nature genetics

Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire

Ryan O Emerson[1,4], William S DeWitt[1,2,4], Marissa Vignali[1], Jenna Gravley[3], Joyce K Hu[1], Edward J Osborne[1], Cindy Desmarais[1], Mark Klinger[1], Christopher S Carlson[3], John A Hansen[3], Mark Rieder[1,5] & Harlan S Robins[1,2,5]

An individual's T cell repertoire dynamically encodes their pathogen exposure history. To determine whether pathogen exposure signatures can be identified by documenting public T cell receptors (TCRs), we profiled the T cell repertoire of 666 subjects with known cytomegalovirus (CMV) serostatus by immunosequencing. We developed a statistical classification framework that could diagnose CMV status from the resulting catalog of TCRβ sequences with high specificity and sensitivity in both the original cohort and a validation cohort of 120 different subjects. We also confirmed that three of the identified CMV-associated TCRβ molecules bind CMV *in vitro*, and, moreover, we used this approach to accurately predict the *HLA-A* and *HLA-B* alleles of most subjects in the first cohort. As all memory T cell responses are encoded in the common format of somatic TCR recombination, our approach could potentially be generalized to a wide variety of disease states, as well as other immunological phenotypes, as a highly parallelizable diagnostic strategy.

The ability of the adaptive immune system to adequately address an infection relies on the presence of T cells that generate αβ-heterodimeric antigen-specific TCRs through V(D)J recombination. TCR specificity is mediated by primary sequence diversity: each mature TCRβ gene is randomly rearranged at its complementarity-determining region 3 (CDR3) by combining noncontiguous variable (V), diversity (D), and joining (J) region gene segments of the germline locus. Nucleotide deletions and template-independent insertions at the V_β–D_β and D_β–J_β junctions further add to the diversity of the encoded receptors, resulting in highly diverse TCRβ CDR3 regions[1,2]. The TCRα chain is generated in a similar process and combines with the TCRβ chain to form a TCR that binds its cognate antigen in the context of specific cell surface major histocompatibility complex (MHC) class I proteins. These are encoded by the highly polymorphic human leukocyte antigen (HLA) loci *HLA-A*, *HLA-B*, and *HLA-C*, and a TCR's antigen specificity is therefore further modulated by HLA context. Upon antigen recognition, activated T cells proliferate by clonal expansion and some become part of the memory compartment, where they can reside for many years as clonal populations of cells with identical TCR rearrangements by virtue of their descent from a common naive T cell[3–5].

Healthy adults express approximately 10^7 unique TCRβ chains on their ~10^{12} circulating T cells, which are drawn from a much larger pool of possible rearrangements[5]. Observing the same TCRβ chain independently in two individuals is thousands of times more common than would be expected if all rearrangements were equally likely[6]. It is expected that many TCRβ sequences (especially those with few or no junctional insertions) are present in the naive T cell repertoires of most humans at any given time and the corresponding T cells will proliferate upon exposure to their target antigen in the proper MHC context[7]. Public T cell responses, in which a particular antigen is targeted by the same TCR sequence in multiple individuals, result when the space of potential high-avidity TCRβ chains that could bind to a particular antigen–MHC complex includes one or more TCRβ chains that also have a high likelihood of existing in the naive repertoire at any given time[7,8]. TCRβ sequences associated with a public T cell response to a particular antigen will only be intermittently observed in the naive compartment of subjects who have not been exposed to that antigen. However, T cells carrying these TCRβ sequences undergo clonal expansion upon antigen encounter, which increases the probability that these sequences will be detected in the repertoire of exposed subjects, thus providing the basis for characterizing immunological memory across different individuals.

Despite historical limitations on sequencing depth and the limited size of investigational cohorts, there are many examples of public T cell responses to infectious diseases, such as CMV, Epstein–Barr virus (EBV), *Clostridium tetani*, parvovirus, herpes simplex virus (HSV), HIV, and influenza, as well as in malignancies and autoimmunity[7,8]. Typically, these public T cell responses have been studied in the context of single antigens in a single HLA context by isolation

[1]Adaptive Biotechnologies, Seattle, Washington, USA. [2]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [3]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [4]These authors contributed equally to this work. [5]These authors jointly directed this work. Correspondence should be addressed to R.O.E. (remerson@adaptivebiotech.com).

Received 15 September 2016; accepted 28 February 2017; published online 3 April 2017; doi:10.1038/ng.3822

of specific antigen-bound T cells followed by low-throughput sequencing of the variable regions of the TCRβ (and, in some cases, the TCRα) chains.

With the goal of identifying statistically significant associations between sets of TCRβ sequences and phenotypes of interest, we immunosequenced the variable region of the TCRβ chain to generate sizable libraries of public TCRβ chains. By measuring the presence or absence of individual TCRβ sequences in a large investigational cohort and statistically assessing their concordance with phenotypes of interest, we demonstrate that CMV serostatus and the presence of particular *HLA-A* and *HLA-B* alleles can be predicted with high confidence solely on the basis of the TCRβ repertoire data generated from peripheral blood. We believe that our approach has the potential to be applied as a diagnostic strategy for a range of immune-related phenotypes.

RESULTS

Data acquisition

We immunosequenced the rearranged CDR3 TCRβ region in a cohort of 666 healthy bone marrow donors (cohort 1), generating a total of 89,840,865 unique TCRβ sequences, which are defined in this study as a unique combination of a V gene, a CDR3 amino acid sequence, and a J gene. This level of similarity does imply an identical TCRβ protein sequence, although, owing to HLA restriction and the potential to pair with different TCRα chains in different T cells, it does not guarantee that two such receptors will have identical antigen specificities. We observed an average of 192,515 (±80,630 s.d.) unique TCRβ sequences per subject; the level of variation observed between subjects is in line with the natural variation of T cell levels in the peripheral blood of healthy adults[9,10]. For our proof-of-principle study, we selected CMV, a chronic virus that has been extensively studied as a model system for public T cell responses[7,8,11] and which infects 30–90% of adults[12], thus providing high statistical power. We performed CMV serotyping for 641 subjects: 352 subjects were CMV negative (CMV^-) and 289 were CMV positive (CMV^+). For validation purposes, we also immunosequenced the CDR3 region of TCRβ in an independent cohort of 120 subjects (cohort 2), which resulted in an average of 202,918 (±108,603) unique TCRβ sequences per subject, and we performed CMV serotyping for all subjects in this cohort. Finally, we performed HLA typing for 626 subjects in cohort 1 to obtain *HLA-A* and *HLA-B* major allele calls. Demographic data for both cohorts are summarized in **Table 1** and **Supplementary Table 1**.

Table 1 Cohort demographics

		Cohort 1	Cohort 2
Sex	Female	297	73
	Male	345	47
	Unknown	24	0
Age (in years)	≤10	22	0
	11–20	25	7
	21–30	87	90
	31–40	141	23
	41–50	155	0
	51–60	93	0
	>60	32	0
	Unknown	111	0
Ancestry	American Indian or Alaska Native, not Hispanic or Latino	9	1
	Asian, not Hispanic or Latino	17	25
	Asian, Hispanic or Latino	0	1
	Black or African American, not Hispanic or Latino	8	3
	Native Hawaiian or other Pacific Islander, not Hispanic or Latino	3	0
	White, not Hispanic or Latino	377	84
	White, Hispanic or Latino	0	3
	Other, not Hispanic or Latino	0	2
	Other, Hispanic or Latino	0	1
	Unknown, Hispanic or Latino	26	0
	Unknown	226	0
CMV status	Positive	289	51
	Negative	352	69
	Unknown	25	0
HLA status	Known	626	0
	Unknown	40	120
Total		666	120

Sex, age, ancestry, known CMV status, and known HLA status are indicated, when available, for the 666 subjects in cohort 1 and the 120 subjects in cohort 2.

