genotyping assays were obtained from Applied Biosystems and genotyping was carried out as recommended by the manufacturer. The patient DNA samples were processed in 384-well plates, each plate containing 16 genotyping controls [four duplicates of four DNA samples from the Centre d'Etude de Polymorphisme Humain (CEPH)]. All SNPs were successfully genotyped in more than 95% of all samples. We had 99% concordance between our genotype data and the CEPH data available from HapMap. Genotyping of the controls was performed on either the Illumina Human 610-Quad or 670-Quad-custom Beadchips, following the manufacturer's protocol (Illumina Inc., San Diego, CA, USA). Quality control on this data was performed by excluding all SNPs that were out of Hardy–Weinberg equilibrium (HWE) [*P*-value (HWE) <0.001] and only including SNPs that were successfully genotyped in 99% of all the samples.

All selected SNPs in the control population were in HWE.

Statistical analysis

Differences in allele and genotype distribution between cases and controls of the individual cohorts were tested for significance by the χ^2 test. The significance threshold for *P*-values was set at 0.05. ORs were calculated and the CIs were approximated using Woolf's method with Haldane's correction. A combined analysis of the initial analysis and of the replication phases was performed with the METAL program (http://www.sph.umich.edu/csg/abecasis/metal).

A second meta-analysis of both the replication phases and the publicly available Barrett *et al*. database was performed. Only *P*-values were available, so a weighted *z*-score meta-analysis was performed. This analysis was performed separately because the Barrett *et al*. database is based on a meta-analysis which contains the data of the GWAS performed by the NIDDK.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank all the patients and controls who participated in this study and Jackie Senior for correcting the manuscript.

Conflict of Interest statement. None declared.

FUNDING

This study was supported by a clinical fellowship grant (90.700.281) to R.K.W., E.A.M.F. is supported by a clinical traineeship research grant (92.003.533), a VICI grant (918.66.620) to C.W., a VENI grant (863.09.007) to J.F. and a VENI grant (916.10.135) to L.F., all from the Netherlands Organization for Scientific Research (NWO). Further support was provided by a grant from the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009) to C.W. and a Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031) to L.F.

REFERENCES

1. Loftus, E.V. Jr (2004) Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. *Gastroenterology*, **126**, 1504–1517.
2. Baumgart, D.C. and Carding, S.R. (2007) Inflammatory bowel disease: cause and immunobiology. *Lancet*, **369**, 1627–1640.
3. Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., Cardon, L., Takazoe, M., Tanaka, T., Ichimori, T. *et al.* (2005) Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum. Mol. Genet.*, **14**, 3499–3506.
4. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, H.A., Abraham, C., Regueiro, M., Griffiths, A. *et al.* (2006) A genomewide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**, 1461–1463.
5. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Wu Datta, L. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.*, **39**, 596–604.
6. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A. *et al.* (2007) 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.
7. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
8. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, L.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.
9. Ioannidis, J.P., Thomas, G. and Daly, M.J. (2009) Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.*, **10**, 318–329.
10. Gilad, Y., Rifkin, S.A. and Pritchard, J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.
11. Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A.R., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D. *et al.* (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.*, **40**, 395–402.
12. Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E. *et al.* (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, **448**, 470–473.
13. Dubois, P.C., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A., Adány, R., Aromaa, A. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295–302.
14. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. and Cox, N.J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **4**, e1000888.
15. Heap, G.A., Trynka, G., Jansen, R.C., Bruinenberg, M., Swertz, M.A., Dinesen, L.C., Hunt, K.A., Wijmenga, C., vanHeel, D.A. and Franke, L. (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics*, **2**, 1.
16. Rosenstiel, P., Sina, C., End, C., Renner, M., Lyer, S., Till, A., Hellmig, S., Nikolaus, S., Fölsch, U.R., Helmke, B. *et al.* (2007) Regulation of DMBT1 via NOD2 and TLR4 in intestinal epithelial cells modulates bacterial recognition and invasion. *J. Immunol.*, **178**, 8203–8211.
17. Fortier, J.M. and Kornbluth, J. (2006) NK lytic-associated molecule, involved in NK cytotoxic function, is an E3 ligase. *J. Immunol.*, **176**, 6454–6463.
18. Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., Hu, Z., Xu, J.H., Cai, Z.M., Huang, W., Zhao, G.P. *et al.* (2009) Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.*, **41**, 1234–1237.
19. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A.S., Zhernakova, A., Hinks, A. *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.*, **42**, 508–514.
20. Zhernakova, A., van Diemen, C.C. and Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.*, **10**, 43–55.
21. Muhlbauer, M., Chilton, P.M., Mitchell, T.C. and Jobin, C. (2008) Impaired Bcl3 upregulation leads to enhanced lipopolysaccharide-induced interleukin (IL)-23P19 gene expression in IL-10(−/−) mice. *J. Biol. Chem.*, **283**, 14182–14189.
22. Kobayashi, T., Okamoto, S., Hisamatsu, T., Kamada, N., Chinen, H., Saito, R., Kitazume, M.T., Nakazawa, A., Sugita, A., Koganei, K. *et al.* (2008) IL23 differentially regulates the Th1/Th17 balance in ulcerative colitis and Crohn's disease. *Gut*, **57**, 1682–1689.
23. Arnott, I.D.R., Ho, G.T., Nimmo, E.R. and Satsangi, J. (2005) Toll-like receptor 4 gene in IBD: further evidence for genetic heterogeneity in Europe. *Gut*, **54**, 308–309.
24. Weersma, R.K., Stokkers, P.C., van Bodegraven, A.A., van Hogezand, R.A., Verspaget, H.W., de Jong, D.J., van der Woude, C.J., Oldenburg, B., Linskens, R.K., Festen, E.A.M. *et al.* (2009) Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. *Gut*, **58**, 388–395.
25. van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C.N.M., Bethel, G., Holmes, G.K.T. *et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.*, **39**, 827–829.
26. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolia, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

