Supplementary Note

Estimating the influence of germline-encoded sequences

The aim of our cdr3-QTL analysis is to detect whether *HLA* allelic variation influences amino acid usage in the randomly recombined region of TCR that contacts antigenic peptides. These residues localize to complementary determining region 3 (CDR3), which is the product of V, D, and J gene joining during random recombination. Though most CDR3 residues arise from random nucleotide additions between V and J genes, flanking CDR3 residues are clearly influenced by these germline segments (**Extended Data Figure 1**).

To estimate the potential bias in our cdr3-QTL results from these germline-encoded sequences, we compared our initial linear regression model (LM) results to those from two linear mixed models (LMM; one for the V gene and one for the J gene). Notably, this "initial" LM analysis does not exclude germline-encoded amino acids and is not the main results in our manuscript. For these LMMs, instead of calculating CDR3 amino acid frequencies within each donor, we calculated CDR3 amino acid frequencies within the subset of TCRs expressing each V gene (or J gene) within each donor (**Extended Data Figure 2**). For each group of TCRs, we included the V gene (or J gene) as a fixed effect, and the donor ID as a random effect, since there were multiple observations per donor in this design. We restricted our analysis to the 435 CDR3 phenotypes (length-position-amino acid combinations) which had at least one significant association in the initial linear regression (LM) analysis ($P < 0.05/1{,}262{,}664$ total tests). For each CDR3 phenotype, we used the *HLA* amino acid allele that had the lowest *P* value for that phenotype in the LM analysis.

Among 435 CDR3 phenotypes, we observed some discordant effects between the initial LM and the subsequent LMM: four phenotypes had significant heterogeneities ($P_{het} <$ 0.05/total tests; **Supplementary Figure 1a**). Discrepancies were enriched at CDR3 positions 107 and 108 (**Supplementary Figure 1b**), where amino acids are moderately correlated with V genes (**Extended Data Figure 1d**). We observed the strongest heterogeneity for serine at position 107 of CDR3 length 12 (L12:P107:S). To better understand this heterogeneity, we conducted cdr3-QTL analyses of L12:P107:S

stratified by V gene. These results confirmed that cdr3-QTL signals for L12:P107:S are highly dependent on the V gene used by the TCR: we observed large effects on serine usage for TCRs with certain V genes, and small effects for TCRs with others (**Supplementary Figure 1b**).

We next conducted the same analyses for J genes (**Supplementary Figure 2**), comparing the initial LM results to those from the LMM conditioned on J genes. As in the V gene analysis, we observed some discordant effects between LMM and LM: four phenotypes had significant heterogeneities (P_{het} < 0.05/all tests). Discrepancies were enriched at CDR3 positions ≧ 113, where amino acids are moderately correlated with J genes (**Extended Data Figure 1d**).

These analyses revealed that a small subset of cdr3-QTL signals, mostly in the flanking positions of CDR3, were dependent on the V/J gene usage. *HLA*-V gene associations[1] have been previously reported, and could possibly mediate the cdr3-QTL signals of interest in our study. Of note, the discovery dataset does not distinguish V and J gene alleles, and so analysis of V/J allelic effects was not possible.

Thus, to filter out these V/J gene-dependent cdr3-QTL signals, we excluded germline-encoded amino acids from each of the individual CDR3 sequences in our main analysis. Our strategy is illustrated in **Supplementary Figure 3**. Most signal within the CDR3 region was unchanged, but many of the initially significant associations in CDR3 flanking positions lost significance (**Supplementary Figure 4**). Moreover, this exclusion resolved the discrepancies between the LMM that conditioned on V/J gene usage and the LM that did not (**Supplementary Figure 1c** and **2c**). Thus, our reported cdr3-QTL associations that exclude germline-encoded sequences are independent of V/J gene usage and orthogonal to *HLA*-V gene associations.

Replication analysis using a multivariate multiple linear regression model

To reproduce the results from the discovery dataset, we obtained a replication data set of 169 healthy individuals consisting of RNA-seq data from sorted naïve $CD4^+$ T cells (**Table1**)[2]. This dataset offered several advantages: 1) a pure $CD4^+$ T cell population, 2) the exclusion of antigen-experienced memory T cells, 3) homogenous continental ancestry (European) and 4) genome-wide genotyping. Therefore, in this analysis we were able to strictly control for the effects of population stratification using principal

components of genome-wide genotype data. For this dataset, we inferred TCRs from bulk RNA-seq, and hence there were fewer CDR3 sequences per individual than in the discovery dataset (around 0.7% of those observed in the discovery dataset; **Table 1**). The lower number of observations reduced power, particularly in our ability to study low-frequency CDR3 amino acids. We applied the same analysis to the replication dataset, testing only *HLA* class II genes for these $CD4^+$ data. From 24,360 tests in the discovery dataset, 11,620 tests localized to *HLA* class II genes, and among these the majority (n= 9,735) were frequent enough to test for association (see **Methods**). We observed that the variance explained by *each* HLA site was similar in the replication dataset and the discovery dataset (Pearson's r = 0.65). We again observed the strongest association between *HLA-DRB1* site 13 and L13-CDR3 position 109 (largest variance explained, **Extended Data Figure 5**).

Replication analysis using a linear regression model

In our replication data, we sought to test the replicability of the strongest *HLA* association for each of the 388 significant CDR3 phenotypes (length-position-amino acid combinations, $P < 0.05/1,249,742$ total tests). Out of these 388 phenotypes, a total of 375 phenotypes were testable in the replication dataset (some *HLA* alleles and CDR3 phenotypes were missing due to low frequencies in the replication dataset). Because the replication dataset consisted of $CD4^+$ TCRs, we tested the 369 of these 388 CDR3 phenotypes whose lead associations localized to class II *HLA* genes. The effect sizes in the discovery and replication datasets were significantly correlated (r = 0.76; $P = 5.4 \times 10^{-70}$; **Extended Data Figure 5; Supplementary Table 7**); 314 of 369 phenotypes replicated in the same allelic direction (sign test *P* value = 3.5×10^{-45}). When we restricted this analysis to the 85 phenotypes for which we found nominally significant associations in the replication dataset ($P < 0.05$), 84 of them replicated in the same allelic direction (sign test *P* value = 4.4×10^{-24}). Thus, we suspect that the majority of replication failures are due to insufficient statistical power from the fewer number of observations in the replication dataset.

cdr3-QTL signals for TCR alpha chains

Using the replication dataset, we also tested cdr3-QTL signals for TCR alpha chains. Since the sequencing depth was very shallow, this analysis was underpowered. However, we did find some cdr3-QTL signals; intriguingly, the HLA site that explained the most variance in CDR3 amino acid compositions was again HLA-DRB1 site 13 (**Extended Data Figure 5**).

Strategy of handling CDR3 length in this study

Our primary strategy in defining CDR3 phenotypes was to stratify by CDR3 length, such that every length-position pair is evaluated separately (the length-position model). CDR3 phenotypes at the same position are correlated across CDR3 lengths, however, and the alternative strategy of aligning CDR3s of different lengths produced generally stronger associations with the same HLA alleles (the position model; **Supplementary Figure 19a**). However, the length-position model detected several length-specific associations (**Supplementary Figure 19b-e**) and revealed that cdr3-QTLs were generally stronger in shorter CDR3 lengths (**Extended Data Figure 6b**, **Supplementary Figure 6**, and **Supplementary Figure 19f**). Thus, we found the length-position model to be more comprehensive.

Thymic selection may drive HLA-CDR3 associations

Since we observed consistent cdr3-QTL signals in PBMCs (including naïve and memory T cells) and sorted naïve T cells, we hypothesized that cdr3-QTL effects might be driven by thymic selection. Alternative possibilities included that cdr3-QTLs were driven by genetic mechanisms prior to thymic selection (phase 1 in **Figure 1a**), or by antigen presentation in the periphery by *HLA* alleles (phase 4 in **Figure 1a**).

To investigate the possibility of a genetic mechanism prior to thymic selection, we analyzed non-productive CDR3 sequences. Although they are generated by the same random recombination process as productive CDR3s, they are not expressed on T cell surfaces and thus are not subjected to thymic selection. If thymic selection is driving cdr3-QTLs, we should not observe HLA-CDR3 associations in non-productive sequences. Indeed, when we tested individual *HLA* sites to assess if they explained variance in CDR3 amino acid frequencies at each position (MMLM analysis), we observed no significant

signals in non-productive sequences (minimum $P = 3.8 \times 10^{-5} > 0.05/24{,}360$ total tests; **Figure 4a**). Since non-productive sequences were only 17.9% of all unique CDR3 sequences, we considered the possibility that the lack of signal was due to reduced power. Thus, we down-sampled the productive sequences to match the number of non-productive sequences and repeated the analysis. We observed that productive CDR3s still had substantial evidence of cdr3-QTL (**Figure 4a**). Consistent with these findings, effect size directions from the non-productive CDR3 analysis were randomly distributed rather than concordant with those from the productive CDR3 analysis (**Figure 4b**).

If peripheral antigen presentation by MHC and memory formation drives the observed cdr3-QTL effects, then T cells with CDR3 favored by specific *HLA* alleles should be expanded due to proliferation. Weighting each unique CDR3 sequence by its expansion level should then augment cdr3-QTL signals, relative to our primary analysis in which we treated each unique sequence equally. To test this possibility, we reanalyzed the discovery data, weighting each unique CDR3 sequence by its read count to emphasize clonally expanded cells. We still observed evidence of cdr3-QTL effects but with a substantially lower magnitude (**Figure 4c-d**). Consistent with the strong replication in naïve T cells, these results suggest that our cdr3-QTL results reflect thymic selection favoring individual CDR3 sequences, and that these signals are mitigated (rather than augmented) by peripheral clonal expansion.

The influence of correlations between *HLA* alleles in our analysis

Although we kept all the correlated *HLA* alleles, our strategy using a Bonferroni corrected *P* value cutoff stringently controlled type I error rate, which was confirmed by extensive permutation analyses (**Extended Data Figure 3** and **Supplementary Figure 10**). Due to correlated genotypes, some TCR phenotypes were associated with multiple *HLA* alleles. This is a common characteristic in QTL analyses: multiple alleles are usually associated with a given gene[3–5]. To mitigate potential problems resulting from this issue, we only used the most significantly associated allele for each TCR phenotype in many parts of our manuscript (e.g., **Figure 4** and **Supplementary Figure 1-2**).

Explained variance in a five-fold cross validation

To confirm that there was no overfitting in our analysis, we performed five-fold cross validation. We used the *HLA* site and CDR3 position pair that showed the strongest association in the MMLM analysis (variance explained = 0.093): alleles at *HLA-DRB1* site 13 were explanatory variables and CDR3 amino acid frequencies at position 109 of L13-CDR3 were response variables. In each round of cross validation, we conducted a linear regression for each CDR3 amino acid using training samples (80% of all samples) to prepare a predictive model. Then, we applied this model to the validation samples (the remaining 20%) and compared the predicted and observed frequencies of the target amino acid. Mean r^2 in each round of cross validation was comparable to the expected value estimated from the MMLM analysis (explained variance = 0.093; **Supplementary Figure 20**). Thus, we confirmed that the strong association at *HLA-DRB1* site 13 was not due to overfitting.

