Sciencexpress Report

Rare Variants of *IFIH1*, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes

Sergey Nejentsev,[1,2]* Neil Walker,[1] David Riches,[3] Michael Egholm,[3] John A. Todd[1]

[1]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, CB2 0XY, UK. [2]Department of Medicine, University of Cambridge, CB2 2QQ, UK. [3]454 Life Sciences, A Roche Company, Branford, CT, 06405, USA.

*To whom the correspondence should be addressed. E-mail: sn262@cam.ac.uk

Genome-wide association studies (GWAS) are widely used to map genomic regions contributing to common human diseases but they often do not identify the precise causative genes and sequence variants. To identify causative type 1 diabetes (T1D) variants we re-sequenced exons and splice sites of ten candidate genes in pools of DNA from 480 patients and 480 controls and tested their disease association in over 30,000 subjects. We discovered four rare variants that lowered T1D risk independently of each other (OR = 0.51 – 0.74; $P = 1.3 \times 10^{-3} – 2.1 \times 10^{-16}$) in *IFIH1*, a gene located in a region previously associated with T1D by GWAS. These variants are predicted to alter the expression and structure of IFIH1 (MDA5), a cytoplasmic helicase that mediates induction of interferon response to viral RNA. This firmly establishes the role of *IFIH1* in T1D and demonstrates that re-sequencing studies can pinpoint disease-causing genes in genomic regions initially identified by GWAS.

Genome-wide association studies (GWAS) of common multifactorial diseases have identified dozens of loci harboring disease-causing sequence variants (*1*, *2*). However, because the human genome contains regions of strong linkage disequilibrium, a disease-associated locus sometimes encompasses several genes and multiple tightly associated polymorphisms, making it difficult to pinpoint the causal variant by association mapping. Moreover, in many instances, the single nucleotide polymorphisms (SNPs) showing the most significant disease association map to genomic regions with no obvious function, thus providing few clues as to how causal variants affect the disease gene.

One way to overcome this limitation is to search for sequence variants that are rare in the population (frequency < 3%) but that reside in exons and other genomic regions of known function to identify polymorphisms that likely alter expression of the gene and/or the function of the protein product (*3*). If rare disease-associated variants with obvious functional effects are found in a candidate gene that harbors a common disease-associated variant, then the gene is likely to be causal. Recent technological advances in high-throughput sequencing (*4*) provide an opportunity to re-sequence multiple genetic regions in hundreds of subjects and discover rare sequence variants (*5*–*7*). Here we used 454 Sequencing (*8*) to search for rare variants in ten candidate genes and study their association with type 1 diabetes (T1D), previously known as insulin-dependent diabetes mellitus (IDDM). T1D is a common disorder that develops as a result of a complex interaction of genetic and environmental factors leading to the immune-mediated destruction of the insulin-producing pancreatic β-cells. To date 15 loci associated with T1D have been identified in the human genome (*9*–*13*).

Of the ten genes that we selected, six genes contain common T1D-associated polymorphisms: *PTPN22*, *PTPN2*, *IFIH1*, *SH2B3*, *CLEC16A* and *IL2RA* (*10*, *11*, *14*–*16*). We also studied two genes that contain rare mutations causing monogenic syndromes that may include immune-mediated diabetes: *FOXP3*, which is responsible for X-linked syndrome of Immunodysregulation-Polyendocrinopathy-Enteropathy (IPEX, OMIM #304790); and *AIRE*, which is responsible for the Autoimmune PolyEndocrinopathy-Candidiasis-Ectodermal Dystrophy syndrome (APECED, OMIM #240300). Finally, we studied *KCNJ11* because mutations in this gene cause permanent neonatal diabetes, an insulin-dependent diabetes of the non-immune etiology that can be misdiagnosed as T1D in young children (*17*), and *IAN4L1*, because the orthologue of this gene is associated with immune-mediated diabetes in the rat model of T1D (*18*, *19*).