Identification of CMV-associated TCRs

To determine whether CMV-associated TCRβ sequences could be identified by their differential incidence among CMV^+ and CMV^- subjects, we followed the experimental and analytical procedure outlined in **Figure 1**. After immunosequencing peripheral blood samples, we analyzed each unique TCRβ chain identified for the 641 subjects in cohort 1 with known CMV serostatus for statistically significant enrichment among CMV^+ subjects in comparison to CMV^- subjects. At a significance threshold of $P < 1 \times 10^{-4}$ (established as optimal via a cross-validation procedure) and a false discovery rate (FDR) of 0.14 (estimated by permutation of CMV status), we identified 164 CMV-associated TCRβ chains (**Supplementary Table 2**) that displayed increased incidence among CMV^+ subjects in cohort 1 in comparison to CMV^- subjects (**Fig. 2a**). The statistical approach used to identify these CMV-associated TCRβ chains is described in detail in the Online Methods.

Next, we investigated the HLA restriction of each of these 164 CMV-associated TCRβ sequences. As TCR–antigen binding occurs in the context of MHC proteins, which are encoded by the highly polymorphic HLA loci, the affinity of a given TCR for a given antigen is modulated by the HLA type of the subject[2]. Hence, for each CMV-associated TCRβ and each *HLA-A* and *HLA-B* allele identified for the subjects in cohort 1, we performed a Fisher's exact test[13] to assess the significance of the enrichment of that TCRβ among subjects positive for that HLA allele. At a significance threshold of $P < 1 \times 10^{-4}$, 45 of the 164 CMV-associated TCRβ chains identified in this study were associated with at least one *HLA-A* or *HLA-B* allele (**Supplementary Table 2**). Several TCRs were significantly associated with both an *HLA-A* and an *HLA-B* allele, and no TCRs were associated with more than one *HLA-A* or *HLA-B* allele.

Several studies have identified TCRs that recognize CMV antigens through low-throughput, *in vitro* methods—mainly tetramer sorting of T cells with particular CMV epitopes followed by sequencing of a portion of the TCRβ and/or TCRα locus—and report both private and public TCR sequences. **Supplementary Table 3** includes a list of 1,054 TCRβ sequences, 917 of which are unique[14–47], that have been reported in the literature as being able to recognize CMV antigens (termed 'CMV reactive') from 34 such publications. Most of these TCRβ sequences were identified on the basis of their reactivity to a small subset of well-defined epitopes from the CMV pp65 or E11 proteins, although a few of the studies report TCRβ sequences that recognize collections of overlapping peptides from these and other CMV proteins. Many of these publications include HLA restriction information.

We first compared the list of 917 CMV-reactive TCRβ sequences to the cumulative list of 89,840,865 unique TCRβ sequences observed

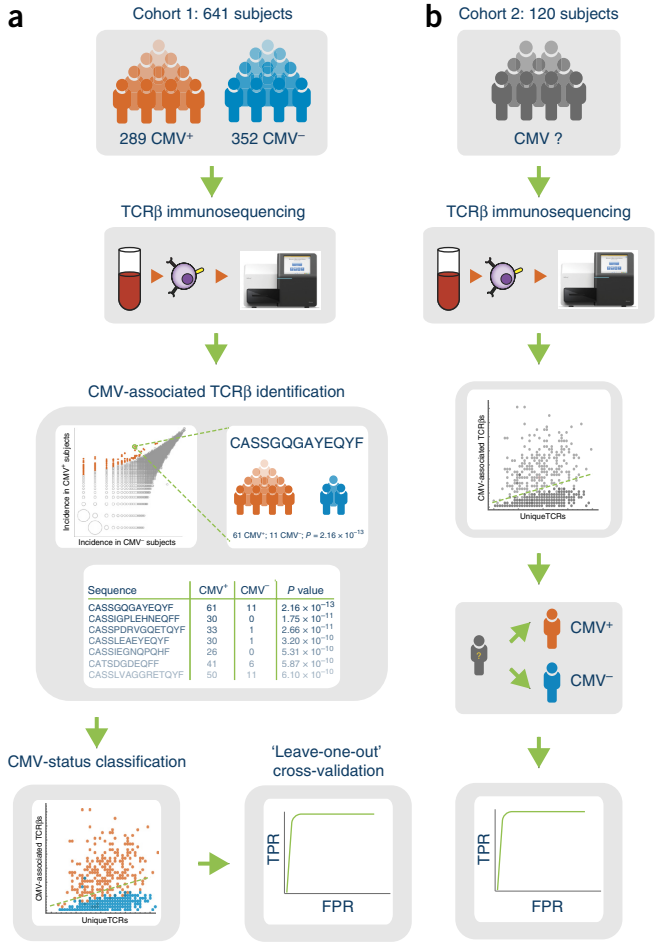

Figure 1 Experimental and analytical overview. (**a**) For cohort 1, we obtained peripheral blood samples from 641 subjects with known CMV serostatus and the immunosequenced variable region of the TCRβ locus. We then identified CMV-associated TCRβs, demonstrated that CMV+ subjects have more CMV-associated TCRβs in their repertoires than CMV− subjects, and used a leave-one-out cross-validation approach to assess the diagnostic potential of screening for public TCRβs. (**b**) For cohort 2, we obtained peripheral blood samples from 120 subjects and immunosequenced the variable region of the TCRβ locus in the same manner. Next, using the model resulting from analysis of the first cohort, we inferred CMV serostatus for all subjects and compared these results to the experimentally determined CMV status a posteriori, thus validating the diagnostic accuracy of the method. TPR, true positive rate; FPR, false positive rate.

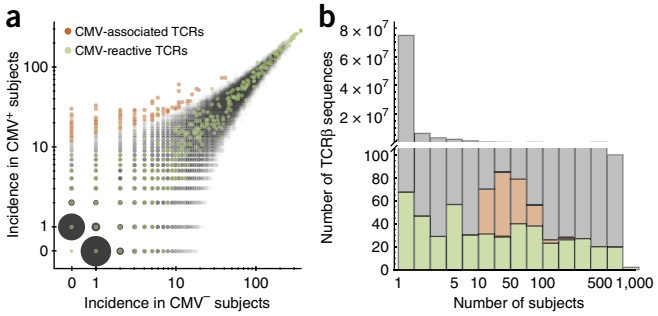

Figure 2 Identification of CMV-associated TCRβs. (**a**) Scatterplot showing the incidence of each TCRβ sequence among CMV+ and CMV− subjects in cohort 1. The number of TCRβs in each position is represented by both the transparency and size of the spot, with opacity and area mapped to the logarithm of frequency. CMV-associated TCRβ sequences identified in this study are represented by orange dots. Previously reported CMV-reactive TCRβs are represented by green dots. All other TCRβs are shown in gray. (**b**) Stacked bar chart depicting the distribution of the incidence of TCRβ sequences in subjects in cohort 1 (the number of subjects in whom each TCRβ is found). Previously reported CMV-reactive TCRβs span the whole range of incidence, whereas the CMV-associated TCRβs identified in this study occupy a narrower range. Colors correspond to those defined in **a**.

in the repertoires of the subjects in cohort 1: 488 of 917 (53.2%) sequences were observed in our cohort of 666 subjects. For each of the 488 CMV-reactive TCRβ sequences observed in our data set, we checked whether it had also been identified in this study as CMV associated (**Supplementary Table 2**) and identified 9 matches: 6 of these were significantly HLA restricted ($P < 1 \times 10^{-4}$) in our cohort, and in all 6 cases the HLA restriction we observed was consistent with previous reports (**Supplementary Table 4**). However, the vast majority of the CMV-reactive TCRβ sequences displayed similar incidence among CMV+ and CMV− subjects (**Fig. 2a**).