Public clonotypes and cdr3-QTL signals

Previous studies have reported *HLA*-allele-associated public clonotypes (a pair of V gene and CDR3 observed in multiple individuals), which represent a small fraction of the entire repertoire. Their associations were enriched in the clonally expanded and pathogen-experienced T cell populations[6,7]. In contrast, our study analyzed the entire repertoire, and detected robust associations in naïve T cells that were attenuated by the inclusion of clonally expanded TCRs. To confirm that our cdr3-QTL signals are independent from public clonotypes, we excluded all public clonotypes and repeated our same analyses. We observed almost identical results, suggesting that the signal is not driven by the public repertoire (**Supplementary Figure 21**). Evidently, our approach to detecting *HLA*-TCR associations has captured novel biology in non-public clonotypes.

Identifying cdr3-QTL loci outside of the *MHC* regions

The moderate correlation between V/J gene usage and CDR3 amino acids raises the possibility that cis-regulatory variants of V/J genes within the *TCR* locus indirectly affect CDR3 amino acid composition. In the replication dataset for which genome-wide genotype data was available, we searched for the cis-regulatory variants of V/J genes and CDR3 amino acid compositions of beta chains among the 940

variants in the *TCR* locus (Chr. 7:141,998,851-142,510,972 in the GRCh37 genomic coordinates). Among 48 V and 13 J genes we detected, we observed significant associations for 22 V genes and nine J genes ($P < 8.7 \times 10^{-7} = 0.05 / (940 \times (48 + 13))$); **Supplementary Table 2** and **Supplementary Figure 5**). Among 915 CDR3 phenotypes (length-position-amino acid combinations; potentially 7 x 10 x 20 = 1,400 phenotypes but we could not detect rare amino acid in the replication dataset), we observed significant associations only for five phenotypes ($P < 5.8 \times 10^{-8} = 0.05 / (940 \times 915)$; **Supplementary Table 3** and **Supplementary Figure 5**). Accounting for nine J genes with significant cis-regulatory variants completely obviated associations with CDR3 amino acid composition (**Supplementary Figure 5**). Of note, we might have failed to detect some variants within the *TCR* locus and thereby underestimated these cis-regulatory effects. In summary, these results suggest that cis-regulatory effects for CDR3 amino acid compositions are mainly driven by their correlation with J genes.

CDR3 risk score

Our CDR3 risk score is analogous to the well-known polygenic risk score (PRS). For a given CDR3 sequence, the CDR3 risk score is the sum of *HLA* risk score effect sizes (analysis summarized in **Figure 6b**) for which the target amino acid exists in the CDR3 sequence (**Supplementary Figure 17**). The *P* value threshold for including effect sizes is flexible and should be optimized for downstream analysis. We defined the performance of the CDR3 risk score by the correlation between the *HLA* risk score and the average CDR3 risk score for each individual. Using five-fold cross validation in the discovery dataset, we tested nine different *P* value thresholds; 0.05, 0.01, 0.001, 1×10^{-4}, 3.6×10^{-5} (= 0.05/1,354 total tests, Bonferroni corrected *P* value), 1×10^{-5}, 1×10^{-7}, 1×10^{-8}, and 1×10^{-10}. We decided to use the *P* value threshold of 3.6×10^{-5} since this threshold had the best performance in the cross validation (**Supplementary Figure 17**).

Embedding of the 3-D TCR structure

In order to embed points into a two-dimensional space, we obtained protein structure data on TCR, HLA-DR, and antigen structures (1J8H, 1YMM, 2IAM, 2IAN, and 4E41). Using the centroids of each amino

acid residue, we calculated distances, $d_{i,j}$, between amino acid residues *i* and *j*, between HLA-DRB1 residues and antigenic peptide residues, between antigenic peptide residues and CDR3 residues, and between HLA-DRB1 residues and CDR3 residues. We examined only polymorphic HLA residues. We averaged distances across all five structures.

To focus on nearby structures, we excluded residue centroids with an averaged pairwise distance >20 angstroms for all measured distances. For the remaining points, we sought to embed the centroids into a two-dimensional plot by assigning x and y coordinates that minimized the difference between distances in embedded space and distances in three-dimensional structural space. For each pairwise distance, we used a weight w_{ij} to emphasize certain distances, and de-emphasize others. We sought to minimize the following objective function:

$$F(\vec{x},\vec{y}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{i,j}\left(\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} - d_{i,j}\right)^2$$

To emphasize short distances (d<6 angstroms) and de-emphasize longer distances, we defined w_{ij} as follows:

$$w_{i,j} = f(x) = \begin{cases} 5, & if\ d_{ij} < 6, one\ of\ i,j\ is\ from\ antigen \\ \max\left(0.1, (4.1667) * \left(1 - \frac{d_{ij}}{15}\right)\right), & if\ 6 \le d_{ij} < 20, one\ of\ i,j\ is\ from\ antigen \\ 2.5, & if\ d_{ij} < 6, i,j\ are\ from\ TCR\ and\ HLA-DR \\ \max\left(0.05, (2.08335) * \left(1 - \frac{d_{ij}}{15}\right)\right), & if\ 6 \le d_{ij} < 20, i,j\ are\ from\ TCR\ and\ HLA-DR \end{cases}$$

To keep residues from collapsing in on themselves, we made certain arbitrary assignments. If *i,j* were from two residues of the same molecule, or were from two residues >20 angstroms apart, we assigned $d_{i,j}$ to be 50 angstroms, and w(i,j) to be 0.01.

To identify the best fit, we used random start positions and optimized the fit of the points by using both gradient descent and Newton's method. For initialization, each point was randomly assigned to twenty points. Newton's method needs more computational time per iteration, since the Jacobian of F needs to be calculated. Therefore, we applied 90 iterations of gradient descent first:

$$\begin{bmatrix}\vec{x}\\ \vec{y}\end{bmatrix}_{i+1} = \begin{bmatrix}\vec{x}\\ \vec{y}\end{bmatrix}_{i} - s\nabla F = \begin{bmatrix}\vec{x}\\ \vec{y}\end{bmatrix}_{i} - s\begin{bmatrix}\frac{\partial F(\vec{x},\vec{y})}{\partial x_1}\\ \vdots\\ \frac{\partial F(\vec{x},\vec{y})}{\partial x_n}\\ \frac{\partial F(\vec{x},\vec{y})}{\partial y_1}\\ \vdots\\ \frac{\partial F(\vec{x},\vec{y})}{\partial y_n}\end{bmatrix}$$

Here ∇F is the gradient of F. For s, we tried values ranging from 2^(1-h), where h ranged from 1 to 21, and hence s ranged from 1 to 9e-7. We selected the value of h that resulted in the lowest value of F. Next, we applied 20 iterations of Newton's method after gradient descent to quickly find local minima:

$$\begin{bmatrix}\vec{x}\\ \vec{y}\end{bmatrix}_{i+1} = \begin{bmatrix}\vec{x}\\ \vec{y}\end{bmatrix}_{i+1} - s\cdot J(F)^{-1}\cdot\nabla F = \begin{bmatrix}\vec{x}\\ \vec{y}\end{bmatrix}_{i+1} - s\cdot\begin{bmatrix}\frac{\partial^2 F(\vec{x},\vec{y})}{\partial^2 x_1} & \cdots & \frac{\partial^2 F(\vec{x},\vec{y})}{\partial x_1 \partial y_n}\\ \vdots & & \vdots\\ \frac{\partial^2 F(\vec{x},\vec{y})}{\partial x_n \partial y_1} & \cdots & \frac{\partial^2 F(\vec{x},\vec{y})}{\partial^2 y_n}\end{bmatrix}^{-1}\cdot\begin{bmatrix}\frac{\partial F(\vec{x},\vec{y})}{\partial x_1}\\ \vdots\\ \frac{\partial F(\vec{x},\vec{y})}{\partial x_n}\\ \frac{\partial F(\vec{x},\vec{y})}{\partial y_1}\\ \vdots\\ \frac{\partial F(\vec{x},\vec{y})}{\partial y_n}\end{bmatrix}$$

Here J(F) is the Jacobian of F. For s, we tried values ranging from 2^(1-h), where h ranged from 1 to 21. We selected the value of h that resulted in the lowest value of F. To calculate the gradient, we calculate the derivative of the objective function at x and y for each point:

$$\frac{\partial F(\vec{x},\vec{y})}{\partial x_k} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 2\left(\delta_{i=k}+\delta_{j=k}\right)w_{i,j}\frac{\left(\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2}-d_{i,j}\right)}{\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2}}\left(x_k-\delta_{i\neq k}x_i-\delta_{j\neq k}x_j\right)$$

$$\frac{\partial F(\vec{x},\vec{y})}{\partial y_k} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 2\left(\delta_{i=k}+\delta_{j=k}\right)w_{i,j}\frac{\left(\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2}-d_{i,j}\right)}{\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2}}\left(y_k-\delta_{i\neq k}y_i-\delta_{j\neq k}y_j\right)$$

Here $\delta_{i=j}$ is the Dirac delta function which is 1 if i and j are equal to each other, and 0 otherwise. Similarly, $\delta_{i\neq k}$ is 1 if i and k are equal to each other, and 0 otherwise. To calculate the Jacobian, we calculate each of the second derivatives empirically setting delta = 0.0000001:

