Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins

Mikhail V. Pogorelyy[a,b], Anastasia A. Minervina[a], Maximilian Puelma Touzel[c], Anastasiia L. Sycheva[a], Ekaterina A. Komech[a,b], Elena I. Kovalenko[d], Galina G. Karganova[e,f], Evgeniy S. Egorov[a,b], Alexander Yu. Komkov[a,g], Dmitriy M. Chudakov[a,b,h,i], Ilgar Z. Mamedov[a,b], Thierry Mora[j,1,2], Aleksandra M. Walczak[c,1,2], and Yuri B. Lebedev[a,k,1,2]

[a]Department of Genomics of Adaptive Immunity, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, 117997 Moscow, Russia; [b]Department of Molecular Technologies, Pirogov Russian National Research Medical University, 117997 Moscow, Russia; [c]Laboratoire de Physique Théorique, CNRS, Sorbonne Université, École Normale Supérieure (PSL), 75005 Paris, France; [d]Department of Immunology, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, 117997 Moscow, Russia; [e]Laboratory of Biology of Arboviruses, Chumakov Institute of Poliomyelitis and Viral Encephalitides, 142782 Moscow, Russia; [f]Department of Virology, Sechenov First Moscow State Medical University, 119146 Moscow, Russia; [g]Laboratory of Cytogenetics and Molecular Genetics, Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, 117198 Moscow, Russia; [h]Center for Data-Intensive Biomedicine and Biotechnology, Skoltech, 121205 Moscow, Russia; [i]Central European Institute of Technology, Masaryk University, 62500 Brno, Czech Republic; [j]Laboratoire de Physique Statistique, CNRS, Sorbonne Université, Université Paris-Diderot, École Normale Supérieure (PSL), 75005 Paris, France; and [k]Biological Faculty, Moscow State University, 119991 Moscow, Russia

Edited by Rafi Ahmed, Emory University, Atlanta, GA, and approved October 17, 2018 (received for review June 5, 2018)

T cell receptor (TCR) repertoire data contain information about infections that could be used in disease diagnostics and vaccine development, but extracting that information remains a major challenge. Here we developed a statistical framework to detect TCR clone proliferation and contraction from longitudinal repertoire data. We applied this framework to data from three pairs of identical twins immunized with the yellow fever vaccine. We identified 600 to 1,700 responding TCRs in each donor and validated them using three independent assays. While the responding TCRs were mostly private, albeit with higher overlap between twins, they could be well-predicted using a classifier based on sequence similarity. Our method can also be applied to samples obtained postinfection, making it suitable for systematic discovery of new infection-specific TCRs in the clinic.

vaccination | high-throughput sequencing | twins | T cell receptor | RepSeq

The extremely diverse repertoire of T cell receptor (TCR) sequences allows the immune system to develop a specific response to almost any possible pathogen. In recent years, huge progress has been made in the deep profiling of TCR repertoires by high-throughput sequencing (RepSeq), allowing for the identification of millions of TCR sequences in a single experiment (1). TCR sequences, phenotypes, and the relative abundances of the corresponding T cell clones encode both the history of previous infections and the protection against yet unseen pathogens. However, despite recent large-scale efforts (2, 3), it is still impossible to predict systematically which antigen is recognized by a TCR with a given sequence. The most reliable method to identify antigen-specific TCR—MHC-multimer staining (4)—is restricted by the choice of HLA alleles (which are the most polymorphic genes in the human population) (5) and by the knowledge of immunodominant peptides. HLA-independent methods for quantitative monitoring of challenge-specific TCRs in vivo are still lacking.

Here we developed a methodology for identifying responding TCRβ clonotypes from time-dependent RepSeq data and applied it to yellow fever immunization: a classical model of acute viral infection in humans. The yellow fever vaccine (YFV strain 17D) is one of the most efficient and safe vaccines ever made (6). Because YFV 17D is a live attenuated virus, vaccination leads to viremia and very intense T cell and humoral responses (7). Tracking of activated T cell subsets such as CD8+ CD38+HLA–DR+ (7) and fluorescent MHC-tetramer staining (8) make it possible to quantitatively describe the kinetics of the T cell response to primary YFV 17D immunization. The response peaks around day 14 after immunization, when activated T cells responding to vaccination occupy 2% to 13% of the CD8+ subpopulation (7–9) and 3% to 4% of the CD4+ subpopulation (10–12). Several immunodominant peptides were identified, and the corresponding pMHC-multimers have made it possible to track YFV-specific T cells years and even decades after immunization (8, 12, 13). The only RepSeq study of yellow fever immunization available to our knowledge (14) reports thousands of CD8+ T cell clones expanding after yellow fever immunization and preferential recruitment of highly expanded CD8+ T cell clones to the memory subpopulation. However, the

Significance

T cells play a key role in the adaptive immune system. The broad repertoire of unique receptors expressed by T cells is in principle able to recognize a huge diversity of pathogens, but how to extract that information from blood samples remains unclear. By sequencing and analyzing the statistics of T cell receptors of subjects vaccinated against yellow fever, we identified vaccine-specific receptors that expanded following vaccination. We show that each individual has a unique response, which is similar yet across subjects in its sequence composition, with a slightly higher similarity between twins. Our method can be used in the clinic to track disease-specific T cell clones expanding or contracting after infection, vaccination, or therapy.

Author contributions: D.M.C., I.Z.M., T.M., A.M.W., and Y.B.L. designed research; M.V.P., A.A.M., M.P.T., A.L.S., E.A.K., E.I.K., G.G.K., E.S.E., A.Y.K., and I.Z.M. performed research; M.V.P., M.P.T., T.M., and A.M.W. contributed new reagents/analytic tools; M.V.P., A.A.M., M.P.T., T.M., and A.M.W. analyzed data; and M.V.P., A.A.M., T.M., and A.M.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequence reported in this paper has been deposited in the NCBI Sequence Read Archive (accession no. PRJNA493983). The code for the clonal expansion model has been deposited in GitHub, https://github.com/mptouzel/pogorelyy_et_al_2018.

[1]T.M., A.M.W., and Y.B.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: tmora@lps.ens.fr, awalczak@lpt.ens.fr, or lebedev_yb@ibch.ru.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1809642115/-/DCSupplemental.

clonal structure of the T cell immune response, how personalized this response is, and what the impact of genetic factors on the response is still remain poorly understood. Studying monozygous twins allows us to quantify the impact of genetic factors using a small cohort of donors (15–17).

In this study, we identified nearly 5,000 YFV 17D-responding TCRs in three pairs of identical twin donors, using RepSeq profiling of T cell repertoires at different timepoints analyzed with advanced statistical modeling. We validated the yellow fever specificity of expanded clones using three independent functional tests. The detailed analysis of the TCRβ sequences of the expanded clones showed both a highly personalized response and high sequence similarity across individuals, especially between twins. This convergence allowed us to develop a supervised classifier that predicted YFV-reactive TCRβ sequences with high specificity. Remarkably, the dynamics of clonal contraction in the month after the peak response specifically predicts the YFV-reactive TCRβ clonotypes. Thus, our methodology can be used during the postinfection period in the clinic to identify TCR sequences that recognized and responded to an acute infection of interest, even without prior knowledge of donor MHC alleles or pathogen epitopes.

Results

Detection of Significantly Expanded TCRβ Following YFV 17D Vaccination. Blood samples were obtained for three pairs of identical twin volunteers aged 20 to 23. We collected peripheral blood samples on five different timepoints (two before and three after immunization) with the live attenuated YFV 17D vaccine (see Fig. 1*A*). On each timepoint, we collected two biological replicates—two independent tubes of blood—to isolate bulk peripheral blood mononuclear cells (PBMCs) and another tube of blood to isolate CD4+ and CD8+ T cell subpopulations. Additional portions of blood were used for other subpopulation isolation (CD45RO+) and functional tests on several timepoints (see *Materials and Methods* for details). From each sample, cDNA libraries of TCRβ chains were prepared as previously described (18) and sequenced on the Illumina HiSeq platform.

We developed a Bayesian statistical framework that identifies clonotypes both significantly ($p < 0.05$) and strongly (fold-change > 32) expanded between different timepoints compared with the expected variability between replicates (see *Materials and Methods*). We reproduced the results with edgeR (19), a widely used method to analyze differential gene expression in RNA-sequencing experiments (see *SI Appendix*, Fig. S1). Since differential expression analysis software such as edgeR was developed for RNA-seq data, its applicability to RepSeq data cannot be assumed a priori. By explicitly modeling the difference between true clone size, sampled clone size, and read count through a two-step process, our model better captures the low-number pair-count statistics than single-step models used by existing methods (see *SI Appendix*, Fig. S2). Another advantage is that our Bayesian approach allows us to transparently assess the uncertainty of the estimated fold-change for each clone, by providing its complete posterior distribution (see *SI Appendix*, Fig. S3). In Fig. 1*B*, we show the number of TCRβ clonotypes identified as expanded with respect to day 0. In all donors, we observe many more expanded clonotypes between day 0 and day 15 than for any other pair of timepoints, despite variations between donors. In the following, we simply call "expanded clonotypes" TCRβ sequences that significantly increased in fraction between days 0 and 15. Clonotypes with a significantly higher frequency on the prevaccination timepoint relative to the vaccination timepoint (day 0 vs day −7, so actually corresponding to a contraction) were relatively few for all donors except for patient 1 (P1), whom we speculate was undergoing another transient immune response.

Most expanded clones were not detected before immunization, and often not even on day 7, due to their low frequency. We report expansion rates of up to 2,000- to 3,000-fold in 7 d, although this estimate is only a lower bound due to lack of detection before day 15 (see *SI Appendix*, Fig. S4).

