Identification of a type 1 diabetes-associated T cell receptor repertoire signature from the human peripheral blood

Puneet Rawat[1], Melanie R. Shapiro[2,H], Leeana D. Peters[2,H], Michael Widrich[3,H], Koshlan Mayer-Blackwell[4,H], Keshav Motwani[2,6,H], Milena Pavlović[5,7,L], Ghadi al Hajj[5,L], Amanda L. Posgai[2,L], Chakravarthi Kanduri[5,L], Giulio Isacchini[8,L], Maria Chernigovskaya[1,L], Lonneke Scheffer[5], Kartik Motwani[2], Leandro Octavio Balzano-Nogueira[2], Camryn M. Pettenger-Willey[4,9], Sebastiaan Valkiers[4,10], Laura Jacobsen[11], Michael J. Haller[11], Desmond A. Schatz[11], Clive H. Wasserfall[2], Ryan O. Emerson[12], Andrew J Fiore-Gartland[4], Mark A. Atkinson[2,11], Günter Klambauer[3,13], Geir Kjetil Sandve[5], Victor Greiff[1,8,C], and Todd M. Brusko[2,11,14,C]

[1] Department of Immunology, University of Oslo, Oslo, Norway
[2] Department of Pathology, Immunology, and Laboratory Medicine, Diabetes Institute, College of Medicine, University of Florida, Gainesville, FL, USA
[3] Institute for Machine Learning, Johannes Kepler University Linz, Austria
[4] Fred Hutchinson Cancer Center, Seattle, WA, USA
[5] Department of Informatics, University of Oslo, Oslo, Norway
[6] Department of Biostatistics, University of Washington, Seattle, WA
[7] UiO:RealArt Convergence Environment, University of Oslo, Oslo, Norway
[8] Imprint Labs, LLC. New York, NY, USA
[9] Whitman College, Walla Walla, WA, USA
[10] University of Antwerp, Antwerp, Belgium
[11] Department of Pediatrics, College of Medicine, University of Florida, Gainesville, FL, USA
[12] Seattle, WA, USA
[13] LIT AI Lab, Johannes Kepler University Linz, Austria
[14] Department of Biochemistry and Molecular Biology, College of Medicine, University of Florida, Gainesville, FL, USA

[H] Equal contribution
[L] Equal contribution
[C] Corresponding authors
Email: tbrusko@ufl.edu, victor.greiff@medisin.uio.no

Abstract

Type 1 Diabetes (T1D) is a T-cell mediated disease with a strong immunogenetic HLA dependence. HLA allelic influence on the T cell receptor (TCR) repertoire shapes thymic selection and controls activation of diabetogenic clones, yet remains largely unresolved in T1D. We sequenced the circulating TCRβ chain repertoire from 2250 HLA-typed individuals across three cross-sectional cohorts, including T1D patients, and healthy related and unrelated controls. We found that HLA risk alleles show higher restriction of TCR repertoires in T1D individuals. Machine learning analysis yielded AUROC of 0.77 on test cohorts for T1D classification. T1D-specific TCR features predominantly localized to the subsequence motifs, indicating absence of T1D-associated public clones. These TCR motifs were also observed in independent TCR cohorts residing in pancreas-draining lymph nodes of T1D individuals. Collectively, our data demonstrate T1D-related TCR motif enrichment based on genetic risk, offering a potential metric for autoreactivity and basis for TCR-based diagnostics and therapeutics.

Introduction

Antigen-specific recognition of both foreign and auto-antigens is enabled by T-cell receptors (TCR). Elucidating the TCR sequences implicated in type 1 diabetes (T1D) is crucial for advancing our understanding of disease pathogenesis and facilitating the development of reliable biomarkers. Human leukocyte antigen (HLA) loci shape the TCR repertoire [1–3], and HLA risk alleles have been shown to influence TCR clonal sequences in various autoimmune diseases, including rheumatoid arthritis (RA), T1D, and celiac disease (CD) [4,5]. In T1D, HLA class II loci account for most genetic risk [6,7], with DR3 and DR4 alleles conferring the highest risk [8]. Seropositivity for islet autoantibodies in combination with genetic risk can provide population-level estimates of the rate of T1D progression, with longitudinal studies illustrating the influence of HLA on type of initial seroconversion (e.g., insulin autoantibody [IAA] vs GAD autoantibody [GADA] first) and progression to multiple autoantibodies [9]. The development of T-cell biomarkers could enhance the monitoring of disease progression and improve current predictive methods for pre-symptomatic disease detection [10]. However, efforts to develop T1D biomarkers have been constrained by a reliance on pre-existing knowledge of autoimmune targets and cellular mechanisms in T1D [11]. There remains a critical need for genetic studies that explore the link between TCRs and HLA in the context of T1D and for the identification of TCR-based biomarkers [12].

Analyses of the adaptive immune receptor repertoire (AIRR) in infectious and autoimmune disease contexts have demonstrated that the TCR repertoire may be used to develop novel diagnostics [13–19]. Longitudinal studies of healthy adult cohorts showed repertoire stability, indicating the robustness of TCR-based signatures for identifying sustained disease repertoire perturbations [20]. However, autoimmunity-associated immune receptor signals, in contrast to those associated with infection and cancer, were initially found to be small or nearly undetectable when considering global repertoire metrics [13,21–24]. Therefore, novel analytical techniques, such as machine learning (ML) platforms, have been used to detect even very small shifts in signal within adaptive immune repertoires. For example, repertoire alterations have been documented in response to cytomegalovirus (CMV) [17,25], cancer (e.g., lymphoma, tumor-infiltrating lymphocytes) [26,27], therapeutic responses to checkpoint inhibitors in melanoma [28], as well as systemic autoimmune diseases including multiple sclerosis (MS) [29], systemic lupus erythematosus (SLE), and RA [16,30] using statistical analyses and ML approaches on AIRR sequencing data with reasonable accuracy (area under the receiver operating curve [AUROC] > 0.75). The ML methods have demonstrated the capability to classify immune repertoires for a given clinical status and to recover immune signals associated with specific clinical conditions [31–33].

While previous reports indicate that islet autoantigen reactive cells are present at similar frequencies in the peripheral blood of control and T1D study participants [34–38], others have observed increased expansion of antigen-specific clonotypes in T1D, indicating a disease signature in the periphery [39]. The majority of islet-antigen specific complementarity determining region 3 beta chain (CDR3β) sequences were found to be private [40], or observed at the individual level, demonstrating a need for short consensus motif-based biomarkers.

We sought to identify a T1D-associated T cell repertoire signature from bulk peripheral blood mononuclear cells (PBMC) to provide a translationally relevant biomarker. To this end, we sequenced 2250 circulating

TCRβ repertoires across the natural history of T1D to: (i) investigate the existence of HLA-restricted, T1D-associated CDR3β sequences and motifs, (ii) assess the feasibility to classify individuals as having T1D based on the TCR repertoire, and (iii) identify T1D-associated TCR signatures within the TCR repertoires. Our study reveals that the TCR repertoire lacks shared public clones across clinical groups. However, a significant enrichment of TCR repertoires at the subsequence level was observed, both with and without accounting for genetic risk factors. These findings suggest that TCR repertoire analysis may offer valuable insights into autoreactivity, potentially serving as a basis for developing TCR-based diagnostic tools and therapeutic strategies.

Results

Overview of the dataset and reproducibility assessment

To study T1D-associated TCR repertoire alterations, we immunosequenced the rearranged CDR3 TCRβ region in bulk PBMCs of three different cross-sectional cohorts (Fig. 1). Cohort 1 contains 1393 repertoires (103176 ± 26850 unique CDR3β sequences) distributed across the natural history of T1D from the University of Florida Diabetes Institute (UFDI) biobank. These include individuals diagnosed with T1D (n=426); first-degree relatives of T1D patients (FDR, n=625); second-degree relatives of T1D patients (SDR, n=59); unrelated healthy control individuals (CTRL, n=188), and islet autoantibody-positive individuals without diabetes who have increased risk of developing T1D (AAb+, n=95). Although statistically significant differences in unique CDR3β sequences were observed between clinical groups (<0.05), the difference in mean, median, and standard deviation was minimal (Fig. S1). Cohorts 2 and 3 respectively contain 679 T1D repertoires (112637 ± 55091 unique CDR3β sequences) and 178 deep sequenced CTRL repertoires (366271 ± 143197 unique CDR3β sequences), the latter sequenced with a different protocol than cohort 1 (see **Methods**). We obtained high reproducibility and adequate sequencing depth to capture the clonal diversity of TCRs using technical replicates (Fig. S2).

In addition to TCRβ sequencing, the dataset has extensive metadata in the form of whole genome (>978K single nucleotide polymorphism [SNP]) data derived from the UFDIchip, a custom Affymetrix array [41], clinical data (HbA1c, C-peptide, gender, age, T1D duration), and serological data (islet AAb number and specificity [IAA, GADA, zinc transporter 8 (ZnT8A), insulinoma associated protein-2 (IA-2A)] [42]) (Fig. S1). We employed precision genotyping [41] and HLA imputation [43] techniques to obtain 4-digit classical alleles for the major histocompatibility complex (MHC) class I and II genes. Specifically, we obtained 4-digit HLA for 1332 (95.6%) individuals including all AAb+ repertoires in cohort 1, 645 (95%) individuals in cohort 2, and 178 (100%) individuals in cohort 3. The majority of the primary analyses in this work were based on cohort 1 because it is the largest cohort among the three that also contains all clinical statuses. In contrast, cohorts 2 and 3 served as test cohorts for ML analysis unless specified otherwise.

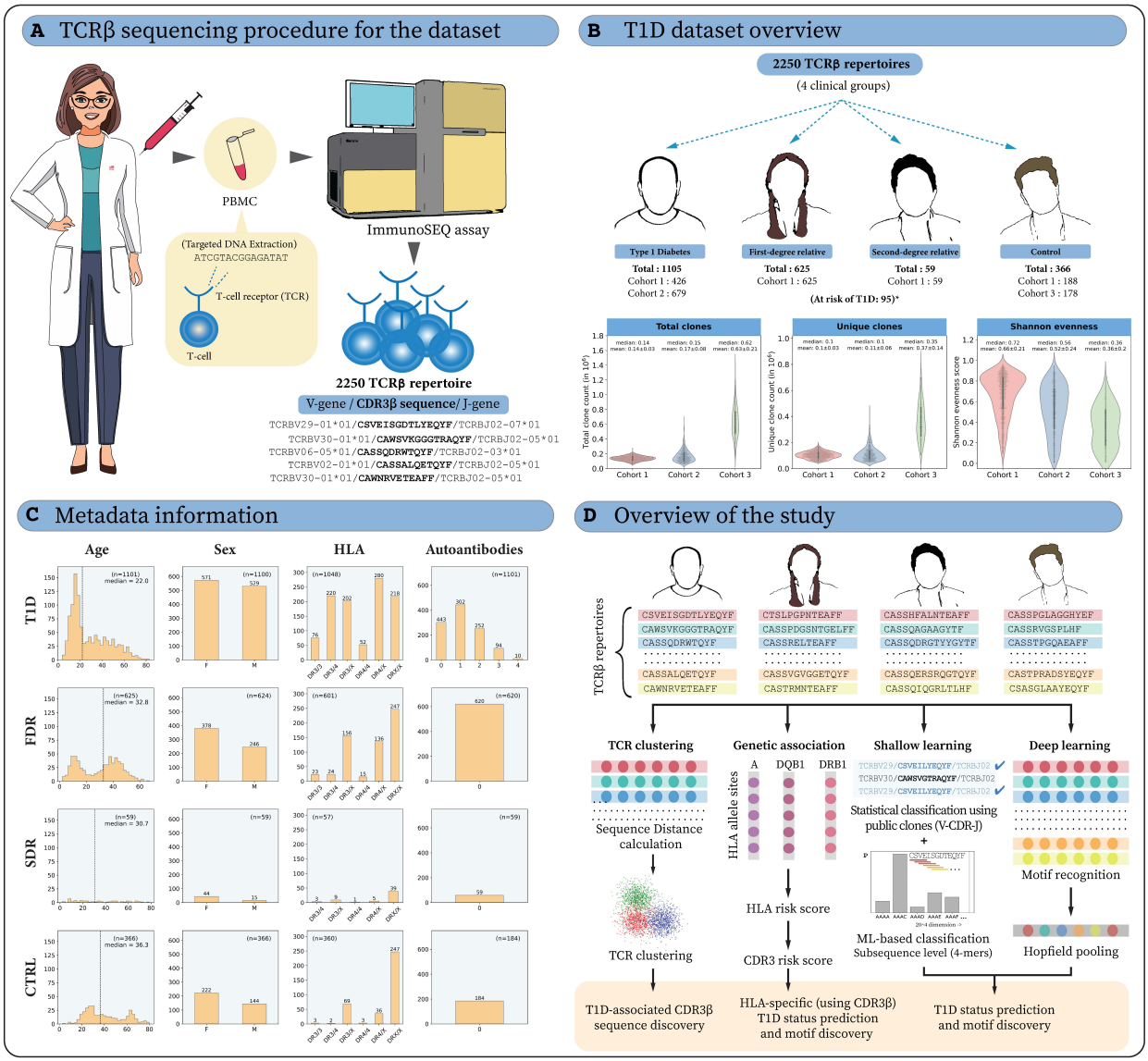

Fig. 1 | Overview of T1D TCRβ cohorts and the study. (A) DNA was isolated from peripheral blood mononuclear cells (PBMC) to conduct TCRβ chain sequencing. **(B)** The dataset of 2250 TCRβ repertoires contains four clinical groups: T1D (T1D), first-degree relatives (FDR), second-degree relatives (SDR), and non-related controls (CTRL) (*excluding 95 individuals without diabetes who have an increased risk of developing T1D, i.e. AAb+), which were sequenced in three different cohorts (see **Methods** section for cohort-specific sequencing details). The unique number of clones, total number of clones (unique clones and their clonal frequency) and Shannon evenness are shown for each cohort. Mean and median values are presented above each violin plot. **(C)** The distribution of age (dotted line shows median value), sex distribution, high-risk human leukocyte antigen (HLA) and number of autoantibodies (IAA, GADA, ZnT8A, IA-2A) present (for cohort 1 and 2) are shown for the four clinical groups across all three cohorts. **(D)** In this study, we leveraged available HLA information to conduct HLA and V-gene-restricted TCR clustering and to elucidate the extent to which high-risk HLAs restrict CDR3β sequences. The influence of HLA-mediated T1D risk was further translated into TCR repertoire-based T1D risk and identification of CDR3β motifs. Furthermore, we applied machine learning and deep learning (ML/DL) techniques to classify T1D and healthy related and unrelated controls. These computational approaches were also employed to identify T1D-associated TCRβ sequences and/or sequence motifs.

Supplementary Figures: Fig. S1, Fig. S2

Repertoire-level similarity and diversity do not differ between T1D clinical groups

Repertoire-level similarity and diversity analyses are widely used for cross-patient comparisons of AIRRs [15,21,44–47]. We assessed immune repertoire diversity and similarity across clinical status within cohort 1 and restricted our analysis to the 779 individuals under age 30 to mitigate potential age-related confounding factors [48]. We assessed *TRBV*-gene usage (Fig. 2A) as previous studies in other autoimmune diseases (e.g., SLE, RA) revealed distinct V-gene usage patterns that may be associated with different disease states [16,49]. We observed a small number of statistically significant differences in V-gene usage, but absolute differences were biologically insubstantial. For example, we observed a significant difference in TCRBV19 (p-value: $1.14 * 10^{-04}$) gene usage across clinical groups. Yet, the difference in average frequency of V-genes in T1D (5.49 ± 0.58) and CTRL (5.75 ± 0.66) repertoires was minimal, in line with previous reports on other T1D cohorts [36].

To compare repertoires on the sequence level, we defined public clones as identical CDR3β sequences shared among two repertoires, ignoring the V- and J-gene information. To compare clonal expansion, we calculated the diversity of the immune repertoires using Hill-based diversity profiles [44]. The α-parameterized Diversity profile consolidates many previously established diversity indices (SR, ${}^{\alpha=0}D$ [50,51]; Shannon, ${}^{\alpha=1}D$ [52,53]; Simpson's, ${}^{\alpha=2}D$ [54,55]; Berger-Parker, ${}^{\alpha=\infty}D$ [50,56]). We calculated the correlation between the diversity profile of each repertoire and clustered them according to clinical groups, conventional high-risk HLA types (DR3, DR4), and age (Fig. 2B, C). However, the diversity profile did not cluster based on these parameters.

Furthermore, we calculated the similarity between repertories based on two metrics: (i) **percentage of shared public clones** for all repertoire pairs with respect to the total number of clones present in the smaller repertoire (based on clonal overlap) [57] and (ii) **Morisita-Horn (MH) index** (based on clonal overlap and clonal frequency). Neither the public clones percentage nor MH index clustered based on clinical status, age, or high-risk HLA types (DR3, DR4) (Fig. 2D, F). Although the observations were statistically significant across clinical groups, the differences in median values were minor and lacked biological relevance for percentage of public clones (median values range from 0.033% to 0.036%) and MH index (median values range from 0.002 to 0.008). Similarly, conventional high-risk HLA alleles were not associated with biologically relevant differences in percentages of public clones (median values range from 0.034 to 0.036) and MH index (median values range from 0.005 to 0.01) (Fig. 2E, G). In summary, repertoire diversity and similarity analyses did not reveal biologically meaningful distinctions across clinical groups or HLA genetic risk factors.

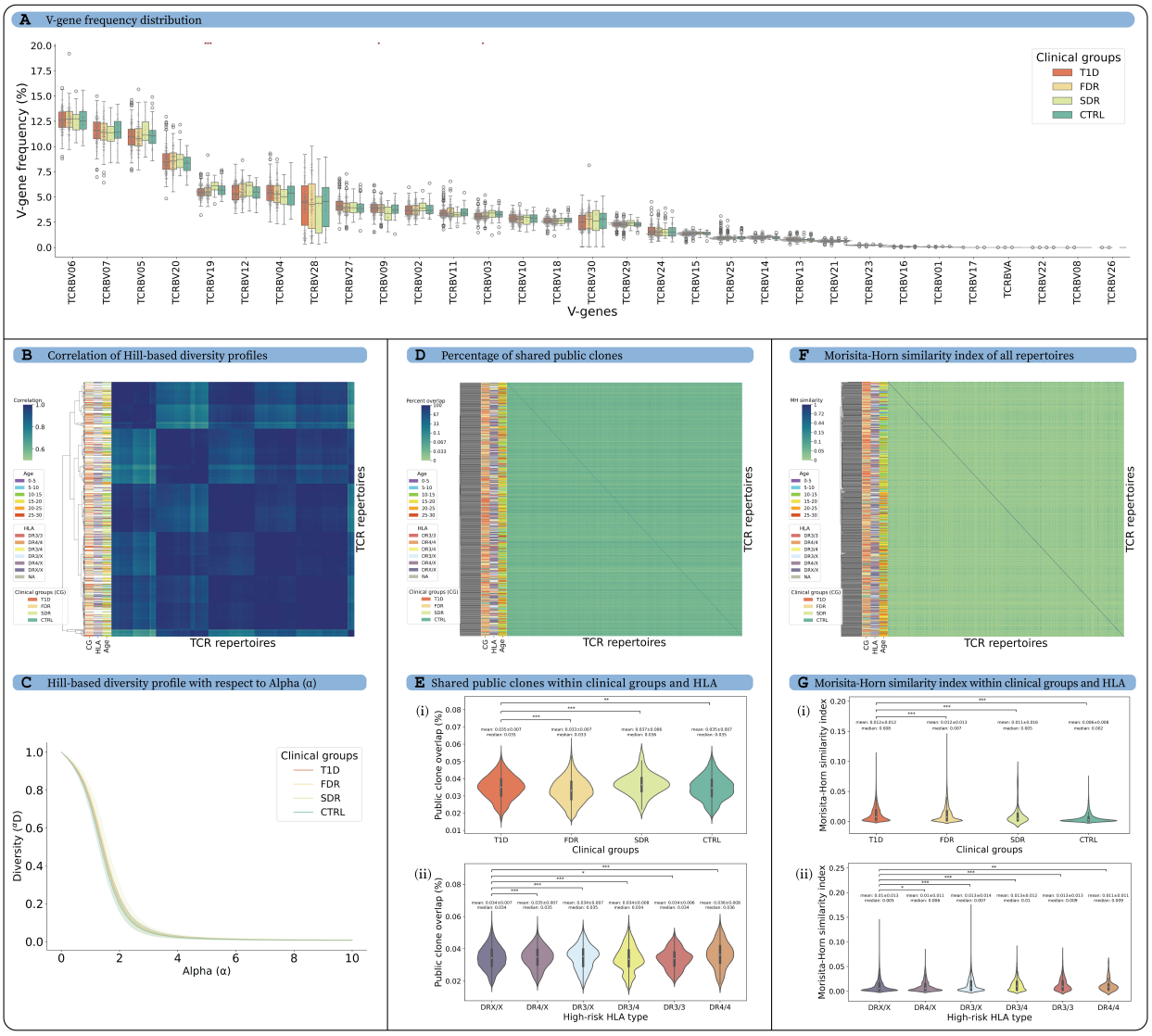

Fig. 2 | The V-gene distribution, diversity and clonal overlap of the TCRβ repertoires were similar across T1D, FDR, SDR, and CTRL groups. The analysis was performed on individuals in cohort 1 younger than 30 years of age, to minimize age-based confounding factors. **(A)** The V-gene distribution was similar across TCRβ repertoires of different clinical groups. The frequency of each V-gene was averaged for all repertoires in each clinical group. The difference in V-gene distribution was tested for all clinical groups using Kruskal–Wallis test, and p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure. **(B)** Hill-based diversity profiles were calculated for alpha (α) value range [0,10] and step size 0.2, and Shannon evenness was used for the undefined value of α=1 for each repertoire. The heatmap shows the Pearson correlation of Hill-based diversity profiles for each pair of repertoires (see **Methods**). **(C)** The average Hill-based diversity profile for each α value of TCRβ repertoires calculated for different clinical groups. **(D)** The calculation of the percentage overlap of public clones involved determining the number of shared clones between two repertoires and dividing it by the total number of clones present in the smaller repertoire. The resulting values were used to create a heatmap that represents the percentage of public clones shared between each pair of repertoires. There was very low overlap among pairs of repertoires (max Public clones overlap = 0.08%). **(E)** The percentage of public clone overlap (i) among the clinical groups and (ii) among the high-risk HLA types (DR3/DR4). **(F)** The Morisita-Horn (MH) similarity index calculated for each pair of repertoire shows low overlap among TCRβ repertoires (Max MH similarity index = 0.34) and no clustering was observed based on age, clinical group, or high-risk HLA. **(G)** The MH similarity index (i) within the clinical groups and (ii) within the high-risk HLA types. All cluster heatmaps were generated using the UPGMA clustering method and Euclidean distance matrix. A universal color scheme was used to show T1D, FDR, SDR and CTRL in red, yellow, light green and dark green, respectively. The significance of difference in distribution was calculated using two tailed Mann–Whitney U tests with respect to T1D or non-high-risk HLA individuals (DRX/X), in respective plots. The p-values were adjusted using Benjamini–Hochberg method and described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001. Only significant p-values were displayed. The

number of individuals in each group were as follows: (i) clinical groups: T1D: 367, FDR: 282, SDR: 29, CTRL: 101 and (ii) high-risk HLA types: DRX/X: 293, DR4/X: 150, DR3/X: 157, DR3/4: 77, DR3/3: 45, DR4/4: 23. In figure **D** and **F**, the linear color scale of the heat map was divided into 2 parts, where midpoint segregates the scale into two separate scales. The lower half encompasses the most common range of values (contains >90% data points) and the upper half accommodates values reaching up to the maximum.

