HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2019 February 06.

Published in final edited form as:

Nat Genet. 2018 October ; 50(10): 1366–1374. doi:10.1038/s41588-018-0216-7.

Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes

Harm-Jan Westra[1,2,3,4,5,20], **Marta Martínez-Bonet**[4,20], **Suna Onengut-Gumuscu**[6,7], **Annette Lee**[8], **Yang Luo**[1,2,3,4], **Nikola Teslovich**[1,2,3,4], **Jane Worthington**[9,10], **Javier Martin**[11], **Tom Huizinga**[12], **Lars Klareskog**[13], **Solbritt Rantapaa-Dahlqvist**[14], **Wei-Min Chen**[6,7], **Aaron Quinlan**[6,15,16], **John A. Todd**[17], **Steve Eyre**[9,10], **Peter A. Nigrovic**[4,18], **Peter K. Gregersen**[8], **Stephen S. Rich**[6,7], and **Soumya Raychaudhuri**[1,2,3,4,9,19,*]

[1]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [2]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [4]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [5]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [6]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA. [7]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [8]The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY, USA. [9]Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. [10]NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester, UK. [11]Instituto de Parasitología y Biomedicina López-Neyra, Consejo Superior de Investigaciones Científicas, Granada, Spain. [12]Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands. [13]Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden. [14]Department of Public Health and Clinical Medicine, Division of Rheumatology, Umeå University, Umeå, Sweden. [15]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [16]Department of Biomedical

* soumya@broadinstitute.org.

Author contributions

H.-J.W., Y.L. and S.R. performed the analyses. M.M.-B. and P.A.N. performed the functional assays. H.-J.W., M.M.-B., P.A.N. and S.R. designed the study. S.O., A.L., N.T., J.W., J.M., T.H., L.K., S.R.-D., W.-M.C., A.Q., J.A.T., S.E., P.K.G., S.S.R. and S.R. acquired the data. H.-J.W., M.M.-B., Y.L., J.A.T., P.A.N., P.K.G., S.S.R. and S.R. wrote and edited the manuscript.

URLs. Bios eQTL browser, http://genenetwork.nl/biosqtlbrowser/; Roadmap Epigenomics datasets, http://www.roadmapepigenomics.org/; ChromHMM enhancers and promoters, http://egg2.wustl.edu/roadmap/web_portal/; 1000 Genomes reference panel, http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a/; Haplotype Reference Consortium panel, http://www.haplotype-reference-consortium.org/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0216-7.

Competing interests

The authors declare no competing interests.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0216-7.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Informatics, University of Utah, Salt Lake City, UT, USA. [17]JDRF/Wellcome Diabetes and Inflammation Laboratory, Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [18]Division of Immunology, Boston Children's Hospital, Boston, MA, USA. [19]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [20]These authors contributed equally: Harm-Jan Westra, Marta Martínez-Bonet.

Abstract

To define potentially causal variants for autoimmune disease, we fine-mapped[1,2] 76 rheumatoid arthritis (11,475 cases, 15,870 controls)[3] and type 1 diabetes loci (9,334 cases, 11,111 controls)[4]. After sequencing 799 1-kilobase regulatory (H3K4me3) regions within these loci in 568 individuals, we observed accurate imputation for 89% of common variants. We defined credible sets of ≤5 causal variants at 5 rheumatoid arthritis and 10 type 1 diabetes loci. We identified potentially causal missense variants at *DNASE1L3, PTPN22, SH2B3,* and *TYK2*, and noncoding variants at *MEG3, CD28-CTLA4*, and *IL2RA*. We also identified potential candidate causal variants at *SIRPG* and *TNFAIP3*. Using functional assays, we confirmed allele-specific protein binding and differential enhancer activity for three variants: the *CD28-CTLA4* rs117701653 SNP, *MEG3* rs34552516 indel, and *TNFAIP3* rs35926684 indel.

Rheumatoid arthritis is an autoimmune disease with citrullinated peptide reactivity where chronic inflammation leads to joint destruction[5]. Type 1 diabetes (T1D) arises through autoimmune reactivity to proinsulin[6] and glutamic acid decarboxylase[7], leading to destruction of pancreatic β cells and loss of insulin production. Genome-wide association studies have identified 101 rheumatoid arthritis loci[3,8] and 53 T1D loci[4]; these alleles implicate $CD4^+$ T-cell function in autoimmunity[9–11]. However, causal variants for most loci have yet to be defined. Pinpointing them will enable mechanistic investigation to identify the specific genes, regulatory structures, and genetic mechanisms central to autoimmunity.

Bayesian fine-mapping has been successfully applied to prioritize associated variants in complex diseases[1,12–14]. We fine-mapped 76 autosomal non-major-histocompatibility-complex loci in rheumatoid arthritis (11,475 cases, 15,870 controls)[3] and T1D (9,334 cases, 11,111 controls) (Supplementary Table 1)[4], covering 46 and 49 loci with known rheumatoid arthritis and T1D associations, respectively (Supplementary Table 2). To enable accurate comprehensive imputation, we used individuals genotyped on ImmunoChip, with dense SNP coverage in selected autoimmune disease loci. Our sample size is smaller than the largest of previously published association analyses[4,8]. However, Okada et al.[8] used imputed genotypes but did not define credible sets, and while Onengut-Gumuscu et al.[4] determined credible sets, they did not use imputation so their study included fewer than half of the variants assessed here. We fine-mapped rheumatoid arthritis and T1D together since potential causal variants for both diseases overlap functional elements in $CD4^+$ T cells[11].

Since fine-mapping methods are highly sensitive to missing data, we benchmarked different imputation strategies. After sequencing 799 1-kilobase regulatory (H3K4me3) regions in 568 individuals within these loci, we observed that the 1000 Genomes cosmopolitan reference panel yielded the best results (Fig. 1), while indels and multi-allelic variants

remain challenging to impute (Supplementary Note, Supplementary Tables 3–7, and Supplementary Figs. 1–4).

After imputation, we calculated association statistics for 66,923 variants for rheumatoid arthritis and 66,942 variants for T1D (minor allele frequency (MAF) > 1%, imputation quality score (INFO) > 0.3; Hardy–Weinberg $P > 10^{-5}$) at 76 loci. In rheumatoid arthritis and T1D, respectively, we identified 20 and 34 significant loci ($P < 7.5 \times 10^{-7}$), mostly consistent with previous studies (Supplementary Note and Supplementary Table 8). Using approximate Bayesian fine-mapping, we assigned posterior probabilities and defined a 95% credible set for each locus[1,2].

Seven loci were significantly associated with both diseases (*PTPN22*, *AFF3*, *CD28-CTLA4*, *BACH2*, *RASGRP1*, *PTPN2*, and *TYK2*). Rheumatoid arthritis and T1D variant effect sizes were positively correlated in 64% of the tested loci (Methods, Supplementary Table 9, and Supplementary Fig. 5), suggesting shared signals. To prioritize possible shared causal variants, we analyzed a combined dataset (20,787 rheumatoid arthritis or T1D cases, and 18,616 controls; Methods) and observed significant associations at 28 loci. Most (62%) associated variants in these loci were in linkage disequilibrium with the strongest associated variant in either rheumatoid arthritis or T1D (coefficient of determination (R^2) > 0.8; Supplementary Table 10). The combined analysis indicated a potential shared signal at the *PRKCQ* locus (Supplementary Note) that was not significant in rheumatoid arthritis or T1D alone. We did not identify additional significant loci. We did not observe residual population stratification by combining both datasets (Supplementary Note, Supplementary Table 11, and Supplementary Fig. 6). Multinomial Bayesian fine-mapping accounting for opposing effects[15] did not identify additional loci (Supplementary Note and Supplementary Table 12). Because six out of seven shared loci had smaller credible sets in the combined analysis than for the individual disease analyses (Supplementary Note), we decided to use the combined analysis for those six loci.

We narrowed down the list of probable causal variants to ≥5 in 5 out of 20 significant rheumatoid arthritis loci and 10 out of 34 significant T1D loci. We subsequently limited the analysis to 9 (rheumatoid arthritis), 14 (T1D), and 11 (combined) loci with ≤10 variants in the credible sets (Fig. 2a,b and Supplementary Table 13). Credible sets for these loci were not markedly altered when considering variants unique to the Haplotype Reference Consortium (HRC) reference panel (Supplementary Fig. 7 and Supplementary Table 14). To systematically investigate candidate causal variants for both diseases, we selected those with a posterior probability of >0.2 (Table 1). We considered variants to be potentially causal if they: (1) were a missense coding variant; or (2) were in a region with evidence of enhancer activity, and demonstrated allele-specific binding in electrophoretic mobility shift assays (EMSAs) and allele-specific enhancer function in luciferase assays (Supplementary Table 15). Approximate Bayesian fine-mapping assumes a single causal variant per locus, although multiple independent causal variants may be present. Therefore, we applied conditional analysis, exhaustive testing of all variant pairs, FINEMAP[16] (which assumes multiple variants), and haplotype analysis for regions where there was evidence of a secondary effect ($P < 1.9 \times 10^{-5}$).

