HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors

Kazuyoshi Ishigaki1,2,3,4, Kaitlyn A. Lagattuta1,2,3,5,6, Yang Luo1,2,3, Eddie A. James1,7, Jane H. Buckner7 and Soumya Raychaudhuri1,2,3,5,8

Autoimmune diseases include a broad set of disorders where the immune system targets host-derived proteins (autoantigens) for destruction. For all autoimmune diseases, the major histocompatibility complex (MHC) locus, which harbors the HLA genes, accounts for more risk than any other locus in the genome1–5. For example, 12.7% of the phenotypic variance of rheumatoid arthritis (RA), a prototypical autoimmune disease, can be explained by 5 amino acid polymorphisms within the HLA locus that encodes the antigen-binding pocket of HLA. In sharp contrast, all established non-MHC risk alleles in aggregate explain 4% of the phenotypic variance of RA. Antigenic peptides presented by HLA proteins are recognized by T cells, which initiate antigen-specific immune responses. Defining the mechanisms by which HLA risk alleles influence autoimmune risk is a critical ongoing challenge.

One potential explanation for autoimmune risk in the MHC locus is that HLA proteins encoded by risk alleles may increase the presentation of critical autoantigens to the immune system (‘peripheral hypothesis’, Fig. 1a)6–9. For example, citrullinated self-peptides have higher binding affinity for HLA proteins encoded by the HLA-DRB1 RA risk alleles than those encoded by protective alleles10,11. Similar findings have been reported for type 1 diabetes (T1D)12 and celiac disease (CD)13–15. However, there is an alternative, nonmutually exclusive hypothesis: HLA risk alleles may modulate the risk of autoimmunity by influencing thymic T cell selection, resulting in an increased frequency of autoreactive TCRs (‘central hypothesis’, Fig. 1a)16–18. T cell antigen specificity is defined by the hypervariable CDR3. During T cell development in the thymus, a highly diverse CDR3 repertoire is generated through random V(D)J recombination in immature T cells19. Thymic epithelial cells present self-peptides on HLA proteins and T cells that cannot generate substantial TCR signaling from any HLA–peptide complex die by neglect (positive selection). However, to protect against autoimmunity, T cells die by apoptosis if TCR signaling from any HLA–peptide complex is too strong (negative selection)20–22. Although multiple studies have demonstrated the importance of thymic T cell selection in autoimmunity16–18,23,24, the potential role of the HLA risk alleles in shaping the T cell repertoire during thymic selection has yet to be demonstrated in humans.

In this study, we sought to assess whether there is genetic evidence supporting the central hypothesis. We treated the amino acid composition of CDR3 as a quantitative trait and tested its association with HLA alleles; we call this CDR3 quantitative trait locus (QTL) analysis (cdr3-QTL). We then investigated how HLA autoimmune risk alleles modify the amino acid compositions of the CDR3 repertoire. Finally, we assessed whether the CDR3 features favored by HLA autoimmune risk alleles are enriched in candidate pathogenic TCRs than control TCRs (for example, citrullinated epitope-specific TCRs in patients with rheumatoid arthritis). Together, these results provide genetic evidence supporting the central hypothesis.
Articles NaTURe GENETICS

Post-disease onset

Inflammation sites

To focus on the antigen reactivity of the TCR, we restricted our cdr3-QTL analysis to the 10 positions (107–116) of CDR3 that directly contacted antigens. To filter out HLA-V gene associations, we excluded germ line-encoded sequences from CDR3 (Supplementary Figs. 1–4 and Supplementary Note). We refer to locations in HLA amino acid sequences as ‘sites’ and in CDR3 sequences as ‘positions’. For each CDR3 position, we created a multidimensional phenotype vector representing the frequencies of all amino acids. We used a multivariate multiple linear regression model (MMLM) to detect associations between this vector and all alleles at an HLA site; we then assessed their significance with the multivariate analysis of variance (MANOVA) test (question 1 in Fig. 1b and Extended Data Fig. 2). Intuitively, this model estimates how much interindividual variation in CDR3 amino acid frequencies at a CDR3 position can be predicted by all alleles at an HLA site. We conducted downstream investigations using the cdr3-QTL results, which include application of the CDR3 risk score (a score indicating the enrichment of CDR3 features favored by HLA risk alleles) to clinical TCR repertoire datasets (n = 22 patients in total).

To understand the sequence patterns in CDR3, we calculated diversity and higher pairwise mutual information (Extended Data Fig. 1d).

15 amino acids (L15-CDR3) was the most frequent (Extended Data Fig. 1a–c).

To understand the sequence patterns in CDR3, we calculated the diversity of amino acids at each CDR3 position and the mutual information between all pairs of positions. The middle positions (109–112) were generated by random recombination in the thymus; unsurprisingly, these positions had high diversity and little evidence of pairwise mutual information. In contrast, the flanking positions (104–108 and 113–118) were almost exclusively defined by germ line-encoded V or J genes; hence, these positions have small diversity and higher pairwise mutual information (Extended Data Fig. 1d).

Position-level cdr3-QTL signals. We first assessed whether HLA alleles explain amino acid frequencies at specific CDR3 positions. To focus on the antigen reactivity of the TCR, we restricted our cdr3-QTL analysis to the 10 positions (107–116) of CDR3 that directly contacted antigens. To filter out HLA-V gene associations,

Fig. 1 | Underlying hypotheses and study overview. a, Pathogenic roles of HLA in autoimmune disease can be explained by two nonexclusive hypotheses: the central and peripheral hypotheses. This figure illustrates five T cell maturation phases: (1) thymic T cells preselection (during T cell development in the thymus, a highly diverse TCR repertoire is generated); (2) thymic T cells postselection; (3) naïve T cells in the peripheral blood; (4) memory T cells in the peripheral blood; and (5) pathogenic T cells at sites of inflammation. In the central hypothesis, HLA proteins encoded by risk alleles allow more autoreactive TCRs (in red) to survive thymic selection (phase 2). In the peripheral hypothesis, HLA proteins encoded by risk alleles have a higher affinity to critical autoantigens; therefore, they can induce autoimmunity more efficiently (phase 5). Using nonproductive CDR3s, which do not experience thymic selection pressure, we can observe T cell biology in phase 1. Using peripheral blood data, we can observe T cell biology in phases 3 and 4. b, Study overview. We asked five major questions and conducted cdr3-QTL analysis using different models in both discovery and replication datasets (n = 797 donors in total). We conducted downstream investigations using the cdr3-QTL results, which include application of the CDR3 risk score (a score indicating the enrichment of CDR3 features favored by HLA risk alleles) to clinical TCR repertoire datasets (n = 22 patients in total).
stochastic processes in the thymus, HLA-DRB1 site 13 explained a striking 9.3% of interindividual variance in amino acid usage at L13-CDR3 position 109, which is comparable to the variance in V gene usage explained by HLA alleles29. For each CDR3 length, the HLA site that explained the most variance in amino acid usage was HLA-DRB1 site 13 (Supplementary Fig. 6). Even when we controlled the effect of potential confounders (ancestry, age, sex and cytomegalovirus (CMV) infection status, which can influence the TCR repertoire28), HLA-DRB1 site 13 continued to explain the most variance (Supplementary Fig. 7). CDR3 amino acid frequencies were less significantly associated with classical HLA alleles than to HLA-DRB1 site 13 (Supplementary Table 4 and Supplementary Fig. 8). Intriguingly, HLA-DRB1 site 13 is known to drive the risk of multiple autoimmune diseases. This is the residue that explains the most heritability for RA1,29 and for juvenile idiopathic arthritis30. It represents the strongest association to T1D after HLA-DQB1 site 57 (ref. 3).

To reproduce these effects, we obtained a replication dataset of 169 healthy individuals consisting of RNA sequencing (RNA-seq) data from sorted naïve CD4+ T cells (1,883 unique CDR3 Beta chain sequences per individual; Table 1) (ref. 31). We observed that the variance explained in the MMLM analysis was similar between replication and discovery datasets (Pearson’s r = 0.65) and the strongest association was again between HLA-DRB1 site 13 and L13-CDR3 position 109 (Extended Data Fig. 5a and Supplementary Table 5).