The Majority of Expanded TCRβs Are YFV 17D-Specific. We hypothesized that most of the expanded clonotypes proliferated specifically in response to the YFV 17D vaccine. To check this hypothesis, three independent functional tests with subsequent TCRβ repertoire sequencing were performed on donor S1: (*i*) interferon (IFN)-gamma secretion assay after stimulation with YFV 17D vaccine; (*ii*) fluorescent sorting of the activated CD8+CD38+HLA–DR+ subset, which was reported to be largely YFV 17D-specific at the peak of the response (7); (*iii*) staining with a MHC-dextramer loaded with an immunodominant epitope. For *i* and *ii*, we considered the clonotype validated if it was enriched in the IFN-gamma positive (*i*) or CD8+CD38+HLA–DR+ (*ii*) fractions, by comparison with the bulk PBMC or CD8+ populations, respectively (using a one-sided exact Fisher test; see *SI Appendix*). We found that out

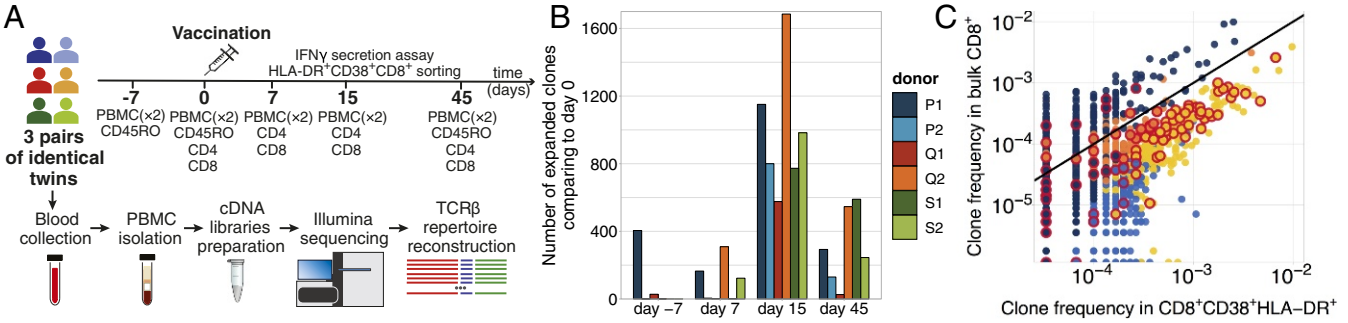

Fig. 1. (*A*) Yellow fever vaccination (YFV 17D) study design. (*Top*) Experimental timeline with the list of samples collected at each timepoint. Unsorted PBMC samples were collected in two biological replicates at each timepoint. (*Bottom*) Method overview. Peripheral blood samples were subjected to PBMC isolation, synthesis of TCRβ cDNA libraries, Illumina sequencing, and reconstruction of TCRβ repertoires. (*B*) Number of significantly expanded clonotypes in comparison with day 0. The number peaks at day 15 for all donors. The comparison with day −7 corresponds to a contraction reflecting the normal dynamics of a healthy repertoire in absence of vaccination. (*C*) Activated CD8+CD38+HLA–DR+ subpopulation is enriched with clonotypes expanded between days 0 and 15. Relative abundance of a TCRβ sequence in the CD8+CD38+HLA–DR+ activated subpopulation (*x* axis) versus its abundance in the bulk CD8+ population isolated at the same timepoint (*y* axis). Yellow dots indicate clonotypes that strongly expanded between day 0 and day 15. Brighter shades of blue and yellow indicate clonotypes significantly enriched in the activated subpopulation. Black line shows identity. Red circles indicate sequences found in the A02-NS4b$_{214-222}$-dextramer-positive fraction 2 y later.

of 773 clonotypes expanded in donor S1 (see Fig 1*B*), 331 were enriched in the CD8+CD38+HLA–DR+ fraction (see Fig. 1*C*) and 64 were enriched in the IFN-gamma secretion assay (see *SI Appendix*, Fig. S5). For *iii*, we isolated T cells positive for the HLA-A*02 dextramer loaded with the immunodominant YFV epitope $NS4b_{214-222}$ (LLWNGPMAV) from a sample collected 2 y after immunization of donor S1 and sequenced their TCRβ. We found 68 clonotypes in the dextramer-positive fraction that were labeled as expanded for this donor by our statistical method. Their cumulative frequency accounted for at least 22% of the CD8+ expanded fraction at day 15. We also sequenced unsorted PBMCs for the 2-y timepoint of this donor and found that the fraction of the repertoire occupied by the YFV-responding clones largely declined 2 y after immunization—from 2.1% (on day 45) to 0.07%—but still exceeded prevaccination levels by nearly 2 orders of magnitude.

The total number of clonotypes validated by any of the three methods was 395 out of 773. Notably, the remaining unvalidated clonotypes had significantly lower average frequencies than the validated ones on day 15 ($p < 10^{-16}$, t test; *SI Appendix*, Fig. S6). To check whether our method did not simply pick up large clones at day 15 (YFV-specific or not), we repeatedly subsampled 773 random TCRβs according to their abundance at day 15 and asked whether they were found in any of the validation datasets. On average, only 29.2 ± 0.2 of those were found in the CD8+CD38+HLA–DR+ subset (vs. 331 of the expanded clonotypes), 8.3 ± 0.1 in the IFN-gamma assay (vs. 64), and 6.4 ± 0.1 in the dextramer assay (vs. 68).

To further validate the expanded clones for YFV 17D specificity, we used previously published TCRβ sequences from the VDJdb database (20) (https://vdjdb.cdr3.net; see Dataset S1) specific for the $NS4b_{214-222}$ YFV 17D immunodominant epitope ($n = 264$) and an unrelated epitope as a control ($n = 370$): the cytomegalovirus (CMV) immunodominant HLA-A*02–restricted $pp65_{495-503}$ peptide (NLVPMVATV). All of our donors are HLA-A*02–positive (complete HLA genotypes can be found in *SI Appendix*, Table S1). We checked if published A02-$NS4b_{214-222}$–specific TCRβ sequences could be found in sets of expanded clonotypes identified in each donor by our model (*SI Appendix*, Table S2). We found multiple exact matches for published A02-$NS4b_{214-222}$–specific sequences and no exact matches for CMV-specific A02-$pp65_{495-503}$ sequences. We also found an increase of cumulative frequency of published YFV-specific (but not CMV-specific) TCRβs in CD8+ repertoires of our donors after immunization (see *SI Appendix* and *SI Appendix*, Fig. S7).

In summary, the expanded clonotypes were validated for their YFV specificity and were also consistent with published YFV-specific TCRs.

The CD8+ T Cell Response Is Sustained Longer than the CD4+ Response. To track the fractions of the CD4+ and CD8+ repertoires involved in the response over time, we calculated the cumulative frequency of expanded clonotypes in the CD4+ and CD8+ compartments at each timepoint. We found similar dynamics in all donors, with both CD8+ and CD4+ responses peaking on day 15. In all donors, the CD4+ response fell off more quickly than the CD8+ response from day 15 to day 45, with a mean $12.6(\pm 3)$-fold decrease for CD4+ versus $5.4(\pm 1.4)$ for CD8+ ($p = 0.027$, paired one-sided t test). We checked the significance of this difference in decay for each donor separately using our statistical framework (see *SI Appendix*). The responding clonotypes at the peak of the response occupied up to 8% of the repertoire (cumulative frequency) for CD8+ and up to 5% for CD4+ T cells. Almost all these clonotypes were undetected before immunization.

The previous analysis described the dynamics of the CD4+ and CD8+ responses separately. What is the relative importance and diversity of the CD4+ and CD8+ responses within the overall T cell population? To answer this question, we associated expanded clonotypes from the unsorted PBMC sample to either the CD4+ or CD8+ subsets in the following way: We labeled an expanded clonotype from the unsorted PBMC sample CD4+ if it had a larger concentration in the CD4+ subpopulation sequenced than in the CD8+ subpopulation and vice versa. This procedure gave unambiguous in silico phenotypes for each expanded clonotype sequence from the PBMC at day 15: The relative abundance of sequences in the CD4+ versus CD8+ compartments was strongly bimodal, with two peaks close to 100% CD4+ and 100% CD8+ (see *SI Appendix*, Fig. S8). This analysis revealed that both CD4+ and CD8+ clones strongly expanded in response to the vaccination, with no strong preference for CD4+ or CD8+ expanded clonotypes in terms of the cumulative frequency or diversity (see *SI Appendix*, Fig. S9). We next asked if the presence of expanded clonotypes in the memory (CD45RO+) compartment 45 d after immunization depends on their CD4+ or CD8+ phenotype. We detected on average 49% ($\pm 7\%$) of CD4+ expanded clonotypes in memory repertoires 45 d after immunization versus only 21% ($\pm 3\%$) of CD8+ expanded clonotypes. This may be explained by different levels of CD45RO + expression in CD4+ and CD8+ memory T cells reactive to yellow fever.

TCR Response Is Highly Personalized Even Among Twins. Our method not only reconstructs the dynamics of immune response but also enables the analysis of TCR sequences of the responding T cell clones. For each donor, we found 600 to 1,700 YFV 17D-reactive clonotypes expanded between day 0 and day 15. We compared the pairwise sharing of responding amino acid TCRβ sequences between different individuals. We found more overlap in twins than in nonrelated individuals (see Fig. 2*A*). However, absolute numbers of identical expanded TCRβ clonotypes were low even in twin pairs (up to 21 out of 1,685 amino acid sequences), indicating that each individual developed an almost unique response. We also observed that the same nucleotide variants encoded some of the amino acid TCRβ sequences shared between the twins (see *SI Appendix*, Table S3). This identity of nucleotide variants was more frequent in twins than expected from convergent recombination alone ($p < 0.014$; see *SI Appendix*).

