nature genetics

Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers

Suna Onengut-Gumuscu[1,2,13], Wei-Min Chen[1,3,13], Oliver Burren[4,13], Nick J Cooper[4], Aaron R Quinlan[1,3], Josyf C Mychaleckyj[1,3], Emily Farber[1], Jessica K Bonnie[1], Michal Szpak[1], Ellen Schofield[4], Premanand Achuthan[4], Hui Guo[4], Mary D Fortune[4], Helen Stevens[4], Neil M Walker[4], Lucas D Ward[5,6], Anshul Kundaje[5–8], Manolis Kellis[5,6], Mark J Daly[6,8], Jeffrey C Barrett[9], Jason D Cooper[4], Panos Deloukas[9], Type 1 Diabetes Genetics Consortium[10], John A Todd[4,14], Chris Wallace[4,11,14], Patrick Concannon[1,12,14] & Stephen S Rich[1,3,14]

Genetic studies of type 1 diabetes (T1D) have identified 50 susceptibility regions[1,2], finding major pathways contributing to risk[3], with some loci shared across immune disorders[4–6]. To make genetic comparisons across autoimmune disorders as informative as possible, a dense genotyping array, the Immunochip, was developed, from which we identified four new T1D-associated regions ($P < 5 \times 10^{-8}$). A comparative analysis with 15 immune diseases showed that T1D is more similar genetically to other autoantibody-positive diseases, significantly most similar to juvenile idiopathic arthritis and significantly least similar to ulcerative colitis, and provided support for three additional new T1D risk loci. Using a Bayesian approach, we defined credible sets for the T1D-associated SNPs. The associated SNPs localized to enhancer sequences active in thymus, T and B cells, and CD34$^+$ stem cells. Enhancer-promoter interactions can now be analyzed in these cell types to identify which particular genes and regulatory sequences are causal.

T1D results from the autoimmune destruction of pancreatic β cells, leading to absolute dependence on exogenous insulin to regulate blood glucose levels[7]. In the present study, we designed and used the Immunochip, a custom Illumina Infinium high-density genotyping array, to (i) identify additional risk loci for T1D, (ii) refine mapping of T1D risk loci to their sets of most associated credible SNPs in order to (iii) analyze the locations of the credible SNPs with respect to regulatory sequences in tissues and cell types, and (iv) assemble summary genome-wide association study (GWAS) and Immunochip results from multiple immune diseases to allow comparisons of the genetic risk profiles of these diseases.

The T1D SNP and indel content selected for inclusion on the Immunochip was chosen on the basis of the 41 T1D-associated regions known at the time (February 2010)[1] and 3,000 'wildcard' SNPs that tagged candidate genes or other SNPs with suggestive evidence of association ($5 \times 10^{-8} < P < 1 \times 10^{-5}$) from GWAS of T1D. In parallel, we collected and curated all available association results for immune diseases for which the Immunochip was designed. To allow efficient comparison and downstream analysis by the research community, we created a publicly available, integrated, web-based portal (ImmunoBase; see URLs) containing complete association summary statistics that are available for querying, browsing or bulk download.

After data cleaning and quality control[8,9], a total of 138,229 SNPs were scored in 6,670 T1D cases[10], 6,523 controls from the British 1958 Birth Cohort[11], 2,893 controls from the UK National Blood Service[12], 2,846 controls from the NIHR Cambridge Biomedical Research Centre Cambridge BioResource[13], 2,601 Type 1 Diabetes Genetics Consortium (T1DGC) affected sibling pairs (ASPs)[14] and 69 T1DGC trio families. Case-control and family data were analyzed independently and combined by meta-analysis. We obtained evidence for T1D association in 44 regions at $P \leq 3.23 \times 10^{-7}$ (Immunochip Bonferroni-corrected $P < 0.05$; **Table 1**). Thirty-eight of these are recognized T1D-associated regions (T1DBase and ImmunoBase), and four are newly identified regions (genome-wide $P < 5 \times 10^{-8}$): 1q32.1 (index SNP rs6691977), 2q13 (rs4849135), 4q32.3 (rs2611215)

[1]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA. [2]Department of Medicine, Division of Endocrinology, University of Virginia, Charlottesville, Virginia, USA. [3]Department of Public Health Sciences, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, Virginia, USA. [4]Juvenile Diabetes Research Foundation (JDRF)/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, National Institute for Health Research (NIHR) Biomedical Research Centre, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [5]Department of Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. [6]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [7]Department of Genetics, Stanford University, Stanford, California, USA. [8]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [9]Wellcome Trust Sanger Institute, Hinxton, UK. [10]A complete list of members and affiliations is provided in the **Supplementary Note**. [11]Medical Research Council (MRC) Biostatistics Unit, Institute of Public Health, University Forvie Site, Cambridge, UK. [12]Present address: University of Florida Genetics Institute and Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, Florida, USA. [13]These authors contributed equally to this work. [14]These authors jointly supervised this work. Correspondence should be addressed to S.S.R. (ssr4n@virginia.edu).

Received 4 May 2014; accepted 13 February 2015; published online 9 March 2015; doi:10.1038/ng.3245

