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

Suna Onengut-Gumuscu*[1,2], Wei-Min Chen*[1,3], Oliver Burren*[4], 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], Luke D. Ward[5,6], Anshul Kundaje[5,6,7,8], 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]#, Chris Wallace[4,11]#, Patrick Concannon[1,12]#, and Stephen S. Rich[1,3]#.

[1] Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA, [2] Department of Medicine, Division of Endocrinology, University of Virginia, Charlottesville, VA, USA, [3]Department of Public Health Sciences, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, VA, USA, and [7]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA, USA. [4]JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, NIHR Biomedical Research Centre, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 0XY, UK. [5]Department of Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA; [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [7]Department of Genetics,

Stanford University, Stanford, CA, USA; [8]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [9]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [10]Authors listed in Supplementary Information.[11]MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, CB2 0SR, Cambridge, United Kingdom. [12]Current address: University of Florida Genetics Institute and Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA

* These authors contributed equally

# These authors jointly supervised this work.

Corresponding author: Stephen S. Rich

Center for Public Health Genomics

University of Virginia

P.O. Box 800717

Charlottesville, VA 22908 USA

Tel: 434-243-7356 Fax: 434-982-1815

email: ssr4n@virginia.edu

Genetic studies of type 1 diabetes (T1D) have identified 50 susceptibility regions[1,2] (www.T1DBase.org) revealing major pathways contributing to risk[3], with some loci shared across immune disorders[4-6]. In order to make genetic comparisons across autoimmune disorders as informative as possible a dense genotyping array, the ImmunoChip, was developed, from which four novel T1D regions were identified ($P < 5 \times 10^{-8}$). A comparative analysis with 15 immune diseases (www.ImmunoBase.org) revealed that T1D is more similar genetically to other autoantibody-positive diseases, most significantly to juvenile idiopathic arthritis and least to ulcerative colitis, and provided support for three additional novel T1D loci. Using a Bayesian approach, we defined credible sets for the T1D SNPs. These T1D 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.

Type 1 diabetes (T1D) results from the autoimmune destruction of the 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, in order to (i) identify additional risk loci, (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 GWAS and ImmunoChip results from multiple immune diseases to allow comparisons of their genetic risk profiles.

The T1D single nucleotide polymorphisms (SNPs) and indel content selected for inclusion on ImmunoChip was based on the 41 T1D regions known at the time (February, 2010)[1] and on 3,000 "wildcard" SNPs that tagged candidate genes or other SNPs with suggestive evidence ($5 \times 10^{-8} < P < 10^{-5}$) of association from T1D GWAS. In parallel, we collected and curated all available association results for immune diseases for which the ImmunoChip was designed. For efficient comparison and downstream analysis by the research community, we created a publicly available, integrated, web-based portal (ImmunoBase) that contains 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 sib-pair (ASP)[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 $P \leq 3.23 \times 10^{-7}$ (an ImmunoChip Bonferroni-corrected $P < 0.05$; **Table 1**). Thirty-eight of these are recognized T1D 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, and 5p13.2/rs11954020. rs11954020 is close to the multiple sclerosis (MS) candidate immune response gene, *IL7R*[15]. Two remaining loci, 17q21.31 and 21q22.3, were marginally associated ($P > 5 \times 10^{-8}$) and, as we describe later, additional support for 17q21.31 comes 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 associated or index SNP in each region, identified five loci with more than one independently associated SNP (**Table 1**). Four were already known to encode for more than one causal variant but the fifth region, 11p15.5 (*INS, INS-IGF2* candidate genes), was surprising as *INS* was the first non MHC region in T1D to be discovered[17], and therefore the region has been examined intensively. The likely causal candidates in this region are SNPs rs689/-23*Hph*I, rs3842753/+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 following quality control. We integrated pre-existing rs689 data with

ImmunoChip data in the 6,670 UK GRID cases and 6,304 British 1958 Birth Cohort controls, and found rs689 to be the most 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 annotation using VEP[19] (Ensembl v75) identifies rs3842753 as an *INS* non-synonymous SNP (His-Pro). However, we found limited evidence for the annotation of the underlying transcript isoform and it is more likely to be a non-coding 3´UTR SNP.

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 February 13, 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 and 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] (Supplementary Information). A difference in P-value distributions indicated that T1D showed stronger (or weaker) association with regions according to their association with the index disease.

This comparison clearly delineated diseases with characteristic autoantibodies (*e.g.*, juvenile idiopathic arthritis (JIA), rheumatoid arthritis (RA) and

T1D) compared to auto-inflammatory disorders (*e.g.*, ulcerative colitis (UC) and Crohn's disease (CD); **Table 2**; **Fig. 1A**). The strongest positive and negative enrichments were observed with JIA (**Fig. 1B**; P = 2 x 10^{-13}) and UC (**Fig. 1C**; P = 5.4 x 10^{-5}), respectively. It should be noted that the susceptibility loci for each disease remain incomplete and the extent of the incompleteness varies between diseases. This limitation prevents us from drawing any conclusion that 'T1D is more like RA than ATD'; however, individually significant results are likely valid representations of disease overlap. The overlap between T1D and JIA was driven, in part, by sharing ($P < 10^{-20}$) at 1p13.2/*PTPN22*, 12q24.11/*SH2B3*, and 10p15.1/*IL2RA* (**Fig. 1B** and **Fig. 1C**) whereas, for UC, 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 lesser associated ($5 \times 10^{-8} < P < 10^{-4}$) with T1D were considered T1D associated, and *vice versa*[5]. Here, we demonstrate (Supplementary Information) that a SNP with $P < 5 \times 10^{-8}$ in any ImmunoChip disease study requires $P < 10^{-5}$ for T1D to obtain a Bayesian posterior probability of T1D association > 0.9, given that different ImmunoChip disease studies shared many control samples. Using this analysis, we identified three additional T1D regions, bringing the number of known T1D regions to 57: 14q24.1/rs911263, 17q21.31/rs17564829 (that achieved Bonferroni correction, but not genome-wide significance in the primary analysis), and 6q23.3/rs17264332/rs6920220 **(Table 3)**.

The 6q23.3 region contains the well-recognized candidate gene *TNFAIP3*, linking T1D susceptibility with the proinflammatory tumour 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 gene. The gene content of the 17q21.31/rs17564829 region, containing a megabase-long 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 the processing of the major autoantigen, preproinsulin, into peptide epitopes some of which contain signal peptide amino acids[22].

We surveyed the NHGRI GWAS catalogue[23] to determine overlap between diseases and traits with the seven novel loci. After removing diseases curated in ImmunoBase, we found that 17q21.31/ rs17564829, in intron 1 of the *MAPT* (microtubule-associated protein tau) gene, is in strong LD (r^2>0.9) with the index SNP for several neurodegenerative diseases, including Parkinson's disease. We also examined two eQTL datasets in relevant tissues[24,25] for overlap with our seven newly identified T1D associations. rs17564829 in the 17q21.31 region associated with expression of *NSF*, *KANSL1*, *ARHGAP27* and *MGC5736*. This region overlaps a set of haplotypes in high LD that incorporate duplication and inversion events[21], complicating further interpretation. No other identified genes have strong functional candidacy.