To assess whether there were independent effects outside of HLA-DRB1 site 13, we conducted serial conditional haplotype analyses within HLA-DRB1 in the discovery dataset (Methods). In order of descending significance, sites 71, 32, 74, 86 and 30 of HLA-DRB1 showed independently significant signals (Fig. 3 and Extended Data Fig. 6a). Including site 13, these 6 sites in total explained up to 20% of the variance in CDR3 middle position amino acid frequencies, with about half of this variance explained by site 13 (Extended Data Fig. 6b). Among these sites, three (13, 71 and 74) face the P4 antigen-binding pocket of HLA-DRB1, suggesting that the HLA-DRB1 P4 pocket plays a critical role in shaping the TCR repertoire (Extended Data Fig. 6a). Further conditional analyses outside of HLA-DRB1 revealed signals at both class I and II HLA genes. In order of descending significance, we observed independently significant associations at HLA-B, HLA-DQB1, HLA-DPB1 and HLA-DQA1 (Extended Data Fig. 7).

MHC–peptide–TCR complex structures. To understand whether these associations are related to the positioning of residues in the MHC–peptide–TCR complex, we analyzed five X-ray crystallography-based structures32–34. As expected, the antigenic peptide was closer to the middle positions of CDR3, where the productive CDR3 ratio, explained by HLA alleles than were less significantly associated with classical HLA alleles than to HLA-DRB1 site 13 (Supplementary Table 4 and Supplementary Note).

### Table 1 Characteristics of the datasets used in this study

| Discovery dataset (628 healthy donors) | Replication dataset (169 healthy donors) |
|---------------------------------------|----------------------------------------|
| Reference Emerson et al.26 Chen et al.27 |
| Sample Peripheral blood Naïve CD4+ T cells |
| Sequecning method TCR target sequencin RNA-seq |
| TCR chains Beta chain Alpha and Beta chains |
| Material DNA Messenger RNA |
| n of unique CDR3s per individual, mean (s.d.) 242,461 (102,355) 1,883 (689) |
| Productive CDR3 ratio, mean (s.d.) 82.1% (2.0%) 94.4% (0.96%) |
| Genotype data Genome-wide No Yes |
| HLA genotypes Direct typing Imputed |
| Age Mean 39.5 57.1 |
| s.d. 14.0 11.4 |
| Sex, n Male 324 72 |
| Female 282 97 |
| Unknown 22 0 |
| CMV infection status, n Positive 271 No data |
| Negative 334 |
| Unknown 23 |

### Notes

To reproduce these effects, we obtained a replication dataset of 169 healthy individuals consisting of RNA sequencing (RNA-seq) data from sorted naïve CD4+ T cells (1,883 unique CDR3 Beta chain sequences per individual; Table 1) (ref. 31). We observed that the variance explained in the MMLM analysis was similar between replication and discovery datasets (Pearson’s r = 0.65) and the strongest association was again between HLA-DRB1 site 13 and L13-CDR3 position 109 (Extended Data Fig. 5a and Supplementary Table 5). To assess whether there were independent effects outside of HLA-DRB1 site 13, we conducted serial conditional haplotype analyses within HLA-DRB1 in the discovery dataset (Methods). In order of descending significance, sites 71, 32, 74, 86 and 30 of HLA-DRB1 showed independently significant signals (Fig. 3 and Extended Data Fig. 6a). Including site 13, these 6 sites in total explained up to 20% of the variance in CDR3 middle position amino acid frequencies, with about half of this variance explained by site 13 (Extended Data Fig. 6b). Among these sites, three (13, 71 and 74) face the P4 antigen-binding pocket of HLA-DRB1, suggesting that the HLA-DRB1 P4 pocket plays a critical role in shaping the TCR repertoire (Extended Data Fig. 6a). Further conditional analyses outside of HLA-DRB1 revealed signals at both class I and II HLA genes. In order of descending significance, we observed independently significant associations at HLA-B, HLA-DQB1, HLA-DPB1 and HLA-DQA1 (Extended Data Fig. 7).

MHC–peptide–TCR complex structures. To understand whether these associations are related to the positioning of residues in the MHC–peptide–TCR complex, we analyzed five X-ray crystallography-based structures32–34. As expected, the antigenic peptide was closer to the middle positions of CDR3, where the productive CDR3 ratio, explained by HLA alleles than were less significantly associated with classical HLA alleles than to HLA-DRB1 site 13 (Supplementary Table 4 and Supplementary Note).

### Notes

To reproduce these effects, we obtained a replication dataset of 169 healthy individuals consisting of RNA sequencing (RNA-seq) data from sorted naïve CD4+ T cells (1,883 unique CDR3 Beta chain sequences per individual; Table 1) (ref. 31). We observed that the variance explained in the MMLM analysis was similar between replication and discovery datasets (Pearson’s r = 0.65) and the strongest association was again between HLA-DRB1 site 13 and L13-CDR3 position 109 (Extended Data Fig. 5a and Supplementary Table 5). To assess whether there were independent effects outside of HLA-DRB1 site 13, we conducted serial conditional haplotype analyses within HLA-DRB1 in the discovery dataset (Methods). In order of descending significance, sites 71, 32, 74, 86 and 30 of HLA-DRB1 showed independently significant signals (Fig. 3 and Extended Data Fig. 6a). Including site 13, these 6 sites in total explained up to 20% of the variance in CDR3 middle position amino acid frequencies, with about half of this variance explained by site 13 (Extended Data Fig. 6b). Among these sites, three (13, 71 and 74) face the P4 antigen-binding pocket of HLA-DRB1, suggesting that the HLA-DRB1 P4 pocket plays a critical role in shaping the TCR repertoire (Extended Data Fig. 6a). Further conditional analyses outside of HLA-DRB1 revealed signals at both class I and II HLA genes. In order of descending significance, we observed independently significant associations at HLA-B, HLA-DQB1, HLA-DPB1 and HLA-DQA1 (Extended Data Fig. 7).
1,354 CDR3 phenotypes were associated with at least one HLA for 1,249,742 total tests; Supplementary Table 6). A total of 388 so far focused on interindividual variance in overall CDR3 combinations), 15,060 of them were significant ($P < 0.05$ for 24,360 total tests). We observed that in the embedded space, the CDR3 loop and position 109 was mediated by an indirect physical interaction set of antigen positions, arguing that the association between site 13 and CDR3 position 109 were close to a common HLA-DRB1 Supplementary Note). We observed that in the embedded space, the paths were well preserved in this embedding. c, Structure of HLA-DR, the antigenic peptide, and CDR3 (PDB ID: 2IAM). HLA-DRB1 sites 13 and 71 and CDR3 position 109 are highlighted in magenta, orange and green, respectively. On the left, we depict the antigen (red) and Beta chain CDR3 (dark blue) overlaid onto HLA-DR molecules, looking into the binding groove. On the right, we depict the same complex from a side view. The shortest paths between site 13 and antigenic peptide and those between position 109 to the antigenic peptide are shown as black lines. d, Two-dimensional embedding plot based on the pairwise distances between the amino acids of HLA-DRB1, the CDR3 loop and the antigenic peptide. We down-weighted the distances between HLA and CDR3 so that their antigen-mediated indirect interaction was highlighted (Methods). The shortest paths between site 13 and antigenic peptide and those between position 109 and antigenic peptide are shown as black lines as in c, the paths were well preserved in this embedding.