Table 1 T1D-associated loci on the Immunochip

New	Chromosome	Position	SNP	Alleles	MAF	OR	*P*	Conditioning	Candidate gene(s)	Previous index SNPs (r^2)
	1p13.2	114,377,568	rs2476601	G>A	0.09	1.89	$<10^{-100}$		*PTPN22*	rs2476601 (1)
*	1q32.1	200,814,959	rs6691977	T>C	0.19	1.13	4.3×10^{-8}			–
	1q32.1	206,939,904	rs3024505	G>A	0.16	0.86	6.4×10^{-8}		*IL10*	rs3024493 (1), rs3024505 (1)
	2q11.2	100,764,087	rs13415583	T>G	0.35	0.90	1.1×10^{-7}		*AFF3*	rs6740838 (0.32), rs9653442 (0.41)
*	2q13	111,615,079	rs4849135	G>T	0.29	0.89	4.4×10^{-8}			–
	2q24.2	163,110,536	rs2111485	G>A	0.39	0.85	3.8×10^{-18}		*IFIH1*	rs1990760 (0.91)
	2q24.2	163,124,637	rs35667974	T>C	0.02	0.59	9.3×10^{-9}	rs2111485	*IFIH1*	rs1990760 (<0.1)
	2q24.2	163,136,942	rs72871627	A>G	0.01	0.61	2.4×10^{-6}	rs2111485, rs35667974	*IFIH1*	rs1990760 (0.0094)
	2q33.2	204,738,919	rs3087243	G>A	0.45	0.84	7.4×10^{-21}		*CTLA4*	rs3087243 (1), rs11571316 (<0.1)
	3p21.31	46,457,412	rs113010081	T>C	0.11	0.85	4.6×10^{-8}		*CCR5*	rs333 (0.34)
	4q27	123,243,596	rs75793288	C>G	0.36	1.15	5.6×10^{-13}		*IL2*, *IL21*	rs6827756 (0.98), rs4505848 (0.85)
*	4q32.3	166,574,267	rs2611215	G>A	0.15	1.18	1.8×10^{-11}			–
*	5p13.2	35,883,251	rs11954020	C>G	0.39	1.11	4.4×10^{-8}		*IL7R*	–
	6q15	90,976,768	rs72928038	G>A	0.17	1.20	6.4×10^{-14}		*BACH2*	rs11755527 (0.194), rs597325 (0.13)
	6q22.32	126,752,884	rs1538171	C>G	0.45	1.12	7.4×10^{-10}			rs9375435 (0.96), rs9388489 (0.98)
	7p12.2	50,465,830	rs62447205	A>G	0.28	0.89	2.5×10^{-8}		*IKZF1*	rs10272724 (0.97)
	7p12.1	51,028,987	rs10277986	A>T	0.04	0.76	1.4×10^{-7}			rs4948088 (0.86), rs10231420 (<0.1)
	9p24.2	4,290,823	rs6476839	A>T	0.40	1.12	1.0×10^{-9}		*GLIS3*	rs10758593 (0.98), rs7020673 (0.66)
	10p15.1	6,094,697	rs61839660	C>T	0.10	0.62	2.8×10^{-39}		*IL2RA*	rs7090530 (<0.1), rs12251307 (0.61)
	10p15.1	61,08,340	rs10795791	A>G	0.41	1.16	5.6×10^{-11}	rs61839660	*IL2RA*	rs7090530 (<0.1), rs12251307 (<0.1)
	10p15.1	6,129,643	rs41295121	C>T	0.01	0.49	4.9×10^{-8}	rs61839660, rs10795791	*IL2RA*	rs7090530 (<0.1), rs12251307 (<0.1)
	10q23.31	90,035,654	rs12416116	C>A	0.28	0.85	3.9×10^{-15}			rs10509540 (0.79)
	11p15.5	2,182,224	rs689	T>A	0.30	0.42	$<10^{-100}$		*INS*	rs7111341 (0.265)
	11p15.5	2,198,665	rs72853903	C>T	0.38	0.85	6.2×10^{-10}	rs689	*INS*	rs7111341 (0.26)
	12p13.31	9,905,851	rs917911	A>C	0.36	1.10	1.9×10^{-7}		*CD69*	rs4763879 (1), rs10492166 (0.470)
	12q13.2	56,435,504	rs705705	G>C	0.34	1.25	4.4×10^{-32}		*IKZF4*	rs2292239 (0.87), rs705704 (0.99)
	12q24.12	112,007,756	rs653178	T>C	0.48	1.30	1.6×10^{-44}		*SH2B3*	rs3184504 (0.99)
	13q32.3	100,081,766	rs9585056	T>C	0.24	1.12	3.3×10^{-8}		*GPR183*	rs9585056 (1)
	14q32.2	98,488,007	rs1456988	T>G	0.27	1.12	2.9×10^{-8}			rs4900384 (0.98)
	14q32.2	101,306,447	rs56994090	T>C	0.41	0.88	1.1×10^{-11}			rs941576 (0.91)
	15q14	38,847,022	rs72727394	C>T	0.19	1.15	3.6×10^{-10}		*RASGRP1*	rs12908309 (<0.1)
	15q25.1	79,234,957	rs34593439	G>A	0.10	0.78	9.0×10^{-14}		*CTSH*	rs3825932 (0.26), rs12148472 (0.79)
	16p11.2	28,505,660	rs151234	G>C	0.12	1.19	4.8×10^{-11}		*IL27*	rs4788084 (0.1), rs9924471 (0.54)
	16p13.13	11,194,771	rs12927355	C>T	0.32	0.82	3.0×10^{-22}		*DEXI*	rs12927355 (1), rs12708716 (0.86), rs12928822 (<1)
	16p13.13	11,351,211	rs193778	A>G	0.25	1.14	4.4×10^{-10}		*DEXI*	rs12927355 (<0.1), rs12708716 (0.069), rs12928822 (<0.1)
	16q23.1	75,252,327	rs8056814	G>A	0.07	1.32	3.0×10^{-19}		*BCAR1*	rs7202877 (0.86), rs8056814 (1)
	17q12	38,053,207	rs12453507	G>C	0.49	0.90	1.0×10^{-8}		*IKZF3*, *ORMDL3*, *GSDMB*	rs2290400 (0.97)
	17q21.2	38,775,150	rs757411	T>C	0.36	0.90	1.1×10^{-7}		*CCR7*	rs7221109 (0.95)
*	17q21.31	44,073,889	rs1052553	A>G	0.24	0.89	8.2×10^{-8}			–
	18p11.21	12,809,340	rs1893217	A>G	0.16	1.21	1.2×10^{-15}		*PTPN2*	rs1893217 (1)
	18p11.21	12,830,538	rs12971201	G>A	0.39	0.89	2.1×10^{-6}	rs1893217	*PTPN2*	rs1893217 (0.13)
	18q22.2	67,526,644	rs1615504	C>T	0.47	1.13	1.8×10^{-11}		*CD226*	rs763361 (0.99)
	19p13.2	10,463,118	rs34536443	G>C	0.04	0.67	4.4×10^{-15}		*TYK2*	rs2304256 (<0.1)
	19p13.2	10,469,975	rs12720356	A>C	0.09	0.82	3.7×10^{-7}	rs34536443	*TYK2*	rs2304256 (0.26)
	19q13.32	47,219,122	rs402072	T>C	0.16	0.87	4.7×10^{-8}			rs425105 (0.98)
	19q13.33	49,206,172	rs516246	T>C	0.49	0.87	5.2×10^{-14}		*FUT2*	rs601338 (1)
	20p13	1,616,206	rs6043409	G>A	0.35	0.88	3.0×10^{-10}			rs2281808 (0.91)
	21q22.3	43,825,357	rs11203202	C>G	0.33	1.16	1.2×10^{-15}		*UBASH3A*	rs11203203 (0.42)
*	21q22.3	45,621,817	rs6518350	A>G	0.18	0.88	9.6×10^{-8}		*ICOSLG*	–
	22q12.2	30,531,091	rs4820830	T>C	0.38	1.14	1.2×10^{-12}			rs5753037 (0.99)
	22q12.3	37,587,111	rs229533	A>C	0.43	1.11	1.8×10^{-8}		*C1QTNF6*, *RAC2*	rs229541 (0.98), rs229526 (0.39)

The most strongly associated SNP in a region is shown, together with the effect of the minor allele relative to the major allele. Where secondary associations were found, they are conditional on the SNPs shown in the column "Conditional". For previously known loci, the r^2 between our lead SNP and the previously reported index SNPs is shown. New loci (at $P < 3.23 \times 10^{-7}$) are indicated by an asterisk. Alleles are shown as major allele>minor allele. rs689 (11p15.5, *INS*) data were obtained from previous TaqMan genotyping. Named candidate genes are genes for which there was additional evidence that they might be causal or which encode proteins with known immune functions that are part of the immune pathways already identified as being involved in T1D pathogenesis. Because SNPs may alter enhancer sequences distant from the target gene, we have not named a gene (or a noncoding RNA) if the only evidence for a causal role was that the peak of SNP association lies in or very near a gene (unless the SNPs alter coding sequence or splicing signals in a potentially functional way). For example, *RNLS* at 10q23.31 has no established role in the immune system, and there is currently no specific functional data linking this gene to T1D etiology.

and 5p13.2 (rs11954020). rs11954020 is close to the multiple sclerosis candidate immune response gene *IL7R*[15]. Two additional loci (at 17q21.31 and 21q22.3) were marginally associated ($P > 5 \times 10^{-8}$), and, as we describe later, additional support for the 17q21.31 locus came from genome-wide significant association of the same SNP, rs1052553, with primary biliary cirrhosis (PBC)[16].

At each of the 44 loci, we investigated whether additional SNPs were independently associated with T1D. Logistic regression analyses conditional on the most strongly associated SNP, or index SNP, in each region identified five loci with more than one independently associated SNP (**Table 1**). Four of these loci were already known to harbor more than one causal variant, but the fifth region, 11p15.5 (*INS* and *INS-IGF2* candidate genes), was surprising as *INS* was the first non-MHC (major histocompatibility complex) region to be discovered for T1D[17], and the region has therefore been examined intensively. The likely causal candidate variants in this region were SNPs rs689 (−23HphI), rs3842753 (c.1140A>C) and the 5′ variable-number tandem repeat (VNTR) polymorphism. In European-ancestry populations, these three sites are in near-perfect linkage disequilibrium (LD)[18]. SNPs rs689 and rs3842753 were assayed on the Immunochip, but both were eliminated after quality control. We integrated pre-existing rs689 data with Immunochip data for the 6,670 UK GRID (UK Genetic Resource Investigating Diabetes) cases and 6,304 British 1958 Birth Cohort controls and found rs689 to be the most strongly associated SNP. After conditioning on rs689, SNP rs72853903 still exhibited significant evidence for an independent association with T1D ($P = 5.4 \times 10^{-10}$; **Table 1**). We did not have sufficient data to integrate rs3842753 or the *INS* VNTR in these analyses, but rs689 is known to tag the VNTR precisely[18]. We note that annotation using VEP[19] (Ensembl v75) identifies rs3842753 as an *INS* nonsynonymous SNP (encoding p.His77Pro). However, we found limited evidence for annotation of the transcript isoform, and rs3842753 is more likely to be a noncoding SNP in the 3′ UTR of *INS*.

