nature genetics

Genetic variation in MHC proteins is associated with T cell receptor expression biases

Eilon Sharon[1,2,8], Leah V Sibener[3–5,8], Alexis Battle[6], Hunter B Fraser[2], K Christopher Garcia[3,4,7] & Jonathan K Pritchard[1,2,7]

In each individual, a highly diverse T cell receptor (TCR) repertoire interacts with peptides presented by major histocompatibility complex (MHC) molecules. Despite extensive research, it remains controversial whether germline-encoded TCR–MHC contacts promote TCR–MHC specificity and, if so, whether differences exist in TCR V gene compatibilities with different MHC alleles. We applied expression quantitative trait locus (eQTL) mapping to test for associations between genetic variation and TCR V gene usage in a large human cohort. We report strong *trans* associations between variation in the MHC locus and TCR V gene usage. Fine-mapping of the association signals identifies specific amino acids from MHC genes that bias V gene usage, many of which contact or are spatially proximal to the TCR or peptide in the TCR–peptide–MHC complex. Hence, these MHC variants, several of which are linked to autoimmune diseases, can directly affect TCR–MHC interaction. These results provide the first examples of *trans*-QTL effects mediated by protein–protein interactions and are consistent with intrinsic TCR–MHC specificity.

MHC proteins are an essential component of the adaptive immune system because of their role in presenting self and foreign processed peptides for inspection by T cells[1]. Human MHC genes—also referred to as HLA (human leukocyte antigen) genes—are extremely polymorphic, and variants of these genes have been associated with many traits, including most autoimmune diseases[2,3]. It has been suggested that in some cases the increased disease risk conferred by particular MHC alleles is due to differences in the peptides they present or to differences in the intrinsic stability of the MHC variants[4]. However, beyond the importance of these variants in shaping the sequence repertoire of antigenic peptides presented, in most cases, understanding of the functional implications of different MHC alleles and their interplay with TCR diversity is still limited[3,4].

One possible functional effect of MHC genotypes could be to influence usage of the paralogous genes that encode the TCR repertoire. Each individual has a highly diverse TCR repertoire that is able to recognize and respond to a huge variety of foreign peptides when they are presented on MHC proteins. Each TCR is a heterodimer, usually comprising α (*TRA*) and β (*TRB*) chains (1–5% of T cells instead carry γ (*TRG*) and δ (*TRD*) chains[5]). Each T cell clone expresses a unique pair of TCR chains resulting from somatic V(D)J recombination of one of each of the paralogous variable (V), joining (J) and, in β and δ chains, diversity (D) genes. During this recombination, the joints are partially digested and nucleotides are randomly added to form the highly variable and non-germline-encoded complementarity-determining region 3 (CDR3) loop that recognizes presented peptides[6]. Additional contacts with the MHC are formed by the CDR1 and CDR2 loops of the TCR, which are encoded by the V genes[6–10]. T cells subsequently undergo both positive and negative selection in the thymus to ensure specificity for foreign, but not self, peptides[11].

The TCR repertoire is reshaped in response to infection[8,12] and varies between individuals[13]. However, little is known about the extent to which the usage of different V genes in the TCR repertoire is shaped by host genetics, apart from limited observations of greater repertoire similarity among close relatives[14,15] and a report that usage of V_α genes in response to an Epstein–Barr virus (EBV) epitope depends on *HLA-B* genotype[16]. Moreover, although it is intuitive that MHC genotype might affect the TCR repertoire, the precise nature of the TCR–MHC interaction remains controversial. In contrast to B cell–secreted antibodies (which are also generated by V(D)J recombination), TCRs interact specifically with peptide–MHC (pMHC) complexes. Yet, despite numerous structural, *in vitro* and mouse *in vivo* studies, there is still active debate about whether germline-encoded TCR–MHC contacts help to promote this specificity[10,17–23] or are merely bystanders[23–26]. Recent studies have reported conflicting conclusions on this point[26,27]. If germline-encoded contacts influence TCR–MHC interaction, then we might expect different TCR V genes to differ in their compatibilities with different MHC alleles. Such differences might bias V gene usage in the post-thymic repertoire, as both thymic selection and clonal expansion of T cells are dependent on TCR–MHC interactions[21].

Here we address the question of how host genotype influences the makeup of the TCR repertoire using eQTL analysis[28] of a large

[1]Department of Genetics, Stanford University, Stanford, California, USA. [2]Department of Biology, Stanford University, Stanford, California, USA. [3]Department of Molecular Physiology, Stanford University School of Medicine, Stanford, California, USA. [4]Department of Structural Biology, Stanford University School of Medicine, Stanford, California, USA. [5]Immunology Program, Stanford University, Stanford, California, USA. [6]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA. [7]Howard Hughes Medical Institute, Stanford University, Stanford, California, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to K.C.G. (kcgarcia@stanford.edu) or J.K.P. (pritch@stanford.edu).

Received 16 January; accepted 23 June; published online 1 August 2016; doi:10.1038/ng.3625

human cohort[29] for which both RNA-seq data from peripheral blood and genotyping data are available (**Fig. 1**). We took an undirected approach to testing, across the genome, for *trans* associations between genetic variation and expression of TCR V genes (we will also use the term 'bias' to refer to genotype-dependent shifts in V gene usage). Our results suggest that MHC genotypes have an important role in determining the V gene usage profiles of each individual's TCR repertoire.

RESULTS

Expression of TCR V genes is associated with MHC variation

We analyzed RNA-seq data collected from the peripheral blood of 922 individuals[29] of European ancestry. To estimate the relative expression of each V gene, we counted the number of reads that mapped uniquely to each V gene while controlling for the total expression of each TCR chain and other relevant covariates (**Fig. 1**, Online Methods, **Supplementary Figs. 1** and **2**, and **Supplementary Table 1**). After removing genes and individuals with low numbers of mapped reads, we had expression measurements for 44 V_α, 40 V_β, 11 V_γ and 3 V_δ genes in each of 895 individuals (**Supplementary Figs. 3–5** and **Supplementary Tables 2** and **3**). As ordinarily only one functional TCR is expressed for each T cell, estimated expression levels will be determined by the fraction of T cells expressing each TCR, as well as the expression level of the TCR in each cell. As a control, we applied a similar pipeline to analyze the V genes from B cell–secreted antibodies (immunoglobulins), which are not expected to interact with MHC.

To test for associations between genotype and the expression of TCR V genes, we used genome-wide genotype measurements in the same individuals[29] (**Fig. 1**). We initially tested for short-range eQTLs, that is, ones within 1 Mb of each V gene. We excluded from this analysis a small number of genes in which read mappability varied across haplotypes (Online Methods and **Supplementary Fig. 6**). As expected, we found many short-range eQTLs, including for 78% of TCR V genes and 46% of immunoglobulin V genes at a 5% false discovery rate (FDR) (**Supplementary Fig. 7a** and **Supplementary Table 4**), presumably reflecting *cis*-acting effects on gene regulation.

We next tested across the genome for long-range eQTLs. Notably, we found multiple highly significant associations between the MHC locus and expression of TCR V genes. Expression of 47.7% and 22.5% of the V_α and V_β genes, respectively, was associated at a stringent threshold of $P < 5 \times 10^{-8}$, which accounts for genome-wide significance testing (**Fig. 2a**,**b**, **Supplementary Fig. 8** and **Supplementary Table 5**). Restricting the analysis to the extended MHC locus reduced the multiple-testing burden such that expression of 66% and 35% of the V_α and V_β genes, respectively, was significantly associated with variation in the MHC locus at a 5% FDR (Online Methods, **Supplementary Fig. 7b** and **Supplementary Table 4**). The MHC locus stands out, as we observed just one other genome-wide-significant *trans* association with any TCR V gene (*TRVB24-1*, associated with variation near *ZNF443*) (**Fig. 2a**,**b** and **Supplementary Fig. 9**). Interestingly, despite the lack of MHC restriction of γδ TCRs[30], the few significant associations with expression of V_δ genes also mapped to the MHC locus (**Fig. 2c**,**d**). We speculate that these associations might be caused by the small population of δβ T cells that recognize MHC-presented peptides[30]. Notably, there were no genome-wide-significant associations between the MHC locus and the expression of immunoglobulin V genes, which encode antibodies (**Fig. 2e–g** and **Supplementary Figs. 7b** and **8**), and there was just one association between MHC variants and expression of any other gene (**Fig. 2h**), highlighting the specific relationship between the MHC and TCR.

