On the selection of unconventional CD8+ T-cell during human thymopoiesis

Valentin Quiniou

Sorbonne Université

Ecole doctorale ED394 : Physiologie, Physiopathologie et Thérapeutique.

I3 Laboratory, Immunology, Immunopathology, Immunotherapy

On the selection of unconventional CD8$^+$ T-cell during human thymopoiesis

by Valentin Quiniou

Thesis of Immunology

Directed by Professor David Klatzmann

Presented and publicly defended on the

Evaluated by a jury composed of:

Pr Jérémie Sellam, MD, PhD, thesis examiner, president

Dr Victor Greiff, PhD, thesis referee.

Pr Simon Fillatreau, MD, PhD, thesis referee

Pr Philippe Kourilsky, PhD, thesis examiner

Pr Alain Fischer, MD, PhD, thesis examiner

Pr David Klatzmann, thesis director

This work is dedicated to my beloved grandmother,
Marie-Thérèse Routier.
1+1=2, …

Acknowledgements

First and foremost, I would like to thank Professor David Klatzmann. Thank you for welcoming me in your laboratory. When I finished my PharmD at the university, I thought I had learned enough to be able to work as a scientist. In reality, it was in your laboratory that, while, working I learned and understood what Science is. Science is not just about findings, results or papers. Science is resiliency. We rule out the wrong tracks and try, modestly, to give the right direction to future generations. Thank you for the TriPod project and for your confidence. Thanks to this one, I think am one of the only PhD students who has made his professional trips by private jet! It was a wonderful scientific and human adventure. Through this project, you have displayed all of the qualities that make up great scientist. I am still impressed how you have the energy to always see the positivity in our results. Your energy is the bone marrow of the lab and I will draw a great deal of inspiration from it for all my future projects. The way you manage your student, the famous "fuzzy management", could appear unconventional for classical scientists, but this leads us to push out our comfort zone, to have a translational approach of scientific questions and results. In one word, it makes us becoming pleioscientists.

Professor Adrien Six. I am always impressed by your encyclopaedic knowledge in Immunology and your relevant remarks for statistical analysis. Your rigor and your scientific objectivity do not have equal.

Encarnita Mariotti-Ferandiz. For what I remember, we arrived nearly at the same time at the laboratory. We went through this project together and I thank you for all the involvement you have shown. The diversity of the TCR repertoire is matched only by the number of things that you have on your schedule!

I would like to thank the members of the jury to have evaluated my work.

To my office-mate, bar-mate and friends: Federica, Vanessa and Pierre, from the HIF group:
Ladies first (I know you are sensible to gallantry…). Federica, you arrived by surprise in our office and it was a great one! I am pretty sure that your job description mentioned

the necessity (i) to be good in math, (ii) to have tongue-in-cheek humour and (iii) to like drinking beers. Always remember Federica: “Tutto va bene quando facciamo l’amore!”
Vanessa. Respect! You are one of the few persons that are able to speak with PHP! I am pretty sure that your paper will be a recognize as a worthful work in the field. Of course, as soon as it will be published ☺.
Pierre Barrenes, a.k.a. Pierre Bernard. What a great meet! I am really proud of you for your multicentric method comparison. Just a few are able to do a “Nature biotech” as first paper. You are a great engineer. You will become a great scientist. I am always available for other collaborations and of course for our “multicentric restaurants and wines comparison”. I hope both will continue for a long time.
To those who are not aware about it, Pierre Barennes is also a great actor. You can judge by yourself: https://www.youtube.com/watch?v=LtmAAGdiiHk.

Many people from the lab have helped me throughout the course of my PhD for which I am truly thankful. A very special thanks for the TriPod team of course. You all have been a source of emulation. Gwladys, you have conducted our “Dear mouse mapping experiments” with the rigor and discipline that characterize you.

To the Husson Mourrier team. Dr Claude Bernard, PHP Phuong Hoi!, Djamel Nehar, Nicolas Derian, Wahiba Chaara. When do we go on vacation together to see the prank monkeys?

To my grandparents, my brothers and sister of course, always here for me. A very special thanks to my mother and father. I could not have done it without your genetical materials!

Some words of love for my children: César and Margaux. Your nocturnal awakes and others classical baby’s illness did not help me very much along this thesis. However, even if I love my job, I prefer to be with you a thousand times rather than in the lab. Please, don’t tell it to David…
Some others words for you, which are warm and cosy in your mother’s womb and do not have an official name at the time of writing. Welcome to earth little baby. Your sister

and your brother are impatient to play with you. I look forward to show you all the extraordinary things you can see on earth when you are curious.

Finally, I must thank my dear and lovely wife, Marie, for **SUPPORTING** me throughout this adventure. I never could have done it without you. Maybe you should appear on the paper's authors! You will be very pleased to know that it is finally ended now.
Well, almost… Just after the next projects!

Synopsis

Figures

Abbreviations

AA: amino acid
APC: antigen presenting cell
APECED: autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy
AIRE: autoimmune regulator
APS-1: autoimmune polyendocrine syndrome type 1
amTreg: activated memory T regulatory cell
C. jejuni: Campylobacter jejuni
cTEC: cortical thymic epithelial cell
CTL: Cytotoxic T-cell
cmCD8: central memory CD8 T-cell
emCD8: effector memory CD8 T-cell
emraCD8: effector memory CD8 T-cell
ERV: endogenous retrovirus
HLA: human leukocyte antigen
IFNγ: interferon gamma
IL= interleukin
KO: knock out
LCMV: lymphocytic choriomeningitis virus
LTR: long terminal repeat
mAb: monoclonal antibodies
MBP: myelin basic protein
MCMV: murine cytomegalovirus
MHC: major histocompatibility complex
MOG: myelin oligodendrocyte glycoprotein
mTEC : medullary thymic epithelial cell
nCD8: naive CD8 T-cell
nTconv: naïve T conventional cell
nTreg: naïve T regulatory cell
pMHC: peptide-MHC
PV: Pichnide virus
RSSs: Recombination signal sequences
Tdt: Terminal deoxynucleotidyl transferase
Teff: effector T lymphocyte
TL: T lymphocyte
TRAC: TCR α chain
TRBC: TCR β chain
Treg: regulatory T lymphocyte
Tssp: thymus-specific serine protease
UMI: unique molecular identifier
VV: Vaccina virus

Introduction

Human possess a remarkably adaptive immune system that can recognize in a very specific manner pathogens like bacteria, fungi, parasites and viruses. It is also able to remember pathogens that have been previously encountered and quickly eliminates the recurrent pathogens by mobilizing a more efficient secondary immune response. In this way, the composition of the adaptive immune system is a snapshot of a subject's state, but also a representation of his immune history [1].

One of the cell lineages that composed the adaptive immune system is T lymphocytes (TL). The ontogenesis of TLs, also known as thymopoiesis, begins in the bone marrow and ends at the exit of the thymus. In the thymus, TLs acquire at their surface a receptor that recognizes antigens: the T-Cell Receptor (TCR). The TCR defines the specificity of these cells as it allows TLs to recognize antigens presented by antigen presenting cells (APC) in the form of a peptide bound to the major histocompatibility complex (MHC). Each TL expresses a single TCR and the set of all TCRs is known as the TCR repertoire. During thymopoiesis, the haematopoietic precursors undergo several steps, like random genomic recombination or thymic selection to generate a pool of functional lymphocytes that are exported in the periphery. A unique TCR equip each of these lymphocytes, as a weapon for a soldier, through its lifespan. To constitute an army that will be able to fight the different pathogens that will be encountered during life, the admitted dogma based on the clonal selection model [2] claimed that the different steps of production and selection eliminated lymphocytes bearing TCR specific for self-antigens, allowing the selection of a TCR repertoire specific for the non-self, composed of clones with unique specificity. So that, the TCR repertoire is able to recognize a multitude of antigens and the activation of lymphocytes through the recognition of their cognate antigens results in the constitution of a pool of activated memory T lymphocytes protecting our self. One of the most fundamental questions in adaptive immunity is how T-cells discriminate antigens? How do they recognize an autologous from an allogeneic one? A dangerous from a harmless? A malign from a foetal? In a Manichean model, if we consider that T effector cells (Teff) are inflammatory and the T regulatory cells (Treg) control this inflammatory reaction: do they recognize the same antigens? Deciphering the specificity of TCR repertoire for antigens is one of the major problems of modern immunology [3]. Only the number of TCR matches the number of questions arising from the TCR repertoire and their implications in human biology and medicine.

The major aim of this work is to investigate the TCR repertoire generation. In this manuscript, the current knowledge about TCR repertoire generation is summarized. This knowledge presents different “grey zone” that found explanation in overlooked published data concerning the heterologous immunity. Together, the data concerning heterologous immunity and this work converge to the conclusion of a highly specific immune response based on T-cell receptor not so specific.

1. T-cell receptor: stochastic to specific.

Discovered in the early 1980s, a few years after the discovery of MHC to which it is closely related, the TCR was first detected thanks to the development of monoclonal antibodies. They allowed the recognition of the TCR and their specific binding on this CD3-associated surface heterodimer, resulting in activation of the T lymphocytes [4,5]. In parallel of the detection of this receptor, researchers focused on the transcripts and the genes encoding the TCR. The conserved structure of these proteins with some variant sequences, like immunoglobulins, oriented research teams to genes having variable and constant regions, with a recombination mechanism. Sequences analysis of genes encoding the TCR confirmed these hypotheses, both in mice and in humans [6–8]. As for immunoglobulins, the recombination mechanisms allow the generation of unique TCR hypothetically able to recognize all the antigens. This hypothesis is the one admitted and tough. It is the one that I have learnt. However, considering previously published papers, this assumption should have to be moderated.

1.1. How TCR can be so diverse?

TCR is a transmembrane heterodimer composed of two glycoprotein chains. There are four types of chains: nearly 95% of lymphocytes carry the α and β chains dimer and 5% carry on the γ and δ chains [9]. Each chain is divided in four segments (Fig. 1a):

-Cytoplasmic C-terminal region: Its length is about 5 to 12 amino acid that is too short to transfer a cytoplasmic activation signal.

-Transmembrane constant region: this predominantly hydrophobic sequence has a lysine residue in its centre. This positively charged and highly conserved amino acid allows interaction with the CD3 coreceptor chains.

-Extracellular constant region (TRC): it is specific for α and β. For the α chain, there is only one TRC and two for the β chain, discriminated by 6 aa. Each chain contains a cysteine proximal to the transmembrane region that is involved in the formation of inter-chain disulphide bond.

-N terminal extracellular variable region: length of about 100aa, with an intra-disulphide loop.

The N-terminal domain contains the hypervariable or "complementary determining region" (CDR) involved in contact with the peptide complexed with the molecular histocompatibility complex (MHC) and therefore in the antigen recognition. There are 3

CDRs by chain. The CDR1 and CDR2 sequences are germline encoded and define each V gene, while the CDR3 is the product of a somatic recombination, forming the whole V domain (Fig. 1b):

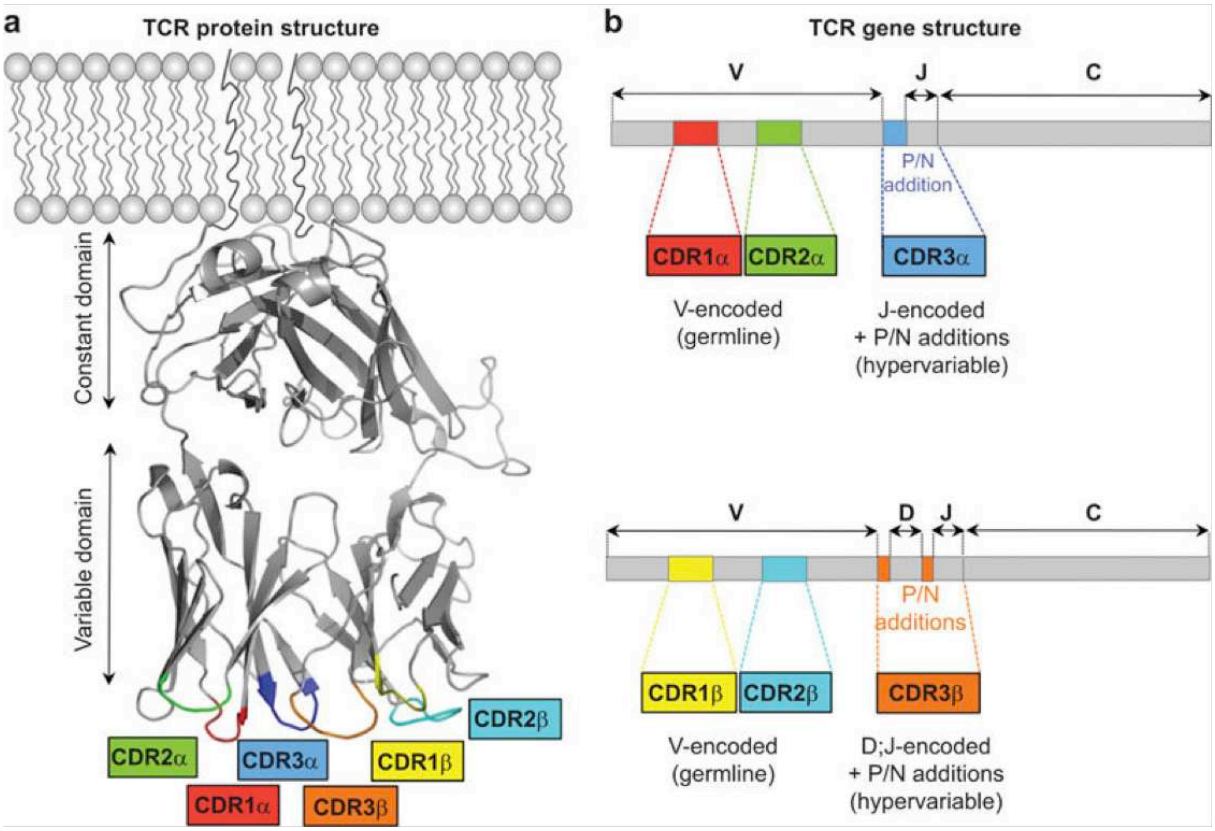


Figure 1. Protein and gene structure of the TCR.

A. TCR protein. CDRs from α and β TCR consist of hair pin loops linking β-strands. B. CDR1 and CDR2 are situated in the V region, whereas CDR3 lies at the junction between the rearranged V and J segments (αTCR) and V, D and J segments (βTCR). CDR1 and CDR2 are germline encoded, whereas CDR3 is made by random addition and deletion of nucleotides (blue for αTCR and orange for βTCR)[10].

A highly diverse TCR repertoire is seen to be a fundamental and necessary property of an effective immune system to efficiently control pathogens infections [11]. Through the recombination process, junctional diversity and combinatory diversity, the TCR repertoire has the potential to be composed of more TCR than the number of cells that composed the human body [12]. In this part, we will describe evidences showing why the TCR repertoire is one of the most diverse and complicated biologic objects.