It is well established that 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 loci. Although putative causal variants cannot be identified without further experimentation, identification of the most 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 associated region to define, for 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**).

We used the set of credible SNPs to interrogate 15 chromatin states across 127 tissues derived from the Epigenomics RoadMap and 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 of enhancer sequences is more relevant to T1D. Our Bayesian approach is more informative in selecting the relevant SNPs than the conventional r^2-based approach that focuses on SNPs with 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 in defining enhancers or their focus on enhancers generally, rather than tissue-specific enhancers, a failure to adjust for potential confounding by minor allele frequency, or reliance on the r^2-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 type 2 diabetes and fasting glucose GWAS signals in a subset of those enhancers, but not of T1D[28].

We also investigated whether analysis of available chromatin state data and its annotation could narrow our credible SNP lists and point to certain genes and SNPs. We focused on credible SNPs that were either non-synonymous/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 **Fig. 2** (**Supplementary Data Set**). While credible SNP sets can be large, this filtering reduced their median size from 28 to eight SNPs (**Supplementary Figure 1**). In **Supplementary Table 2**, we highlight 29 SNPs corresponding to 12 regions for which the size of filtered sets is relatively small (< 5). The analyses did not identify any new candidate gene, other than the known candidate causal genes containing high confidence missense variants: *PTPN22*, *IFIH1*, *CTSH*, *TYK2* and *FUT2*. Nevertheless, this analysis does identify SNPs that overlap 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 have 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[33] in the most enriched primary cell types using our credible SNP positions. There is a discrete cluster of enhancer credible 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 both). Genome-wide analysis of promoter-enhancer interactions will help identify new candidate causal genes[34,35] Notwithstanding the current lack of data on promoter-enhancer interactions, these analyses identify *AFF3* (2q11.2) and *BCAR1* (16q23.1) as novel 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, http://genenetwork.nl/bloodeqtlbrowser/2012-12-21-CisAssociationsProbeLevelFDR0.5.zip accessed 20/19/2014; NHGRI GWAS catalogue, http://www.genome.gov/admin/gwascatalog.txt accessed 20/19/2014; Epigenomic Roadmap annotations, https://sites.google.com/site/anshulkundaje/projects/epigenomeroadmap.

Accession codes. ImmunoChip data for UKGRID cases, T1DGC ASP and trio families are deposited in dbGaP and are available from http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000180.v2.p2; ImmunoChip data for British 1958 Birth Cohort, UK National Blood Service and the NIHR Cambridge Biomedical Research Centre Cambridge BioResource are deposited in the European Genome-phenome Archive (EGA) and are available from https://www.ebi.ac.uk/ega/home.

ACKNOWLEDGEMENTS

This research utilizes 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), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and JDRF and supported by U01 DK062418. Further support was provided by grants from the NIDDK (DK46635 and DK085678) to PC and a joint JDRF and Wellcome Trust (WT061858/09115) grant to the Diabetes and Inflammation Laboratory at Cambridge, which also received support from the National Institute for Health Research Cambridge Biomedical Research Centre. ImmunoBase receives support from Eli-Lilly and Company. CW and HG 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 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 SAB and Management Committee for their support of our study. We acknowledge the National Institute for Health Research Cambridge Biomedical Research Centre for funding. Access to Cambridge BioResource volunteers and their data and samples is 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 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, and a full list of the investigators who contributed to the generation of the data is available from 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 paper 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 to the final paper.

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**, 703-7 (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**, 2767-77 (2008).

6. Wellcome Trust Case Control, C. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294-301 (2012).

7. Genuth, S. *et al.* Follow-up report on the diagnosis of diabetes mellitus. *Diabetes Care* **26**, 3160-7 (2003).

8. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-73 (2010).

9. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (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**, 857-64 (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, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (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**, 1011-5 (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**, 1018-22 (2009).

15. Zhang, Z. *et al.* Two genes encoding immune-regulatory molecules (LAG3 and IL7R) confer susceptibility to multiple sclerosis. *Genes Immun* **6**, 145-52 (2005).

16. Liu, J.Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet* **44**, 1137-41 (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**, 176-83 (1984).

18. Barratt, B.J. *et al.* Remapping the insulin gene/IDDM2 locus in type 1 diabetes. *Diabetes* **53**, 1884-9 (2004).

19. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).

20. Heinig, M. *et al.* A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* **467**, 460-4 (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**, 881-5 (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 beta-cells. *Diabetes* **61**, 1752-9 (2012).

23. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (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**, 502-10 (2012).

25. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (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**, D930-4 (2012).

27. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-61 (2014).

28. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46**, 136-43 (2014).

29. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193-201 (2011).

30. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).

31. 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**, 246-52 (2011).

32. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* **44**, 1336-40 (2012).

33. 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**, 322-33 (2012).

34. Dryden *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Research* **24**:1854-1868 (2014).

35. Hughes *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* **46**:205-212 (2014).

Figure legends

Fig. 1: T1D ImmunoChip p-value enrichment analysis. Panel (a) shows 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 ImmunoChip. Those with negative scores indicate evidence for negative association. Each bar is labelled with the Wilcoxon rank sum test p-value and coloured by disease autoantibody positive/negative status. The MHC region (chr6:25Mb..35Mb GRCh37) was excluded from analysis. AA- Alopecia Areata, AS - Ankylosing Spondylitis ATD - Autoimmune thyroid disease,, CEL- Celiac disease, CD - 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. Panels (b) and (c) show *P'* = min(-log(p.t1d.meta)) for each densely typed region accessible on the ImmunoChip excluding the MHC and autosomal regions. Regions that overlap known T1D susceptibility regions are identified by blue bars, whereas yellow and pink show JIA and UC overlap respectively (http://www.ImmunoBase.org – accessed February 13, 2014). Red bars denote shared overlap between T1D and focal disease. The y-axis is truncated for clarity. A fully interactive version of panels (b) and (c), along with further supporting resources are available at http://www.immunobase.org/poster/type-1-diabetes-immunochip-study-onengut-gumuscu/.

Fig. 2: Heat map showing chromatin state enrichment analysis of T1D 99% credible SNP set in ImmunoChip densely mapped regions versus the complement set, within Epigenomic Roadmap and ENCODE tissues. The top coloured row groups cell-types into 4 anatomical categories with relevance to type 1 diabetes, subsequent rows use a red (enrichment) to blue (depletion) scale to illustrate enrichment in a particular chromatin state (1_TssA – Active Tss, 2_TssAFlnk – Flanking Active TSS, 3_TxFlnk – Transcribed at gene 5' and 3', 4_Tx – Strong transcription, 5_TxWk – Weak transcription, 6_EnhG – Genic Enhancer, 7_Enh - Enhancer, 8_ZNF/Rpts – ZNF genes & repeats, 9_Het - Heterochromatin, 10_TssBiv- Bivalent/Poised TSS, 11_BivFlnk – Flanking Bivalent TSS/Enhancer, 12_EnhBiv – Bivalent enhancer, 13_RepPC – Repressed PolyComb, 14_RepPCWk – Weak repressed polycomb, 15_Quies – Quiescent/Low).

