HLA-DQβ57, anti-insulin T cells and insulin mimicry in autoimmune diabetes

Arcadio Rubio García[1,10], Athina Paterou[1], Rebecca D. Powell Doherty[2], Laurie G. Landry[4], Mercede Lee[1], Amanda M. Anderson[4], Hubert Slawinski[3], Ricardo C. Ferreira[1], Dominik Trzupek[1], Agnieszka Szypowska[7], Linda S. Wicker[1], Luc Teyton[8], Nicola Ternette[2], Maki Nakayama[4-6], John A. Todd[1,10], and Marcin L. Pekalski[1,9,10]

Abstract

Type 1 diabetes (T1D) is caused by a T-cell-mediated destruction of insulin-secreting pancreatic islet β cells. The T1D-predisposing human leukocyte antigen (HLA) class II molecule, DQ8, binds and presents insulin B chain peptides in the thymus producing autoreactive CD4$^+$ T cells[1–12]. Here, we show that this process is driven by negatively-charged T cell receptor (TCR) complementarity-determining region 3β (CDR3β) sequences interacting with alanine at position 57 of the DQ8 β chain. Since T1D aetiology is linked to gut microbiota dysbiosis[13–18], we hypothesized that the commensal proteome contains mimics of the primary insulin B:9-23 epitope that control TCR selection and tolerance. We identified a large set of bacterial proteins with significant similarity to insulin B:9-25, particularly from the transketolase (TKT) superfamily. We isolated a CD4$^+$ TCR with a negatively-charged CDR3β from the pancreas of a DQ8-positive patient that was cross-reactive with one of these TKT peptides and insulin B:9-23. The T1D-protective molecule, DQ6, with the negatively-charged aspartic acid (D) at DQβ57$^{(12,19)}$, showed strong TKT mimotope binding, supporting a role for TKT-specific regulatory T cells in resistance to T1D. We propose that in a DQ8$^+$DQ6$^-$ child with a proinflammatory dysbiotic gut microbiota, cross-reactive TKT-insulin B chain peptide T effector cells escape from the thymus and initiate T1D. TKT is a strong candidate because it is highly upregulated during weaning, a key period in T1D aetiology, and hence a prominent target for an autoimmune-prone immune system. Inhibiting gut dysbiosis and improving immune tolerance to TKT and other mimotopes, especially before and during weaning, could be a route to primary prevention of T1D and other common diseases.

---

[1]JDRF/Wellcome Diabetes and Inflammation Laboratory, Wellcome Centre for Human Genetics, Nuffield Department of Medicine, NIHR Biomedical Research Centre, University of Oxford, Oxford, UK. [2]The Jenner Institute, Centre for Immuno-Oncology, Old Road Campus Research Building, University of Oxford, Oxford, UK. [3]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, NIHR Biomedical Research Centre, University of Oxford, Oxford, UK. [4]Barbara Davis Center for Childhood Diabetes, University of Colorado School of Medicine, Aurora, Colorado, USA. [5]Department of Pediatrics, University of Colorado School of Medicine, Aurora, Colorado, USA. [6]Department of Immunology and Microbiology, University of Colorado School of Medicine, Aurora, Colorado, USA. [7]Department of Pediatrics, Medical University of Warsaw, Warsaw, Poland. [8]Department of Immunology and Microbiology, Scripps Research Institute, La Jolla, California, USA. [9]Department of Paediatrics, University of Oxford, Oxford, UK. [10]Correspondence: john.todd@well.ox.ac.uk, arcadio.rubiogarcia@bnc.ox.ac.uk, marcin.pekalski@paediatrics.ox.ac.uk.

Main

In type 1 diabetes (T1D) the first autoantibodies to appear are most frequently against insulin, with a peak incidence between 9 and 18 months of age, particularly in children carrying the predisposing DR4-DQ8 haplotype[20]. In contrast, the DR15-DQ6 haplotype provides strong dominant protection from T1D[6,10,11]. Many sequence differences exist between these two haplotypes, with the strongest single effect at position 57 of the β chain of the DQ molecule, with a non-charged amino acid (alanine/A, valine/V, or in the nonobese diabetic [NOD] T1D mouse model, serine/S) associated with increased T1D susceptibility and the negatively-charged aspartic acid (D) conferring T1D protection. DQ8 has a primary role in disease initiation through presentation of peptide epitopes to the autoreactive $CD4^+$ T-cell receptors (TCR) from the N-terminal region of the insulin B chain, including residues 9-23 and 9-25 (B:9-23; B:9-25; Methods)[1–7]. It is already established that TCR repertoire differences from the thymus, mediated in part through variable α and β complementarity-determining region (CDR3) 3 regions and their interactions with peptide epitopes and amino acids in the peptide-binding cleft of HLA class II molecules, are determined by HLA class II genotype[21–27]. However, the effect of DQβ57 on CDR3 antigen-recognition sequences in humans and in T1D is unknown.

Alterations of the gut microbiota preceding the appearance of anti-islet autoantibodies have been hypothesised as a primary causal factor in HLA class II susceptible children developing T1D[13–18,28,29]. This is consistent with associations between HLA class II genotypes and the gut microbiota[14,30], and is supported by evidence from the NOD mouse model of T1D[31]. It is possible that immune tolerance to bacterial mimotopes, defined as sequences in bacterial proteins that are cross-reactive with host proteins, is compromised in an environment of microbiota dysbiosis and reduced gut epithelial integrity and functions. Here, we provide evidence that such a mechanism exists coupled to HLA-DQ-regulated TCR CDR3β antigen recognition. These findings suggest therapeutic opportunities for prevention of the earliest pathogenic events in T1D aetiology long before insulin deficiency and diagnosis.

DQ and T cell receptor CDR3 sequences

We purified $CD4^+$ T cells (n=349,623) from PBMCs from selected donors (n=48, median age=11 years) carrying the lowest (protected), low risk and highest risk HLA-DR-DQ risk diplotypes (susceptible; Methods; Supplementary Tables S1-S4). Cells were stimulated and loaded into a single-cell platform for preparing paired 5' gene expression and TCR libraries (Methods). After sequencing, gene expression data analysis and TCR repertoire assembly, we obtained 288,903 cells with both a valid gene expression profile and at least one productive TCR chain (Methods). $CD4^+$ T conventional cells (Tconv; n=258,387; TCR clonotypes,

255,952), defined as those not belonging to either the CD4$^+$ recent thymic emigrant (RTE; 18,194; 18,127 clonotypes) or CD4$^+$ Treg (12,322; 12,260 clonotypes)[32] clusters (Extended Data Fig. 1), we examined which specific differences of CDR3β frequencies could be explained by HLA-DR-DQ risk (Methods). For single amino acids (Fig. 1(a)) negatively-charged residues (D and glutamic acid/E) were more likely to be present in CDR3β chains from susceptible donors carrying the hydrophobic amino acid, A, at DQβ57. Conversely, in protected donors, with a D at β57 of the DQβ chain, there were much fewer D/E CDR3βs, which were replaced by amino acids, V, leucine/L and isoleucine/I that have high interaction potentials[33,34] with the negatively-charged βD57. The 2-mers, AD, GD, EE and EY, were increased in CDR3βs from susceptible donors (Fig. 1(b)), and similarly for 3-mers such as DTE and EET (Fig. 1(c); 4-mers shown in Extended Data Fig. 2). Moreover, a multilevel regression model predicted an increase of D as the T1D OR became higher (Fig. 1(d)), and the opposite effect in case of V (Fig. 1(e)).

We obtained some evidence for the same effects for CD4$^+$ RTEs and Tregs, although the support for this was limited owing to the much lower number of cells available for analysis compared to Tconvs (Extended Data Fig. 3 (a) RTEs and (b) Tregs).

For the TCRα chain sequences, we also estimated systematic differences at low local false sign rates (LFSR) between DR-DQ protective and susceptible diplotypes (Extended Data Fig. 4-6). These corresponded to biases in the usage of particular Vα and Jα genes, which are carried onto the CDR3α region due to the lower recombination diversity of the TCRα. In particular, top estimated differences in k-mers match fragments of SGTYK and GGSYI, which is a sequence encoded in the TRAJ40 and TRAJ6 genes whose usage increased in individuals carrying susceptible DQ versus protected alleles (Extended Data Fig. 7). A previous study analysed TCR Vα gene usage by HLA genotypes by bulk sequencing of RNA from blood and reported an association between DQβ57 and certain Vα genes[35]. We observed the same Vα genes differed in frequency between susceptible and protective DQ, consistent with the reported DQβ57 effects (Extended Data Fig. 8).

Susceptible donors select anti-insulin TCRs

We then investigated whether these observed repertoire differences correlated with an immune response against the primary epitope in T1D, insulin B:9-23[1,2,7,36]. We purified activated circulating CD4$^+$ T cells (n=19,969), defined as HLA-DR$^+$CD38$^+$, from five children newly diagnosed with T1D carrying susceptible DR-DQ diplotypes (Supplementary Tables S5 and S6). This activated cell subpopulation is enriched in autoreactive CD4$^+$ T cells in patients with coeliac disease and other autoimmune disorders[37]. We also sourced CD4$^+$ TCR clonotype sequences (n=1,428) isolated from the islets of five T1D patients from the Network of

Pancreatic Organ Donors (nPOD)[38] who carried susceptible DQ (Supplementary Tables S7 and S8). We compared these TCR sequences against those from individuals with highest susceptibility, DQ2/8 (n=23) from our original cohort (Supplementary Table S1). Both CDR3β sequences from islets and circulating activated cells had an increased frequency of the D residue when compared to non-activated circulating $CD4^+$ T cells from susceptible donors (Fig. 2(a)). Note that owing to the low number of cells available from islet-infiltrating cells, we cannot rule out a negative fold (90% credible interval (0.974, 1.120)). In case of circulating activated cells, where the number of cells sequenced is larger, the posterior estimate only includes fold changes above 1 (90% credible interval (1.114, 1.183); Fig. 2(a)).

We also analysed TCR clonotype sequences (n=159) from nPOD donors (n=5; Supplementary Tables S7 and S8) for whom reactivity against preproinsulin (PPI) peptides had been tested[2]. A binomial regression was consistent with insulin-reactive TCRs being increased in repertoires proportionally to the presence of negatively-charged CDR3βs (posterior probability of a positive effect, p=0.81). Without considering noise in D frequencies, the posterior distribution predicted strong positive effects (Fig. 2(b); p=0.97).

Insulin mimicry and transketolase

In order to assess the existence of insulin mimotopes in the gut microbiome, we assumed bacterial proteins of interest would exhibit a significant degree of similarity to the primary epitope B:9-25 (Methods). We measured similarity between B:9-25 and any given protein as the maximum pairwise local alignment score between both sequences. We chose B:9-25 instead of B:9-23 because peptides have to have sufficient length for productive searches and the importance of the two C-terminal phenylalanine/F residues is established for T cell recognition[9].

Taking block maxima of gapless pairwise alignment scores yields an extreme value distribution. This distribution emphasizes tail events and reflects the expected biology of mimicry, where only an extreme degree of similarity should be relevant. We estimated a null distribution of scores by drawing random permutations from B:9-25 and calculating pairwise local alignment scores against every reviewed entry in the human proteome. We then calculated a parametric approximation to this empirical distribution (Methods). Finally, we aligned B:9-25 to every protein in a gut metagenome reference catalogue which comprised millions of common bacterial and archaeal proteins (MGnify[39,40]). This allowed us to derive precise p-values for exceedingly rare events, and also false discovery rates (FDRs) to account for the large set of multiple comparisons performed.

We found 134 microbial proteins to be significantly similar to B:9-25 at FDR < 0.2 (Fig. 3). The largest association signal was from proteins with a TKT domain (Fig. 3; Extended Data Table

1), although there are significant associations located in other protein families. Both the number and the variety of significant associations suggest tolerance to insulin is regulated by a large niche of gut commensals.

T cell cross-reactivity between TKT and B:9-23 peptides

An important part of our proof for functional cross-reactive bacterial mimotopes is the demonstration that certain $CD4^+$ T cells have TCRs that recognise both B:9-25 peptides in the canonical B:9-23 register and one or more of our top commensal peptide epitopes. We therefore screened 179 cloned $CD4^+$ TCRs detected in islets from six T1D organ donors carrying the DR4-DQ8 haplotype with ten TKT mimotope peptides (Supplementary Table S9) and insulin B:9-23 (Methods). We identified three islet $CD4^+$ TCRs, including two previously reported TCRs[1,41], that were reactive to insulin B:9-23 (Fig. 4(a)), and one new TCR, GSE.166H9 (Vα TRAV3, Jα TRAJ21, CDR3α CAVMYNFNKFYF, Vβ TRBV12-3, Jβ TRBJ2-5, CDR3β CASSLGGRETQYF), from a DQ2/DQ8-heterozygous newly-diagnosed patient (nPOD 6533, age < 5 years old, Supplementary Tables S7 and S8) reacted strongly to one of the TKT peptides (peptide 8, Supplementary Table S9), GHSVEALYCILADRG, from *Clostridium leptum* (Fig. 4(b); Extended Data Table 1). As expected, the GSE.166H9 TCR recognised the TKT peptide 8 and B:9-23 presented by HLA-DQ molecules (Fig. 4(c)). Further analysis determined that the TCR responds to both peptides presented by DQ8-trans, consisting of DQ2α and DQ8β, most strongly, while the TCR can also recognise the peptides presented by DQ8 (Fig. 4(d)). Intensities of responses to the insulin B:9-23 and TKT peptide 8 presented by DQ8-trans were similar (Fig. 4(e)), whereas GSE.166H9 T cells responded to insulin B:9-23 more strongly than TKT peptide 8 when presented by DQ8 (Fig. 4(f)). The most potent response was induced by both insulin B:9-23 and TKT peptide 8 presented by DQ2α-DQ8β-trans, indicating that immune responses by the 166.H9 TCR were initiated by these peptide-MHC complexes.

Transketolase peptide binding to DQ6

If TKT sequences are functionally mimicking insulin B-chain peptides *in vivo* then they should have similar DQ binding properties as insulin B-chain peptides, which bind strongly to DQ6 and weakly to DQ8[5,19]. To investigate this we pulsed DQ6- and DQ8-positive B cell lines with whole, recombinant *Blautia caecimuris* TKT, immunoprecipitated the DQ and DR molecules, eluted the bound peptides and sequenced them by mass spectrometry (Methods). Two TKT peptides bound to DQ6 corresponded to the insulin mimotope sequences, residues, 61-73, FVMSKGHSVEALY, and 66-78, GHSVEALYAVLAE (Extended Data Fig. 9; Supplementary Table S10).

Discussion

In this study we report associations of HLA-DQ with the amino acid charge and interaction potential of CDR3β TCR sequences in anti-insulin autoreactivity and molecular mimicry between insulin and the commensal enzyme, TKT. Protective DQ molecules, with a D residue at DQβ57 (DQ6 and DQ7/DQB*0301) are associated with reduced frequencies of CDR3β D/E residues and, conversely, with increased frequencies of D/E when the A residue is present at DQβ57 (DQ8 and DQ2). These results are consistent with previous data from coeliac disease[42] and from the T1D model, the NOD mouse[12,27,43,44], in which mutation of I-Ag7 β chain (orthologue of DQβ) at position 57 from S to D resulted in T1D protection and reduced D and E amino acids in the CDR3 sequences of B:12-20-specific $CD4^+$ T cells[27]. Furthermore, known insulin-reactive clones isolated from human islets contain negative residues within their CDR3β[2]. The same directions of CDR3β effects were observed in Tregs and RTEs (Extended Data Fig. 3), suggesting that early DQ-mediated positive and negative selection events in the thymus dictate the frequencies of anti-insulin Teffs and Tregs that leave the thymus.

We also conclude that functional TCR repertoire differences associated with T1D DQ diplotypes can influence recognition of mimics of the insulin B:9-23 epitope expressed in the microbiome, most notably, but not only (Fig. 3), from the metabolic enzyme superfamily TKT[45]. TKT enzymes are among the most upregulated bacterial genes during infant weaning to help metabolise the incoming fibre with the introduction of solid food, prior to the appearance of insulin autoantibodies[45,46]. TKT amino acids occupying highly conserved residues and their spacing seem very well suited for evolution to develop insulin mimotopes employing TKT as a template. This is facilitated by the lack of conservation in the remaining residues, which can be mutated without disrupting enzyme function (Extended Data Fig. 10). Given that insulin mimotopes encoded in TKT are present across two bacterial phylums, Firmicutes and Actinobacteria, horizontal or lateral gene transfer events are probable. Phylogenetic tree reconstruction supports the existence of these events[47], as well as some observations derived from insect models[48].

Some of these microbial commensals, which harbor sequences more similar to B:9-25 than what would be expected by chance, have been previously associated with T1D in a variety of longitudinal and cross-sectional studies. *Clostridium leptum,* encoded a TKT mimotope peptide that stimulated a DQ8-restricted TCR from an islet-resident T cell (Fig. 4), was diminished in NOD mice that progressed to T1D[49]. A recent faecal transplant trial in adults with active T1D halted the disease in some patients[50]. Notably, *C. leptum* abundance changes were among the best predictors of response to the transplant[50]. *C. leptum* was also the most positively associated bacteria with less healthy plant-based food diets, maltose, sucrose,

starch and other carbohydrate intake in a large observational study with deeply phenotyped individuals[51].

Blautia caecimuris expresses a TKT mimotope peptide that we demonstrated to be bound by the protective DQ6 molecule. The Blautia genus is a well-known group of anaerobic bacteria that produce short-chain fatty acids (SCFA). The relative overabundance of this genus was measured at the prediabetic and progressive stages of T1D in the longitudinal DIABIMMUNE cohort[52]. The same association was also found in other cross-sectional studies[53]. *Acetatifactor sp.* containing the most significant match to insulin B:9-25, also within the TKT mimotope domain (Extended Data Table 1), was found to be higher in NOD mice compared to non-autoimmune prone ICR mice[54].

Anaerobutyricum (previously *Eubacterium*) *hallii*, is an example of a species that contains a potential non-TKT mimic. We identified a perfect 9-mer match to one of the preferred insulin B:9-23 DQ registers (EALYLVCGE) within an open-reading frame that encoded a protein with a CH3/CHASE3 domain[55]. *A. hallii* was found to be significantly less abundant in children at T1D onset by a recent case-control study[56].

We propose that in a child with microbiota dysbiosis leak of microbial epitopes from the gut and inflammatory context of DQ8, TKT presentation at weaning could initiate anti-insulin autoimmunity via TCR cross reactivity and negatively-charged CDR3βs. We predict that carriage of DQ8 or DQ6 will alter the composition of the microbiota and the abundance of microbial genes such carrying cross-reactive mimotopes. The T1D susceptibility allele at the insulin gene decreases insulin expression in the thymus[57,58] and consequently increasing numbers of circulating anti-insulin $CD4^+$ T cells, which may further control content of the microbiome[59] in a mimotope-dependent manner. Given cross-reactivity with TKT mimotopes, we might expect in DQ8-positive children that the insulin gene polymorphism might also affect microbiota composition, in addition to insulin's susceptibility allele allowing more B:9-23-TKT cross-reactive $CD4^+$ T cells to escape from the thymus[60]. Gut bacteria can be transported into the thymus by a subset of dendritic cells thereby altering the T cell repertoire[61]. In a child with high HLA class II risk of T1D delivery of bacterial TKT into the thymus and by DQ8 could lead to increased numbers of insulin cross-reactive pathogenic T cells escaping from the thymus mediating the increased T1D risk. In contrast, a child carrying DQ6 will delete those cross-reactive cells in the thymus alongside increased production of insulin-specific Tregs, explaining the strong dominant protection against T1D associated with this allotype and its D residue at DQβ57. This mechanism is supported by our demonstration that TKT mimotope peptides from *B. caecimuris* bind to DQ6 (Extended Data Fig. 9; Supplementary Table S10).