Overall, shared sequences (across at least two donors) accounted for a small fraction of the response in each donor. Only 2.5% to 4.4% of unique responding TCRβ sequences were public (present in more than one donor). These sequences

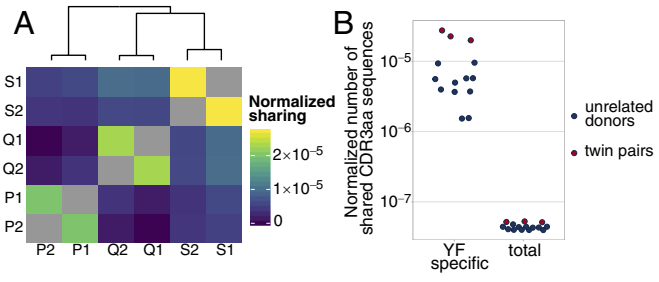

Fig. 2. Sharing of YFV-responding TCRβs across donors. (*A*) Number of expanded TCRβ amino acid clonotypes shared between pairs of donors, divided by the product of the numbers of expanded clonotypes in each donor. All three twin pairs (S, P, and Q) show higher numbers of shared expanded clonotype TCRβs than unrelated individuals and are clustered together by hierarchical clustering (dendrogram on top). (*B*) The normalized sharing of expanded clonotype TCRβs (*Left*) is much higher than the normalized sharing in the whole TCRβ repertoire (*Right*). Sharing in twins (red) always exceeds sharing in unrelated individuals (blue).

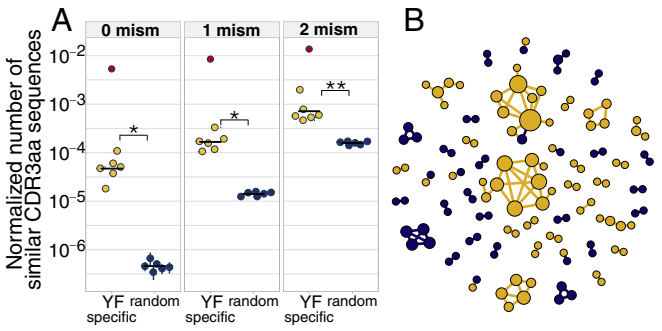

Fig. 3. Convergence of amino acid sequences in the YFV-responding TCRβ repertoire. (*A*) Number of pairs of similar clones (*Left*: exact same CDR3 amino acid sequence; *Middle*: up to one mismatch; *Right*: up to two mismatches) normalized by the number of possible pairings in each individual (see *SI Appendix*). The number of similar clones in the data (yellow dots) is larger than the number of similar clones in randomly drawn samples (dark-blue dots) of the same size ($p = 0.022,\ 0.015,\ 0.008$ for zero, one, and two mismatches; one-sided paired *t* test). As a reference, the red dots show an example of a restricted and specific repertoire for the yellow fever virus immunodominant epitope NS4b$_{214-222}$ [data from VDJdb (20); see Dataset S1]. (*B*) One thousand most abundant TCRs from donor S1 at day 15. Each vertex corresponds to a TCRβ amino acid sequence; edges connect sequences differing by two or fewer amino acids in their CDR3. Only vertices with neighbors are plotted. Yellow clonotypes indicate expanded clonotypes, while blue clonotypes were present before immunization at similar frequencies as on day 15. The vast majority of edges (95 out of 103) are formed between TCRβ clonotypes of the same status (expanded or not expanded).

correspond to 2.7% to 8.2% of YF-specific cells in all donors except Q1, which show an outlying value of 33% due to large expansion of few public TCRβ sequences (see *SI Appendix*, Table S4). Despite the low numbers of shared TCRβs, normalized sharing of YFV 17D-reactive amino acid clonotype sequences was more than 2 orders of magnitude higher than in the total TCRβ repertoire for both twins and unrelated individuals (Fig. 2*B*). This may result from convergent selection of the same TCRβ variants recognizing the same epitopes in different donors.

TCR Sequence Analysis Reveals a Mixture of Convergent and Private Response. It was previously shown that in many cases TCRs recognizing the same antigens have restricted sequence diversity (2, 21). To analyze the sequence diversity of responding TCRs, we performed a pairwise comparison of all expanded TCRβ sequences within each donor at the amino acid level. In each individual, we identified many more pairs of expanded TCRs with the same or highly similar CDR3β amino acid sequences (up to zero, one, or two amino acid mismatch) on day 15 than in a random subset of equal size (see Fig. 3*A*). Interestingly, our expanded TCRs were still more diverse than published TCRs selected for their specificity to a single immunodominant YFV epitope NS4b$_{214-222}$ [red dots; data from VDJdb (20); see Dataset S1], suggesting that the response is directed against multiple epitopes in each donor. Expanded TCRs form multiple dense clusters of highly similar sequences (Fig. 3*B*, yellow circles), suggesting they are responding to multiple epitopes. Further experiments need to test this conjecture. By contrast, prevaccination abundant TCRs form fewer, sparser, and smaller clusters (blue circles).

Sequence Score Based on Distance to Expanded TCR Predicts YFV 17D Specificity. The similarity of sequences in the expanded repertoire makes it possible to build a simple classifier to identify novel YFV 17D-specific TCRβ sequences. For each TCRβ of interest, we defined a YFV specificity score as the Hamming distance to the closest amino acid sequence neighbor among the expanded clonotypes from all six donors. To test how informative this score is about YFV specificity, we first applied it to published TCRs [from VDJdb (20); Dataset S1] specific to the NS4b$_{214-222}$ epitope from YFV 17D and to the pp65$_{495-503}$ epitope from CMV as a negative control. We found that TCRβs of published YFV-specific clonotypes were much closer to our set of expanded sequences than CMV-specific clonotypes (*SI Appendix*, Fig. S10*A*), with some exact matches for YFV-specific but none for CMV-specific sequences (*SI Appendix*, Table S2). Accordingly, using the score to discriminate YFV-specific from CMV-specific published sequences yields good specificity and sensitivity (*SI Appendix*, Fig. S10*B*).

We then asked if the score could be used to identify reactive clonotypes in the repertoire of an individual after immunization. To test this capability, we used a leave-one-out approach. Information about expanded clonotypes in five individuals was used to build a score as described earlier. The score was then used to predict expanded clonotypes among the 1,000 most abundant ones at day 15 in the sixth individual. The score performed with similar accuracy as on published epitope-specific clonotypes (*SI Appendix*, Fig. S10*C*).

Retrospective Detection of YFV 17D-Reactive TCRs Using Postvaccination Data. So far we have identified responding TCRs as those that significantly expanded between days 0 and 15. While prechallenge timepoints are easy to collect in vaccination studies, this is not the case for acute infections, where the first samples can usually be obtained only after the onset of symptoms, when it is too late to detect TCR expansion. However, the clonal contraction dynamics (Fig. 4, day 15 to day 45) can in principle be used to identify responding clonotypes, by comparing a timepoint taken on the peak of the response to a timepoint taken several weeks or months after the infection.

To demonstrate the feasibility of this detection on our data, we identified significantly ($p < 0.05$, fold-change > 1; see *Materials and Methods*) contracted clonotypes between day 15 and day 45 using our model. We computed the overlap between this set of candidate TCRβ clonotypes with the subset of expanded clonotypes obtained before (*SI Appendix*, Fig. S11). Strikingly, 74% to 97% of the significantly contracted TCRs were also present in the expanded subset, showing that the contraction dynamics can help to identify YFV-reactive clonotypes with high specificity. This method is also sensitive: 45% to 81% of expanded TCRs could be identified by contraction. This shows that contraction dynamics alone is sufficient to identify a large fraction of the responding TCRs. Thus, our method could be used to

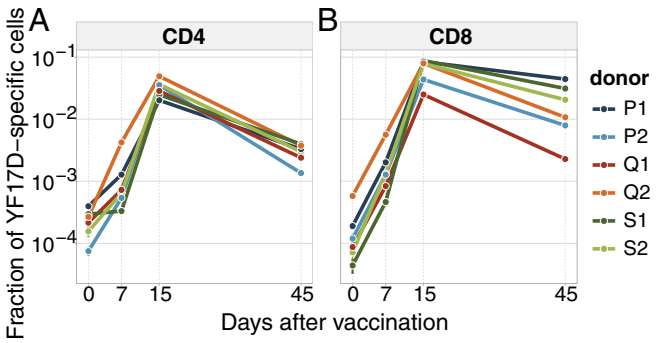

Fig. 4. Dynamics of YFV-responding T cells in the CD4+ and CD8+ compartments. Total fraction of (*A*) CD4+ and (*B*) CD8+ repertoires occupied by clonotypes significantly expanded from day 0 to day 15 for different timepoints. CD4+ and CD8+ T cell subpopulations show similar dynamics, although the CD8+ response degrades more slowly. Error bars are smaller than one line width.

identify clonotypes responding to infections in the clinic, when preinfection timepoints are not available.

Discussion

In this study, we used high-throughput TCRβ repertoire profiling to identify major changes occurring in the repertoire after immunization with the live attenuated YFV 17D vaccine. We found several hundreds of unique TCR clones in each donor expanded in response to vaccination. A strong clonal expansion of up to >2000-fold occurred between days 7 and 15 following vaccination. This proliferation corresponds to at least 11 divisions in 7 d, with an average of 15 h per cell cycle. Similar division rates (doubling times of 8 to 15 and 11 to 17 h for CD4+ and CD8+ T cells, respectively) were observed for lymphocytic choriomeningitis virus (LCMV) infection (22) and also in adoptive transfer experiments in mice (18 h) (23). However, in our case, the actual expansion rate could be much higher, because when a clonotype is not found on day 7, its initial concentration is unknown, providing only a lower bound estimate of the fold-change. Higher sequencing depth or shorter intervals between timepoints are needed to refine this estimate of expansion rates. This extension would allow us to study the impact of the T cell clone phenotype and the TCR sequence on the clonal expansion rate. Initial low concentrations of YFV-reactive clonotypes suggest their naive phenotype.

Our longitudinal approach can be applied to identify antigen-specific clonotypes for poorly characterized pathogens with unknown epitopes, which may be useful to study the immune response to emerging viral infections. One could also use this approach to track the response to infection and vaccination of unconventional T cell subsets for which antigens are still unknown (i.e., $\gamma\delta$ T cells).