Novel	Chromosome	Position	SNP	Alleles	MAF	OR	*P*	Condition	Candidate gene	Previous index SNPs (r^2)
	1p13.2	114377568	rs2476601	G>A	0.09	1.89	$< 10^{-100}$		*PTPN22*	rs2476601(1)
*	1q32.1	200814959	rs6691977	T>C	0.19	1.13	4.3 x 10^{-8}			--
	1q32.1	206939904	rs3024505	G>A	0.16	0.86	6.4 x 10^{-8}		*IL10*	rs3024493(1),rs3024505(1)
	2q11.2	100764087	rs13415583	T>G	0.35	0.90	1.1 x 10^{-7}		*AFF3*	rs6740838(0.32),rs9653442(0.41)
*	2q13	111615079	rs4849135	G>T	0.29	0.89	4.4 x 10^{-8}			--
	2q24.2	163110536	rs2111485	G>A	0.39	0.85	3.8 x 10^{-18}		*IFIH1*	rs1990760(0.91)
	2q24.2	163124637	rs35667974	T>C	0.02	0.59	9.3 x 10^{-9}	rs2111485	*IFIH1*	rs1990760(<0.1)
	2q24.2	163136942	rs72871627	A>G	0.01	0.61	2.4 x 10^{-6}	rs2111485, rs35667974	*IFIH1*	rs1990760(0.0094)
	2q33.2	204738919	rs3087243	G>A	0.45	0.84	7.4 x 10^{-21}		*CTLA4*	rs3087243(1),rs11571316(<0.1)
	3p21.31	46457412	rs113010081	T>C	0.11	0.85	4.6 x 10^{-8}		*CCR5*	rs333(0.34)
	4q27	123243596	rs75793288	C>G	0.36	1.15	5.6 x 10^{-13}		*IL2,IL21*	rs6827756(0.98),rs4505848(0.85)
*	4q32.3	166574267	rs2611215	G>A	0.15	1.18	1.8 x 10^{-11}			--
*	5p13.2	35883251	rs11954020	C>G	0.39	1.11	4.4 x 10^{-8}		*IL7R*	--
	6q15	90976768	rs72928038	G>A	0.17	1.20	6.4 x 10^{-14}		*BACH2*	rs11755527(0.194),rs597325(0.13)
	6q22.32	126752884	rs1538171	C>G	0.45	1.12	7.4 x 10^{-10}			rs9375435(0.96),rs9388489(0.98)
	7p12.2	50465830	rs62447205	A>G	0.28	0.89	2.5 x 10^{-8}		*IKZF1*	rs10272724(0.97)
	7p12.1	51028987	rs10277986	A>T	0.04	0.76	1.4 x 10^{-7}			rs4948088(0.86),rs10231420(<0.1)
	9p24.2	4290823	rs6476839	A>T	0.40	1.12	1.0 x 10^{-9}		*GLIS3*	rs10758593(0.98),rs7020673(0.66)
	10p15.1	6094697	rs61839660	C>T	0.10	0.62	2.8 x 10^{-39}		*IL2RA*	rs7090530(<0.1),rs12251307(0.61)
	10p15.1	6108340	rs10795791	A>G	0.41	1.16	5.6 x 10^{-11}	rs61839660	*IL2RA*	rs7090530(<0.1),rs12251307(<0.1)
	10p15.1	6129643	rs41295121	C>T	0.01	0.49	4.9 x 10^{-8}	rs61839660, rs10795791	*IL2RA*	rs7090530(<0.1),rs12251307(<0.1)

	10q23.31	90035654	rs12416116	C>A	0.28	0.85	3.9×10^{-15}			rs10509540(0.79)
	11p15.5	2182224	rs689	T>A	0.30	0.42	$< 10^{-100}$		*INS*	rs7111341(0.265)
	11p15.5	2198665	rs72853903	C>T	0.38	0.85	6.2×10^{-10}	rs689	*INS*	rs7111341(0.26)
	12p13.31	9905851	rs917911	A>C	0.36	1.10	1.9×10^{-7}		*CD69*	rs4763879(1),rs10492166(0.470)
	12q13.2	56435504	rs705705	G>C	0.34	1.25	4.4×10^{-32}		*IKZF4*	rs2292239(0.87),rs705704(0.99)
	12q24.12	112007756	rs653178	T>C	0.48	1.30	1.6×10^{-44}		*SH2B3*	rs3184504(0.99)
	13q32.3	100081766	rs9585056	T>C	0.24	1.12	3.3×10^{-8}		*GPR183*	rs9585056(1)
	14q32.2	98488007	rs1456988	T>G	0.27	1.12	2.9×10^{-8}			rs4900384(0.98)
	14q32.2	101306447	rs56994090	T>C	0.41	0.88	1.1×10^{-11}			rs941576(0.91)
	15q14	38847022	rs72727394	C>T	0.19	1.15	3.6×10^{-10}		*RASGRP1*	rs12908309(<0.1)
	15q25.1	79234957	rs34593439	G>A	0.10	0.78	9.0×10^{-14}		*CTSH*	rs3825932(0.26),rs12148472(0.79)
	16p11.2	28505660	rs151234	G>C	0.12	1.19	4.8×10^{-11}		*IL27*	rs4788084(0.1),rs9924471(0.54)
	16p13.13	11194771	rs12927355	C>T	0.32	0.82	3.0×10^{-22}		*DEXI*	rs12927355(1),rs12708716(0.86),rs12928822(<1)
	16p13.13	11351211	rs193778	A>G	0.25	1.14	4.4×10^{-10}		*DEXI*	rs12927355(<0.1),rs12708716(0.069),rs12928822(<0.1)
	16q23.1	75252327	rs8056814	G>A	0.07	1.32	3.0×10^{-19}		*BCAR1*	rs7202877(0.86),rs8056814(1)
	17q12	38053207	rs12453507	G>C	0.49	0.90	1.0×10^{-8}		*IKZF3, ORMDL3 ,GSDMB*	rs2290400(0.97)
	17q21.2	38775150	rs757411	T>C	0.36	0.90	1.1×10^{-7}		*CCR7*	rs7221109(0.95)
*	17q21.31	44073889	rs1052553	A>G	0.24	0.89	8.2×10^{-8}			--
	18p11.21	12809340	rs1893217	A>G	0.16	1.21	1.2×10^{-15}		*PTPN2*	rs1893217(1)
	18p11.21	12830538	rs12971201	G>A	0.39	0.89	2.1×10^{-6}	rs1893217	*PTPN2*	rs1893217(0.13)
	18q22.2	67526644	rs1615504	C>T	0.47	1.13	1.8×10^{-11}		*CD226*	rs763361(0.99)
	19p13.2	10463118	rs34536443	G>C	0.04	0.67	4.4×10^{-15}		*TYK2*	rs2304256(<0.1)
	19p13.2	10469975	rs12720356	A>C	0.09	0.82	3.7×10^{-7}	rs34536443	*TYK2*	rs2304256(0.26)
	19q13.32	47219122	rs402072	T>C	0.16	0.87	4.7×10^{-8}			rs425105(0.98)
	19q13.33	49206172	rs516246	T>C	0.49	0.87	5.2×10^{-14}		*FUT2*	rs601338(1)
	20p13	1616206	rs6043409	G>A	0.35	0.88	3.0×10^{-10}			rs2281808(0.91)

