The imprinted *DLK1-MEG3* gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes

Chris Wallace, Deborah J Smyth, Meeta Maisuria-Armer, Neil M Walker, John A Todd & David G Clayton

Genome-wide association (GWA) studies to map common disease susceptibility loci have been hugely successful, with over 300 reproducibly associated loci reported to date[1]. However, these studies have not yet provided convincing evidence for any susceptibility locus subject to parent-of-origin effects. Using imputation to extend existing GWA datasets[2–4], we have obtained robust evidence at rs941576 for paternally inherited risk of type 1 diabetes (T1D; ratio of allelic effects for paternal versus maternal transmissions = 0.75; 95% confidence interval (CI) = 0.71–0.79). This marker is in the imprinted region of chromosome 14q32.2, which contains the functional candidate gene *DLK1*. Our meta-analysis also provided support at genome-wide significance for a T1D locus at chromosome 19p13.2. The highest association was at marker rs2304256 (odds ratio (OR) = 0.86; 95%CI = 0.82–0.90) in the *TYK2* gene, which has previously been associated with systemic lupus erythematosus[5] and multiple sclerosis[6].

We used imputation to assess association with T1D across 2.6 million polymorphic SNPs from the International HapMap Project in a total of 7,514 cases and 9,405 controls of European ancestry from three existing GWA studies: Wellcome Trust Case-Control Consortium (WTCCC; UK)[2], Genetics of Kidneys in Diabetes–National Institute of Mental Health (GoKinD-NIMH; USA)[3] and Type 1 Diabetes Genetic Consortium (T1DGC; UK)[4] (**Supplementary Table 1**). We used the R package snpMatrix[7] to conduct the imputation and calculate single SNP association score tests for each HapMap SNP. The score tests were based on the Cochran-Armitage test, with a Mantel extension to allow combination over different strata (UK region in the case of the WTCCC and T1DGC samples, and estimated ancestry score derived from principal components in the case of the GoKinD-NIMH samples[3]). For imputed SNPs, we calculated the score statistics using the expected value of the imputed SNP, given observed SNPs, with the expectation calculated under the null hypothesis.

Overall, there was some overdispersion of test statistics ($\lambda = 1.14$ and 1.09 for 1 and 2 degrees of freedom, respectively). This was consistent with the large sample size (almost 17,000 samples) and the overdispersion observed in earlier analysis of these data without HapMap imputation[4]. It has been argued that the greater contributor to overdispersion in these data is bias (such as differential genotyping error), rather than population structure[4]; we therefore carefully examined cluster plots for all SNPs used to impute associated SNPs. Three loci showed suggestive evidence for association ($P < 10^{-7}$) in regions not previously associated with T1D (**Supplementary Fig. 1** and **Supplementary Table 2**). One SNP, rs229484, was proximal (30 kb) to a nearby known T1D locus (most associated SNP, rs229541)[3], also at 22q13.1, but was separated by two moderate recombination hotspots, and there was low linkage disequilibrium (LD) between the two markers ($r^2 = 0.1$; $D' = 0.4$).

To replicate these potential effects, we carried out direct genotyping of the three SNPs using TaqMan in a subset of the GWA samples, additional case-control and family samples, and we obtained evidence for association in two of the three loci (**Table 1** and **Supplementary Table 3**). In these two loci, the overall levels of significance were $<10^{-8}$ ($P = 4.13 \times 10^{-9}$ for rs2304256 and $P = 1.62 \times 10^{-10}$ for rs941576).

Table 1 Association testing of two SNPs using direct genotyping in case-control and family samples

	rs2304256:C>A, chromosome 19p13.2					rs941576:A>G, chromosome 14q32.2				
Cohort	*N*	Fq (A)	OR (A:C)	95% CI	*P* value	*N*	Fq (G)	OR (G:A)	95% CI	*P* value
WTCCC	1,766/1,384	0.299	0.84	(0.75–0.94)	2.68×10^{-3}	1,798/1,406	0.43	0.90	(0.81–1.00)	0.049
T1DGC	3,838/3,883	0.294	0.85	(0.80–0.92)	1.45×10^{-5}	3,754/3,736	0.43	0.88	(0.82–0.94)	9.3×10^{-5}
Additional	2,686/4,794	0.290	0.87	(0.81–0.94)	6.02×10^{-4}	2,670/4,840	0.43	0.92	(0.86–0.99)	0.030
Families	3,099	0.266	0.96	(0.90–1.03)	0.290	4,057	0.45	0.87	(0.82–0.93)	1.8×10^{-5}
Case-control combined	8,290/1,0061	0.293	0.86	(0.82–0.90)	1.43×10^{-10}	8,222/9,982	0.43	0.90	(0.86–0.94)	9.8×10^{-7}
Families and case-control	(See above)	—	—	—	4.13×10^{-9}	(See above)	—	—	—	1.62×10^{-10}

