# Independent and Population-Specific Association of Risk Variants at the IRGM Locus with Crohn's Disease

*Human Molecular Genetics, 2010, Vol. 19, No. 9* **1828–1839**
*doi:10.1093/hmg/ddq041*
*Advance Access published on January 27, 2010*

# Independent and population-specific association of risk variants at the *IRGM* locus with Crohn's disease

**Natalie J. Prescott[1,*], Katherine M. Dominy[1], Michiaki Kubo[2], Cathryn M. Lewis[1], Sheila A. Fisher[1], Richard Redon[3], Ni Huang[3], Barbara E. Stranger[3,10], Katarzyna Blaszczyk[1], Barry Hudspith[4], Gareth Parkes[5], Naoya Hosono[2], Keiko Yamazaki[2], Clive M. Onnie[6], Alastair Forbes[7], Emmanouil T. Dermitzakis[3,11], Yusuke Nakamura[8], John C. Mansfield[9], Jeremy Sanderson[5], Matthew E. Hurles[3], Roland G. Roberts[1] and Christopher G. Mathew[1]**

[1]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK, [2]Laboratory for Genotyping Development, Center of Genomic Medicine, RIKEN Yokohama Institute, Yokohama City, Japan, [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [4]Nutritional Sciences Division, King's College London, Waterloo Campus, London SE1 9NH, UK, [5]Department of Gastroenterology, Guy's & St Thomas' NHS Foundation Trust, St Thomas' Hospital, London SE1 7EH, UK, [6]Department of Gastroenterology, Whittington Hospital NHS Trust, London NW11 6BJ, UK, [7]Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK, [8]Laboratory of Molecular Medicine, Human Genome Centre, Institute of Medical Science, University of Tokyo, Tokyo 108, Japan, [9]Department of Gastroenterology and Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK, [10]Division of Genetics, Harvard Medical School/Brigham and Women's Hospital, Boston MA, USA and [11]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

Received October 12, 2009; Revised January 12, 2010; Accepted January 25, 2010

**DNA polymorphisms in a region on chromosome 5q33.1 which contains two genes, immunity related GTPase related family, M (*IRGM*) and zinc finger protein 300 (*ZNF300*), are associated with Crohn's disease (CD). The deleted allele of a 20 kb copy number variation (CNV) upstream of *IRGM* was recently shown to be in strong linkage disequilibrium (LD) with the CD-associated single nucleotide polymorphisms and is itself associated with CD ($P < 0.01$). The deletion was correlated with increased or reduced expression of *IRGM* in transformed cells in a cell line-dependent manner, and has been proposed as a likely causal variant. We report here that small insertion/deletion polymorphisms in the promoter and 5′ untranslated region of *IRGM* are, together with the CNV, strongly associated with CD ($P = 1.37 \times 10^{-5}$ to $1.40 \times 10^{-9}$), and that the CNV and the 5′-untranslated region variant $-308(GTTT)_5$ contribute independently to CD susceptibility ($P = 2.6 \times 10^{-7}$ and $P = 2 \times 10^{-5}$, respectively). We also show that the CD risk haplotype is associated with a significant decrease in *IRGM* expression ($P < 10^{-12}$) in untransformed lymphocytes from CD patients. Further analysis of these variants in a Japanese CD case–control sample and of *IRGM* expression in HapMap populations revealed that neither the *IRGM* insertion/deletion polymorphisms nor the CNV was associated with CD or with altered *IRGM* expression in the Asian population. This suggests that the involvement of the *IRGM* risk haplotype in the pathogenesis of CD requires gene–gene or gene–environment interactions which are absent in Asian populations, or that none of the variants analysed are causal, and that the true causal variants arose after the European–Asian split.**

*To whom correspondence should be addressed at: Department of Medical and Molecular Genetics, King's College London School of Medicine, 7th Floor Tower Wing, Guy's Hospital, London SE1 9RT, UK. Tel: +44 2071883713; Fax: +44 2071882585; Email: natalie.prescott@genetics.kcl.ac.uk

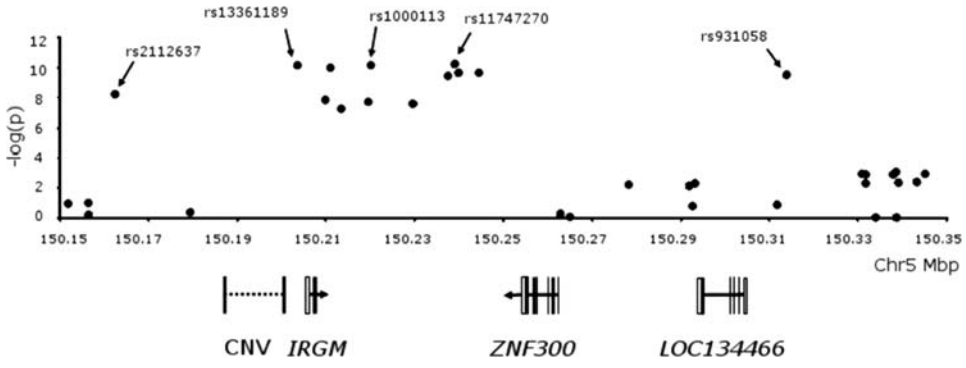


**Figure 1.** Association of 35 SNPs at the *IRGM-ZNF300* locus with CD (data from Ref. 16) indicated by −log of *P*-value. The physical location of the SNPs on chromosome 5 in Mbp is given on the *x*-axis with the relative position of the genes and CNV underneath.

## INTRODUCTION

Genome-wide association scans (GWAS) have been very successful in identifying susceptibility loci for Crohn's disease (CD), one form of chronic inflammatory bowel disease [reviewed in (1)]. The discovery by the Wellcome Trust Case Control Consortium (WTCCC) that single nucleotide polymorphisms (SNPs) near the immunity related GTPase related family, M (*IRGM*) gene on chromosome 5q33.1 were associated with CD provided a potentially important clue to its pathogenesis (2,3). *IRGM* is an atypical member of the IRG family of p47 immunity-related GTPase genes (4,5) which are characteristically induced by interferon and provide resistance to intracellular pathogens. The gene has had an unusual evolutionary history, with disruption of the open reading frame generating a non-functional pseudogene in Old and New World monkeys and apparent restoration of a truncated version in humans and African great apes (5). Although human *IRGM* lacks interferon-inducible elements in its promoter, reduction of its expression in culture was associated with impairment of induction of autophagy and clearance of intracellular bacteria (6,7). The region of association with CD also includes *ZNF300*, a gene whose product is reported to bind the promoter of the gene encoding interleukin 2 receptor beta-chain (8) involved in T cell-mediated immunity. *ZNF300* is expressed predominantly in heart, skeletal muscle and brain (9), with weaker expression in the small intestine. In addition to *IRGM* itself and *ZNF300*, which is transcribed in the opposite direction to *IRGM* (right to left in Fig. 1), the region also contains LOC134466, a pseudogene of *ZNF300* (Fig. 1). The nearest gene other than *IRGM* of functional interest at this locus is *TNIP1*, which is located 80 kb distal to the region of association. *TNIP1* encodes the tumour necrosis factor alpha inducing protein 3 (*TNFAIP3*) interacting protein, which inhibits NF-kB activation by tumour necrosis factor (10), and could conceivably be regulated by sequences within the region of association with CD. The original report (2) and replication (3) of the association of this locus with CD have been confirmed in several other studies (7,11–16).

The association of this locus with CD is clearly robust, but significant questions remain regarding the nature of its contribution to pathogenesis. In particular, we need to establish whether the association is driven by the *IRGM* gene itself or by other genes in the region, and to identify the causal variants in order to understand what effects they have on gene expression and function. Identification of causal variants also has the potential to provide more precise genetic markers of disease susceptibility (1,17). We reported previously (2) that extensive re-sequencing of the *IRGM* coding region did not reveal any obvious causal variants. A recent study by McCarroll *et al*. (7) showed that the deletion allele of a 20 kb copy number variant (CNV) that maps 1.6 kb upstream of *IRGM* is completely correlated ($r^2 = 1.0$) with the CD risk allele at the SNP rs13361189 (3). They also showed that the CNV deletion allele itself was significantly associated with CD in 172 cases and 344 controls ($P < 0.01$), and that the risk haplotype was correlated with altered expression levels of *IRGM* in cultured cells. *IRGM* expression from the risk haplotype was reduced in HeLa cells and in lymphoblastoid cell lines from 10 individuals, but increased in a colon carcinoma cell line and in smooth muscle cells. They therefore proposed that the CD association results from altered regulation of *IRGM*, and that the common deletion polymorphism is likely to be the causal variant. This was further supported by the fact that the strongest association with CD from this region in a North American GWAS was with rs13361189 (p $3.02 \times 10^{-4}$) just upstream of *IRGM* (7,18).