Associations between dysbiosis (altered gut microbiota derived context signals), microbial taxa composition and T1D in industrialised countries are widely reported[13–18,28,29], consistent with the increasing incidence of T1D over the past decades and the proposed extinction of certain metabolically-beneficial gut bacteria[17,18], particularly the efficient metabolisers of human milk oligosaccharides (HMO) such as *Bifidobacterium longum* subsp. *infantis* (*B. infantis*). Recently, it has been shown that proinflammatory cytokines are present in the gut of in exclusively breast-fed babies and that this can be corrected by supplementation with a probiotic strain of *B. infantis*[62]. Efficient HMO metabolism leads to the production of microbial metabolites such as indolelactate and SCFAs[63] that promote anti-inflammatory T-cell function and gut epithelial integrity[62]. Hence, in countries with high incidence of T1D and the lowest HMO-metabolising capacity gut dysbiosis and inflammation could be a causal factor in the development of T1D[18]. In industrialised countries a parallel rise in childhood obesity, which Mendelian randomisation studies show is a causal factor in T1D, which could be mediated by suboptimal dysbiotic metabolism[64]. We also note that vitamin B1 (thiamine) deficiency is a frequent metabolic complication in T1D[65] and thiamine is a co-factor of TKT[66]. Changes in gut microbiome taxa abundances and transcribed microbial pathways of T1D patients have been shown to precede thiamine deficiency[67].

Our results have mechanistic and potential therapeutic implications for a wide range of diseases associated with the same HLA class II genotypes. For example, it is possible that the commensal proteome influences the TCR repertoire of other DR15-DQ6 associated diseases such as multiple sclerosis[68], narcolepsy[69], lupus[70] and Alzheimer disease[71] in which this haplotype is predisposing. In Parkinson's disease the DR4-DQ8 haplotype is protective and the prodromal phase of this neurological disorder has been associated with an altered microbiota[72]. In multiple sclerosis and lupus examples of microbial mimicry have been reported[68,73]. Interactions between a child's genetic, microbiota and gut health metabolism and diet in early life could therefore have profound predisposing or protective effects for a whole range of diseases later in life: understanding these mechanisms at the molecular, cellular and whole organism levels will be part of future primary prevention efforts for several common diseases.

Methods

HLA class II associations with T1D and donor sample selection

The T1D associations with HLA class II DR and DQ alleles, individually, or in *cis* haplotypes or *trans* diplotypes, are complex[10,11,74–76]. Here we used DR-DQ diplotypes and their T1D risks[10] to investigate a possible TCR repertoire mechanism underlying their associations with T1D. A large proportion of susceptibility to, and protection from, T1D has been mapped three specific amino acid positions in the β chains of the DR and DQ molecules, DQβ57 and DRβ13 and DRβ71[(11,75)]. For DQβ57 the negatively-charged aspartic acid (D) is associated with dominant protection from T1D and neutral amino acids such as alanine (A), or serine (S) in the nonobese diabetic (NOD) mouse model of T1D[12,27,77], encoding increased risk of the disease. The HLA-DQB1 allele, DQB1*0302, encodes the most T1D-predisposing allotype, DQ8, with A at β57, in contrast to the D^{57}-positive DQ6 β chain (DQB1*0602), which is the most protective HLA class II allotype encoded by the DRB1*1501-DQB1*0602 haplotype. As can be seen in the crystal structures of DQ8 with the primary T1D autoantigenic epitope from the insulin B chain, B:9-23[(4)], and with B:11-23 and its TCR[5], DQ8 presents the peptide to $CD4^+$ T cells in a trimolecular interaction between it, the peptide, and the TCR CDR3β contacting the acidic N-terminal amino acids of the peptide, and pocket 9 of DQ8, containing A at β57. When the three most T1D susceptible amino acids are all present in DQ and DR molecules, as in the DRB1*0401-DQ8 haplotype, the haplotypic risk is at its highest, with individuals homozygous for this haplotype at over 32-fold increased risk of T1D[10]. Additional susceptibility is encoded in a trans interaction between the two main susceptible haplotypes, DRB1*0401-DQ8 and DRB1*03-DQB1*02 (DR3-DQ2/DR4-DQ8), where there is a strong disease-predisposing synergy, most likely due to the structure and peptide-binding properties of the trans-heterodimer of the DQ α chain from the DR3 haplotype, DQA1*0501, and the DQ8 β chain from the DR4 haplotype (DQ2αDQ8β). Hence, DQ2/DQ8 heterozygous individuals are at the highest risk of T1D, at over 63-fold increased risk of T1D[10]. DQ2 also has A at β57, and known to present proinsulin peptides, including B:9-23, and strongly predisposing to T1D[1,2,36].

Previously, we assembled a cohort of T1D families, as part of the JDRF Centre, Diabetes-Genes, Autoimmunity and Prevention Centre (D-GAP)[78], including a collection of peripheral blood mononuclear cells (PBMCs) from blood samples from children with T1D and their unaffected siblings in order to define the immune system before and during T1D. Using this resource, we investigated possible associations between DR-DQ diplotype and the TCR repertoire by selecting donors with the most susceptible diplotypes (DQ2/DQ8) to compare with those with the protective DQ6 haplotype. In order to assess the role of DQ more specifically we also analysed PBMCs from donors who carried the D^{57}-positive DQ7 allotype (DQB1*0301), which when present with the most predisposing haplotype, DR4-DQ8, reduces

DR4-DQ8 T1D risk by over 6-fold[10]; DQ7 is not as protective against T1D as DQ6 because it often occurs on DR haplotypes where the amino acids at DRβ13 and β71 are both highly predisposing to T1D. Nevertheless, the protection encoded by DQ7 is dominant over the susceptible DR molecules[11,75,76].

Donors (Supplementary Table S1) were obtained from the JDRF D-GAP cohort which comprised T1D cases and unaffected siblings (REC Ref:08/H072025). This cohort provided PBMCs and genomic DNA samples. Genomic DNA was prepared from PBMCs or whole blood using QiaAmp DNA Blood kit (Qiagen), or phenol/chloroform extraction.

Initially, we selected two groups with an equal number of donors in the two extremes of the T1D susceptibility-protection axis, namely the most protected DR15-DQ6 (Supplementary Table S2), the most susceptible, DR3-DQ2/DR4-DQ8 (Supplementary Table S4). This choice was performed using data from Taqman genotyping (Applied Biosystems) of four SNPs (rs2187668, rs660895, rs9271366 and rs7454108) and RELI SSO (DYNAL Biotech) classical HLA typing. HLA class II types were further confirmed with ImmunoArray-24 BeadChip v2.0 (Infinium) or HumanImmuno BeadChip v1.0 (Illumina) and HLA imputation[79], along with further SSP classical HLA class II typing (MC Diagnostics and Oxford Transplant Centre).

Owing to the limited number of DR15-DQ6 homozygotes, we also included DR15-DQ6 heterozygotes with a neutral haplotype - a residue other than the susceptible Ala (A) at DQB1 position 57, either neutral Val (V) or Ser (S).

In each batch loaded on a single-cell Chromium V(D)J cassette (10x Genomics), wherever possible, we matched individuals for age (< 20 years old) and homozygosity for DQβ57. For example, in each batch all susceptible DR3-DQ2/DR4-DQ8 individuals were homozygotes for DQβ57 A, and the protected DQ6 were an equal proportion of homozygotes and heterozygotes of DQβ57 D. We also included two batches of DR4/DQ8 versus DR4/DQ7 or lower risk homozygotes, providing a focus on the specific effect of DQβ57 (Supplementary Table S3).

PBMC processing

PBMC isolation, cryopreservation, and thawing were performed as previously described[32]. PBMC isolation was carried out using Lympholyte (CEDARLANE) and were cryopreserved in heat-inactivated, filtered human AB serum (Sigma-Aldrich) and 10% DMSO (Hybri-MAX, Sigma-Aldrich) at concentration between 2 to 10 × 10^6/ml and were stored in liquid nitrogen. PBMCs were thawed in a 37°C water bath for 2 minutes and then washed by adding 1 ml of AB serum to cells dropwise, followed by adding 10 ml of cold (4°C) X-VIVO (Lonza) containing

10% AB serum per up to 10 × 10^6 cells, in a drop-wise fashion. PBMCs were then washed again with 10 ml of cold (4°C) X-VIVO containing 1% AB serum per 10 × 10^6 cells.

10x single $CD4^+$ TCR sequencing from HLA-DQ selected cohort

$CD4^+$ T cells were purified from thawed PBMCs using negative selection (EasySep Human $CD4^+$ T Cell Enrichment Cocktail, STEMCELL Technologies), and washed with X-VIVO medium (Lonza) containing 5% human AB serum and resuspended at the 100,000 cells per well (96 well plate) in a final volume of 200 μl X-VIVO medium (Lonza) containing 5% human AB serum. Cells were activated with the PMA/Ionomycin (eBioscience) for 2 hours and then harvested, washed, resuspended in PBS, counted and 5,000 cells were transferred to the 10x Genomics platform for single-cell immune profiling with version 1.1 kit, which provides paired gene expression and TCR sequences.

cDNA library preparation and sequencing (10x Genomics)

We washed $CD4^+$ T cells in PBS with 0.04% BSA and re-suspended them at a concentration of 800-1,200 cells/μl, before capturing single cells in droplets using the Chromium platform (10x Genomics). Generation of paired gene expression and TCR libraries was performed using the Chromium Single Cell V(D)J Reagent Kits v1 and v1.1b. Quantification of libraries was carried out using Qubit dsDNA HS Assay Kit (Life Technologies) and D1000 ScreenTape (Agilent). Libraries were sequenced on HiSeq 4000 and NovaSeq 6000 (Illumina) to achieve an average of 20,000 reads per cell for gene expression libraries and 5,000 read pairs per cell for TCR libraries.

cDNA library preparation and sequencing (BD Rhapsody)

Single-cell capture and cDNA library preparation, including TCR libraries, was performed using the BD Rhapsody Express Single-cell analysis system (BD Biosciences) using the VDJ CDR3 protocol to generate the mRNA, TCR, AbSeq and Sample Tag libraries. The targeted mRNA panel used in this assay was based on the pre-designed Human T-cell Expression primer panel (BD Biosciences), combined with a custom designed primer panel (containing and additional 306 primer pairs), as previously described[80].

cDNA was initially amplified for 11 cycles (PCR1) to amplify the mRNA, TCR, AbSeq and Sample Tag products. The resulting PCR1 products were purified by double-sized selection using AMPure XP magnetic beads (Beckman Coulter), to separate the shorter AbSeq (~170 bp) and Sample Tag (~250bp) products from the longer mRNA (350-800bp) and TCR (600-1,000 bp) products. The purified mRNA (10 cycles) and Sample Tag (10 cycles) and TCR (15 cycles) PCR1 products were then further amplified using their respective nested PCR primer panels (PCR2) on separate reactions. The resulting mRNA, Sample Tag and TCR PCR2 products were purified by size selection. The concentration, size and integrity of the PCR

products was assessed using both Qubit (High Sensitivity dsDNA kit; ThermoFisher Scientific) and the Agilent 4200 Tapestation system (High Sensitivity D1000; Agilent). The final products were normalised to 2.5 ng/µl (mRNA), 1 ng/µl (Sample Tag & AbSeq) and 0.5 ng/µl (TCR) and underwent a final round of amplification (six cycles for mRNA, Sample Tag and AbSeq; seven cycles for the TCR libraries) using indexes for Illumina sequencing to prepare the final libraries. Final libraries were quantified using Qubit and Agilent Tapestation and pooled (~25/14/57/4% mRNA/TCR/AbSeq/Sample Tag ratio) to achieve a final concentration of 5 nM. Final pooled libraries were spiked with 15% PhiX control DNA to increase sequence complexity and sequenced (75 × 225 bp paired-end) on a NovaSeq sequencer (Illumina).

Cell preparation and fluorescence-activated cell sorting (FACS)

Fresh whole-blood samples (Supplementary Table S5 and S6; Fig. 2) from newly diagnosed with T1D patients were shipped overnight. From each donor, $CD4^+$ T cells were isolated from 10 ml of blood using RosetteSep (STEMCELL Technologies) according to the manufacturers' instructions. Negatively selected $CD4^+$ T cells were washed with PBS + 2% FBS and incubated with the following fluorochrome conjugated antibodies: CD38-BV421 (Biolegend), HLA-DR-AF700 (Biolegend), CD3-BV510 (BD Biosciences) and CD4-BUV395 (BD Biosciences) in Brilliant Stain Buffer (BD Biosciences). Following incubation for 30 minutes at 4ºC, cells were washed two times and resuspended in PBS + 1% FBS for cell sorting at 4ºC in a BD FACSAria Fusion sorter (BD Biosciences). $CD3^+CD4^+$ HLA-DR^+CD38^+ and control pools of $CD3^+CD4^+$ HLA-DR^- and $CD3^+CD4^+$ HLA-DR^+CD38^- were FACS-purified (between 20,000 and 35,000 cells per donor), washed and processed for BD Rhapsody Express Single-cell analysis (BD Biosciences). FACS-sorted cells were incubated with a different oligo-conjugated sample barcoding antibody (sample multiplexing kit; BD Biosciences) for 20 min on ice. Barcoded cells from each batch of donors (total 3 batches) were then pooled together a 5 ml FACS tube (Falcon) and washed in cold PBS + 2% FBS. Cell pools were then incubated for 5 min at 4ºC with Human Fc block (BD Biosciences) and then immediately incubated with a mastermix of 62 oligo-conjugated AbSeq antibodies (BD Biosciences)[80,81] for 45 minutes on ice. Following AbSeq incubation, cells were washed three times in cold BD Sample Buffer (BD Biosciences) to remove any residual unbound antibody, filtered and resuspended in 620 µl of cold BD Sample Buffer for cell capture. Each of the three patient pools was loaded on a BD Rhapsody cartridge (BD Biosciences), and we aimed to retrieve approximately 20,000 cells from each pool.

Single-cell data processing

We preprocessed all single-cell RNA and TCR sequence libraries separately using Cell Ranger v4.0.0 (10x Genomics) to obtain gene counts and receptor assemblies for each donor.

These were then merged into a single gene expression matrix and a single TCR database, which mapped gene counts or TCR chains to cells and donors.

Subsequently, we called cells from read counts with a minimum of 300 genes expressed. We also removed genes not present in at least 50 cells to keep the expression matrix tractable. Furthermore, we applied batch-dependent cutoffs to remove outliers suspected to be cell doublets or multiplets. We also filtered cells with more than 15% of mitochondrial expression to discard those undergoing apoptosis. After data cleanup, we normalized all expression values to 10,000 reads per cell and applied a logarithmic transformation. Next, we discarded all but the top 5,000 most variable genes and regressed out differences due to sequencing depth and mitochondrial gene expression.

Lastly, we aligned cells from each sample using batch-balanced nearest neighbors[82], reduced the dimensionality[83], called clusters[84], and performed a multivariate differential expression[85] to find population markers. The initial run yielded two low-frequency clusters with non-$CD4^+$ contaminants. We discarded cells mapping to these, reran all data processing steps, found another cluster with contaminants, and iterated through the same process one last time to remove another non-$CD4^+$ cluster. This led to 12 different cell subpopulations with distinct RTE and Treg clusters (Extended Data Fig. 1).

We filtered TCRs called by Cell Ranger to retain consensus assemblies with productive rearrangements only. Finally, we performed an inner join between gene expression and receptor assembly data using cell barcodes to obtain TCR chains paired with gene expression cluster information.

T-cell stimulation assay

TCR sequences were identified from CD4 T cells in the islets or pancreas slices of T1D organ donors having the DR4-DQ8 haplotype distributed from the nPOD program (nPOD 69, 6323, 6342, 6367, 6472, 6533; Supplementary Tables S7 and S8) as described previously[86]. The total of 171 TCRs were expressed in 5KC T-hybridoma cells that are devoid of endogenous TCR expression and have been engineered with a ZsGreen-1 reporter gene preceded by the nuclear factor activated T cells (NFAT) binding sequences[87,88]. TCR-expressing 5KC cells (20,000 cells per well) were cultured with 75 μM or designated concentrations of peptides in the presence of antigen presenting cells, Epstein-Barr virus-transformed autologous B cells (100,000 cells per well) or K562 cells transduced with each designated HLA molecule[86] (50,000 cells per well) in round-bottom 96-well plate for 16-22 hours as described in figure legends. Peptides at >95% purity were purchased from Genemed Synthesis, and peptide sequences are included in Supplementary Table 9. Cultures with and without anti-CD3ε antibody at 5 μg/ml (Clone 125-2C11) were included in each assay as positive and negative

control, respectively. To determine HLA molecules presenting antigen to T cells, anti-HLA-DR (clone L243/G46-6), anti-HLA-DQ (clone REA303), or anti-DP (B7/21) antibodies were added at 12.5 µg/ml.

Peptide elution from DQ molecules

Cell culture and pulsing[89,90]

Recombinant TKT protein was made by Centre for Medicines Discovery (CMD) University of Oxford, UK. Recombinant transketolase from *Blautia caecimuris* (IGC[91] entry MH0370_GL0036213) was produced in *Escherichia coli* (BL21(DE3)-pRARE2) and purified by nickel affinity chromatography and size exclusion chromatography with endotoxin removal. Among all our hits in MGnify, *Blautia caecimuris* was the first entry in IGC, a large cohort assembly with estimated abundances, present in more than 10% of the individuals. IGC[91] entry MH0370_GL0036213 corresponds to MGnify entry MGYG000164756_00723, with the latter lacking a nine amino acid sequence in the N-terminus probably arising due to an alternative transcription start site prediction. EBV cells were grown to a total cell count of 1×108 in complete RPMI (R10). Spent R10 was removed, and cells were incubated with 500 µg TKT or insulin in 5 ml of R10 for 2 hours at 37°C. R10 was added to full plate volume (15 ml) and cells further incubated at 37°C for 18 hours. Cells were collected, pelleted (400 × g, 5 min) and washed with PBS. Washed cells were pelleted again and frozen at -20°C for further use.

Immunoprecipitation and HPLC

Harvested pellets were washed in PBS then lysed by mixing for 30 minutes with 2 ml of lysis buffer (0.5% (v/v) IGEPAL 630, 50 mM Tris pH 8.0, 150 mM NaCl) and one tablet Complete Protease Inhibitor Cocktail EDTA-free (Roche) per 10 ml buffer at RT. Lysate was clarified by centrifugation at 1,000 × g for 10 minutes followed by a 20,000 × g spin step for 45 mins at 4°C. Two mg of anti-HLA-DQ SPVL3 antibody-PAS was incubated with lysate overnight with gentle rotation at 4°C. Resin was collected by gravity flow and flow-through lysate was collected for sequential incubation with IVA12-PAS. Antibody-resin-HLA complexes were sequentially washed (15 ml of 0.005% IGEPAL, 50 mM Tris pH 8.0, 150 mM NaCl, 5 mM EDTA, 15 ml of 50 mM Tris pH 8.0, 150 mM NaCl, 15 ml of 50 mM Tris pH 8.0, 450 mM NaCl, and 15 ml of 50 mM Tris pH 8.0), and 5 ml of 10% acetic acid was use to elute bound HLA-DQ complexes from the PAS-antibody resin. Elutions were vacuum centrifuged for drying, dissolved in loading buffer (0.1% v/v trifluoroacetic acid (TFA), 1% v/v acetonitrile in water), and injected by an Ultimate 3000 HPLC system (ThermoFisher Scientific) and separated across a 4.6 mm × 50 mm ProSwift RP-1S column (ThermoFisher Scientific). Peptides were eluted using a 1 ml/min gradient over 5 min from 1-35% Acetonitrile in 0.1% TFA and fractions were collected every 30 s for 18 fractions. Peptide fractions 1-12 were combined into odd and even fractions, then dried.