Comorbidity between T1D and other immune-mediated diseases has been reported widely through epidemiological and clinical studies, but evidence for shared genetic etiology has not been assessed in a uniform manner across multiple diseases. We sought to compare the underlying genetic susceptibilities to T1D and each of 15 immune diseases curated in ImmunoBase (accessed 13 February 2014). We first divided the densely mapped regions of the Immunochip into two sets according to whether there was published association with the index disease for that region. We then tested whether T1D single-SNP *P* values differed between the two sets of regions using a variant set enrichment method that accounts for LD between SNPs[20] (Online Methods). A difference in *P*-value distributions indicated that T1D showed stronger (or weaker) association with a particular region according to association with the index disease.

This comparison clearly delineated diseases with characteristic autoantibodies (for example, juvenile idiopathic arthritis (JIA), rheumatoid arthritis and T1D) relative to autoinflammatory disorders (ulcerative colitis and Crohn's disease; **Fig. 1a** and **Table 2**). We observed the strongest positive and negative enrichments with JIA ($P = 2 \times 10^{-13}$; **Fig. 1b**) and ulcerative colitis ($P = 5.4 \times 10^{-5}$; **Fig. 1c**), respectively. We note that the susceptibility loci for each disease remain incomplete, with the extent of the incompleteness varying across diseases. This limitation prevents us from drawing any conclusion such as 'T1D is more like rheumatoid arthritis than autoimmune thyroid disease'; however, individually significant results are likely to be valid representations of disease overlap. The overlap between

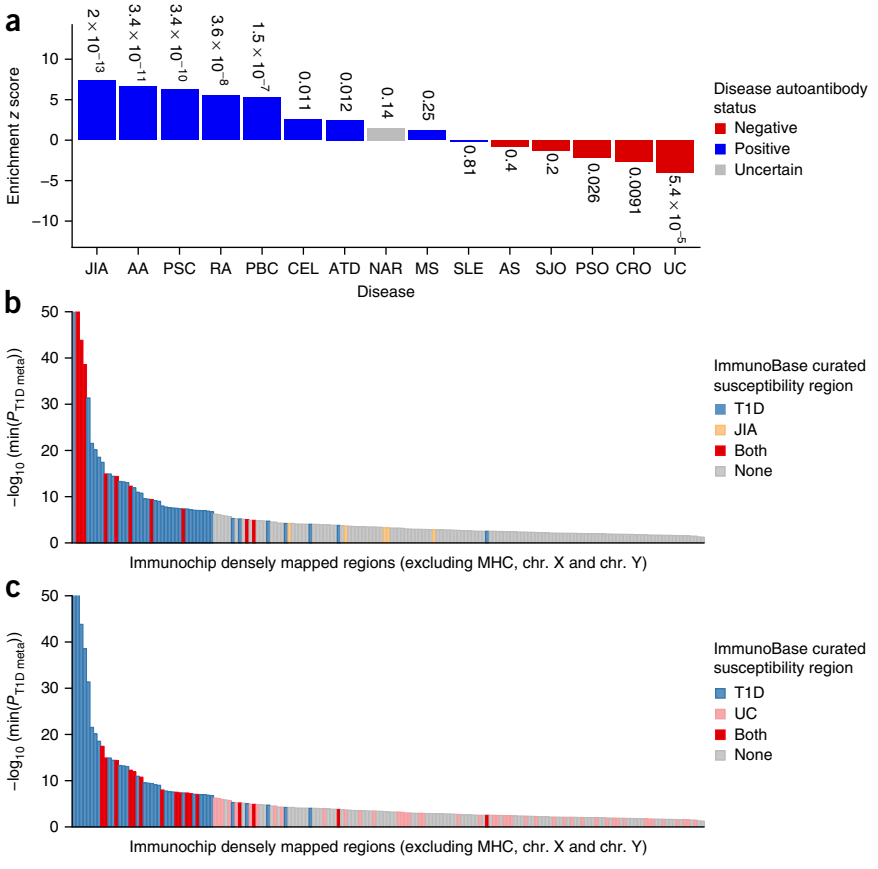

Figure 1 T1D Immunochip *P*-value enrichment analysis. (**a**) Plot showing *z* scores for densely typed regions against diseases curated in ImmunoBase. Diseases with positive *z* scores indicate evidence for overall genetic overlap with T1D, within densely typed regions accessible on the Immunochip. Diseases with negative scores indicate evidence for negative association. Each bar is labeled with the Wilcoxon rank-sum test *P* value and colored by disease autoantibody positive/negative status. The MHC region (chr. 6: 25–35 Mb, GRCh37) was excluded from analysis. AA, alopecia areata; AS, ankylosing spondylitis; ATD, autoimmune thyroid disease; CEL, celiac disease; CRO, Crohn's disease; JIA, juvenile idiopathic arthritis; MS, multiple sclerosis; NAR, narcolepsy; PBC, primary biliary cirrhosis; PSC, primary sclerosing cholangitis; PSO, psoriasis; RA, rheumatoid arthritis; SJO, Sjogren's syndrome; SLE, systemic lupus erythematosus; UC, ulcerative colitis. (**b**,**c**) Plots showing $P' = \min(-\log(P_{\text{T1D meta}}))$ for each densely typed region accessible on the Immunochip, excluding the MHC region and the sex chromosomes. Regions that overlap known T1D susceptibility regions are identified by blue bars, whereas yellow and pink bars show overlap with JIA and ulcerative colitis, respectively. Red bars denote shared overlap between T1D and focal disease. The *y* axis is truncated for clarity. Fully interactive versions of **b** and **c**, along with further supporting resources, are available at http://www.immunobase.org/poster/type-1-diabetes-immunochip-study-onengut-gumuscu/.

Table 2 Enrichment analysis of evidence for T1D association across densely genotyped non-MHC loci associated with other autoimmune or autoinflammatory diseases

		SNPs in regions		Enrichment result	
Index disease	Associated regions	Disease associated	Not disease associated	*z*	*P*
Juvenile idiopathic arthritis	15	2,527	22,725	7.35	2.00×10^{-13}
Areata alopecia	4	763	24,489	6.63	3.40×10^{-11}
Primary sclerosing cholangitis	10	1,866	23,386	6.28	3.40×10^{-10}
Rheumatoid arthritis	27	4,382	20,870	5.51	3.60×10^{-8}
Primary biliary cirrhosis	16	2,289	22,963	5.26	1.50×10^{-7}
Celiac disease	29	4,512	20,740	2.55	1.10×10^{-2}
Autoimmune thyroid disease	9	1,622	23,630	2.50	1.20×10^{-2}
Narcolepsy	2	217	25,035	1.49	1.40×10^{-1}
Multiple sclerosis	57	8,312	16,940	1.15	2.50×10^{-1}
Systematic lupus erythematosus	14	2,528	22,724	–0.23	8.10×10^{-1}
Ankylosing spondylitis	21	3,103	22,149	–0.84	4.00×10^{-1}
Sjogren's syndrome	6	985	24,267	–1.29	2.00×10^{-1}
Psoriasis	25	4,457	20,795	–2.22	2.60×10^{-2}
Crohn's disease	83	13,225	12,027	–2.61	9.10×10^{-3}
Ulcerative colitis	58	9,336	15,916	–4.04	5.40×10^{-5}

Immunochip densely mapped regions were assigned as associated or not associated with each index disease according to publications curated in ImmunoBase (accessed 13 February 2014). We then tested whether the distribution of T1D *P* values differed between these sets of regions. The numbers of SNPs that passed quality control in our T1D study in the two sets of regions are shown. A positive (negative) *z* score implies that T1D shows stronger (weaker) evidence of association in regions known to associate with the index disease.

T1D and JIA was driven, in part, by shared loci ($P < 1 \times 10^{-20}$) at 1p13.2 (*PTPN22*), 12q24.11 (*SH2B3*) and 10p15.1 (*IL2RA*) (**Fig. 1b**,**c**), whereas, for ulcerative colitis, no shared loci reached this level of significance.

We exploited this pleiotropy to identify additional T1D associations. Previously, T1D was compared with celiac disease, and SNPs robustly associated ($P < 5 \times 10^{-8}$) with celiac disease and more weakly associated ($5 \times 10^{-8} < P < 1 \times 10^{-4}$) with T1D were considered to be associated with T1D and vice versa[5]. Here we demonstrate that a SNP with $P < 5 \times 10^{-8}$ in any Immunochip disease study requires $P < 1 \times 10^{-5}$ for T1D to obtain a Bayesian posterior probability of T1D association >0.9, given that different Immunochip disease studies shared many control samples (Online Methods). Using this analysis, we identified 3 additional T1D-associated regions, bringing the number of known T1D susceptibility regions to 57: 14q24.1 (rs911263), 17q21.31 (rs17564829) (which achieved Bonferroni-corrected but not genome-wide significance in the primary analysis) and 6q23.3 (rs17264332 and rs6920220) (**Table 3**).

