

# A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene

Hakon Hakonarson<sup>1,3\*</sup>, Struan F. A. Grant<sup>1,3\*</sup>, Jonathan P. Bradfield<sup>1\*</sup>, Luc Marchand<sup>5</sup>, Cecilia E. Kim<sup>1</sup>, Joseph T. Glessner<sup>1</sup>, Rosemarie Grabs<sup>5</sup>, Tracy Casalunovo<sup>1</sup>, Shayne P. Taback<sup>6</sup>, Edward C. Frackelton<sup>1</sup>, Margaret L. Lawson<sup>7</sup>, Luke J. Robinson<sup>1</sup>, Robert Skraban<sup>1</sup>, Yang Lu<sup>5</sup>, Rosetta M. Chiavacci<sup>1</sup>, Charles A. Stanley<sup>4</sup>, Susan E. Kirsch<sup>8</sup>, Eric F. Rappaport<sup>9</sup>, Jordan S. Orange<sup>10</sup>, Dimitri S. Monos<sup>2,10</sup>, Marcella Devoto<sup>3,11</sup>, Hui-Qi Qu<sup>5</sup> & Constantin Polychronakos<sup>5</sup>

Type 1 diabetes (T1D) in children results from autoimmune destruction of pancreatic beta cells, leading to insufficient production of insulin<sup>1</sup>. A number of genetic determinants of T1D have already been established through candidate gene studies, primarily within the major histocompatibility complex<sup>2–4</sup> but also within other loci<sup>5–12</sup>. To identify new genetic factors that increase the risk of T1D, we performed a genome-wide association study in a large paediatric cohort of European descent. In addition to confirming previously identified loci<sup>2–9</sup>, we found that T1D was significantly associated with variation within a 233-kb linkage disequilibrium block on chromosome 16p13. This region contains *KIAA0350*, the gene product of which is predicted to be a sugar-binding, C-type lectin. Three common non-coding variants of the gene (rs2903692, rs725613 and rs17673553) in strong linkage disequilibrium reached genome-wide significance for association with T1D. A subsequent transmission disequilibrium test replication study in an independent cohort confirmed the association. These results indicate that *KIAA0350* might be involved in the pathogenesis of T1D and demonstrate the utility of the genome-wide association approach in the identification of previously unsuspected genetic determinants of complex traits.

The risk for T1D is strongly influenced by multiple genetic loci and environmental factors. The disease is heritable, with first-degree relatives of patients with T1D being at 15-fold greater risk for developing the condition than the general population.

Variation in four loci has already been established to account for a significant proportion of the familial clustering of T1D. These include the major histocompatibility complex (MHC) region on chromosome 6p21 (mostly residing in the *HLA-DRB1*, *-DQA1* and *-DQB1* genes<sup>2–4</sup>), the insulin locus (*INS*) on chromosome 11p15 (refs 5–7), the protein tyrosine phosphatase-22 (*PTPN22*) gene on chromosome 1p13 (refs 8, 9) and the gene that encodes the cytotoxic T-lymphocyte-associated protein 4 (*CTLA4*) on chromosome 2q31 (refs 10–12). The interleukin-2 receptor alpha (*CD25 IL2RA*) locus on chromosome 10p15 (ref. 13) has also been implicated, and a report that T1D is associated with a non-synonymous variant in the innate immunity gene *IFIH1* (ref. 14) remains to be independently replicated. Several other reported associations<sup>15–17</sup> have not been convincingly replicated and remain controversial, and linkage

studies<sup>18–21</sup> have established that there is no other locus with an effect size approaching that of the *HLA* genes.

The established genetic associations with T1D explain little more than half of the genetic risk for T1D, indicating that other loci exist, although their number and effect size remain unknown. To search systematically for the remaining loci, we performed a two-stage genome-wide association (GWA) study.

In stage 1, we genotyped 550,000 single nucleotide polymorphisms (SNPs) with the Illumina Human Hap550 Genotyping BeadChip<sup>22</sup>, from 563 patients with T1D and 1,146 controls of European ancestry (based on self-report) plus 483 complete T1D family trios of the same ancestry. All patients had clinically proven T1D and were using insulin. We rejected 2,239 SNPs because of call rates <90% and 19,065 SNPs were removed for minor allele frequencies below 1%; as such, 534,071 SNPs remained in the analysis. Following this process, 16 trios, 2 cases and 3 controls were removed from further consideration.