Mass Spectrometry

HPLC fractions were dissolved in loading buffer and analysed by an Ultimate 3000 HPLC system coupled to a high field Q-Exactive (HFX) Orbitrap mass spectrometer (ThermoFisher Scientific). Peptides were initially trapped in loading buffer, before RP separation with a 60 min linear acetonitrile in water gradient of 2-35% across a 75 µm × 50 cm PepMap RSLC C18 EasySpray column (ThermoFisher Scientific) at a flow rate of 250 nl/min. An EasySpray source was used to ionise peptides at 2000 V, and peptide ions were introduced to the MS at an on-transfer tube temperature of 305°C. Ions were analysed by data-dependent acquisition. Initially a full-MS1 scan (120,000 resolution, 60 ms accumulation time, AGC 3×10^6) was followed by 20 data-dependent MS2 scans (60,000 resolution, 120 ms accumulation time, AGC 5×10^5), with an isolation width of 1.6 m/z and normalized HCD energy of 25%. Charge states of 2–4 were selected for fragmentation.

LC-MS data analysis

Raw data files were analysed with PEAKS X (Bioinformatic Solutions) using a protein sequence database containing 20,606 reviewed human UniProt entries, supplemented with the sequence for *Blautia caecimuris* protein. No enzyme specificity was set, peptide mass error tolerances were set at 5 ppm for precursors and 0.03 Da for MS2 fragments. Additionally, post translational modifications were identified utilizing the PEAKS PTM inbuilt *de novo*-led search for 303 common modifications. FDR was calculated using the decoy database search built into PEAKS.

Estimation of TCR repertoire differences

We excluded an N-terminal prefix and a C-terminal suffix from each CDR3 to avoid the HLA class II binding bias that constrains gene usage[35,92].

Regression of TCR repertoire differences

We used a large hierarchical or multi-level model (Supplementary Algorithm S1) to regress all observed k-mers using the log-transformed T1D OR due to HLA class II diplotypes[10] as an explanatory variable (Fig. 1). This model was instantiated for each combination of CDR3 chain (α or β) and $k \in [1,3]$. Joint estimation of all k-mers for a given chain and k-mer length has the advantage of providing better moderated predictions by partial pooling[93].

We assumed k-mer counts to be binomially distributed. We modelled the relationship between the explanatory variable and each observed k-mer count as a linear function with normally-distributed random effects, composed with a logit link function, also known as a binomial-normal model. We used weakly informative hierarchical hyperpriors. Mean of slopes and intercepts was assumed to be distributed as Normal(μ=0, σ=5). Standard deviations for slopes, intercepts and random effects were set to a Cauchy(μ=0, σ=2.5) truncated at 0. For

additional regularization, we assumed a weak correlation between slopes and intercepts, modelled as a Lewandowski-Kurowicka-Joe(η=2)[94].

We reported fold changes and absolute differences derived from estimated regressions. These were calculated as the difference or the ratio between predicted rates for the maximum and the minimum OR in our cohort, respectively. Here, the minimum OR served as reference, i.e. was the second term in the subtraction or the denominator in the quotient. We reported LFSRs as the probability of the slope having a sign different than the median one.

Estimation of aspartic acid/D proportions

We used a beta-binomial model (Supplementary Algorithm S2) to estimate the proportion of D in each of the three susceptible repertoire types (Fig. 2(a)): peripheral blood, islet infiltrating and activated peripheral blood. The peripheral blood group consisted of all DQ2/8 donors (n=23, Supplementary Table S1). The islet infiltrating group consisted of available donors with evidence of active disease, i.e. nPOD 6323, 6342, 6414, 6533 and 6536 (Supplementary Tables S7 and S8). The activated peripheral blood was formed by all recently diagnosed patients we had recruited (Supplementary Tables S5 and S6). Our prior belief, derived from our previous inference on whole repertoires, was modelled with an informative prior Beta(α=50, β=1000). Observations were modelled as binomial variables k ~ Binomial(n, p), where k was the number of aspartic acid residues in a given donor out of a total n CDR3β amino acids counted, after postprocessing TCRs as indicated previously. We reported the mean and standard deviation for the posterior distribution of p.

Regression of PPI reactivity

We used a linear model (Supplementary Algorithm S3) with a logit link function to regress the dependence between aspartic acid frequency and PPI reactivity (Fig. 2(b)). To this end, we used all donors with available PPI information (Supplementary Table S8) except nPOD 69, whose repertoire had just six productive CDR3β chains and was therefore excluded from further consideration. The slope and intercept were assigned weakly informative priors Normal(μ=0, σ=100) and Normal(μ=0, σ=1000), respectively. Observations were modelled as binomial variables k ~ Binomial(n, p), where k was the number of PPI reactive clonotypes in a given donor out of a total n infiltrating clonotypes measured. The latent probability p was generated by a linear model with D residue counts as an explanatory variable. We reported the posterior probability of the regression slope being greater than zero.

Markov-chain Monte Carlo inference

Since all density functions were differentiable, posterior distributions were estimated by running four independent traces with a No-U-Turn Hamiltonian Monte Carlo sampler for 1,000 iterations, with half of them used as warmup and standard control parameters[95,96]. We verified

the validity of inference by visual inspection of traces to ensure adequate mixing, checked diagnostic metrics ($\hat{r}$, ESS and MCSE) and performed posterior predictive checks.

Proteome-wide similarity with a given epitope

We assumed potential mimotopes to be a subset of those proteins that are more similar to a given epitope of interest than what would be expected by random chance. We measured similarity as the maximum local pairwise alignment score between two protein sequences. We used a linear or affine alignment cost model defined by infinite gap penalties and a BLOSUM 80 substitution matrix[97].

We estimated the null distribution of uninteresting or non-sufficiently similar scores by drawing 10^5 random permutations of the epitope of interest and aligning these with a Smith-Waterman algorithm against all 20,375 reviewed canonical isoforms from *Homo sapiens* stored in the UniProt database. This yielded an empirical null distribution with more than 2×10^9 scores.

We obtained a parametric approximation to the empirical null by fitting a generalized extreme value (GEV) distribution using a maximum likelihood approach (Supplementary Equation S1) and employing an interior point search filter algorithm[98] to estimate the three parameters in GEV(μ, σ, ξ). With this approximation, we derived a p-value for each alignment score of the actual query epitope against a large gut microbiome protein assembly (MGnify[39]), with protein similarity reduced by clustering at 95% of similarity and 95% of mutual coverage and selecting one protein representative per cluster (Fig. 3). Subsequently, we estimated the proportion of discoveries and FDRs using an empirical Bayes procedure[99,100].

To speed up convergence and to obtain a more conservative approximation, we added an additional constraint requiring the score of a perfect alignment to the query epitope to be included in the support of the GEV distribution. In case of insulin B:9-25, this value is 145. The resulting GEV shape parameter estimate was ξ=-0.042015 with a 95% confidence interval of (-0.042012, -0.042017), estimated using a set of 1,000 bootstrap samples. Without this constraint, the estimate ξ=-0.070.

A negative ξ parameter is important because it indicates the GEV distribution falls within the reverse Weibull family, also known as type III domain of attraction. This family has finite support, i.e. scores beyond certain value have probability zero. This agrees with the biology of epitopes and local alignments. Scores larger than a perfect match should be impossible and scores of identical matches should have non-zero probability. However, the widely used Basic Local Alignment Search Tool (BLAST[101]) and its implementation variants use a null approximated by a GEV from the Gumbel family or type II[102,103], which has a shape parameter ξ=0 and therefore support for infinite scores.

We have previously identified many of the same insulin mimotopes using a different generative model-based method[28]. Other approaches to the same problem have relied on human-curated results from BLAST searches performed on fully assembled genomes from cultured species[29,104]. Consequently, prioritized peptides lack statistical significance or come from organisms that do not have humans as a host and therefore do not have any clinical interest. For instance, the antigen presented in by Huang et al[29], RILVELLYLVCSEYL, is far from significance according to our local alignments model (FDR=1) when using B:9-23 (score=68) or B:9-25 (score=68) as query epitope. In comparison, many of the bacterial sequences we have prioritized surpass insulin-like growth factor II (IGF2; score=81) and one surpasses insulin-like growth factor I (IGF1; score=90) in their similarity to insulin B:9-25. If the number of identities is used as a statistic, RILVELLYLVCSEYL again yields a non-significant result when using either B:9-23 (score=9) or B:9-25 (score=9) as a query epitope.

Author Contributions

ARG, JAT and MLP designed and co-led the study.

ARG conceived and developed TCR repertoire and epitope mimicry statistical models.

ARG, MLP and JAT drafted the manuscript.

MN contributed data, comments on manuscript drafts, and designed and produced the T cell cross-reactivity results with the assistance of LGL and AMA.

MLP and AP led experiments with help of ML, HS and LGL.

AP, DT, MLP, ARG, JAT and LSW selected HLA haplotypes.

RF, MLP, AP and ML performed additional single-cell Rhapsody experiments.

AP and LSW supervised QC of D-GAP's DNAs for iChip and HLA genotyping.

DT processed BD Rhapsody single-cell data and genotype data queries.

RPD and NT conducted immunopeptidomics experiments and mass spectrometry data analysis.

LSW, JAT, MLP and AS designed the clinical cohort collections.

LT discussed the results.

All authors contributed to the final version.

Acknowledgements

We thank all volunteers participating in this study. We are grateful to the study volunteers and staff associated with D-GAP, including the Wellcome Clinical Research Facility (Addenbrooke's Clinical Research Centre, Cambridge, UK) and DMech hospital sites, including the Institute of Mother and Child (Warsaw, Poland). We are grateful to nPOD donors and their families. We thank Aaron Michels (Barbara Davis Center for Childhood Diabetes, University of Colorado School of Medicine, Aurora, Colorado, USA) for helpful discussion and for providing EBV cell lines expressing various T1D variants of HLA class II molecules. We are grateful to Wojciech Szypowski (Polish Society for Autoimmune Diseases, Warsaw, Poland) for sample logistics. Diabetes, Genes, Autoimmunity and Prevention (DGAP): London Hampstead Research Committee of the NHS Health Research Authority gave ethical approval for this work; ethics reference number 08/H0720/25. Samples were transferred to ethics 08/H0308/153 Investigating Genes and Phenotypes of Type 1 Diabetes (Cambridgeshire 2 Research Ethics Committee) upon closure of the DGAP study. DMech: Investigating underlying causal mechanisms in type 1 diabetes, South Central Oxford A Research Ethics committee of the NHS Health Research Authority gave ethical approval for this work; ethics reference number 18/SC/0559. The Network for Pancreatic Organ donors with Diabetes (nPOD): studies involving human participants were reviewed and approved by the University of Florida Institutional Research Board (IRB201600029). We thank Nicola Burgess-Brown and Alejandra Fernández Cid (University of Oxford, Oxford, UK) for helpful discussions and production of recombinant microbial proteins. We thank members of the Diabetes and Inflammation Laboratory (University of Oxford): Heather McMurray, Shannah Donhou, Sarune Kacinskaite, Michael Ellis, Sandra Banks, Georgina Burton and past members at the University of Cambridge (Cambridge, UK) led by Helen Stevens for blood sample processing. We thank Claire Scudder, Raqeem Mahmood and Sylwia Kopijasz for managing ethics and blood donor recruitment; Hong Harper for managing the patients' registry and D-GAP data; Florent Yvon for HLA imputation of SNP data from D-GAP patient DNA; Jamie Inshaw for information concerning D-GAP donor metadata; and Olga Platonova for managing finance and funding. We thank Moustafa Attar (Oxford Genomics Centre, University of Oxford), for assistance with single cell sequencing. This research was performed with the support of the Network for Pancreatic Organ donors with Diabetes (nPOD; RRID:SCR_014641), a collaborative T1D research project. The content and views expressed are the responsibility of the authors and do not necessarily reflect the official view of nPOD. Organ Procurement Organizations (OPO) partnering with nPOD to provide research resources are listed at http://www.jdrfnpod.org/for-partners/npod-partners. This work has been supported by a JDRF (4-SRA-2017-473-A-N) and Wellcome (107212/A/15/Z) Strategic Award to JAT and LSW. D-GAP was a centre grant funded by the JDRF (1-2007-1803) to Mark Peakman, Tim Tree, JAT,

LSW, Polly J. Bingley and David B. Dunger. ARG was also supported by the DARPA Probabilistic Programming for Advanced Machine Learning (PPAML) program. The work performed in the University of Colorado has been supported by the National Institutes of Diabetes and Digestive and Kidney Diseases (R01DK099317, R01DK032083, P30DK116073) and JDRF (5-SRA-2018-557-Q-R to nPOD).

We dedicate this work to the memory of Professor David B. Dunger FRCP FMedSci (1948-2021) and of Professor Hugh O. McDevitt MD ForMemRS (1930-2022).

1. Michels, A. W. *et al.* Islet-derived CD4 T cells targeting proinsulin in human autoimmune diabetes. *Diabetes* **66**, 722–734 (2017).

2. Landry, L. G. *et al.* Proinsulin-reactive CD4 T cells in the islets of type 1 diabetes organ donors. *Front. Endocrinol.* **12**, 622647 (2021).

3. Daniel, D., Gill, R. G., Schloot, N. & Wegmann, D. Epitope specificity, cytokine production profile and diabetogenic activity of insulin-specific T cell clones isolated from NOD mice. *Eur. J. Immunol.* **25**, 1056–1062 (1995).

4. Lee, K. H., Wucherpfennig, K. W. & Wiley, D. C. Structure of a human insulin peptide–HLA-DQ8 complex and susceptibility to type 1 diabetes. *Nat. Immunol.* **2**, 501–507 (2001).

5. Wang, Y. *et al.* How C-terminal additions to insulin B-chain fragments create superagonists for T cells in mouse and human type 1 diabetes. *Sci. Immunol.* **4**, eaav7517 (2019).

6. Wen, X. *et al.* Increased islet antigen–specific regulatory and effector $CD4^+$ T cells in healthy individuals with the type 1 diabetes–protective haplotype. *Sci. Immunol.* **5**, eaax8767 (2020).

7. James, E. A., Mallone, R., Kent, S. C. & DiLorenzo, T. P. T-cell epitopes and neo-epitopes in type 1 diabetes: a comprehensive update and reappraisal. *Diabetes* **69**, 1311–1335 (2020).

8. Bettini, M. L. & Bettini, M. Understanding autoimmune diabetes through the prism of the tri-molecular complex. *Front. Endocrinol.* **8**, 351 (2017).

9. Wan, X. *et al.* The MHC-II peptidome of pancreatic islets identifies key features of autoimmune peptides. *Nat. Immunol.* **21**, 455–463 (2020).

10. Sharp, S. A. *et al.* Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care* (2019).

11. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015).

12. Todd, J. A., Bell, J. I. & McDevitt, H. O. HLA-DQ β gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* **329**, 599–604 (1987).

13. Davis-Richardson, A. G. *et al. Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front. Microbiol.* **5**, (2014).

14. Russell, J. T. *et al.* Genetic risk for autoimmunity is associated with distinct changes in the human gut microbiome. *Nat. Commun.* **10**, 3621 (2019).

15. Abdellatif, A. M. & Sarvetnick, N. E. Current understanding of the role of gut dysbiosis in type 1 diabetes. *J. Diabetes* **11**, 632–644 (2019).

16. Siljander, H., Honkanen, J. & Knip, M. Microbiome and type 1 diabetes. *EBioMedicine* **46**, 512–521 (2019).

17. Insel, R. & Knip, M. Prospects for primary prevention of type 1 diabetes by restoring a disappearing microbe. *Pediatr. Diabetes* **19**, 1400–1406 (2018).

18. Vatanen, T. *et al.* Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).

19. Ettinger, R. A. & Kwok, W. W. A peptide binding motif for HLA-DQA1*0102/DQB1*0602, the class II MHC molecule associated with dominant protection in insulin-dependent diabetes mellitus. *J. Immunol. Baltim. Md 1950* **160**, 2365–2373 (1998).

20. Krischer, J. P. *et al.* The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: the TEDDY study. *Diabetologia* **58**, 980–987 (2015).

21. Borg, N. A. *et al.* The CDR3 regions of an immunodominant T cell receptor dictate the 'energetic landscape' of peptide-MHC recognition. *Nat. Immunol.* **6**, 171–180 (2005).

22. Logunova, N. N. *et al.* MHC-II alleles shape the CDR3 repertoires of conventional and regulatory naïve $CD4^+$ T cells. *Proc. Natl. Acad. Sci.* **117**, 13659–13669 (2020).

23. Lu, J. *et al.* Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. *Nat. Commun.* **10**, 1019 (2019).

24. Ishigaki, K. *et al.* HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors. *Nat. Genet.* (2022).

25. Simone, E. *et al.* T cell receptor restriction of diabetogenic autoimmune NOD T cells. *Proc. Natl. Acad. Sci.* **94**, 2518–2521 (1997).

26. Baker, F. J., Lee, M., Chien, Y. & Davis, M. M. Restricted islet-cell reactive T cell repertoire of early pancreatic islet infiltrates in NOD mice. *Proc. Natl. Acad. Sci.* **99**, 9374–9379 (2002).

27. Gioia, L. *et al.* Position β57 of I-Ag7 controls early anti-insulin responses in NOD mice, linking an MHC susceptibility allele to type 1 diabetes onset. *Sci. Immunol.* **4**, (2019).

28. Rubio García, A. *et al.* Peripheral tolerance to insulin is encoded by mimicry in the microbiome. *bioRxiv* 2019.12.18.881433 (2019).

29. Huang, Q. *et al. Parabacteroides distasonis* enhances type 1 diabetes autoimmunity via molecular mimicry. *bioRxiv* 2020.10.22.350801 (2020).

30. Paun, A. *et al.* Association of HLA-dependent islet autoimmunity with systemic antibody responses to intestinal commensal bacteria in children. *Sci. Immunol.* **4**, eaau8125 (2019).

31. Silverman, M. *et al.* Protective major histocompatibility complex allele prevents type 1 diabetes by shaping the intestinal microbiota early in ontogeny. *Proc. Natl. Acad. Sci.* **114**, 9671–9676 (2017).

32. Pekalski, M. L. *et al.* Neonatal and adult recent thymic emigrants produce IL-8 and express complement receptors CR1 and CR2. *JCI Insight* **2**, 93739 (2017).

33. Kosmrlj, A. *et al.* Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* **465**, 350–354 (2010).

34. Miyazawa, S. & Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644 (1996).

35. Sharon, E. *et al.* Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.* **48**, 995–1002 (2016).

36. Ihantola, E.-L. *et al.* Characterization of proinsulin T cell epitopes restricted by type 1 diabetes–associated HLA class II molecules. *J. Immunol.* **204**, 2349–2359 (2020).

37. Christophersen, A. *et al.* Distinct phenotype of CD4+ T cells driving celiac disease identified in multiple autoimmune conditions. *Nat. Med.* **25**, 734–737 (2019).