As the CMV-associated TCRβ sequences identified in this study had to fall within a narrow range of incidence to occur frequently in CMV+ subjects but rarely in CMV− subjects, thus disqualifying most truly public TCRβ sequences from being identified by this method (**Fig. 2b**), we also performed a more thorough analysis to investigate the overlap with the 488 CMV-reactive TCRβ sequences observed in our data set. In brief, we used a Mann–Whitney *U* test to determine whether these TCRβ sequences were significantly more abundant in CMV+ subjects than in CMV− subjects, restricting the analysis by HLA type for CMV-reactive TCRβ sequences with reported HLA restriction. For each TCRβ sequence, the fractional abundance in a subject was computed as the number of templates counted for that TCRβ divided by the total number of template TCRβ molecules in the sample. We then compared the list of fractional abundances in CMV+ subjects to that in CMV− subjects. This analysis combines both presence or absence (as, by definition, the abundance metric must be larger than zero), the relative abundance of the TCRβ in CMV− and CMV+ subjects (a total of 372 such TCRβ sequences were observed in at least one CMV+ and at least one CMV− subject), and HLA restriction (as the lists of CMV+ and CMV− subjects that contained a specific TCRβ sequence in their repertoire were restricted to those who possessed the HLA allele that restricted that CMV-reactive TCRβ in their genome). We identified six CMV-reactive TCRβ sequences that were significantly enriched in abundance among CMV+ subjects at a *P*-value threshold of 1×10^{-4}. Five of these six TCRβ sequences were present in both CMV+ and CMV− subjects with the given HLA restriction, and four of them were also present in our list of CMV-associated TCRβ sequences that were identified solely by the presence/absence criterion. Three of the six were reported as restricted to HLA-A2, and the other three were restricted to HLA-B7. By extension, five of our nine matches among CMV-associated TCRβ sequences were identified by the incidence approach but not by the HLA-aware abundance comparison. In conclusion, the approach based on the presence or absence of a TCRβ sequence in our data set is sufficient to recapitulate or even improve the results of an HLA- and abundance-aware analysis, and most CMV-reactive TCRβ sequences previously reported in the literature are not more likely to be observed in—or be more abundant in—CMV+ subjects in our cohort.

Inference of CMV serostatus from public TCRβ sequences

To control for differences in repertoire sampling depth, we plotted the number of CMV-associated TCRβ sequences found in each of the 641 CMV-serotyped subjects in cohort 1 versus the total number of

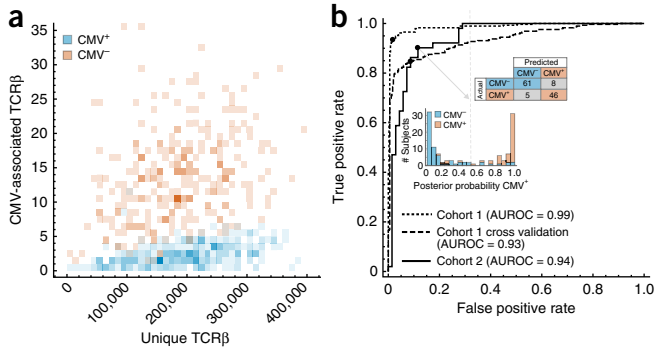

Figure 3 The incidence of CMV-associated TCRβs is diagnostic of CMV serostatus. (**a**) Scatterplot comparing the distribution of the number of CMV-associated TCRβs to the total number of unique TCRβs sampled for CMV+ (orange) and CMV− (blue) subjects in cohort 1. (**b**) ROC curves showing the classification performance of a classifier trained on data from cohort 1 and tested on all the data from cohort 1 (dotted line), by cross-validation in cohort 1 (dashed line), or by independent validation in cohort 2 (solid line). The black circle on each line corresponds to the maximum a posteriori (MAP) decision threshold. The inset depicts the MAP classification results for cohort 2 with the confusion matrix showing a sensitivity of 0.90 and a specificity of 0.89.

unique TCRβ sequences observed in that subject (**Fig. 3a**). Although there was a clear separation between CMV+ and CMV− subjects, the number of CMV-associated TCRβ sequences increased with increased sampling depth for both CMV+ and CMV− subjects, as expected. We hypothesized that the presence of many CMV-associated TCRβ sequences in a subject would be diagnostic of CMV seropositivity. To test this hypothesis, we constructed a generative binary classifier that infers CMV serostatus from the number of CMV-associated TCRβ sequences. The general framework for inferring subject-level phenotypes from immunosequencing data is described in the Online Methods. We trained this classifier on the CMV-serotyped subjects from cohort 1, tuning the *P*-value threshold for finding CMV-associated TCRβ sequences while simultaneously estimating the performance of our classifier using exhaustive 'leave-one-out' cross-validation. The performance of this classifier, as measured by the area under the receiver operating characteristic curve (AUROC; **Fig. 3a**), was 0.99 for all data (dotted line) and 0.93 for the cross-validation data set (dashed line). Although the first model was overfitted, the robust performance observed in the cross-validation approach is encouraging.

To fully validate our method for inference of an individual's CMV serostatus, we collected peripheral blood samples from an independent set of 120 subjects who were not used to train the classifier (cohort 2), and we performed immunosequencing and CMV serotyping as described (**Table 1** and **Fig. 1b**). We observed very strong discrimination performance for our classifier on this validation cohort, as indicated by an AUROC value of 0.94 (**Fig. 3b**, solid line). This value is no less accurate than the one obtained from the cross-validation approach described above, suggesting that the cross-validation scheme was a fair estimate of model accuracy. The application of a maximum a posteriori decision threshold for CMV serostatus classification in cohort 2 resulted in a sensitivity of 0.90, a specificity of 0.88, and a diagnostic odds ratio of 70. Taken together, these results demonstrate the excellent diagnostic power of our approach.

In vitro confirmation of CMV antigen specificity

In order to confirm that the approach described above can identify TCRβ sequences from TCRs that recognize CMV antigens, we performed a MIRA (multiplexed identification of TCR antigen specificity) assay on a healthy HLA-A2+ adult[29,48]. We screened approximately 200 million peripheral blood mononuclear cells (PBMCs) to identify TCRβ sequences specific for any of 38 characterized HLA-A2-restricted peptide antigens, including CMV pp65 peptide NLVPMVATV (amino acids 495–503) and 37 unrelated epitopes (**Supplementary Table 5**).

Briefly, we found 1,840 antigen-reactive TCRβ sequences in total, of which 69 (3.75%) were specific for CMV peptide CMV-pp65$_{(495–503)}$ and 1,771 were specific for the other non-CMV-derived epitopes (**Supplementary Table 6**). Of these 1,840 TCRβ sequences and 69 CMV-specific TCRβ sequences, 652 and 32, respectively, were observed in any of the subjects comprising cohort 1 in this study—consistent with the hypothesis that most TCRs are private.

When comparing these results with the list of 164 CMV-associated TCRβ sequences generated by our association study in cohort 1, we found three TCRβ sequences in common with the MIRA assay results; all three were specific for CMV in the MIRA experiment (3/3 overlapping TCRβ sequences specific for CMV, while 69/1,840 were CMV specific overall: $P = 5.3 \times 10^{-5}$ by binomial test). Of these three, two (CASSLAPGATNEKLFF and CASASANYGYTF) were associated with HLA-A2 in both our study and previous studies[32,36,41–43]. Thus, of the four CMV-associated TCRβ chains that were confidently assigned to HLA-A2 in our study (**Supplementary Table 2**), half were seen in an experimental results derived from a single individual's T cells (**Supplementary Table 6**).

Supplementary Figure 1 shows the prevalence of each of the 652 MIRA TCRβ sequences found in any subject from cohort 1 within CMV+ and CMV− individuals. Of the 32 CMV-reactive TCRβ sequences also seen in cohort 1, some of which had high prevalence in the data, most appeared equally in CMV+ and CMV− subjects. In addition to this incidence analysis, **Supplementary Table 5** includes the results of a Mann–Whitney *U* test (with Bonferroni correction for 32 hypotheses) assessing clonal abundance in CMV+ and CMV− subjects. Aside from the three TCRβ sequences already identified as CMV associated by incidence alone, this analysis identified one additional TCRβ that was more abundant in CMV+ subjects (CASSSANYGYTF). In conclusion, using this antigen–TCR matching assay, we have demonstrated that (i) several of the TCRβ sequences we identified as CMV associated using an association study do react to CMV epitopes *in vitro* and (ii) most TCRβ sequences we identified as reacting to CMV epitopes *in vitro* do not appear to be more prevalent or more abundant in CMV-seropositive individuals in an epidemiological sense.

Inference of HLA type from public TCRs

As described above, we determined that our approach can find public TCRβ sequences that have strong statistical association with CMV serostatus and that many of these TCRβ sequences were restricted to particular HLA types in cohort 1. Thus, we hypothesized that, because HLA type plays a strong role in shaping the T cell repertoire[49], we should be able to find public TCRβ sequences associated with particular HLA alleles using a similar framework. For each *HLA-A* and *HLA-B* allele present in cohort 1, we performed an association analysis similar to that used for CMV status determination, using *HLA-A* and *HLA-B* allele presence as the Boolean phenotype of interest. Notably, the analyst who conducted this study was not made aware of the HLA status of the subjects until after the analysis was completed.

