Autoimmune interactions between the HLA-DQβ1$_{57}$ polymorphism, T cell receptors, and microbial mimics of insulin in type 1 diabetes

Arcadio Rubio García[a,*], Athina Paterou[a], Rebecca D. Powell Doherty[b], Laurie G. Landry[c], Mercede Lee[a], Amanda M. Anderson[c], Claire L. Scudder[a], Hubert Slawinski[d], Ricardo C. Ferreira[a], Dominik Trzupek[a], Agnieszka Szypowska[e], Luc Teyton[f], Nicola Ternette[b], Maki Nakayama[c,g,h], Linda S. Wicker[a], John A. Todd[a,*], Marcin L. Pekalski[a,i,*]

[a]*JDRF/Wellcome Diabetes and Inflammation Laboratory, Centre for Human Genetics, Nuffield Department of Medicine, NIHR Biomedical Research Centre, University of Oxford, Oxford, UK*
[b]*The Jenner Institute, Centre for Immuno-Oncology, Old Road Campus Research Building, University of Oxford, Oxford, UK*
[c]*Barbara Davis Center for Childhood Diabetes, University of Colorado School of Medicine, Aurora, Colorado, USA*
[d]*Centre for Human Genetics, Nuffield Department of Medicine, NIHR Biomedical Research Centre, University of Oxford, Oxford, UK*
[e]*Department of Paediatric Diabetology and Paediatrics, Medical University of Warsaw, Warsaw, Poland*
[f]*Department of Immunology and Microbiology, Scripps Research Institute, La Jolla, California, USA*
[g]*Department of Pediatrics, University of Colorado School of Medicine, Aurora, Colorado, USA*
[h]*Department of Immunology and Microbiology, University of Colorado School of Medicine, Aurora, Colorado, USA*
[i]*Military Medical Institute, Warsaw, Poland*

Abstract

Insulin autoimmunity and pancreatic islet β-cell destruction, mediated by T cells, underlie the pathogenesis of type 1 diabetes (T1D). However, the mechanisms by which genetic and environmental risk factors interact to cause T1D remain incompletely understood. Here, we show how variation in the human leukocyte antigen (HLA) class II region, notably DQβ1$_{57}$, the strongest T1D genetic risk factor, induces a positive thymic selection bias that favors amino acid motifs with negatively charged side chains in the peptide-binding CDR3β region of CD4$^+$ T cell receptors (TCRs). This bias was enriched in TCRs from T1D patients and predicted the proportion of anti-insulin infiltrates in pancreatic islets. We link host genetics and immunity to the environment by identifying over 100 gut bacterial proteins with significantly similar sequences to a primary epitope in T1D, insulin B$_{9–25}$ peptide. We show that CD4$^+$ T cells isolated from the pancreas of a newly diagnosed child with T1D recognised both insulin and a bacterial mimotope peptide, restricted by the most predisposing DQ molecule. Our results point to the evolution of certain microbial antigens towards mimicry in the regulation of immune tolerance to insulin.

Keywords: Autoimmunity, type 1 diabetes (T1D), insulin (INS), human leukocyte antigen (HLA), major histocompatibility complex (MHC), T cell receptor (TCR), complementarity determining region 3 (CDR3), thymic selection, molecular mimicry, gut microbiome

Introduction

Type 1 diabetes (T1D) is a common autoimmune disease characterized by a T cell-mediated destruction of pancreatic islet β cells, resulting in a lifelong reliance on exogenous insulin for metabolic control and survival.[1] Autoantibodies against insulin often emerge as the first pre-clinical sign of the disease, peaking in incidence between 9 and 18 months of age.[2,3] Before this period, the immune system and the pancreas undergo major developmental changes, regulated by host genetic variation and shaped by successive waves of mucosal microbial colonization. This process is particularly relevant in the gut, which harbors the greatest number of T cells in the body and shares lymph nodes with the pancreas, leading to immune crosstalk and co-regulation between both organs.[4,5] Such an immune crosstalk, influenced by the interaction between host genes and the gut microbiome, might play a crucial role in the initiation of T1D.

The strongest genetic risk factor for T1D is located in the human leukocyte antigen (HLA) class II region, within the major histocompatibility complex (MHC) on chromosome 6p21. This region encodes molecules that display peptides, derived from extracellular protein antigens, to CD4$^+$ T cells for their recognition. In populations with European ancestry, HLA-DQβ1 position 57 and DRβ1 positions 13 and 71 explain ~15%, 11%, and 1% of the overall disease risk, respectively, and 80% of the risk contributed by the entire HLA region,[6] including HLA class I associations.[7] Inherited as DR-DQ haplotype pairs or diplotypes, these variants may increase or decrease predisposition to T1D up to a factor of 50 by altering the peptide-binding properties of DR and DQ (α, β) heterodimers.

HLA class II haplotypes that protect from T1D are associated with the presence of aspartic acid (D), a residue with a negatively charged side chain, at DQβ1$_{57}$,[8] which is located in a key peptide-binding pocket of the DQ molecule. Conversely, haplotypes conferring near-neutral to strong susceptibility include valine (V), serine (S), and alanine (A) at DQβ1$_{57}$, all lacking such an electrically charged side chain. The association between DQβ1$_{57}$

*Correspondence to: john.todd@well.ox.ac.uk, arcadio@pm.me, and marcin.pekalski@well.ox.ac.uk.

and T1D is conserved in the murine ortholog of DQβ, I-Aβ.[9] Specific mutation of I-Aβ_{57} (S→D) in the non-obese diabetic (NOD) mouse model of T1D supports a causal role for this residue in disease by reducing the incidence of T cell-mediated islet autoimmunity.[10]

Outside the HLA region, polymorphism of the insulin (*INS*) gene is one of the leading T1D risk factors in terms of effect size, explaining around 3% of the phenotypic variance.[6] Predisposing common alleles of causal variants located in the *INS* gene promoter region reduce its thymic expression, thereby increasing the frequency of anti-insulin CD4$^+$ T cells in the periphery.[11–13] Insulin, and its precursor preproinsulin, particularly peptides from positions 9–25 of its B chain ($B_{9–25}$), is a primary T1D autoantigen in humans and NOD mice.[14,15] Furthermore, the appearance of anti-insulin autoantibodies before 18 months of age is associated with the most T1D predisposing DR-DQ haplotype, DR4-DQ8.[2,3] Taken together, these results indicate HLA class II-restricted anti-insulin CD4$^+$ T cells are responsible for the destruction of the islet β cells.

Here, we show that the overall T1D risk contributed by DR-DQ diplotypes explains the thymic output of anti-insulin CD4$^+$ T cell receptors (TCR). This thymic output is driven by biases during positive thymic selection that arise from charge interactions between peptides displayed by cortical epithelial cells of the thymus, T1D-associated DR and DQ β-chain residues, most notably DQ$\beta1_{57}$, and TCR complementary-determining regions (CDR3) of maturing thymocytes.

However, carriers of the most T1D-predisposing diplotype, DR3-DQ2/DR4-DQ8, which encodes the highest risk DQ molecule, the transdimer (DQ2_α, DQ8_β), have an estimated probability of developing T1D during their life course of only 5%. Therefore, identifying environmental factors and gene-environment interactions that modulate disease risk is of critical importance not only to improve our understanding of the primary causal disease mechanisms, but also to design disease prevention interventions. During industrialization most populations have experienced a sharp increase in T1D incidence, accompanied by a decrease in the relative incidence of high-risk HLA class II diplotypes, including DR3-DQ2/DR4-DQ8, suggesting rapid changes in the prevalence of protective or predisposing environmental factors.[16]

Different environmental factors associated with a Western lifestyle, ultimately leading to an increase in gut dysbiosis during the first months of life, could be responsible for some of these rapid epidemiological changes. Children who develop persistent islet autoantibodies and progress to T1D diagnosis typically exhibit a reduction in overall gut microbiome diversity before any autoantibody-related events, along with a decrease in the abundance of commensal bacteria from the *Bifidobacterium* genus and an increase of other phyla,[17] in a geographical-dependent manner. This results in a reduced capacity to metabolize human milk oligosaccharides (HMOs) and an upregulation of pathways linked to inflammation, with likely implications for the development of the host immune system. Other observational studies have identified similar patterns in children that eventually developed T1D.[18,19]

In order to link disruptions to gut microbial ecology and the activation of anti-insulin T cells that leads to islet autoimmunity, we have modeled the similarity between insulin $B_{9–25}$ and every protein in the gut microbiome. Our work demonstrates the existence of over 100 bacterial proteins with a statistically significant similarity to insulin $B_{9–25}$, indicating evolutionary selection pressure to mimic insulin and a central role in the regulation of its immune tolerance.

We have identified a bacterial peptide mimic of insulin $B_{9–25}$, displayed in the canonical $B_{9–23}$ register, from the transketolase (TKT) enzyme family, which is highly upregulated during the transition from breastfeeding to solid food known as weaning.[20] This mimic is recognized by one of the few anti-insulin CD4$^+$ T cells isolated from recently diagnosed T1D donor pancreata, suggesting that a targeted intervention to preserve host-microbiome symbiosis during early life could be a potential disease prevention strategy,[4,21] particularly in children that are genetically predisposed to T1D.[22,23]

Results

HLA-DR-DQ and thymic T cell receptor selection

We analyzed the TCR α and β chain repertoires from selected donors carrying DR-DQ diplotypes in the whole protection-susceptibility risk continuum (n = 48, Table S1), with T1D odds ratio (OR) as the explanatory variable.[24] Purified CD4$^+$ T cells (n = 349 623) from PBMCs were stimulated and loaded into a single-cell platform for preparing paired 5' gene expression and TCR libraries. After sequencing, gene expression data analysis, and TCR assembly, we obtained 288 903 cells with both a valid gene expression profile and at least one productive TCR chain.

We partitioned the CD4$^+$ T cell compartment into conventional (Tconv), recent thymic emigrant (RTE),[25] and regulatory (Treg) cell clusters by gene expression analysis (Figure S1 and Table S5), since these could be subject to different receptor selection constraints. As expected, the Tconv cluster was the largest (n = 258 387, clonotypes = 255 952). Owing to the low age of the cohort, this was followed by RTEs (n = 18 194, clonotypes = 18 127) and Tregs (n = 12 322, clonotypes = 12 260).

We split TCR CDR3 region sequences from α and β chains into their constituent k-mers, with $k \in [1, 4]$, and aggregated counts for each (subset, chain, donor, clonotype) combination. CDR3 sequences are hypervariable and determine the peptide-binding affinity of each TCR. Counts were regressed against the log-transformed OR using a Bayesian hierarchical model to estimate the effect of DR-DQ diplotypes on each k-mer. This model takes into consideration the high variability of count data and estimates moderated effects by sharing information across variables.

We report effects as fold changes between the estimated k-mer frequencies for the most susceptible diplotype, DR3-DQ2/DR4-DQ8, compared to the most protective, DR15-DQ6/DR15-DQ6. We use the false-sign rate (FSR) as an error estimate, defined as the probability that the direction of the effect, most commonly fold change, is different from the posterior median. We also report 90% credible intervals (CI) as a summary of the posterior distribution of fold changes.

For k-mers present in the CDR3β region of Tconv TCRs, those that included at least one residue with a negatively charged side chain, aspartic acid (D) and glutamic acid (E), represented most of the sequences that exhibited fold changes greater than one as T1D OR increased, at FSR ≤ 0.05 (Figure 1). Susceptible donors carrying the hydrophobic amino acid alanine (A) at DQβ1$_{57}$ exhibited the highest frequency of D and E residues in CDR3β. Conversely, for protected donors, with at least one D at DQβ1$_{57}$, there were fewer D and E residues in CDR3β. Instead, these residues with negatively charged side chains were replaced by amino acids of high interaction potential with D at DQβ1$_{57}$, such as valine (V), leucine (L), and isoleucine (I).

For RTEs and Tregs we observed the same biases in the amino acid composition of CDR3β as a function of T1D OR (Figure S2), suggesting that genetic variation effects of HLA-DR and DQ on the immune repertoire are mediated by alterations to the positive thymic selection threshold, which is common to all T cell subsets. T cell clonotypes that contain amino acids with negatively charged side chains in the CDR3β region are more likely to form a stable complex with HLA molecules expressed in the surface of thymic epithelial cells for diplotypes with a higher T1D OR as these typically lack D, a residue with a negatively charged side chain, at DQβ1$_{57}$.

For TCRα chain sequences, we also estimated systematic differences at FSR ≤ 0.05 associated with T1D risk (Figure S3). These correspond to biases in the usage of particular Vα and Jα genes, which are carried into the CDR3α region owing to the lower recombination diversity of TCRα genes. In particular, top estimated differences in k-mers match fragments of SGTYK and GGSYI, sequences encoded in TRAJ40 and TRAJ6 genes, respectively. Usage of these fragments increased in individuals carrying susceptible DR-DQ diplotypes compared to protective ones. In NOD mice, studies assessing α chain patterns have reported an expansion of anti-insulin clones with a double glycine plus serine motif (GGS) in the CDR3α region.[26].

Negatively charged CDR3β sequences and anti-insulin responses

Next, we investigated whether these observed repertoire differences are linked to an immune response against insulin peptides. We purified activated circulating CD4$^+$ T cells (n = 19 969), defined as HLA-DR$^+$ CD38$^+$, from children with susceptible DR-DQ diplotypes and newly diagnosed T1D (n = 5, Table S6). This activated subpopulation is enriched in autoreactive CD4$^+$ T cells from patients with celiac and other autoimmune diseases.[27] We also sourced CD4$^+$ TCR clonotype sequences (n = 1428) isolated from the islets of another group of T1D patients with susceptible DR-DQ diplotypes (n = 7, median time since diagnosis = 2 years, Table S7) from the Network for Pancreatic Organ Donors with Diabetes (nPOD).[28] We compared these TCR sequences against those from individuals with the highest T1D susceptibility (DR4-DQ2/DR3-DQ8, n = 23, Table S4) from our original cohort.

Both CDR3β sequences from circulating activated cells and islets had an increased frequency of D residues when compared to non-activated circulating CD4$^+$ T cells from susceptible donors (Figure 2a). In the case of circulating activated cells, the posterior estimate only included fold changes above one, 90% CI = (1.115, 1.183) and FSR ≈ 0. Since few clonotypes were available from islet-infiltrating cells and our baseline was conservative, in this case the fold estimate had more uncertainty, 90% CI = (0.974, 1.120) and FSR = 0.151, but most of the posterior distribution density supported fold changes greater than one.

