Meta-analysis of genome-wide association study data identifies additional type 1 diabetes loci

Jason D Cooper[1], **Deborah J Smyth**[1], **Adam M Smiles**[2], **Vincent Plagnol**[1], **Neil M Walker**[1], **James Allen**[1], **Kate Downes**[1], **Jeffrey C Barrett**[1], **Barry Healy**[1], **Josyf C Mychaleckyj**[3], **James H Warram**[2], and **John A Todd**[1]

[1]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, CB2 0XY, UK

[2]Research Division, Joslin Diabetes Center, Boston, MA, USA

[3]Center for Public Health Genomics, University of Virginia, VA, USA

Abstract

To provide more power to detect type 1 diabetes (T1D) loci, we performed a meta-analysis of data from three genome-wide association (GWA) studies. We tested 305,090 SNPs in 3,561 T1D cases and 4,646 controls of European ancestry. We obtained further support for 4q27/*IL2-IL21* ($P = 1.9 \times 10^{-8}$) and, after genotyping 6,225 cases, 6,946 controls and 2,828 families, convincing evidence for four previously unknown and distinct loci in chromosome regions 6q15/*BACH2* (4.7×10^{-12}), 10p15/*PRKCQ* (3.7×10^{-9}), 15q24/*CTSH* (3.2×10^{-15}) and 22q13/*C1QTNF6* (2.0×10^{-8}).

In the present study, we undertook a meta-analysis of three GWA studies, combining the British Wellcome Trust Case Control Consortium (WTCCC)1 T1D case-control data with 1,785 American T1D cases from the Genetics of Kidneys in Diabetes (GoKinD) study2,3 and 1,727 American controls from the National Institute of Mental Health (NIMH). All samples had been genotyped using the Affymetrix 500K SNP chipset, but each study had used a different scoring algorithm - potentially introducing a differential bias in genotype calling between American cases and controls, resulting in false-positive associations4. Consequently, we started the analysis by re-scoring the American case-control data (**Supplementary Methods**).

After re-scoring the American case-control data, for consistency between the studies, we also updated the WTCCC and NIMH SNP information to NCBI Human Genome Build 36 and aligned the SNP alleles between studies. We then applied SNP and sample quality control filters to each study (**Supplementary Methods**). After applying clustering quality and minimum minor allele frequency filters, 330,183 WTCCC and 335,565 GoKinD/NIMH SNPs remained. We note that inspection of allele signal intensity plots was still required for the GoKinD/NIMH SNPs (**Supplementary Methods**). The sample quality control filters consisted of excluding duplicate samples, first or second degree relatives, samples with low heterozygosity and samples with substantial non-European ancestry (**Supplementary**

Methods), all deduced from their genotype. This resulted in 3,561 cases (1,960 British and 1,601 American) and 4,646 controls (2,942 British and 1,704 American).

We analysed the American case-control data separately in order to gauge their quality and suitability for inclusion in the meta-analysis. To control for population structure within the American data, we used a propensity score derived from principal components (**Supplementary Methods**), which reduced the inflation of the test statistic from 18% to 14%. We convincingly detected (least significant $P = 9.2 \times 10^{-4}$) eight of the ten confirmed T1D associated regions[5] (1p13/*PTPN22*, 2q33/*CTLA4*, 6p21/*HLA*, 10p15/*IL2RA*, 12q13/ *ERRB3*, 12q24/*C12orf30*, 16p13/*CLEC16A* and 18p11/*PTPN2*; Supplementary Table 1). The remaining regions were 2q24/*IFIH1* ($P = 0.020$) and 11p15/*INS*, as the closest SNP to the INS gene on the Affymetrix 500K SNP chipset[1] was not available for American cases (**Supplementary Note**). Although no new T1D loci at genome-wide levels of significance ($P < 5.0 \times 10^{-7}$ (ref.[1])) were evident (Supplementary Figure 1), we did find additional support for the previously detected but unconfirmed[5] 4q27/*IL 2-IL21* region at $P = 4.8 \times 10^{-3}$ (Supplementary Table 2).