The fact that *IRGM* plays a role in autophagy, and that SNPs in another autophagy-related gene, autophagy 16-like isoform 1 (*ATG16L1*), are also associated with CD (2,18,19), add weight to the hypothesis that *IRGM* is the causal gene at this locus. However, given the extent of the association signal and the lack of experimental evidence that the CNV itself is directly responsible for the regulation of *IRGM* expression, we have undertaken a detailed genetic analysis of the contribution of this locus to susceptibility to CD. We have used the results of a large meta-analysis of three GWAS in CD which combined data from 3230 cases and 4829 controls (16) to provide a more robust estimate of the extent of the association across this locus. In addition we have carried out fine mapping in the region of association, and screening of all exon sequences, including *ZNF300* and the previously neglected *IRGM* promoter and exon 1, for novel genetic variants. This was followed by an association study and conditional analysis of novel and known variants in a large UK-based case–control (1800 versus 2000) cohort. Finally, we investigated the expression of candidate genes and the association of candidate variants in different populations, and examined *IRGM* expression in a physiologically relevant primary tissue (lymphocytes) from CD patients of known risk

genotypes. Our results provide novel insights into the contribution of sequence variants at this locus to disease susceptibility.

## RESULTS

### Disease association at the IRGM locus

The WTCCC study (2) found strong association of 11 SNPs with CD, spanning a 110 kb region of chromosome 5 (from rs13361189 at 150 203 580 bp to rs931058 at 150 313 891 bp, NCBI build 36). In addition, a meta-analysis of three GWAS in CD (16) included 35 SNPs in the 200 kb interval from 150 150 000 bp to 150 350 000 bp on chromosome 5. The results, which are plotted onto the physical map of the region in Figure 1, show strong association with CD from the SNP rs2112637 at position 150 162 627 bp to SNP rs931058 at 150 313 891 bp. The most significant SNPs were rs11747270 and rs1000113 ($P = 6.37 \times 10^{-11}$ and $7.5 \times 10^{-11}$), which are both located within the 42 kb of non-coding DNA between *IRGM* and *ZNF300*, and rs13361189 ($P = 8.17 \times 10^{-11}$) just upstream of *IRGM*. These data suggest that, purely on the basis of physical location, both *IRGM* and *ZNF300* should be considered as candidates for the source of the association signal.

In order to evaluate the extent of the association signal and to detect any possible additional associated common haplotypes not well tagged on the Affymetrix 500K SNP array, genotypes from the HapMap panel, from Caucasian Europeans from Utah (CEU), were used to identify eight additional SNPs which provided more complete tagging of the region. These were genotyped in 931 CD cases and 976 controls (Supplementary Material, Table S1). The only new tagging SNP that was associated with CD was rs12659118, located within LOC134466 ($MAF_{CD} = 0.116$, $MAF_{CON} = 0.084$, $P = 0.0017$). This SNP is in strong linkage disequilibrium (LD) with rs13361189 in controls ($r^2 = 0.79$), and was not associated with CD in individuals who did not carry the risk allele at rs13361189 ($MAF_{CD} = 0.019$, $MAF_{CON} = 0.016$, $P = 0.55$).

### Sequencing IRGM and ZNF300

The strong association across the region suggested that re-sequencing the *IRGM* and *ZNF300* genes to screen for possible causal variants was warranted. The six exons and adjacent splice sites of the *ZNF300* gene were sequenced (see Materials and Methods) in 45 cases. The only variant detected was a synonymous SNP, rs17800771, which is in strong LD with the SNP rs2290989 ($r^2 = 0.88$) that was genotyped in the WTCCC scan and was not associated with CD ($P = 0.42$).

Previous re-sequencing of the coding regions of *IRGM* in more than 700 CD cases detected two non-synonymous SNPs, E17D and T94K, which were not associated with CD in a sample of 769 cases and 705 controls, and an exonic synonymous SNP (L105 or rs10065172), which was associated (3). We extended this study by genotyping E17D and T94K in an expanded panel of 1400 cases and 1800 controls. Neither E17D ($MAF_{CD} = 0.0028$, $MAF_{CON} = 0.0013$) nor T94K ($MAF_{CD} = 0.044$, $MAF_{CON} = 0.042$) were associated with CD ($P = 0.19$ and 0.86, respectively).

In view of the lack of association of potential functional coding variants we investigated the upstream region of this gene for variants that might affect *IRGM* expression. The human and ape versions of *IRGM* are unusual in that the ancestral promoter has been supplanted by the promoter element of an inserted endogenous retrovirus (ERV9) long terminal repeat (LTR) ~1.6 kb upstream of the initiation codon (4,5). This has also introduced an upstream exon (exon 1) which encodes the first 695 bp of the 1.11 kb 5′ untranslated region (UTR) of *IRGM* and contains the ERV9 U5 repeat elements (Fig. 2). A 2.9 kb region spanning the *IRGM* initiation codon, the entire 5′-UTR including exon 1 and intervening intron, ERV9 LTR and promoter were sequenced in 94 unrelated individuals, including 43 cases of CD. Two insertion/deletion (indel) polymorphisms were detected (Fig. 2). One is a 4 bp insertion in the promoter region of the ERV9 LTR (−1644insTGGG) and the other is a 12 bp insertion in exon 2 (308 bp upstream of the initiation codon) which has also been detected in the Ghanaian population (20). The −308 variant is a microsatellite which has a common allele $(GTTT)_2$ and two additional alleles, $(GTTT)_4$ and $(GTTT)_5$. The −1644ins is located between three closely juxtaposed transcription factor (TF) binding sites for nuclear factor gamma, myeloid zinc finger 1 and GATA binding protein 2. The region upstream of *IRGM* also contains a CNV (21–23) which is correlated with altered expression of *IRGM* (6). Fine-mapping of the CNV on a high-resolution array and sequencing of the breakpoint revealed a deletion of 20 101 bp, spanning from 150 183 354 to 150 203 455 on chromosome 5 (NCBI 36, Fig. 2). The position and size of the CNV, as we identified it, is in close agreement with what has been reported previously (5,7).

### Association of IRGM promoter and CNVs with CD

The association of the *IRGM* promoter indel, microsatellite and upstream CNV with CD was investigated by analysis of these variants together with a strongly associated and replicated SNP from the WTCCC study (rs13361189) (2,3) and the synonymous coding SNP (L105 or rs10065172) in 1848 CD cases and 2025 population controls. Mapping of the CNV breakpoints enabled the design of a robust, qualitative assay with a common forward primer positioned immediately upstream of the CNV and two allele-specific primers, one located in the deleted region and the other immediately downstream of the CNV [Materials and Methods and (24)]. The results (Table 1) show that all these variants are strongly associated with CD, with the most significant signals coming from the deletion allele of the CNV and from the WTCCC SNP, rs13361189 ($P_{allele} = 1.4 \times 10^{-9}$ and $3.7 \times 10^{-9}$, respectively). However, allele frequencies and odds ratios for all the variants with the exception of the CD-associated allele $-308(GTTT)_5$ are very similar. As reported by McCarroll *et al*. (7), the CNV was in strong LD with rs13361189 and rs10065172 (Fig. 3). The promoter variant −1644ins was also in strong LD with the CNV and both of these SNPs ($r^2 > 0.9$ in cases and controls), whereas $-308(GTTT)_5$ was in moderate LD with the other four variants.