In the case-control analysis, we compared single-marker allele frequencies using  $\chi^2$  statistics for all markers. We used the transmission disequilibrium test (TDT) to calculate the *P*-values of transmission distortion from heterozygous parents in affected parent-child trios. The resulting *P*-values from the case-control and family-based analyses were then combined using Fisher's method<sup>23</sup> to quantify the overall evidence for association. As anticipated, the MHC region was strongly positive, with 392 markers above the threshold for Bonferroni correction (Supplementary Table S1). As this locus is well established, and as a much denser marker coverage is needed to deal with the particularities of this region, we elected not to address this locus further. However, it should be noted that allele A of the most significant MHC-associated SNP, rs2647044, tags *HLA-DRB1* as efficiently as a previously identified SNP<sup>24</sup> and was observed to be in epistasis ( $P < 10^{-10}$ ) with rs3117098, which is also associated with T1D, at the butyrophilin-like 2 (*BTNL2*) locus within the MHC. We found no other significant epistasis with significantly associated SNPs.

Eleven non-MHC SNPs were the next most significant markers, and remained significant at the 0.05 level after Bonferroni correction (Table 1). One of these eleven markers, rs2476601 ( $P = 1.11 \times 10^{-12}$ ) and another five markers, rs1004446, rs6356, rs10770141, rs7111341

<sup>1</sup>Center for Applied Genomics, and <sup>2</sup>Department of Pathology and Laboratory Medicine, Abramson Research Center, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>3</sup>Department of Pediatrics and Division of Human Genetics, and <sup>4</sup>Division of Endocrinology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>5</sup>Departments of Pediatrics and Human Genetics, McGill University, Montreal H3H 1P3, Québec, Canada. <sup>6</sup>Department of Pediatrics and Child Health, University of Manitoba, Winnipeg R3E 0Z2, Manitoba, Canada. <sup>7</sup>Division of Endocrinology, Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa K1H 8L1, Ontario, Canada. <sup>8</sup>Markham-Stouffville Hospital, Markham L3P 7P3, Ontario, Canada. <sup>9</sup>The Children's Hospital of Philadelphia Nucleic Acid and Protein Core, Philadelphia, Pennsylvania 19104, USA. <sup>10</sup>Department of Pediatrics, University of Pennsylvania, School of Medicine, Philadelphia, Pennsylvania 19104, USA. <sup>11</sup>Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

\*These authors contributed equally to this work.

**Table 1 | TDT and case-control association study results for GWA significant markers**

Chr.	SNP	Allele	Case-control cohort				Triad cohort (n = 467)					Locus
			Aff. allele freq. (n = 561)	Ctrl allele freq. (n = 1,143)	OR (95% CI)	P-value	Alleles	Trans:untrans	TDT P-value	P-value combined		
1	rs2476601	A	0.1471	0.08757	1.80 (1.44, 2.24)	$1.32 \times 10^{-7}$	A:G	137:64	$2.62 \times 10^{-7}$	$1.11 \times 10^{-12}$	<i>PTPN22</i>	
11	rs1004446	T	0.254	0.3539	0.62 (0.53, 0.73)	$4.38 \times 10^{-9}$	T:C	160:228	$5.56 \times 10^{-4}$	$6.75 \times 10^{-11}$	<i>INS</i>	
16	rs2903692	A	0.2834	0.3782	0.65 (0.56, 0.76)	$4.77 \times 10^{-8}$	A:G	170:251	$7.89 \times 10^{-5}$	$1.03 \times 10^{-10}$	<i>KIAA0350</i>	
11	rs6356	A	0.4602	0.3593	1.52 (1.31, 1.76)	$1.78 \times 10^{-8}$	A:G	255:197	0.00637	$2.70 \times 10^{-9}$	<i>INS</i>	
16	rs725613	C	0.3004	0.3898	0.67 (0.58, 0.78)	$3.24 \times 10^{-7}$	C:A	178:248	$6.95 \times 10^{-4}$	$5.23 \times 10^{-9}$	<i>KIAA0350</i>	
7	rs10255021	A	0.06667	0.1095	0.58 (0.44, 0.77)	$1.16 \times 10^{-4}$	A:G	18:57	$6.69 \times 10^{-6}$	$1.71 \times 10^{-8}$	<i>COL1A2</i>	
11	rs10770141	A	0.2799	0.373	0.65 (0.56, 0.76)	$7.20 \times 10^{-8}$	A:G	186:234	0.01917	$2.95 \times 10^{-8}$	<i>INS</i>	
1	rs672797	T	0.2257	0.1589	1.54 (1.29, 1.85)	$2.67 \times 10^{-6}$	T:G	177:119	$7.49 \times 10^{-4}$	$4.20 \times 10^{-8}$	<i>LPHN2</i>	
16	rs17673553	G	0.2023	0.2791	0.66 (0.55, 0.78)	$1.30 \times 10^{-6}$	G:A	146:203	0.00228	$6.12 \times 10^{-8}$	<i>KIAA0350</i>	
11	rs7111341	T	0.1843	0.2631	0.63 (0.53, 0.76)	$3.77 \times 10^{-7}$	T:C	138:185	0.008919	$6.90 \times 10^{-8}$	<i>INS</i>	
11	rs10743152	T	0.271	0.3574	0.67 (0.57, 0.78)	$4.73 \times 10^{-7}$	T:C	179:233	0.007805	$7.53 \times 10^{-8}$	<i>INS</i>	