1.1.1. V(D)J recombination and genomic organisation

The diversity of the TCR is in part due to the organisation of the genes encoding its primary protein entity. Dr Marie Paul Lefranc and her team have done the major work for the description of the TCR loci and the genes that encode the TCR [13]. Notably, IMGT implemented an international nomenclature, recognized by WHO and adopted by NCBI, the major gene bank database. The aim of IMGT is to provide standardized reference of TCR annotation and, more generally, for all immunogenic object. TCR loci are divided into different families of genes: V, D and J. Unlike many other genes, TCRs are not functionally encoded in the genome. Genes encoding the TCR consist of different non-contiguous loci, which by recombination will form a complete TCR sequence. This is the first mechanism underlying the diversity of the TCR repertoire.

In human, genes encoding the α chain (TRA) are located on the chromosome 14, loci 14q11-12 and lay on nearly 1000kb. There are 49 V genes (TRAV) over 41 groups [14], 61 J genes (TRAJ) [15] et 1 C genes (TRAC). TRAV 5' genes are close to the centromere and TRAC 3' are the most telomeric ones. Within these sequences, only 43 are functional for TRAV, 58 for TRAJ and only one for TRAC. TRAVJ recombination can rearrange these 101 genes into 2494 unique germline encoded TRAVJ gene combinations.
Similarly, genes encoding the β chain (TRB) are located on the chromosome 7, loci 7q35 and lay on nearly 620kb. There are 68 V genes (TRBV), 2 D genes (TRBD), 14 J genes (TRBJ) and 2 C genes (TRBC). All TRBV loci, except TRBV30, precede a duplicated cluster D-J-C. The first cluster is made of 1 TRBD, 6 TRBJ, and TRBC1 gene. Second one is made of 1 TRBD, 7 TRBJ and TRBC2 gene. TRBV1 gene is the most centromeric in 5' whereas TRBV30 in 3' is the most telomeric. The total number of functional genes comprises up to 54 TRBV, 2 TRBD, up to 13 TRBJ and 2 TRBC. V, D and J recombination can rearrange these 71 genes into 2808 unique germline encoded gene combinations.
Even if the germline possibilities represent more than 7.10^6 gene combinations, it is only 5% of the possible diversity of the $\alpha\beta$TCR[16].

All these genes are the roots of the TCR repertoire germinal diversity. During thymopoiesis, these genes are rearranged to produce a functional protein. As for immunoglobulin genes, each of the V, D and J genes are flanked by recombination signal sequences (RSSs). They are composed of a heptamer and a nonamer, of conserved

structure, separated by a random sequence of 12 or 23 nucleotides. The rearrangement mechanism operates according to the 12/23 rule. The V genes have at their 3 'end an RSS, the J segments at their 5' end and the D segments at their 5 'and 3' end.

For the beta chain, two successive recombinations are observed. First, a D gene assembles with a J gene with a deletion of the intermediate portion to create a preliminary DJ gene. Second, one of the V genes is joined to the rearranged DJ segment with deletion of the intermediate fraction to create a rearranged V(D)J gene (Figure 2).

The somatic rearrangement of the alpha chains is different due to the absence of D genes. The synthesis of the alpha chain begins with the assembly of a V gene with a J gene and a deletion of the DNA segments between these two loci. This mechanism is called V-J rearrangement and is performed thanks to the enzymatic complex RAG (recombination activating genes). This complex is composed of 2 proteins RAG1 [17] and RAG2 [18] identified in the early 90's. They are expressed in T lymphocytes at the early stage of their development, although few studies indicated peripheral (re)-expression[19]. The presence of this complex is crucial for the adaptive immune system because it is also this system that allows the synthesis of BCR. In humans, RAG gene mutations cause severe combined immunodeficiencies[20].

Selection of the different V, D and J genes for recombination has been for a while supposed to be stochastic. However, recent results have shown that these genes are targeted for recombination based on different factors[21]. Indeed, there are wide variation in the heptamer, nonamer and spacer quality and these variations in RSS could contribute significantly to non-stochastic recombination frequency[22]. Moreover, the accessibility of the chromatin and the epigenetic states seems also to influence the recombination process of the TCR[23,24].

1.1.2. Junctional diversity

The second mechanism involved in repertoire diversity is the junctional diversity. During the rearrangement of the different gene segments V, D and J of the TCR, the double-strand breaks are formed in hairpin loops and must joined together to form a single DNA molecule. To do so, deletion and addition of nucleotides at the junctional regions of V, D and J genes are performed under the action of different enzymes. This expands the diversity by junctional diversification arising from the addition of

nucleotides by Artemis and the Terminal deoxynucleotidyl Transferase (TdT) and from the exonuclease trimming of the recombining gene ends.

Just after the action of the RAG proteins, DNA repair protein Artemis are responsible for single-stranded cleavage of the hairpin loops and addition of a series of palindromic nucleotides P that are inserted as a result of imprecise joining during V(D)J recombination[25]. The action of Artemis is a crucial step of V(D)J recombination and mutations in *Artemis* gene also cause severe combined immunodeficiencies[26]. Once Artemis has added nucleotides, the addition of random nucleotides by a specific enzyme: the terminal deoxynucleotidyl transferase (TdT), is performed. TdT is produced during the maturation of lymphocytes in the thymus. The gene encoding TdT is expressed during the early stages of thymic development and is down-regulated with the maturation of cellular clones [27]. The TdT is a template-independent DNA polymerase that adds N nucleotides to the coding end junctional site. The N nucleotides are string of random nucleotides inserted at each V(D)J joining site and are responsible for the appearance in TCR chains of amino acids that are not germline-encoded. The addition is mostly random, but TdT seems to exhibit a preference for deoxyguanosine and deoxycytosine nucleotides[28]. Could this phenomenon imply bias in TCR repertoire generation?

Finally, exonucleases remove nucleotides from the coding ends including any P or N nucleotides that may have formed. If it is necessary, DNA polymerases can make the two-end compatible for joining with the insertion of additional nucleotides. The processed coding ends are then ligated together by DNA ligase IV[29]. All of these processes are represented in Figure 2:

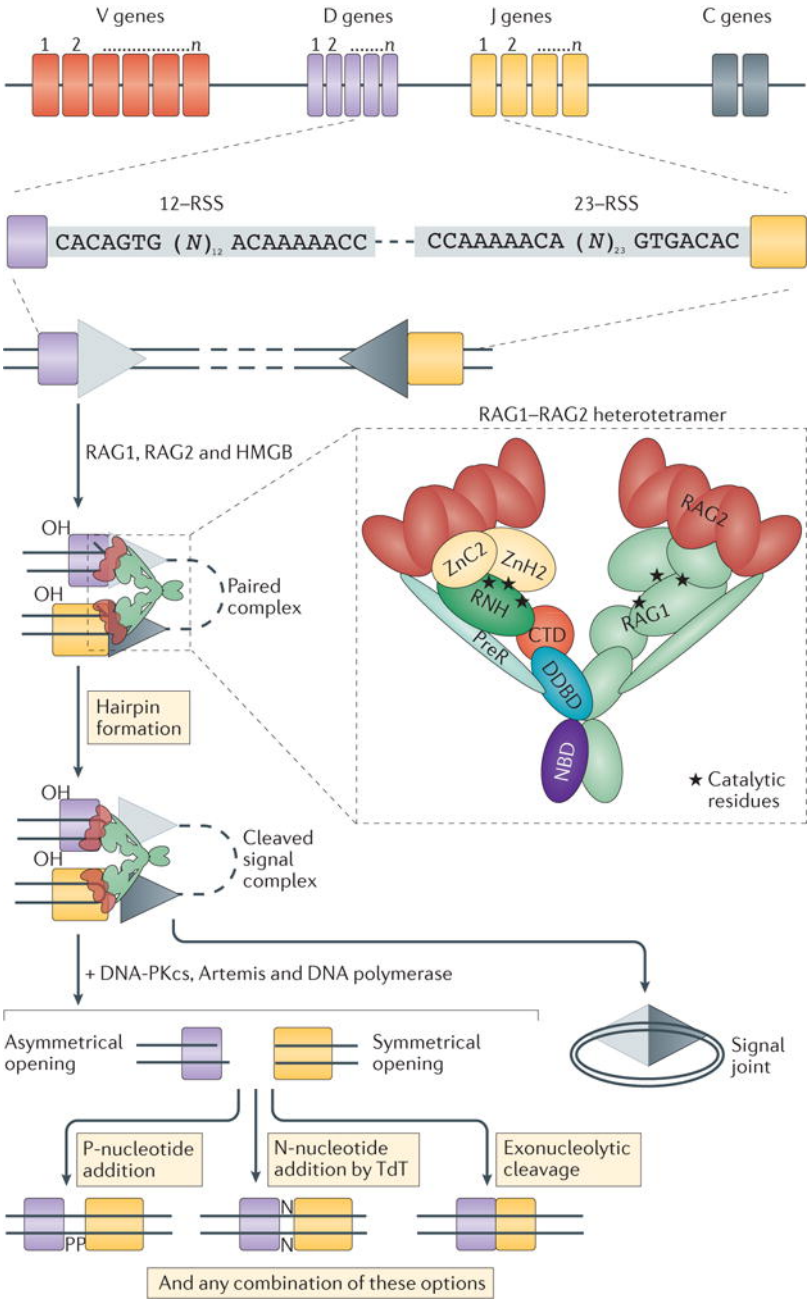


Figure 2. Mechanism of V(D)J recombination.

The recombination signal sequence (RSS) frame each antigen receptor gene segment and contains two conserved DNA sequences: the heptamer and the nonamer, separated by a spacer region of 12 or 23 nucleotides. The RAG complex binds to the RSS to cleave the DNA by forming a synapsis of one 12-RSS and one 23-RSS (the 12-23 rule). Subsequently to the hairpin formation, Artemis opens it. The asymmetrical opening allows the incorporation of the P nucleotides. Furthermore, the TdT introduces the N nucleotides. Finally, exonucleases trim nucleotides between the two coding ends [20].

These mechanisms allow the formation of two chain α and β that are paired at the cell surface to form a complete TCR.

1.1.3. Chain pairing diversity

The third mechanism involved in repertoire diversity is the association between the two chains. Indeed, once the chains are produce based on the previous mechanisms, the alpha and beta chain are combined and joined with bound to form an entire TCR molecule. The combination of alpha and beta chain forms the pairing diversity of the TCR. Assuming that all these process are stochastic and independent between the two chains, the overall diversity of the pool from which TCRs are generated is theoretically about 10^{61} possible TCR[30] and was admitted to be the source of the specificity against all pathogens that could be encountered during life. But most of the studies admit that the diversity of the TCR is approximately 10^{19}[12]. Moreover, the number of TCR in an individual has been estimated to 10^{8}[31,32]. Indeed, there are not enough T-cells in the periphery to bear all the theoretical diversity of the repertoire and many cells have the same TCR, due to clonal expansion following the recognition of the antigen. These different mechanisms are represented in Figure 3.

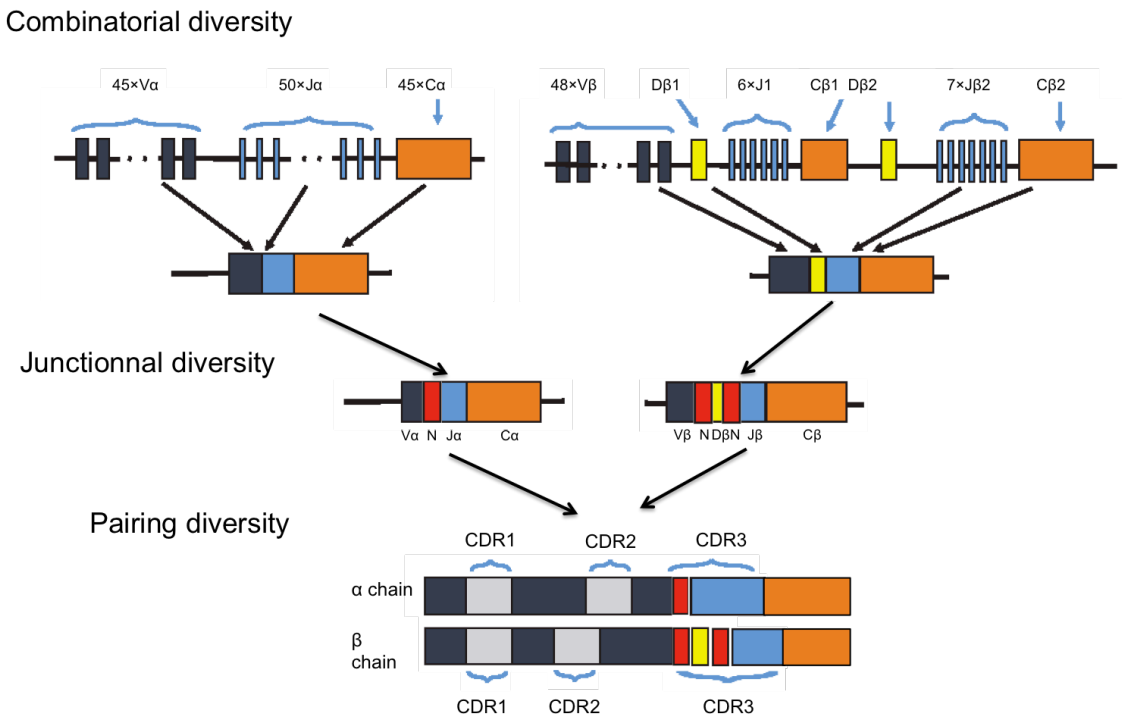


Figure 3. Generation of TCR diversity.

The TCR α and β chain genes are composed of discrete segments that are joined by somatic recombination during thymopoiesis. This step generates a diversity of 6.10^{6}. The junctional diversity generated by the TdT and others enzyme generated a potential of diversity of 2.10^{11}. Finally, the estimated biological diversity after pairing is about 10^{19} [12].

The tremendous set of TCR shaped equips the TLs that are present in the periphery and act in all the immune response. TLs are able to recognize endo- and exogenous antigen and this recognition results in the activation of the cell and expression of different cytokines, receptors and others active and informative molecules. This recognition is done by the TCR.

1.2. Perception of the antigens

Every day, at every time, the immune system permanently deals with the recognition of antigens with the TCR. T-cells selected in the thymus, under mechanisms that we will describe after, migrate to the periphery to act as critical component of the adaptive immunity system. They circulate in the vascular system to one of the various peripheral secondary lymphoid organs. These organs are either well-defined structures such as the spleen and lymph nodes, or more diffuse structures such as lymphoid tissue associated with intestinal mucosa. In this structure append the presentation of antigen by APC to T-cell.