Inflammatory-disease-associated CDR3β sequences are overrepresented in T1D repertoires

Since no discernible repertoire-level similarity or diversity differences were observed across clinical groups without TCR-level antigen information, we next investigated if incorporating publicly available information on T1D and other disease-associated CDR3β sequences from the McPAS and VDJdb databases (Table S1) would show differences between clinical groups [58–60]. Using McPAS data, we found that T1D repertoires had a slightly higher percentage of T1D-associated CDR3β sequences (0.096 ± 0.047) compared to other clinical groups (FDR: 0.093 ± 0.073; SDR: 0.09 ± 0.061; CTRL: 0.094 ± 0.043 with Kruskal–Wallis test p-value of 0.0036; Extended Data Fig. 1A, Table S2). However, these T1D-associated CDR3β sequences were not sufficient to cluster the repertoires based on clinical groups, high-risk HLA, or age (Extended Data Fig. 1B). Similarly, we observed a significant overrepresentation of CD and influenza-associated CDR3β sequences in T1D repertoires [61–63] (Extended Data Fig. 1A, Table S2). T1D and CD have shared genetic drivers, the most notable being shared high-risk HLA (DQ2.5/DQ8) [64]. Clinical guidelines emphasize vaccination for individuals with T1D, and the overrepresentation of influenza-associated CDR3β sequences might reflect increased vaccination compliance at the population level. Using VDJdb data, we observed a significant enrichment in hepatitis C virus (HCV), human immunodeficiency virus (HIV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) associated sequences in T1D repertoires compared to other clinical groups (Extended Data Fig. 1C and Table S2). Despite the absence of a direct genetic link, several case reports have shown potential associations of T1D with HCV [65,66], HIV [67,68], and SARS-CoV-2 [69]. The SARS-CoV-2-specific CDR3β sequences observed in our cohort likely represent potential cross-reactive CDR3β sequences with other human coronaviruses, as all patient samples were collected before the COVID-19 pandemic.

The association of CMV with T1D is still unclear, with prior reports suggesting a lower [70] and higher prevalence [71] of T1D in CMV patients. 2118 CMV-associated CDR3β sequences in McPAS observed no significant association with T1D status in our study. We further utilized a collection of 25,508 unique CDR3β sequences associated with CMV exposure curated from the literature [72] and observed a significant reduction in the presence of CMV-associated CDR3β sequences in the T1D repertories compared to other groups in cohort 1 (Extended Data Fig. 1D). Our observation supports findings of a recent study performed on a larger cohort [70] suggesting early childhood CMV infection decelerates the progression to clinical T1D. In summary, we validated that CDR3β sequences previously linked with T1D are associated with clinical groups within our cohort. However, the strength of these associations was minimal, highlighting the need for a more comprehensive analysis using advanced methodologies. Furthermore, we identified potential links between T1D progression and several other diseases, including CD, influenza, HCV, HIV, SARS-CoV-2, and CMV.

HLA risk alleles demonstrate higher restriction of TCR repertoire diversity in T1D individuals

Limited T1D association of HLA-associated public TCRs

Several autoimmune diseases, including T1D, RA, CD, and SLE, have a substantial component of genetic risk driven by HLA class II genotype [8,73–76]. Given the potential implications of high-risk HLA alleles in modulating thymic selection of autoreactive clones [77] or in promoting expansion of these clones in the periphery [78] (Fig. 3A), several computational models and statistical methods were recently developed to predict the association between the TCRs and HLA [2,3,79]. Thus, in an attempt to control for HLA differences among clinical groups and probe for T1D-specific associations of TCRβ receptor features, we first identified strongly HLA-associated TCRβ receptor features within cohort 2 with the intent of testing those features prevalence and enrichment with T1D status in cohort 1 (see Q1, Fig. 3B).

We tabulated the publicity of all unique cohort 2 TCRs across the cohort 2 repertoires, defining detection as an exact or near-exact match (1 CDR3 AA variation), and computed the odds ratio (OR) of detection in participants with and without each common HLA allele. A TCR feature was assigned to the HLA allele with the lowest p-value (Fisher's exact test). To limit the number of hypotheses tested in the next stage, we filtered TCR features with p-values < 1e-8 and detected in at least 5% of HLA-matched individuals and not more than 10% of HLA-mismatched individuals. This yielded 20,037 unique V-gene family-CDR3β centroids assigned to either MHC class I allele (n = 7,255) or MHC class II (n = 12,782) allele(s) (Fig. S4, Table S3). To determine whether these features were associated with T1D status, we then searched for these features in cohort 1 among participants with the hypothesized restricting HLA allele (Fig. S3). After correcting for multiple hypothesis testing, 39 HLA-DRB1*03 associated TCRs were significantly underrepresented (False discovery rate < 0.2) among the T1D group (Fig. 3C, Extended Data Fig. 2A, Extended Data Fig. 2C). Twelve strongly HLA-associated (B*40:01, B*15:01, DQA1*01:02, DQA1*05:01, DQB1*02:01) TCR features were also overrepresented (False discovery rate < 0.2) in the HLA-matched T1D repertoires versus controls (Extended Data Fig. 2B).

The TCRβ feature most underrepresented in cohort 1 T1D repertoires was TRBV7 CASSLSLAGSNNEQFF, with it or its near-exact neighbor detected in 31.5% (80/254) control and 9.3% (18/192) T1D repertoires of persons expressing the DRB1*03:01 and/or DQA*05:01:DQB*02:01. Many of the other TCR features, strongly underrepresented in T1D repertoires (blue circles in Fig. 3C), also had a similar CDR3β sequence (1-3 mutation distance from TRBV7 CASSLSLAGSNNEQFF) as shown in motifs (Extended Data Fig. 2C). Two of these sequences, V07,CASSLSLAGTYNEQFF and V07,CASSLSLAGAYNEQFF, were previously identified as TCRβ sequences associated with DRB1*03:01 status by statistical testing in a healthy population [1]. These features were observed in repertoires across all age groups among controls; however, because the population without diabetes (FDR/SDR/CTRL) is older than the T1D diagnosed group, we cannot exclude the possibility that the TCR features most depleted in cohort 1 T1D repertoires may reflect age-dependent acquisition of T cell memory due to recurrent exposures and vaccinations. Overall, despite finding many TCRs associated with known T1D-associated HLA risk alleles, we noted little definitive signal of HLA-restricted TCRs enriched in repertoires based on T1D clinical status after conditioning to persons with the relevant allele.

Amino acid (AA) variation in HLA risk alleles restricts the position-wise AA frequency of the TCR repertoire

A statistical framework to assess HLA-based TCR repertoire restriction

Investigations into the genetic basis of type 1 diabetes (T1D) have increasingly highlighted the HLA region as pivotal in shaping immune tolerance and susceptibility [80]. A previous study on 18,832 T1D case-control samples identified three AA sites in HLA-DQβ1 (site 57) and HLA-DRβ1 (sites 13 and 71) as the main drivers of T1D risk [74]. These three positions together explained 90% of the T1D-specific phenotypic variance in the HLA-DRB1–HLA-DQA1–HLA-DQB1 locus and 80% of the variance explained by the entire MHC region. Ishigaki et al. [5] further observed that the polymorphisms in HLA alleles alter the T-cell repertoire. The variation of AA in the above three HLA sites (in DQβ1 and DRβ1) and associated log odd scores were used to calculate the genetic risk of T1D from HLA in our cohorts, termed the HLA risk score [74] (see **Methods**). A TCR sequence-based disease-associated risk score (termed CDR3 risk score) was also established, which quantifies the association of AA variation in HLA alleles and positional AA variation in CDR3β sequences [5].

We computed CDR3 risk scores for all TCR sequences in our most comprehensive cohort 1 containing all clinical groups and then, tested our observations in cohorts 2 and 3 comprised of T1D and CTRL individuals, respectively: (i) To examine the association of CDR3β sequences with HLA allele-based genetic risk, (ii) to identify the high-risk HLA-associated CDR3β AA, utilize them to calculate a repertoire-level CDR3 risk score, and assess whether this score exhibits an association with clinical groups, and (iii) to identify CDR3β sequence motifs that encapsulate HLA-based T1D risk within TCR repertoires (see Q2-4; Fig. 3B).

In keeping with previous work [5], we refer to AA locations in HLA as 'sites' and AA locations within CDR3β as 'positions'. Across eight HLA genes, we identified 398 HLA sites that exhibited AA variation (Extended Data Fig. 3A). The observed variations in the HLA sequences were distributed almost equally between HLA class I genes (218 variations, 54.8%) and HLA class II genes (180 variations, 45.2%). For the computation of a CDR3 risk score, only the CDR3β lengths 12 to 18 (denoted as L12 to L18) were considered for all HLA-CDR3β association analyses as they were the most frequent CDR3β lengths in all repertoires (Extended Data Fig. 3F). To compute the CDR3 risk score, the AA frequencies were calculated for each position of the CDR3β sequences, where 7 AA positions were present for L12 with increments of one position up to L18, due to removal of TRBV- and J-gene encoded regions, leading to a total of 70 CDR3β positions (Extended Data Fig. 3B). We divided the cohort 1 repertoires into four subsets: (i) total cohort 1 (contains all clinical statuses, 1242 repertoires), (ii) T1D only (402 repertoires), (iii) FDR only (601 repertoires), and (iv) CTRL only (182 repertoires). All the analyses in this section were performed for all four cohort 1 subsets. SDR (57 repertoires) were not analyzed as no significant T1D-associated CDR3β AAs were observed in the dataset due to low repertoire count.

HLA sites restrict the position-specific AA diversity of CDR3β sequences

We first validated that CDR3β sequence diversity was restricted as a function of HLA sites in cohort 1, as observed in the previous study [5]. We obtained AA variants for each HLA site from HLA genotyping. To quantify the association between each HLA site and CDR3β position, multivariate multiple linear

regression (MMLR) was used, where a multidimensional vector of CDR3β position-specific AA frequency (response variable) was used to predict the association with multidimensional vector containing polymorphisms at each HLA site (explanatory variable) [5] (Extended Data Fig. 3C). To test interindividual variance in CDR3 AA frequencies explained by the HLA genotype, we subsequently evaluated the MMLR model using a multivariate analysis of variance (MANOVA) (Fig. 3D). A total of 27860 MMLR-MANOVA tests were performed (398 HLA sites x 70 CDR3β positions), and 13313 significant associations (47.8% of total tests; FDR<0.05) were observed after false discovery rate correction using Benjamini-Hochberg adjustment (Table S4). There were 5022 (37.7%) significant associations from class I HLA types and 8291 (62.3%) from class II HLA types. The number of significant associations was more than two times higher than those reported previously [5], as our dataset was also almost double in size. The lowest p-value was observed between *HLA-DRB1* site 13 (p-value = $3.1 * 10^{-214}$), located within the T1D risk-associated peptide binding groove [74], and the CDR3β position 111 for L15. Position 109 in L13 (5th position in our analysis with p-value = $4.7 * 10^{-180}$) was reported in a previous study as having the lowest p-value [5]. Both CDR3β positions were in the antigen-binding region. Separately, we performed MMLR analysis on each T1D, FDR, and CTRL repertoire subset to observe similar HLA-based CDR3β restriction in each clinical group (Fig. S5, Fig. S6 and **Supplementary Note**). Taken together, (i) we confirmed the observation from Ishigaki et al. [5] that AA polymorphisms in HLA restrict the AA positional frequencies of the CDR3β sequences and (ii) extended these observations to T1D, FDR, and CTRL.

HLA-risk-allele-linked CDR3 risk score is highest in T1D individuals

Building on observed associations between variation in HLA sites and AA frequencies within the CDR3β sequences, we next investigated the influence of HLA risk score on position-specific CDR3β AAs (hereafter referred to as CDR3 phenotypes) in TCR repertoires. We first calculated the HLA risk score for each repertoire in cohort 1 (see **Methods**, Extended Data Fig. 3D, Table S5). To identify CDR3 phenotypes associated with the HLA risk score, we conducted a total of 1400 (70 CDR3β positions x 20 AA frequencies at each position) linear regression (LR) tests with HLA risk score (Extended Data Fig. 3E, Table S6). There were a total of 529 CDR3 phenotypes with significant association with T1D risk for cohort 1 (p-value ≤0.05) (Fig. S5). The correlation coefficients from these LR tests were treated as the effect sizes for the respective CDR3 phenotypes, which were further used to calculate the CDR3 risk score by summing up the effect sizes of each CDR3 phenotype for each CDR3β sequence and calculating the average value for the whole repertoire. The CDR3 risk score reflects HLA-based T1D risk in the TCR repertoire, and this relationship was validated by correlating it with the HLA risk score across different CDR3β sequence lengths (correlation coefficient, r = 0.35 to 0.51, Fig. 3E). The HLA-associated CDR3 phenotypes (AAs) exhibit both positive and negative effect sizes. CDR3 risk scores based solely on positive or negative effect sizes showed comparable correlations (r = 0.27 to 0.49) to the combined CDR3 risk score (Fig. S5, Fig. S6). Additionally, average CDR3 risk score was highest for the T1D repertoires followed by FDR, SDR, and CTRL repertoires, respectively, across different lengths (Fig. 3F). A similar observation was also exhibited by classical high-risk HLA allele types, where presence of at least one DR3 or DR4 allele had higher CDR3 risk score compared to other non-T1D risk allele types (DRX/X) (Fig. S7). We assessed the robustness and factors influencing the CDR3 risk score on cohort 1 by analyzing the overlap of CDR3 phenotypes across clinical groups and assessing the significance of CDR3 phenotype effect sizes, respectively (see **Supplementary Note** and Fig. S8, Table S7). CDR3 phenotypes obtained from cohort 1 were also tested on cohorts 2 and 3, where the CDR3 risk score was significantly higher for the T1D repertoires present in cohort 2 compared to CTRL repertoires present in cohort 3, across different CDR3β length (Fig. S9). The

significant correlation with HLA risk score and association with clinical groups were also validated for CDR3 risk score, across each subset of cohort 1 (Fig. S5, Fig. S6).

To summarize, we identified 529 HLA-associated CDR3 phenotypes that were selectively restricted by high-risk HLA alleles associated with T1D. These CDR3 phenotypes and their effect sizes were used to calculate the CDR3 risk score exclusively based on TCR repertoires. A robust association between CDR3 risk score, HLA risk score, and T1D clinical status validates the role of HLA-mediated restriction in shaping TCR repertoires. Additionally, these CDR3 phenotypes were also validated in cohorts 2 and 3.

HLA-associated CDR3β motifs provide a simplified representation of high-risk HLA-based CDR3 restriction

We identified a consistent recurring pattern in charged and aromatic AA in both positively- and negatively-associated CDR3 phenotypes of different CDR3β length (see **Methods**, Fig. S5A, Table S8). The pattern was utilized to derive a simplified representation of the HLA-associated T1D signature within TCR repertoires as positively-associated HLA-motif (pHLA-motif) and negatively-associated HLA-motif (nHLA-motif) (Fig. 3G). We further utilized respective HLA-motifs to calculate an HLA-motif score for each repertoire based on the total number of CDR3β sequences containing these motifs. The pHLA-motif score was significantly higher for T1D (0.00463 ± 0.00056, p-value<0.001) versus FDR, SDR, and CTRL groups (values ranging from 0.00406 to 0.0043) (Fig. 3H, Extended Data Fig. 4). A similar trend was observed for the T1D individuals in cohort 2 (0.00451 ± 0.00057) and CTRL in cohort 3 (0.00404 ± 0.00053). On the other hand, nHLA-motif scores were lower for T1D (0.01028 ± 0.00071) repertoires compared to other clinical groups (values ranging from 0.01049 to 0.01079) (Fig. 3I, Extended Data Fig. 5). Again, a similar trend was observed for the T1D individuals present in cohort 2 (0.0103 ± 0.0008) and CTRL in cohort 3 (0.011 ± 0.0008).

The expected trends for pHLA-motif score (T1D>FDR>SDR>CTRL) and nHLA-motif score (T1D<FDR<SDR<CTRL) were observed, even when at least one HLA allele was non-risk (DRX) and when no islet autoantibody presence was detected (Extended Data Fig. 4, Extended Data Fig. 5). However, the differences observed for the nHLA-motif score were less pronounced. pHLA-motif score showed an expected negative correlation with T1D duration and T1D patient age (Pearson correlation ranging from -0.3 to -0.44), likely due to a reduction in β-cell antigens [81]. However, the nHLA-motif score had a low and inconsistent correlation (Pearson correlation ranging from -0.12 to 0.11) with T1D duration and/or patient age. In conclusion, the distinct pHLA- and nHLA-motif scores reflect the influence of HLA-associated T1D risk on TCR repertoires, suggesting these motifs as potential markers of HLA-driven susceptibility and protection in T1D.

The presence of heterozygous HLA alleles restricts TCR diversity

A recent murine study reported that the presence of heterozygous MHC class II alleles constrains the diversity of the TCR repertoire [82]. To investigate this observation in humans, we tested the hypothesis in our dataset. First, we calculated the generation probability (Pgen) using Igor, a tool that models the V(D)J recombination process to estimate the likelihood of generating a specific TCR or B-cell receptor (BCR) sequence [83]. Pgen reflects the probability of a sequence being produced purely by the random recombination of V, D, and J gene segments, along with nucleotide insertions and deletions. Subsequently, we calculated

the post-selection probability (Ppost) using SoNNia, which distinguishes sequences that have undergone selection. In our study, Pgen and Ppost were used as proxies to assess TCR repertoire diversity [84].

