To adjust for an unknown genetic heterogeneity of our results, we also calculated λ-corrected P values (P_{GC})[19] for the GWAS results.

As a result of the GWAS, we identified 37 SNPs as significantly[20] associated with prostate cancer at $P < 1.0 \times 10^{-7}$, and these SNPs were located at eight independent loci. Among them, multiple loci on chromosome *8q24* (Block5/Region1: rs1447295 with $P = 6.0 \times 10^{-20}$ and rs7837688 with 1.2×10^{-25}; Block2/Region2: rs1016343 with $P = 2.7 \times 10^{-13}$, rs1456315 with 1.6×10^{-29} and rs16901966 with 1.9×10^{-12}), *3p12* (rs9284813, $P = 5.1 \times 10^{-9}$), *NKX3.1* (rs1512268, $P = 4.3 \times 10^{-11}$), *MSMB* (rs10993994, $P = 3.4 \times 10^{-8}$) and *HNF1B* (rs7501939, $P = 1.2 \times 10^{-12}$) have been reported[6–8,10,13,15] (**Supplementary Fig. 2e**). In addition, we identified new two loci for prostate cancer susceptibility on chromosome *5p15* (rs12653946, $P = 8.3 \times 10^{-10}$) and *6q22* (rs339331, $P = 6.0 \times 10^{-8}$) in our GWAS (**Supplementary Fig. 2e**).

To search for further susceptibility loci for prostate cancer, we conducted a replication study using 3,001 independent individuals with prostate cancer and 5,415 independent control subjects (**Supplementary Fig. 1**). Among the 263 SNPs that showed significant association ($P < 1.0 \times 10^{-4}$) in our GWAS, 80 SNPs were located at the previously reported loci and excluded from further analysis. We calculated the linkage disequilibrium (LD) coefficient (r^2) between the remaining 183 SNPs and selected the 91 SNPs with the lowest P value within each r^2 of ≥ 0.8 (**Supplementary Fig. 1**; maximum coefficient was 0.35). In the replication study, three (rs12653946, rs339331 and rs9600079 on *13q22*) out of the 91 SNPs were replicated with Bonferroni corrected P values $<5.5 \times 10^{-4}$ (**Supplementary Table 2**). When we combined both stages using the inverse method, these three SNPs were significantly associated with prostate cancer at $P_{GC} = 3.9 \times 10^{-18}$ – 2.8×10^{-9} (**Table 1**). Furthermore, although rs13385191 on *2p24* and rs1983891 on *6p21* were marginally replicated with P values $>5.5 \times 10^{-4}$ but <0.05, they showed association beyond the genome-wide significance threshold of $P < 1.0 \times 10^{-7}$ in the analysis of the combined sample of the two stages (**Table 1**; $P_{GC} = 7.5 \times 10^{-8}$ for rs13385191 and $P_{GC} = 7.6 \times 10^{-8}$ for rs1983891). In total, we identified five SNPs that have not previously been associated with prostate cancer.

We assessed the association between the genotypes of these five SNPs and family history of prostate cancer, Gleason score or cancer stage by a case-only analysis (**Supplementary Table 3**). Although the odds ratio of the men with prostate cancer who had a family history of the illness tended to be higher than that of those without family history in the GWAS, we found no evidence for an association of these SNPs with family history when the cases were subdivided by positivity of prostate cancer family history. Also, we found no significant association with Gleason score (Gleason score < 7 compared with ≥ 7) or cancer stage (non-advanced cancer compared with advanced cancer) in the case-only analysis. We found no difference in the per-allele odds ratio between individuals with prostate cancer when they were subdivided by age at diagnosis (**Supplementary Table 4**). Removing all female control subjects from the GWAS yielded essentially identical results for these five SNPs, as did excluding all individuals with other types of cancer from the control subjects in the GWAS (**Supplementary Table 5**). We also investigated whether the disease status of control subjects could affect the significance of these SNPs. There was no significant interaction between the genotype of these five SNPs and the disease status of the control subjects ($P > 0.05$, chi-square test with 25 degrees of freedom in the GWAS and with 9 degrees of freedom in the replication study).

To search for susceptibility gene(s) for prostate cancer within these five loci, we performed a further mapping analysis of these regions. We used Haploview[21] to select tagging SNPs to capture ($r^2 \geq 0.9$) all SNPs with minor allele frequencies (MAFs) of 5% or higher in the regions of interest based on Phase II of the HapMap Japanese JPT data (**Fig. 1**). The most significantly associated locus (rs12653946, $P_{GC} = 3.9 \times 10^{-18}$) was located on chromosome *5p15*, and further mapping analysis of the candidate region (Chr. 5: 1.90–1.99 Mb) using 36 tag-SNPs revealed that rs12653946 represented an associated region spanning 20 kb (**Fig. 1b**). There are no known genes within this region. The second locus (rs339331, $P_{GC} = 1.6 \times 10^{-12}$) on chromosome *6q22* was located in intron 4 of *RFX6* (encoding Regulatory factor X6*)*. Mapping analysis of a 580-kb region (Chr. 6: 117.0–117.6 Mb) using 59 tag-SNPs showed that rs339331 was located in a 200-kb associated region which included two genes, *RFX6* and *GPRC6A* (encoding G protein-coupled receptor, family C, group 6, member A; **Fig. 1d**). *RFX6* belongs to the RFX family of transcription factors and is expressed almost exclusively in the pancreatic islets[22]. By contrast,

Table 1 Summary results of newly identified loci associated with prostate cancer susceptibility

					MAF			Odds ratio[e]
SNP ID	Location[a]	Region[b]	Allele[c] (risk allele[d])	Study	Case	Control	*P*	(95% CI)
rs13385191	20751746	*2p24*	G/A	GWAS	0.406	0.455	1.1×10^{-5}	1.22 (1.12–1.33)
		C2orf43	G	Replication	0.407	0.439	7.4×10^{-4}	1.12 (1.05–1.19)
				Combined	0.407	0.442	7.5×10^{-8}	1.15 (1.10–1.21)
rs12653946	1948829	*5p15*	C/T	GWAS	0.500	0.434	2.8×10^{-9}	1.31 (1.20–1.42)
			T	Replication	0.498	0.454	1.2×10^{-10}	1.23 (1.16–1.31)
				Combined	0.499	0.443	3.9×10^{-18}	1.26 (1.20–1.33)
rs1983891	41644405	*6p21*	C/T	GWAS	0.452	0.402	4.3×10^{-6}	1.23 (1.13–1.34)
		FOXP4	T	Replication	0.440	0.414	1.5×10^{-3}	1.11 (1.04–1.18)
				Combined	0.444	0.410	7.6×10^{-8}	1.15 (1.09–1.21)
rs339331	117316745	*6q22*	T/C	GWAS	0.316	0.372	1.6×10^{-7}	1.28 (1.17–1.40)
		GPRC6A / RFX6	T	Replication	0.324	0.361	1.0×10^{-6}	1.18 (1.11–1.27)
				Combined	0.321	0.366	1.6×10^{-12}	1.22 (1.15–1.28)
rs9600079	72626140	*13q22*	G/T	GWAS	0.423	0.381	1.3×10^{-4}	1.19 (1.09–1.30)
			T	Replication	0.420	0.380	2.2×10^{-6}	1.17 (1.10–1.25)
				Combined	0.421	0.382	2.8×10^{-9}	1.18 (1.12–1.24)

Detailed data including genotype counts are shown in **Supplementary Table 2**. *P* values in GWAS indicate P_{GC}. Combined results were estimated by using the inverse method.
[a]Chromosomal location based on NCBI Human Genome Build 36 coordinates. [b]Relative to SNP position. SNPs are included in the region of a gene if they are located within 20 kb of its transcription start site or within 10 kb of its last exon. [c]Major/minor allele, based on the frequencies in the GWAS controls. [d]SNP allele that confers susceptibility to prostate cancer. [e]Allelic odds ratio (OR) with 95% CI.