38. Campbell-Thompson, M. *et al.* Network for pancreatic organ donors with diabetes (nPOD): developing a tissue biobank for type 1 diabetes: tissue biobank for diabetes. *Diabetes Metab. Res. Rev.* **28**, 608–617 (2012).

39. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* gkz1035 (2019).

40. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).

41. Landry, L. G. *et al.* Proinsulin-reactive CD4 T cells in the islets of type 1 diabetes organ donors. *Front. Endocrinol.* **12**, 622647 (2021).

42. Hovhannisyan, Z. *et al.* The role of HLA-DQ8 beta57 polymorphism in the anti-gluten T-cell response in coeliac disease. *Nature* **456**, 534–538 (2008).

43. Singer, S. M. *et al.* Prevention of diabetes in NOD mice by a mutated I-Ab transgene. *Diabetes* **47**, 1570–1577 (1998).

44. Yoshida, K. *et al.* The diabetogenic mouse MHC class II molecule I-Ag7 is endowed with a switch that modulates TCR affinity. *J. Clin. Invest.* **120**, 1578–1590 (2010).

45. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).

46. Pinto, E. *et al.* The intestinal proteome of diabetic and control children is enriched with different microbial and host proteins. *Microbiology,* **163**, 161–174 (2017).

47. Rogers, M. B. *et al.* A complex and punctate distribution of three eukaryotic genes derived by lateral gene transfer. *BMC Evol. Biol.* **7**, 89 (2007).

48. Brown, B. P. & Wernegreen, J. J. Genomic erosion and extensive horizontal gene transfer in gut-associated Acetobacteraceae. *BMC Genomics* **20**, 472 (2019).

49. Sysi-Aho, M. *et al.* Metabolic regulation in progression to autoimmune diabetes. *PLoS Comput. Biol.* **7**, e1002257 (2011).

50. de Groot, P. *et al.* Faecal microbiota transplantation halts progression of human new-onset type 1 diabetes in a randomised controlled trial. *Gut* **70**, 92–105 (2021).

51. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).

52. Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).

53. Qi, C.-J. *et al.* Imbalance of fecal microbiota at newly diagnosed type 1 diabetes in Chinese children. *Chin. Med. J. (Engl.)* **129**, 1298–1304 (2016).

54. Wu, Y., You, Q., Fei, J. & Wu, J. Changes in the gut microbiota: a possible factor influencing peripheral blood immune indexes in non-obese diabetic mice. *Antonie Van Leeuwenhoek* **114**, 1669–1682 (2021).

55. Zhulin, I. B., Nikolskaya, A. N. & Galperin, M. Y. Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in *Bacteria* and *Archaea*. *J. Bacteriol.* **185**, 285–294 (2003).

56. Biassoni, R. *et al.* Gut microbiota in T1DM-onset pediatric patients: machine-learning algorithms to classify microorganisms as disease linked. *J. Clin. Endocrinol. Metab.* **105**, e3114–e3126 (2020).

57. Pugliese, A. *et al.* The insulin gene is transcribed in the human thymus and transcription levels correlate with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nat. Genet.* **15**, 293–297 (1997).

58. Vafiadis, P. *et al.* Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nat. Genet.* **15**, 289–292 (1997).

59. Assfalg, R. *et al.* Oral insulin immunotherapy in children at risk for type 1 diabetes in a randomised controlled trial. *Diabetologia* **64**, 1079–1092 (2021).

60. Durinovic-Belló, I. *et al.* Insulin gene VNTR genotype associates with frequency and phenotype of the autoimmune response to proinsulin. *Genes Immun.* **11**, 188–193 (2010).

61. Zegarra-Ruiz, D. F. *et al.* Thymic development of gut-microbiota-specific T cells. *Nature* **594**, 413–417 (2021).

62. Henrick, B. M. *et al. Bifidobacteria*-mediated immune system imprinting early in life. *Cell* S0092867421006607 (2021).

63. Tsukuda, N. *et al.* Key bacterial taxa and metabolic pathways affecting gut short-chain fatty acid profiles in early life. *ISME J.* **15**, 2574–2590 (2021).

64. Richardson, T. G. *et al.* Childhood body size directly increases type 1 diabetes risk based on a lifecourse Mendelian randomization approach. *Nat. Commun.* **13**, 2337 (2022).

65. Rosner, E. A., Strezlecki, K. D., Clark, J. A. & Lieh-Lai, M. Low thiamine levels in children with type 1 diabetes and diabetic ketoacidosis: a pilot study. *Pediatr. Crit. Care Med.* **16**, 114–118 (2015).

66. Racker, E., Haba, G. D. L. & Leder, I. G. Thiamine pyrophosphate, a coenzyme of transketolase. *J. Am. Chem. Soc.* **75**, 1010–1011 (1953).

67. Kaysen, A. *et al.* Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic hematopoietic stem cell transplantation. *Transl. Res. J. Lab. Clin. Med.* **186**, 79-94.e1 (2017).

68. Wang, J. *et al.* HLA-DR15 Molecules jointly shape an autoreactive T cell repertoire in multiple sclerosis. *Cell* **183**, 1264-1281.e20 (2020).

69. Luo, G. *et al.* Autoimmunity to hypocretin and molecular mimicry to flu in type 1 narcolepsy. *Proc. Natl. Acad. Sci.* **115**, E12323–E12332 (2018).

70. Graham, R. R. *et al.* Specific combinations of HLA-DR2 and DR3 class II haplotypes contribute graded risk for disease susceptibility and autoantibodies in human SLE. *Eur. J. Hum. Genet.* **15**, 823–830 (2007).

71. Hollenbach, J. A. *et al.* A specific amino acid motif of *HLA-DRB1* mediates risk and interacts with smoking history in Parkinson's disease. *Proc. Natl. Acad. Sci.* **116**, 7419–7424 (2019).

72. Heintz-Buschart, A. *et al.* The nasal and gut microbiome in Parkinson's disease and idiopathic rapid eye movement sleep behavior disorder. *Mov. Disord. Off. J. Mov. Disord. Soc.* **33**, 88–98 (2018).

73. Greiling, T. M. *et al.* Commensal orthologs of the human autoantigen Ro60 as triggers of autoimmunity in lupus. *Sci. Transl. Med.* **10**, eaan2306 (2018).

74. Erlich, H. *et al.* HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. *Diabetes* **57**, 1084–1092 (2008).

75. Lenz, T. L. *et al.* Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat. Genet.* **47**, 1085–1090 (2015).

76. Zhao, L. P. *et al.* Motifs of three HLA-DQ amino acid residues (α44, β57, β135) capture full association with the risk of type 1 diabetes in DQ2 and DQ8 children. *Diabetes* (2020).

77. Acha-Orbea, H. & McDevitt, H. O. The first external domain of the nonobese diabetic mouse class II I-A beta chain is unique. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 2435–2439 (1987).

78. Arif, S. *et al.* Blood and islet phenotypes indicate immunological heterogeneity in type 1 diabetes. *Diabetes* **63**, 3835–3845 (2014).

79. Zheng, X. *et al.* HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).

80. Trzupek, D. *et al.* Single-cell multi-omics analysis reveals IFN-driven alterations in T lymphocytes and natural killer cells in systemic lupus erythematosus. *Wellcome Open Res.* **6**, 149 (2021).

81. Trzupek, D. *et al.* Discovery of CD80 and CD86 as recent activation markers on regulatory T cells by protein-RNA single-cell analysis. *Genome Med.* **12**, 55 (2020).

82. Polański, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinforma. Oxf. Engl.* **36**, 964–965 (2020).

83. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).

84. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

85. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16**, 163–166 (2019).

86. Anderson, A. M. *et al.* Human islet T cells are highly reactive to preproinsulin in type 1 diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2107208118 (2021).

87. Mann, S. E. *et al.* Multiplex T Cell stimulation assay utilizing a T cell activation reporter-based detection system. *Front. Immunol.* **11**, 633 (2020).

88. Landry, L. G., Mann, S. E., Anderson, A. M. & Nakayama, M. Multiplex T-cell stimulation assay utilizing a T-cell activation reporter-based detection system. *Bio-Protoc.* **11**, e3883 (2021).

89. Purcell, A. W., Ramarathinam, S. H. & Ternette, N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat. Protoc.* **14**, 1687–1707 (2019).

90. Parker, R. *et al.* Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. *Cell Rep.* **35**, 109179 (2021).

91. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

92. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).

93. Gelman, A., Hill, J. & Yajima, M. Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* **5**, 189–211 (2012).

94. Lewandowski, D., Kurowicka, D. & Joe, H. Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **100**, 1989–2001 (2009).

95. Hoffman MD & Gelman A. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).

96. Carpenter, B. *et al. Stan*: A probabilistic programming language. *J. Stat. Softw.* **76**, (2017).

97. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**, 10915–10919 (1992).

98. Nocedal, J., Wächter, A. & Waltz, R. A. Adaptive barrier update strategies for nonlinear interior methods. *SIAM J. Optim.* **19**, 1674–1693 (2009).

99. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

100. Benjamini, Y. & Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25**, 60–83 (2000).

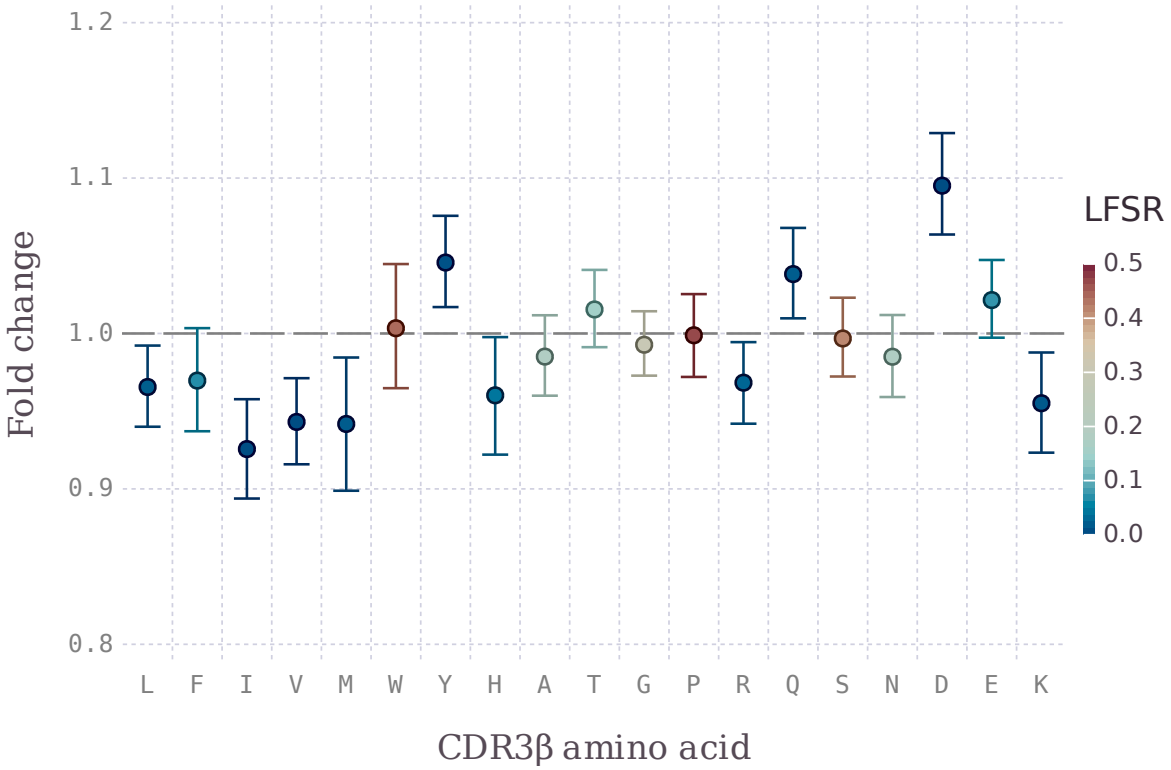
101. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

102. Karlin, S. & Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.* **87**, 2264–2268 (1990).

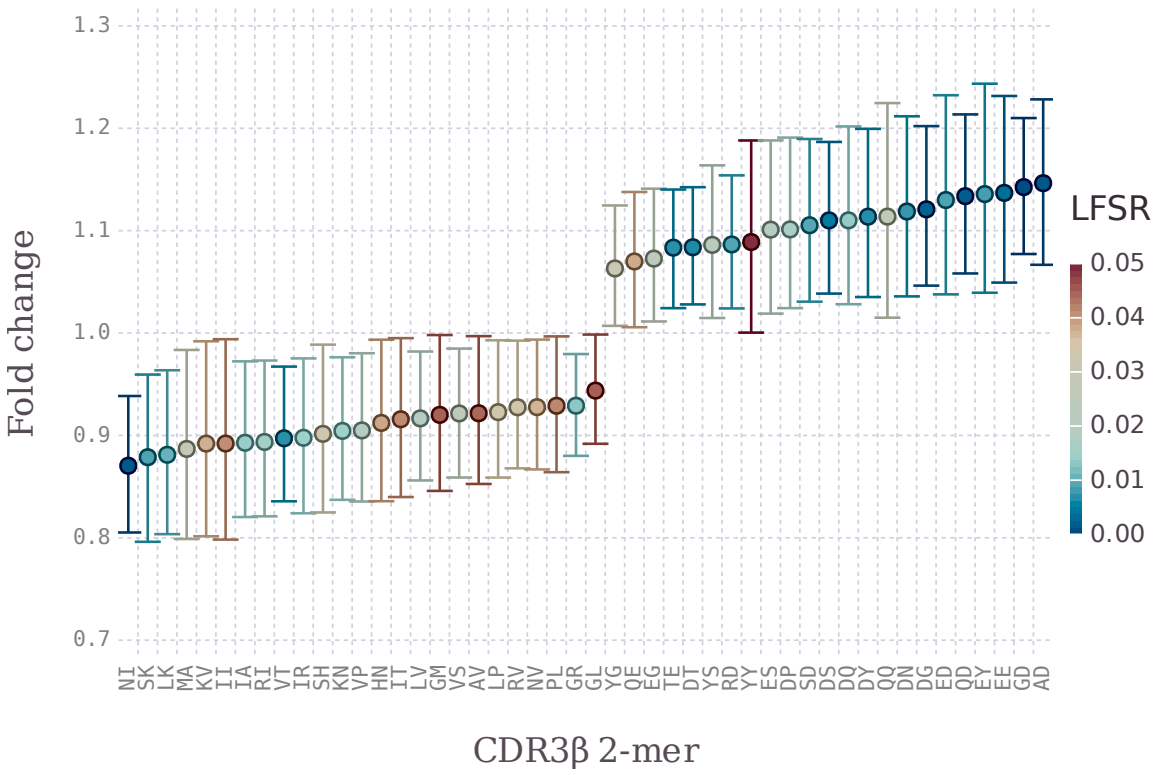
103. Altschul, S. F. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.* **29**, 351–361 (2001).

104. Yang, J. *et al.* Autoreactive T cells specific for insulin B:11-23 recognize a low-affinity peptide register in human subjects with autoimmune diabetes. *Proc. Natl. Acad. Sci.* **111**, 14840–14845 (2014).

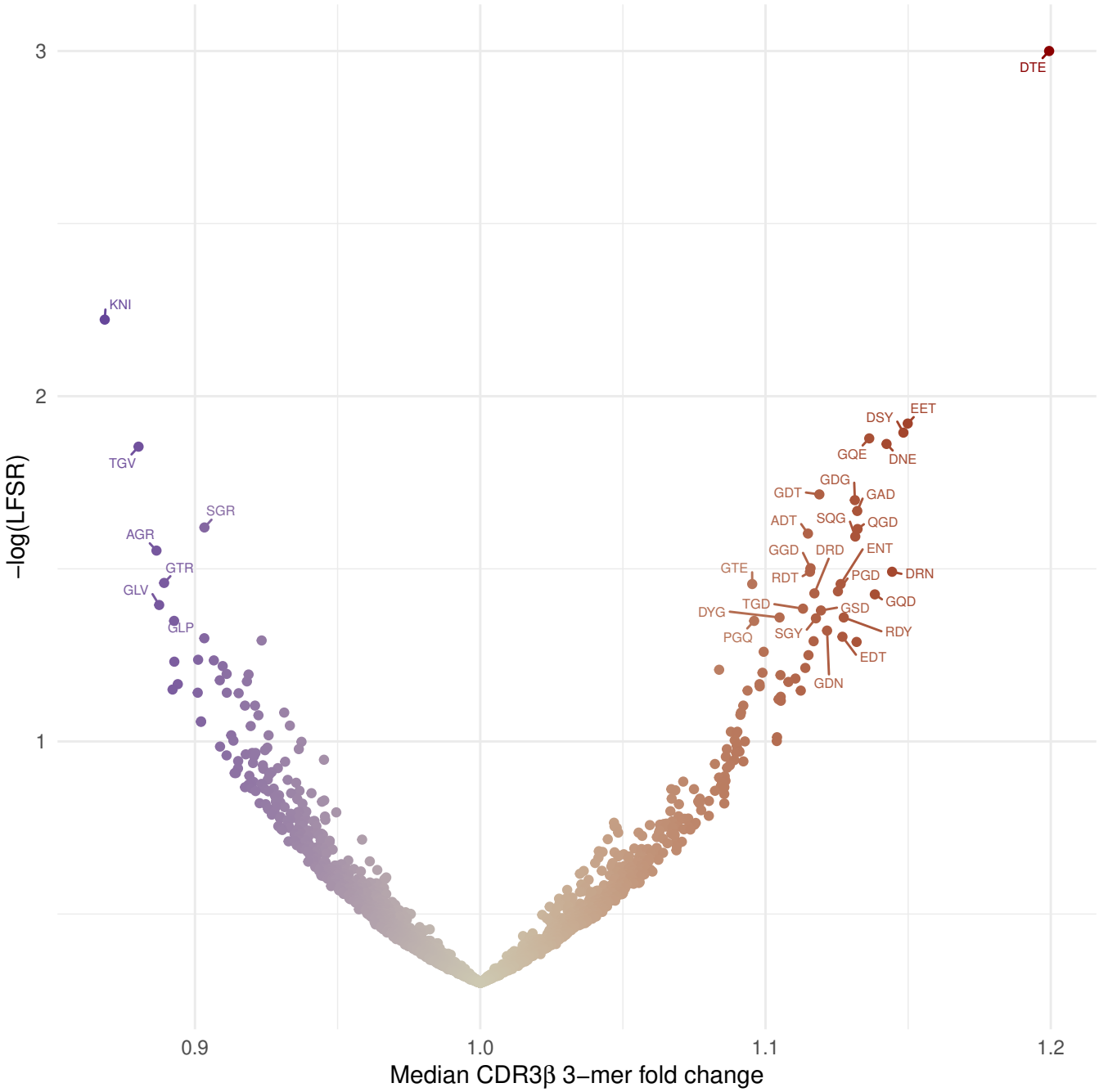
Fig. 1: Estimates of CDR3β k-mer fold changes across HLA class II risk extremes.