Both positive and negative thymic selection based on interactions between TCRs and MHC proteins are known to affect T cell fate and, consequently, the presence of specific T cells in an adult human's peripheral blood[49,50]. We therefore employed a two-tailed Fisher's

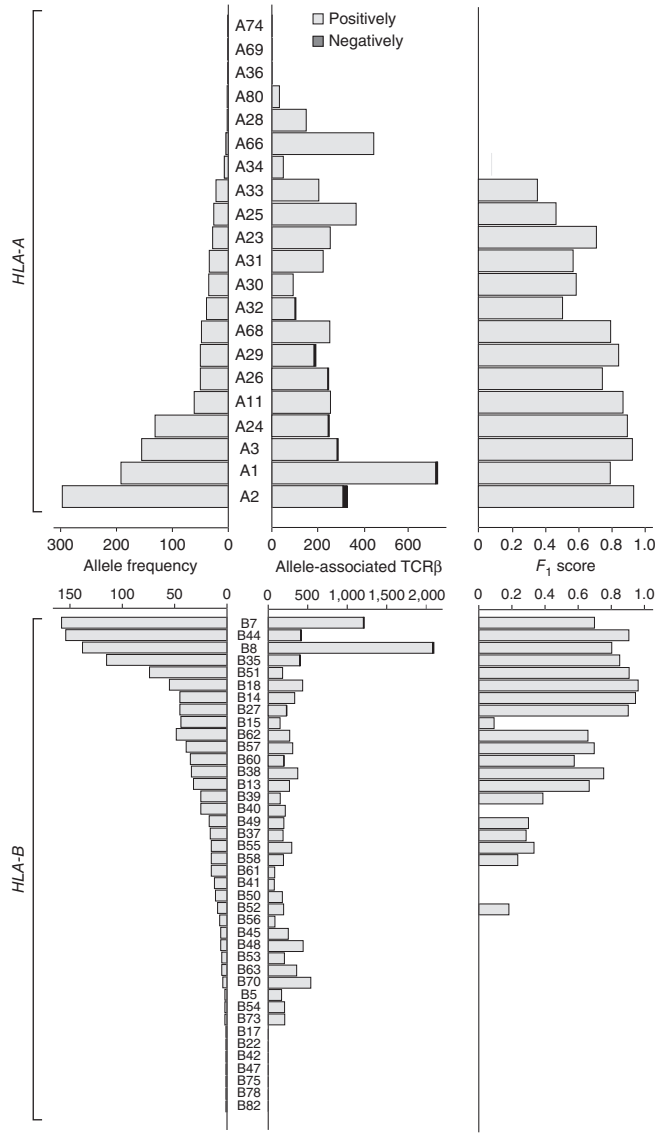

Figure 4 Identification of HLA-allele-associated TCRβs. For all *HLA-A* (top) and *HLA-B* (bottom) alleles observed in cohort 1, the graphs show the frequency of each allele (left), the number of allele-associated TCRβs identified (middle; the direction of the association is indicated as follows: light gray, positive association; dark gray, negative association), and the classification accuracy based on cross-validation (right; measured using F_1 score).

exact test to identify TCRβ sequences that were either positively associated with specific HLA alleles (i.e., enriched in subjects carrying that HLA allele) or negatively associated with specific HLA alleles (i.e., suppressed in subjects carrying that HLA allele). This analysis resulted in the identification of 15,601 HLA-allele-associated TCRβ sequences for the 61 *HLA-A* and *HLA-B* alleles observed in cohort 1, with all but 87 having positive associations in our data (**Fig. 4**, left and middle).

Next, for each allele, we performed exhaustive leave-one-out cross-validation to assess the ability of our classification framework to successfully infer the presence of HLA alleles by using a model that performs allele-associated TCRβ identification and classifier training for each subject, including all of the subjects other than the one under consideration. This approach is described in detail in the Online Methods. Classification was highly sensitive and specific for the more common alleles, with accuracy (measured by the F_1 score) diminishing with decreasing allele frequency (**Fig. 4**, right).

Finally, we inferred an HLA type for each subject, consisting of the set of HLA alleles whose individual classification model suggested that the allele was present in that subject, without explicitly enforcing homo- or heterozygosity at each locus. **Supplementary Table 2** includes the resulting HLA type inferences: of 626 HLA-typed subjects in cohort 1, 332 subjects were assigned cross-validated *HLA-A* allele inferences that exactly matched their known *HLA-A* alleles, 234 subjects were assigned exactly matched *HLA-B* allele inferences, and 138 subjects had both exactly matched *HLA-A* and *HLA-B* allele inferences.

These results demonstrate, in principle, the feasibility of HLA typing by immunosequencing. HLA typing was achieved for most HLA alleles in the data set, including those observed at lower frequencies. The inference of rare HLA types would require the acquisition of more data so that those alleles and their associated TCRβ sequences would be better represented in the data set. Also, despite our initial hypothesis that the TCR repertoire would be shaped by both positive and negative thymic selection, we observed very few TCRβ sequences negatively associated with HLA alleles. Although this observation warrants further study, it could suggest that very few TCRβ sequences recombine frequently and are reliably deleted by negative selection. The relative contributions of positive and negative selection to the shaping of the TCR repertoire remain unclear. For example, the contribution of positive and negative intrathymic selection to the peripheral T cell repertoire might depend on the avidity of thymocyte interactions with selecting endogenous peptide–MHC (pMHC) complexes and their structural relationships with pMHC encountered in the periphery[51], and it has been shown that expression of the MHC-restricting allele has, at most, a mild impact on T cell frequencies at the level of thymic selection, has barely an effect on homeostatic periphery expansion, and varies for different antigens[52].

DISCUSSION

Using a very large set of immunosequencing data from over 650 healthy subjects, we performed a high-throughput screen for public TCRβ sequences whose presence is associated with CMV serostatus or with particular *HLA-A* and *HLA-B* alleles. First, we identified a set of 164 TCRβ sequences that could be used to correctly predict the CMV serostatus of subjects in the training data set (by using a leave-one-out cross-validation strategy) and, more importantly, of subjects from an unrelated cohort, with very high specificity and sensitivity. For a small set of these TCRβ sequences, we confirmed their specificity for CMV-pp65$_{(495–503)}$ *in vitro*, thus showing that, despite being blind to the biochemistry of actual TCR–MHC–antigen interactions, our statistical association method can find TCRβ sequences that bind CMV epitopes. We believe that this approach can be generalized to identify exposure to other pathogens, as they should imprint a similar signature in the TCR repertoire of exposed individuals. Future experiments will extend this approach to other pathogens and to vaccination. Only a few of the CMV-reactive TCRβ sequences previously reported in the literature were either seen more often in CMV$^+$ subjects or were more abundant in CMV$^+$ subjects than in CMV$^-$ subjects in our cohort. This result is not unexpected: first, it has been established that a large majority of T cell responses are private rather than public, so it is not surprising that only half of the previously reported CMV-reactive TCRβ sequences were observed in our study[43]. Second, for a TCRβ to be part of our list of CMV-associated clones, it must rearrange frequently enough to constitute

a public response to CMV but not so frequently that it appears in the TCRβ repertoires of many naive subjects, such that truly public TCRβ sequences present in most repertoires at high frequency would not have been identified by our study. Third, the TCRβ must belong to a TCR that does not react to any common stimulus other than CMV. T cell cross-reactivity is an essential characteristic of antigen recognition[53,54], with some estimates proposing that a single T cell can recognize up to a million different pMHC complexes[55], and thus it is possible that some of the previously reported TCRβ sequences could be cross-reactive with epitopes from other pathogens or other proteins. For example, CMV-reactive clones have been shown to recognize peptides derived from human HLA proteins in the context of allotransplant[56–58]. Finally, as we only profiled TCRβ, the apparent lack of overlap with previously reported TCRs could represent commonly rearranged TCRβ sequences that have a broad array of antigen specificities in different HLA contexts or that are paired with different TCRα chains, resulting in a different antigen specificity.