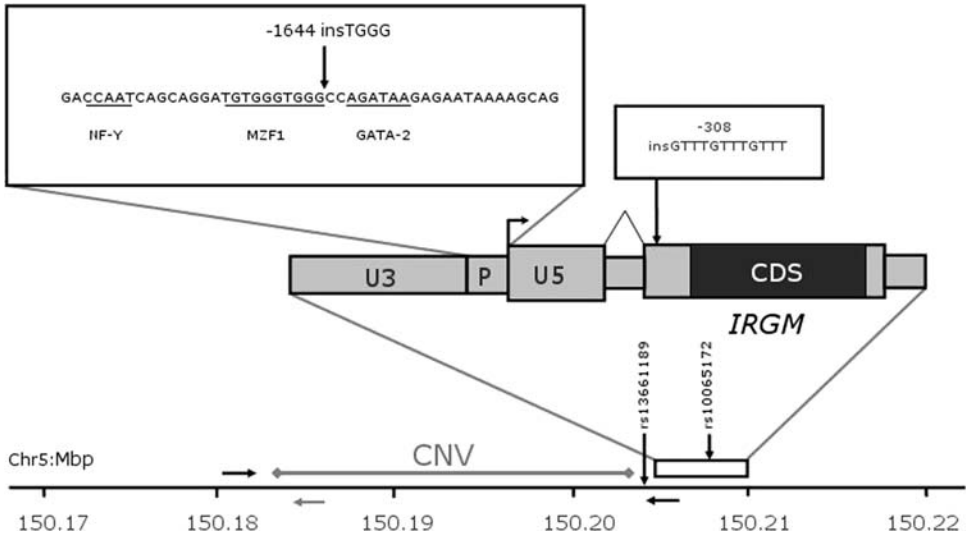


**Figure 2.** The structure of *IRGM* showing the location of the upstream CNV on chromosome 5. Horizontal arrows represent the relative positions of the three primers for the CNV PCR assay which uses the common left primer (black), and either the insertion right primer (grey) or deletion right (black). The position of the two risk SNPs are indicated by vertical arrows. The *IRGM* region consisting of the single coding exon 2 and upstream un-translated exon 1 containing the promoter (P) and U3/U5 repeats of the ERV9LTR is expanded above to indicate the relative positions of the two insertion/deletion polymorphisms (boxed). Potential TF binding sites adjacent to the −1644ins are underlined. The coding sequence (CDS) of *IRGM* is shaded in dark grey.

The existence of multiple highly correlated variants (some of which have potential functional significance) which are all associated with disease risk raises the question as to which, if any, of these might be a causal variant and thus driving the association at this locus. We investigated this by conditional logistic regression analysis across the five loci (Table 2). The analysis showed that the CNV remained highly significantly associated with disease when conditioned on the variants at −1644 and −308; that is to say, when all the association at either of the two variants was accounted for, there remained significant independent association with the CNV ($P = 1.6 \times 10^{-5}$ and $2.6 \times 10^{-7}$, respectively). However, the effect of the CNV on disease was not significant when conditioned on the two SNPs rs13361189 or rs10065172 ($P = 0.221$, 0.251). Thus the effect of the CNV could not be distinguished from the effect of either SNP, which is consistent with the strong LD between these three variants. Similarly, the two SNPs showed an association that was independent of the variants at −1644 and −308 but not of each other or the CNV. The −1644 variant showed significant independent association when conditioned on either the CNV ($P = 3.1 \times 10^{-4}$) or on −308 ($P = 9.3 \times 10^{-6}$), but was not significant or marginally so when conditioned on the two SNPs ($P =$ 0.142, 0.047). However, −308 showed highly significant independent association when conditioned on any of the other four variants ($P = 9.4 \times 10^{-6}$ to $2.38 \times 10^{-4}$). The apparently independent effect from −1644 appeared to be due to two very rare haplotypes (haplotypes 10 and 11 in Table 3) which had a combined frequency of 0.0035 in CD cases but were not present in controls; one of these (haplotype 10 in Table 3) contained the risk (del) allele at the CNV and the common (del) allele at −1644. In our case–control study the $-308(GTTT)_4$ (rare 8 bp insertion) allele was detected in only four CD cases and in none of the controls (haplotype 11 in Table 3). It is possible that these very rare haplotypes, that were only seen in CD cases ($n = 7$), were over-inflating the test statistic. The conditional regression analysis was therefore repeated with the exclusion of rare haplotypes with a frequency $<0.005$ (Table 4). In this analysis the independent effect observed previously for −1644 disappeared and thus the effects of the CNV, −1644 and the two SNPs were indistinguishable. However, all remained significant when conditioned on the −308, and conversely, −308 retained highly significant independent association with CD when conditioned against all other four variants ($P = 4.24 \times 10^{-5}$ to $3.9 \times 10^{-4}$). At least part of the independent effect for the −308 variant appeared to be due to a haplotype that had the non-risk (non-deleted) allele at the CNV but the high risk $(GTTT)_5$ allele at -308 (haplotype 3 in Table 3); this haplotype had a frequency of 5.2% in CD cases and 4.1% in controls and was associated with CD ($P = 0.038$; Table 3).

The analysis was repeated on a subset (75%) of cases (1265) and controls (1609) with complete genotypes for all five loci and produced very similar results (not shown). The regression analysis suggests that the haplotype represented by the CNV, the two SNPs and −1644ins constitutes one independent effect on disease risk, and that $-308(GTTT)_5$ may be another.

### Analysis of IRGM variants in other populations

The difficulties in identifying the origin of association signals at the *IRGM* locus in European populations led us to investigate the frequency of these variants and LD structure in other populations, since weaker LD might facilitate fine-mapping studies. We genotyped the CNV, promoter variant and microsatellite in five of the HapMap3 populations for whom genotype data were available for the original associated SNP rs13361189. These were the Han Chinese from Beijing (CHB), the Japanese from Tokyo, Japan (JPT), the Yoruba in Ibadan, Nigeria (YRI) and two other African populations, the Maasai in Kinyawa, Kenya (MKK) and Luhya in Webuye, Kenya (LWK). We found that the frequencies of

**Table 1.** Association of *IRGM* variants with Crohn's disease analysed in 1848 CD cases and 2025 population controls

| Variant | Risk allele | Risk allele frequency | | $P_{allele}$ | $P_{trend}$ | OR (95% CI) |
|---|---|---|---|---|---|---|
| | | Cases | Controls | | | |
| CNV | Del | 0.115 | 0.073 | $1.40 \times 10^{-9}$ | $1.9 \times 10^{-9}$ | 1.66 (1.41–1.95) |
| rs13361189 | C | 0.115 | 0.075 | $3.73 \times 10^{-9}$ | $5.10 \times 10^{-9}$ | 1.65 (1.40–1.95) |
| −1644 | Ins | 0.108 | 0.073 | $4.20 \times 10^{-7}$ | $4.20 \times 10^{-7}$ | 1.53 (1.38–1.80) |
| −308 | $(GTTT)_5$ | 0.149 | 0.111 | $1.37 \times 10^{-5}$ | $2.77 \times 10^{-6}$ | 1.37 (1.19–1.58) |
| | $(GTTT)_4$ | 0.001 | 0.000 | 0.0048[a] | n/a | n/a[b] |
| rs10065172 | T | 0.107 | 0.071 | $1.04 \times 10^{-7}$ | $1.01 \times 10^{-7}$ | 1.56 (1.33–1.85) |

Results show *P*-values for allele specific association tests and the genotype trend test which can also account for multi-allelic genotypes.
[a]Fishers exact test.
[b]Not possible due to absence in controls.