We then performed a meta-analysis of the evidence for 305,090 SNPs available in both the British and American studies. To produce an overall score test for these SNPs, we summed the score statistics and score variances from the British and American case-control analyses. As expected, the meta-analysis was dominated by known T1D associated regions (Supplementary Table 1 and Supplementary Figure 1); the test statistic inflation was 12% (Supplementary Figure 2). Despite very different population backgrounds (Great Britain, GB, versus USA) and ascertainment (paediatric T1D clinics versus longstanding T1D cases with or without diabetic nephropathy) there was no evidence of heterogeneity in the ten previously established T1D loci (Supplementary Table 1). The combined evidence for 4q27/ *IL2-IL21* was $P = 1.9 \times 10^{-8}$, suggesting that this is a true T1D locus (Supplementary Table 2).

Although no new T1D loci were associated at genome-wide levels of significance, we followed up the most associated SNPs by genotyping an additional 6,225 case and 6,946 control samples from GB. We removed the known T1D loci and SNPs with an r^2 0.1 with them, and shortlisted the top 30 ranked SNPs (least significant $P = 1.2 \times 10^{-5}$ and corrected for 12% inflation, $P = 3.4 \times 10^{-5}$) for follow-up (Supplementary Table 3). A further 11 SNPs from 12q24 and five SNPs from 1p13 were removed as their associations were explained by known T1D loci, rs3184504/*SH2B3*(ref.[5]) (**Supplementary Methods**) and rs2476601/ *PTPN22* (ref.[6]) respectively. A SNP, rs947474/*PRKCQ*, on 10p15 which was 260 kb centromeric of the T1D associated *IL2RA* region[7] proved to be independently associated by regression analysis (Supplementary Table 4) and separated by a number of recombination hotspots (Supplementary Figure 3). In addition, we note that we were unable to find any evidence of an extended haplotype connecting the *IL2RA* region with rs947474/*PRKCQ* (data not shown). A further four SNPs, despite passing SNP quality control filters (**Supplementary Methods**), were excluded after inspection of the genotype signal intensity plots (Supplementary Table 3). As four of the remaining ten SNPs were from 6q15/*BACH2*, we genotyped seven SNPs (Table 1).

We obtained additional support for four of the seven associations: rs11755527/*BACH2* on 6q15 ($P = 6.9 \times 10^{-7}$; OR for minor allele G = 1.13); rs947474/*PRKCQ* on 10p15 ($P = 5.5 \times 10^{-3}$; OR for minor allele G = 0.91); rs3825932/*CTSH* on 15q24 ($P = 8.7 \times 10^{-8}$; OR for minor allele C = 0.86); and, rs229541/*C1QTNF6* on 22q13 ($P = 8.1 \times 10^{-5}$; OR for minor allele T = 1.11) (Table 1). In addition, we regenotyped the seven SNPs in a minimum of 1,771 case and 2,756 control samples used in the WTCCC to validate the genotyping, which showed concordance was 99.3% or better.

We also genotyped the four most associated SNPs, rs11755527/*BACH2*, rs947474/*PRKCQ*, rs3825932/*CTSH* and rs229541/*C1QTNF6*, in 871 multiplex and 1,957 simplex families, resulting in *P*-values of 0.019, 1.3×10^{-3}, 2.3×10^{-4} and 0.15 respectively (Table 1). The overall, combined *P*-values (2.0×10^{-8} to 3.2×10^{-15}; Table 1) provided convincing support for four, previously undetected and distinct, T1D loci.

