Human thymopoiesis selects unconventional CD8$^+$ α/β T cells that respond to multiple viruses.

Valentin Quiniou[1,2], Pierre Barennes[1,2], Federica Martina[2#], Vanessa Mhanna[1#], Helene Vantomme[1,2#], Hang Phuong Pham[3], Mikhail Shugay[4], Adrien Six[1], Encarnita Mariotti-Ferrandiz[1,2], David Klatzmann[1,2]*

[1]Sorbonne Université, INSERM, Immunology-Immunopathology-Immunotherapy (i3), Paris, France

[2]AP-HP, Hôpital Pitié-Salpêtrière, Clinical Investigation Center for Biotherapies (CIC-BTi) and Immunology-Inflammation-Infectiology and Dermatology Department (3iD), Paris, France

[3]ILTOO pharma, Statistical department, Paris, France

[4]Center of Life Sciences, Skoltech, Moscow, Russia

# these authors have contributed equally

* Address correspondence and reprint requests to:
Prof. D. Klatzmann, Pitié-Salpêtrière Hospital, 83 boulevard de l'Hôpital, F-75013, Paris, France.
E-mail: david.klatzmann@sorbonne-universite.fr

T cell receptors (TCRs) are formed by stochastic gene rearrangements, theoretically generating >>10^{19} sequences[1]. They are selected during thymopoiesis, which releases a repertoire of about 10^8 unique TCRs[2,3] per individual. How evolution shaped a process that produces TCRs that would effectively respond to infectious agents is a central question of immunology. The paradigm is that a diverse enough repertoire of TCRs should always provide a proper, though rare, specificity for any given need. Expansion of such rare T cells would provide enough fighters for an efficacious immune response and enough antigen-experienced cells for memory[3,4]. We show here that thymopoiesis releases a large population of $CD8^+$ T cells harbouring diverse α/βTCRs with innate-like properties. These TCRs (i) have high generation probabilities and a preferential usage of some V and J genes, (ii) are shared between individuals, (iii) are highly enriched for viral antigen recognition and (iv) have a fuzzy rather than tight specificity. In vitro, T cells expressing these TCRs bind to and are activated by multiple unrelated viral peptides; in vivo, they respond to vaccination and infection, being notably found in bronchoalveolar lavages of COVID-19 infected patients. Our results support an evolutionary selection of pleiospecific α/βTCRs for broad antiviral responses and heterologous immunity.

We analysed the TCR repertoire dynamics of developing thymocytes. We first focused on the hypervariable CDR3 region of the TCR that interacts with the antigenic peptide, while CDR1 and CDR2 usually interact with HLA molecules[5]. CDR3 analyses can therefore be used to investigate the sharing of TCR specificities between individuals with distinct HLA molecules. We analysed the repertoire of purified $CD4^+CD8^+CD3^-$ ($DPCD3^-$), $CD4^+CD8^+CD3^+$ ($DPCD3^+$) and $CD4^-CD8^+CD3^+$ ($CD8^+$) thymocytes. $DPCD3^-$ thymocytes represent the earliest stage of TCR β-chain gene recombination and their repertoire embodies the unaltered outcome of the TCR generation process; $DPCD3^+$ thymocytes are at an early stage of the selection process and their repertoire should be minimally modified; $CD8^+$ thymocytes have passed the selection process and bear a fully selected repertoire. We analysed and represented the structure of these repertoires by connecting CDR3s (nodes) differing by at most one single amino acid (AA) (Levenshtein distance less than or equal to one: LD≤1) as such similar CDR3s most often bind the same peptide[6–12] (Supplementary Figure 1. In these networks, connected CDR3s are designated as clustered nodes and the others as dispersed nodes. For normalisation, we represented the first 18,000 most expressed β or α CDR3s from each sample. We observed a marked increase in the number of clustered CDR3s from $DPCD3^-$ to $CD8^+$ thymocytes (Fig. 1a and c, Supplementary Figure 2), which was remarkably consistent among all individuals studied, independently of their age, sex or HLA (Supplementary Figure 3). The node degree, i.e. its number of connections, was also significantly increased during T cell differentiation for clustered CDR3s (Fig. 1b and d, Supplementary Figure 2). The major and statistically significant ($p<0.0001$) increase in the proportion of clustered TCRs from $DPCD3^+$ to $CD8^+$ thymocytes, which was also accompanied by a significant increase ($p<0.0001$) in the node degree of clustered TCRs, reveals a positive selection of TCRs with shared recognition properties during thymopoiesis.

The probability of generation (*Pgen*) of a given TCR varies enormously from one TCR to the other, spanning over 10 orders of magnitude[1]. The clustered TCRs from both $DPCD3^+$ and $CD8^+$ thymocytes have a significantly higher *Pgen* ($p<0.0001$) than the dispersed ones (Fig. 1e). *Pgen* also increased significantly ($p<0.0001$) from $DPCD3^+$ to $CD8^+$ thymocytes (Fig. 1e). Moreover, the *Pgen* of $CD8^+$ thymocytes is significantly correlated with the node degree (Supplementary Figure 4). Clustered TCRs have a preferential usage of some V and J genes, resulting in a markedly different VJ recombination usage, notably similar across individuals (Fig. 1f, Supplementary Figure 5). As there is a remarkably shared clustered structure of $CD8^+$

thymocytes βCDR3s across individuals (Supplementary Figure 3), we investigated their private or public nature (Fig. 1g, Supplementary table 1). We found a significant increase of public (i.e. shared between at least 2 individuals) βCDR3s in $CD8^+$ versus $DPCD3^+$ thymocytes, which is mostly that of the clustered βCDR3s. For $CD8^+$ thymocytes, up to 31.7% of clustered βCDR3s are public compared to barely 1% of the dispersed ones (Fig. 1g, Supplementary Figure 6). These results are independent of HLA alleles sharing (Supplementary Figure 7). The βCDR3s with the highest *Pgen* values and degree were the most shared between individuals (Supplementary Figure 8). Moreover, there is a convergence of specificities between individuals' repertoires, as many βCDR3s of one individual are connected to those of other individuals (up to twelve), and more frequently in $CD8^+$ versus $DPCD3^+$ thymocytes (Supplementary Figure 9). Altogether, these results indicate that the mechanisms for TCR generation and for their further thymic selection are biased to shape a public repertoire of connected βCDR3s with shared recognition properties.

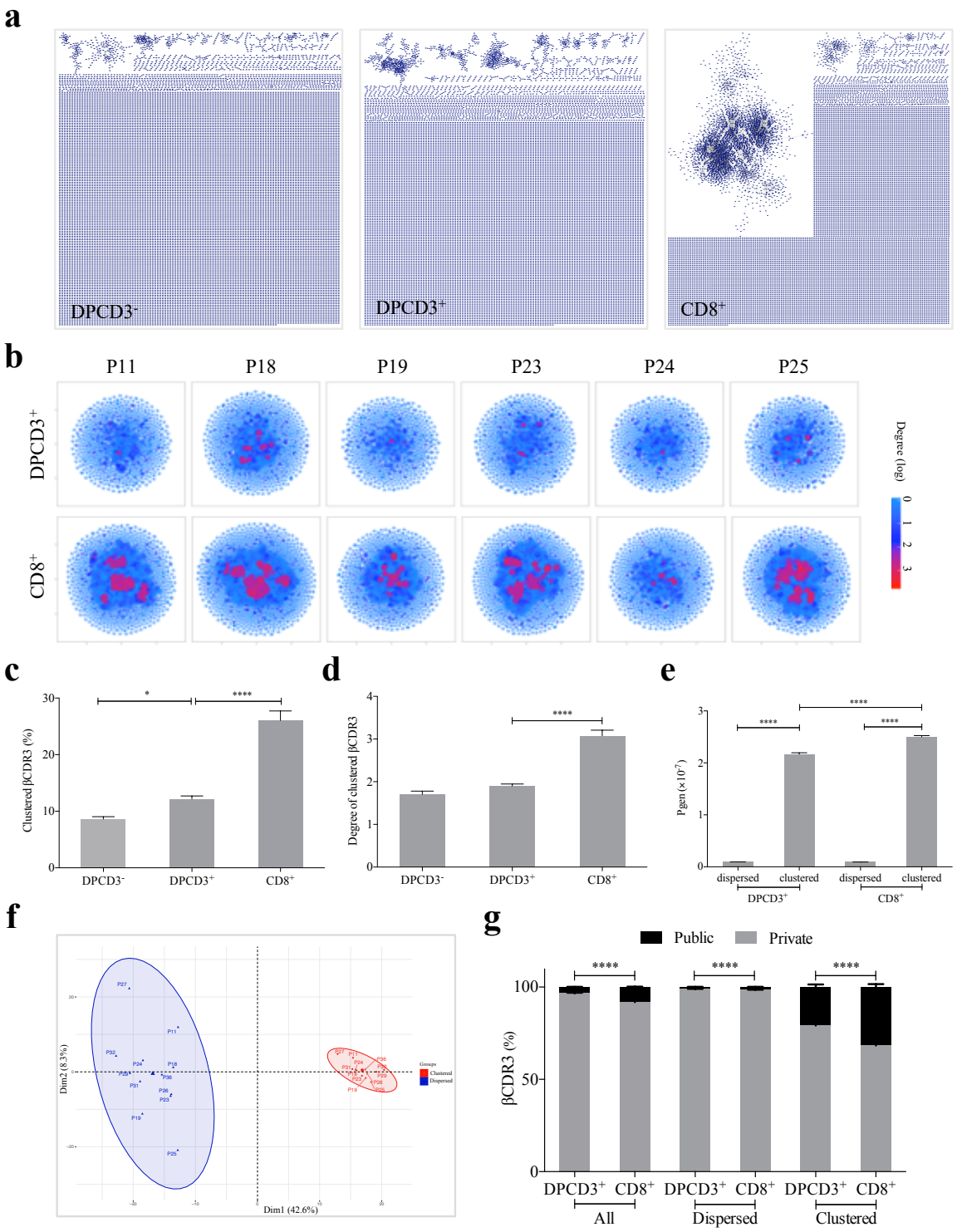


Figure 1. Thymocyte differentiation produces clustered and public CDR3s with high generation probability. Representations and analysis are performed on the first 18,000 most frequent βCDR3s. **a.** Representation of βCDR3 LD≤1 networks from DPCD3$^-$, DPCD3$^+$ and CD8$^+$ thymocytes for one representative donor. Each node represents a single βCDR3. **b.** Node degree (number of connections) of clustered βCDR3 for DPCD3$^+$ and CD8$^+$ thymocytes from 6 donors (P11 to P25). Each node represents a single βCDR3, the colour of which represents its degree (log scale). **c-e.** Data are mean±s.e.m. from two DPCD3$^-$, ten DPCD3$^+$ and twelve CD8$^+$ thymocyte samples; **c.** Percentage of clustered βCDR3s (*p=0.0152 and ****p<0.0001, Mann-Whitney test); **d.** Node degree for clustered βCDR3s

(****p<0.0001, Mann-Whitney test); **e.** Generation probability (Pgen) of dispersed and clustered βCDR3s (****p<0.0001, Mann-Whitney test). **f**. PCA analysis of TRB VJ gene combinations in CD8 thymocytes. Blue: dispersed nodes; Red: clustered nodes. **g.** Mean percentages of public (black) or private (grey) βCDR3s in all, dispersed or clustered nodes. (****p<0.0001, Mann-Whitney test).

The preferential selection of clustered public TCRs that could represent over 8% of the sampled repertoire (Fig. 1g) raises the question of their specificities[13]. As the main function of $CD8^+$ T cells is cytotoxicity towards virally infected cells, we investigated whether the clustered TCRs could be associated with virus recognition. We curated databases of βCDR3s specific for human infectious pathogens[14,15] to retain only those 5,437 that had been identified by tetramer-based selection, i.e. binding to a soluble HLA bound to a defined peptide. We detected an enrichment of these βCDR3s in $CD8^+$ versus $DPCD3^+$ thymocytes (p<0.0001) and in clustered versus dispersed $CD8^+$ thymocytes (p<0.0001) (Fig. 2a and b, Supplementary Figure 10. , Supplementary table 2). Moreover, these virus-specific βCDR3s were significantly enriched within βCDR3s with the highest *Pgen* and node degree (Fig. 2c, Supplementary Figure 11.) and were highly shared between individuals (Fig. 2d).

We aimed to confirm these observations with virus-specific paired α and β TCR chains obtained from single-cell TCR sequencing. These sequences were obtained from 160,914 blood $CD8^+$ T cells isolated from four healthy donors and incubated simultaneously with barcoded dextramers complexed with peptides from CMV, EBV, HIV, HPV, HTLV and influenza[16]. We observed a significant increase in the representation of the virus-specific α and β TCR chains in $CD8^+$ versus $DPCD3^+$ thymocytes (Fig. 2e and f), with a higher representation of the virus-specific alpha chains than that of the beta chains. Noteworthily, these TCRs are also mostly represented in clustered rather than dispersed TCRs from $CD8^+$ thymocytes (Fig. 2g, Supplementary table 3). Altogether, these results indicate that the selection of clustered TCRs with high generation probabilities corresponds, at least in part, to the selection of virus-associated TCRs whose CDR3s are remarkably conserved between individuals independently of their HLA.