Fig. 2 | HLA-DRB1 site 13 strongly influences CDR3 amino acid composition. a, MANOVA test $P$ values in the MMLM analysis ($n=628$; the discovery dataset). At each HLA site, $P$ values for all CDR3 phenotypes were plotted. The HLA site with the lowest $P$ value (HLA-DRB1 site 13) is highlighted by a diamond. The dashed line indicates the significance threshold with Bonferroni multiple testing correction ($P < 0.05$ for 24,360 total tests). b, Variance explained in the MMLM analysis ($n=628$; the discovery dataset). For each HLA site-CDR3 position pair, the largest variance explained across all CDR3 lengths is depicted in a heatmap showing all polymorphic sites of HLA-DRB1 (other HLA genes are shown in Extended Data Fig. 4). The pair with the largest variance is indicated by an asterisk. Above the heatmap, we present the shortest distance between each CDR3 position and any antigenic peptide residue; middle positions of CDR3 are highlighted in magenta. To the right of the heatmap, we present the shortest distance between each HLA-DRB1 site and any antigenic peptide residue (the bar plot on the left) and that between each HLA-DRB1 site and any CDR3 position (the bar plot on the right); site 13 is highlighted in magenta. Distances were averaged across the five X-ray crystallography structures (Methods). c, Structure of HLA-DR, the antigenic peptide, and CDR3 (PDB ID: 2IAM). HLA-DRB1 sites 13 and 71 and CDR3 position 109 are highlighted in magenta, orange and green, respectively. On the left, we depict the antigen (red) and Beta chain CDR3 (dark blue) overlaid onto HLA-DR molecules, looking into the binding groove. On the right, we depict the same complex from a side view. The shortest paths between site 13 and antigenic peptide and those between position 109 to the antigenic peptide are shown as black lines. d, Two-dimensional embedding plot based on the pairwise distances between the amino acids of HLA-DRB1, the CDR3 loop and the antigenic peptide. We down-weighted the distances between HLA and CDR3 so that their antigen-mediated indirect interaction was highlighted (Methods). The shortest paths between site 13 and antigenic peptide and those between position 109 and antigenic peptide are shown as black lines as in c, the paths were well preserved in this embedding.

TCR and antigens (Pearson’s $r > 0.9$; Supplementary Fig. 9b and Supplementary Note). We observed that in the embedded space, HLA-DRB1 site 13 and CDR3 position 109 were close to a common set of antigen positions, arguing that the association between site 13 and position 109 was mediated by an indirect physical interaction through antigenic peptide residues (Fig. 2d).

Amino acid-level cdr3-QTL signals. Our cdr3-QTL analyses have so far focused on interindividual variance in overall CDR3 composition explained by HLA alleles. We next applied a linear regression model (LM) to examine more specific relationships: for all possible pairs of CDR3 amino acids and HLA alleles, we tested how much interindividual variation in the CDR3 amino acid frequency could be predicted by the HLA allele count (question 2 in Fig. 1b and Extended Data Fig. 2). Permutation analyses using the LM similarly demonstrated that the $P$ values were well calibrated (Supplementary Fig. 10). Of the 1,249,742 total tests ($923$ HLA amino acid alleles×$1,354$ CDR3 phenotypes: length-position-amino acid combinations), $15,060$ of them were significant ($P < 4.0 \times 10^{-8} = 0.05$ for 1,249,742 total tests; Supplementary Table 6). A total of $388$ of 1,354 CDR3 phenotypes were associated with at least one HLA amino acid allele; we detected multiple CDR3 modification patterns (Supplementary Fig. 11). The effect sizes from the discovery and replication datasets were significantly correlated (Pearson’s $r = 0.76$; $P = 5.4 \times 10^{-70}$; Extended Data Fig. 5 and Supplementary Table 7). Using this model, we reevaluated the potential impacts of founders and confirmed that those effects on cdr3-QTL results were minimal (Supplementary Fig. 12).

Our results suggested that the HLA–peptide complex drives the associations between HLA alleles and CDR3 phenotypes. However, some associations could be mediated by HLA interactions with V and J genes. To rule out this possibility, we used linear mixed models (LMMs) to reestimate cdr3-QTL effects while accounting for the potential effects of V and J genes. In these models, the cdr3-QTL signals were unchanged, confirming that these associations were independent of V and J gene usage (Supplementary Figs. 1c and 2c).

Thymic selection may be driving HLA–CDR3 associations. To investigate the biological site and timing of HLA allelic effects on TCR composition, we considered five T cell phases: (1) thymic T cells preselection; (2) thymic T cells postselection; (3) naïve T cells in the peripheral blood; (4) memory T cells in the peripheral blood;
and (5) memory T cells after disease onset (Fig. 1a). The consistent cdr3-QTLs between peripheral blood (the discovery dataset with both naïve and memory T cells) and naïve T cells (the replication dataset) suggested cdr3-QTL influence before the development of the naïve TCR repertoire (phase 1 or phase 2). To assess the possibility of cdr3-QTL effects during phase 1, we tested for cdr3-QTLs among nonproductive TCRs and productive CDR3s with TRBV21-1, a pseudogene that renders the TCR nonfunctional and we observed no evidence of association (Fig. 4a,b and Supplementary Fig. 13). These results argue against the hypothesis that cdr3-QTLs reflect genetic biases in random recombination before thymic selection. Next, we evaluated the possibility that cdr3-QTLs are driven by antigen presentation in the periphery, during T cell phase 4. cdr3-QTL signals were not enriched among clonally expanded cell populations (Supplementary Fig. 14). Moreover, including clone size as a weight in fact attenuated cdr3-QTL signals (explained variance reduced by 47.3% on average; Fig. 4c,d). Taken together, these results argue that cdr3-QTLs reflect thymic selection favoring different CDR3 sequence features in the context of different HLA alleles (see Supplementary Note for more detailed discussion).

**CDR3 patterns associated with autoimmunity risk.** Since the HLA site that explained the most variance in CDR3 composition (HLA-DRB1 site 13) was the site with the strongest association to RA risk, we hypothesized that HLA risk for RA could be partially mediated by TCR composition. If HLA risk for RA is mediated by cdr3-QTLs, the effect sizes of the six possible amino acid alleles at HLA-DRB1 site 13 on RA risk should track with their
Aspartic acid; the magnitude of these effects was strongly correlated while those that protect against RA decrease the frequency of aspartic acid (a negatively charged amino acid), site 13 amino acids that raise the risk for RA increase HLA-DRB1 risk score is a risk score for RA, T1D and CD. Briefly, this HLA risk score is a genetic risk score that is a product of two parameters: (1) the number of disease-associated HLA haplotypes of each individual; and (2) the effect sizes of each haplotype estimated in the previous genetic studies1,2,5 (Fig. 6a and Supplementary Table 8). We calculated HLA risk scores in the discovery dataset and tested for CDR3 associations using an LM (question 3 in Fig. 1b). Out of 1,354 CDR3 phenotypes (combinations of length, positions and amino acid), we observed significant associations for 83, 187 and 119 phenotypes for RA, T1D and CD risk scores, respectively (P < 3.7 × 10^{-5} = 0.05 for 1,354 total tests; Supplementary Table 9). We observed weaker associations to V/J gene usage than to CDR3 patterns (Supplementary Fig. 16), suggesting that the main target of autoimmune risk is CDR3 composition rather than V/J genes.

To illustrate the CDR3 amino acid patterns associated with autoimmune risk alleles, we created sequence logos (Fig. 6b and Extended Data Fig. 9). Interestingly, we noted that amino acids with similar biochemical features demonstrated similar trends. This suggested that there may be latent biochemical features driving these disease-specific cdr3-QTLs. To quantify these trends, we examined the five broad aggregate amino acid features35: charge, hydrophobicity, refractivity, propensity for canonical secondary structures and molecular size (Methods). At each position of each length of CDR3, quantitatively ranked amino acids according to these features and raised the hypothesis that a negative charge at position 110 is partially observed in other CDR3 lengths (Supplementary Fig. 15). Interestingly, we observed the opposite finding for lysine, a positively charged amino acid (r = 0.76). Interestingly, we observed the opposite finding for lysine, a positively charged amino acid (r = 0.76). We observed the opposite finding for lysine, a positively charged amino acid (r = 0.76).