Association testing using observed (not imputed) genotypes in a subset of GWA samples, additional case control samples and family samples. SNP names are followed by alleles, ordered as major > minor. N, number of cases/controls or informative transmissions; Fq, frequency of minor allele in controls or parents; OR, odds ratio; CI, confidence interval.

Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. Correspondence should be addressed to C.W. (chris.wallace@cimr.cam.ac.uk).

Received 13 August; accepted 8 October; published online 6 December 2009; doi:10.1038/ng.493

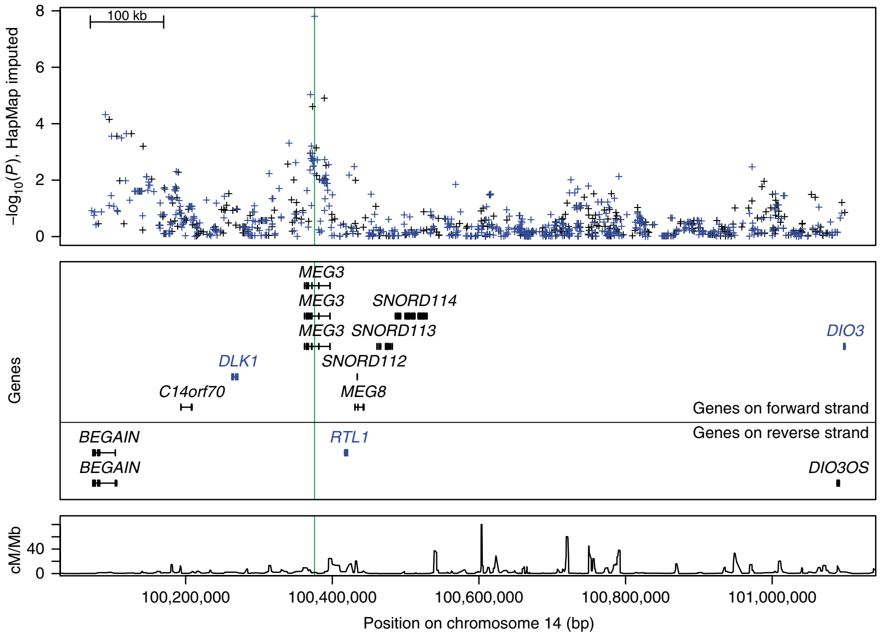

Figure 1 Imprinted region on chromosome 14q32.2. Region shown is delimited by most distant genes known to be imprinted[10] with positions according to the NCBI36 assembly of the human genome. Top panel shows –log$_{10}$(*P*) from 1–degree of freedom tests of association with SNPs across the region. Black, SNPs directly genotyped; blue, SNPs imputed from HapMap. Middle panel shows location and orientation of genes in the region. Blue, paternally expressed genes; black, maternally expressed genes. Bottom panel shows recombination rates (cM/Mb) from HapMap. Solid green line indicates location of rs941576 in all panels.

rs2304256:C>A (OR for A versus C = 0.86) is located within the *TYK2* gene at chromosome 19p13.2, which is implicated in interferon-α, interleukin (IL)-6, IL-10 and IL-12 signaling. This is a region of wide LD containing several functional candidate genes (**Supplementary Fig. 2**). rs2304256 is one of six SNPs in the 1000 Genomes Project database (pilot 1, April 2009) in mutual tight LD (r^2 > 0.9); two are located within *TYK2* (rs34725611 and rs11085725 in introns 6 and 23, respectively) and the remaining three (not yet in dbSNP) are downstream of *TYK2* and upstream of *ICAM3*. No other SNPs had r^2 > 0.62 with any of these six SNPs. There is evidence that *TYK2* is involved in multiple autoimmune diseases: a low-frequency nonsynonymous *TYK2* SNP (rs34536443:G>C, P1104A, minor allele frequency 0.04 versus 0.29 for rs2304256) has been convincingly associated with multiple sclerosis[6] and, in smaller samples and at lower statistical significance, with ankylosing spondylitis and autoimmune thyroid disease[8]. The T1D associated rs2304256 is itself a nonsynonymous SNP (V362F) that has also been associated with systemic lupus erythematosus[5]. In all five diseases, the minor—and inferred nonancestral[9]—allele (A, encoding phenylalanine, for rs2304256 and T1D and lupus; C, encoding alanine, for rs34536443 and multiple sclerosis, ankylosing spondylitis and autoimmune thyroid disease)[5,6,8].