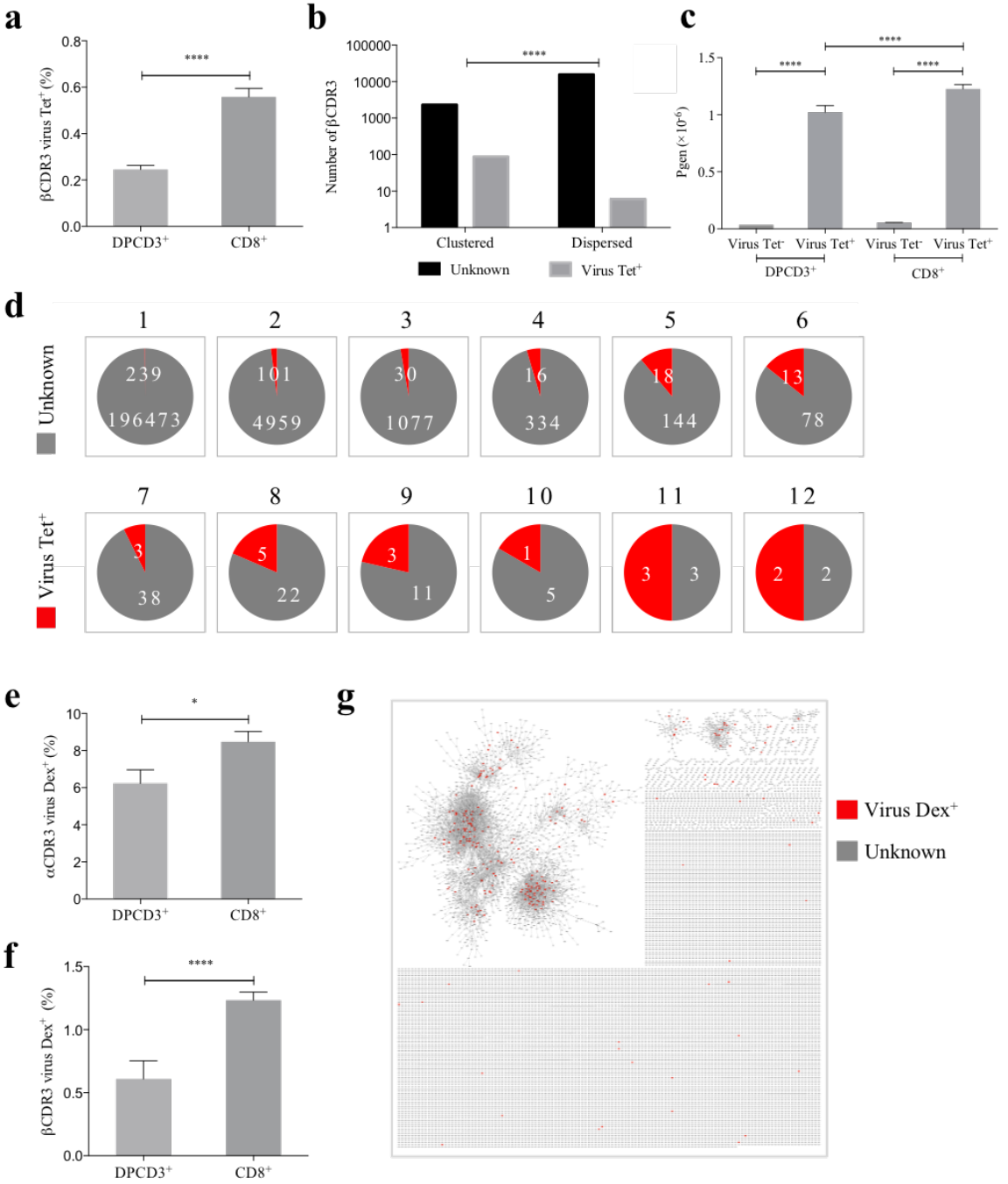


Figure 2. Clustered public TCRs are enriched for virus-specific TCRs.

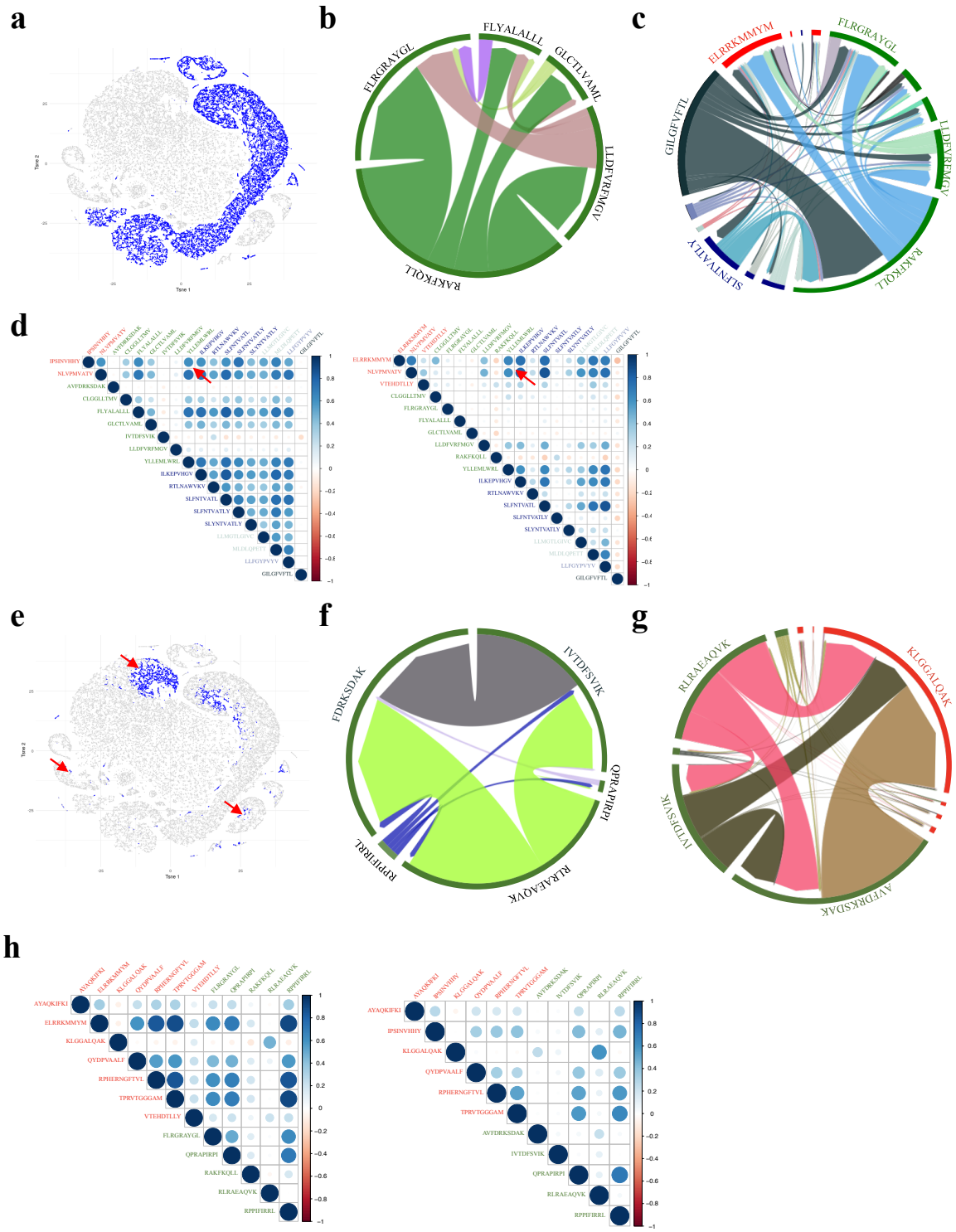
a-d. Analyses of virus-specific βCDR3s from public databases[14,15]. **a.** Mean percentages of virus-specific βCDR3s in DPCD3$^+$ (n=10) vs CD8$^+$ (n=12) thymocytes (****p<0.0001, Mann-Whitney test, mean±s.e.m.). **b.** Virus-specific βCDR3 enrichment in clustered vs dispersed nodes in CD8$^+$ thymocytes from one representative donor (p<0.0001; Chi-square test). Data for all CD8$^+$ thymocytes are in Supplementary table 2. **c.** Mean generation probability of virus-specific βCDR3s in dispersed and clustered DPCD3$^+$ or CD8$^+$ thymocytes (****p<0.0001, Mann-Whitney test, mean±s.e.m.). **d.** Sharing of virus-specific βCDR3s in CD8$^+$ thymocytes. Pie charts represent the βCDR3s from private (1) to shared by all donors (12), in grey for βCDR3s with unknown specificity or in red for those with a virus specificity. **e-g.** Identification of virus-specific TCRs from single-cell sequencing dataset[16]. **e.** Mean percentages of virus-specific αCDR3s (*p=0.0496, Mann-Whitney test, mean±s.e.m.). **f.** Mean percentages of virus-specific βCDR3s (****p<0.0001, Mann-Whitney test, mean±s.e.m.). **g.** Overlay of the βCDR3 network

132 of CD8$^+$ thymocytes from one individual with virus-specific βCDR3s identified in single-cell sequencing
133 datasets[16].
134

135 Intriguingly, we noted that TCRs with different assigned specificities could be detected
136 within close proximity in single clusters (Supplementary Figure 12), although this should
137 suggest similar specificities[6–12] (Supplementary Figure 1.). We thus investigated this
138 observation in greater detail, at the single-cell level[16]. We first analysed the different TCRs
139 assigned to recognise EBV peptides, i.e. binding dextramers (Dex$^+$) matching the HLA of a given
140 individual loaded with EBV peptides (Fig. 3a). Single TCRs were found to bind multiple
141 dextramers loaded with distinct peptides from EBV (Fig. 3b) and some TCRs also bound
142 dextramers loaded with peptides from unrelated viruses such as CMV, HIV, HPV, HTLV-1 or
143 influenza (Fig. 3c); TCRs able to bind dextramers specific for both EBV and CMV were even
144 found in a CMV and EBV seronegative patient (Supplementary Figure 13). TCRs binding multiple
145 viral peptides have binding scores for the different viral peptides that are highly positively
146 correlated (Fig. 3d); for example, the different TCRs that bind a dextramer expressing the
147 "IPSINVHHY" CMV peptide have a strong positive correlation (blue dot) for the binding of a
148 dextramer expressing the "YLLEMLWRL" EBV peptide (Fig. 3d; red arrows), indicating that most
149 of the TCRs that bind one of these peptides have an equivalent binding score for the other.
150 Noteworthily, there is only rare negative correlation (red dots) for the binding of dextramer
151 harbouring different peptides.

152 We also analysed the binding of HLA-mismatched dextramers. When HLA-mismatch
153 EBV Dex$^+$ cells are overlaid on a TSNE representation based on single-cell specificity (Fig. 3e),
154 some of them corresponded to cells also labelled by HLA-matched EBV dextramers (Fig. 3a),
155 while some others did not (Fig. 3e, red arrow). As for the HLA-matched dextramers, single TCRs
156 were found to bind HLA-mismatched dextramers loaded with distinct unrelated peptides from
157 EBV (Fig. 3f) and some TCRs even bound HLA-mismatched dextramers loaded with epitopes
158 from different viruses such as CMV, HIV, HPV, HTLV-1 or influenza (Fig. 3g). There were also
159 mostly positive correlations for the binding to different HLA-mismatched dextramer
160 specificities (Fig. 3h). Altogether, within a dataset of >160,000 single CD8$^+$ T cells, among the
161 66,191 that did bind dextramers, 24,083 could bind more than one viral-derived peptide from
162 either the same or different viruses, and presented by HLA-matched or even HLA-mismatched
163 dextramers. Thus, there are pleiospecific CD8$^+$ T cells (psT cells) whose TCRs are diverse and

bind HLA class-I based dextramers, but are not strictly constrained by HLA matching and the presented peptide. Such binding properties markedly differ from the classical cross-reactivity of TCRs for mimotopes[17] and from those of innate-like MAIT and NKT cells[18,19]. The latter have a restricted diversity, with an invariant TCRα chain and a constrained TCRβ repertoire, and are MR1- or CD1-restricted, respectively[18,19].