	21q22.3	43825357	rs11203202	C>G	0.33	1.16	1.2×10^{-15}	*UBASH3A*	rs11203203(0.42)
*	21q22.3	45621817	rs6518350	A>G	0.18	0.88	9.6×10^{-8}	*ICOSLG*	--
	22q12.2	30531091	rs4820830	T>C	0.38	1.14	1.2×10^{-12}		rs5753037(0.99)
	22q12.3	37587111	rs229533	A>C	0.43	1.11	1.8×10^{-8}	*C1QTNF6,RAC2*	rs229541(0.98),rs229526(0.39)

Table 1: T1D associated loci on ImmunoChip. The most associated SNP in a region is shown, together with the effect of the minor allele relative to major. Where secondary associations are found, they are conditional on SNPs shown in the column "condition". For previously known loci, the r^2 between our lead SNP and previously reported index SNPs is shown. Novel loci, at $P < 3.23 \times 10^{-7}$, are indicated by "*". Alleles are shown major > minor. MAF=minor allele frequency. rs689 (11p15.5, *INS*) data obtained from previous TaqMan genotyping. Named candidate genes are genes for which there is additional evidence that they might be causal, or that they encode proteins with known immune functions that are part of the immune pathways already identified as involved in T1D pathogenesis. Since SNPs may alter enhancer sequences distant from the target gene, we have not named a gene (or a non-coding RNA) if the only evidence for a causal role is that the peak of SNP association lies in or very near a gene (unless those SNPs alter coding-sequence or splice 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 linked this gene to T1D etiology.

Index disease	**Associated regions**	**SNPs in regions**		**Enrichment result**	
		disease assoc'd	**not disease assoc'd**	**Z**	***P***
juvenile idiopathic arthritis	15	2527	22725	7.35	2.00 x 10^{-13}
areata alopecia	4	763	24489	6.63	3.40 x 10^{-11}
primary sclerosing cholangitis	10	1866	23386	6.28	3.40 x 10^{-10}
rheumatoid arthritis	27	4382	20870	5.51	3.60 x 10^{-8}
primary biliary cirrhosis	16	2289	22963	5.26	1.50 x 10^{-7}
celiac disease	29	4512	20740	2.55	1.10 x 10^{-2}
autoimmune thyroid disease	9	1622	23630	2.50	1.20 x 10^{-2}
narcolepsy	2	217	25035	1.49	1.40 x 10^{-1}
multiple sclerosis	57	8312	16940	1.15	2.50 x 10^{-1}
systematic lupus erythematosus	14	2528	22724	-0.23	8.10 x 10^{-1}
ankylosing spondylitis	21	3103	22149	-0.84	4.00 x 10^{-1}
Sjogren's syndrome	6	985	24267	-1.29	2.00 x 10^{-1}
psoriasis	25	4457	20795	-2.22	2.60 x 10^{-2}
Crohn's disease	83	13225	12027	-2.61	9.10 x 10^{-3}
ulcerative colitis	58	9336	15916	-4.04	5.40 x 10^{-5}

Table 2: Enrichment Analysis of evidence for T1D association across densely genotyped non-MHC loci associated with other autoimmune or auto-inflammatory diseases. ImmunoChip densely mapped regions were assigned as associated or not associated with each index disease accorded to publications curated in ImmunoBase (accessed February 13, 2014) and tested whether the distribution of T1D P-values differed between these sets of regions. The numbers of SNPs that passed QC in our T1D study in the two sets of regions are shown. A positive (negative) Z-score implies T1D shows stronger (weaker) evidence of association in regions known to associate with the index disease.

	Index SNP	Chr	Position	MAF	Alleles	Index Disease	Disease association		T1D association		Candidate genes	Reference
							OR	*P*	OR	*P*		
#	rs17264332	6q23.3	1.38E+08	0.22	A>G	CEL	1.29	$5.00x10^{-30}$	1.12	$8.26x10^{-6}$	*TNFAIP3*	29
#	rs6920220	6q23.3	1.38E+08	0.22	G>A	UC	1.16	$1.40x10^{-21}$	1.12	$7.26x10^{-6}$	*TNFAIP3*	30,31
#	rs6920220	6q23.3	1.38E+08	0.22	G>A	RA	1.2	$2.30x10^{-13}$	1.12	$7.26x10^{-6}$	*TNFAIP3*	32
	rs911263	14q24.1	68753593	0.29	T>C	PBC	0.79	$9.95x10^{-11}$	0.89	$4.93x10^{-6}$		16
*	rs17564829	17q21.31	44006601	0.195	T>C	PBC	1.25	$2.15x10^{-9}$	0.89	$6.77x10^{-6}$		16

Table 3: Pleiotropic SNPs associated with T1D. We show all genome-wide significant index SNPs for immune mediated diseases[15,29-32] that are in regions not associated with T1D at genome-wide significance, but have $P < 10^{-5}$ in the case-control analysis presented here. Shown are the index SNPs and diseases, and the single SNP association test statistics for each index disease and T1D. Chromosome positions are given according to GRCh37. MAF=minor allele frequency, OR=odds ratio. # rs17264332 is in LD with rs6920220, r^2=1. * rs17564829 is in LD with rs1052553 in **Table 1**, r^2=0.99.