We identified 42 variants at 20 loci with a posterior probability of >0.2 (Table 1), including 5 missense variants (*PTPN22*, *DNASE1L3*, *SH2B3*, *TYK2*, and *SIRPG*), 3 indels (*ANKRD55*, *TNFAIP3*, and *MEG3*), and 34 noncoding SNP variants (*PTPN22*, *IL10*, *IFIH1*, *CD28-CTLA4*, *ANKRD55*, *BACH2*, *CCL21*, *IL2RA*, *INS*, *SH2B3*, Chr13, *MEG3*, *CTSH*, *TYK2*, *SIRPG*, *UBASH3A*, and *C1QTNF6*). We observed a 16-fold posterior probability enrichment for missense variants. Potentially causal missense variants at *PTPN22*, *SH2B3*, and *TYK2* are well described in the literature[4,17–19] (Supplementary Note and Supplementary Figs. 8–10). We also noted potentially causal missense variants in *DNASE1L3* and *SIRPG*.

DNASE1L3 encodes a nuclease that cleaves double-stranded DNA during apoptosis[20]. The 3p14 *DNASE1L3* locus confers rheumatoid arthritis susceptibility without evidence of a T1D effect ($P > 0.02$; Supplementary Fig. 11). The reported[3] lead SNP rs35677470, encoding an p.Arg206Cys change in DNASE1L3, has a high posterior probability ($P = 1.8 \times 10^{-8}$; posterior probability = 0.82; Supplementary Table 13) and is in linkage disequilibrium with another reported[8] lead variant, rs73081554 ($R^2 = 0.79$). Conditioning on p.Arg206Cys obviates any evidence of independent risk variants ($P > 5 \times 10^{-4}$; Supplementary Table 16). p.Arg206Cys has been implicated in systemic sclerosis[21]; other loss-of-function DNASE1L3 mutations have been reported in familial forms of systemic lupus erythematosus[22]. p.Arg206Cys is a loss-of-function variant that abolishes the protein's nuclease activity[23].

Within 20p13, we identified a p.Val263Ala signal-regulatory protein gamma (SIRPG) missense variant with modest posterior probability (rs6043409; $P = 3.9 \times 10^{-10}$; posterior probability = 0.24) for T1D (Supplementary Fig. 12 and Supplementary Table 13). Conditional analysis using rs6043409 obviated any association signal in the locus ($P > 2 \times 10^{-3}$). p.Val263Ala is in the D3 domain[24]. While D3 has unknown function, homologous D1 and D2 immunoglobulin-like domains mediate interaction with CD47 ligands and SIRPG dimerization[25–27]. We observed linkage disequilibrium ($R^2 > 0.8$) with protein quantitative trait loci (QTL) and expression QTL (eQTL) SNPs (Supplementary Table 17), suggesting that p.Val263Ala might cause a SIRPG conformational change that alters stability, structure, or function. Since p.Val263Ala has unknown function, the two noncoding variants in the *SIRPG* credible set with comparable posterior probabilities and regulatory effects may potentially be causal. However, given the enrichment of missense variants in our data and elsewhere[28,29], we consider p.Val263Ala as the strongest candidate causal variant at this locus.

We identified noncoding alleles with allele-specific function at *CTLA4–CD28*, *TNFAIP3*, and *MEG3* in regions with evidence of $CD4^+$ T-cell enhancer function (Table 1). We identified the rs61839660 variant at *IL2RA* with a high posterior probability (0.85), which was recently demonstrated to have allele-specific regulatory and cellular functions confirmed by CRISPR assays[30]. Detailed analyses of all other loci with ≥10 variants in the 95% credible set are presented in the Supplementary Note and Supplementary Figs. 13–24.

CD28 and *CTLA4* are central to the regulation and differentiation of T cells[20]. The 2q33.2 *CD28–CTLA4* locus is shared by rheumatoid arthritis and T1D[31]. In the combined analysis, a single variant had a high posterior probability (rs3087243; $P = 1.4 \times 10^{-16}$; posterior

probability = 0.91) near *CTLA4*, which also has the largest posterior probability in T1D ($P = 1.6 \times 10^{-15}$; posterior probability = 0.48; Fig. 3a, Supplementary Fig. 25a, and Supplementary Table 13), but not in rheumatoid arthritis ($P = 1.6 \times 10^{-8}$; posterior probability = 0.02). Incontrast,rs117701653near*CD28* ($R^2 = 0.03$ with rs3087243) carried the highest posterior probability in rheumatoid arthritis ($P = 3.5 \times 10^{-10}$; posterior probability = 0.67). In conditional analysis on rs3087243, rs117701653 demonstrated an independent effect ($P = 4.0 \times 10^{-8}$; Fig. 3a and Supplementary Table 16). The FINEMAP[16] program identified the rs117701653 + rs3087243 pair as having the highest posterior probability (0.05; Supplementary Table 18), consistent with an analysis where we tested all SNP pairs exhaustively in rheumatoid arthritis (Fig. 3b and Supplementary Fig. 25b). Haplotype analysis demonstrated independent protective effects of rs3087243-A and rs117701653-C alleles in rheumatoid arthritis and T1D (Fig. 3c), suggesting that rs117701653 might similarly influence T1D risk ($P = 0.03$ in the conditional haplotype analysis). A previous association study identified rs1980422 as the variant with the strongest association in the *CD28* region[8] ($R^2 = 0.02$ with rs117701653 and $R^2 = 0.04$ with rs3087243). Haplotype analysis indicated that rs1980422 imperfectly tags high-frequency haplotypes defined by rs3087243 and rs117701653 (Supplementary Fig. 25c). Both rs117701653 and rs3087243 overlap H3K4me3 peaks in immune cells and disrupt protein-binding motifs (Supplementary Tables 19–25 and Supplementary Note). Only rs3087243 was in linkage disequilibrium with an eQTL (*CTLA4* in $CD4^+$ and $CD8^+$ cells and testis; $R^2 > 0.8$; Supplementary Table 17).

We observed allele-specific protein binding and enhancer activity for rs117701653, but not rs3087243 (Fig. 3d). The rs117701653-C allele showed higher specific binding than the A allele in Jurkat T-cell nuclear extracts by EMSA (Supplementary Fig. 25d) and conferred higher luciferase expression ($P = 0.0017$; Fig. 3e). The binding is lineage specific: it was absent in THP-1 monocytic cells (Supplementary Fig. 25d). We observed peak overlap in an assay for transposase-accessible chromatin sequencing (ATAC-Seq) for rs117701653 in $CD4^+$ T cells (Supplementary Table 25) and a subtle increase in luciferase expression with the C allele after anti-CD3/CD28 cell stimulation ($P = 0.02$; Supplementary Fig. 25e), suggesting that binding may be stimulation dependent. While the rheumatoid arthritis credible set variant rs55686954 ($R^2 = 0.91$ with rs117701653, posterior probability = 0.27) showed allele-specific protein binding, it had no evidence of allele-specific enhancer function (Supplementary Fig. 25d,e). Promoter-capture Hi-C assays[32] demonstrated genomic contacts between the rs117701653 region and the *CTLA4* promoter and a region downstream of *RAPH1* (Supplementary Fig. 26), suggesting that, despite its proximity to *CD28*, the allele may influence *CTLA4* or *RAPH1* gene regulation.

MEG3 is a noncoding RNA tumor suppressor gene whose transcript binds p53 (ref. [33]). Paternal alleles carry greater risk[34] in this 14q32.2 T1D locus. It shows no association to rheumatoid arthritis ($P > 0.04$). Two variants with >0.2 probabilities in the credible set were the rs34552516 indel ($P = 1.1 \times 10^{-9}$; posterior probability = 0.42) and rs56994090 intronic variant ($P = 1.1 \times 10^{-9}$; posterior probability = 0.44; linkage disequilibrium with rs34552516, $R^2 = 0.99$; Fig. 4a, Supplementary Fig. 27a, and Supplementary Table 12). While we observed no evidence of independent variants conditioning on rs34552516 ($P > 0.04$; Supplementary Table 13), FINEMAP analysis could not exclude the possibility of a

secondary association (Supplementary Table 18 and Supplementary Note). Both *MEG3* variants overlap DNAse-I hypersensitive sites (DHS) and H3K4me3 regions in multiple cell types (Supplementary Tables 20 and 21), but do not overlap ATAC-Seq peaks after stimulation (Supplementary Table 25) and are not in high linkage disequilibrium ($R^2 > 0.8$) with QTL SNPs.

We observed that the rs34552516-TC allele demonstrated specific Jurkat cell nuclear extract binding (Fig. 4b) and increased luciferase activity compared with empty vector ($P = 0.01$) and the T allele ($P < 0.05$; Fig. 5c). We observed no specific binding in THP-1 cells (Supplementary Fig. 27b), indicating cell-type specificity. In contrast, we observed no allele-specific binding for rs56994090 (Fig. 4b). The region harboring rs34552516 in promoter-capture Hi-C data[32] showed contacts to *DIO3* and *RP11–1029J19* promoters (Supplementary Fig. 26), suggesting multiple downstream genes. We favor rs34552516 as potentially causal based on our functional evidence, but acknowledge that these assays are limited and cannot exclude rs56994090 function that may occur in other unexamined cellular contexts.