Our analysis revealed a weak trend in cohort 1, where increased HLA heterozygosity appeared to be associated with reduced TCR diversity, as indicated by both Pgen and Ppost (Extended Data Fig. 6A, B). This trend was not observed in cohorts 2 and 3, potentially due to HLA biases arising from presence of only T1D repertoires in significantly larger cohort 2 and fewer CTRL repertoires in cohort 3, respectively. Additionally, we observed that frequencies of certain V-genes (e.g., TRBV28, TRBV4-3, TRBV3-2, TRBV3-1) clustered into a distinct group (Extended Data Fig. 6C). We used TRBV28 as a case study and observed V-gene frequencies as low-frequency ($f < 0.015$), mid-range frequency ($0.015 \leq f \leq 0.04$), and high-frequency ($f > 0.04$) clusters. Although a direct association between V-gene frequency clusters and specific clinical groups was not identified, a potential link with certain HLA alleles was observed. (Extended Data Fig. 6D). However, these V-gene frequency clusters could be attributed to confounding factors, such as age, ethnicity, infection history, or sequencing artifacts, which fall outside the scope of the current study. These observations also contrast a previous study on 666 individuals, indicating HLA (class I)-heterozygous individuals present a broader immunopeptidome for recognition by cytotoxic T cells [85]. Altogether, our findings were consistent with the observations from the murine study [82], where HLA heterozygosity was shown to limit TCR repertoire diversity.

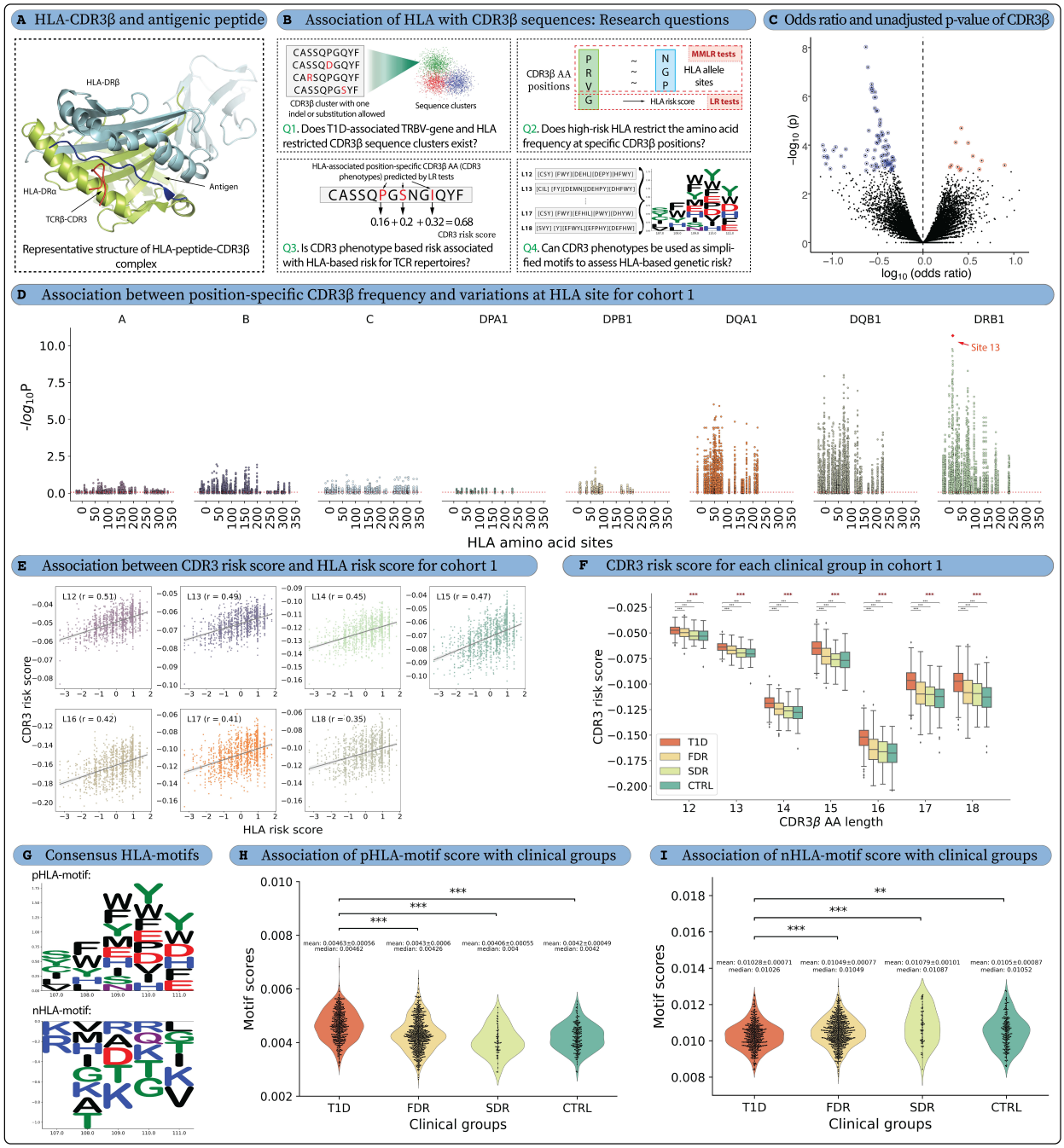

Fig. 3 | High-risk HLA alleles restrict the AA frequency of CDR3β sequences. (A) A representative complex structure of the human CDR3β region in close proximity to Influenza HA Antigen Peptide and MHC Class II Molecule, HLA-DR4 **(**PDB id: 1J8H), showcasing the potential impact of HLA on CDR3β sequences **(B)** This study addressed four key questions related to HLA restriction: **Q1** finds the T1D-associated TRBV-gene and HLA-restricted full length CDR3β sequences that are overrepresented or depleted in the TCR repertoires, **Q2** investigates the restriction of CDR3β AA frequencies by HLA leading to identification of HLA-associated position-specific CDR3β AAs (CDR3 phenotypes), **Q3** calculates the risk score for each repertoire based on CDR3 phenotypes to compares it with HLA-based risk score and clinical groups, and **Q4** identifies the positively- and negatively-associated HLA-motifs from CDR3 phenotypes and assesses their association with clinical groups, classical high-risk HLA alleles (DR3/DR4), autoantibody presence and T1D duration. The first three questions were addressed for all TCR repertoires in cohort 1 with available HLA information (n=1242) and individually for each clinical group, including T1D repertoires (n=402), FDR repertoires (n=601), and CTRL repertoires (n=182) to assess the generalisability and robustness of the analysis. The number of SDR (n=57) repertoires was insufficient to obtain significant CDR3 phenotype and, therefore, SDR were excluded from this analysis. **(C)** We observed few TRBV-gene family- and HLA-associated CDR3β sequences depleted (blue) or overrepresented

(red) in the T1D repertoires vs FDR, SDR, CTRL in Cohort 1, when considering nearly-identical CDR3 with one indel or substitution allowed (Q1). The plot shows odds ratio on x-axis and unadjusted p-values of y-axis. **(D)** The MANOVA test p-values from the MMLR analysis for the 1242 repertoires (cohort 1) showed association between CDR3 positions and HLA sites, where variation in CDR3β AA frequency at each CDR3β position was plotted for each HLA site containing mutation(s) (Q2). An LR model was also developed with HLA risk score and CDR3β AA frequency to identify CDR3 phenotypes. These CDR3 phenotypes were utilized to compute CDR3 risk score and HLA-motifs from TCR repertoire. **(E)** A significant correlation was observed between T1D-associated HLA risk scores and CDR3 risk scores calculated for each repertoire in cohort 1 (Q3). **(F)** CDR3 risk score was also higher for the T1D clinical group and observed the expected trend where T1D>FDR>SDR>CTRL (Q3) **(G)** The positively- (pHLA-motif) and negatively- (nHLA-motif) associated HLA-motif obtained from the CDR3 phenotypes, where **(H)** pHLA-motif showed higher presence in T1D repertoires and **(I)** nHLA-motif showed expected opposite trend with higher presence in CTRL repertoires. In the figure, Multiple testing was performed using Kruskal–Wallis test (denoted with red stars), and p-values were adjusted between different CDR3β lengths. p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests, and p-values were adjusted between different clinical groups. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. p-values were described as * for [0.01,0.05], ** for [0.001,0.01], *** for <0.001, and no stars plotted for non-significant values.
Supplementary Figures: Extended Data Fig. 2, Extended Data Fig. 3, Extended Data Fig. 4, Extended Data Fig. 5, Fig. S4, Fig. S3, Fig. S5, Fig. S6, Fig. S7, Fig. S8, Fig. S9

Machine learning enabled TCR repertoire classification

The classification of immune repertoire for disease status prediction is based on the principle that disease signatures (identical or similar) are shared across individuals affected with the same disease [15]. However, there are several approaches to identify disease signatures associated with disease [86]. In our study, we used previously utilized HLA risk scores [74] and three different ML/deep learning (DL) approaches, namely, (i) public clone (V-CDR3-J)-based shallow ML, (ii) k-mer based shallow ML, and (iii) attention-based DL, to classify the T1D repertoires. Across all ML approaches, T1D status was considered the positive class and the remaining clinical groups (FDR, SDR, CTRL, together termed "no diabetes" [ND]) were considered non-T1D or negative class for training, tuning, and testing within cohort 1. The combined set of cohorts 2 and 3 (only T1D and only CTRL repertoires, respectively) was used as a held-out test set. Due to class imbalance, we used AUROC as the primary metric for evaluating predictive performance. Moreover, a confounder adjustment procedure was applied for DeepRC and k-mer based logistic regression (LogReg) models through a series of steps aimed at adjusting the impact of age on the classification of subjects based on their T1D status (see **Methods**).

T1D status classification based on HLA only or HLA-based TCR features achieve an AUROC ranging from 0.66 to 0.85 on test cohorts

We first used classical high-risk HLA alleles, DR3 and DR4, to assess T1D status classification. As expected, we found that the presence of high-risk HLA alleles (DR3 and/or DR4) alone is insufficient to accurately predict the clinical status of T1D (see **Supplementary Note,** Extended Data Fig. 7A), reflective of the complex, multifactorial nature of this disease [87]. However, homozygosity for either DR3 or DR4 is a strong predictor of T1D status (91.4% in cohort 1 and 97.2% in cohort 2). We also utilized HLA risk score [74] to classify the T1D status (Fig. 4A). The AUROC curve was 0.73 for cohort 1 and 0.85 for the held-out test set (cohort 2 and 3) (Fig. 4B). However, balanced accuracy was ~67% for both training and held-out test datasets, due to class imbalance.

We further investigated the performance of the CDR3 risk score, which was essentially an effect of high-risk HLA on the TCR repertoire. As the CDR3 risk score was calculated for different CDR3β lengths, we obtained AUROC values between 0.65 to 0.75 (balanced accuracy ranging from 50.4% to 65.8%) on cohort

1 and 0.66 to 0.80 (balanced accuracy ranging from 52% to 72.7%) on cohort 2+3 (Fig. S10). The pHLA and nHLA motifs can be considered a simplified output of genetic association study, and therefore, tested for their classification performance. The pHLA motif obtained AUROC of 0.68 (balanced accuracy 64.8%) on cohort 1 and 0.73 (balanced accuracy 68.6%) on cohort 2+3. Whereas, nHLA motifs expectedly had comparatively lower AUROC of 0.59 (balanced accuracy 50.2%) on cohort 1 and 0.73 (balanced accuracy 50%) on cohort 2+3 (Fig. S10). The performance metrics for the aforementioned classifiers are also summarized in the supplementary (Table S9).

The presence of public clones is insufficient for the classification of T1D status

Emerson and colleagues [17] developed a statistical classification framework that could diagnose CMV status from peripheral blood TCRβ sequences. Their approach identified the statistically significant enrichment of TCRβ sequences in CMV^{+} compared to CMV^{-} individuals. We applied this approach on the T1D dataset using our open-source ImmuneML [88] platform (Fig. 4C). The model performance was tested on 5-fold cross-validation (CV), where log-loss value was predicted to be the least at p-value threshold of 0.0001 (log-loss 0.417 for training and 0.4 for test; Fig. S11, Table S10). The balanced accuracy and AUC values of the model at the optimal p-value threshold were low for the training cohort 1 (average AUROC: 0.615 and balanced accuracy: 56.1% across 5 splits) (Fig. 4D). Testing the model on cohort 2+3 showed relatively high AUROC (0.956) and balanced accuracy (62.2%). However, increased performance in the test dataset was likely due to the sequencing depth bias between cohorts 2 and 3 since test cohort precision (100%) was significantly higher than recall (24.4%) (Fig. S11). A total of 140 TCR clones were predicted to be associated with T1D (Table S11). A low overlap, ranging from 21% to 28%, was observed among T1D-associated TCRβ sequences across each CV split (Fig. 4D). In summary, the statistical classification approach based on public clones was insufficient to classify repertoires, due to the low prevalence of public clones.

K-mer based logistic regression achieves an AUROC of 0.78 on test cohorts

In the k-mer based repertoire classification approach, a sliding window of 4-mers was employed for each CDR3β sequence, and the occurrence of k-mers was calculated for each immune repertoire (see **Methods**). The k-mer frequency matrix was utilized as a feature in logistic regression (LogReg)-based ML models. The final model was obtained by averaging the output of 5-folds into a single model after sigmoidal activation (Fig. 4E). There can be several confounding factors related to T1D affecting the TCR repertoires [89]. The dataset was age-corrected by applying sample weights to avoid more prominent age-related confounding factors (see **Supplementary Notes**) [90]. The LogReg model was fitted on the training dataset (cohort 1) and yielded an AUROC of 0.78 on the held-out test dataset (cohort 2 and 3) (Fig. 4F). The model showed balanced accuracy of 73.25% on the held-out test dataset with a sensitivity of 70.1% and specificity of 76.4% (Table S9). We also observed that the prediction score obtained from the LogReg model did not demonstrate any association with the clinical groups, high-risk HLA types (DR3 and/or DR4), or islet autoantibody status (Fig. S12). Contrary to common observation, LogReg scores exhibited a positive correlation (r = 0.12) with T1D duration. Whereas, patient age had negative correlation with LogReg score across all clinical groups (r = -0.45 to -0.5) except for T1D (r = 0.15). It is important to note that we observed overfitting on the training dataset even after regularising the coefficients (see **Methods**). We subsequently employed a more sophisticated DL model to enhance the classification accuracy of immune repertoires.

DL-based multiple instance learning achieves an AUROC of 0.77 on test cohorts

Significant developments have been made in applying DL methods for the classification of immune repertoires based on disease status [17,31,91,92]. The Deep Repertoire Classification (DeepRC) model is one such modular and customizable method particularly suited for large-scale multiple-instance learning problems including immune repertoire classification [31] (Fig. 4G). Similar to the LogReg approach, the DeepRC model was also trained using a 5-fold CV, where three splits were designated for training, one for tuning, and the remaining one for testing, in a recursive manner. The best-performing model for each CV fold was selected based on the AUROC of the tuning set, resulting in five selected models. The AUROC of these five best-performing models ranged from 0.69 to 0.76 (Table S9) on the remaining testing set in cohort 1. We further trained an ensemble LogReg model to combine the predictions of the 5 best models, which improved the AUROC to 0.79 (balanced accuracy 72.5%). The ensemble model was considered the final DeepRC model and applied to the held-out test set, resulting in the AUROC 0.77 (balanced accuracy 72.9%) (Fig. 4H).

The DeepRC model exhibited performance comparable to the LogReg model on the held-out test set and observed association with clinical groups, high-risk HLA types (DR3/DR4), autoantibody status, and age in cohort 1 (Extended Data Fig. 8). Notably, DeepRC predictions based solely on the TCR repertoire demonstrated improved classification performance compared to HLA-associated TCR features (such as CDR3 risk score, pHLA- and nHLA-motif scores). However, its performance was lower when compared to HLA-based genetic risk (HLA risk score). Interestingly, DeepRC's predictions also showed a positive correlation with T1D duration, similar to the k-mer-based LogReg model, while exhibiting a negative correlation with the age of T1D patients. Additionally, DeepRC allowed for the extraction of a simplified sequence motif representation from the trained model, facilitating a biologically meaningful interpretation of the DL model, as discussed in the section below.

DL-identified T1D-associated CDR3β motifs achieve an AUROC of 0.7 on training and test cohort

Interpretability of the above presented DL model is a challenging task; however, DeepRC [31] supports different methods of interpretability, via the attention values and contribution analysis method known as Integrated Gradients [93]. We extracted a low-complexity DeepRC-motif to identify a biological signature of T1D (see **Methods**) and calculated the DeepRC-motif score by normalizing the number of motif-containing sequences with the total number of CDR3β sequences. The DeepRC-motif achieved an AUROC of around 0.7 on both training dataset (balanced accuracy 65.67%) and held-out test dataset (balanced accuracy 66.48%) (Table S9). The DeepRC-motif showed association with clinical groups, high-risk HLA types (DR3 and/or DR4), autoantibody status, and age in cohort 1, similar to the DeepRC model. However, T1D duration showed a negative correlation (r = -0.13) with the DeepRC-motif score (Extended Data Fig. 9). Taken together, DeepRC provided a low-complexity alternative in the form of a motif, which demonstrated an association with clinical groups, despite being derived solely from TCR repertoires and without HLA information. It is noteworthy that the DeepRC-motif is a gapped motif with two optional positions out of total five positions, which may exhibit a higher rate of false positives in identifying T1D-associated CDR3β sequences, as ~17% of the CDR3β sequences in a repertoire contained a DeepRC motif, compared to only ~0.4% for the pHLA motif.

The generation of pHLA-motifs by aggregating position-specific AAs from CDR3β sequences of varying lengths presents considerable challenges. Nevertheless, only ~0.4% of the repertoires contained these motifs. To enhance specificity, we obtained refined consensus HLA-associated CDR3β sequences by identifying the CDR3β sequences that contain both pHLA- and DeepRC-motif, termed "consensus-motif". This approach reduced the proportion of HLA-associated CDR3β sequences to approximately 0.25%, without reducing the performance compared to pHLA-motifs. Additionally, AUROC of consensus-motif (0.74) was marginally improved compared to pHLA-motif (0.73) (Fig. S13). Expectedly, the consensus-motif score was higher in T1D donors and increased alongside the number of autoantibodies and risk HLA (Fig. S13). It also showed negative correlation with both patient age and T1D duration (Fig. S13).