Notably, the newly identified locus with the strongest association with T1D susceptibility occurred in a well-established imprinted region on chromosome 14q32.2 (ref. 10), marked by SNP rs941576: A>G (OR for G versus A = 0.9). Beyond the insulin T1D susceptibility locus, marked by rs7111341 (ref. 4), we do not know of any other T1D SNPs in established imprinted genes. Within this imprinted region of just over 1 Mb, a mixture of paternally derived (*DLK1*, *RTL1* and *DIO3*) and maternally derived (*BEGAIN*, *MEG3*, *MEG8* and *DIO3OS*) genes are expressed[10] (**Fig. 1**). We therefore tested for a parent-of-origin effect, expecting to see excess transmissions of the risk allele from either fathers or mothers (but not both) if the SNP is influencing one of these imprinted genes. A simple way to do this is to consider separately the paternal and maternal transmissions in a transmission disequilibrium testing framework; this revealed strong evidence for reduced paternal transmission of the protective G allele ($P = 6.3 \times 10^{-8}$). Although the maternal transmissions are distorted in the same direction, and a small effect of the maternal copy cannot be discounted, there was no significant evidence for such an effect (P = 0.11; **Table 2**).

Effects resulting from the action of maternal genotype *in utero* are confounded with imprinting effects[11], so we fitted a model allowing for both maternal genotype and imprinting effects. This has been approached in case-parent trio data by log-linear modeling of counts of trios by parental and affected offspring genotype. We extended this method to allow for the fact that many of our families had multiple affected offspring (see Online Methods). The imprinting-only model was preferred (**Supplementary Table 4**); under that model, the imprinting effect was highly significant ($P = 1.85 \times 10^{-8}$; ratio of allelic effects for paternally to maternally inherited alleles = 0.75). This test gains power by using information on parental asymmetry induced by parent-of-origin effects. Asymmetry was clearly shown in our data: the protective allele (G) was less common among fathers of affected offspring than among mothers (0.43 versus 0.47, respectively; $P = 6.53 \times 10^{-7}$). To confirm that the results were not falsely positive, driven by unusual patterns in a subset of the data, we reanalyzed the families subdivided by broad geographical region and found consistent effect estimates across all regions (**Table 3**).

The SNP rs941576 lies within intron 6 of the maternally expressed noncoding RNA gene *MEG3*. However, our observation that only paternal transmissions alter T1D risk suggests that the causal variant influences one of the paternally expressed imprinted genes in its neighborhood (*DLK1*, *RTL1* or *DIO3*). rs941576 is between and downstream of both *DLK1* and *RTL1* and upstream of *DIO3*, at distances of 105 kb, 41 kb and 721 kb respectively. Unusually for a locus identified from GWA data, the signal is restricted to rs941576, and there are no SNPs in HapMap or the current prerelease of the 1000 Genomes Project (pilot 1, April 2009) that are in strong or moderate LD with rs941576 (all r^2 < 0.5; data not shown). Although that does not preclude the existence

Table 2 Transmission disequilibrium tests of rs941576:A>G

Transmissions from	Fq	G untransmitted	G transmitted	*P* value
All parents	0.45	2,166	1,891	1.6×10^{-5}
Fathers	0.43	869	657	6.3×10^{-8}
Mothers	0.47	793	730	0.11

Parental frequency (Fq) and transmissions of the rs941576 protective G allele, overall and separated by parent of origin. Frequencies are calculated using all parents. Because only transmissions from heterozygous (informative) parents are shown, transmission of a G allele implies that A was not transmitted (and *vice versa*). The sum of maternal and paternal transmissions is less than the number of transmissions from all parents because it is not always possible to identify which parent transmitted which allele.