We resequenced 144 target regions that covered exons and regulatory sequences of the ten genes, 31 kb in total (table S1 and T1DBase: http://www.t1dbase.org/page/PosterView/454Resequencing), in DNA of 480 T1D patients and 480 healthy controls from Great Britain arranged in 20 DNA pools (*20*). We generated 9.4 million reads with an average length of 250 bases and identified a total of 212 SNPs (*20*). We classified 33 of them as common because their estimated minor allele frequency

(MAF) was >3% (table S2), and 179 as rare, because their estimated MAF was <3%. Of the 179 rare SNPs 156 were new (table S3). In the pooled samples it was impossible to distinguish rare insertion/deletion polymorphisms from sequencing errors, so here we studied nucleotide substitutions only.

Our goal was not only to discover new rare variants but also to test their association with T1D in the same experiment, comparing allele frequency in DNA pools of patients and controls. Therefore, it was important that sequence reads generated from the DNA pools estimated accurately allele frequency among individuals that contributed DNA to these pools. To test this, we analyzed eight SNPs from the sequenced regions that had been genotyped previously. We found good correlation between allele frequency in the individual samples and its estimate in the DNA pools ($r = 0.99$, fig. S1), demonstrating that high-throughput sequencing of the DNA pools can be used to accurately measure allele frequencies. We then tested association of all 212 SNPs with T1D comparing pooled samples of cases and controls. As expected, we confirmed previously known association of the common SNPs with T1D ($P = 0.02 - 5 \times 10^{-7}$, χ^2 test; table S2). Among rarer SNPs that had not previously been studied for association with T1D, we noted that the two most associated variants, rs35667974 and rs35337543 ($P = 0.0049$ and 0.000044, exact test; Table 1), reside within the *IFIH1* gene. We did not find evidence of association for rare variants in other genes, except for potential associations of the two SNPs located in introns of the *CLEC16A* gene (Table 1 and table S3).

We next studied two *IFIH1* and two *CLEC16A* SNPs in individual DNA samples from 8379 T1D patients and 10,575 controls from Great Britain. *IFIH1* SNPs were also studied in 3,165 families from Europe and USA comprising one or more offspring with T1D and their parents. The two rare intronic *CLEC16A* SNPs were not associated (table S4), while both rare *IFIH1* SNPs demonstrated strong statistical evidence of association with T1D, showing consistent effect in the case-control and family collections (combined $P = 2.1 \times 10^{-16}$ for rs35667974 and 1.4×10^{-4} for rs35337543, score test; table 2). SNP rs35667974 in exon 14 changes a conserved amino acid from Ile923 to Val (fig. S2), while SNP rs35337543 resides within a conserved splice donor site at position +1 in intron 8. Apart from these two SNPs, our sequencing study identified other rare *IFIH1* SNPs, including three non-synonymous SNPs (nsSNPs), ss107794691/Lys349Arg, ss107794690/Thr702Ile and rs10930046/His460Arg, another SNP in a conserved splice donor site at position +1 in intron 14 (rs35732034) and a nonsense mutation in exon 10 (rs35744605). We genotyped these rare SNPs and found evidence of T1D association for the nonsense mutation rs35744605 and SNP rs35732034 located in the conserved splice site (Table 2), but not for nsSNPs Lys349Arg, Thr702Ile or His460Arg (table S5). We did not genotype *IFIH1* intronic and synonymous SNPs or very rare nsSNP (MAF ≤ 0.2%).