We also analyzed TCR clonotype sequences (n = 159) from nPOD T1D donors for which reactivity against preproinsulin (PPI) peptides had been measured (n = 5, median time since diagnosis = 2 years, Table S7).[29] The posterior fold change in PPI-reactivity for the highest observed D CDR3β content compared to the lowest had a median = 15.36 (Figure 2b), which supports a strong effect, FSR = 0.03. The posterior credible interval covered a wide range of folds over one due to the relatively small amount of data available and the sharp increase in reactivity for high D content, 90% CI = (1.28, 320.45).

Gut microbiome immune mimics of insulin

Differences in the proportion of anti-insulin TCRs, driven by the effects of HLA and other genetic risk variants that regulate thymic selection, may require environmental factors to cause a loss of immune homeostasis and initiate an immune response against islet β cells. Insulin mimics, encoded within microbial proteins present in the human gut, might play a fundamental role in this process. To identify these epitope mimics, or mimotopes, we investigated whether any commensal protein exhibits a statistically significant similarity to insulin B$_{9–25}$.

We assessed the similarity between insulin B$_{9–25}$ and any given protein as the maximum pairwise local alignment score between both sequences. This score represents the log-likelihood ratio between two models that consider amino acid substitutions as evolutionarily related or random, respectively. Log-likelihood ratios have been previously estimated using conserved regions in protein domain families, and represented as substitution matrices.[30] We estimated a null distribution of scores by drawing random permutations from the insulin B$_{9–25}$ sequence and calculating pairwise local alignment scores against every reviewed entry in the human proteome.

Then, we calculated a parametric approximation to this empirical distribution and aligned insulin B$_{9–25}$ to every protein in a gut metagenome reference catalog comprising millions of common bacterial and archaeal proteins.[31] This allowed the derivation of *p*-values for rare events, and false discovery rates (FDRs) to account for the large set of multiple comparisons performed. Of note, BLAST and other pairwise aligner derivatives use precalculated null distributions to reduce search runtime, and make assumptions that may not hold in case of a short query epitope.[32]

We found over 100 microbial proteins (Figure 3) that have a significant similarity to insulin B$_{9–25}$ at FDR ≤ 0.2, including 15 at FDR ≤ 0.05 (Table S8). The strongest association signal was found in proteins with a transketolase (TKT) N-terminal domain (Pfam PF00456). Significant associations were also present in other protein families, including YqeN DNA replication protein and hydroxyacylglutathione hydrolase (GloB). Both the number and the variety of associations suggest tolerance to insulin is regulated by a large niche of gut commensals.

A motif analysis of the TKT N-terminal domain revealed that conserved residues, required for enzymatic function and cofactor binding, are similar or identical to those from the corresponding insulin $B_{9–25}$ sequence (Figure S4). Therefore, this domain could have served as an ideal template for the evolution of insulin mimotopes.

T cell cross-reactivity between microbial mimics and insulin

Functional cross-reactive bacterial mimotopes require that certain $CD4^+$ T cells have TCRs that recognize both insulin $B_{9–25}$ peptides, in the canonical $B_{9–23}$ register, and one or more commensal epitopes. Therefore, we screened cloned $CD4^+$ TCRs (n = 179), detected in islets from nPOD T1D organ donors carrying at least one copy of DQ8 (n = 6, Table S7), with TKT mimotope peptides (n = 10, Table S10) and insulin $B_{9–23}$.

Only three islet $CD4^+$ TCRs that react to insulin displayed in the $B_{9–23}$ register are currently known (Figure 4a), including two previously reported[33] and one novel TCR clonotype, nPOD GSE.166H9 (Table S9). This TCR originated from a newly diagnosed patient, nPOD 6533 (DR3-DQ2/DR4-DQ8, age ∈ [0,5) years), and reacted to one of the TKT peptides (peptide 8 = `GHSVEALYCILADRG`) from *Clostridium leptum*. Notably, clonotype GSE.166H9 has an `ET` motif in the C-terminal side of its CDR3β chain, which increases in abundance proportionally to HLA class II T1D OR, FSR = 0.036.

TCR GSE.166H9 recognized TKT peptide 8 and insulin $B_{9–23}$ (Figure 4b), presented by HLA-DQ molecules (Figure 4c). This TCR responded most strongly to both peptides presented by DQ2-$DQ8_{trans}$, consisting of DQ2α and DQ8β transdimers (Figure 4d). This TCR also recognized peptides presented by a DQ heterodimer with both chains encoded by DQ8. Intensities of responses to insulin $B_{9–23}$ and TKT peptide 8 presented by DQ2-$DQ8_{trans}$ were similar, whereas GSE.166H9 T cells responded to insulin $B_{9–23}$ more strongly than TKT peptide 8 when presented by DQ8.

Therefore, the most potent response was induced by both insulin $B_{9–23}$ and peptide 8 presented by DQ2-$DQ8_{trans}$, indicating that immune responses by TCR GSE.166H9 were initiated by these peptide-DQ complexes. The demonstration of cross-reactivity to insulin and commensal peptides, initiated by the DQ2-$DQ8_{trans}$ molecule, helps explain why heterozygous individuals carrying both DQ2 and DQ8 have the highest predisposition to T1D.

We also sought evidence of natural processing and presentation of TKT proteins similar to those that contain putative mimotopes. We pulsed $DQ6^+$ and $DQ8^+$ B cell lines with a recombinant TKT protein. We then immunoprecipitated DQ and DR molecules, eluted bound peptides, and sequenced them by mass spectrometry. The top ranked eluted peptides bound to DQ6 corresponded to the insulin mimotope region (Table S11 and Figure S5). Another peptide from the mimotope region appeared to have a very similar presentation register to the only insulin B epitope that was displayed by DQ6. Bacterial TKT in the predicted mimotope area and insulin B were below detection levels in the case of DQ8. This is consistent with the properties of DQ8 peptidomes, which differ from DQ6 peptidomes in many physicochemical features[34] and bind insulin $B_{9–23}$ weakly.[35]

Discussion

Our results support a specific biological mechanism linking two major genetic and environmental risk factors in T1D etiology: HLA class II-TCR interactions[36] in the thymic selection process and mimicry of insulin epitopes in the gut microbial proteome leading to increases in the frequency of T cells that can recognise both insulin and bacterial peptides, in which the interacting components have co-evolved over millions of years. Immune cross-reactivity could strengthen tolerance to an essential, highly expressed self antigen, such as insulin, as well as providing protection against the host immune system to the bacteria that encode functional insulin mimotopes and also provide essential metabolic capacity to the host.

A determinant role of positive thymic selection during interactions between DR and DQ residues, notably DQβ1_{57}, and CDR3β, which alter the stability of HLA-epitope-TCR complexes, is supported by the presence of a similar bias in all three T cell subsets studied, Tconv, RTE, and Treg. These three subsets are under the same positive selection threshold in early thymocyte development. As a consequence of this bias, in case of diplotypes conferring protection, the frequency of amino acids with a negatively charged side chain in the epitope-binding CDR3β region is reduced. Conversely, in the presence of T1D-predisposing molecules such as DQ8, negatively charged CDR3β sequences are increased. These sequences have been observed to lead to highly immunogenic DQ-epitope-TCR complexes in the context of anti-gluten responses in celiac disease.[37]

Structural models suggest CDR3 loops contribute most of the energy when TCRs form complexes with HLA and peptides, with CDR1 and CDR2 having secondary roles.[38] These intermolecular interactions are dominated by electrostatic forces. Contact energy models predict phenylalanine (`F`), isoleucine (`I`), and valine (`V`) to be favorable interaction partners of aspartic acid (`D`).[39] Conversely, interactions between aspartic acid pairs, or aspartic (`D`) and glutamic acid (`E`), are particularly unfavorable owing to charge repulsion.

Previous T1D studies employing NOD mice have observed an enrichment of amino acids with negatively charged side chains in the CDR3β region of islet T cell infiltrates.[10,40] These sequences typically include aspartic acid preceded by glutamine (`QD`). Mutation of serine at I-Aβ_{57} (`S`→`D`) provides direct support for a role of this residue in the regulation of highly immunogenic responses against insulin.[10] Our results from the analyses of human repertoire samples extend our understanding of why negatively charged CDR3β sequences were less frequent in the presence of `D` at I-Aβ_{57} in this mouse model of T1D.

Other studies have reported autoreactive TCRs are associated with the presence of hydrophobic residues in the CDR3β region.[41,42] This is a consequence of such residues having a high average interaction potential with other amino acids, which leads to more stable HLA-epitope-TCR complexes and influences downstream TCR signaling. However, epitopes from insulin $B_{9–25}$ have an unusual amino acid composition, with negatively charged side chains in the N- and C-terminal regions, which suggests average trends may not be applicable to this particular case. We note that the only three available anti-insulin TCRs isolated

from the pancreas of children recently diagnosed with T1D lack hydrophobic doublets at positions 6–7 of the CDR3β region, which have been described as a common pattern in self-reactive TCR repertoires.

In addition, previous studies assessing TCR α chain patterns have reported an expansion of anti-insulin clones with a double glycine plus serine motif (GGS) in the CDR3α region.[26] These two sequence patterns coincide with some of the k-mers that exhibit the highest fold increases as a function of T1D OR in the present study (Figures 1 and S3), consistent with HLA class II-TCR interactions increasing the frequency of anti-insulin T cells in the periphery.

Position 57 at DQβ1 is also associated with other autoimmune diseases such as celiac disease, asthma, and polyglandular autoimmunity. Position 57 in the paralog DRβ1 molecule lies in a structurally analogous pocket to $DQ\beta1_{57}$, close to the epitope C-terminus and CDR3β. $DR\beta1_{57}$ has been recently fine-mapped as a strong association with several infectious and autoimmune disorders,[43] The existence of wide pleiotropy in these two genes, DRB1 and DQB1, is probably a consequence of the deep repertoire selection biases induced by the presence or absence of a negatively charged side chain at position 57, involving the thymic selection mechanism indicated by our results.

The existence of a higher proportion of anti-insulin TCRs in the periphery, driven solely by host genetics, is not sufficient to cause a loss of tolerance to insulin leading to a diagnosis of T1D. This is evidenced, for example, by the continuous increase of T1D incidence in multiple geographical locations that cannot be explained by changes in the gene pool, and suggests alterations to some environmental factors.[16]

Molecular mimicry of host antigens has been widely explored as a causal mechanism of common autoimmune diseases mediated by environmental factors. In the case of T1D, early efforts were mainly centered around enteroviral mimics of a primary islet autoantigen, glutamate decarboxylase 65 (GAD65).[44,45] Possible instances of insulin B chain mimicry by microbial antigens, with varying levels of support, have been reported. These include environmental bacteria that cross-react with human tetramer-sorted $CD4^+$ T cells,[46] a gut bacterial antigen that cross-reacts with a NOD $CD4^+$ clonotype,[47,48] and fungal sequences that cross-react with a human $CD8^+$ clonotype.[49] Other reported findings involve the insulin signal peptide[50] and defective ribosomal products (DRiPs) as a target antigen.[51]

The degeneracy of TCRs, meaning some clonotypes can cross-react with thousands or even millions of different peptides,[52] has led to questions about the validity of some evidence and the overall role of mimicry in disease pathogenesis. For example, islet-specific $CD8^+$ T cells have been reported to gain effector function from by-stander activation, and not molecular mimicry.[53] Generating high-grade evidence is further complicated by the paucity of islet-specific human clones available.[54]

By initially framing this question as a statistical sequence analysis problem, we were able to circumvent some of these issues, and reject the null hypothesis of a lack of evolutionary relatedness between more than 100 bacterial sequences and insulin $B_{9–25}$. This suggests symbiotic co-evolution between hosts and commensals,[55] and a role in the regulation of immune tolerance to insulin. We have previously reported analogous results using a Bayesian mixture model.[22]

We obtained functional evidence for this mimicry by demonstrating cross-reactivity to one of the only three identified anti-insulin DQ-restricted TCRs isolated so far from the pancreata of recently diagnosed T1D patients. Our findings draw parallels to other emerging autoimmune disease mechanisms driven by epitope mimicry, such as those observed in systemic lupus erythematosus, multiple sclerosis, rheumatoid arthritis, and narcolepsy.[56]

Some of the identified microbial commensals, which harbor sequences more similar to insulin $B_{9–25}$ than what would be expected by random chance, have been previously associated with T1D in a variety of longitudinal and cross-sectional studies. *Clostridium leptum*, which encoded the TKT mimotope that stimulated the TCR from an islet-resident T cell (Figure 4), was diminished in NOD mice that progressed to T1D.[57]

A fecal transplant trial in adults with active T1D halted the disease in some patients. *C. leptum* abundance increases were among the predictors of response to the transplant.[58] *C. leptum* also exhibited the strongest correlation with certain plant-based food diets, consisting of maltose, sucrose, starch, and other carbohydrate sources, in a large observational study that profiled gut microbiomes.[59]

TKT mimics contain a N-terminal transketolase domain and exhibit sequence similarity to a few members of the apulose-4-phosphate transketolase enzyme subfamily. These are involved in the catabolism of D-apiose, found in pectins, which could point to imbalanced plant-based nutrients in the development of gut dysbiosis during early life. Diets with pectins increased the ratio of Bacteroidota (Bacteroidetes) to Bacillota (Firmicutes) in the gut and promoted T1D in NOD mice.[60]

Bacterial TKT is among the most highly upregulated commensal enzymes during the first year of life,[20] probably to aid processing of incoming fiber and other carbohydrates from solid foods. Successive colonization waves, particularly at the time of weaning, which precedes the peak of insulin autoantibody positivity between 9 and 18 months of age, represent a period of physiological stress when immune tolerance to a myriad of antigens is facilitated by interactions with the microbiome.

Further studies are required to determine the role of insulin mimics in the maintenance of immune homeostasis and susceptibility to disease. Gut microbiome dysbiosis, increased gut permeability, and chronic low-grade inflammation, owing to a myriad of modern lifestyle factors,[61–63] may interfere with the role of commensal mimics in the induction of sufficient insulin-reactive Tregs.[64]

These environmental factors could also lead to an excessive growth of certain commensal species that encode insulin mimics causing the expansion of insulin-reactive Tconvs, the number of which is increased in carriers of susceptible HLA class II diplotypes, as shown here. These results and other associations between the microbiome and T1D have motivated a phase II clinical trial that is investigating whether a daily probiotic supplementation may reduce the incidence of islet autoimmunity.[23]

Acknowledgements

We thank all volunteers participating in this study. We are grateful to the study volunteers and staff associated with JDRF Diabetes, Genes, Autoimmunity and Prevention (D-GAP), including the Wellcome Clinical Research Facility (Addenbrooke's Clinical Research Centre, Cambridge, UK) and Investigating Underlying Causal Mechanisms in Type 1 Diabetes (DMech) hospital sites, including the Institute of Mother and Child (Warsaw, Poland). We thank Network for Pancreatic Organ Donors with Diabetes (nPOD) and their families.