171 **Figure 3. Innate-like TCR binding properties.**
172 **a.** TSNE representation of the single-cell TCR specificities from one individual. HLA-matched EBV-Dex$^+$
173 cells are in blue. **b.** Chord diagram showing TCR binding to multiple HLA-matched EBV Dextramers
174 loaded with different peptides. Each segment represents TCR binding to the peptides marked above.
175 The size of the segments corresponds to the number of TCRs binding to these peptides. The link
176 between segments identifies multiple TCR binding to different peptides. **c.** Chord diagram showing TCR
177 binding to HLA-matched dextramers loaded with peptides from unrelated viruses. The colours of the
178 segments represent the different viruses: CMV (red), EBV (green), HIV (dark blue), HPV (light green),
179 HTLV (purple) and influenza (dark grey). The same colour code is used in d, g and h. The full list of the
180 different peptides is in Supplementary table 4. **d.** Correlation between the binding scores for the
181 different HLA-matched virus-specific dextramers from 2 patients. Significant correlations (p-value <
182 0.01, Pearson correlation) are represented by coloured circles. The intensity of the colours and the size
183 of the circles are proportional to the correlation coefficients. Positive correlations are displayed in blue
184 and negative correlations in red. **e.** TSNE representation of the single-cell TCR specificities from one
185 individual. HLA-mismatched EBV-Dex$^+$ cells are in blue. Red arrows highlight cells only labelled by HLA-
186 mismatched EBV dextramers. **f.** Chord diagram showing TCR binding to HLA-mismatched dextramers
187 loaded with peptides from EBV. **g.** Chord diagram showing TCR binding to HLA-mismatched dextramers
188 loaded with peptides from unrelated viruses. The full list of the different peptides is in Supplementary
189 table 5. **h.** Correlation between the binding scores for the different HLA-mismatched virus-specific
190 dextramers from 2 patients.
191

192 These peculiar properties led us to assess the functional relevance of the innate-like
193 recognition of psT cells' TCRs. We first evaluated the in vitro cross-activation of T cells with
194 different peptides. Human effector memory CD8$^+$ T cells (emCD8$^+$) were purified according to
195 their binding of CMV or EBV fluorescent dextramers; the sorted cells were then stimulated by
196 either the peptide that was used to purify them, or by different ones, and their activation was
197 measured by their IFNγ production (Fig. 4a). All T cells were efficiently non-specifically activated
198 by PMA/ionomycin. T cells that did not bind any dextramers could not be stimulated by any
199 peptide. In contrast, dextramer-sorted cells could be activated by their cognate peptide and
200 almost as well by the other (Fig. 4b). Thus, the binding of multiple dextramers appears
201 functionally relevant, translating into proper psT cells activation by multiple unrelated peptides.
202 We then investigated whether we could detect the involvement of psT cells during *in vivo*
203 immune responses. We first analysed individuals vaccinated against yellow fever (YFV) (Fig.
204 4c)[20] or influenza (Flu) (Fig. 4d), identifying their TCR repertoires responding to YFV and Flu
205 using the ALICE[21] and TCRNET algorithms[22,23], respectively. Within these repertoires, we looked
206 for known viral-specific TCR sequences. Besides YFV- and Flu-specific β and α CDR3s (Fig. 4c &
207 d), we could also detect CDR3s of psT cells, i.e. assigned to one or even multiple other viral
208 specificities, notably for CMV and EBV. We then also analysed the TCR repertoire of T cells from
209 bronchoalveolar lavages (BAL) from patients with COVID-19 pulmonary infections[24], i.e. cells

210 responding to the local infection. We observed that many of these BAL T cells have the
211 characteristics of psT cells: their CDR3s (i) could be detected within the clusters of psT cells
212 responding to YFV and Flu infection (Fig. 4 c & d); (ii) they are highly connected to virus-specific
213 sequences from databases (Fig. 4 e & f) and (iii) many harbour a TCR assigned to at least two
214 specificities from CMV, EBV or Flu. Thus, psT cells migrate to the sites of primary antiviral
215 immune responses.
216

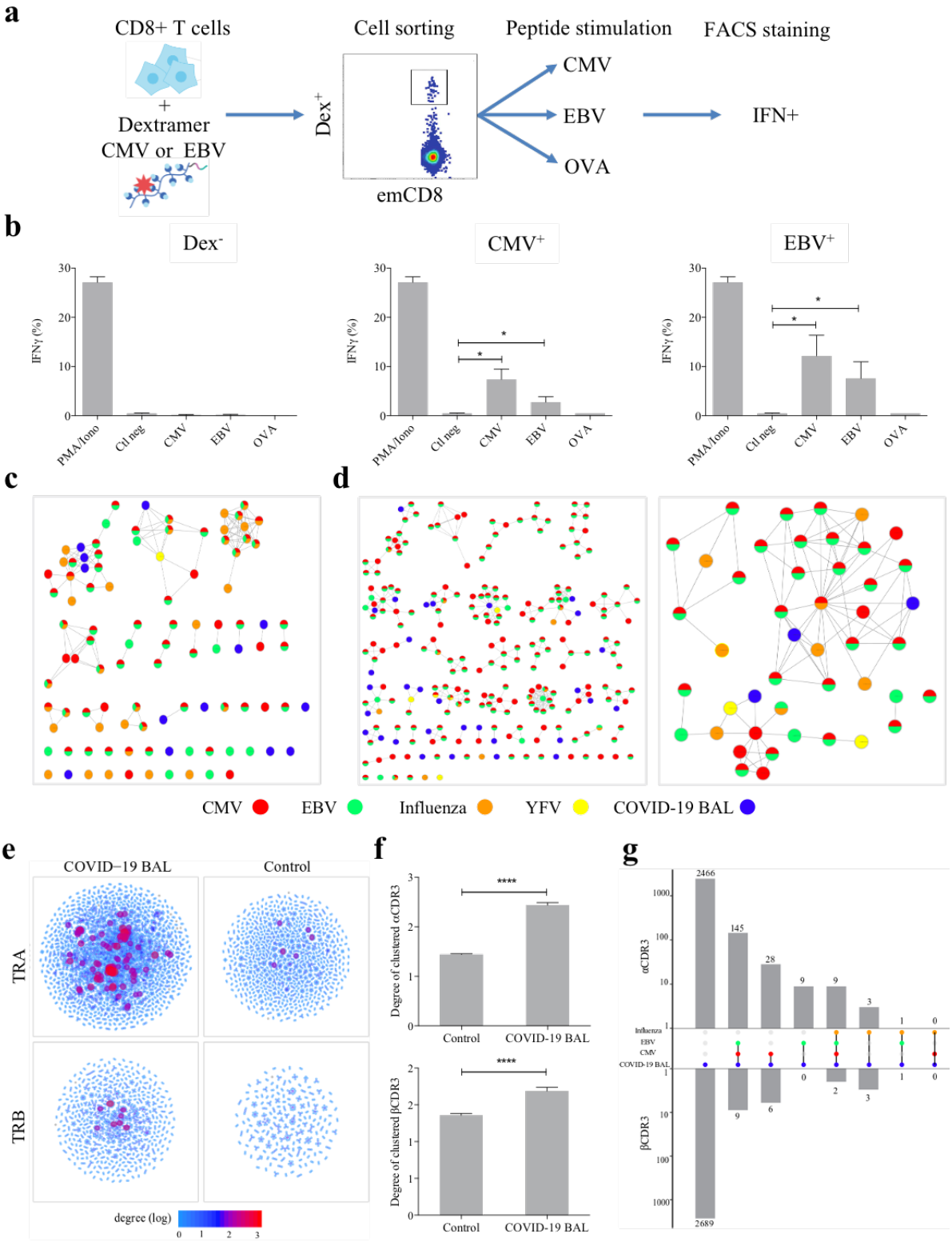


Figure 4. In vitro activation and in vivo recruitment of psT cells. a. Schematic representation of the *in vitro* cross-activation experiment. **b.** In vitro activation of innate-like T cells. Percentage of IFNγ producing emCD8$^+$ cells after activation with PMA/ionomycin (positive control), or CMV, EBV and OVA peptides (mean±s.e.m. *p<0.05, Mann-Whitney test,) **c.** Identification of unconventional T cells responding to yellow fever vaccination[20] (Patient P1 in[20]). The identification of βCDR3s significantly recruited following yellow fever vaccination was performed using the ALICE algorithm[21] and represented as LD≤1 networks. CDR3s assigned to known specificities are coloured as indicated. **d.**

Identification of unconventional T cells responding to influenza vaccination. The identification of α and β CDR3 significantly recruited following influenza vaccination was performed using the TCRNET algorithm[22,23] for one representative individual. **e.** Node degree α and β CDR3 of TCRs from bronchoalveolar lavage (BAL) of COVID-19 patients[24] and from control repertoire clustered with virus-specific α and β CDR3 from databases[14–16]. Each node represents a single CDR3, the colour of which represents its degree (log scale). **f.** Statistical analysis of node degree from **e**. (****p<0.0001, Mann-Whitney test). **g.** Multiple viral specificities of the α and β CDR3s from BAL of COVID-19 patients. Every combination of specificities is represented by the middle-coloured plot (same colour codes as in c & d). The occurrence of each combination is shown for α (top bar plot) and β CDR3s (bottom bar plot).

Our findings have important implications for the study of the adaptive immune response in health, diseases and immunotherapies. They prompt reconsideration of the paradigm of highly diverse adaptive immune repertoires driving a highly antigen-specific antiviral immune response. The immune response may instead proceed through tinkering, as evolution does[25]. For life-threatening situations, the initial recruitment of frequent pleiospecific effector T cells might be more efficient and rapid than having to rely on rare cells with stringent specificity. This would be another mechanism of preparedness of the immune system, reminiscent of the role of (i) other unconventional T cells like MAIT and NKT cells[4], (ii) TCR activation by bacterial superantigen[26] and (iii) natural antibodies specific for microbial determinants[27]. Our findings would also explain the overlooked observation that a very restricted repertoire of only about 1,000 different TCRs arising from a single T cell progenitor was sufficient to cope with viral infections in a child with severe combined immunodeficiency[28].

A fuzzy recognition by pleiospecific TCRs would explain the so-called "heterologous immunity"[29,30] in which T cell responses to one pathogen can have a major impact on the course and outcome of a subsequent infection with an unrelated pathogen[31]. In support of this concept, in humans, (i) there are abundant virus-specific memory-phenotype T cells in unexposed adults[32], (ii) vaccination against measles provides better overall survival independently of measles infection[33] and (iii) CMV infection enhances immune responses to influenza[31]. Individuals' histories of fuzzy immune responses may create "antigenic sins" that might be responsible for the diverse immune responses to viruses, from inapparent infection to fulminant immunopathology[34–36]. Interestingly, heterologous immunity has rarely been linked to B cell/antibody responses, which might thus be the mediators of more specific immune responses. In this regard, it is noteworthy that B cells have a machinery for somatic mutations of their BCRs that ultimately allows them to generate antibodies with increased affinity (specificity) for antigens. While TCR generation and BCR generation share many common mechanisms, the fact that T cells did not evolve to use this available machinery is

260 another indication that T cell recognition could have been selected to be more fuzzy than
261 stringent. Further studies will have to evaluate the contribution of fuzzy immune responses to
262 the efficacy but also the immunopathology of antimicrobial responses and to autoimmunity.

263 Materials and methods:

264

265 Human samples:

266 Thirteen human thymus samples were obtained from organ donors undergoing surgery
267 (Department of Cardiac Surgery, Pitié-Salpêtrière Hospital, France) after approval by the
268 *Agence de Biomédecine* and the *Ministry of Research*. Their age at the time of sampling ranged
269 from 19 to 65 years old. The male-to-female sex ratio was 2.6.

270 For cross-activation experiments, six leukapheresis samples were freshly collected from healthy
271 donors at EFS Paris Saint-Antoine-Crozatier (Etablissement Français du Sang, Paris, France)
272 after informed consent and according to institutional guidelines. Donor selection was based on
273 matching HLA-A2 class I allele.

274 For the influenza vaccination protocol, two unrelated healthy individuals were vaccinated with
275 inactivated influenza vaccine (Influvac Tetra, Mylan) after written informed consent. The blood
276 was collected with informed consent.

277

278 Isolation of thymocytes and extraction of RNA:

279 Single-cell suspensions were prepared from the thymus by mechanical disruption through
280 nylon mesh (cell strainer). Single-cell suspensions from whole thymus were stained with
281 antibodies anti-CD3 (AF700), anti-CD4 (APC), anti-CD8 (FITC). Cells were sorted by fluorescent
282 activated cell sorting (Becton Dickinson™ FACSAria II) with purity >95% to collect populations
283 based on the following labelling: DPCD3$^-$ were gated as CD3$^-$CD4$^+$CD8$^+$, DPCD3$^+$ were gated as
284 CD3$^+$CD4$^+$CD8$^+$ and CD8$^+$ were gated as CD3$^+$CD4$^-$CD8$^+$. RNA was isolated from sorted
285 populations by means of lysis buffer with the RNAqueous-Kit (Invitrogen®) extraction kit,
286 according to the manufacturer's protocol. The RNA concentration and sample integrity were
287 determined on NanoDrop (Thermo Fisher®).