(a)

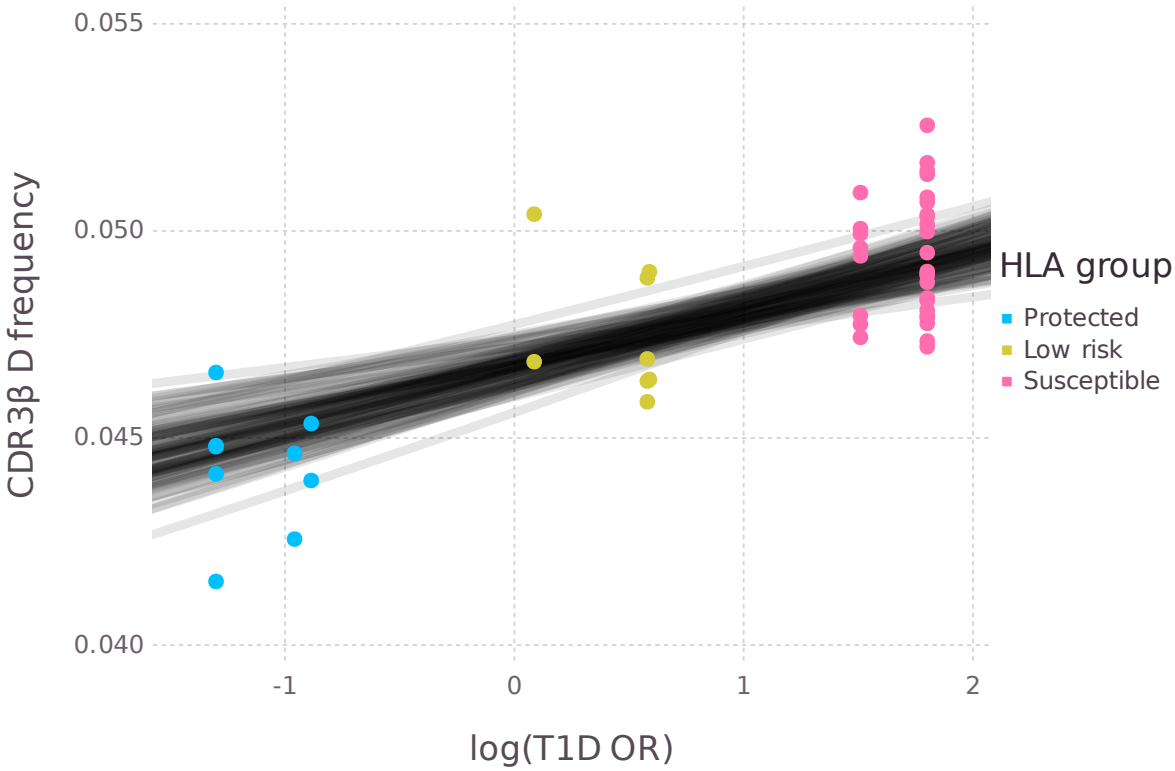


(b)

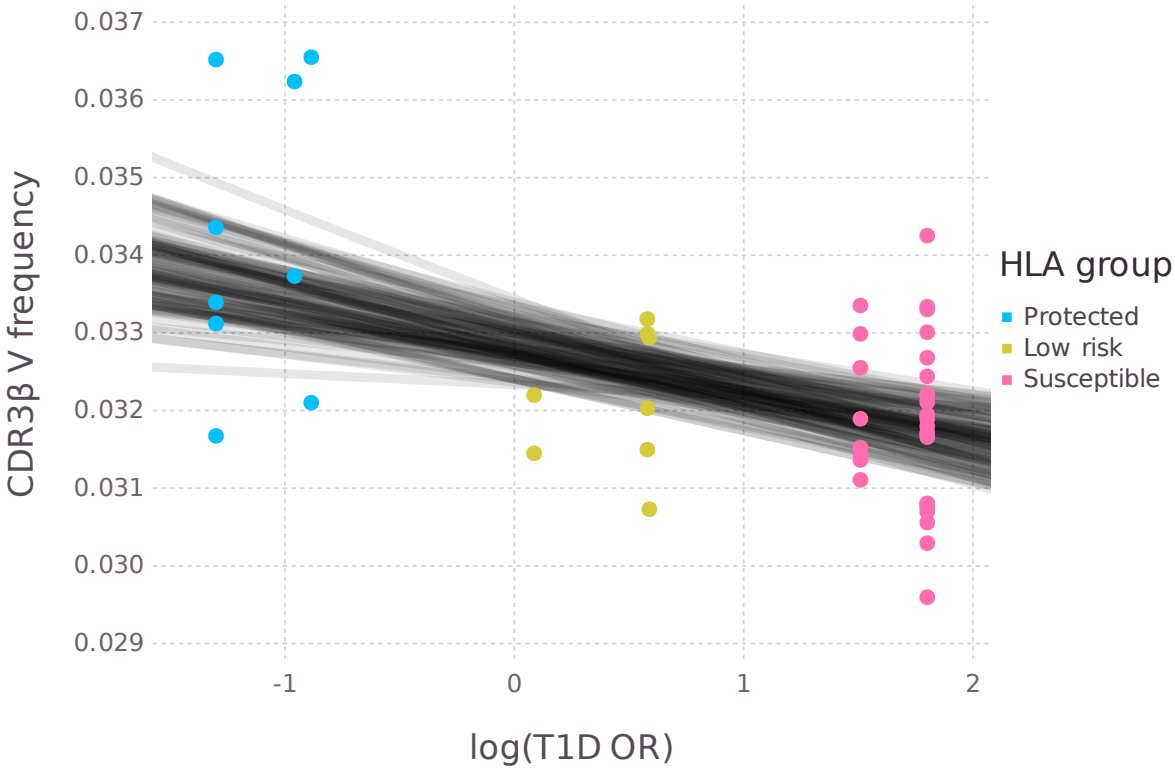


(c)

HLA class II risk extremes are DQ6 (negatively-charged D at DQβ57, used as baseline) versus DQ2/8 (DQβ with non-charged A at position 57 which includes the DQ2α/DQ8β trans HLA class heterodimer). 90% credible intervals and median predicted local false sign rates (LFSR) are shown for (a) k=1 or single amino acid counts, where amino acids are sorted by decreasing average interaction potential,[33] (b) k=2 and (c) k=3 $CD4^+$ T conventional cells (Tconv). Amino acids with low average interaction potential and negatively-charged side chains are enriched in the presence of DQ2/8, whereas amino acids of high interaction potential are enriched in DQ6, as evidenced by multilevel regression for (d) aspartic acid/D and (e) valine/V. Regression lines represent 100 random draws from the posterior distribution. Each point in (d) and (e) denotes a single donor.

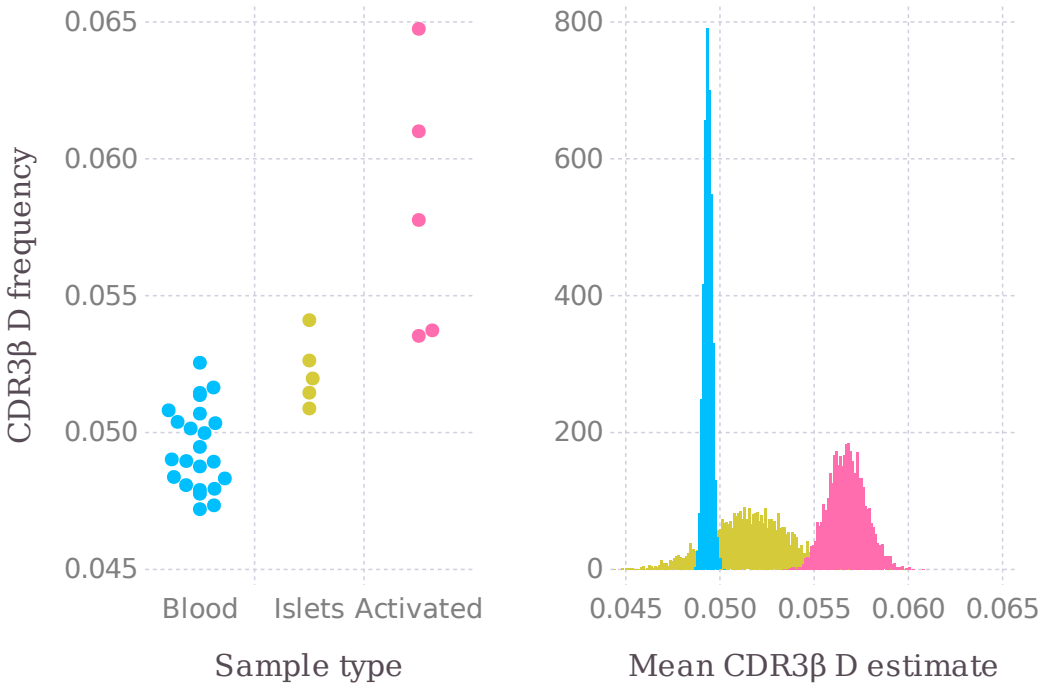

0.055
0.050
0.045
0.040
CDR3β D frequency
-1
0
1
2
log(T1D OR)
HLA group
Protected
Low risk
Susceptible

(d)

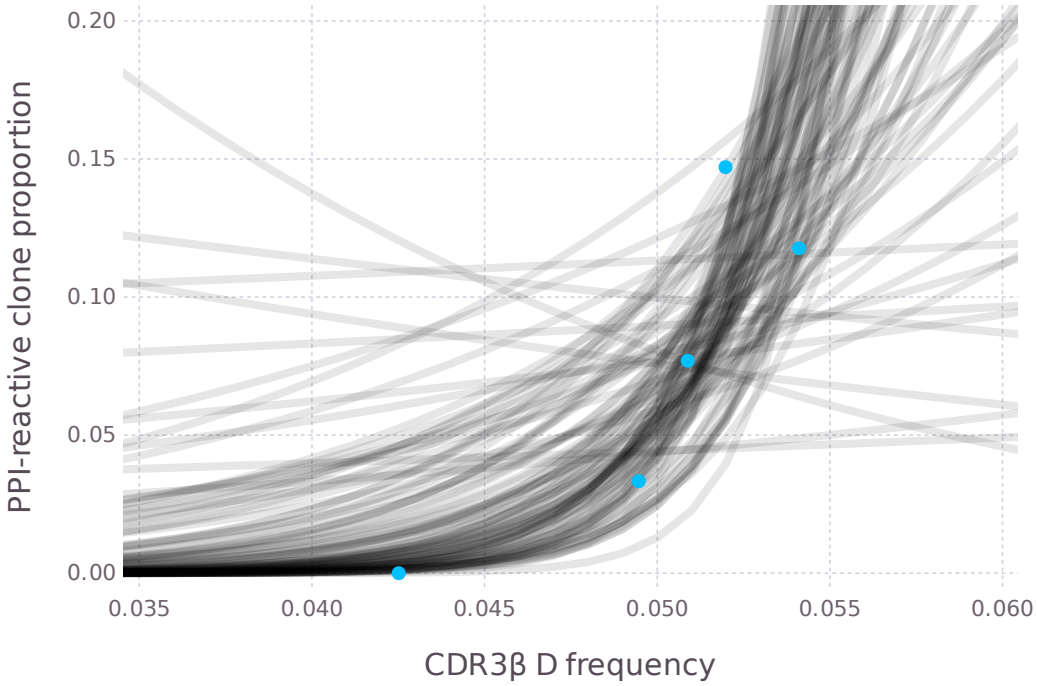

0.037
0.036
0.035
0.034
0.033
0.032
0.031
0.030
0.029
CDR3β V frequency
-1
0
1
2
log(T1D OR)
HLA group
Protected
Low risk
Susceptible

(e)

Fig. 2: CDR3β aspartic acid/D frequency association with CD4⁺ T cell repertoire type on susceptible HLA class II donors.



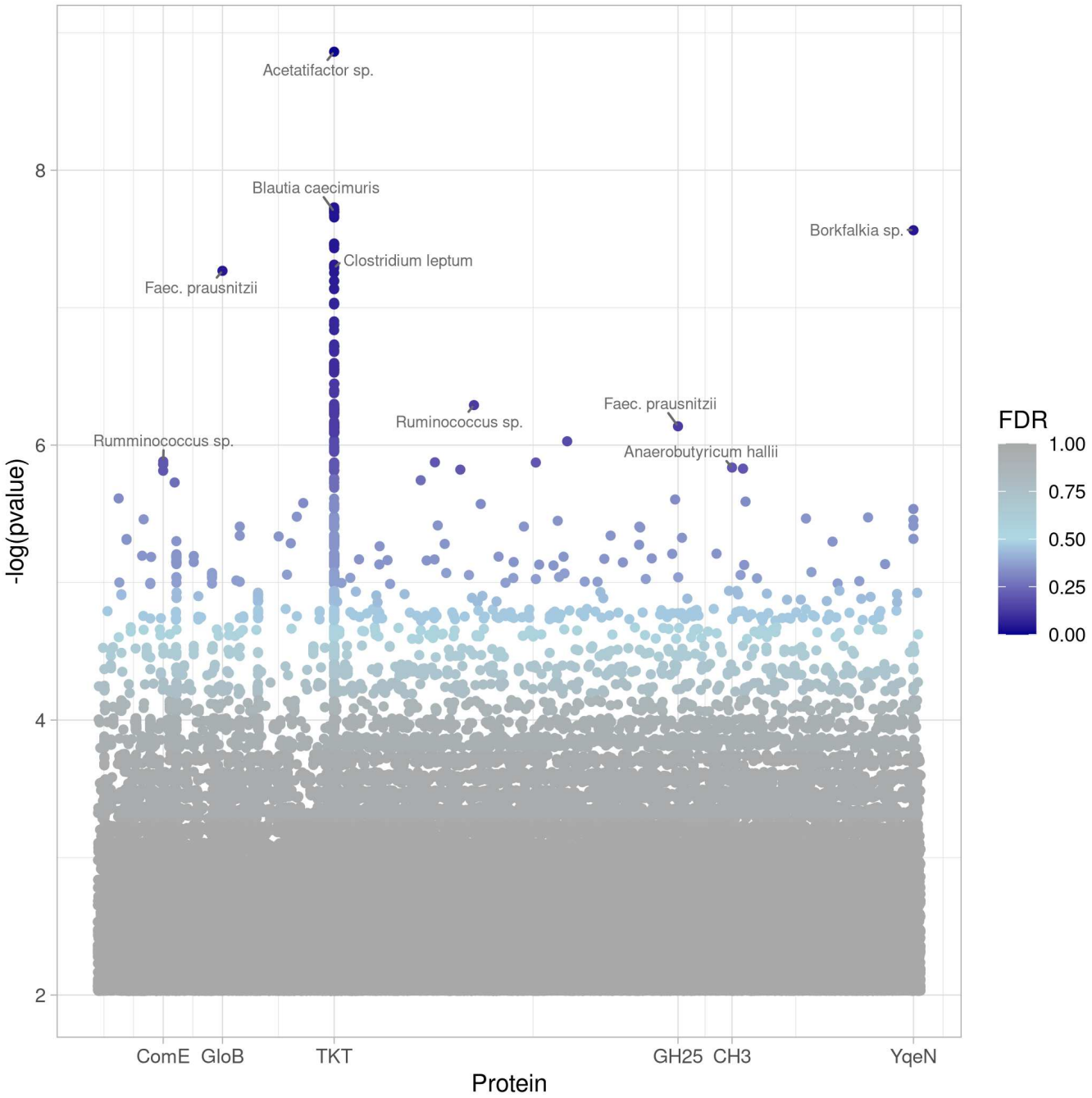
(a)



(b)

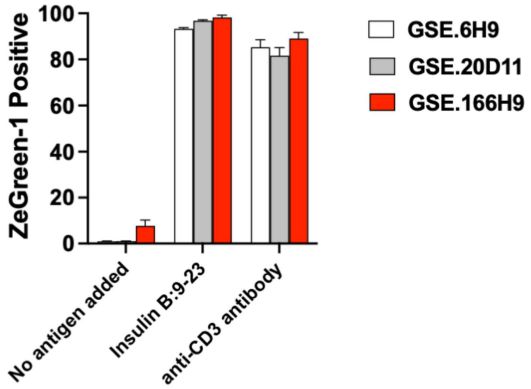
(a) Bayesian posterior CDR3β aspartic acid/D mean estimates in cells from peripheral blood, infiltrating islets and activated circulating cells. (b) Pre-proinsulin-reactive (PPI) clone proportion measured in islets as a function of aspartic acid repertoire frequency.

Fig. 3: Gut microbial proteome-wide similarity with insulin B:9–25.

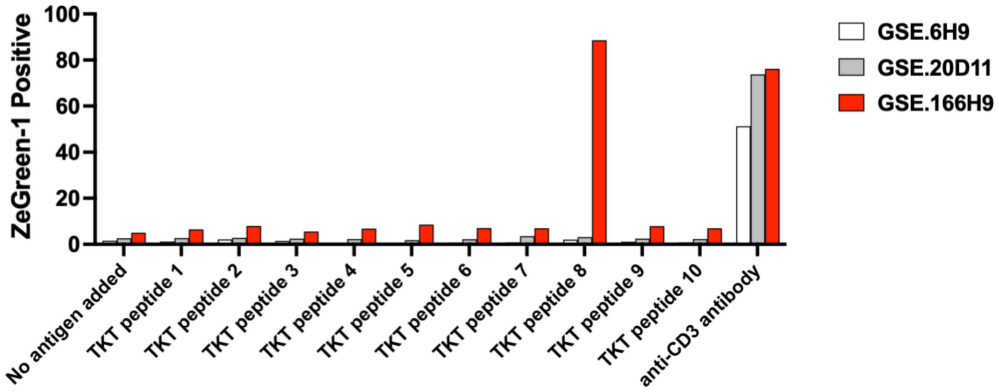


Log-transformed p-values denote the significance of similarity scores for each of the ~1×10^7 predicted proteins in a reference gut metagenome assembly. Data points are labelled with false discovery rate (FDR) and predicted bacterial species. The top association signals are present in proteins with a transketolase (TKT) domain. Other protein superfamilies and domains include late competence operon (ComE), hydroxy-acylglutathione hydrolase (GloB), glycosyl hydrolase family 25 (GH25), cyclases/histidine kinases associated sensory extracellular domain 3 (CH3) and yqeN DNA replication protein (YqeN).

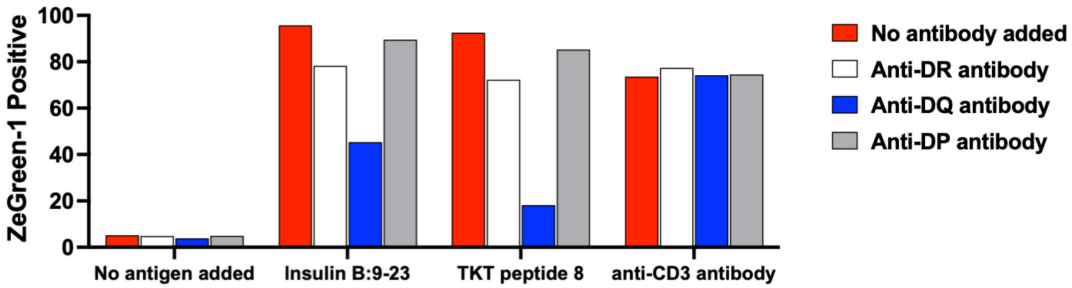
Fig. 4: Cross-reactivity to insulin B:9-23 and TKT peptides.