The 6q23.3 *TNFAIP3* gene encodes A20, which inhibits nuclear factor-κB signaling and prevents apoptosis[20]. It is associated with multiple autoimmune diseases[35–41], including rheumatoid arthritis, but not T1D ($P > 2.3 \times 10^{-4}$). The indel rs35926684 carries the highest posterior probability ($P = 6.5 \times 10^{-12}$; posterior probability = 0.24; Fig. 5a, Supplementary Table 13, and Supplementary Fig. 28a) of 9 variants in the credible set and is in linkage disequilibrium with the previously identified SNPs rs17264332 ($R^2 = 0.86$)[8] and rs6920220 ($R^2 = 0.88$)[3]. Conditional analysis identified independent association at rs58721818 ($P = 3.6 \times 10^{-5}$; $R^2 = 0.05$ with rs35926684; Fig. 5a and Supplementary Table 16). A previous study[3] identified secondary signals at rs5029937 (linked to rs58721818; $R^2 = 0.84$) and rs13207033. Exhaustive pairwise analysis demonstrated comparable association for the rs35926684 + rs58721818 pair ($-\log_{10}[P] = 13.95$) and the most strongly associated rs6920220 + rs58721818 pair ($-\log_{10}[P] = 14.21$; Fig. 5b and Supplementary Fig. 28b). Haplotypes with the rs35926684-G allele increased rheumatoid arthritis risk, even in absence of the highly linked rs6920220-A risk allele (that is, GGGC versus GAGC; Fig. 5c), although this effect was not significant in conditional haplotype analysis ($P = 0.14$). Consistent with our exhaustive pairwise search, FINEMAP identified the rs35926684 + rs58721818 combination as having the thirteenth highest posterior probability (Supplementary Table 18).

The rs35926684 indel alters binding motifs, overlaps enhancer marks and Encyclopedia of DNA Elements transcription factor binding site (TFBS) chromatin immunoprecipitation sequencing peaks in immune cell types (Supplementary Note and Supplementary Tables 20, 21, and 23–25), and is in linkage disequilibrium ($R^2 > 0.8$) with a methylation QTL in neutrophils (Supplementary Table 17). EMSA identified stronger specific binding of Jurkat-cell nuclear extract for the rs35926684-GA allele than the G allele (Fig. 5d and Supplementary Fig. 28c). Luciferase assays demonstrated increased enhancer activity with the GA allele compared with the empty vector ($P = 7 \times 10^{-4}$) and G allele ($P = 0.053$; Fig. 5e). We observed no specific binding in THP-1 cells, indicating cell-type specificity (Supplementary Fig. 28c). We observed no allele-specific binding for rs6920220

(Supplementary Fig. 28c) or the other four alleles in partial linkage disequilibrium with rs35926684 (Posterior probability > 0.1), except for rs62432712. We did not observe allele-specific enhancer activity for rs62432712 (Supplementary Fig. 28d). Hence, for this locus, we favor rs35926684 as the potentially causal variant since it has the best evidence of allele-specific activity. Nevertheless, the presence of multiple alleles in linkage disequilibrium (R^2 > 0.86) provides ambiguity at this locus. Interestingly, in promoter-capture Hi-C data, the rs35926684 region contacts the *TNFAIP3* promoter[42] and the *IL22RA* and *IFNGR1* promoters (Supplementary Fig. 26)[32], suggesting that multiple genes may be influenced by this rheumatoid arthritis risk allele.

Our study illustrates some challenges of fine-mapping. First, only a few loci had ≤10 plausible causal variants, and in even fewer was it possible to identify promising candidates. Nonetheless, identifying plausible candidates in even a few instances is valuable. Second, we identified multiple potentially causal variants that were indels. Indels are the most likely to be missed or poorly imputed by current imputation reference panels (Supplementary Note), although coverage could improve with more complete reference panels based on high-depth whole-genome sequencing data. Third, since most loci have many plausible variants with low posterior probabilities, strategies to accurately predict causal variation from functional annotations are critical. This will require more precise noncoding maps that define regulatory elements central to the functions of pathogenic cell types.

We used a posterior probability of >0.2 to prioritize variants for functional follow-up, which allowed us to include variants that might have been excluded by fluctuations in calculated probabilities caused by quality control, genotyping error, imputation quality, and parameter choices. This threshold is relatively stringent (only 42 variants passed it in our entire study), and variants with a posterior probability of <0.2 may also be worthy of further investigation. We focused on loci with ≤10 variants in the 95% credible set; 12 loci had >10 variants in the 95% credible set and at least 1 variant with a posterior probability of >0.2 (Supplementary Table 26). In the *RASGRP1* and *PRKQC* credible sets, we observed a single variant with tenfold higher posterior probability than the remaining variants, but we did not investigate these loci in detail since they had weaker association and their credible set size suggests extensive linkage disequilibrium, making determination of the functional impact more difficult.

We used Jurkat T-cell lines for EMSA and luciferase assays since T cells are critical to the genetic etiology of rheumatoid arthritis and T1D. We acknowledge that, in vivo, many contexts may be relevant, some of which may not be captured by these assays. Consequently, many of the remaining variants with a posterior probability of >0.2 are plausible candidates that cannot be ruled out. Furthermore, our study is limited in identifying causal molecular mechanisms and genes. First, variants may be linked to multiple genes. For example, the region harboring the rs117701653 variant shows chromatin contacts with the *CTLA4* promoter and the *RAPH1* gene. Second, only a limited number of prioritized variants were in linkage disequilibrium with different molecular QTLs. As such, the specific gene(s) accounting for disease susceptibility remain(s) to be determined. Despite these limitations, we believe that the combination of statistical evidence with functional follow-up is a powerful way to prioritize potentially causal variants. Defining cellular models that best

represent the pathogenic cellular contexts is an essential component for this functional follow-up.

Methods

Ethics.

This study complies with all relevant ethical regulations. The study protocol was approved as an exempt study by Brigham and Women's Hospital. Institutional Review Board approval for the original genotyping studies is described separately[3,4].

Patient collections.

We used genotyping data from samples collected on the ImmunoChip platform (Supplementary Table 1)[3,4]. For rheumatoid arthritis, we used data for 11,475 cases and 15,870 controls from 6 different cohorts (from the UK, the Swedish Epidemiological Investigation of Reumatoid Arthritis, the USA, the Netherlands, Umeå (Sweden), and Spain)[3]. For T1D, we used data for 12,241 cases and 14,636 controls from 2 different cohorts[4]: the Type 1 Diabetes Genetics Consortium (T1DGC) family collection, and the UK Genetic Resource Investigating Diabetes (GRID), British 1958 Birth Cohort, and UK Blood Service collection. To include trios from the Type 1 Diabetes Genetics Consortium cohort in the case-control analysis, we generated pseudocontrol pairs for each affected individual using the untransmitted alleles from the parents of that individual. As a consequence, the final numbers of individuals for T1D were 9,334 cases and 11,111 controls (including 1,661 pseudocontrols). Genotype quality control was performed as described in the previously published studies. Additionally, we merged the genotype data for the different cohorts within T1D and rheumatoid arthritis using PLINK[43], and converted genomic coordinates using the University of California, Santa Cruz liftOver tool[44] and the hg18ToHg19 chain file. Variants unable to liftOver were removed. We then replaced the variant identifiers using National Center for Biotechnology Information Database of Single Nucleotide Polymorphisms (dbSNP) build 138[45]. Finally, we removed variants with a MAF of <0.5%.

Imputation.

To assess the imputation strategy best suited for fine-mapping, we tested three reference panels: (1) the European subpopulation of 1000 Genomes (n = 503); (2) the cosmopolitan panel of 1000 Genomes (n = 2,504); and (3) the HRC version 1.1 reference panel (n = 32,611). We matched variants to each reference panel, removed variants absent in the reference panel, and aligned the strands of the remaining ImmunoChip genotypes. We extended the disease loci upstream and downstream by 1 Mb. We excluded variants when alleles could not be matched. For C/G and A/T variants, we removed the variant when the minor allele was unequal and the MAF was <45%. For multi-allelic variants, we ensured that the allele encoding was identical to the reference panel variant. This resulted in a different number of input variants for each imputation strategy (Supplementary Table 6). We imputed genotypes into rheumatoid arthritis and T1D separately. We phased and imputed the 1000 Genomes reference panels using Beagle version 4.1 (22Apr16.1cf)[46]. To accommodate computational constraints of Beagle, we split the rheumatoid arthritis and T1D datasets into 30 batches, randomizing cases and controls between batches, while maintaining trio

structure in the T1D dataset. Since the HRC version 1.1 reference panel genotype data are not publicly available, we evaluated different imputation servers and settings for the T1D dataset, to determine their effects on imputation output. On the Sanger Institute imputation server (date of access: 11 May 2016), we used prephasing with either EAGLE (version 2.3.4)[47] or SHAPEIT (version 2.r837)[48], followed by imputation with PBWT (version 3.0)[49]. On the Michigan University server (date of access: 5 July 2016), we split the dataset into three batches and used prephasing with EAGLE[47] and imputation by MiniMac[50]. For rheumatoid arthritis, we performed HRC imputation on the Sanger imputation server using EAGLE prephasing followed by PBWT imputation. Finally, we locally performed 1000 Genomes imputation by first phasing with EAGLE, and subsequently imputing using PBWT. We then merged the imputed dosages and probabilities from each batch (if any) for each imputation reference panel and replaced the variant identifiers in the imputed output using National Center for Biotechnology Information dbSNP build 138. We replaced genotypes for variants genotyped on ImmunoChip with the original genotypes. Genotyped variants correlated with genotypes after imputation ($R^2 > 0.99$). Finally, we recalculated the imputation quality scores for each imputed variant in each dataset: we used the INFO score for biallelic variants and Beagle version 4.1 allelic-R^2 for multi-allelic variants.