Table 3 Imprinting analysis of rs941576:A>G

Region	N	$\exp(-\hat{\theta})$	95% CI	P value
UK	361	0.792	0.724–0.866	9.40×10^{-3}
Asia-Pacific	32	0.88	0.656–1.18	0.662
Other Europe	257	0.725	0.644–0.815	6.08×10^{-3}
USA	184	0.764	0.676–0.863	0.028
Finland	397	0.697	0.632–0.769	2.25×10^{-4}
Overall	1,231	0.749	0.712–0.789	1.85×10^{-8}

Imprinting analysis using family data divided by broad geographical region. N, number of informative families (which is less than the total number of families available, as only transmissions from asymmetric parents are informative); $\exp(-\hat{\theta})$, ratio of allelic effect for a paternally inherited risk allele compared to a maternally inherited allele; CI, confidence interval.

of an as-yet-unknown variant (SNP or structural variant) in tighter LD, rs941576 lies within a region conserved across mammalian species, including opossum. This is notable because the region in opossum is not imprinted and shows no sequence homology to *MEG3*, and although it shows some sequence homology to mouse *Rtl1* and human *RTL1*, the opossum *Rtl1* sequence seems to be extensively degraded[12]. Thus, if the region is conserved because it contains regulatory elements of nearby genes, these must regulate one of the genes common to all mammals (*DLK1* or *DIO3*).

Although rs941576 lies some distance from the paternally expressed genes in the region, regulatory regions can lie >100 kb from their target genes, particularly in imprinted regions[13]. This region is already subject to long-range *cis*-acting regulation from the intergenic differentially methylated region located 12.5 kb upstream of *MEG3* (ref. 14). Insertion of a transgene in the mouse downstream of this differentially methylated region causes loss of imprinting on the paternal chromosome, biallelic expression of *Meg3* (previously known as *Gtl2*) and reduced expression of *Dlk1* (ref. 15). Thus, it is plausible that this SNP (or another unknown variant nearby) alters the regulation of the paternally expressed *DLK1* or *RTL1*.

Of the paternally expressed genes, only *DLK1* has a strong functional candidacy. It is most strongly expressed in human heart, pancreatic islet cells, pituitary tissue, ovaries, placenta and testes (T1DBase, BioGPS), is related to members of the Notch-Delta family of signaling molecules and encodes a membrane-bound protein that can be cleaved to form fetal antigen-1 (FA1)[16]. FA1 is involved in differentiation of many cell types[17], including pancreatic beta cells, where FA1 immunoreactivity has been localized to glucagon-negative cells in the mature pancreas[18]. FA1 is also involved in hematopoiesis, including differentiation and function of B lymphocytes[19,20], and has been shown to increase expression of proinflammatory cytokines in human bone marrow mesenchymal stem cells and promote B-cell proliferation in human peripheral blood[21]. Thus, there are a number of ways in which variation in *DLK1* expression could alter susceptibility to T1D, which is caused by autoimmune destruction of insulin-producing beta cells in the pancreas.

The mechanisms underlying imprinting are not fully understood but are known to involve epigenetic processes, including DNA methylation and histone acetylation. The causal variant underlying this association could directly alter the expression of the paternally inherited copy of a nearby gene (*DLK1* seems to be the strongest candidate), or it could interfere subtly with the imprinting mechanism and in turn alter expression of either the paternally or maternally inherited copies of a target gene. Although rs941576 may be tagging an unknown causal variant, there is support for the hypothesis that this SNP is itself the causal variant, given its isolation from other SNPs in terms of LD and its location in a conserved and presumably regulatory region.

Rare disorders related to imprinting defects are known (such as Prader-Willi syndrome, MIM#176270). For common complex diseases, over 300 reproducibly associated[1] loci have been reported, but we are not aware of any convincing evidence for another susceptibility locus subject to parent-of-origin effects. At least one common disease locus overlaps a known imprinted region: the T1D-associated region of chromosome 11p15 contains the genes encoding insulin and IGF2, but a previous report by our group of potential parent-of-origin effects at this locus in T1D[22] has not yet been substantiated. We are aware of only one other report of a parent-of-origin effect, in basal cell carcinoma[23], although this was only shown in a single population and at a relatively modest level of statistical significance ($P = \sim 0.01$).

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

This work was funded by the Juvenile Diabetes Research Foundation International, the Wellcome Trust and the National Institute for Health Research Cambridge Biomedical Centre. The Cambridge Institute for Medical Research is in receipt of a Wellcome Trust Strategic Award (079895). We thank all study participants and family members.

We acknowledge use of DNA from the 1958 British Birth Cohort collection, funded by the Medical Research Council (grant G0000934) and the Wellcome Trust (grant 068545/Z/02). We thank The Avon Longitudinal Study of Parents and Children laboratory in Bristol and the British 1958 Birth Cohort team, including S. Ring, R. Jones, M. Pembrey, W. McArdle, D. Strachan and P. Burton, for preparing and providing the control DNA samples.