288

289 TCR repertoire library preparation and sequencing

290 T cell receptor (TCR) beta libraries were prepared on 100 ng of RNA from each sample with the
291 SMARTer Human TCR a/b Profiling Kit (Takarabio®) following the provider's protocol. Briefly,
292 the reverse transcription was performed using TRBC reverse primers and further extended with
293 a template-switching oligonucleotide (SMART-Seq® v4). cDNAs were then amplified following
294 two semi-nested PCRs: a first PCR with TRBC and TRAC reverse primers as well as a forward

295 primer hybridising to the SMART-Seqv4 sequence added by template-switching and a second
296 PCR targeting the PCR1 amplicons with reverse and forward primers including Illumina Indexes
297 allowing for sample barcoding. PCR2 were then purified using AMPure beads (Beckman-
298 Coulter®). The cDNA samples were quantified and their integrity was checked using DNA
299 electrophoresis performed on an Agilent 2100 Bioanalyzer System in combination with the
300 Agilent DNA 1000 kit, according to the manufacturer's protocol. Sequencing was performed
301 with Hiseq 2500 (Illumina®) SR-300 protocols using the LIGAN-PM Genomics platform (Lille,
302 France).
303
304 TCR deep sequencing data processing
305 FASTQ raw data files were processed for TRB sequence annotation using MiXCR[37] software
306 (v2.1.10) with RNA-Seq parameters. MiXCR extracts TRBs and provides corrections of PCR and
307 sequencing errors.
308
309 Network generation and representation
310 To construct a network, we computed a distance matrix of pairwise Levenshtein distances
311 between CDR3s using the "stringdist"[38] R package. When two sequences were similar under
312 the defined threshold, LD>1 (i.e., at most one amino acid difference), they were connected and
313 designated as "clustered" nodes. CDR3s with more than one amino acid difference from any
314 other sequences are not connected and were designated as "dispersed" nodes.
315 Layout of networks for Fig. 1b and Supplementary Fig. 9a were obtained by using the graphopt
316 algorithm of the "Igraph"[39] R package and plotted in 2D with "ggplot2" to generate figures[40].
317 Only clustered nodes are represented, edges are not shown and colours represent the node
318 degree (log scale).
319 Layouts of detailed networks in Fig. 1a, Fig. 2g, Fig. 4c, Fig. 4d & Supplementary Fig. 1a,
320 Supplementary Fig. 3 & Supplementary Fig. 12 were done with Cytoscape[41].
321
322 Statistical analysis and visualisation
323 Normalisation was performed by sampling on the top α or β 18,000 CDR3s based on their
324 frequency in each sample. The repertoires with less than 18,000 α or β CDR3s were not
325 included in the statistical analysis. The numbers of samples included in the statistical analysis
326 for the β repertoire were: two for DPCD3$^-$, ten for DPCD3$^+$ and twelve for CD8$^+$. The numbers

327 of samples included in the statistical analysis of the α repertoire were: six for DPCD3$^+$ and ten
328 for CD8$^+$. Statistical tests used to analyse data are included in the figure legends. Comparisons
329 of two groups were done using the Mann-Whitney test (Fig. 1c, Fig. 1d, Fig. 1e, Fig. 1g, Fig. 2a,
330 Fig. 2c, Fig. 2e Fig. 2f, Fig. 4b) and multiple t-test (Supplementary Fig. 5, Supplementary Fig. 9).
331 The correlation coefficient was calculated using the Pearson correlation coefficient
332 (Supplementary Fig. 4). Enrichment of public CDR3s or virus-associated CDR3s was done using
333 the two-tailed Chi-square test with Yate's correction (Fig. 2f, Supplementary table 1, 2 & 3) and
334 the Fisher test (Supplementary Fig. 8, Supplementary Fig. 11). Statistical comparisons and
335 multivariate analyses were performed using Prism (GraphPad Software, La Jolla, CA) and using
336 R software version 3.5.0 (www.r-project.org). PCA was performed on the frequency of VJ
337 combination usage frequency within each donor using the factoextra R package. Tsnes were
338 generated using the binding scores of each cell across all the antigens present in the dataset,
339 disregarding the HLA matching with the donor. The function Rtsne of the homonymous R
340 package[42] was applied with the perplexity parameter set to 10. Correlograms were generated
341 using the cells that have a significant binding score (i.e. >10) for at least one of the virus-specific
342 dextramers tested. The correlation was calculated across all the antigens present on each
343 correlogram (Pearson test). Correlograms were plotted using the corrplot R package.

344

345 Probability of generation calculation

346 The generation probability (Pgen) of a sequence is inferred using the Olga [33] algorithm, which
347 is inferred by IGoR [44], for Fig. 1e, Fig. 2c. IGoR uses out-of-frame sequence information to infer
348 patient-dependent models of VDJ recombination, effectively bypassing selection. From these
349 models, the probability of a given recombination scenario can be computed. The generation
350 probability of a sequence is then obtained by summing over all the scenarios that are
351 compatible with it. We also used OLGA to generate a random repertoire of 500,000 sequences
352 for each α or β repertoire and 3 down-sampling (of unique sequences) to get control repertoire
353 equal to the size of COVID-19 BAL dataset used in fig. 4c, 4d. The control repertoire was
354 parametrized by the predefined genomic templates provided with the package.

355

356 CDR3 connections between individuals

357 In Supplementary Fig. 9, the top 1,500 βCDR3s were sampled from each of the 12 datasets of
358 DPCD3$^+$ and CD8$^+$, then merged to obtain two datasets of 18,000 βCDR3s for DPCD3$^+$ and CD8$^+$.
359 We generated and represented networks, as described above, to investigate the βCDR3 inter-
360 individual network structure.
361

362 **Virus-specific βCDR3 tetramer public databases**

363 The virus-associated CDR3 databases used for the search for specificity was compiled from the
364 most complete previously published McPAS-TCR[14] and VDJdb[15] databases. Virus-associated
365 βCDR3s were selected from the original datasets only when derived from a TCR of sorted CD8
366 T cells that were bound by a specific tetramer. A total of 5,437 such unique tetramer-associated
367 βCDR3s were identified and used. Peptides used for tetramer sorting were from
368 cytomegalovirus (CMV), Epstein-Barr virus (EBV), hepatitis C virus (HCV), herpes simplex virus
369 2 (HSV2), human immunodeficiency virus (HIV), influenza and yellow fever virus (YFV).
370

371 **Virus-specific CDR3 single-cell dextramer public dataset**

372 This dataset contains single-cell alpha/beta TCRs from 160 914 CD8$^+$ T cells isolated from
373 peripheral blood mononuclear cells (PBMCs) from 4 healthy donors. Briefly, 30 dCODE™
374 Dextramer® reagents (Immudex®) with antigenic peptides derived from infectious diseases (9
375 from CMV, 12 from EBV, 1 for influenza, 1 for HTLV, 2 for HPV and 5 for HIV) were
376 simultaneously used to mark cells. Each Dextramer® reagent included a distinct nucleic acid
377 barcode. A panel of fluorescently labelled antibodies was used to sort pure Dextramer®-
378 positive cells within the CD8$^+$ T cell population using an MA900 Multi-Application Cell Sorter
379 (Sony Biotechnology) in a reaction mix containing RT Reagent Mix and Poly dT RT primers. The
380 Chromium Single Cell V(D)J workflow generates single cell V(D)J and Dextramer® libraries from
381 amplified DNA derived from Dextramer®-conjugated barcode oligonucleotides, which are
382 bound to TCRs. Chromium Single Cell V(D)J enriched libraries and cell surface protein libraries
383 were quantified, normalised, and sequenced according to the user guide for Chromium Single
384 Cell V(D)J reagent kits with feature barcoding technology for cell surface protein. We used this
385 dataset to study the presence of multiple specificities in TCR and CDR3. There were 139,378
386 unambiguous TCRs (with only one α and one β chain). We set the threshold defining positive
387 binding at UMI counts greater than 10 for any given dextramer. This identified 15,195 unique
388 virus-specific TCRs with at least one binding.

389

390 Single-cell sequencing of TCRs from bronchoalveolar lavages from COVID-19 patients

391 This dataset contains single-cell alpha/beta TCRs from T cells isolated from bronchoalveolar
392 lavage (BAL) of 9 patients infected by COVID-19[24]. We excluded the TCRs from cells in which
393 more than 1 CDR3aa alpha and 1 CDR3aa beta were detected.

394

395 Cross-activation experiment

396 PBMCs were separated on Ficoll gradient. CD8+ T cells were isolated from PBMCs by positive
397 isolation using the DYNABEADS® CD8 Positive Isolation Kit (Thermo Fisher Scientific) according
398 to the manufacturer's instructions. emCD8 T cells were purified after staining with CD3-AF700,
399 CD8-KO, CD45RA-PeCy7 according to the manufacturer's instructions. The samples were also
400 stained either with CMV pp65 NLVPMVATV or with EBV BMLF-1 GLCTLVAML PE-conjugated
401 Dextramers (Immudex®). emCD8$^+$Dex$^+$ cells were sorted by FACS (FACS Aria II®; BD Biosciences)
402 with a purity >95%. Sorted cells were cultured at a maximum of 5×10^5 cells/mL in round-
403 bottom 96-well plates in RPMI 1640 medium supplemented with 10% FCS, 1%
404 penicillin/streptomycin and glutamate at 37°C with 5% CO_2. *In vitro* stimulation was performed
405 24 hours after cell sorting. Sorted cells were stimulated for 6 hours with either nothing or 1
406 µg/mL of SIINFEKL ovalbumin peptide (OVA), NLVPMVATV cytomegalovirus pp65 peptide
407 (CMV) or GLCTLVAML Epstein-Barr virus BMLF-1 peptide (Ozyme®). The positive control (Ctl
408 PMA/Iono) was performed with 50 ng/mL phorbol myristate acetate (PMA) and 1 mM
409 ionomycin. Intracellular IFN-γ production with an IFN-γ-FITC antibody (BD Pharmingen) was
410 detected in the presence of Golgi-Plug (BD Pharmingen®) after fixation and permeabilisation
411 (BD Cytofix/Cytoperm). Data were acquired using a Navios flow cytometer and analysed with
412 Kaluza analysis software (Beckman Coulter).

413

414 Influenza vaccination protocol

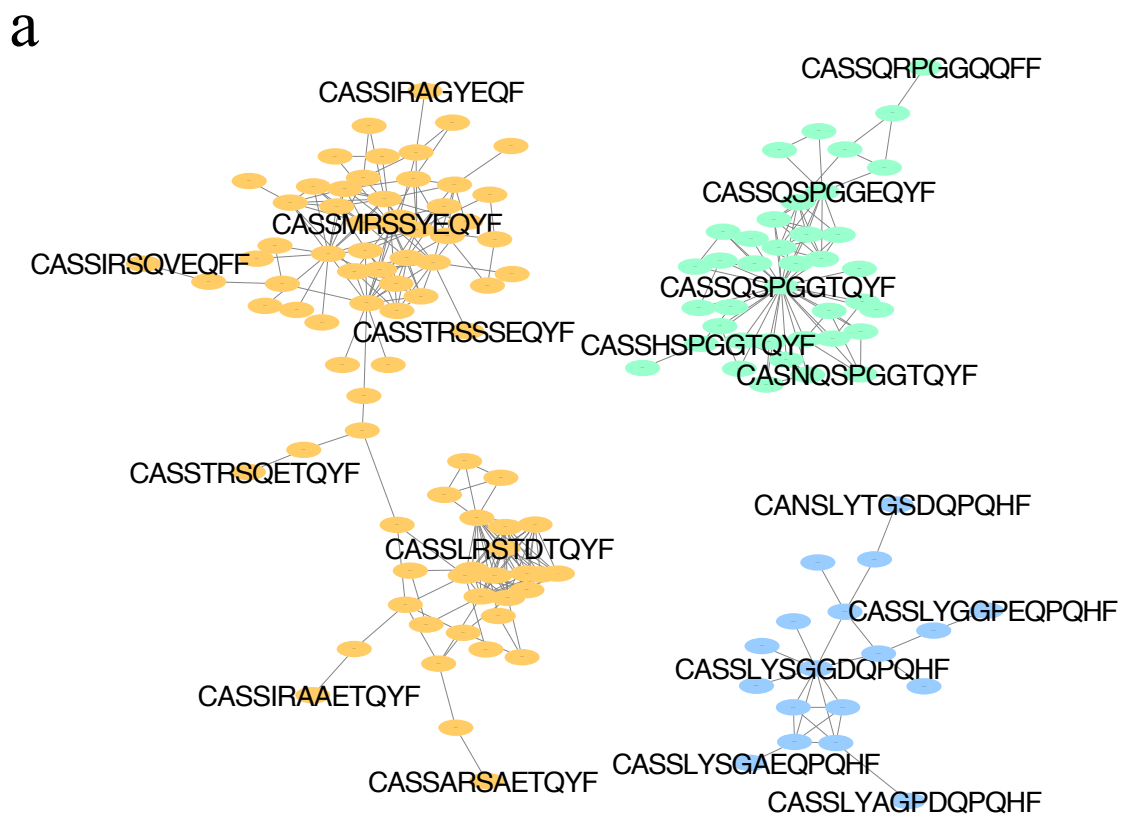
415 Peripheral blood was obtained before vaccination and on day 14 after vaccination. PBMCs were
416 stained with antibodies anti-CD3 (AF700), anti-CD8 (FITC), CD45RA (PeCy7), CCR7 (BV421). Cells
417 were sorted by fluorescent activated cell sorting (Becton Dickinson™ FACSAria II) with purity
418 >95% to collect populations based on the following labelling: naïve CD8 were gated as CD3$^+$CD8$^-$
419 CD45RA$^+$CCR7$^+$ and effector memory CD8 were gated as CD3$^+$CD8$^-$CD45RA$^-$CCR7$^-$. RNA

420 extraction, library preparation, sequencing and raw data processing were performed as
421 described above.
422
423 Identification of TCR enrichment after vaccination
424 Homologous groups of TCRs that are specifically recruited during an antigen-specific response
425 following yellow fever or influenza vaccination were sought using with the previously published
426 algorithms ALICE[21] and TCRNET[22,23], respectively. The difference between the two algorithms
427 is that ALICE uses the VDJ rearrangement model as a control[45] while TCRNET uses real control
428 samples as background.
429