1.2.1. Antigen recognition by TCR

The evidence of the recognition capacity of the TCRs began with the discovery of the MHC restriction. Indeed, in 1974, Peter Doherty and Rolf Zinkernagel discovered that an antigen could be recognized by a TCR only if it is complexed to a MHC, also known as human leukocyte antigens (HLA) in humans[33,34]. This discovery earned them the Nobel Prize of medicine in 1996. This condition of presentation is known as the MHC-restriction. It means that TCR can only recognize peptide if they are first processed and presented by APC as peptide-MHC (pMHC) complex. APC respectively present to CD8 T lymphocytes a peptide fixed with MHC class I (MHC-I), and to CD4 T lymphocytes, peptides fixed with MHC class II (MHC-II).

The HLA is one of the more diverse biological systems. HLA molecules are divided in two classes: HLA-I and HLA-II. HLA-I comprises 3 major substitutes: HLA-A, HLA-B and HLA-C. HLA-II comprises 3 major substitutes: HLA-DR, HLA-DQ and HLA-DP. In humans, more than 18,000 different alleles composing the HLA-A, HLA-B, and HLA-C alleles, have been identified. They can differ by just 1 up to more than 30 amino acids, which can profoundly affect the repertoire of bound peptides, the pMHC structure, and, of course, TCR interaction and recognition of the antigen[35,36].

HLA-I is present on the surface of all nucleated cells of the body, in varying concentrations. It is composed of two chains: the heavy chain and the β2-microglobulin chain. The heavy chain is composed of two alpha helices that surround a beta sheet. This is the place where the peptide is complexed, usually a length of 8-10 AA. Proteins synthesized in the cell, like endogenous or viral proteins if the cell is infected, are degraded by a proteasome into peptides which are then supported and transported in the endoplasmic reticulum by the TAP complex (transporter associated with antigen presentation). HLA-II is mostly present on APC and is composed of a α and β chains that forms a pocket in which the peptide is fixed. Peptides that are fixed on MHC-II have longer than in MHC-I, of the order of 15-20 AA[37].

The first TCR-pMHC crystal structures provided information on the structural basis of T-cell recognition of the complexed antigen[38]. The TCR is diagonally oriented above the long axis of the pMHC binding cleft with a range from 35°[39] to 67°[40], with the CDR3 region sitting above the peptide, which exhibit the greatest degree of genetic variability. The CDR1 and CDR2 loops mediate MHCα contacts[41]. The Vα domain is above the N-terminal portion of the antigenic peptide, whereas the Vβ domain is over the C-terminal portion of it. As soon as the TCR is in contact with is cognate antigen complexed with the correct MHC, the activation signal is trigger via the immune synapse.

1.2.2. The immune synapse and the coreceptors

Considering the short size of the cytoplasmic domain of the TCR, the transduction of the signal resulting from the recognition of the antigen is carried out thanks to a multi-molecular complex: the CD3[42]. This structure is present on all TLs. It plays both a role in signal transduction and in the formation of the TCR as it allows its migration to membrane surface of the cell. The CD3 complex is composed of different chains whose transmembrane parts are negatively charged, which allows them to associate with transmembrane domain of the positively charged TCR.

In addition to the CD3 complex, two other co-receptors were identified to interact for the formation of the immune synapse: CD4 and CD8. These molecules interact with the MHC molecule that carries the antigen to be recognized. The CD4 is a monomer present on the surface of Teff and Treg. The extracellular part is made of 4 immunoglobulins domains. The CD8 molecule is a heterodimer consisting of two alpha and beta homologous chains linked by disulphide bridges and of a single immunoglobulin

domain. The CD4 and CD8 molecules will respectively bind the HLA-II and HLA-I with low affinity and independently. They are associated to the TCR at the time of antigen recognition[43].

Different models have been proposed to explain how TCR recognise peptides. Engagement of the TCR by pMHC initiates TCR signalling and results in the formation of an immune synapse. It is a nano scale region of close contact between TL and APC. The immune synapse is centred on the TCR, which is surrounded by several coreceptors that contribute to TCR signal transduction[44].

First of all, TCR is mainly accepted to works as a mechanosensor, based on the idea that ligand binding results in conformational changes in the TCR complex, which is responsible of signal transduction. Different teams have demonstrated that the TCR engagement leads to a conformational change that make the CD3 chains more prone to phosphorylation[45,46].

The kinetic proofreading is the second mechanism proposed for activation via the TCR. It suggests that the interaction of TL with TCR and APC with pMHC leads to the recruitment of molecules on the cell membrane that result in the initiation of TCR signalling. It supposed that thymocytes detect ligand affinity by measuring how long pMHC complexes remain bound to a TCR. This model is supported by the observation that high-affinity TCR–pMHC interactions are characterized by slow dissociation rates and long half-lives[47,48]. This dwell time provide a window of opportunity for the CD3, the TCR and others co-receptors to be fully assembled[49,50]. In contrast, TCR–pMHC complex with low affinity interactions have fast dissociation kinetics, resulting in incompletely associated CD3, TCR and others co-receptors and a lack in the transmission of the activation signal[51].

Even if the initial mechanism of activation is not fully understood, the transduction of the signal is clearly much more identified. Recognition of the antigen by TL first results in the engagement of co-receptors (such as CTLA-4 or CD28) and leads to the recruitment of the CD4 or CD8, which bind to conserved regions on the MHC molecules. The cytosolic parts of CD4 and CD8 bind the LCK and promotes the phosphorylation of ITAM on CD3[52]. These phosphorylations lead to the recruitment of the tyrosine kinase

Zap-70. The Zap-70 kinase propagates, thanks to its enzymatic activity, the signals induced after recognition of its specific ligand by the TCR.

The outcome of these signalling events is the activation of multiple transcription factors, which coordinate multiple T-cell responses, including proliferation, migration, and cytokines production. Immune cells carrying an antigen-specific TCR will be activated and, depending on their type, will have regulatory, helper or cytotoxic action. Some of them will even become memory TLs, characteristics of the adaptive immune response.

1.2.3. From naive to memory T-cell

The pool of naive TLs is diverse and contains cells bearing TCR that differ in their affinity for the same antigen due to the previous mechanism of generation. Once the presented antigen is recognized, naïve T-cell undergo clonal expansion and generate activated T-cell. This step is known as specificity signal of activation. Indeed, lymphocyte activation occurs if, and only if, the TCR recognizes the pMHC complex presented by APC. The greater the binding affinity between the TCR CDR3 and the immuno-dominant peptide epitope present in the MHC groove, the longer the binding between the two cells will be. This prolonged high affinity signal allows activation of the lymphocytes. In the context of viral infections, the CD8 cells expand more than CD4 with a ratio of 3:1[53]. These activated cells down regulate the CD45RA and expressed the CD45RO. Following the pathogen clearance, these activated cells undergo an apoptosis contraction step characterized by the death of nearly 90% of previously activated cells. Activated T-cells are much more susceptible to apoptosis occurring as a consequence of growth factor deprivation or triggering through the TCR in a process known as activation-induced cell death[54].

Thereafter, a long-term persistence T-cells called pathogen-specific memories are maintain by low level homeostatic proliferation [55]. As the pool of memory T-cell is not extensible, IFN type 1 cells that are generated during infection cause attrition of previously existing memory T-cell[56].

The clonal selection and expansion of TL after antigen recognition leads to a discussion of the impact of TCR affinity as a determinant of TL response intensity. Analyses of both Th and CTL responses have indicated that clonal prevalence within epitope-specific CTL responses may be reflective of TCR affinity for pMHC [57,58]. The repertoire modification during antigen recognition seems to be dependant of the affinity of the TCR [59].

In addition to determining proliferation, the way that TCR binds to pMHC complexes seems to regulate the clonal apoptosis during the contraction phase. Indeed, TL with TCR with a higher affinity for pMHC-I are the first to go in apoptosis during the immune response [60]. Treg cells may control clonal expansion by preferentially inhibiting the low-affinity CTL[61].

There are numerous other ways to study the evolution of the phenotype from naive to activated or memory. What we show across these different examples is that the TCR is directly implicated in the evolution of the TL from the naïve phenotype to the subsequent one. How this implication works remain an open question of immunology.

1.3. TCR is the key tool of the adaptive immune system

TLs are fundamental in the orchestration of the immune adaptive response. Upon encounter with their cognate antigen, naive TLs become activated, proliferate and are induced to differentiate into specific memory T-cell subsets. This differentiation depends on numerous factors including the type of APC, cytokines and the microenvironment. Both CD4 and CD8 T-cells activation requires three sequential signals: TCR triggering, co-stimulator signal (example CD28-B7) and IL-2 mediated signalling. The expression of IL-2 and its receptors is induced upon TCR signalling[62]. Also, the affinity of the TCR for its cognate antigen and the concentration of the antigen have emerged as a key factor in determining TL cell fate.

1.3.1. Helper T-cells

The variety of pathogens present in our environment has pressured the immune system evolution toward different mechanisms for tailored protection. Helper T-cells (Th), characterized by the expression of the CD4 co-receptor, their cytokine signatures, the expression of specific transcription factors and distinct migration patterns, orchestrate the immune response. Naïve Th become activated when they recognize, via the TCR, their cognate antigen bound on MHC-II. They divide rapidly and undergo clonal expansion, expanding up to 50000 times a week. This activation triggers them to release cytokines that regulates the activity of many types of cells[63].

They secrete cytokines that act as enhancer of the immune response. Depending on the cytokine released in the environment, notably by APCs, Th will differentiate into dedicated subsets, named Th1, Th2, Th9 and Th17, polarizing the immune response and triggering activation/inhibition of other immune cells. For example, Th1 produce IL-2,

IFN-γ, TNF-β and mediate the defence against intracellular bacteria and viruses. Th2 produce IL-5, IL-4 and IL-13 and are important in worm infections and are involved in allergic immunopathology(Cousins et al., 2002). For Th1 and Th2, it was observed that peptide affinity and dose played an important role in determining the differentiation of naive CD4 T-cells[64,65]. Th1 differentiation of naive TL required strong TCR signals, whereas Th2 cells differentiation exhibited weaker signals[66]. Furthermore, the antigen dose has an impact on the stimulation of two different pathways: ERK and GATA-3, involve in Th1 and Th2 differentiation respectively. TLs stimulated with low dose of antigen exhibit an up-regulation of GATA-3, whereas stimulation of TL with high dose of antigen induces an up-regulation of ERK, which reduces the ability of cells to respond to IL-2 and inhibits GATA-3 pathway[67].

1.3.2. T CD8 cytotoxic lymphocytes

Cytotoxic T lymphocytes (CTL), characterized by CD8 molecule on their membrane, are highly specific killer cells. The role of CTL is to monitor all the cells of the body and to destroy any of them that are considered to be a threat for the integrity of the host. CTL kill infected cells, preventing the production and dissemination of the intracellular pathogens. They are also providing a certain degree of protection against spontaneous malignant tumours, by their ability to detect quantitative and qualitative antigenic differences in malignant T-cells. In both cases, the protein repertoire presented by the target T-cells is altered. The TCR of CTL recognizes these antigens presented by class HLA-I molecule on the target T-cell surface[68].

The CD8 responses against viruses such as Influenza, CMV, EBV and HIV are critical for control of viral replication. Selective reduction of T-cell responses, as observed in immunosuppressed transplant recipients, is strongly associated with reactivation of latent viruses such as human herpes viruses[69,70]. However, it can be prevented or treated by adoptive transfer of virus-specific $CD8^+$ T-cells[71]. Indeed, CTLs kill infected cells by several pathways that involve direct T-cell–cell contacts between CTL and target T-cells. In one case, the Fas-ligand of the CTL binds the Fas-receptor of the target T-cells. This binding triggers apoptosis through the classical caspase cascade[72]. In the second mechanism, the CTL release highly cytotoxic proteins: perforin and granzymes, into the immune synapse of the target T-cells. These cytotoxic proteins are pre-synthesized and stored in lysosomes that are exocytosed following the TCR binding to the target T-cell[73].

Even the antiviral cytokines mediated mechanisms, involving INF-γ of TNF-α, works only as long as TCR stimulation continues.

After the immune response, most of the activated CTL undergo apoptosis in the contraction phase and are cleared by phagocytic cells once the infection is resolved. Subsets of activated CTL differentiate into memory cells that can more efficiently respond to the same antigen in a subsequent exposition.

1.3.3. For T regulatory cells

The discovery of Treg has been a breakthrough in immunology[74]. Indeed, Treg are essential players in the control of all immune responses including to self, tumours, infectious agents, grafts and inflammatory disorders[75]. Treg ensure the regulation of many different types of cells of the innate and adaptive immune response by bringing into play many different mechanisms that can inhibit the targeted population, as the specific recognition of antigen via the TCR [76]. As for the other TL, the development of Treg cells in the thymus is dependent on signalling via the TCR[77–79]. Moreover, the mechanism of action in the periphery is also TCR dependant[80]. This mechanism has been challenged in a KO mouse model, which allows the presence of Treg in the absence of TCR expression in the periphery[81]. Indeed, in this model, most of the Treg cells lacking the TCR, expressed normal amounts of Foxp3 as well as other Treg signature molecules such as CD25 or GITR. However, there was no observable up-regulation of CTLA-4, as it is induced by TCR signalling. The absence of TCR on Treg in periphery resulted in the loss of activated memory Treg (amTreg). These results fit with the necessity of a TCR signal for the differentiation as well as maintenance of activated Treg. Finally, mice that completely lack the amTreg developed severe autoimmunity. As IL-2 is required for Treg survival and activation, the measurement of IL-2 signalling was normal in TCR KO mice and the administration of IL-2 did not improve the autoimmune phenotype of these mice. Thus, TCR signalling is a necessity for the suppressive function of Treg.

T-cells are central drivers of the immune system both as effector and regulators and complete their functions thanks to their TCR. They are activated by signal transduction when their TCR, which recognizes in a very specific manner an epitope, carried by the MHC expressed at the membrane surface of an APC.

The TCR repertoire, resulting from the collection of unique TL clones is therefore dynamic and characterized by a high degree of plasticity. Indeed, during the immune response, because of the expansion, the contraction, the apoptosis and the memory pool formation, the TCR repertoire is constantly reshaped. Studying TCR repertoire can provide a better understanding on immune response mechanisms, and it is now seen as a mandatory approach in immunology. However, TCR repertoire analysis is far from being standardized, especially when considering human repertoire, since every individual has its own immune history and its own capacity to respond against antigens.

1.4. TCR repertoire specificity analysis

In the last decade, the sequencing technologies have made massive progresses with the next-generation sequencer. We are now able to obtain millions of raw sequences from the TCR repertoire with next-generation sequencing (NGS) but studies are still focusing on CDR3 numbers and clonality. Indeed, even with this huge amount of information, the complexity of the TCR repertoire is only glimpsed due to the paucity of powerful analytical methods[82].