Classical MHC genes drive most MHC locus–TCR associations

To localize functional elements in the MHC region that associate with V_α or V_β gene expression, we used the genotyped SNPs to impute genotype estimates for all known MHC variants[31] (**Fig. 1**). We then performed forward stepwise regression to identify independent associations with any SNPs within the MHC locus or with amino acid polymorphisms in classical MHC genes. We iteratively added polymorphic positions to a predictive model of each gene's expression until no additional position significantly improved the model fit (*F* test, *P*-value threshold = 0.05 under Bonferroni correction; **Fig. 3a**, Online Methods, **Supplementary Fig. 10** and **Supplementary Table 6**).

Using this conservative threshold, we identified 66 independent MHC–TCR associations for 28 TCR V_α and 15 V_β genes. These associations explained from 2.2 to 37% of the variation in expression for each V gene (**Supplementary Fig. 11** and **Supplementary Table 7**). It is often difficult to precisely locate causal sites in QTL mapping because of linkage disequilibrium (LD; the property that genotypes at linked sites tend to be correlated). Nonetheless, despite the fact that LD tends to spread association signals, we observed strong enrichment of signals within the transcribed regions of classical MHC genes, especially in *HLA-DRB1*. Forty-two of 66 associations were in MHC

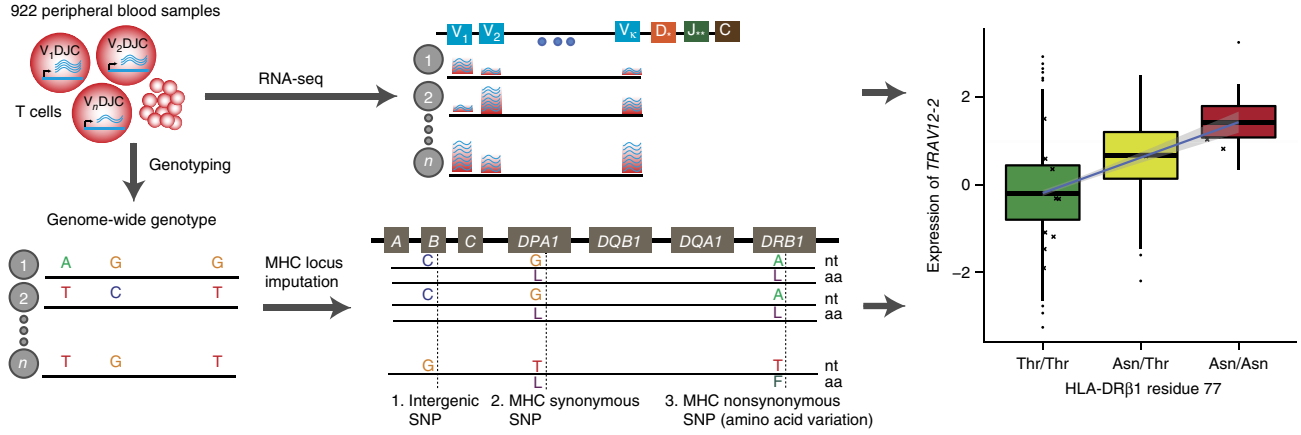

Figure 1 Illustration of our approach. Expression of TCR V genes in peripheral blood was estimated by mapping whole-blood RNA-seq[29] reads to V genes while controlling for relevant individual-level covariates and the total number of reads mapped for each TCR chain in that individual. Genotypes were measured across the genome using the Illumina HumanOmni1-Quad BeadChip[29]. MHC genotypes were imputed with SNP2HLA[31]. Associations between nucleotide or amino acid genotypes and V gene expression were tested using Pearson correlations. nt, nucleotide; aa, amino acid.

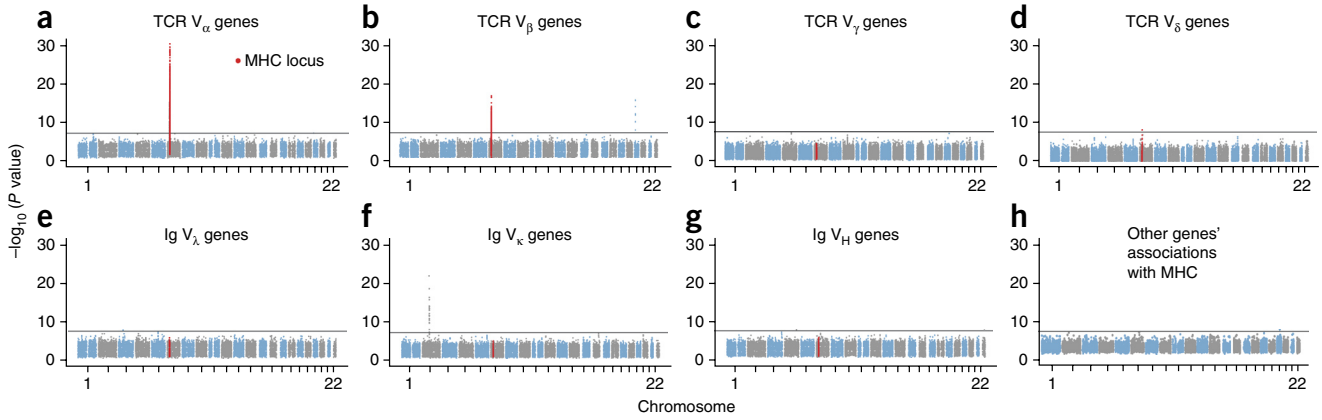

Figure 2 Expression of TCR V_α and V_β genes is significantly associated with genetic variation in the MHC locus. (**a**–**d**) Manhattan plots showing, for each of 649,863 measured SNPs, the most significant association across all V genes in the TCR α chain (**a**), β chain (**b**), γ chain (**c**) and δ chain (**d**). (**e**–**g**) Manhattan plots are shown as in **a**–**d** for genes in the immunoglobulin (Ig) λ chain (**e**), κ chain (**f**) and heavy (H) chain (**g**). (**h**) A Manhattan plot showing, for each of 13,732 tested genes, the most significant association between its expression and any of the 9,967 measured SNPs in the extended MHC locus. Excluded from the plot were all immunoglobulin and TCR V genes, genes within 1 Mb of the MHC locus and one pseudogene with high sequence similarity to the MHC (*RNF5P1*). The black horizontal lines correspond to a *P* value of 5×10^{-8}.

genes; of these, 37 associations were in class II genes and 25 associations were in the class II gene *HLA-DRB1* alone (representing 3.5-fold, 7.5-fold and 15.8-fold enrichment, respectively, relative to all variants). In addition, many of the 24 remaining associations outside genes were near classical MHC protein-encoding regions and may thus be in LD with causal variants in genes. The larger number of associations with variation in MHC class II protein-encoding regions than in MHC class I protein-encoding regions may be biologically meaningful, but this difference might also reflect greater power in our data set to detect class II interactions due to the higher abundance of $CD4^+$ than $CD8^+$ T cells in peripheral blood[32].

To test the robustness of our results, we conducted two further analyses. First, we tested for independent associations using classical four-digit MHC haplotypes instead of nucleotide and amino acid variation. This analysis yielded qualitatively similar results. Seventy-five of 92 associations were with MHC class II haplotypes; of these, 32 associations were with *HLA-DRB1* haplotypes (representing 1.6-fold and 2.4-fold enrichment, respectively, relative to all classical MHC haplotypes (**Supplementary Figs. 12** and **13**, and **Supplementary Table 8**). HLA-DRB1*03:01 was associated with expression of the largest number of different V genes (12 genes). Second, we performed a joint multiple-phenotype regression analysis. This analysis also indicated that MHC class II genes and especially *HLA-DRB1* contributed the most signals (Online Methods and **Supplementary Figs. 14**–**16**).

To assess the specific contribution of coding variants to these associations, we used a variance-components method designed for genomic data, GCTA[33], to estimate the fraction of the variation in expression of TCR V genes that can be explained by coding variants in classical MHC genes. We fit a model with genetic components representing the amino acid variation for each MHC gene and a component for variants in the MHC locus outside the transcribed regions of the classical MHC genes (**Fig. 3b**,**c** and **Supplementary Figs. 17**–**22**). We found that amino acid variants for MHC genes explained a significant fraction of variation in expression for 33 of 44 TCR V_α genes and 16 of 40 V_β genes and explained 92% and 88%, respectively, of the total variance in gene expression for these classes (5% FDR; **Fig. 3b**,**c**, Online Methods and **Supplementary Fig. 21**). In a negative-control analysis, only one immunoglobulin V gene of 149 was significant at a 5% FDR. For significant TCR V genes, the fraction of variance explained ranged from 5–75% (additional variance may be due to environmental or random factors, as well as measurement noise—especially for genes expressed at low levels; **Supplementary Fig. 22**). Thus, in summary, we conclude that the vast majority of the TCR expression variation explained by the MHC locus is due to amino acid variation in MHC proteins, with major contributions of MHC class II β chains and especially HLA-DRβ1.