We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS collection), funded by the Wellcome Trust (grant 076113/C/04/Z), the Wellcome Trust/Juvenile Diabetes Research Foundation (grant 061858) and the National Institute of Health Research of England. The collection was established as part of the Wellcome Trust Case-Control Consortium.

We thank D. Dunger, B. Widmer and the British Society for Paediatric Endocrinology and Diabetes for the TID case collection.

We acknowledge use of DNA from the Human Biological Data Interchange and Diabetes UK for the USA and UK multiplex families, respectively; the Norwegian Study Group for Childhood Diabetes (D. Undlien and K. Ronningen) for the Norwegian families; D. Savage, C. Patterson, D. Carson and P. Maxwell for the Northern Irish families; the Genetics of Type 1 Diabetes in Finland (GET1FIN); J. Tuomilehto, L. Kinnunen, E. Tuomilehto-Wolf, V. Harjutsalo and T. Valle for the Finnish families; and C. Guja and C. Ionescu-Tirgoviste for the Romanian families.

This research used resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the US National Institute of Diabetes and Digestive and Kidney Diseases, the US National Institute of Allergy and Infectious Diseases, the US National Human Genome Research Institute, the US National Institute of Child Health and Human Development and the Juvenile Diabetes Research Foundation International, and was supported by U01 DK062418.

This study made use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

We acknowledge the US National Institute of Mental Health for generously allowing the use of their control CEL and genotype data. Control subjects from the National Institute of Mental Health Schizophrenia Genetics Initiative, data and biomaterials are being collected by the Molecular Genetics of Schizophrenia II collaboration. The investigators and coinvestigators are as follows: P.V. Gejman (collaboration coordinator) and A.R. Sanders (Evanston Northwestern Healthcare–Northwestern University, grant MH059571); F. Amin (Emory University School of Medicine, MH59587); N. Buccola (Louisiana State University Health Sciences Center, MH067257); W. Byerley (University of California, Irvine, MH60870); C.R. Cloninger (Washington University in St. Louis, U01 and MH060879); R. Crowe (principal investigator) and D. Black (University of Iowa, MH59566); R. Freedman (University of Colorado, MH059565); D. Levinson (University of Pennsylvania, MH061675); B. Mowry (University of Queensland, MH059588); and J. Silverman (Mt. Sinai School of Medicine, MH59586). The samples were collected by

V.L. Nimgaonkar's group at the University of Pittsburgh as part of a multi-institutional collaborative research project with J. Smoller and P. Sklar (Massachusetts General Hospital, MH63420).

We acknowledge the US National Institutes of Health for allowing the use of their control-allele signal intensity and genotype data. The data sets used for the analyses described in this manuscript were obtained from the GAIN database, controlled through dbGaP accession number phs000018.v1.p1.

We thank H. Stevens, P. Clarke, G. Coleman, S. Duley, D. Harrison, S. Hawkins, T. Mistry and N. Taylor for preparation of DNA samples; and A. Ferguson-Smith for helpful advice on the chromosome 14q32 region.

AUTHOR CONTRIBUTIONS

C.W. contributed to the design and interpretation of the study, conducted the statistical analyses and drafted the manuscript. D.J.S. conducted genotyping of the three SNPs for replication. M.M.-A. was responsible for the preparation and quality control of DNA samples. N.M.W. was responsible for data management. J.A.T. and D.G.C. contributed to the design and interpretation of the study and drafting of the manuscript. D.G.C. also contributed to the statistical analysis and development of methods for parent-of-origin effects testing.