ONLINE METHODS

Samples

Affected sib-pair 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 Genetic Resource Investigating Diabetes (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 (UK NBS, N=3,060)[12], and the NIHR Cambridge Biomedical Research Centre Cambridge BioResource (CBR, 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 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, ImmunoChip (Illumina, Inc; CA) according to manufacturer's protocols. The ImmunoChip, a custom Illumina Infinium HD array, was designed to densely genotype, using 1000 Genomes 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 criteria ($P < 5\times10^{-8}$) from twelve 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, type 1 diabetes, and ulcerative colitis). All 1000 Genomes Project pilot phase[37] CEU population variants (Sept 2009 release) within 0.1cM (HapMap3 CEU) recombination blocks around each GWAS region lead marker were submitted for array design. No filtering on correlated variants (linkage disequilibrium) 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 affected sib-pair (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 for the family data set and the case, control data set was performed separately. Initial sample quality control metrics included sample call rate, heterozygosity, and sex concordance check of reported versus genotyped. 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 ethnicity data (see sections on population inference and population structure for more detail). 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 based upon the genetic data. Individuals with low X chromosome heterozygosity and a large number of Y chromosome SNPs were defined as ‘males’; individuals with a high X chromosome heterozygosity and a small number of Y chromosome SNPs were defined as ‘females’. Inconsistency between the self-reported sex and the 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, and 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 in T1DGC ASP and trio family data were rejected due to failure to attain at least 95% genotyping rate. An additional 618 SNPs in the case and control data were removed due to 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 HWE $P < 10^{-6}$) in controls and 527 SNPs failed (with HWE $P < 10^{-10}$) 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-QC was undertaken, whereby the call-rate cut off was raised to 99%, the HWE cut off was lowered to $P < 10^{-4}$. A further 8,349 SNPs were removed for lower call-rate and 10,708 for violation of HWE, and 34 for manually identified poor signal clouds. This strategy reduced the total number of SNPs analysed 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_1000 = 1.09, which was expected as the ImmunoChip was designed to target robustly defined immune-mediated disease susceptibility loci. Excluding SNPs from regions reported in this paper, lambda_1000 was reduced to 1.07; excluding all densely genotyped regions reduced lambda_1000 to 1.03.

Relationship Inference

Cryptic relatedness can confound the result of population structure and association analyses and 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 based on their SNP data. Since only SNPs of these two individuals are used when the kinship coefficient is estimated for a pair of individuals, the estimation accuracy is independent of the population structure in 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 (1st- and 2nd-degree) with extremely low false positives (*i.e.*, to separate unrelated pairs from close relatives)[7]. After the 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 1st and 2nd-degree relationships.

A total of 30 individuals were removed in family data due to the inconsistency between the estimated and documented relationships, and ~500 pairs of 1st-degree relatives that were not reflected in the documented pedigree have been incorporated in the pedigree data by pedigree reconstruction. **Supplementary Figure 2** shows all pair-wise relationships in families after QC. The estimated kinship coefficient of each pair of relatives is plotted against the proportion of zero IBS, with the documented relationships being indicated by colour. All 42 pairs of documented identical twins have estimated kinship coefficient > 0.4. Among 16,292 documented 1st-degree relative pairs, 16,270 pairs have estimated kinship coefficient between 0.177 and

0.36 (criteria to be inferred as 1st-degree relative in KING), 21 pairs have estimated kinship coefficient between 0.150 and 0.177, and 1 pair has estimated kinship coefficient 0.137. After pedigree reconstruction, there was no 1st-degree relatedness across any two families, and there were only 3 pairs of documented unrelated pairs with estimated kinship coefficient > 0.1 (all 3 kinship coefficients < 0.139). In the analysed data, a total of 10,796 individuals from 2,682 nuclear families have genotypes available. There were 1,670 families with both parents available, 652 with only one parent and 360 with neither parent. The distribution of affected siblings was 69 families with one affected, 2490 with two, 104 with three, 5 with four, and 2 with five.

In the T1D cases and the UK control data, 159 controls and 48 cases were removed for being close relatives. After this level of QC, no remaining "unrelated" pairs in the case or control data have estimated kinship coefficient > 0.09, indicating all individuals are indeed unrelated. We also checked the UK T1D case and UK control for relatedness in the T1DGC ASP and trio family data set, since one of the four T1DGC collection sites was in the UK. A total of 5 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 III data (1097 unrelated individuals were used[39], with 215 of European ancestry) with each cohort. We kept those SNPs that are present on both HapMap and ImmunoChip panels, and removed SNPs with $r^2 > 0.5$ with other SNPs. After applying the QC filters, ~30,000 SNPs were used for the structure analysis. PCA was first carried out among the HapMap individuals only, and then each ImmunoChip individual was projected to the space that was expanded by the principal components of HapMap individuals. The projected principal components for each individual represent its ancestry relative to the HapMap populations. Using this algorithm, we obtained the principal components for case-control individuals by cohort, projected to either the entire HapMap III populations (**Supplementary Figure 3**), or the European ancestry populations only including CEU and TSI (**Supplementary Figure 4**); we also obtained the principal components for individuals in the family data (**Supplementary Figure 5**).

In **Supplementary Figure 3**, population structure of our case-control data was compared with all HapMap III populations. A total of 69 individuals were identified to be greater than 3 standard deviations (SD) from the average of the second principal components in European populations, and these outliers were excluded from analysis. The principal components of all case-control individuals from four cohorts (GRID and 1958 British Cohort that were genotyped at UVA, 1958 British Cohort and National Blood Service that were genotyped at Sanger) are in the range of the European ancestry populations, clearly separated from non-European populations. In **Supplementary Figure 4**, case-control individuals were compared with European populations only, including CEU and TSI. The cluster on the left is for CEU that represents the northern European, and the cluster on the right is for TSI which represents the southern European. A total of 55 “outliers” were identified in this analysis to cluster with the southern European and have been excluded prior to analysis. **Supplementary Figure 5** suggests that there is no substructure difference between our cases (UVA GRID) and controls (UVA 1958 BC, Sanger 1958 BC, and UK NBS). **Supplementary Figure 5** shows the population structure in the family data, compared with the HapMap populations. 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* utility. For each variant, the SNP alleles have been normalized so the reference and alternate alleles are reported on the reference (top) strand.

Single SNP Association Analysis

In order to test association between each SNP and T1D, we applied the Generalized Disequilibrium Test (GDT) method[39] to the T1DGC ASP and trio families, 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 utilizes the information of all discordant relative pairs, including those nuclear families that are not efficiently used in family-based tests such as Transmission/Disequilibrium Test (TDT) or Family Based Association Test (FBAT). To estimate the effect at each variant, we carried out the TDT at each region and approximated the odds ratio 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 the family data[42]. An overall beta coefficient and standard error were computed as the weighted average of the individual beta statistics, and a corresponding P-value for that statistic was computed. The weights were proportional to the inverse variance (1 divided by the standard error squared) in each study and

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

scaled by the meta-variance (σ^2_{meta}, equation above) so the weights summed to 1. For the family data, instead of using the total number of family members, we used twice of the number of parent-affected-offspring trios as the effective sample size for the meta-analysis.