By extending our approach to HLA typing, we also showed that, for HLA alleles present at a high enough frequency in the training data set, we could correctly predict the *HLA-A* and *HLA-B* alleles of the majority of individuals in cohort 1. As more TCR sequencing data across thousands of subjects are accumulated and compiled, we expect that our approach will be useful in HLA typing all but the rarest of alleles. Therefore, low-resolution HLA typing may soon be an additional benefit of any immunosequencing experiment, achieved simply by consulting a database of known allele-associated TCR sequences.

Our study proves that particular immunological phenotypes (in particular, CMV infection and HLA type) qualitatively mold the T cell repertoire, leaving an imprint that can be read using immunosequencing. Because high-throughput sequencing of TCRs captures all T cell responses equally and because hosts store immunological memory in this common format regardless of the stimulus, we believe that reading T cell memory by looking for known public responses will be a viable strategy for simultaneously diagnosing a wide range of immunological conditions using a single peripheral blood sample and a simple, unified assay. While the initial effort of collecting and analyzing a large training cohort for each phenotype of interest may be substantial, our diagnostic approach has the unique and useful property that, following routine immunosequencing and once such a database of associated TCRs is available, the incremental effort required to test a sample for an additional phenotype is negligible. We believe that further exploration of this approach is warranted, both to refine our methodology and to include other applications, such as other chronic infections and autoimmune conditions, and to refine HLA inference. One caveat of our approach is that it will also detect TCRβ sequences specific for any condition that is highly correlated with CMV seropositivity. The use of this approach for the diagnosis of multiple infections simultaneously in the context of correlated seropositivity will require additional experiments and new methodologies.

In summary, using an *ab initio* approach and a very large data set with high power, we demonstrated that TCR sequencing can provide a sensitive and specific diagnostic. We expect that, once a sufficient number of TCR association studies have been completed to enable a highly multiplexed assay, immunosequencing will become a cost-competitive alternative to current diagnostic methods.

METHODS

Methods, including statements of data availability and any associated accession codes and 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

The authors would like to thank M. Chung and other technical staff in the Adaptive Biotechnologies immunosequencing laboratory for their work on this project, S. House for helping compile the list of CMV-reactive TCRβ sequences from the literature, and C. Linkem and K. Boland for assistance with sample tagging for the immuneACCESS project. This work was funded in part by an award from the W.M. Keck Foundation Medical Research Program to H.S.R. and C.S.C.

AUTHOR CONTRIBUTIONS

J.G. and J.A.H. obtained the DNA samples and determined the CMV status and HLA type of the subjects. R.O.E., C.S.C., M.R., and H.S.R. conceived and designed the experiments. M.R. generated the sequence data. R.O.E., W.S.D., M.V., and C.D. analyzed the results. R.O.E. and W.S.D. performed the statistical analyses. M.V. and C.D. performed the literature searches of CMV-specific TCRs. J.K.H., E.J.O., and M.K. performed and analyzed *in vitro* confirmation experiments. R.O.E., W.S.D., M.V., M.K., and H.S.R. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cabaniols, J.P., Fazilleau, N., Casrouge, A., Kourilsky, P. & Kanellopoulos, J.M. Most α/β T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J. Exp. Med.* **194**, 1385–1390 (2001).
2. Davis, M.M. & Bjorkman, P.J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
3. Arstila, T.P. *et al.* A direct estimate of the human αβ T cell receptor diversity. *Science* **286**, 958–961 (1999).
4. Neller, M.A., Burrows, J.M., Rist, M.J., Miles, J.J. & Burrows, S.R. High frequency of herpesvirus-specific clonotypes in the human T cell repertoire can remain stable over decades with minimal turnover. *J. Virol.* **87**, 697–700 (2013).
5. Robins, H.S. *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107 (2009).
6. Robins, H.S. *et al.* Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci. Transl. Med.* **2**, 47ra64 (2010).
7. Venturi, V., Price, D.A., Douek, D.C. & Davenport, M.P. The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8**, 231–238 (2008).
8. Li, H., Ye, C., Ji, G. & Han, J. Determinants of public T cell responses. *Cell Res.* **22**, 33–42 (2012).
9. Peters, R.E. & al-Ismail, S. Immunophenotyping of normal lymphocytes. *Clin. Lab. Haematol.* **16**, 21–32 (1994).
10. Reichert, T. *et al.* Lymphocyte subset reference ranges in adult Caucasians. *Clin. Immunol. Immunopathol.* **60**, 190–208 (1991).
11. Hanley, P.J. & Bollard, C.M. Controlling cytomegalovirus: helping the immune system take the lead. *Viruses* **6**, 2242–2258 (2014).
12. Gandhi, M.K. & Khanna, R. Human cytomegalovirus: clinical aspects, immune regulation, and emerging treatments. *Lancet Infect. Dis.* **4**, 725–738 (2004).
13. Fisher, R. On the interpretation of χ^2 from contingency tables, and the calculation of *P. J. R. Stat. Soc.* **85**, 87–94 (1922).
14. Arakaki, A. *et al.* TCR-β repertoire analysis of antigen-specific single T cells using a high-density microcavity array. *Biotechnol. Bioeng.* **106**, 311–318 (2010).
15. Babel, N. *et al.* Clonotype analysis of cytomegalovirus-specific cytotoxic T lymphocytes. *J. Am. Soc. Nephrol.* **20**, 344–352 (2009).
16. Brennan, R.M. *et al.* Predictable αβ T-cell receptor selection toward an HLA-B*3501-restricted human cytomegalovirus epitope. *J. Virol.* **81**, 7269–7273 (2007).
17. Brennan, R.M. *et al.* The impact of a large and frequent deletion in the human TCR β locus on antiviral immunity. *J. Immunol.* **188**, 2742–2748 (2012).
18. Day, E.K. *et al.* Rapid CD8+ T cell repertoire focusing and selection of high-affinity clones into memory following primary infection with a persistent human virus: human cytomegalovirus. *J. Immunol.* **179**, 3203–3213 (2007).
19. Dziubianau, M. *et al.* TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell–related pathology. *Am. J. Transplant.* **13**, 2842–2854 (2013).
20. Giest, S. *et al.* Cytomegalovirus-specific CD8+ T cells targeting different peptide/HLA combinations demonstrate varying T-cell receptor diversity. *Immunology* **135**, 27–39 (2012).
21. Hamel, Y. *et al.* Characterization of antigen-specific repertoire diversity following *in vitro* restimulation by a recombinant adenovirus expressing human cytomegalovirus pp65. *Eur. J. Immunol.* **33**, 760–768 (2003).
22. Hanley, P.J. *et al.* CMV-specific T cells generated from naïve T cells recognize atypical epitopes and may be protective *in vivo*. *Sci. Transl. Med.* **7**, 285ra63 (2015).
23. Heemskerk, M.H. *et al.* Efficiency of T-cell receptor expression in dual-specific T cells is controlled by the intrinsic qualities of the TCR chains within the TCR–CD3 complex. *Blood* **109**, 235–243 (2007).