Minor allele frequencies, P-values and odds ratios (OR) are shown. The ORs shown are for the minor alleles (as observed in the controls). Combined P-values are also shown, together with the gene in which the markers reside or which they are nearest to. P-values are two-sided in each instance. Aff. allele freq., allele frequency in affected individuals; Chr., chromosome; CI, confidence interval; Ctrl allele freq., allele frequency in unaffected individuals; Trans:untrans, ratio of transmitted to untransmitted alleles.

and rs10743152 ( $P$ -value range  $7.53 \times 10^{-8}$  to  $6.75 \times 10^{-11}$ ), are in two known T1D susceptibility loci, *PTPN22* and *INS*, respectively. Three common non-coding variants (rs2903692 allele A, rs725613 allele C and rs17673553 allele G), in strong linkage disequilibrium (LD) in the *KIAA0350* gene on chromosome 16p13.13, also attained genome-wide significance for T1D association ( $P$ -value range  $6.12 \times 10^{-8}$  to  $1.03 \times 10^{-10}$ ; case-control odds ratio (OR) range 0.65–0.66). The minor allele is protective, with a frequency of 0.28–0.39 in controls. In addition, eleven other markers in the *KIAA0350* LD block showed association with  $P < 0.00001$  in the family trios and case-control cohort combined (Supplementary Table S2). We found no significant interaction between allele A of rs2903692 and known HLA subtypes (Breslow-Day test for heterogeneity of the allelic odds ratio  $P = 0.67$ ; Supplementary Table S3).

Two other loci, namely the gene for collagen type 1  $\alpha 2$  (*COL1A2*; rs10255021) and rs672797, in the vicinity of latrophilin 2 (*LPHN2*), were also significantly associated following Bonferroni correction in stage 1 (Table 1); however, they failed to replicate in follow-up studies. Thus, in stage 1 we have confirmed the association with the three established T1D loci, and uncovered *KIAA0350* as a potential T1D locus of genome-wide significance. This locus was fast-tracked to Stage 2.

Many reported associations with common variants have not been replicated owing to such factors as population stratification, inadequate statistical power and genotyping errors<sup>25</sup>. Using TDT for the family-based analysis provides adequate protection against stratification, and the clustering of significant SNPs at the *KIAA0350* locus makes genotyping error extremely unlikely; in addition, the application of EIGENSTRAT<sup>26</sup> to the case-control data set indicated that population stratification had little impact on our results (see Supplementary Information). Nevertheless, we sought to confirm the association between T1D and the locus in an additional unrelated sample of affected parent–offspring trios. We used TDT to calculate the level of significance of differences between transmitted and untransmitted allele counts in 1,333 affected offspring from 549 nuclear families from the Type 1 Diabetes Genetics Consortium (T1DGC) plus an additional 390 Canadian trios. Using the SNPlex platform from Sequenom, we confirmed the association of rs17673553, rs725613 and rs2903692 ( $P = 0.023–0.0022$ ) and found that several other markers in the LD block also showed association (Supplementary Table S2). All of these SNPs were in LD, with the

minor allele again conferring protection, except for the minor A allele of rs7200786, which conferred risk, yielding an OR of 1.33 and a PAR of 12.6% (combined  $P$  for all three cohorts =  $9.12 \times 10^{-7}$ ). These SNPs have frequencies in our controls that are very close to those observed in the International HapMap, all are in Hardy–Weinberg equilibrium and all survive all quality control measures for high-quality SNPs.