There are several interesting functional candidate genes located within the four new associated regions (**Supplementary Methods**). The 365 kb associated region on 6q15 contains only one gene, *BACH2*, intron 3 of which contains rs11755527. *BACH2* encodes BTB and CNC homology 1, basic leucine zipper transcription factor 2, which has a role as key regulator of nucleic acid-triggered antiviral responses in human cells8 and is highly expressed in B cells9 (GNF SymAtlas10) (Supplementary Figure 4). In the 234 kb associated region of 10p15, the gene protein kinase C, theta (*PRKCQ*) is 79 kb telomeric of rs947474 (Supplementary Figure 3). PRKCQ controls several fundamental processes in T cell biology, including integration of T cell receptor (TCR) and CD28 signalling, leading to activation of transcription factors (NF-κB and AP-1)11. PRKCQ deficient mice display defects in the differentiation of T helper subsets, particularly in Th2 and Th17 mediated inflammatory responses11. Furthermore, its selective role in T cell effector function, makes *PRKCQ* an attractive therapeutic target in T cell mediated disease processes12.

In the 660 kb associated region of 15q24, rs3825932 is located in intron 1 of cathepsin H (*CTSH*), along with eight other genes (Supplementary Figure 5). On 22q13, rs229541 is located between the genes: C1q and tumor necrosis factor related protein 6 (*C1QTNF6*) and somatostatin receptor 3 (*SSTR3*), in a 125 kb associated region that contains two other genes (Supplementary Figure 6). The meta-analysis results suggest that coding and intronic sequences of the strong candidate gene *IL2RB*, 56 kb centromeric of rs229541, are not associated with T1D, as we previously reported5; although causal variants could affect the function of regulatory sequences that control the expression of genes hundreds of kb away.

To conclude, we present convincing evidence for four, previously undetected and distinct, T1D loci (6q15/*BACH2,* 10p15/*PRKCQ,* 15q24/*CTSH* and 22q13/*C1QTNF6*). In addition, we provide further support for 4q27/*IL2-IL21*, increasing the total of T1D loci with convincing evidence from ten(ref.[5,13,14]) to 15 (including the HLA region). The evidence for these new T1D loci was obtained by forming an American case-control GWA study from existing data and incorporating this data into a meta-analysis with the British (WTCCC) data, followed by independent replication in cases and controls, and with some success, in families.

Recently, an additional locus on chromosome 21q22.3, including the UBASH3A gene, was reported by a Type 1 Diabetes Genetics Consortium (T1DGC) study15. In this study, we have demonstrated the effectiveness of combining the evidence from GWA studies to find disease loci with typical effect sizes (OR < 1.2), and that GWA studies can be successfully formed using case and control data from different studies, provided that allele signal intensity data are available for recalling and checking the SNP genotypes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Wellcome Trust Case Control Consortium. Nature. 2007; 447:661–678. [PubMed: 17554300]