The 6q23.3 region contains the well-recognized candidate gene *TNFAIP3*, linking T1D susceptibility with the proinflammatory tumor necrosis factor (TNF) pathway. The three genes most proximal to the index SNP in the 14q24.1 region (*RAD51B*, *ZFP36L1* and *ACTN1*) do not provide obvious insights into the biology of T1D, nor do genes near the index SNPs in the three other regions: 1q32.1 (*CAMSAP2-GPR25-C1orf106*), 2q13 (*ACOXL*) and 4q32.3 (*LINC01179-CPE-TLL1*). *CPE* encodes carboxypeptidase E, a protease active in the neuroendocrine system and therefore could be considered a candidate T1D-associated gene. The gene content of the 17q21.31 (rs17564829) region, containing a 1-Mb inversion polymorphism with several copy number variants[21], is also not informative, although *SPPL2C*, encoding signal peptide peptidase–like 2C, could be considered a candidate gene. Antigen presentation and associated proteolysis is important in the autoimmune process in T1D, including processing of the major autoantigen, pre-proinsulin, into peptide epitopes, some of which contain signal peptide amino acids[22].

We surveyed the National Human Genomes Research Institute (NHGRI) GWAS catalog[23] to determine overlap of diseases or traits with the seven newly discovered loci. After removing diseases curated in ImmunoBase, we found that the 17q21.31 variant (rs17564829), in intron 1 of the *MAPT* gene (encoding microtubule-associated protein τ), was in strong LD ($r^2 > 0.9$) with the index SNP for several neurodegenerative diseases, including Parkinson disease. We also examined two expression quantitative trait locus (eQTL) data sets in relevant tissues[24,25] for overlap with our seven newly identified T1D associations. SNP rs17564829 in the 17q21.31 region was associated with expression of *NSF*, *KANSL1*, *ARHGAP27* and the RNA gene *MGC57346*. This region overlaps a set of haplotypes in high LD that incorporate duplication and inversion events[21], complicating further interpretation. No other identified genes had strong functional candidacy.

It is well established that the SNPs showing the strongest association with disease in any region are not necessarily the causal variants, owing to a combination of sampling variation and LD. Nevertheless, the dense coverage of the Immunochip increases the likelihood that causal variants are among the SNPs genotyped in the T1D-associated loci. Although putative causal variants cannot be identified without further experimentation, identification of the most strongly associated SNPs in each region allowed us to integrate the location of these SNPs and their flanking sequences with emerging knowledge of the regulatory sequences of the genome. Focusing on primary and conditional signals in each of the 44 loci listed in **Table 1**, we used a Bayesian approach similar to that described previously[6] to define the 99% credible set of SNPs within which the causal variants are most likely to be present (**Supplementary Table 1**).

Table 3 Pleiotropic SNPs associated with T1D

						Disease association		T1D association			
Index SNP	Chr.	Position (bp)	MAF	Alleles	Index disease	OR	*P*	OR	*P*	Candidate gene(s)	Reference
rs17264332[a]	6q23.3	138,005,515	0.22	A>G	CEL	1.29	5.00×10^{-30}	1.12	8.26×10^{-6}	*TNFAIP3*	32
rs6920220[a]	6q23.3	138,006,504	0.22	G>A	UC	1.16	1.40×10^{-21}	1.12	7.26×10^{-6}	*TNFAIP3*	33,34
rs6920220[a]	6q23.3	138,006,504	0.22	G>A	RA	1.20	2.30×10^{-13}	1.12	7.26×10^{-6}	*TNFAIP3*	35
rs911263	14q24.1	68,753,593	0.29	T>C	PBC	0.79	9.95×10^{-11}	0.89	4.93×10^{-6}		16
rs17564829[b]	17q21.31	44,006,601	0.195	T>C	PBC	1.25	2.15×10^{-9}	0.89	6.77×10^{-6}		16

We show all genome-wide significant index SNPs for immune-mediated diseases[15,32–35] that are in regions not associated with T1D at genome-wide significance but associated at $P < 1 \times 10^{-5}$ in the case-control analysis presented here. Shown are the index SNPs, diseases and the single-SNP association test statistics for each index disease and T1D. Chromosome positions are given according to GRCh37. Chr., chromosome; MAF, minor allele frequency; OR, odds ratio; CEL, celiac disease; UC, ulcerative colitis; RA, rheumatoid arthritis; PBC, primary biliary cirrhosis.
[a]rs17264332 is in LD with rs6920220, $r^2 = 1$. [b]rs17564829 is in LD with rs1052553 in **Table 1**, $r^2 = 0.99$.

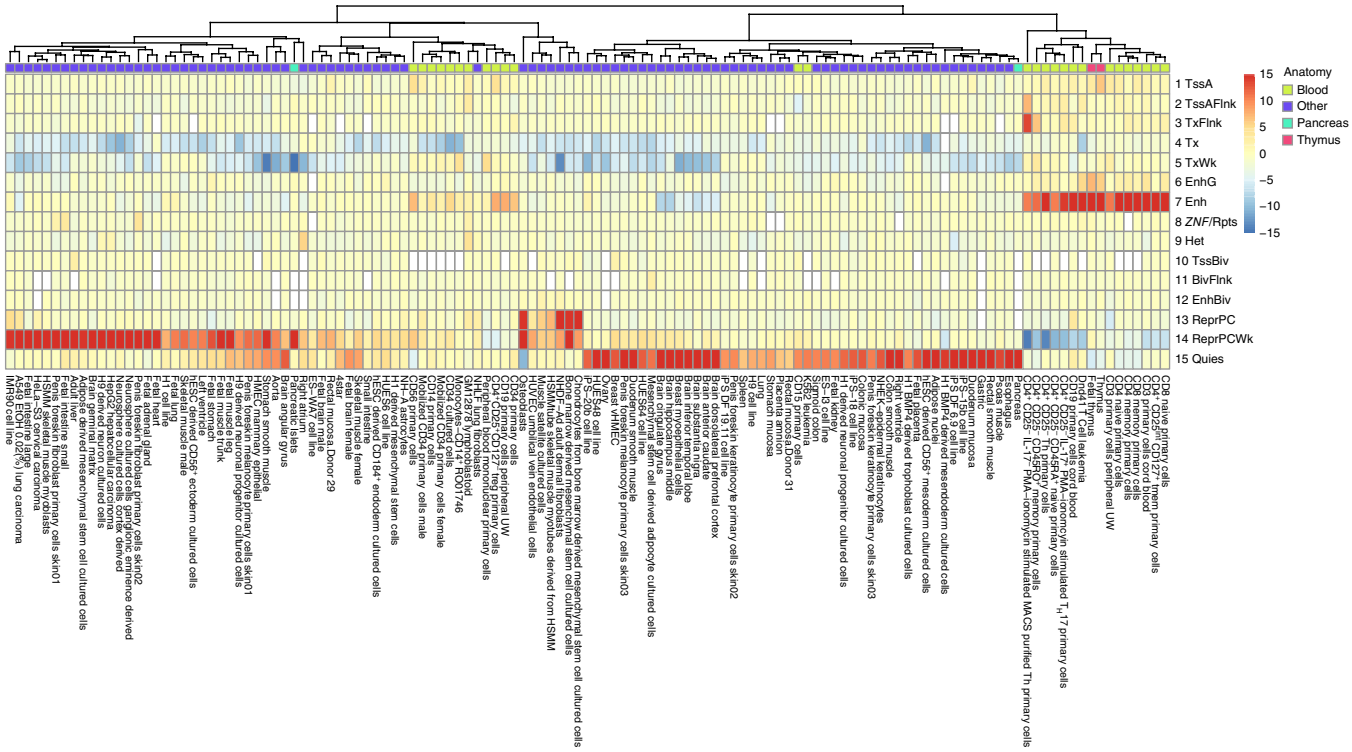

Figure 2 Heat map showing chromatin state enrichment analysis of the T1D 99% credible SNP set in Immunochip densely mapped regions versus the complement set within the Epigenomic Roadmap and ENCODE tissues. The top colored row groups cell types into four anatomical categories with relevance to T1D. Subsequent rows use a red (enrichment) to blue (depletion) scale to illustrate enrichment in a particular chromatin state (1, TssA (active transcription start site); 2, TssAFlnk (flanking active transcription start site); 3, TxFlnk (transcribed at the 5′ and 3′ ends of the gene); 4, Tx (strong transcription); 5, TxWk (weak transcription); 6, EnhG (genic enhancer); 7, Enh (enhancer); 8, *ZNF*/Rpts (*ZNF* genes and repeats); 9, Het (heterochromatin); 10, TssBiv (bivalent/poised transcription start site); 11, BivFlnk (flanking bivalent transcription start site or enhancer); 12, EnhBiv (bivalent enhancer); 13, ReprPC (repressed Polycomb); 14, ReprPCWk (weakly repressed Polycomb); 15, Quies (quiescent or low expression)).