In an analysis that combined all three independent cohorts (563 cases against 1,146 controls; 483 stage 1 trios and 1,333 T1D offspring from 939 nuclear families for stage 2) for these three intragenic *KIAA0350* markers, the combined  $P$ -values for their association with T1D ranged from  $2.74 \times 10^{-9}$  to  $6.7 \times 10^{-11}$ . The results remain significant when limited to the 839 nuclear families that self-report as Caucasian (Table 2, Supplementary Table S2).

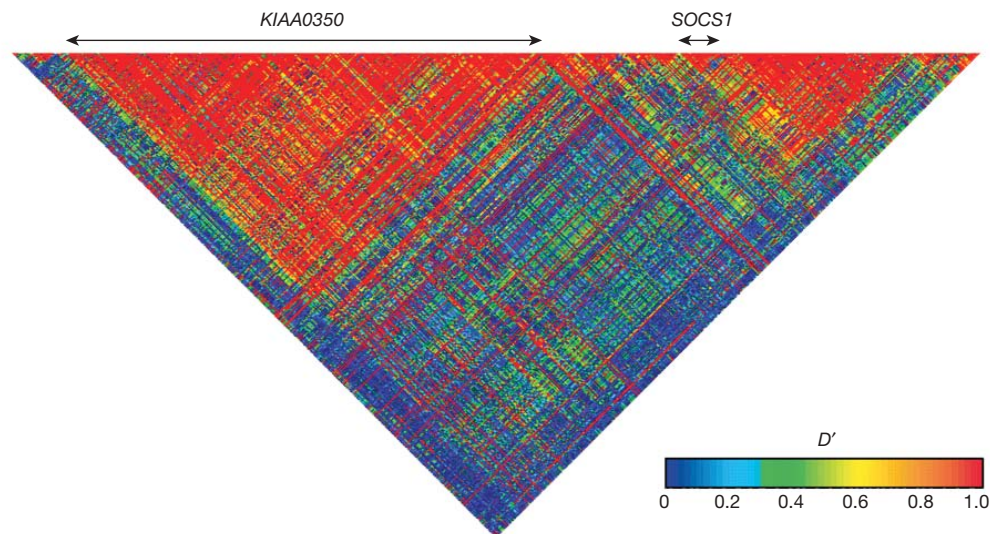
This locus resides in a 233-kb block of LD that contains only *KIAA0350* and no other genes, making this gene a prime candidate for harbouring the causative variant. *KIAA0350* encodes a protein of unknown function and its genomic location is next to the suppressor of cytokine signalling 1 (*SOCS1*) gene. The almost exclusive expression specificity of *KIAA0350* in immune cells (<http://symatlas.gnf.org/SymAtlas>), including dendritic cells, B lymphocytes and natural killer (NK) cells, all of which are pivotal in the pathogenesis of T1D<sup>27,28</sup>, indicates that the variant probably contributes to the disease by modulating immunity. The predicted protein product of *KIAA0350* bears similarities to a subset of adhesion and immune function signalling molecules. Pfam<sup>29</sup> prediction indicates that this gene probably encodes a protein with a calcium-dependent, or C-type, lectin-binding domain structure. Proteins of this type are known to be involved with calcium current flux, and the predicted function of the protein encoded by *KIAA0350* includes sugar binding, according to the Gene Ontology project (GO: 0005529; <http://www.geneontology.org>). The C-type lectins are known for their recognition of various carbohydrates and are crucial for processes that range from cell adhesion to pathogen recognition<sup>30</sup>.