MHC residues that bias TCR expression contact the TCR

Our next goal was to infer which amino acid positions are most likely responsible for expression biases of the TCR V genes (Online Methods). As many of the positions are in strong LD, it is often unclear which position is causal for a particular association. To quantify the uncertainty, we implemented a Bayesian Markov chain Monte Carlo (MCMC) approach that sampled over the joint distribution of potential causal positions that were consistent with the association data for each V gene. In comparison to frequentist approaches such as variable selection methods, which generally make firm choices about which sites to include, Bayesian models are generally better at quantifying and reflecting the uncertainty due to LD[34–37]. Our model started with a low, uniform prior probability that any given amino acid position would be causal and incorporated the intuition that if a particular position is causal for one V gene then it may be more likely to be causal for others as well (separately for V_α and V_β genes). The model outputs a posterior probability that any given amino acid position is causally associated with expression of a particular V gene or with any V gene, while accounting for correlations across sites due to LD.

The results identified several MHC amino acid positions with high posterior probabilities of influencing expression of TCR V_α genes (**Fig. 4** and **Supplementary Table 9**), although for some associations the posterior was shared across multiple potential causal positions in strong LD (**Supplementary Fig. 23**). Three of the top 15 amino acid positions influencing expression of TCR V_α genes (HLA-DRβ1 residues 71 and 86 and HLA-DQβ1 residue 57) are strongly associated with several autoimmune diseases. Alleles for these three amino acids were strongly correlated with expression biases in TCR V_α genes (**Supplementary Fig. 24**). For example, HLA-DRβ1 residue 71, the third-ranked amino acid position, is strongly associated with seropositive rheumatoid arthritis, type 1 diabetes and multiple sclerosis[38–40]. Different variants of the amino acid at this position were correlated with increased expression of different V genes. It is

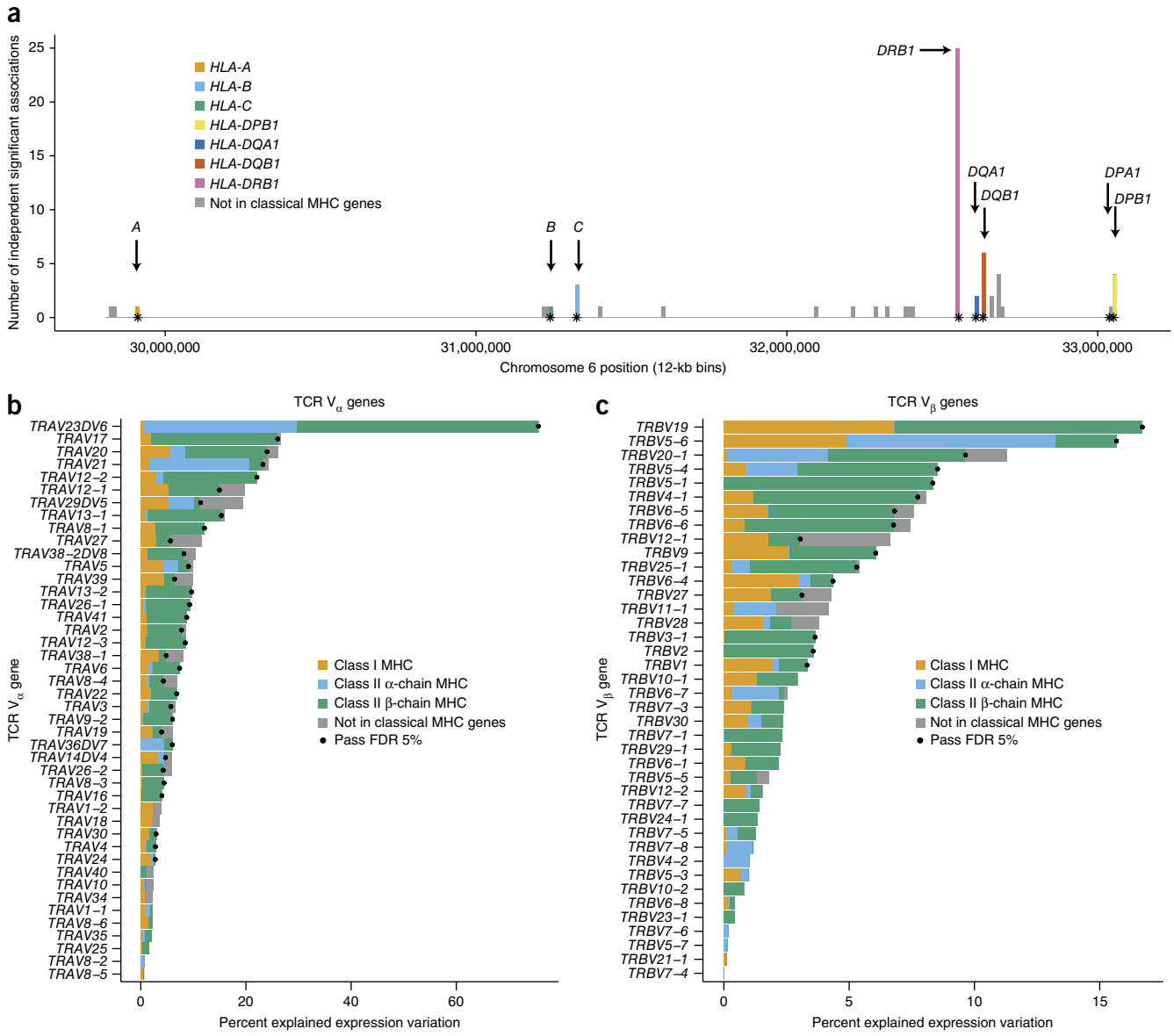

Figure 3 Expression of TCR V_α and V_β genes is associated with amino acid variation in MHC proteins. (**a**) Independent associations between V_α or V_β gene expression and nucleotide or amino acid variation in the MHC locus ($P < 0.05$ with Bonferroni correction). SNPs are binned according to their genomic position (in 12-kb bins); asterisks correspond to the center positions of the classical MHC genes. For binning by equal numbers of SNPs, see **Supplementary Figure 10**. (**b**,**c**) Variation in TCR V_α (**b**) and V_β (**c**) gene expression explained by amino acid variation for classical MHC genes and genetic variability in the MHC locus outside the transcribed regions of the classical MHC genes. Values were computed using GCTA[33]. A dot indicates that the total fraction of variation explained by the MHC gene components was significant at a 5% FDR.

possible that biases in TCR V gene expression, previously shown to affect the outcome of autoimmunity and infection[41,42], are related to some MHC associations with autoimmune and infectious disease risk. Consistent with the analysis above, we found fewer positions that likely influence expression of TCR V_β genes (**Supplementary Fig. 25** and **Supplementary Table 10**).

If the detected associations result from MHC residues influencing the TCR–pMHC interaction, then the relevant residues should cluster near the contact interface involved in TCR interaction with MHC or the presented peptide (influencing the TCR indirectly). A priori, MHC residues that affect TCR germline interactions could be direct pairwise structural contacts or could cause indirect effects relayed through subtle conformational changes of the MHC or peptide from within the MHC groove, as seen for alloreactive MHC in graft rejection[43].

To test whether the associated residues tend to be at or near the TCR–pMHC interface, we first superimposed our genetic mapping results for HLA-DRβ1 residues onto protein structures of class II MHC–TCR interactions (**Fig. 5a**,**b**). The residues with high posterior probability of influencing V gene usage tended to be either physically near or in direct contact with the TCR in structures that contained HLA-DRβ1 (**Fig. 5d**). We next aligned structures of TCRs bound to human class II MHCs and identified all contacts between the HLA-DR, HLA-DQ or HLA-DP β chains and any of the TCR α chains or the presented peptides (**Fig. 5e** and **Supplementary Figs. 26** and **27**; see the Online Methods for details).