2. Mueller PW, et al. J Am Soc Nephrol. 2006; 17:1782–1790. [PubMed: 16775037]

3. Manolio TA, et al. Nat Genet. 2007; 39:1045–1051. [PubMed: 17728769]

4. Clayton DG, et al. Nat Genet. 2005; 37:1243–1246. [PubMed: 16228001]

5. Todd JA, et al. Nat Genet. 2007; 39:857–864. [PubMed: 17554260]

6. Smyth DJ, et al. Diabetes. 2008; 57:1730–1737. [PubMed: 18305142]

7. Lowe CE, et al. Nat Genet. 2007; 39:1074–1082. [PubMed: 17676041]

8. Hong SW, et al. Biochem Biophys Res Commun. 2008; 365:426–432. [PubMed: 17991429]

9. Muto A, et al. EMBO J. 1998; 17:5734–5743. [PubMed: 9755173]

10. Su AI, et al. Proc Natl Acad Sci U S A. 2002; 99:4465–4470. [PubMed: 11904358]

11. Hayashi K, Altman A. Pharmacol Res. 2007; 55:537–544. [PubMed: 17544292]

12. Chaudhary D, Kasaian M. Curr Opin Investig Drugs. 2006; 7:432–437.

13. Hakonarson H, et al. Diabetes. 2008; 57:1143–1146. [PubMed: 18198356]

14. Smyth DJ, et al. Nat Genet. 2006; 38:617–619. [PubMed: 16699517]

15. Concannon P, et al. Diabetes advance online publication. Jul 22.2008 doi:10.2337/db08-0753.

Table 1

...esults for the seven most associated SNPs found in the meta-analysis and followed-up in additional samples. We report the maximum
...trol and family samples with genotype data. The MAF was estimated in a minimum of 9,647 controls from the British 1958 Birth
...od Services. The signal intensity plots are shown in Supplementary Figure 7. A 2-df test is reported when there was a significant
...genotypic effects model and the multiplicative allelic effects model. The SNP genotypes are reported in Supplementary Table 5. We
... for the family data (that is, the null hypothesis was not rejected unless the effect was in the same direction as the original study).

Gene region	SNP	MAF in British controls	GWA studies						Follow-up				Combined P-values (1-df)
			British 1,960 cases and 2,942 controls		American 1,601 cases and 1,704 controls		British and American		British 6,225 cases and 6,946 controls		2,828 families (3,064 parent-child trios)		
			P (1-df)	OR (95% c.i.)	P (1-df)	OR (95% c.i.)	P (1-df)		P (1-df)	OR (95% c.i.)	P (1-df)	RR (95% c.i.)	
...ergenic region	rs165738 G>C	0.270	0.0157	1.12 (1.02-1.23)	6.09×10^{-5}	1.26 (1.12-1.40)	1.03×10^{-5}		0.0147	1.07 (1.01-1.13)		N/A	3.65×10^{-6}
gene desert	rs887079 T>C	0.498	1.68×10^{-3}	1.14 (1.05-1.24)	5.60×10^{-4}	1.19 (1.08-1.31)	3.95×10^{-6}		0.454	1.02 (0.97-1.07)		N/A	0.465
(2-df test)			8.56×10^{-4}										
BACH2 [†]	rs1755527 C>G	0.465	3.00×10^{-3}	1.13 (1.04-1.23)	4.16×10^{-4}	1.20 (1.08-1.33)	6.07×10^{-6}		6.93×10^{-7}	1.13 (1.08-1.19)	0.0185	1.08 (1.01-1.16)	4.66×10^{-12}
(2-df test)									2.35×10^{-8}				
...× PRKCQ	rs947474 A>G	0.187	2.03×10^{-4}	0.81 (0.73-0.91)	4.56×10^{-3}	0.83 (0.73-0.94)	3.34×10^{-6}		5.49×10^{-3}	0.91 (0.85-0.97)	1.32×10^{-3}	0.86 (0.78-0.95)	3.65×10^{-9}
CTSH	rs825932 T>C	0.318	3.56×10^{-3}	0.87 (0.80-0.96)	5.47×10^{-4}	0.83 (0.74-0.92)	8.93×10^{-6}		8.67×10^{-8}	0.86 (0.82-0.91)	2.29×10^{-4}	0.86 (0.80-0.93)	3.17×10^{-15}
...75, PRM3, TNP2	rs116603 A>T	0.440	8.01×10^{-4}	0.87 (0.80-0.94)	2.02×10^{-3}	0.85 (0.77-0.94)	5.61×10^{-6}		0.0158	0.94 (0.89-0.99)		N/A	2.63×10^{-6}
C1QTNF6	rs229541 C>T	0.427	1.06×10^{-3}	1.15 (1.06-1.25)	3.36×10^{-3}	1.16 (1.05-1.29)	1.16×10^{-5}		8.10×10^{-5}	1.11 (1.05-1.16)	0.146	1.04 (0.97-1.12)	1.98×10^{-8}

...CH2 SNP was rs6191928 meta-analysis P-value = 3.98×10^{-6}, which was in perfect linkage disequilibrium ($r^2 = 1$) with rs11755527, genotyped based on an initial meta-analysis.

...ency, df = degree-of-freedom, OR = odds ratio for minor allele, c.i. = confidence interval, N/A = not attempted.