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

1. Hindorff, L.A., Junkins, H.A., Mehta, J.P. & Manolio, T.A. *A catalog of published genome-wide association studies* (Office of Population Genomics, National Human Genome Research Institute, Bethesda, Maryland, USA, accessed 9 November 2009) <http://www.genome.gov/gwastudies/>.
2. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **47**, 661–678 (2007).
3. Cooper, J.D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.* **40**, 1399–1401 (2008).
4. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
5. Suarez-Gestal, M. *et al.* Replication of recently identified systemic lupus erythematosus genetic associations: a case-control study. *Arthritis Res. Ther.* **11**, R69 (2009).
6. Mero, I.L. *et al.* A rare variant of the *TYK2* gene is confirmed to be associated with multiple sclerosis. *Eur J. Hum. Genet.* published online, doi:10.1038/ejhg.2009.195 (4 November 2009).
7. Clayton, D. & Leung, H.T. An R package for analysis of whole-genome association studies. *Hum. Hered.* **64**, 45–51 (2007).
8. Wellcome Trust Case Control Consortium. *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39**, 1329–1337 (2007).
9. Spencer, C.C.A. *et al.* The influence of recombination on human genetic diversity. *PLoS Genet.* **2**, e148 (2006).
10. Hagan, J.P. *et al.* At least ten genes define the imprinted *Dlk1-Dio3* cluster on mouse chromosome 12qF1. *PLoS One* **4**, e4352 (2009).
11. Weinberg, C.R. *et al.* A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am. J. Hum. Genet.* **62**, 969–978 (1998).
12. Edwards, C.A. *et al.* The evolution of the *DLK1–DIO3* imprinted domain in mammals. *PLoS Biol.* **6**, e135 (2008).
13. Arney, K.L. *H19* and *Igf2* – enhancing the confusion? *Trends Genet.* **19**, 17–23 (2003).
14. Lin, S.P. *et al.* Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the *Dlk1-Gtl2* imprinted cluster on mouse chromosome 12. *Nat. Genet.* **35**, 97–102 (2003).
15. Steshina, E.Y. *et al.* Loss of imprinting at the *Dlk1-Gtl2* locus caused by insertional mutagenesis in the *Gtl2* 5′ region. *BMC Genet.* **7**, 44 (2006).
16. Jensen, C.H. *et al.* Protein structure of fetal antigen 1 (FA1). A novel circulating human epidermal-growth-factor-like protein expressed in neuroendocrine tumors and its relation to the gene products of dlk and pG2. *Eur. J. Biochem.* **225**, 83–92 (1994).
17. Laborda, J. The role of the epidermal growth factor-like protein dlk in cell differentiation. *Histol. Histopathol.* **15**, 119–129 (2000).
18. Tornehave, D. *et al.* FA1 immunoreactivity in endocrine tumours and during development of the human fetal pancreas; negative correlation with glucagon expression. *Histochem. Cell Biol.* **106**, 535–542 (1996).
19. Sakajiri, S. *et al.* Dlk1 in normal and abnormal hematopoiesis. *Leukemia* **19**, 1404–1410 (2005).
20. Raghunandan, R. *et al.* *Dlk1* influences differentiation and function of B lymphocytes. *Stem Cells Dev.* **17**, 495–507 (2008).
21. Abdallah, B.M. *et al.* dlk1/FA1 regulates the function of human bone marrow mesenchymal stem cells by modulating gene expression of pro-inflammatory cytokines and immune response-related factors. *J. Biol. Chem.* **282**, 7339–7351 (2007).
22. Bennett, S.T. *et al.* Insulin VNTR allele-specific effect in type 1 diabetes depends on identity of untransmitted paternal allele. the IMDIAB group. *Nat. Genet.* **17**, 350–352 (1997).
23. Stacey, S.N. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat. Genet.* **41**, 909–914 (2009).

ONLINE METHODS

Sample selection and genotyping. A total of 7,514 cases and 9,045 control samples were included from three GWA studies: WTCCC (UK), T1DGC (UK) and GoKinD-NIMH (USA). The samples and their genotyping have been described[2–4]. Numbers of samples from each study and genotyping platform are given in **Supplementary Table 1**. SNP and sample-exclusion criteria were as applied previously[4]. Briefly, all subjects were of self-reported white European ancestry; samples were excluded if they showed evidence of non-European ancestry, or if they duplicated or were closely related to another sample in the study. SNPs were excluded if the minor allele frequency fell below 1% in cases or controls, if they deviated from Hardy-Weinberg equilibrium ($P < 5.7 \times 10^{-7}$), if the call rate fell below 95% (WTCCC and T1DGC) or if a genotype-calling metric indicated insufficient separation of the signal clouds (GoKinD-NIMH)[24].

SNPs showing suggestive association in the imputed analysis were genotyped directly using TaqMan (Applied Biosystems) on a subset of the GWA samples (the T1DGC, all WTCCC cases and about half the WTCCC controls were available to us), additional case-control samples and a set of family samples with T1D-affected offspring (**Supplementary Table 1**). The additional case and control samples have also been described[4]. The family samples were drawn from across Europe and America and were predominantly of self-reported white European origin; we did not exclude subjects who self-reported a nonwhite European origin, as testing for transmissions within families is equivalent to a pseudo–case-control approach with ethnically matched controls. All TaqMan genotyping data were scored twice to minimize error; the second operator was unaware of case-control status and family structure.