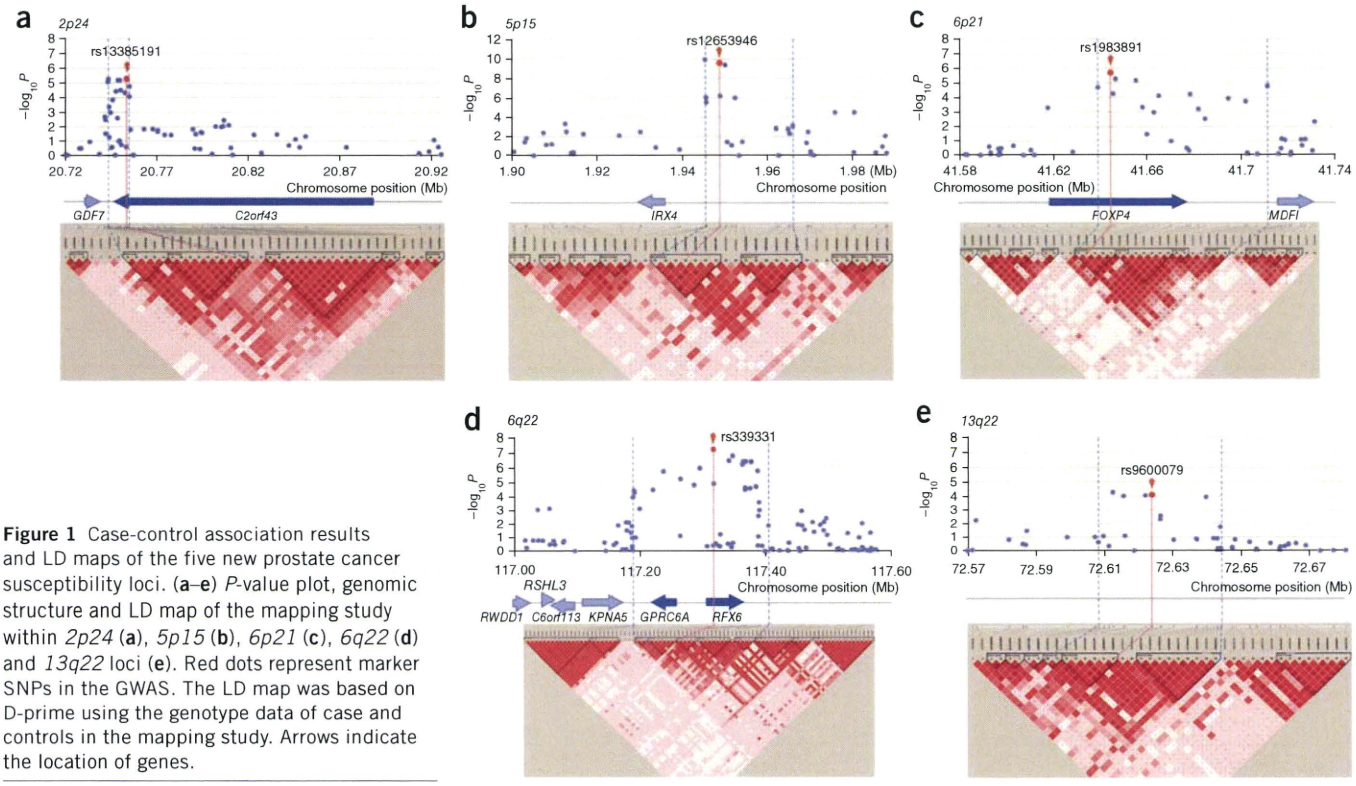

Figure 1 Case-control association results and LD maps of the five new prostate cancer susceptibility loci. (**a**–**e**) *P*-value plot, genomic structure and LD map of the mapping study within *2p24* (**a**), *5p15* (**b**), *6p21* (**c**), *6q22* (**d**) and *13q22* loci (**e**). Red dots represent marker SNPs in the GWAS. The LD map was based on D-prime using the genotype data of case and controls in the mapping study. Arrows indicate the location of genes.

GPRC6A encodes an orphan G-protein-coupled receptor that is highly expressed in the Leydig cells of the testis, and *Gprc6a*-null mice show male feminization and a metabolic syndrome that includes increased circulating levels of estradiol and reduced levels of testosterone[23]. These hormones are crucial for the initiation and progression of prostate cancer, and genetic variations at the *GPRC6A/RFX6* locus might affect susceptibility to prostate cancer by altering *GPRC6A*-mediated sexual hormone production. Although further investigation is required to determine the expression and function of these genes in the prostate, current annotation suggests that *GPRC6A* is the more plausible susceptibility gene in this locus. The third locus (rs9600079, $P_{GC} = 2.8 \times 10^{-9}$) was located on chromosome *13q22*. Mapping analysis of this region (Chr. 13: 72.57–72.68 Mb) using 36 tag-SNPs showed that rs9600079 was located in a 35-kb associated region (**Fig. 1e**) that contained no genes. The fourth locus (rs13385191, $P_{GC} = 7.5 \times 10^{-8}$) was located in intron 6 of *C2orf43* (encoding chromosome 2 open reading frame 43). Subsequent mapping of this region (Chr. 2: 20.72–20.93 Mb) with 45 tag-SNPs narrowed down the susceptibility region to a 12.3-kb associated area in the 3′ region of *C2orf43* (**Fig. 1a**). The fifth locus (rs1983891, $P_{GC} = 7.6 \times 10^{-8}$) was located in intron 2 of *FOXP4* (forkhead box P4). Mapping analysis of this region (Chr. 6: 41.58–41.74 Mb) identified a 72 kb prostate cancer susceptibility region (**Fig. 1c**), with *FOXP4* being the only gene in this region. *FOXP4* belongs to subfamily P of the FOX transcription factor family, which has key roles in embryonic development, cell cycle regulation and oncogenesis[24]. The functions of *C2orf43* and *FOXP4* are unknown and further studies will be required to clarify their functional associations with prostate carcinogenesis.