(a)

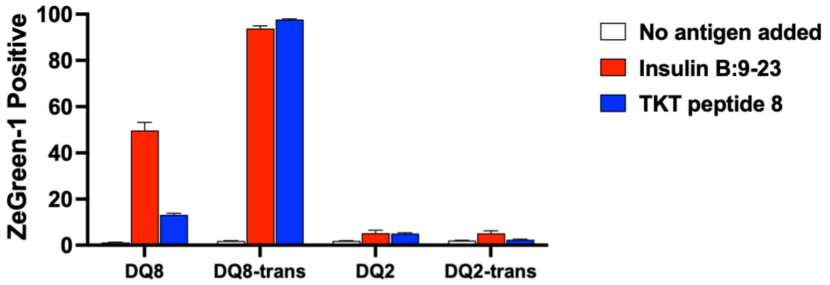


(b)

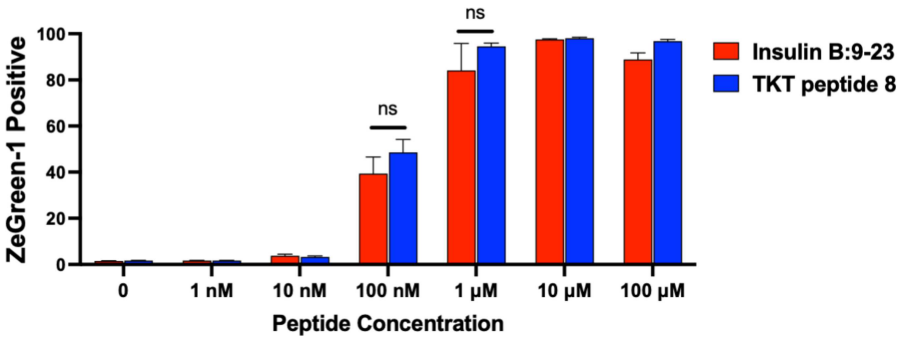


(c)

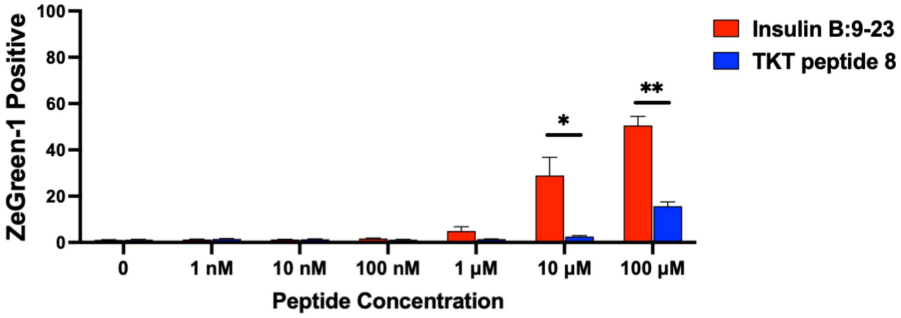
5KC T-hybridoma reporter cells expressing TCRs identified from CD4$^+$ T cells in the islets of T1D organ donors were generated to test for the response to antigen stimulation, and activation of 5KC cells was assessed by evaluating expression of ZsGreen-1, which is driven via NFAT signaling upon stimulation. 5KC cells expressing GSE.6H9, GSE.20D11, or GSE.166H9 were cultured with the insulin B:9–23 peptide (a) and TKT peptides (b) at 100 mM in the presence of Epstein-Barr virus (EBV)-transformed autologous B cells. Cultures without peptide and with anti-CD3 antibody were included as negative and positive control, respectively. ZsGreen-1 expression was assessed by flow cytometry. (c) 5KC cells expressing the GSE.166H9 TCR and EBV-transformed autologous B cells were co-cultured with the insulin B:9–23 peptide and the TKT peptide 8 (100 mM) in the presence or absence of anti-DR, DQ, or DP antibodies, followed by evaluation of ZsGreen-1 expression by flow cytometry.



(d)



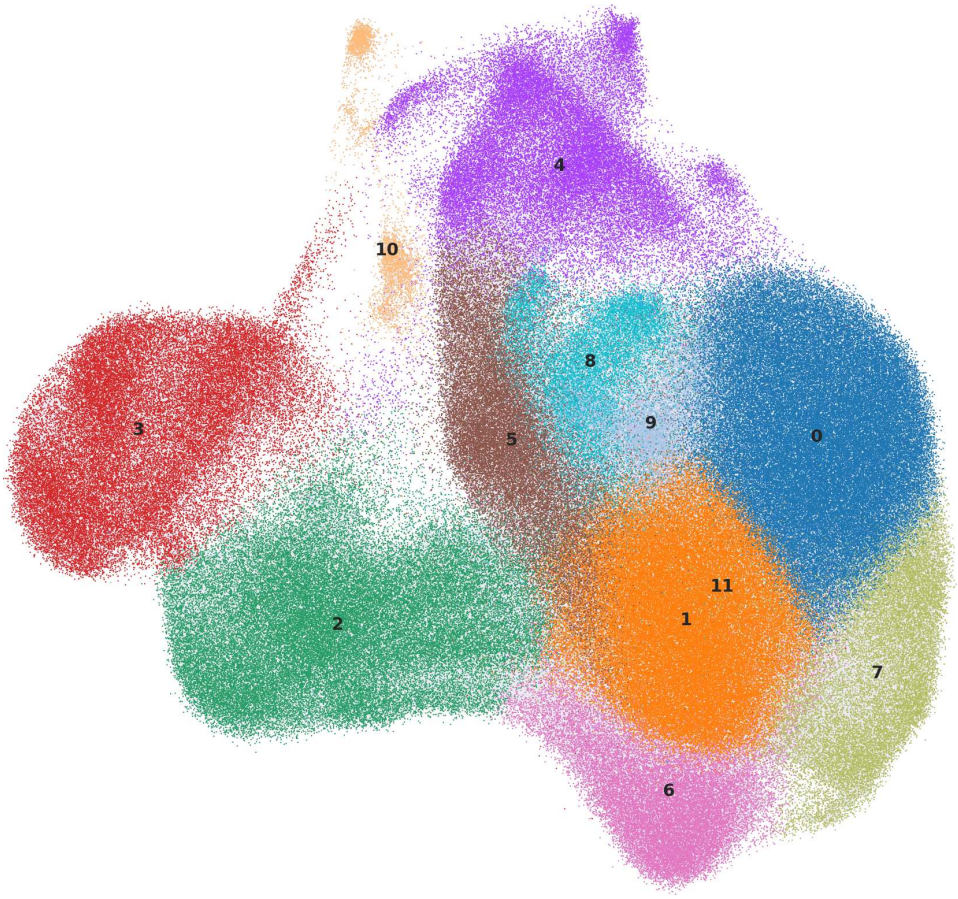
(e)



(f)

(d) 5KC cells expressing the GSE.166H9 TCR were cultured with the peptides (100 mM) in the presence of K562 myeloma cells expressing DQ8, DQ8-trans, DQ2, or DQ2-trans molecules, followed by evaluation of ZsGreen-1 expression. (e) and (f) 5KC cells expressing the GSE.166H9 TCR were cultured with different concentrations of the insulin B:9–23 peptide and the TKT peptide 8 in the presence of K562 myeloma cells expressing DQ8-trans (e) or DQ8 (f), followed by evaluation of ZsGreen-1 expression. Screening experiment results shown in panels (b) and (c) were performed once. Experiments in panels (a), (d), (e), and (f) were repeated two (a), three (d), or four times (e and f), and mean values ± standard error of the mean are shown. ns: not significant, *$P < 0.05$ and **$P < 0.01$ (calculated by a two-tailed paired t-test).

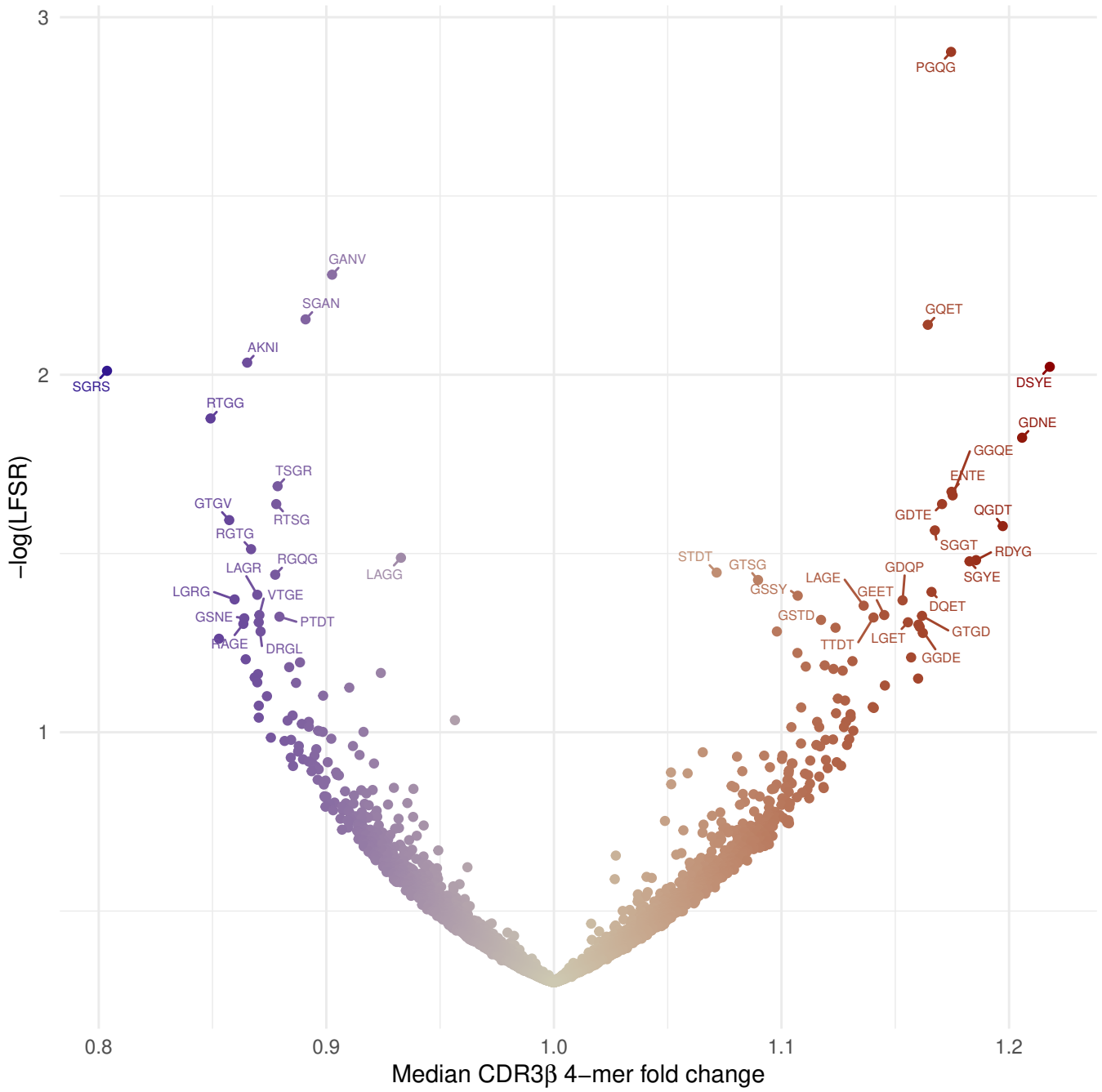
Extended Data Fig. 1: Single-cell subpopulations in the D-GAP cohort.



	Gene marker rank				
	1	2	3	4	5
0	*CD40LG*	*NFKBID*	*SLA*	*H3F3B*	*PRNP*
1	AC083862*	AL590652*	AP001160*	*HSD17B1*	*LINC00926*
2	*FOS*	*DNA1B1*	*TSC22D3*	*JUN*	*DUSP1*
3	*TXNIP*	*CXCL3*	*TMSB10*	*CCL4*	*SH2D1B*
4	*IL13*	*IFNG*	*TNF*	*CCL5*	*IL2*
5	*C17orf100*	*CYTOR*	*CNTLN*	AG01000058*	*CLDND1*
6	*RPS2*	*RPS27*	*RGCC*	*PIM3*	*DUSP2*
7	*CXCL3*	*CXCL8*	*CXCL2*	*CLDN1*	*CCR4*
8	*FOXP3*	*CTLA4*	*TNFRSF9*	*IKZF2*	*TIGIT*
9	*MALAT1*	*SRGN*	*CD69*	*IL7R*	*HLA-B*
10	*CRTAM*	*XCL2*	*XCL1*	*KLRK1*	*CSF2*
11	*CLPTM1L*	*CRBN*	*ZNF519*	*AKNA*	*SAMD10*

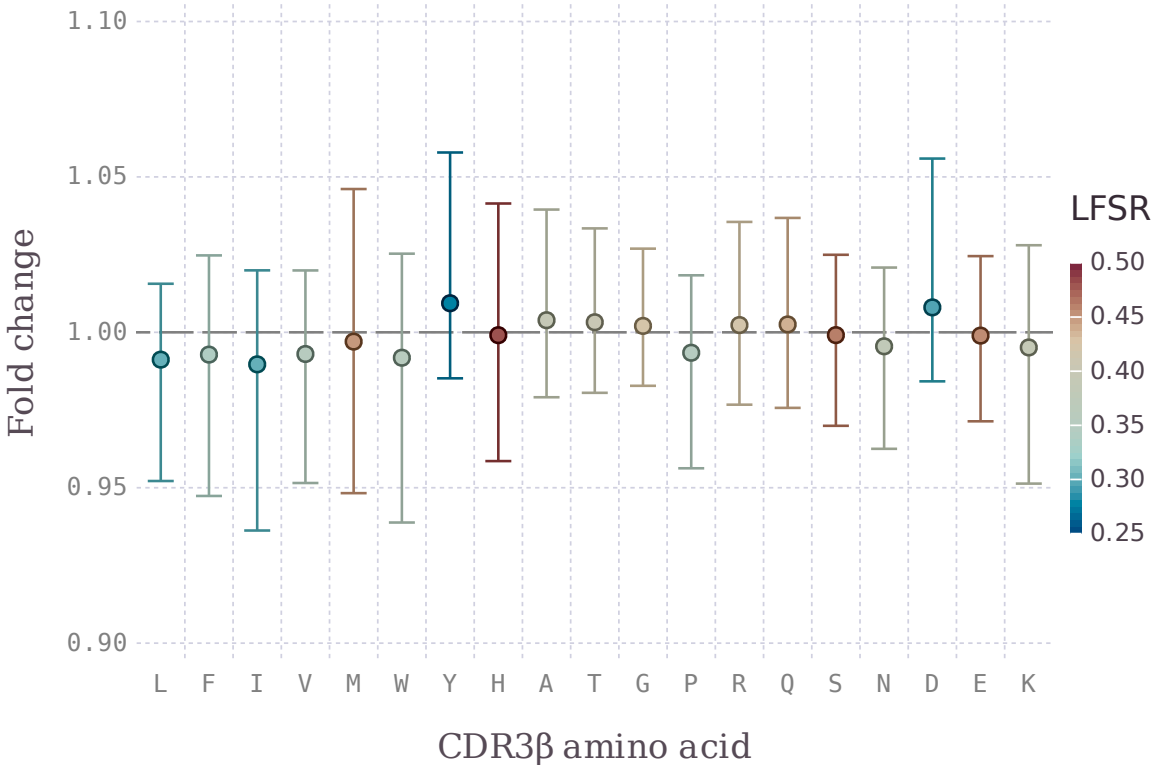
Gene markers for single $CD4^+$ T cell clusters 0–11, depicted in UMAP projection. Cluster 7 corresponds to recent thymic emigrants (RTEs) and cluster 8 to regulatory T cells (Tregs). * indicates no standard gene name is currently assigned, and the identifier of the corresponding genomic region has been used as a placeholder.

Extended Data Fig. 2: Estimates of CDR3β 4-mer fold changes across HLA class II risk extremes for Tconv cells.

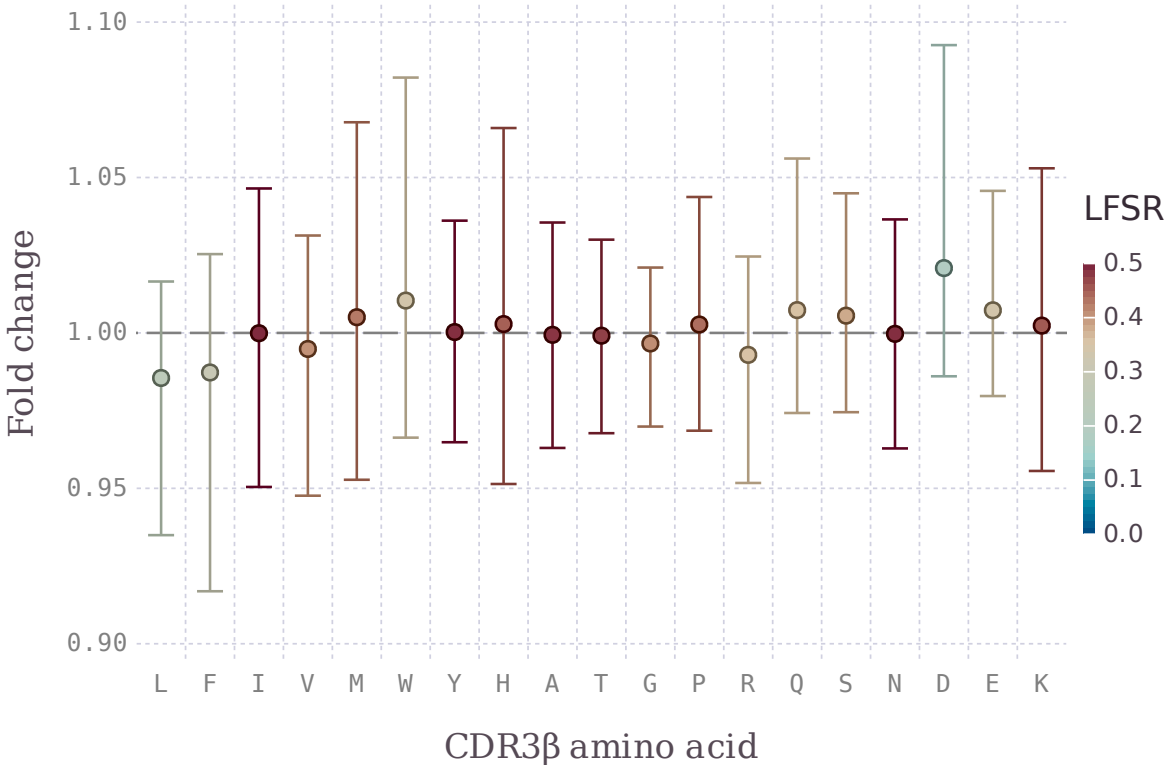


CDR3β 4-mer fold change and local false sign rate (LFSR) estimates for CD4$^+$ T conventional (Tconv) cells across HLA class II risk extremes, DQ6 (baseline) vs. DQ2/8.

Extended Data Fig. 3: Estimates of CDR3β k-mer fold changes across HLA class II risk extremes for RTE and Treg cells.