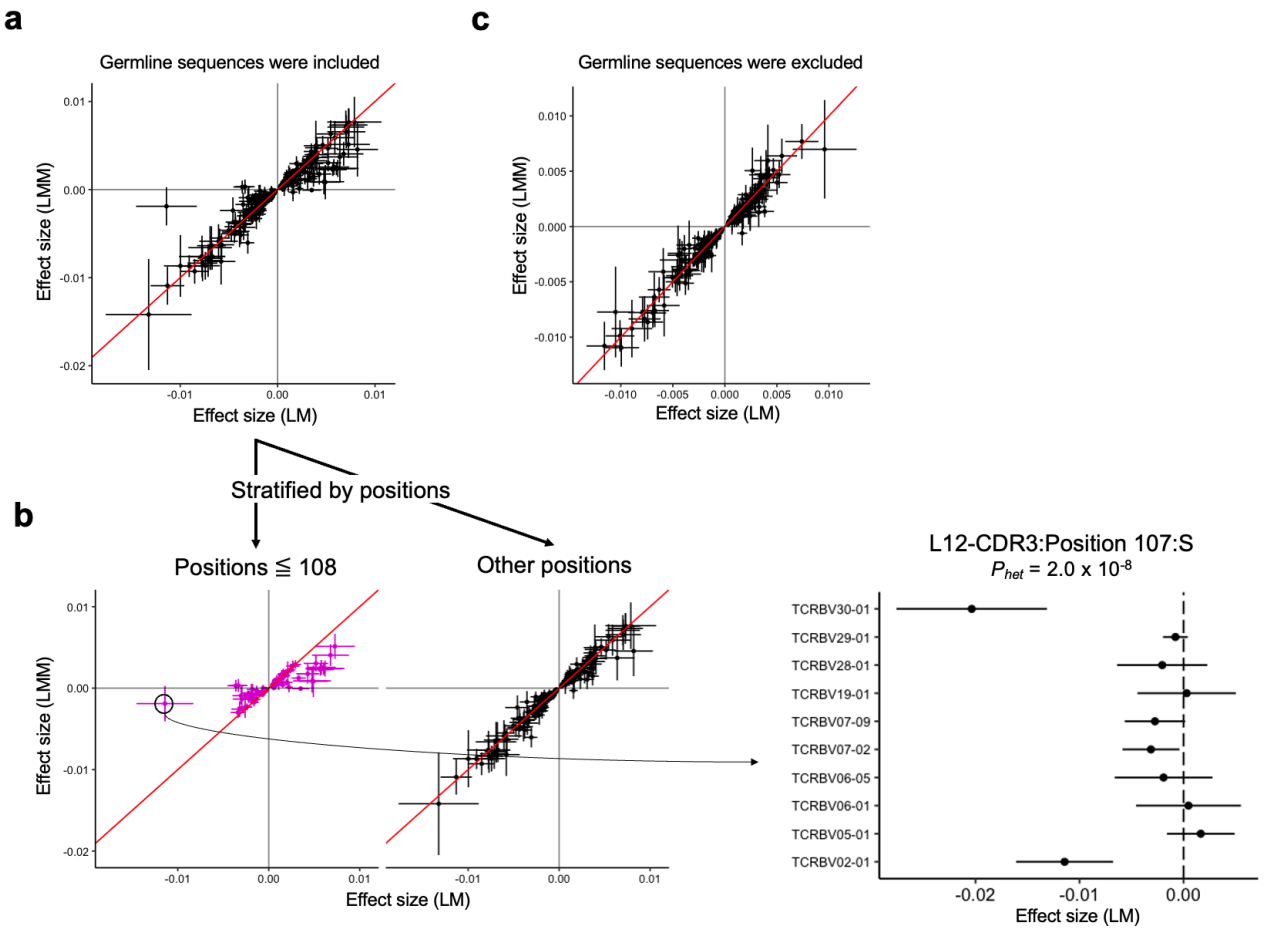
$$\frac{\partial^2 F(\vec{x},\vec{y})}{\partial y_l \partial y_k} = \frac{\frac{\partial F(\vec{x},\vec{y})}{\partial y_k}\Big|_{yl+delta} + \frac{\partial F(\vec{x},\vec{y})}{\partial y_k}\Big|_{yl-delta}}{2 * delta}$$

We calculate other partial derivatives for all pairs *i* and *j* for both x and y coordinates similarly.

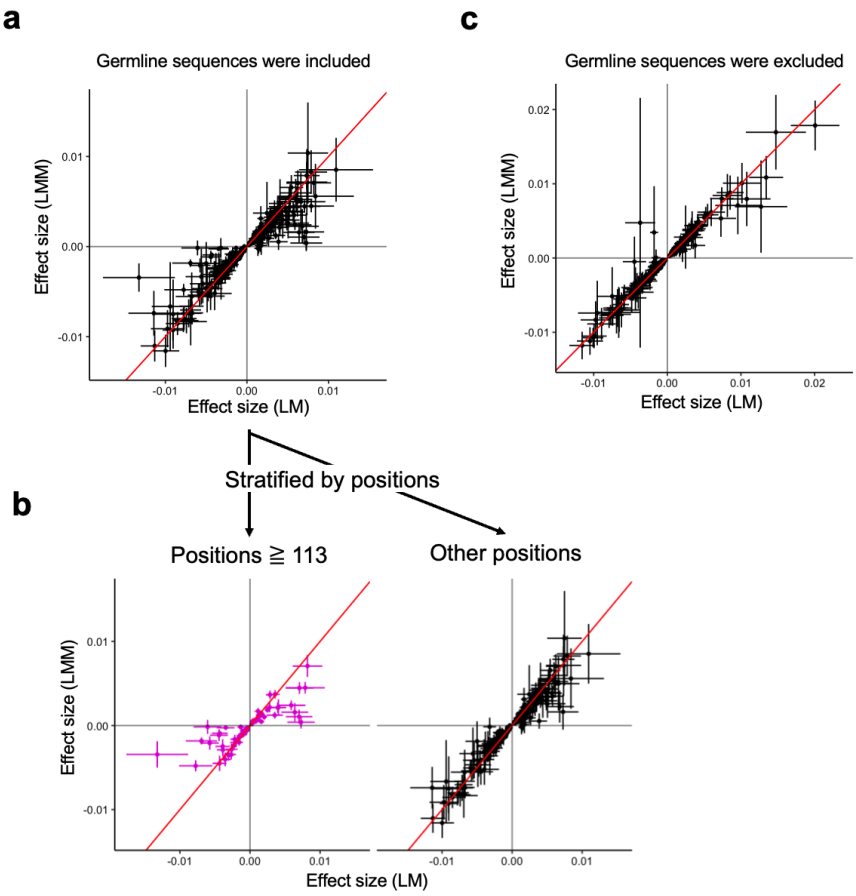
References:

1. Sharon, E. *et al.* Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.* **48**, 995–1002 (2016).
2. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
3. Ishigaki, K. *et al.* Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nat. Genet.* **49**, 1120–1125 (2017).
4. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
5. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715.e16 (2018).
6. 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).
7. DeWitt, W. S. *et al.* Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife* **7**, 1–39 (2018).

Supplementary Figures



Supplementary Figure 1. The effect of germline-encoded V gene sequences on cdr3-QTL analysis.
(a,b) Germline-encoded sequences were included in this analysis. We compared the effect sizes from the linear regression model (LM) that did not adjust for V gene usage to the effect sizes from the linear mixed regression model (LMM) that did model the effect of the V gene. The analysis was restricted to the 435 CDR3 phenotypes (length-position-amino acid combinations) that had at least one significant association in LM analysis ($P < 0.05/1,262,664$ total tests). For each CDR3 phenotype, we used the *HLA* amino acid allele that had the lowest *P* value for that phenotype in the LM analysis. We used *P* values from two-sided linear regression test. The error bar indicates +/- 2 x S.E.
(a) Effect sizes for non-transformed phenotype are provided.
(b) Left panel: the same plot as in **(a)** but stratified by CDR3 position. Right panel: LM association test results for serine (S) at position 107 of length 12 CDR3, stratified by V gene usage for the ten most frequent V genes.
(c) Germline-encoded sequences were excluded in this analysis (the primary analysis in the manuscript). The analysis was restricted to the 388 CDR3 phenotypes (length-position-amino acid combinations) that had at least one significant association in LM analysis ($P < 0.05/1,249,742$ total tests). For each CDR3 phenotype, we used the *HLA* amino acid allele that had the lowest *P* value for that phenotype in the LM analysis. Effect sizes for non-transformed phenotypes are provided. The error bar indicates +/- 2 x S.E. We used *P* values from two-sided linear regression test.



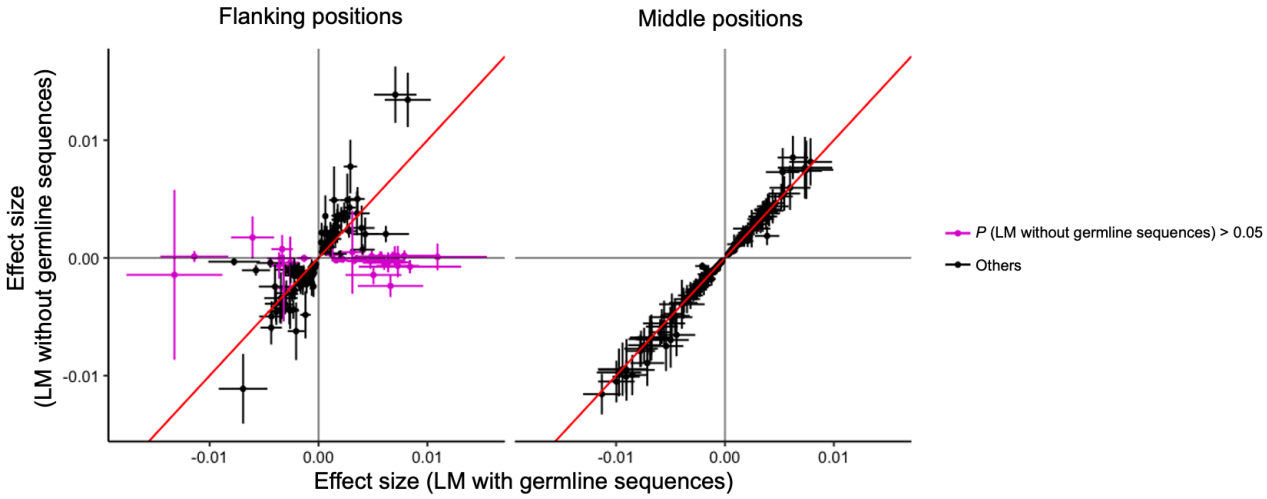
Supplementary Figure 2. The effect of germline-encoded J gene sequences on cdr3-QTL analysis.
(a,b) Germline-encoded sequences were included in this analysis. We compared the effect sizes from the linear regression model (LM) that did not adjust for J gene usage to the effect sizes from the linear mixed regression model (LMM) that did model the effect of the J gene. The analysis was restricted to the 435 CDR3 phenotypes (length-position-amino acid combinations) that had at least one significant association in LM analysis ($P < 0.05/1,262,664$ total tests). For each CDR3 phenotype, we used the *HLA* amino acid allele that had the lowest *P* value for that phenotype in the LM analysis. We used *P* values from two-sided linear regression test. The error bar indicates +/- 2 x S.E.
(a) Effect sizes for non-transformed phenotype are provided.
(b) The same plot as in **(a)** but stratified by CDR3 position.
(c) Germline-encoded sequences were excluded in this analysis (the primary analysis in the manuscript). The analysis was restricted to the 388 CDR3 phenotypes (length-position-amino acid combinations) that had at least one significant association in LM analysis ($P < 0.05/1,249,742$ total tests). For each CDR3 phenotype, we used the *HLA* amino acid allele that had the lowest *P* value for that phenotype in the LM analysis. Effect sizes for non-transformed phenotypes are provided. The error bar indicates +/- 2 x S.E. We used *P* values from two-sided linear regression test.