To investigate whether genotype influences *KIAA0350* expression, we used a synonymous SNP in exon 19, rs2286973, ( $r^2 = 0.72$  with rs725613;  $r^2 = 0.67$  with rs2903692) to evaluate the relative abundance of each allele in steady-state messenger RNA from ten lymphoblastoid cell lines. The allele ratios in the mRNA, determined through bi-directional sequencing, were not different from those in the DNA, indicating that expression is not influenced by genotype

**Table 2 | Replication of stage 1 results for *KIAA0350* in a family-based analysis of an independent cohort derived from 939 nuclear families**

Chr.	SNP	All trios (939 nuclear families)			Caucasians only (839 nuclear families)		Stage 1 and replication
		Alleles	Trans: untrans	TDT P-value	Trans: untrans	TDT P-value	Combined P-value
16	rs2903692	A:G	466:538	0.023	438:504	0.032	$6.70 \times 10^{-11}$
16	rs725613	C:A	461:559	$2.15 \times 10^{-3}$	435:520	$5.95 \times 10^{-3}$	$8.86 \times 10^{-11}$
16	rs17673553	G:A	371:448	$7.13 \times 10^{-3}$	348:422	$7.66 \times 10^{-3}$	$2.74 \times 10^{-9}$

Family-based association P-values were computed using TDT. P-values are two-sided in each instance.



**Figure 1 | Pairwise linkage disequilibrium diagram of the *KIAA0350* locus on 16p13.13.** This 'gold plot' is derived from HapMap CEPH Utah (CEU) data corresponding to a region spanning from 10,899,122 (rs8063850) to 11,395,501 (rs12597032) base pairs on chromosome 16 (build 35); intensity of shading is proportional to  $D'$  (the fraction of observed linkage

disequilibrium over the maximal possible). The relative genomic location of the *KIAA0350* gene is shown; it is contained within a single LD block; no other gene resides within this LD block of association. The most pertinent gene in the adjacent region is the *SOCS1* gene (indicated).

(Supplementary Fig. S1). In an additional attempt to identify the functional variant, we sequenced exon 23 in 20 affected individuals who were homozygous for the risk allele, to determine whether the association might be due to a strong predisposing effect from the rare non-synonymous SNP, rs2241100; however, all individuals were homozygous for the common allele. We also investigated the expression of *KIAA0350* in four different NK cell lines and found a trend towards higher expression in the NKL cell line; interestingly, this cell line is the only one that is homozygous for allele A of rs2903692 (Supplementary Fig. S2).

Studies are underway to characterize the functional role of *KIAA0350*. In light of the crucial role of the MHC genetic repertoire in antigen presentation involving sugar groups, such as lectin, we hypothesize that a genetic variant in the binding site for such a molecule on the activating cytotoxic T-cell could elicit an autoimmune response that results in the destruction of the islet cells of the pancreas, as seen in T1D.

Finally, it is worth noting that haplotype-based coverage with 550-K markers did not reveal loci with effect sizes equal to or stronger than those of *INS* and *PTPN22*. This might lead to the conclusion that such loci, if they exist, cannot be numerous. Although *INS* was unequivocally detected despite the fact that the most associated haplotype at that locus was tagged at an  $r^2 < 0.8$ , it is still possible that our stage 1 missed loci in that order of effect magnitude because of imperfect tagging. Our approach has a high likelihood of discovering such loci in the course of a full stage 2 and subsequent fine mapping of the loci that will be discovered.

*Note added in proof:* After the acceptance of this manuscript, an independent GWA was reported that also identified *KIAA0350* as a T1D locus<sup>31,32</sup>.

## METHODS SUMMARY

Cases and family trios for stage 1, as well as 390 of the families that were used in stage 2, were identified through paediatric diabetes clinics in Philadelphia, Montreal, Toronto, Ottawa and Winnipeg; these cases are unique to our study. The remaining stage 2 families were provided by the T1DGC and originated from Europe, North America and Australia. T1DGC families are available to many investigators but we are not aware of any GWA studies that have used them. Controls for stage 1 were drawn from the Children's Hospital of Philadelphia Health Care Network. Genotyping for stage 1 was conducted using Illumina Infinium Hap500 high-density oligonucleotide microarrays. Genotyping for

stage 2 was conducted with matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) mass spectrometry using the Sequenom iPLEX system. All statistical tests for association were carried out using the software package *plink*. The single-marker analysis for the genome-wide data was carried out using a  $\chi^2$  test on allele count differences in cases and controls. ORs and the corresponding 95% confidence intervals were calculated for the association analysis. A TDT was used to evaluate differences between transmitted and untransmitted allele counts in T1D trios in stage 1 and in nuclear families in stage 2, using the standard TDT implemented in the Haploview software package (<http://www.broad.mit.edu/mpg/haploview>). The *P*-values from the case-control and family-based analyses in stage 1 were combined using Fisher's method to quantify the overall evidence for association.