Targeted sequencing.

To test the accuracy of imputation, we sequenced 900 regions of 1,000 base pairs (bp) around H3K4me3 peak centers overlapping loci associated with either disease in 864 individuals: 384 unrelated rheumatoid arthritis and 480 T1D individuals (160 trios), of which 149 and 480 were on ImmunoChip, respectively. We generated 100-bp paired-end reads with the Illumina MiSeq platform. Sequencing was performed at The Feinstein Institute for Medical Research at Northwell Health, and at The Center for Public Health Genomics, at the University of Virginia. We used BWA-mem[51] (version 0.7.12) to align reads to the hg19 reference genome. We tagged and removed duplicate reads using Picard MarkDuplicates. We removed 101 regions where >50% of the samples had <20× coverage at >80% of sequenced bases, and removed 86 samples having <20× coverage at 90% of sequenced bases. We called genotypes using GATK version 3.4, following the recommended guidelines for using HaplotypeCaller[52] in a joint genotype-calling approach. To determine the impact of local alignment on indel calls, we also called variants using the UnifiedGenotyper present in GATK. We then set genotypes with <10× coverage and genotype quality (QUAL) <30 to missing, and excluded variants with >5% missingness. We correlated called genotypes with ImmunoChip genotypes to identify and remove (when the coefficient of correlation (r) < 0.95) possible sample swaps and mismatched samples, resulting in 568 final samples (439 for T1D and 129 for rheumatoid arthritis). Finally, we selected variants with MAF > 1%, resulting in 1,862 variants within the 76 rheumatoid arthritis- and T1D-associated regions.

Combined dataset.

Before the association analysis, we merged the data for the rheumatoid arthritis and T1D dataset, imputed with the cosmopolitan reference panel of 1000 genomes. We identified shared controls between datasets by generating a list of linkage disequilibrium pruned variants from the ImmunoChip genotypes using PLINK[43] (using --indep-pairwise 1000 100

0.2) and then used this list to determine the genetic similarity (unified additive relationship)[53] between each pair of samples across both datasets. We considered sample pairs with a unified additive relationship of >0.2 genetically related, and randomly selected one sample of the pair to be included. We considered the remaining sample pairs unrelated. We finally merged genotypes and imputation probabilities from the selected samples and recalculated the imputation INFO scores for the merged genotypes as described earlier.

Fine-mapping and statistical analysis.

We limited our association analysis to variants with an overall MAF of >1%, a Hardy-Weinberg P value of $> 10^{-5}$ in controls, and an overall INFO score >0.3. The Hardy–Weinberg P value was calculated using an exact test for biallelic variants and a Pearson's chi-squared test for multi-allelic variants. We then split multi-allelic variants, creating a single variant for each alternative allele. To test each variant for association with disease, we used logistic regression, assuming a log-linear relation between the number of alternative alleles and case-control status. In the rheumatoid arthritis dataset, the null model included the first ten principal components calculated over the genotype covariance matrix, as described previously[3], and five additional covariates indicating the cohort. For T1D, we included 12 regional indicator variables in the null model, as described previously[4], and an additional variable indicating the cohort. For the joint analysis, the null model included all covariates for the T1D and rheumatoid arthritis datasets and an additional covariate indicating the sample originating dataset. We recoded the imputation probabilities to a dosage value ranging between 0 and 2 (that is $P(\mathrm{AB}) + 2 \times P(\mathrm{BB})$). Finally, we calculated the P value for the association as the difference in deviance between the null model and alternative model containing the imputation probabilities, which follows a chi-squared distribution with one degree of freedom. We corrected for multiple testing using a study-wide Bonferroni threshold using the maximum number of tests across datasets ($P < 7.5 \times 10^{-5} \sim 0.05/67{,}156$). To test whether our model was properly adjusting for population stratification when performing the combined analysis, we also evaluated using the first 20 principal components as covariates. We obtained principal components with the PLINK[43] –pca command using the non-imputed and pruned combined ImmunoChip genotypes. By also including a covariate indicating the source dataset of each individual, we accounted for any residual technical differences caused by rheumatoid arthritis and T1D samples being genotyped and imputed independently.

Definition of credible sets.

To define potentially causal variants for each locus, we calculated posterior probabilities using the approximate Bayesian factor (ABF)[1,2] under the assumption of a single causal variant per locus. This framework assumes that the association effect sizes follow a distribution of $N(0, V)$ under H_0, with V being the squared standard error. Under H_1, the framework assumes a distribution following $N(0, V + W)$, where W is $(\ln[1.5]/1.96)^2$, reflecting the prior of observing an odds ratio of 1.5. The ABF for an observed effect size β is then calculated as the ratio of $P(\beta|H_0)/P(\beta|H_A)$. Using the sum of the ABF for all variants in the locus, we calculate the posterior probability (PP) for variant i as:

$$PP_i = \frac{ABF_i}{\sum_{k=0}^{n} ABF_k}$$

Following calculation of the posterior probability, we created credible sets by sorting associations descending on the basis of their posterior probability and including associations such that the sum of posterior probability was >0.95.

Detecting secondary associations.

To determine the presence of multiple independent effects, we performed conditional analyses using logistic regression. For each locus with a significant association, we included the top-associated variant as a covariate in the null and alternative models and repeated the association analysis for that locus. We considered secondary associations significant when $P < 1.9 \times 10^{-5}$ (Bonferroni correction for maximum number of variants in significant loci: 0.05/2,704). We then performed exhaustive pairwise association analyses to test whether the primary and secondary associations together provided the strongest pairwise association signal given all possible pairs of variants in the locus. We calculated a *P* value using the difference in deviance between the null and alternative models, following a chi-squared distribution with two degrees of freedom.

Finally, for loci with two or more independent associations, we assessed whether the risk alleles for the associated variants were located on the same haplotypes. We derived haplotypes from the phased imputation output (for example, four haplotypes for two independent variants). We removed all haplotypes with a frequency of <1% and individuals having those haplotypes, and used the haplotype with the highest frequency as a reference. We then used logistic regression to test remaining haplotypes for association, assuming a log-linear relationship between the number of haplotype copies and disease status. To correct for population differences, our null model included covariates as described above.

We also performed fine-mapping using FINEMAP version 1.1[16], which allows multiple independent associations per locus. As input, we used the summary statistics for the individual disease association analyses and genotype correlation matrices as linkage disequilibrium estimates. Finally, we assessed whether opposite effects between rheumatoid arthritis and T1D may cause lower power in the combined analysis by applying Trinculo version 0.96[15]—a method that implements multinomial logistic regression. As covariates for this analysis, we used the 20 principal components and indicator variable described above, to prevent multicollinearity.

Overlap with eQTLs, H3K4me3 peaks, DNAse-I hypersensitive sites, enhancers, and motifs.

To provide functional annotation for the identified variants, we assessed overlap with eQTLs, H3K4me3 peaks, DNAse-I hypersensitive sites, promoters, and enhancers. We used eQTLs from an RNA sequencing-based eQTL meta-analysis of 2,116 whole blood samples[54], a study assessing eQTLs in $CD4^+$ T cells from 461 individuals[55], a study assessing eQTLs in $CD4^+$ and $CD8^+$ cells from 313 individuals[56], and tissue-specific eQTLs

from the Genotype–Tissue Expression (GTEx) project[57]. We also included molecular QTLs, such as histone QTLs, methylation QTLs, and protein QTLs[56,58]. For each variant in a credible set, we considered a QTL to be overlapping when it was in high linkage disequilibrium ($R^2 > 0.8$) with the top QTL for a given gene, methylation probe, histone mark, or protein. For calculation of linkage disequilibrium, we used the European subpopulation of 1000 Genomes.

For further annotation, we determined the overlap of variants with a posterior probability of >0.2 with H3K4me3 peaks, DNAse-I peaks, and ChromHMM[59] genome segmentations from 12 imputed epigenetic marks from the Roadmap Epigenomics Consortium[60], consisting of 127 consolidated epigenomes from different cell types. Furthermore, we determined whether candidate causal variants affected protein-binding motifs or transcription factor binding sites using HaploReg[61]. Finally, we determined overlap with TFBSs using Encyclopedia of DNA Elements project transcription factor chromatin immunoprecipitation sequencing[62] and determined whether these variants overlapped conserved TFBS motifs by defining a 40-bp region around each variant, and using the Homer[63] software to test all known motifs in vertebrates.

ATAC-Seq time series.