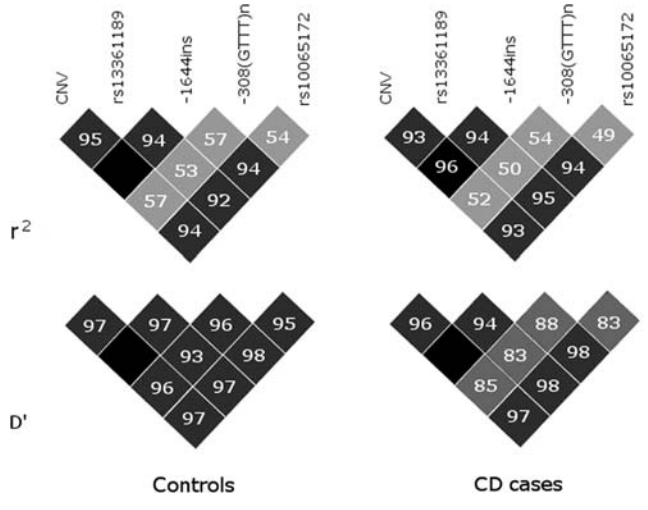


**Figure 3.** Linkage disequilibrium of *IRGM* variants in UK CD cases and controls.

all four CD risk variants were much higher in all these populations compared with the white UK population (Fig. 4). Indeed, the CNV deletion (CD risk) allele is the common allele in the YRI population, whereas the $-308(GTTT)_5$ CD risk allele is the common allele in both Asian populations. Also of interest is that the $-308(GTTT)_4$ allele, which is very rare in European populations, has a frequency of 8–18% in the African populations; the frequencies of the −308 alleles in the Yoruba group are similar to those reported in the West African population of Ghana (20). We observed that LD between the CNV and rs13361189 was lower in the JPT ($r^2 = 0.88$) and CHB ($r^2 = 0.84$) compared with Europeans ($r^2 = 0.95$), and was substantially lower in two of the three African populations (YRI: $r^2 = 0.70$, LWK: $r^2 = 0.66$). LD between the CNV and both the −1644 and the −308 variants was also lower in Asian and African populations as compared with Europeans (Supplementary Material, Fig. S1). The reduction in LD observed across this locus in Asian and African populations is consistent with the substantial differences in allele frequencies, and suggested that investigation of the association of these variants with CD in other populations might provide further insight into their contribution to CD.

## Association study of IRGM variants in Japanese CD cases and controls

An African case/control sample was not available, so we focused our analysis on a well-studied Japanese collection. A recent analysis of 484 Japanese CD cases and 470 controls from this collection found no association of SNPs rs13361189 or rs4958847 at the *IRGM* locus with CD (25). We genotyped the CNV, variants at −1644 and −308 and SNPs rs13361189 and rs10065172 in the same 484 CD cases and in an expanded set of 933 Japanese controls. The results (Table 5) show no association of these variants with CD, with the exception of a weak protective effect for the $-308(GTTT)_5$ allele ($P = 0.03$), which is the risk allele in Europeans. As in the UK population, there was very strong LD between the CNV, −1644, rs13361189 and rs10065172 ($r^2 > 0.95$), with −308 again being in weaker LD with the other four variants ($r^2 = 0.48$–$0.55$). One obvious explanation for this lack of association is inadequate power to detect small effects. Reported odds ratios for the *IRGM* locus were 1.33 and 1.34 in two meta-analyses (15,16) and 1.44 in a recent study of a large Dutch–Belgian cohort (13). Power to detect association at the CNV in this Japanese case/control sample with an allele frequency in controls of 0.38 is >90% for an OR = 1.35, $P = 0.05$, so the study is well powered assuming that the effect size in Japanese is similar to that in European populations.

## Effect of the risk haplotype on IRGM and ZNF300 expression

In view of the lack of obvious pathogenic or disease-associated variants in the coding regions of *IRGM* and the presence of several potential regulatory variants upstream of the ATG start codon (including −1644, −308 and the CNV), we then addressed the question of whether a correlation existed between the risk haplotype and expression levels of *IRGM*. McCarroll *et al.* (7) reported variable effects on *IRGM* expression in different cell lines and cell types (7). We analysed *IRGM* expression from the high- and low-risk haplotypes by sequencing cDNA and genomic DNA prepared from primary lymphocytes of eight CD patients who were heterozygous for the risk haplotype, and comparing the relative expression of the C (low-risk) and T (high-risk) alleles of the exonic SNP rs10065172 in these individuals (Supplementary Material, Fig. S2). This showed that expression of the T allele was mark-

**Table 2.** Conditional logistic regression and haplotype analysis of *IRGM* variants in CD cases/controls, all haplotypes

| | | Conditional locus CNV | rs13361189 | −1644 | −308 | rs10065172 |
|---|---|---|---|---|---|---|
| Test locus | CNV | 1.0 | 0.221 | $1.6 \times 10^{-5}$ | $2.6 \times 10^{-7}$ | 0.251 |
| | rs13361189 | 0.361 | 1.0 | 0.005 | $3.88 \times 10^{-6}$ | 0.388 |
| | −1644 | $3.1 \times 10^{-4}$ | 0.142 | 1.0 | $9.3 \times 10^{-6}$ | 0.047 |
| | −308 | $9.4 \times 10^{-6}$ | $2.38 \times 10^{-4}$ | $5.8 \times 10^{-5}$ | 1.0 | $1.50 \times 10^{-5}$ |
| | rs10065172 | 0.392 | 0.49 | $1.58 \times 10^{-3}$ | $1.12 \times 10^{-7}$ | 1.0 |

This is a test of multiple haplotypes (see Materials and Methods).

**Table 3.** Haplotypes analysis (all haplotypes)

| Haplotype | CNV | rs13361189 | −1644 | −308 | rs10065172 | Count[a] Case | Control | Frequency Case | Control | OR(95%CI) | *P* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | INS | T | DEL | $(GTTT)_2$ | C | 1668 | 2762 | 0.831 | 0.884 | ref | ref |
| 2 | *DEL* | *C* | *INS* | $(GTTT)_5$ | *T* | 191.5 | 207.8 | 0.095 | 0.066 | 1.53 (1.24–1.88) | $6.18 \times 10^{-5}$ |
| 3 | INS | T | DEL | $(GTTT)_5$ | C | 104.5 | 130.3 | 0.052 | 0.042 | 1.33 (1.02–1.73) | 0.038 |
| 4 | INS | *C* | DEL | $(GTTT)_2$ | C | 6.003 | 6.006 | 0.003 | 0.002 | 1.66 (0.53–5.14) | 0.386 |
| 5 | *DEL* | *C* | *INS* | $(GTTT)_2$ | *T* | 28.46 | 5.214 | 0.014 | 0.002 | 9.04 (3.47–23.53) | $4.76 \times 10^{-8}$ |
| 6 | *DEL* | *C* | *INS* | $(GTTT)_5$ | C | 0 | 4.906 | 0 | 0.002 | n/a | 0.032 |
| 7 | *DEL* | T | *INS* | $(GTTT)_5$ | *T* | 1.003 | 3.006 | 0.0005 | 0.001 | 0.55 (0.06–5.32) | 0.591 |
| 8 | INS | T | DEL | $(GTTT)_2$ | *T* | 0 | 3.005 | 0 | 0.001 | n/a | 0.092 |
| 9 | *DEL* | *C* | *INS* | $(GTTT)_2$ | C | 0 | 2.099 | 0 | 0.0007 | n/a | 0.161 |
| 10 | *DEL* | T | DEL | $(GTTT)_2$ | C | 3 | 0 | 0.0015 | 0 | n/a[b] | $5.20 \times 10^{-3}$ |
| 11 | *DEL* | *C* | DEL | $(GTTT)_4$ | *T* | 4 | 0 | 0.002 | 0 | n/a[b] | $1.55 \times 10^{-2}$ |

Potential risk alleles are italicized.
[a]Non-integer haplotype counts are due to maximum likelihood estimation.
[b]OR over-inflated due to lack of controls with this haplotype ($OR \gg 4 \times 10^6$).

**Table 4.** Conditional logistic regression and haplotype analysis of *IRGM* variants in CD cases/controls but excluding rare haplotypes ($f < 0.005$)

| | | Conditional locus CNV | rs1331189 | −1644 | −308 | rs10065172 |
|---|---|---|---|---|---|---|
| Test locus | CNV | 1 | 1 | 1 | $2.6 \times 10^{-7}$ | 1 |
| | rs13361189 | 1 | 1 | 1 | $3.88 \times 10^{-6}$ | 1 |
| | −1644 | 1 | 1 | 1 | $9.3 \times 10^{-6}$ | 1 |
| | −308 | $2.0 \times 10^{-5}$ | $9.76 \times 10^{-4}$ | $3.9 \times 10^{-4}$ | 1 | $4.24 \times 10^{-5}$ |
| | rs10065172 | 1 | 1 | 1 | $1.12 \times 10^{-7}$ | 1 |

This is a test of multiple haplotypes (see Materials and Methods).

edly lower than the C allele (C/T peak height ratio in cDNA: 1.82–353.44) in seven of eight samples tested ($P = 0.015$). We also analysed primary lymphocytes in 25 CD patients of defined *IRGM* genotype and measured expression of both genes by real-time quantitative RT–PCR. The risk haplotype is relatively rare in European populations, and expression levels of *IRGM* varied widely between individuals. Nonetheless, we found significantly lower *IRGM* expression ($P < 10^{-12}$) in homozygotes and heterozygotes for the risk haplotypes at all three loci, i.e. the CNV, −1644ins and $-308(GTTT)_5$ than in homozygotes for the absence of the risk haplotype (Fig. 5). Most individuals tested had the same genotype for all three variants as a result of the strong LD between them, so we could not test for independent effects of the variants on expression.