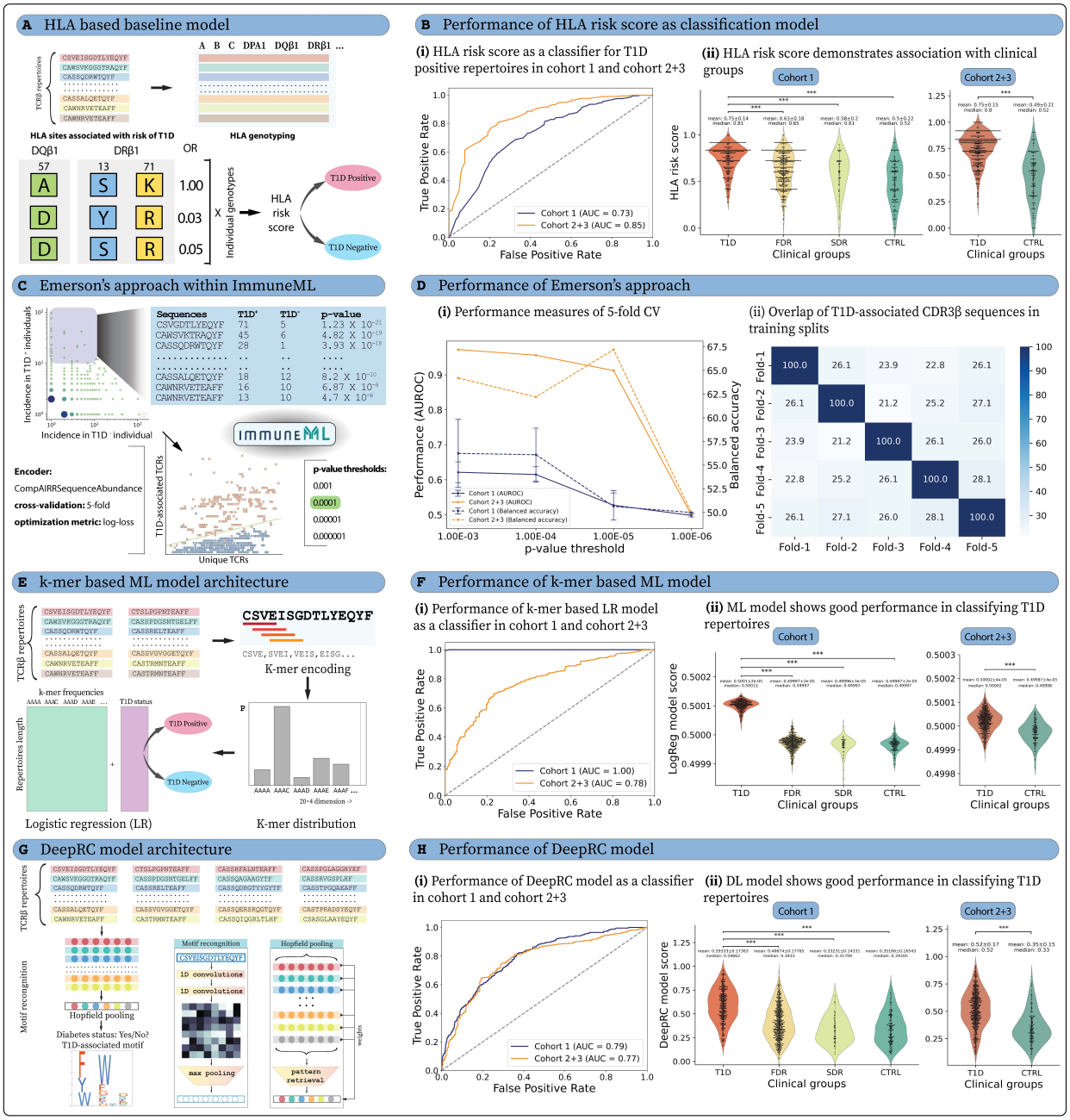

Fig. 4 | The application of different repertoire classification methods to classify T1D and non-T1D repertoires shows varying levels of prediction performance. (A) Schematic representation of T1D status classification using the HLA risk score (baseline model). HLA risk score calculated based on 3 AA positions (HLA-DQβ1: site 57 and HLA-DRβ1: site 13 and 71) was utilized to classify T1D and non-T1D repertoires. **(B)** HLA risk score was used for classification in cohort 1 and cohort 2+3, where (i) AUROC was 0.73 for cohort 1 and 0.85 for cohort 2+3 and (ii) HLA risk score was higher for T1D clinical group with expected trend of T1D>FDR>SDR>CTRL. **(C)** The immuneML platform was utilized to replicate the approach utilized by Emerson et al., [17], where T1D-associated TCRβs were identified and used in a classification model. **(D)** The model was trained on Cohort 1 based on log-loss calculations and 5-fold cross-validation (CV) and tested on cohort 2+3, where (i) AUROC was significantly lower for cohort 1 compared to cohort 2+3 for all p-value thresholds and (ii) overlap of identified T1D-associated sequences across different splits was also relatively low (21% - 28%). **(E)** Schematic representation of repertoire classification using k-mer frequency as a feature in machine learning models. We calculated the frequencies of 4-mers to create a matrix of $n*20^4$, where n is the number of repertoires. We further used Logistic regression (LogReg) for the classification of T1D positive and negative repertoires. **(F)** The LogReg prediction showed overfitting in cohort 1, where (i) AUROC of cohort 1 was 1.0 and reduced to 0.775 for cohort 2+3. (ii) Overfitting was also visible when LogReg predictions in cohort 1 were stratified based on clinical groups. T1D score was significantly higher compared to the non-T1D clinical groups and almost no difference was observed between FDR, SDR and

CTRL in cohort 1. **(G)** Schematic representation of the Deep learning approach (DeepRC) used for T1D disease status classification and DeepRC-motif identification. TCRβ repertoires were represented as collections of CDR3β sequences, which exhibit variability between individuals. A convolutional neural network (CNN) model was implemented to recognize motif patterns in each of the sequences and map them to sequence-representations. Subsequently, a pooling function aggregated these sequence representations into a repertoire-level representation. **(H)** DeepRC was more reliable compared to the LogReg model. (i) The AUROC values were similar for both cohort 1 (0.79) and cohort 2+3 (0.77) and (ii) DeepRC predictions also observed the expected trend of T1D>FDR>CTRL with SDR having the least score. In all violin plots, p-values for pairwise testing with respect to the T1D clinical group were calculated using two tailed Mann-Whitney U tests and p-values were adjusted between different clinical groups using the Benjamini–Hochberg method. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values.
Supplementary Figures: Extended Data Fig. 7, Extended Data Fig. 8, Extended Data Fig. 9, Fig. S10, Fig. S11, Fig. S12

T1D-associated TCR motifs as predictive markers in independent cohorts

T1D genetic risk loci are associated with TCR motifs in peripheral blood and pancreatic lymph nodes

Despite significant progress in defining genetic loci that contribute to T1D risk via Genome-wide association studies (GWAS), our understanding of the impact for such variants on immune function remains limited. We have shown that high-risk HLA alleles restrict the TCR repertoires, allowing for potentially autoreactive TCR presence in genetically-predisposed individuals. Thus, we hypothesized that T1D risk variants, particularly in the HLA region, may be associated with increased frequency of the pHLA-motif and DeepRC-motif. To address this, we performed microarray-based precision medicine genotyping [41] of 716 unrelated living participants from cohort 1 (ND, n=489; T1D, n=227).

To understand which T1D risk variants contributed to the association with the enriched TCRβ motifs, we performed quantitative trait locus (QTL) analysis. *HLA-DQA1**05:01-*DQB1**02:01 (p=3.91e-10), *HLA-DRB1**0301 (p=5.83e-10), *HLA-DQA1**01:02-*DQB1**06:02 (p=3.47e-09), *HLA-DRB1**15:01 (p=2.38e-07), *HLA-DQA1**03:0X-*DQB1**03:01 (p=8.5e-05), and *HLA-DQA1**05:05-*DQB1**03:01 (p=3.23e-4) haplotypes were significantly associated with DeepRC-motif frequency independent of disease status such that risk haplotype correlated with an increased motif score (Fig. 5A, B). The risk alleles of a T1D-associated variant tagging the *XL9* super enhancer (p=3.91e-10), known to regulate *HLA-DRB1* and *HLA-DQA1* expression [94], and an intergenic *HLA-DRA1-DRB1* variant (p=1.85e-3) were likewise associated with increased peripheral blood DeepRC-motif score (Fig. 5A, B). An intergenic deletion nearby *CTLA4* carrying risk for T1D that is in linkage with reduced *CTLA4* expression quantitative trait loci (eQTL) [95,96] was also weakly associated with increased DeepRC-motif score (p=0.015) (Fig. 5A, B). Analysis of the pHLA-motif revealed enrichment in risk allele-carrying subjects for *HLA-DRB1**0301 (p=0.018), *HLA-DQA1**05:01-*DQB1**02:01 (p=0.018), and *HLA-DQA1**05:05-*DQB1**03:01 (p=0.021) (Extended Data Fig. 10A, B) similar to that observed for the DeepRC-motif. The nHLA-motif frequency was decreased in those with the T1D risk haplotypes *HLA-DQA1**02:01-*DQB1**02:02 (p=9.14e-08), *HLA-DQA1**03:0X-*DQB1**03:02 (p=9.85e-06), and HLA-DR4 (p=1.1e-05) (Fig. S14A, B). Thus, variants affecting HLA class II type in addition to HLA class II and *CTLA4* expression levels may influence the frequency of T1D-enriched TCRβ motifs in blood.

Observations in peripheral blood were tested in TCRβ-sequenced sorted conventional $CD4^+CD127^+$ T cells (Tconv), $CD4^+CD127^-CD25^+$ regulatory T cells (Treg), and $CD8^+$ T cells from pancreatic lymph nodes (pLN) as well as spleen from 27 Network for Pancreatic Organ donors with Diabetes (nPOD) donors (data

previously published [37]: ND, n=7; T1D, n=15; T2D, n=3; other diabetes, n=2) in order to understand matters of cell type and tissue specificity. DeepRC- and pHLA-motif scores were significantly increased in $CD8^+$ T cells in the spleen and pLN as compared to Tconv and Treg in the spleen (Fig. 5C, Extended Data Fig. 10C). Treg DeepRC- and pHLA-motif scores were also higher than those of Tconv in the spleen (Fig. 5C, Extended Data Fig. 10C). nHLA-motif score was enriched in splenic Tconv and Treg as compared to spleen and pLN $CD8^+$ T cells (Fig. S14C). T1D subjects showed a trend for increased DeepRC (p=0.095, Fig. 5C) and pHLA (p=0.067, Extended Data Fig. 10C) motif scores and increased nHLA (p=0.025, Fig. S14C) motif score in pLN $CD8^+$ T cells as compared to those with ND. Of the peripheral blood findings, *HLA-DRB1**15:01 (p=0.004) was also significantly associated with decreased DeepRC TCR motif frequency in pLN $CD8^+$ T cells. nHLA motif score was decreased in *HLA-DQA1**02:01-*DQB1**02:02 pLN $CD8^+$ T cells (p=0.226, Fig. S14C), in contrast to the increased frequency in peripheral blood (Fig. S14A). These data support the notion that HLA class II-mediated protection from T1D may limit the frequency of potentially autoreactive TCRβ motifs at the site of autoimmunity.

Similar analysis was performed in TCRβ-sequenced sorted $CD45RO^-CD27^+CCR7^+CD95^-$ naive (TN), $CD45RO^-CD27^+CCR7^+CD95^+$ stem cell memory (TSCM), $CD25^{hi}CD127^{lo/-}$Treg, and $CD45RO^+CD27^+$ central memory (TCM) $CD4^+$ T cell subsets from peripheral blood of 28 living study participants (data previously published [36]: ND, n=14; T1D, n=14). DeepRC- and pHLA-motif scores were significantly higher in TSCM and Treg than TN and TCM (Fig. 5D, Extended Data Fig. 10D). nHLA motif score was increased in Treg as compared to TN and TCM, while TSCM also showed higher frequency than TCM (Fig. S14D). Despite showing lower DeepRC-motif score compared to other cell subsets, TCM DeepRC-motif score was significantly enriched in T1D as compared to ND subjects (p=0.0047, Fig. 5D), which can partially be accounted for by increased motif score in *HLA-DQA1**05:01-*DQB1**02:01-carrying subjects (p=0.028, Fig. 5D). *HLA-DQA1**05:01-*DQB1**02:01 was also associated with enrichment of the pHLA-motif score in Treg (p=0.031) and TCM (p=0.027) (Extended Data Fig. 10D). Together, these data suggest that T1D HLA risk genetics may confer selection pressure for T1D-enriched TCRβ motifs in peripheral blood, particularly in $CD4^+$ TCM and Treg.

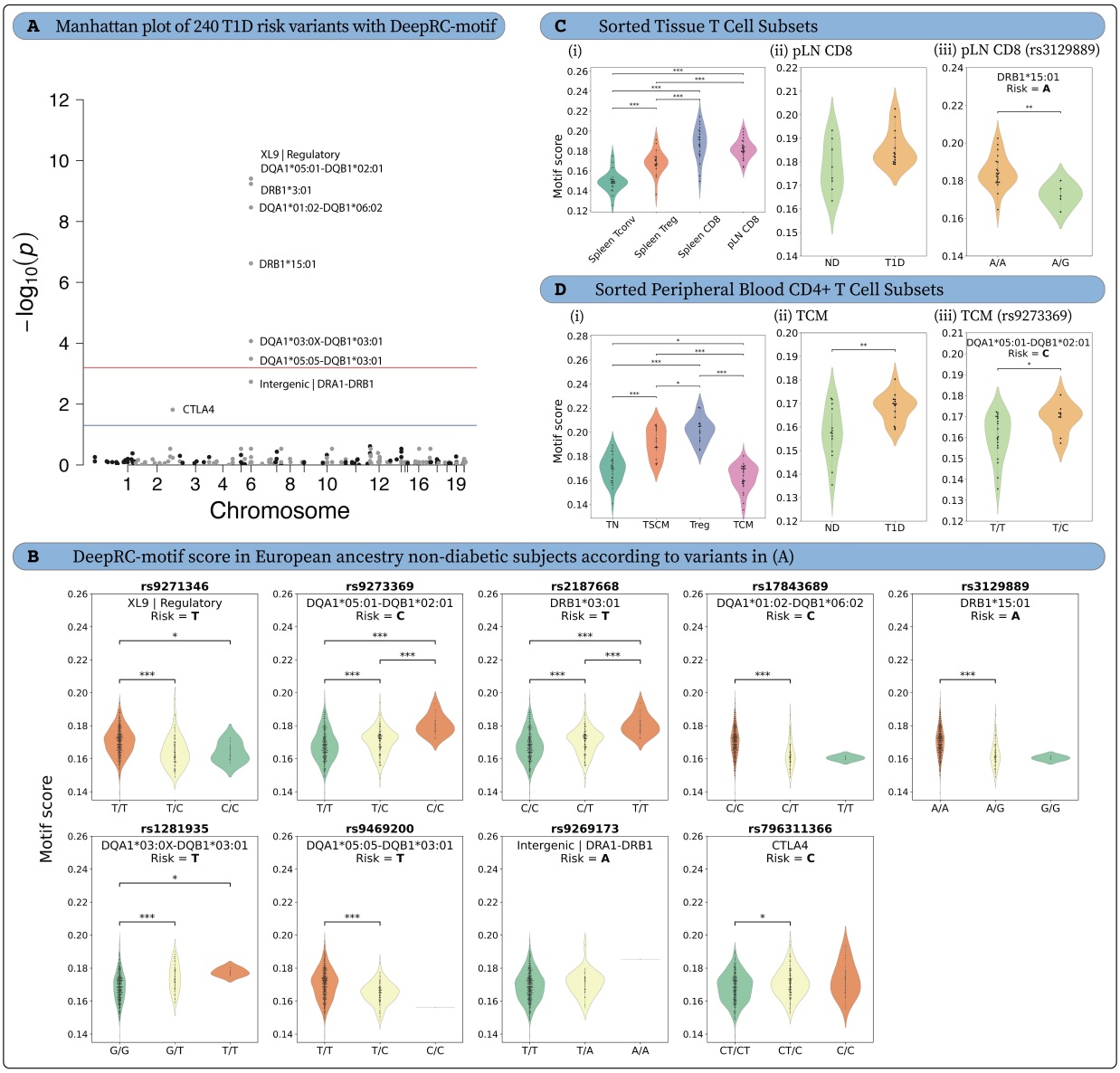

Fig. 5 | DeepRC motif frequency enriched in carriers of T1D risk genetics in bulk and sorted central memory CD4⁺ T cells from peripheral blood as well as pancreatic lymph node CD8⁺ T cells. (A) Manhattan plot of 240 T1D risk variants versus DeepRC-motif score in bulk peripheral blood. Linear regression assuming additive genotypic effect with age, sex, T1D status, predicted probability of CMV infection, and 10 multidimensional scaling components as covariates. Benjamini-Hochberg false discovery rate threshold at p=5.73e-4 (red line) and a conventional threshold of p = 0.05 (blue line). **(B)** Violin plots of DeepRC-motif score in European ancestry non-diabetic (ND) subjects, according to variants in **(A)**. Gene or tagged HLA genotype and risk allele annotated above each plot. **(C)** DeepRC-motif score across sorted T cell subsets in spleen and pancreatic lymph node (pLN). **(i)** Mixed-effects analysis with Tukey's multiple comparisons test. **(ii)** pLN CD8⁺ T cell DeepRC-motif score according to T1D status. **(iii)** pLN CD8⁺ T cell DeepRC-motif score according to DRB1*15:01 tag SNP. Linear regression with age, sex, diabetes status, and 10 multidimensional scaling components as covariates. **(D)** DeepRC-motif score in sorted peripheral blood CD4⁺ T cell subsets. **(i)** Mixed-effects analysis with Tukey's multiple comparisons test. **(ii)** Central memory CD4⁺ T cell DeepRC-motif score according to T1D status. **(iii)** Central memory CD4⁺ T cell DeepRC-motif score according to DQA1*05:01-DQB1*02:01 tag SNP. Linear regression with age, sex, and T1D status as covariates. In all violin plots, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests and p-values were adjusted between different clinical groups using the Benjamini–Hochberg method. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values.

Supplementary Figures: Extended Data Fig. 10, Fig. S14

Discussion

The adaptive immune repertoire is shaped by a myriad of factors, including aging, environmental exposures, and genetics. T1D natural history studies identified islet autoantibody seropositivity as early autoimmune signals in genetically predisposed populations [97,98]. Yet there is a need to improve disease prediction and monitoring through the validation of additional cellular biomarkers. It has been proposed that the TCR repertoire, as a record of response to foreign and self-antigens, could potentially distinguish a T1D-specific signal from repertoire shifts caused by pathogen exposures and vaccinations. In this study, we have comprehensively assessed the HLA-based TCR repertoire restriction, TCR repertoire classification, and immune signal identification on a cross-sectional dataset of 2250 TCR repertoires spanning various stages through the natural history of T1D. We initially observed that repertoire-level diversity, similarity, and V-gene usage do not differentiate T1D donors from their relatives without diabetes or unrelated controls. However, mapping known T1D antigen-specific clones [59] in our dataset revealed an enrichment of T1D-associated sequences in T1D repertoires. This mirrors recent data that demonstrated stable enrichment of CDR3β sequences matching known preproinsulin reactive clones during disease progression in a genetically at-risk cohort, thus providing support for T1D reactive sequences as a biomarker in T1D [35]. Despite this, we also found T1D-associated sequences frequent in non-T1D donors. This is in accordance with previous reports suggesting phenotypic rather than frequency differences in circulating autoreactive cells in T1D [99], and may also be impacted by an incomplete understanding of antigen reactivities and receptor sequences in T1D, altogether complicating the utility of *a priori* sequences and necessitating a more comprehensive analyses of our bulk repertoire data.

We next examined how high-risk HLA molecules, which account for most of the genetic risk in T1D [8,74], influence T1D-associated TCR features. Interestingly, clustering of public TCRs yielded more significantly underrepresented hits associated with HLA risk allele *HLA-DRB1**0301 in T1D, while we identified overrepresented TCR features associated with risk HLA-DQ molecules *DQA1**0501 and *DQB1**0201 in T1D. The observation of a majority of significant underrepresented features is intriguing given previous data suggesting persistent low grade viral infections are associated with autoimmunity development [100], or that certain infections may be protective [70,101]. Alternatively, our observations could reflect deficits in regulatory subsets in T1D; however, while differences in diversity have been shown for other circulating T cell subsets in T1D, this is not the case for Tregs [36]. Thus, further investigation of the identity and phenotypes of T cells expressing these differentially abundant sequences is warranted.

We further analyzed the HLA-DQβ1 and HLA-DRβ1 variants associated with high risk for T1D [74] to examine their effect on TCR repertoire restriction [5]. We identified several TCR features, including a repertoire-level CDR3 risk score and subsequence-level CDR3 phenotypes, which indicate HLA-based T1D-susceptibility imparted solely by TCR repertoires. Additionally, we identified HLA-associated CDR3β motifs by aggregating CDR3 phenotypes, which were positively (pHLA) and negatively (nHLA) associated with T1D. Importantly pHLA-motifs were enriched in T1D donors and exhibited an increased frequency of hydrophobic residues, particularly in the CDR3β middle region, a feature of self-reactivity [102,103]. Interestingly, pHLA-motif scores exhibited a negative correlation with disease duration and patient age, potentially indicative of a waning autoimmune response associated with concomitant loss of β cell antigens [81]. To test the potential for HLA alleles to influence the diversity of the TCR repertoire, we modeled T cell selection by calculating the probability of generation and selection of TCR sequences. In accordance with a recent murine study [82], we observed, although weakly, that heterozygosity limits repertoire diversity. This runs contrary to some observations that the highest risk HLA DQ2/DQ8 genotype allows for transdimer formation, thus potentially allowing for presentation of a more diverse pool of antigens [104] or permitting cross-reactivity [105]. Moreover, recent work suggests that HLA-DM mediated

editing is reduced for HLA-DQ2 and HLA-DQ8 pMHC molecules [106]. While it is thought that lack of DM editing would favor presentation of peptides with more rapid dissociation rates [107] and thus, allow for broader peptide presentation, it is still unclear how lower versus higher stability of pMHC complexes affect the diversity of the repertoire [108]. These analyses exemplify the multifaceted influence of HLA and the complexity of immune regulation in T1D, while providing further insights into the interplay between genetic risk and TCR repertoire diversity.

Considering that most prior analyses have focused on HLA as a primary determinant of T1D risk, we applied several ML strategies to identify features indicative of T1D status independent of HLA information. Although the performance of HLA-dependent TCR features (pHLA and nHLA motifs; AUROC: ~0.73) was lower than the HLA risk score (AUROC: 0.85), the classification of T1D status without explicitly utilizing HLA allele information achieved notable performance with the K-mer-based LogReg strategy (AUROC: 0.78) and DeepRC (AUROC: 0.77). While the K-mer-based LogReg model exhibited overfitting on the training dataset, the DeepRC model proved more interpretable and capable of identifying the input sequence with the highest contribution to the T1D prediction per repertoire. These identified sequences were subsequently analyzed to derive a T1D-associated motif. We found motif scores to be increased in T1D donors and to be increased alongside the number of islet autoantibodies and risk HLA, altogether providing support for this receptor motif as being involved in T1D pathogenesis. Similar to our pHLA-motif, we once again found a negative association of DeepRC-motif score with T1D duration. A caveat for our analysis is the potential for false positives and non-specific sequences included in our ML motif. The frequency of CDR3β sequences containing the DeepRC motif was approximately 17%, whereas the estimated precursor frequency of T1D antigen specific T cells is around 1-10 per million cells [109], closer to the frequency of sequences possessing the pHLA-motif. It is possible that the high frequency of DeepRC-motif-containing sequences could be reflective of incorporation of sequences that are not restricted by high-risk HLA alleles, cross-reactivity [110], or bystander activation, any of which could be important in T1D progression. Thus, further investigation is needed to examine the epitope reactivity of sequences containing this motif. Moreover, we expect incorporation of additional information such as paired TCRαβ sequences and cell phenotype features (e.g., activated effector memory states) to further improve the classification accuracy.