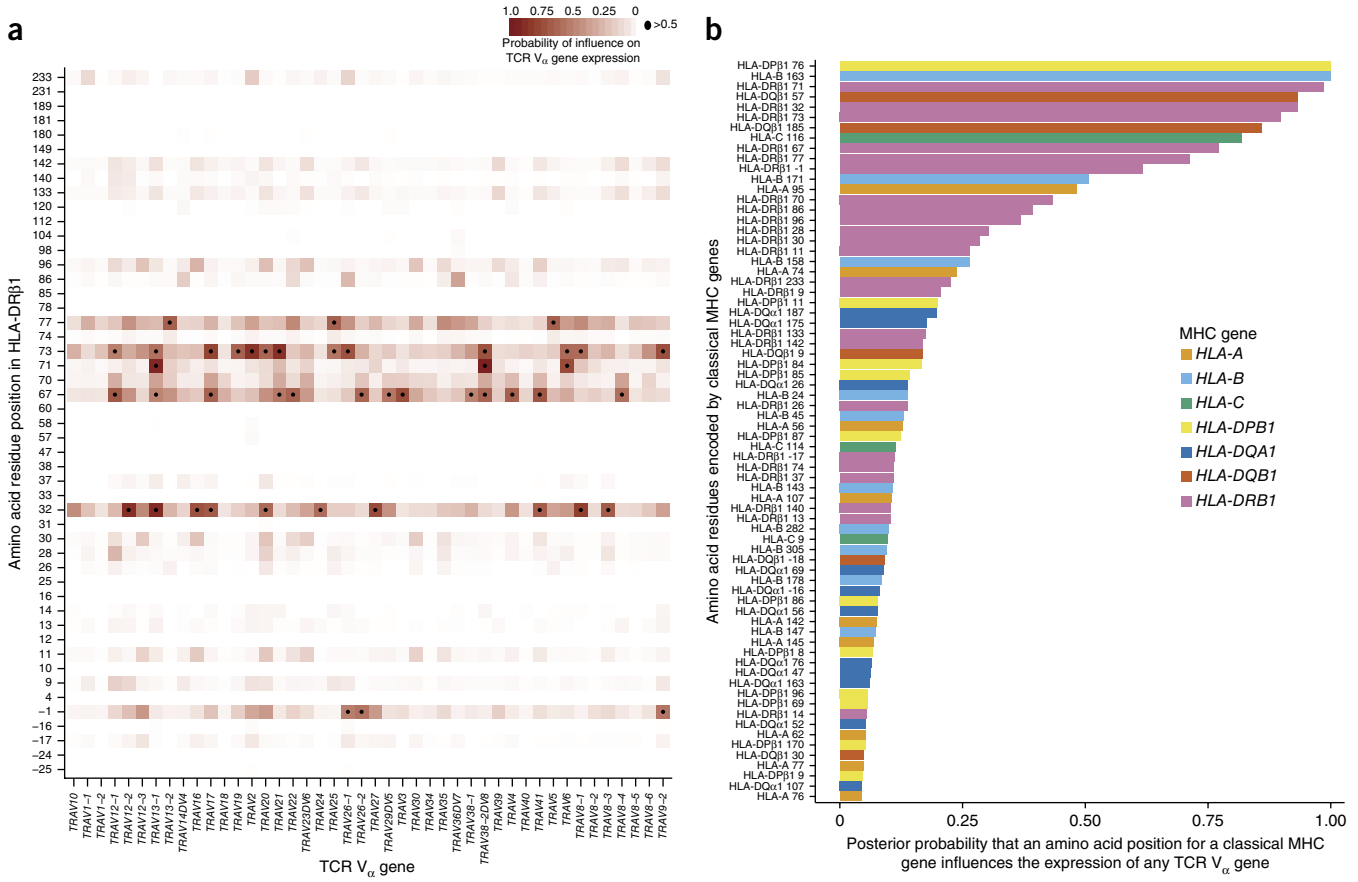

Figure 4 Bayesian inference of amino acid residues encoded in MHC genes that influence expression of TCR V_α genes. (**a**) Estimated posterior probabilities that amino acid residues in HLA-DRβ1 (*y* axis) influence expression of each TCR V_α gene (*x* axis). Dots indicate positions with >50% probability of influencing expression of a particular V gene. (**b**) Posterior probabilities that particular MHC amino acid residues influence expression of any TCR V_α gene.

Despite the large diversity in TCR α chain–MHC β chain interaction chemistries, a small subset of MHC residues contacted the TCR α chain in a large fraction of the structures, in agreement with similar analyses of TCR–MHC class I complexes[25,44]. For example, residue 77, predicted by our model to influence TCR V_α gene expression (posterior probability of 0.71), showed diverse but consistent contact with germline-encoded TCR residues in all analyzed TCR–MHC class II complexes (**Fig. 5c** and **Supplementary Table 11**). To quantitatively test whether MHC amino acid residues near the TCR interaction surface tend to be associated with TCR V gene expression, we correlated our model posterior probabilities with the mean distance between the centroid for each MHC residue and the centroid for the closest TCR residue (**Supplementary Fig. 28**) or peptide residue (**Supplementary Fig. 29**) in all analyzed complexes. Although inter-residue proximity did not strictly correlate with energetic importance in protein–protein interactions, we found that the association probabilities and TCR distances were significantly correlated (Pearson $r = -0.28$, $P = 0.022$; Spearman $r = -0.35$, $P = 0.0044$; **Supplementary Fig. 30a**). This correlation would be higher if not for one outlier, residue 185, which may either be a false prediction or indirectly affect the structural integrity of the MHC protein[45]. The results were similar when distances between amino acid C_α atoms were used instead of distances between centroids (**Supplementary Fig. 30b**). Hence, our structural analysis indicates that several of the residues in close proximity to the TCR or peptide affect expression of TCR V genes.

DISCUSSION

Our results show that MHC genotype has a key role in shaping the TCR repertoire in a broad population sample, even in the absence of a shared immune challenge. We see an excess of association signals in MHC class II loci, and within *HLA-DRB1* in particular; however it is unclear whether this reflects a greater role for class II genes in shaping the TCR repertoire or simply differences in power due to the greater abundance of $CD4^+$ T cells in whole blood[32]. Many of the observed associations are linked to MHC residues, implying a direct role for protein–protein interactions in mediating these effects. We suggest that germline-encoded TCR–MHC compatibilities may bias thymic selection of some V genes relative to others, in an MHC-dependent manner.

Further, we were able to fine-map some of these signals to specific amino acids that lie at the TCR–pMHC contact interface. Overall, we find evidence for the intuitive result that positions near the interaction surface have higher probability of influencing expression of V genes. That said, not all of the most strongly associated MHC positions are in direct physical contact with the TCR. More distant associations—for example, at HLA-DQβ1 residue 185—may reflect LD with other more proximal sites or may be responsible for longer-range effects within the protein complex. It is known that amino acids distal to contact interfaces do sometimes have important effects on interaction energetics, even when they are not in direct contact[45]. In particular, this has been shown previously for alloreactive MHC in graft