We applied ATAC-Seq[64] to measure chromatin accessibility in a time series after stimulation. We used a leukopak (30 ml whole blood) from a healthy anonymous donor to isolate peripheral blood mononuclear cells using Ficoll tubes, which were stored in 500 μl aliquots of 100×10^6 cells in liquid nitrogen. Cells were subsequently thawed and stained with anti-biotin microbeads to magnetically select $CD4^+$ Tmem cells. Cells were resuspended and transferred to 24-well plates in 3 ml aliquots of 6×10^6 cells and stimulated using Dynabeads (Human T-Activator CD3/CD28 for T Cell Expansion and Activation; Life Technologies) in a ratio of two cells per bead. Samples of 100,000 cells were taken at 0, 1, 2, 4, 8, 12, 24, and 48 h after stimulation. Nucleosome isolation and ATAC-Seq open chromatin sequencing were performed as described earlier[64]. Sequenced reads were mapped to the hg19 reference genome using BWA-mem. Reads mapping to the mitochondrial genome or to multiple genomic locations, and duplicate reads (labeled by Picard MarkDuplicates (version 1.128)) were removed, and reads were shifted +4 and −5 bp for the reverse and forward strands, respectively. Enrichment for open chromatin was determined by calling peaks using MACS version 2.1.0 (ref. [65]) (default settings).

Cell lines.

Jurkat and THP-1 cell lines were obtained from the American Type Culture Collection (TIB-152 and TIB-202). Jurkat cells were grown in complete Roswell Park Memorial Institute (RPMI) (RPMI-1640; Gibco, with 10% decomplemented fetal bovine serum, penicillin and streptomycin) and THP-1 cells in complete RPMI supplemented with 2-mercaptoethanol to a final concentration of 0.05 mM. Both cell lines were grown in a 37°C incubator with 5% CO_2.

EMSAs.

EMSAs were performed using the LightShift Chemiluminiscent EMSA Kit (Thermo Scientific). Single-stranded oligonucleotides corresponding to 30–32 nucleotide fragments of the human genome with the SNP of interest in the middle were purchased from Integrated DNA Technologies (Supplementary Table 27) and biotinylated using the Biotin 3′ End DNA Labeling Kit (Thermo Scientific). Double-stranded oligonucleotides were generated by mixing equal amounts of biotin-labeled (for the probe) or unlabeled (for the competitor) complementary oligonucleotides and incubated for 5 min at 95°C and then 1 h at room temperature.

Nuclear extract from Jurkat and THP-1 cells was obtained using the NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific). Protein extracts were dialyzed using a dialysis membrane with a molecular weight cutoff of 12–14 kDa (Spectrum Spectra) against 1 l of dialysis buffer (10 mM Tris pH 7.5, 50 mM KCl, 200 mM NaCl, 1 mM dithiothreitol, 1 mM phenylmethane sulfonyl fluoride, and 10% glycerol) for 16 h at 4°C with slow stirring. Protein inhibitor cocktail (Sigma) was added to a final concentration of 2.5×. The protein concentration was measured using the Pierce BCA Protein Assay Kit (Thermo Scientific).

The standard binding reaction contained 2 μl of 10× Binding Buffer (100 mM Tris pH 7.5, 500 mM KCl and 10 mM dithiothreitol), 2.5% glycerol, 5 mM $MgCl_2$, 0.05% NP40, 50 ng Poly(dI:dC), 20 fmol biotin-labeled probe, and 16 μg nuclear extract in a final volume of 20 μl. For competition experiments, a 200-fold molar excess (4 pmol) of unlabeled probe was added. Variations to these conditions are indicated in the corresponding figure (Supplementary Figures 25D, 27B, and 28C).

Binding reactions were incubated at room temperature for 30 min and loaded onto a 6% polyacrylamide 0.5× TBE Gel. After sample electrophoresis and transfer to a nylon membrane, transferred DNA was crosslinked for 10 min, and the biotinylated probes were detected by chemiluminescence followed by film exposure. Original films are presented in Supplementary Fig. 29.

Luciferase reporter assay.

The double-stranded oligonucleotide containing the SNP of interest (obtained as described above) was cloned downstream from the luciferase gene in the luciferase reporter vector pGL3 promoter (Promega). Unlabeled double-stranded oligonucleotides containing the corresponding SNPs were amplified with specific primers containing the BamHI restriction site obtained from Integrated DNA Technologies (Supplementary Table 28). The PCR was carried out in 50 μl reaction volume under the following program: 94 °C for 3 min; 10 cycles at 94 °C for 30 s, 60 °C for 40 s, and 68 °C for 30 s; 15 cycles at 94 °C for 30 s, 60 °C for 40 s, and 68 °C for 30 s; and 72 °C for 10 min (AccuPrime Taq; Invitrogen). Both the PCR products and pGL3 promoter vector were digested with BamHI (New England Biolabs) for 1h at 37 °C, and linearized vector was then dephosphorylated for 30 min at 37 °C with the Quick Dephosphorylation kit (New England Biolabs). Digestion products were purified with the QIAquick Gel Extraction Kit (Qiagen) from 1.2% agarose gels. Ligation of SNP-

containing fragments into the pGL3 promoter plasmid was performed in a ratio of 1:50 (vector:insert) with T4 DNA ligase (New England Biolabs) at 16 °C overnight and transformed into JM109 competent cells (Promega). Plasmids from independent colonies were isolated using a Wizard Plus SV Minipreps DNA purification system and sequenced using RV primer 4 (Promega), selecting those harboring the SNP-construct cloned 'in sense' in the pGL3 promoter vector for further HighPure plasmid isolation (Invitrogen).

At least three independent transfection experiments for each construct were performed, each in duplicate. 2×10^4 Jurkat cells in 0.1 ml of complete RPMI were transfected with 0.8 μg of pGL3-Promoter vector along with 0.2 μg of pRL-TK *Renilla* luciferase vector (Promega) using 1.5 μl of Lipofectamine LTX Reagent and 1 μl of PLUS Reagent (both from Invitrogen) diluted in Opti-MEM (Gibco). After 16 h of transfection, luciferase activity was measured using the Dual-Glo Luciferase Assay System (Promega). Firefly luciferase activity was expressed as relative luciferase units (RLU) after correction for *Renilla* luciferase activity to adjust for transfection efficiency. Data were normalized to those cells transfected with empty pGL3-Promoter vector.

For the rs117701653 variant at the *CD28-CTLA4* locus, we also investigated the luciferase signal under stimulatory conditions. We transfected Jurkat cells as described above, and 18 h after transfection, cells were left untreated or stimulated with αCD3/αCD28 (0.5 μg αCD3 coated to the plate and 5 μg ml^{-1} of soluble α CD28) or phytohemagglutinin (2 μg ml^{-1}). Luciferase activity was measured 6 h after stimulation. Four independent experiments were performed in duplicate. The results of individual assays are presented in Supplementary Table 29.

Statistical analysis of functional studies.

Luciferase activity levels were compared by unpaired two-sided t-test. Error bars represent s.d.

Reporting Summary.

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability.

Associated computer code for this manuscript can be found at the following GitHub repositories: https://github.com/immunogenomics/harmjan/tree/master/FinemappingPaper and https://github.com/immunogenomics/harmjan/tree/master/FinemappingTools.

Data availability

Summary statistics for all variants are available through the following GitHub repository: https://github.com/immunogenomics/harmjan/tree/master/RA-T1D-Finemap-SummaryStats. Genotype data have been previously published[3,4] and are available from Rheumatoid Arthritis Consortium International and the Type 1 Diabetes Genetics Consortium upon

request. The ATAC-Seq data discussed in this publication have been deposited in the Gene Expression Omnibus under accession number GSE116497.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported in part by funding from the National Institutes of Health (U01GM092691, UH2AR067677, 1U01HG009088, and 1R01AR063759 to S.R.), and Doris Duke Charitable Foundation Grant number 2013097. This work is part of the research program Rubicon ALW with project number 825.14.019 (H.-J.W), which is partly financed by the Netherlands Organization for Scientific Research. Further support was provided by Wellcome (107212/Z/15/Z) and the Juvenile Diabetes Research Foundation (5-SRA-2015-130-A-N) to the Diabetes and Inflammation Laboratory, and by Wellcome (203141/Z/16/Z) to the Wellcome Centre for Human Genetics (J.A.T.). P.K.G. was supported in part by the Feinstein Institute and a generous gift from E. L. Greenland. P.A.N. is supported by a Rheumatology Research Foundation Disease Targeted Research Grant, NIH P30 AR070253 and R01 AR065538, and the Fundación Bechara. S.S.R., W.-M.C. and S.O. were supported in part by funding from the National Institutes of Health (R01DK096926). This research makes use of 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, National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development and Juvenile Diabetes Research Foundation International, and is supported by grant U01DK062418 (S.S.R.).