Motivated by these results, we aimed to extend our understanding of CDR3 amino acid patterns associated with autoimmune disease risk. Since autoimmune risk is driven by multiple HLA alleles, analysis of a single HLA amino acid allele might fail to detect important CDR3 patterns. To directly infer the comprehensive influence of HLA alleles on CDR3 restriction, we defined a multiallelic HLA risk score for RA, T1D and CD. Briefly, this HLA risk score is a genetic risk score that is a product of two parameters: (1) the number of disease-associated HLA haplotypes of each individual; and (2) the effect sizes of each haplotype estimated in the previous genetic studies1,2,5 (Fig. 6a and Supplementary Table 8). We calculated HLA risk scores in the discovery dataset and tested for CDR3 associations using an LM (question 3 in Fig. 1b). Out of 1,354 CDR3 phenotypes (combinations of length, positions and amino acid), we observed significant associations for 83, 187 and 119 phenotypes for RA, T1D and CD risk scores, respectively (P < 3.7 × 10^{-5} = 0.05 for 1,354 total tests; Supplementary Table 9). We observed weaker associations to V/J gene usage than to CDR3 patterns (Supplementary Fig. 16), suggesting that the main target of autoimmune risk is CDR3 composition rather than V/J genes.

To illustrate the CDR3 amino acid patterns associated with autoimmune risk alleles, we created sequence logos (Fig. 6b and Extended Data Fig. 9). Interestingly, we noted that amino acids with similar biochemical features demonstrated similar trends. This suggested that there may be latent biochemical features driving these disease-specific cdr3-QTLs. To quantify these trends, we examined the five broad aggregate amino acid features35: charge, hydrophobicity, refractivity, propensity for canonical secondary structures and molecular size (Methods). At each position of each length of CDR3,
Antigens may have higher CDR3 risk scores. Although our TCRs recognize candidate pathogenic antigens (question 5 in Fig. 6c), we observed an increased hydrophobicity at position 109 was associated with the HLA risk score of all 3 diseases (Fig. 6c, Extended Data Fig. 10). Interestingly, negative charge at CDR3 position 110 might be involved in the pathogenesis of RA, as we observed that TCRs specific to citrullinated epitopes with HLA-DQ2 tetramers, we collected TCR sequences specific to these three epitopes and control nonspecific TCRs (Supplementary Table 11). We observed that gliadin-specific TCRs had a higher CD-CDR3 risk score than control TCRs (one-sided t-test P = 0.0058; Fig. 7a). When we restricted this analysis to individual epitopes, we observed that only α-II gliadin-specific TCRs had a significantly higher CD-CDR3 risk score than control TCRs (one-sided t-test P = 0.0021; Supplementary Table 11). The inter-epitope differences in CDR3 risk score might be useful to differentiate causal epitopes from those targeted by epitope spreading. However, the limited number of TCRs specific for individual epitopes warrants cautious interpretation of these results. Interestingly, the CDR3 sequence with the highest score was reactive to α-II gliadin and featured arginine at position 109, which is important for the recognition of α-II gliadin (Fig. 7a). Recognizing that subtle HLA genotypic differences could affect CDR3 scores, we next conducted an intradividual analysis by restricting the analysis to TCRs from the three individuals for whom control TCR data were available. Even in this stringent analysis, we still observed significant differences in CD-CDR3 risk score between α-II gliadin-specific and control TCRs (P = 0.04; an LM adjusted for donor-level effects).

We then analyzed TCR data from patients with RA. The hallmark of RA is the immune reaction to citrullinated antigens. We sequenced 6 TCRs specific to citrullinated aggrecaulgin (n = 5 patients), 5 TCRs specific to citrullinated cartilage intermediate layer protein (CILP) (n = 2 patients), 1 TCR specific to citrullinated vimentin (n = 1 patient) and one TCR specific to citrullinated enolase (n = 1 patient), which were identified by HLA-DRB1*0401 or HLA-DRB1*0404 tetramers (Supplementary Table 12). Since we did not have control TCR data from the same individuals, we prepared 1,753 control TCR sequences from an individual homozygous for HLA-DRB1*0401, the allele with the highest HLA risk for RA (Methods). We observed that TCRs specific to citrullinated epitopes had higher RA-CDR3 risk scores than control TCRs (one-sided
HLA risk increases the frequency of TCRs reactive to candidate pathogenic antigens.

**Discussion**

Our study demonstrated large effect size associations between HLA alleles and CDR3 amino acid compositions using a quantitative trait analysis framework for the TCR. We identified CDR3 amino acid patterns associated with MHC-wide risk for autoimmune diseases, which were enriched in T cells reactive to candidate pathogenic antigens. In future studies, CDR3 risk scoring can complement tetramer-based analyses by prioritizing pathogenic T cell populations solely based on TCR sequencing.

Three important points clarify the new elements of our study. First and most importantly, we identified that the same HLA site (HLA-DRB1 site 13) substantially influences autoimmune risk and TCR-CDR3 amino acid composition. This specific site’s influence on CDR3 composition was not previously described. Second, our cdr3-QTL signals are independent of V gene usage and the presence of public clonotypes (Supplementary Note). Therefore, previous studies that only analyzed V genes or public clonotypes could not have detected these associations. Third, our study demonstrates that HLA alleles influence specific amino acids at precise CDR3 positions. These HLA effects on individual TCR positions cannot be observed in clonotype- or V gene-level analyses that treat the entire TCR sequence as a single event.

Cumulative evidence supports the peripheral hypothesis, where HLA risk alleles increase the affinity for pathogenic antigens. Although our work provides genetic evidence to support the central hypothesis, these results by no means exclude the peripheral hypothesis. Rather, the combination of central and peripheral biology probably synergizes to drive autoimmunity risk.

Although our work provides various insights into HLA autoimmunity risk, we should address several limitations of this study. First, it is important to recognize that our investigation in the discovery dataset was limited to the TCR Beta chain. Although the Beta chain is more important for antigen specificity, future work is needed to assess whether there are also disease-relevant cdr3-QTLs.
3. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and further clarify the role of the central hypothesis.

2. Hu, X. et al. Additive and interaction effects at three amino acid positions in nearly twice as prevalent as CD8+ T cells in peripheral blood, our analysis was better powered to detect class II HLA-CD4+ associations. Future studies in sorted CD8+ T cells will be necessary to resolve the relative contribution of MHC class I. Lastly, the number of antigen-specific TCRs was limited to currently available data. Advances in high-throughput experimental systems will likely expand the ability to detect disease-relevant antigens and their specific T cell populations in the near future. We anticipate that this will expand our knowledge of pathogenic TCR patterns and will further clarify the role of the central hypothesis.

It is now clear that HLA risk alleles modulate the process of thymic selection and give rise to TCR repertoires that may be poised for autoreactivity. This finding reinvigorates a role for the central hypothesis in mediating interindividual differences in autoimmune disease risk.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of information, acknowledgements, peer review information; details of online content are available at https://doi.org/10.1038/s41588-022-01032-z.