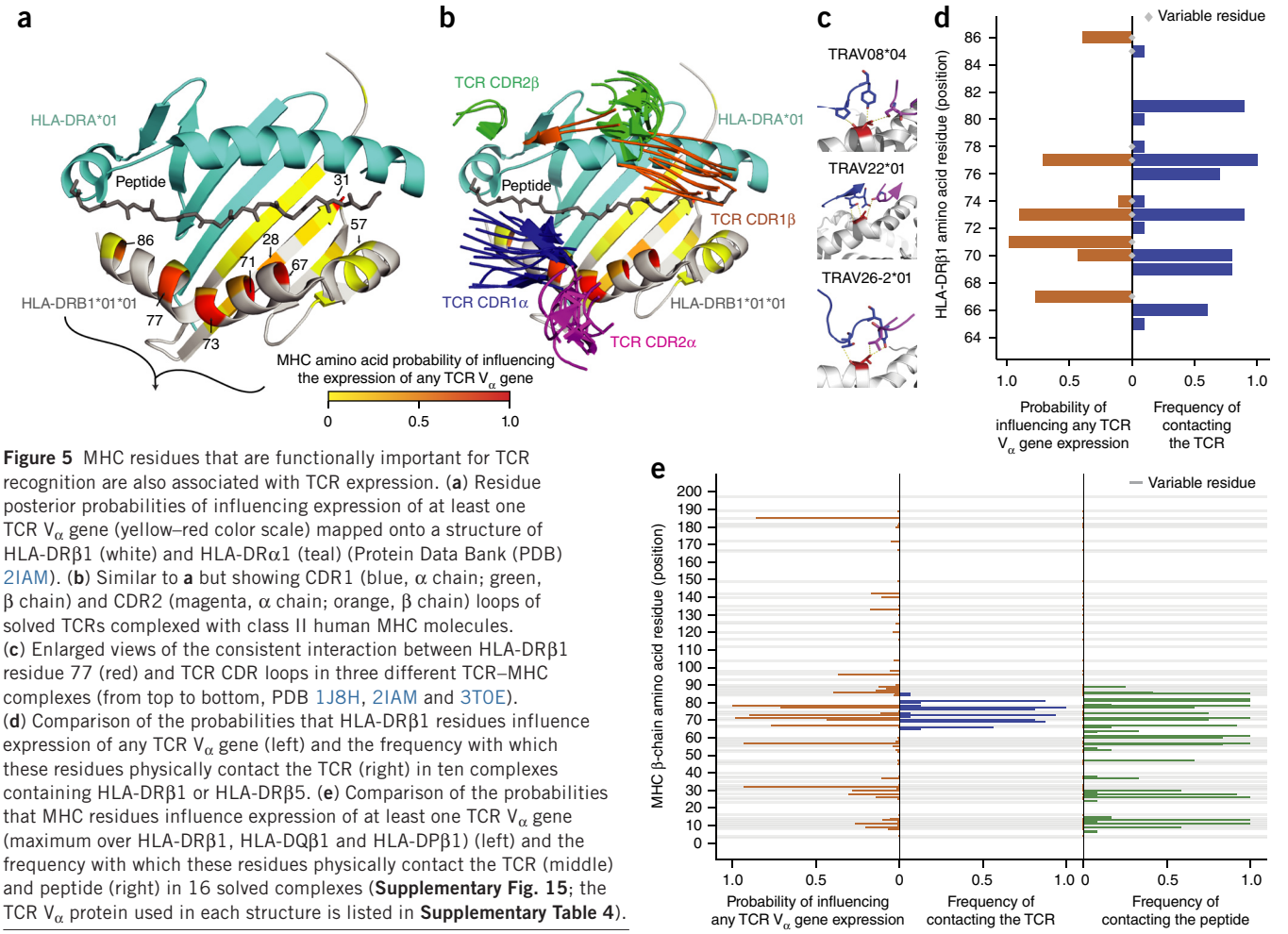

Figure 5 MHC residues that are functionally important for TCR recognition are also associated with TCR expression. (**a**) Residue posterior probabilities of influencing expression of at least one TCR V_α gene (yellow–red color scale) mapped onto a structure of HLA-DRβ1 (white) and HLA-DRα1 (teal) (Protein Data Bank (PDB) 2IAM). (**b**) Similar to **a** but showing CDR1 (blue, α chain; green, β chain) and CDR2 (magenta, α chain; orange, β chain) loops of solved TCRs complexed with class II human MHC molecules. (**c**) Enlarged views of the consistent interaction between HLA-DRβ1 residue 77 (red) and TCR CDR loops in three different TCR–MHC complexes (from top to bottom, PDB 1J8H, 2IAM and 3T0E). (**d**) Comparison of the probabilities that HLA-DRβ1 residues influence expression of any TCR V_α gene (left) and the frequency with which these residues physically contact the TCR (right) in ten complexes containing HLA-DRβ1 or HLA-DRβ5. (**e**) Comparison of the probabilities that MHC residues influence expression of at least one TCR V_α gene (maximum over HLA-DRβ1, HLA-DQβ1 and HLA-DPβ1) (left) and the frequency with which these residues physically contact the TCR (middle) and peptide (right) in 16 solved complexes (**Supplementary Fig. 15**; the TCR V_α protein used in each structure is listed in **Supplementary Table 4**).

rejection[43]. Nonetheless, our Bayesian approach does highlight several amino acid residues positioned in the TCR–pMHC contact interface as being important for biasing V gene usage. To the best of our knowledge, these are the first examples of *trans* associations mediated by protein–protein interactions.

Our observations also have implications for a long-standing debate about the basis of TCR specificity for MHC molecules and the molecular forces responsible for MHC restriction. In 1971, Niels Jerne postulated that germline TCR and MHC genes coevolved to be predisposed toward interacting[46]—a phenomenon also referred to as 'germline bias'. Our observation that MHC genotype has a direct association with TCR V gene usage in the broader population implies that germline-encoded TCR–MHC contacts influence interaction specificity. This orthogonal genetic evidence is in agreement with a variety of structural and functional data supporting a model of intrinsic specificity between TCR and MHC proteins. Other examples of supportive data include structural studies of numerous TCR–pMHC complexes that show persistent germline-encoded contacts between V regions and MHC molecules[20,47–49] and compatible residues between TCR loops and HLA-A*02:01 (ref. 50). Some of these contacts are necessary for functional recognition of MHC molecules and, when mutated, can lead to drastically altered outcomes of thymic selection[20,21]. Germline-derived V_α elements have been shown to influence MHC class I versus class II selection[51]. Moreover, T cells have recently been shown to recognize MHC molecules independently of the MHC allele and peptide[27], supplying functional evidence that the TCR has intrinsic specificity for the MHC via germline V genes. Other studies argue against the intrinsic specificity model. For example, it has been reported that specific MHC residues are not essential for TCR recognition[25], and there are no known constraints on the CDR sequences[24]; TCRs in mice that lack MHC class I and II molecules, CD4 and CD8 are activated by a non-MHC ligand[23], and two TCR–pMHC structures composed of proteins from the same V genes were recently shown to have reversed polarity for MHC binding[26]. Our observations that TCR–MHC compatibilities exist within a large cohort of individuals likely reflect how most (although not necessarily all) TCR–MHC interactions occur. This leads us to view the studies that are discordant with the coevolution model as representing bona fide deviations from a 'canonical' continuum of TCR–MHC recognition modes. Some level of 'non-canonical' recognition may be expected, considering the enormous repertoire of TCR sequences possible through recombination. In summary, our genetic results are in agreement with some degree of 'hard-wired' TCR–MHC recognition and are supportive of the Jerne Hypothesis.

One key limitation of our approach is that it is based on RNA-seq analysis of peripheral whole blood. Therefore, our measurements reflect average V gene usage across different subpopulations of T cells, most notably aggregating across the $CD4^+$ and $CD8^+$ T cell subsets. Previous work has identified several variants in the MHC region that are associated with individual-level variation in CD4:CD8 ratios[52]; moreover, V gene usage differs between $CD4^+$ and $CD8^+$ T cells[53]. Therefore, one concern is whether the variants associated with

CD4:CD8 ratio might drive the signals reported here. However, we find that these SNPs are only modestly associated with our signals (**Supplementary Fig. 31a**,**b**) and are not selected by our conditional analyses (**Fig. 3a** and **Supplementary Figs. 12** and **15**). Controlling for these SNPs results in similar, highly significant associations with expression of V genes (**Supplementary Fig. 31c**,**d**). Although we cannot rule out some effect on the fine-mapping, especially in the *HLA-B* gene, which reportedly harbors the strongest signal for CD4:CD8 ratio, the localization of high posteriors to the TCR–pMHC interface suggests that this effect is not very strong. Additionally, some measurement noise may also result from the expansion of particular T cell clones on an individual-specific basis. These and other sources of variation are implicitly modeled by our Bayesian fine-mapping approach; however, future studies with longer-read sequencing of sorted cell populations may be able to improve mapping resolution. A final concern is the possibility that we may not be able to identify causal residues that are poorly imputed. Most residues in our data have high imputation quality, and we find no correlation between imputation quality and posterior probability (**Supplementary Fig. 32**); however, it remains possible that we may have overlooked poorly imputed causal sites.

In summary, we have found that usage of TCR V_α and V_β genes is associated with the genotype of MHC proteins. Our structural analysis suggests that these associations result from differences in the specificity of different TCR V genes for different MHC variants. Our results provide a compelling example of strong *trans* associations that are mediated by protein–protein interaction between a receptor and its ligand and shed light on the basis of TCR–MHC recognition.

URLs. International Immunogenetics Information System (IMGT), http://www.imgt.org/; RCSB Protein Data Bank (PDB), http://www.rcsb.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Golan, D. Knowles, A. Fu, M. Birnbaum, M. Gee, J. Mendoza, A. Bhaskar and T. Raj for helpful discussions and the anonymous referees for valuable comments. This work was supported by NIH grants HG0070736, 1R01GM097171-01A1, RO1AI03867 and U19AI057229, the Howard Hughes Medical Institute, the EMBO Long-Term Fellowship and a National Science Foundation Graduate Research Fellowship.