We thank Aaron Michels (Barbara Davis Center for Childhood Diabetes, University of Colorado School of Medicine, Aurora, Colorado, USA) for helpful discussions and for providing EBV cell lines expressing various HLA class II molecules, and Wojciech Szypowski (Polish Society for Autoimmune Diseases, Warsaw, Poland) for sample logistics. 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 JDRF/Wellcome Diabetes and Inflammation Laboratory (DIL, 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 Raqeeb 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; Olga Platonova for managing finance and funding; Moustafa Attar (Oxford Genomics Centre, University of Oxford) for assistance with single cell sequencing; and Holly Roach for contributions to the manuscript.

This research was performed with the support of 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 https://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 Dr Robert Goldstein MD (1941–2018), Professor David B. Dunger FRCP FMedSci (1948–2021), and Professor Hugh O. McDevitt MD ForMemRS (1930–2022).

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 class II diplotypes.

RF, MLP, AP, and ML performed additional single-cell BD 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 contributed to the planning of the study and discussed the results. CLS was responsible for ethics and sponsor approvals. All authors approved the final manuscript.

Declaration of interests

JAT consults for GSK, Immunocore, Vesalius, Avammune Therapeutics, and Precion. Other authors declare no competing interests.

Data and code availability

All raw data and source code will be deposited in public repositories ahead of publication following FAIR guidelines. Pre-processed data will also be made available to ease reproducibility.

References

[1] American Diabetes Association Professional Practice Committee, 2. Classification and diagnosis of diabetes: Standards of care in diabetes—2024, Diabetes Care 47 (2024) S20–S42.

[2] A.-G. Ziegler, E. Bonifacio, Babydiab-Babydiet Study Group, Age-related islet autoantibody incidence in offspring of patients with type 1 diabetes, Diabetologia 55 (2012) 1937–1943.

[3] J. Ilonen, A. Hammais, A.-P. Laine, J. Lempainen, O. Vaarala, R. Veijola, O. Simell, M. Knip, Patterns of β-cell autoantibody appearance and genetic associations during the first years of life, Diabetes 62 (2013) 3636–3640.

[4] Z. Al Nabhani, S. Dulauroy, R. Marques, C. Cousu, S. Al Bounny, F. Déjardin, T. Sparwasser, M. Bérard, N. Cerf-Bensussan, G. Eberl, A weaning reaction to microbiota is required for resistance to immunopathologies in the adult, Immunity 50 (2019) 1276–1288.

[5] H. Brown, M. R. Komnick, P. H. Brigleb, T. S. Dermody, D. Esterházy, Lymph node sharing between pancreas, gut, and liver leads to immune crosstalk and regulation of pancreatic autoimmunity, Immunity 56 (2023) 2070–2085.

[6] X. Hu, A. J. Deutsch, T. L. Lenz, S. Onengut-Gumuscu, B. Han, W.-M. Chen, J. M. Howson, J. A. Todd, P. I. De Bakker, S. S. Rich, 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 (2015) 898–905.

[7] S. Nejentsev, J. M. Howson, N. M. Walker, J. Szeszko, S. F. Field, H. E. Stevens, P. Reynolds, M. Hardy, E. King, J. Masters, et al., Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A, Nature 450 (2007) 887–892.

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

[9] H. Acha-Orbea, H. O. McDevitt, The first external domain of the nonobese diabetic mouse class II IA beta chain is unique, Proc. National Acad. Sci. 84 (1987) 2435–2439.

[10] L. Gioia, M. Holt, A. Costanzo, S. Sharma, B. Abe, L. Kain, M. Nakayama, X. Wan, A. Su, C. Mathews, 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) eaaw6329.

[11] P. Vafiadis, S. T. Bennett, J. A. Todd, J. Nadeau, R. Grabs, C. G. Goodyer, S. Wickramasinghe, E. Colle, C. Polychronakos, Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus, Nat. Genet. 15 (1997) 289–292.

[12] A. Pugliese, M. Zeller, A. Fernandez Jr, L. J. Zalcberg, R. J. Bartlett, C. Ricordi, M. Pietropaolo, G. S. Eisenbarth, S. T. Bennett, D. D. Patel, 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 (1997) 293–297.

[13] I. Durinovic-Belló, R. Wu, V. Gersuk, S. Sanda, H. Shilling, G. Nepom, Insulin gene VNTR genotype associates with frequency and phenotype of the autoimmune response to proinsulin, Genes & Immun. 11 (2010) 188–193.

[14] M. Nakayama, N. Abiru, H. Moriyama, N. Babaya, E. Liu, D. Miao, L. Yu, D. R. Wegmann, J. C. Hutton, J. F. Elliott, et al., Prime role for an insulin epitope in the development of type 1 diabetes in NOD mice, Nature 435 (2005) 220–223.

[15] X. Wan, A. N. Vomund, O. J. Peterson, A. V. Chervonsky, C. F. Lichti, E. R. Unanue, The MHC-II peptidome of pancreatic islets identifies key features of autoimmune peptides, Nat. Immunol. 21 (2020) 455–463.

[16] K. M. Gillespie, S. C. Bain, A. H. Barnett, P. J. Bingley, M. R. Christie, G. V. Gill, E. A. Gale, The rising incidence of childhood type 1 diabetes and reduced contribution of high-risk HLA haplotypes, Lancet 364 (2004) 1699–1700.

[17] A. D. Kostic, D. Gevers, H. Siljander, T. Vatanen, T. Hyötyläinen, A.-M. Hämäläinen, A. Peet, V. Tillmann, P. Pöhö, I. Mattila, et al., The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes, Cell Host & Microbe 17 (2015) 260–273.

[18] A. G. Davis-Richardson, A. N. Ardissone, R. Dias, V. Simell, M. T. Leonard, K. M. Kemppainen, J. C. Drew, D. Schatz, M. A. Atkinson, B. Kolaczkowski, et al., *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes, Front. Microbiol. 5 (2014) 678.

[19] M. Bélteky, P. L. Milletich, A. P. Ahrens, E. W. Triplett, J. Ludvigsson, Infant gut microbiome composition correlated with type 1 diabetes acquisition in the general population: the ABIS study, Diabetologia 66 (2023) 1116–1128.

[20] T. Vatanen, E. A. Franzosa, R. Schwager, S. Tripathi, T. D. Arthur, K. Vehik, Å. Lernmark, W. A. Hagopian, M. J. Rewers, J.-X. She, et al., The human gut microbiome in early-onset type 1 diabetes from the TEDDY study, Nature 562 (2018) 589–594.

[21] F. Guarner, R. Bourdet-Sicard, P. Brandtzaeg, H. S. Gill, P. McGuirk, W. Van Eden, J. Versalovic, J. V. Weinstock, G. A. Rook, Mechanisms of disease: the hygiene hypothesis revisited, Nat. Rev. Gastroenterol. & Hepatol. 3 (2006) 275–284.

[22] A. Rubio García, A. Paterou, M. Lee, H. Slawinski, L. S. Wicker, J. A. Todd, M. L. Pekalski, Peripheral tolerance to insulin is encoded by mimicry in the microbiome, bioRxiv (2019) 2019.12.18.881433.

[23] A.-G. Ziegler, S. Arnolds, A. Kölln, P. Achenbach, R. Berner, E. Bonifacio, K. Casteels, H. E. Larsson, M. Gündert, J. Hasford, et al., Supplementation with *Bifidobacterium longum* subspecies *infantis* EVC001 for mitigation of type 1 diabetes autoimmunity: the GPPAD-SINT1A randomised controlled trial protocol, BMJ Open 11 (2021) e052449.

[24] S. A. Sharp, S. S. Rich, A. R. Wood, S. E. Jones, R. N. Beaumont, J. W. Harrison, D. A. Schneider, J. M. Locke, J. Tyrrell, M. N. Weedon, et al., Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis, Diabetes Care 42 (2019) 200–207.

[25] M. L. Pekalski, A. Rubio García, R. C. Ferreira, D. B. Rainbow, D. J. Smyth, M. Mashar, J. Brady, N. Savinykh, X. Castro Dopico, S. Mahmood, et al., Neonatal and adult recent thymic emigrants produce IL-8 and express complement receptors CR1 and CR2, JCI Insight 2 (2017) e93739.

[26] E. Simone, D. Daniel, N. Schloot, P. Gottlieb, S. Babu, E. Kawasaki, D. Wegmann, G. Eisenbarth, T cell receptor restriction of diabetogenic autoimmune NOD T cells, Proc. National Acad. Sci. 94 (1997) 2518–2521.

[27] A. Christophersen, E. G. Lund, O. Snir, E. Solà, C. Kanduri, S. Dahal-Koirala, S. Zühlke, Ø. Molberg, P. J. Utz, M. Rohani-Pichavant, et al., Distinct phenotype of CD4+ T cells driving celiac disease identified in multiple autoimmune conditions, Nat. Med. 25 (2019) 734–737.

[28] M. Campbell-Thompson, C. Wasserfall, J. Kaddis, A. Albanese-O'Neill, T. Staeva, C. Nierras, J. Moraski, P. Rowe, R. Gianani, G. Eisenbarth, et al., Network for Pancreatic Organ Donors with Diabetes (nPOD): developing a tissue biobank for type 1 diabetes, Diabetes / Metab. Res. Rev. 28 (2012) 608–617.

[29] L. G. Landry, A. M. Anderson, H. A. Russ, L. Yu, S. C. Kent, M. A. Atkinson, C. E. Mathews, A. W. Michels, M. Nakayama, Proinsulin-reactive CD4 T cells in the islets of type 1 diabetes organ donors, Front. Endocrinol. 12 (2021).

[30] S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. National Acad. Sci. 89 (1992) 10915–10919.

[31] L. Richardson, B. Allen, G. Baldi, M. Beracochea, M. L. Bileschi, T. Burdett, J. Burgin, J. Caballero-Pérez, G. Cochrane, L. J. Colwell, et al., MGnify: the microbiome sequence data analysis resource in 2023, Nucleic Acids Res. 51 (2023) D753–D759.

[32] S. F. Altschul, R. Bundschuh, R. Olsen, T. Hwa, The estimation of statistical parameters for local alignment score distributions, Nucleic Acids Res. 29 (2001) 351–361.

[33] A. W. Michels, L. G. Landry, K. A. McDaniel, L. Yu, M. Campbell-Thompson, W. W. Kwok, K. L. Jones, P. A. Gottlieb, J. W. Kappler, Q. Tang, et al., Islet-derived CD4 T cells targeting proinsulin in human autoimmune diabetes, Diabetes 66 (2017) 722–734.

[34] B. Reynisson, C. Barra, S. Kaabinejadian, W. H. Hildebrand, B. Peters, M. Nielsen, Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data, J. Proteome Res. 19 (2020) 2304–2315.

[35] M. T. Tran, P. Faridi, J. J. Lim, Y. T. Ting, G. Onwukwe, P. Bhattacharjee, C. M. Jones, E. Tresoldi, F. J. Cameron, N. L. La Gruta, et al., T cell receptor recognition of hybrid insulin peptides bound to HLA-DQ8, Nat. Commun. 12 (2021) 5110.

[36] M. L. Bettini, M. Bettini, Understanding autoimmune diabetes through the prism of the tri-molecular complex, Front. Endocrinol. 8 (2017) 299527.

[37] Z. Hovhannisyan, A. Weiss, A. Martin, M. Wiesner, S. Tollefsen, K. Yoshida, C. Ciszewski, S. A. Curran, J. A. Murray, C. S. David, et al., The role of HLA-DQ8 β57 polymorphism in the anti-gluten T-cell response in coeliac disease, Nature 456 (2008) 534–538.

[38] N. A. Borg, L. K. Ely, T. Beddoe, W. A. Macdonald, H. H. Reid, C. S. Clements, A. W. Purcell, L. Kjer-Nielsen, J. J. Miles, S. R. Burrows, et al., The CDR3 regions of an immunodominant T cell receptor dictate the 'energetic landscape' of peptide-MHC recognition, Nat. Immunol. 6 (2005) 171–180.

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

[40] F. J. Baker, M. Lee, Y.-h. Chien, M. M. Davis, Restricted islet-cell reactive T cell repertoire of early pancreatic islet infiltrates in NOD mice, Proc. National Acad. Sci. 99 (2002) 9374–9379.

[41] B. D. Stadinski, K. Shekhar, I. Gómez-Touriño, J. Jung, K. Sasaki, A. K. Sewell, M. Peakman, A. K. Chakraborty, E. S. Huseby, Hydrophobic CDR3 residues promote the development of self-reactive T cells, Nat. Immunol. 17 (2016) 946–955.

[42] B. D. Stadinski, S. B. Cleveland, M. A. Brehm, D. L. Greiner, P. G. Huseby, E. S. Huseby, I-Ag7 β56/57 polymorphisms regulate non-cognate negative selection to CD4+ T cell orchestrators of type 1 diabetes, Nat. Immunol. 24 (2023) 652–663.

[43] S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T. Konuma, K. Yamamoto, M. Akiyama, et al., A cross-population atlas of genetic associations for 220 human phenotypes, Nat. Genet. 53 (2021) 1415–1424.

[44] D. L. Kaufman, M. Erlander, M. Clare-Salzler, M. Atkinson, N. Maclaren, A. Tobin, Autoimmunity to two forms of glutamate decarboxylase in insulin-dependent diabetes mellitus, J. Clin. Investig. 89 (1992) 283–292.

[45] G. R. Vreugdenhil, A. Geluk, T. Ottenhoff, W. Melchers, B. Roep, J. Galama, Molecular mimicry in diabetes mellitus: the homologous domain in coxsackie B virus protein 2C and islet autoantigen GAD65 is highly conserved in the coxsackie B-like enteroviruses and binds to the diabetes associated HLA-DR3 molecule, Diabetologia 41 (1998) 40–46.

[46] J. Yang, I.-T. Chow, T. Sosinowski, N. Torres-Chinn, C. J. Greenbaum, E. A. James, J. W. Kappler, H. W. Davidson, W. W. Kwok, Autoreactive T cells specific for insulin B:11–23 recognize a low-affinity peptide register in human subjects with autoimmune diabetes, Proc. National Acad. Sci. 111 (2014) 14840–14845.

[47] Q. Huang, I.-T. Chow, C. Brady, A. Raisingani, D. Li, D. A. Ostrov, M. A. Atkinson, W. W. Kwok, C. R. Kahn, E. Altindis, *Parabacteroides distasonis* enhances type 1 diabetes autoimmunity via molecular mimicry, bioRxiv (2020) 2020.10.22.350801.