Received: 3 August 2021; Accepted: 4 February 2022; Published online: 24 March 2022

References
1. Raychaudhuri, S. et al. Five amino acids in three HLA proteins explain most of the association between MHC and sero-positive rheumatoid arthritis. Nat. Genet. 44, 291–296 (2012).
2. 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).
3. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381 (2014).
4. Lenz, T. L. et al. Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat. Genet. 47, 1085–1090 (2015).
5. Gutierrez-Achury, J. et al. Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. Nat. Genet. 47, 577–578 (2015).
6. Stahl, E. A. et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42, 508–514 (2010).
7. Gebe, J. A., Swanson, E. & Kwok, W. W. HLA Class II peptide-binding and autoimmunity. Tissue Antigens 59, 78–87 (2002).
8. Busch, R., Kollberger, S. & Mellins, E. D. HLA associations in inflammatory arthritis: emerging mechanisms and clinical implications. Nat. Rev. Rheumatol. 15, 364–381 (2019).
9. Koning, F., Thomas, R., Rossjohn, J. & Toes, R. E. Coeliac disease and rheumatoid arthritis: similar mechanisms, different antigens. Nat. Rev. Rheumatol. 11, 450–461 (2015).
10. Scally, S. W. et al. A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. J. Exp. Med. 210, 2569–2582 (2013).
11. Ting, Y. T. et al. The interplay between citrullination and HLA-DRB1 polymorphism in shaping peptide binding hierarchies in rheumatoid arthritis. J. Biol. Chem. 293, 3256–3251 (2018).
12. Kwok, W. W., Domier, M. L., Raymond, M. L., Byers, P. & Nepom, G. T. Allele-specific motifs characterize HLA-DQ interactions with a diabetes-associated peptide derived from glutamic acid decarboxylase. J. Immunol. 156, 2171–2177 (1996).
13. Jabri, B. & Solliid, L. M. T cells in celiac disease. J. Immunol. 198, 3005–3014 (2017).
14. Molberg, Ø. et al. Tissue transglutaminase selectively modifies gluten peptides that are recognized by gut-derived T cells in celiac disease. Nat. Med. 4, 713–717 (1998).
15. Kim, C. Y., Quarsten, H., Bergseng, E., Khosla, C. & Solliid, L. M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc. Natl Acad. Sci. USA 101, 4175–4179 (2004).
16. Gough, S. C. L. & Simmonds, M. J. The HLA region and autoimmune disease: associations and mechanisms of action. Curr. Genomics 8, 453–465 (2007).
17. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. Nat. Rev. Immunol. 18, 325–339 (2018).
18. Crux, N. B. & Elahi, S. Human leukocyte antigen (HLA) and immune regulation: how do classical and non-classical HLA alleles modulate immune response to human immunodeficiency virus and hepatitis C virus infections? Front. Immunol. 8, 832 (2017).
19. Jung, D. & Alt, F. W. Unraveling V(D)J recombination: insights into gene regulation. Cell 116, 299–311 (2004).
20. Dupic, T., Marcou, Q., Walczak, A. M. & Mora, T. Genesis of the αβ T-cell receptor. PLoS Comput. Biol. 15, e1006874 (2019).
21. Lu, J. et al. Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. Nat. Commun. 10, 1019 (2019).
22. Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nat. Rev. Immunol. 14, 377–391 (2014).
23. Sakaguchi, N. et al. Altered thymic T-cell selection due to a mutation of the ZAP-70 gene causes autoimmune arthritis in mice. Nature 426, 454–460 (2003).
24. Kishimoto, H. & Sprent, J. A defect in central tolerance in NOD mice. Nat. Immunol. 2, 1025–1031 (2001).
25. Liston, A., Lesage, S., Wilson, J., Peltonen, L. & Goodnow, C. C. Aire regulates negative selection of organ-specific T cells. Nat. Immunol. 4, 350–354 (2003).
26. 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).
27. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. Nature 547, 94–98 (2017).
28. Sharon, E. et al. Genetic variation in MHC proteins is associated with T cell receptor expression biases. Nat. Genet. 48, 995–1002 (2016).
29. Okada, Y. et al. Contribution of a non-classical HLA gene, HLA-DOA, to the risk of rheumatoid arthritis. Am. J. Hum. Genet. 99, 366–374 (2016).
30. Hinks, A. et al. Fine-mapping the MHC locus in juvenile idiopathic arthritis (JIA) reveals genetic heterogeneity corresponding to distinct adult inflammatory arthritic diseases. Ann. Rheum. Dis. 76, 765–772 (2017).
31. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell 167, 1398–1414.e24 (2016).
32. Hennecke, J. & Wiley, D. C. Structure of a complex of the human αβ T cell receptor (TCR) HA1.7, Influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. J. Exp. Med. 195, 571–581 (2002).
33. Hahn, M., Nicholson, M. J., Pyrdol, J. & Wucherpfennig, K. W. Unconventional topology of self peptide–major histocompatibility complex binding by a human autoimmune T cell receptor. Nat. Immunol. 6, 490–496 (2005).
34. Deng, L. et al. Structural basis for the recognition of mutant self by a tumor-specific, MHC class II-restricted T cell receptor. Nat. Immunol. 8, 398–408 (2007).
35. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. Proc. Natl Acad. Sci. USA 102, 6395–6400 (2005).
36. Stadinski, B. D. et al. Hydrophobic CDR3 residues promote the development of self-reactive T cells. Nat. Immunol. 17, 946–955 (2016).
37. Christophersen, A. et al. Distinct phenotype of CD4+ T cells driving celiac disease identified in multiple autoimmune conditions. Nat. Med. 25, 734–737 (2019).
38. Qiao, S.-W. et al. Posttranslational modification of gluten shapes TCR usage in celiac disease. J. Immunol. 187, 3064–3071 (2011).
39. Dahal-Koirala, S. et al. TCR sequencing of single cells reactive to DQ2.5-glia-α2 and DQ2.5-glia-α2 reveals clonal expansion and epitope-specific V-gene usage. Mucosal Immunol. 9, 587–596 (2016).
40. Han, A. et al. Dietary gluten triggers concomitant activation of CD4+ and CD8+ αβ T cells and γδ T cells in celiac disease. Proc. Natl Acad. Sci. USA 110, 13073–13078 (2013).
41. Petersen, J. et al. T-cell receptor recognition of HLA-DQ2–gliadin complexes associated with celiac disease. Nat. Struct. Mol. Biol. 21, 480–488 (2014).
42. DeWitt, W. S.3rd. et al. Human γδ T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. eLife 7, e38358 (2018).
43. Madi, A. et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. eLife 6, e22057 (2017).
44. Kula, T. et al. T-Scan: a genome-wide method for the systematic discovery of T cell epitopes. Cell 178, 1016–1028.e13 (2019).
45. Bentzen, A. K. et al. Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. Nat. Biotechnol. 34, 1037–1045 (2016).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022
HLA genotypes. We downloaded genotype data for the discovery dataset from the Adaptative Immunotools immuneACCESS site (http://www.immunetools.com/pub/emerson-2017-natgen). The discovery dataset has the CDR3 sequences of 70 HLA loci from peripheral blood of healthy individuals (n = 663). This dataset has 242,461 unique CDR3 sequences per individual on average (Table 1). For the main analysis, we included the 628 individuals whose four-digit classical alleles were available for all of HLA-A, HLA-B, HLA-C, HLA-DQB1 and HLA-DRB1. As for the demographic data, age was available for 555 samples, sex was available for 642 samples and ancestry data was available for 414 samples (detailed information is provided in Table 1). We defined amino acid sequences with stop codons or frameshifs as nonproductive amino acid sequences. We used productive amino acid sequences of the CDR3 and V/J gene information as reported in the original data. For the amino acids of nonproductive CDR3 sequences, we used the MIXCR software v2.1.11 with default parameters since the original data did not report these sequences. For the primary analysis, we treated each unique CDR3 sequence as a single event irrespective of read depth to exclude the influence of clonal expansion. We considered the read depth of each CDR3 sequence as a weight to calculate amino acid frequencies only when treated each unique TCR sequence as a single event irrespective of read depth used productive amino acid sequences of the CDR3 and V/J gene information with stop codons or frameshifts as nonproductive amino acid sequences. We used the following LM:

\[ Y_i = \sum_{k=1}^{3} \pi_k PC_{k,i} + \theta \]

where \( Y_i \) indicates a scalar phenotype for individual \( i \), \( \pi_k \) is the value for the kth principal component and \( PC_{k,i} \) is the value for individual \( i \) for the kth principal component. \( \theta \) is an n-dimensional parameter that represents the intercept.