AUTHOR CONTRIBUTIONS

E.S., J.K.P. and K.C.G. conceived the project. E.S. performed genetic analyses with input from A.B., H.B.F. and J.K.P. E.S. and L.V.S. performed structural analyses with input from K.C.G. E.S., L.V.S., K.C.G. and J.K.P. wrote the manuscript. The work was supervised by K.C.G. and J.K.P. All authors reviewed, revised and provided feedback on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Neefjes, J., Jongsma, M.L.M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836 (2011).
2. McDevitt, H.O. & Bodmer, W.F. HL-A, immune-response genes, and disease. *Lancet* **1**, 1269–1275 (1974).
3. Gutierrez-Arcelus, M., Rich, S.S. & Raychaudhuri, S. Autoimmune diseases—connecting risk alleles with molecular traits of the immune system. *Nat. Rev. Genet.* **17**, 160–174 (2016).
4. Miyadera, H. & Tokunaga, K. Associations of human leukocyte antigens with autoimmune diseases: challenges in identifying the mechanism. *J. Hum. Genet.* **60**, 697–702 (2015).
5. Vantourout, P. & Hayday, A. Six-of-the-best: unique contributions of γδ T cells to immunology. *Nat. Rev. Immunol.* **13**, 88–100 (2013).
6. Rossjohn, J. *et al.* T cell antigen receptor recognition of antigen-presenting molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).
7. Rudolph, M.G., Stanfield, R.L. & Wilson, I.A. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–466 (2006).
8. Turner, S.J., Doherty, P.C., McCluskey, J. & Rossjohn, J. Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.* **6**, 883–894 (2006).
9. Housset, D. & Malissen, B. What do TCR–pMHC crystal structures teach us about MHC restriction and alloreactivity? *Trends Immunol.* **24**, 429–437 (2003).
10. Garcia, K.C. *et al.* A closer look at TCR germline recognition. *Immunity* **36**, 887–888 (2012).
11. 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).
12. Roudier, J. Association of MHC and rheumatoid arthritis. Association of RA with *HLA-DR4*: the role of repertoire selection. *Arthritis Res.* **2**, 217–220 (2000).
13. Robins, H.S. *et al.* Overlap and effective size of the human $CD8^+$ T cell receptor repertoire. *Sci. Transl. Med.* **2**, 47ra64 (2010).
14. Zvyagin, I.V. *et al.* Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proc. Natl. Acad. Sci. USA* **111**, 5980–5985 (2014).
15. Gulwani-Akolkar, B. *et al.* Do HLA genes play a prominent role in determining T cell receptor V_α segment usage in humans? *J. Immunol.* **154**, 3843–3851 (1995).
16. Miles, J.J. *et al.* TCRα genes direct MHC restriction in the potent human T cell response to a class I–bound viral epitope. *J. Immunol.* **177**, 6804–6814 (2006).
17. Garcia, K.C. Reconciling views on T cell receptor germline bias for MHC. *Trends Immunol.* **33**, 429–436 (2012).
18. Garcia, K.C., Adams, J.J., Feng, D. & Ely, L.K. The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat. Immunol.* **10**, 143–147 (2009).
19. Castro, C.D., Luoma, A.M. & Adams, E.J. Coevolution of T-cell receptors with MHC and non-MHC ligands. *Immunol. Rev.* **267**, 30–55 (2015).
20. Marrack, P., Scott-Browne, J.P., Dai, S., Gapin, L. & Kappler, J.W. Evolutionarily conserved amino acids that control TCR–MHC interaction. *Annu. Rev. Immunol.* **26**, 171–203 (2008).
21. Scott-Browne, J.P., White, J., Kappler, J.W., Gapin, L. & Marrack, P. Germline-encoded amino acids in the αβ T-cell receptor control thymic selection. *Nature* **458**, 1043–1046 (2009).
22. Van Laethem, F. *et al.* Lck availability during thymic selection determines the recognition specificity of the T cell repertoire. *Cell* **154**, 1326–1341 (2013).
23. Van Laethem, F. *et al.* Deletion of CD4 and CD8 coreceptors permits generation of αβ T cells that recognize antigens independently of the MHC. *Immunity* **27**, 735–750 (2007).
24. Holland, S.J. *et al.* The T-cell receptor is not hardwired to engage MHC ligands. *Proc. Natl. Acad. Sci. USA* **109**, E3111–E3118 (2012).
25. Burrows, S.R. *et al.* Hard wiring of T cell receptor specificity for the major histocompatibility complex is underpinned by TCR adaptability. *Proc. Natl. Acad. Sci. USA* **107**, 10608–10613 (2010).
26. Beringer, D.X. *et al.* T cell receptor reversed polarity recognition of a self-antigen major histocompatibility complex. *Nat. Immunol.* **16**, 1153–1161 (2015).
27. Parrish, H.L., Deshpande, N.R., Vasic, J. & Kuhns, M.S. Functional evidence for TCR-intrinsic specificity for MHCII. *Proc. Natl. Acad. Sci. USA* **113**, 3000–3005 (2016).
28. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
29. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
30. Sottini, A., Imberti, L., Fiordalisi, G. & Primi, D. Use of variable human V_δ genes to create functional T cell receptor α chain transcripts. *Eur. J. Immunol.* **21**, 2455–2459 (1991).
31. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
32. Sinclair, C., Bains, I., Yates, A.J. & Seddon, B. Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proc. Natl. Acad. Sci. USA* **110**, E2905–E2914 (2013).
33. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
34. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
35. Wallace, C. *et al.* Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping. *PLoS Genet.* **11**, e1005272 (2015).
36. Wellcome Trust Case Control Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).

37. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).
38. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).
39. 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).
40. Patsopoulos, N.A. *et al.* Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. *PLoS Genet.* **9**, e1003926 (2013).
41. Messaoudi, I., Guevara Patiño, J.A., Dyall, R., LeMaoult, J. & Nikolich-Zugich, J. Direct link between MHC polymorphism, T cell avidity, and diversity in immune defense. *Science* **298**, 1797–1800 (2002).
42. Price, D.A. *et al.* Public clonotype usage identifies protective Gag-specific $CD8^+$ T cell responses in SIV infection. *J. Exp. Med.* **206**, 923–936 (2009).
43. Luz, J.G. *et al.* Structural comparison of allogeneic and syngeneic T cell receptor–peptide–major histocompatibility complex complexes: a buried alloreactive mutation subtly alters peptide presentation substantially increasing V_β interactions. *J. Exp. Med.* **195**, 1175–1186 (2002).
44. Murray, J.S. An old Twist in HLA-A: CDR3α hook up at an R65-joint. *Front. Immunol.* **6**, 268 (2015).
45. Levin, A.M. *et al.* Exploiting a natural conformational switch to engineer an interleukin-2 'superkine'. *Nature* **484**, 529–533 (2012).
46. Jerne, N.K. The somatic generation of immune recognition. *Eur. J. Immunol.* **1**, 1–9 (1971).
47. Dai, S. *et al.* Crossreactive T cells spotlight the germline rules for αβ T cell-receptor interactions with MHC molecules. *Immunity* **28**, 324–334 (2008).
48. Adams, J.J. *et al.* Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR–peptide–MHC cross-reactivity. *Nat. Immunol.* **17**, 87–94 (2016).
49. Feng, D., Bond, C.J., Ely, L.K., Maynard, J. & Garcia, K.C. Structural evidence for a germline-encoded T cell receptor–major histocompatibility complex interaction 'codon'. *Nat. Immunol.* **8**, 975–983 (2007).
50. Blevins, S.J. *et al.* How structural adaptability exists alongside HLA-A2 bias in the human αβ TCR repertoire. *Proc. Natl. Acad. Sci. USA* **113**, E1276–E1285 (2016).
51. Sim, B.C., Zerva, L., Greene, M.I. & Gascoigne, N.R. Control of MHC restriction by TCR V_α CDR1 and CDR2. *Science* **273**, 963–966 (1996).
52. Ferreira, M.A.R. *et al.* Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control. *Am. J. Hum. Genet.* **86**, 88–92 (2010).
53. Klarenbeek, P.L. *et al.* Somatic variation of T-cell receptor genes strongly associate with HLA class restriction. *PLoS One* **10**, e0140815 (2015).