On completion of our full stage 2, we will make summary data publicly available (that is, genotype counts for all sets of subjects and transmission counts from heterozygous parents in the family data, at all loci that passed quality control).

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 30 April; accepted 11 June 2007.

Published online 15 July 2007.

- Devendra, D., Liu, E. & Eisenbarth, G. S. Type 1 diabetes: recent developments. *Br. Med. J.* **328**, 750–754 (2004).
- Cucca, F. *et al.* A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins. *Hum. Mol. Genet.* **10**, 2025–2037 (2001).
- Nerup, J. *et al.* HL-A antigens and diabetes mellitus. *Lancet* **2**, 864–866 (1974).
- Noble, J. A. *et al.* The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families. *Am. J. Hum. Genet.* **59**, 1134–1148 (1996).
- Bell, G. I., Horita, S. & Karam, J. H. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33**, 176–183 (1984).
- Bennett, S. T. *et al.* Susceptibility to human type 1 diabetes at *IDDM2* is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nature Genet.* **9**, 284–292 (1995).
- Vafiadis, P. *et al.* Insulin expression in human thymus is modulated by *INS VNTR* alleles at the *IDDM2* locus. *Nature Genet.* **15**, 289–292 (1997).
- Bottini, N. *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type 1 diabetes. *Nature Genet.* **36**, 337–338 (2004).
- Smyth, D. *et al.* Replication of an association between the lymphoid tyrosine phosphatase locus (*LYP/PTPN22*) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes* **53**, 3020–3023 (2004).
- Kristiansen, O. P., Larsen, Z. M. & Pociot, F. CTLA-4 in autoimmune diseases—a general susceptibility gene to autoimmunity? *Genes Immun.* **1**, 170–184 (2000).

11. Ueda, H. *et al.* Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
12. Anjos, S. M., Tessier, M. C. & Polychronakos, C. Association of the cytotoxic T lymphocyte-associated antigen 4 gene with type 1 diabetes: evidence for independent effects of two polymorphisms on the same haplotype block. *J. Clin. Endocrinol. Metab.* **89**, 6257–6265 (2004).
13. Vella, A. *et al.* Localization of a type 1 diabetes locus in the *IL2RA/CD25* region by use of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76**, 773–779 (2005).
14. Smyth, D. J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nature Genet.* **38**, 617–619 (2006).
15. Guo, D. *et al.* A functional variant of SUMO4, a new I $\kappa$ B $\alpha$  modifier, is associated with type 1 diabetes. *Nature Genet.* **36**, 837–841 (2004).
16. Mirel, D. B. *et al.* Association of *IL4R* haplotypes with type 1 diabetes. *Diabetes* **51**, 3336–3341 (2002).
17. Biason-Lauber, A. *et al.* Association of childhood type 1 diabetes mellitus with a variant of PAX4: possible link to beta cell regenerative capacity. *Diabetologia* **48**, 900–905 (2005).
18. Davies, J. L. *et al.* A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371**, 130–136 (1994).
19. Concannon, P. *et al.* A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nature Genet.* **19**, 292–296 (1998).
20. Mein, C. A. *et al.* A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nature Genet.* **19**, 297–300 (1998).
21. Cucca, F. *et al.* A male-female bias in type 1 diabetes and linkage to chromosome Xp in MHC HLA-DR3-positive patients. *Nature Genet.* **19**, 301–302 (1998).
22. Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genet.* **37**, 549–554 (2005).
23. Fisher, R. A. *Statistical Methods for Research Workers* edn 13 (Hafner, New York, 1958).
24. de Bakker, P. I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genet.* **38**, 1166–1172 (2006).
25. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
26. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
27. Poirot, L., Benoist, C. & Mathis, D. Natural killer cells distinguish innocuous and destructive forms of pancreatic islet autoimmunity. *Proc. Natl Acad. Sci. USA* **101**, 8102–8107 (2004).
28. Rodacki, M. *et al.* Altered natural killer cells in type 1 diabetic patients. *Diabetes* **56**, 177–185 (2007).
29. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).
30. Cambi, A. & Figdor, C. G. Levels of complexity in pathogen recognition by C-type lectins. *Curr. Opin. Immunol.* **17**, 345–351 (2005).
31. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* advance online publication, doi: 10.1038/ng2068 (6 June 2007).
32. The 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).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We acknowledge the use of DNA samples from the T1DGC, funded by the NIH. We thank all participating subjects and families. A. Belisle, A. W. Eckert, A. Estevez, K. Fain, R. Frechette, P. Kline, C. C. Onyiah, G. Otieno, E. Santa, J. L. Shaner, R. M. Smith, A. Thomas and L. Williams helped with genotyping or data collection and management. We thank D. Laforde and the PRUDENT team for subject recruitment and all T1DGC coordinating teams. We also thank S. Kristinsson, L. A. Hermannsson and A. Krisbjörnsson for their software design and contribution. This research was financially supported by the Children's Hospital of Philadelphia, Genome Canada through the Ontario Genomics Institute and the Juvenile Diabetes Research Foundation.