1.4.1. Overview of TCR repertoire analytical tools

The first technique to investigate the TCR repertoire in human was cytometry. This analysis consists of co-staining of T-cells with a panel of TRBV and/or TRAV specific monoclonal antibodies (mAb). It provided some information on the variable region usage during the TCR repertoire modification related to treatment, aging or infection[83]. However, this method has low-resolution as it is only specific of the V region of the TCR. In addition, not all TRBV and almost none of TRAV gene families can be identified therefore providing a partial view of the TCR repertoire. Moreover, it does not provide any information about the CDR3 region, which recognize the antigens.

Then emerged the molecular-based approach of TCR repertoire analysis involving PCR amplification of DNA or RNA, also known as Immunoscope® [84,85] or Spectratyping[86]. Proposed simultaneously in France and in the US by two independent teams, this technic is based on the analysis of DNA fragment length analysis by electrophoresis, taking advantage of the variation in length of the CDR3 region. In naive repertoires, TLs are normally polyclonal and Immunoscope® analysis typically yields eight-peak following a Gaussian curve, each peak corresponding to a given CDR3-length. During immune response, this regular polyclonal display can be perturbed: one can see one or several

prominent peaks that correspond to the oligoclonal or clonal expansion of lymphocytes. This technique has been used in many different contexts such as lymphocytes selection[87], viral infections[88] or autoimmunity[89]. Since the Immunoscope, some alternative technologies such as single-strand conformation polymorphism [90], heteroduplex analysis [91] have been developed but the major breakthrough was the advent of the next generation sequencing. While the previous technics were time and cost extensive, and have allowed for the analysis of 10^2 to 10^3 sequences, the NGS is now able to generates billion of sequence and with an affordable cost. However, the methods used to generate the library and to sequence these libraries seems to have an impact on the results[92–94]. This point has to be investigated if we plan to use immune repertoire analysis in clinic.

Moreover, such data require precise identification of genes and sequences as well as mutations, and a standardized approach of nomenclature is necessary for TCR sequences identification[82]. Several robust softwares have been developed for processing of T-cell repertoire raw data. MiXCR is one of the most popular[95]. MiXCR provides an alignment of the V and J genes that only requires the order of 150 base pairs. It can also provide an assignment to specific D regions but this region is generally short making accurate assignment difficult, and in fact most studies of TCR repertoire simply include the D region within the CDR3 sequence. It also contains error-correction algorithms based on different statistical models, which allow removing the reverse transcriptase mismatch or errors in the early cycles of PCR. Another major advance in error correction has come from the incorporation of UMIs[96]. UMIs are short random oligonucleotide sequence incorporated before PCR amplification. After PCR, UMIs can be used to identify which sequences derived from the same single starting mRNA molecule, as they would all have been tagged with the same UMI and, thus, can be counted together. In conclusion, this new technology seems theoretically interesting and needs to be systematically evaluated.

Driven by NGS, the rapid increase in immune repertoire sequencing has generated a lot of data. From the earliest analysis, it was clear that antigen-specific repertoires were enriched for shared features, whether TRBV or CDR3 length. Attempts to statistically characterize epitope-specific repertoires for comparative purposes have taken many forms. First of all, some tools used in ecology have been used. The analogy between TCR repertoire and ecological is based on the measures of diversity that incorporate

information on both the number of different species and the number of individuals of each species. So some diversity measures in repertoire analysis were originally used to quantify the distribution of species in ecology [97,98]. The principal example for such diversity measures is the Shannon entropy, even if it first comes from the information theory, and Simpson diversity. These tools allow to performed some comparisons between repertoires quantitatively, with a goal of identifying features associated with immune response, even if these did not necessarily correlate with clear functional profiles[99].

For the analysis epitope-specific repertoire, NGS and more recently single-cell, paired with cell sorting of tetramer-specific T-cells have enabled the studies of molecular patterns of TCR sequence in epitope-specific responses[100]. By using all these new public data, different teams made effort to develop cluster based approaches and machine learning to predict the epitope specificity of TCRs[101,102].

Regardless of the analytical approach described here, it clearly emerged that within an antigen-specific repertoire, a significant portion of the responding receptors shared motifs. The shared repertoire can represent a majority of the response. This bias observed in different immune response was the subject of a standardisation to classify this bias.

1.4.2. Standardization of the TCR repertoire specificity:

NGS techniques have been employed to investigate T-cell repertoires across plenty of human diseases. The major goal of these investigations was often to find “the” specific clones of “the” disease. Indeed, several papers reported “skewing”, “immunodominance” or “restriction” in the TCR repertoire. Research of TCR specificity is one of the leitmotivs of immunologist.

From the first analysis by cytometry, epitope-specific population have been observed during immune response. These populations shared different characteristics across individuals as Vβ usage, CDR3 length or AA motifs. The presence of these public TCR repertoires, i.e. TCR repertoire that shared the same characteristics during the same antigen stimulation in two different individuals, was first surprising due to infinite diversity of the TCR repertoire theory. These observations imply either the presence of bias in the TCR repertoire generation, either the presence of some antigens involved in the increase of public TCR. In both there is a problem of interpretation as there is no

standard. To standardize these observations, a classification has been proposed based on the structure of the public TCR repertoire[103,104]:

-Type I bias: conservation of TRBV or TRAV gene usage, TRBJ or TRAJ gene usage is not conserved. There is no AA identity across the CDR3 region.

-Type 2 bias: conservation in TRBV or TRAV gene usage, TRBJ or TRAJ gene usage is not conserved. In the CDR3 region, AA residue motifs are present.

-Type 3 bias: Complete identity of TCRβ or TCRα. TCR presents identical AA structures encoded by either identical transcripts or redundant codons in the CDR3 region. These TCRs are classically referred to as 'public'.

-Type 4 bias: conservation of TRBV or TRAV gene usage and of TRBJ or TRAJ gene usage. TCR presents conserved motif across the CDR3 region but not strictly identic as it differs by only one or two AA within the CDR3 loop.

This classification is represented in Figure 4:

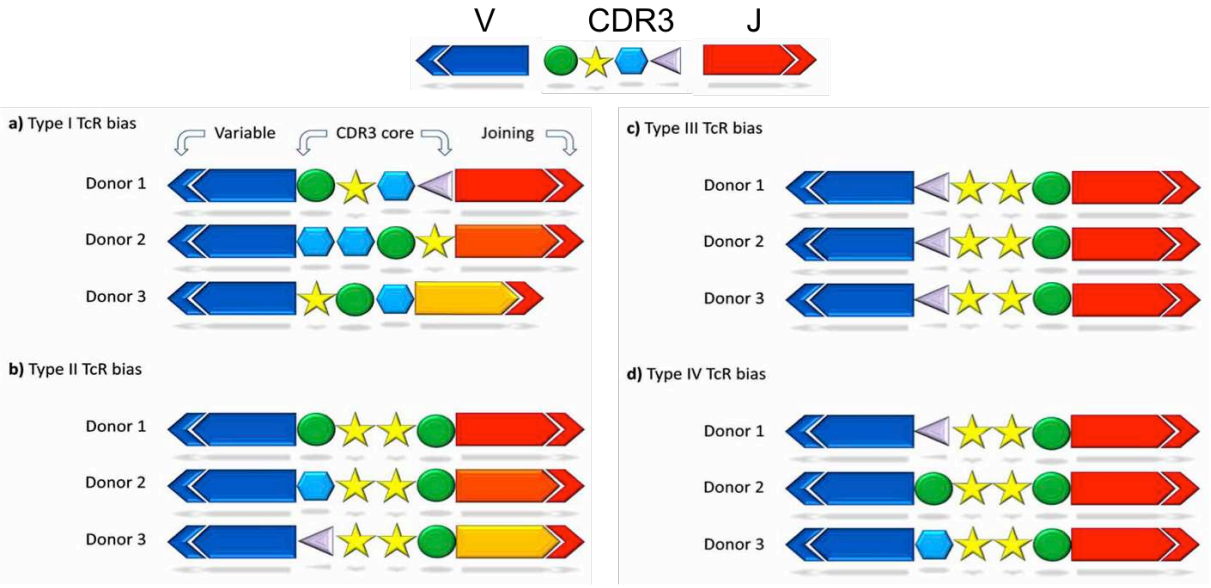


Figure 4. TCR repertoire bias classification.

Schematic model of the four categories of TCR repertoire bias. A. Type 1 bias: Identical uses of a specific Variable region across different individuals but no conservation in CDR3, Joining region. B. Type 2 bias: Identical uses of a specific Variable region across different individuals but no conservation in CDR3, Joining region. C. Type 3 bias: identical TCR chains across different individuals. D. Type 4 bias: near identical CDR3 across different individuals, differentiate by 1 or 2 AA within an invariant motif[103].

These TCR biases are present in the Teff, Treg and CTL[103]. This point out that the general rules of immune response are driven by universal mechanisms in all T-cells bearing a TCR and do not depend of the function of the cell.

This model covers most of the TCR repertoire perturbations during immune response that have been observed in many fields including infectious disease, malignancy and autoimmunity. The main purpose of this classification is to determine how the specificity of the immune response works through the TCR repertoire.

1.4.3. TCR repertoire network analysis

The preservation of specific protein sequences usually has important functional origins. The classical example is the drepanocytosis. But this seems different with the TCR. In fact, we have seen that the bias in TCR repertoire allow the presence of TCR that differ from their CDR3 sequences but have the same capacity of recognition. For example, the public TCR repertoire JM22, which engages the HLA-A2 restricted GIL peptide from

Influenza virus[105], is highly biased toward TRBV19 usage but it is not strictly public as the CDR3 residue composition and length are variable. Moreover, the α-chain is not fixed, exhibiting variable TRAV usage.

Moreover, when analysing TCR repertoire in human, you have to deal with the infinite diversity of the HLA system. The HLA molecule presents antigen and interacts with the TCR at the VJ level. Consequence of this interaction is a direct impact on the VJ usage. So, it is difficult to compare clonotypes, composed of V-CDR3-J segment, between individuals because you would not have the same HLA. So, the focus on CDR3 based on network analysis is a justified strategy.

Niels Jerne proposed the first network theory about the immune system[106]. This theory states that the immune system is interacting as a network of cells and molecules. They recognize not only things that are foreign, but also other elements of their own immune system. The immune system is therefore seen as a network, with the components connected to each other by different kind of interactions.

Network helps us to understand how TCR could have, more or less, the same specificity. Indeed, the same or very similar CDR3 sequences (Bias 2, 3 and 4) are frequently observed within repertoires of T-cells or B cells specific for a given antigen[107–110]. Using network analysis of TCRs that differ by one AA in their CDR3 sequences shows cluster of CDR3 that can be annotated with known sequence to define specificity[111]. Moreover, antigen selection account for the enhanced network connectivity, so this analysis can be used for immune response monitoring[112–114]. In fact, the TCR bias and its possible use as network is often observed in TCR repertoire during immune response.

1.4.4. TCR repertoire as biomarkers:

TCR are implicated at the early beginning of immune response, so they are good candidate to monitor the immune system state in individuals. Indeed, chronic diseases such as autoimmune diseases and cancers have in their great majority a relatively slow development. For example, in the case of autoimmune diseases like Type 1 diabetes (T1D), the first symptoms of the disease can occur only years after the beginning of the disease. Identified potential deleterious or disease-associated TCR repertoire could help preventing the development of these pathologies

Autoimmune diseases:

Qualitative or quantitative deficiency in the immune tolerance can lead to the appearance of autoimmune pathologies[115]. The destruction of cells, tissues and organs of the individuals, by their own immune system, characterize these diseases. The thymic selection steps are intended to suppress self-reactive clones but the system is not a perfect one and that some self-reactive clones are found in periphery. It can lead to the destruction of cells and tissues, and the production of autoantibodies via plasmocytes consequently of both a T and B cells collaboration and absence of control from Treg. How are the TCR of Teff and Treg involved in this mechanism?

Type 1 diabetes is a chronic autoimmune disease, commonly diagnosed in children and young adults and is characterized by metabolic disorders caused by the destruction of Langerhans islets beta cells. Several studies corroborate the involvement of autoreactive T-cells in type 1 diabetes, in particular by demonstrating major lymphocyte infiltrates in Langerhans islet of T1D [116,117]. There is a strong predisposition to T1D due to HLA-I and HLA-II haplotypes like HLA-A24, HLA-B39, HLA-B57[118] and HLA-DR3, DR4, DQ2 [119] as there are involved in presentations of autoantigen epitopes, like preproinsulin, to T-cells[120,121]. Due to this presentation, autoreactive T-cells more easily survive thymic selection and are better stimulated in periphery. Even if autoreactive T-cells are present in healthy individuals, autoreactive T-cells in T1D patients are enriched in activated/memory cells[122] indicating that they are stimulated by their cognate antigens. This is confirmed by the presence of a type 3 bias in pancreatic islet from T1D patients[123].

Multiple sclerosis is a disease of the central nervous system. It involves an autoimmune process in which CD8+ T-cells predominate to cause inflammation and destruction of myelin and nerve fibres. CD8+ T-cells were isolated from the brain by micro dissection and analysed for TCR repertoire. These results showed a clonal expansion of a small number of clones where some of them accounted for up to 35% of CD8 brain-infiltrating cells[124] exhibiting type 3 and 4 bias. Indeed, their CDR3 regions were very similar, suggesting common antigen specificities. Some brain infiltrating clones were also detected in the cerebrospinal fluid and in the blood[125,126]. The specificity of these public

clones, which appeared to be directly involved in multiple sclerosis pathogenesis, was the myelin basic protein[127,128].

Rheumatoid arthritis (RA) is also a chronic autoimmune disease characterized by the inflammation and destruction of articulations. The TCR repertoire from the synovial fluid of inflammatory articulations in recent onset patients was dominated by a small number of highly expanded clones. These results were not observed in established RA. This could be explained by the kinetic of this disease. In established RA, the articulation has been destructed by a specific repertoire and maybe there is no more auto-antigen to activate a specific repertoire but much more the presence of T-cells stimulated by « bystander activation » and cytokines. Moreover, further investigations of the shared TCR between affected articulations highlighted that the most expanded clone in each articulation was in fact the same. These results indicate that it is reasonable to think that auto-antigens in the synovial fluid are involved in the disease[129]. A recent study confirms this hypothesis. TCR repertoires of PBMC from more than 200 RA patients were compared to healthy controls. Significant differences were identified between RA and healthy controls at the V, J and VJ levels[130].