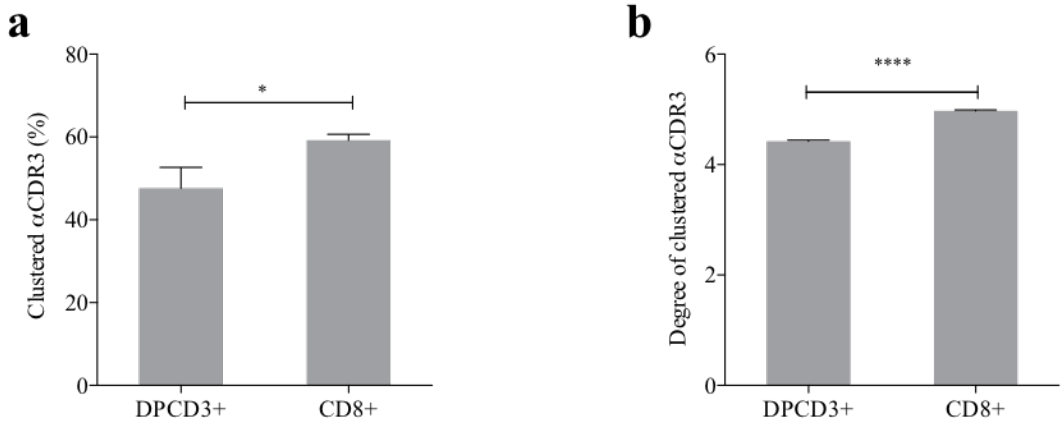
1. Bradley, P. & Thomas, P. G. Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. *Annu. Rev. Immunol.* **37**, 547–570 (2019).
2. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci.* **111**, 13139–13144 (2014).
3. Nikolich-Žugich, J., Slifka, M. K. & Messaoudi, I. The many important facets of T-cell repertoire diversity. *Nat. Rev. Immunol.* **4**, 123–132 (2004).
4. Godfrey, D. I., Uldrich, A. P., McCluskey, J., Rossjohn, J. & Moody, D. B. The burgeoning family of unconventional T cells. *Nat. Immunol.* **16**, 1114–1123 (2015).
5. Rossjohn, J. *et al.* T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).
6. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
7. 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).
8. Madi, A. *et al.* T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife* **6**, (2017).
9. Chen, G. *et al.* Sequence and Structural Analyses Reveal Distinct and Highly Diverse Human CD8 + TCR Repertoires to Immunodominant Viral Antigens. *Cell Rep.* **19**, 569–583 (2017).
10. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
11. Qi, Q. *et al.* Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Sci. Transl. Med.* **8**, 332ra46-332ra46 (2016).
12. Meysman, P. *et al.* On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics* **35**, 1461–1468 (2019).
13. Thomas, P. G. & Crawford, J. C. Selected before selection: A case for inherent antigen bias in the T-cell receptor repertoire. *Curr. Opin. Syst. Biol.* **18**, 36–43 (2019).
14. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
15. Shugay, M. *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* **46**, D419–D427 (2018).
16. 10x_AN047_IP_A_New_Way_of_Exploring_Immunity_Digital (1).pdf.
17. Nelson, R. W. *et al.* T Cell Receptor Cross-Reactivity between Similar Foreign and Self Peptides Influences Naive Cell Population Size and Autoimmunity. *Immunity* **42**, 95–107 (2015).
18. Toubal, A., Nel, I., Lotersztajn, S. & Lehuen, A. Mucosal-associated invariant T cells and disease. *Nat. Rev. Immunol.* **19**, 643–657 (2019).
19. Mori, L., Lepore, M. & De Libero, G. The Immunology of CD1- and MR1-Restricted T Cells. *Annu. Rev. Immunol.* **34**, 479–510 (2016).
20. Pogorelyy, M. V. *et al.* Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12704–12709 (2018).
21. Pogorelyy, M. V. *et al.* Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLOS Biol.* **17**, e3000314 (2019).
22. Shugay, M. *et al.* VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Comput. Biol.* **11**, e1004503 (2015).
23. Ritvo, P.-G. *et al.* High-resolution repertoire analysis reveals a major bystander activation of Tfh and Tfr cells. *Proc. Natl. Acad. Sci.* **115**, 9604–9609 (2018).

24. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* (2020) doi:10.1038/s41591-020-0901-9.
25. Jacob, F. (1977). Evolution and tinkering. Science, 196(4295), 1161-1166.
26. Hayday, A. C. & Vantourout, P. The Innate Biologies of Adaptive Antigen Receptors. *Annu. Rev. Immunol.* **38**, annurev-immunol-102819-023144 (2020).
27. Panda, S. & Ding, J. L. Natural Antibodies Bridge Innate and Adaptive Immunity. *J. Immunol.* **194**, 13–20 (2015).
28. Bousso, P. *et al.* Diversity, functionality, and stability of the T cell repertoire derived in vivo from a single human T cell precursor. *Proc. Natl. Acad. Sci.* **97**, 274–278 (2000).
29. Welsh, R. M. & Selin, L. K. No one is naive: the significance of heterologous T-cell immunity. *Nat. Rev. Immunol.* **2**, 417–426 (2002).
30. Sewell, A. K. Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).
31. Furman, D. *et al.* Cytomegalovirus infection enhances the immune response to influenza. *Sci. Transl. Med.* **7**, 281ra43-281ra43 (2015).
32. Su, L. F., Kidd, B. A., Han, A., Kotzin, J. J. & Davis, M. M. Virus-Specific CD4+ Memory-Phenotype T Cells Are Abundant in Unexposed Adults. *Immunity* **38**, 373–383 (2013).
33. Aaby, P. *et al.* Non-specific effects of standard measles vaccine at 4.5 and 9 months of age on childhood mortality: randomised controlled trial. *BMJ* **341**, c6495–c6495 (2010).
34. Peteranderl, C., Herold, S. & Schmoldt, C. Human Influenza Virus Infections. *Semin. Respir. Crit. Care Med.* **37**, 487–500 (2016).
35. Bertoletti, A. & Ferrari, C. Adaptive immunity in HBV infection. *J. Hepatol.* **64**, S71–S83 (2016).
36. Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. *Nat. Rev. Immunol.* (2020) doi:10.1038/s41577-020-0311-8.
37. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
38. van der Loo, M. P. J. The stringdist package for approximate string matching. *The R Journal, 6(1), 111-122.*
39. Csardi, G. & Nepusz, T. The igraph software package for complex network research.
40. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).
41. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
42. Krijthe, J. H. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R Package Version 013 URL Httpsgithub ComjkrijtheRtsne* (2015).
43. Sethna, Z., Elhanati, Y., Jr, C. G. C. & Mora, T. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. 8.
44. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, (2018).
45. Murugan, A., Mora, T., Walczak, A. M. & Callan, C. G. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci.* **109**, 16161–16166 (2012).

SUPPLEMENTARY MATERIALS



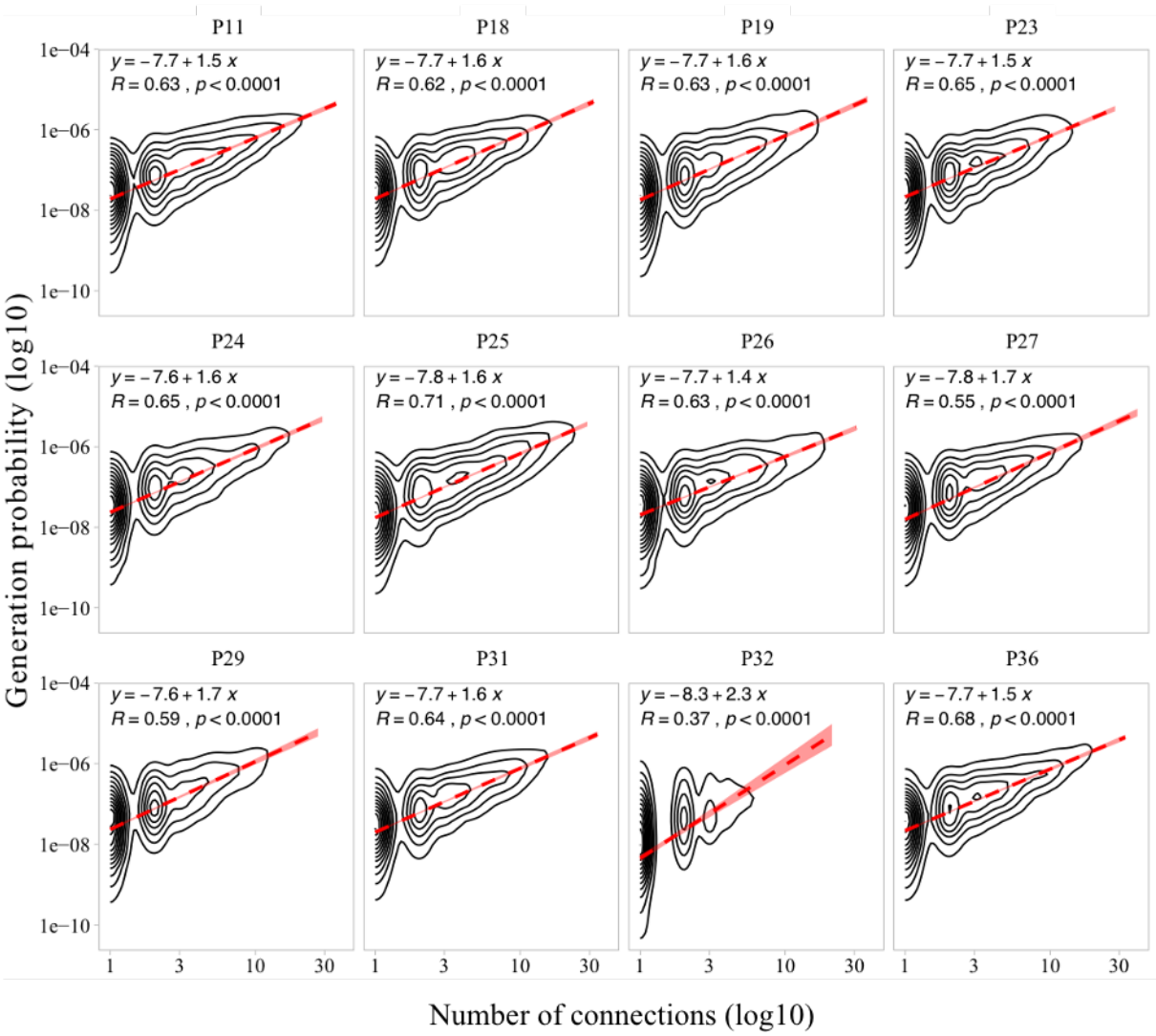
Supplementary Figure 1. TCRs with the same specificity form clusters. a. Networks of βCDR3s specific for GILGFVFTL from influenza (orange), GLCTLVAML from Epstein-Barr virus (green) and FPRPWLHGL from human immunodeficiency virus (blue) are shown. These βCDR3s are from TCRs identified on CD8 T lymphocytes isolated with class I tetramer[14,15] loaded with the indicated peptides[14,15]. Each node represents a clonotype. Two different clonotypes are connected if their βCDR3s differ by at most one amino acid (LD≤1).



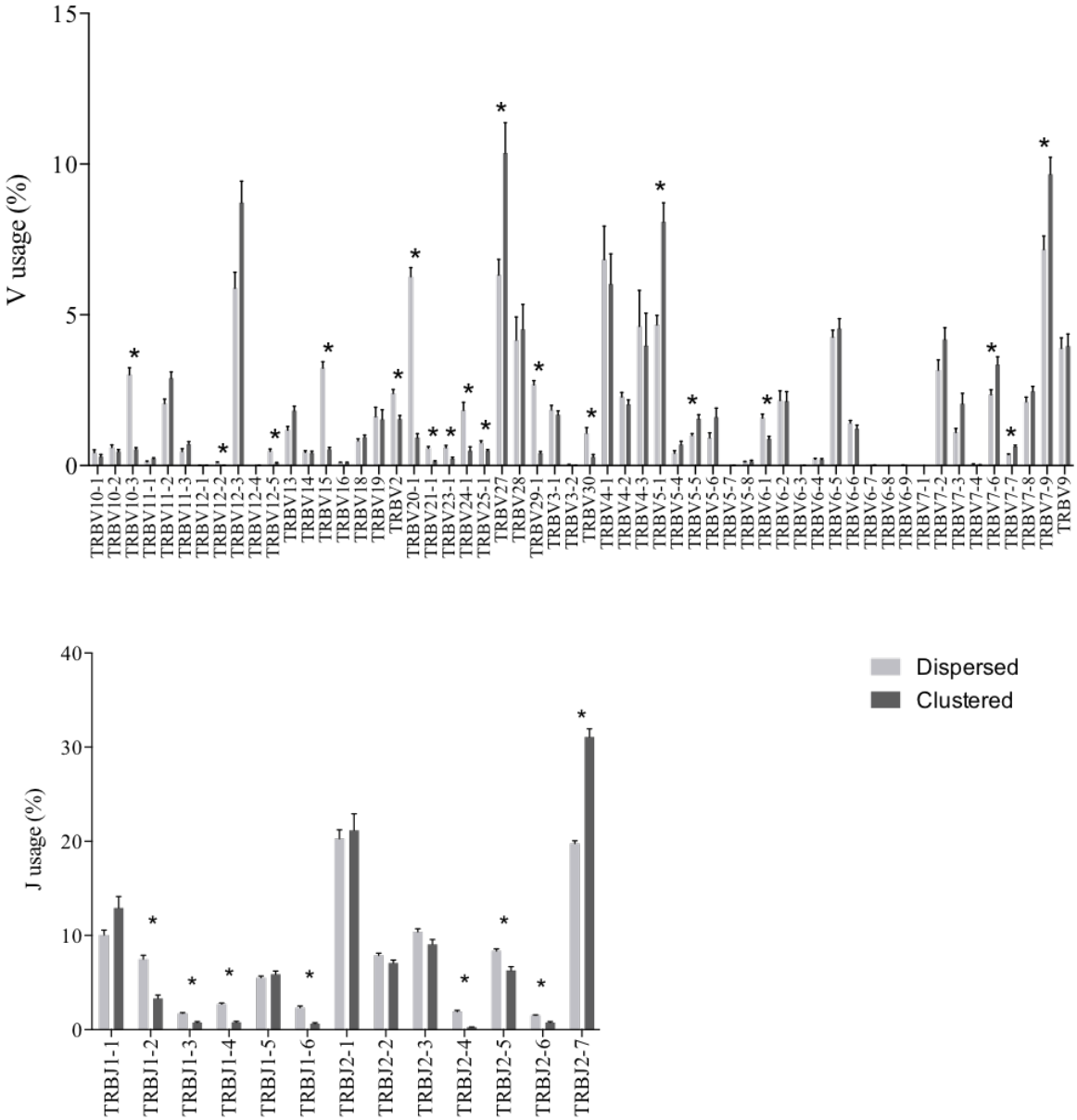
535
536 **Supplementary Figure 2. Clustered αCDR3s from DPCD3⁺ and CD8⁺ thymocytes.** Analyses were
537 performed on the first 18,000 most frequent αCDR3s per sample (n=6 for DPCD3⁺ and n=10 for
538 CD8⁺ thymocytes). **a.** Percentage of clustered αCDR3s. (mean±s.e.m., *p=0.016, Mann-
539 Whitney test). **b.** Degree of clustered αCDR3s. (mean±s.e.m., ****p<0.0001, Mann-Whitney
540 test).