TRBV2-1: CASSE
TRBV4-1: CASSQ

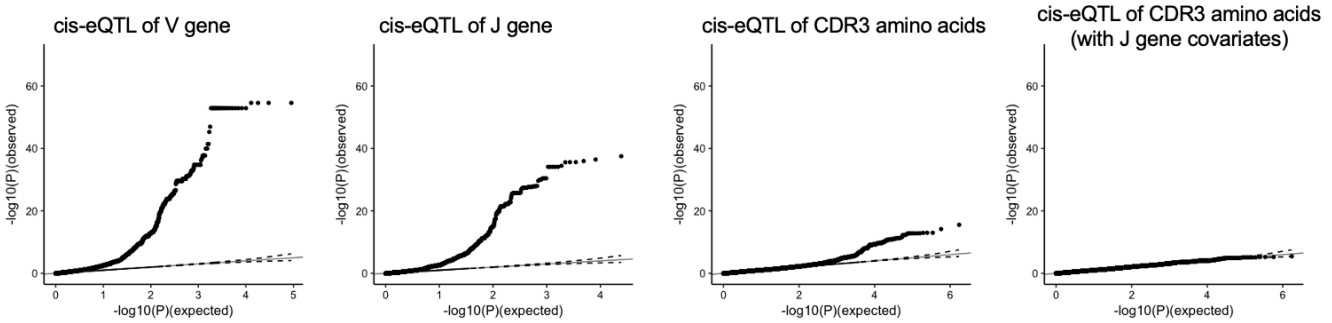
	TCR-CDR3 position						
V gene	104	105	106	107	108	109	...
TRBV2-1	~~C~~	~~A~~	~~S~~	~~S~~	~~E~~	Q	...
TRBV2-1	~~C~~	~~A~~	~~S~~	~~S~~	F	P	...
TRBV2-1	~~C~~	~~A~~	~~S~~	**D**	~~E~~	D	...
TRBV2-1	~~C~~	~~A~~	~~S~~	~~S~~	~~E~~	R	...
TRBV2-1	~~C~~	~~A~~	~~S~~	Q	~~E~~	P	...
TRBV2-1	~~C~~	~~A~~	~~S~~	~~S~~	G	Q	...
TRBV2-1	~~C~~	~~A~~	~~S~~	G	~~E~~	G	...
TRBV2-1	~~C~~	~~A~~	~~S~~	~~S~~	D	E	...
TRBV2-1	~~C~~	~~A~~	~~S~~	**D**	Q	D	...
TRBV4-1	~~C~~	~~A~~	~~S~~	~~S~~	~~Q~~	F	...
TRBV4-1	~~C~~	~~A~~	~~S~~	**D**	~~Q~~	F	...
TRBV4-1	~~C~~	~~A~~	~~S~~	~~S~~	F	Y	...
TRBV4-1	~~C~~	~~A~~	~~S~~	~~S~~	~~Q~~	Q	...
TRBV4-1	~~C~~	~~A~~	~~S~~	G	**E**	P	...
TRBV4-1	~~C~~	~~A~~	~~S~~	~~S~~	~~Q~~	G	...
TRBV4-1	~~C~~	~~A~~	~~S~~	~~S~~	~~Q~~	D	...
TRBV4-1	~~C~~	~~A~~	~~S~~	~~S~~	**E**	Y	...
TRBV4-1	~~C~~	~~A~~	~~S~~	G	**E**	D	...

D: 3/7 E: 3/8

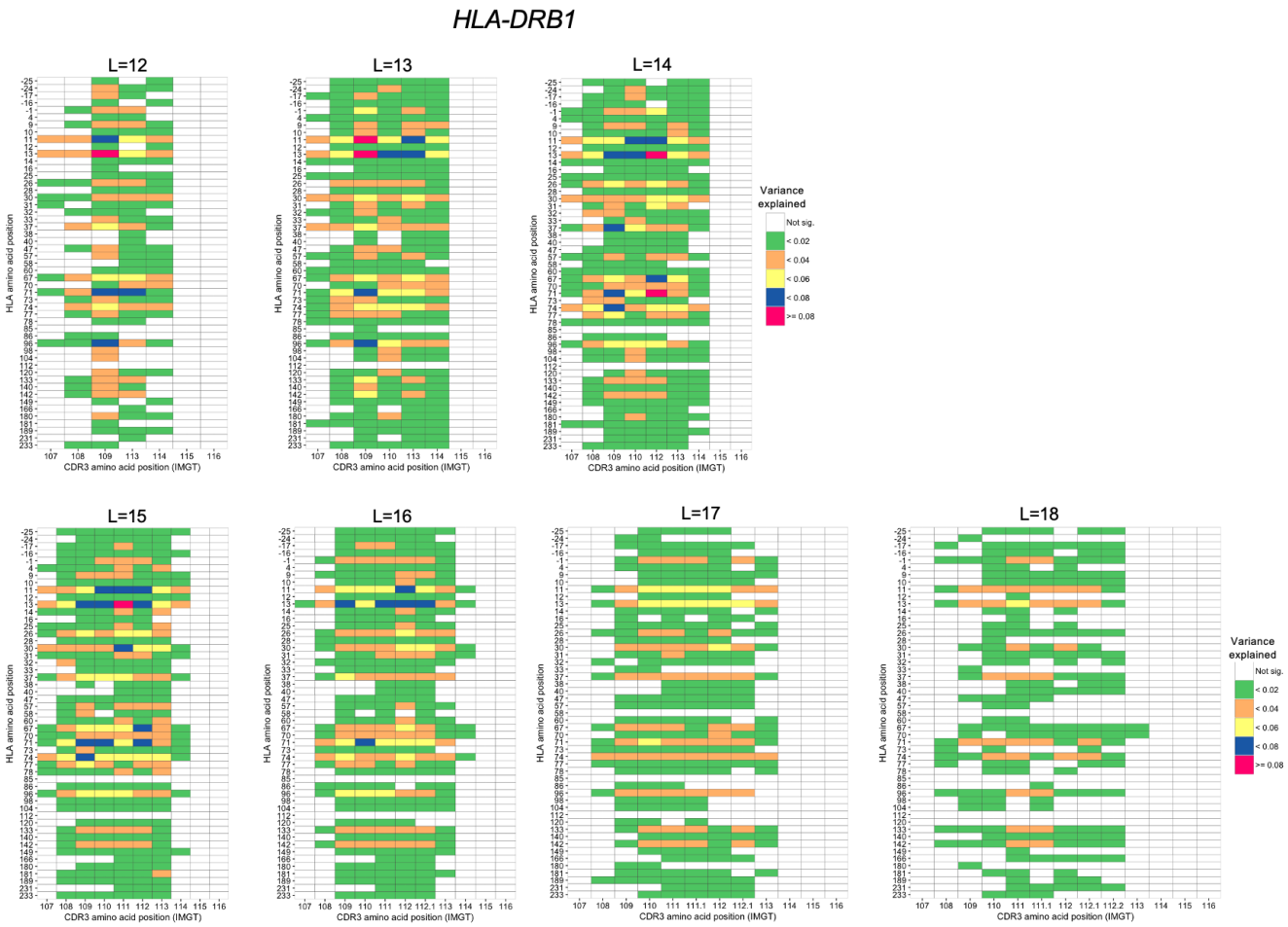
Supplementary Figure 3. The strategy to exclude germline-encoded amino acids from each CDR3 sequence.
This is a schematic figure of TCR data with two V genes (TRBV2-1 and TRBV4-1); their germline-encoded sequences in the CDR3 region (positions 104-108) are provided above the table. When a clonotype had an encoded amino acid at a CDR3 position, we excluded that amino acid from our analysis (erased with a horizontal line). We included non-encoded amino acids (highlighted by red), even when they localized to positions that were germline-encoded in other TCRs. Then, we calculated amino acid frequencies at each position; examples for aspartate (D) at position 107 and glutamate (E) at position 108 are provided.



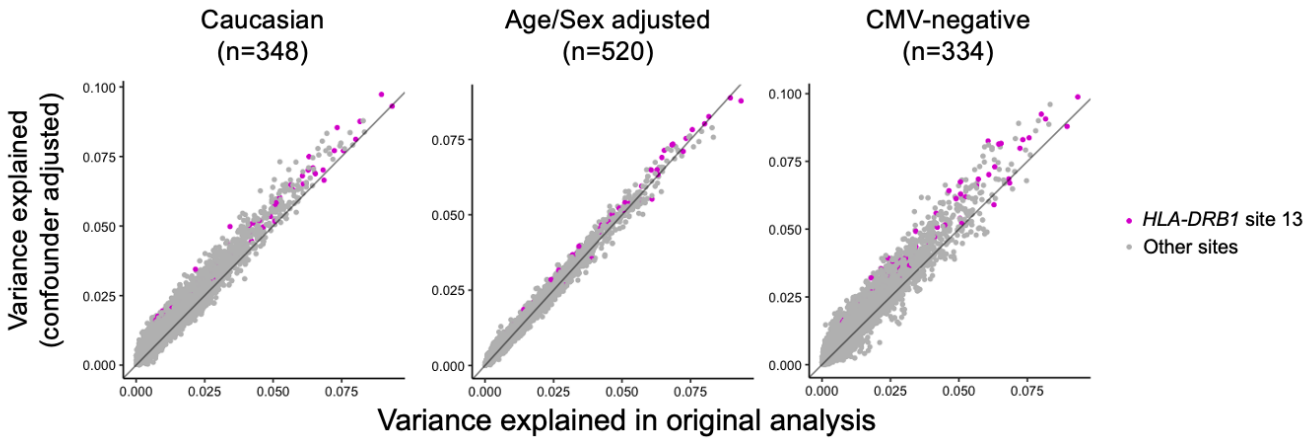
Supplementary Figure 4. The effect of germline-encoded V and J gene sequences on cdr3-QTL analysis.
The effect size estimates from the linear regression model (LM) are compared between two conditions: the analyses including germline-encoded amino acids and those excluding them. The analysis was restricted to the 435 CDR3 phenotypes (length-position-amino acid combinations) which had at least one significant association in LM analysis including germline-encoded amino acids ($P < 0.05/1,262,664$ total tests), and we used the *HLA* amino acid allele that had the lowest *P* value for each phenotype. Effect sizes for non-transformed phenotype are provided. The CDR3 middle positions are positions 109-112; the flanking positions are positions 107, 108, and 113-116. We used *P* values from two-sided linear regression test. The error bar indicates +/- 2 x S.E.