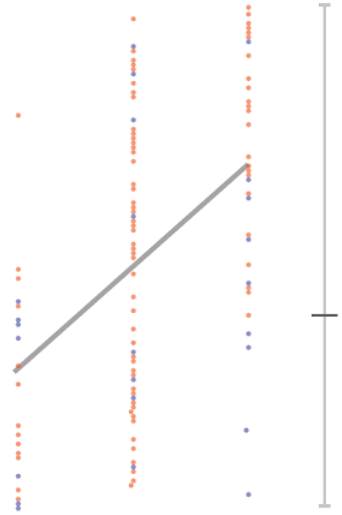
Supplementary figures

eQTL effects of five known Crohn's disease associated loci. (a) *GSMDL* (b) *SLC22A5* (c) *ORMDL3* (d) *UBE1L* (e) *RNASET2*

On the X-axis the three different genotypes are displayed, on the Y-axis the level of expression for each gene. Each dot represents the expression level of the gene for one individual; the individuals are grouped per genotype. The level of expression of each gene is correlated to the different genotypes. Data for this analysis were obtained from publicly available expression data from patients with celiac disease and HapMap.

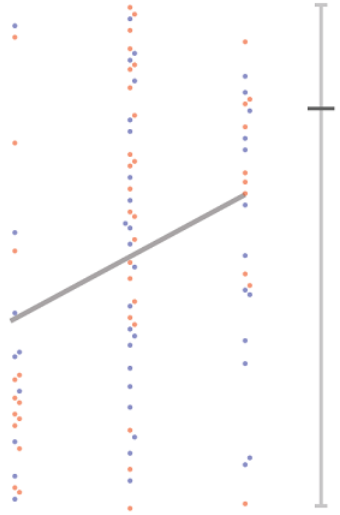
SNP rs2872507 (chr. 17, 35294289 bp)
Probe GI_8924175 (chr. 17, 35314498 - 35314547 bp), GSDML
P-Value 5.2E-9 **P-Value Abs.** 5.2E-9

Celiac



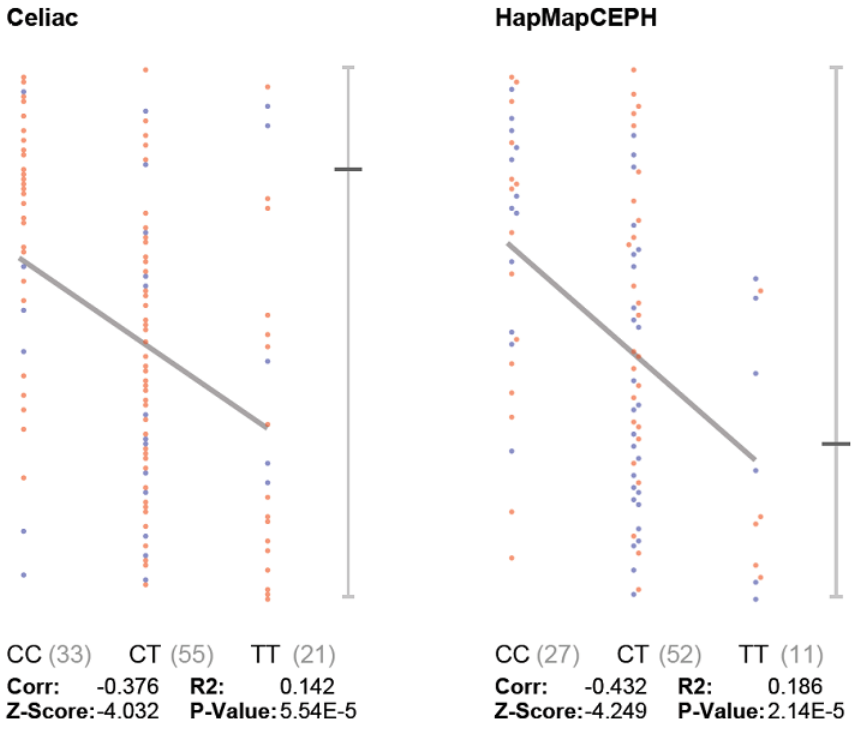
AA (20) AG (57) GG (32)
Corr: 0.48 **R2:** 0.23
Z-Score: 5.282 **P-Value:** 1.28E-7