Graft versus host disease (GVHD) is a complication in hematopoietic stem cell transplantation patients. It resemble autoimmune disease as it results of the inflammation of different organs [131] and gastrointestinal GVHD account for most mortality[132]. By comparing the TCR repertoires of patients undergoing GVHD who either responded or not to first-line corticosteroid, different biases in T-cells repertoire isolated from gastrointestinal biopsies were observed and these similarities were much more observed in patients, which are refractory to treatment. Moreover, when TCR initially identified in these biopsies were tracked in peripheral blood samples, their frequency was observed to expand in the treatment refractory patients, unlikely to the responsive patients. These results could be used to develop biomarkers to stratify patients at risk for steroid refractory response[133].

Malignancy:

The studies of TCR repertoire in malignancy is of interest because tumour-associated antigens (TAA) are able to activate the tumour infiltrated lymphocytes (TIL). These TAA are specifically expressed by numerous cancers, including lung, bladder or ovarian[134–

[136]. Numerous examples of TCR repertoires bias has been reported in cancer. For example in melanoma, the TCR found around metastatic lesions and melanoma-infiltrated lymph nodes presented a type 3 bias and were specific for the TAA Melan-A peptide[137]. Another type 3/4 bias, specific for the TAA NY-ESO-1 peptide was observed in prostate cancer and melanoma [138]. The identification of these specific repertoire and cognate TAA are of interest because they can be used to design treatments. Indeed, adoptive T-cell transfer of specific TAA can mediate *in vivo* tumour regression as demonstrated by clinical responses in trials using autologous TIL derived from human melanomas[139].

When T-cells are part of the malignant tissue, the TCR repertoire could also be biased. For example, in patients with T-large granular lymphocyte leukaemia, >50% of all monoclonal T-cells exhibited type 3/4 bias with an over expression of TRBV6-5 and glutamine-glycine motif in the CDR3 region[140]. Either there is a bias in the generation of this TCR, either this observation involves that some TCR may be more prompt to degenerate into malignant T-cells.

Infectious diseases:

The first referenced cases of TCR repertoire type 3 bias in human was observed in EBV infected patients whom shared the HLA-B*0801 allele, but were otherwise genetically divergent. In this study, HLA-B*0801-restricted CD8+ T-cells specific for the FLR EBV-associated peptide (FLRGRAYGL) expressed a public motif encoded by TRBV7-6/TRBJ2-7 and TRAV26-2/TRAJ52[141]. These results were first attributed to PCR contaminations, as the probability to found the same TCR in unrelated individuals was statistically improbable according to the knowledge of this period. But, to exclude PCR cross-contamination, the authors provided both forward and reverse sequencing of the CDR3, revealing redundant nucleotide sequences coding for the same protein chain across individuals. These papers paved the way to rethink the TCR repertoire, not as a random and chaotic TCR lottery, but much more as an organised and predictable system.

TCR repertoire type 4 bias in infectious diseases was also first reported in the 90's from influenza-infected patient. CD8 T-cells specific for the GIL influenza-associated peptide were observed to be composed of TRBV19/TRBJ2-7 TCRs[112]. Since these first papers,

many other proofs of TCR repertoire bias have been highlighted in infectious diseases like EBV, CMV and HIV[58,142,143].

The presence of bias in the repertoire, including public TCR, questioned the so-called stochastic generation of the repertoire. Regarding these results, it seems obvious that some TCR may have a higher frequency of production/selection as they are found in multiple individuals. Moreover, if some TCR are more prompt to be present in the adaptive immune system, do they have the same properties than the others? What are their cognate antigens? Those are the questions that this thesis work aims at answering. The different TCR repertoires reported in the different examples of bias were all generated and selected in the thymus. The basic mechanisms of selection of a functional TCR repertoire from a randomly generated one still remain unresolved. In the next part, I introduce the current knowledge and theories on thymic selection.

2. Thymic selection: many are called but few are chosen

Located in the mediastinum, the thymus consists of 2 pyramidal lobes, themselves made up of different lobules, each having an internal medulla surrounded by a cortex. The thymus has given its name to the TL because it plays a fundamental role in T-cell ontogeny, also named thymopoiesis. Numerous studies have focused on how T-cells learn to recognize antigen that are endogenous or exogenous. Many studies demonstrated that (i) TLs are educated within the thymus and (ii) auto-reactive TLs are eliminated to prevent autoimmune disorders. The general mechanisms of the thymic selection are represented in Figure 5:

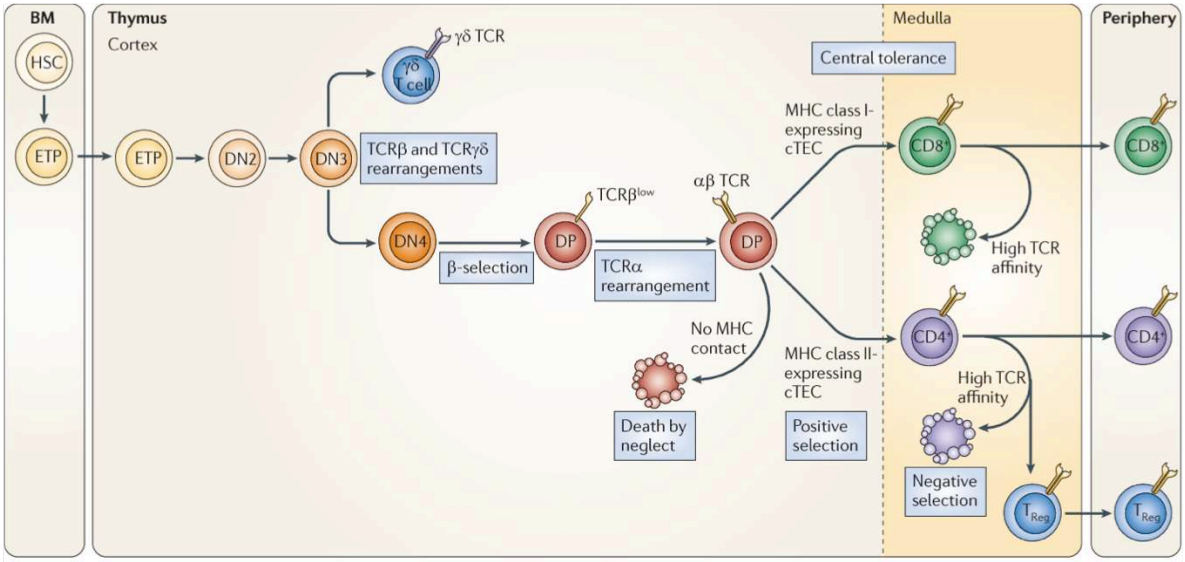


Figure 5. General mechanisms of thymic selection.

Haematopoietic stem cells (HSCs) from the bone marrow give rise to DN cells, which do not express the TCR. They undergo rearrangements of βTCR genes and β selection leads to the generation of DP with both the CD4 and CD8 co-receptors expression. At the same time, it is followed by the rearrangement of the TCR α-chain and the expression of the αβ TCR. During positive selection in the cortex, cells that failed to interact with pMHC complex results in death by neglect and those that succeed migrate to the medulla. Cells with TCRs that bind to MHC-I molecules retain expression of CD8 and inhibit CD4, whereas cells that bind to MHC-II retain CD4 and inhibit CD8. In the medulla, potential autoreactive thymocytes are deleted if the avidity of binding to pMHC is over a certain threshold. This mechanism is not perfect and some T-cell with high avidity are not deleted. Indeed, a small percentage of CD4+ thymic cells with an avidity for MHC class II

molecules just below the high limit threshold for negative selection upregulate the transcription factor forkhead box P3 (FOXP3) and become Treg.

2.1. Orientation of a common progenitor

T lymphocytes generation begins with the migration of common precursors: the hematopoietic stem cells (HSCs)[144] HSC first arise from the yolk sac and then colonized the foetal liver. The bone marrow is colonized at 11 post-conception weeks and becomes the principal site of haematopoiesis after 20 post-conception weeks[145]. It is classically admitted that these precursors are common to lymphoid and myeloid lineages and the commitment of these precursors in T lineage will begin after their entry into the thymus even if some papers challenge this affirmation[146,147]. The thymocytes encounter a lot of checkpoint for the selection of the TCR in the two distinct anatomical thymic regions: the cortex and the medulla, which are composed of two different microenvironments, specialized for TCR repertoire selection.

First of all, HSCs which arise into the thymus are characterized by the expression of $CD34^+$ and the absence of expression of CD4 and CD8 T-cell lineage markers. They are referred to as $CD4^-CD8^-$ (double negative cells (DN)). The DN step is divided in different sub-steps, from DN1 to DN4 with different phenotypes [148]. After the DN step, the second phenotype observed is $CD3^-CD4^+$ or simple positive (SPI). Then come cells expressing $CD3^+CD4^+CD8^+$ (double positive (DP))[149]. Three populations of DP thymocytes have been identified that are at distinct stages of development: DP1 that are the most abundant and that express low levels of TCR and CD5 ($TCR^{low}CD5^{low}$). DP1 represents the preselected thymic repertoire; DP2 cells are $TCR^{int}CD5^{high}$ and consist of class I- and class II-restricted thymocytes in the first 12–48 h of development; and DP3 thymocytes are $TCR^{high}CD5^{int}$ and consist entirely of CD8 lineage cells [150]. The final step is the generation of the two major type of T-cells composing the adaptive system: $CD3^+CD4^+CD8^-$ (simple positive $CD4^+$ ($SPCD4^+$)) and $CD3^+CD4^-CD8^+$ (simple positive $CD8^+$ ($SPCD8^+$))[151]. At the end of the thymic differentiation, the selected $SPCD4^+$ or $SPCD8^+$ migrate to the periphery as naive cells with the $CD45RA^+$ phenotype. About 10% of new thymic emigrants SPCD4+ are regulatory T-cells (Treg) defined by the $CD4^+CD25^+$ $CD127^-$ phenotype and expressing the $FoxP3^+$ transcription factor. The other $SPCD4^+$ are T effectors cells[152,153]. In periphery, when these cells will encounter their

specific antigens, they will down regulate the expression of CD45RA and expressed the CD45RO marker as memory cells.

The shaping and selection of the TCR repertoire starts at the DN stage through the β selection. Once the rearrangement of α and β chains is done, the DPCD3+ cells expressing a functional TCR undergoes a double process of selection. These steps have attracted much research because of the implications of T-cells in the immune response. There are several theories about thymic selection but this process still remains an open question. Because of the highly stochastic mechanism of generation of the TCR, a mechanism of selection is needed to allow survival of only functional and safe TCR.

2.2. Thymic checkpoints of the TCR repertoire:

2.2.1. Beta selection: the first breath of the TCR repertoire

Thymocytes enter the thymus through blood vessels at the corticomedullary junction and then migrate to the thymic cortex, where they undergo the β selection. At the beginning, the TCR is not produced in is final form. Indeed, the beta chain is the first one to be produces in thymocytes with V(D)J rearrangement of the βTCR genes.

In fact, there are three different types of TCR on thymocytes: the αβTCR, the γδTCR and the pre-TCR. Indeed, the pre-TCR has its own structure and function. It is composed of a newly rearranged βTCR chain combined with the pre-TCR α chain. The pre-TCR α is an invariant chain, not rearranged, comprising a single immunoglobulin-like domain that is structurally distinct from the constant domain of the TCR α chain [154].

Nevertheless, the association between the pre-TCR α and the β chain is nearly identical to that the rearranged α and rearranged β chains. The β chain is paired with the pre-TCRα chain and co-expressed with CD3 molecules on the cell membrane of DN cells. This is the first checkpoint to verify that the beta chain rearrangement generated a functional βTCR chain. It is known as β-selection and allows (i) DN cells with a productive βTCR rearrangement to differentiate into DP cells, (ii) inhibition of βTCR gene rearrangement and (iii) initiation of αTCR gene rearrangement. β-selection is driven by the pre-TCR α and is the first checkpoint of T-cell development, as DN cells that fail to generate a functional β-chain do not proceed along the differentiation pathway.

The β-selection detects successful rearrangements in the β chain loci to generate a diverse set of βTCR chains. βTCR-expressing cell populations expand and subsequently undergo αTCR rearrangement and thus generate a large TCR repertoire.

It was previously admitted that the β-selection requires recognition of pMHC through the pre-TCR [155]. More recent studies have suggested that the pre-TCR triggers differentiation signals in a ligand independent way[156]. However, how the extracellular domain of the unique pre-TCR α chain is able to recognize a productive TCR β-gene rearrangement and/or proper β-chain structure through pairing with any successfully rearranged TCR β-chain, comprising variable V(D)J rearrangement and junctional diversity, is unclear and the molecular basis for this mode of signalling remains unresolved [157,158]. One hypothesis is that DN thymocytes express a self-oligomerizing pre-TCR on T-cell membrane that requires only weak signals for survival and differentiation. Membrane expression of pre-TCR complexes containing successfully rearranged *β*TCR proteins is sufficient for providing the signals, without ligand engagement, to induce the expression of coreceptors like CD4 and CD8 as well as V-J rearrangement of αTCR genomic region. Therefore, it is speculate that the threshold for activation is shifted to a larger amount and the receptors lose their self-oligomerizing potential by replacement of pre-α chain with *α*TCR in DP thymocytes, in which a selecting peptide with appropriate affinity is required for further maturation[156,158]. More recently, X-ray crystallographic based study's authors speculated that the pre-TCR might form a superdomain of two pre-TCRs sitting close to the membrane in an antiparallel way with ligand binding precluded[159].

It allows the system to avoid loss of clones that have successfully generated functional *β*TCR whatever their specificities. Indeed, it has been found that the Vβ frequency is identical among productively rearranged β genes before and after β selection[160]. The specificity and avidity of TCR are strictly verified after, during positive and negative selection steps, to eliminate non-functional or autoreactive clones in the periphery.

Within the massive amount of DP cells that have been stochastically generated, only a subset will give rise to T-cells able to function and to protect each individual against the antigen encounters in the future. These selected cells form the functional repertoire that will be exported to periphery and form the thymic recent emigrant T-cell subset. To select this sub-repertoire within the stochastic one, two major processes have been

highlighted: positive and negative selection. These processes are different of the β-selection in the way that they involve the specificity of the TCR for pMHC complex and so the recognition of antigens.

2.2.2. Positive selection: a functional repertoire

The stochastic generation of the repertoire leads to the generation of many non-functional TCRs. The positive selection processes the stochastic repertoire to select only TCRs that are functional. The productive rearrangement of an α chain is not sufficient to trigger the end of recombination. Indeed, only TCRs with an α chain able to form a MHC-restricted and functional receptor when paired with the β chain will inhibit recombination(Brandle et al., 1992). T-cells that get a survival signal based on a sufficient but not alarming interaction with pMHC are positively selected. This process is known as positive selection, is initiated in cortical DP thymocytes and takes several days to finalize[161]. DP precursors expressing a TCR able to recognize, in a restricted and acceptable manner, a MHC Class I or II restricted receptor produce (i) a negative signal for RAG gene transcription, (ii) long-term survival and (iii) migration into the medulla.