[48] K. Girdhar, Q. Huang, I.-T. Chow, T. Vatanen, C. Brady, A. Raisingani, P. Autissier, M. A. Atkinson, W. W. Kwok, C. R. Kahn, et al., A gut microbial peptide and molecular mimicry in the pathogenesis of type 1 diabetes, Proc. National Acad. Sci. 119 (2022) e2120028119.

[49] T. Whalley, G. Dolton, P. E. Brown, A. Wall, L. Wooldridge, H. Van den Berg, A. Fuller, J. R. Hopkins, M. D. Crowther, M. Attaf, et al., GPU-accelerated discovery of pathogen-derived molecular mimics of a T-cell insulin epitope, Front. Immunol. 11 (2020) 296.

[50] D. K. Cole, A. M. Bulek, G. Dolton, A. J. Schauenberg, B. Szomolay, W. Rittase, A. Trimby, P. Jothikumar, A. Fuller, A. Skowera, et al., Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity, J. Clin. Investig. 126 (2016) 2191–2204.

[51] K.-Y. Ma, A. A. Schonnesen, C. He, A. Y. Xia, E. Sun, E. Chen, K. R. Sebastian, Y.-W. Guo, R. Balderas, M. Kulkarni-Date, et al., High-throughput and high-dimensional single-cell analysis of antigen-specific CD8+ T cells, Nat. Immunol. 22 (2021) 1590–1598.

[52] L. Wooldridge, J. Ekeruche-Makinde, H. a. Van Den Berg, A. Skowera, J. J. Miles, M. P. Tan, G. Dolton, M. Clement, S. Llewellyn-Lacey, D. A. Price, et al., A single autoimmune T cell receptor recognizes more than a million different peptides, J. Biol. Chem. 287 (2012) 1168–1177.

[53] M. Okada, V. Zhang, J. D. Loaiza Naranjo, B. J. Tillett, F. S. Wong, R. J. Steptoe, A.-S. Bergot, E. E. Hamilton-Williams, Islet-specific CD8+ T cells gain effector function in the gut lymphoid tissues via bystander activation not molecular mimicry, Immunol. Cell Biol. 101 (2023) 36–48.

[54] E. A. James, R. Mallone, S. C. Kent, T. P. DiLorenzo, T-cell epitopes and

neo-epitopes in type 1 diabetes: a comprehensive update and reappraisal, Diabetes 69 (2020) 1311–1335.
[55] G. Rook, F. Bäckhed, B. R. Levin, M. J. McFall-Ngai, A. R. McLean, Evolution, human-microbe interactions, and life history plasticity, Lancet 390 (2017) 521–530.
[56] W. E. Ruff, T. M. Greiling, M. A. Kriegel, Host–microbiota interactions in immune-mediated diseases, Nat. Rev. Microbiol. 18 (2020) 521–538.
[57] M. Sysi-Aho, A. Ermolov, P. V. Gopalacharyulu, A. Tripathi, T. Seppänen-Laakso, J. Maukonen, I. Mattila, S. T. Ruohonen, L. Vähätalo, L. Yetukuri, et al., Metabolic regulation in progression to autoimmune diabetes, PLoS Comput. Biol. 7 (2011) e1002257.
[58] P. de Groot, T. Nikolic, S. Pellegrini, V. Sordi, S. Imangaliyev, E. Rampanelli, N. Hanssen, I. Attaye, G. Bakker, G. Duinkerken, et al., Faecal microbiota transplantation halts progression of human new-onset type 1 diabetes in a randomised controlled trial, Gut 70 (2021) 92–105.
[59] F. Asnicar, S. E. Berry, A. M. Valdes, L. H. Nguyen, G. Piccinno, D. A. Drew, E. Leeming, R. Gibson, C. Le Roy, H. A. Khatib, et al., Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals, Nat. Med. 27 (2021) 321–332.
[60] R. K. Toivonen, R. Emani, E. Munukka, A. Rintala, A. Laiho, S. Pietilä, J.-P. Pursiheimo, P. Soidinsalo, M. Linhala, E. Eerola, et al., Fermentable fibres condition colon microbiota and promote diabetogenesis in NOD mice, Diabetologia 57 (2014) 2183–2192.
[61] H. K. Åkerblom, O. Vaarala, H. Hyöty, J. Ilonen, M. Knip, Environmental factors in the etiology of type 1 diabetes, Am. J. Med. Genet. 115 (2002) 18–29.
[62] M. Rewers, J. Ludvigsson, Environmental risk factors for type 1 diabetes, Lancet 387 (2016) 2340–2348.
[63] T. G. Richardson, D. J. Crouch, G. M. Power, F. Morales-Berstein, E. Hazelwood, S. Fang, Y. Cho, J. R. Inshaw, C. C. Robertson, C. Sidore, et al., Childhood body size directly increases type 1 diabetes risk based on a lifecourse Mendelian randomization approach, Nat. Commun. 13 (2022) 2337.
[64] K. Atarashi, T. Tanoue, T. Shima, A. Imaoka, T. Kuwahara, Y. Momose, G. Cheng, S. Yamasaki, T. Saito, Y. Ohba, et al., Induction of colonic regulatory T cells by indigenous *Clostridium* species, Science 331 (2011) 337–341.
[65] X. Zheng, J. Shen, C. Cox, J. C. Wakefield, M. G. Ehm, M. R. Nelson, B. S. Weir, HIBAG—HLA genotype imputation with attribute bagging, Pharmacogenomics J. 14 (2014) 192–200.
[66] D. Trzupek, M. Dunstan, A. J. Cutler, M. Lee, L. Godfrey, L. Jarvis, D. B. Rainbow, D. Aschenbrenner, J. L. Jones, H. H. Uhlig, et al., Discovery of CD80 and CD86 as recent activation markers on regulatory T cells by protein-RNA single-cell analysis, Genome Med. 12 (2020) 1–22.
[67] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, et al., An integrated catalog of reference genes in the human gut microbiome, Nat. Biotechnol. 32 (2014) 834–841.
[68] A. W. Purcell, S. H. Ramarathinam, N. Ternette, Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics, Nat. Protoc. 14 (2019) 1687–1707.
[69] A. M. Anderson, L. G. Landry, A. A. Alkanani, L. Pyle, A. C. Powers, M. A. Atkinson, C. E. Mathews, B. O. Roep, A. W. Michels, M. Nakayama, Human islet T cells are highly reactive to preproinsulin in type 1 diabetes, Proc. National Acad. Sci. 118 (2021) e2107208118.
[70] S. E. Mann, Z. Zhou, L. G. Landry, A. M. Anderson, A. K. Alkanani, J. Fischer, M. Peakman, R. Mallone, K. Campbell, A. W. Michels, et al., Multiplex T cell stimulation assay utilizing a T cell activation reporter-based detection system, Front. Immunol. 11 (2020) 633.
[71] K. Polański, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, J.-E. Park, BBKNN: fast batch alignment of single cell transcriptomes, Bioinformatics 36 (2020) 964–965.
[72] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nat. Biotechnol. 37 (2019) 38–44.
[73] V. A. Traag, L. Waltman, N. J. Van Eck, From Louvain to Leiden: guaranteeing well-connected communities, Sci. Reports 9 (2019) 5233.
[74] V. Ntranos, L. Yi, P. Melsted, L. Pachter, A discriminative learning approach to differential expression analysis for single-cell RNA-seq, Nat. Methods 16 (2019) 163–166.
[75] M. D. Hoffman, A. Gelman, The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, J. Mach. Learn. Res. 15 (2014) 1593–1623.
[76] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: A probabilistic programming language, J. Stat. Softw. 76 (2017) 1–32.
[77] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, P.-C. Bürkner, Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC, arXiv (2019).
[78] X.-L. Meng, Posterior predictive *p*-values, Ann. Stat. 22 (1994) 1142–1160.
[79] A. Rubio García, Causal Mechanisms in Type 1 Diabetes, Ph.D. thesis, University of Oxford, 2022.
[80] T. F. Smith, M. S. Waterman, et al., Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[81] L. Zhang, B. A. Shaby, Uniqueness and global optimality of the maximum likelihood estimator for the generalized extreme value distribution, Biometrika 109 (2022) 853–864.
[82] J. Nocedal, A. Wächter, R. A. Waltz, Adaptive barrier update strategies for nonlinear interior methods, SIAM J. on Optim. 19 (2009) 1674–1693.
[83] T. D. Schneider, R. M. Stephens, Sequence logos: a new way to display consensus sequences, Nucleic Acids Res. 18 (1990) 6097–6100.

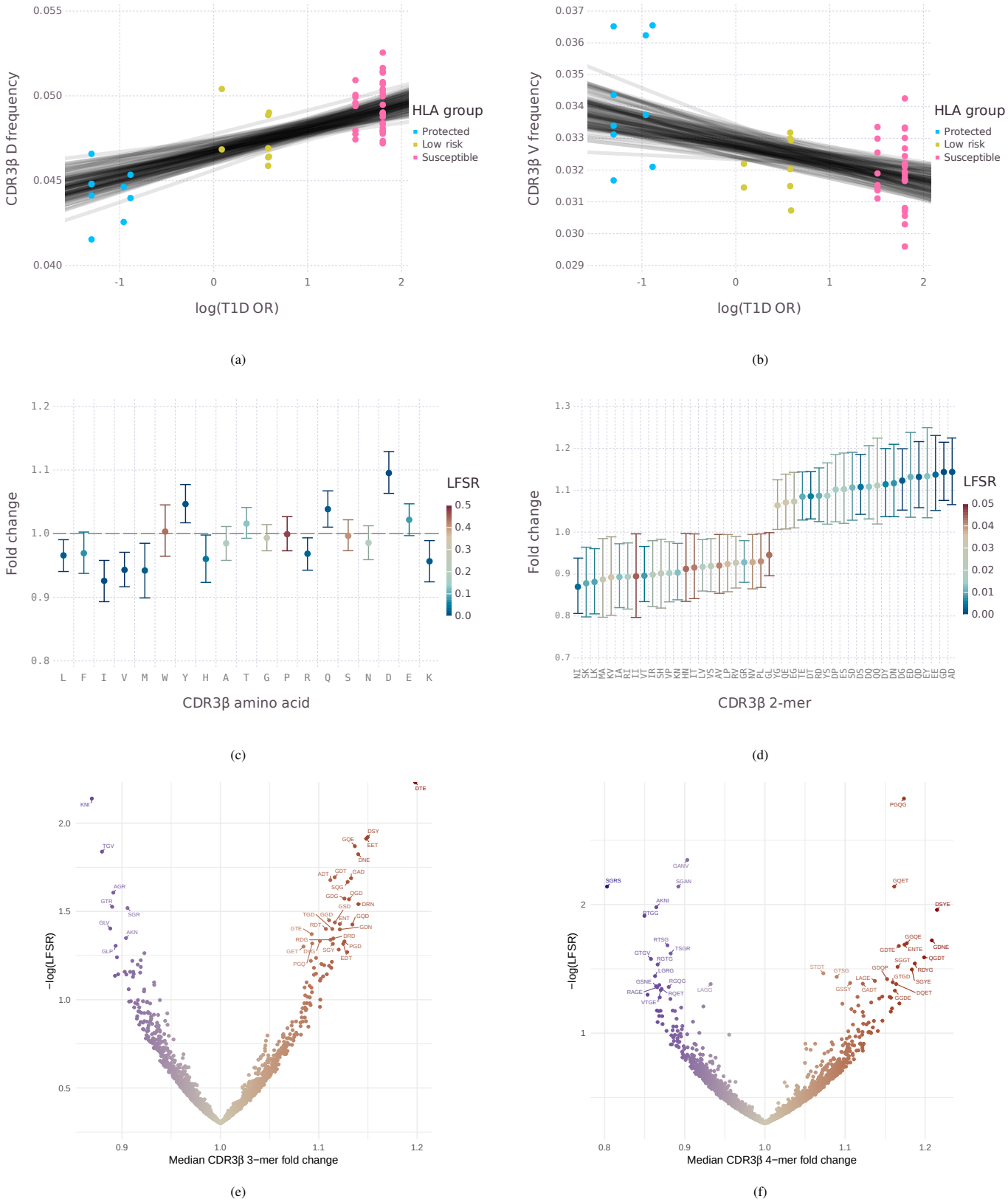
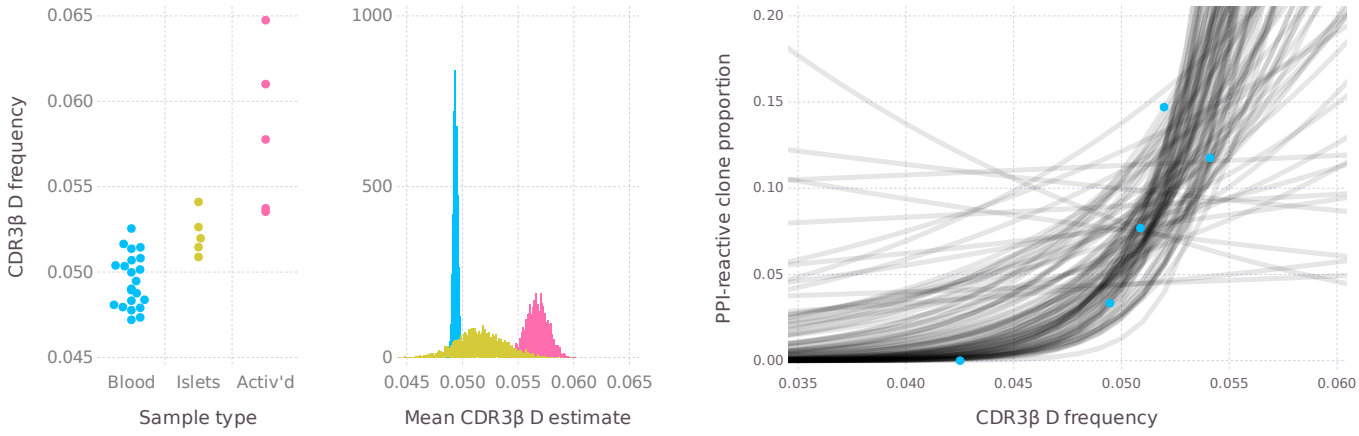


Figure 1: Estimates of T1D odds ratio effects, contributed by HLA DR-DQ diplotypes, on CDR3β k-mers from CD4$^+$ T conventional cells. Regression lines represent 100 random draws from the posterior distribution. Error bars summarize the effects on multiple k-mers by depicting the median and 90% credible intervals of fold changes across HLA class II extremes, with the most protective DQ6 diplotypes as baseline (Table S1). Funnel plots only represent the median and the local false sign rate (LFSR). (a) Aspartic acid. (b) Valine. (c) All amino acids, or 1-mers, except cysteine. (d) 2-mers, LFSR ≤ 0.05. (e) All 3-mers with non-zero counts. 3-mers on the y axis boundary have an estimated LFSR ≈ 0. (f) All 4-mers with non-zero counts.