While analyzing expansion between a prevaccination timepoint and the peak of the T cell response (day 15 in our case) is the the most natural choice, for real-world, nonexperimental acute infections, acquiring a pre-infection timepoint is often impossible or impractical. In this case, we showed how to use the dynamics of clonal contraction to identify reactive clonotypes, by comparing their abundance at the peak of the response to a timepoint taken a few weeks later. Implementing such a protocol in the clinic for acute primary viral infections with predictable kinetics [e.g., tick-borne encephalitis (24), hantavirus infection (25), and dengue fever (26) in which the T cell response peak occurs 1 wk after symptom onset] could lead to the rapid growth of the number of annotated TCR sequences specific to a range of infectious diseases, facilitating diagnostics and vaccine design. For secondary immune responses and chronic infections, further studies are required to determine the set of timepoints and sequencing depth needed to identify responding clonotypes.

Simultaneous sequencing of bulk PBMC, CD4+, and CD8+ subsets allowed us to determine the phenotype of the responding clones and to describe the kinetics of the response inside each compartment. While Blom et al. (10) found that YFV-specific CD4+ T cell concentration peaked slightly earlier than CD8+ T cells, we did not detect any difference in the expansion kinetics with our limited temporal resolution. More timepoints between days 7 and 15 could have been helped to detect such differences. However, we did find differences in the contraction kinetics, with CD4+ cells contracting faster than CD8+ cells. We also found differential recruitment to the CD45RO+ memory compartment, with many more CD4+ than CD8+ TCRβs detected in the CD45RO + fraction on day 45. CD8+ memory formed in response to YFV immunization was shown to be largely CD45RA+ (8, 10, 13), whereas James et al. (27) found that CD4+ memory cells were mainly CD45RA–. Our results suggest that CD4+ and CD8+ YFV-reactive memory populations also differ in their CD45RO expression.

Our study describes the reaction of the T cell repertoire in a model of an acute infection in genetically identical individuals. It was previously shown that the T cell repertoires of twins have more sequence overlap in abundant clonotypes (15, 16). Part of this extensive overlap may be explained by in utero sharing of T cells (18). The authors in ref. 17 also found more overlap in vaccinia virus-specific CD4+ repertoires (isolated by in vitro cultivation with the antigen) in twins. Consistent with these findings, we report more TCRβ amino acid sequence overlap in the YFV-reactive repertoires of twins than in those of unrelated individuals. The maximum normalized number of shared YFV-reactive clonotypes that we observe between twins (3×10^{-5}) is very close to the values measured by Qi et al. (17) for VZV-specific clonotypes (4×10^{-5}). Twin YFV-specific TCRβ clonotypes also show higher nucleotide sequence overlap than those of unrelated individuals, even relative to their higher amino acid sequence overlap (see *SI Appendix*, Table S4). To assess the significance of this observation, we used a generative model (28) to ask how likely it is to produce such an amount of shared nucleotide sequences by convergent recombination (see *SI Appendix*). In the vast majority of simulations ($p < 0.014$), the model underestimates the number of shared sequences between twin donors. We speculate that some of these shared clonotypes were exchanged in utero. Yet two-thirds of YFV-reactive TCRβ amino acid sequences shared between twins have different nucleotide sequences and can only be explained by convergent recombination and selection. Our results also suggest a mechanism for the previously reported extensive sharing of abundant TCRs in twins (15). Under this mechanism, twins would share more TCR sequences expanded in response to the same infection.

We showed that the response to the vaccine is very diverse, with few TCRβs shared even between donors with identical genetic and environmental backgrounds. Nevertheless, using a simple similarity measure, it is possible to identify YFV-specific clonotypes from yet unseen repertoires with high specificity, using datasets of TCRs with known YFV specificity. The sensitivity of this classifier could be improved by collecting more examples of YFV-specific TCRs from more donors. We also show how published antigen-specific sequences can be used for functional repertoire annotation. On day 15 after immunization, we found a much higher cumulative frequency of published A02-NS4b$_{214-222}$–specific sequences compared with prevaccination levels. Yet only a few significantly expanded clones matched those published sequences. This could be explained in two possible ways. First, our significantly expanded clones may be specific to epitopes other than NS4b$_{214-222}$. Second, the A02-NS4b$_{214-222}$–specific TCRβ repertoire may be so diverse that little overlap is to be expected between random subsamples of it. Further accumulation of antigen-specific TCR sequence data—acquired by sequencing of multimer-specific cells, longitudinal studies as done here, and disease-associated studies with large cohorts (29, 30)—will provide the means for disease diagnostics and extraction of clinically relevant information from T cell repertoire data.

Materials and Methods

Donors and Samples. Detailed description of all experimental procedures could be found in *SI Appendix*. Three pairs of healthy identical twins participated in this study. The blood was collected with informed consent in a certified diagnostics laboratory. The study was approved by the institutional review board (IRB) of Pirogov Russian National Research Medical University. PBMCs were isolated with the Ficoll–Paque method, and CD4+, CD8+, and CD45RO+ subpopulations were immunomagnetically positively isolated from PBMCs. Staining with the HLA-A*02 dextramer loaded with the NS4b$_{214-222}$ peptide (LLWNGPMAV) from YFV 17D (Immudex) was performed according to the manufacturer's protocol. The IFN-gamma secretion assay was performed using the IFN-gamma Secretion Assay-Cell Enrichment and Detection Kit (Myltenyi Biotec) according to the manufacturer's protocol. To stimulate IFN-gamma production,

whole blood was incubated with viral particles and anti-CD28 antibody for 5 h. All isolated cells were immediately lyzed with Trizol reagent (Invitrogen).

TCRβ cDNA Library Preparation and Sequencing. TCRβ cDNA library preparation, sequencing, and data analysis were performed as described in ref. 18. Briefly, total RNA was isolated from cells using TRIzol reagent. The 5′RACE cDNA synthesis with primer specific for TCRβ constant segments was followed by two-step PCR amplification. All libraries were sequenced on Illumina HiSeq 2500 platform. Raw data were demultiplexed and clustered by UMI (unique molecular identifiers) using the MIGEC software (31); alignment of V, D, and J templates was then performed by MiXCR (32).

Statistical Analysis. We identified expanded clones probabilistically using Bayesian statistics. We first assumed and inferred from data (using maximum likelihood) a generative model with power-law clone size distribution (33), followed by selective expansion of a subset of clones after vaccination. The number of cells of each clone in each sample was modeled by a negative binomial and the number of sequenced mRNAs for each cell by a Poisson distribution. We then applied Bayes rule to compute the posterior probability distribution of the fold-change in concentration for each clone. Clones that expanded >32-fold with posterior probability >0.95 were selected for further analysis. Mathematical details are given in *SI Appendix*.

Data and Code Availability.

The code for the clonal expansion model is available on github: https://github.com/mptouzel/pogorelyy_et_al_2018. Raw data are available on NCBI Sequence Read Archive (accession no. PRJNA493983).

ACKNOWLEDGMENTS. We thank Dr. I. V. Zvyagin for his help in finding appropriate donors for this study and Drs. L. V. Gmyl and M. F. Vorovitch for YFV 17D purification. TCRβ libraries sequencing, raw sequencing data processing, and reconstruction of TCRβ repertoires were supported by Russian Science Foundation Grant 15-15-00178. M.V.P. and E.S.E. are supported by Skoltech Systems biology fellowships. This work was partially supported by European Research Council Consolidator Grant 724208.

1. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135: 183–191.
2. Dash P, et al. (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547:89–93.
3. Glanville J, et al. (2017) Identifying specificity groups in the T cell receptor repertoire. *Nature* 547:94–98.
4. Davis MM, Altman JD, Newell EW (2011) Interrogating the repertoire: Broadening the scope of peptide–MHC multimer analysis. *Nat Rev Immunol* 11:551–558.
5. Robinson J, et al. (2013) The IMGT/HLA database. *Nucleic Acids Res* 41:D1222–D1227.
6. Monath TP, Vasconcelos PF (2015) Yellow fever. *J Clin Virol* 64:160–173.
7. Miller JD, et al. (2008) Human effector and memory CD8+ T cell responses to smallpox and yellow fever vaccines. *Immunity* 28:710–722.
8. Akondy RS, et al. (2009) The yellow fever virus vaccine induces a broad and polyfunctional human memory CD8+ T cell response. *J Immunol (Baltimore, Md.: 1950)* 183:7919–7930.
9. Akondy RS, et al. (2015) Initial viral load determines the magnitude of the human CD8 T cell response to yellow fever vaccination. *Proc Natl Acad Sci USA* 112:3050–3055.
10. Blom K, et al. (2013) Temporal dynamics of the primary human T cell response to yellow fever virus 17D as it matures from an effector- to a memory-type response. *J Immunol* 190:2150–2158.
11. Kohler S, et al. (2012) The early cellular signatures of protective immunity induced by live viral vaccination. *Eur J Immunol* 42:2363–2373.
12. Kongsgaard M, et al. (2017) Adaptive immune responses to booster vaccination against yellow fever virus are much reduced compared to those after primary vaccination. *Sci. Rep* 7:1–14.
13. Fuertes Marraco SA, et al. (2015) Long-lasting stem cell-like memory CD8+ T cells with a naïve-like profile upon yellow fever vaccination. *Sci Transl Med* 7:282ra48.
14. DeWitt WS, et al. (2015) Dynamics of the cytotoxic T cell response to a model of acute viral infection. *J Virol* 89:4517–4526.
15. Zvyagin IV, et al. (2014) Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proc Natl Acad Sci USA* 111:5980–5985.
16. Rubelt F, et al. (2016) Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun* 7:11112.
17. Qi Q, et al. (2016) Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Sci Transl Med* 8:332ra46.
18. Pogorelyy MV, et al. (2017) Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput Biol* 13:e1005572.
19. Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.
20. Shugay M, et al. (2017) VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* 46:419–427.
21. Miles JJ, Douek DC, Price Da (2011) Bias in the $\alpha\beta$ T-cell repertoire: Implications for disease pathogenesis and vaccination. *Immunol Cell Biol* 89:375–387.
22. De Boer RJ, Homann D, Perelson AS (2003) Different dynamics of CD4+ and CD8+ T cell responses during and after acute lymphocytic choriomeningitis virus infection. *J Immunol* 171:3928–3935.
23. Buchholz VR, et al. (2013) Disparate individual fates compose robust CD8+ T cell immunity. *Science* 340:630–635.
24. Blom K, et al. (2015) Specificity and dynamics of effector and memory CD8 T cell responses in human tick-borne encephalitis virus infection. *PLoS Pathog* 11:e1004622.
25. Lindgren T, et al. (2011) Longitudinal analysis of the human T cell response during acute hantavirus infection. *J Virol* 85:10252–10260.
26. Rivino L (2018) Understanding the human T cell response to dengue virus. *Dengue and Zika: Control and Antiviral Treatment Strategies*, eds. Hilgenfeld R, Vasudevan SG (Springer Singapore, Singapore), pp 241–250.
27. James EA, et al. (2013) Yellow fever vaccination elicits broad functional CD4+ T cell responses that recognize structural and nonstructural proteins. *J Virol* 87: 12794–12804.
28. Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA* 109:16161–16166.
29. Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 49:659–665.
30. Pogorelyy MV, et al. (2018) Method for identification of condition-associated public antigen receptor sequences. *eLife* 7:e33050.
31. Shugay M, et al. (2014) Towards error-free profiling of immune repertoires. *Nat Methods* 11:653–655.
32. Bolotin DA, et al. (2015) MiXCR: Software for comprehensive adaptive immunity profiling. *Nat Methods* 12:380–381.
33. Mora T, Walczak A (2018) Quantifying lymphocyte receptor diversity. *System Immunology*, eds. Das JD, Jayaprakash C (CRC Press, Boca Raton, FL), pp 185–199.