ONLINE METHODS

Genotype and TCR V gene expression data. We analyzed whole-blood RNA-seq and genotyping data from 922 individuals from the Depression Genes and Networks Project reported by Battle *et al.*[29] (National Institute of Mental Health grant 5RC2MH089916). Measured SNPs were filtered as described previously, resulting in genotype data at 649,863 SNPs[29]. The expression of each V gene relative to total chain expression was estimated from peripheral blood RNA-seq analysis (~70 million 51-bp reads per individual). Sequencing reads were mapped as in Battle *et al.* (using Bowtie2 (ref. 54) with TopHat[55] default parameters), and the number of unique reads that mapped to each V, J or C TCR or immunoglobulin gene was determined using a modified version of HTSeq[56] that allows reads to map to sequences of more than one V, D, J or C gene (see **Supplementary Fig. 1** for the average number of reads mapped in an individual to each V gene). Individuals and genes with low read counts were removed: specifically, we removed from our analysis individuals with <80,000 reads mapping uniquely to TCR or immunoglobulin V, D, J or C genes (31% of the median). For analysis, we required that each V gene have on average >1 read per individual, at most 100 samples with zero reads and at least 1% of the individuals with at least 20 reads. Next, the read counts were log transformed (0.1 pseudoreads were added to avoid zeroes) and regressed on known technical and biological confounding factors as in Battle *et al.*[29] and on the log-transformed total number of reads mapped to each TCR or immunoglobulin chain (including reads that mapped to J and C genes; **Supplementary Fig. 3** and **Supplementary Table 1**). The median coefficient of variation for each V gene was 6% (**Supplementary Fig. 3**). Finally, to avoid the effect of outliers, the residuals were quantile normalized to a normal distribution (although in practice this made little difference to the detected associations). Similar methodology was applied to immunoglobulin V genes. The final data set contained measurements for 44 TCR V_α, 40 TCR V_β, 11 TCR V_γ, 3 TCR V_δ, 31 immunoglobulin V_κ, 38 immunoglobulin V_λ and 41 immunoglobulin V_H genes in 895 individuals.

As read mappability may vary by genotype, we analyzed the mappability of 25-mers from the reference and imputed alternative haplotypes (**Supplementary Fig. 6**). We found that in nine V genes >0% and up to 10% of 25-mers had better mappability in the reference sequence than in alternative haplotypes. We thus excluded all of these genes from the short-range (*cis*) association analysis (thus dropping six, one and three TCR V_α, V_β and V_γ genes, respectively). We note that mapping differences may increase measurement noise and reduce power for detecting long-range (*trans*) associations, but they should not create false positives; therefore, these genes were not removed from the *trans* analysis. Moreover, we avoided comparing expression of different V genes, as it is difficult to estimate the mappable lengths of the V genes because of possible differences between the partial digestion of V genes and J gene pairing during V(D)J recombination.

eQTL detection. Associations between genotypes and TCR V gene normalized expression values were tested using Pearson correlations. Similarly, in testing for *trans* associations between MHC and other genes, we used Pearson correlations to test all genes at least 1 Mb away from the MHC locus against all variants genotyped in the extended MHC locus. The extended MHC locus is defined as the region from the *SLC17A2* gene at the telomeric end to the *DAXX* gene at the centromeric end of chromosome 6 (hg19 coordinates 25,912,984–33,290,793)[57]. For genome-wide *trans*-QTL mapping, we used a significance threshold of $P < 5 \times 10^{-8}$, which accounts for genome-wide significance testing.

Identification of V genes that are significantly associated with genotypes in *cis* or in the MHC locus. The empirical significance of the association between the expression of each V gene and short-range (*cis*) or MHC locus genotype was evaluated by comparing the *P* value of the most significantly associated SNP to the most significant *P* values in each of 10,000 random permutations of expression values across individuals. A threshold of 5% FDR was then used to control for the testing of multiple V genes (**Supplementary Fig. 7**). The SNPs used for the *cis* analysis were imputed using SHAPEIT[58] (for prephasing) and IMPUTE2 (ref. 59). The 1000 Genomes Project Phase 1 (ref. 60) panel was used for imputation, requiring European (EUR) minor allele frequency (MAF) >0.01 and genotype likelihood >0.9. In testing for *trans* associations with MHC genotypes, we used genotyped SNPs within the extended MHC locus[57].

Imputation of MHC locus genotypes. SNP2HLA[31] was used to impute the MHC locus genotypes, with a reference panel of 5,225 individuals of European descent collected by the Type 1 Diabetes Genetics Consortium[31]. The software imputes two- and four-digit classical alleles for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1* and their corresponding amino acid polymorphisms and SNPs (**Fig. 1**, bottom). SNPs and polymorphic amino acids with a maximum allele frequency >97.5 and alleles with a frequency <0.5% were removed. The imputation quality (Beagle r^2 value with predicted true genotypes) was higher than 0.95 and 0.99 for 97% and 80% of the imputed amino acids, respectively (excluding alleles with MAF <2.5%). There was no significant correlation between the imputation quality of amino acid residues and the posterior probability that these residues influence expression of TCR V genes (**Supplementary Fig. 3**).

Conditional analysis of associations between expression of TCR V_α and V_β genes and genetic variation in the MHC locus. The imputed genotype of each variable amino acid position or nucleotide position was encoded by allelic dosage variables while omitting the most common allele (a similar encoding was used by Raychaudhuri *et al.*[38]). Conditional analysis was performed separately for each V gene using forward stepwise linear regression. In each step of the regression and for every amino acid residue or nucleotide position, we considered an expanded model that included the allelic dosage variables for that position. The significance was tested using *F* tests. If the most significant position had a *P* value $<0.05/n$, where *n* was the number of tested positions, then this position was added to the model as a covariate. Regression stopped when the *P* value of the most significant position was greater than $0.05/n$ (this threshold was $P \sim 1 \times 10^{-5}$). We performed two additional variations of this type of analysis. First, instead of testing for associations with nucleotide or amino acid positions in the MHC locus, we tested for associations with four-digit classical MHC haplotypes. Second, instead of testing for associations of MHC genetic variability (at the position or four-digit-haplotype level) with expression of each V gene separately, we tested for associations with joint expression of all V genes from a single chain. As the expression for each gene was quantile normalized to a normal distribution, using the joint expression of V genes gives equal weight to explaining the expression of each gene in the chain.

Estimating the fraction of variation in expression for each TCR V gene that is explained by genetic variation in the MHC proteins. We used the GCTA tool developed by Yang *et al.*[33]. We fit a model with a component for each variable classical MHC protein (HLA-A, HLA-B, HLA-C, HLA-DRβ1, HLA-DQα1, HLA-DQβ1, HLA-DPα1 and HLA-DPβ1) representing the variability at the amino acid level and a component for all SNPs in the MHC locus outside the transcribed regions of the classical MHC genes. Amino acid–level variation is used in place of genotype because the annotation of transcribed SNPs in MHC genes is often ambiguous and allele dependent. To adapt GCTA to multiallelic positions, we calculated the genetic relationship between individuals at a multiallelic position by averaging the relationship over dosage variables for all possible alleles. To estimate the significance of the total variation explained by the MHC protein components, we ran GCTA on 100 permutations of the expression data for each TCR V gene. The fraction of variation explained for these permuted expression vectors follows a truncated normal distribution[61]. We estimated the variance of this distribution and used it to compute a *P* value for the fraction of variation in the expression of each TCR V gene that was explained by genetic variation in MHC proteins. We corrected for multiple-hypothesis testing using a 5% FDR[62].

Inferring which MHC protein amino acid residues are associated with expression biases of TCR V genes. Identifying which amino acid residues drive associations of MHC proteins with the expression of TCR V genes is challenging because of the LD between nearby variants. To test the association of each residue while accounting for the genotype of other residues, we used a Bayesian variable selection approach with a spike-and-slab prior[63,64]. The Bayesian modeling framework has a number of practical advantages for this problem: in particular, it allows us to appropriately account for the joint uncertainty when there are multiple causal positions and extensive LD. We used priors that reflected our expectation that most residues do not directly

influence expression and that a residue that influences expression of one V gene is more likely to be relevant for expression of other V genes. The priors were set such that the model required strong evidence to return high posterior probabilities of association.

Specifically, we modeled the expression of each V_α or V_β gene as a linear combination of the imputed alleles of amino acid residues of classical MHC genes. Each residue's allelic dosage variable can either be included in the model or excluded (coefficient = 0) from the model of expression for each V gene. We sampled the space of possible models to estimate the posterior probability that each residue has at least one of its allelic dosage variables in the model of a specific V gene. Assuming that associations are caused by the effects of residues on V gene expression, we refer to this posterior as the probability that a residue influences expression of a TCR V gene.