During this process, more than 90% of the T-cells are not able to interact with pMHC so, they are deprived of extrinsic signals that maintain cellular homeostasis and will therefore die by neglect. Death by neglect is a form of apoptosis that appends when TCRs failed to engage with pMHC as they express TCR either non-productive or with inappropriate conformation unable to mediate activation signals [162,163].

At the end, it is not so surprising that very few TCR can finally engage pMHC. First of all, the generation of an MHC-restricted receptor from stochastic TCR generation process is thought to occur relatively infrequently[164]. As an optimized mechanism, the system is able to test multiple combination of αβTCR. Indeed, DP cells in the cortex express high level of the RAG complex. The α locus is made such that several V/J recombination can append on the same allele, each time resulting in excision of the prior recombined DNA[165].

This process, potentially infinite, allows different productive TCR α gene rearrangements to be tested per cell and provides an optimized process to screen rare clonotypes while expending a minimal amount of metabolic energy. This process is only limited by the DP cell, which has an average life span of 3–4 days.

Secondly, the TCR must be pMHC specific and studies have shown that positive selection involves both recognition of peptide and MHC[166]. Historically, it was suggested that positive selection ensure that only T-cells able to interact with antigen restricted to the MHC alleles would migrate to the periphery to defend the host[167]. But, it seems that positive selection is not just defined by MHC interaction whatever the peptide complexed with the MHC. TCR are MHC restricted, but MHC molecules need to be associated with peptides to be presented at the cell membrane and positive selection involves both recognition of peptide and MHC, as does T-cell activation in the periphery.

The involvement of MHC-associated self-peptides in the positive selection was discovered in experiments that used polyclonal T-cells selected by various MHC class I variants. The substitution of amino acid residues in the deep peptide-binding groove of MHC molecules, which can interact with the peptides but not TCRs, modified the binding of the peptides and so the repertoire of positively selected T-cells, suggesting the role of both peptides and MHC recognition in the process of positive selection[168]. If some variations of peptides that are complexed and presented by MHC are so important, the TCR specificity and /or affinity for these peptides seems to be one of the main requirements for a TCR to be selected or not. A fundamental question in the field regards the constitution of the peptides that stimulates thymic selection.

Indeed, the thymopeptidome, which is the set of peptides specifically presented by endemic antigen presenting cells implicated in the positive selection: the cortical thymic epithelial cells (cTEC), is a small part of the pMHC possible numbers. The cTEC are present in the cortex of the thymus and are responsible of the positive selection[169]. The positive selection requires the thymopeptidome that is generated by proteolytic enzymes complex exclusively expressed by cTEC: the thymoproteasome. The thymoproteasome exhibits altered proteolytic activity to produce a repertoire of self-peptides linked to the MHC-I and are implicated in the positive selection of CD8 cells[170]. cTECs in thymoproteasome KO mice expressed an altered set of pMHC repertoire that fail to positively select most of the CD8. Moreover, CD8 selected in these KO mice displayed a defective TCR repertoire that is incompetent in promoting allogeneic or antiviral responses. Thus, this study revealed a unique role of the thymopeptidome expressed by cTEC in the positive selection of a functional repertoire of CD8+ T-cells[171].

For the selection of CD4 thymocytes in the thymus, another proteasome complex is implicated. The thymus-specific serine protease (Tssp) and cathepsin L are highly, but

not exclusively, expressed in cTEC. Experiments with Tssp KO mice indicate its necessity for an optimal positive selection of CD4[172,173].

These observations question the dogma of the highest diversity to be produce in the thymus. Indeed, it is unlikely that the stochastic generation of the TCR repertoire of DP unselected cells would produce many TCR that matched to the particular pMHC that are present at a given time in the thymus. Moreover, this hypothesis is questionable if we considerate that a given TCR bind a single pMHC.

2.2.3. Negative selection: an inoffensive repertoire

The previous cortical positive selection by TCR ligation does not provide the end of the selection process. Once they have been positively selected, cells increased the numbers of medulla chemokine receptors and migrate into the medulla for the negative selection[174], even if it seems that these two processes could be concomitant[175].

The basic principle of negative selection is that thymocytes expressing TCRs with a too high affinity for self-peptide are deleted because they are potentially self-reactive. The purpose is to generate a largely self-tolerant peripheral T-cell repertoire. T-cells that obtain too strong signals from the pMHC complex get an apoptosis signal to avoid presence of autoimmune cells in the periphery.

It is important to note that the TCR is extremely discriminating in its responses to pMHC. A clear but tiny threshold between positive and negative selection has been defined, where the affinity difference between the weakest ligand that induces negative selection is barely higher than one that induces positive selection[176,177]. Considering the importance of bystander activation and cytokines impact on thymocytes activation, this observation is a clue for a potential mechanism explaining involvement of TCR in autoimmunity even if they have been negatively selected.

As positive selection in the cortex, negative selection is triggered by specific antigen presenting cells: the medullary thymic epithelial cell (mTEC). They are responsible of the deletion of auto-reactive thymocytes in the thymus and are crucial for maintaining self-tolerance. Like the cTEC, the mTEC generate a specific repertoire of peptides presented to thymocytes for negative selection. The mTEC uniquely express the transcriptional factor autoimmune regulator (AIRE) giving rise to the promiscuous gene expression, which contributes to a low amount expression of many genes including tissue-restricted antigens (TRA)[178]. Indeed, transcripts encoding a diversity of TRA are

expressed specifically in mTEC[179]. AIRE complex is necessary for thymic selection as it is involved in the generation of the specific peptidome found in the mTEC. The absence of a functional AIRE is the cause of a disorder known as autoimmune polyendocrine syndrome type 1 (APS-1) or autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED)[180]. Thus, the hypothesis is that AIRE regulates the thymic expression and presentation of TRA and thereby controls negative selection and consequently autoimmunity.

The negative selection has yet several limits. The potentially infinite number of self-peptides presented by MHC prompts the question of whether it is possible to select thymocytes against the recognition of all self-peptides during negative selection. The diversity of self-peptides that could be presented is about 10^{15} based on simple combinatory arguments[181] but it is more closely aligned with 10^5 different self-peptides that are expected to bind to one human MHC class I allele[182]. The probability that a thymocyte specific for only one self-antigens escapes negative selection has been calculated based on the different parameters: the probability that a given self-peptide is presented by any given APC in sufficient numbers, the number of unique APC encountered during selection and the number of copies of a given self-peptide that an APC needs to present in order to cause deletion. The probability that a TCR specific of a single self-peptide escapes negative selection was estimated between 10^{-11} to 0,8[183]. The author concluded that the impact of negative selection on the repertoire diversity against self-antigens seems to be very weak. Instead, they suggest thymic selection operates on a restricted subset of self-pMHC, a constraint imposed by the number of APC present in the thymus and encountered during the selection process. This imposes that additional tolerogenic mechanisms act in the periphery to prevent autoimmunity of autoreactive TCR that could be activated by self-antigens not encountered in the thymus. Indeed, under this medulla specific environment, a strong TCR signal could also lead to the selection of an important T-cell subpopulation: T regulatory cells[184]. Foxp3 Treg cells were shown to develop in the thymus from $CD4^+$ T-cells that express an autoreactive αβ TCR[152,185]. They lie at the higher end of the spectrum of acceptable self-reactivity. However, it appears that the fixed-threshold model to include a fixed range of affinity or avidity for Treg selection is not sufficient to explain many experimental observations[186].

In conclusion, the model accepted about negative selection depends of the level of activation based on the recognition of self-peptides. A too strong activation signal induces apoptosis and a signal below the apoptosis level leads to negative selection. Between these two levels of activation, Treg are selected. This model is based on a simplistic "key-lock system" of the TCR recognition and is not fully consistent with the observation made on the lineage fate.

2.2.4. Lineage fate

In the periphery, there is a predominance of CD4 over CD8 T-cell that is known for a while but the causes of this bias have remained obscure. Thymocytes CD4/CD8 ratio is about 4/1 in thymus, which is highly conserved across several species, suggesting that the selection mechanisms involved are fundamental to the processes that give rise to peripheral T-cells. But what is the source of this bias? First, there may be a significant difference in the numbers of precursor DP thymocytes expressing functional TCR that are able to recognize MHC class I vs MHC class II during selection steps. The second hypothesis is that the thymic selection process of the two lineages are different and their efficiencies and the probabilities with which cells are selected or die at each checkpoint are different for TCR restricted MHC-I or MHC-II[150].

A first explanation seems to emerge with interesting experiments using tetracycline-inducible zap-70 (TetZap70) transgene mouse model to induce the selection of SP thymocytes in zap-70 knockout mice. As described above (1.2.2), ZAP-70 is a protein tyrosine kinase that is associated with the TCR signalling subunits CD3 and it has been shown that ZAP-70 plays a fundamental role in T-cell development[187,188]. Indeed, in human, several severe combined immunodeficiencies (SCID) present mutations in the zap-70 gene resulting in the absence of ZAP-70 expression[189]. The zap-70 KO mouse model has thymocytes development blocked at the DP stage and can be reconstituted following injection of tetracycline to induce tetracycline-inducible zap-70 transgene expression. The authors generated chimeric mice in which TetZap70 development occurred in the absence of either MHC-I or MHC-II. In this way, they were able to check both simultaneous and independent development of SPCD4 and/or SPCD8. Thymocytes from DP1 to DP2 were very similar between class MHC-I- and MHC-II deficient TetZap70 mice, suggesting that TCR MHC-I restricted precursors are not less generated in the randomly generated TCR repertoire and are not limiting for the selection. Moreover, the

estimation of the DP2 precursor ratio with mathematical models supports the fact that the principal source of bias toward CD4 and CD8 is not a preferential generation of TCR MHC-II restricted. In conclusion, CD4 and CD8 asymmetry do not derive from a significant asymmetry in the generation of TCR repertoire to recognize MHC-I or MHC-II.

In a second time, the author analysed the developmental course of thymocytes in vivo. They revealed that TCRs MHC-I restricted have a higher death rate among DP2 thymocytes than those MHC-II restricted. By rating the number of cells starting the selection compared with those compelling it, they estimate that the MHC-I restricted cell selection was much more stringent. These results are summarized in Figure 6. Together, these results highlight clear evidence that the selection bias towards CD4 instead of CD8 is influenced by a much more stringent mechanism of selection for TCR MHC-I restricted cells[150].

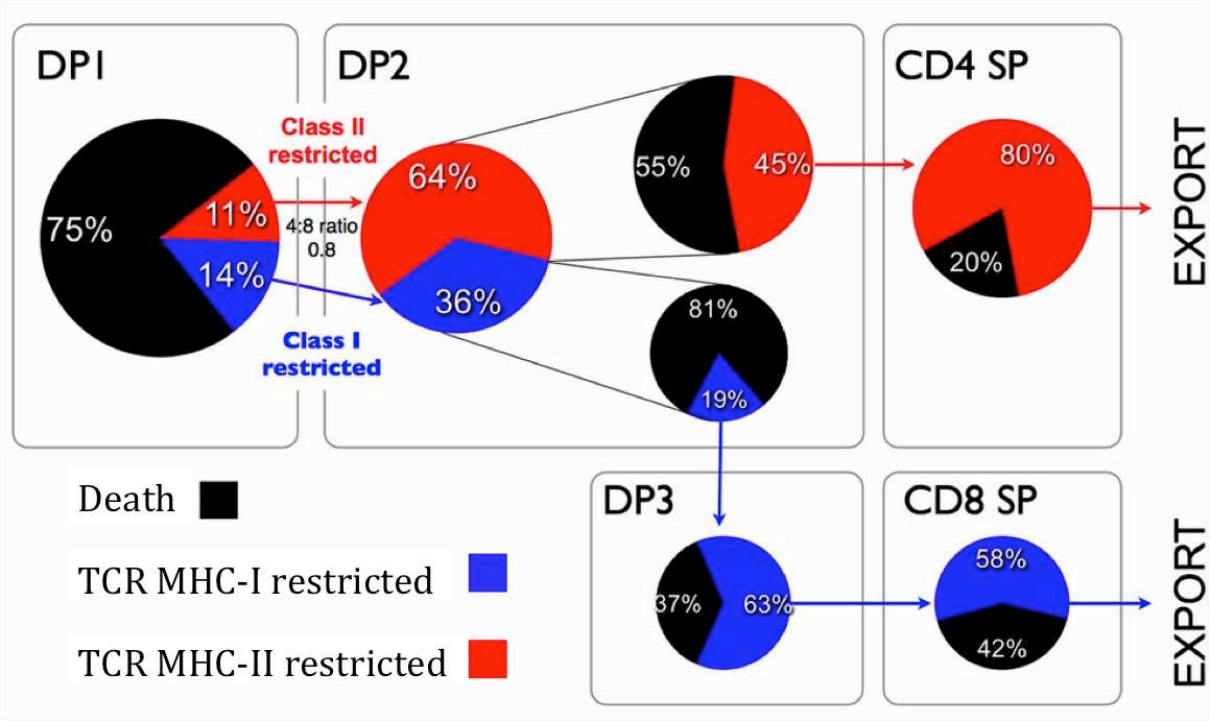


Figure 6. Lineage fate estimation during thymic selection.

Pie charts show the fate of MHC-I and MHC-II restricted thymocytes that initiate selection to the periphery (From Sinclair et al. 2013).

For a while, it has been accepted that the lineage fate is determined with accuracy by the MHC specificity of the TCRs that DP thymocytes express. DP cells with TCRs that are specific for MHC-I fall into CD8+ T-cells whereas DP cells with TCRs that are specific for MHC-II direct differentiation into CD4+ T-cells[190]. But these datas are also contested by

the fact that there is an overlap between the TCR repertoire of the different thymocyte populations that exclude the strength of the TCR signal as a unique factor of T-cell fate in the thymus. In fact, it fits with what we have been exposed bellow (1.2.2): the TCR does not act as a "lone wolf". The implication of coreceptor, microenvironment and others molecules are known to be fundamental in the activation of T-cells and are also implicated in the process of lineage determination. Indeed, TCR specificity participates in thymocytes lineage fate during thymic selection by whether TCR signalling persists throughout selection or is disrupted, allowing selected thymocytes to be activated by cytokines. This is the kinetic model of fate determination[191]. In this model, persistent TCR signalling induces expression of a specific CD4 lineage transcription factor: Th-POK[191], whereas cytokine signalling induces expression of a specific CD8 lineage transcription factor : Runx3d[192]. Thymocytes that are selected based on the recognition of MHC-II express a lower level of CCR7 chemokine receptor than those selected on MHC-I. But this model is not fully right as a study challenges the admitted theory that DP thymocytes differentiate into SPCD4 or SPCD8 by selectively silencing the transcription of those coreceptors based on the affinity with MHC. Indeed, it demonstrates that DP always silence CD8 coreceptor transcription, even if they later differentiate into CD8SP T-cells. This CD8 late differentiation, also name "coreceptor reversal" is under the control of cytokines, notably IL-7[193].