Conditional Analysis to Identify Secondary Signals

To determine if additional SNPs within a region were significantly associated with T1D, independent of the most 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 fit 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 February 13, 2014). We excluded IBD as this is a combination of UC and Crohn's which are summarised individually. The MHC region(chr6:25Mb..35Mb GRCh37) was excluded from analysis. For each disease in turn, we used the index SNPs to label each of densely mapped regions of the ImmunoChip as associated with the index disease and that region or not. After LD pruning (r^2 <= 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 using 10,000 permutations of case *vs* 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 dense region.

eQTL and GWAS Catalogue overlap in seven novel regions

To define a query SNP set we took a 2Mb window centred on each novel index SNP and then filtered overlapping SNPs based on a linkage disequilibrium (LD) threshold of $r^2 \geq 0.9$ with the index SNP, using 1000 genomes data. To identify potential cis eQTL overlap we downloaded summary statistics from Fairfax *et al*.[44] (their Table S7) and Westra et al.[25] (Blood eQTL browser) and computed overlap with the query SNP set. For each significant overlap we computed the LD with the top eQTL SNP for that probe/tissue, again using 1000 genome data To look for trait/disease overlap outside ImmunoBase scope we used the query SNP set to examine overlap between NHGRI GWAS catalogue[45].

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 relation

$$-2 \log(ABF_{ij}) = BIC_i - BIC_j$$

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; *e.g.*, 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 all alternative two SNP models containing rs61839660 and another SNP within a 50 kb window of rs10795791.

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

$$\mathrm{PP}_j = \mathrm{BIC}_j/\mathrm{sum}(\mathrm{BIC}_j)$$

and thus we create a 99% credible set of SNPs as the smallest set of SNPs with a total PP $\geq$ 99%.

Enrichment Analysis

Epigenomic Roadmap annotations were downloaded from the web portal. These were processed using R and Bioconductor packages to annotate those ImmunoChip SNPs overlapping tissue specific functional elements. According to the credible sets formed above, the ImmunoChip SNPs that passed QC could be divided into two sets :

A: those that are in any credible set, within ImmunoChip densely mapped regions - potential causal variants (n=1,256)

B: their complement, within ImmunoChip densely mapped regions - 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 2x2 contingency tables, stratified by a SNP’s MAF in controls (<0.05, <0.1, <0.2, <0.3, <0.4, <0.5) showing the overlap of SNPs in 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 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 which could be targeted in future functional studies, we first expanded the sets by considering all neighbouring SNPs in 1000 Genomes CEU release that were did not pass genotyping on the ImmunoChip. These 1000 Genomes SNPs were assigned to credible sets if the ImmunoChip SNP with which they should strongest LD according to r^2 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 overlap enhancers in tissues which showed enrichment according to **Fig 2**, and the number which are non-synonymous. These 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 associated with T1D on the basis of $p<10^{-4}$ for a SNP that also shows $p<5x10^{-8}$ in another autoimmune disease[5]. Here, we explore the strength of evidence these thresholds provide, based on previous work[46]. For any individual SNP and two diseases, there exist four hypotheses:

H_0: Not associated with either disease

H_1: Associated with only disease 1

H_2: Associated with only disease 2

H_{12}: Associated with both disease 1 and disease 2

Realistic prior probabilities[46] are:

$\pi_0 = 1 - 2x10^{-4} – 10^{-5}$ $\pi_1 = 10^{-4}$

$\pi_2 = 10^{-4}$ $\pi_{12} = 10^{-5}$

that imply we expect about 1 in 1000 SNPs show association to either disease and, of SNPs associated to one disease, we expect about 1 in 10 to be associated with both diseases.

Posterior probabilities for independent datasets

We use the approximate Bayes Factors presented previously[47] to estimate ϕ_i, the Bayes Factor for association to disease *i* compared to no association to disease *i* given only single SNP p-values and the minor allele frequency (MAF) of the SNP in controls. If we assume the case and control datasets for each disease are independent, they can be combined to calculate Bayes Factors for each hypothesis

$$BF_0 = 1 \qquad BF_1 = \phi_1$$

$$BF_2 = \phi_2 \qquad BF_{12} = \phi_1\phi_2$$

Thus, the posterior probability for each hypothesis is given as

$$PP_0 = \pi_0/B \qquad PP_1 = \pi_1\phi_1/B$$

$$PP_2 = \pi_2\phi_2/B \qquad PP_{12} = \pi_{12}\phi_1\phi_2/B$$

where $B = 1 + \phi_1 + \phi_2 + \phi_{12}$. The conditional probability of association to disease 2, given we believe there is association to 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[48], so BF_{12} cannot be expressed as a simple product of disease-specific Bayes Factors. Methods to account for this appear conservative[48], as they do not allow for the reasonable assumption that related diseases share genetic susceptibility variants. Instead, we use simulation to explore the effect of non-independence on $PP_{2|1}$. We use multinomial models and the approximate Bayes Factor[49] 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 WTCCC and the ImmunoChip papers (**Supplementary Table 3**). Using p_i to denote the single SNP p-value for disease *I*, the results (**Supplementary Figure 6**) show that, for independent controls, $PP_{2|1} > 0.9$ (median 0.97) whenever $p_2 < 10^{-4}$. 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 $PP_{2|1}$, while 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 $PP_{2|1}$ in the interval (0.37, 0.61)(median 0.46) at $p_2 = 10^{-4}$ for all scenarios. When $p_2= 10^{-5}$, $PP_{2|1}$ is in the interval (0.87,0.90)(median 0.89), suggesting that a $p_2 = 10^{-5}$ threshold may be more suitable for convincing evidence of association to a second autoimmune disease.

The 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 at https://github.com/mdfortune/colocCommonControl.

Online Methods References

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 *et al*. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (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, WM, Manichaikul, A. & Rich, S.S. A generalized family-based association test for dichotomous traits. *Am J Hum Genet* **85**, 364-76 (2009).

40. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75. (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**, 2190-1 (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), DOI: 10.1126/science.1246949.