We found T1D risk alleles, primarily within the HLA region, to control the frequency of the DeepRC-motif. Notably, risk genotypes at *HLA-DQ* and *HLA-DR* loci, as well as regulatory element *XL9*, were associated with a higher motif score. We recently reported that high-risk HLA alleles contributed to increased HLA expression on circulating monocytes [111]. Our motif score eQTL [95] results provide further support for the notion that enhanced HLA class II expression modulated by risk alleles could promote the activation and expansion of lower affinity autoreactive clones in T1D. Interestingly, the only eQTL hit outside of the HLA region passing the significance threshold was located within *CTLA4*, and was previously linked with reduced CTLA4 expression. Notably, other T1D risk-conferring variants within *CTLA4* are known to negatively impact *CTLA4* expression and function [112] and thus, are thought to result in reduced TCR activation threshold or impaired regulatory signaling.

HLA-DQ restricted proinsulin and HLA-DR restricted GAD-specific T cells have been identified within the insulitic lesion [113] or peripheral blood [38] as well as the pLN and spleen of T1D donors [37]. To explore the tissue relevance of the identified DeepRC-motif and the cell type localization of motif bearing sequences, we interrogated the motif in our previously published sorted bulk-sequenced nPOD tissue dataset [37]. We found that motif bearing sequences were increased in $CD8^+$ T cells and Tregs in the spleen and pLN, with increased motif score in pLN $CD8^+$ T cells of T1D donors, which was modulated by genotype at protective HLA class II loci. The observation of HLA class II loci modulating $CD8^+$ T cell phenotype likely indicates an indirect effect whereby risk alleles promote inflammatory $CD4^+$ T cell phenotype or alter Treg phenotype

to permit this enrichment [114], though the mechanisms controlling this need further defining. Accordingly, in a separate validation cohort [36] consisting of a sorted $CD4^+$ T cell subset from the PBMC of T1D and CTRL individuals, we identified an enrichment in motif score among $CD4^+$ TCM cells in T1D donors, which was impacted by HLA-DQ risk genotype. $CD4^+$ TCM cells have the capacity to produce cytokines such as IL-2 and IL-21, which drive $CD8^+$ T cell proliferation and effector phenotype [115]. Notably, alterations in the peripheral blood $CD4^+$ TCM compartment, specifically involving increases in T follicular helper (TFH) cells, have been reported in T1D and prior to overt disease [116,117]. In fact, an altered TFH-like signature is detectable as early as infancy in genetically at-risk children who later progressed to T1D, implicating genetics as a driving factor in enhancing helper and effector T cell function in T1D. Thus, we have identified a replicable TCRβ motif which encodes T1D status, is relevant in circulation as well as disease relevant tissue and is controlled by T1D risk loci.

One key aspect of our study is that HLA-based TCR restriction analysis did not account for clinical group information, while FDR and SDR repertoires were classified as ND in the ML models. Nevertheless, most analyses, including the DeepRC model, HLA and CDR3 risk scores, and DeepRC- and pHLA-motif scores, revealed a trend where the scores were highest for T1D repertoires and lowest for CTRL, following the expected pattern: T1D > FDR > CTRL. Furthermore, both the DeepRC-motif and pHLA-motif were overrepresented in individuals with genetic risk for T1D across several independent cohorts, including bulk and sorted $CD4^+$ TCM cells from peripheral blood as well as pLN $CD8^+$ T cells. Notably, we observed a limited presence of public TCRs in our dataset compared to previous studies focused on other autoimmune diseases [17,35,118]. This discrepancy may be attributed to high variability in confounding factors, including age, T1D duration, genetic background, and ethnicity of the individuals. A large longitudinal HLA-stratified cohort could potentially enhance the identification of T1D-specific public TCRs. In this study, we integrated the assessments: (i) the influence of genetic risk on TCR repertoires with HLA-based restriction analysis, and (ii) influences of T1D-associated genetic and environmental factors on TCR repertoires through ML-based methodologies. This integration underscores the critical role of HLA in T1D risk and highlights adaptive immune motifs associated with T1D risk.

Limitations of this study

As with all AIRR studies, sequencing depth remains a critical concern, particularly for autoimmune diseases such as T1D, where immune signals in peripheral blood may be relatively low [21,22]. The detection of an immune signature is further hindered by limited information on cell populations and antigen specificity. While the repertoires in the current study maintain an average of unique CDR3β sequence count > 100,000 (cohorts 1 and 2), deep sequencing in cohort 3 yielded nearly a 3.5-fold increase in unique CDR3β sequences. Although analysis of replicates demonstrated a high MH index and strong Pearson correlation between deep and shallow sequencing data, the loss of additional unique CDR3β sequences at lower sequencing depth may impact the identification of rare, T1D-specific signals. Furthermore, repetition in data collection centers introduced duplicate repertoires with different donor identification numbers into the dataset. While we applied an MH similarity threshold to exclude duplicates, there remains a possibility that some duplicate repertoires were retained in the dataset.

Furthermore, repetition in data collection centers resulted in the inclusion of duplicate repertoires associated with different donor identification numbers in the dataset. To address this, we applied an MH index

similarity threshold of 0.4 to exclude duplicates. Nevertheless, there remains a possibility that some duplicate repertoires were not excluded from the dataset.

The HLA risk score is derived from three HLA positions with the strongest association with T1D risk. A more comprehensive HLA risk score incorporating additional HLA positions could improve T1D risk prediction. HLA risk score is fixed and thus, cannot serve as a biomarker for T1D disease progression, diagnosis, or assessing the effects of T1D treatment.

Islet autoantibody seroconversion peaks early in childhood, at around two years of age [119]. As this represents a critical period of exposure to pathogens and vaccinations, it will be extremely important to understand the longitudinal stability of the TCR repertoires in children to effectively identify TCR-based signatures and monitor their shifts over time in diseases such as T1D. A more extensive longitudinal sampling could also provide insights into how other diseases or infections influence the TCR repertoire at specific time points. Additionally, a larger cohort of HLA-matched individuals would be ideal for identifying HLA-specific signals associated with T1D.

It is also essential to acknowledge the limitations associated with derivation of HLA-motifs. HLA-motifs were constructed by aggregating position-specific AA of different CDR3β lengths, which complicates the interpretation of positional dependence. Furthermore, motifs derived from aggregated positional information do not permit the retrieval of AA-specific positional dependencies and are primarily suited to produce ungapped motifs. The DeepRC-derived motifs had unexpectedly higher proportions within each repertoire, indicating a substantial rate of false positive CDR3β sequences and emphasizing the need for more effective motif extraction methods. A comprehensive study is also warranted to examine the role of underrepresented protective motifs within the T1D cohort, potentially enhancing T1D risk prediction.

A future strategy toward identifying disease-relevant biomarkers may benefit from combining PBMC-based sorting with tissue-based, spatially resolved, and cell-population-specific sampling. That being said, we showed that the T1D-associated TCR signals identified were also present in entirely different datasets [34,120], tracking with HLA-risk status. Additionally, T1D antigen-specific TCRs were also observed in CTRL individuals, warranting further studies on low-affinity, cross-reactive signals and physiological autoreactive processes [121]. Biomarker specificity could be further enhanced by incorporating paired alpha and beta chain information [122]. Identifying epitopes reactive with TCRs exhibiting overrepresentation or depletion in T1D repertoires could provide valuable insights into underlying disease mechanisms and help refine associated motifs. To circumvent large-scale experimental screening for antigen specificity (which is currently unfeasible), there is a need for AI-based deorphanization of TCR epitopes [123–127].

Conclusions

There is growing interest in harnessing the immune history encoded within TCR repertoires to develop TCR-based diagnostics and therapeutics for future clinical applications, including immunization [128,129], viral infections [17,130], and autoimmune diseases [40,131]. We conducted a large-scale, comprehensive analysis of T1D-associated TCR repertoire alterations by immunosequencing the CDR3β regions and genotyping HLA using sampling amenable to young donor samples for screening efforts. The HLA-associated T1D risk was leveraged to extract robust HLA-mediated TCR features, including repertoire-level CDR3 risk score and

CDR3β subsequence-level HLA-motifs, to predict T1D risk. Additionally, we predicted T1D status directly from TCR repertoires using ML and DL approaches. DeepRC performed comparably to HLA-mediated TCR features and identified a simplified DeepRC-motif from TCR repertoires alone, without needing HLA information. These potential motif-based biomarkers, derived from cross-sectional samples, reflect the natural history and pathogenesis of T1D, and show associations with HLA, disease duration, and islet autoantibody status. These findings not only support future longitudinal studies but also, introduce disease-associated motifs that can be tracked in response to clinical interventions aimed at halting disease progression, such as T cell-targeting therapies like teplizumab (anti-CD3, Tzield), anti-thymocyte globulin (ATG), Alefacept (LFA3-Ig fusion protein), and Abatacept (CTLA4-Ig fusion protein) [132]. As studies expand in organ donor tissues from individuals with T1D, this work is expected to provide a valuable reference dataset for identifying antigen reactivities and convergent or public receptor sequences enriched in the pLN and pancreas.

Methods

Study participants and sample collection

Study participants or their legal guardians provided written informed consent with pediatric and adolescent participants also providing assent prior to enrollment. Cohort 1 was comprised of individuals from outpatient clinics at the University of Florida (UF; Gainesville, FL), Nemours Children's Hospital (Orlando, FL), and Emory University in accordance with Institutional Review Board (IRB) approved protocols at each site. Peripheral blood samples were collected into sodium heparin-coated and serum separator vacutainer tubes by venipuncture from non-fasted individuals (i.e., unknown prandial state or time of day), then shipped or rested overnight prior to processing at the UF Diabetes Institute (UFDI). At the time of blood draw, participants were generally healthy with no reported malignancy or infection, and sample collection occurred between 2010-2018 prior to the coronavirus disease 2019 (COVID-19) pandemic. All samples and their associated data and metadata were deidentified in accordance with UF IRB201400703. Cohort 2 included subjects with T1D from the T1D Exchange Clinical Network [133,134] and cohort 3 included controls sequenced by Adaptive Biotechnologies.

DNA isolation and TCRβ sequencing

Genomic DNA (gDNA) was isolated from peripheral blood mononuclear cells (PBMCs) of 2250 subjects. *TRB* (TCRβ) CDR3 region sequencing was performed via Adaptive Biotechnologies immunosequencing assay. Briefly, gDNA was amplified using bias-controlled multiplex PCR prior to sequencing [135,136]. Cohort 1, containing 1393 repertoires, comprising 188 CTRL, 59 SDR, 625 FDR, 426 T1D repertoires, was sequenced using hsTCRB_v4 Service at shallow sequencing depth [135,136]. There were 95 additional autoantibody positive non-T1D repertoires (AAb+) in cohort 1, which were excluded to avoid potential bias in the study. Cohort 2, containing 679 T1D repertoires, was sequenced using hsTCRB_v4b Service at shallow depth. Cohort 3, containing 178 CTRL repertoires, was sequenced using hsTCRB_v4b Service at deep sequencing depth. The shallow and deep technical replicates of cohort 1 were studied separately and were not part of the above mentioned cohort.

Genotyping

Subjects from cohorts 1 (n=1242), 2 (n=645), and 3 (n=178) were genotyped using the UFDIchip custom microarray, with processing on an Affymetrix GeneTitan instrument and a BioMek FX dual arm robotic workstation [41]. The UFDIchip includes >9,000 markers from the Axiom™ Precision Medicine Research Array (Thermo Fisher Scientific) covering the HLA region. Raw data were converted to genotype calls using Axiom™ Analysis Suite software (v3.0, Thermo Fisher Scientific) "Best Practices Workflow" with "Human.legacy.v5" settings. Four-digit HLA genotypes were imputed using Axiom™ HLA Analysis software (v1.2.0.38) [137]. Imputation results were used for analysis if probability scores > 0.7. HLA class II haplotypes considered to carry T1D risk were defined as DR3 (*HLA-DRB1**03:01–*HLA-DQA1**05:01–*HLA-DQB1**02:01) and DR4 (*HLA-DRB1**04:01/02/04/05/08–*HLA-DQA1**03:01–*HLA-DQB1**03:02/04) [8]. Subjects were grouped based on HLA class II risk diplotypes, with non-risk haplotypes designated as "DRX", resulting in the following groups: DRX/X, DR3/X, DR3/3, DR4/X, DR4/4, or DR3/4.

Islet autoantibody measurement

GADA, IA-2A, and ZnT8A were measured from serum using commercial ELISA kits, which consistently demonstrate high sensitivity and specificity in the Islet Autoantibody Standardization Program (IASP) workshops [138]. IAA was measured from serum by chemiluminescence assay, as recently reported and evaluated by the IASP [139].

Dataset preprocessing

Firstly, we filtered the Out-of-frame sequences from each repertoire and selected repertoire, which contains a minimum of 50,000 distinct sequences. In this study, we defined a clone or clonotype as a unique CDR3β sequence, unless specified otherwise. The clone counts for duplicate CDR3β amino acid sequences were merged. Additionally, V/D/J-gene information from different columns was consolidated to address unresolved V/D/J-gene annotations. Repertoires with a potential common donor origin were excluded by applying a Morisita-Horn (MH) similarity threshold of less than 0.40, based on thresholds established in previous studies on replicate repertoires [140,141].

Statistical analysis

All cluster heatmaps were generated using the UPGMA clustering method and Euclidean distance matrix [142]. The p-values for the multiple testing were performed using Kruskal–Wallis test. Similarly, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method.

Repertoire-level similarity and diversity analysis

V-gene usage analysis

The V-gene distribution within each immune repertoire was determined by assessing the frequency of occurrence for each V-gene. This frequency was obtained by dividing the count of a specific V-gene by the total number of unique clones present in the repertoire. The V-gene frequencies were calculated without considering the clonal frequency.

Hill-based diversity profile analysis

The diversity of the TCR repertoires were calculated using Hill-based diversity profile, which is based on Rényi's definition of generalized entropy [143–145]. It is defined as:

$$^{\alpha}D(f) = \left(\sum_{i=1}^{n} f_i^{\alpha}\right)^{\frac{1}{1-\alpha}},$$

Where f is the clonal frequency distribution with f_i being the frequency of each clone and n the total number of clones. The α-values represent weights, which means as α increases, higher frequency clones are weighted more. The α-parameterized Diversity generates a diversity index profile for a given array of α values. The diversity profile is not defined for $\alpha = 1$. However, diversity tends towards Shannon entropy when α tends to 1 (based on L'Hospital's rule). We calculated the diversity profile of each repertoire based on α value ranging [1,10] with step size of 0.2. The above analyses were performed on individuals with age less than 30 in cohort 1 to avoid confounding factors due to age [48].

Public clone analysis

In order to evaluate the similarity between TCR repertoires, we defined public clones as identical CDR3β sequences shared among two repertoires. The percentage of shared public clones is defined as:

$$Public\ clones\ (\%) = \frac{|X \cap Y| * 100}{min(|X|, |Y|)}$$

Where, $|X|$ and $|Y|$ are the repertoire sizes (number of unique clones) of the repertoire X and Y. The number of shared clones between both repertoires are $|X \cap Y|$.

Morisita-Horn similarity index analysis

We employed the MH similarity index [146] to assess the degree of similarity among TCR repertoires, taking into account the clonal frequency attributed to each unique clonotype. The MH index between a pair of repertoires is defined as:

$$MH = \frac{2\sum_{i=1}^{S} x_i y_i}{\sum_{i=1}^{S} x_i^2 + \sum_{i=1}^{S} y_i^2}$$

Where, S is the number of unique clones, and x and y denote the frequency of ith clone in either repertoire. The MH index ranges between 0 (no overlap) and 1 (complete clonal overlap and identical clonal frequencies).

Overrepresentation analysis of different disease-associated CDR3β sequences

We downloaded the McPAS [59] and VDJdb [60] datasets from the respective websites (April, 2024). The unavailable, redundant and non-human sequences were removed from the dataset. Only pathology (in McPAS)/antigen species (in VDJdb) with more than 30 sequences were selected for analysis. The final statistics of the datasets is given in Table S1. We obtained the overlapping CDR3β sequences between cohort 1 and McPAS/VDJdb datasets using CompAIRR [147]. We considered the clonal frequency of the overlapping sequences in each repertoire in cohort 1 and normalized the count using the following formula and grouped them by the clinical groups.

$$CDR3\beta\ frequency\ =\ Log_e(CDR_f/R_T)$$

Where, CDR_f is the clonal frequency of the overlapping CDR3β clone in a repertoire in cohort 1 and R_T is the sum of clonal frequencies in the same repertoire.

HLA-conditional T1D-association testing in single AA variation neighborhood

We first considered each unique TCRβ sequence within T1D positive repertoires from Cohort 2. The sequences were partitioned by the IMGT V-gene family and all receptor sequences from 669 repertoires were concatenated to map each sequence to its exact and near-exact (single AA variation distance) neighbors. To avoid $O(n^2)$ comparisons on such a large set of sequences, we identified nearly identical sequences using a fuzzy clustering technique. This method bins highly similar sequences into shared memory locations using keys with a single masked position, inspired by prior work on TCR clustering [129,147–150]. Briefly, each CDR3 was represented as the set of possible sequences with a single masked wildcard position for either a substitution or insertion in a Python dictionary. Each unique key was then assigned an integer index and two arrays link CDR sequence indices to masked CDR3 key indices. Group-by key and permutation operations were then applied for rapid identification of all links between identical and near-identical CDR3 sequences. After identifying single AA variation sequence neighbors of each unique CDR3 centroid, only CDR3s with at least five total neighbors in Cohort 2 were retained for further analysis. Next, we considered the donors associated with each CDR3 sequence index and computed the odds ratio (OR) of detecting a CDR3 or its near-exact single AA variation neighbor across donors with class I or class II HLA alleles, computing a p-value by a Fisher's exact test.

A TCR feature centroid was assigned to the most statistically significant HLA allele if it had an OR > 1 with a p-value < 1e-8, was detected in > 5% of HLA-matched individuals and not in more than 10% of HLA-mismatched individuals. From cohort 2, this procedure yielded 20,037 unique V-gene family-CDR3β centroids assigned to either a MHC class I allele (n = 7,255) or MHC class II allele(s) (n = 12,782). Next, for each strongly HLA-associated feature, we tabulated detection in Cohort 1 repertoires of participants with the hypothesized restricting HLA allele. (We considered a detection if the query TCR had an exact or near exact match to a TCR with a with-in repertoires frequency greater than 1 in 500,000 templates). Based on detections in HLA-matched repertoires we computed the odds ratio of observing the centroid or its neighbor in participants based on T1D clinical status and computed a p-value based on Fisher's exact test (with the Python package *fishersapi* based on *fast-fisher*), and computed and adjusted p-value based on the

Benjamini-Hochberg procedure [151] to control false discovery rate using the Python package *statsmodels*. To visualize potentially T1D underrepresented and overrepresented sequence motifs derived from sequences with a false discovery rate < 0.2 we clustered these sequences into graphs using the Python package *tcrdist3* [152] and *networkx* [153] at an edge threshold of 24 TCRdist units and plotted motifs based on aligned sequences within each graph connected component.

Quantification of HLA-based restriction of CDR3β sequence diversity

We studied the association of the HLA and CDR3β sequences using published methods [5] and observed the impact of T1D-risk HLA alleles on selective restriction of CDR3β sequences. The analysis was primarily performed on cohort 1 and observations were validated on cohort 2+3. To assess the robustness of the findings, we also performed the analogous analysis on T1D, FDR and CTRL repertoires separately. SDRs (59 repertoires) were not included in this analysis due to insufficient sample number. We considered A, B, C, DPA1, DPB1, DQA1, DQB1, DRB1 HLA haplotypes in the study and variations in the HLA alleles were obtained from genotyping experiments. The rare or common alleles for each HLA haplotype were removed from the analysis based on their frequency in cohort 1 (frequency ≥ 99% or <1%). Furthermore, four-digit classical alleles were partitioned based on AA polymorphism at the HLA site (Extended Data Fig. 3A).

Position-specific CDR3β AA frequency analysis

Firstly, CDR3β sequences were segregated based on the length. We considered CDR3β lengths ranging from 12 to 18 (described as L12–L18 in the manuscript) as most of the CDR3β sequences (94.17% of total sequences) were within this range (Extended Data Fig. 3B). All analyses focused on the international ImMunoGeneTics information system (IMGT) [154] CDR3 positions 107 to 116 that directly contact antigens [129] (Extended Data Fig. 3B, F). The flanking positions in CDR3β sequences defined by germline-encoded V- or J-genes were identified using IMGT/HighV-QUEST for each length group and removed from the CDR3β sequences. The frequency of each AA was computed at every position within the CDR3β sequence for length L12 to L18. This computation involved dividing the total count of a particular AA by the count of all AAs at a specific position of CDR3β sequence (Extended Data Fig. 3B).