References

1. Maller JB et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet 44, 1294–1301 (2012). [PubMed: 23104008]
2. Wakefield J A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am. J. Hum. Genet 81, 208–227 (2007). [PubMed: 17668372]
3. Eyre S et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet 44, 1336–1340 (2012). [PubMed: 23143596]
4. Onengut-Gumuscu S et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat. Genet 47, 381–386 (2015). [PubMed: 25751624]
5. Klareskog L, Catrina AI & Paget S Rheumatoid arthritis. Lancet 373, 659–672 (2009). [PubMed: 19157532]
6. Palmer JP et al. Insulin antibodies in insulin-dependent diabetics before insulin treatment. Science 222, 1337–1339 (1983). [PubMed: 6362005]
7. Baekkeskov S et al. Identification of the 64K autoantigen in insulin-dependent diabetes as the GABA-synthesizing enzyme glutamic acid decarboxylase. Nature 347, 151–156 (1990). [PubMed: 1697648]
8. Okada Y et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381 (2014). [PubMed: 24390342]
9. Hu X et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am. J. Hum. Genet 89, 496–506 (2011). [PubMed: 21963258]
10. Finucane HK et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet 50, 621–629 (2018). [PubMed: 29632380]
11. Trynka G et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet 45, 124–130 (2013). [PubMed: 23263488]
12. Huang H et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 547, 173–178 (2017). [PubMed: 28658209]
13. Gaulton KJ et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat. Genet 47, 1415–1425 (2015). [PubMed: 26551672]

14. Farh KK et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015). [PubMed: 25363779]
15. Jostins L & McVean G Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. Bioinformatics 32, 1898–1900 (2016). [PubMed: 26873930]
16. Benner C et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501 (2016). [PubMed: 26773131]
17. Begovich AB et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am. J. Hum. Genet 75, 330–337 (2004). [PubMed: 15208781]
18. Bottini N et al. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. Nat. Genet 36, 337–338 (2004). [PubMed: 15004560]
19. Diogo D et al. TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. PLoS One 10, e0122271 (2015). [PubMed: 25849893]
20. Sisirak V et al. Digestion of chromatin in apoptotic cell microparticles prevents autoimmunity. Cell 166, 88–101 (2016). [PubMed: 27293190]
21. Zochling J et al. An ImmunoChip-based interrogation of scleroderma susceptibility variants identifies a novel association at *DNASE1L3*. Arthritis Res. Ther 16, 438 (2014). [PubMed: 25332064]
22. Al-Mayouf SM et al. Loss-of-function variant in *DNASE1L3* causes a familial form of systemic lupus erythematosus. Nat. Genet 43, 1186–1188 (2011). [PubMed: 22019780]
23. Ueki M et al. Caucasian-specific allele in non-synonymous single nucleotide polymorphisms of the gene encoding deoxyribonuclease I-like 3, potentially relevant to autoimmunity, produces an inactive enzyme. Clin. Chim. Acta 407, 20–24 (2009). [PubMed: 19559017]
24. Nettleship JE et al. Crystal structure of signal regulatory protein gamma (SIRPγ) in complex with an antibody Fab fragment. BMC Struct. Biol 13, 13 (2013). [PubMed: 23826770]
25. Brooke G, Holbrook JD, Brown MH & Barclay AN Human lymphocytes interact directly with CD47 through a novel member of the signal regulatory protein (SIRP) family. J. Immunol 173, 2562–2570 (2004). [PubMed: 15294972]
26. Piccio L et al. Adhesion of human T cells to antigen-presenting cells through SIRPbeta2-CD47 interaction costimulates T-cell proliferation. Blood 105, 2421–2427 (2005). [PubMed: 15383453]
27. Liu Y et al. Functional elements on SIRPalpha IgV domain mediate cell surface binding to CD47. J. Mol. Biol 365, 680–693 (2007). [PubMed: 17070842]
28. Gusev A et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet 95, 535–552 (2014). [PubMed: 25439723]
29. Sveinbjornsson G et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nat. Genet 48, 314–317 (2016). [PubMed: 26854916]
30. Simeonov DR et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature 549, 111–115 (2017). [PubMed: 28854172]
31. Fortune MD et al. Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. Nat. Genet 47, 839–846 (2015). [PubMed: 26053495]
32. Javierre BM et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 167, 1369–1384. e19 (2016). [PubMed: 27863249]
33. Zhou Y et al. Activation of p53 by MEG3 non-coding RNA. J. Biol. Chem 282, 24731–24742 (2007). [PubMed: 17569660]
34. Wallace C et al. The imprinted *DLK1-MEG3* gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat. Genet 42, 68–71 (2010). [PubMed: 19966805]
35. Tsoi LC et al. Identification of 15 new psoriasis susceptibility loci highlight: the role of innate immunity. Nat. Genet 44, 1341–1348 (2012). [PubMed: 23143594]
36. Jostins L et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124 (2012). [PubMed: 23128233]

37. Beecham AH et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat. Genet 45, 1353–1360 (2013). [PubMed: 24076602]
38. Lessard CJ et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. Nat. Genet 45, 1284–1292 (2013). [PubMed: 24097067]
39. Cordell HJ et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat. Commun 6, 8019 (2015). [PubMed: 26394269]
40. Bentham J et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat. Genet 47, 1457–1464 (2015). [PubMed: 26502338]
41. 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). [PubMed: 22057235]
42. McGovern A et al. Capture Hi-C identifies a novel causal gene, *IL20RA*, in the pan-autoimmune genetic susceptibility region 6q23. Genome. Biol 17, 212 (2016). [PubMed: 27799070]
43. 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). [PubMed: 17701901]
44. Hinrichs AS et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34, D590–D598 (2006). [PubMed: 16381938]
45. Sherry ST et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311 (2001). [PubMed: 11125122]
46. Browning BL & Browning SR Genotype imputation with millions of reference samples. Am. J. Hum. Genet 98, 116–126 (2016). [PubMed: 26748515]
47. Loh P-R et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet 48, 1443–1448 (2016). [PubMed: 27694958]
48. Delaneau O, Marchini J & Zagury J-F A linear complexity phasing method for thousands of genomes. Nat. Methods 9, 179–181 (2011). [PubMed: 22138821]
49. Durbin R Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics 30, 1266–1272 (2014). [PubMed: 24413527]
50. Fuchsberger C, Abecasis GR & Hinds DA minimac2: faster genotype imputation. Bioinformatics 31, 782–784 (2015). [PubMed: 25338720]
51. Li H & Durbin R Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009). [PubMed: 19451168]
52. Van der Auwera GA et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33 (2013). [PubMed: 25431634]
53. Powell JE, Visscher PM & Goddard ME Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet 11, 800–805 (2010). [PubMed: 20877324]
54. Zhernakova DV et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet 49, 139–145 (2016). [PubMed: 27918533]
55. Raj T et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science 344, 519–523 (2014). [PubMed: 24786080]
56. Yang J, Fritsche LG, Zhou X & Abecasis G A scalable Bayesian method for integrating functional information in genome-wide association studies. Am. J. Hum. Genet 101, 404–416 (2017). [PubMed: 28844487]
57. Lonsdale J et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet 45, 580–585 (2013). [PubMed: 23715323]
58. Sun BB et al. Genomic atlas of the human plasma proteome. Nature 558, 73–79 (2018). [PubMed: 29875488]
59. Ernst J & Kellis M ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9, 215–216 (2012). [PubMed: 22373907]
60. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]

61. Ward LD & Kellis M HaploRegv4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 44, D877–D881 (2016). [PubMed: 26657631]

62. Bernstein BE et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]

63. Heinz S et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010). [PubMed: 20513432]

64. Buenrostro JD, Giresi PG, Zaba LC, Chang HY & Greenleaf WJ Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013). [PubMed: 24097267]

65. Zhang Y et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008). [PubMed: 18798982]

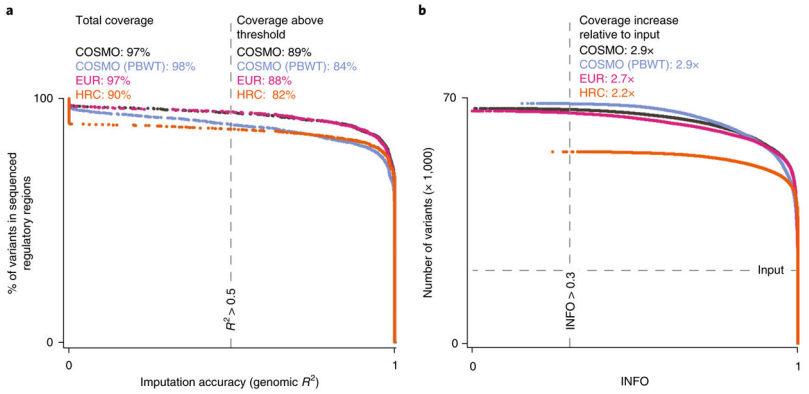

Fig. 1 |. Imputation accuracy and quality of datasets.

Datasets were imputed with different reference panels: the European subpopulation of 1000 Genomes (EUR), full 1000 Genomes (COSMO), full 1000 Genomes imputed with PBWT (COSMO(PBWT)), and HRC. **a**, We sequenced 799 1-kilobase regions in 568 individuals with ImmunoChip genotypes and called 1,854 common (MAF>1%) variants. We calculated the imputation accuracy (genomic R^2) by correlating imputed genotypes using each reference panel with genotypes called from the sequencing experiment. **b**, INFO scores for each reference panel in the rheumatoid arthritis dataset.