PNAS
www.pnas.org

Supplementary Information for

Precise tracking of vaccine-responding T-cell clones reveals convergent and personalized response in identical twins

M.V. Pogorelyy, A.A. Minervina, M. Puelma Touzel, A.L. Sycheva, E.A. Komech, E.I. Kovalenko, G.G. Karganova, E.S. Egorov, A.Y. Komkov, D.M. Chudakov, I.Z.Mamedov, T. Mora, A.M. Walczak, Y.B. Lebedev

To whom correspondence should be addressed. E-mail: tmora@lps.ens.fr; awalczak@lpt.ens.fr; lebedev_yb@ibch.ru

This PDF file includes:

Supplementary text
Figs. S1 to S11
Tables S1 to S4
Captions for Databases S1 to S2
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Databases S1 to S2

www.pnas.org/cgi/doi/10.1073/pnas.1809642115

Supporting Information Text

Analysis of sequencing data of functional assays. All three functional assays (IFN-gamma secretion assay, CD8+CD38+HLA-DR+ staining and A02-NS4b$_{214-222}$ dextramer staining) provided us with TCRβ repertoire sequencing information. However, both fluorescence activated cell sorting after dextramer staining, activation marker staining and magnetic column enrichment are not completely precise. These methods enrich resulting samples with YF-specific cells, but abundant irrelevant cells are still found in the corresponding repertoires. To detect clonotypes enriched in the CD8+CD38+HLA-DR+ population, the IFN-gamma producers population or the tetramer-positive population in comparison to the bulk population at the same timepoint we used Fisher's exact test, as suggested in Ref. (1). For each clone the number of barcodes occupied by it in the activated population, the number of barcodes occupied by other clonotypes in the activated population, the number of barcodes occupied by the clone in the bulk population and the number of barcodes occupied by all other clones in the bulk population formed a 2×2 contingency table, and then one way Fisher's exact test was applied. This procedure was repeated for each clone found in the activated population, clones with BH-corrected p-values<0.05 were considered significantly enriched in the activated population.

Identification of expanded clones using edgeR. To check robustness of our clonal expansion model we used an independent approach to detect expanded clones. edgeR (2) is a package used to analyse a variety of count data produced with HTS. It was applied for differential gene expression analysis, differential splicing analysis and bisulfite sequencing. Here we apply statistical tests implemented in the edgeR package to identify clones expanded after YFV immunization. Two biological replicates were used for each timepoint, TMM-normalization and trended dispersion estimates was performed as described in the edgeR manual. An exact test based on the quantile-adjusted conditional maximum likelihood (qCML) was used to identify clones significantly expanded between pairs of timepoints. A clonotype was considered significantly expanded, if it's $\log_2$ fold change estimate $\log_2$FC>5 and it's p-value after multiple testing correction was lower than 0.01. We reproduced the main results described in the manuscript text using clonotypes identified as significant by edgeR (see SI Fig. S1), so two differential clone expansion models agree with each other.

Screening of CD8+ repertoires for known YFV and CMV specific clonotypes. We screened the CD8+ TCRβ repertoires of our donors at different timepoints for A02-NS4b$_{214-222}$ specific sequences, as well as for sequences specific for an unrelated epitope as a control: the cytomegalovirus (CMV) immunodominant HLA-A*02-restricted pp65$_{495-503}$ peptide (NLVPMVATV). All of our donors are HLA-A*02-positive (complete HLA genotypes can be found in SI Table S1). We found that both the number of hits and their cumulative frequency (SI Fig. S7) were significantly higher for published A02-NS4b$_{214-222}$-specific sequences on day 15 than on day 0 (one-sided t-test p=0.019). By contrast, in the control no significant difference was found (one-sided t-test p=0.27).

Testing for difference in CD4 and CD8 YFV-specific clonotypes contraction rates using Negative Binomial model for cDNA counts. For each clonotype we redrew a random number of counts on days 15 and 45, with the same variability as inferred from replicate experiments to mimic experimental noise. Repeating this process many times gave a p-value corresponding to the fraction of numerical experiments where the ordering between CD4+ and CD8+ decay was opposite to the actual observation. CD4+ and CD8+ fold decreases significantly differed ($p < 0.0001$) for 5 donors out of 6, with the exception of donor Q1 ($p = 0.15$) who was an outlier in the CD8+ response strength and the number of significantly expanded clones.

Sequence similarity analysis. We used the Hamming distance to quantify CDR3 sequence similarity in each YFV-reactive repertoire. We counted the number of CDR3 amino acid sequence pairs, which are 0 mismatches (same amino acid, but different nucleotide sequences), 1 mismatch or less, 2 mismatches or less from each other. Since the number of YFV-reactive clones varied among donors, the number of similar clones inside each YFV-reactive repertoire needed to be normalized. We divided it by $n(n-1)/2$, which is the total number of possible clone pairs. We randomly sampled clonotypes from the same timepoint for comparison, and found a smaller number of similar pairs. In Fig. 4A. we plot the normalized sharing in log scale to compare these values.

Nucleotide sequence sharing analysis of YFV-reactive clonotypes. To see, if the excess of nucleotide sequence sharing in twins may be explained by convergent recombination, we performed the following simulation: for each donor, for each shared YFV-specific TCRβ amino acid sequence we generated nucleotide variants using a previously published TCR recombination model (3) and compared the resulting simulated nucleotide sequence variants between donors. Then the total number of shared nucleotide sequences between twins in this simulation was compared to the total number of shared nucleotide sequences between the twins in the data. The simulation was performed 1000 times: in 986 cases out of 1000 the number of nucleotide sequences shared between twins was less than in the data, which gives a p-value=0.014. This result suggests, that at the least some of the nucleotide sequences of YF-reactive clones may be shared prenatally between twin donors, because this extensive sharing of nucleotide sequences cannot be explained by convergent recombination and selection. On the contrary, the number of shared nucleotide sequences between unrelated donors in the simulation was rarely less than in the data (in 156 cases out of 1000), which means that the nucleotide sequence sharing between unrelated individuals can easily be achieved by convergent recombination alone (p-value=0.84).

Statistical analysis. Candidate responding clones were identified using a model of mRNA count statistics accounting for differential expression and the sequencing process. The normalized clone size, f, is distributed according to the probability density function $\rho(f)$, bounded by $M_T^{-1} \leq f \leq 1$, where $M_T = 10^{11}$ is an estimate of the total number of lymphocytes in an individual. Based on previous observations (4–6), $\rho(f)$ is set as a power-law, i.e. $\rho(f) \propto f^{-\gamma}$. A clone of size f appears in a sample containing M lymphocytes on average as fM cells. To account for overdispersed count statistics, the number of cells of a clone, m, is set to be Negative-Binomial distributed with mean fM and variance $fM + a(fM)^{\beta}$, with $a > 0$ the coefficient and $\beta > 1$ the power controlling the over-dispersion. For each clone, the number n of detected mRNA molecules (i.e. UMI) is distributed according to a Poisson distribution with mean mN/M, where N/M is the average number of UMI per cell, obtained using the observed total number of molecules, N. The two-step model (negative-Binomial distributed cell counts and Poisson distributed molecule counts) captures inter-replicate count noise better than a simple model of negative binomial distributed molecule counts (see Fig. S2).