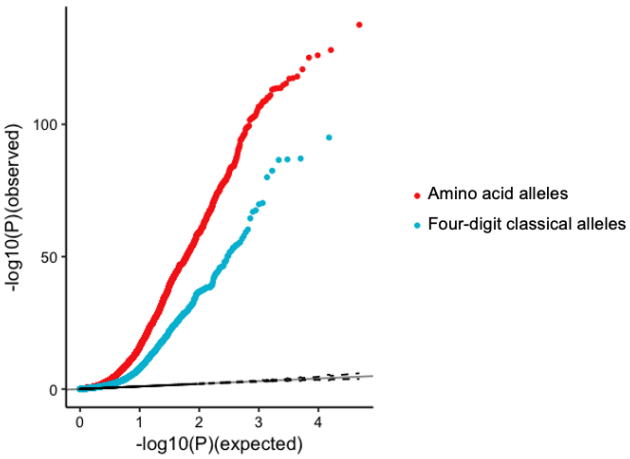
Supplementary Figure 5. Cis-regulatory effects of V/J genes and CDR3 amino acid compositions. Using the replication dataset (n = 169), we tested associations between the allelic dosages of TCR locus variants and V/J usage or CDR3 amino acid composition. We used *P* values from two-sided linear regression test. For CDR3 amino acid composition, we conducted a second analysis that included the nine J genes with significant cis-regulatory effects as covariates.



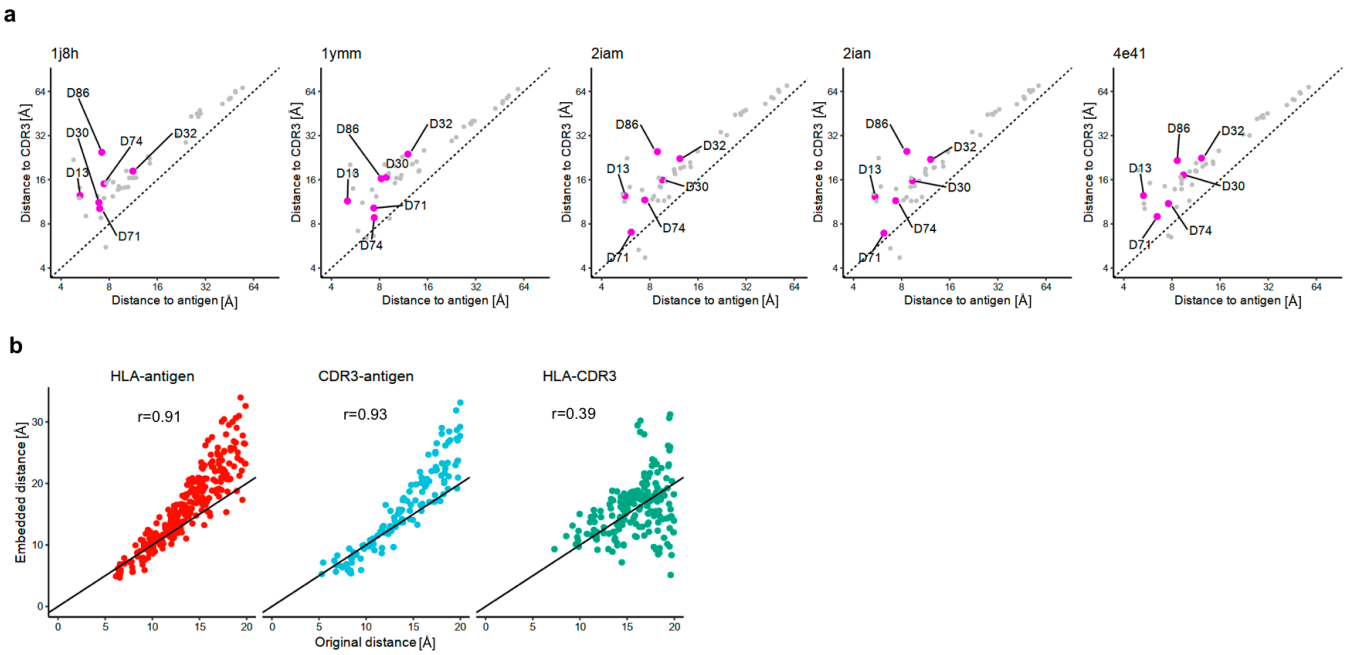
Supplementary Figure 6. Variance explained in cdr3-QTL analysis for each length of CDR3. Variance explained in the MMLM analysis for different lengths of CDR3 (n=628; the discovery dataset). The results for *HLA-DRB1* are provided.



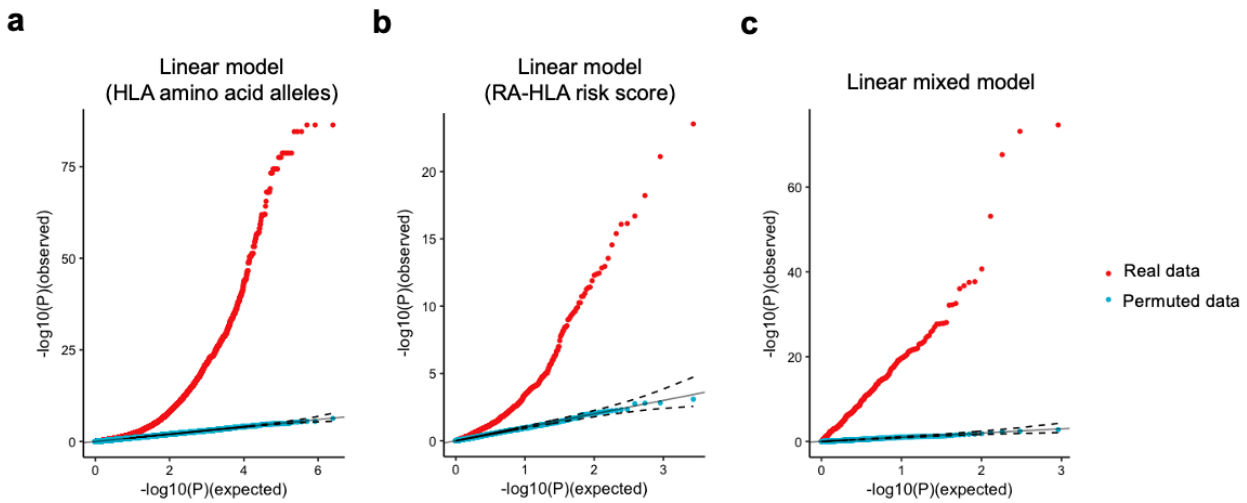
Supplementary Figure 7. Evaluation of the confounders in cdr3-QTL analysis (MMLM analysis). To evaluate the effect of potential confounders in the MMLM analysis, we tested the variance explained in three different conditions. In each plot, the X-axis corresponds to the variance explained in the primary analysis (n=628; the discovery dataset) and the Y-axis corresponds to the variance explained in one of the three conditions. First, we conducted the analysis only using European ancestry samples (n=348) to test potential bias driven by ancestry. Second, we conducted analyses modeling age and sex as covariates (n=520; sample size decreased due to missing data in covariates) to test potential bias driven by age or sex. Third, to test potential bias driven by cytomegalovirus infection status, we restricted the analysis to non-infected samples (n=334). For each condition, we analyzed all tests in the primary analysis (24,360 total tests).



Supplementary Figure 8. cdr3-QTL results using four-digit classical allele genotypes.
QQ plot of the MMLM analysis comparing two different methods (n = 628; the discovery dataset): associations with *HLA* amino acid alleles and those with four-digit classical *HLA* alleles. We used MANOVA test *P* values.

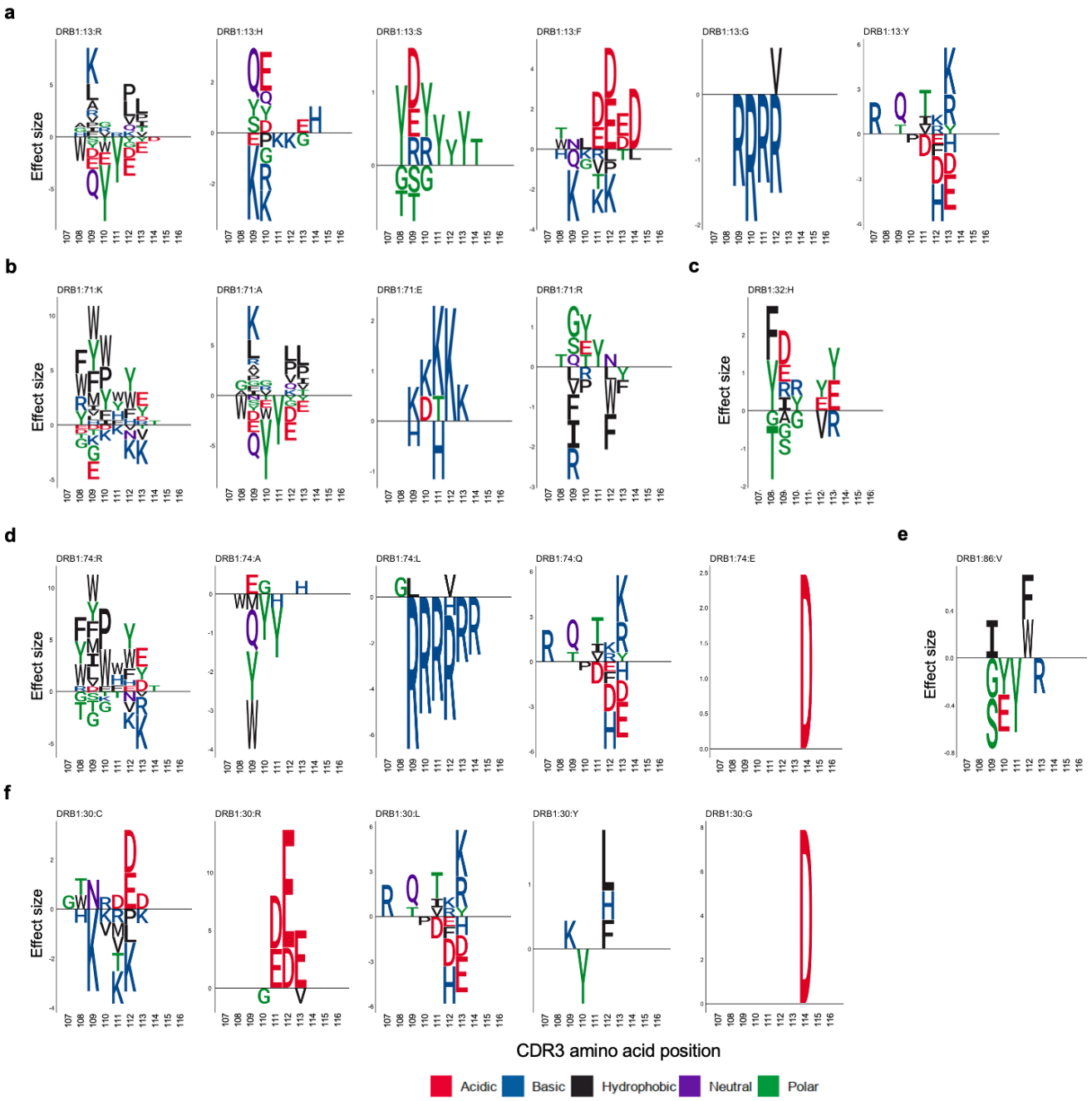


Supplementary Figure 9. Observed and embedded pair-wise distances of amino acids in MHC-peptide-TCR complexes.
(a) The shortest distances between each *HLA-DRB1* site and all positions of the antigen (X-axis) and those between each *HLA-DRB1* site and all positions of the CDR3 (Y-axis) are provided for each protein structure. The sites with independently significant cdr3-QTL effects are highlighted in magenta.
(b) We embedded all pairwise distances in the pMHC-TCR complex into a two-dimensional space, down-weighting the distances between HLA and TCR to highlight antigen-related interaction. Embedding pair-wise distances (Y-axis) compared with those observed in structural data (X-axis). Visualized are averaged values across the five structures. Pearson's correlation coefficients are provided.