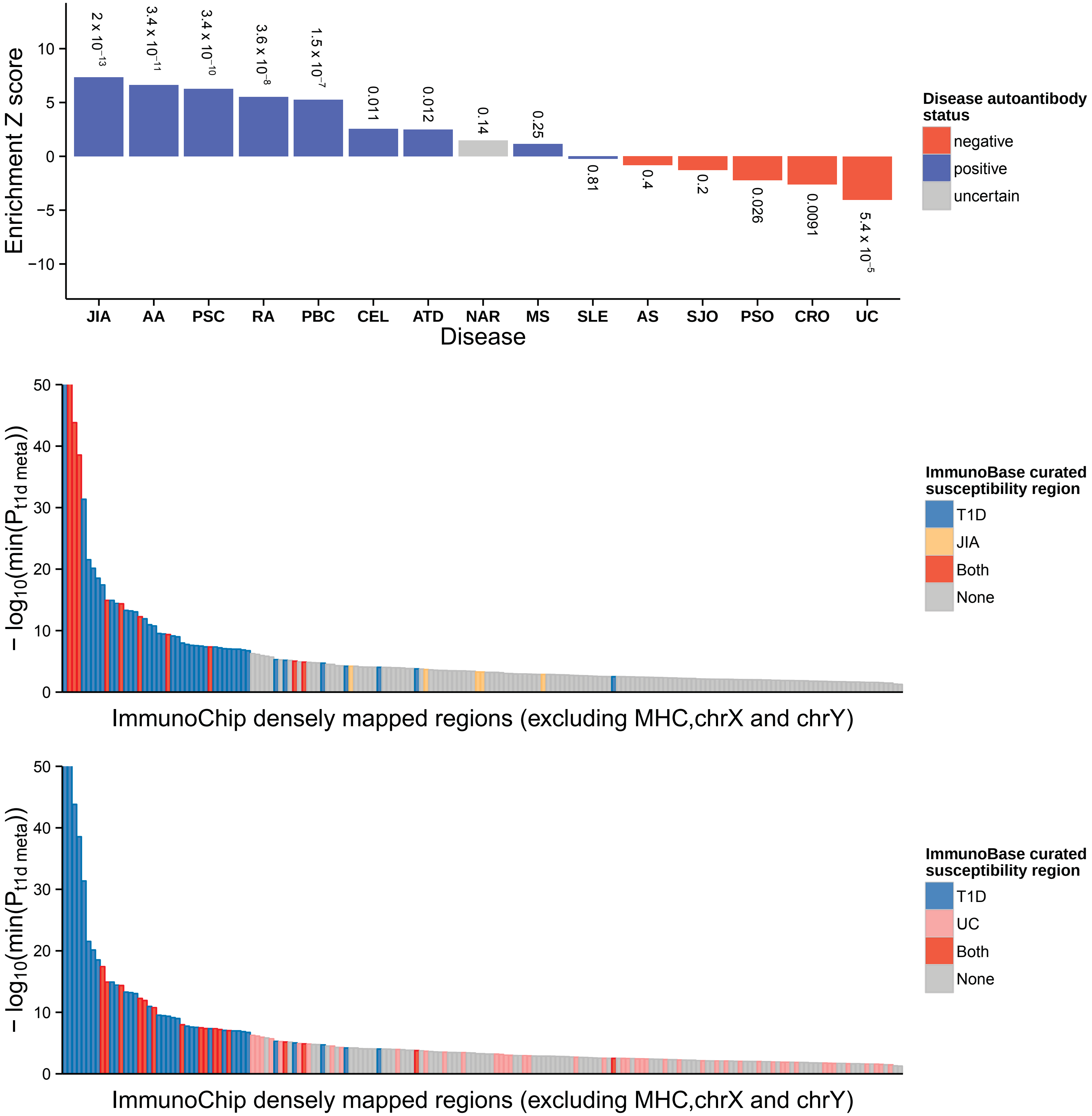
45. Welter D*. et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–D1006 (2013).

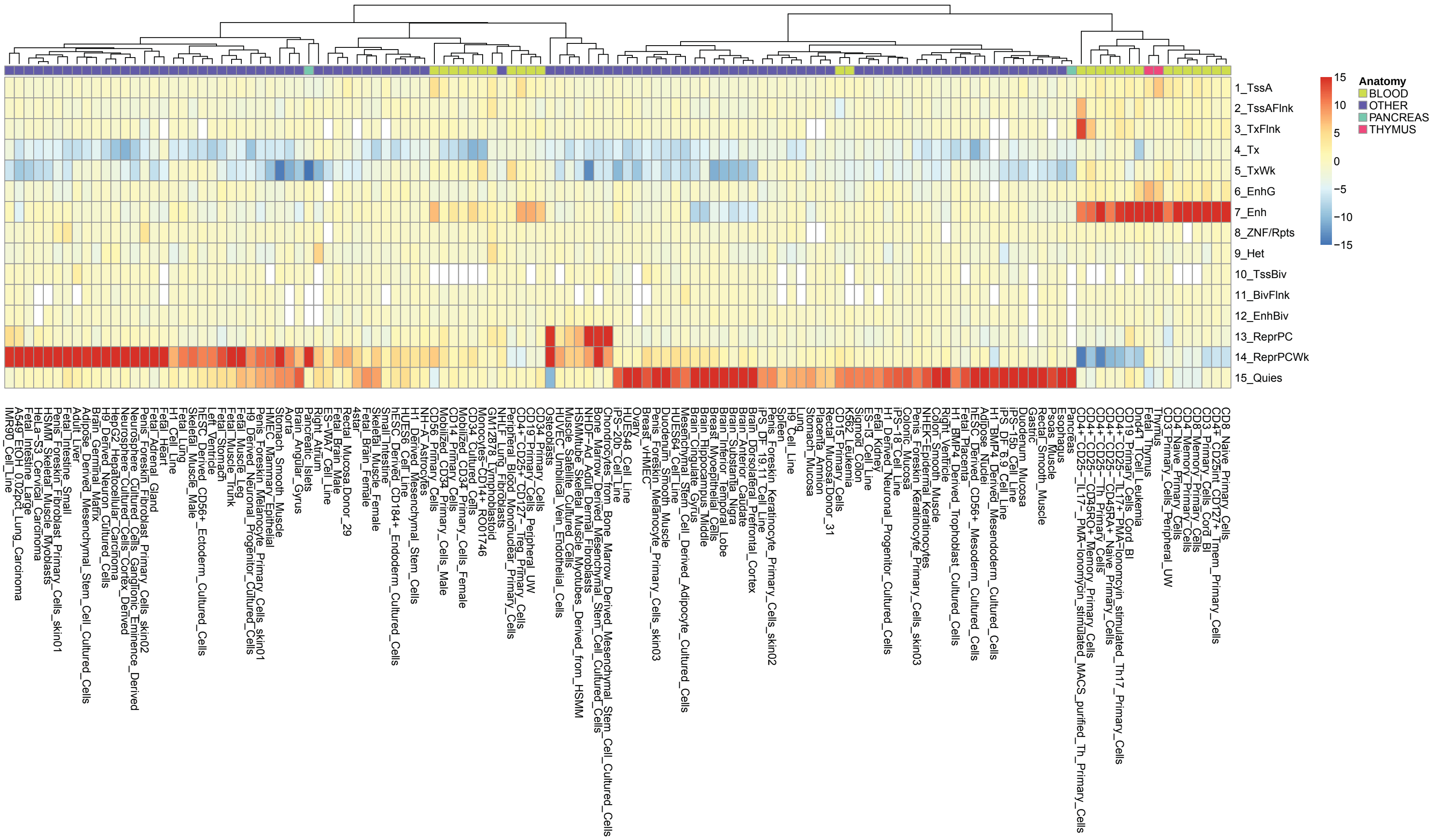
46. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS. Genet* **10**, e1004383 (2014).

47. Wakefield, J. Bayes factors for genome-wide association studies: comparison of p-values. *Genetic Epidemiology* **33**, 79-86 (2009).

48. Zaykin, D.V. & Kozbur, D.O. P-value based analysis for shared controls design in genome-wide association studies. *Genetic Epidemiology* **34**, 725-38 (2010).