We calculated linkage disequilibrium and found that it is low ($r^2 < 0.04$) between all four associated rare variants, indicating that association of one SNP cannot be explained by any of the other SNPs. We also genotyped two common nsSNPs rs3747517/Arg843His and rs1990760/Thr946Ala (MAF > 25%) that had been found associated with T1D by GWAS (*10*, *12*, *21*) and confirmed their association (table S5). We also used logistic regression analyses (*22*) and found that all four rare variants rs35667974, rs35337543, rs35732034, and rs35744605 were associated with T1D independently of each other and of the common nsSNP rs1990760/Thr946Ala (table S6) and so do not account for association of rs1990760/Thr946Ala detected previously by GWAS. Two common nsSNPs were in strong linkage disequilibrium with each other ($r^2 = 0.60$), and association of rs1990760/Thr946Ala explained the effect of rs3747517/Arg843His. Thus, in the *IFIH1* gene four rare polymorphisms and one common nsSNP rs1990760/Thr946Ala show independent association with T1D (fig. S3), although we cannot exclude a possibility that additional variants with weaker effects also exist in this gene. Importantly, here we demonstrated T1D association and measured effects of each of the newly discovered rare variants separately, without grouping them (*5*, *6*).

In the previous GWAS of 12,000 common nsSNPs we identified T1D-associated locus on chromosome 2q24 that included *IFIH1* along with *FAP* and *GCA* genes and part of *KCNH7* gene (fig. S4) (*10*). Although *IFIH1* is a biologically plausible candidate gene, there was no evidence indicating which of these genes is causative for T1D. Discovery of multiple rare T1D-associated variants in *IFIH1* now points to its etiological role in T1D, because it is highly unlikely that multiple untested variants elsewhere in the region could explain association of the rare *IFIH1* variants via linkage disequilibrium. We did not resequence the *FAP*, *GCA* and *KCNH7* genes and so we cannot formally exclude that they might also contain rare T1D-associated variants. This possibility is unlikely, but if true, would not negate the role of *IFIH1*, instead implying that *IFIH1* is not the only T1D gene in this region.

All four associated rare *IFIH1* variants have predicted biological effects, either truncating the protein (nonsense mutation rs35744605) or affecting essential splicing positions (rs35337543 and rs35732034) or a highly conserved amino acid (rs35667974/Ile923Val; fig. S2). These rare *IFIH1* variants have stronger protective effects on T1D risk (Odds Ratio, OR = 0.51 – 0.74) than the common nsSNP rs1990760/Thr946Ala (OR = 0.86; table S5). For example, rare subjects

carrying Valine at position 923 of the IFIH1 protein have only ~50% risk of developing T1D comparing to those who carry Isoleucine. Our results suggest that in complex diseases, such as T1D, there may be no, or very few, low frequency variants with very strong effects (e.g. allele OR > 3), even if such variants have large impacts on a certain molecule's function, possibly because in complex multifactorial diseases such a molecule and its biological pathway are just one of many contributing to the pathogenesis. Nevertheless, discovery of such rare variants using high-throughput sequencing will help to pinpoint disease genes in the associated loci found by GWAS in various complex diseases.

IFIH1 (interferon induced with helicase C domain 1), also known as MDA5 (Melanoma differentiation-associated protein 5), is a 1025 amino acid cytoplasmic protein that recognizes RNA of picornaviruses and mediates immune activation (*23*). Interestingly, infection with enteroviruses, which belong to the picornavirus family, is more common among newly diagnosed T1D patients and prediabetic subjects than in the general population and precedes the appearance of autoantibodies—markers of prediabetes (*24*). Enteroviruses are small RNA viruses that include coxsackie A and B, polioviruses and echoviruses, and cause common and often asymptomatic infections. Upon infection IFIH1 senses the presence of viral RNA in the cytoplasm, triggers activation of NF-κB and IRF pathways and induces antiviral IFN-β response (*25*). Although the mechanisms by which *IFIH1* polymorphisms contribute to T1D pathogenesis remain to be explored, we note that one of the protective variants is a nonsense mutation leading to a truncated 626 amino acid protein lacking the C-terminal helicase domain (fig. S3), while two other protective variants localize to the conserved splice donor sites and probably disrupt normal splicing of the IFIH1 transcript. This suggests that variants, which are predicted to reduce function of the IFIH1 protein, would decrease the risk of T1D, while normal IFIH1 function is associated with T1D. To elucidate biological mechanism linking enterovirus infection with T1D future functional experiments should test whether normal immune activation caused by enterovirus infection and mediated by IFIH1 protein may stimulate autoreactive T cells leading to T1D and whether blocking IFIH1 can disrupt this pathogenic mechanism.