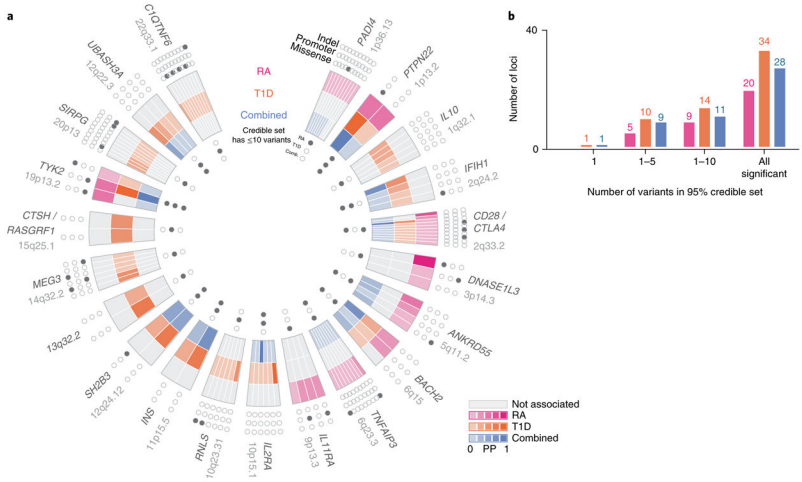

Fig. 2 | Variants in the 95% credible sets of significant loci determined by the Bayesian factor. **a**, The inner ring of dots indicates whether the locus has ≤10 variants in the credible set and a significant association signal (filled circles). Comb., combined; RA, rheumatoid arthritis. The middle ring shows variants in each credible set. Highlighted segments indicate loci with a candidate causal variant. The color intensity indicates the posterior probability (PP), gray representing a lack of significance. The outer ring shows indel, promoter, and missense coding annotation for each variant in the credible set. **b**, Number of variants in the 95% credible sets within significant loci. We narrowed down the list of probable causal variants to ≤5 in 5 out of 20 significant RA loci, and 10 out of 34 significant T1D loci.

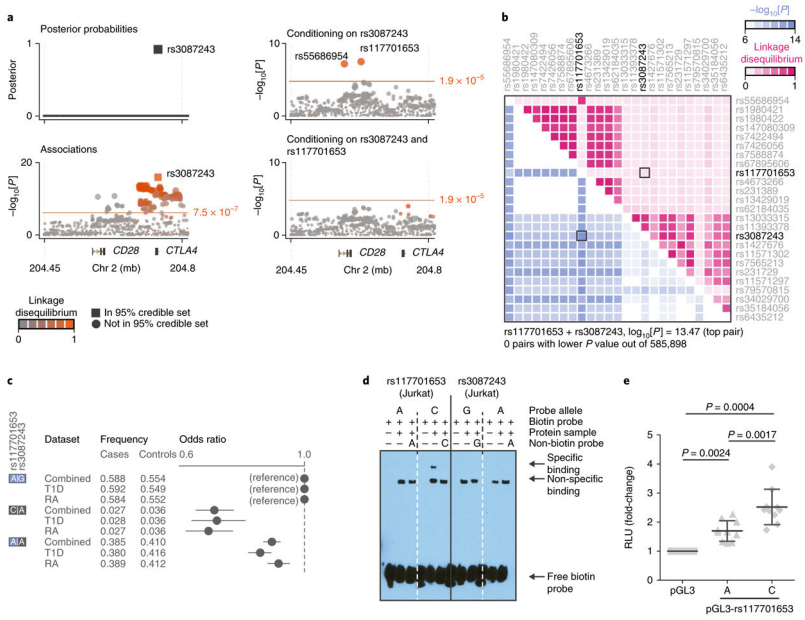

Fig. 3 | Analysis of the *CD28-CTLA4* locus.

a, A regional association plot for the combined analysis (20,787 rheumatoid arthritis or T1D cases, and 18,616 unique controls; Methods) shows a single variant (rs3087243) near *CTLA4* in the credible set. Conditioning on rs30872043 identifies rs117701653 as an independent association. Logistic regression $-\log_{10}[P]$ values are from a two-sided χ^2 test (n = 39,403). **b**, An exhaustive pairwise analysis for rheumatoid arthritis shows that the rs3087243 + rs117701653 pair has the strongest association. Logistic regression $-\log_{10}[P]$ values are from a two-sided χ^2 test (n = 27,345). **c**, Haplotype analysis using rs30872043 and rs117701653, with the AG haplotype as a reference. The C allele of rs117701653 shows the largest decrease in risk for rheumatoid arthritis (RA), and the A allele of rs30872043 shows the largest decrease in risk for T1D. Logistic regression odds ratios (dots) and 95% confidence intervals (bars) are from a two-sided χ^2 test (combined n = 39,403; T1D n = 20,445; RA n = 27,345). **d**, EMSA with Jurkat nuclear extract using probes containing rs117701653 and rs3087243. A representative blot of three independent experiments is shown. **e**, Luciferase assay in Jurkat T cells transfected with pGL3 plasmids containing rs117701653. RLUs are normalized to cells transfected with the empty plasmid (pGL3). Dots represent independent experiments, each of which was performed in duplicate. Means ±s.d. and two-tailed P values of the grouped comparisons (unpaired t-test) are shown.

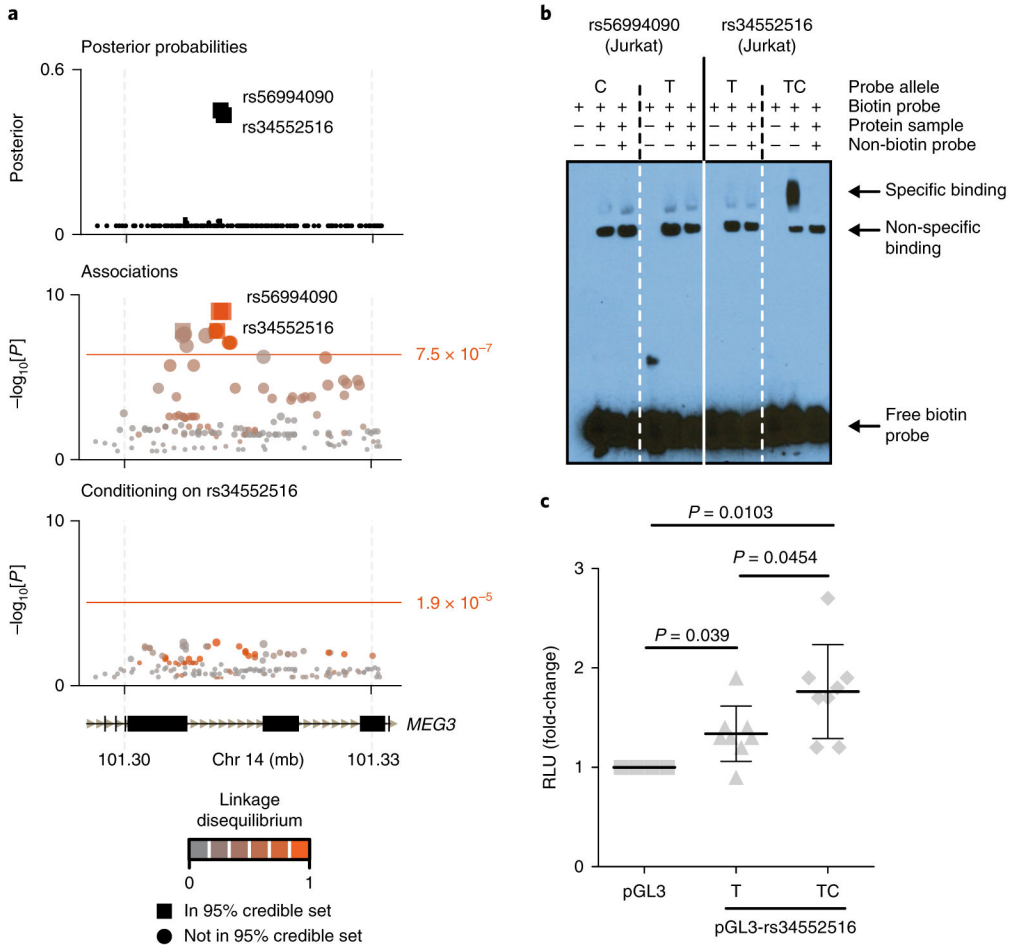

Fig. 4 |. Analysis of the *MEG3* locus.

a, Regional plot for the *MEG3* locus in T1D (9,334 cases, 11,111 controls). We observe two variants in the credible set (rs56994090 and the rs34552516 indel). We did not observe secondary signals when conditioning on rs56994090. Logistic regression $-\log_{10}[P]$ values are from a two-sided χ^2 test (n = 20,445). **b**, EMSA with Jurkat nuclear extract using probes containing rs354552516 and rs56994090. A representative blot of three independent experiments is shown. **c**, Luciferase assay in Jurkat T cells transfected with pGL3 plasmids containing rs34552516. RLUs are normalized to cells transfected with the empty plasmid (pGL3). Dots represent independent experiments, each of which was performed in duplicate. Means±s.d. and two-tailed P values of the grouped comparisons (unpaired t-test) are shown.