Imputation. For each of the three GWA studies, we divided SNPs from HapMap version 2 (release 24) into two sets: those that were genotyped and passed quality control thresholds in the study (X), and those that were not genotyped or failed quality control (Y). The R package snpMatrix[7] from the BioConductor project[25] was to used calculate imputation ′rules′ for prediction of each SNP in Y from nearby SNPs in X using HapMap genotypes and to carry out association tests for the imputed SNPs. The algorithms used in snpMatrix, together with the parameter settings we used, are described below.

In regions of high LD, the genotype of one SNP can be related to the genotypes of others by a linear regression[26–28]. The first step in calculating an imputation rule is to select a set of ′tag′ SNPs by forward stepwise regression of the Y SNP on the nearest 50 X SNPs (subject to a maximum missing-data requirement). New SNPs are added to the regression until either (i) $R^2 > 0.95$, (ii) the change in R^2 is <0.05 or (iii) the number of tag SNPs reaches four. Regression calculations are carried out at the genotype level, with each SNP genotype coded 0, 1 or 2. If a prediction of $R^2 \geq 0.95$ cannot be achieved using this stepwise regression approach, then an alternative imputation rule is attempted using the set of tag SNPs selected by the forward stepwise procedure. Using the conventional expectation maximization algorithm, frequencies are estimated for the haplotypes of the Y SNP plus the selected tags. Conditional probabilities of the Y allele given the tag SNP haplotype are calculated and provide the imputation rule. This rule is used in preference to the regression rule if the improvement in R^2 exceeds 0.1.

These imputation rules are then applied to the main study data set to calculate the expectation of each Y SNP conditional on typed SNPs. This expectation is not generally an integer, and the Cochran-Armitage test then becomes a t test comparing the mean imputation score in cases with that in controls. Extension to allow for stratified comparisons and to combine information from different studies is straightforward: differences between mean scores are simply averaged over strata (and studies), with weights inversely proportional to their variances. These procedures are all implemented in snpMatrix.

This imputation method is computationally faster than those based on hidden Markov models[29] or variable-length Markov chains[30]. For a subset of our data, we compared our imputation results with those from IMPUTE[29] and found them to be very similar. It has an additional advantage over such methods in that, because each imputation is based on a small number of tag SNPs, it is easier to differentiate between genuine associations and those caused by poor clustering and differential measurement error; for each putative association, allele signal plots for all tags were visually inspected.

Association analysis. Single SNP association score tests were conducted for each HapMap SNP within each cohort using direct genotypes if available, or imputed genotypes if not. The score is calculated using the equation

$$\sum_i (Y_i - \overline{Y})(X_i - \overline{X})$$

where Y_i and X_i are the phenotype (case or control) and genotype data, respectively, for subject i. When a SNP is not directly observed, X_i is replaced by its expected value calculated under the null hypothesis as described above. When it is poorly imputed, this expected value is shrunk toward $\overline{X}$ and contributes little to the test statistic. The permutation variance (the variance under random permutation of Y) is used to calculate the χ^2 test. The score statistics were combined first across strata within cohorts and finally across cohorts using the method proposed by Mantel[31]. The scores (U_i, where i denotes cohort or stratum) and the variances (V_i) are summed to form an overall test of association, $(\Sigma\ U_i)^T\ (\Sigma\ V_i)^{-1}(\Sigma\ U_i)$. Strata were defined by UK region in the case of the WTCCC and T1DGC samples, and by an estimated ancestry score derived from principal components in the case of the GoKinD-NIMH samples[3]. Testing for association with SNPs on the X chromosome was carried out using a previously proposed method[32]. Overdispersion of the test statistics was calculated after removal of known T1D loci[4], and these parameters were used to calculate the adjusted P values given in **Supplementary Table 2**.

SNPs showing overall association ($P < 1 \times 10^{-7}$) in regions not previously reported[4] were subject to further screening. Cluster plots of each SNP used for imputation were examined manually, and the results were discarded unless all cluster plots for all cohorts were considered clearly separated. One of the cohorts studied (USA) was not designed as a T1D case-control study and was serendipitously assembled after cases and controls were genotyped on different versions of the Affymetrix 500K chip and to different protocols. This cohort was subject to greater differential bias than were the other cohorts. As a result, many SNPs were found that showed (often extreme) association in the USA samples ($P < 1 \times 10^{-7}$) but no association in the T1DGC and WTCCC samples combined ($P > 1 \times 10^{-3}$); for these SNPs, only the data from T1DGC and WTCCC were combined.