Multivariate multiple linear regression model (MMLR)-multivariate analysis of variance (MANOVA) analysis

In simplest terms, the MMLR analysis is performed using n-dimensional independent variables to predict the n-dimensional dependent variables. In our study, the frequency of all AAs at a specific position of the CDR3β sequence was used to calculate the association with counts of all AA polymorphism at an HLA site (Extended Data Fig. 3C) [155]. We also included the top three principal components of all HLA genotypes in this analysis. Therefore, the full MMLR model was the following equation.

$$y_i = \sum_{a=1..m-1} \beta_a g_{a,i} + \sum_{k=1,2,3} \pi_k PC_{k,i} + \theta$$

Here, y_i is n-dimensional CDR3β amino acid frequency vector of ith repertoire, β_a is an n-dimensional parameter that represents the additive effect per group of classical alleles containing the same AA at specific HLA site (denoted by a in equation), $g_{a,i}$ is the allele count of allele group a in i^{th} repertoire. We included

m − 1 group of classical alleles, casting 1 group as the reference. π_k is an n-dimensional parameter that represents the effect of the kth principal component and $PC_{k,i}$ is the value of the kth principal component of ith repertoire. θ is an n-dimensional parameter that represents the intercept.
The null model only had the terms for covariates without allelic effects.

$$y_i = \sum_{k=1,2,3} \pi_k PC_{k,i} + \theta$$

The improvement in the model fit between the full and null model was estimated using MANOVA and the significance of the improvement was assessed using Pillai's trace. Codes from a previous study (https://github.com/immunogenomics/cdr3-QTL) were modified and reused accordingly in the study using Python and R languages [5].

HLA risk score

The mutations in HLA-DQB1 at site 57 and HLA-DRB1 at site 13 and 71 have previously been associated with T1D disease progression based on 18,832 case-control samples of T1D [74]. In this analysis, we grouped the classical four-digit HLA-DQB1 and HLA-DRB1 alleles based on the AAs present in these three sites along with the T1D risk in terms of OR (Table S5). The HLA risk score was calculated based on the sum of the individual OR scores multiplied by the number of alleles containing a specific AA polymorphism (Extended Data Fig. 3D). In case multiple combinations were possible for the three AAs, then we consider the combination with the highest OR score.

CDR3 risk score

The association of CDR3β amino acids with HLA risk score is referred to as CDR3 risk score. We developed linear regression (LR) models between the HLA risk score and CDR3β frequency of each AA at specific CDR3β position (for all lengths running from L12-L18) as described previously [5,155]. The p-values obtained from the LR model were further adjusted for multiple testing using the Benjamini–Hochberg method. The correlation coefficients of the significant association between the AA frequency and HLA risk score were denoted as effect size of that particular T1D-associated CDR3β AA or CDR3 phenotype (Extended Data Fig. 3E). The number of significant associations were dependent on the number of repertoires in the study. Therefore, we observed a higher number of associations for cohort 1.

The cumulative effect sizes of CDR3 phenotypes were calculated for the whole CDR3β sequence, and CDR3 risk score for the entire repertoire was derived by averaging the cumulative effect sizes of all CDR3β sequences (Extended Data Fig. 3E). Further, we calculated the correlation between HLA risk score and CDR3 risk score for each length and also compared CDR3 risk scores across clinical groups.

High-risk HLA-associated motif identification from CDR3 phenotypes

The CDR3 phenotypes, derived from the HLA risk score, were transformed into sequence motifs to evaluate their association with clinical groups. The AAs were grouped based on the IMGT positions, and a cutoff of effect sizes ≥0.1 was employed to select significant associations. Positive and negative effect sizes were analyzed separately to identify positively-associated (pHLA-motif) and negatively-associated (nHLA-motif) motifs with HLA-based T1D risk. The analysis was focused on CDR lengths L13 to L16, which had

a higher occurrence in the repertoires (see Table S8). Furthermore, IMGT position aligned AAs showed the repetitive behavior of AAs on the C-terminal end, leading to selection of the conserved regions (IMGT positions 107-111). The effect sizes of consensus motifs in the logo plot represent the average effect sizes of CDR3 phenotypes for length L13 to L16.

Comparative analysis of CDR3 risk score across T1D clinical statuses

We calculated CDR3 phenotypes and their effect sizes across repertoires from different clinical groups, including T1D, FDR and CTRL. Notably, an overlap was found in CDR3 phenotypes across the clinical groups. To appropriately test for association between CDR3 risk score and clinical groups, the CDR3 phenotypes and their corresponding effect sizes, derived from a specific clinical group, were applied to calculate the CDR3 risk score across all clinical groups. It is crucial to emphasize that the CDR3 phenotypes for each clinical group were computed independently of other clinical groups. As a control, we shuffled the effect sizes of the CDR3 phenotypes obtained from cohort 1 to highlight the importance of effect sizes and position-specific nature of the CDR3 phenotypes.

Analysis of heterozygous HLA allele based restriction of TCR diversity

We employed IGoR, a probabilistic model of V(D)J recombination, to model the recombination processes of unproductive TCR sequences. The sequences were aligned, and the recombination parameters were inferred through iterative learning. Multiple rounds of inference were performed to ensure robust parameter estimation. The inferred models were saved, and gene usage patterns were visualized for each sample. Note that raw sequences without any filtering criteria were used in the IGoR model [83]. Further, the productive TCR sequences were analyzed using SoNNia [84], a selection model that builds on the recombination probabilities learned by IGoR. Sonia inferred selection pressures by modeling the observed TCR sequences and generating a large number of synthetic sequences for comparison. The model was trained over 80 epochs to optimize selection parameters, and its performance was evaluated by comparing the generative (Pgen) and posterior (Ppost) sequence probabilities. Model performance was assessed by calculating the entropy of the sequence distributions both pre- and post-selection, which provided insights into the diversity and selection pressures within the TCR repertoire.

The HLA genes A, B, C, DPA1, DPB1, DQA1, DQB1, DRB1 were considered to assess the heterozygosity of HLA alleles. We calculated the total count of homozygous alleles in HLA genes (ranging from 0 being heterozygous and 8 being homozygous), and assuming the equal contribution of the HLA genes in the TCR diversity.

To assess the clustering of the v-gene frequencies, TRBV28 was used as a case study, where it was segregated as low-frequency ($f < 0.015$), mid-range frequency ($0.015 \leq f \leq 0.04$), and high-frequency ($f > 0.04$). For each subset, bootstrapped samples were generated by resampling 30% of the data multiple times, allowing for robust estimation of frequency distributions with respect to HLA alleles. The mean and standard deviation of allele frequencies were computed across the resampled datasets to capture the variability in the data.

Machine learning approaches for repertoire classification and motif identification

ML analysis data preprocessing and general ML workflow

We employed a combination of machine learning and deep learning approaches to classify T1D status and identify T1D-associated motifs within immune repertoires. Cohort 1 served as the training set for all methods, while repertoires from Cohort 2 and 3 were used as a held-out test set. The T1D class was considered a positive class and remaining clinical statuses, FDR, SDR and CTRL, were considered non-T1D or negative class.

In the k-mer based logistic regression (LogReg) and DeepRC model, the same 5-fold cross-validation (CV) dataset was used to train both model parameters and hyper-parameters in cohort 1. The data was split into 5 random subsets, with T1D and control samples drawn separately without replacement for each fold. Each fold contained an approximately equal number of samples and maintained a consistent ratio of control to T1D samples. Moreover, each repertoire was assigned a weight based on the prevalence of their age group and the distribution of disease states across these groups, thereby normalizing the influence of age on the classification model (see **Supplementary Notes** for more detail). The first and last four AAs in each CDR3β sequence were also excluded to avoid the effect of V- and J-genes in ML/DL methods.

Statistical classification framework for T1D repertoires using public clones

We utilized the ImmuneML platform [88] to implement the statistical classification framework [17] on the T1D dataset, which utilizes unique V-gene, CDR3 and J-gene information. The YAML file was prepared following the immuneML documentation (https://docs.immuneml.uio.no/latest/usecases/emerson_reproduction.html). In summary, we employed the 'CompAIRRSequenceAbundance' encoding based on the CompAIRR tool [147] to enable faster repertoire level comparisons. Model performance was assessed across varying significance thresholds (p-values = [0.001, 0.0001, 0.00001, 0.000001]). The optimal model was selected based on optimization of log-loss values in 5-fold cross-validation. Log-loss measures the divergence between predicted probabilities and the actual class value, with lower log-loss indicating better performance. We further calculated the balanced accuracy and area under the ROC curve (AUROC) to analyze the performance of the model. The optimal model was further tested on cohort 2 and 3 to assess its predictive performance in unseen datasets.

K-mer based classification model using Logistic regression

The linear subsequence information at the 4-mer level was employed as a feature set to classify T1D status. A 4-mer frequency matrix of 160,000 (20^4; number of possible 4-mers) by 1298 (number of repertoires in cohort 1) was generated. To populate this matrix, we used a sliding window approach, where a window of 4 consecutive AA moved across each CDR3β sequence. The frequency of occurrence of each unique 4-mer was recorded in the matrix for its respective repertoire. This 4-mer frequency matrix was then used as a feature set in a LogReg to classify the T1D statuses of immune repertoires. For model training, we applied a 5-fold CV on Cohort 1, where 3 folds were used for training, 1 fold for tuning, and 1 fold for testing, with the process repeated recursively. The outputs of the 5 models, one from each fold, were then assembled

into a single model by averaging their predictions after applying the sigmoidal activation function. This ensembled model was subsequently tested on a held-out dataset consisting of Cohorts 2 and 3. The following hyperparameter ranges were used for the training of LogReg model:

Penalty	:	L1, L2, Elastic-Net penalty or no penalty
C	:	random value in [0.1, 10)
Solver algorithms	:	Newton-CG, LBFGS, Liblinear, SAG, SAGA
Number of iterations	:	{100, 500, 1000, 5000}

Deep learning-based multiple instance learning model for classification of T1D status

Development of DeepRC model

We utilized DeepRC [31], a DL method based on continuous modern Hopfield networks (MHN) [156], within a multiple instance learning framework to classify immune repertoires based on T1D status. In this method, each TCR repertoire was treated as a sample, classified as either T1D or CTRL, while individual CDR3β sequences within the repertoire were viewed as instances. Thus, each sample is represented as a "bag" of CDR3β sequences with a single class label.

The DeepRC architecture used a convolutional neural network (CNN) as an encoder for processing individual sequences. Input vectors were structured with dimensions corresponding to sequence_length × n_amino_acids, where the sequence length may vary across sequences. A learned kernel of shape (kernel_size x n_amino_acids) was convolved along the AA positions of each CDR3β sequence, and feature-wise max-pooling was applied to obtain a fixed-sized feature vector of shape (n_sequences x kernel_size), where n_sequences may differ between repertoires. A learned MHN attention pooling mechanism, as described in [31], was used to pool the instances in a repertoire of shape (n_sequences x kernel_size) into a single feature vector of shape (kernel_size). This results in a single feature vector of shape (kernel_size) as representation for a single repertoire. Finally, the feature vector was passed to a fully connected output layer, consisting of one output unit, to predict the class label of the repertoire.

All trainable parameters in the DeepRC architecture were optimized end-to-end using the Adam optimizer in PyTorch, with a binary cross-entropy loss function. Similar to LogReg, we assigned 3 folds for training, 1 fold as tuning, and the remaining 1 fold as test set, recursively, in a 5 fold CV setup. We train the DeepRC model on the training set while using the tuning set loss as early stopping criteria. The best-performing model for each CV fold was selected based on the area under the receiver operating characteristic curve (AUROC) of the tuning set, resulting in 5 distinct "best" models from the CV process. Subsequently, a LogReg model was trained to aggregate the predictions from these 5 “best” models, forming the final DeepRC model, which was then evaluated on the held-out test set. Following parameters were used for the training of DeepRC model:

Learning rate	:	5e-5
Number of weight updates	:	3e5
Number of CNN kernels	:	{8, 16, 32, 64}
Number of CNN layers	:	1
CNN kernel sizes	:	{5, 7, 9}

Number of attention layers	:	2
Number of neurons per attention layer	:	{32}
Number of output network layers	:	1
Number of units per output network layer	:	64
Weight decay penalties ({l1 term, l2 term})	:	{{0, 0},{1e-5, 1e-3},{1e-4, 1e-2}}

T1D-associated motif identification from DeepRC

DeepRC supports different ways of interpretability, via the attention values and via the contribution analysis method Integrated Gradients (IG) [93]. We applied the IG method to the trained DeepRC model from each cross-validation fold as described in the original paper [31]. It allows visualization of the contribution of inputs and weights to the prediction of the DeepRC models, which could then be used to manually extract motifs. For each DeepRC model in the cross-validation, we compute the Integrated Gradients (IG) such that we obtain the contribution of the input sequences to the DeepRC model output. In this step, we compute IG on the T1D positive samples of the tuning set samples that were also used for early stopping. The training split samples were not used here as they have already been overfitted by the DeepRC model. Subsequently, for each DeepRC model, we identified the input sequence with the highest contribution to the T1D prediction per repertoire. The DeepRC model with null coefficient in the ensemble LogReg model was excluded from the motif calculation. The identified sequences from each DeepRC model, corresponding to T1D-positive repertoires in the tuning set, were collected into a FASTA file and subsequently analyzed with GLAM2 [157] to compute motifs through sequence alignment.

$$glam2\ p\ -p\ -b\ 5\ -D\ 0.01\ -I\ 0.0001\ -s\ 1$$

We chose a motif width of -b 5, as it corresponds to the smallest motif width matching the CNN kernel size. The first and last 4 AA of every input sequence were cropped before alignment. The final selected motif [FWY],([AFHILMPQRSTVY]){0,1},[EFGHSWY],([AFQT]){0,1},[ACDEGHIKNPQRSTVY] contains two deletions at position 2 and 4 in the motif (Extended Data Fig. 9).

Consensus sequences containing both the pHLA-motif and DeepRC-motif were obtained by sequentially filtering the T1D dataset: first by the presence of DeepRC-motif, followed by further filtering based on presence of pHLA-motif. The resulting intersection of the pHLA-motif and DeepRC-motif is termed the "consensus-motif". Similar to other motifs, the consensus-motif score was calculated by normalizing the number of motif-containing sequences with the total CDR3β sequence count in the respective repertoire.

Quantitative Trait Locus Analysis

QC measures were performed to remove data from individuals that were sex discordant, related, and/or showed unusual levels of heterozygosity, as previously described [111]. Genetic ancestry was inferred using Admixture software [158] for projection analysis on the 1000 Genomes cohort[159]. Quantitative trait loci (QTL) analysis was performed to detect associations between TCR motif frequency versus 240 T1D risk variants [111] that were directly genotyped or imputed, as previously described. For PBMC data reported in this publication, linear regression was performed assuming additive genotypic effect with age, sex, T1D status, predicted probability of CMV infection [111], and 10 multidimensional scaling components as covariates. For nPOD tissue data, age, sex, diabetes status, and 10 multidimensional scaling components were included and for sorted peripheral blood data, age, sex, and T1D status were included as covariates.

Data availability

Sequence data are publicly available through Adaptive Biotechnologies ImmuneACCESS site. The data will be published upon completion of the review process. The code and data used in this study are available at the following GitHub repository: https://github.com/csi-greifflab/T1D-TCR.

Acknowledgement

We thank the study participants for the gift of blood sample donation, without which this work would not have been possible.

Disclosure statement

V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Absci, Omniscope, and Diagonal Therapeutics. V.G. is a consultant for Adaptyv Biosystems, Specifica Inc, Roche/Genentech, immunai, Proteinea, LabGenius, and FairJourney Biologics.

Funding

This work was supported by grants from The Leona M. and Harry B. Helmsley Charitable Trust (#2019PG-T1D011, to TMB and VG), the National Institutes of Health (P01 AI042288, to TMB), the American Diabetes Association (11-23-PDF-78, to LDP), Breakthrough T1D (formerly JDRF, 3-PDF-2022-1137-A-N, to MRS), UiO World-Leading Research Community (to VG), UiO: LifeScience Convergence Environment Immunolingo (to VG), EU Horizon 2020 iReceptorplus (#825821) (to VG), a Norwegian Cancer Society Grant (#215817, to VG), Research Council of Norway projects (#300740, #311341, #331890 to VG), a Research Council of Norway IKTPLUSS project (#311341, to VG). This project has received funding (to VG) from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 101007799 (Inno4Vac). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This communication reflects the author's view and neither IMI nor the European Union, EFPIA, or any Associated Partners are responsible for any use that may be made of the information contained therein. Funded by the European Union (ERC, AB-AG-INTERACT, 101125630, to VG). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801133 (to PR). This research was performed with the support of the Network for Pancreatic Organ donors with Diabetes (nPOD; RRID:SCR_014641), a collaborative type 1 diabetes research project supported by Breakthrough T1D and The Leona M. & Harry B. Helmsley Charitable Trust (Grant#3-SRA-2023-1417-S-B). 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://npod.org/for-partners/npod-partners/.

Author contributions

PR: Software, Formal Analysis, Investigation, Writing-Original Draft, Writing-Review & Editing, Visualization.
MRS: Software, Formal Analysis, Writing-Original Draft, Writing-Review & Editing, Visualization.
LDP: Software, Formal Analysis, Writing-Original Draft, Writing-Review & Editing, Visualization.
MW: Methodology, Software, Formal Analysis, Writing-Review & Editing.
KMB: Software, Formal Analysis, Writing-Original Draft, Writing-Review & Editing
KeM: Investigation, Data Curation, Writing-Review & Editing.
MP: Software, Writing-Review & Editing
GAH: Software
ALP: Writing-Review & Editing, Project Administration.
CK: Software, Writing-Review & Editing
GI: Software, Writing-Review & Editing
MC: Methodology, Writing-Review & Editing
LS: Software
KaM: Methodology, Writing-Review & Editing
LOB-N: Software, Writing-Review & Editing
CPW: Software, Writing-Review & Editing
SV: Software, Writing-Review & Editing
AD: Data Curation
EY: Data Curation
MAA: Writing-Review & Editing, Funding Acquisition.
MJH: Resources, Writing-Review & Editing.
DAS: Resources, Writing-Review & Editing.
CHW: Writing-Review & Editing.
SH: Writing-Review & Editing, Supervision
ROE: Data Curation, Supervision
AJFG: Writing-Review & Editing
PB: Conceptualization, Writing-Review & Editing, Supervision
GK: Writing-Review & Editing, Supervision
GKS: Conceptualization, Writing-Review & Editing, Supervision
VG: Conceptualization, Writing-Review & Editing, Supervision, Funding Acquisition.
TMB: Conceptualization, Resources, Writing-Review & Editing, Supervision, Funding Acquisition.

Extended data Figures

Extended Data Fig. 1 | The publicly available inflammatory disease and T1D-associated CDR3β sequences demonstrate association with clinical groups in cohort 1. The unique clone count of CDR3β sequences in cohort 1, which were also present in the McPAS/VDJdb database for given pathology (as detailed in Table S1), were averaged for each repertoire and normalized by the total clone count. These normalized values were further clustered based on the clinical groups and visualized in log scale. (**A**) Sequence matches were calculated for different pathologies in McPAS database. (**B**) The distribution of T1D pathology-associated CDR3β sequences in each repertoire. individual T1D-associated CDR3β sequences revealed a small subset of potential false positives in the McPAS database, which were observed in almost all repertoires. Manually curated databases are subject to bias based on a priori knowledge and experimental design. Thus, our results provide further support for utilization of machine learning and other unbiased methodologies to classify T1D repertoires. (**C**) Sequence matches were calculated for different antigen species in the VDJdb database. **(D)** Association of 25,508 unique CDR3β sequences associated with cytomegalovirus (CMV) exposure with clinical groups. In the figure, multiple testing was performed using Kruskal–Wallis test (denoted with red stars) and p-values were adjusted between pathology/antigen species. p-values for pairwise testing were calculated using two tailed Mann-Whitney U test and p-values were adjusted between different clinical groups. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. Cluster plot was generated using the UPGMA clustering method and Euclidean distance matrix. A universal color scheme was used to show type 1 diabetes (T1D), first degree relatives (FDR), second degree relatives (SDR) and control (CTRL) in red, yellow, light green and dark green, respectively.