24. Iancu, E.M. *et al.* Clonotype selection and composition of human CD8 T cells specific for persistent herpes viruses varies with differentiation but is stable over time. *J. Immunol.* **183**, 319–331 (2009).
25. Janbazian, L. *et al.* Clonotype and repertoire changes drive the functional improvement of HIV-specific CD8 T cell populations under conditions of limited antigenic stimulation. *J. Immunol.* **188**, 1156–1167 (2012).
26. Khan, N., Cobbold, M., Keenan, R. & Moss, P.A. Comparative analysis of CD8+ T cell responses against human cytomegalovirus proteins pp65 and immediate early 1 shows similarities in precursor frequency, oligoclonality, and phenotype. *J. Infect. Dis.* **185**, 1025–1034 (2002).
27. Khan, N. *et al.* Cytomegalovirus seropositivity drives the CD8 T cell repertoire toward greater clonality in healthy elderly individuals. *J. Immunol.* **169**, 1984–1992 (2002).
28. Klarenbeek, P.L. *et al.* Deep sequencing of antiviral T-cell responses to HCMV and EBV in humans reveals a stable repertoire that is maintained for many years. *PLoS Pathog.* **8**, e1002889 (2012).
29. Klinger, M. *et al.* Combining next-generation sequencing and immune assays: a novel method for identification of antigen-specific T cells. *PLoS One* **8**, e74231 (2013).
30. Koning, D. *et al. In vitro* expansion of antigen-specific CD8+ T cells distorts the T-cell repertoire. *J. Immunol. Methods* **405**, 199–203 (2014).
31. Liang, X. *et al.* A single TCRα-chain with dominant peptide recognition in the allorestricted HER2/neu-specific T cell repertoire. *J. Immunol.* **184**, 1617–1629 (2010).
32. Miconnet, I. *et al.* Large TCR diversity of virus-specific CD8 T cells provides the mechanistic basis for massive TCR renewal after antigen exposure. *J. Immunol.* **186**, 7039–7049 (2011).
33. Nakasone, H. *et al.* Single-cell T-cell receptor-β analysis of HLA-A*2402-restricted CMV-pp65-specific cytotoxic T-cells in allogeneic hematopoietic SCT. *Bone Marrow Transplant.* **49**, 87–94 (2014).
34. Nguyen, T.H. *et al.* Recognition of distinct cross-reactive virus-specific CD8+ T cells reveals a unique TCR signature in a clinical setting. *J. Immunol.* **192**, 5039–5049 (2014).
35. Peggs, K. *et al.* Characterization of human cytomegalovirus peptide–specific CD8+ T-cell repertoire diversity following *in vitro* restimulation by antigen-pulsed dendritic cells. *Blood* **99**, 213–223 (2002).
36. Price, D.A. *et al.* Avidity for antigen shapes clonal dominance in CD8+ T cell populations specific for persistent DNA viruses. *J. Exp. Med.* **202**, 1349–1361 (2005).
37. Retière, C. *et al.* Generation of cytomegalovirus-specific human T-lymphocyte clones by using autologous B-lymphoblastoid cells with stable expression of pp65 or IE1 proteins: a tool to study the fine specificity of the antiviral response. *J. Virol.* **74**, 3948–3952 (2000).
38. Scheinberg, P. *et al.* The transfer of adaptive immunity to CMV during hematopoietic stem cell transplantation is dependent on the specificity and phenotype of CMV-specific T cells in the donor. *Blood* **114**, 5071–5080 (2009).
39. Schub, A., Schuster, I.G., Hammerschmidt, W. & Moosmann, A. CMV-specific TCR–transgenic T cells for immunotherapy. *J. Immunol.* **183**, 6819–6830 (2009).
40. Schwele, S. *et al.* Cytomegalovirus-specific regulatory and effector T cells share TCR clonality—possible relation to repetitive CMV infections. *Am. J. Transplant.* **12**, 669–681 (2012).
41. Trautmann, L. *et al.* Selection of T cell clones expressing high-affinity public TCRs within Human cytomegalovirus–specific CD8 T cell responses. *J. Immunol.* **175**, 6123–6132 (2005).
42. van Bockel, D.J. *et al.* Persistent survival of prevalent clonotypes within an immunodominant HIV gag-specific CD8+ T cell response. *J. Immunol.* **186**, 359–371 (2011).
43. Venturi, V. *et al.* TCR β-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. *J. Immunol.* **181**, 7853–7862 (2008).
44. Wang, G.C., Dash, P., McCullers, J.A., Doherty, P.C. & Thomas, P.G. T cell receptor αβ diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection. *Sci. Transl. Med.* **4**, 128ra42 (2012).
45. Weekes, M.P., Wills, M.R., Mynard, K., Carmichael, A.J. & Sissons, J.G. The memory cytotoxic T-lymphocyte (CTL) response to human cytomegalovirus infection contains individual peptide-specific CTL clones that have undergone extensive expansion *in vivo*. *J. Virol.* **73**, 2099–2108 (1999).
46. Weekes, M.P., Wills, M.R., Sissons, J.G. & Carmichael, A.J. Long-term stable expanded human CD4+ T cell clones specific for human cytomegalovirus are distributed in both $CD45RA^{high}$ and $CD45RO^{high}$ populations. *J. Immunol.* **173**, 5843–5851 (2004).
47. Wynn, K.K. *et al.* Impact of clonal competition for peptide–MHC complexes on the CD8+ T-cell repertoire selection in a persistent viral infection. *Blood* **111**, 4283–4292 (2008).
48. Klinger, M. *et al.* Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS One* **10**, e0141561 (2015).
49. Goldrath, A.W. & Bevan, M.J. Selecting and maintaining a diverse T-cell repertoire. *Nature* **402**, 255–262 (1999).
50. 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).
51. Legoux, F. *et al.* Impact of TCR reactivity and HLA phenotype on naive CD8 T cell frequency in humans. *J. Immunol.* **184**, 6731–6738 (2010).
52. Hesnard, L. *et al.* Role of the MHC restriction during maturation of antigen-specific human T cells in the thymus. *Eur. J. Immunol.* **46**, 560–569 (2016).
53. Gras, S. *et al.* A structural voyage toward an understanding of the MHC-I-restricted immune response: lessons learned and much to be learned. *Immunol. Rev.* **250**, 61–81 (2012).
54. Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol. Today* **19**, 395–404 (1998).
55. Wooldridge, L. *et al.* A single autoimmune T cell receptor recognizes more than a million different peptides. *J. Biol. Chem.* **287**, 1168–1177 (2012).
56. Amir, A.L. *et al.* Allo-HLA reactivity of virus-specific memory T cells is common. *Blood* **115**, 3146–3157 (2010).
57. Burrows, S.R., Khanna, R., Burrows, J.M. & Moss, D.J. An alloresponse in humans is dominated by cytotoxic T lymphocytes (CTL) cross-reactive with a single Epstein–Barr virus CTL epitope: implications for graft-versus-host disease. *J. Exp. Med.* **179**, 1155–1161 (1994).
58. Rist, M., Smith, C., Bell, M.J., Burrows, S.R. & Khanna, R. Cross-recognition of HLA DR4 alloantigen by virus-specific CD8+ T cells: a new paradigm for self-/nonself-recognition. *Blood* **114**, 2244–2253 (2009).

ONLINE METHODS

Experimental cohort and study approval. For cohort 1, human peripheral blood samples were obtained from the Fred Hutchinson Cancer Research Center Research Cell Bank biorepository of healthy bone marrow donors. Donors underwent routine HLA typing and CMV serostatus testing at the time the samples were taken. For cohort 2, human peripheral blood samples were taken from healthy volunteers under a protocol to examine past and present exposure to infectious agents. In both cases, protocols were approved and supervised by the Fred Hutchinson Cancer Research Center Institutional Review Board, following written informed consent.

Immunosequencing. Genomic DNA was extracted from peripheral blood samples using the Qiagen DNeasy Blood Extraction kit (Qiagen). The CDR3 region of rearranged TCRβ genes, defined according to IMGT[59], was amplified and sequenced using previously described protocols[5,60]. Briefly, a multiplexed PCR method that uses a mixture of 60 forward primers specific to TCR V_β gene segments and 13 reverse primers specific to TCR J_β gene segments was employed. An average of 2.5 μg of input DNA was used for each sample (range, 0.5–3.5 μg). Reads of 87 bp were obtained using the Illumina HiSeq system. Raw HiSeq sequence data were preprocessed to remove errors in the primary sequence of each read and to compress the data. To remove both PCR and sequencing errors, a nearest-neighbor algorithm was used to collapse the data into unique sequences by merging closely related sequences.

Identification of phenotype-associated TCRs and inference of phenotype status. We developed a statistical learning framework for the identification of TCRs associated with particular subject phenotypes, as well as for the inference of phenotype status in novel subjects. In this study, the phenotypes analyzed were CMV status and presence of *HLA-A* and *HLA-B* alleles.

In brief, we performed an association analysis to identify a set of TCRβs that had significantly increased incidence among phenotype-positive subjects. We then defined a subject's phenotype burden as the number of these phenotype-associated TCRβs that were found among all unique TCRβs immunosequenced from that subject. Phenotype burden was modeled using a beta binomial likelihood and separately trained on phenotype-negative and phenotype-positive subjects (modeling a common beta prior for all phenotype-negative subjects and a common beta prior for all phenotype-positive subjects). The probability that a novel subject was phenotype positive or negative was taken as the posterior probability using the trained likelihoods and priors estimated from phenotype prevalence in the training data. Phenotype status inference was taken as the maximum a posteriori estimate (i.e., the phenotype status that maximizes this posterior probability). Each of these steps is described in more detail below.