Of the previously reported loci in European GWAS (**Supplementary Table 6**), six loci showed association beyond the genome-wide significance threshold of $P < 1.0 \times 10^{-7}$ in our Japanese cohorts. Six variants on five independent loci showed association after multiple testing correction, with $P < 1.6 \times 10^{-3}$ (*THADA*, *8q24* Block 1, *8q24* Block 3/Region 3, *8q24* Block 4/Region 3 and *TTLL1/BIK*). *TET2* (rs7679673), one independent SNP on *8q24* Block 3 (rs16902094), *KLK2/KLK3* (rs2735839) and *NUDT10/NUDT11* (rs5945619) showed suggestive association ($1.6 \times 10^{-3} < P < 0.05$) with prostate cancer susceptibility in the Japanese population. The remaining 12 out of the 31 variants identified as susceptible loci for prostate cancer by European GWAS did not show any significant association ($P > 0.05$) in the Japanese population. When we calculated statistical power on the basis of MAF in the HapMap JPT data, the odds ratio reported in each study, the prevalence of prostate cancer in the Japanese population and the sample size in the GWAS, four of them can be explained by insufficient statistical power due to the difference in MAF among the populations and five loci can be explained by the difference in LD patterns among different populations. However, three loci, *ITGA6*, *11p15* and *17q24*, were not associated with Japanese prostate cancers even with a sufficient statistical power of >0.9 and a similar LD pattern between Japanese and European populations (**Supplementary Table 6**). Further detailed studies are needed to clarify the genetic difference in prostate cancer risk among different populations.

A recent GWAS for prostate cancer and follow-up studies using more than 10,000 samples of mainly European ancestry identified several new loci for prostate cancer susceptibility[12,13]. However, these world-wide meta-analyses and other previous GWAS for prostate cancer failed to detect the association of our five new loci even when using the same platform. The MAFs of these five SNPs are more than 0.2 in European populations and the HapMap database indicates that they are not markedly different from those in the Japanese population, so we speculate that the LD patterns between unknown causative variants and marker SNPs might be different between Japanese and European populations. It is also possible that unknown causative variants at these loci arose after the

European-Asian split, as is the case for the Crohn's disease susceptibility gene *NOD2* (also known as *CARD15*) (ref. 25).

In summary, we conducted a GWAS of prostate cancer in the Japanese population and identified five new prostate cancer susceptibility loci. Our study not only advances our understanding of the genetic basis of prostate cancer susceptibility but also highlights the genetic heterogeneity of prostate cancer susceptibility among various ethnic populations. Further studies on non-European or other ethnic populations are necessary for better understanding of the genetic etiology and heterogeneity of common complex diseases such as prostate cancer, as well as to promote effective clinical translation of prostate cancer risk assessment and improvement of screening protocols for prostate cancer.

METHODS

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

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

ACKNOWLEDGMENTS

We thank K. Ashikawa, H. Amitani and other staff from the Laboratory for Genotyping Development, Center for Genomic Medicine, RIKEN for their contribution to SNP genotyping, K. Boroevich for comments on the manuscript, and members of the BioBank Japan project and the Rotary Club of Osaka-Midosuji District 2660 Rotary International in Japan for supporting our study. This work was conducted as a part of the BioBank Japan Project, which was supported by the Ministry of Education, Culture, Sports, Sciences and Technology of the Japanese government.

AUTHOR CONTRIBUTIONS

Y.N. conceived the study; Y.N., R.T., S.A., M.K. and H.N. designed the study; R.T., S.A., N.H. and M.K. performed genotyping; R.T., S.A., M.K., Y.N. and H.N. wrote the manuscript; A.T., T.K., T.T. and N.K. performed data analysis at the genome-wide phase; Y.N., H.N. and M.K. managed DNA samples; R.T. and H.N. summarized the results; Y.N., O.O., T.F. and J.I. obtained funding for the study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Parkin, D.M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA Cancer J. Clin.* **55**, 74–108 (2005).
2. Matsuda, T. & Saika, K. Comparison of time trends in prostate cancer incidence (1973–2002) in Asia, from cancer incidence in five continents, Vols IV–IX. *Jpn. J. Clin. Oncol.* **39**, 468–469 (2009).
3. Kurahashi, N. *et al.* Dairy product, saturated fatty acid, and calcium intake and prostate cancer in a prospective cohort of Japanese men. *Cancer Epidemiol. Biomarkers Prev.* **17**, 930–937 (2008).
4. Schaid, D.J. The complex genetic epidemiology of prostate cancer. *Hum. Mol. Genet.* **13**, R103–R121 (2004).
5. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
6. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
7. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
8. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.* **39**, 977–983 (2007).
9. Gudmundsson, J. *et al.* Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.* **40**, 281–283 (2008).
10. Eeles, R.A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40**, 316–321 (2008).
11. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40**, 310–315 (2008).
12. Gudmundsson, J. *et al.* Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1022–1026 (2009).
13. Eeles, R.A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41**, 1116–1121 (2009).
14. Yeager, M. *et al.* Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **41**, 1055–1057 (2009).
15. Al Olama, A.A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1058–1060 (2009).
16. Witte, J.S. Prostate cancer genomics: towards a new understanding. *Nat. Rev. Genet.* **10**, 77–82 (2009).
17. Freedman, M.L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).
18. Yamaguchi-Kabata, Y. *et al.* Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am. J. Hum. Genet.* **83**, 445–456 (2008).
19. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
20. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
21. Barrett, J., Fry, B., Maller, J. & Daly, M. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
22. Aftab, S., Semenec, L., Chu, J.S. & Chen, N. Identification and characterization of novel human tissue-specific RFX transcriptional factors. *BMC Evol. Biol.* **8**, 226 (2008).
23. Pi, M. *et al.* GPRC6A null mice exhibit osteopenia, feminization and metabolic syndrome. *PLoS One* **3**, e3858 (2008).
24. Teufel, A., Wong, E.A., Mukhopadhyay, M., Malik, N. & Westphal, H. FoxP4, a novel forkhead transcription factor. *Biochim. Biophys. Acta* **1627**, 147–152 (2003).
25. Croucher, P.J. *et al.* Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. *Eur. J. Hum. Genet.* **11**, 6–16 (2003).

ONLINE METHODS

Samples. All case samples were obtained from the BioBank Japan at the Institute of Medical Science, the University of Tokyo[26]. This project was started in 2003 with the goal of collecting samples from a total of 300,000 patients who have at least one of 47 diseases by a collaborative network of 66 hospitals across Japan. The registration of cases was based on diagnoses made by physicians at the affiliated hospitals.

From the registered cases in BioBank Japan, we selected individuals for the GWAS and the replication study that were clinically diagnosed as having prostate cancer based on the pathological evaluation of prostatic biopsy. For the GWAS cases, we preferentially selected subjects who had a family history of prostate cancer or a high-grade Gleason score (**Supplementary Table 1**). Among 1,583 cases for the GWAS, 229 subjects had a family history of prostate cancer and 952 subjects had a Gleason score of 7 or more. We used genotype data from the GWAS for thirteen other diseases as controls (**Supplementary Table 1**). Control subjects in the GWAS consisted of 2,480 individuals that were registered in the BioBank Japan as subjects with thirteen diseases other than prostate cancer and 906 healthy volunteers recruited from the Osaka-Midosuji Rotary Club (MRC), Osaka, Japan. The controls in the replication study were male subjects whose samples were subjected to GWAS for five diseases (**Supplementary Table 1**). Detailed mapping of the five candidate loci was performed using GWAS samples. All participants provided written informed consent. This project was approved by the ethical committees at the Institute of Medical Science, the University of Tokyo and RIKEN Yokohama Institute.