We used the set of credible SNPs to interrogate 15 chromatin states across 127 tissues derived from the Epigenomics Roadmap and Encyclopedia of DNA Elements (ENCODE) projects[26]. We observed a strong enrichment of SNPs in enhancer chromatin states in immunologically relevant tissues (**Fig. 2**). Thymus, CD4$^+$ and CD8$^+$ T cells, B cells and CD34$^+$ stem cells exhibited the strongest enrichment in more than one sample of each tissue or cell type. There was less evidence of enrichment in promoter sequences (**Fig. 2**), suggesting that variation in enhancer sequences is more relevant to T1D. Our Bayesian approach is more informative in selecting relevant SNPs than the conventional r^2-based approach that focuses on SNPs showing r^2 >0.8 with index SNPs—the r^2-based approach only identified enhancer enrichment in one subtype of CD4$^+$ T cells (data not shown). Recently, an analysis of active gene enhancers across multiple tissues reported enrichment of T1D GWAS SNPs in promoters, not enhancers[27]. This difference could be attributable to the empirical technique used in defining enhancers or the focus of this analysis on enhancers in general (rather than tissue-specific enhancers), a failure to adjust for potential confounding by minor allele frequency (MAF) or reliance on the r^2-based approach rather than establishing a credible set of putatively causal SNPs. Our analyses found no evidence of enrichment in pancreatic islet enhancers, a result supported by a recent detailed analysis of pancreatic islets that found evidence for enrichment of GWAS signals for type 2 diabetes and fasting glucose levels in a subset of those enhancers but not T1D signals[28].

We also investigated whether analysis of available chromatin state data and accompanying annotation could narrow our credible SNP lists and highlight certain genes and SNPs. We focused on credible SNPs that were either nonsynonymous or missense (as annotated by VEP[19] Ensembl v75) or that overlapped enhancer regions in the tissues that showed an enrichment for T1D-associated SNPs in **Figure 2** (**Supplementary Data Set**). Although credible SNP sets can be large, this filtering reduced their median size from 28 to 8 SNPs (**Supplementary Fig. 1**). We highlight 29 SNPs corresponding to 12 regions for which the size of the filtered sets was relatively small (<5 SNPs) in **Supplementary Table 2**. The analyses did not identify any new candidate genes other than the known candidate causal genes containing high-confidence missense variants: *PTPN22*, *IFIH1*, *CTSH*, *TYK2* and *FUT2*. Nevertheless, these analyses did identify SNPs that overlapped potential enhancers near *CTSH*, *TYK2* and *UBASH3A* that are worthy of specific laboratory investigations. In addition, we identified candidate enhancer SNPs in four other regions (6q22.32, 7p12.1, 10q23.31 and 16q23.1), none of which had obvious candidate genes (**Table 1** and **Supplementary Data Set**). Chromosome conformational capture can be used to directly determine the presence of physical interactions between promoters and potential enhancer sequences[29] in the most enriched primary cell types using our credible SNP variants. There was a discrete cluster of credible enhancer SNPs 5′ of the functional candidate gene *IL10* (**Supplementary Data Set**), yet this potential regulatory sequence could interact with the promoter of the adjacent candidate gene *IL19* (or with both). Genome-wide analysis of promoter-enhancer interactions will help identify new candidate causal genes[30,31]. Notwithstanding the current lack of data on promoter-enhancer interactions, these analyses identify *AFF3* (2q11.2) and *BCAR1* (16q23.1) as new candidate genes for T1D.

URLs. ImmunoBase, http://www.immunobase.org/; T1DBase, http://www.t1dbase.org/; wgsea, http://cran.r-project.org/web/packages/wgsea/index.html; Blood eQTL browser (data deposition 21 December 2012), http://genenetwork.nl/bloodeqtlbrowser/2012-12-21-CisAssociationsProbeLevelFDR0.5.zip; NHGRI GWAS catalog (accessed 19 February 2014), http://www.genome.gov/admin/gwascatalo.txt; Epigenomic Roadmap annotations, https://sites.google.com/site/anshulkundaje/projects/epigenomeroadmap.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Immunochip data for UK GRID cases and for T1DGC ASP and trio families have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession phs000180.v2.p2. Immunochip data for the British 1958 Birth Cohort, the UK National Blood Service and the NIHR Cambridge Biomedical Research Centre Cambridge BioResource have been deposited in the European Genome-phenome Archive (EGA) under accession EGAS00000000038.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research uses resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the National Institute of Allergy and Infectious Diseases (NIAID), the National Human Genome Research Institute (NHGRI), the National Institute of Child Health and Human Development (NICHD) and JDRF and supported by grant U01 DK062418 from the US National Institutes of Health. Further support was provided by grants from the NIDDK (DK046635 and DK085678) to P.C. and by a joint JDRF and Wellcome Trust grant (WT061858/09115) to the Diabetes and Inflammation Laboratory at Cambridge University, which also received support from the NIHR Cambridge Biomedical Research Centre. ImmunoBase receives support from Eli Lilly and Company. C.W. and H.G. are funded by the Wellcome Trust (089989). The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust Strategic Award (100140).

We gratefully acknowledge the following groups and individuals who provided biological samples or data for this study. We obtained DNA samples from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council and the Wellcome Trust. We acknowledge use of DNA samples from the NIHR Cambridge BioResource. We thank volunteers for their support and participation in the Cambridge BioResource and members of the Cambridge BioResource Scientific Advisory Board (SAB) and Management Committee for their support of our study. We acknowledge the NIHR Cambridge Biomedical Research Centre for funding. Access to Cambridge BioResource volunteers and to their data and samples are governed by the Cambridge BioResource SAB. Documents describing access arrangements and contact details are available at http://www.cambridgebioresource.org.uk/. We thank the Avon Longitudinal Study of Parents and Children laboratory in Bristol, UK, 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. This study makes use of data generated by the Wellcome Trust Case Control Consortium, funded by Wellcome Trust award 076113; a full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/.