We estimated the improvement in model fit between the full and null model using MANOVA and assessed the significance of the improvement using Pillai’s trace. The interindividual variance in the n-dimensional phenotype vector explained by \( m - 1 \) genotype groups was estimated using the R package MVLM (v0.1.4). We calculated the variance explained by the full and null models separately and we defined the variance explained by genotypes as the difference between the two values. Although we conducted separate association tests for each CDR3 length, we provided the largest variance explained across all CDR3 lengths for each HLA site-CDR3 position pair when we prepared summary figures to provide concise results (for example, Fig. 2). In null datasets with permuted sample labels, the variance explained was minimal and the MANOVA \( P \) values were well calibrated (Extended Data Fig. 3).

For each amino acid residue at a given position of CDR3 and we used this information to create a multidimensional phenotype vector to represent all amino acids at that position. For each component of this phenotype vector, proportions were transformed into a standard normal distribution across individuals (rank-based).

At a given HLA site with m possible amino acid residues, we partitioned the classical alleles into m groups, each with identical residues at the given site. We then calculated the allele count of each group (Supplementary Fig. 3). We included the top three principal components of genotypes in this analysis. As a result, the full model is the following MMLM:

\[ Y_i = \sum_{a=1}^{A} \beta_a g_{a,i} + \sum_{k=1}^{3} \pi_k PC_{k,i} + \theta \]

where \( Y_i \) indicates the n-dimensional phenotype vector of individual \( i \) (assuming there are \( n \) possible amino acids at that CDR3 position), \( \beta_a \) is the ratio of the allele count of allele \( a \) in individual \( i \) in the replication dataset, we used the estimated allele dosage instead of the allele count since genotype data was imputed, \( \beta_a \) is an n-dimensional parameter that represents the additive effect per allele. We included \( m - 1 \) group of classical alleles, casting 1 group as the reference. \( \pi_k \) is an n-dimensional parameter that represents the effect of the kth principal component and \( PC_{k,i} \) is the value for individual \( i \) for the kth principal component. \( \theta \) is an n-dimensional parameter that represents the intercept.
HLA site-level conditional analysis. To test for independent cdr3-QTL signals within a given HLA gene, we conducted a conditional haplotype analysis using an MMLM and controlling all effects coming from specific sites of that HLA gene. The strongest cdr3-QTL signal was found at site 13 of HLA-DRB1. Therefore, in the first round of the conditional analysis, we conducted cdr3-QTL analysis by controlling the effects coming from site 13. The null model consisted of haplotypes defined only by alleles at site 13. The full model consisted of haplotypes defined by the combination of alleles at site 13 and the target site; addition of the target site resulted in a statistically significant unique haplotype if the site is independent from site 13. We tested whether the creation of 6 additional haplotype groups improved the model fit. In this analysis, the strongest signal was found within site 71 of HLA-DRB1. Therefore, in the second round of the conditional analysis, we conducted cdr3-QTL analysis by controlling the effects coming from sites 13 and 71. We repeated this process iteratively within HLA-DRB1 until we did not observe further significant signal (P < 0.05 for 24,430).

Structural analysis of the MHC–peptide–TCR complex. We downloaded the structural analysis results of MHC–peptide–TCR complexes from the Protein Data Bank (PDB) (https://www.rcsb.org). We restricted our analysis to results for HLA-DRB1 that had all positional data for all three molecules: the HLA proteins; the antigenic peptide; and the TCR Beta chain. These PDB entries include 1H18, 1YMM, 2IAM, 2IAN and 4E41. For each amino acid, we first calculated the centroids of every atom using the XYZ orthogonal coordinates reported in the database; next we calculated the pairwise distances between amino acids using the centroid positions. We calculated pairwise distances between HLA, TCR and antigenic peptide amino acids and embedded them into a 2D space. In this embedding, we preserved distances important for antigen recognition (distances between HLA-DRB1 and antigens and those between TCR and antigens) and down-weighted the distances between HLA-DRB1 and TCR (Supplementary Note).

We then visualized the HLA-DRB1 amino acid sites that had independent significance in the conditional haplotype analyses. We calculated the pairwise distances between HLA-DRB1 and TCR and antigenic peptides using the Chimera software (v1.14) to create a three-dimensional view of the MHC–peptide–TCR complex based on PDB entry 2IAM.

Replication analysis using naïve CD4+ T cells. The replication dataset was generated by the BLUEPRINT consortium and was downloaded from the European Genome-phenome Archive (EGA) under accession nos. EGAD00001002671 and EGAD00001002663 (https://www.ebi.ac.uk/ega/home). We analyzed the FASTQ files of the RNA-seq data of naïve CD4+ T cells (n = 169 healthy donors). Demographic data are provided in Table 1. Reads were mapped to GRCh38 human reference sequences with GENCODE v26 gene models by the STAR software v2.5.3. Using reads mapped to the TCR loci (chromosome 7: 142,299,011–142,813,287 for the Beta chain) and unmapped reads, we analyzed the TCR sequences using MIXCR v2.1.11 with default parameters. Using genotype data around the MHC locus, we imputed HLA genotypes using the SNP2HLA software (v1.0.3) using the T1DGC reference panel (n = 5,225 European samples) and excluded poor-quality genotypes (r² < 0.5). We conducted PCA using the PLINK software v.2.10 using linkage disequilibrium-pruned genome-wide variants (r² > 0.2).

HLA risk score. To evaluate the MHC-wide risk of RA, T1D and CD for each individual, we defined an HLA risk score for each disease. First, we defined the critical MHC types necessary for the CD analysis or the RA analysis using the results from the previous studies. For the RA analysis, we included four amino acid sites: HLA-DRB1 sites 11, 13, 71 and 74. For T1D, we included 3 amino acid sites: HLA-DQB1 site 57 and HLA-DRB1 sites 13 and 71. For CD, we included four-digit classical alleles of HLA-DRB1 and HLA-DQB1. Since many samples did not have HLA-DQA1 genotypes (n = 144), we did not include the genotype of HLA-DQA1. We used the odds ratio (OR) of each haplotype according to multivariate regression as reported in previous genome-wide association studies (GWAS) (Supplementary Table 8). We then calculated the product of effect sizes (log(OR)) and the count of those haplotypes in each individual; the sum of two products was defined as the HLA risk score of that individual. The HLA risk score has a different distribution for each disease. When we conducted regression using the HLA risk score, we used a Poisson regression model. The HLA risk score was scaled the HLA risk scores to have an s.d. = 1. Therefore, the effect sizes for the HLA risk scores of different diseases are comparable to each other.

Amino acid features. Amino acids have multiple complex physicochemical and biological properties. To analyze amino acid features comprehensively, we used five previously reported features that capture the entire constellation of amino acid physicochemical properties (each amino acid has a unique value for each feature). Briefly, these 5 features were derived from factor analysis based on 494 amino acid indices, which include general attributes (for example, molecular volume) as well as more specific measures (for example, side chain orientation angle). Based on the original report, we annotated factor I as charge, factor II as hydrophobicity, factor III as refractivity, factor IV as propensity for secondary structure and factor V as molecular size. For factor II, the original value indicated hydrophilicity and we flipped the sign so that the value represents hydrophobicity. When we conducted the cdr3-QTL analysis using these amino acid features, we calculated the weighted average of a given amino acid feature. (We multiplied the frequency of each amino acid by its value for a given feature and calculated the sum of the product.) When we conducted the cdr3-QTL analysis for a given feature, we added the other four features as covariates to handle the correlations between these features.

CDR3 risk score. We developed a scoring system that indicates the extent of each CDR3 sequence association with HLA risk score and we refer to this as the CDR3 risk score. This is analogous to polygenic risk score (PRS). We used the effect sizes of the HLA risk score on each amino acid residue in each position of CDR3 with each length (L12–L18) from an LM (that is, the effect sizes of the cdr3-QTL analysis based on the HLA risk score). The risk score is the sum of the effect sizes of amino acids in a given CDR3 sequence (Supplementary Fig. 17).