We next analysed the effect of *IRGM* genotype on *IRGM* expression using microarray expression data from lymphoblastoid cell lines in the Asian and extended African HapMap populations (26; Stranger *et al*., in preparation). The risk alleles for the CNV, rs13361189 and $-308(GTTT)_5$ were associated with a highly significant reduction in expression of *IRGM* in the YRI population and in pooled data from the three African populations (YRI, MKK, LWK), with much weaker association for −1644ins (Table 6). However, there was no association of any of the loci with altered *IRGM* expression in the Japanese or Chinese populations. Interestingly, overall expression of *IRGM* was higher in the JPT and CHB samples than in the three African populations, with the lowest expression across all populations observed in Europeans ($P < 10^{-15}$; Fig. 6).

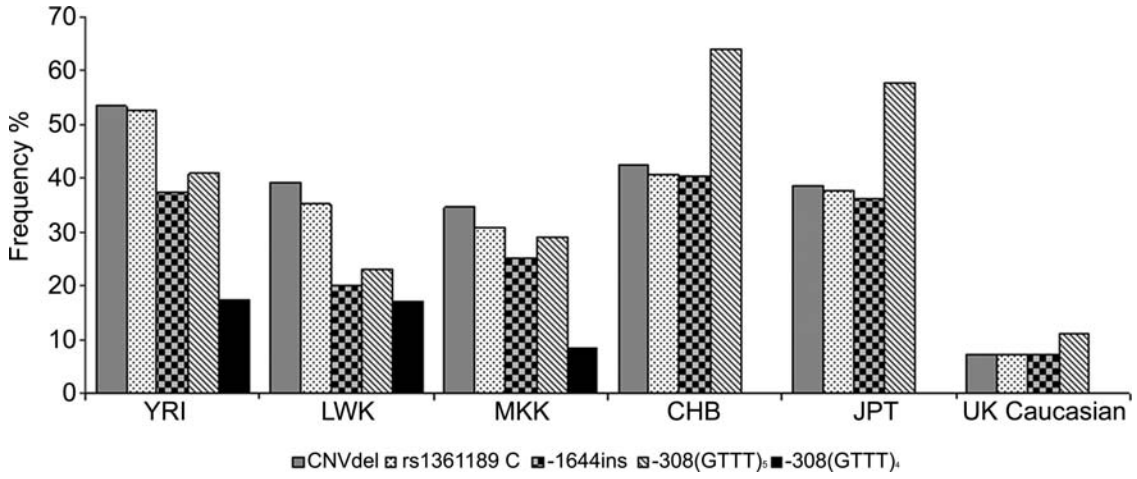


**Figure 4.** Frequency of *IRGM* variants in five HapMap populations from Asia and Africa (YRI 120, MKK 145, LWK 145, JPT 88, CHB 88) and 2025 white population controls from the 1958 MRC British birth cohort (BC1958).

In addition to these quantitative effects, we noted that the $-308(GTTT)_5$ allele potentially strengthens an alternative splice acceptor site 139 bp [or 151 bp on the $(GTTT)_5$ allele] downstream of the canonical splice site (Fig. 2 and Supplementary Material, Fig. S3) by extending its polypyrimidine tract. The splicing of *IRGM* mRNA was therefore investigated by RT–PCR in four individuals with three possible genotypes at −308: $(GTTT)_5/(GTTT)_5$, $(GTTT)_5/(GTTT)_2$ and $(GTTT)_2/(GTTT)_2$. The identity of each mRNA was determined by sequencing of gel-extracted products. The $-308(GTTT)_5$ resulted, as predicted, in use of the alternative splice site with removal of 139 bp from the 5′ untranslated region of the *IRGM* transcript (Supplementary Material, Fig. S3).

Finally, we investigated the expression of the other gene at this locus, *ZNF300*, since it remains a possible source of the association with CD. We looked for a correlation between the risk haplotype and *ZNF300* expression in lymphocytes from the same 25 CD patients that showed a correlation for *IRGM* but found none ($P = 0.10$ for patients typed for the CNV, data not shown). Similarly, analysis of microarray expression data from 141 HapMap samples did not detect significant correlation between the risk haplotype and *ZNF300* expression ($P = 0.45$). This does not exclude possible qualitative effects of the risk haplotype on *ZNF300* expression.

## DISCUSSION

In this study, we have addressed a question generic to the follow-up of GWAS in complex disease, which is how to define the causal genes and variants that are driving an association at a specific locus. At the *IRGM* locus, very significant association of SNPs with CD is seen across an interval which includes two known genes, *IRGM* and *ZNF300*. The biological evidence for the role of *IRGM* in autophagy, coupled with the correlation of the risk haplotype with altered *IRGM* expression, constitutes strong support for its role in the pathogenesis of CD. However, since the strongest association signals extend from just upstream of *IRGM* to a position midway between *IRGM* and *ZNF300*, and since *IRGM* does not contain an obvious functional variant, the *ZNF300* gene cannot be formally excluded as the source of the association signal. The strong LD between the CNV upstream of *IRGM* and SNPs associated with CD, and the correlation of the deletion allele with altered *IRGM* expression (7), further supports a primary role for *IRGM* in pathogenesis and for the deletion as the causal variant. However, functional evidence for the role of this gene in intestinal inflammation and for a direct regulatory effect of the CNV (as opposed to association with altered *IRGM* expression) is needed. We have sought to address the question of which gene and which variants might be driving the association by sequencing the upstream region of *IRGM* and the coding region of *ZNF300* to look for other potential causal variants, and by conditional analysis of the most strongly associated variants in a well powered sample of CD cases and controls. In addition we have investigated association of these variants with *IRGM* expression in CD cases from the UK as well as in five other populations from Africa and East Asia and carried out a case–control analysis in a Japanese population.

Sequencing of all the exons and adjacent splice sites of *ZNF300* did not reveal any functionally relevant or CD-associated variants. However, sequencing of the 5′-UTR, intron and complex ERV9-derived promoter region of *IRGM* identified two insertion/deletion polymorphisms, one of which is novel and located within the proximal promoter (c.*IRGM* −1644) between three TF binding sites. This variant is strongly associated with CD but is in tight LD with the CNV. Another indel at c.*IRGM* −308, has been previously described as a tetranucleotide repeat (microsatellite) in a Ghanian population (20). We found that the $-308(GTTT)_5$ allele was common in the UK population and was also significantly associated with CD. We also detected the 8 bp insertion allele $-308(GTTT)_4$, which had a frequency of 0.1% in CD cases and 0% in controls. This suggested that both alleles were possible new CD risk alleles, which was supported by the multi-allelic association test. Conditional regression analysis of our data provided highly significant support for an independent effect for the −308 microsatellite polymorphism. The $-308(GTTT)_5$ allele reinforces an alternative splice site

**Table 5.** Association analysis of *IRGM* variants in 484 Japanese CD cases and 933 Japanese controls

| Variant | Risk allele | Risk allele frequency Cases | Controls | $P_{allele}$ | OR (95%CI) |
|---|---|---|---|---|---|
| CNV | Del | 0.370 | 0.380 | 0.76 | 0.98 (0.83–1.14) |
| −1644 | Ins | 0.370 | 0.380 | 0.62 | 0.95 (0.81–1.12) |
| −308 | $(GTTT)^a_5$ | 0.510 | 0.553 | 0.03 | 0.84 (0.72–0.98) |
| rs13361189 | C | 0.393 | 0.379 | 0.43 | 1.07 (0.91–1.26) |
| rs10065172 | T | 0.366 | 0.379 | 0.52 | 0.94 (0.81–1.11) |

[a]No $(GTTT)_4$ alleles were observed in this population.