AUTHOR CONTRIBUTIONS

The study was conceptually designed by M.J.D., J.C.B., P.D., J.A.T., C.W., P.C. and S.S.R. The study was implemented by S.O.-G., E.F., H.S., N.M.W., P.D., T1DGC, J.A.T., C.W., P.C. and S.S.R. DNA samples were managed by S.O.-G., E.F. and H.S. Genotyping and laboratory quality control were conducted by S.O.-G., E.F. and P.D. Statistical quality control methods were implemented by W.-M.C., M.S., N.J.C., H.G. and J.C.M. Statistical analyses were performed by W.-M.C., A.R.Q., J.C.M., J.D.C., O.B., J.K.B., N.J.C., M.D.F. and C.W. Chromatin state analyses were conducted by O.B., L.D.W., A.K. and M.K. ImmunoBase is maintained by O.B., E.S. and P.A. The manuscript was written by S.O.-G., W.-M.C., A.R.Q., O.B., J.A.T., C.W., P.C. and S.S.R. All authors reviewed and contributed on the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. 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).
2. Bradfield, J.P. *et al.* A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **7**, e1002293 (2011).
3. Virgin, H.W. & Todd, J.A. Metagenomics and personalized medicine. *Cell* **147**, 44–56 (2011).
4. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7**, e1002254 (2011).
5. Smyth, D.J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **359**, 2767–2777 (2008).
6. Wellcome Trust Case Control Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
7. Genuth, S. *et al.* Follow-up report on the diagnosis of diabetes mellitus. *Diabetes Care* **26**, 3160–3167 (2003).
8. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
9. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
10. Todd, J.A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39**, 857–864 (2007).
11. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.* **35**, 34–41 (2006).
12. 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).
13. Dendrou, C.A. *et al.* Cell-specific protein phenotypes for the autoimmune locus *IL2RA* using a genotype-selectable human bioresource. *Nat. Genet.* **41**, 1011–1015 (2009).
14. Concannon, P. *et al.* Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the Type 1 Diabetes Genetics Consortium. *Diabetes* **58**, 1018–1022 (2009).
15. Zhang, Z. *et al.* Two genes encoding immune-regulatory molecules (*LAG3* and *IL7R*) confer susceptibility to multiple sclerosis. *Genes Immun.* **6**, 145–152 (2005).
16. Liu, J.Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44**, 1137–1141 (2012).
17. 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).
18. Barratt, B.J. *et al.* Remapping the insulin gene/*IDDM2* locus in type 1 diabetes. *Diabetes* **53**, 1884–1889 (2004).
19. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
20. Heinig, M. *et al.* A *trans*-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* **467**, 460–464 (2010).
21. Boettger, L.M., Handsaker, R.E., Zody, M.C. & McCarroll, S.A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**, 881–885 (2012).
22. Kronenberg, D. *et al.* Circulating preproinsulin signal peptide–specific CD8 T cells restricted by the susceptibility molecule HLA-A24 are expanded at onset of type 1 diabetes and kill β-cells. *Diabetes* **61**, 1752–1759 (2012).
23. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
24. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
25. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
26. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
27. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
28. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
29. Davison, L.J. *et al.* Long-range DNA looping and gene expression analyses identify *DEXI* as an autoimmune disease candidate gene. *Hum. Mol. Genet.* **21**, 322–333 (2012).
30. Dryden, N.H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* **24**, 1854–1868 (2014).
31. Hughes, J.R. *et al.* Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
32. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
33. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
34. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
35. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).

ONLINE METHODS

Samples. ASP families were collected by the T1DGC from five geographic regions through four recruitment networks. Recruitment criteria for the families have been discussed previously[36]. A total of 6,808 T1D case samples were ascertained from the UK GRID cohort[10]. Control samples were obtained from the British 1958 Birth Cohort ($n = 6,929$)[11] and the UK National Blood Services collection ($n = 3,060$)[12] and the NIHR Cambridge Biomedical Research Centre Cambridge BioResource ($n = 2,846$)[13]. Many of these samples (98% of cases, 59% of controls and 57% of family samples) were also used in an earlier GWAS meta-analysis that initially identified many of the T1D-associated regions[1]. All samples included in this analysis have reported or self-declared European ancestry. All DNA samples were collected after approval from relevant institutional research ethics committees. Review boards of all contributing institutions approved all protocols and informed consent for sharing of data and sample collection; appropriate informed consent was obtained from all subjects and families.

Genotyping and quality control. Genotyping was performed using a custom high-density genotyping array, the Immunochip (Illumina), according to the manufacturer's protocols. The Immunochip, a custom Illumina Infinium HD array, was designed to densely genotype, using 1000 Genomes Project data and any other available disease-specific resequencing data, immune-mediated disease loci identified by common variant GWAS. The Immunochip Consortium selected 186 distinct loci containing markers meeting genome-wide significance ($P < 5 \times 10^{-8}$) from 12 such diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, T1D and ulcerative colitis). All 1000 Genomes Project pilot phase[37] CEU population variants (September 2009 release) within 0.1-cM recombination blocks (HapMap 3 CEU) around the lead marker for each GWAS region were submitted for array design. No filtering on correlated variants (LD) was applied. Additional content included regional resequencing data (submitted by several groups) as well as a small proportion of investigator-specific undisclosed content including intermediate GWAS results.

All individuals from T1DGC ASP and trio families ($n = 11,584$), T1D cases ($n = 6,808$) and British 1958 Birth Cohort controls ($n = 5,452$) were genotyped at the Genome Sciences Laboratory within the Center for Public Health Genomics at the University of Virginia. An additional 1,477 control samples from the British 1958 Birth Cohort, 2,846 samples from the NIHR Cambridge Biomedical Research Centre Cambridge BioResource and 3,060 UK National Blood Service samples were genotyped at the Wellcome Trust Sanger Institute. The Illumina GeneTrain2 algorithm was used to cluster genotypes.

Sample and SNP quality control analyses for the family data set and the case-control data set were performed separately. Initial sample quality control metrics included sample call rate, heterozygosity and concordance for reported versus genotyped sex. Relationship and population structure inference analyses were performed, and the inferred relationship and population membership for each individual determined from the genetic data were compared to the self-reported pedigree and ancestry data (see sections on population inference and population structure for more details). A total of 34 cases, 192 controls and 20 individuals in T1DGC ASP families were removed for missing rate >5%. Approximately 2,000 SNPs on the X chromosome and Y chromosome were used to infer sex on the basis of the genetic data. Individuals with low X chromosome heterozygosity and a large number of Y-chromosome SNPs were defined as 'males'; individuals with high X-chromosome heterozygosity and a small number of Y-chromosome SNPs were defined as 'females'. Inconsistency between self-reported sex and genetically determined sex for any individual was considered an error in sex. From this analysis, 39 T1D cases, 79 controls and 59 individuals in T1DGC ASP families were removed. Samples with heterozygosity outside the range of 19–23.5% were removed, including 7 cases and 19 controls. A further 75 cases and 201 controls were removed for other reasons, comprising sample duplication, inability to map sample IDs to demographic information, relatedness (see below) and population structure. A total of 6,683 cases, 12,173 controls, 2,601 ASP families and 69 trio families (10,796 total individuals) were used for analysis following quality control.

Monomorphic SNPs (~23,000) were identified and removed. A total of 527 SNPs in cases, 2,405 SNPs in controls and 1,387 SNPs in T1DGC ASP and trio family data were rejected owing to failure to attain a genotyping rate of at least 95%. An additional 618 SNPs in the case and control data were removed owing to a low genotyping rate at less frequent and rare variants (genotyping rate <99% for SNPs with MAF <1% or genotyping rate less than (1 – MAF) for SNPs with MAF <5%). In the case and control collections, 1,432 SNPs failed Hardy-Weinberg equilibrium tests (with Hardy-Weinberg equilibrium $P < 1 \times 10^{-6}$) in controls, and 527 SNPs failed (with Hardy-Weinberg equilibrium $P < 1 \times 10^{-10}$) in cases. In the ASP families, 2,939 SNPs failed with mendelian inconsistency (MI) errors (with a standard MI error rate >0.5% or an adjusted MI error rate >5% for rare variants). A total of 163,924 SNPs passed quality control metrics in the case and control collections, and 164,643 SNPs passed quality control metrics in the families. Of these sets of SNPs, 154,939 SNPs overlapped and were used for initial analyses. The first iteration of identifying the best markers for dense regions produced a large number of markers with visually identified noisy signal clouds. As a result, further SNP quality control was undertaken, whereby the call rate cutoff was increased to 99% and the Hardy-Weinberg equilibrium cutoff was decreased to $P < 1 \times 10^{-4}$. A further 8,349 SNPs were removed for lower call rate, 10,708 were removed for violation of Hardy-Weinberg equilibrium and 34 were removed for manually identified poor signal clouds. This strategy reduced the total number of SNPs analyzed to 135,870 and produced top SNPs with much cleaner signal cloud data.

We observed inflation of test statistics across all SNPs that passed quality control ($\lambda_{1,000} = 1.09$), which was expected as the Immunochip was designed to target robustly defined immune-mediated disease susceptibility loci. Excluding SNPs from regions reported here, λ_{1000} was reduced to 1.07; excluding all densely genotyped regions reduced λ_{1000} to 1.03.

Relationship inference. Cryptic relatedness can confound the result of population structure and association analyses and can lead to inflated type I error rates. We used the relationship inference method that was implemented in KING[8] to estimate the kinship coefficient between every pair of individuals on the basis of their SNP data. Because only SNPs for these two individuals were used when the kinship coefficient was estimated for a pair of individuals, the estimation accuracy was independent of the population structure for the entire data.

Twenty-two autosomes are well covered on the Immunochip array; thus, the SNP density provides sufficient power to correctly identify close relationships (first and second degree) with extremely low numbers of false positives (i.e., to separate unrelated pairs from close relatives)[7]. After cryptic relatedness was identified, pedigree errors were resolved by removing problematic individuals (within families) and/or by reconstructing the pedigree (both within and across families) incorporating the newly identified first- and second-degree relationships.