Immunosequencing of a peripheral blood sample from each subject in both cohorts yielded a list of unique TCRβ CDR3 sequences (cohort 1 mean = 192,515 ± 80,630, cohort 2 mean = 202,918 ± 108,603) identified by a V gene, a J gene, and the amino acid sequence of the CDR3. Our learning problem in its raw form consists of Boolean phenotype and TCR presence data (**Supplementary Fig. 2a**).

For each of N subjects, the phenotype is indicated as present (1) or absent (0), and each of the M TCRβs is indicated as present or absent in each subject. The training cohort in this study consisted of $N = 666$ subjects, from which $M = 89{,}840{,}865$ unique TCRβs were identified. We wished to define a binary classifier, train on these data, and infer the classes (phenotype statuses) of novel subjects from TCRβ immunosequencing data.

Owing to the stochastic nature of the V(D)J recombination process, the set of possible TCRβs is extremely large, and any TCRβ repertoire (i.e., each immunosequencing sample) very sparsely occupies this space. Also, many TCRβs from a novel subject are expected to be novel with respect to the M unique TCRβs identified in any training cohort. An important complication is that the binding affinity of a given TCR for a given peptide antigen is modulated across individuals by HLA type. Therefore, the features relevant for discrimination of phenotype status will segregate according to latent HLA variables. We introduce a feature selection and dimensionality reduction approach that attempts to accommodate the idiosyncrasies of immunosequencing data with a minimum of model complexity.

Identification of phenotype-associated TCRs. For each of the M features (TCRβs) present in the training data with a subject incidence of at least two, we assess the significance of association with class assignment (phenotype status) by performing Fisher's exact test[13] on a 2 × 2 contingency table, counting the number of subjects in each class according to the presence and absence of the TCRβin question (**Supplementary Fig. 2b**). Defining a rejection region by setting a maximum P value, we identify a set of phenotype-associated TCRβs. **Figure 2a** shows the incidence of all TCRβs from cohort 1 among CMV^+ and CMV^- subjects and highlights significantly CMV-associated TCRs.

To identify CMV-associated TCRβs, we performed a one-tailed Fisher's exact test to identify TCRβs enriched in samples from CMV^+ subjects, presumably owing to their specificity for CMV antigen and clonal expansion following activation. To identify HLA-allele-associated TCRβs, we used a two-tailed Fisher's exact test. Thymic positive selection is expected to favor specific TCRβs that promote MHC binding given an HLA context. Therefore, these TCRβs are enriched in allele-positive subjects. Conversely, thymic negative selection is expected to censor specific TCRβs that lead to excess TCR–MHC affinity in the same HLA context. These TCRβs are expected to be suppressed in allele-positive subjects.

Given a P-value threshold, the false discovery rate (FDR) among phenotype-associated TCRβs may be determined by permutation of class labels (shuffling the second column in **Supplementary Fig. 2a**) according to the method of Storey and Tibshirani[61]. Let r_0 denote the number of rejected null hypotheses (phenotype-associated TCRβs) in the unpermuted data. Suppose we perform b random permutations of the class labels and, for each permutation i, we perform significance tests for all TCRβs and find r_i rejected null hypotheses at the same significance threshold set for unpermuted data. Assuming that most null hypotheses are true (i.e., most TCRβs are not associated with the phenotype under study), the FDR may be approximated as follows.

$$\mathrm{FDR} \approx \frac{1}{b}\sum_{i=1}^{b}\frac{r_i}{r_0}$$

To determine FDR among CMV-associated TCRβs at various significance thresholds, $b = 100$ permutations were performed (**Supplementary Fig. 3**).

Quantification of phenotype burden on the TCR repertoire. Having identified a catalog of phenotype-associated TCRβs as described above, we can quantify the phenotype's burden on a subject's TCRβ repertoire by comparing this catalog to the list of TCRβs observed in the subject. Suppose immunosequencing data from subject i resulted in n_i unique TCRβs. Instead of a multidimensional feature representation that considers the incidence of each phenotype-associated TCRβ in this subject, we simply count the number k_i of unique TCRβs out of the total n_i that are in our catalog of phenotype-associated TCRβs. This transformation reduces feature space dimensionality to 2 and has a simple interpretation as a measure of how much of the TCRβ repertoire is devoted to the phenotype. We refer to this as 'phenotype burden'. **Figure 3a** depicts the distribution in this reduced feature space for CMV^+ and CMV^- subjects, which shows strong discrimination.

In the case of disease-exposure phenotypes, each of the phenotype-associated TCRβs is likely HLA restricted and power to identify phenotype association will vary with HLA allele frequency among subjects in the training cohort. Therefore, this dimensionality reduction eliminates the segregation of subjects into discriminative subspaces according to HLA type in exchange for variation in univariate discrimination power according to HLA type.

Modeling of phenotype burden. We approach the learning problem on the phenotype burden data by observing that the reduced feature vector may be interpreted as a binomial pair of k_i successes in n_i trials (as long as k_i is much less than the total number of phenotype-associated TCRβs, avoiding saturation). Instead of employing a standard discriminative binary classifier on the k_i/n_i ratio, we preserve sampling depth information by constructing a generative model for k_i conditioned on n_i and class assignment. Let $c_i \in \{0,1\}$ denote the class assignment of subject i. The probability that any unique TCRβ in the repertoire of subject i is associated with the phenotype is modeled as a binomial proportion p_i. We suppose that the p_i values are independently and identically beta-distributed random variables within each class as follows.

$$p_i \sim \mathrm{Beta}\left(\alpha_{c_i}, \beta_{c_i}\right)$$

doi:10.1038/ng.3822

With conditioning on class assignment, we thus have the beta binomial distribution for subject i

$$p(k_i|n_i,c_i)=\binom{n_i}{k_i}\frac{\mathrm{B}(k_i+\alpha_{c_i},n_i-k_i+\beta_{c_i})}{\mathrm{B}(\alpha_{c_i},\beta_{c_i})}$$

where $\mathrm{B}(\cdot,\cdot)$ denotes the beta function. The parameters $\{\alpha_0, \beta_0\}$ and $\{\alpha_1, \beta_1\}$ parameterize the beta-distributed prior for phenotype-negative and phenotype-positive subjects, respectively. These may be determined by maximizing the joint likelihood over all subjects.

$$p(k|n,c)=\prod_{i=0}^{N}p(k_i|n_i,c_i)=\left(\prod_{i:c_i=0}p(k_i|n_i,0)\right)\left(\prod_{i:c_i=1}p(k_i|n_i,1)\right)$$

Forming the log likelihood, dropping terms dependent only on the data, and exploiting parameter separability, we maximize

$$\ell_l(\alpha,\beta)=-N_l\log\,\mathrm{B}(\alpha,\beta)+\sum_{i:c_i=l}\log\mathrm{B}(k_i+\alpha,n_i-k_i+\beta),\ l=0,1$$

where N_l denotes the number of subjects with class l. These objective functions have gradients

$$\begin{aligned}\frac{\partial\ell_l}{\partial\alpha}&=-N_l\left(\psi(\alpha)-\psi(\alpha+\beta)\right)+\sum_{i:c_i=l}\left(\psi(k_i+\alpha)-\psi(n_i+k_i+\alpha+\beta)\right)\frac{\partial\ell_l}{\partial\beta}\\&=-N_l\left(\psi(\beta)-\psi(\alpha+\beta)\right)\\&\quad+\sum_{i:c_i=l}\left(\psi(n_i-k_i+\beta)-\psi(n_i+k_i+\alpha+\beta)\right)\end{aligned}$$

where $\psi(\cdot)$ denotes the digamma function. Standard numerical gradient ascent methods were used to determine the class-wise beta priors.

$$\{\alpha_l,\beta_l\}=\underset{\{\alpha,\beta\}\in\mathbb{R}_+^2}{\operatorname{argmax}}\ \ell_l(\alpha,\beta),\ l=0,1$$

Laplace smoothing of the most deeply sampled subject in each class (largest n_i) was used to regularize the likelihood. Densities for estimated priors from the CMV data are shown in **Supplementary Figure 4**.