SNP genotyping. For the GWAS, we genotyped 1,594 prostate cancer patients using Illumina Human610-Quad BeadChip and 3,386 control subjects using the Illumina HumanHap550v3 Genotyping BeadChip. After excluding 11 cases with a call rate of <0.98, we applied SNP quality control criteria (call rate of ≥ 0.99 in both cases and controls and *P* value of Hardy-Weinberg equilibrium test of ≥ 1.0×10^{-6} in controls). Among the common SNPs in both BeadChips, 497,172 autosomal SNPs and 13,515 SNPs on the X chromosome passed the quality control filters and were further analyzed. For the replication study, we selected 263 SNPs with a *P*-value of < 1.0×10^{-4} in the GWAS. Among them, 80 SNPs were located at previously reported loci[6–15] and excluded from further study. We calculated the LD coefficient (r^2) between the remaining SNPs, and selected the 91 SNPs with the lowest *P* value within each r^2 of ≥ 0.8. In the replication study, we genotyped an additional panel of 3,001 prostate cancer patients using the multiplex PCR-based Invader assay[27] (Third Wave Technologies). The concordance rate between genotypes determined by the Illumina Human610-Quad BeadChip and those determined by the Illumina HumanHap550v3 BeadChip among 182 duplicated samples was 0.99998. The concordance rate of the five SNPs selected for the replication study was 0.998 between the Illumina BeadChip and the multiplex PCR-based Invader assay. Tagging SNPs were selected by Haploview[21] to capture (r^2 of ≥ 0.9) all SNPs with MAFs of 5% or higher in the region of interest based on Phase II of the HapMap JPT data.

Statistical analysis. In all stages, associations of each autosomal SNP were assessed under an additive model, and also referenced dominant and recessive models. For the X chromosome, associations were assessed under a two-by-two chi-square test with male samples. To select SNPs for the replication study, we selected SNPs with a minimum *P* value <1.0×10^{-4} for at least one of the three models. Combined analysis of the GWAS and the replication study was conducted using the inverse method under the same genetic model. Heterogeneity among studies was examined by using the Breslow-Day test. For the mapping study, LD was defined according to published criteria[28] using Haploview software[21].

Software. For general statistical analysis, we used R statistical environment version 2.6.1 or PLINK1.03[29]. To draw the LD map, we used Haploview software[21].

URLs. PLINK1.03, http://pngu.mgh.harvard.edu/~purcell/plink/; R package haplo.stats, http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm; R package Epi, http://staff.pubhealth.ku.dk/~bxc/Epi/; Center for Cancer Control and Information Services, National Cancer Center, Japan, http://ganjoho.ncc.go.jp/professional/statistics/statistics.html.

[26] Nakamura, Y. The BioBank Japan project. *Clin. Adv. Hematol. Oncol.* **5**, 696–697 (2007).

[27] Ohnishi, Y. *et al.* A high-throughput SNP typing system for genome-wide association studies. *J. Hum. Genet.* **46**, 471–477 (2001).

[28] Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).

[29] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

Journal of Human Genetics (2010), 1–4
© 2010 The Japan Society of Human Genetics All rights reserved 1434-5161/10 $32.00
www.nature.com/jhg

SHORT COMMUNICATION

Association study of the polymorphisms on chromosome 12p13 with atherothrombotic stroke in the Japanese population

Tomonaga Matsushita[1,2], Junji Umeno[1,2], Yoichiro Hirakawa[3], Koji Yonemoto[3], Kyota Ashikawa[1], Hanae Amitani[1], Toshiharu Ninomiya[3], Jun Hata[3], Yasufumi Doi[3], Takanari Kitazono[2], Mitsuo Iida[2], Yusuke Nakamura[4], Yutaka Kiyohara[3] and Michiaki Kubo[1,2,3]

Recent genome-wide association study using four prospective population-based cohorts identified two single-nucleotide polymorphisms (SNPs) on chromosome 12p13, rs12425791 and rs11833579, to be significantly associated with the incidence of atherothrombotic stroke. To examine the association of these SNPs with atherothrombotic stroke in the Japanese population, we carried out a case–control association study using a total of 3784 cases and 3102 controls. We also examined the effect of these SNPs on the subtypes of ischemic stroke. Association analysis was carried out using logistic regression model after adjustment of age, sex and cardiovascular risk factors. Rs12425791 was significantly associated with atherothrombotic stroke (*P*=0.0084, odds ratio (OR)=1.15). When we analyzed effects of rs12425791 on ischemic stroke subtypes, rs12425791 was significantly associated with both small-artery occlusion (*P*=0.015, OR=1.15) and large-artery atherosclerosis (*P*=0.024, OR=1.19). Rs11833579 showed no association with atherothrombotic stroke or its subtypes in our population. Our data suggest that rs12425791 on chromosome 12p13 is a genetic marker for atherothrombotic stroke in multiethnic population.

Journal of Human Genetics advance online publication, 7 May 2010; doi:10.1038/jhg.2010.45

Keywords: atherothrombotic stroke; replication study; SNP

INTRODUCTION

Genome-wide association study (GWAS) has emerged as a powerful new approach to identify many susceptibility variants with moderate genetic risk on various common diseases, such as diabetes[1] and coronary heart diseases.[2,3] As for ischemic stroke, few GWASs have been carried out and genetic components of common forms of ischemic stroke are still largely unknown. Recently, the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium reported two single-nucleotide polymorphisms (SNPs), rs11833579 and rs12425791, to be significantly associated with the incidence of ischemic stroke in a GWAS of four population-based cohorts, which included 19 602 white persons with an average of 11 years of follow-up data.[4] These SNPs were located in close proximity to ninjurin 2 (*NINJ2*) gene on chromosome 12p13. Both SNPs showed genome-wide significance, however, only rs12425791 was replicated in both the African-American cohort and the white case–control sample.[4] Although this study has an advantage of a prospective study design, these SNPs were merely the marker and the true causative variant(s) have not been identified yet. Moreover, this study did not analyze the effects of these SNPs on ischemic stroke subtypes probably because of small number of events.

As the association of these SNPs in Asian population remains unknown, we examined the association of these SNPs with atherothrombotic stroke using two Japanese case–control sets with a sufficient sample size. We also examined the effect of these SNPs on the subtypes of ischemic stroke.