541

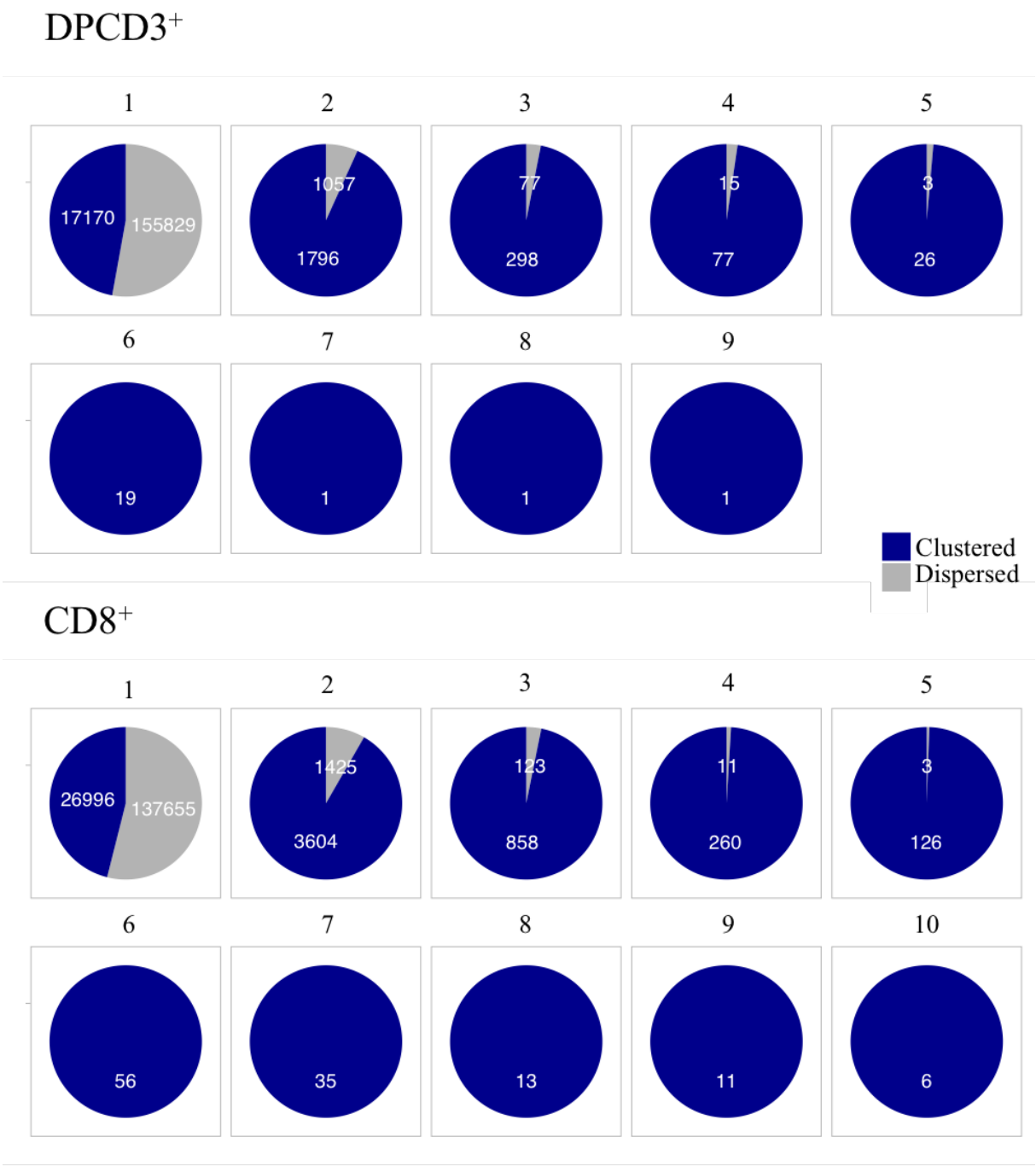
Supplementary Figure 3. βCDR3 network during thymopoiesis. Representation of the 18,000 most frequent βCDR3 networks from DPCD3⁺ and CD8⁺ thymocytes of eight donors (Pn).



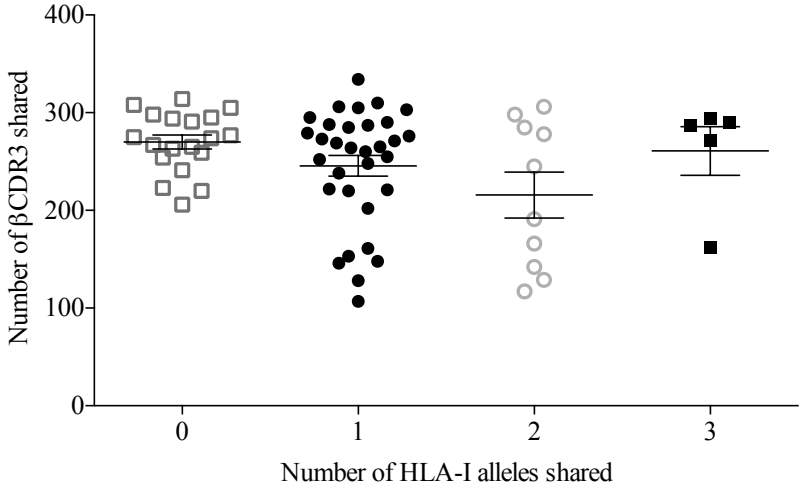
Supplementary Figure 4. Correlation between *Pgen* and βCDR3 number of connections in the CD8⁺ thymocyte repertoire. The contour plots represent the generation probability as a function of βCDR3 connections in the CD8⁺ thymocytes for donors P11 to P36. Linear regression curves between *Pgen* and number of connections are represented as red dashed lines ("y" represents the regression curve's equation). The Pearson correlation coefficient "R" and p-value "p" are calculated for each individual.



Supplementary Figure 5. Clonogram representation of TCR Vβ and Jβ usage in clustered versus dispersed CD8⁺ thymocytes. The bar plots represent the mean percentage of TCR Vβ (up) and Jβ (down) in dispersed (light grey) versus clustered (dark grey). (* $p<0.01$, multiple t-test).



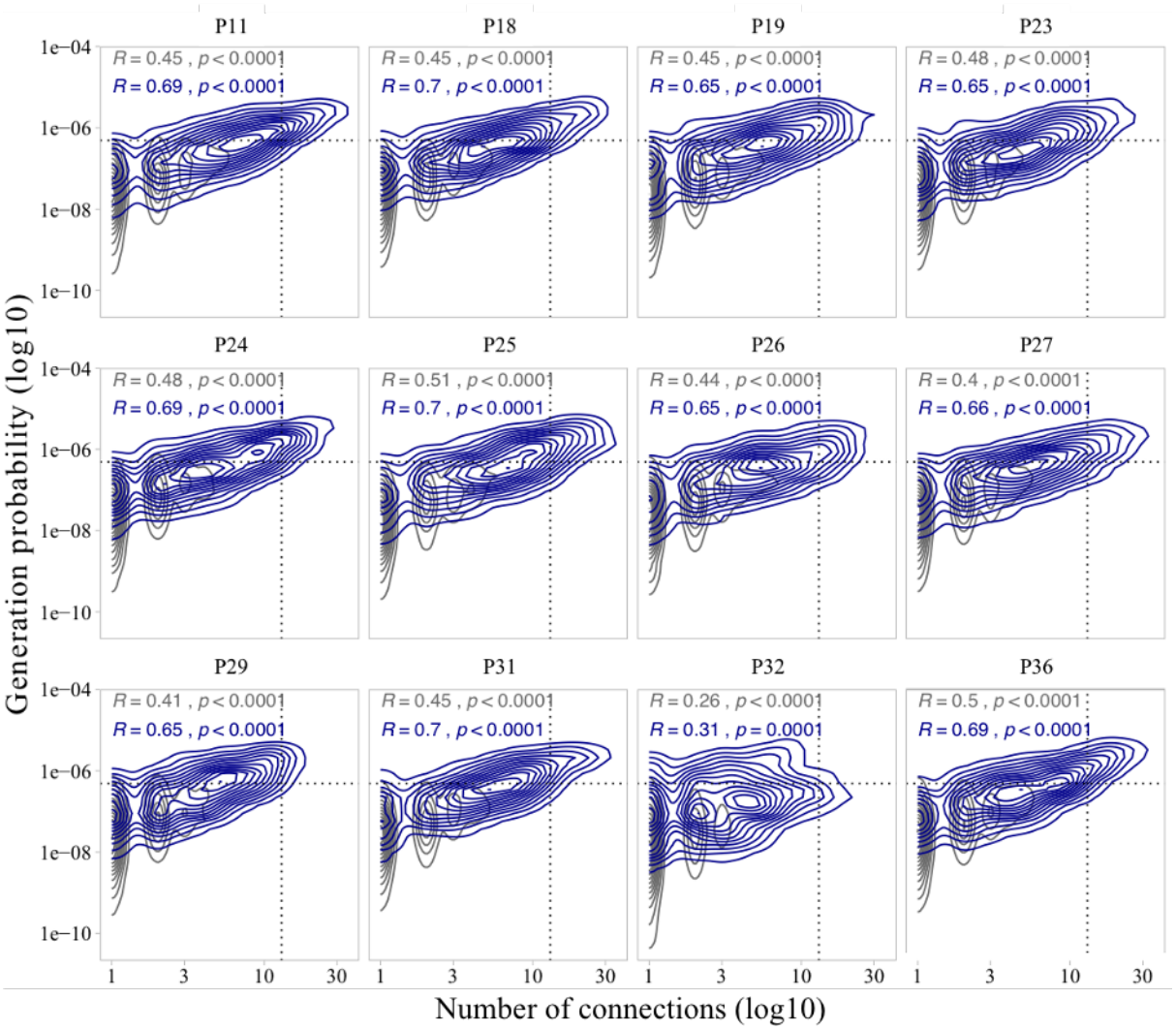
Supplementary Figure 6. βCDR3 sharing between individuals. Pie charts represent the sharing between individuals before (DPCD3$^+$) and after thymic selection (CD8$^+$). Colours represent the dispersed (grey) or clustered (blue) CDR3s. Sharing was analyzed within the 10 donors for which there were at least 18,000 βCDR3s in DPCD3$^+$ and in CD8$^+$ thymocytes.



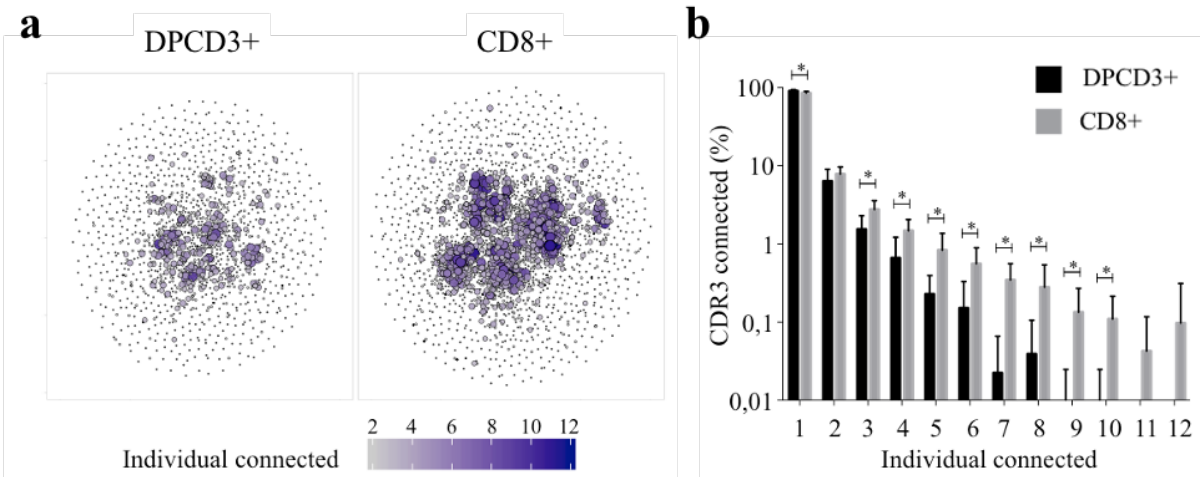
562

563 **Supplementary Figure 7. The number of public βCDR3s in CD8⁺ thymocytes is independent of**
564 **the number of HLA-I alleles shared.** Each dot represents the number of βCDR3s shared between
565 two donors in the first 18,000 CD8⁺ thymocytes. There is no significant difference in the number
566 of public βCDR3s according to the number of HLA-I alleles shared. The number of public βCDR3s
567 is independent of the number of HLA alleles shared.