In conclusion, these different studies suggest that the highly stringent thymic selection process of thymocytes is a cost effective but necessary process. It balances the generation of stochastic TCR with the presentation of a narrow peptides repertoire to select a sub-repertoire of T-cells that will be efficient in periphery but not too harmful. The models of positive and negative selection are now widely accepted based on the strength of the signal. But, models about how the strength of the signal is determined by the thymocytes based on TCR interaction are still under discussion and investigation.

2.3. Models for TCR thymic selection

2.3.1. The avidity model of selection

The avidity model predicts that the selection is derived from a stimulation gradient related to the concentration of the antigenic peptide and through its capacities to trigger transcription factors activation in thymocytes via its TCR recognition.

The first experiments on the nature of thymic selection examined the effects of low concentrations of cognate antigenic peptides in *in vitro* foetal thymi of TCR transgenic mice. These studies revealed that low concentrations of peptides induce differentiation into SP thymocytes. On the other hand, high concentrations of the same antigens induced negative selection. These finding paved the way of the avidity theory to explain thymic selection. It suggest that positive selection requires subthreshold engagement of TCRs due to the limited avidity of pMHC ligands[194,195].

In other models using a fixed concentration of thymic presented peptides, like the OVA-specific TCR transgenic mouse model, the avidity theory was confirmed by the fact that peptides with low capacity of stimulation would favour positive selection, while higher affinity interactions favour negative selection[196].

We have shown previously that coreceptor can influence the duration and strength of contact with TEC, the TCR signalling and therefore have an impact on thymocytes fate through the transduction of the TCR signal. However, these co-receptors have a conserve structure and are constitutive of all TLs, suggesting that the only one factor that fate TLs is the TCR. In this model, multiple TCR/pHLA interactions form a signalling gradient that defines cell fate. Numerous weak contacts or, on the contrary, limited high-affinity interactions provide sufficient avidity to induce signals that reaches a positively selecting threshold to fate thymocytes into CD4 or CD8. Stronger signals are required to reach a tolerizing threshold and fate thymocytes into Treg.

The thymic selection avidity model fits well with the kinetic model of thymocytes activation previously evocated.

2.3.2. The qualitative model of thymic selection

This model is much more a Manichean selection model. It is mostly based on the AA sequence and physico-chemical properties of the antigenic peptides. It suggests that peptides promote always the same positive or negative signal of selection for a specific TCR. In this model, positive selection is achieved when TCR interactions generate a unique and weak signal[197,198].

Historically, this model was supported by the demonstration that antagonist peptides, i.e. peptides that inhibit the activation of mature thymocytes, are promoters of the positive selection, while, agonist peptides, i.e. peptides that activate mature thymocytes, were shown to promote clonal deletion[199–201]. These experiments support the thesis of

the qualitative model that different TCR corresponding to different antigens and subsequent signals define different thymocytes fates. This model was supported by other studies using altered peptide ligands. They have shown that the transduction signal was modified with these peptides, leading to a lack of ZAP-70 activity[202]. However, numerous others experiments mentioned the contrary based on the fact that some antagonist peptides have been shown to induce clonal deletion[203] or to inhibit thymocytes development[204]. Moreover, *in vivo* experiments have shown that there is no correlation between antagonist peptides and positive selection[205]. The demonstration that antagonist ligands are responsible of positive selection is not absolute. Indeed, thymocytes could be either positively or negatively selected in the presence of antagonist or agonist peptides. Others studies looking at the intracellular transduction signals associated with altered peptide ligands have also challenged the concept of antagonist peptides inducing unique signal. Indeed, they have shown that different altered peptide ranging from agonist to antagonist ligands can transmit a gradient of intracellular signals. The phosphorylation profiles obtained with antagonist peptides could be reproduced with low concentrations of agonist antigen[206]. From this observation, it remains difficult to accept the existence of antagonist specific signals that trigger the selection.

The capacities of a peptide to be antagonist or agonist are relative characteristics that have to be moderated. It seems that these characteristics are distributed along a gradient for every antigen[207]. This is a combination of concentration and affinity that contributes to the distinct effects of antagonist or agonist ligands on thymocytes.

The relationship between positively and negatively selecting peptides distinguishes a qualitative model from a quantitative model. A qualitative model stresses the distinction between these two forms of selection through the peptide-mediated generation of unique signals, whereas an avidity model views selection as a gradient based on TCR-pMHC affinity in combination with coreceptor signalling. The observation that low-affinity peptides induce positive selection more efficiently than high-affinity ligands is consistent with either model. However, studies reporting an overlap between positively and negatively selecting ligands clearly favour an avidity model. The majority of experiments support the avidity model for thymocytes development.

Conclusions of these different observations are that both negative and positive selection depends, for a major part, on TCR signalling. How this signal is interpreted by thymocytes is one of the more haunted questions in the field of immunology and T-cell development. Deciphering the signals that drive the development of CTL, Teff or Treg is a field of intense research. It is largely accepted that positive and negative selections allow only cells with functional TCR that will not be self-reactive in the periphery. But the different phenotypes observed in individuals with a quantitative/qualitative defect in T regulatory cells, always leads to the apparition of an autoimmune phenotype. It is a proof that there are, in periphery, presence of auto-reactive T-cells that have passed the thymic selection. The composition of the T-cell repertoire after thymic selection arises from the use of self-ligands in the selection of the TCRs that compose the peripheral T-cell population. It seems like the thymopeptidome acts as an essential "test set" predictive of the capacity of a T-cell to recognize future presented foreign antigens and not too much autoreactive.

2.4. Dynamic of TCR repertoire during thymic selection:

Based on the previously exposed theories and experiments, it is difficult to have a clear view of the evolution of the TCR repertoire in the thymus during selection. The difficulties to access human thymus samples limits our knowledge on the formation and the selection of the human TCR repertoire. There are few previous publications on the repertoire in human thymus but nearly none on sorted cells. The only ones that have explored the TCR repertoire during human thymic selection were by using humanized mice [208–211]. One paper published in 2019 investigates a bit more deeply the thymic selection in humanized mice with cell sorting and TCR repertoire analysis at different level[212].

First of all, at the V and J level, a conserved usage of several rearrangements in humans and humanized mice has been observed during thymic selection. The distributions of V-J rearrangements in DPCD3- cells in the thymus of humanized mice, that represent the TCR repertoire before selection, and in mature T-cells CD3+ of the periphery, that represent the repertoire after selection, have been studied. The distribution of V-J rearrangements was very similar in DPCD3- thymocytes and in peripheral T-cells[209 212]. It suggests a common mechanism for generation and selection of these rearrangements.

One hypothesis is that the distribution of these conserved rearrangements might be genetically programmed during the "not so much" stochastic generation of the TCR repertoire, independently of TCR-pMHC interactions during thymic selection.

At the clonotypes level, the clonality of the different thymocytes population, DP, SPCD4 and SPCD8, was also investigated. Their clonalities were very low compared with peripheral populations. Furthermore, the clonality is higher in SP thymocytes versus DP thymocytes. This reflects (i) the production and selection of a highly diverse repertoire in the thymus for DPCD3 thymocytes, (ii) the narrowing of the repertoire at the clonotypic level due to thymic selection and (iii) antigen-driven expansions in periphery.

At the CDR3-β level, the number of shared CDR3 was investigated in humanized mice that have received the same human thymus. The sharing between mice was first higher at the AA level compared to the nucleotide level. The Shannon divergence index, based on the frequency of shared sequence, was decreased for selected thymocytes compared with unselected. Together, these observations suggest that thymic selection results in selection of shared CDR3-β sequences in humanized mice[212]. The highest proportion of sequences that has been observed between two autologous mice accounted for 3,5% of the SPCD4 repertoire.

The convergence characteristic of these selected CDR3 was compared for shared and unshared sequences. The purpose was to investigate the number of nucleotide sequences corresponding to each AA sequence present in the different thymocyte populations and shared or not between mice. As predicted, the average nucleotide sequences per AA sequences was close to 1 for unshared sequences, but it was significantly higher for the shared CDR3. This indicates a convergence mechanism to generate shared sequences and a preferential thymic selection of these sequences. What was really surprising is that the proportion of shared CDR3 was not different between mice with allogenic versus autologous thymi[212].

In the same paper, the authors tried to characterize the CDR3 that are shared versus the other. They revealed that the shared CDR3 sequences were significantly shorter than the other. Also, the number of inserted nucleotides at the V-D and/or D-J junctions was lower in shared sequence.

As these TCR were specifics in their structures, the authors investigated the potential functionality of the shared sequences. They compared the repertoires of shared and unshared with a dataset of CDR3 that have been defined as cross-reactive. The definition

of cross-reactivity in this paper is related to in vitro expansion greater than 2-fold in mixed lymphocytes reactions of a human peripheral blood sample against 2 different allogenic donors sharing no HLA alleles. This definition is open to criticisms, as authors do not know if the expansion is due to a specific activation of the cell via TCR recognition or due to bystander activation. However, more than thousand CDR3 were identified to be "cross-reactive". The analysis revealed a significantly higher presence of these cross-reactive sequences in shared repertoire versus unshared.
Overall, these observations demonstrate that the thymic selection narrow the TCR repertoire. The shared sequences after thymic selection is consistent with the existence of public CDR3 sequences with convergent recombination and with fewer nucleotide insertions[213]. Moreover, these shared sequences seem to have specific functionalities.

Thymic selection mechanism is still an open question. How the immune system is able to shape and sub-select a functional repertoire in the infinite universe of stochastic CDR3 generation is one of the greatest mysteries of immunology. The models of positive and negative selection are admitted to be the two fundamental aspects of thymic selection. However, the mechanisms by which the TCR recognize endogenous peptides, the threshold of selection and the composition of antigen repertoire are far to be resolved. Some clues are emerging for recent studies including the last paper described. Indeed, the selection process does not seem to be done with the higher specificity has it been though before. These results fit well with the fact that the antigen repertoire is also restricted by TEC in the thymus. Moreover, the presence of public sequences in periphery in different individuals indicates the existence of universal mechanisms of generation and/or selection of TCR with higher probability. These results imply functional characteristics of these sequences, as there are more often presents in the TCR repertoire.

3. On the necessity of unconventional CD8 T-cells

A high level of antigen specificity is an accepted feature of T-cell activation, and moreover, of adaptive immunity. This affirmation is the actual dogma. Indeed, a single TCR has been for a while, considered to be restricted to a single epitope[214,215]. Let's do a military parallel considering this dogma. It is as if an infantryman could defend his country only against a single, unique and specific belligerent, his military career boils down to trying to find him... Moreover, based on these admitted affirmations, the generation of a useful TCR repertoire needs three conditions:

First, the number of individuals peptides presented by the host, the peptidome repertoire, has to be sufficient, otherwise the useful repertoire could not be shape in the thymus and some pathogens could not be recognize in the periphery.

Second, the specificity of the TCR repertoire has to be high enough to respond to foreign antigen but not to endogenous peptides.

Third, the frequency of T-cells specific for a foreign antigen has to be sufficient to generate a protective immune response before the infectious agent overwhelms the host.

Considering that the maximum number of T-cells is 10^{12} in an individual and that the number of peptides that can be presented by self-MHC is estimated to be greater 10^{15} at the same time[216,217], the model of one TCR responding to one antigen seems unlikely. Indeed, more than 10^{15} T-cells, which would weigh more than 500 kg, would be needed to provide efficient coverage of the peptidome[218]. Moreover, it has been suggested that the immune system would be incompetent to defend the host against injuries if only one TCR would be able to recognize only a single peptide presented in a by HLA[219]. While it is evident that a certain degree of specificity is necessary for lymphocyte activation, overlooked studies reveals that one TCR could recognize numerous different peptides that do not necessarily show related sequences. Several observations described below highlight the fact that evolution has selected not a stochastic but much more a public TCR repertoire with improved functional properties.

The unconventional T-cells are classically defined as T-cells that (i) recognize non-polymorphic antigen-presenting molecules. Some of these are encoded by genes outside the MHC locus (CD1 or MR1) or within the HLA locus, (ii) they do not recognize classical peptide antigens; and (iii) they are mainly public. Functionally, they are abundant populations of T cells that are rapidly recruited during immunological responses.

Moreover, these cells are limited in their TCR diversity and are defined with an antigen specificity different of the classical pHLA specificity. For example, invariant NKT are specific for lipid antigens[220]. and harbour an invariant TCR α chain (TRAV10/ TRAJ18 in humans) and a limited, but not invariant, range of TCR β chains[221]. In the next part of this manuscript and in our results, we highlighted the existence of, and not previously defined, unconventional CD8 T-cells. These cells are defined by (i) their capacities to interact with multiple HLA, (ii) they harboured CDR3 sequences that are "hard-wired" for innate-like response to multiple viruses and (iii) these CDR3s are mainly public, independently of the age, the sex or the HLA of the individuals, likely to indicate that these naive T-cell repertoires contains particular TCRs with a selective advantage. To summarize our findings, we propose that a whole set of αβ CD8+ T cells harbour pleiospecific TCRs that have an innate-like sensing of pHLA complexes, a concept far beyond mimotopic cross-reactivity.

3.1. Evolution has gathered a loyal following

The adaptive immune system has been shaped and selected trough evolution to respond a variety of pathogen-derived molecules. This capacity to recognize diverse antigens is conferred by the huge diversity of TCR that are made by different stochastic process, generating a random repertoire. The randomness and diversity of this recognition is often observed in T-cell responses, in which the TCR repertoire responding to a particular antigenic epitope consist of many different TCRs. Indeed, in most T-cell responses, the responding TCR repertoire consists of some private TCRs [222,223]. These observations seem logical on a mathematical point of view. Indeed, to contribute to a public TCR repertoire, we consider that (i) TCR must be produced by genetic recombination and stochastic junction from among the 10^{23} even more recently estimated from the 10^{61} possible combination of TCRs generated within the DP stage in the thymus[224]; (ii) TCR have to be part of the only ~3% of TL that accomplish the thymic selection, representing a potential peripheral diversity laying from 10^{21} to 10^{59}. These estimates are at least larger than the estimated number of T-cells in a human, which is 10^{12} [225]; (iii) to survive in the periphery, clonotypes need to have sufficient frequency and avidity to compete effectively with the other available TCR of the repertoire specific for a given antigen. Thus, if there were an equal probability of producing each of the

different possible TCRs, one would only rarely expect the same TCRs to be present in the repertoire of 7×10^9 individuals.