Family data were analyzed by transmission disequilibrium testing, splitting multiplex families into parent offspring trios and using a pseudo–case-control framework to estimate allelic effects. A score statistic was also generated, and a score test for association in case-controls and families combined was conducted by summing the scores and variances as described above.

Imprinting test. We used a logistic regression approach to test for imprinting and maternal genotype effects on risk in offspring. This approach was originally proposed by Weinberg[10,33] for data consisting of trios of an affected individual and both parents, but we required extension to deal with our data, which included families with multiple affected offspring. Weinberg's approach is to analyze counts of case-parent trios classified by genotype of mother (M), father (P) and affected offspring (O) in a 3 × 3 × 3 table. Of the 15 cells in this table consistent with Mendelian transmission, five concern families in which the genotypes of the two parents are concordant; these are not informative in the analysis. The remaining ten cells can be organized by mating type and offspring genotype into five pairs in which the maternal and paternal genotypes are considered interchangeable (**Supplementary Table 5**). In the absence of maternal genotype and imprinting effects, and assuming that, in the population from which families are drawn, the two possible parental genotype combinations within each mating type are equally frequent, their frequencies in case-parent trios will also not differ systematically. However, maternal genotype and imprinting effects will distort these ratios. In **Supplementary Table 5**, pairs of genotype configurations are set out with the configuration in which the mother carries more copies of the ′2′ allele than the father appearing first. The table also sets out the predictions of a multiplicative model for relative risk conditional on genotype and on parents; the genotype relative risk for the offspring ($\gamma_{1/1}$, $\gamma_{1/2}$ and $\gamma_{2/2}$) is modified by multiplicative effects of the maternal genotype ($\varphi_{1/2}$ and $\varphi_{2/2}$, $\varphi_{1/1}$ being taken as 1) and by a factor θ if a '2' allele was received from the mother rather than from the father. The ratio of these two risks for each mating type gives the ratio of expected frequencies in case-parent trios. This model can be fitted to the observed pairs of case-parent trio frequencies using any standard logistic regression program, thus allowing estimation and testing of maternal genotype and imprinting effects.

Extension of this method to deal with families in which there may be several affected offspring is relatively straightforward. Again, we tabulated counts of families by genotype of mothers, fathers and offspring, but there were then more possible cells in the tabulation. For example, with two affected offspring, there are seven informative pairs of genotype configurations (**Supplementary Table 6**). Under the assumption that the SNP under observation is the sole causal variant or has $r^2 = 1$ with a sole causal variant, disease occurrences in the offspring are conditionally independent given their genotypes and their parents, and the ratio of expected frequencies is given by the ratio of products of predicted relative risks for the two offspring. Extension to the case of more than two affected offspring follows similar principles. For families with three affected offspring, there are 9 informative pairs of genotype configuration, for four affected offspring, 11, and so on. Logistic regression can then be used to estimate and test for effects of maternal genotype and imprinting in the general case, as in our study, where the data consist of families with varying numbers of affected offspring.

In the case where the SNP tested is not the sole causal variant (or in perfect LD with it), disease occurrences in offspring are not conditionally independent and there may be some bias. We would expect this to be small when the SNP has high r^2 with the causal variant. Moreover, the type 1 error rate will be unaffected by departure from conditional independence when testing the hypothesis of no imprinting and no maternal genotype effect against presence of either (or both) effects, although the method may then not be fully efficient.

URLs. 1000 Genomes, http://www.1000genomes.org; BioGPS, http://biogps.gnf.org; International HapMap Project, http://www.hapmap.org; T1DBase, http://www.t1dbase.org.

24. Plagnol, V. *et al.* A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet.* **3**, e74 (2007).
25. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
26. Chapman, J.M. *et al.* Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31 (2003).
27. Clayton, D. *et al.* Use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol.* **27**, 415–428 (2004).
28. 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).
29. Marchini, J. *et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
30. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
31. Mantel, N. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.* **58**, 690–700 (1963).
32. Clayton, D. Testing for association on the X chromosome. *Biostatistics* **9**, 593–600 (2008).
33. Weinberg, C.R. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am. J. Hum. Genet.* **65**, 229–235 (1999).