We have found that rare alleles of all associated *IFIH1* polymorphisms consistently protect from T1D, while *IFIH1* alleles carried by the majority of the population predispose to the disease. This observation suggests that variants that disrupt IFIH1 function in the host antiviral response have been negatively selected, rather than positively selected because they confer protection from T1D.

References and Notes

1. M. I. McCarthy *et al*., *Nat. Rev. Genet.* **9**, 356 (2008).
2. D. Altshuler, M. J. Daly, E. S. Lander, *Science* **322**, 881 (2008).
3. W. Bodmer, C. Bonilla, *Nat. Genet.* **40**, 695 (2008).
4. E. R. Mardis, *Trends Genet.* **24**, 133 (2008).
5. J. C. Cohen *et al*., *Science* **305**, 869 (2004).
6. W. Ji *et al*., *Nat. Genet.* **40**, 592 (2008).
7. S. Romeo *et al*., *Nat. Genet.* **39**, 513 (2007).
8. M. Margulies *et al*., *Nature* **437**, 376 (2005).
9. S. Nejentsev *et al*., *Nature* **450**, 887 (2007).
10. D. J. Smyth *et al*., *Nat. Genet.* **38**, 617 (2006).
11. J. A. Todd *et al*., *Nat. Genet.* **39**, 857 (2007).
12. P. Concannon *et al*., *Diabetes* **57**, 2858 (2008).
13. J. D. Cooper *et al*., *Nat. Genet.* **40**, 1399 (2008).
14. Wellcome Trust Case Control Consortium, *Nature* **447**, 661 (2007).
15. N. Bottini *et al*., *Nat. Genet.* **7**, 337 (2004).
16. C. E. Lowe *et al*., *Nat. Genet.* **39**, 1074 (2007).
17. R. Murphy, S. Ellard, A. T. Hattersley, *Nat. Clin. Pract. Endocrinol. Metab.* **4**, 200 (2008).
18. L. Hornum, J. Romer, H. Markholst, *Diabetes* **51**, 1972 (2002).
19. A. J. MacMurray *et al*., *Genome Res.* **12**, 1029 (2002).
20. Materials and methods are available as supporting material on *Science* Online.
21. S. Liu *et al*., *Hum. Mol. Genet.* **18**, 358 (2009).
22. H. J. Cordell, D. G. Clayton, *Am. J. Hum. Genet.* **70**, 124 (2002).
23. H. Kato *et al*., *Nature* **441**, 101 (2006).
24. H. Hyoty, K. W. Taylor, *Diabetologia* **45**, 1353 (2002).
25. E. Meylan, J. Tschopp, M. Karin, *Nature* **442**, 39 (2006).
26. D. Smyth *et al*., *Diabetes* **53**, 3020 (2004).

27. We thank the patients, control subjects, and family members for participating in the study. S.N. held the Diabetes Research and Wellness Foundation Non-Clinical Fellowship at the early stages of the project and now holds the Royal Society University Research Fellowship. The Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory is funded by the Juvenile Diabetes Research Foundation, the Wellcome Trust, and the National Institute for Health Biomedical Research Centre. The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust Strategic Award (079895). Full acknowledgements are in the supporting online material. New SNPs were submitted to dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/); their submission (ss) numbers are in table S3.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1167728/DC1
Materials and Methods
SOM Text
Figs. S1 to S5

Tables S1 to S6
References

27 October 2008; accepted 19 February 2009
Published online 5 March 2009; 10.1126/science.1167728
Include this information when citing this paper.

Table 1. Association analysis of rare variants in sequenced pools of DNA from T1D patients and controls.