The model produces an ensemble of count pairs (n_1, n_2), where $n_i \sim \text{Poisson}(m_i N_i/M)$, $m_i \sim \text{NegBin}(fM, fM + a(fM)^{\beta})$, for each replicate $i = 1, 2$, and $f \sim \rho$ is common to both replicates. For a given pair, the likelihood, $P(n_1, n_2|\theta)$, is obtained by marginalizing over m_1, m_2, and f:

$$P(n_1, n_2|\theta) = \int_{f_{min}}^{1} \rho(f)\mathrm{d}f \prod_{i=1}^{2} \sum_{m_i=0}^{\infty} \text{NegBin}\left[fM, fM + a(fM)^{\beta}\right](m_i) \times \text{Poisson}\left[m_i N/M\right](n_i). \quad [1]$$

The parameters of this model, $\theta = (\gamma, M, a, \beta)$, are learned for a given data set of two RepSeq replicates taken at the same time point, $\mathcal{D} = (n_{1,j}, n_{2,j})_{j=1,\ldots,N_{\text{clones}}}$, by maximizing the likelihood over all observed count pairs in the replicates (say at day 0):

$$\mathcal{L}(\theta) = \prod_{j=1}^{N_{\text{clones}}} P(n_{1,j}^{\text{day.0}}, n_{2,j}^{\text{day.0}}|\theta). \quad [2]$$

To account for differential expression, we introduce a selection factor s defined as the log-fold change between a clone's frequency on one day, f, and that on another, fe^s, and define $P(n_1, n_2|s, \theta)$ as before, but replacing f by fe^s in the definition of m_2. Given a prior distribution over s, $P(s|\theta')$, parametrized by a set of parameters θ' distinct from θ, we used Bayes rule to obtain the posterior log fold-change probability function given an observed count pair,

$$P(s|n_1, n_2, \theta, \theta') = \frac{P(n_1, n_2|s, \theta)P(s|\theta')}{P(n_1, n_2|\theta, \theta')}, \quad [3]$$

with molecule pair count distribution in this differentially expressed case given by $P(n_1, n_2|\theta, \theta') = \int P(n_1, n_2|s, \theta)P(s|\theta')\mathrm{d}s$.

We parametrize our prior as $P(s|\theta') = \alpha \exp(-|s|/\bar{s})/(2\bar{s}) + (1-\alpha)\delta(s)$, with $0 \leq \alpha \leq 1$ the fraction of clones that respond to the change, and $\bar{s} > 0$ their typical effect size. We set the values of the parameters $\theta' = (\alpha, \bar{s})$ of this prior by maximizing the likelihood of the count pair data taken at two given time points (say day 0 and day x), given the model over θ':

$$\mathcal{L}(\theta'|\theta) = \prod_{j=1}^{N_{\text{clones}}} P(n_{1,j}^{\text{day.0}}, n_{2,j}^{\text{day.x}}|\theta, \theta'), \quad [4]$$

where θ is fixed to the value inferred previously using two replicates at the same time point.

The differential expression model described above was implemented using custom python scripts available at https://github.com/mptouzel/pogorelyy_et_al_2018. The program consisted of three main steps. First, the four null model parameters (θ) were jointly learned on the pair of day-0 replicates. In the computation of the likelihood, $P(n_1, n_2|\theta)$, only a finite number of terms contribute significantly to the sum over m_1 and m_2 and so we computed the sum over a finite range that shifted with f and n, the latter only taking values appearing in the dataset. The integral over the density $\rho(f)$ was discretized into 500 bins and performed using trapezoid integration methods. The likelihood (2) was maximized using a simplex method contained as an option in the `minimize` function of the `scipy` optimize package (http://www.scipy.org/). Next, these parameters were used to generate the conditional distribution $P(n_1, n_2|s, \theta)$ for a range of values of s centered on $s = 0$, using the same code but with an s-dependent range of frequencies (e.g. the range for $s = 0$ is the same as in the null model). To maximize (4), grid search was then made over the $\theta' = (\alpha, \bar{s})$ parameter space of the prior distribution, $P(s, \theta')$. For better numerical control over the parameter space, we implemented s over a discrete set of values of step size $\Delta s \approx 0.1$ from $s_{\min} \approx -25$ to $s_{\max} \approx 25$, reducing the integral over s to a sum. The approximations in these values arise from the fact that they were taken as multiples of the discretized frequency step, so that variation in s could be achieved using quantities computed over f. The prior parameter combination θ' giving the largest averaged likelihood was then used in the final step of the program in which the posteriors, $P(s|n_1, n_2, \theta, \theta')$ using (3). These posteriors varied systematically with (n_1, n_2), for example in the relative weight of the unaffected ($s = 0$) to affected ($s \neq 0$) component (see Fig. S3).

In analogy with p-values, we used the posterior probability corresponding to the null hypothesis that they are not expanded, $p = P(s \leq 0|n_1, n_2, \theta, \theta') < 0.05$ to select the significantly expanded clones. A second threshold on the median s, $\text{argmax}_s[P(s|n_1, n_2, \theta, \theta')] > \ln(32)$ was applied to select for clones that were also greatly expanded. For the contraction from day 15 to day 45, no threshold on the fold change was applied.

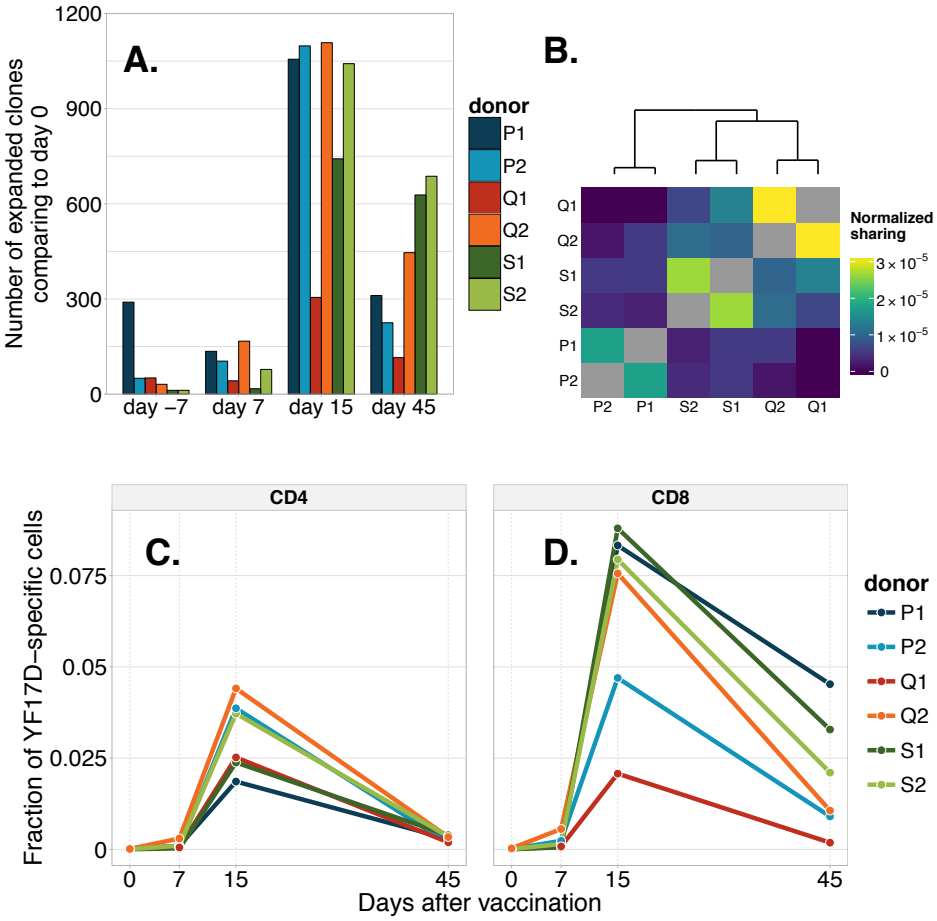

Fig. S1. **Results of our analysis could be reproduced using the edgeR differential gene expression analysis software.** **A.** Number of significantly expanded clonotypes in comparison to day 0. **B.** The heatmap shows the number of the same amino acid clonotypes between two sets of clones significantly expanded from day 0 to day 15, divided by the product of the sizes of these sets. **C. and D.** Dynamics of T-cells in the CD4+ and CD8+ compartments. Total fractions of CD4+ (**C.**) and CD8+ (**D.**) repertoires occupied by clonotypes significantly expanded from day 0 to day 15 is plotted at different timepoints.

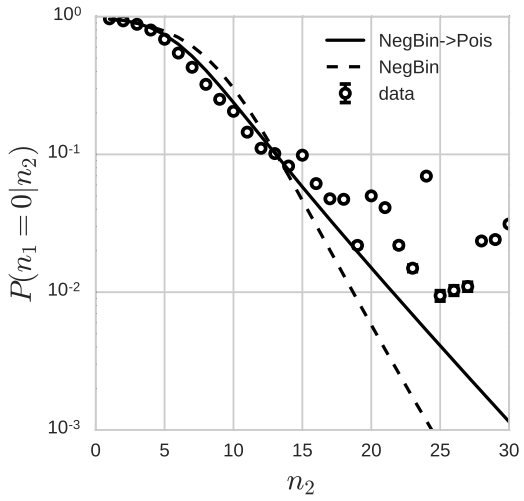

Fig. S2. **Conditional count distributions of learned one and two step models for donor S1 day 0 timepoint biological replicates.** Conditional RNA molecule count distributions for the best one-step negative binomial distributed RNA molecule count distribution (dashed line), and the best two-step negative binomial distributed cell count distribution and Poisson distributed RNA molecule count distribution (line). Error bars on data (dots) are standard error.

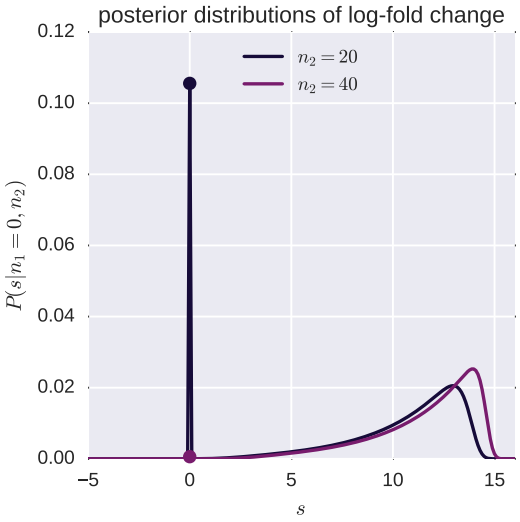

Fig. S3. Posterior distributions with and without a salient $s = 0$ mode. Shown are the posterior expansion probability distributions, $P(s|n_1, n_2)$, for two observed clones of donor S2, each of which were unseen at day 0, i.e. $n_1 = 0$, and appearing at $n_2 = 20$ and $n_2 = 40$, respectively, at day 15. Note the presence and absence, respectively, of a salient $s = 0$ mode.