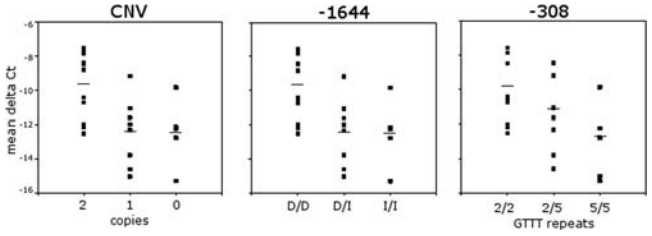


**Figure 5.** Quantitative analysis of *IRGM* expression by real-time PCR in lymphocytes from CD patients of defined genotype for CNV (2, 1 and 0 copies of the CNV where 1 and 0 are the heterozygous and homozygous risk genotypes, respectively) −1644 (D = del allele, I = risk insertion allele) and −308 (where 2× GTTT repeats represent the non-risk allele and 5× GTTT repeats represent the risk allele. Relative expression is measured by, and inversely proportional to, mean ΔCt (see Materials and Methods).

which removes 139 bp from the 5′ untranslated region of the *IRGM* transcript. The consequence of this interstitial deletion is not known, but it may affect the stability of the transcript or the rate at which it is translated.

A previous study (7) has shown that the risk haplotype at the *IRGM* locus is associated with either a reduction or an increase in *IRGM* expression, depending on the cell line analysed. We find that the risk alleles $-308(GTTT)_5$, CNVdel and −1644ins are all significantly associated with a down-regulation of *IRGM* expression in untransformed lymphocytes from CD patients. Given the strong LD between all three variants, a very large sample of individuals of known genotype would be required to determine whether each of the variants had independent effects on gene expression. It is possible that other as yet unknown variants at this locus may have different effects; the SNP −261C>T (rs9637876) has recently been reported to confer protection from tuberculosis caused by infection with *Mycobacterium tuberculosis* (27). Direct functional analysis of the effect of all these candidate causal variants on *IRGM* expression is likely to be required to fully resolve their contribution to CD pathogenesis.

The lack of association of the index SNPs or candidate causal variants with CD in the Japanese population is intriguing. This does not appear to be due to phenotypic differences, since CD in Japanese patients is clinically indistinguishable from Europeans. It is also unlikely to be due to insufficient statistical power, unless the effect size is significantly smaller in this population. An alternative explanation is that the contribution of *IRGM* to pathogenesis requires gene–gene or gene–environment interactions which are absent in the Japanese. It is also possible that none of the variants tested, including the CNV, are causal and that the causal variant at this locus arose after the European–Asian split, as is the case for the major CD susceptibility gene NOD2/CARD15 (28,29). A further possibility relates to the much higher expression of *IRGM* we observed in CHB and JPT HapMap samples than in the CEU samples. If the variants studied here result in only a modest decrease in *IRGM* levels, then the relative effect may be insufficient to influence disease risk in Japanese individuals, who show significantly higher expression than Europeans. Conversely, in European individuals, for whom we observed a much lower baseline expression, these variants may result in a larger relative effect that is sufficient to influence disease risk. This would be consistent with the lack of correlation between *IRGM* expression levels and risk haplotype in the JPT and CHB cell lines.

In conclusion, we have shown that multiple sequence variants upstream of the *IRGM* gene with potential gene regulatory effects are strongly associated with CD and with reduced *IRGM* expression in untransformed cells from CD patients. The lack of association of these variants with CD in the Japanese population suggests that they may have population-specific effects on the pathogenesis, or that more recent, un-described mutations may be responsible for the association in European populations. A combination of genetic and functional approaches will be required to fully understand the contribution of this locus to the development of this form of chronic inflammatory bowel disease.

## MATERIALS AND METHODS

### Patients and controls

More than 1800 patients with CD were recruited from specialist IBD clinics in London and Newcastle (30) after informed consent and ethical review (REC 05/Q0502/127). 2000 Population controls were obtained from the 1958 British Birth Cohort, which includes subjects born in 1 week of March 1958 in England, Scotland and Wales (31). UK case–control studies were restricted to white Caucasian individuals. Japanese CD cases (484) and controls (933) are described elsewhere (32). HapMap DNA samples were purchased from Coriell Cell Repositories, Camden, NJ, USA.

### Sequencing of the IRGM promoter region

A 2.9-kb region of genomic DNA upstream of *IRGM* encompassing the entire 5′-UTR including the upstream exon 1 and intervening intron, ERV9 LTR and promoter was amplified in 94 unrelated individuals (including 43 CD individuals with known risk haplotype, 29 UC and 22 unaffected) using 8 pmol of each primer 5′-ACAATGAGTGTGTGAAACA GACCT-3′ and 5′-CATAGTGATGTTAACTGGTGTCCTG-3′, 1× PCR Master mix (Promega) and 25 ng of template genomic DNA in a 10 μl reaction. PCR conditions were as follows: 5 min at 95°C followed by 35 cycles of 30 s at 95°C, 30 s at 62°C and 3 min at 72°C with a final extension step of 10 min at 72°C. Subsequent ExoSAP-IT clean up (USB Europe, Staufen, Germany) followed by forward and reverse cycle sequencing was performed in ten independent

**Table 6.** Quantitative trait analysis of *IRGM* risk variants and expression in HapMap populations

| Locus | All African | YRI | MKK | LWK | CHB | JPT |
|---|---|---|---|---|---|---|
| CNV | $8.15 \times 10^{-6}$ | $5.41 \times 10^{-5}$ | 0.240 | 0.010 | 0.307 | 0.264 |
| rs13361189 | $3.74 \times 10^{-7}$ | $2.29 \times 10^{-6}$ | 0.033 | 0.016 | 0.797 | 0.349 |
| −1644 | 0.024 | 0.087 | 0.461 | 0.094 | 0.371 | 0.294 |
| −308 | $5.66 \times 10^{-5}$ | $4.38 \times 10^{-6}$ | 0.257 | 0.040 | 0.262 | 0.578 |

reactions using 8 pmol of each of the overlapping nested sequencing primers (Supplementary Material, Table S2) and 0.25 μl of ABI BigDye v3.1 (Applied Biosystems) in a 5 μl reaction volume and using standard conditions. Products were analysed on an ABI3730xl DNA sequencer (Sequence analysis, Applied Biosystems) and aligned to the published genomic sequence using the Sequencher 4.7 package (Gene-Codes).

### Fine-mapping of the CNV

Fine-mapping of the CNV was performed on a custom NimbleGen 385k array across a 300-kb interval encompassing the BAC on the WGTP array on which the CNV was first identified (22). The median spacing between probes was 45-bp. This custom array also targeted a number of other CNVs, which are not described here. The results confirmed that the CNV is a bi-allelic polymorphism that comprises either the presence or absence of 20 kb of sequence on chromosome 5q. The non-ancestral deletion spans from 150 183 354 to 150 203 455 on chromosome 5 (NCBI36).These data were subsequently used to design PCR primers, 5′-TTGCTGATGGCATGATCTTC-3′ and 5′-ATA TGGCGAGAGCAGCAACT-3′ for amplifying and sequencing the deletion breakpoint.

### Genotyping

The regions flanking the *IRGM* −1644 and −308 polymorphisms were amplified independently using 4 pmol of each of the following primer pairs 5′-AAATGGACCAATCAGCAGG A-3′ (5′ labelled with 6-FAM fluorescent dye): 5′-AGGGG CCAGGTATTTGAGAC-3′ and 5′-TGCCCACAGATACG ACAGAG-3′ (5′ labelled with HEX fluorescent dye): 5′-GG ACGCAGATATTGCAGTGA-3′, respectively. The reaction mix also included 1× PCR Mastermix (Promega) and 25 ng of genomic DNA in a 10 μl reaction volume. PCR conditions were as follows; 2 min at 95°C followed by 30 cycles of 20 s at 95°C, 30 s at 60°C and 30 s at 72°C, with a final extension step of 5 min at 72°C.