As in PRS, the P value threshold in cdr3-QTL analysis is an important tuning parameter for CDR3 risk score. Therefore, we determined an appropriate P value threshold with fivefold cross-validation in the discovery dataset (Supplementary Note). We conducted each round of cross-validation, we conducted cdr3-QTL analysis based on HLA risk score using 80% of samples; we used the effect sizes from this analysis to calculate CDR3 risk score in the testing samples (the remaining 20% of samples). The CDR3 risk score was expected to be correlated with HLA risk score based on its definition. Therefore, we evaluated the performance of the CDR3 risk score by using its correlation with the HLA risk score. Using the RA CDR3 risk score, we confirmed that the Bonferroni-corrected P < 0.05 for 1,368) provided the best performance (Supplementary Fig. 17). Therefore, for the main analysis of the CDR3 risk score, we included the 83, 187 and 119 CDR3 phenotypes (length-position-amino acid combinations) for RA, T1D and CD, respectively, that passed the Bonferroni-corrected P threshold (P < 0.05 for 1,368).

TCR sequences of tetramer-positive T cells. We analyzed TCR sequences that were reactive to candidate pathogenic epitopes for RA and CD. We restricted our analyses to CDR3s with a length between 12 and 18 amino acids and calculated the CDR3 risk scores for RA and CD. We used a one-sided t-test to assess the significance of the difference in CDR3 risk scores between the two groups.

For the RA analysis, we collected the peripheral blood of patients with RA (n = 7 in total; Supplementary Table 12) and expanded T cells using peptides corresponding to relevant citrullinated epitopes46,47. We conducted FACS using HLA-DRB1*0401- or HLA-DRB1*0404-tetramer loaded with citrullinated aggrecan (cit-aggrecan), citrullinated CILP (cit-CILP), citrullinated vimentin (cit-vimentin) or citrullinated enolase (cit-enolase), isolating and expanding single cells and then sequencing the TCRs of these T cells. We thus identified six cit-aggrecan-specific TCR sequences from five patients with RA, five cit-CILP-specific TCR sequences from two patients with RA, one cit-vimentin-specific TCR from one patient with RA and one cit-enolase-specific TCR from one patient with RA. The patients in this analysis had at least one allele of HLA-DRB1*0401 or HLA-DRB1*0404, although their HLA genotypes were not necessarily identical (Supplementary Table 12). Since we did not have control TCR sequences from the same patients, we utilized TCR sequences from the replication dataset (naïve CD4+ T cells in the peripheral blood of healthy individuals). We restricted the samples in the replication dataset to individuals who were regressive for HLA-DRB1*0401 (the allele with the highest HLA risk score). Since the CDR3 risk score is positively correlated with HLA risk score by definition, this strategy should identify individuals who have TCR repertoires with the highest CDR3 risk scores; hence, this strategy is a conservative approach. One individual with 1,753 TCR sequences met this criterion. Among the 20% control TCR samples that we observed, we searched the literature for studies that utilized tetramers to identify gladin epitope-specific CD4+ T cells and reported their sequences. Three studies met these criteria39,40 and included 7 TCRs specific to α-gliadin (n = 4 patients), 92 TCRs specific to α-gliadin (n = 13 patients), 8 TCRs specific to α-gliadin (n = 2 patients) and 49 control TCRs (n = 3 patients). Patients in these reports had at least one HLA-DQ2 haplotype (HLA-DQA1*0501 -HLA-DQB1*0201), although their HLA genotypes were not necessarily identical.

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

Data availability
All raw TCR sequence and genotype data of the discovery and the replication datasets are available at the Aderma Biotechnologies immuneACENS site and EGA under accession nos. EGAD00001002671 and EGAD00001002663. All summary statistics of the cdr3-QTL analysis are available at https://github.com/immunogenomics/cdr3-QTL.

Code availability
All code used in this study is available at https://github.com/immunogenomics/cdr3-QTL and deposited at Zenodo (https://zenodo.org/badge/latestdoi/306212645).

References
James, J. et al. Citrulline-specific Th1 cells are increased in rheumatoid arthritis and their frequency is influenced by disease duration and therapy. Arthritis Rheumatol. 66, 1712–1722 (2014).
47. Rims, C. et al. Citrullinated aggrecan epitopes as targets of autoreactive CD4+ T cells in patients with rheumatoid arthritis. *Arthritis Rheumatol.* 71, 518–528 (2019).

48. 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).

Acknowledgements
This work was supported in part by funding from the National Institutes of Health (grant nos. 1-U01-HG012009 to S.R., AR063759-05 to S.R., U01-HG09379-04 to S.R., U19-AI111224-06 to S.R. and T32GM007753 to K.A.L.). K.I. was supported by The Uehara Memorial Foundation. We thank M. Brenner, A. H. Jonsson and D. Rao for helpful feedback.

Author contributions
K.I. and S.R. conceived and designed the study. K.I. conducted all the analyses with support from K.A.L., Y.L. and S.R. E.A.J. and J.H.B. generated and managed the TCR data from patients with RA. K.I. and S.R. wrote the initial draft of the manuscript. All co-authors contributed to the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01032-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01032-z.

Correspondence and requests for materials should be addressed to Soumya Raychaudhuri.

Peer review information *Nature Genetics* thanks Corey Watson, Giulio Isacchini and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1  |  TCR structure in the discovery dataset.

a, The amino acid positioning scheme used in this study. Additional amino acids in longer CDR3s align to middle positions. 

b, Schematic explanation of the structure of CDR3. During T cell development in thymus, TCRs are generated by randomly recombining component genes (V, D, and J gene for beta chain). In addition, several nucleotides are randomly added or deleted at the junctional regions.

c, The distribution of CDR3 amino acid length in the discovery dataset.