			Reads (*n*) | Reads (%) | Estimated chr (*n*)*		*P*-value†
SNP	**Location**	**Alleles**	**T1D**	**Controls**	
rs35337543	*IFIH1*, intron 8, IVS8+1	G>C	35/9,719 | 0.36 | 3/960	221/8,808 | 2.51 | 24/960	0.000044
rs35667974	*IFIH1*, exon 14, Ile923Val	A>G	261/36,095 | 0.72 | 7/960	906/37,475 | 2.42 | 23/960	0.0049
ss107794688	*CLEC16A*, intron 23	C>T	168/33,712 | 0.50 | 5/960	450/25,138 | 1.79 | 17/960	0.016‡
ss107794687	*CLEC16A*, intron 11	C>T	431/40,186 | 1.07 | 10/960	808/32,947 | 2.45 | 24/960	0.023‡

Rare SNPs (MAF < 3%) associated with T1D with $P < 0.05$ are shown in this table. Results for all rare SNPs are in table S3.

*Number and proportion of reads and estimated number of chromosomes carrying minor allele.

†*P*-value was calculated using Fisher's exact test for the estimated number of chromosomes carrying minor alleles in the pools of 960 chromosomes from T1D patients and controls.

‡We genotyped two SNPs located in introns 11 and 23 of the *CLEC16A* (C-type lectin domain family 16, member A) gene in the overall case-control collection but found no association with T1D (table S4).

Table 2. Association analysis of the four rare *IFIH1* polymorphisms in T1D patients and controls and in families comprising one or more offspring with T1D and their parents.

	Allele* 1>2		Case – control study									Family study			Combined
			11	(%)	12	(%)	22	(%)	MAF, %	OR (95% CI)†	*P*-value‡	T/NT	RR (95% CI)†	*P*-value§	*P*-value‖
rs35667974/Ile923Val	A>G	T1D	7853	(97.8)	172	(2.1)	3	(0.04)	1.1	0.51	1.3×10^{-14}	67/	0.60	5.9×10^{-4}	2.1×10^{-16}
Exon 14		Controls	9166	(95.7)	404	(4.2)	4	(0.04)	2.2	(0.43 – 0.61)		111	(0.45 – 0.82)		
rs35337543/IVS8+1	G>C	T1D	7945	(98.0)	163	(2.0)	0	(0.0)	1.0	0.68	1.1×10^{-4}	51/	0.85	0.20	1.4×10^{-4}
Intron 8, splice site		Controls	9330	(97.1)	280	(2.9)	0	(0.0)	1.5	(0.56 – 0.83)		60	(0.59 – 1.23)		
rs35744605/Glu627X	G>T	T1D	8109	(99.1)	76	(0.9)	0	(0.0)	0.46	0.69	9.0×10^{-3}	17/	0.55	2.8×10^{-2}	1.3×10^{-3}
Exon10		Controls	9621	(98.7)	131	(1.3)	0	(0.0)	0.67	(0.52 – 0.91)		31	(0.30 – 0.99)		
rs35732034/IVS14+1	G>A	T1D	8047	(98.6)	109	(1.3)	2	(0.03)	0.69	0.74	1.2×10^{-2}	35/	0.63	2.1×10^{-2}	1.1×10^{-3}
Intron 14, splice site		Controls	9552	(98.1)	180	(1.9)	1	(0.01)	0.93	(0.59 – 0.94)		56	(0.41 – 0.95)		

Results for additional *IFIH1* SNPs are available in table S5.

*Major allele is coded 1, minor allele is coded 2.

†Odds ratios (OR) and relative risks (RR) for minor (rarer) alleles are shown.

‡Two-tailed *P*-values were calculated using logistic regression.

§One-tailed *P*-values were calculated using transmission disequilibrium test with robust variance estimates.

‖Combined *P*-values for the case-control and family data were calculated using a score test as described previously (*26*).

95% CI, 95% confidence interval; MAF, minor allele frequency; T/NT, number of alleles transmitted and non-transmitted to the affected offspring.