(a) (b)

Figure 2: Aspartic acid (D) and anti-insulin T cell receptors. Each point denotes the summary statistics of all clonotypes in the repertoire of one donor. (a) CDR3β aspartic acid observed frequencies and posterior mean frequency estimates in $CD4^+$ T cells from peripheral blood, infiltrating islets, and activated circulating cells. (b) Preproinsulin-reactive (PPI) clone proportion measured in islets as a function of CDR3β aspartic acid frequency in infiltrating $CD4^+$ T cells. Regression lines represent 100 random draws from the posterior distribution.

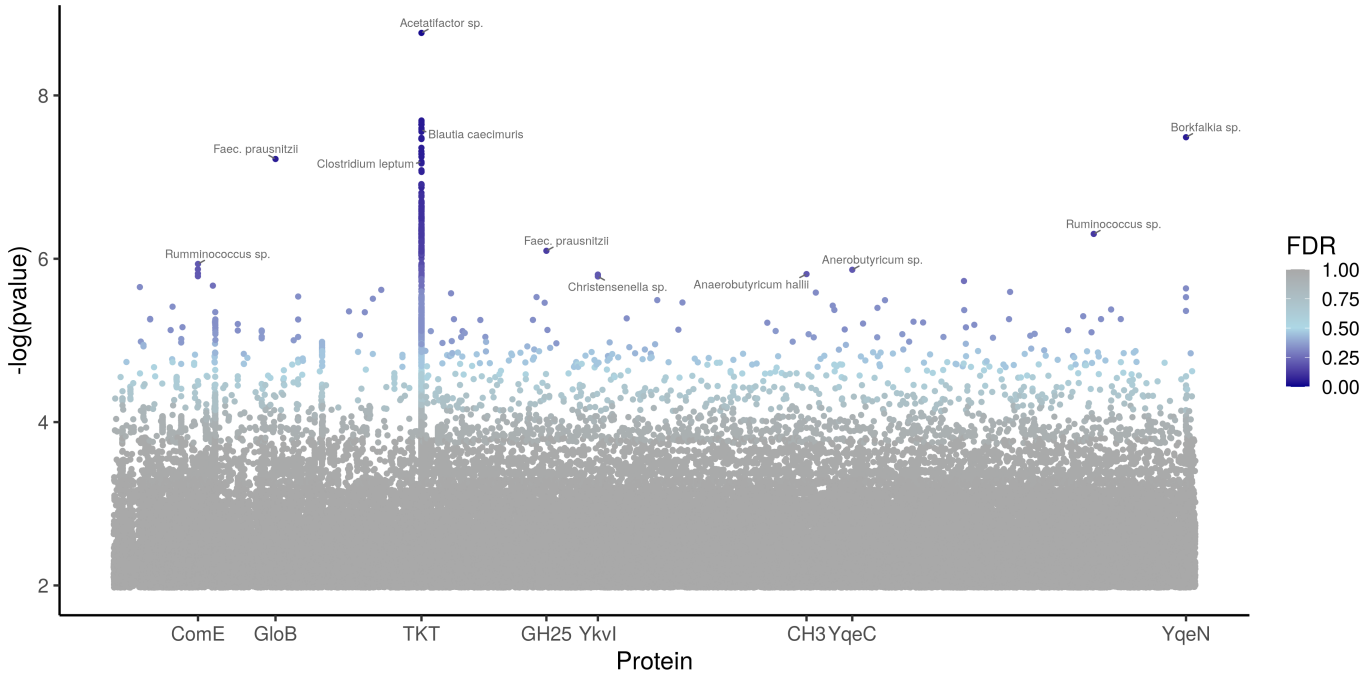


Figure 3: Gut microbial proteome-wide similarity with insulin B_{9-25}. Each point denotes one gut microbiome protein represented against a *p*-value derived by contrasting its maximal local pairwise alignment score with insulin B_{9-25} versus a null distribution of alignment scores. Color indicates the estimated probability of false discoveries (FDR) when considering all proteins with a *p*-value lower or equal than a given threshold. Points high in the y axis denote proteins whose similarity to insulin B_{9-25} deviates from what would be expected by random chance (Table S8). Protein domains and superfamilies listed include N-terminal transketolase (TKT), yqeN DNA replication protein (YqeN), hydroxyacylglutathione hydrolase (GloB), glycosyl hydrolase family 25 (GH25), uncharacterized membrane protein (YkvI), late competence operon (ComE), selenium cofactor biosynthesis protein (YqeC), and cyclases/histidine kinases associated sensory extracellular domain 3 (CH3).

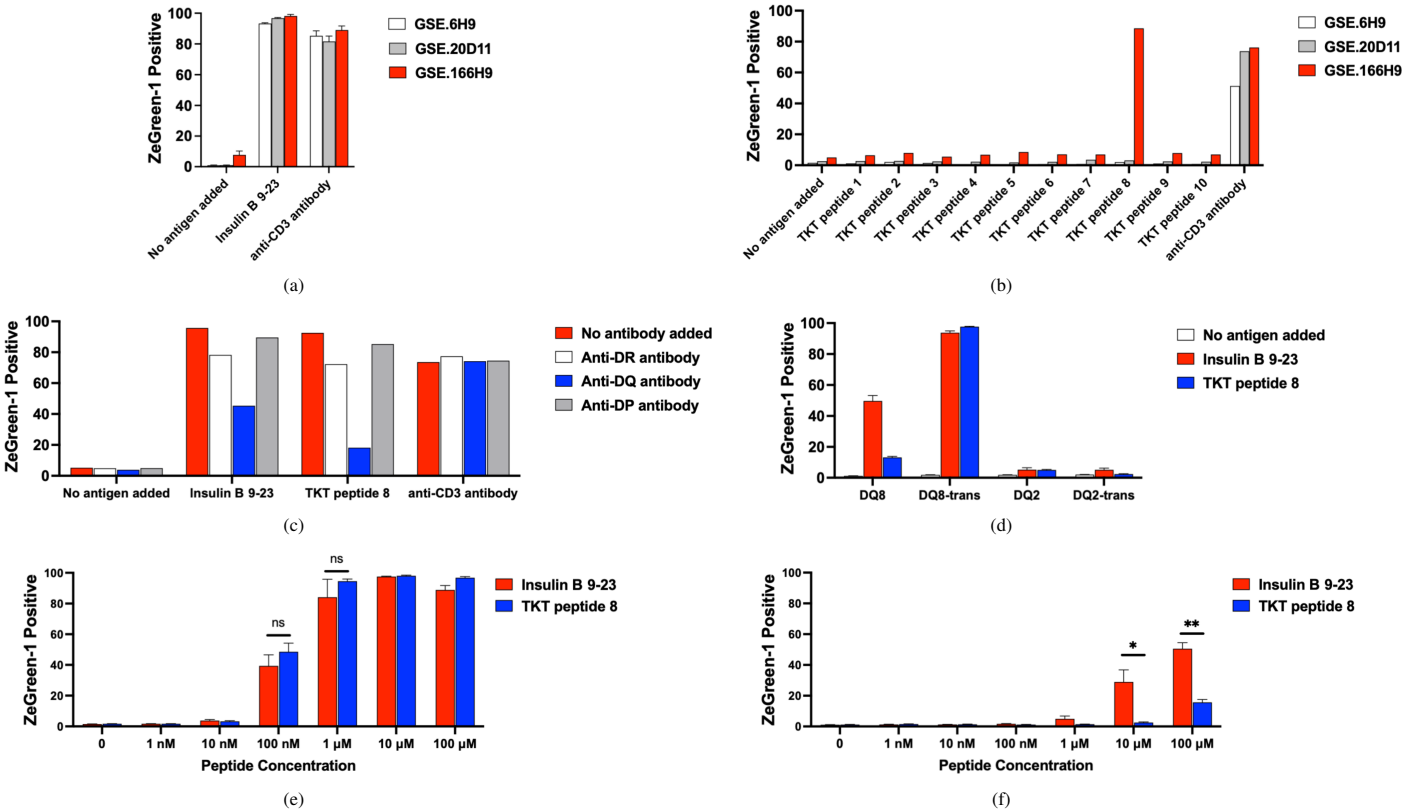


Figure 4: CD4$^+$ T cell stimulation assays on TCRs isolated from the pancreas of T1D patients from the Network of Pancreatic Organ Donors (nPOD) and validated to recognize insulin $B_{9\text{–}23}$. Peptide 8, which stimulates nPOD TCR GSE.166H9 (Table S9), corresponds to transketolase (TKT) from *Clostridium leptum* (Table S10). Measurements of stimulation on TCRs were obtained from 5KC T-hybridoma cells that were devoid of endogenous TCR expression and have been engineered with a ZsGreen-1 reporter gene. (a) Insulin $B_{9\text{–}23}$ reactivity. (b) TKT epitope scan. (c) HLA class II antibodies on TCR GSE.166H9. (d) DQ presentation assay on TCR GSE.166H9. Stimulation dependent on different peptide concentrations and (e) DQ2-DQ8$_{trans}$ or (f) DQ8 molecules. Experiments were repeated 2 (a), 3 (d), or 4 times (e and f). Mean and standard error are shown. ns: $P > 0.05$, *: $P \leq 0.05$, and **: $P \leq 0.01$ on a two-tailed paired t-test.

Methods

Method details

Circulating CD4$^+$ T cell repertoires

Donor selection. Donors were sourced from the JDRF Diabetes, Genes, Autoimmunity, and Prevention (D-GAP) cohort. Donors were selected to cover the whole protection-susceptibility risk continuum (Table S1), according to T1D risk contributed by HLA DR-DQ diplotypes.[24] London Hampstead Research Committee of the NHS Health Research Authority gave ethical approval for this work, ref. no. 08/H0720/25. Upon closure of the D-GAP study, samples were transferred to Investigating Genes and Phenotypes of Type 1 Diabetes, Cambridgeshire 2 Research Ethics Committee, ref. no. 08/H0308/153.

Genotyping. Genomic DNA was prepared from PBMCs or whole blood using QiaAmp DNA Blood kit (Qiagen), or phenol-chloroform extraction. Haplotypes were measured using Taqman sequencing (Applied Biosystems) of four SNPs (rs2187668, rs660895, rs9271366, and rs7454108), complemented with 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,[65] along with further SSP classical HLA class II typing (MC Diagnostics and Oxford Transplant Centre).

PBMC processing. PBMC isolation, cryopreservation, and thawing were performed as previously described.[25] PBMCs were isolated using Lympholyte (CEDARLANE). PBMCs were cryopreserved in heat-inactivated filtered human AB serum (Sigma-Aldrich) and 10% DMSO Hybri-MAX (Sigma-Aldrich) at a concentration of 2–10 $\times$ 10^6 / ml, and were stored in liquid nitrogen. PBMCs were thawed in a 37 °C water bath for 2 min, and then washed by adding 1 ml of AB serum to cells, followed by adding 10 ml of cold (4 °C) X-VIVO (Lonza), containing 10% AB serum per up to 1 $\times$ 10^7 cells, both in a drop-wise fashion. PBMCs were then washed again with 10 ml of cold (4 °C) X-VIVO containing 1% AB serum per 1 $\times$ 10^7 cells.

CD4$^+$ cell purification. CD4$^+$ T cells were purified from thawed PBMCs using negative selection with EasySep Human CD4$^+$ T Cell Enrichment Cocktail (STEMCELL Technologies), then washed with X-VIVO medium (Lonza) containing 5% human AB serum, and resuspended at 1 $\times$ 10^5 cells / well (96 well plate) in a final volume of 200 µl X-VIVO medium (Lonza) containing 5% human AB serum. Cells were activated with PMA/ionomycin (eBioscience) for 2 h, then harvested, washed, resuspended in PBS, and counted.

Single-cell sequencing. CD4$^+$ T cells were washed in PBS with 0.04% BSA and re-suspended them at a concentration of 800–1200 cells / µl, before capturing ~5 $\times$ 10^3 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 2 $\times$ 10^4 read pairs / cell for gene expression libraries, and 5 $\times$ 10^3 read pairs / cell for TCR libraries.

Single-cell data preprocessing. Single-cell RNA and TCR sequencing libraries were processed separately using Cell Ranger v4.0.0 (10x Genomics) to obtain gene counts and receptor assemblies. These were then merged into a single gene expression matrix and a single TCR database, which mapped gene counts and TCR chains to unique molecular identifiers and donor barcodes.

Circulating activated CD4$^+$ T cell repertoires

Donor selection. Donors were sourced from the study Investigating Underlying Causal Mechanisms in Type 1 Diabetes (DMech). Donors were selected among those that had been recently diagnosed and carried a susceptible HLA DR-DQ diplotype (Table S6), conferring a T1D OR > 0.9 in $\log_{10}$ scale. Donors were genotyped as described in the previous section. South Central Oxford A Research Ethics committee of the NHS Health Research Authority gave ethical approval for this work, ref. no. 18/SC/0559.

CD4$^+$ cell purification. Fresh whole-blood samples from newly diagnosed T1D patients were shipped overnight at room temperature. 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 plus 2% FBS.

Fluorescence-activated cell sorting. Cells were 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 min at 4 °C, cells were washed two times and resuspended in PBS + 1% FBS for cell sorting, at 4 °C in a BD FACSAria Fusion sorter (BD Biosciences). CD3$^+$ CD4$^+$ HLA$^-$ DR$^+$ CD38$^+$, control pools of CD3$^+$ CD4$^+$ HLA-DR$^-$, and CD3$^+$ CD4$^+$ HLA-DR$^+$ CD38$^-$ were FACS-purified (2–3.5 $\times$ 10^4 cells / donor), washed, and processed for BD Rhapsody Express Single-Cell analysis (BD Biosciences). FACS-sorted cells were incubated with different oligo-conjugated sample barcoding antibodies (Sample Multiplexing Kit, BD Biosciences) for 20 min on ice.

Cell incubation. Barcoded cells from each batch of donors (total three batches) were then pooled together in a 5 ml FACS tube (Falcon) and washed in cold PBS + 2% FBS. Cell pools were then incubated for 5 min at 4 °C with Human Fc block (BD Biosciences) and then immediately incubated with a mastermix of 62 oligo-conjugated AbSeq antibodies (BD Biosciences)[66] for 45 min 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 2×10^4 cells / pool.