The CNV upstream of *IRGM* was genotyped via allele-specific PCR with a common forward primer and two allele specific reverse primers. The common forward primer (5′-AACAGTGACCTATCTGAAAAGGAAA-3′) was 5′ labelled with 6-FAM fluorescent dye and complementary to sequence immediately upstream of the copy number region. Of the two allele-specific reverse primers, the one complementary to sequence within the copy number region immediately adjacent to the forward primer (5′-TTGAAA TTTTGTAGAGATTGCATTG-3′) will only amplify if the 20 kb copy number variant sequence is present, and the other complementary to sequence immediately downstream of the copy number region (5′-TGCAGGGTACTGACTG TCCA-3′) will only amplify if the 20 kb copy number sequence is absent (deleted). The assay was validated by analysis of eight HapMap samples of known CNV status (22) (2 copies: NA07000, NA07348; 1 copy: NA11995, NA12874, NA18501; 0 copies: NA18545, NA18547, NA18555). All samples gave genotypes consistent with CGH data. PCR products for all three variants (CNV, −308, −1644) were diluted 1 in 50, pooled for each individual and separated via by capillary electrophoresis on the ABI3730xl Genetic Analyser with 10 μl of HiDi formamide and 0.125 μl of GS500LIZ size standard (both Applied Biosystems). SNPs rs10065172 (L105) was genotyped using validated Taqman assays (ABI), and allelic discrimination was carried out via endpoint read on ABI7900HT Sequence detection system. All genotypes at all variants were in Hardy–Weinberg equilibrium ($P > 0.01$).

### Quantitative analysis of IRGM expression

Lymphocytes from patients with CD, genotyped for all *IRGM* and upstream variants, were harvested from 40 ml of peripheral blood. Peripheral blood mononuclear cells were isolated by Lymphoprep (Axis-Shield, UK) and cultured at a density of 2 × 106 cells/ml in RPMI (Sigma Aldrich, UK) supplemented with 2 mM glutamine (Sigma Aldrich, UK) and 10% FCS (Sigma Aldrich, UK) in 24-well plates for 2 h at 37°C in a humidified atmosphere with 5% $CO_2$. After this time the non-adherent cell fraction (lymphocytes) were removed and washed twice in PBS. The cell pellet was then re-suspended in 0.5 ml RNAlater (Sigma Aldrich, UK), incubated for 24 h at 4°C and then stored at −80°C. Whole RNA was extracted from primary lymphocytes using the Ribopure kit (Ambion) and quantified using the Agilent Bioanalyser RNA 6000 Nano chip (Agilent Technologies UK Limited). cDNA synthesis was performed on 500 ng per sample of whole RNA using iScript cDNA Synthesis Kit (BIO RAD Laboratories, CA, USA). HapMap RNA was purified from cells purchased from Coriell Cell Repositories, Camden, NJ, USA.

*Allelic imbalance assay*. Sequencing of the exonic SNP rs10065172 in cDNA and genomic DNA (gDNA) samples was performed using standard procedure (see above) with exonic primers flanking the SNP (sequences available on request). Mean C and T peak heights in duplicate samples of cDNA and gDNA from eight CD individuals sequenced for SNP rs10065172 were estimated from sequence electropherograms by Sequence Scanner Software v1.0 (Applied Biosystems, Foster City, CA, USA). Comparison of the mean ratio of C:T peak heights in cDNA versus gDNA were calculated via the Wilcoxon signed rank test. A ratio of >1 indicates higher expression of the C allele (33).

*Real-time RT–PCR assays*. Quantitative fluorescent real time RT–PCR was carried out in triplicate on 1 μl of cDNA from primary lymphocyte samples of 24 CD patients using custom 6-FAM labelled fluorigenic Taqman MGB probe (5′-TG CCCACAGATACGAC-3′) and flanking primers 5′-CCCG CCTGATGAGCTTACTC-3′ 5′-AAGAGGTTAAGGATGCA GCTAATAGAG-3′ and a parallel reaction with a GAPDH

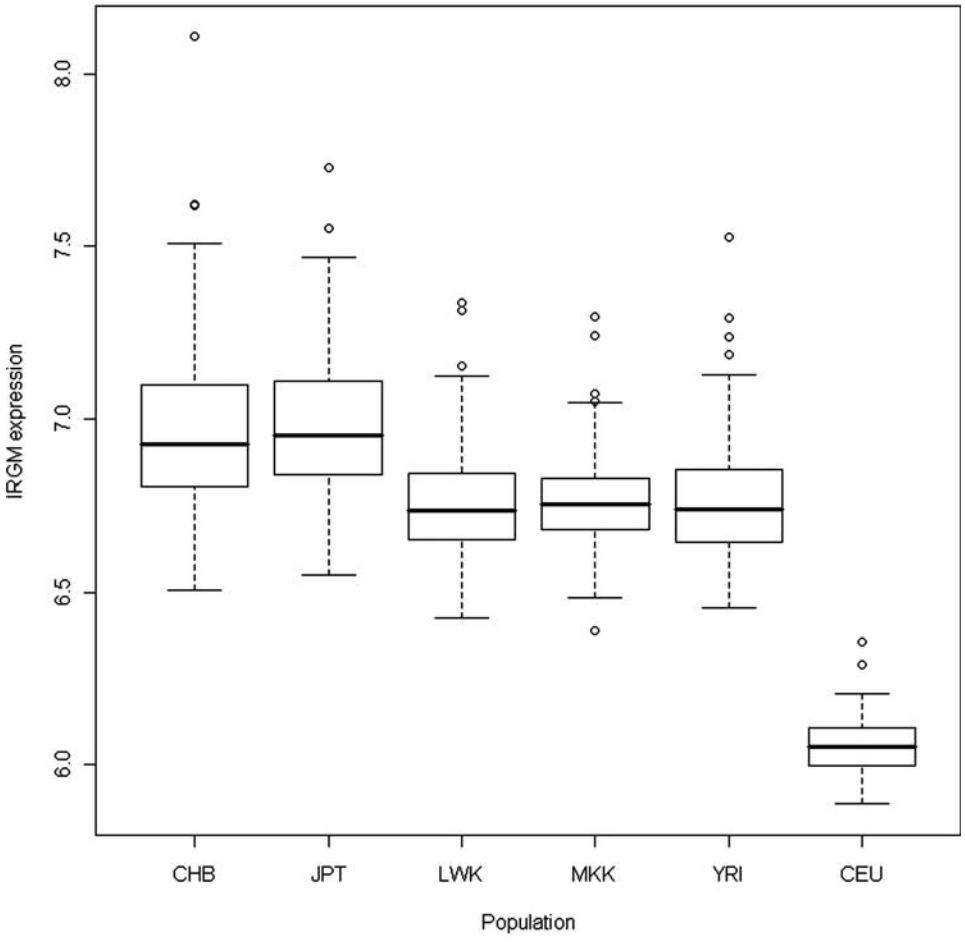


**Figure 6.** Relative expression of *IRGM* in HapMap populations from Asia (JPT & CHB), Africa (YRI, LWK & MKK) and Europe (CEU) by microarray analysis (26).

endogenous control (Eurogentech Ltd, Southampton, UK). The *IRGM* genotype of the 24 patients had been previously determined for each of the risk variants −1644 (11/8/5) −308 (9/7/5) and CNV (10/8/5). Real-time Quantitative-PCR was carried out on ABI7900HT system. Results were analysed via the ΔCt method for relative cDNA quantitation (http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_042380.pdf). Briefly, a threshold fluorescence level was selected at which PCR amplification of the target sequence was in the logarithmic phase and the cycle number at which each sample PCR reaction crossed that threshold was recorded (the threshold cycle or Ct). The relative quantity of cDNA for the target gene in each individual (ΔCt) was calculated as the difference between the mean Ct value for the target and the mean Ct value for the endogenous control and all were calibrated (normalised) with Ct values of cDNA from a low-level expressing placental sample.

Alternative splicing of *IRGM* mRNA was investigated by RT−PCR of cDNA from four individuals with three different genotypes for the *IRGM*-308 variant [$(GTTT)_5/(GTTT)_5$, $(GTTT)_5/(GTTT)_2$, $(GTTT)_2/(GTTT)_2$] using forward primer 5′-GTCTCAAATACCTGGCCCCT-3′ and reverse primer *IRGM*PROM_PCR_rev (Supplementary Material, Table S2). The identity of each cDNA species was confirmed by sequencing of gel-extracted product.

*Microarray expression analysis*. Expression of *IRGM* and *ZNF300* in HapMap3 RNA samples was analysed on Illumina human whole-genome expression arrays as previously described (26), but using Illumina WG-6 v2 arrays.