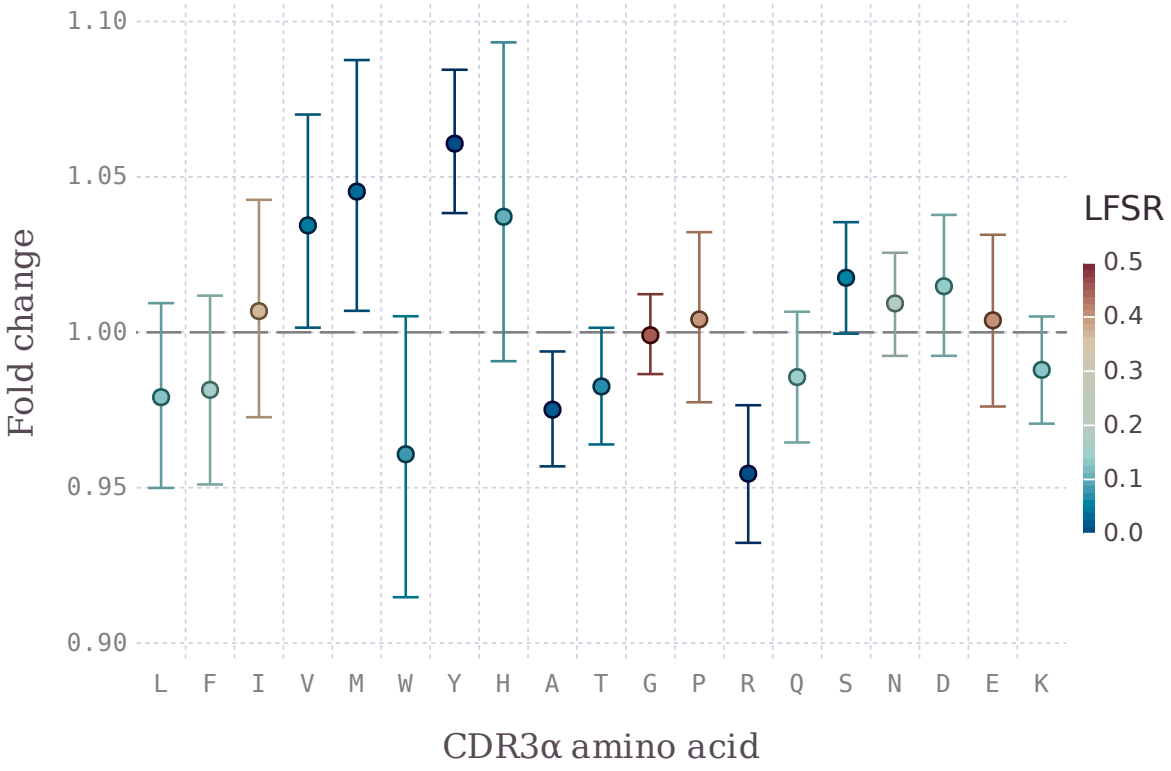
(a)



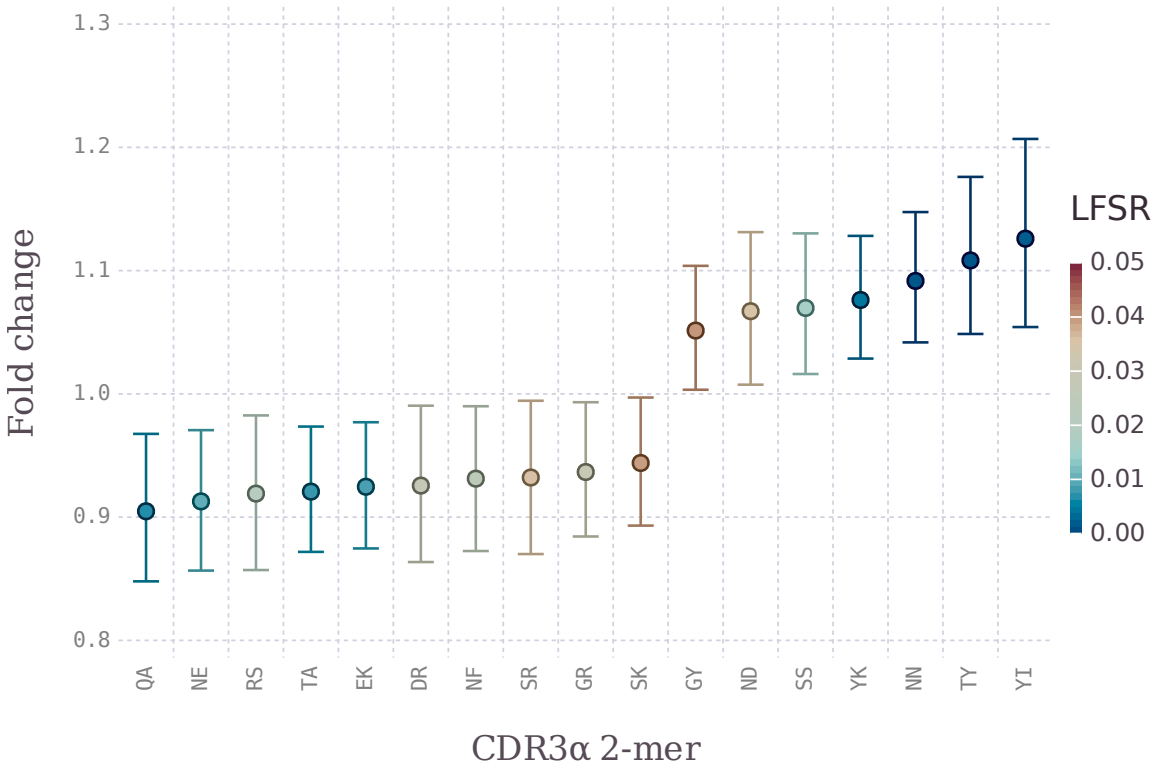
(b)

CDR3β k-mer fold change and local false sign rate (LFSR) estimates for $CD4^+$ (a) recent thymic emigrants (RTEs) and (b) regulatory T cells (Tregs), DQ6 (baseline) vs. DQ2/8. Error bars represent the posterior median and a 90% credible interval.

Extended Data Fig. 4: Estimates of CDR3α k-mer fold changes across HLA class II risk extremes for Tconv cells.



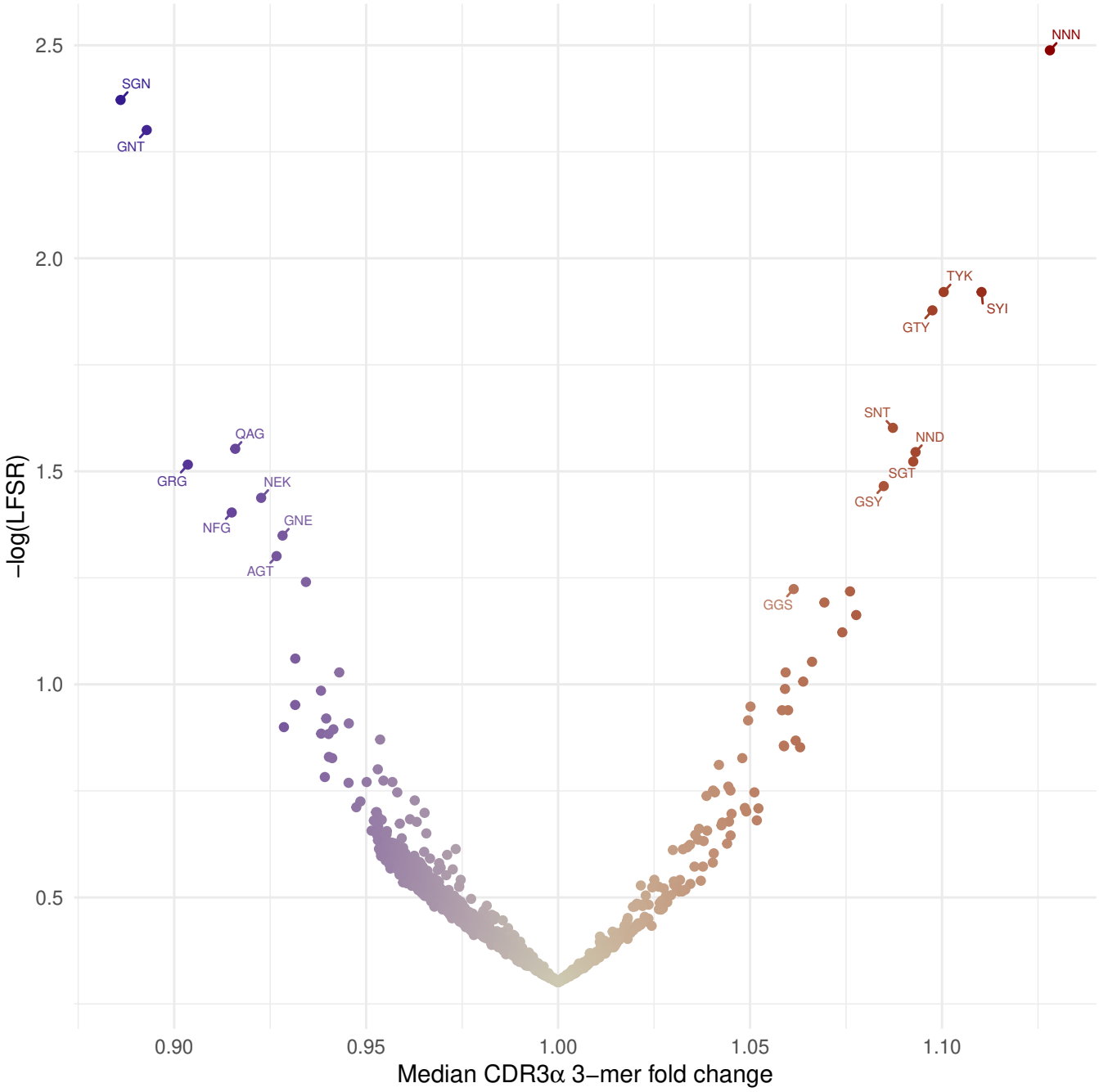
(a)



(b)

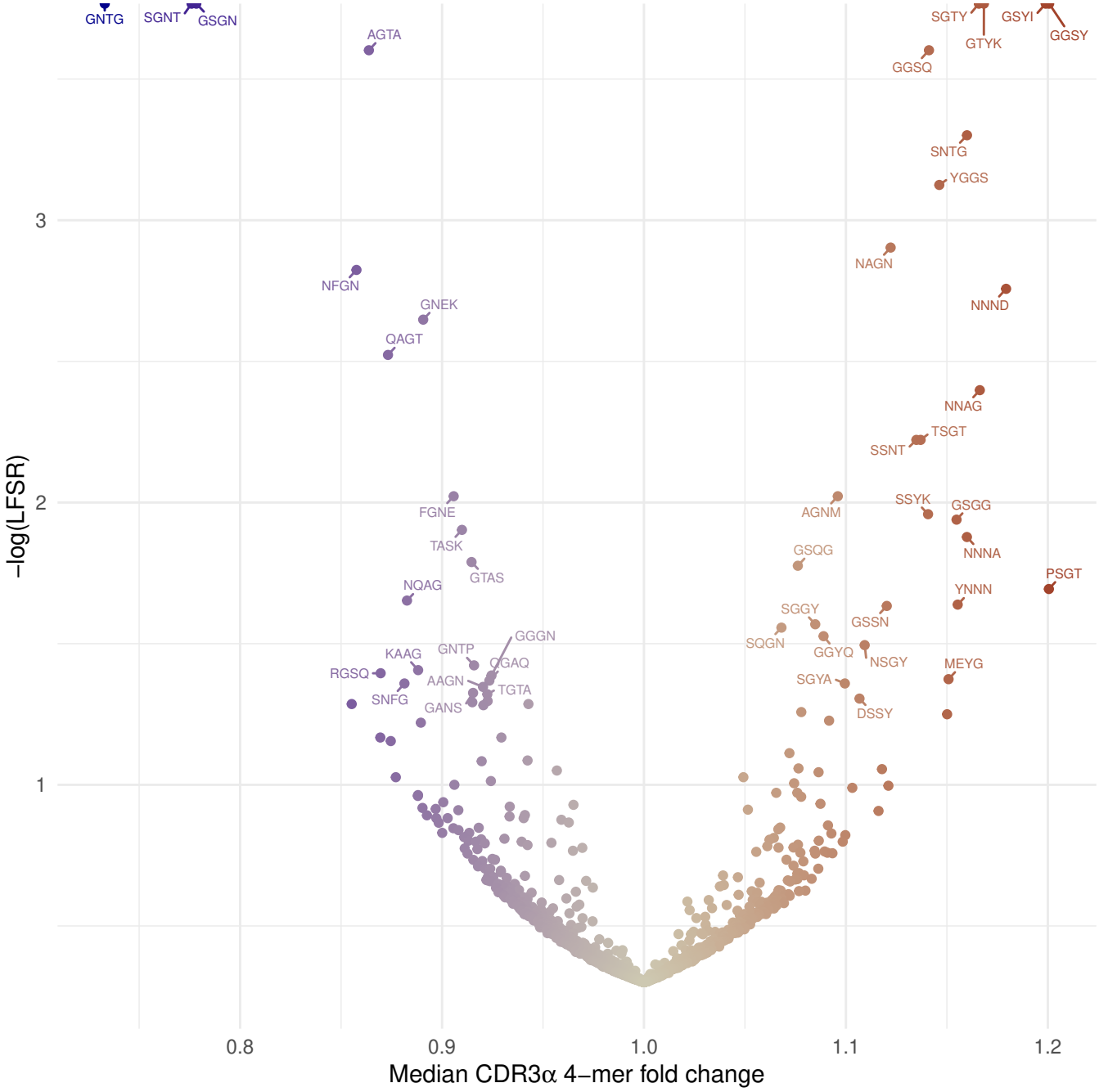
CDR3α k-mer fold change and local false sign rate (LFSR) estimates for CD4$^+$ T conventional (Tconv) cells (a) k=1 and (b) k=2, DQ6 (baseline) vs. DQ2/8. Error bars represent the posterior median and a 90% credible interval.

Extended Data Fig. 5: Estimates of CDR3α 3-mer fold changes across HLA class II risk extremes for Tconv cells.



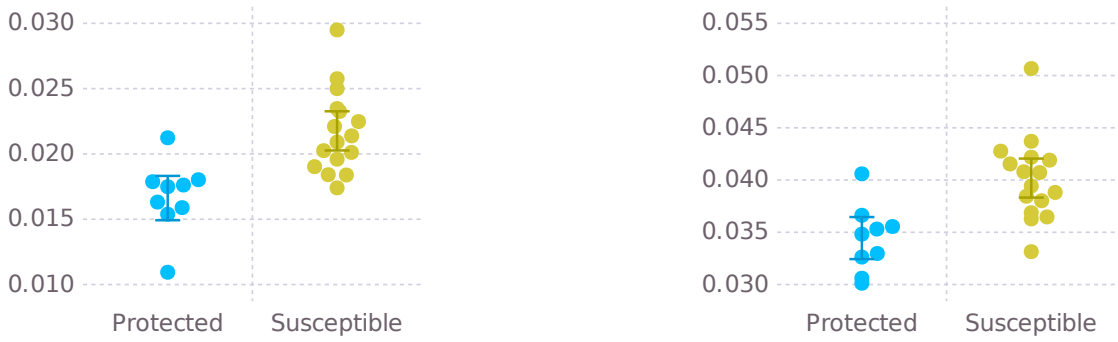
CDR3α 3-mer fold change and local false sign rate (LFSR) estimates for $CD4^+$ T conventional (Tconv) cells across HLA class II risk extremes, DQ6 (baseline) vs. DQ2/8.

Extended Data Fig. 6: Estimates of CDR3α 4-mer fold changes across HLA class II risk extremes for Tconv cells.



CDR3α 4-mer fold change and local false sign rate (LFSR) estimates for CD4$^+$ T conventional (Tconv) cells across HLA class II risk extremes, DQ6 (baseline) vs. DQ2/8. 4-mers on the y axis boundary have an estimated LFSR ≈ 0.

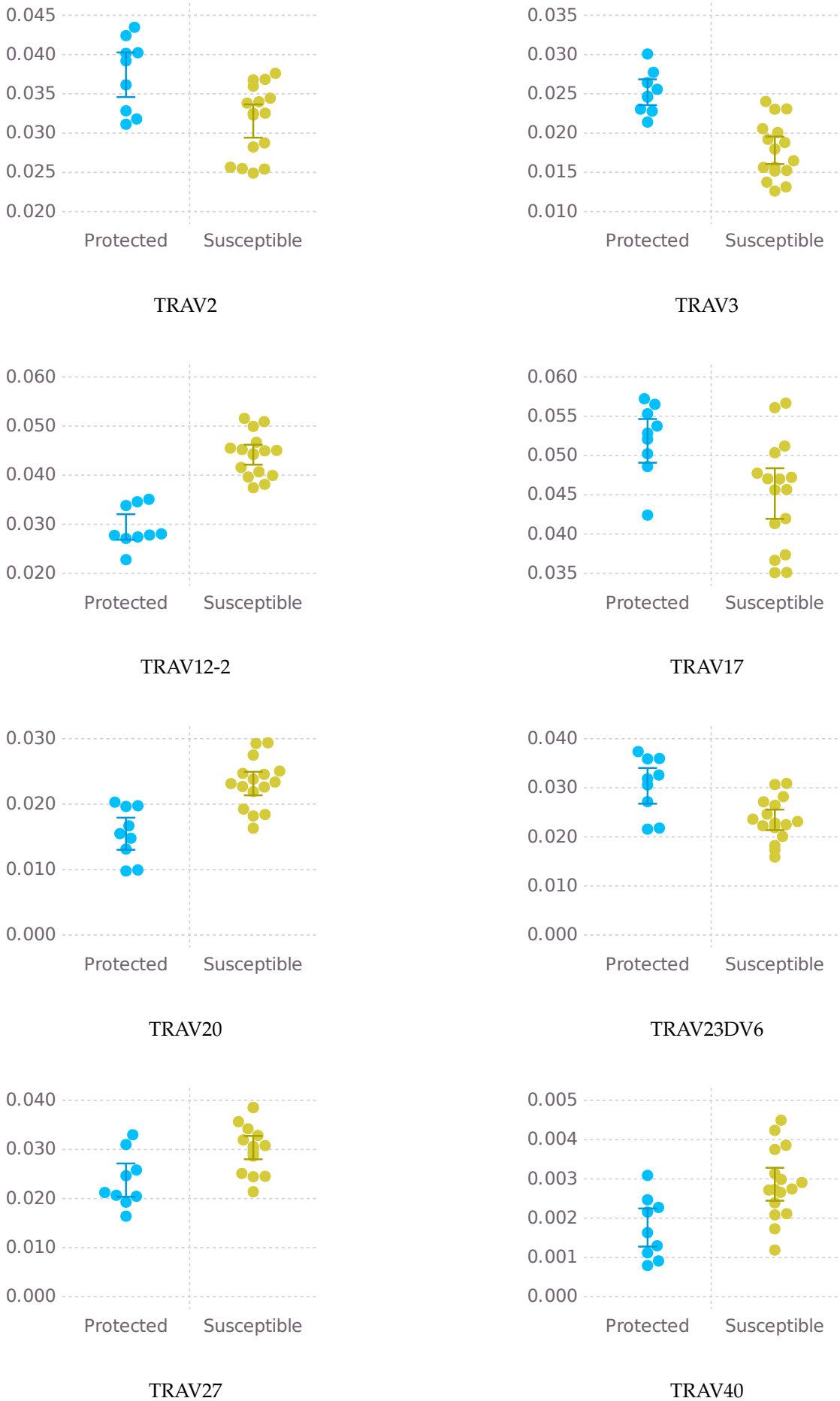
Extended Data Fig. 7: Differentially used TCR α chain joining genes.



TRAJ6 TRAJ40

Results filtered at FDR < 5%. Error bars represent a 95% confidence interval for the mean frequency. This comparison was made using a simpler two-group design, with nine DQ6 (Protected) and 16 DQ2/DQ8 donors (Susceptible).

Extended Data Fig. 8: Differentially used TCR α chain variable genes.