MATERIALS AND METHODS

We used two independent Japanese case–control sets for this study. One case–control set (set-1) is consisted of 860 cases of atherothrombotic stroke and 860 age- and sex-matched controls. Details of the registration and case ascertainment were previously described.[5] We selected 860 cases of atherothrombotic stroke on the basis of the classification as in the CHARGE study[4] and subdivided them into 491 small-artery occlusion (SAO) and 369 large-artery atherosclerosis (LAA) according to the TOAST criteria.[6] Age- (within 5 years) and sex-matched controls were selected from the 3196 participants of the

[1]Laboratory for Genotyping Development, Center for Genomic Medicine, RIKEN Yokohama Institute, Kanagawa, Japan; [2]Department of Clinical Sciences, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan; [3]Department of Environmental Medicine, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan and [4]Laboratory for Molecular Medicine, Human Genome Center, The Institute of Medical Science, University of Tokyo, Tokyo, Japan
Correspondence: Dr T Matsushita, Laboratory for Genotyping Development, Center for Genomic Medicine, RIKEN, 1-7-22, Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan.
E-mail: matsushita@src.riken.jp

Received 7 January 2010; revised 25 March 2010; accepted 29 March 2010

Hisayama screening survey between 2002 and 2003. Another case–control set (set-2) consisted of 2924 atherothrombotic stroke and 2242 controls. Cases were selected from the BioBank Japan Project[7] based on the similar criteria as in the set-1 cases. These cases were classified into 2256 SAO and 668 LAA. The Hisayama participants who did not have ischemic stroke and did not enrolled in the set-1 were used as controls in the set-2. Clinical characteristics of the study populations are shown in Table 1.

Written informed consent was obtained from all study subjects, and this study was approved by the ethics committees of the Graduate School of Medical Sciences, Kyushu University and RIKEN Yokohama Institute.

We genotyped SNPs using the multiplex PCR-based Invader assay (Third Wave Technologies, Madison, WI, USA).[8] Crude association analysis was carried out using χ^2-test under allele model. We also assessed the association after adjustment of age, sex, body mass index, hypertension (yes/no), diabetes (yes/no) and dyslipidemia (yes/no) using logistic regression analysis under additive model. In a combined analysis, pooled estimates of the odds ratio (OR) for two case–control sets were obtained using inverse-variance-weighting analysis.[4] Heterogeneities across the population were estimated formally using Cochran's *Q*-test.[9]

RESULTS

We carried out association analysis for atherothrombotic stroke using two case–control sets (Table 2). In the crude analysis, we found a weak association of rs12425791 with atherothrombotic stroke in the combined sample (*P*=0.041), although each case–control set did not show significant association. This association became stronger after adjusted for various cardiovascular risk factors (*P*=0.0084, OR=1.15, 95% confidence interval=1.04–1.27). In contrast, rs11833579 showed no association with atherothrombotic stroke even in the combined sample set (*P*=0.58).

When we examined these associations by ischemic stroke subtypes, rs12425791 showed no association with SAO (*P*=0.072) or LAA (*P*=0.13) in the crude analysis. However, after adjustment of cardiovascular risk factors, rs12425791 was significantly associated with SAO (*P*=0.015, OR=1.15, 95% confidence interval=1.03–1.28) and LAA (*P*=0.024, OR=1.19, 95% confidence interval=1.02–1.39) in the combined sample set (Table 3). We found no significant association of rs11833579 with either SAO or LAA.

We also carried out the association analysis stratified by sex. After adjustment of cardiovascular risk factors, rs12425791 did not show significant association with atherothrombotic stroke in men (*P*=0.086, OR=1.14), whereas it showed a weak association in women (*P*=0.027, OR=1.17). When we examined these associations by ischemic stroke subtypes, rs12425791 was associated with SAO (*P*=0.022, OR=1.19), but not with LAA (*P*=0.080, OR=1.25), in women. Rs12425791 did not show any association with SAO (*P*=0.19, OR=1.11) or LAA (*P*=0.075, OR=1.20) in men. Rs11833579 showed no association with

Table 1 Clinical characteristics of the study population

	Set-1		*Set-2*	
	Case	*Control*	*Case*	*Control*
N	860	860	2924	2242
Male sex (%)	60.7	60.7	64.0	36.3
Age (years)	70.3±9.9	70.2±10.0	69.1±9.2	58.2±11.7
Body mass index (kg m^{-2})	22.0±3.9	22.7±3.3	23.5±3.4	23.2±3.4
Ischemic stroke subtype				
Small-artery occlusion	491		2256	
Large-artery atherosclerosis	369		668	
Cardiovascular risk factors				
Hypertension (%)	82.6	53.8	72.2	39.8
Diabetes mellitus (%)	32.8	20.6	15.6	16.2
Dyslipidemia (%)	50.9	41.1	21.9	47.5

Data are shown in mean±s.d. or percentage except for ischemic stroke subtypes.

Table 2 Association between the SNPs reported in the CHARGE study and atherothrombotic stroke among Japanese

		Case					*Control*					*Crude*			*Adjusted*		
SNP (allele 1/2)	*Set*	*11*	*12*	*22*	*Total*	*MAF*	*11*	*12*	*22*	*Total*	*MAF*	P-*value*	*OR*	*(95% CI)*	P-*value*	*OR*	*(95% CI)*
rs12425791 (G/A)	Set-1	342	419	93	854	0.35	392	360	107	859	0.33	0.22	1.09	(0.95–1.26)	0.69	1.07	(0.76–1.51)
	Set-2	1200	1353	361	2914	0.36	976	999	262	2237	0.34	0.099	1.07	(0.99–1.16)	0.0084	1.15	(1.04–1.28)
	Combined											0.041	1.08	(1.00–1.16)	0.0084	1.15	(1.04–1.27)
rs11833579 (G/A)	Set-1	264	455	136	855	0.43	292	422	146	860	0.42	0.55	1.04	(0.91–1.19)	0.98	1.00	(0.71–1.40)
	Set-2	942	1469	507	2918	0.43	749	1082	403	2234	0.42	0.77	1.01	(0.94–1.09)	0.58	1.03	(0.93–1.14)
	Combined											0.58	1.02	(0.95–1.09)	0.60	1.03	(0.93–1.13)

Abbreviations: CHARGE study, the Cohorts for Heart and Aging Research in Genomic Epidemiology study; CI, confidence interval; MAF, minor allele frequency; OR, odds ratio; SNP, single-nucleotide polymorphism.
Alleles for the SNPs on the forward strand of the human genome reference sequence (NCBI build 36.3) are shown. Crude analysis was carried out using χ^2-test under allele model. Adjusted analysis was carried out using logistic regression model after adjustment of cardiovascular risk factors.