## Statistical analysis

Association analysis for qualitative (CD) and quantitative (*IRGM* expression) trait loci, including conditional regression analysis was performed using UNPHASED v3.0.12 (34), the latter assuming a full haplotype model. Haploview v4.1 (35) was used to calculate linkage disequilibrium coefficients ($r^2$). All other statistical analysis was performed using R v2.7.0 (www.r-project.org). Linear regression with repeated measures was used to analyse *IRGM* expression (as estimated by ΔCt values from Q-RT−PCR) with multiple replicates for each individual. The relationship between *IRGM* expression (from Illumina microarray data) and *IRGM* genotype in the different HapMap populations was also analysed using linear regression.

## SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

*Conflict of Interest statement*. None declared.

## FUNDING

This work was supported by the Wellcome Trust (081808/C.G.M.), the National Institutes of Health Research Biomedical Research Centres at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London and University College Hospital Trust with University College London; and the Guy's and St Thomas' Charity. We acknowledge use of the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. Funding to pay the Open Access charge was provided by the Wellcome Trust.

## REFERENCES

1. Mathew, C.G. (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat. Rev. Genet.*, **9**, 9–14.
2. The Wellcome Trust Case Control Consortium (2007) Genome-wide association studies of 14,000 cases of seven common human diseases and 3,000 shared controls. *Nature*, **447**, 661.
3. Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M., Anderson, C.A., Fisher, S.A., Roberts, R.G., Nimmo, E.R., Cummings, F.R., Soars, D. *et al.* (2007) Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.*, **39**, 830–832.
4. Bekpen, C., Hunn, J.P., Rohde, C., Parvanova, I., Guethlein, L., Dunn, D.M., Glowalla, E., Leptin, M. and Howard, J.C. (2005) The interferon-inducible p47 (IRG) GTPases in vertebrates: loss of the cell autonomous resistance mechanism in the human lineage. *Genome. Biol.*, **6**, R92.
5. Bekpen, C., Marques-Bonet, T., Alkan, C., Antonacci, F., Leogrande, M.B., Ventura, M., Kidd, J.M., Siswara, P., Howard, J.C. and Eichler, E.E. (2009) Death and resurrection of the human *IRGM* gene. *PLoS Genet.*, **5**, e1000403.
6. Singh, S.B., Davis, A.S., Taylor, G.A. and Deretic, V. (2006) Human *IRGM* induces autophagy to eliminate intracellular mycobacteria. *Science*, **313**, 1438–1441.
7. McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H. *et al.* (2008) Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat. Genet.*, **40**, 1107–1112.
8. Gou, D., Wang, J., Gao, L., Sun, Y., Peng, X., Huang, J. and Li, W. (2004) Identification and functional analysis of a novel human KRAB/C2H2 zinc finger gene *ZNF300*. *Biochim. Biophys. Acta.*, **1676**, 203–209.
9. Qiu, H., Xue, L., Gao, L., Shao, H., Wang, D., Guo, M. and Li, W. (2008) Identification of the DNA binding element of the human *ZNF300* protein. *Cell. Mol. Biol. Lett.*, **13**, 391–403.
10. Verstrepen, L., Carpentier, I., Verhelst, K. and Beyaert, R. (2009) ABINs: A20 binding inhibitors of NF-kappa B and apoptosis signaling. *Biochem. Pharmacol.*, **78**, 105–114.
11. Franke, A., Balschun, T., Karlsen, T.H., Sventoraityte, J., Nikolaus, S., Mayr, G., Domingues, F.S., Albrecht, M., Nothnagel, M., Ellinghaus, D. *et al.* (2008) Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.*, **40**, 1319–1323.
12. Roberts, R.L., Hollis-Moffatt, J.E., Gearry, R.B., Kennedy, M.A., Barclay, M.L. and Merriman, T.R. (2008) Confirmation of association of *IRGM* and NCF4 with ileal Crohn's disease in a population-based cohort. *Genes Immun.*, **9**, 561–565.
13. Weersma, R.K., Stokkers, P.C., Cleynen, I., Wolfkamp, S.C., Henckaerts, L., Schreiber, S., Dijkstra, G., Franke, A., Nolte, I.M., Rutgeerts, P. *et al.* (2009) Confirmation of multiple Crohn's disease susceptibility loci in a large Dutch–Belgian cohort. *Am. J. Gastroenterol.*, **104**, 630–638.
14. Van Limbergen, J., Russell, R.K., Nimmo, E.R., Drummond, H.E., G, D., Wilson, D.C. and Satsangi, J. (2009) Germline variants of *IRGM* in childhood-onset Crohn's disease. *Gut*, **58**, 610–611.
15. Palomino-Morales, R.J., Oliver, J., Gomez-Garcia, M., Lopez-Nevot, M.A., Rodrigo, L., Nieto, A., Alizadeh, B.Z. and Martin, J. (2009) Association of *ATG16L1* and *IRGM* gene polymorphisms with inflammatory bowel disease: a meta-analysis approach. *Genes Immun.*, **10**, 356–364.
16. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.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.
17. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P. and Hirschhorn, J.N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
18. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W. *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.
19. Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J. *et al.* (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat. Genet.*, **39**, 207–211.
20. Intemann, C.D., Thye, T., Sievertsen, J., Owusu-Dabo, E., Horstmann, R.D. and Meyer, C.G. (2009) Genotyping of *IRGM* tetranucleotide promoter oligorepeats by fluorescence resonance energy transfer. *Biotechniques*, **46**, 58–60.
21. de Smith, A.J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N.A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R.J., Bruhn, L. *et al.* (2007) Array CGH analysis of copy number variation identifies 1284 new gene variants in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.*, **16**, 2783–2794; Epub 2007 Jul 2731.
22. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
23. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
24. Prescott, N.J., Fisher, S.A., Dominy, K.M., Blaszczyk, K., Redon, R., Huang, N., Onnie, C.N., Lewis, C.M., Sanderson, J., Forbes, A. *et al.* (2008) Association of a promoter variant and copy number polymorphisms at the *IRGM* locus with Crohn's disease. *J. Med. Genet.*, **45**, S23.
25. Yamazaki, K., Takahashi, A., Takazoe, M., Kubo, M., Onouchi, Y., Fujino, A., Kamatani, N., Nakamura, Y. and Hata, A. (2009) Positive association of genetic variants in the upstream region of NKX2-3 with Crohn's disease in Japanese patients. *Gut*, **58**, 228–232.
26. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
27. Intemann, C.D., Thye, T., Niemann, S., Browne, E.N., Amanua Chinbuah, M., Enimil, A., Gyapong, J., Osei, I., Owusu-Dabo, E., Helm, S. *et al.* (2009) Autophagy gene variant *IRGM* -261T contributes to protection from tuberculosis caused by *Mycobacterium tuberculosis* but not by *M. africanum* strains. *PLoS Pathog.*, **5**, e1000577.
28. Croucher, P.J., Mascheretti, S., Hampe, J., Huse, K., Frenzel, H., Stoll, M., Lu, T., Nikolaus, S., Yang, S.K., Krawczak, M. *et al.* (2003) Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. *Eur. J. Hum. Genet.*, **11**, 6–16.
29. Yamazaki, K., Takazoe, M., Tanaka, T., Kazumori, T. and Nakamura, Y. (2002) Absence of mutation in the NOD2/CARD15 gene among 483 Japanese patients with Crohn's disease. *J. Hum. Genet.*, **47**, 469–472.
30. Prescott, N.J., Fisher, S.A., Franke, A., Hampe, J., Onnie, C.M., Soars, D., Bagnall, R., Mirza, M.M., Sanderson, J., Forbes, A. *et al.* (2007) A Nonsynonymous SNP in ATG16L1 Predisposes to Ileal Crohn's Disease

and is Independent of CARD15 and IBD5. *Gastroenterology*, **132**, 1665–1671.
31. Power, C. and Elliott, J. (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.*, **35**, 34–41.
32. 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.
33. Ge, B., Gurd, S., Gaudin, T., Dore, C., Lepage, P., Harmsen, E., Hudson, T.J. and Pastinen, T. (2005) Survey of allelic expression using EST mining. *Genome. Res.*, **15**, 1584–1591.
34. Dudbridge, F. (2003) Pedigree disequilibrium tests for multilocus haplotypes. *Genet. Epidemiol.*, **25**, 115–121.
35. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.