Extended Data Fig. 2 | T1D-association of TCR features observed based on assessment of HLA-associated, TRBV-gene family constrained and single amino acid mutation distance neighbor CDR3β sequences. (**A**) TRBV-gene family and HLA-associated CDR3β sequences depleted (blue) or overrepresented (red) in the T1D repertoires (false discovery rate < 0.2), when considering detection of a CDR3 or a nearly-identical CDR3 with only one indel or substitution. (**B**) Gene usage and CDR3β(AA) motifs formed from top sequence clusters overrepresented in T1D repertoires. **(C**) Gene usage and CDR3β(AA) motifs formed from top sequence clusters underrepresented in T1D repertoires. HLA association determined by Fisher Exact Test p-value < 1E-8, HLA-prevalence > 0.05, non-HLA-prevalence < 0.1 in Cohort2. Relates to Fig. 3.

Extended Data Fig. 3 | The methodology used to assess the association between HLA and CDR3β. (A) The information related to amino acid (AA) polymorphism present at the HLA site was obtained from the HLA genotyping experiments. Here, HLA-DRB1 is given as an example, which has 4 AA variants (K, A, E, R) at site 71. All possible AA variant counts were calculated for each repertoire. **(B)** The histogram (left) shows the distribution of the CDR3β length in cohort 1 (for lengths containing more than 30,000 CDR3β sequences). The position-specific AA frequency was calculated for each CDR3β sequence with length ranging from L12 to L18 after removing the V- and J-germline-encoded flanking regions (right). In the example, we have removed the N-terminal sequences based on specific V-genes ("CASSE" for TRBV02-01 and "CASSQ" for TRBV04-01). The V and J germline-encoded sequences were removed using IMGT/HighV-QUEST tool. Further, the frequency of all AAs was calculated for each position and for each length (L12-L18) of CDR3β sequences. **(C)** The multivariate multiple linear regression model (MMLR) was calculated for each HLA site polymorphism, where count of AA variants at the site was explanatory variable and AA frequency of specific CDR3β position was response variable. This was done for each IMGT position of CDR3β sequence for all lengths (L12-L18), recursively. **(D)** The HLA risk score was calculated based on the 3 AA positions: HLA-DQβ1 (site 57) and HLA-DRβ1 (site 13 and 71). The odds ratio (OR) for possible combinations of three AAs was provided in a previous study [74], which was multiplied by the individual's genotype. In case multiple combinations were possible for the three AAs, then we consider the combination with the highest OR score. **(E)** In the first step to calculate CDR3 risk score, we considered HLA risk score and position-specific frequency of one AA to develop several linear regression models (LR) for each CDR3β position and length, recursively. The p-values were further adjusted for multiple testing using the Benjamini–Hochberg method. The AAs which showed significant correlations with HLA risk score (p-value≤0.05) were considered "T1D-associated position-specific CDR3β AA" or "CDR3 phenotypes" and respective correlation coefficients were considered the "effect sizes" associated with T1D. Within a repertoire, the effect sizes of each CDR3 phenotype were summed up for the whole CDR3β sequence, and averaged for the whole repertoire to calculate CDR3 risk score. **(F)** In our study, the CDR3β positions for CDR3β length L12 to L18 were defined using the IMGT numbering scheme, which are highlighted in blue color in the figure. The examples in the figures are taken directly or inspired from Iskigaki et al [5]. Relates to Fig. 3.

Extended Data Fig. 4 | The assessment of the pHLA-motif, derived from the HLA risk score, shows positive association with the clinical groups, high-risk HLA alleles (DR3/DR4) and autoantibody presence. (A) A sequence logo representation of the pHLA-motif**.** The consensus motif was derived from the CDR3β lengths L13-16 and IMGT positions 107-111. **(B)** The motif score was higher for T1D repertoires in cohort 1, following the expected trend of T1D > FDR > SDR > CTRL, with SDR being an exception. A similar pattern was also observed in cohorts 2 and 3. The pHLA-motif score represents the number of CDR3β sequences containing the pHLA-motifs normalized by total number of CDR3β sequences per repertoire, where higher pHLA-motif score signifies higher risk of T1D. **(C)** The correlation between T1D duration and motif score shows a pearson correlation of -0.18 and spearman correlation of -0.16. **(D)** There was no trend observed in V-gene usage across TCRβ repertoires of different clinical groups, when considering the CDR3β sequences containing pHLA-motifs. **(E)** The pHLA-motif score was significantly higher in

the high-risk HLA type (DR3 or DR4) individuals than other HLA types (DRX/X). The p-values in the figure were calculated with respect to the DRX/X HLA type. **(F)** The pHLA-motif score was plotted for each high-risk HLA type and grouped by the clinical groups.The pHLA-motif score observed the expected trend T1D>FDR>CTRL in all cases irrespective of high-risk HLA allele type. Here, pairwise testing was performed with respect to the T1D clinical group. **(G)** The pHLA-motif score also showed positive association with the number of autoantibodies present. The p-values were calculated with respect to no presence of autoantibodies. **(H)** The pHLA-motif score was high for autoantibody count >0. However, in case of no presence of autoantibody, motif score followed the expected trend where T1D>FDR>CTRL. **(I)** The pHLA-motif score was plotted against age of the patient for different clinical groups (for cohort 1). Each T1D status showed negative correlation with the age of the patient. In the figure, the p-values for the multiple testing were performed using Kruskal–Wallis test (denoted with red stars). Similarly, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. Relates to Fig. 3.

Extended Data Fig. 5 | The assessment of the nHLA-motif (protective motif), derived from the HLA risk score, shows negative association with the clinical groups, high-risk HLA alleles (DR3/DR4) and autoantibody presence. (A) A sequence logo representation of the nHLA-motif**.** The consensus motif was derived from the CDR3β lengths L13-16 and IMGT positions 107-111. **(B)** The motif score was higher for non-T1D repertoires in cohort 1, following the expected trend of T1D < FDR < SDR < CTRL, with SDR being an exception. A similar pattern was also observed in cohorts 2 and 3. The nHLA-motif score represents the number of CDR3β sequences containing the nHLA-motifs normalized by total number of CDR3β sequences per repertoire, where higher nHLA-motif score signifies lower risk of T1D. **(C)** The nHLA-motif score from the protective motif showed no correlation with the T1D duration (pearson correlation of 0.05 and spearman correlation of 0.07). **(D)** The V-genes with higher frequency had relatively lower usage in the T1D repertoires and had an observable trend of T1D<FDR<CTRL, when considering the CDR3β sequences containing the nHLA-motif. **(E)** The repertoires containing at least one DR3 allele exhibited nHLA-motif scores comparable to non-risk repertoires, indicating that DR3 alleles may possess more protective motifs. **(F)** The nHLA-motif score was plotted for each high-risk HLA type and grouped by the clinical groups. The nHLA-motif score was significantly higher for CTRL repertoires in cases both HLA alleles were not associated with high-risk of T1D (DRX/X) and observed the expected trend T1D<FDR<CTRL. Here, pairwise testing was performed with respect to the T1D clinical group. **(G)** The nHLA-motif score was higher for the repertoires with no or just one autoantibody present. The p-values were calculated with respect to no autoantibody presence. **(H)** The nHLA-motif score observed association with clinical groups in repertoires with no autoantibody presence with an expected trend where T1D<FDR<CTRL. **(I)** The nHLA-motif score was plotted against age of the patient for different clinical groups (for cohort 1). No correlation was observed for the nHLA-motif score with age of the patient. In the figure, the p-values for the multiple testing were performed using Kruskal–Wallis test (denoted with red stars). Similarly, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. Relates to Fig. 3.

Extended Data Fig. 6 | Heterozygous HLA alleles showed restriction of the TCR repertoire diversity in cohort 1. (A) The pre-selection TCR repertoire diversity (Pgen, IGoR) and **(B)** post-selection repertoire diversity (Ppost, SoNNia) calculated with respect to count of homozygous HLA alleles for 8 HLA genes (A, B, C, DPA1, DPB1, DQA1, DQB1, DRB1) showed a trend where repertoires with more heterozygous HLA alleles had lower diversity in both pre and post selection. **(C)** V-gene usage frequency calculated from the IGoR model. Some of the V-genes including TRBV28, TRBV4-3, TRBV3-2, TRBV3-1 showed clustering of v-gene frequency. **(D)** as a case study, the three v-gene frequency clusters of TRBV28 were assessed for their potential link with the HLA genes. The HLA alleles with frequency >0.01 were considered in the plot. The error bars were plotted by bootstrapping samples, where 30% of the data resampled multiple times. In figure **A** and **B**, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests with respect to most heterozygous HLA individuals (number of homozygous alleles = 0). p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method.

Extended Data Fig. 7 | Assessment of HLA risk score demonstrates that HLA alleles are robust predictors of T1D status (A) Confusion matrices were generated using high-risk HLA allele types (DR3 and DR4) as predictor of T1D status under two scenarios: (i) considering the presence of at least one high-risk HLA allele as indication of T1D status, and (ii) considering the presence of both high-risk HLA allele as indication of T1D status. The confusion matrices were generated separately for cohort 1 and cohort 2+3 (see **Supplementary Note** for more detail). **(B)** The HLA risk score was calculated based on the 3 AA positions: HLA-DQβ1 (site 57) and HLA-DRβ1 (site 13 and 71). Further, HLA risk score was used for binary classification of T1D positive and negative repertoires. **(C)** Expectedly, T1D duration does not show correlation with HLA risk score (pearson correlation: 0.11 and spearman correlation: 0.14). **(D)** HLA risk score in the repertoires with high-risk HLA alleles type (DR3 or DR4) was significantly higher compared to other HLA types (DRX/X). The p-values in the figure were calculated with respect to the DRX/X HLA type. **(E)** The HLA risk score was plotted for each high-risk HLA type and grouped by the clinical groups. The HLA risk score was significantly high in all clinical groups containing both alleles associated with high-risk (DR3 and DR4). However, it observed the expected trend T1D>FDR>SDR>CTRL in cases where at least one HLA allele was not associated with high-risk of T1D. Here, pairwise testing was performed with respect to the T1D clinical group. **(F)** The HLA risk score also observed positive association with the number of autoantibodies present. The p-values were calculated with respect to no presence of autoantibodies.

(G) HLA risk score was high for all individuals when autoantibody count was >0. However, in case of no presence of autoantibody, HLA risk score followed the expected trend where T1D>FDR>SDR>CTRL. **(H)** The HLA risk score was plotted against age of the patient for different clinical groups (for cohort 1) and as expected, it did not show any correlation with the age of the patient. In the figure, the p-values for the multiple testing were performed using Kruskal–Wallis test (denoted with red stars). Similarly, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. Relates to Fig. 4.

Extended Data Fig. 8 | Assessment of the DeepRC model performance demonstrates that TCR repertoires can be robust predictors of T1D risk. (A) DeepRC model score in the repertoires with high-risk HLA alleles type (DR3 or DR4) was significantly higher compared to other HLA types (DRX/X). The p-values in the figure were calculated with respect to the DRX/X HLA type. **(B)** The DeepRC model score was plotted for each high-risk HLA type and grouped by the clinical groups. The DeepRC model score was significantly high in all clinical groups containing both alleles associated with high-risk (DR3 and DR4). However, it observed the expected trend T1D>FDR>CTRL in cases where at least one HLA allele was not associated with high-risk of T1D. Here, pairwise testing was performed with respect to the T1D clinical group. **(C)** The DeepRC model score also observed positive association with the number of autoantibodies present. The p-values were calculated with respect to no presence of autoantibodies. **(D)** DeepRC model score was high for all individuals when autoantibody count was >0. However, in case of no presence of autoantibody, DeepRC model score followed the expected trend where T1D>FDR>CTRL. **(E)** The DeepRC model score was plotted against age of the patient for different clinical groups (for cohort 1) and a negative correlation was observed for all clinical groups. **(F)** Interestingly, T1D duration observed a positive correlation with DeepRC model score (pearson correlation: 0.19 and spearman correlation: 0.09). **(G)** The final ensemble logistic regression (LR) model, built using the outputs of each DeepRC model, assigned weights to each fold. These ensemble weights, along with the performance (AUROC curve) of the DeepRC model for each fold, were also presented. In the figure, the p-values for the multiple testing were performed using Kruskal–Wallis test (denoted with red stars). Similarly, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. Relates to Fig. 4.

Extended Data Fig. 9 | Assessment of the TCR repertoires-derived, interpretable version of the DeepRC model, referred to as DeepRC-motif, demonstrates a robust association with T1D status. The DeepRC-motif score represents the number of CDR3β sequences containing the DeepRC-motif normalized by total number of CDR3β sequences per repertoire, where higher DeepRC-motif score signifies higher risk of T1D. **(A)** The sequence logo representation of the DeepRC-motif. **(B)** DeepRC-motif score was significantly higher for T1D repertoires and observed the expected trend T1D>FDR>SDR>CTRL. **(C)** T1D duration observed a negative correlation with DeepRC-motif score (pearson correlation: -0.13 and spearman correlation: -0.12). **(D)** DeepRC-motif score in the repertoires with high-risk HLA alleles type (DR3 or DR4) was significantly higher compared to other HLA types (DRX/X). The p-values in the figure were calculated with respect to the DRX/X HLA type. **(E)** The DeepRC-motif score was plotted for each high-risk HLA type and grouped by the clinical groups. The DeepRC-motif score was significantly high in all clinical groups containing both alleles associated with high-risk (DR3 and DR4). However, it observed the expected trend T1D>FDR>CTRL in cases where at least one HLA allele was not associated with high-risk of T1D. Here, pairwise testing was performed with respect to the T1D clinical group. **(F)** The DeepRC-motif score also observed positive association with the number of autoantibodies present. The p-values were calculated with respect to no presence of autoantibodies. **(G)** DeepRC-motif score was high for all individuals when autoantibody count was >0. However, in case of no presence of autoantibody, DeepRC-motif score followed the expected trend where T1D>FDR>CTRL. **(H)** The DeepRC-motif score was plotted against age of the patient for different clinical groups (for cohort 1) and a negative correlation was observed for all clinical groups. In the figure, the p-values for the multiple testing were performed using Kruskal–Wallis test (denoted with red stars). Similarly, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. All p-values were adjusted for multiple testing using the Benjamini–Hochberg method. Relates to Fig. 4.

Extended Data Fig. 10 | Positively-associated HLA-motif (pHLA-motif) enriched in carriers of T1D risk genetics in bulk and sorted $CD4^+$ Treg and TCM from peripheral blood as well as pancreatic lymph node $CD8^+$ T cells. (**A**) Manhattan plot of 240 T1D risk variants versus pHLA-motif score in bulk peripheral blood. Linear regression assuming additive genotypic effect with age, sex, T1D status, predicted probability of CMV infection, and 10 multidimensional scaling components as covariates. Benjamini-Hochberg false discovery rate threshold at p=2.58e-4 (red line) and a conventional threshold of p = 0.05 (blue line). **(B)** Violin plots of pHLA-motif score in European ancestry non-diabetic (ND) subjects, according to significantly associated variants in **(A)**. Tagged HLA genotype and risk allele annotated above each plot. (**C**) pHLA-motif score across sorted T cell subsets in spleen and pancreatic lymph node (pLN). **(i)** Mixed-effects analysis with Tukey's multiple comparisons test. (**ii**) pLN $CD8^+$ T cell pHLA-motif score according to T1D status. (**D**) pHLA-motif score in sorted peripheral blood $CD4^+$ T cell subsets. **(i)** Mixed-effects analysis with Tukey's multiple comparisons test. (**ii**) Treg and (**iii**) $CD4^+$ TCM pHLA-motif score according to DQA1*05:01-DQB1*02:01 tag SNP. Linear regression with age, sex, and T1D status as covariates. In all violin plots, p-values for pairwise testing were calculated using two tailed Mann-Whitney U tests and p-values were adjusted between different clinical

groups using the Benjamini–Hochberg method. p-values were described as * for [0.01,0.05], ** for [0.001,0.01] and *** for <0.001 and no stars plotted for non-significant values. Relates to Fig. 5.

References

1. DeWitt, W. S., 3rd *et al.* Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife* **7**, (2018).
2. Ortega, M. R. *et al.* Learning predictive signatures of HLA type from T-cell repertoires. *bioRxivorg* 2024.01.25.577228 (2024).
3. Zahid, H. J. *et al.* Large-scale statistical mapping of T-cell receptor*β*sequences to Human Leukocyte Antigens. *bioRxiv* 2024.04.01.587617 (2024) doi:10.1101/2024.04.01.587617.
4. Nagafuchi, Y. *et al.* Control of naive and effector CD4 T cell receptor repertoires by rheumatoid-arthritis-risk HLA alleles. *J. Autoimmun.* **133**, 102907 (2022).
5. Ishigaki, K. *et al.* HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors. *Nat. Genet.* **54**, 393–402 (2022).
6. Noble, J. A. *et al.* The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families. *Am. J. Hum. Genet.* **59**, 1134–1148 (1996).
7. Lambert, A. P. *et al.* Absolute risk of childhood-onset type 1 diabetes defined by human leukocyte antigen class II genotype: a population-based study in the United Kingdom. *J. Clin. Endocrinol. Metab.* **89**, 4037–4043 (2004).
8. Noble, J. A. & Valdes, A. M. Genetics of the HLA region in the prediction of type 1 diabetes. *Curr. Diab. Rep.* **11**, 533–542 (2011).
9. Krischer, J. P. *et al.* Predictors of the Initiation of Islet Autoimmunity and Progression to Multiple Autoantibodies and Clinical Diabetes: The TEDDY Study. *Diabetes Care* **45**, 2271–2281 (2022).
10. Greissl, J. *et al.* Immunosequencing of the T-cell receptor repertoire reveals signatures specific for diagnosis and characterization of early Lyme disease. *bioRxiv* (2021) doi:10.1101/2021.07.30.21261353.
11. Ross, J. J. *et al.* Exocrine pancreatic enzymes are a serological biomarker for type 1 diabetes staging and pancreas size. *Diabetes* **70**, 944–954 (2021).
12. Hanna, S. J. *et al.* The Type 1 Diabetes T Cell Receptor and B Cell Receptor Repository in the AIRR Data Commons: a practical guide for access, use and contributions through the Type 1 Diabetes AIRR Consortium. *Diabetologia* 1–17 (2024).
13. Greiff, V., Yaari, G. & Cowell, L. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology* (2020) doi:10.1016/j.coisb.2020.10.010.
14. Arnaout, R. A., Prak, E. T. L., Schwab, N., Rubelt, F. & Adaptive Immune Receptor Repertoire Community. The Future of Blood Testing Is the Immunome. *Front. Immunol.* **12**, 626793 (2021).
15. Mhanna, V. *et al.* Adaptive immune receptor repertoire analysis. *Nature Reviews Methods Primers* **4**, 1–25 (2024).
16. Liu, X. *et al.* T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Ann. Rheum. Dis.* **78**, 1070–1078 (2019).
17. Emerson, R. O. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
18. Snyder, T. M. *et al.* Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels. *medRxiv* (2020) doi:10.1101/2020.07.31.20165647.
19. Yu, X. *et al.* Quantifiable TCR repertoire changes in prediagnostic blood specimens among patients with high-grade ovarian cancer. *Cell Rep. Med.* **5**, 101612 (2024).
20. Chu, N. D. *et al.* Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. *BMC Immunol.* **20**, 19 (2019).
21. Weber, C. R. *et al.* Reference-based comparison of adaptive immune receptor repertoires. *Cell Rep*