In detail, our response vectors are the relative expression of each V gene from a specific chain across the $N = 895$ individuals. For a TCR chain with T paralogous V genes, we define a multiple-response matrix $\boldsymbol{Y}_{N\times T}$ with $y_{i,t}$ being the standardized relative expression of TCR V gene $t \in \{1,2,\ldots,T\}$ in individual $i \in \{1,2,\ldots,N\}$. Our features are the imputed amino acid residues in each individual for each of the classical MHC genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*). The allelic dosage of a specific amino acid variant ($a \in \{1,2,\ldots,A_r\}$) of residue r of MHC gene g in individual i is represented by a continuous variable ($x_{i,(g,r,a)} \in [0,2]$). For clarity, we also use below $h \in \{1,2,\ldots,H\}$ as the index for all possible combinations of (g,r,a). Our genotype matrix is therefore $\boldsymbol{X}_{N\times H}$. We filtered out residues in which the maximal allele frequency was greater than 97.5% and dropped alleles with frequency lower than 0.5%. We define $z_{t,h} \in \{0,1\}$ to be an indicator of whether genotype x_h is included (has a coefficient $\beta_{t,h} \sim N(0, \sigma^2_\beta)$) or excluded (has a zero coefficient) from the model of y_t. Together, the expression of gene t in individual i is modeled by

$$y_{i,t} = \sum_h x_{i,h} z_{t,h} \beta_{t,h} + \varepsilon_i \tag{1}$$

where

$$z_{t,h} \sim \mathrm{Ber}(\pi_h) \tag{2}$$

$$\beta_{t,h} \sim N(0,\sigma^2_\beta)$$

$$\varepsilon_{t,i} \sim N(0,\sigma^2_\varepsilon)$$

independently and

$$\pi_h \sim \mathrm{Beta}(a,b), a_0 = 0.002, b_0 = 1.998 \tag{3}$$

$$\sigma^2_\beta \sim \mathrm{invGammas}(a,b), a_0 = 1.5, b_0 = 1.5$$

$$\sigma^2_\varepsilon \sim \mathrm{invGammas}(a,b), a_0 = 1.5, b_0 = 1.5$$

Previous structural analysis of TCR–pMHC complexes has found that a limited and consistent set of MHC residues interacts with the TCR and the docking orientation of the TCR is semiconserved. This is reflected by the choice of the prior probability of $z_{t,h} = 1$ to be small and shared across V genes ($p(z_{t,h} = 1) = p_h$, for $t \in \{1,2,\ldots,T\}$) and the modeling of α-chain and β-chain genes separately.

We consider that residue r of gene g influences the expression of TCR V gene t if at least one of the allelic dosage variables that represent it is in the model (its indicator variable $z_{t,(g,r,a)}$ is equal to 1). We denote this event $F(Z,t,g,r)$ as

$$F(Z,t,g,r) = I(f(Z,t,g,r) > 0) \tag{4}$$

where

$$f(Z,t,g,r) = \left(\sum_{\{a \in 1,2,\ldots,A_{(g,r)}\}} z_{t,(g,r,a)} \right) \sim \mathrm{Bin}\left(A_{(g,r)}, \pi\right) \tag{5}$$

To avoid an a priori preference for residues with different numbers of possible alleles, we modified the prior beta distribution parameters of the indicator prior π_h in equation (3) such that the a priori expectation and variance of $F(Z,t,g,r)$ are equal for residues with different numbers of possible variants and match the prior in equation (3) for a residue with two possible alleles.

$$\mathrm{E}^{\mathrm{prior}\,\theta}(F(Z,t,g_1,r_1)) = \mathrm{E}^{\mathrm{prior}\,\theta}(F(Z,t,g_2,r_2)),\ \forall((g_1,r_1) \text{ and } (g_2,r_2)) \tag{6}$$

$$\mathrm{Var}^{\mathrm{prior}\,\theta}(F(Z,t,g_1,r_1)) = \mathrm{Var}^{\mathrm{prior}\,\theta}(F(Z,t,g_2,r_2)),\ \forall((g_1,r_1) \text{ and } (g_2,r_2)) \tag{7}$$

To find suitable values for the prior parameters, we assume that x_h is an allelic dosage variable of a residue r that has only two possible alleles and therefore is represented by a single dosage variable. We define $\pi_h \sim \mathrm{Beta}(a,b)$ to be the prior of $z_h \sim \mathrm{Ber}(\pi_h)$. Now, let x_{h^*} be one of A dosage variables that represent a residue r^* and $\pi_{h^*} \sim \mathrm{Beta}(a^*,b^*)$ be the prior of $z_{h^*} \sim \mathrm{Ber}(\pi_{h^*})$. We seek (a^*,b^*) such that the a priori expectation and variance of F over residue r^* are equal to the expectation and variance of F over residue r.

$$\mathrm{E}^{\mathrm{prior}\,\theta}(F(Z,t,g,r)) = \mathrm{E}^{\mathrm{prior}\,\theta}(F(Z,t,g^*,r^*)) \tag{8}$$

$$\mathrm{Var}^{\mathrm{prior}\,\theta}(F(Z,t,g,r)) = \mathrm{Var}^{\mathrm{prior}\,\theta}(F(Z,t,g^*,r^*)) \tag{9}$$

Equation (8) is equal to

$$\frac{a}{a+b} = \mathrm{E}\left(1-(1-\pi_{h^*})^A\right) \tag{10}$$

$$\frac{a}{a+b} = 1 - \frac{1}{\mathrm{B}(a^*,b^*)} \int_0^1 (1-\pi_{h^*})^A \pi_{h^*}^{a^*-1} (1-\pi_{h^*})^{b^*-1} d\pi_{h^*} \tag{11}$$

$$\frac{b}{a+b} = \frac{\mathrm{B}(a^*,b^*)}{\mathrm{B}(a^*,b^*+A)} \tag{12}$$

where B is the beta function. In a similar manner, equation (9) is equal to

$$\frac{a+b(a+b+1)}{(a+b)(a+b+1)} = \frac{\mathrm{B}(a^*,b^*+2A)}{\mathrm{B}(a^*,b^*+A)} \tag{13}$$

We then solved equations (12) and (13) numerically using the R nleqslv package to achieve (a^*,b^*) for various values of A.

In our analyses, the prior probability of any MHC residue interacting with a specific TCR V gene was set at 0.1%, and the prior probability of each position of the MHC interacting with any of the 44 V_α and 40 V_β genes was 4.3% and 3.9%, respectively.

To compute the posterior probabilities with which each MHC residue influences TCR V gene expression (that is, $\mathrm{E}^{\theta|X,Y}(F(Z,t,g,r))$ for all r), we sampled the space of possible models using an efficient Gibbs sampler[65] where the likelihood is integrated over the coefficients (β)

$$\mathrm{L}\left(Z,\sigma^2_\beta,\sigma^2_\varepsilon \mid Y,X\right) = \mathrm{P}\left(Y \mid Z,X,\sigma^2_\beta,\sigma^2_\varepsilon\right) = \int_\beta \mathrm{P}\left(Y \mid Z,X,\sigma^2_\varepsilon\right) \mathrm{P}\left(\beta \mid \sigma^2_\beta\right) d\beta \tag{14}$$

Altogether, we sampled 100,000 samples from ten different starting points (excluding the first 2,000 samples).

Structural analysis. A full list of TCR-pMHC structures and intermolecular interactions is curated by the International Immunogenetics Information System (IMGT)[66]. We used all structures that were available in PDB[67] on 15 June 2015. To compare intermolecular interactions to genetic analysis, IMGT numbering was converted to PDB numbering for all residues (PDB is indexed in the same manner as proteins in the UniProt database[68]). For the structure 4P4K, we converted the PDB numbering for the HLA-DP residues to the HLA-DQ and HLA-DR numbering using IMGT labeling (for example, residue 72AG in the IMGT numbering equates to PDB residue 75 for HLA-DP structures, but residue 77 for all HLA-DP and HLA-DQ structures). This allowed for a more consistent comparison between the different alleles. The IMGT structural contact algorithm determined the structural contacts, which were defined as polar, non-polar and hydrogen bonds.

doi:10.1038/ng.3625

Pairwise distance analysis. The distance between a pair of amino acid residues were calculated as the distance between the centroids of the amino acids using PDB *xyz* coordinates for every atom in a given amino acid. This calculation accounts for directionality of the side chains in the structure. For each MHC β-chain residue, the minimum distance to any TCR α-chain residue and the minimum distance to every peptide residue were calculated. For comparison, we also calculated the distance between the C_α atom of the residues using the PDB *xyz* coordinates of the C_α atoms (**Supplementary Fig. 28b**). The results using the two distance measurements were similar. A list of the PDB files used can be found in **Supplementary Table 11**.

54. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
55. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
56. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
57. de Bakker, P.I.W. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
58. Delaneau, O. & Marchini, J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
59. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
60. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
61. Visscher, P.M., Yang, J. & Goddard, M.E. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang *et al.* (2010). *Twin Res. Hum. Genet.* **13**, 517–524 (2010).
62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
63. Mitchell, T.J. & Beauchamp, J.J. Bayesian variable selection in linear regression. *J. Am. Stat.* **83**, 1023–1032 (1988).
64. Ishwaran, H. & Rao, J.S. Spike and slab gene selection for multigroup microarray data. *J. Am. Stat.* **100**, 764–780 (2005).
65. Geman, S. & Geman, D. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984).
66. Lefranc, M.-P. *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2015).
67. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
68. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).