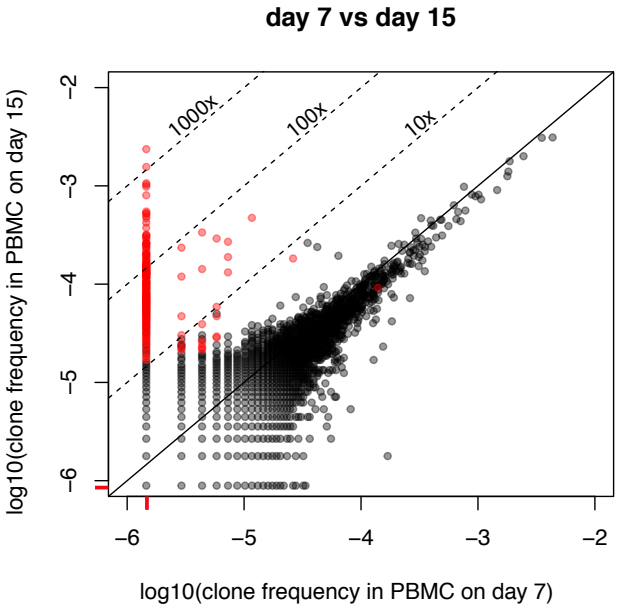

Fig. S4. **Scatterplot of clone frequencies in donor S1 on day 7 and day 15.** Log_{10} of clone frequency (defined as the ratio of the UMI for given clone and the sum of UMIs in sample) on day 7 (x-axis) is plotted against $\log_{10}$ of clone frequency on day 15. Pseudocounts of 1 UMI are used to show 0 on logarithmic scale. Red ticks on axes mark pseudocount size, these clones have 0 concentration at the corresponding timepoint. Red dots on the plot indicate clones significantly expanded between day 0 and day 15.

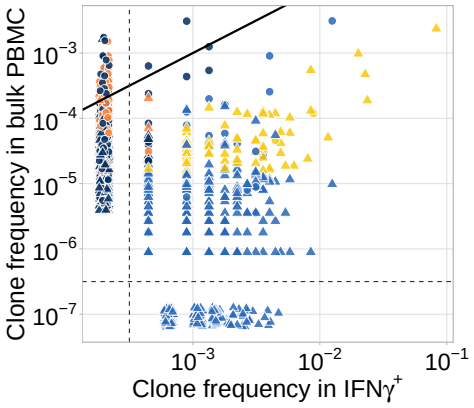

Fig. S5. C. IFN-gamma producing subpopulation is enriched with clonotypes expanded between days 0 and 15. The relative abundance of a clonotype in the IFN-gamma producing subpopulation (x-axis) isolated from whole blood stimulated with vaccine by interferon gamma secretion assay is plotted against the relative abundance in the bulk PBMC isolated at the same timepoint (y-axis). Yellow dots indicate clonotypes that strongly expanded between day 0 and day 15. Bright and dark shades indicate clonotypes enriched and unenriched in activated subpopulation respectively. Circles indicate clonotypes identified as CD8+ by their abundance in repertoire of CD8+ T-cells sequenced on same timepoint, and triangles are clonotypes with CD4+ phenotype. In this stimulation conditions mostly CD4+ T-cells produce IFN-gamma. Black line shows identity. Behind dash lines we show clonotypes absent in either IFN-γ positive or sequenced PBMC fractions.

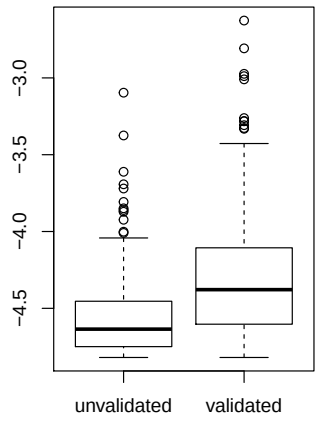

Fig. S6. Log_{10} **frequencies of YFV-reactive clones unvalidated and validated by functional tests.** Unvalidated clones have significantly lower frequencies (t-test p$< 10^{-16}$)

.

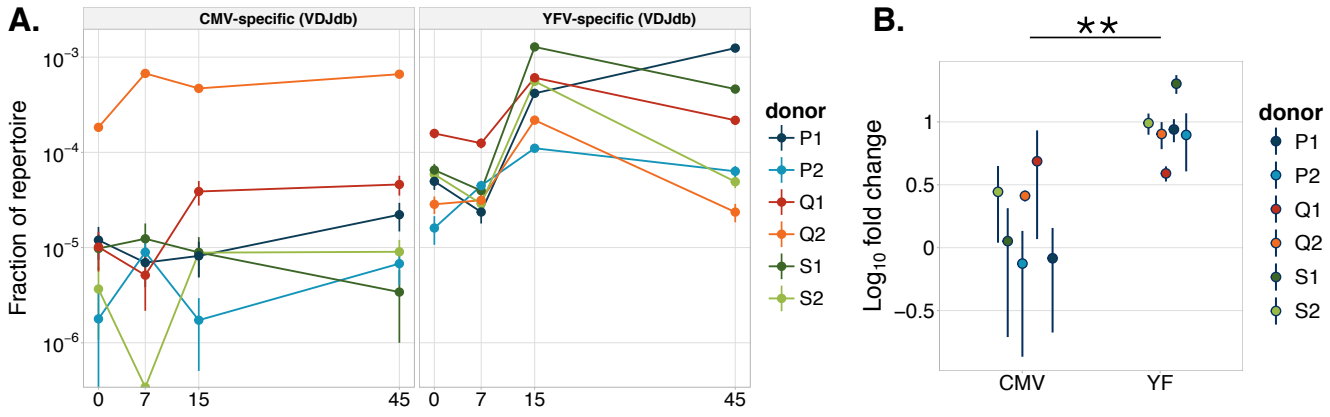

Fig. S7. Log_{10} **cumulative frequency of published YFV- and CMV-specific clonotypes in our CD8+ datasets. A.** We found much higher relative abundance of known YFV-reactive clonotypes on day 15 comparing to day 0 for each donor, however no such effect was found for known CMV-reactive clonotypes. **B.** $\log_{10}$ fold-change between days 0 and 15 for cumulative concentrations of CMV- and YF-specific clonotypes from VDJDB. Every donor show increase for YF-specific clonotypes from VDJdb, and there is significant difference (Welch two sample t-test, p=0.002) between CMV- and YF-specific clonotype concentration $\log_{10}$ fold-changes.

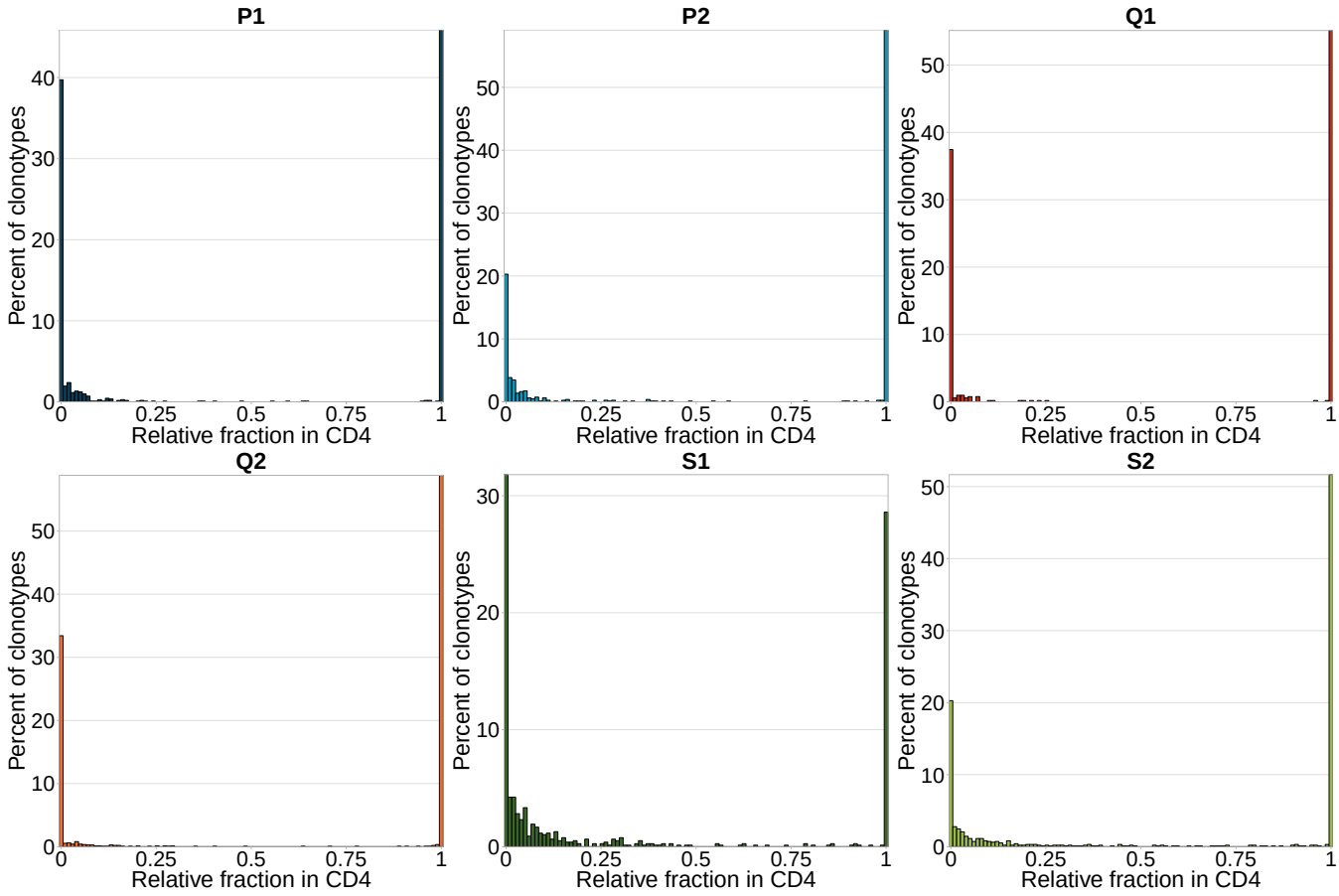

Fig. S8. **Bimodality of YF-specific clone concentrations in the CD4+ and CD8+ compartments.** For each expanded clonotype from bulk PBMC we show relative fraction in CD4+ repertoire: fraction of 0 means that clonotype is found exclusively in CD8+ repertoire, fraction of 1 means clonotype is found exclusively in CD4+ repertoire.