HapMapCEPH



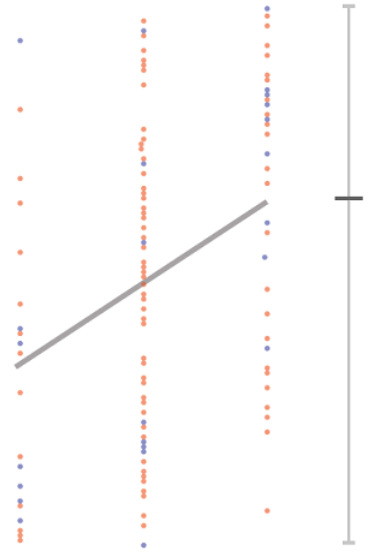
AA (22) AG (45) GG (23)
Corr: 0.3 **R2:** 0.09
Z-Score: 2.876 **P-Value:** 0.004

SNP rs2188962 (chr. 5, 131798704 bp)
Probe GI_24497491 (chr. 5, 131758857 - 131758906 bp), SLC22A5
P-Value 5.18E-9 **P-Value Abs.** 5.18E-9



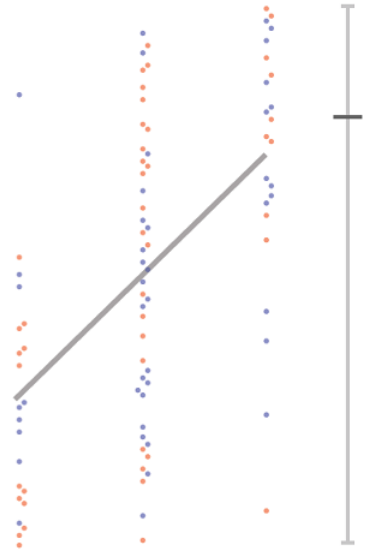
SNP rs2872507 (chr. 17, 35294289 bp)
Probe GI_27544926 (chr. 17, 35331073 - 35331122 bp), ORMDL3
P-Value 6.94E-11 **P-Value Abs.** 6.94E-11

Celiac



AA (20) AG (57) GG (32)
Corr: 0.356 **R2:** 0.127
Z-Score: 3.801 **P-Value:** 1.44E-4

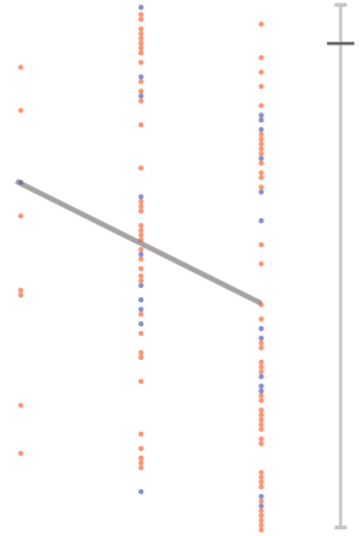
HapMapCEPH



AA (22) AG (45) GG (23)
Corr: 0.542 **R2:** 0.294
Z-Score: 5.521 **P-Value:** 3.37E-8

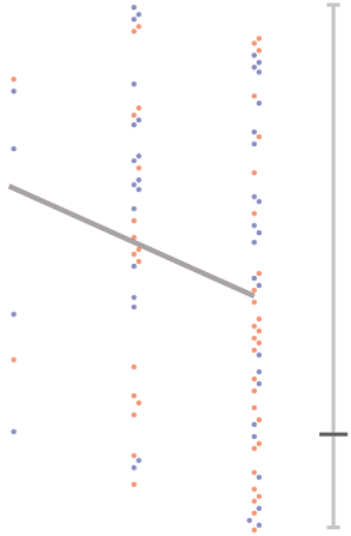
SNP rs3197999 (chr. 3, 49696536 bp)
Probe GI_38045947 (chr. 3, 49817752 - 49817801 bp), UBE1L
P-Value 9.79E-4 **P-Value Abs.** 9.79E-4

Celiac



AA (8) AG (46) GG (55)
Corr: -0.246 **R2:** 0.06
Z-Score: -2.577 **P-Value:** 0.01

HapMapCEPH



AA (6) AG (33) GG (51)
Corr: -0.218 **R2:** 0.047
Z-Score: -2.064 **P-Value:** 0.039

SNP rs2301436 (chr. 6, 167357978 bp)
Probe GI_38683865 (chr. 6, 167263061 - 167263110 bp), RNASET2
P-Value 6.52E-5 **P-Value Abs.** 6.52E-5