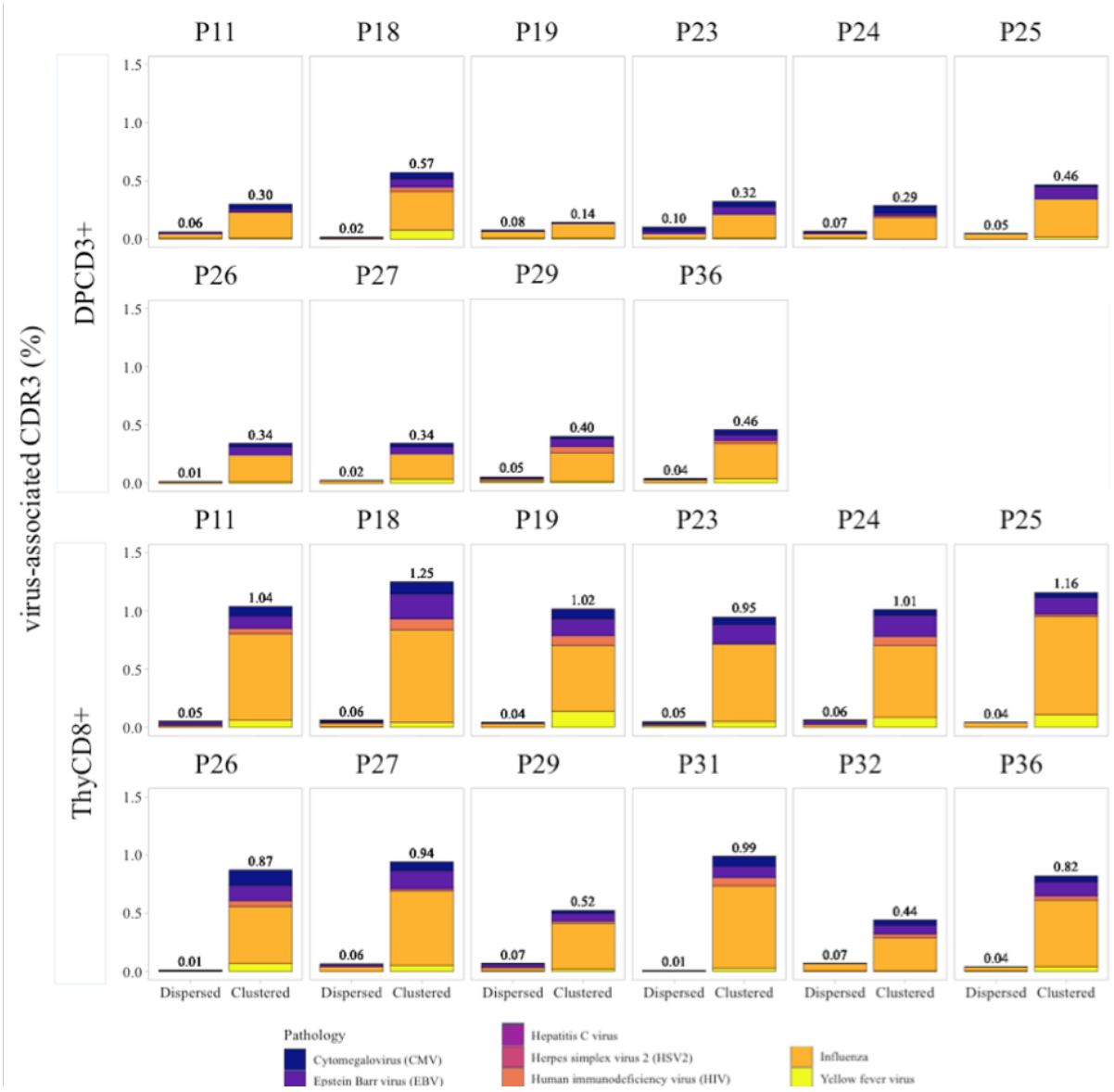
568
569



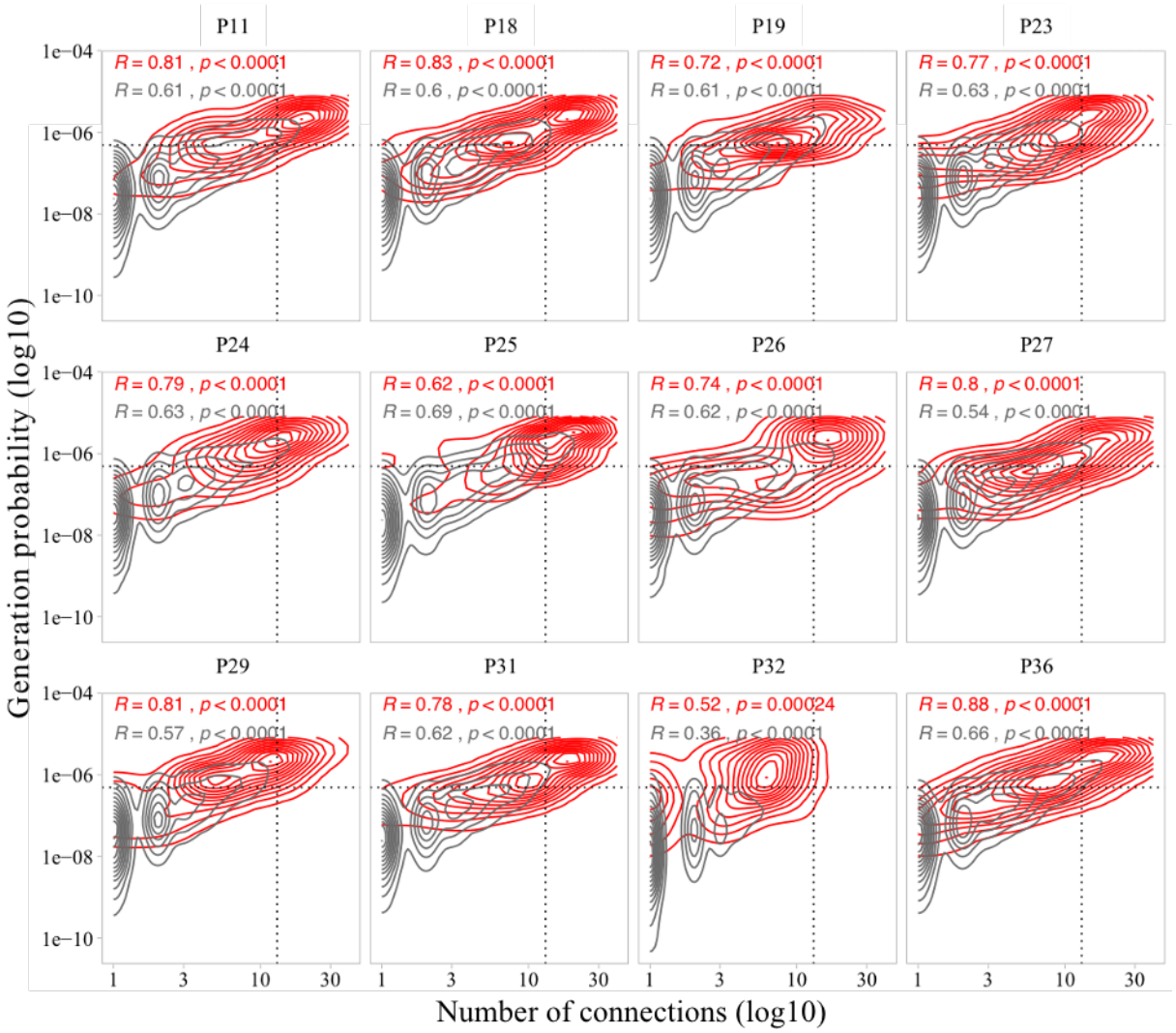
Supplementary Figure 8. Enrichment of public βCDR3s in the CD8⁺ thymocyte repertoires. Representation of the generation probability as a function of βCDR3 connections in individuals (Pn). The contour plots represent shared (blue) or private (grey) βCDR3s. The Pearson correlation coefficient "R" and p-value "p" are calculated for each group. The black dotted lines delimit the threshold for the 2.5% sequences with the higher *Pgen* and connection. βCDR3s with both the highest *Pgen* and connections are also the most public for 12 out of 12 individuals (p<0.0001, two-tailed Fisher test).



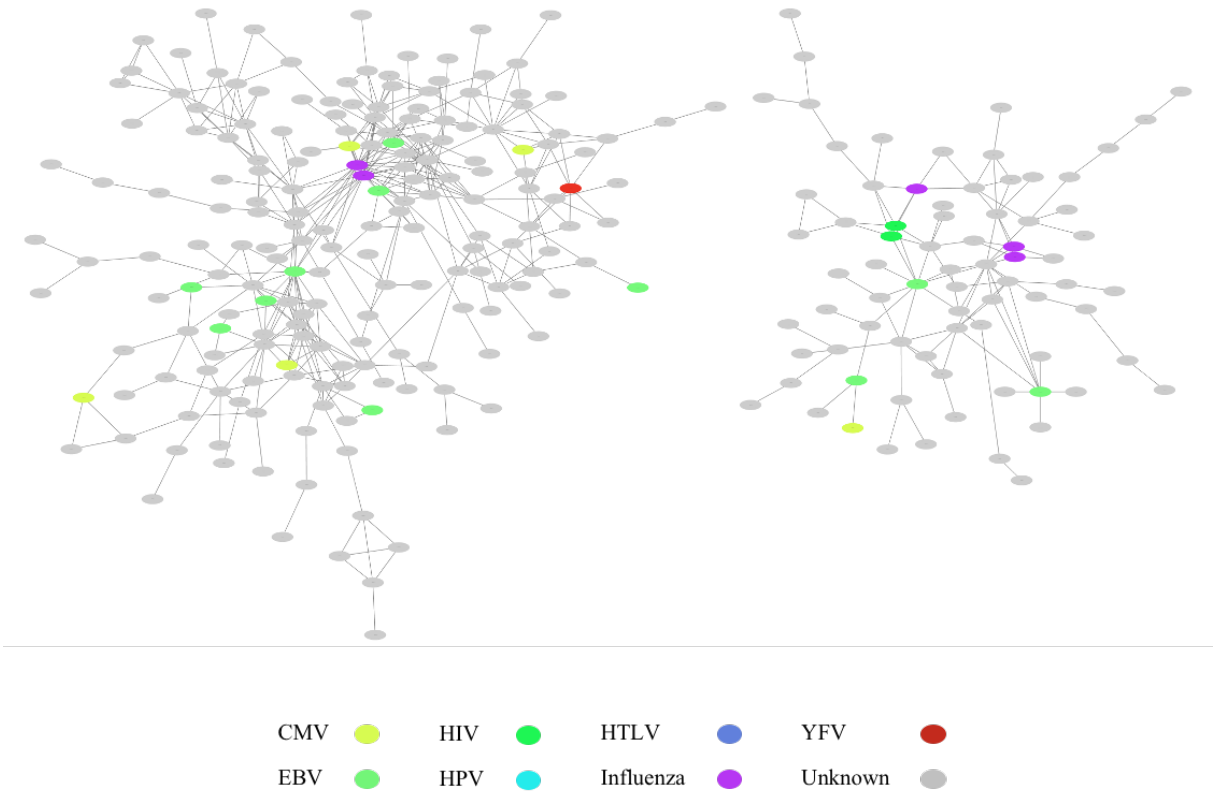
Supplementary Figure 9. Convergence of public βCDR3 specificities during thymopoiesis. a. CDR3 connections between individuals. The top 1,500 βCDR3s were sampled from DPCD3$^+$ (left) and CD8$^+$ (right) cells from each individual and pooled. The CDR3s are clustered based on LD≤ 1 with colour and size both representing the level of sharing between individuals for each CDR3. **b.** Bar plots representing the percentage of CDR3s from an individual that are connected to CDR3s of other individuals, for DPCD3$^+$ and CD8$^+$ thymocytes. The first two bars represent CDR3s that are not connected (n=1). The number of unconnected nodes in DPCD3$^+$ is higher than in CD8$^+$ (*p=0.002). The other bars represent the percentage of CDR3s connected between individuals. The number of nodes connected to 3 to 10 individuals is significantly higher in CD8$^+$ than in DPCD3$^+$ cells (*p<0.01, multiple t-test).