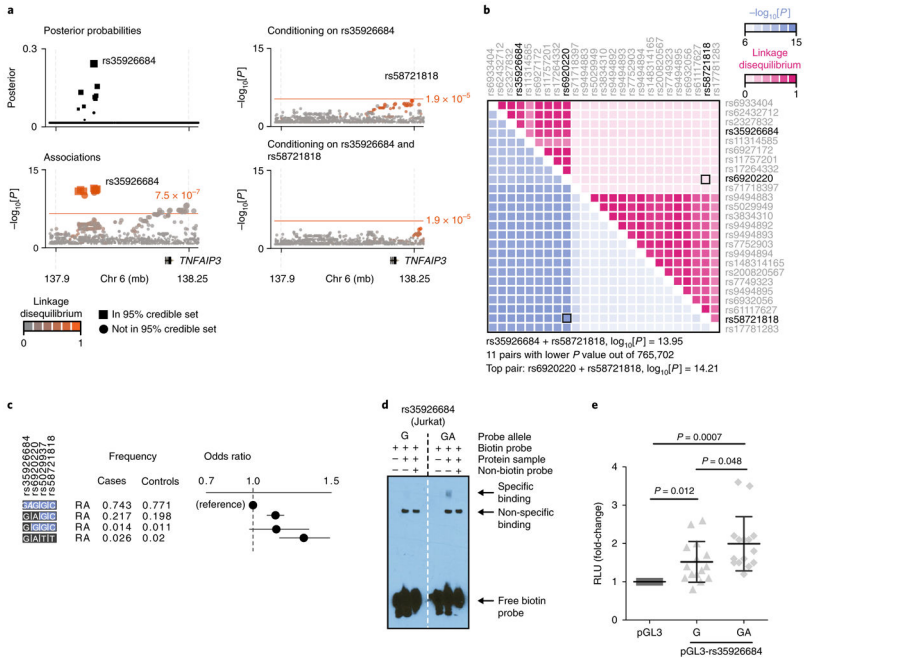

Fig. 5 | Analysis of the *TNFAIP3* locus.

a, Regional plot for the *TNFAIP3* locus in rheumatoid arthritis (11,475 cases, 15,870 controls). The variant with the strongest posterior probability in this locus is rs35926684, a G/GA indel. Conditional on rs35926684, we observe a significant secondary association with rs58721818. Logistic regression $-\log_{10}[P]$ values are from a two-sided χ^2 test (n = 27,345). **b**, Exhaustive pairwise association analysis in rheumatoid arthritis indicates that there are 11 pairs with a lower *P* value than rs35926684 + rs58721818, although the top-associated pair (rs69220220 + rs58721818) has an equivalent *P* value $-\log_{10}[P]$ =13.95 versus 14.21. Logistic regression $-\log_{10}[P]$ values are from a two-sided χ^2 test (n = 27,345). **c**, Haplotype analysis with rs35926684 + rs58721818 and previously reported variants rs6920220 and rs5029937 shows that rs35926684 and the previously reported top variant rs6920220 are often located on the same haplotype (GAGC), although a rare haplotype exists with only the alternative allele of rs35926684, which causes a similar increase in risk, but with a larger confidence interval. Logistic regression odds ratios (dots) and 95% confidence intervals (bars) are from a two-sided χ^2 test (n = 27,345). RA, rheumatoid arthritis **d**, EMSA with Jurkat nuclear extract using probes containing the G or GA allele of rs35926684. A representative blot of three independent experiments is shown. **e**, Luciferase assay in Jurkat T cells transfected with pGL3 plasmids containing rs35926684. RLUs are normalized to cells transfected with the empty plasmid (pGL3). Dots represent independent experiments, each of which was performed in duplicate. Means±s.d. and two-tailed *P* values of comparisons (unpaired *t*-test) are shown.

Table 1 |

Overview of potentially causal variants at loci having significant association and 95% credible sets with ≤10 variants

PP > 0.2	Potentially causal	Locus	Marker gene(s)	Variant	Disease association	Alleles	Previously identified as top variant	OR for RA	OR for T1D	PP for RA	PP for T1D	PP combined	Variant type
RA		1p13.2	*PTPN22*	rs6679677	RA + T1D	C, A		1.60	1.95	0.48			Noncoding
RA+T1D	RA + T1D			rs2476601	RA + T1D	G, A	Ref. [8]	1.60	1.95	0.52	0.92	0.90	p.Arg620Trp
T1D		1q32.1	*IL10*	rs3024505	T1D	G, A	Ref. [4]	1.02	0.85		0.29		Noncoding
T1D				rs3024495	T1D	C, T		1.02	0.85		0.31		Noncoding
T1D				rs3024493	T1D	C, A		1.02	0.85		0.32		Noncoding
T1D		2q24.2	*IFIH1*	rs2111485	T1D	G, A	Ref. [4]	0.96	0.85		0.46	0.74	Noncoding
T1D				rs984971	T1D	A, G		0.97	0.85		0.31		Noncoding
RA		2q33.2	*CD28 CTLA4*	rs55686954	RA	G, A		0.75	0.80	0.27			Noncoding
RA	RA			rs117701653	RA	A, C		0.74	0.79	0.67			Noncoding
T1D				rs3087243	T1D; RA (secondary)	G, A	Refs. [4,8]	0.90	0.85		0.48	0.91	Noncoding
RA	RA	3p14.3	*DNASE1L3*	rs35677470	RA	G, A		1.21	1.01	0.82			p.Arg206Cys
T1D		5q11.2	*ANKRD55*	rs10213692	RA + T1D	T, C		0.81	0.90		0.34		Noncoding
				rs11377254	RA + T1D	A, AT		0.79	0.89			0.25	Indel
RA				rs7731626	RA + T1D	G, A	Ref. [8]	0.80	0.90	0.41		0.29	Noncoding
		6q15	*BACH2*	rs10944479	T1D	G, A		1.12	1.19			0.21	Noncoding
RA				rs72928038	RA + T1D	G, A	Ref. [4]	1.13	1.20	0.37	0.27	0.61	Noncoding
RA				rs6908626	RA + T1D	G, T		1.14	1.21	0.29			Noncoding
RA	RA	6q23.3	*TNFAIP3*	rs35926684	RA	GA, G		1.16	1.06	0.24			Indel
RA		9p13.3	*CCL21*	rs10972201	RA	G, A		1.12	0.97	0.34			Noncoding
RA				rs2812378	RA	A, G		1.12	0.98	0.36			Noncoding
RA				rs11574914	RA	G, A	Ref. [8]	1.12	0.97	0.25			Noncoding
T1D	T1D	10p15.1	*IL2RA*	rs61839660	T1D	G, A		0.93	0.64		0.85		Noncoding
RA				rs706778	T1D	C, T	Ref. [8]	1.09	1.22	0.27		0.89	Noncoding
T1D		10q23.31	*RNLS*	rs12416116	T1D	C,A	Ref. [4]	1.01	0.85		0.46		Noncoding
T1D		11p15.5	*INS-IGF2*	rs3842753	T1D	T, G		0.98	2.22		0.73	0.77	Noncoding
T1D				rs689	T1D	A, T	Ref. [4]	0.98	2.21		0.27	0.23	Noncoding

PP > 0.2	Potentially causal	Locus	Marker gene(s)	Variant	Disease association	Alleles	Previously identified as top variant	OR for RA	OR for T1D	PP for RA	PP for T1D	PP combined	Variant type
T1D	T1D	12q24.12	*SH2B3*	rs3184504	T1D	A, G		1.04	0.76		0.33	0.45	p.Arg262Trp
T1D				rs653178	T1D	G, A	Ref. [4]	1.04	0.76		0.66	0.54	Noncoding
T1D		13q32.3	-	rs9517712	T1D	T, C		1.03	0.87		0.34		Noncoding
T1D				rs9585056	T1D	A, G	Ref. [4]	0.96	1.16		0.65		Noncoding
T1D		14q32.2	*MEG3*	rs56994090	T1D	T, C	Ref. [4]	1.01	0.88		0.44		Noncoding
T1D	T1D			rs34552516	T1D	TC, T		1.00	0.88		0.42		Indel
T1D		15q25.1	*CTSH*	rs34843303	RA	A, G		0.98	0.78		0.45		Noncoding
RA				rs34593439	RA	G, A	Ref. [4]	0.99	0.77		0.51		Noncoding
RA+T1D	RA+T1D	19p13.2	*TYK2*	rs34536443	RA + T1D	C, G	Refs. [4, 8]	0.71	0.64	0.41	1.00	1.00	p.Pro1104Ala
	RA			rs35018800	RA(secondary)	G, A							p.Ala928Val
RA	T1D			rs12720356	Combined(secondary)	A, C	Ref. [4]						p.Ile684Ser
RA				rs45524632	RA	C, A		0.54	0.69	0.53			Noncoding
T1D		20p13	*SIRPG*	rs6043405	T1D	T, C		1.02	1.15		0.26		Noncoding
T1D				rs6110697	T1D	T, C		1.02	1.15		0.23		Noncoding
T1D	T1D			rs6043409	T1D	A, G	Ref. [4]	1.02	0.87		0.24		p.Val263Ala
T1D		21q22.3	*UBASH3A*	rs80054410	T1D	A, G		1.06	1.16		0.41	0.59	Noncoding
T1D				rs11203203	T1D	G, A		1.05	1.16		0.39		Noncoding
T1D		22q12.3	*C1QTNF6*	rs229533	T1D	A, C	Ref. [4]	0.99	1.15		0.23		Noncoding

We declare variants at these loci as potentially causal when they are coding and have a posterior probability (PP) of >0.2, or when they are noncoding, have a PP of >0.2, and have functional evidence from the EMSA and luciferase assays. Logistic regression odds ratio (OR) values are from a two-sided χ^2 test (combined, n = 39,403; T1D, n = 20,445; rheumatoid arthritis (RA), n = 27,345).