Having determined these likelihood parameters from a joint model for all training subjects, we now consider a novel subject with the phenotype burden k', n'. Approximating class priors from Laplace-regularized class counts in the training data, the posterior probability of each class assignment for the novel subject is

$$p(c'=x|n',k')=\binom{n'}{k'}\frac{\mathrm{B}(k'+\alpha_x,n'-k'+\beta_x)}{\mathrm{B}(\alpha_x,\beta_x)}\frac{N_x+1}{N+2},\ x=0,1$$

We form the log-posterior odds ratio for class assignment as

$$\begin{aligned}F(k',n')&=\log p(c'=1|k',n')-\log p(c'=0|k',n')\\&=\log(N_1+1)-\log(N_0+1)+\log\mathrm{B}(\alpha_0,\beta_0)\\&\quad-\log\mathrm{B}(\alpha_1,\beta_1)+\log\mathrm{B}(k'+\alpha_1,n'-k'+\beta_1)\\&\quad-\log\mathrm{B}(k'+\alpha_0,n'-k'+\beta_0).\end{aligned}$$

A decision function is defined by a threshold θ on this quantity.

$$\hat{c}(k',n',\theta)=\begin{cases}0,\ F(k',n')\le\theta\\1,\ F(k',n')>\theta\end{cases}$$

The maximum a posteriori (MAP) classification corresponds to $\theta=0$.

$$\hat{c}(k',n',0)=\underset{x\in\{0,1\}}{\operatorname{argmax}}\ p(c'=x|n',k')$$

With the classifier now defined, we may address a model selection question. In identifying phenotype-associated TCRβs, we must apply a P-value threshold. This threshold is a hyperparameter for the classifier. To select an optimal P-value threshold, we perform exhaustive leave-one-out cross-validation in the training data over a range of P-value thresholds and assess the cross-entropy loss at each P value used. For each subject held out, we recompute phenotype-associated TCRβs at the given threshold (feature selection) and fit likelihood parameters ($\{\alpha_0, \beta_0\}$ and $\{\alpha_1, \beta_1\}$) using only the remaining subjects. We then use this classifier to estimate the class probabilities for the subject held out. Iterating this process over all subjects gives the cross-validated class probabilities for each subject in the training set.

Let $q_i(\phi)$ denote the probability that $c_i = 1$ under the classifier built with subject i held out from feature selection and training, and using a P-value threshold of ϕ for identifying phenotype-associated TCRβs. The average cross-entropy loss over all subjects using P-value threshold ϕ is defined as

$$L(\phi)=-\frac{1}{N}\sum_{i=1}^{N}\left[c_i\log q_i(\phi)+(1-c_i)\log(1-q_i(\phi))\right]$$

By computing this loss function over a discrete set of P values ϕ, we may approximate a minimum.

$$\hat{\phi}=\underset{\phi\in\ddot{\mathrm{O}}}{\operatorname{argmin}}\ L(\phi)$$

Cross-entropy loss is a preferable metric to classification error, as it utilizes the probabilistic information provided by the generative model rather than the decision function per se. Minimizing $L(\phi)$ is tantamount to maximizing a joint likelihood of the class vector c, where the marginal probability for each subject i is computed from a model where subject i was held out. In modeling CMV data, a logarithmic grid of P values was used, indicating $\hat{\phi}=10^{-4}$ as approximately optimal (**Supplementary Fig. 3**). Cross-entropy loss on this grid was also computed using a single model that was trained on all data. In this case, $L(\phi)$ monotonically decays with increasing ϕ, highlighting the importance of cross-validation to avoid overfitting to spuriously phenotype-specific TCRβs.

Assessing classification performance. To validate this learning framework, we acquired a second cohort of CMV-typed subjects as a testing cohort. To assess the performance of the method in classifying the presence of HLA alleles, leave-one-out cross-validation was performed. For each HLA allele, we held out each subject and built classifiers using the remaining subjects. HLA-allele-associated clones may be both positively and negatively associated. As the vast majority of clones were positively associated, we excluded these clones in our dimensionality reduction. Our model could be augmented from beta binomial to Dirichlet multinomial to separately count the incidence of TCRβs negatively and positively associated with HLA alleles, but little improvement in discrimination is to be expected, as negative associations were relatively rare. In this case, $\hat{\phi}=10^{-4}$ (from optimization on CMV data) was used as a fixed parameter, rather than optimized separately for each allele. This was necessary to avoid bias created by performing cross-validation for both model selection and model evaluation. Fixing this parameter is expected to provide a conservative assessment of classification performance, as the wide variation in HLA allele frequency entails variation in optimal P-value thresholds across alleles (which could be determined with more expensive nested cross-validation for both model selection and model evaluation). To call the HLA type of each subject, results from the classification on each allele were aggregated, with zygosity at each locus not explicitly enforced. Accuracy for these cross-validated HLA-type inferences is described in the Results section.

Overlap with the literature. We compiled an exhaustive list of CMV-reactive TCRβs from 34 publications. Most of the published CDR3s conformed to the standard nomenclature (i.e., they extended from a cysteine to a phenylalanine, or C–F), but some did not. For those that were longer, we trimmed to the C–F fragment before comparing to the list of TCRβs identified in this study; those that were shorter or incomplete were used as published. In our overlap analysis, a match consisted of either the published TCRβ or the TCRβ identified in this study constituting a substring of the corresponding sequence.

MIRA assays. Peptide-based MIRA was performed as previously described[29]. Briefly, approximately 200 million PBMCs from a single CMV$^+$ HLA-A2$^+$ donor were divided into 8 aliquots and incubated with different pools of 19 dextramers (Immudex) out of a total of 38 available dextramers. These included one dextramer with a CMV-derived epitope and 37 dextramers with non-CMV-derived peptides as a negative control to show the specificity of our assay. Dextramer-positive and dextramer-negative CD8$^+$ T cells were sorted and immunosequenced as described above. Antigen-specific TCRβ sequences were identified on the basis of the following criteria: first, we selected TCRβ sequences that were significantly enriched in the positive sorted population in a subset (3, 4 or 5) of the 8 initial PBMC aliquots on the basis of a binomial model of differential abundance[62]. Next, we examined observed occupancy by using a maximum-likelihood framework to discern the best possible antigen address given the clone enrichment patterns. Significantly associated clones were then identified using a likelihood-ratio test comparing the best occupancy hypothesis to a null model of no enrichment across all aliquots. TCRβ sequences responding to CMV-pp65$_{495–503}$ using MIRA were then compared to associated sequences identified using population-based sequencing.

Statistics. CMV-associated TCRβs were identified using a one-tailed Fisher's exact test[13]. For each TCRβ in the training cohort that was found in more than one subject, we formed a 2×2 contingency table counting the number of subjects in the training cohort according to CMV serostatus and presence of the TCRβ in question. HLA-allele-associated TCRβs were found similarly by using allele presence and presence of the TCRβ in question to generate a 2×2 contingency table. A two-tailed test was used to detect both positively and negatively associated TCRβs (corresponding to HLA modulation of positive and negative thymic selection, respectively). FDR was computed by permutation, and a *P*-value threshold of 1×10^{-4} (FDR = 0.14) was selected by cross-validation.

Enrichment of pp65-specific TCRs among overlaps between the MIRA assay results and our catalog of CMV-associated TCRβs was calculated as follows. We assumed that the population frequency of pp65-specific TCRβs among all TCRβs identified by the MIRA experiment was 3.75% (empirically, using 69 pp65-specific TCRβs out of 1,840 MIRA TCRβs overall). Noting that three of three hits to our CMV-associated TCR catalog were pp65 specific in the MIRA experiment, we performed a one-tailed binomial test assessing the hypothesis that more than two of three trials were successes given a frequency of 3.75%. This led to a *P* value for enrichment of 5.3×10^{-5}.

Data availability. All immunosequencing data underlying this study are freely available from https://doi.org/10.21417/B7001Z and can also be analyzed and downloaded from the Adaptive Biotechnologies immuneACCESS site at https://clients.adaptivebiotech.com/pub/Emerson-2017-NatGen.

59. Yousfi Monod, M., Giudicelli, V., Chaume, D. & Lefranc, M.P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. *Bioinformatics* **20** (Suppl. 1), i379–i385 (2004).
60. Carlson, C.S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
61. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
62. DeWitt, W.S. *et al.* Dynamics of the cytotoxic T cell response to a model of acute viral infection. *J. Virol.* **89**, 4517–4526 (2015).