**Author Contributions** H.H. and C.P. designed the study and supervised the data analysis and interpretation. S.F.A.G., J.P.B. and M.D. conducted the statistical analyses. C.E.K., T.C., E.C.F. and R.S. directed the genotyping of stage 1. H-Q.Q. and C.P. coordinated the genotyping and data analysis for stage 2. Y.L. and H-Q.Q. performed the resequencing and allelic-imbalance experiments. J.S.O. and E.F.R. carried out the work on NKT expression. J.P.B., J.T.G. and L.J.R. provided bioinformatics support. The remaining authors coordinated the studies used in stage 1 and 2. H.H., S.F.A.G., J.P.B., H-Q.Q. and C.P. drafted the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.H. ([hakonarson@chop.edu](mailto:hakonarson@chop.edu)) or C.P. ([constantin.polychronakos@mcgill.ca](mailto:constantin.polychronakos@mcgill.ca)).

## METHODS

**Type 1 diabetes cohort from Canada.** The Canadian cohort consisted of 1,120 nuclear family trios (one affected child and two parents) and 267 independent T1D cases, collected in paediatric diabetes clinics in Montreal, Toronto, Ottawa and Winnipeg. The median age at onset is 8 yr with lower and upper quartiles at 4.6 yr and 11 yr. All patients were diagnosed under the age of 18 and have been treated with insulin since diagnosis, and none has stopped treatment for any reason during that time. Disease diagnosis was based on these clinical criteria, rather than any laboratory tests. Ethnic backgrounds were of mixed European descent, with the largest single subset (409 families) being French Canadian. The Research Ethics Board of the Montreal Children's Hospital and other participating centres approved the study, and written informed consent was obtained from all subjects.

**Type 1 Diabetes Genetics Consortium cohort.** The Type 1 Diabetes Genetics Consortium cohort consisted of 549 families (2,350 individuals) with at least two children diagnosed with diabetes and both parents available as of the July 2005 data freeze. Criteria were age at diagnosis below 35 yr and had uninterrupted treatment with insulin within six months of diagnosis. For siblings of probands diagnosed under the age of 35 yr, the age-at-diagnosis limit was extended to 45 yr if they were lean and had positive antibodies and/or low C-peptide levels at diagnosis. The median age of onset is 8 yr with quartiles at 4 yr and 13 yr. The samples were collected in Europe, North America and Australia and most subjects were of European ancestry. Autoantibody results are available but were not used to substantiate the diagnosis, except as noted above.

**Type 1 diabetes cohort from Philadelphia.** The T1D cohort consisted of 103 children recruited at the Children's Hospital of Philadelphia (CHOP) since September 2006. All patients were diagnosed under the age of 18 yr. Of those, 49 T1D patients (32 female, 17 male) were caucasian by self-report (average age of onset 7.07 yr; range 9 months–14 yr) and were included in the analysis. All have been treated with insulin since diagnosis and none has stopped treatment for any reason since. The Research Ethics Board of CHOP approved the study and written informed consent was obtained from all subjects.

**Control subjects from Philadelphia.** The control group included 1,146 children with self-reported caucasian status, mean age 9.42 yr; 53.05% male and 46.95% female, who did not have diabetes or a first-degree relative with T1D. These individuals were recruited by CHOP's clinicians and nursing staff within the CHOP's Health Care Network, including four primary care clinics and several group practices and outpatient practices that included well child visits. The Research Ethics Board of CHOP approved the study, and written informed consent was obtained from all subjects.