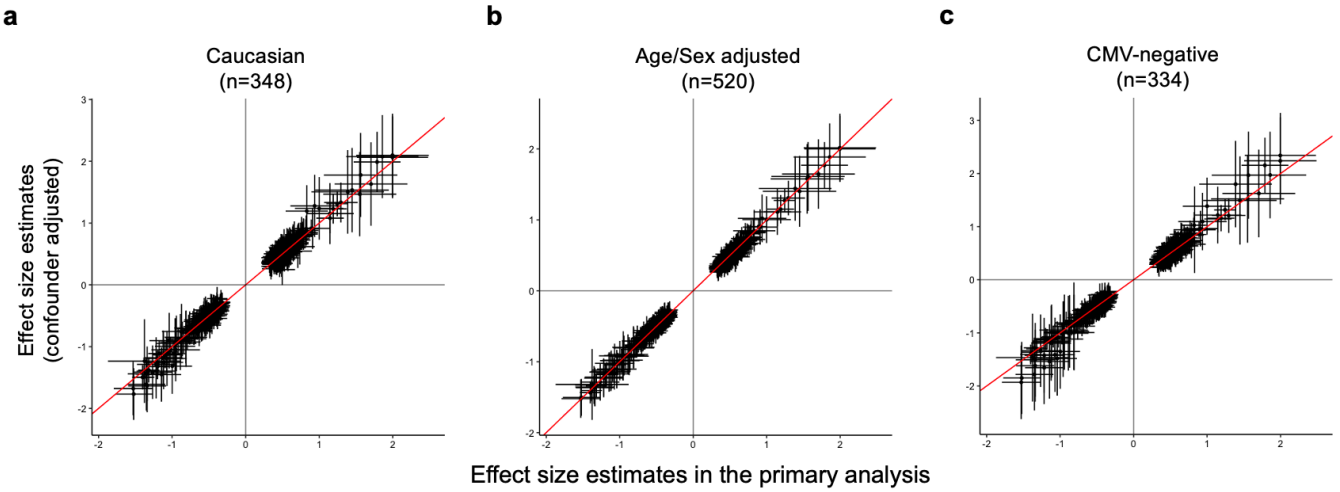
Supplementary Figure 10. Permutation analyses using the linear regression model and the linear mixed model.

QQ plots for the real dataset and the permuted dataset using three different conditions: **(a)** the LM using *HLA* amino acid alleles (1,249,742 total tests), **(b)** the LM using RA-*HLA* risk score (1,354 total tests), and **(c)** the LMM using *HLA* amino acid alleles (388 tests total tests). For the LMM, we restricted our analysis to 388 CDR3 phenotypes (CDR3 length, position, amino acid combinations) that had at least one significant association in the linear regression analysis ($P < 0.05/1{,}249{,}742$ total tests) and used *HLA* amino acid alleles that had the lowest *P* value for each phenotype (**Methods**). For all plots, we used *P* values from two-sided linear regression test.

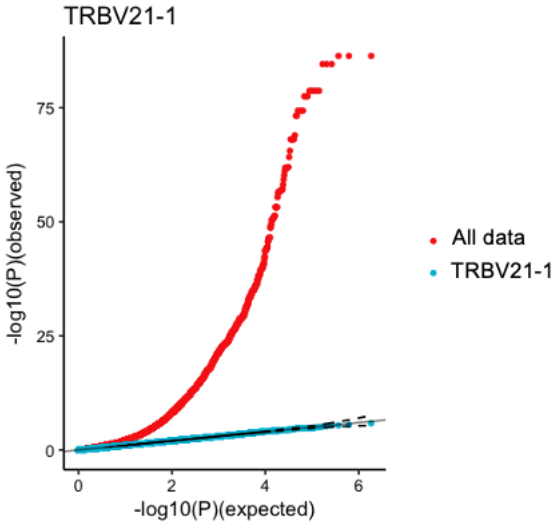


Supplementary Figure 11. CDR3 modification patterns associated with *HLA-DRB1* amino acid alleles.

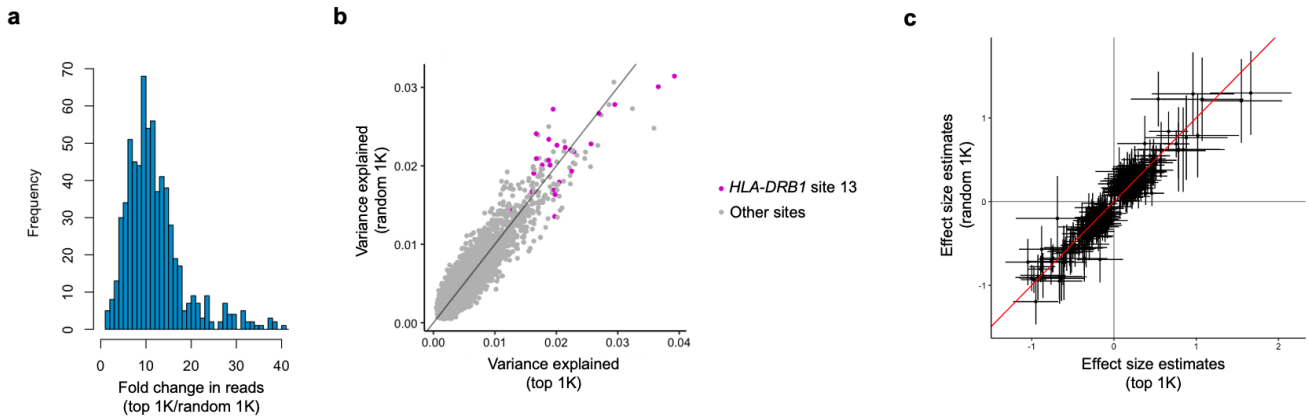
We conducted the LM analysis using the *HLA-DRB1* amino acid alleles at sites 13, 71, 32, 74, 86, and 30 (n=628; the discovery dataset). These six positions within *HLA-DRB1* showed independent associations in the MMLM analysis (**Extended Data Figure 5**). To create a sequence logo for each allele, the effect sizes of significant ($P < 0.05/1{,}249{,}742$ total tests) associations for each amino acid at a given position were summed across L12-L18 CDR3s. We used *P* values from two-sided linear regression test.



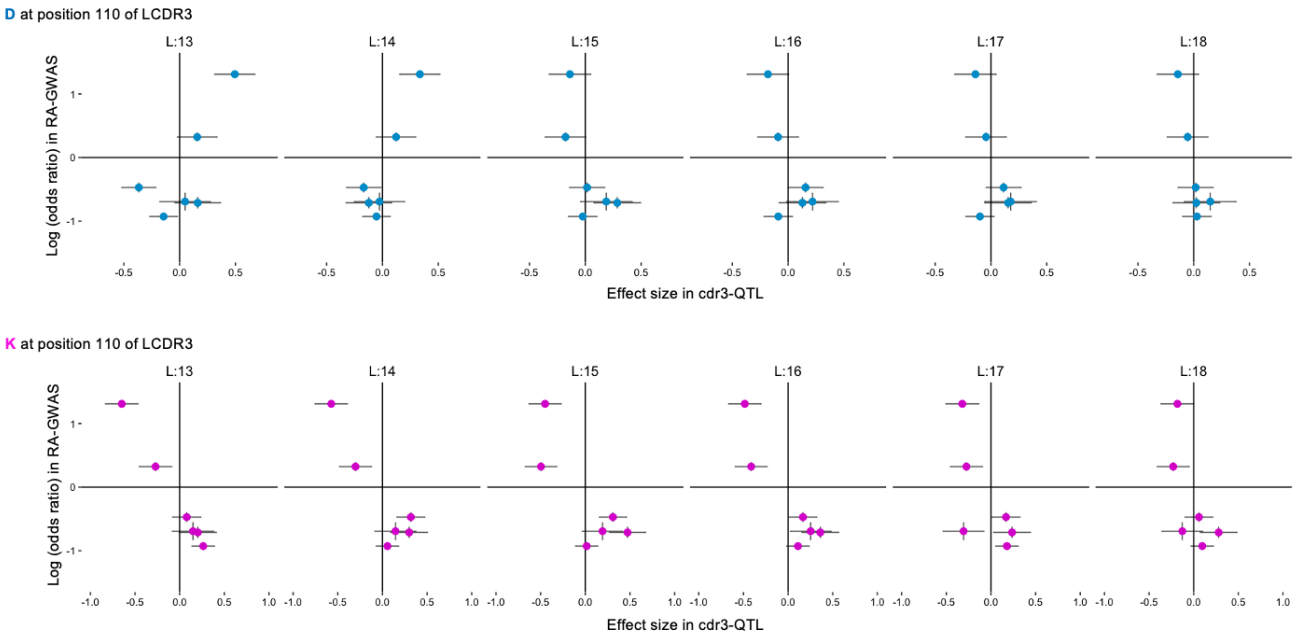
Supplementary Figure 12. Evaluation of the confounders in the LM analysis.
The analysis was restricted to the 388 CDR3 phenotypes (length-position-amino acid combinations) which had at least one significant association in the primary analysis ($P < 0.05/1,249,742$ total tests; the discovery dataset), and we used the *HLA* amino acid allele that had the lowest *P* value for each phenotype. We used *P* values from two-sided linear regression test. The error bar indicates +/- 2 x S.E.
(a) Effect sizes from the LM analysis in which all samples were used (n=628; the primary analysis) compared to those from the analysis restricted to European ancestry samples (n=348).
(b) Effect sizes from the LM analysis that did not adjust for age or sex effects (n=628; the primary analysis) compared to those from the LM analysis that did adjust for age and sex effects (n=520). Including these covariates in the model decreased the sample size due to missing values in covariate data.
(c) Effect sizes from the LM analysis in which all samples were used (n=628; the primary analysis) compared to those from the analysis restricted to donors not infected by cytomegalovirus (n=334).