Table 3 Association between the SNPs reported in the CHARGE study and the subtypes of ischemic stroke among Japanese

			Case					*Control*					*Crude*			*Adjusted*		
Subtype	*SNP (allele 1/2)*	*Set*	*11*	*12*	*22*	*Total*	*MAF*	*11*	*12*	*22*	*Total*	*MAF*	P-*value*	*OR*	*(95% CI)*	P-*value*	*OR*	*(95% CI)*
SAO																		
	rs12425791 (G/A)	Set-1	197	238	54	489	0.35	230	204	56	490	0.32	0.14	1.15	(0.95–1.39)	0.58	1.13	(0.74–1.72)
		Set-2	931	1046	272	2249	0.35	976	999	262	2237	0.34	0.19	1.06	(0.97–1.16)	0.017	1.15	(1.02–1.28)
		Combined											0.072	1.07	(0.99–1.16)	0.015	1.15	(1.03–1.28)
	rs11833579 (G/A)	Set-1	153	256	80	489	0.43	162	252	77	491	0.41	0.59	1.05	(0.88–1.26)	0.77	0.94	(0.61–1.44)
		Set-2	728	1142	380	2250	0.42	749	1082	403	2234	0.42	0.99	1.00	(0.92–1.09)	0.73	1.02	(0.91–1.14)
		Combined											0.81	1.01	(0.94–1.09)	0.79	1.01	(0.91–1.13)
LAA																		
	rs12425791 (G/A)	Set-1	145	181	39	365	0.35	162	156	51	369	0.35	0.83	1.02	(0.83–1.27)	0.97	0.99	(0.54–1.82)
		Set-2	269	307	89	665	0.36	976	999	262	2237	0.34	0.10	1.11	(0.98–1.26)	0.019	1.21	(1.03–1.42)
		Combined											0.13	1.09	(0.98–1.21)	0.024	1.19	(1.02–1.39)
	rs11833579 (G/A)	Set-1	111	199	56	366	0.42	130	170	69	369	0.42	0.77	1.03	(0.84–1.27)	0.80	1.08	(0.60–1.93)
		Set-2	214	327	127	668	0.43	749	1082	403	2234	0.42	0.42	1.05	(0.93–1.19)	0.26	1.09	(0.94–1.27)
		Combined											0.40	1.05	(0.94–1.16)	0.25	1.09	(0.94–1.27)

Abbreviations: CHARGE study, the Cohorts for Heart and Aging Research in Genomic Epidemiology study; CI, confidence interval; MAF, minor allele frequency; OR, odds ratio; SAO, small-artery occlusion; LAA, large-artery atherosclerosis; SNP, single-nucleotide polymorphism.
Alleles for the SNPs on the forward strand of the human genome reference sequence (NCBI build 36.3) are shown. Crude analysis was carried out using χ^2-test under allele model.
Adjusted analysis was carried out using logistic regression model after adjustment of cardiovascular risk factors.

atherothrombotic stroke or ischemic stroke subtypes in both sexes (data not shown).

DISCUSSION

Using two independent Japanese case–control sets, we examined the association of two SNPs on chromosome 12p13 recently identified by a Caucasian GWAS of stroke. Rs12425791 was significantly associated with atherothrombotic stroke, whereas rs11833579 was not. Rs12425791 was also associated with both SAO and LAA, and its effect on the risk of SAO and LAA were similar. These results suggest that rs12425791 is a genetic marker for the incidence of atherothrombotic stroke in multiethnic populations including Japanese and might equally affect the risk of both SAO and LAA.

Similar ORs of rs12425791 on both SAO and LAA indicate that this SNP may be a marker for common pathogenesis of both ischemic stroke subtypes, probably for atherosclerosis. Rs12425791 is located at ~10 kb proximal from the 5′ untranslated region of the *NINJ2* gene. On the basis of the Hapmap JPT data, rs12425791 is linked to the promoter region of *NINJ2*. Although fine mapping is needed, SNPs linked to rs12425791 might regulate the expression level of *NINJ2*. Ninjurin2, a gene product of *NINJ2*, is a cell surface adhesion molecule and is highly expressed in the bone marrow, peripheral leukocyte, lung and lymph node in human.[10] Although ninjurin2 is reported to be upregulated after nerve injury in Schwann cells and promotes neurite outgrowth,[10] the function of ninjurin2 on the ischemic stroke is largely unknown. Further functional studies are needed to clarify this issue.

Assuming the sample size of our study population using the allele frequencies in our controls and the hazard ratios in the CHARGE study, the statistical power to detect the associations at a significance level of 0.05 would be >99% for both SNPs. However, we found a significant association of atherothrombotic stroke only in rs12425791. Similarly, the CHARGE consortium showed that the association of ischemic stroke for rs12425791 was replicated in the African-American cohort, but the association for rs11833579 was not significant. This might be due to the difference in the linkage disequilibrium between the two SNPs and true causative variant among different populations. On the basis of the Hapmap data, linkage disequilibrium between the two SNPs was different among populations (r^2=0.69 for JPT, r^2=0.34 for YRI and r^2=0.75 for CEU). There is a possibility that rs12425791 is closely linked to the true causative variant of atherothrombotic stroke among different populations. In contrast, the linkage disequilibrium between true causative variant and rs11833579 will be strong in Caucasian population, but it may be weak in Japanese and African-American populations.

The association between rs12425791 and ischemic stroke in this study was much weaker than that in the CHARGE study. The relationship of rs12425791 with atherothrombotic stroke or stroke subtypes was observed in the case–control set-2 but not in the case–control set-1. Furthermore, the relationship of rs12425791 with atherothrombotic stroke or stroke subtypes in the set-2 was not detected by the χ^2-test of allele frequencies. These results suggest that the impact of rs12425791 to atherothrombotic stroke or stroke subtypes in Japanese individuals is relatively low as compared with Caucasian population. Another possible explanation is that the effect size obtained from GWAS overestimates the true effect (winner's course). Indeed, CHARGE study showed that the genetic risk of rs12425791 in the replication study is lower than that in GWAS: in the GWAS, rs12425791 showed the strong association with atherothrombotic stroke (P=3.3×10^{-8}, hazard ratio=1.37), whereas it showed the P-value of 0.0052 and OR of 1.29 in the Dutch case–control study using 652 cases and 3613 controls.

In conclusion, our study suggests that rs12425791 or linked variations would be the true causative variant(s) for the genetic risk of atherothrombotic stroke in multiethnic population.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the residents of Hisayama town and the patients with ischemic stroke for their participation; T Omae and the staff of the Division of Health and Welfare of Hisayama for their cooperation; many members of the Hisayama study for assistance; T Ago, H Ooboshi, M Kamouchi, H Sugimori, J Kuroda, Y Kumai, N Hagiwara, S Yoshimura (Kyushu University Hospital), K Tamaki, Y Wakugawa (Hakujyuji Hospital), K Fujii (Fukuoka Red Cross Hospital), Y Okada, K Toyoda (National Hospital Organization, Kyushu Medical Center), T Nagao (Imazu Red Cross Hospital), H Nakane (National Hospital Organization, Fukuoka Higashi Medical Center), S Ibayashi, Y Yamashita, K Kusuda (Seiai Rehabilitation Hospital) for clinical sample collection. We thank all the patients who participated in BioBank Japan project. We also thank all members of BioBank Japan, Institute of Medical Science, the University of Tokyo, and of the Center for Genomic Medicine, RIKEN Yokohama Institute, for their contribution to the completion of our study. This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

1 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).

2 Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* **41,** 334–341 (2009).

3 McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316,** 1488-1491 (2007).

4 Ikram, M. A., Seshadri, S., Bis, J. C., Fornage, M., DeStefano, A. L., Aulchenko, Y. S. *et al.* Genomewide association studies of stroke. *N. Engl. J. Med.* **360,** 1718–1728 (2009).