Library preparation. mRNA, TCR, AbSeq, and sample tag cDNA products were initially amplified for 11 cycles (PCR1). 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 (~250 bp) products from the longer mRNA (350–800 bp) and TCR (600–1000 bp) products. The purified mRNA (10 cycles), 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 normalized to 2.5 ng / µl (mRNA), 1 ng / µl (sample tag and AbSeq), and 0.5 ng / µl (TCR). These underwent a final round of amplification (6 cycles for mRNA, sample tag, and AbSeq; 7 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% mRNA : 14% TCR : 57% AbSeq : 4% sample tag ratio) to achieve a concentration of 5 nM.

Single-cell sequencing. 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 an additional 306 primer pairs), as previously described.[66] Final pooled libraries were spiked with 15% PhiX control DNA to increase sequence complexity and were sequenced (75 bp, 225 bp paired-end) on a NovaSeq system (Illumina).

Single-cell data preprocessing. Single-cell RNA and TCR sequencing libraries were processed jointly using the cloud-based BD Rhapsody Targeted Analysis Pipeline (Seven Bridges Genomics). A notable difference with other platforms is the absence of a N-terminal cysteine in CDR3 chain calls, which needs to be considered when counting k-mers in downstream data analysis.

Infiltrating CD4$^+$ T cell repertoires

Donor selection. Donors were sourced from the Network for Pancreatic Organ Donors with Diabetes (nPOD, Table S7). Studies involving human participants were reviewed and approved by the University of Florida Institutional Research Board, ref. no. IRB201600029.

Peptide elution from HLA-DQ molecules

Protein production. Recombinant transketolase (TKT) protein was made by the Centre for Medicines Discovery (University of Oxford). Recombinant TKT from *B. caecimuris*, Integrated Gene Catalog (IGC)[67] entry MH0370_GL0036213, was produced in *Escherichia coli* BL21(DE3)-pRARE2, then purified by nickel affinity chromatography and size exclusion chromatography with endotoxin removal. Among all our hits in MGnify,[31] *B. caecimuris* was the first entry in IGC, a large cohort assembly with estimated abundances, and present in more than 10% of the individuals. IGC 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.

Cell culture and pulsing. Peptide elution from HLA-DQ molecules was performed as previously described.[68] EBV cells were grown to a total cell count of 1×10^8 in complete RPMI (R10). Spent R10 was removed, and cells were incubated with 500 µg of TKT or insulin (Biomm SA) in 5 ml of R10 for 2 h at 37 °C. R10 was added to full plate volume (15 ml) and cells further incubated at 37 °C for 18 h. 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. Harvested pellets were washed in PBS then lysed by mixing for 30 min with 2 ml of lysis buffer (0.5% v/v IGEPAL 630, 50 mM Tris pH 8.0, 150 mM NaCl) and 1 tablet Complete Protease Inhibitor Cocktail EDTA-free (Roche) / 10 ml buffer at room temperature. Lysate was clarified by centrifugation at 1000 × g for 10 min followed by a 20 000 × g spin step for 45 min at 4 °C. 2 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 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 used to elute bound HLA-DQ complexes from the PAS-antibody resin.

High-performance liquid chromatography. Elutions were vacuum centrifuged for drying, dissolved in loading buffer (0.1% v/v trifluoroacetic acid [TFA], 1% v/v acetonitrile in water), 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 analyzed 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 2×10^3 V, and peptide ions were introduced to the MS at an on-transfer tube temperature of 305 °C. Ions were analyzed by data-dependent acquisition. Initially a full-MS1 scan (1.2×10^5 resolution, 60 ms accumulation time, AGC 3×10^6) was followed by 20 data-dependent MS2 scans (6×10^4 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.

$CD4^+$ T cell stimulation assay

T-hybridoma cells. TCR sequences were identified from $CD4^+$ T cells in the islets or pancreas slices of T1D organ donors carrying the DR4-DQ8 haplotype distributed by nPOD: nPOD 69, 6323, 6342, 6367, 6472, and 6533 (Table S7), as described previously.[69] A total of 171 TCRs were expressed in 5KC T-hybridoma cells, devoid of endogenous TCR expression and engineered with a ZsGreen-1 reporter gene preceded by the nuclear factor activated T cells (NFAT) binding sequences.[70] TCR-expressing 5KC cells (2×10^4 cells / well) were cultured with 100 µM or designated concentrations of peptides in the presence of antigen presenting cells, Epstein-Barr virus-transformed autologous B cells (1×10^5 cells / well) or K562 cells transduced with each designated HLA class II molecule[69] (5×10^4 cells / well) in round-bottom 96-well plate for 16–22 h, as described in figure legends.

Peptide stimulation. Peptides at > 95% purity were purchased from Genemed Synthesis (Table S10). Cultures with and without anti-CD3ε antibody at 5 µg / ml (clone 125-2C11) were included in each assay as positive and negative controls, respectively. To determine which HLA molecules were presenting antigen to T cells, anti-HLA-DR (clone L243/G46-6), anti-HLA-DQ (REA303), or anti-DP (B7/21) antibodies were added at 12.5 µg / ml.

Quantification and statistical analysis

Single-cell data analysis

Quality control. We called cells from unique molecular identifiers 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. We only retained clonotypes originating from a consensus assembly and productive rearrangements.

Normalization. After data cleanup, we normalized all expression values to 10^4 reads / cell and applied a $\log_{10}$ transformation. Next, we discarded all but the top 5×10^3 most variable genes, and regressed out differences due to sequencing depth and mitochondrial expression using a generalized linear model.

Community detection. We aligned cells from all samples using batch-balanced nearest neighbors,[71] reduced the dimensionality to 2D,[72] called clusters,[73] and performed multivariate differential expression to detect subpopulation markers (Table S5).[74]

k-mer quantification. We excluded a N-terminal prefix of 4 residues and a C-terminal suffix of 3 residues from each CDR3 to avoid the HLA class II binding bias that constrains gene usage. Subsequently, we computed k-mer sequence frequencies by aggregating clonotype counts for each combination of k-mer length, $CD4^+$ subpopulation, and donor.

Bayesian inference on $CD4^+$ T cell repertoires

Markov chain Monte Carlo inference. We drew samples from posterior distributions using the No-U-Turn sampler (NUTS) implemented by Stan v2.32.1.[75,76] NUTS was run with default control parameters. In case of the model defined to estimate the effects of HLA on TCR repertoires (Generative model 1), for efficiency reasons, we removed covariates age, sex, autoantibody status, and T1D status since these had little predictive value. We also re-ran inference excluding autoantibody-positive and T1D donors, obtaining analogous differential expression results.

Model checking and diagnostics. We used the split-$\widehat{R}$ convergence diagnostic statistic to flag chain autocorrelation issues.[77] For all inference runs, $\widehat{R}$ deviated from 1 by less than 0.05, the standard convergence bound. We also carried posterior predictive checks with Bayesian p-values on 50 CDR3β 2-mers with the lowest and highest LFSR using sample mean as a test statistic.[78] No tail-area probabilities were near 0 or 1, which indicated replicated datasets were similar to observed data and therefore no major model failures could be identified.

$\mu_i \sim \text{Normal}(\mu{=}0, \sigma{=}1)$
$\zeta \sim \text{HalfCauchy}(\mu{=}0, \sigma{=}2.5)$
$\rho \sim \text{LKJ}(\eta{=}2)$
for $i \in \{1, \ldots, K\}$ **do**
    $\sigma_i \sim \text{HalfCauchy}(\mu{=}0, \sigma{=}2.5)$
    $\xi_i \sim \text{MultiNormal}(\mu{=}[\mu_1, \mu_2], \sigma{=}[\sigma_1, \sigma_2], \rho{=}\rho)$
    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

Generative model 1: Hierarchical mixed HLA effects model. Observed counts k_{ij} on k-mer i and donor j, out of a total n_j k-mers, are used to estimate the effect ξ_i explained by the T1D log-odds ratio or_j in a regression with normally distributed random effects. Regression intercepts, slopes, intercept-slope covariances, and random effects are shrunk by sharing information across k-mers with a hierarchical 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

Generative model 2: Beta-binomial aspartic acid proportion model. Observed aspartic acid counts k_i on donor i, out of a total n_i amino acid counts, are used to estimate the true hidden proportion p on each set of group observations. The prior belief for the frequency of aspartic acid p is modeled with an informative prior, using the posterior inferred previously (Generative model 1).

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

Generative model 3: Binomial preproinsulin-reactivity (PPI) regression model. Observed aspartic acid counts l_i on donor i, out of a total m_i amino acid counts, are used to regress the number of PPI clonotypes k_i, out of a total n_i clonotypes tested, using a measurement error structure. The prior belief for the frequency of aspartic acid p is modeled with an informative prior, using the posterior inferred previously (Generative model 1).

Insulin mimotope discovery

Alignment score model. We can formulate the problem of finding probable mimotopes to an epitope of interest as a statistical inference problem where we have to identify pairwise similarity that exceeds what would be expected by random chance.[79] We measured similarity as the maximum local pairwise alignment score between two protein sequences, computed using the Smith-Waterman algorithm.[80] We used a cost model defined by infinite gap penalties and a BLOSUM 80 substitution matrix.[30]

Empirical null distribution. We drew 10^5 random permutations of the epitope of interest, insulin $B_{9\text{–}25}$. On each iteration, we performed a pairwise local alignment against all 20 375 reviewed canonical protein isoforms from *Homo sapiens* in the UniProt database. This yielded more than 2×10^9 scores.

Non-parametric inference. Equipped with an empirical null distribution, it is straightforward to assess the significance of maximal local pairwise alignment scores of the non-permuted epitope against any set of proteins, e.g. all predicted open-reading frames from a gut microbiome assembly. This is simply done by calculating one-sided p-values, or evaluating the complementary cumulative distribution function (ccdf) for each score.

Parametric inference. Since typically only the highest alignment score from each pairwise alignment is considered, we want to model the statistical behavior of the maximum of a sequence of independent and identically distributed random variables, all scores for a given pairwise alignment, $max\{S_1, \ldots, S_n\}$. The Fisher-Tippett-Gnedenko or extremal types theorem states such a maximum, after proper renormalization, can only converge to the generalized extreme value (GEV) distribution. The cumulative distribution function, defined on the set $\{z \mid 1 + \xi(z-\mu)/\sigma > 0\}$, is given by

$$\mathrm{G}(z) = \exp\left\{-\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right\}$$

Maximum likelihood estimation. We can fit a GEV distribution to the observed data by maximum likelihood. This maximum likelihood fit can be obtained by solving the constrained optimization problem defined below, and is guaranteed to exist and be unique.[81] We used an interior point line search filter algorithm[82] to estimate such a fit to a set of maximal alignment scores and calculate a parametric approximation to the null distribution of local alignment scores. The GEV log-likelihood to maximize is

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

subject to the constraint

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

If $\xi = 0$, we require separate treatment

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

We have previously identified many of the same insulin mimotopes using a different generative model-based method.[22] Other approaches to the same problem have relied on human-curated results from BLAST searches performed on fully assembled genomes from cultured species.[46,48] Consequently, selected peptides lack statistical significance or come from organisms that do not have humans as a host.

Sequence motif conservation

Amino acid information content. The information content is defined as the reduction in entropy after some message is received. In case of a protein motif, the information content I at position i is given by the difference in entropy between a uniform distribution of 20 amino acids and the observed distribution

$$I_i = H_i^0 - H_i^1 = \log_2(20) - H_i^1$$

where the entropy H of residue i is given by

$$H_i = -\sum_{j=1}^{20} P(x_{ij})\log_2 P(x_{ij})$$

Consensus motif logo. We built a consensus motif logo for N-terminal transketolase (TKT) in the region of insulin mimicry by representing the amino acid distribution at each position i, rescaled by its information content I_i (Figure S4).[83] Since the profile hidden Markov model (pHMM) for TKT (Pfam PF00456) in the region of interest has near zero probability of insertions or deletions, this allowed establishing a one-to-one mapping between pHMM states and residues.

Liquid chromatography and mass spectrometry

Peptide identification. Raw data files were analyzed with PEAKS X (Bioinformatic Solutions) using a protein sequence database containing 20 606 reviewed human entries from UniProt, supplemented with the sequence of *B. caecimuris* transketolase protein (Figure S5). 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 inbuilt *de novo*-led search for 303 common modifications. FDR was calculated using the decoy database search built into PEAKS (Table S11).

Supplemental information

JDRF Diabetes, Genes, Autoimmunity, and Prevention (D-GAP) cohort

	Haplotype 1			Haplotype 2						
n	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	AA_1	AA_2	log(OR)	Group
2	15:01	01:02	06:02	11:04	05:05	06:03	DRA	DSR	−1.30	Protected
1	15:01	01:02	06:02	13:02	01:02	06:04	DRA	VSE	−1.30	Protected
2	15:01	01:02	06:02	16:01	01:02	05:02	DRA	SRR	−1.30	Protected
2	15:01	01:02	06:02	15:01	01:02	06:02	DRA	DRA	−0.96	Protected
1	15:01	01:02	06:02	11:04	05:05	03:01	DRA	DSR	−0.89	Protected
1	15:01	01:02	06:02	12:01	05:05	03:01	DRA	DGR	−0.89	Protected
1	04:01	03:01	03:01	11:03	05:05	03:01	DHK	DSE	0.09	Low risk
1	04:07	03:03	03:01	01:03	05:05	03:01	DHR	DFE	0.09	Low risk
3	04:01	03:02	03:01	04:01	03:02	03:01	DHK	DHK	0.58	Low risk
1	04:08	03:02	03:01	04:07	03:02	03:01	DHR	DHR	0.58	Low risk
1	04:01	03:01	03:01	12:02	06:01	03:01	DHK	DGR	0.59	Low risk
1	09:01	03:02	02:02	04:01	03:02	03:01	AFR	DHK	0.59	Low risk
5	04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	1.51	Susceptible
1	04:04	03:01	03:02	04:01	03:01	03:02	AHR	AHK	1.51	Susceptible
2	04:04	03:01	03:02	04:04	03:01	03:02	AHR	AHR	1.51	Susceptible
17	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	1.80	Susceptible
1	04:02	03:01	03:02	03:01	05:01	02:01	AHE	ASK	1.80	Susceptible
1	04:03	03:01	03:02	03:01	05:01	02:01	AHR	ASK	1.80	Susceptible
4	04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	1.80	Susceptible

Table S1: D-GAP sample selection summary, with samples grouped by HLA DR-DQ diplotypes and sorted by T1D risk, measured as the $\log_{10}$-transformed odds ratio (OR) of developing T1D contributed by each diplotype. OR estimates were obtained from UK Biobank,[24] and used as the explanatory variable (Generative model 1). The categorical group variable only has a nominal value. AA_1 and AA_2 encode amino acids at HLA DQ$\beta1_{57}$, DR$\beta1_{13}$, and DR$\beta1_{71}$.[6]