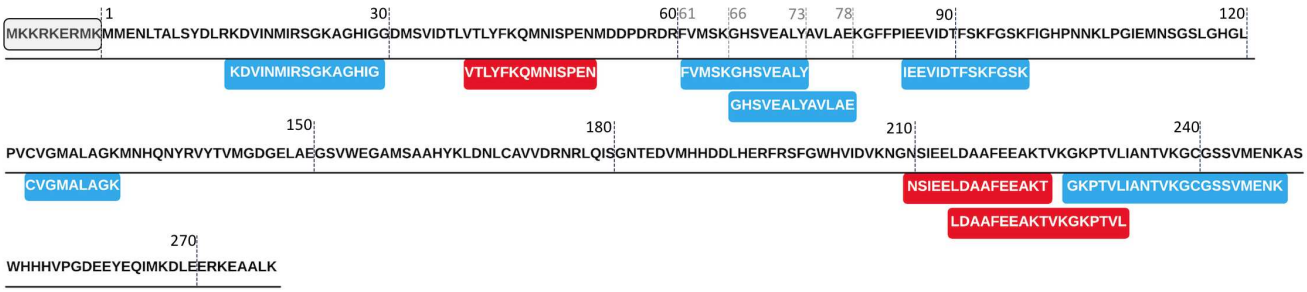
Results filtered at FDR < 5%. Error bars represent a 95% confidence interval for the mean frequency. This comparison was made using a simpler two-group design, with nine DQ6 (Protected) and 16 DQ2/DQ8 donors (Susceptible).

Extended Data Table 1: Top scoring local alignments to insulin B:9–25 in proteins from the gut microbiome.

Species	Sequence	Protein	Score	FDR
Acetatifactor sp.	HCVDALYMVLGEKGFF	TKT	98	0.001
Blautia caecimuris	HSVEALYAVLAEKGFF	TKT	88	0.041
Succinivibrio sp.	HSVEALYAVLAEKGFF	TKT	88	0.041
Acutalibacter sp.	HSVEALYAVLADRGFF	TKT	88	0.041
Pygmaiobacter sp.	HAVEALYSVLADRGFF	TKT	88	0.041
Oscillospiraceae sp.	HSVEALYAVLADRGFF	TKT	88	0.041
Lachnospiraceae sp.	HSVEALYAVLAEKGFF	TKT	88	0.041
Borkfalkia sp.	VEPIYLVCGEDAFF	YqeN	87	0.045
Mitsuokella multacida	HCVEALYAILADRGFF	TKT	86	0.045
Christensenella sp.	HAAPALYAVLGERGFF	TKT	86	0.045
Mobilibacterium sp.	HAVPALYAALGERGFF	TKT	86	0.045
Clostridium leptum	HSVEALYCILADRGFF	TKT	85	0.048
Eisenbergiella sp.	HCVDALYMVLGDLGFF	TKT	85	0.048
Lachnospiraceae sp.	HCVDALYMVLGDLGFF	TKT	85	0.048
Faec. prausnitzii	SHLEEVLYLLCGEK	GloB	85	0.048
⋮	⋮	⋮	⋮	⋮
Ruminococcus sp.	YIVCGERGF	—	78	0.129
Faec. prausnitzii	DAIYLLCGERGL	GH25	77	0.132
Ruminococcus sp.	YLTCGENGFF	ComE	75	0.200
Anaerobutyricum hallii	EALYLVCGE	CH3	75	0.200
Insulin B:9–25	SHLVEALYLVCGERGFF		145	

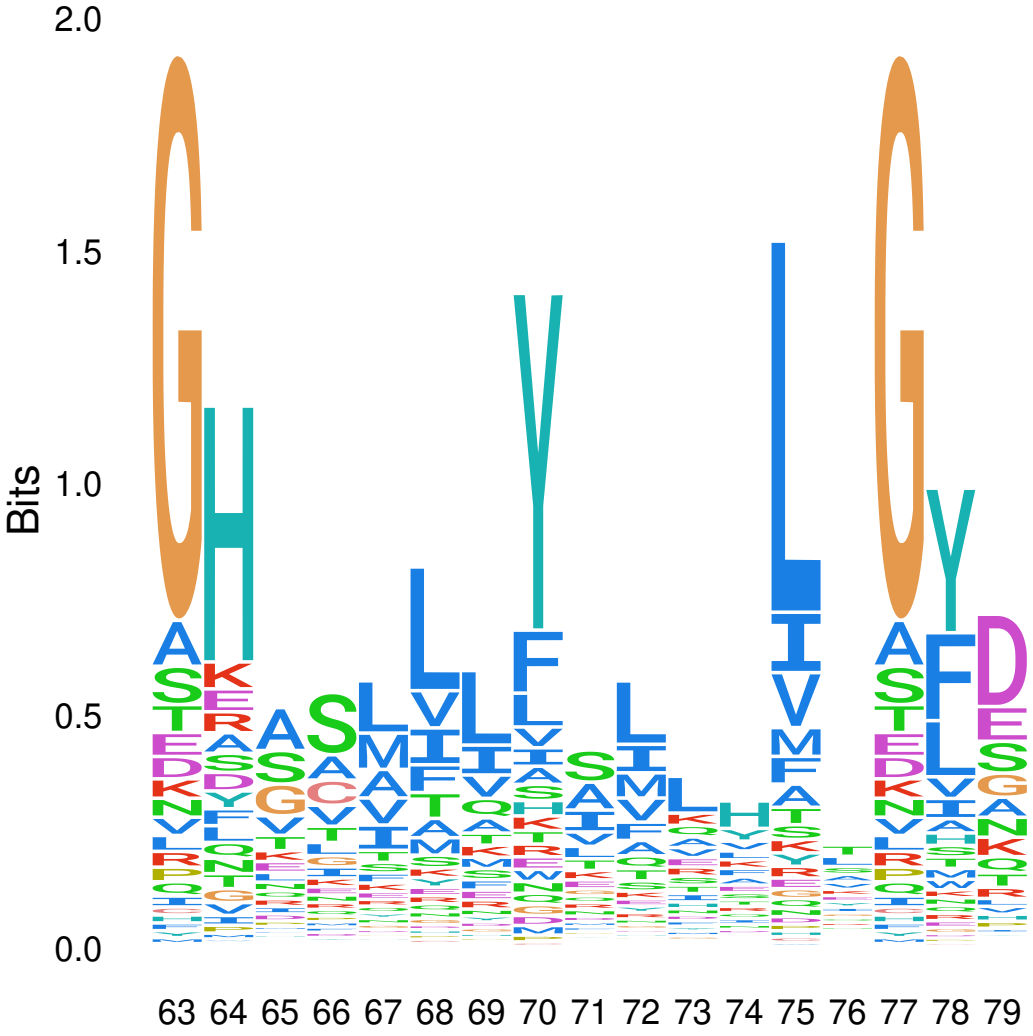
Results filtered at FDR ≤ 0.2. Protein domains and superfamilies listed include transketolase (TKT), yqeN DNA replication protein (YqeN), hydroxyacylglutathione hydrolase (GloB), glycosyl hydrolase family 25 (GH25), late competence operon (ComE) and cyclases/histidine kinases associated sensory extracellular domain 3 (CH3).

Extended Data Fig. 9: Immunopeptidomic protein map.



Peptides identified from *Blautia caecimuris* transketolase (TKT) in HLA-DQ peptide elution mass spectrometry experiments. Using an anti-DQ antibody (SPVL3), peptides presented by EBV B-cell lines were identified from elution experiments with either DQ6-expressing cells (blue) or DQ8-expressing cells (red). The initial protein sequence of nine amino acids (`MKKRKERMK`) at the beginning of the protein sequence originates from the IGC protein assembly[91]. MGnify[39] predicts the start of the protein at the third methionine within the primary sequence of *B. caecimuris* TKT, and this is depicted as position 1.

Extended Data Fig. 10: Transketolase superfamily motif for all domains of life in the region of insulin mimicry.



Most conserved residues in transketolase (TKT) are similar or identical to those from the corresponding insulin B:9–25 (SHLVEALYLVCGERGFF) position. This may facilitate the evolution of insulin B:9–25 mimotopes using TKT as a template.

Supplementary information

	Haplotype 1		Haplotype 2				
n	DQA1	DQB1	DQA1	DQB1	$DQB1_{57}$	OR	Group
23	03:01	03:02	05:01	02:01	AA	1.80	Susceptible
8	03:01	03:02	03:01	03:02	AA	1.51	Susceptible
1	03:02	02:02	03:02	03:01	AD	0.59	Low risk
1	03:01	03:01	06:01	03:01	DD	0.59	Low risk
4	03:02	03:01	03:02	03:01	DD	0.58	Low risk
1	03:03	03:01	05:05	03:01	DD	0.09	Low risk
1	03:01	03:01	05:05	03:01	DD	0.09	Low risk
2	01:02	06:02	05:05	03:01	DD	-0.89	Protected
2	01:02	06:02	01:02	06:02	DD	-0.96	Protected
2	01:02	06:02	05:05	06:03	DD	-1.30	Protected
2	01:02	06:02	01:02	05:02	DS	-1.30	Protected
1	01:02	06:02	01:02	06:04	DV	-1.30	Protected

Table S1: D-GAP sample selection and HLA class II diplotype interactions. T1D odds ratios (OR) for DR-DQ haplotype combinations correspond to $\log_{10}$-transformed estimates from UK Biobank.[10]

Haplotype 1			Haplotype 2						
DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	AA_1	AA_2	Age	Sex
15:01	01:02	06:02	11:04	05:05	03:01	DRA	DSR	[20, 25)	M
15:01	01:02	06:02	11:04	05:05	06:03	DRA	DSR	[15, 20)	M
15:01	01:02	06:02	11:04	05:05	06:03	DRA	DSR	[15, 20)	F
15:01	01:02	06:02	12:01	05:05	03:01	DRA	DGR	[15, 20)	F
15:01	01:02	06:02	13:02	01:02	06:04	DRA	VSE	[15, 20)	F
15:01	01:02	06:02	15:01	01:02	06:02	DRA	DRA	[5, 10)	M
15:01	01:02	06:02	15:01	01:02	06:02	DRA	DRA	[5, 10)	M
15:01	01:02	06:02	16:01	01:02	05:02	DRA	SRR	[10, 15)	M
15:01	01:02	06:02	16:01	01:02	05:02	DRA	SRR	[15, 20)	F

Table S2: D-GAP genetically protected group. AA_1 and AA_2 encode amino acids at HLA DQB1 57, DRB1 13 and 71.[11]

Haplotype 1			Haplotype 2						
DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	AA_1	AA_2	Age	Sex
04:01	03:01	03:01	11:03	05:05	03:01	DHK	DSE	[20, 25)	M
04:01	03:01	03:01	12:02	06:01	03:01	DHK	DGR	[15, 20)	F
04:01	03:02	03:01	04:01	03:02	03:01	DHK	DHK	[10, 15)	M
04:01	03:02	03:01	04:01	03:02	03:01	DHK	DHK	[5, 10)	M
04:01	03:02	03:01	04:01	03:02	03:01	DHK	DHK	[5, 10)	M
04:07	03:03	03:01	01:03	05:05	03:01	DHR	DFE	[10, 15)	M
04:08	03:02	03:01	04:07	03:02	03:01	DHR	DHR	[5, 10)	M
09:01	03:02	02:02	04:01	03:02	03:01	AFR	DHK	[5, 10)	F

Table S3: D-GAP low genetic risk group. AA_1 and AA_2 encode amino acids at HLA DQB1 57, DRB1 13 and 71.[11] The DQA1*03:02 allele is unresolved to 03:02 or 03:03.

Haplotype 1			Haplotype 2						
DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	AA_1	AA_2	Age	Sex
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	M
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	M
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	M
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[10, 15)	M
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[15, 20)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[15, 20)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[15, 20)	M
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[15, 20)	M
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[20, 25)	M
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[5, 10)	F
04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	[5, 10)	F
04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	[10, 15)	M
04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	[10, 15)	M
04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	[10, 15)	F
04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	[15, 20)	M
04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	[5, 10)	F
04:02	03:01	03:02	03:01	05:01	02:01	AHE	ASK	[10, 15)	M
04:03	03:01	03:02	03:01	05:01	02:01	AHR	ASK	[10, 15)	M
04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	[10, 15)	F
04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	[10, 15)	M
04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	[5, 10)	F
04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	[5, 10)	F
04:04	03:01	03:02	04:01	03:01	03:02	AHR	AHK	[15, 20)	M
04:04	03:01	03:02	04:04	03:01	03:02	AHR	AHR	[10, 15)	M
04:04	03:01	03:02	04:04	03:01	03:02	AHR	AHR	[5, 10)	M

Table S4: D-GAP genetically susceptible group. AA_1 and AA_2 encode amino acids at HLA DQB1 57, DRB1 13 and 71.[11]

	Haplotype 1			Haplotype 2				
	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	Age	Sex
11	04:01	03:01	03:02	04:02	03:01	03:02	[5, 10)	M
13	03:01	05:01	02:01	03:01	05:05	02:02	[10, 15)	M
14	04:01	03:01	03:02	04:01	03:01	03:02	[5, 10)	F
16	03:01	05:01	02:01	03:01	05:05	02:02	[10, 15)	M
18	04:01	02:01	02:02	07:01	03:01	03:02	[5, 10)	F

Table S5: DILmech donor HLA DR–DQ diplotypes.

	IAA		GADA		
	Status	U/ml	Status	U/ml	ICA
11	Negative	0.80	Positive	15.9	Positive
13	Negative	0.80	Positive	91.9	Positive
14	Positive	6.34	Positive	86.3	Positive
16	Positive	8.44	Positive	88.2	Positive
18	Positive	6.14	Positive	71.8	Positive

Table S6: DILmech donor autoantibody status at the time of diagnosis and sample collection.

		Haplotype 1			Haplotype 2				
	RRID	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	Age	Sex
69	SAMN15879056	04:01	03:01	03:02	07:01	02:01	02:02	[5, 10)	F
6323	SAMN15879377	03:01	05:01	02:01	04:02	03:01	03:02	[20, 25)	F
6342	SAMN15879396	01:01	01:01	05:01	04:01	03:01	03:02	[10, 15)	F
6367	SAMN15879420	04:01	03:01	03:02	07:01	02:01	02:02	[20, 25)	M
6414	SAMN15879467	03:01	05:01	02:01	09:01	03:03	02:02	[20, 25)	M
6472	SAMN15879525	03:01	05:01	02:01	04:04	03:01	03:02	[10, 15)	F
6533	SAMN18242777	03:01	05:01	02:01	04:01	03:01	03:02	[0, 5)	F
6536	SAMN18242780	DR4	DQ8	DQ8	DR17	DQ2	DQ2	[20, 25)	F

Table S7: nPOD donor HLA DR–DQ diplotypes.

	IAA	IA2A	GADA	ZnT8A	PPI-reactive		Diag.
					Positive	Tested	
69	—	—	—	—	0	7	3
6323	Negative	Positive	Positive	Negative	4	52	6
6342	Positive	Positive	Negative	Negative	5	34	2
6367	Negative	Negative	Negative	Negative	0	9	2
6414	Positive	Negative	Positive	Positive	4	34	½
6472	Positive	Negative	Negative	Negative	1	30	4
6533	Positive	Positive	Negative	Positive	—	—	0
6536	Negative	Negative	Positive	Negative	—	—	4

Table S8: nPOD donor autoantibody status, number of preproinsulin-reactive (PPI) $CD4^+$ T cell clones isolated and years since diagnosis at the time of death.

	Sequence	Species
1	GHSVEALYAVLAEKG	*Blautia caecimuris*
2	HSVEALYAVLAEKGF	*Blautia caecimuris*
3	SVEALYAVLAEKGFF	*Blautia caecimuris*
4	GHCVEALYVTLEAKG	*Phocaeicola dorei*
5	HCVEALYVTLEAKGF	*Phocaeicola dorei*
6	CVEALYVTLEAKGFI	*Phocaeicola dorei*
7	GHTVEALYAVLCQKG	*Eubacterium siraeum*
8	GHSVEALYCILADRG	*Clostridium leptum*
9	GHIAEALYVTLAKRG	*Coprobacillus sp.*
10	GHCVEALYVTLESKG	*Phocaeicola dorei*

Table S9: Peptides employed for $CD4^+$ T cell stimulation assay.

Sample	Peptide	-10 log(P)	FDR	Source
DQ6	KDVINMIRSGKAGHIG	49.23	2.9	PEAKS DB
DQ6	IEEVIDTFSKFGSK	42.79	2.9	PEAKS DB
DQ6	FVMSKGHSVEALY	47.21	2.9	PEAKS DB
DQ6	GHSVEALYAVLAE	35.34	2.9	PEAKS PTM
DQ6	CVGMALAGK	27.34	2.9	PEAKS PTM
DQ6	GKPTVLIANTVKGCGSSVMENK	57.60	2.9	PEAKS PTM
DQ8	VTLYFKQMNISPEN	49.95	2.5	PEAKS PTM
DQ8	NSIEELDAAFEEAKT	24.84	2.5	PEAKS DB
DQ8	LDAAFEEAKTVKGKPTVL	43.62	2.5	PEAKS DB

Table S10: Eluted peptides from *Blautia caecimuris* transketolase that were measured to be bound by DQ molecules using mass spectrometry.

$\mu_i \sim \text{Normal}(\mu{=}0, \sigma{=}5)$
$\sigma_i, \zeta \sim \text{HalfCauchy}(\mu{=}0, \sigma{=}2.5)$
$\rho \sim \text{LKJ}(\eta{=}2)$
for $i \in \{1, \ldots, K\}$ **do**
    $\xi_i \sim \text{MultiNormal}(\mu{=}[\mu_1, \mu_2], \sigma{=}[\sigma_1, \sigma_2], \rho{=}\rho)$
    for $j \in \{1, \ldots, N\}$ **do**
        $\epsilon_{ij} \sim \text{Normal}(\mu{=}0, \sigma{=}\zeta)$
        $k_{ij} \sim \text{Binomial}(n{=}n_j, p{=}\text{logit}^{-1}(\xi_{i1} + \xi_{i2}\, or_j + \epsilon_{ij}))$
    end for
end for

Algorithm S1: Hierarchical mixed HLA effects model.

$p \sim \text{Beta}(\alpha{=}50, \beta{=}1000)$
for $i \in \{1, \ldots, N\}$ **do**
    $k_i \sim \text{Binomial}(n{=}n_i, p{=}p)$
end for

Algorithm S2: Beta-binomial aspartic acid proportion model.

$\alpha \sim \text{Normal}(\mu{=}0, \sigma{=}100)$
$\beta \sim \text{Normal}(\mu{=}0, \sigma{=}1000)$
for $i \in \{1, \ldots, N\}$ **do**
    $p_i \sim \text{Beta}(\alpha{=}50, \beta{=}1000)$
    $l_i \sim \text{Binomial}(n{=}m_i, p{=}p_i)$
    $k_i \sim \text{Binomial}(n{=}n_i, p{=}\text{logit}^{-1}(\alpha + \beta\, p_i))$
end for

Algorithm S3: Binomial preproinsulin-reactivity regression model.

Equation S1: Generalized extreme value (GEV) log-likelihood

$$\ell(\mu, \sigma, \xi) = - m \log(\sigma) - (1 + 1/\xi) \sum_{i=1}^{m} \log \left(1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right) - \sum_{i=1}^{m} \left(1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right)^{-1/\xi}$$

subject to

$$1 + \xi \left(\frac{z_i - \mu}{\sigma}\right) > 0, \ \forall i \in [1, m]$$

If $\xi = 0$, we require separate treatment using the Gumbel limit

$$\ell(\mu, \sigma) = -m \log(\sigma) - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}$$