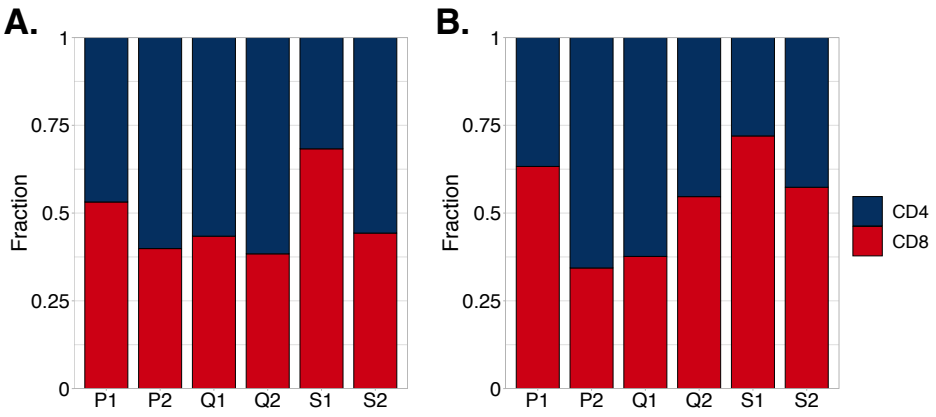

Fig. S9. $\mathbf{Distribution\ of\ yellow\ fever\ reactive\ clonotypes\ between\ CD4+\ and\ CD8+\ subpopulations.}$ **A.** Relative fractions of activated clonotypes attributed to the CD4+ and CD8+ compartments and **B.** relative fractions of cells attributed to the CD4+ and CD8+ activated clonotypes in each repertoire.

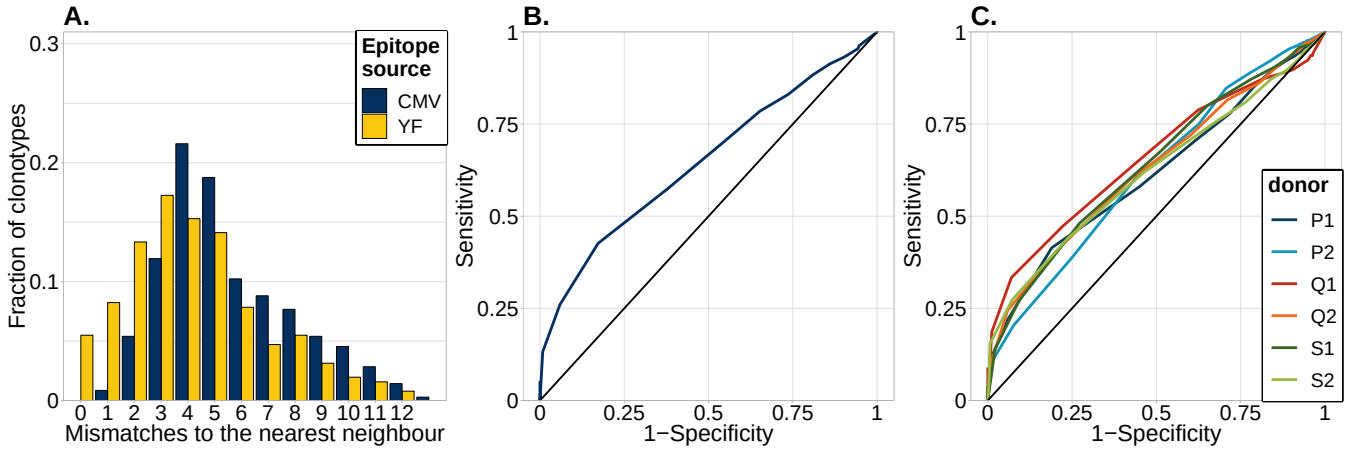

Fig. S10. $\mathbf{Predicting\ the\ YFV\ specificity\ of\ TCR\ sequences.}$ $\mathbf{(A)}$ Histogram of the minimal Hamming distance of expanded TCR amino acid sequences identified in this study to published YFV-specific sequences (YFV immunodominant epitope NS4b$_{214-222}$, yellow) and cytomegalovirus (CMV)-specific sequences (CMV immunodominant epitope pp65$_{495-503}$, blue). A much larger fraction of our expanded TCRs were similar (0, 1, or 2 mismatches) to YFV-specific sequences than to CMV-specific sequences. $\mathbf{(B)}$ Receiver Operating Characteristic (ROC) curve for the classification of published NS4b$_{214-222}$-specific sequences using the minimal distance to the expanded TCRs identified in this study. $\mathbf{(C)}$ ROC curves for "leave-one-out" classifier validation. The top 1000 most abundant TCRs of each individual were classified as YFV-reactive or non YFV-reactive according to their minimal distance to TCRs expanded in the other five individuals.

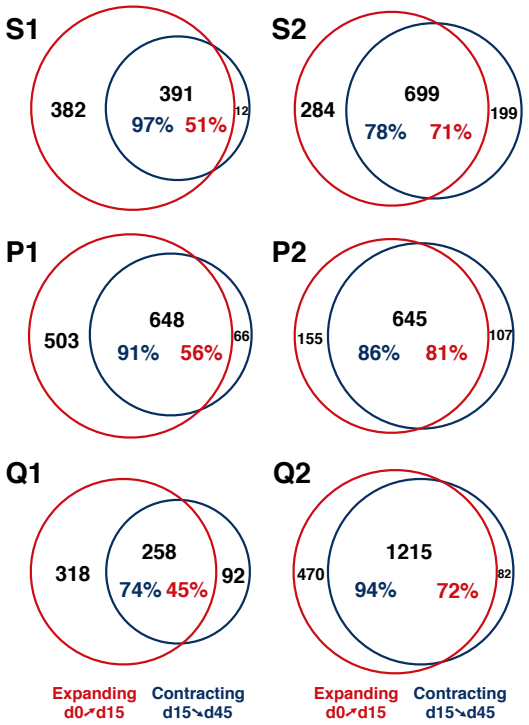

Fig. S11. **Clonal contraction can be used to identify YFV-17D reactive clonotypes with high specificity and good sensitivity.** Venn diagrams show overlap between the subset of TCR clones significantly expanded from day 0 to day 15 (red), and the subset of TCR clones significantly contracted from day 15 to day 45 (blue). All donors show large overlaps between the contracting and expanding subsets.

Table S1. HLA-haplotypes of donors

Locus	S1 and S2	P1 and P2	Q1 and Q2
A	02:01:01/03:01:01	02:01:01/02:01:01	02:01:01/02:01:01
B	38:01:01/07:02:01	51:01:01/07:02:01	13:02:01/44:02:01
C	07:02:01/12:03:01	07:02:01/14:02:01	05:01:01/06:02:01
DQB1	06:03:01/06:02:01	03:01:01/04:02:01	02:02:01/06:03:01
DRB1	13:01:01/15:01:01	08:03:02/08:01:03	07:01:01/13:01:01
DRB3	01:01:02	-	01:01:02
DRB4	-	-	01:03

Table S2. The number of YFV-reactive TCRβ sequences found in this study that are similar to published TCRβ sequences specific for immunodominant YFV and CMV epitopes.

	S2	S1	P2	P1	Q2	Q1
CMV, 0 mism	0	0	0	0	0	0
YFV, 0 mism	5	3	1	2	4	3
CMV, 1 mism	0	0	1	0	2	0
YFV, 1 mism	10	24	9	12	21	5
CMV, 2 mism	5	5	3	5	10	2
YFV, 2 mism	30	27	11	24	40	21

Table S3. The number of unique shared nucleotide/amino acid sequences between YFV-reactive repertoires of our donors. The diagonal shows the total number of unique nucleotide/amino acid TCRβ sequences inside each YFV-reactive repertoire.

	S2	S1	P2	P1	Q2	Q1
S2	983/969	4/20	0/3	0/4	2/9	1/3
S1	4/20	773/762	0/3	0/5	1/12	0/4
P2	0/3	0/3	800/792	1/18	0/2	0/0
P1	0/4	0/5	1/18	1151/1142	0/7	0/1
Q2	2/9	1/12	0/2	0/7	1685/1649	5/21
Q1	1/3	0/4	0/0	0/1	5/21	576/563

Table S4. Contribution of public clones to YFV response. A TCRβ amino acid clonotype is considered public if it is shared with at least one other donor in this study.

	S2	S1	P2	P1	Q2	Q1
Fraction of public YFV-reactive clones	0.03	0.04	0.03	0.03	0.02	0.04
Cumulative freq. of public YFV-reactive clones	0.05	0.08	0.06	0.03	0.03	0.33

Additional data table S1 (SI_dataset_S1_VDJDB.txt)
A selection of CMV- and YFV-specific TCRbeta from VDJdb we used for our analysis.

Additional data table S2 (SI_dataset_S2_expanded_clones.txt)
Table with all YFV-specific clonotypes identified by clonal expansion in this work.

References

1. DeWitt WS, et al. (2015) Dynamics of the Cytotoxic T Cell Response to a Model of Acute Viral Infection. *Journal of Virology* 249(February):JVI.03474–14.
2. Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140.
3. Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences* 109(40):16161–16166.
4. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science* 324(5928):807–810.
5. Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America* 107(12):5405–10.
6. Mora T, Walczak A (2018) Quantifying lymphocyte receptor diversity in *Syst. Immunol.*, eds. Das JD, Jayaprakash C. (CRC Press), pp. 185–199.