5 Kubo, M., Hata, J., Ninomiya, T., Matsuda, K., Yonemoto, K., Nakano, T. *et al.* A nonsynonymous SNP in PRKCH (protein kinase Cη) increases the risk of cerebral infarction. *Nat. Genet.* **39,** 212–217 (2007).

6 Adams, H. P. Jr., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L. *et al.* Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* **24,** 35–41 (1993).

7 Nakamura, Y. The BioBank Japan Project. *Clin. Adv. Hematol. Oncol.* **5,** 696–697 (2007).

8 Ohnishi, Y., Tanaka, T., Ozaki, K., Yamada, R., Suzuki, H. & Nakamura, Y. A high-throughput SNP typing system for genome-wide association studies. *J. Hum. Genet.* **46,** 471–478 (2001).

9 Li, D., Collier, D. A. & He, L. Meta-analysis shows strong positive association of the neuregulin 1 (NRG1) gene with schizophrenia. *Hum. Mol. Genet.* **15,** 1995–2002 (2006).

10 Araki, T. & Milbrandt, J. Ninjurin2, a novel homophilic adhesion molecule, is expressed in mature sensory and enteric neurons and promotes neurite outgrowth. *J. Neurosci.* **20,** 187–195 (2000).

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 1.40×10^{-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

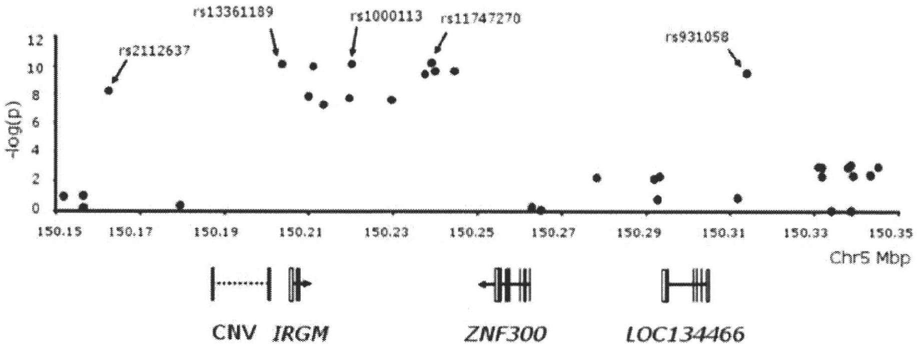

Figure 1. Association of 35 SNPs at the *IRGM-ZNF300* locus with CD (data from Ref. 16) indicated by –log of *P*-value. The physical location of the SNPs on chromosome 5 in Mbp is given on the *x*-axis with the relative position of the genes and CNV underneath.

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 3.02×10^{-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 7.5×10^{-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 3.7×10^{-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.

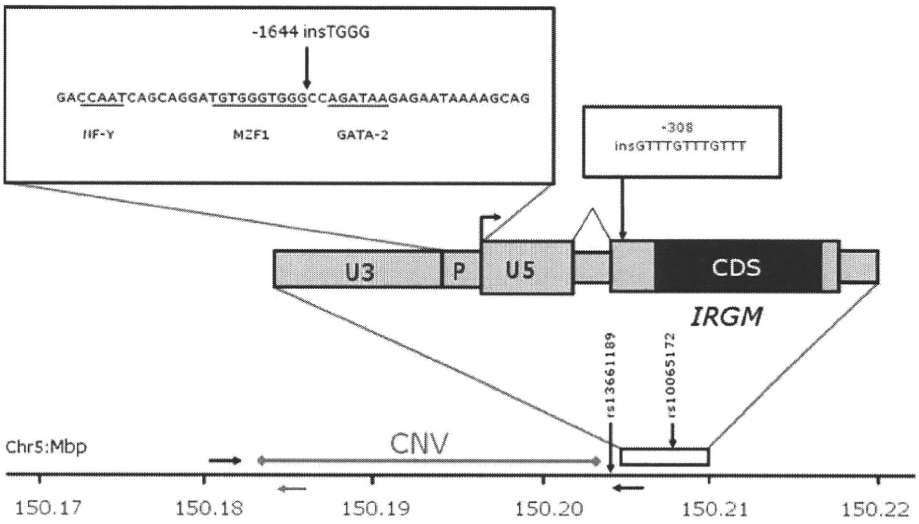

Figure 2. The structure of *IRGM* showing the location of the upstream CNV on chromosome 5. Horizontal arrows represent the relative positions of the three primers for the CNV PCR assay which uses the common left primer (black), and either the insertion right primer (grey) or deletion right (black). 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 2.6×10^{-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 2.38×10^{-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 3.9×10^{-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	1.40×10^{-9}	1.9×10^{-9}	1.66 (1.41–1.95)
rs13361189	C	0.115	0.075	3.73×10^{-9}	5.10×10^{-9}	1.65 (1.40–1.95)
−1644	Ins	0.108	0.073	4.20×10^{-7}	4.20×10^{-7}	1.53 (1.38–1.80)
−308	$(GTTT)_5$	0.149	0.111	1.37×10^{-5}	2.77×10^{-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	1.04×10^{-7}	1.01×10^{-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.

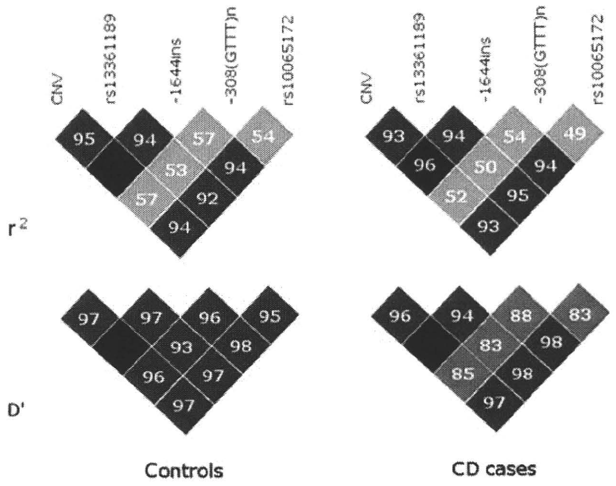

Figure 3. Linkage disequilibrium of *IRGM* variants in UK CD cases and controls.

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	1.6×10^{-5}	2.6×10^{-7}	0.251
	rs13361189	0.361	1.0	0.005	3.88×10^{-6}	0.388
	−1644	3.1×10^{-4}	0.142	1.0	9.3×10^{-6}	0.047
	−308	9.4×10^{-6}	2.38×10^{-4}	5.8×10^{-5}	1.0	1.50×10^{-5}
	rs10065172	0.392	0.49	1.58×10^{-3}	1.12×10^{-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)	6.18×10^{-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)	4.76×10^{-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]	5.20×10^{-3}
11	*DEL*	*C*	DEL	*(GTTT)*$_4$	*T*	4	0	0.002	0	n/a[b]	1.55×10^{-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 ≫ 4×10^6).

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	2.6×10^{-7}	1
	rs13361189	1	1	1	3.88×10^{-6}	1
	−1644	1	1	1	9.3×10^{-6}	1
	−308	2.0×10^{-5}	9.76×10^{-4}	3.9×10^{-4}	1	4.24×10^{-5}
	rs10065172	1	1	1	1.12×10^{-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).