A total of 30 individuals were removed in family data owing to inconsistency between the estimated and documented relationships, and ~500 pairs of first-degree relatives that were not reflected in the documented pedigree have been incorporated in the pedigree data by pedigree reconstruction. All pairwise relationships in families after quality control are shown in **Supplementary Figure 2**. The estimated kinship coefficient of each pair of relatives is plotted against the proportion of zero identity by state (IBS), with the documented relationships being indicated by color. All 42 pairs of documented identical twins had estimated kinship coefficients >0.4. Among 16,292 documented first-degree relative pairs, 16,270 pairs had estimated kinship coefficients between 0.177 and 0.36 (the criterion to be inferred as first-degree relatives in KING), 21 pairs had estimated kinship coefficients between 0.150 and 0.177, and 1 pair had an estimated kinship coefficient of 0.137. After pedigree reconstruction, there was no first-degree relatedness across any two families, and there were only three pairs of documented unrelated pairs with estimated kinship coefficients >0.1 (all three kinship coefficients < 0.139). In the analyzed data, a total of 10,796 individuals from 2,682 nuclear families had genotypes available. There were 1,670 families with both parents available, 652 families with only one parent available and 360 individuals with neither parent available. The distribution of affected siblings was 69 families with 1 affected sibling, 2,490 families with 2 affected siblings, 104 families with 3 affected siblings, 5 families with 4 affected siblings and 2 families with 5 affected siblings.

In the T1D cases and the UK control data, 159 controls and 48 cases were removed for being close relatives. After this level of quality control, no

remaining 'unrelated' pairs in the case or control data had estimated kinship coefficients >0.09, indicating that all individuals were indeed unrelated. We also checked the UK T1D cases and UK controls for relatedness in the T1DGC ASP and trio family data set, as one of the four T1DGC collection sites was in the UK. A total of five pairs of individuals were identified with a genotype concordance rate >99.99%; the related individuals were selectively removed from the T1DGC family data set.

Population structure. We applied the principal-component analysis (PCA) method that is implemented in KING[38] for the identification of the population structure. We combined HapMap 3 data (1,097 unrelated individuals were used[39], with 215 of European ancestry) with each cohort. We kept SNPs that were present on both the HapMap and Immunochip panels and removed SNPs with r^2 >0.5 with other SNPs. After applying the quality control filters, ~30,000 SNPs were used for the structure analysis. PCA was first carried out among the HapMap individuals only, and each Immunochip individual was then projected onto the space that was expanded by the principal components of the HapMap individuals. The projected principal components for each individual sample represent its ancestry relative to the HapMap populations. Using this algorithm, we obtained the principal components for case-control individuals by cohort, projected onto either the entire HapMap 3 populations (**Supplementary Fig. 3**) or the European-ancestry populations only, including CEU and TSI (**Supplementary Fig. 4**); we also obtained the principal components for individuals in the family data set (**Supplementary Fig. 5**).

The population structure of our case-control data is compared with that of all HapMap 3 populations in **Supplementary Figure 3**. A total of 69 individuals were identified as being greater than 3 s.d. from the average of the second principal component in European populations, and these outliers were excluded from analysis. The principal components of all case-control individuals from four cohorts (UK GRID and British 1958 Birth Cohort genotyped at the University of Virginia and British 1958 Birth Cohort and National Blood Service genotyped at the Wellcome Trust Sanger Institute) were in the range of the European-ancestry populations, clearly separated from non-European populations. Case-control individuals are compared with European populations only, including CEU and TSI, in **Supplementary Figure 4**. The cluster on the left is for CEU that represents northern Europeans, and the cluster on the right is for TSI, which represents southern Europeans. A total of 55 outliers were identified in this analysis clustering with the southern Europeans and were excluded before analysis. The results in **Supplementary Figure 5** suggest that there was no substructure difference between cases (UK GRID) and controls (British 1958 Birth Cohort) genotyped on the Immunochip at the University of Virginia with those controls (British 1958 Birth Cohort and UK National Blood Service) genotyped at the Wellcome Trust Sanger Institute. The population structure in the family data, in comparison to the HapMap populations, is shown in **Supplementary Figure 5**. Only individuals of European ancestry were used in the analysis.

SNP annotation. The chromosomal locations of the Immunochip SNPs were standardized to Build 37 (hg19) coordinates using the UCSC liftOver tool. For each variant, the SNP alleles have been normalized so that the reference and alternate alleles are reported on the reference (top) strand.

Single-SNP association analysis. To test the association between each SNP and T1D, we applied the Generalized Disequilibrium Test (GDT) method[39] to the T1DGC ASP and trio family data and fit a logistic regression to the T1D case and control data. We then combined the family and case-control data using meta-analysis.

The GDT method computes the genotype difference between all pairs of phenotypically discordant relatives within each family. This method uses the information from all discordant relative pairs, including data for nuclear families that are not efficiently used in family-based tests such as the Transmission/Disequilibrium Test (TDT) or the Family-Based Association Test (FBAT). To estimate the effect at each variant, we carried out a TDT at each region and approximated the OR of a variant by the transmission/non-transmission ratio at this region observed in parent–affected offspring trios. In the logistic regression model for T1D in the case-control data, association between T1D and an additive genotype score at each SNP was performed with adjustment for sex and regions in UK (12 dummy variables created for the 13 regions)[40]. The snp.rhs.estimates function from package snpS in R 3.0.2 was used for analysis[41].

Meta-analysis. A weighted z score was used to combine results from the case-control and family data[42]. An overall β coefficient and standard error were computed as the weighted average of the individual β statistics, and a corresponding P value for that statistic was computed. Weights were proportional to the inverse variance (1 divided by the standard error squared) in each study and were scaled by the meta-variance (σ^2_{meta}) defined as follows

$$\sigma^2_{\text{meta}} = 1/(1/\sigma^2_{\text{cc}} + 1/\sigma^2_{\text{fam}})$$

where σ^2_{cc} is the variance from the case-control series and σ^2_{fam} is the variance from the family collection, so that the weights summed to 1. For the family data, instead of using the total number of family members, we used twice the number of parent–affected offspring trios as the effective sample size for the meta-analysis.

Conditional analysis to identify secondary signals. To determine whether additional SNPs within a region were significantly associated with T1D, independently of the most strongly associated SNP identified in the primary analysis, we performed conditional analysis using the case-control data. For each T1D region, the conditional analysis started with the SNP that was the most statistically significant as identified in the meta-analysis. A new logistic regression model was fitted to the case-control data, adjusting for the previously identified SNP as a covariate. We repeated this procedure until no SNPs in the region attained our threshold for statistical significance.

Overlap of T1D with other autoimmune diseases. For each disease in ImmunoBase, we downloaded the set of curated index SNPs (http://www.immunobase.org/page/RegionsLanding; accessed 13 February 2014). We excluded inflammatory bowel diseases, as this is a combination of ulcerative colitis and Crohn's disease, which are summarized individually. The MHC region (chr. 6: 25–35 Mb, GRCh37) was excluded from analysis. For each disease in turn, we used the index SNPs to label each densely mapped region of the Immunochip as associated or not with the index disease. After LD pruning ($r^2 \leq 0.95$) to remove excessive correlation, distributions of T1D association meta-analysis P values for SNPs were compared between the two sets or regions using a non-parametric Wilcoxon rank-score test, as implemented in the R package wgsea[43]. LD between SNPs inflates the variance of the test statistic, so we estimated this variance empirically under the null hypothesis of no association using 10,000 permutations of case versus control status. Given overall significant evidence of shared or disparate genetic architecture, we examined which loci were involved by summarizing the evidence for T1D association in a region using $P = \min(-\log(P))$ over all SNPs in a given densely genotyped region.

eQTL and GWAS catalog overlap in seven newly discovered regions. To define a query SNP set, we took a 2-Mb window centered on each newly discovered index SNP and then filtered out overlapping SNPs on the basis of an LD threshold of $r^2 \geq 0.9$ with the index SNP, using 1000 Genomes Project data. To identify potential *cis* eQTL overlap, we downloaded summary statistics from Fairfax *et al.* (Table S7 in ref. 44) and Westra *et al.*[25] (Blood eQTL browser; see URLs) and computed overlap with the query SNP set. For each significant overlap, we computed the LD with the top eQTL SNP for that probe or tissue, again using 1000 Genomes Project data. To look for trait or disease overlap outside the scope of ImmunoBase, we used the query SNP set to examine overlap with the NHGRI GWAS catalog[23].