d, The diversity and mutual information of amino acid composition at CDR3 positions (length = 15 amino acids). Normalized entropy (bar plot) and normalized mutual information (NMI, heatmap) of amino acid usage at each position of CDR3 and V/J gene usage were calculated in each individual, and the averaged values are provided. In the top heatmap, NMI is shown in a linear scale. In the bottom heatmap, NMI is shown in log scale. CDR3 positions 107–116, which directly contact antigenic peptides, are highlighted in red (b and d).
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Statistical models used in this study. a, Strategy to calculate amino acid frequencies in the main analysis. In this example, the alanine (A) usage ratio at CDR3 position 110 is calculated for each individual. b, Strategy to calculate amino acid frequencies for the linear mixed model (LMM) used to adjust the effect of V genes. In this example, alanine (A) usage ratio at CDR3 position 110 was calculated for each individual for each V gene. c, Schematic explanation of LM and MMLM, the two main linear models in this study. Each square indicates the dimensions of the matrix. In LM, the frequency of a single amino acid at a position of CDR3 is the response variable; the count of a single amino acid allele at a site of HLA is the explanatory variable. In MMLM, a vector of frequency of all amino acids at a given position of CDR3 is the response variable; the counts of all amino acid alleles except one at the HLA site are the explanatory variables. When we have 20 CDR3 phenotypes and the five HLA alleles, we need to conduct 100 LM tests to cover all combinations (as shown in d). On the other hand, MMLM model aggregates all 100 combinations into one single association test, maximizing the power of detecting associations. cov, covariates. d, On the left, we provide an association plot between the allele count of arginine (R) in HLA-DRB1 site 13 and the frequency of lysine (K) in position 109 of L13-CDR3. The P value from the LM analysis is provided (n = 628 donors; two-sided linear regression test). On the right, we provide a heatmap showing P values from all 100 LM tests. We also provide a heatmap showing the P values from a permuted dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range (IQR), and the whiskers reflect the maximum and minimum values within each grouping no further than 1.5 x IQR from the hinge.
Extended Data Fig. 3 | Permutation analyses using the MMLM. a, MANOVA test \( P \) values in cdr3-QTL analysis using the MMLM with permuted sample labels (\( n=628 \); the discovery dataset). At each HLA site, \( P \) values of all CDR3 phenotypes are plotted. The black dashed line corresponds to the Bonferroni \( P \) value threshold (\( P=0.05/24,360 \) total tests). b, QQ plots of MANOVA test \( P \) values in cdr3-QTL analysis using the MMLM with the real and the permuted sample labels (\( n=628 \)). Both have 24,360 data points. c, The distribution of minimum \( P \) values (\( P_{\text{min}} \)) using the MMLM in each round of the 1,000 permutations (MANOVA test). We restricted this analysis to alleles at HLA-DRB1 site 13. In each round of permutation, we tested associations for all CDR3 positions (70 length-position combinations). The bottom 5 percentile of \( P_{\text{min}} \) was \( 8.6 \times 10^{-4} \), almost identical to the Bonferroni \( P \) value threshold (\( = 0.05/70 \) total tests \( = 7.1 \times 10^{-4} \)), which indicates that our \( P \) values are well calibrated. d, The distribution of variance in amino acid composition at position 109 of L13-CDR3 explained by the alleles at HLA-DRB1 site 13 in each round of the 1,000 permutations. Red vertical line denotes the observed variance explained in unpermuted data.
Extended Data Fig. 4 | Variance explained in the MMLM analysis summarized across different lengths of CDR3. Variance explained in the MMLM analysis (n = 628; the discovery dataset). The results for all HLA genes except HLA-DRB1 are provided. For each HLA site-CDR3 position pair, the largest variance explained across different CDR3 lengths is shown in a heatmap.
Extended Data Fig. 5 | MMLM and LM results in the replication dataset. **a**, Explained variance in the MMLM analysis in the discovery dataset (n = 628; peripheral blood) compared with that in the replication dataset (n = 169; naïve CD4+ T cells). All pairs of class II HLA sites and CDR3 phenotypes are shown without any filtering (9,735 data points). The results at HLA-DRB1 site 13 and the results with P < 0.05 in the replication dataset are highlighted. **b**, Explained variance in the MMLM analysis in the replication dataset (n = 169; naïve CD4+ T cells). For each HLA site-CDR3 position pair, the largest variance explained across different CDR3 lengths are shown in a heatmap. The results of HLA-DRB1 are provided. Only associations with P < 0.05 are colored in the heatmap. The results both for alpha and beta chains are provided. The pair with the largest variance is indicated by an asterisk. **c**, LM analysis using the replication dataset (n = 169; naïve CD4+ T cells). Effect sizes for non-transformed phenotypes from discovery and replication datasets are provided. The error bar indicates ± 2 × s.e. The nominally significant associations in the replication dataset are highlighted in red (P < 0.05). The analysis was restricted to the 388 CDR3 phenotypes (length-position-amino acid combinations) that had at least one significant association in the LM analysis (P < 0.05/1,249,742 total tests) and were testable in the replication dataset. For each CDR3 phenotype, we used the HLA amino acid allele that had the lowest P value for that phenotype in the LM analysis of the discovery dataset. We used P values from two-sided linear regression test.
Extended Data Fig. 6 | Six sites in HLA-DRB1 have independently significant cdr3-QTL effects. a, Structure of HLA-DRB1 protein and amino acid sites with independently significant cdr3-QTL effects (Protein database 2IAM). Positions 13, 71 and 74 are within the P4 binding pocket. On the left, we depict only HLA-DR molecules looking into the binding groove. In the middle, we depict the antigen (red) and CDR3 (dark blue) overlaid onto HLA-DR molecules. On the right, we depict HLA-DR, antigen, and CDR3 from a side view. b, Variance explained by six HLA-DRB1 amino acid sites with independently significant cdr3-QTL effects (n = 628; MMLM; the discovery dataset). The order of sites on the x-axis indicates the order of significance. c, The distances from HLA-DRB1 amino acid sites to antigen (Ag) or to CDR3. We analyzed five structures and the shortest distances in each structure were used. One-sided paired t test P values are provided (n = 5). Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range (IQR), and the whiskers reflect the maximum and minimum values within each grouping no further than 1.5 × IQR from the hinge.
Extended Data Fig. 7 | Conditional analysis using four-digit classical alleles. Conditional analysis using four-digit classical alleles (n = 628; MMLM; the discovery dataset). In the first conditioning analysis, to assess whether there were independent effects outside of the HLA-DRB1 locus, we conducted cdr3-QTL analysis using all four-digit classical alleles of HLA-DRB1 as covariates, and the strongest signal was found in HLA-B region. In the second conditional analysis, we additionally included all four-digit classical alleles of HLA-B as covariates. We sequentially included as covariates all four-digit classical alleles of the gene with the strongest signal until we did not observe further significant signal (P > 0.05/24,360 total tests). We excluded strongly correlated alleles among covariates (r² > 0.8). We reported MANOVA test P values.
Extended Data Fig. 8 | The pair-wise distances of amino acids in MHC-peptide-TCR complexes. The distances (in Å) between HLA-DRB1 sites and antigen (left), CDR3 amino acids of beta chains and antigen (middle), and HLA-DRB1 sites and CDR3 amino acids of beta chains (right) are shown in heatmaps.
Extended Data Fig. 9 | CDR3 amino acids associated with MHC-wide risk of RA, T1D, and CD. CDR3 amino acids influenced by HLA risk score. We conducted the LM analysis using HLA risk score; the CDR3 phenotypes were each amino acid at each position of each length of CDR3 (n = 628; the discovery dataset). a-c, The effect sizes of significant associations for each amino acid at a given position are illustrated by sequence logo (P < 0.05/1,354 total test), separately for different CDR3 lengths (a, RA; b, T1D; and c, CD). We used P values from two-sided linear regression test.
Extended Data Fig. 10 | Amino acid features at each position of CDR3 influenced by HLA risk score. We conducted the LM analysis using HLA risk score in which the phenotypes were amino acid features at a given position of each length of CDR3 (n=628; the discovery dataset). a, Effect sizes were plotted separately for different lengths of CDR3. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range (IQR), and the whiskers reflect the maximum and minimum values within each grouping no further than 1.5 x IQR from the hinge. b, Meta-analyzed effect sizes were plotted (the results for charge and hydrophobicity are shown in Fig. 6c). The error bar indicates ± 2 x s.e.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a  Confirmed

☑️ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☑️ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☑️ The statistical test(s) used AND whether they are one- or two-sided
  *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☑️ A description of all covariates tested
☑️ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☑️ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☑️ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  *Give P values as exact values whenever suitable.*
☑️ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☑️ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☑️ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection  No software was used.

Data analysis  We used publicly available softwares for the analysis as follows.
  - MIXCR (v2.1.11)
  - STAR (v2.5.3)
  - Plink (v1.90)

All code used in this study is available at our website [https://github.com/immunogenomics/cdr3-QTL] and deposited at Zenodo [https://zenodo.org/badge/latestdoi/306212645].

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
  - Accession codes, unique identifiers, or web links for publicly available datasets
  - A list of figures that have associated raw data
  - A description of any restrictions on data availability

All raw TCR sequence data and genotype data of the discovery dataset and the replication dataset are available at Adaptive Biotechnologies immuneACCESS site.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-list.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | We used all available data without sample size calculation because we utilized publicly available datasets. |
|------------|---------------------------------------------------------------------------------------------------------|
| Data exclusions | We excluded the TCR phenotypes and HLA variants based on the standard quality control procedure in genetics study. Briefly, we excluded rare phenotypes and rare genotypes from our analysis. This is because rare observation makes statistics unstable. The criteria were a part of standard QC steps in human genetics research. Detailed information was described in the method section. |
| Replication | We conducted replication analysis using an independent dataset. Detailed information was described in the method section. |
| Randomization | Randomization is not relevant for our study because it is a retrospective study. |
| Blinding | Blinding was not relevant in our study because it is a retrospective study. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|---------------------------------|---------|
| n/a                             | n/a     |
| ☑ Antibodies                    | ☑ ChiP-seq |
| ☑ Eukaryotic cell lines         | ☑ Flow cytometry |
| ☑ Palaeontology and archaeology | ☑ MRI-based neuroimaging |
| ☑ Animals and other organisms   |         |
| ☑ Human research participants   |         |
| ☑ Clinical data                 |         |
| ☑ Dual use research of concern  |         |