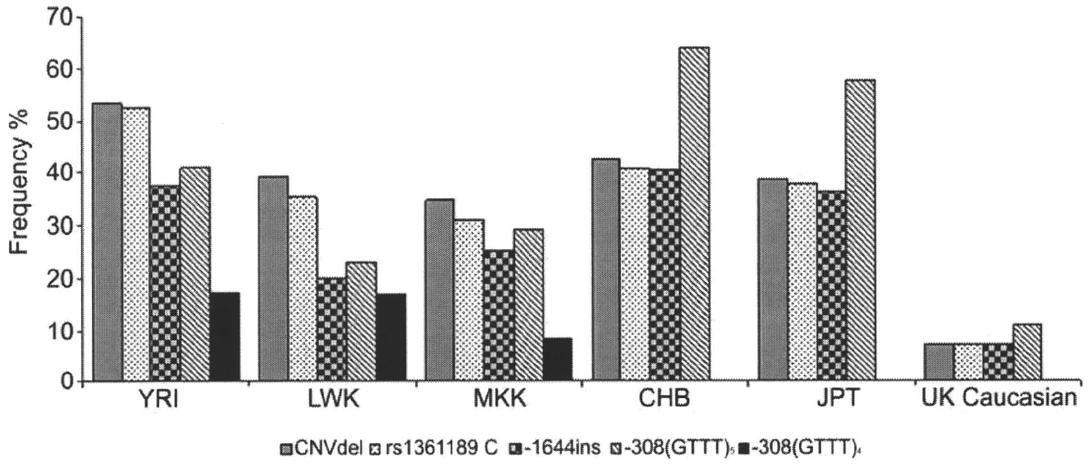

Figure 4. Frequency of *IRGM* variants in five HapMap populations from Asia and Africa (YRI 120, MKK 145, LWK 145, JPT 88, CHB 88) and 2025 white population controls from the 1958 MRC British birth cohort (BC1958).

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		P_{allele}	OR (95%CI)
		Cases	Controls		
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	$(GTTT)_5^a$	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.

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.

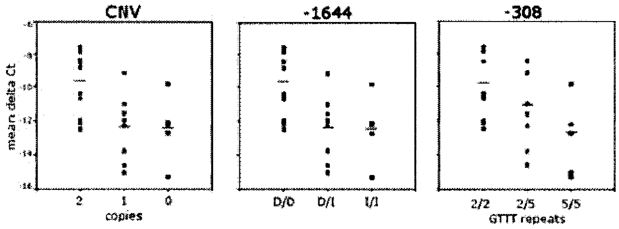

Figure 5. Quantitative analysis of *IRGM* expression by real-time PCR in lymphocytes from CD patients of defined genotype for CNV (2, 1 and 0 copies of the CNV where 1 and 0 are the heterozygous and homozygous risk genotypes, respectively) −1644 (D = del allele, I = risk insertion allele) and −308 (where 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).

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′-ACAATGAGTGTGTGAAACAGACCT-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	8.15×10^{-6}	5.41×10^{-5}	0.240	0.010	0.307	0.264
rs13361189	3.74×10^{-7}	2.29×10^{-6}	0.033	0.016	0.797	0.349
−1644	0.024	0.087	0.461	0.094	0.371	0.294
−308	5.66×10^{-5}	4.38×10^{-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′-ATATGGCGAGAGCAGCAACT-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′-AAATGGACCAATCAGCAGGA-3′ (5′ labelled with 6-FAM fluorescent dye): 5′-AGGGGCCAGGTATTTGAGAC-3′ and 5′-TGCCCACAGATACGACAGAG-3′ (5′ labelled with HEX fluorescent dye): 5′-GGACGCAGATATTGCAGTGA-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′-TTGAAATTTTGTAGAGATTGCATTG-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′-TGCAGGGTACTGACTGTCCA-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′-TGCCCACAGATACGAC-3′) and flanking primers 5′-CCCGCCTGATGAGCTTACTC-3′ 5′-AAGAGGTTAAGGATGCAGCTAATAGAG-3′ and a parallel reaction with a GAPDH

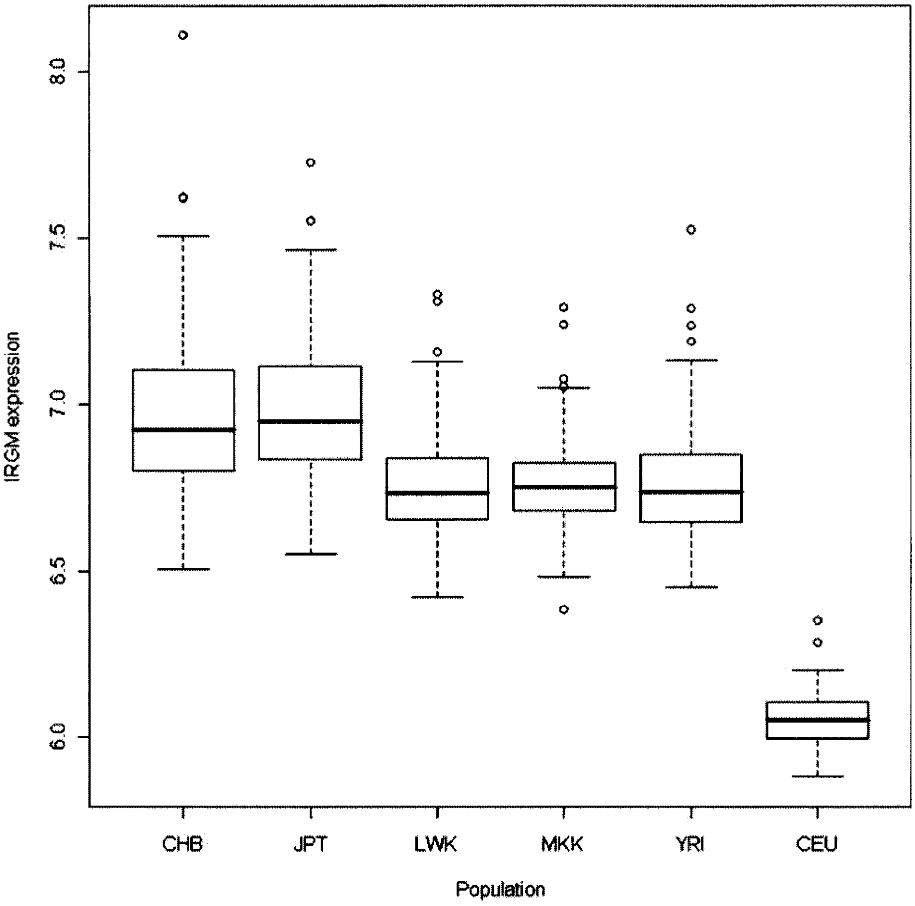

Figure 6. Relative expression of *IRGM* in HapMap populations from Asia (JPT & CHB), Africa (YRI, LWK & MKK) and Europe (CEU) by microarray analysis (26).

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.