Methods **2**, 100269 (2022).
22. Christophersen, A. *et al.* Tetramer-visualized gluten-specific CD4+ T cells in blood as a potential diagnostic marker for coeliac disease without oral gluten challenge. *United European Gastroenterol J* **2**, 268–278 (2014).
23. Christophersen, A. *et al.* Distinct phenotype of CD4+ T cells driving celiac disease identified in multiple autoimmune conditions. *Nature Medicine* **25**, 734–737 (2019).
24. Dahal-Koirala, S. *et al.* TCRpower: quantifying the detection power of T-cell receptor sequencing with a novel computational pipeline calibrated by spike-in sequences. *Brief. Bioinform.* **23**, bbab566 (2022).
25. De Neuter, N. *et al.* Memory CD4+ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus. *Genes Immun.* **20**, 255–260 (2019).
26. Ostmeyer, J., Christley, S., Toby, I. T. & Cowell, L. G. Biophysicochemical motifs in T cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocytes and adjacent healthy tissue. *Cancer Res.* (2019) doi:10.1158/0008-5472.CAN-18-2292.
27. Schmidt-Barbo, P. *et al.* Detection of disease-specific signatures in B cell repertoires of lymphomas using machine learning. *PLoS Comput. Biol.* **20**, e1011570 (2024).
28. Sidhom, J.-W. *et al.* Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. *Sci Adv* **8**, eabq5089 (2022).
29. Schneider-Hohendorf, T. *et al.* Broader anti-EBV TCR repertoire in multiple sclerosis: disease specificity and treatment modulation. *Brain* awae244 (2024).
30. Komech, E. A. *et al.* CD8+ T cells with characteristic T cell receptor beta motif are detected in blood and expanded in synovial fluid of ankylosing spondylitis patients. *Rheumatology* **57**, 1097–1104 (2018).
31. Widrich, M. *et al.* Modern Hopfield networks and attention for immune repertoire classification. *bioRxiv* 18832–18845 (2020) doi:10.1101/2020.04.12.038158.
32. Sidhom, J.-W., Benjamin Larman, H., Pardoll, D. M. & Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications* vol. 12 Preprint at https://doi.org/10.1038/s41467-021-21879-w (2021).
33. Slabodkin, A., Sollid, L. M., Sandve, G. K., Robert, P. A. & Greiff, V. Weakly supervised identification and generation of adaptive immune receptor sequences associated with immune disease status. *bioRxiv* 2023.09.24.558823 (2023) doi:10.1101/2023.09.24.558823.
34. Culina, S. *et al.* Islet-reactive $CD8^+$ T cell frequencies in the pancreas, but not in blood, distinguish type 1 diabetic patients from healthy donors. *Sci Immunol* **3**, (2018).
35. Mitchell, A. M. *et al.* Temporal development of T cell receptor repertoires during childhood in health and disease. *JCI Insight* **7**, (2022).
36. Gomez-Tourino, I., Kamra, Y., Baptista, R., Lorenc, A. & Peakman, M. T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nat. Commun.* **8**, 1792 (2017).
37. Seay, H. R. *et al.* Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight* **1**, e88242 (2016).
38. Eugster, A. *et al.* High diversity in the TCR repertoire of GAD65 autoantigen-specific human CD4+ T cells. *J. Immunol.* **194**, 2531–2538 (2015).
39. Cerosaletti, K. *et al.* Single-Cell RNA Sequencing Reveals Expanded Clones of Islet Antigen-Reactive CD4+ T Cells in Peripheral Blood of Subjects with Type 1 Diabetes. *J. Immunol.* **199**, 323–335 (2017).
40. Nakayama, M. & Michels, A. W. Using the T cell receptor as a biomarker in type 1 diabetes. *Front. Immunol.* **12**, 777788 (2021).
41. Perry, D. J. *et al.* A genomic data archive from the Network for Pancreatic Organ donors with Diabetes. *Sci Data* **10**, 323 (2023).
42. Williams, M. D. *et al.* Genetic Composition and Autoantibody Titers Model the Probability of Detecting C-Peptide Following Type 1 Diabetes Diagnosis. *Diabetes* **70**, 932–943 (2021).
43. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683

(2013).
44. Greiff, V. *et al.* A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* **7**, 49 (2015).
45. Chiffelle, J. *et al.* T-cell repertoire analysis and metrics of diversity and clonality. *Curr. Opin. Biotechnol.* **65**, 284–295 (2020).
46. Vujović, M., Marcatili, P., Chain, B., Kaplinsky, J. & Andresen, T. L. Signatures of T cell immunity revealed using sequence similarity with TCRDivER algorithm. *Commun Biol* **6**, 357 (2023).
47. Miho, E. *et al.* Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Front. Immunol.* **9**, 224 (2018).
48. Britanova, O. V. *et al.* Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192**, 2689–2698 (2014).
49. Britanova, O. V. *et al.* Targeted depletion of TRBV9+ T cells as immunotherapy in a patient with ankylosing spondylitis. *Nat. Med.* **29**, 2731–2736 (2023).
50. Greiff, V. *et al.* Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* **15**, 40 (2014).
51. Gotelli, N. J. & Colwell, R. K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379–391 (2001).
52. Estorninho, M. *et al.* A novel approach to tracking antigen-experienced CD4 T cells into functional compartments via tandem deep and shallow TCR clonotyping. *J. Immunol.* **191**, 5430–5440 (2013).
53. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).
54. Bashford-Rogers, R. J. M. *et al.* Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* **23**, 1874–1884 (2013).
55. Simpson, E. H. Measurement of Diversity. *Nature* **163**, 688–688 (1949).
56. Berger, W. H. & Parker, F. L. Diversity of planktonic foraminifera in deep-sea sediments. *Science* **168**, 1345–1347 (1970).
57. Greiff, V. *et al.* Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immunol.* **199**, 2985–2997 (2017).
58. Amoriello, R. *et al.* TCR repertoire diversity in Multiple Sclerosis: High-dimensional bioinformatics analysis of sequences from brain, cerebrospinal fluid and peripheral blood. *EBioMedicine* **68**, 103429 (2021).
59. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
60. Bagaev, D. V. *et al.* VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* **48**, D1057–D1062 (2020).
61. Cohn, A., Sofia, A. M. & Kupfer, S. S. Type 1 diabetes and celiac disease: clinical overlap and new insights into disease pathogenesis. *Curr. Diab. Rep.* **14**, 517 (2014).
62. Volta, U., Tovoli, F. & Caio, G. Clinical and immunological features of celiac disease in patients with Type 1 diabetes mellitus. *Expert Rev. Gastroenterol. Hepatol.* **5**, 479–487 (2011).
63. Mitchell, A. M. *et al.* Tracking DNA-based antigen-specific T cell receptors during progression to type 1 diabetes. *Sci. Adv.* **9**, eadj6975 (2023).
64. Gutierrez-Achury, J. *et al.* Contrasting the Genetic Background of Type 1 Diabetes and Celiac Disease Autoimmunity. *Diabetes Care* **38 Suppl 2**, S37–44 (2015).
65. Chen, L.-K., Chou, Y.-C., Tsai, S.-T., Hwang, S.-J. & Lee, S.-D. Hepatitis C virus infection-related Type 1 diabetes mellitus. *Diabet. Med.* **22**, 340–343 (2005).
66. Masuda, H. *et al.* Acute onset of type 1 diabetes accompanied by acute hepatitis C: the potential role of proinflammatory cytokine in the pathogenesis of autoimmune diabetes. *Diabetes Res. Clin. Pract.* **75**, 357–361 (2007).
67. Taguchi, M. *et al.* New-onset type 1 diabetes and Graves' disease after antiretroviral therapy in a patient with human immunodeficiency virus infection. *J. Diabetes Investig.* **14**, 489–493 (2023).
68. Min-ChunYeh *et al.* Newly diagnosed type 1 diabetes mellitus in a human immunodeficiency virus-

infected patient with antiretroviral therapy-induced immune reconstitution inflammatory syndrome: a case report. *BMC Infect. Dis.* **23**, 619 (2023).
69. Kendall, E. K., Olaker, V. R., Kaelber, D. C., Xu, R. & Davis, P. B. Association of SARS-CoV-2 Infection With New-Onset Type 1 Diabetes Among Pediatric Patients From 2020 to 2021. *JAMA Netw Open* **5**, e2233014 (2022).
70. Ekman, I. *et al.* Early childhood CMV infection may decelerate the progression to clinical type 1 diabetes. *Pediatr. Diabetes* **20**, 73–77 (2019).
71. Pak, C. Y., Eun, H. M., McArthur, R. G. & Yoon, J. W. Association of cytomegalovirus infection with autoimmune type 1 diabetes. *Lancet* **2**, 1–4 (1988).
72. May, D. H., Woodhouse, S., Howie, B. & Robins, H. S. A Catalog of the Public T-cell Response to Cytomegalovirus. *bioRxiv* 2024.05.08.593237 (2024) doi:10.1101/2024.05.08.593237.
73. Ilonen, J. *et al.* Estimation of genetic risk for type 1 diabetes. *Am. J. Med. Genet.* **115**, 30–36 (2002).
74. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015).
75. Pociot, F. & Lernmark, Å. Genetic risk factors for type 1 diabetes. *Lancet* **387**, 2331–2339 (2016).
76. Smith, C. J. *et al.* Haplotype analysis reveals pleiotropic disease associations in the HLA region. *medRxiv* 2024.07.29.24311183 (2024).
77. Ashby, K. M. & Hogquist, K. A. A guide to thymic selection of T cells. *Nat. Rev. Immunol.* **24**, 103–117 (2024).
78. Pugliese, A. Autoreactive T cells in type 1 diabetes. *J. Clin. Invest.* **127**, 2881–2891 (2017).
79. Liu, S., Bradley, P. & Sun, W. Neural network models for sequence-based TCR and HLA association prediction. *PLoS Comput. Biol.* **19**, e1011664 (2023).
80. Nakayama, M. *et al.* Prime role for an insulin epitope in the development of type 1 diabetes in NOD mice. *Nature* **435**, 220–223 (2005).
81. Williams, C. L. *et al.* The longitudinal loss of islet autoantibody responses from diagnosis of type 1 diabetes occurs progressively over follow-up and is determined by low autoantibody titres, early-onset, and genetic variants. *Clin. Exp. Immunol.* **210**, 151–162 (2022).
82. Brown, A. J. *et al.* MHC heterozygosity limits T cell receptor variability in CD4 T cells. *Sci Immunol* **9**, eado5295 (2024).
83. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, 561 (2018).
84. Isacchini, G., Walczak, A. M., Mora, T. & Nourmohammad, A. Deep generative selection models of T and B cell receptor repertoires with soNNia. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2023141118 (2021).
85. Krishna, C., Chowell, D., Gönen, M., Elhanati, Y. & Chan, T. A. Genetic and environmental determinants of human TCR repertoire diversity. *Immun. Ageing* **17**, 26 (2020).
86. Liu, M. *et al.* TCR-L: an analysis tool for evaluating the association between the T-cell receptor repertoire and clinical phenotypes. *BMC Bioinformatics* **23**, 152 (2022).
87. Noble, J. A. & Erlich, H. A. Genetics of type 1 diabetes. *Cold Spring Harb. Perspect. Med.* **2**, a007732 (2012).
88. Pavlović, M. *et al.* The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nature Machine Intelligence* **3**, 936–944 (2021).
89. Conrad, N. *et al.* Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK. *Lancet* **401**, 1878–1890 (2023).
90. Zaslavsky, M. E. *et al.* Disease diagnostics using machine learning of immune receptors. *bioRxiv* 2022.04.26.489314 (2024) doi:10.1101/2022.04.26.489314.
91. Chen, M., Zhao, Y., Wang, Z., He, B. & Yao, J. A Noisy-Label-Learning Formulation for Immune Repertoire Classification and Disease-Associated Immune Receptor Sequence Identification. *arXiv [cs.LG]* (2023).
92. Katayama, Y. & Kobayashi, T. J. Comparative study of repertoire classification methods reveals data

efficiency of k -mer feature extraction. *Front. Immunol.* **13**, 797640 (2022).
93. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) vol. 70 3319–3328 (PMLR, 2017).
94. Majumder, P. *et al.* A super enhancer controls expression and chromatin architecture within the MHC class II locus. *J. Exp. Med.* **217**, jem.20190668 (2020).
95. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
96. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
97. Achenbach, P., Bonifacio, E., Koczwara, K. & Ziegler, A.-G. Natural history of type 1 diabetes. *Diabetes* **54 Suppl 2**, S25–31 (2005).
98. Atkinson, M. A. & Eisenbarth, G. S. Type 1 diabetes: new perspectives on disease pathogenesis and treatment. *Lancet* **358**, 221–229 (2001).
99. Skowera, A. *et al.* β-cell-specific CD8 T cell phenotype in type 1 diabetes reflects chronic autoantigen exposure. *Diabetes* **64**, 916–925 (2015).
100. Krogvold, L. *et al.* Detection of a low-grade enteroviral infection in the islets of langerhans of living patients newly diagnosed with type 1 diabetes. *Diabetes* **64**, 1682–1687 (2015).
101. Strom, T. B. Can childhood viral infection protect from type 1 diabetes? *J. Clin. Invest.* **119**, 1458–1461 (2009).
102. Lagattuta, K. A. *et al.* Repertoire analyses reveal T cell antigen receptor sequence features that influence T cell fate. *Nat. Immunol.* **23**, 446–457 (2022).
103. Textor, J. *et al.* Machine learning analysis of the T cell receptor repertoire identifies sequence features of self-reactivity. *Cell Syst.* **14**, 1059–1073.e5 (2023).
104. van Lummel, M. *et al.* Type 1 diabetes-associated HLA-DQ8 transdimer accommodates a unique peptide repertoire. *J. Biol. Chem.* **287**, 9514–9524 (2012).
105. Chow, I.-T. *et al.* Discriminative T cell recognition of cross-reactive islet-antigens is associated with HLA-DQ8 transdimer-mediated autoimmune diabetes. *Sci. Adv.* **5**, eaaw9336 (2019).
106. Zhou, Z. *et al.* Type 1 diabetes associated HLA-DQ2 and DQ8 molecules are relatively resistant to HLA-DM mediated release of invariant chain-derived CLIP peptides. *Eur. J. Immunol.* **46**, 834–845 (2016).
107. Weber, D. A., Evavold, B. D. & Jensen, P. E. Enhanced dissociation of HLA-DR-bound peptides in the presence of HLA-DM. *Science* **274**, 618–620 (1996).
108. Denzin, L. K. Inhibition of HLA-DM mediated MHC class II peptide loading by HLA-DO promotes self tolerance. *Front. Immunol.* **4**, 465 (2013).
109. Naik, R. G. *et al.* Precursor frequencies of T-cells reactive to insulin in recent onset type 1 diabetes mellitus. *J. Autoimmun.* **23**, 55–61 (2004).
110. Tran, M. T. *et al.* A structural basis of T cell cross-reactivity to native and spliced self-antigens presented by HLA-DQ8. *J. Biol. Chem.* **300**, 107612 (2024).
111. Shapiro, M. R. *et al.* Human immune phenotyping reveals accelerated aging in type 1 diabetes. *JCI Insight* **8**, (2023).
112. Robertson, C. C. *et al.* Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nat. Genet.* **53**, 962–971 (2021).
113. Pathiraja, V. *et al.* Proinsulin-specific, HLA-DQ8, and HLA-DQ8-transdimer-restricted CD4+ T cells infiltrate islets in type 1 diabetes. *Diabetes* **64**, 172–182 (2015).
114. Stadinski, B. D. *et al.* I-Ag7 β56/57 polymorphisms regulate non-cognate negative selection to CD4+ T cell orchestrators of type 1 diabetes. *Nat. Immunol.* **24**, 652–663 (2023).
115. Zander, R. *et al.* Tfh-cell-derived interleukin 21 sustains effector CD8+ T cell responses during chronic viral infection. *Immunity* **55**, 475–493.e5 (2022).
116. Bettelli, E. & Campbell, D. J. Circulating TFH cells as a marker for early therapeutic intervention in

T1D. *Nat. Immunol.* **21**, 1141–1142 (2020).
117. Ferreira, R. C. *et al.* IL-21 production by CD4+ effector T cells and frequency of circulating follicular helper T cells are increased in type 1 diabetes patients. *Diabetologia* **58**, 781–790 (2015).
118. Mitchell, A. M. & Michels, A. W. T cell receptor sequencing in autoimmunity. *J. Life Sci. (Westlake Village)* **2**, 38–58 (2020).
119. Ziegler, A.-G., Bonifacio, E. & BABYDIAB-BABYDIET Study Group. Age-related islet autoantibody incidence in offspring of patients with type 1 diabetes. *Diabetologia* **55**, 1937–1943 (2012).
120. Mhanna, V. *et al.* Impaired activated/memory regulatory T cell clonal expansion instigates diabetes in NOD mice. *Diabetes* **70**, 976–985 (2021).
121. Eugster, A. *et al.* Physiological and pathogenic T cell autoreactivity converge in type 1 diabetes. *Nat. Commun.* **15**, 9204 (2024).
122. Pogorelyy, M. V. *et al.* TIRTL-seq: Deep, quantitative, and affordable paired TCR repertoire sequencing. *Immunology* (2024).
123. T, R. R., Demerdash, O. N. A. & Smith, J. C. TCR-H: explainable machine learning prediction of T-cell receptor epitope binding on unseen datasets. *Front. Immunol.* **15**, 1426173 (2024).
124. Zdinak, P. M. *et al.* De novo identification of CD4+ T cell epitopes. *Nat. Methods* **21**, 846–856 (2024).
125. Joglekar, A. V. & Li, G. T cell antigen discovery. *Nat. Methods* **18**, 873–880 (2021).
126. Ma, K.-Y. *et al.* High-throughput and high-dimensional single-cell analysis of antigen-specific CD8+ T cells. *Nat. Immunol.* **22**, 1590–1598 (2021).
127. Weber, A., Pélissier, A. & Rodríguez Martínez, M. T-cell receptor binding prediction: A machine learning revolution. *Immunoinformatics (Amst.)* **15**, 100040 (2024).
128. Minervina, A., Pogorelyy, M. & Mamedov, I. T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. *Transpl. Int.* **32**, 1111–1123 (2019).
129. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
130. Shomuradova, A. S. *et al.* SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T cell receptors. *Immunity* **53**, 1245–1257.e5 (2020).
131. Jacobsen, L. M., Posgai, A., Seay, H. R., Haller, M. J. & Brusko, T. M. T cell receptor profiling in type 1 diabetes. *Curr. Diab. Rep.* **17**, 118 (2017).
132. Gitelman, S. E. & Bluestone, J. A. Regulatory T cell therapy for type 1 diabetes: May the force be with you. *J. Autoimmun.* **71**, 78–87 (2016).
133. Beck, R. W. *et al.* The T1D Exchange clinic registry. *J. Clin. Endocrinol. Metab.* **97**, 4383–4389 (2012).
134. Davis, A. K. *et al.* Prevalence of detectable C-Peptide according to age at diagnosis and duration of type 1 diabetes. *Diabetes Care* **38**, 476–481 (2015).
135. Carlson, C. S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
136. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**, 4099–4107 (2009).
137. Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comput. Biol.* **9**, e1002877 (2013).
138. Wasserfall, C. *et al.* Validation of a rapid type 1 diabetes autoantibody screening assay for community-based screening of organ donors to identify subjects at increased risk for the disease. *Clin. Exp. Immunol.* **185**, 33–41 (2016).
139. Marzinotto, I. *et al.* Islet Autoantibody Standardization Program: interlaboratory comparison of insulin autoantibody assay performance in 2018 and 2020 workshops. *Diabetologia* **66**, 897–912 (2023).
140. Chang, Y. M. *et al.* T Cell Receptor Diversity and Lineage Relationship between Virus-Specific CD8 T Cell Subsets during Chronic Lymphocytic Choriomeningitis Virus Infection. *J. Virol.* **94**,

(2020).
141. Niu, J. *et al.* Association of CD8(+) T lymphocyte repertoire spreading with the severity of DRESS syndrome. *Sci. Rep.* **5**, 9913 (2015).
142. Sokal, R. & Michener, C. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin* (1958).
143. Rényi, A. On measures of entropy and information. in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* vol. 4 547–562 (University of California Press, 1961).
144. Hill, M. O. Diversity and evenness: A unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
145. Jost, L. Entropy and diversity. *Oikos* **113**, 363–375 (2006).
146. Horn, H. S. Measurement of 'overlap' in comparative ecological studies. *Am. Nat.* **100**, 419–424 (1966).
147. Rognes, T., Scheffer, L., Greiff, V. & Sandve, G. K. CompAIRR: ultra-fast comparison of adaptive immune receptor repertoires by exact and approximate sequence matching. *Bioinformatics* **38**, 4230–4232 (2022).
148. Valkiers, S., Van Houcke, M., Laukens, K. & Meysman, P. ClusTCR: a python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity. *Bioinformatics* **37**, 4865–4867 (2021).
149. Chotisorayuth, T. & Tiffeau-Mayer, A. Lightning-fast adaptive immune receptor similarity search by symmetric deletion lookup. *arXiv [q-bio.QM]* (2024).
150. Liu, S., Bradley, P. & Sun, W. Neural Network Models for Sequence-Based TCR and HLA Association Prediction. *bioRxiv* 2023.05.25.542327 (2023) doi:10.1101/2023.05.25.542327.
151. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
152. Mayer-Blackwell, K. *et al.* TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *Elife* **10**, (2021).
153. Hagberg, A., Schult, D., Swart, P. & Hagberg, J. M. Exploring Network Structure, Dynamics, and Function using NetworkX. (2008).
154. Lefranc, M.-P. *et al.* IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37**, D1006–12 (2009).
155. Sakaue, S. *et al.* Tutorial: a statistical genetics guide to identifying HLA alleles driving complex disease. *Nat. Protoc.* (2023) doi:10.1038/s41596-023-00853-4.
156. Ramsauer, H. *et al.* Hopfield Networks is All You Need. *arXiv [cs.NE]* (2020).
157. Frith, M. C., Saunders, N. F. W., Kobe, B. & Bailey, T. L. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.* **4**, e1000071 (2008).
158. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
159. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
160. Boughter, C. T. & Meier-Schellersheim, M. Conserved biophysical compatibility among the highly variable germline-encoded regions shapes TCR-MHC interactions. *Elife* **12**, (2023).
161. Aly, T. A. *et al.* Extreme genetic risk for type 1A diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14074–14079 (2006).
162. Gomes, K. F. B. *et al.* The influence of population stratification on genetic markers associated with type 1 diabetes. *Sci. Rep.* **7**, 43513 (2017).