Credible sets of causal variants. For each index SNP (**Table 1**), we considered all SNPs within a 50-kb window and used the case-control data to compare models containing the index SNP i or each alternative SNP j using approximate Bayes factors, by the relationship

$$-2\log(\text{ABF}_{ij}) = \text{BIC}_i - \text{BIC}_j$$

doi:10.1038/ng.3245

where ABF_{ij} is the approximate Bayes factor comparing models containing SNPs i and j, and BIC_i is the Bayesian information criterion (BIC) calculated from a logistic model of case/control status against SNP i. For simplicity, this analysis was performed using only the case-control cohort. For multiple-SNP models, we considered the conditional SNPs as fixed; for example, for chromosome 10p15.1, when considering rs10795791 as an index SNP and conditioning on rs61839660, we calculated BICs for the index model containing rs61839660 and rs10795791 and for all alternative two-SNP models containing rs61839660 and another SNP within a 50-kb window centered on rs10795791.

For any interval, we estimated the probability that any individual SNP j was the causal variant responsible for that signal (again, including conditional models, where appropriate) by the posterior probability

$$PP_j = BIC_j / \text{sum}(BIC_j)$$

and we thus created a 99% credible set of SNPs as the smallest set of SNPs with total posterior probability ≥99%.

Enrichment analysis. Epigenomic Roadmap annotations were downloaded from the web portal. These were processed using R and Bioconductor packages to annotate the Immunochip SNPs overlapping tissue-specific functional elements. According to the credible sets formed above, the Immunochip SNPs that passed quality control could be divided into two sets: A, those that were in any credible set within densely mapped regions on the Immunochip, i.e., potential causal variants (n = 1,256); B, their complement within densely mapped regions on the Immunochip, i.e., variants unlikely to be causal (n = 78,692). We tested for enrichment of T1D signals in enhancers in each cell type in turn by forming a series of 2 × 2 contingency tables, stratified by a SNP's MAF in controls (<0.05, <0.1, <0.2, <0.3, <0.4 or <0.5), showing the overlap of SNPs in sets A and B with functional elements according to physical location. The stratification was important to control for confounding, as both enhancer presence/absence and membership of a SNP in a credible set were associated with MAF. We used Cochran-Armitage tests, with the Mantel extension, to test for association. The sign of the score statistic determined the direction of association.

Filtering of credible SNPs. To create a filtered set of credible SNPs that could be targeted in future functional studies, we first expanded the sets by considering all neighboring SNPs in the 1000 Genomes Project CEU release that did not pass genotyping on the Immunochip. These 1000 Genomes Project SNPs were assigned to credible sets if the Immunochip SNP with which they showed the strongest LD, according to r^2 value, was in a credible set. For each set, we calculated the size of the expanded credible set, the number of SNPs in the credible set that overlapped enhancers in tissues that showed enrichment according to **Figure 2** and the number of SNPs that were nonsynonymous. These results are presented in **Supplementary Table 1**.

Evidence for T1D association conditional on genome-wide significant association in another autoimmune disease. Loci have previously been assigned as being associated with T1D on the basis of association $P < 1 \times 10^{-4}$ for a SNP that also showed association $P < 5 \times 10^{-8}$ in another autoimmune disease[5]. Here we explored the strength of evidence these thresholds provide, on the basis of previous work[45]. For any individual SNP and two diseases, there exist four hypotheses:

H_0: The SNP is not associated with either disease.
H_1: The SNP is associated with only disease 1.
H_2: The SNP is associated with only disease 2.
H_{12}: The SNP is associated with both disease 1 and disease 2.

Realistic prior probabilities[45] are

$$\pi_0 = 1\text{–}2 \times 10^{-4} \text{ to } 1 \times 10^{-5}$$
$$\pi_1 = 1 \times 10^{-4}$$
$$\pi_2 = 1 \times 10^{-4}$$
$$\pi_{12} = 1 \times 10^{-5}$$

which imply that we expect about 1 in 1,000 SNPs to show association with either disease and, of the SNPs associated with 1 disease, we expect about 1 in 10 to be associated with both diseases.

Posterior probabilities for independent data sets. We used the approximate Bayes factors presented previously[46] to estimate ϕ_i, the Bayes factor for association with disease i compared to no association with disease i given only single-SNP P values and the MAF of the SNP in controls. If we assume that the case and control data sets for each disease are independent, they can be combined to calculate Bayes factors for each hypothesis.

$$BF_0 = 1$$
$$BF_1 = \phi_1$$
$$BF_2 = \phi_2$$
$$BF_{12} = \phi_1\phi_2$$

Thus, the posterior probability for each hypothesis is given as

$$PP_0 = \pi_0/B$$
$$PP_1 = \pi_1\phi_1/B$$
$$PP_2 = \pi_2\phi_2/B$$
$$PP_{12} = \pi_{12}\phi_1\phi_2/B$$

where $B = 1 + \phi_1 + \phi_2 + \phi_{12}$. The conditional probability of association with disease 2, given that we believe there is association with disease 1, is

$$PP_{2|1} = PP_{12}/(PP_1 + PP_{12}).$$

Effect of shared versus independent controls. The Immunochip Consortium genotyped a large sample of shared UK controls. This induces correlation between the P values from different diseases[47], so BF_{12} cannot be expressed as a simple product of disease-specific Bayes factors. Methods to account for the effects from shared controls appear conservative[47], as they do not allow for the reasonable assumption that related diseases share genetic susceptibility variants. Instead, we used simulation to explore the effect of non-independence on $PP_{2|1}$. We used multinomial models and the approximate Bayes factor[48] to properly estimate the posterior probabilities of each hypothesis.

To explore the effect of shared controls, we considered two general scenarios, relating the sample sizes available in the Wellcome Trust Case Control Consortium and the Immunochip reports (**Supplementary Table 3**). Using p_i to denote the single-SNP P value for disease i, the results (**Supplementary Fig. 6**) show that, for independent controls, $PP_{2|1} > 0.9$ (median = 0.97) whenever $p_2 < 1 \times 10^{-4}$. However, for shared controls, we cannot be as confident of association. $PP_{2|1}$ is independent of p_1, given that we believe the association with disease 1 is real. The number of cases for each disease has a relatively minor effect on $PP_{2|1}$, whereas the MAF and the number of shared controls have slightly larger effects. Conditional posterior probabilities increase with MAF but decrease with an increasing number of shared controls. The strongest determinant is p_2, with $PP_{2|1}$ being in the interval 0.37–0.61 (median = 0.46) at $p_2 = 1 \times 10^{-4}$ for all scenarios. When $p_2 = 1 \times 10^{-5}$, $PP_{2|1}$ is in the interval 0.87–0.90 (median = 0.89), suggesting that a threshold of $p_2 = 10^{-5}$ may be more suitable for convincing evidence of association with a second autoimmune disease.

The corresponding R code is available at http://dx.doi.org/10.6084/m9.figshare.827246 and is based, in part, on functions from the R package colocCommonControl available at https://github.com/mdfortune/colocCommonControl.

36. Hilner, J.E. *et al.* Designing and implementing sample and data collection for an international genetics study: the Type 1 Diabetes Genetics Consortium. (T1DGC). *Clin. Trials* **7**, S5–S32 (2010).
37. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
38. Manichaikul, A. *et al.* Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. *PLoS Genet.* **8**, e1002640 (2012).
39. Chen, W.M., Manichaikul, A. & Rich, S.S. A generalized family-based association test for dichotomous traits. *Am. J. Hum. Genet.* **85**, 364–376 (2009).

40. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
41. Clayton, D.G. snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.10.0 (2012).
42. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
43. Wallace, C. wgsea: Wilcoxon based gene set enrichment analysis. R package version 1.8. http://CRAN.Rproject.org/package=wgsea (2013).
44. Fairfax, B.P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
45. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
46. Wakefield, J. Bayes factors for genome-wide association studies: comparison of *p*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
47. Zaykin, D.V. & Kozbur, D.O. *P*-value based analysis for shared controls design in genome-wide association studies. *Genet. Epidemiol.* **34**, 725–738 (2010).
48. Rafferty, A.E. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251–265 (1996).