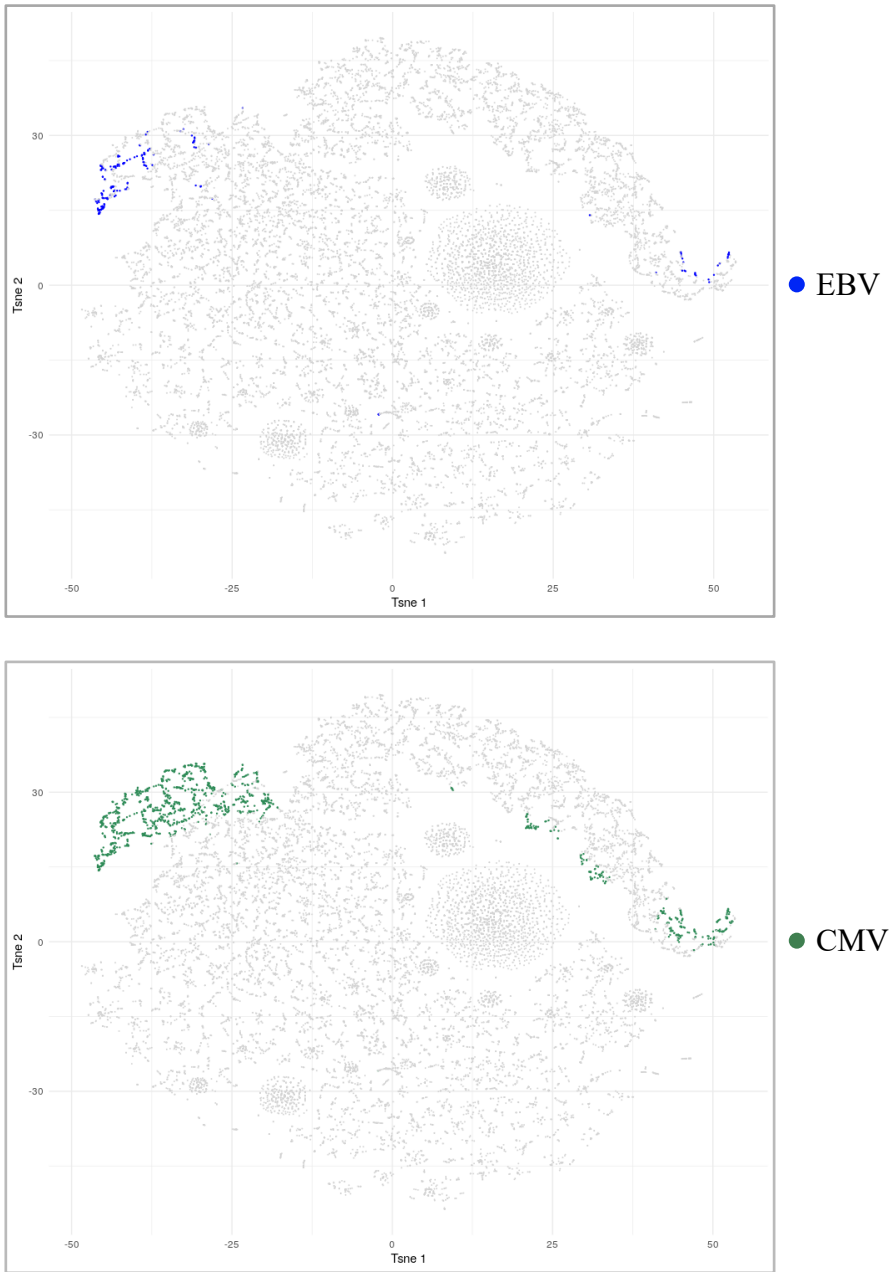
Supplementary Figure 10. Virus-specific CDR3s among DPCD3⁺ and CD8⁺ thymocytes. Bar plots represent the percentage, within the top 18,000 βCDR3s of each donors, of βCDR3s from TCRs identified as virus-specific based on tetramer identification[14,15]. For each panel, the percentage is calculated within dispersed (left boxplot) or clustered (right boxplot) βCDR3s. Colours correspond to different viral specificities.



Supplementary Figure 11. Enrichment of virus-specific βCDR3s in the CD8⁺ thymocyte repertoire. Representation of the generation probability as a function of βCDR3 connections in individuals (Pn). The contour plots represent βCDR3s from TCRs identified as virus-specific based on tetramer identification[14,15] (red) or with unknown specificity (grey). The Pearson correlation coefficient "R" and p-value "p" are calculated for each group. The black dotted lines delimit the threshold for the 2.5% sequences with both higher *Pgen* and degree of connection. βCDR3s with both the highest *Pgen* and connections were also the most virus-specific for 11 out of 12 individuals (p-value <0.0001, two-tailed Fisher test).



Supplementary Figure 12. Single clusters of βCDR3s from CD8⁺ thymocytes comprise TCRs with different viral specificities. The tagged dots represent βCDR3s with known specificity according to the public database[14,15]. βCDR3 specific for unrelated virus epitopes can be found close to or even directly linked in the cluster.



614
615
616 **Supplementary Figure 13. TSNE representation of the single-cell** EBV and CMV **TCR specificities**
617 **from one** seronegative **individual**. CD8$^+$ T cells are able to bind both EBV and CMV HLA-matched
618 dextramers. EBV Dex$^+$-specific cells are in blue and CMV Dex$^+$-specific cells are in green. Some
619 of the cells are specific for both.

620

		DP	CD8	p-value	Odds ratio
P11	Public	590	1630	p<0.0001	0.3403
	Private	17410	16370		
P18	Public	695	1567	p<0.0001	0.4212
	Private	17305	16433		
P19	Public	485	1489	p<0.0001	0.3071
	Private	17515	16511		
P23	Public	649	1489	p<0.0001	0.4148
	Private	17351	16511		
P24	Public	578	1416	p<0.0001	0.3886
	Private	17422	16584		
P25	Public	640	1536	p<0.0001	0.3952
	Private	17360	16464		
P26	Public	584	1470	p<0.0001	0.3771
	Private	17416	16530		
P27	Public	640	1583	p<0.0001	0.3823
	Private	17360	16417		
P29	Public	655	1583	p<0.0001	0.3916
	Private	17345	16417		
P36	Public	653	1468	p<0.0001	0.4239
	Private	17347	16532		

Supplementary table 1. Enrichment of public βCDR3s in CD8$^+$ thymocytes vs DPCD3$^+$. Contingency table for the Chi-square analysis performed with Yates' correction to test the null hypothesis of independence between the sharing of βCDR3s (Public or Private) vs the cell phenotype (DPCD3$^+$ and CD8$^+$). We performed this test in the 10 donors for which we have 18,000 βCDR3s in both DPCD3$^+$ and CD8$^+$ thymocytes. A βCDR3 is defined as public if it is found at least once in the 18,000 βCDR3s of the same cell phenotype from other donors. The results (p-value < 0.0001) rejected the null hypothesis, thereby indicating the interdependency of the two variables.

		Clustered	Dispersed	p-value	Odds ratio
P11	Virus Tet+	106	8	p<0.0001	40.75
	Unknown	4389	13497		
P18	Virus Tet+	126	10	p<0.0001	39.93
	Unknown	4285	13579		
P19	Virus Tet+	102	8	p<0.0001	46.77
	Unknown	3832	14058		
P23	Virus Tet+	93	7	p<0.0001	47.82
	Unknown	3892	14008		
P24	Virus Tet+	96	9	p<0.0001	40.38
	Unknown	3739	14156		
P25	Virus Tet+	112	7	p<0.0001	63.60
	Unknown	3594	14287		
P26	Virus Tet+	83	2	p<0.0001	146.1
	Unknown	3963	13952		
P27	Virus Tet+	99	8	p<0.0001	34.60
	Unknown	4714	13179		
P29	Virus Tet+	55	9	p<0.0001	31.65
	Unknown	2903	15033		
P31	Virus Tet+	104	1	p<0.0001	360.8
	Unknown	4004	13891		
P32	Virus Tet+	44	9	p<0.0001	32.93
	Unknown	2320	15627		
P36	Virus Tet+	86	6	p<0.0001	54.12
	Unknown	3750	14158		

Supplementary table 2. Enrichment of virus-specific βCDR3s from databases[14,15] in clustered CD8⁺ thymocytes. Contingency table for the Chi-square analysis performed with Yates' correction to test the null hypothesis of independence between the specificity of βCDR3s (Virus Tet⁺ and Unknown specificities) vs the connection of βCDR3 ("clustered" and "dispersed") in all the CD8⁺ thymocytes from 12 donors. The results (p-value < 0.0001) rejected the null hypothesis, thereby indicating the interdependency of the two variables.

		Clustered	Dispersed	p-value	Odds ratio
P11	Virus Dex+	223	22	p<0.0001	31.99
	Unknown	4272	13483		
P18	Virus Dex+	225	26	p<0.0001	28.04
	Unknown	4186	13563		
P19	Virus Dex+	200	25	p<0.0001	30.08
	Unknown	3734	14041		
P23	Virus Dex+	211	19	p<0.0001	41.18
	Unknown	3774	13996		
P24	Virus Dex+	194	25	p<0.0001	30.14
	Unknown	3641	14140		
P25	Virus Dex+	264	11	p<0.0001	99.59
	Unknown	3442	14283		
P26	Virus Dex+	216	18	p<0.0001	43.66
	Unknown	3830	13936		
P27	Virus Dex+	220	20	p<0.0001	31.53
	Unknown	4593	13167		
P29	Virus Dex+	143	22	p<0.0001	34.68
	Unknown	2815	15020		
P31	Virus Dex+	200	23	p<0.0001	30.86
	Unknown	3908	13869		
P32	Virus Dex+	102	20	p<0.0001	35.21
	Unknown	2262	15616		
P36	Virus Dex+	187	25	p<0.0001	28.98
	Unknown	3649	14139		

Supplementary table 3. Enrichment of virus-specific βCDR3s from the single-cell sequencing dataset[16] in clustered CD8⁺ thymocytes. Contingency table for the Chi-square analysis performed with the Yates correction to test the null hypothesis of independence between the specificity of βCDR3s (Virus Dex⁺ and Unknown specificity) vs the connection of βCDR3s ("clustered" and "dispersed") in all CD8⁺ thymocytes from 12 donors. The results (p-value < 0.0001) rejected the null hypothesis, thereby indicating the interdependency of the two variables.

649

Peptide	Virus
FLRGRAYGL	EBV
FLYALALLL	EBV
GLCTLVAML	EBV
LLDFVRFMGV	EBV
RAKFKQLL	EBV
RTLNAWVKV	HIV
SLFNTVATL	HIV
SLFNTVATLY	HIV
MLDLQPETT	HPV
LLFGYPVYV	HTLV
GILGFVFTL	Influenza
ELRRKMMYM	CMV
VTEHDTLLY	CMV
SLYNTVATLY	HIV

650 **Supplementary table 4. List of peptides represented in the chord plot of Fig 3c.** The table is
651 organized according to the clockwise order of the chord plot segments.
652

Peptide	Virus
KLGGALQAK	CMV
QYDPVAALF	CMV
RPHERNGFTVL	CMV
TPRVTGGGAM	CMV
AVFDRKSDAK	EBV
IVTDFSVIK	EBV
QPRAPIRPI	EBV
RLRAEAQVK	EBV
RPPIFIRRL	EBV
AYAQKIFKI	CMV
IPSINVHHY	CMV

Supplementary table 5. List of peptides represented in the chord plot of Fig 3g. The table is organized according to the clockwise order of the chord plot segments.

Acknowledgements

The authors would like to express their gratitude to the donors and their families who allowed the collection of samples for research under sad circumstances. The authors would also like to thank Prof. Pascal Leprince, Dr. Guillaume Lebreton, and Dr. Marina Rigolet of the cardiac surgery team, and Prof. Bruno Riou and the graft coordination team, of the Pitié-Salpêtrière hospital for their contribution to sample collection. The authors thank the UMR 8199 LIGAN-PM Genomics platform (Lille, France) for sequencing. We thank Thierry Mora and Aleksandra Walczak of the Ecole Normale Supérieure de Paris for helpful discussion.

Author contributions

VQ performed the experiments with assistance from PB, HV and EMF; VQ, PB, VM and HPP analysed data with contributions from all authors; VQ and DK wrote the manuscript; DK conceptualized and supervised the study.

Data availability statement

Datasets from VDJdb were downloaded from https://vdjdb.cdr3.net. Datasets from McPAS-TCR were downloaded from http://friedmanlab.weizmann.ac.il/McPAS-TCR/. We manually curated these datasets to be sure to use only βCDR3s from CD8 tetramer-specific cells. Single-cell datasets from 10X genomics were downloaded from https://support.10xgenomics.com/single-cell-vdj/datasets ('Application Note - A New Way of Exploring Immunity' section, datasets 'CD8+ T cells of Healthy Donor' 1–4, available under the Creative Commons Attribution license). Single-cell dataset of COVID-19 patient were downloaded from https://www-ncbi-nlm-nih-gov.proxy.insermbiblio.inist.fr/geo/query/acc.cgi?acc=GSE145926.

Dataset repertoires of immunisation with live yellow fever vaccine are available in the NCBI Sequence Read Archive (accession no. PRJNA493983). Only P1 and S1 at day 15 post-vaccination are used and represented.

Data from the donors are available on request to the authors.

Competing interests

The authors declare no competing financial interests.

Funding

690 This work was primarily funded by the TRiPoD ERC-Advanced EU (322856) grant to DK, and by
691 the LabEx Transimmunom (ANR-11-IDEX-0004-02) and RHU iMAP (ANR-16-RHUS-0001) grants.
692 Corresponding author
693 Correspondence to Prof. David Klatzmann
694 Immunology-Immunopathology-Immunotherapy Laboratory (i3) and Clinical Investigation
695 Center for Biotherapies (CIC-BTi)
696 Pitié-Salpêtrière Hospital, 83 boulevard de l'Hôpital, F-75013, Paris, France.
697 E-mail: david.klatzmann@sorbonne-universite.fr