Supplementary Figure 13. cdr3-QTL analysis using productive CDR3s with a nonfunctional V gene.
QQ plots comparing the *P* value distribution from the LM analysis with all productive sequences (red) to the LM analysis restricted to productive sequences with *TRBV21-1*, a pseudogene that renders the TCR nonfunctional. We used *P* values from two-sided linear regression test.

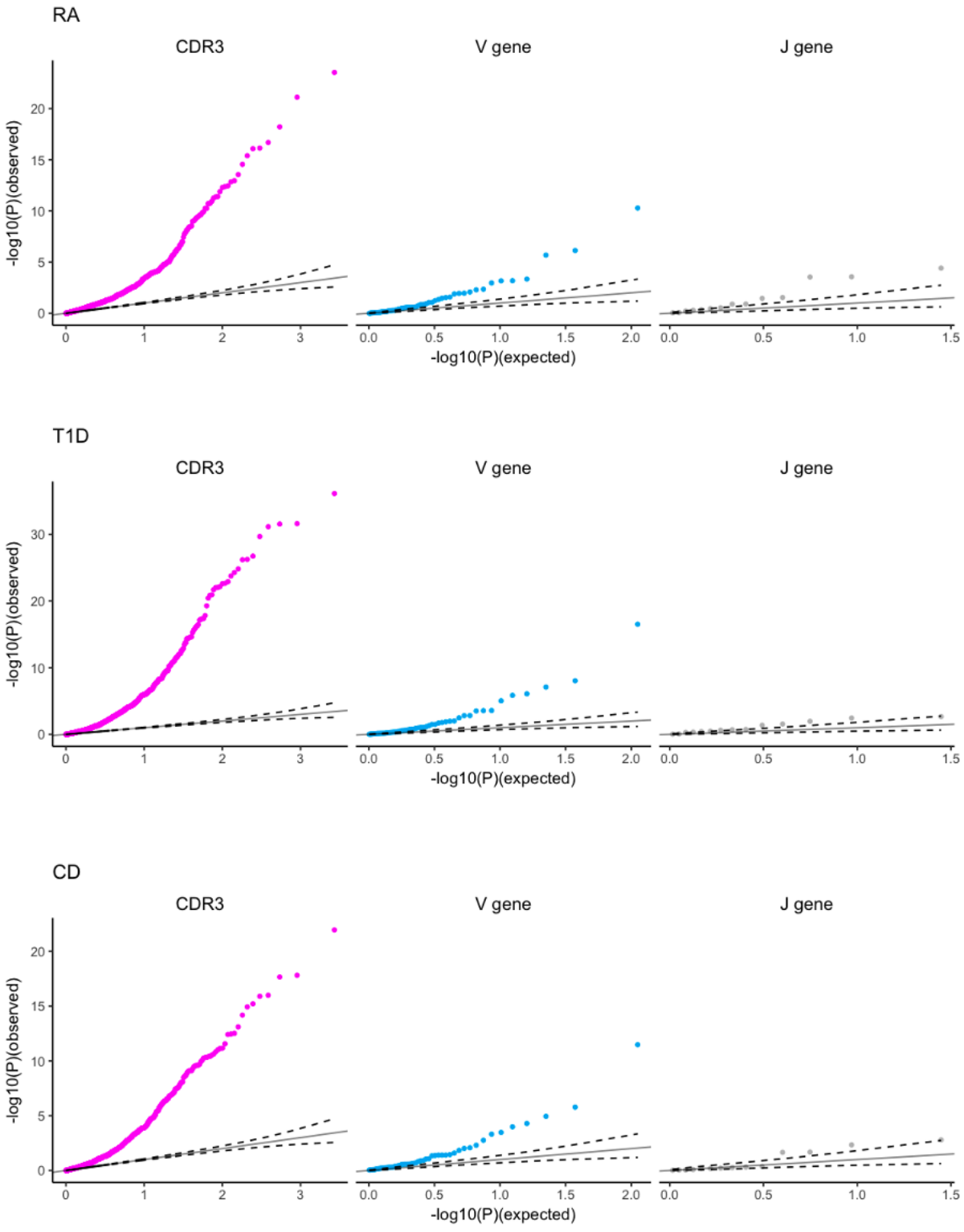


Supplementary Figure 14. cdr3-QTL signals are not enriched in clonally expanded cells.
We used the read count of each clonotype as an index of its clonal expansion. Within each donor, we sorted clonotypes based on read count. We considered the top 1,000 clonotypes to be clonally expanded. We randomly selected 1,000 clonotypes as a control population.
(a) Within each donor, we calculated a ratio of the total read counts between the top and the random 1,000 clonotypes. The ratios of all donors are shown in a histogram. On average, these top 1000 clones had 12-times more sequencing reads compared with the randomly selected TCRs, indicating that the top 1000 clones are substantially expanded and thus appropriate for this analysis.
(b) The variance explained in the MMLM analysis using the top 1,000 (X-axis) and the random 1,000 (X-axis) clonotypes.
(c) Effect size estimates in the LM analysis using the top 1,000 and the random 1,000 clonotypes. We used 388 CDR3 phenotypes (CDR3 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 used the HLA amino acid alleles that had the lowest P value for each phenotype. We used P values from two-sided linear regression test. The error bar indicates +/- 2 x S.E.

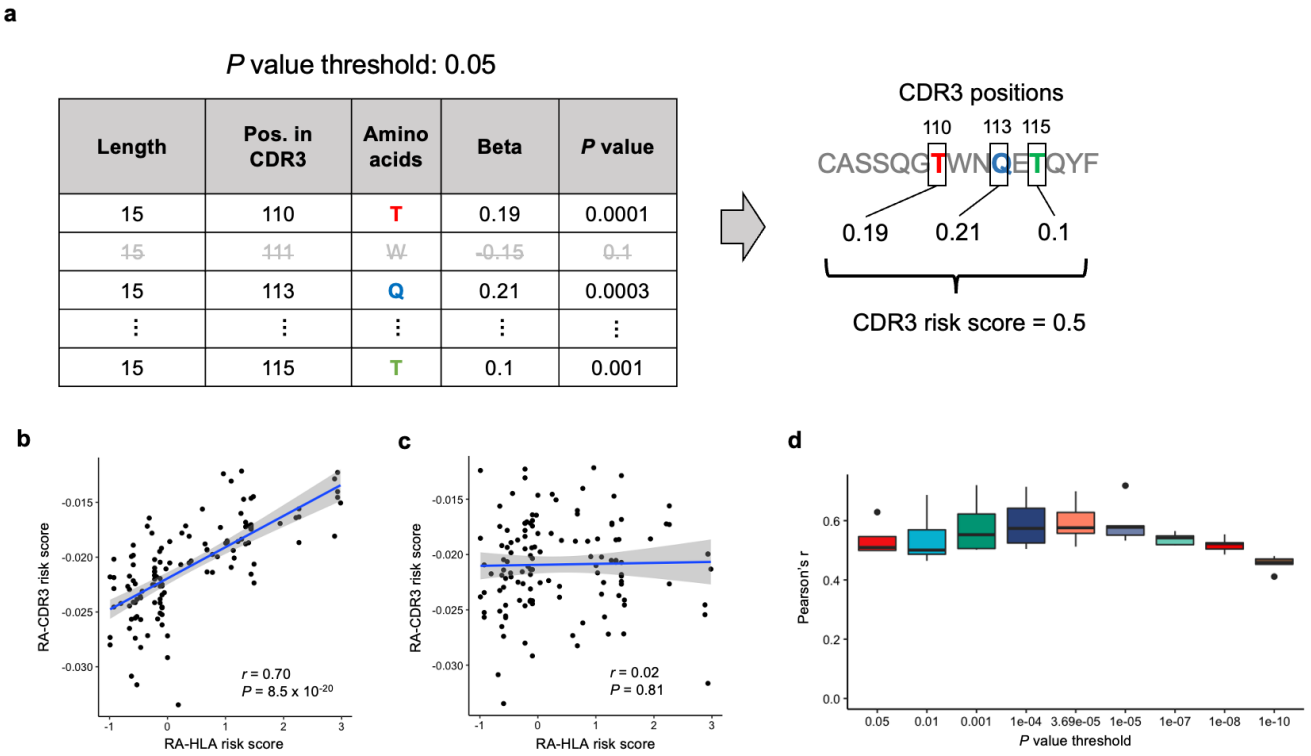


Supplementary Figure 15. The influence of *HLA-DRB1* site 13 amino acid alleles on CDR3 position 110.

We used the linear regression model (LM) with six possible amino acid alleles at *HLA-DRB1* site 13. Their cdr3-QTL effect sizes for aspartic acid (D) and lysine (K) usage at position 110 of CDR3 with different lengths (X-axis; n=628; the discovery dataset) are plotted against their effect size in RA-GWAS (Y-axis). The error bar indicates +/- 2 x S.E. CDR3s of length 12 were excluded from this analysis because position 110 does not exist in L12 CDR3.



Supplementary Figure 16. QQ plot of cdr3-QTL, V/J gene association based on *HLA*-risk scores. We provide QQ plots from the LM analysis with *HLA*-risk scores. The identity line is provided with the 95% confidence interval. We used *P* values from two-sided linear regression test.



Supplementary Figure 17. CDR3 risk score calculation.
(a) Schematic explanation of our strategy to calculate the CDR3 risk score. The table shows effect size estimates from the LM analysis based on *HLA* risk scores. Effect sizes for corresponding amino acid positions are summed to calculate the CDR3 risk score. Only the effects that passed the Bonferroni *P* value threshold were used (0.05/1,354 total tests).
(b, c) We conducted five-fold cross validation in the discovery dataset to evaluate the performance of RA-CDR3 risk score. A representative plot of the correlation between RA-*HLA* risk score and RA-CDR3 risk score in a round of cross validation **b,** using real data. **c,** using permuted data. We used a *P* value threshold of 0.05/1,354 total tests to include effect sizes in CDR3 risk calculation. The error bands indicate 95% confidence interval for predictions from a fitted linear model. Pearson's *r* is provided.
(d) Pearson's correlation coefficient between RA-*HLA* risk score and RA-CDR3 risk score in the five-fold cross validation using nine different *P* value thresholds. Within each boxplot, the horizontal line reflects 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.