49. Rafferty, A.E. Approximate bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251-65 (1996).


Enrichment Z score
10
5
0
−5
−10
2 x 10^{-13}
3.4 x 10^{-11}
3.4 x 10^{-10}
3.6 x 10^{-8}
1.5 x 10^{-7}
0.011
0.012
0.14
0.25
0.81
0.4
0.2
0.026
0.0091
5.4 x 10^{-5}
JIA
AA
PSC
RA
PBC
CEL
ATD
NAR
MS
SLE
AS
SJO
PSO
CRO
UC
Disease
Disease autoantibody status
🟥 negative
🟦 positive
⬜ uncertain
$-\log_{10}(\min(P_{t1d\ meta}))$
50
40
30
20
10
0
ImmunoChip densely mapped regions (excluding MHC,chrX and chrY)
ImmunoBase curated susceptibility region
🟦 T1D
🟧 JIA
🟥 Both
⬜ None
$-\log_{10}(\min(P_{t1d\ meta}))$
50
40
30
20
10
0
ImmunoChip densely mapped regions (excluding MHC,chrX and chrY)
ImmunoBase curated susceptibility region
🟦 T1D
🟥 UC
🟥 Both
⬜ None


1_TssA
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF/Rpts
9_Het
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
15
10
5
0
−5
−10
−15
Anatomy
BLOOD
OTHER
PANCREAS
THYMUS
IMR90_Cell_Line
A549_EtOH_0.02pct_Lung_Carcinoma
Fetal_Intestine_Large
HeLa-S3_Cervical_Carcinoma
HSMM_Skeletal_Muscle_Myoblasts
Penis_Foreskin_Fibroblast_Primary_Cells_skin01
Fetal_Intestine_Small
Adult_Liver
Adipose_Derived_Mesenchymal_Stem_Cell_Cultured_Cells
Brain_Germinal_Matrix
H9_Derived_Neuron_Cultured_Cells
HepG2_Hepatocellular_Carcinoma
Neurosphere_Cultured_Cells_Cortex_Derived
Neurosphere_Cultured_Cells_Ganglionic_Eminence_Derived
Penis_Foreskin_Fibroblast_Primary_Cells_skin02
Fetal_Adrenal_Gland
Fetal_Heart
H1_Cell_Line
Fetal_Lung
Skeletal_Muscle_Male
hESC_Derived_CD56+_Ectoderm_Cultured_Cells
Left_Ventricle
Fetal_Stomach
Fetal_Muscle_Trunk
Fetal_Muscle_Leg
H9_Derived_Neuronal_Progenitor_Cultured_Cells
Penis_Foreskin_Melanocyte_Primary_Cells_skin01
HMEC_Mammary_Epithelial
Stomach_Smooth_Muscle
Aorta
Brain_Angular_Gyrus
Pancreatic_Islets
Right_Atrium
ES-WA7_Cell_Line
Fetal_Brain_Male
Rectal_Mucosa.Donor_29
4star
Fetal_Brain_Female
Skeletal_Muscle_Female
Small_Intestine
hESC_Derived_CD184+_Endoderm_Cultured_Cells
HUES6_Cell_Line
H1_Derived_Mesenchymal_Stem_Cells
NH-A_Astrocytes
CD56_Primary_Cells
Mobilized_CD34_Primary_Cells_Male
CD14_Primary_Cells
Mobilized_CD34_Primary_Cells_Female
CD34_Cultured_Cells
Monocytes-CD14+_RO01746
GM12878_Lymphoblastoid
NHLF_Lung_Fibroblasts
Peripheral_Blood_Mononuclear_Primary_Cells
CD4+_CD25+_CD127-_Treg_Primary_Cells
CD19_Primary_Cells_Peripheral_UW
CD34_Primary_Cells
Osteoblasts
HUVEC_Umbilical_Vein_Endothelial_Cells
Muscle_Satellite_Cultured_Cells
HSMMtube_Skeletal_Muscle_Myotubes_Derived_from_HSMM
NHDF-Ad_Adult_Dermal_Fibroblasts
Bone_Marrow_Derived_Mesenchymal_Stem_Cell_Cultured_Cells
Chondrocytes_from_Bone_Marrow_Derived_Mesenchymal_Stem_Cell_Cultured_Cells
iPS-20b_Cell_Line
HUES48_Cell_Line
Ovary
Breast_vHMEC
Penis_Foreskin_Melanocyte_Primary_Cells_skin03
Duodenum_Smooth_Muscle
HUES64_Cell_Line
Mesenchymal_Stem_Cell_Derived_Adipocyte_Cultured_Cells
Brain_Cingulate_Gyrus
Breast_Myoepithelial_Cells
Brain_Hippocampus_Middle
Brain_Inferior_Temporal_Lobe
Brain_Substantia_Nigra
Brain_Anterior_Caudate
Brain_Dorsolateral_Prefrontal_Cortex
iPS_DF_19.11_Cell_Line
Penis_Foreskin_Keratinocyte_Primary_Cells_skin02
Spleen
H9_Cell_Line
Lung
Stomach_Mucosa
Placenta_Amnion
Rectal_Mucosa.Donor_31
CD15_Primary_Cells
K562_Leukemia
Sigmoid_Colon
ES-I3_Cell_Line
Fetal_Kidney
H1_Derived_Neuronal_Progenitor_Cultured_Cells
iPS-18_Cell_Line
Colonic_Mucosa
Penis_Foreskin_Keratinocyte_Primary_Cells_skin03
NHEK-Epidermal_Keratinocytes
Colon_Smooth_Muscle
Right_Ventricle
H1_BMP4_Derived_Trophoblast_Cultured_Cells
Fetal_Placenta
hESC_Derived_CD56+_Mesoderm_Cultured_Cells
Adipose_Nuclei
H1_BMP4_Derived_Mesendoderm_Cultured_Cells
iPS-15b_Cell_Line
iPS_DF_6.9_Cell_Line
Duodenum_Mucosa
Gastric
Rectal_Smooth_Muscle
Psoas_Muscle
Esophagus
Pancreas
CD4+_CD25-_IL17-_PMA-Ionomycin_stimulated_MACS_purified_Th_Primary_Cells
CD4+_CD25-_CD45RO+_Memory_Primary_Cells
CD4+_CD25-_Th_Primary_Cells
CD4+_CD25-_CD45RA+_Naive_Primary_Cells
CD4+_CD25-_IL17+_PMA-Ionomcyin_stimulated_Th17_Primary_Cells
CD19_Primary_Cells_Cord_BI
Dnd41_TCell_Leukemia
Fetal_Thymus
Thymus
CD3_Primary_Cells_Peripheral_UW
CD4_Naive_Primary_Cells
CD4_Memory_Primary_Cells
CD8_Memory_Primary_Cells
CD3_Primary_Cells_Cord_BI
CD4+_CD25int_CD127+_Tmem_Primary_Cells
CD8_Naive_Primary_Cells