**Illumina Infinium assay.** We performed high-throughput, genome-wide SNP genotyping, using the InfiniumII HumanHap550 BeadChip technology<sup>22,33</sup> (Illumina), at the Center for Applied Genomics at CHOP. We used 750 ng of genomic DNA to genotype each sample, according to the manufacturer's guidelines. On day one, genomic DNA was amplified 1,000–1,500-fold. On day two, amplified DNA was fragmented to ~300–600 bp, then precipitated and resuspended before being hybridized on to a BeadChip. Single-base extension (SBE) uses a single probe sequence, ~50 bp long, that is designed to hybridize immediately adjacent to the SNP query site. After targeted hybridization to the bead array, the arrayed SNP locus-specific primers (attached to beads) were extended with a single hapten-labelled dideoxynucleotide in the SBE reaction. The haptens were subsequently detected by a multi-layer immunohistochemical sandwich assay, as recently described<sup>22,33</sup>. The Illumina BeadArray Reader scanned each BeadChip at two wavelengths and created an image file. As BeadChip images

were collected, intensity values were determined for all instances of each bead type, and data files were created that summarized intensity values for each bead type. These files consisted of intensity data that were loaded directly into Illumina's genotype analysis software, BeadStudio. A bead pool manifest created from the laboratory information management system (LIMS) database containing all the BeadChip data was loaded into BeadStudio along with the intensity data for the samples. BeadStudio used a normalization algorithm to minimize BeadChip to BeadChip variability. Once the normalization was complete, the clustering algorithm was run to evaluate cluster positions for each locus and to assign individual genotypes. Each locus was given an overall score, which was based on the quality of the clustering, and each individual genotype call was given a GenCall score. GenCall scores provided a quality metric that ranges from 0 to 1 assigned to every genotype called. GenCall scores were then calculated using information from the clustering of the samples. The location of each genotype relative to its assigned cluster determined its GenCall score.

**Sequenom iPLEX assay.** Genotypes for the rapid confirmation study were obtained using the iPLEX assay (Sequenom). Locus-specific PCR primers and allele-specific detection primers were designed using the MassARRAY Assay Design 3.0 software (Sequenom). The sample DNAs were amplified in a 34-plex PCR reaction and labelled using a locus-specific single-base extension reaction. The resulting products were desalted and transferred to a 384-element SpectroCHIP array. Allele detection was performed using MALDI-TOF MS. The mass spectrograms were analysed by the MassARRAY TYPER software (Sequenom). The 90 CEU (European-descent individuals genotyped in HapMap) were included as accuracy controls.

**Statistical approaches.** All statistical tests for association were carried out using the software package *plink* (<http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml>)<sup>34</sup>. The single marker analysis for the genome-wide data was carried out using a  $\chi^2$  test on allele count differences between 563 cases and 1,146 controls. Odds ratios and the corresponding 95% confidence intervals were calculated for the association analysis. The TDT was used to calculate *P*-values on differences between transmitted and untransmitted allele counts in 467 T1D trios in Stage 1 and in 939 nuclear families in Stage 2. Counts of untransmitted and transmitted alleles from heterozygous parents to affected offspring were determined using the standard TDT implemented in the Haploview software package (<http://www.broad.mit.edu/mpg/haploview>)<sup>35</sup>. The *P*-values from the case-control and family-based analyses in stage 1 were combined using Fisher's method<sup>23</sup> to quantify the overall evidence for association.

Because some of our samples (mainly the controls) were not genotyped for HLA-DQB1, tags were chosen on the basis of the Wellcome Trust Sanger Institute data (<http://www.sanger.ac.uk/HGP/Chr6/ng2006-data>). One SNP was chosen to tag DQB1\*0201, namely rs2187668, at  $r^2 = 0.51$ . A two-marker haplotype was used to tag DQB1\*0302, namely rs9275184 and rs241448, at  $r^2 = 0.96$ . To test how well these tags worked to predict HLA status with respect to these two surrogates, the subset of samples that were HLA typed were tested, yielding a correct predication rate of 82.7%. The remainder of the sample was then predicted on the basis of these selected tags and thus the stratification analysis was carried out using these predictions.

33. Steemers, F. J. *et al.* Whole-genome genotyping with the single-base extension assay. *Nature Methods* **3**, 31–33 (2006).
34. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* (in the press).
35. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).