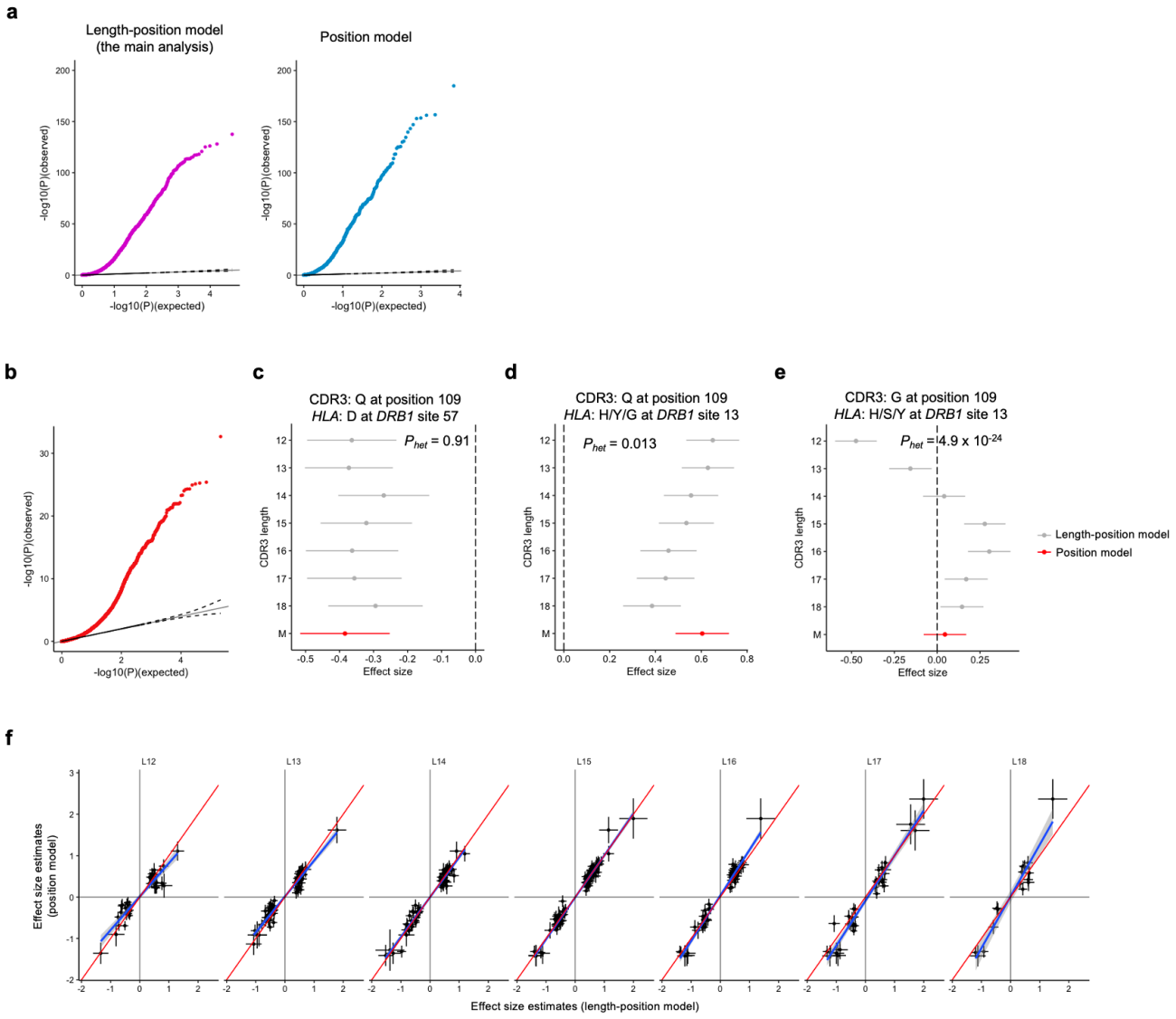
a

Allele group	*HLA-DRB1* site 71 amino acid allele	Four-digit allele
1	K	0301,0401,1303
2	A	1501,1502,1503
3	E	0103,0402,1102,1103,1301,1302
4	R	0101,0102,0403,0404,0405,0407,0408, 0701,0801,0802,0803,0804,0901,1001, 1101,1104,1201,1202,1305,1401,1402, 1404,1407,1454,1601,1602

b

Reference allele	Alternative allele
K	A E R
A	K E R
E	A K R
R	A E K
K A	E R
K E	A R
K R	A E

Supplementary Figure 18. Our strategy to define *HLA* amino acid alleles.
(a) At a given *HLA* site that has *m* possible amino acid residues, we partitioned the four-digit alleles into *m* groups with identical residues at the given site. We then calculated the allele count of each group. This is an example of *HLA-DRB1* site 71 with four possible amino acid residues.
(b) At a given *HLA* site that has *m* amino acid residues, we considered all possible combinations of amino acids and calculated the allele counts for each group. This is an example of *HLA-DRB1* site 71 with four possible amino acids.

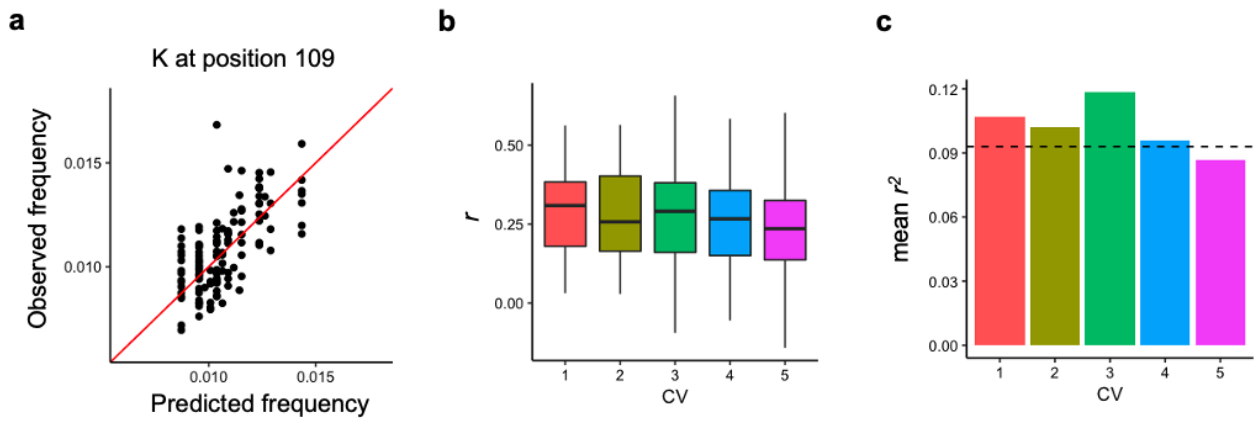


Supplementary Figure 19. CDR3 length affects cdr3-QTL signals.

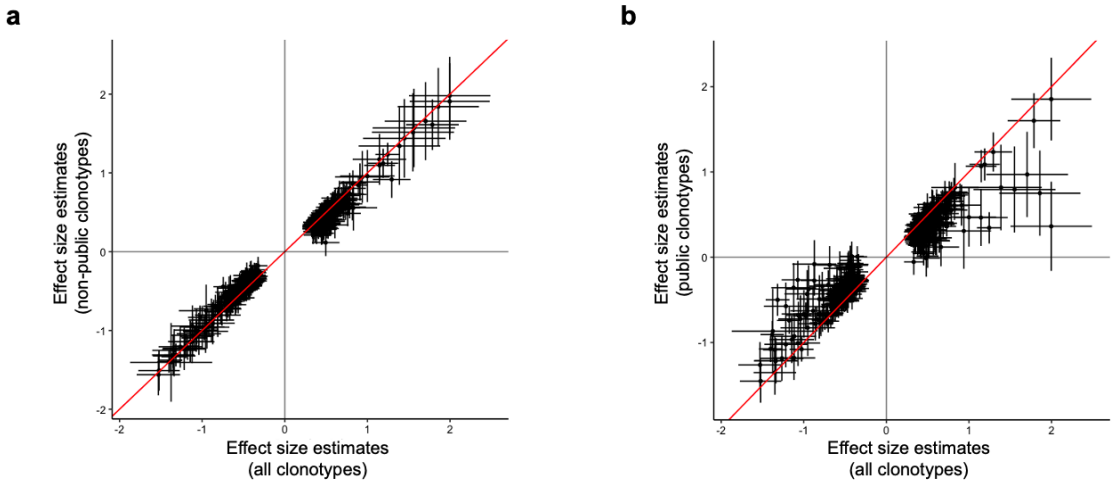
(a) QQ plot in the MMLM analysis showing MANOVA test *P* values from the length-position model (the primary analysis) and the position model (n=628; the discovery dataset).

(b-e) We used 106,145 CDR3 phenotypes (position and amino acid combinations) that were testable in CDR3 lengths 12-18. For each phenotype, we tested for heterogeneity in effect size estimates from the LM analysis across CDR3 lengths using Cochran's Q test (P_{het}). **b,** QQ plot of all P_{het} values. We provide effect size estimates of linear regression tests from three examples: signals with no (**c**), modest (**d**) and strong (**e**) heterogeneity across different CDR3 lengths. The error bar indicates +/- 2 x S.E.

(f) Comparison of effect size estimates from the LM analysis between the length-position model (the primary analysis, X-axis) and the position model (Y-axis). The error bar indicates +/- 2 x S.E. We used the 388 CDR3 phenotypes (CDR3 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 used the *HLA* amino acid alleles that had the lowest *P* value for each phenotype. The blue line indicates a fitted linear regression using the datapoints in each panel (the error bands indicate 95% confidence interval for predictions). The red line is the identity line. For shorter CDR3 lengths, effect sizes from the length-position model tend to be larger than those from the position model; however, the opposite is the case for longer CDR3 lengths. We used *P* values from two-sided linear regression test.



Supplementary Figure 20. Explained variance in a five-fold cross validation.
To confirm that there was no overfitting in our analysis, we performed five-fold cross validation. We used the *HLA* site and CDR3 position pair that showed the strongest association in the MMLM analysis: alleles at *HLA-DRB1* site 13 were explanatory variables and CDR3 amino acid frequencies at position 109 of L13-CDR3 were response variables. In each round of cross validation, we conducted a linear regression for each CDR3 amino acid using training samples (80% of all samples: n=503) to prepare a predictive model. Then, we applied this model to the validation samples (the remaining 20%: n=125) and compared the predicted and observed frequencies of the target amino acid.
(a) an exemplar plot showing the predicted and observed frequencies of lysine (K) in a round of cross validation.
(b) Pearson's *r* for all amino acids in each round of cross validation. 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.
(c) Mean r^2 in each round of cross validation. The expected value that was estimated from the MMLM analysis (explained variance = 0.093) is shown with a dashed line.



Supplementary Figure 21. The influence of public clonotypes on cdr3-QTL signals.
Effect size estimates from the linear regression model (LM) compared across three conditions: the analyses with all clonotypes (X-axis, both plots), the analysis with non-public clonotypes (Y-axis, left), and the analysis with public clonotypes (Y-axis, right). We used the 388 CDR3 phenotypes (length-position-amino acid combinations) that had at least one significant association in the LM analysis using all clonotypes ($P < 0.05/1,249,742$ total tests), and the *HLA* amino acid alleles that had the lowest *P* value for each phenotype. The error bar indicates +/- 2 x S.E. We used *P* values from two-sided linear regression test.