	Haplotype 1			Haplotype 2					Autoantibody count						
	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	AA_1	AA_2	GADA	IA2A	ZnT8A	CIAA	T1D status	Age	Sex
01	15:01	01:02	06:02	15:01	01:02	06:02	DRA	DRA	0	0	0	0	Unaffected	[5,10)	M
03	15:01	01:02	06:02	11:04	05:05	06:03	DRA	DSR	0	0	0	0	Unaffected	[15,20)	M
07	15:01	01:02	06:02	15:01	01:02	06:02	DRA	DRA	0	0	0	0	Unaffected	[5,10)	M
09	15:01	01:02	06:02	11:04	05:05	06:03	DRA	DSR	0	0	0	0	Unaffected	[15,20)	F
18	15:01	01:02	06:02	13:02	01:02	06:04	DRA	VSE	0	0	0	0	Unaffected	[15,20)	F
21	15:01	01:02	06:02	11:04	05:05	03:01	DRA	DSR	0	0	0	0	Unaffected	[20,25)	M
23	15:01	01:02	06:02	16:01	01:02	05:02	DRA	SRR	0	0	0	0	Unaffected	[15,20)	F
29	15:01	01:02	06:02	12:01	05:05	03:01	DRA	DGR	0	0	0	0	Unaffected	[15,20)	F
32	15:01	01:02	06:02	16:01	01:02	05:02	DRA	SRR	0	0	0	0	Unaffected	[10,15)	M

Table S2: D-GAP genetically protected group HLA DR-DQ diplotypes, autoantibody count, and disease status at the time of peripheral blood sample collection. AA_1 and AA_2 encode amino acids at HLA DQ$\beta1_{57}$, DR$\beta1_{13}$, and DR$\beta1_{71}$.[6]

	Haplotype 1			Haplotype 2					Autoantibody count						
	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	AA_1	AA_2	GADA	IA2A	ZnT8A	CIAA	T1D status	Age	Sex
42	09:01	03:02	02:02	04:01	03:02	03:01	AFR	DHK	0	0	0	0	Unaffected	[5,10)	F
43	04:07	03:03	03:01	01:03	05:05	03:01	DHR	DFE	0	0	0	0	Unaffected	[10,15)	M
45	04:01	03:02	03:01	04:01	03:02	03:01	DHK	DHK	0	0	0	0	Unaffected	[10,15)	M
46	04:01	03:01	03:01	12:02	06:01	03:01	DHK	DGR	0	0	0	0	Unaffected	[15,20)	F
51	04:01	03:02	03:01	04:01	03:02	03:01	DHK	DHK	0	0	0	0	Unaffected	[5,10)	M
52	04:08	03:02	03:01	04:07	03:02	03:01	DHR	DHR	0	0	0	0	Unaffected	[5,10)	M
53	04:01	03:01	03:01	11:03	05:05	03:01	DHK	DSE	0	0	0	0	Unaffected	[20,25)	M
55	04:01	03:02	03:01	04:01	03:02	03:01	DHK	DHK	2	0	2	2	Unaffected	[5,10)	M

Table S3: D-GAP low genetic risk group HLA DR-DQ diplotypes, autoantibody count, and disease status at the time of peripheral blood sample collection. AA_1 and AA_2 encode amino acids at HLA DQ$\beta1_{57}$, DR$\beta1_{13}$, and DR$\beta1_{71}$.[6] The DQA1*03:02 allele is unresolved to DQA1*03:02 or DQA1*03:03 due to insufficient HLA typing resolution.

	Haplotype 1			Haplotype 2					Autoantibody count						
	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	AA_1	AA_2	GADA	IA2A	ZnT8A	CIAA	T1D status	Age	Sex
04	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	2	0	0	0	Unaffected	[10,15)	F
05	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[5,10)	F
06	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[10,15)	M
08	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[15,20)	M
11	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[15,20)	M
12	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[15,20)	F
13	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[10,15)	M
14	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[20,25)	M
19	04:02	03:01	03:02	03:01	05:01	02:01	AHE	ASK	1	1	1	—	T1D	[10,15)	M
20	04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	1	1	1	—	T1D	[10,15)	F
22	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	2	2	2	0	T1D	[15,20)	F
24	04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	1	1	1	—	T1D	[5,10)	F
25	04:03	03:01	03:02	03:01	05:01	02:01	AHR	ASK	1	1	1	—	T1D	[10,15)	M
26	04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	1	1	1	—	T1D	[5,10)	F
27	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	1	1	1	—	T1D	[5,10)	F
28	04:04	03:01	03:02	03:01	05:01	02:01	AHR	ASK	1	1	1	—	T1D	[10,15)	M
41	04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	0	0	0	0	Unaffected	[10,15)	M
44	04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	0	0	0	0	Unaffected	[15,20)	M
47	04:04	03:01	03:02	04:04	03:01	03:02	AHR	AHR	0	0	0	0	Unaffected	[5,10)	M
48	04:04	03:01	03:02	04:04	03:01	03:02	AHR	AHR	0	0	0	0	Unaffected	[10,15)	M
49	04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	0	0	0	0	Unaffected	[5,10)	F
50	04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	0	0	0	0	Unaffected	[10,15)	F
54	04:01	03:01	03:02	04:01	03:01	03:02	AHK	AHK	0	0	0	0	Unaffected	[10,15)	M
56	04:04	03:01	03:02	04:01	03:01	03:02	AHR	AHK	0	0	0	0	Unaffected	[15,20)	M
57	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[10,15)	F
59	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	2	0	0	0	Unaffected	[10,15)	M
60	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[10,15)	F
61	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[10,15)	M
62	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	2	0	2	2	Unaffected	[10,15)	F
63	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[10,15)	F
64	04:01	03:01	03:02	03:01	05:01	02:01	AHK	ASK	0	0	0	0	Unaffected	[10,15)	F

Table S4: D-GAP genetically susceptible group HLA DR-DQ diplotypes, autoantibody count, and disease status at the time of peripheral blood sample collection. AA_1 and AA_2 encode amino acids at HLA DQ$\beta1_{57}$, DR$\beta1_{13}$, and DR$\beta1_{71}$.[6] — denotes missing data.

	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

Table S5: Top 5 gene markers for single $CD4^+$ T cell clusters 0–11 in the D-GAP cohort (Figure S1). Genes labeled with * indicate no standard name is currently assigned, and the identifier of the corresponding genomic region has been used as a placeholder.

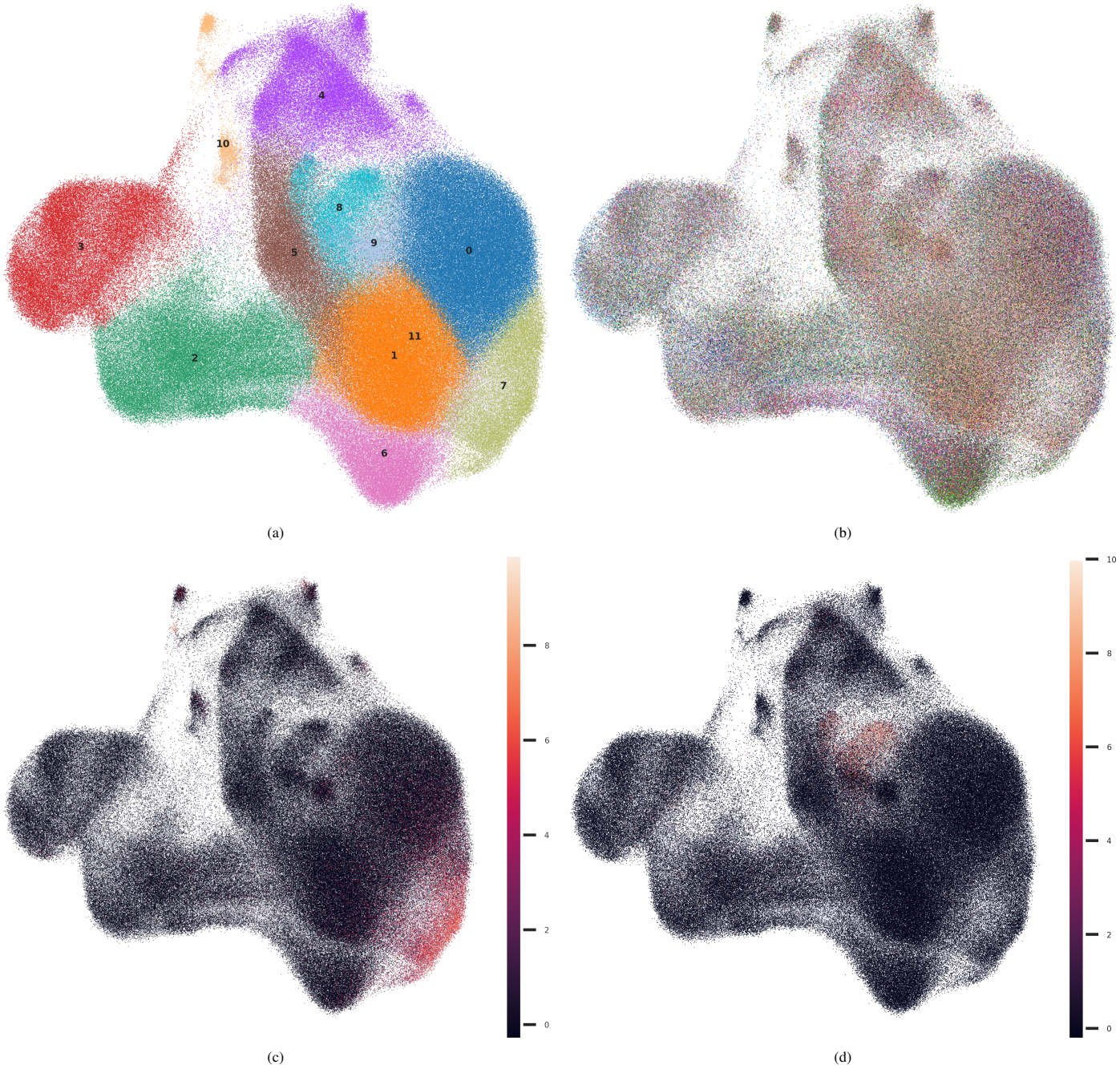
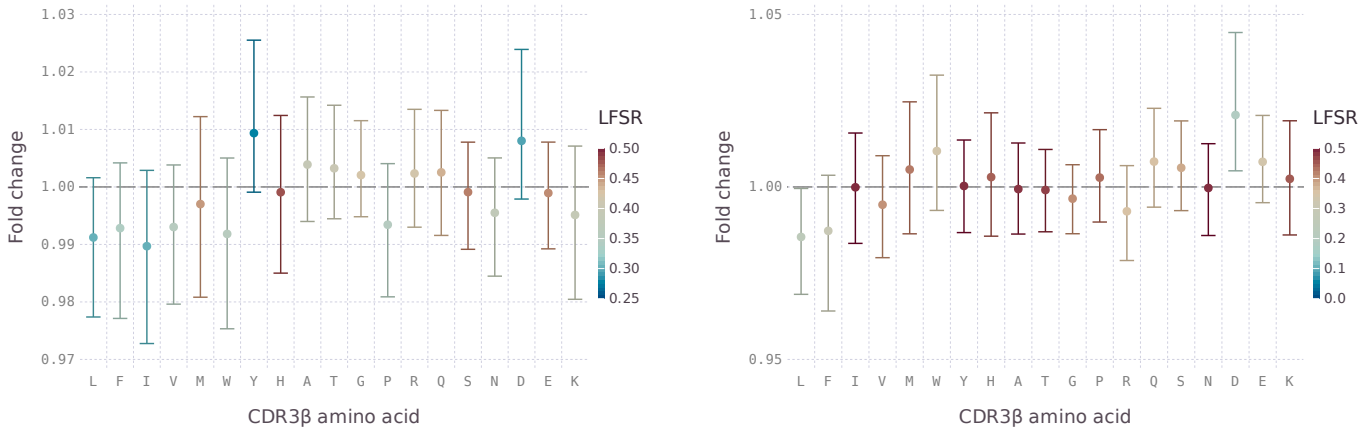
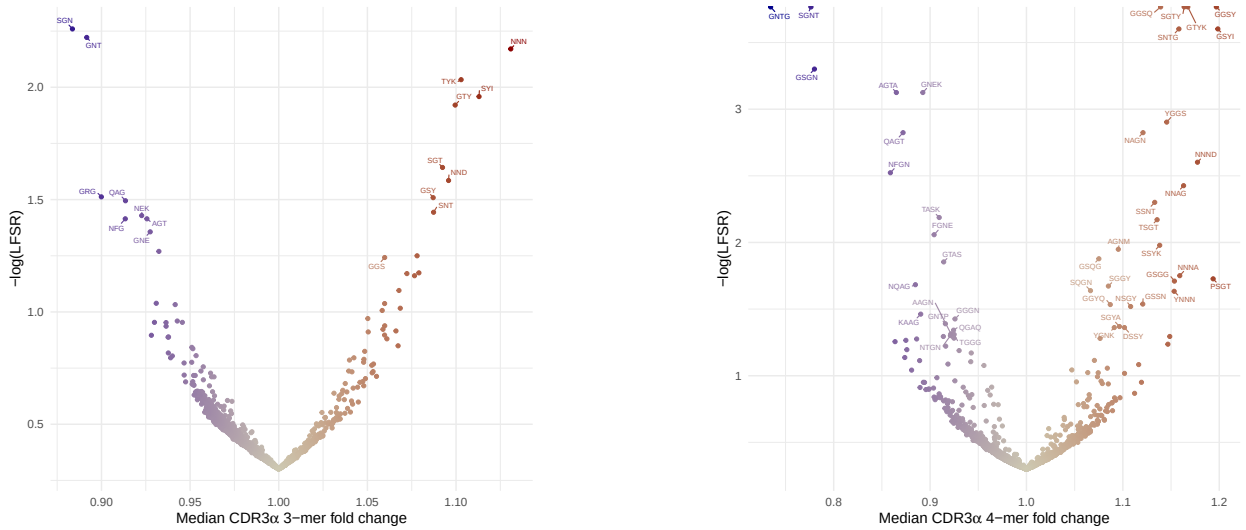


Figure S1: Single-cell RNA subsets in the D-GAP cohort. Each point represents one cell. Expression levels are indicated in $\log_{10}$-transformed read counts, where 1 unit = 10^4 reads prior to the transformation. (a) $CD4^+$ subsets 0–11. (b) Donor distribution. (c) *CXCL8* expression, a marker of recent thymic emigrants (RTEs),[25] identified as subset 7. (d) *FOXP3* expression, a marker of regulatory T cells (Tregs), identified as subset 8. $CD4^+$ T conventional cells (Tconvs) were defined as the union of all subsets except 7 and 8.



(a) (b)

Figure S2: Estimates of T1D odds ratio effects, contributed by HLA DR-DQ diplotypes, on CDR3β 1-mers from recent thymic emigrant (RTE) and regulatory (Treg) CD4$^+$ T cells, excluding cysteine. Error bars summarize the effects on multiple k-mers by depicting the median and 90% credible intervals of fold changes across HLA class II extremes, with the most protective DQ6 diplotypes as baseline (Table S1). (a) RTE cell subset. (b) Treg cell subset.



(a) (b)

Figure S3: Estimates of T1D odds ratio effects, contributed by HLA DR-DQ diplotypes, on CDR3α k-mers from CD4$^+$ T conventional cells. Funnel plots represent the median and the local false sign rate (LFSR) of fold changes across HLA class II extremes, with the most protective DQ6 diplotypes as baseline (Table S1). (a) All 3-mers with non-zero counts. (b) All 4-mers with non-zero counts. 4-mers on the y axis boundary have an estimated LFSR ≈ 0.

Investigating Underlying Causal Mechanisms of Type 1 Diabetes (DMech) donors

	Haplotype 1			Haplotype 2			IAA		GADA				
	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	Status	U/ml	Status	U/ml	ICA	Age	Sex
11	04:01	03:01	03:02	04:02	03:01	03:02	Negative	0.80	Positive	15.9	Positive	[5,10)	M
13	03:01	05:01	02:01	03:01	05:05	02:02	Negative	0.80	Positive	91.9	Positive	[10,15)	M
14	04:01	03:01	03:02	04:01	03:01	03:02	Positive	6.34	Positive	86.3	Positive	[5,10)	F
16	03:01	05:01	02:01	03:01	05:05	02:02	Positive	8.44	Positive	88.2	Positive	[10,15)	M
18	04:01	02:01	02:02	07:01	03:01	03:02	Positive	6.14	Positive	71.8	Positive	[5,10)	F

Table S6: DMech donor HLA-DR-DQ diplotypes and autoantibody status at the time of peripheral blood sample collection.

Network of Pancreatic Organ Donors with Diabetes (nPOD)

	Haplotype 1			Haplotype 2			Autoantibody				PPI-reactive				
	DRB1	DQA1	DQB1	DRB1	DQA1	DQB1	IAA	IA2A	GADA	ZnT8A	Positive	Tested	Diag.	Age	Sex
69	04:01	03:01	03:02	07:01	02:01	02:02	—	—	—	—	0	7	[0,5)	[5,10)	F
6323	03:01	05:01	02:01	04:02	03:01	03:02	Negative	Positive	Positive	Negative	4	52	[5,10)	[20,25)	F
6342	01:01	01:01	05:01	04:01	03:01	03:02	Positive	Positive	Negative	Negative	5	34	[0,5)	[10,15)	F
6367	04:01	03:01	03:02	07:01	02:01	02:02	Negative	Negative	Negative	Negative	0	9	[0,5)	[20,25)	M
6414	03:01	05:01	02:01	09:01	03:03	02:02	Positive	Negative	Positive	Positive	4	34	[0,5)	[20,25)	M
6472	03:01	05:01	02:01	04:04	03:01	03:02	Positive	Negative	Negative	Negative	1	30	[0,5)	[10,15)	F
6533	03:01	05:01	02:01	04:01	03:01	03:02	Positive	Positive	Negative	Positive	—	—	[0,5)	[0,5)	F
6536	DR4	DQ8	DQ8	DR17	DQ2	DQ2	Negative	Negative	Positive	Negative	—	—	[0,5)	[20,25)	F

Table S7: nPOD donor HLA DR-DQ diplotypes, autoantibody status, number of preproinsulin-reactive (PPI) $CD4^+$ T cell clones isolated from the pancreas, and years since diagnosis at the time of death. — denotes missing data.

Insulin mimotope discovery

Species	Sequence	Protein	Score	FDR
Acetatifactor sp.	GHCVDALYMVLGEKGFF	TKT	98	0.001
Blautia caecimuris	GHSVEALYAVLAEKGFF	TKT	88	0.041
Succinivibrio sp.	GHSVEALYAVLAEKGFF	TKT	88	0.041
Acutalibacter sp.	GHSVEALYAVLADRGFF	TKT	88	0.041
Pygmaiobacter sp.	GHAVEALYSVLADRGFF	TKT	88	0.041
Oscillospiraceae sp.	GHSVEALYAVLADRGFF	TKT	88	0.041
Lachnospiraceae sp.	GHSVEALYAVLAEKGFF	TKT	88	0.041
Borkfalkia sp.	TVKVEPIYLVCGEDAFF	YqeN	87	0.045
Mitsuokella multacida	GHCVEALYAILADRGFF	TKT	86	0.045
Christensenella sp.	GHAAPALYAVLGERGFF	TKT	86	0.045
Mobilibacterium sp.	GHAVPALYAALGERGFF	TKT	86	0.045
Clostridium leptum	GHSVEALYCILADRGFF	TKT	85	0.048
Eisenbergiella sp.	GHCVDALYMVLGDLGFF	TKT	85	0.048
Lachnospiraceae sp.	GHCVDALYMVLGDLGFF	TKT	85	0.048
Faec. prausnitzii	SHLEEVLYLLCGEKCAI	GloB	85	0.048
Ruminococcus sp.	NDDLCPDYIVCGERGFQ	—	78	0.129
Faec. prausnitzii	QGDADAIYLLCGERGLT	GH25	77	0.132
Christensenella sp.	DRLVNAVYVLCGYVGFF	YkvI	75	0.200
Ruminococcus sp.	NDYSVCAYLTCGENGFF	ComE	75	0.200
Anaerobutyricum sp.	SHLSEALLLTCGQKKSL	YqeC	75	0.200
Anaerobutyricum hallii	VSFDEALYLVCGESIG*	CH3	75	0.200
Insulin $B_{9–25}$	SHLVEALYLVCGERGFF		145	

Table S8: Top scoring local alignments to insulin $B_{9–25}$ in proteins from the gut microbiome, FDR ≤ 0.05. Top scoring members from other protein families, FDR ≤ 0.2, are also included. Shaded residues do not belong to the optimal local alignment to insulin $B_{9–25}$. Protein domains and superfamilies listed include N-terminal transketolase (TKT), yqeN DNA replication protein (YqeN), hydroxyacylglutathione hydrolase (GloB), glycosyl hydrolase family 25 (GH25), uncharacterized membrane protein (YkvI), late competence operon (ComE), selenium cofactor biosynthesis protein (YqeC), and cyclases/histidine kinases associated sensory extracellular domain 3 (CH3). — indicates no protein family could be predicted. For reference, human insulin paralogs insulin-like growth factor I (IGF-1) and II (IGF-2) scores are 90 and 81, respectively, whereas human transketolase has a score of 43.

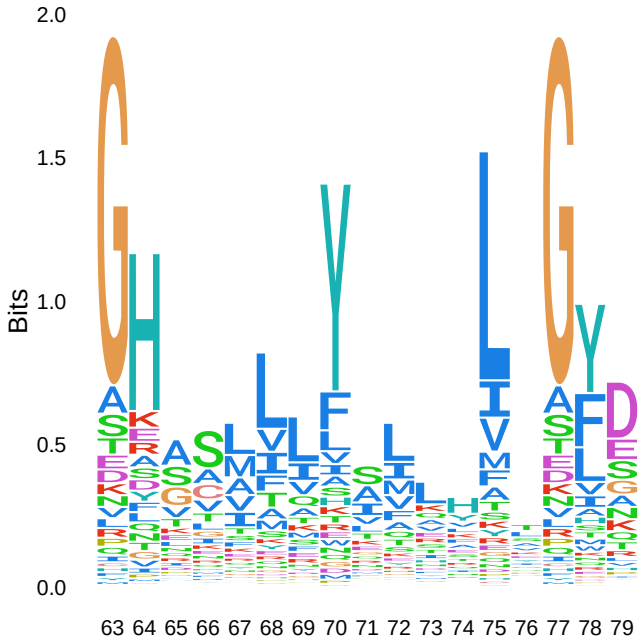


Figure S4: N-terminal transketolase domain (Pfam PF00456) motif for all domains of life in the region of insulin mimicry. A large information content in bits denotes strong conservation. Most conserved residues in transketolase (TKT) are similar or identical to those from the corresponding insulin B_{9-25} (`SHLVEALYLVCGERGFF`) position. This may facilitate the evolution of insulin B_{9-25} mimotopes using TKT as a template.

CD4⁺ T cell stimulation assays

		α chain			β chain		
	nPOD	V gene	J gene	CDR3	V gene	J gene	CDR3
GSE.6H9	6323	TRAV26-1	TRAJ40	`CIVRVDSGTYKYIF`	TRBV7-2	TRBJ2-1	`CASSLTAGLASTYNEQFF`
GSE.20D11	6342	TRAV12-3	TRAJ4	`CAILSGGYNKLIF`	TRBV2	TRBJ2-5	`CASSAETQYF`
GSE.166H9	6533	TRAV3	TRAJ21	`CAVMYNFNKFYF`	TRBV12-3	TRBJ2-5	`CASSLGGRETQYF`

Table S9: nPOD CD4⁺ T cell clonotypes stimulated by insulin B_{9-23} (Figure 4). GSE.6H9 and 20D11 were previously reported.[33] GSE.166H9 is a novel clonotype that cross-reacts with transketolase from *Clostridium leptum*, peptide 8 (Table S10).

	Peptide	Species	Sequence	Protein	Score	FDR
1	`GHSVEALYAVLAEKG`					
2	`HSVEALYAVLAEKGF`	*Blautia caecimuris*	`GHSVEALYAVLAEKGFF`	TKT	88	0.041
3	`SVEALYAVLAEKGFF`					
4	`GHCVEALYVTLEAKG`					
5	`HCVEALYVTLEAKGF`	*Phocaeicola dorei*	`GHCVEALYVTLEAKGFI`	TKT	62	~1.0
6	`CVEALYVTLEAKGFI`					
7	`GHTVEALYAVLCQKG`	*Ruminiclostridium siraeum*	`GHTVEALYAVLCQKGFF`	TKT	78	0.129
8	`GHSVEALYCILADRG`	*Clostridium leptum*	`GHSVEALYCILADRGFF`	TKT	85	0.048
9	`GHIAEALYVTLAKRG`	*Erysipelatoclost. ramosum*	`GHIAEALYVTLAKRGFF`	TKT	82	0.090
10	`GHCVEALYVTLESKG`	*Phocaeicola dorei*	`GHCVEALYVTLESKGFI`	TKT	63	~1.0

Table S10: Peptides employed on CD4⁺ T cell stimulation assay (Figure 4), derived from prioritized mimotopes (Table S8). Shaded residues do not belong to the optimal local alignment to insulin B_{9-25}. Peptide 8 stimulates the novel anti-insulin clonotype GSE.166H9 (Table S9). Peptides were chosen to maximize sequence diversity within transketolase sequences with FDR ≤ 0.2 and also present in >10% individuals from IGC, a population-wide gut microbiome reference catalog.[67] Despite lower similarity to insulin B_{9-25}, *P. dorei* peptides were also included owing to the increase in abundance of this species observed prior to autoimmunity in one cohort study.[18]

Peptide elution in HLA-DQ molecules

Sample	Peptide	P	FDR
DQ6	GKPTVLIANTVKGCGSSVMENK	1.74×10^{-6}	0.029
DQ6	KDVINMIRSGKAGHIG	1.20×10^{-5}	0.029
DQ6	FVMSKGHSVEALY	1.90×10^{-5}	0.029
DQ6	IEEVIDTFSKFGSK	5.26×10^{-5}	0.029
DQ6	GHSVEALYAVLAE	2.92×10^{-4}	0.029
DQ6	CVGMALAGK	1.85×10^{-3}	0.029
DQ8	VTLYFKQMNISPEN	1.01×10^{-5}	0.025
DQ8	LDAAFEEAKTVKGKPTVL	4.35×10^{-5}	0.025
DQ8	NSIEELDAAFEEAKT	1.04×10^{-3}	0.025

Table S11: Eluted peptides from *Blautia caecimuris* transketolase measured to be bound by HLA-DQ molecules using mass spectrometry, FDR ≤ 0.05. Using an anti-DQ antibody, peptides presented by EBV B-cell lines were identified from elution experiments with either DQ6$^+$ or DQ8$^+$ cells. Peptides FVMSKGHSVEALY and GHSVEALYAVLAE, bound by DQ6, correspond to the region of insulin B_{9-25} mimicry (Figure S5).

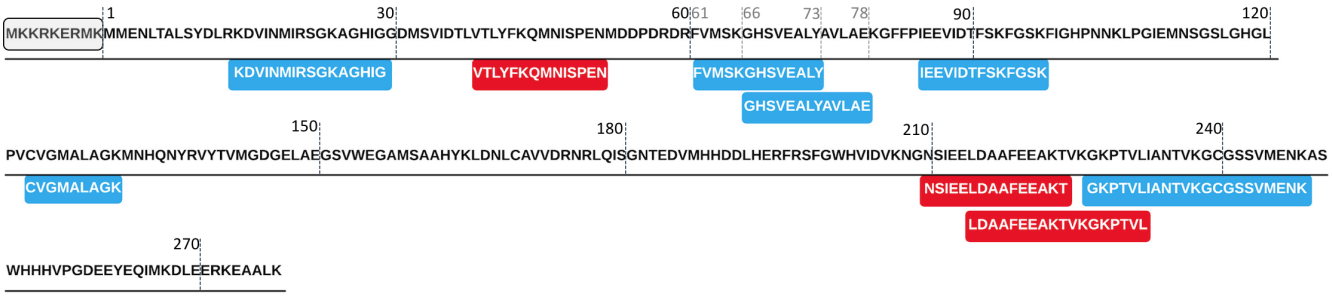


Figure S5: Eluted peptides measured to be bound by HLA-DQ6 (blue) and DQ8 (red) molecules using mass spectrometry, FDR ≤ 0.05 (Table S11), overlaid on the *Blautia caecimuris* transketolase sequence they originate from. Peptides in the region 61–78 correspond to insulin B_{9-25} mimics. The N-terminal protein sequence (MKKRKERMK), present in the IGC gut metagenome assembly,[67] is absent in the alternative assembly constructed by MGnify.[31] The longer version was synthesized and used in the elution assays.