In conclusion, within a stochastic repertoire, the generation and selection of a unique TCR found in multiple individuals, requires many steps and this makes public TCR statistically unexpected. But, how can a large majority of individual respond exactly the same way to common pathogens? Either there a multitude of private TCR repertoires that have a public response, either there are few public TCR in the repertoire with the capacities to respond to multiple pathogens, or maybe a combination of these two hypotheses?

3.1.1. Probabilistic models for TCR generation:

Public TCRs have been observed in numerous TL responses, in different species and in different kind of pathologies. Many example of studies have highlighted the presence of public TCR in CTL response against CMV[143,226], EBV[227,228]. It can be just by chance as these viruses infect a large part of the population. However, public TCR have also been detected in CTL responses to HIV that is hopefully, not a common virus for human species. The presence of such TCR could be the reflect of others/ancient retrovirus infection? One have to keep in mind that 8% of the human genome is composed of endogenous retroviruses…[229]. In both cases, some TCR seem to be not so specific. As the public TCRs are observed in many cases of immune responses, it suggests that there are fundamental biases in the generation and/or selection of the TCR repertoire.

The generation mechanisms of the TCR repertoire has been recently dusted thanks to the development of theoretical models for TCR generation mechanism[213,230–232]. These models allow the assignment of a generation probability (P_{gen}) to the TCR. The P_{gen} reflects the probability to observe a particular sequence in the pre-selected TCR repertoire.

First of all, we have seen that different observations and experiments pointed out the fact that germline gene segment usage in V(D)J recombination is not equal between the different genes combination[22,233–235].

Secondly, it seems that CDR3 do not have the same generation probability. Recombination leads to sequences being generated more frequently than others. For example, in enzyme terminal deoxynucleotidyl transferase (TdT) KO mice, public TCR sequences are observed suggesting that they are easier to generate because they do not

require random nucleotide addition[236]. The generation of a nucleotide sequence that do not need any addition or deletion is likely to occur more often than CDR3 requiring it. Public TCR are made of germline recombination involving any or minimal random nucleotide junctional modifications[227]. However, other studies on public TCR have revealed that their sequence could also include a substantial number of nucleotide additions[143,232].

The theoretical model of generation probability is based on the usage of the V, D and J genes for recombination, their degree of trimming and the nucleotides insertion. This model could be used to estimate the likelihood that different individuals could share CDR3 or to estimate the impact of the immune response on the TCR repertoire[231,237].

3.1.2. Convergent recombination:

Convergent recombination mechanism has been proposed to play a role for the generation of public repertoire[232]. It assumes the fact that the same TCR sequences can be made by different means. Indeed, there are different examples of V(D)J recombination events that produce the same nucleotide sequence and many different nucleotide sequences that encode the same AA sequence[238–241], the latter partly due to the redundancy of the genetic code which allows different CDR3 nucleotide sequences to encode the same AA CDR3 sequences. This can explain the observation of private nucleotide sequences encoding public AA TCR sequences[228,242–244]. Moreover, the convergent recombination mechanisms allow the production of CDR3 sequences that are made of few distinct AA usage, but with a conservation of a global AA pattern which are associated with public T-cell responses, like the type 4 bias previously described[245]. The germline-encoded sequences compose the conserved motifs and one or a few positions in the CDR3 varying between sequences. These varying positions usually appear at the V(D)J or VJ gene junctions. These different mechanisms were proposed to explain the presence of public TCR in the naïve repertoire and are shown in Figure 7:

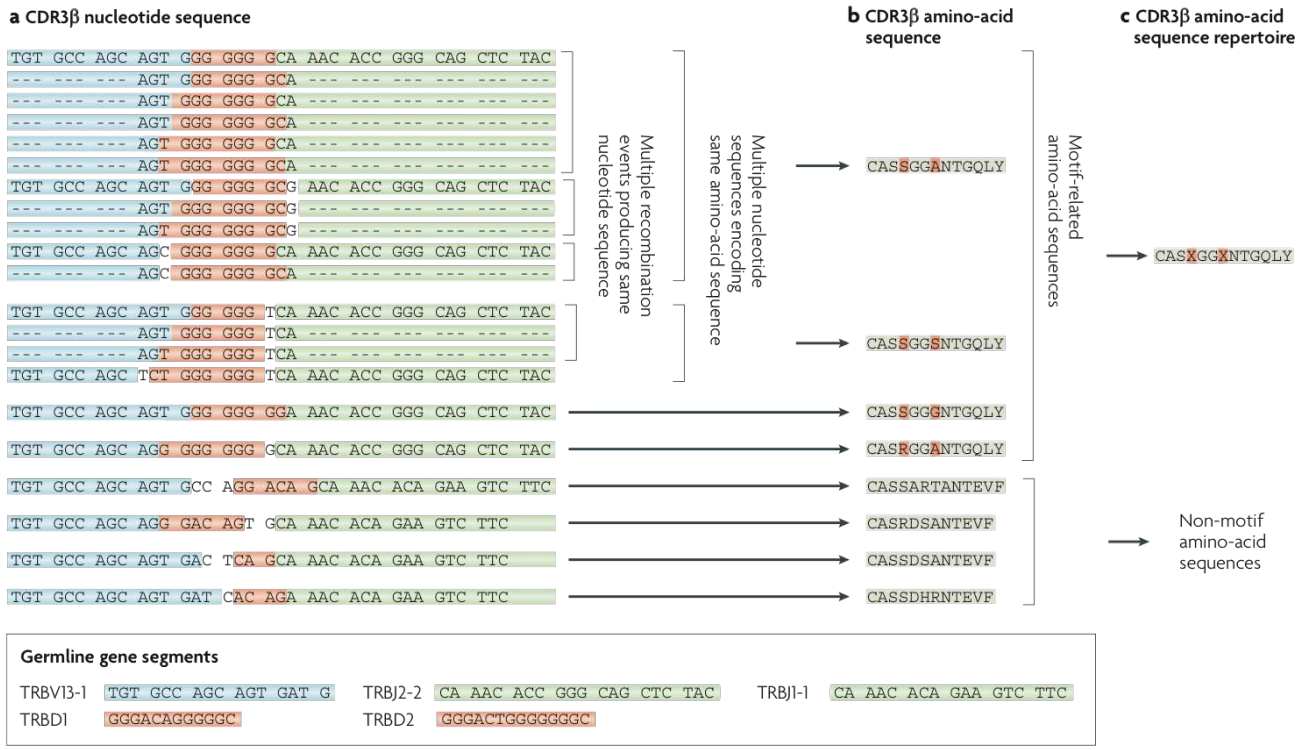


Figure 7. Different mechanisms of CDR3 convergences.

Convergent recombination mechanisms are exampled by CDR3 beta sequences that are expanded during influenza virus infection in response to the NP_{366} epitope[246,247]. Variable gene: blue, Joining gene: green, and Diversity gene: red. A. First level of convergent recombination relies on multiple recombination events, involving different splices of the germline producing the same nucleotide sequences. The random nucleotide additions (not highlighted) could also lead to the same nucleotide sequence. B. Second level of convergent recombination involves different nucleotide sequences encoding the same amino-acid sequence due to redundancy of the genetic code. Protein sequences of the CDR3 that are encoded by many codons can be encoded by different nucleotide sequences. C. Third level of convergent recombination is at the level of the TCR repertoire, where some of the AA sequences form to a specific motif in the sequence, as illustrated in the type 4 bias. In this particular case, the 'XGGX' AA motif (where X are variable AA). The different sequences are still functional to recognize the same antigen[237].

In the periphery, the presence of public TCR in the activated/memory population can be explained by the nature of the pMHC. Indeed, if the conformation of the presented antigen is very specific, it will restrict a narrow repertoire of public TCR [248]. Some studies have suggested that the intrinsic structure of the public TCR determines whether a T-cell response is public. Observations have raised that the specific structural features of the public TCR and its interactions with the pMHC complex may provide an antigen-specificity advantage that drives the public nature of the response[249,250]. The presence of

public AA motifs in different clonotypes sequences are observe in public response against pathogens [239,240]. The encoded germline sequence is just a little modified by one or two AA that don't change the final specific function of the TCR.

Despite the unlimited theoretical diversity of TCR, identical clones are also often observed responding to different pMHC in different individuals. The conclusion of this observation goes against the dogma that the repertoire is generated and selected as diverse as possible. Based on the clonal selection theory[2], the presence of public TCR implied a narrowing of the repertoire that is incompatible with the unique recognition of antigen within the diversity of a universe of peptides. We can explain the presence of public TCR only if we accept that public TCR, at least, are cross-reactive.

3.2. TCR: The Cross-Reactive:

One of the most interrogative mechanisms in immunology is how the thymic selection can shape a functional repertoire specific for pathogens, while these are not present? How can a TCR recognize a peptide while it has never seen it before? It seems more and more evident that the answer leads in favour of cross-reactivity. Surprisingly, the specificity of the immune system is the major way that the immunology is learnt and studied, notably with tetramers, while it is precisely cross-reactivity that is the basis of the adaptive immune system. Indeed, the thymic selection is based on cross-reactivity. Thymic selection of thymocytes is mediated by recognition of pMHC loaded with self-peptides. The TCR are educated based on the recognition of this set of autologous peptides. During the thymic selection, only a part of the peptide universe is believed to be presented to thymocytes, but, it is sufficient to shape a repertoire that is not too much autoreactive and sufficiently functional to defend the host against infections. This is the first cross-reaction of our TCR repertoire.

It is now well established that the thymic selection is under the control of specific enzymatic complex (see 2.2) that present a restricted peptide repertoire. If this peptide repertoire is restricted, why do we consider that the thymic selection selects a TCR repertoire as diverse as possible? Based on a highly specific mechanism of TCR-pMHC recognition, these affirmations cannot be true. But, under the consideration that the cross-reactivity is the hallmark of TCR-pMHC recognition, it becomes clearer. Even if a diverse TCR repertoire is accepted to be a requirement to cover all the pMHC universe,

the number of theoretically pMHC complexes appears to be far greater than the recognition capacities of any individual TCR repertoire.

Different mechanisms are involved in cross-reactivity. It can be due to the intrinsic structure of the TCR or to its interaction with the pMHC complex. Here is a non-exhaustive list of cross-reactivity causes (Figure 8).

3.2.1. Flexible structure of TCR

Different studies have highlighted structural rearrangement occurring when TCR bind to pMHC. Structural flexibility in the different CDR regions may allow a single TCR to recognize different pMHC. An example of flexible fit is provided by TCR BM3.3, which recognizes three distinct peptides bound to H2Kb MHC[251]. Comparison of the corresponding BM3.3-peptide-H2Kb structures showed that cross-reactivity is achieved through changes in the conformation of the flexible CDR loops region, which allow the TCR to adapt its binding to different pMHC. Most of the time, these variations are observed in the randomly generated CDR3 region. The germline encoded CDR1 and CDR2 loops undergoing relatively minor changes. Interestingly, there are less conformational changes on the beta than on the alpha chain, suggesting that the beta chain is the major driver of antigen recognition[252].

3.2.2. Flexible structure of the pMHC complex

Previous works have shown that the pMHC structure shift could also append with the TCR binding. Indeed, recognition by the same TCR of two similar HLA-A2-restricted peptides: Tax from HLTV-1 and Tel1p from *Saccharomyces cerevisiae* show that a mechanism of conformational flexibility in pMHC allows recognition of different ligands by the same TCR[253], even if cross-reactivity between the two peptides is expected given to their similarities. The interface formed by the TCR with the ligands is different and the conformational differences involve not only the peptide but also the HLA-A2 α2 helix. In this study, cross-reactivity appeared to be based on conformational selection by TCR of one upon different conformations of the pMHC complex that is in dynamic equilibrium with an ensemble of different isomers. This mechanism as also been observed for other molecules interactions, including antigen and antibody recognition[254].

3.2.3. Variable docking angle

It is generally admitted that docking angle of the TCR is diagonally oriented with the peptide backbone orientation in the pMHC complex. But, it seems that TCR can engage different pMHC complex via different docking orientation. In some structure, the TCR axis have been reported to be orthogonal with the peptide axis orientation in the pMHC[255]. The 2C TCR binding to different pMHC is one of the examples of such changes. The 2C TCR employ two very different binding modes to recognize a self-ligand (QL9-H-2Ld) or a foreign ligand (dEV8-H-2Kb)[256] bound on different MHCs. This study demonstrated that a TCR could be oriented in different way to cross-react with different ligands with structural similarity. Here is an explanation of how TCR can cross-react with both self and foreign peptides complexed with the MHC during alloreactivity. In conclusion, TCR docking orientation, even if its overall orientation seems conserved during pMHC binding, present a certain degree of variability. This variability is a possible mechanism to explain the ability if the same TCR to recognize different pMHC complex.

3.2.4. Fuzzy recognition

The cross-reactivity can be the consequence of a lack of specific interaction with pMHC complex. An example of fuzzy recognition is provided by autoimmune 3A6 TCR, which recognizes a self-peptide from MBP presented by HLA-DR2, as well as numerous other peptides that are far more stimulatory than MBP itself[257]. The crystal structure showed interactions between TCR and peptide mainly restricted to *Van der Walls* interactions, with limited juxtaposition of hydrophobic interactions. The low level of interactions between the TCR and MBP offers more structural combinations for optimizing the TCR pMHC interface through variations of the peptide. Using combinatorial libraries, peptides with multiple substitutions at TCR contact positions were identified to stimulate 3A6 TCR bearing T-cells far more efficiently than MBP[258].

3.2.5. Molecular mimicry

Molecular mimicry was the first way proposed to consider cross-reactivity. Indeed, it is easy to think that a TCR can recognize two different peptides if there is sufficient homology between them. Originally, the structural definition of mimicry defined an epitope expressed by a pathogen that shared antigenic structures with host peptides but we can enlarge this definition with a sharing with others pathogens. We will see bellow

that this definition can be enlarge from different point of view as immunological recruitment and epidemiological analysis. In structural terms, molecular mimicry is described as a similarity in charge distribution and in the overall shape of the interaction surface[259]. Molecular mimicry is much more the result of evolution that a specific mechanism of cross-reactivity. Indeed, molecular mimicry can lead to cross-reactivity because of the mechanisms described above, even if some studies showed that the mimicry alone can leads to recognition. In humanized mice, the TCR, which recognizes a self-peptide from MBP-HLA-DR2 complex, also recognizes a naturally processed peptide from Escherichia coli. This cross-reactivity, which induced a multiple sclerosis like disease, is attributable to structural mimicry of a binding hotspot shared by MBP and Escherichia coli. Structural superposition of these TCR-peptide-MHC complexes reveals that the Escherichia coli and MBP peptide main chains are positioned identically and have only minor differences. The peptide-residue side chains that protrude from the groove are very similarly positioned in the two complexes; the P2-His and P3-Phe side chains, common to both peptides, superpose exactly[260]. Other well-documented example is the cross reactivity between EBV and MBP[259].

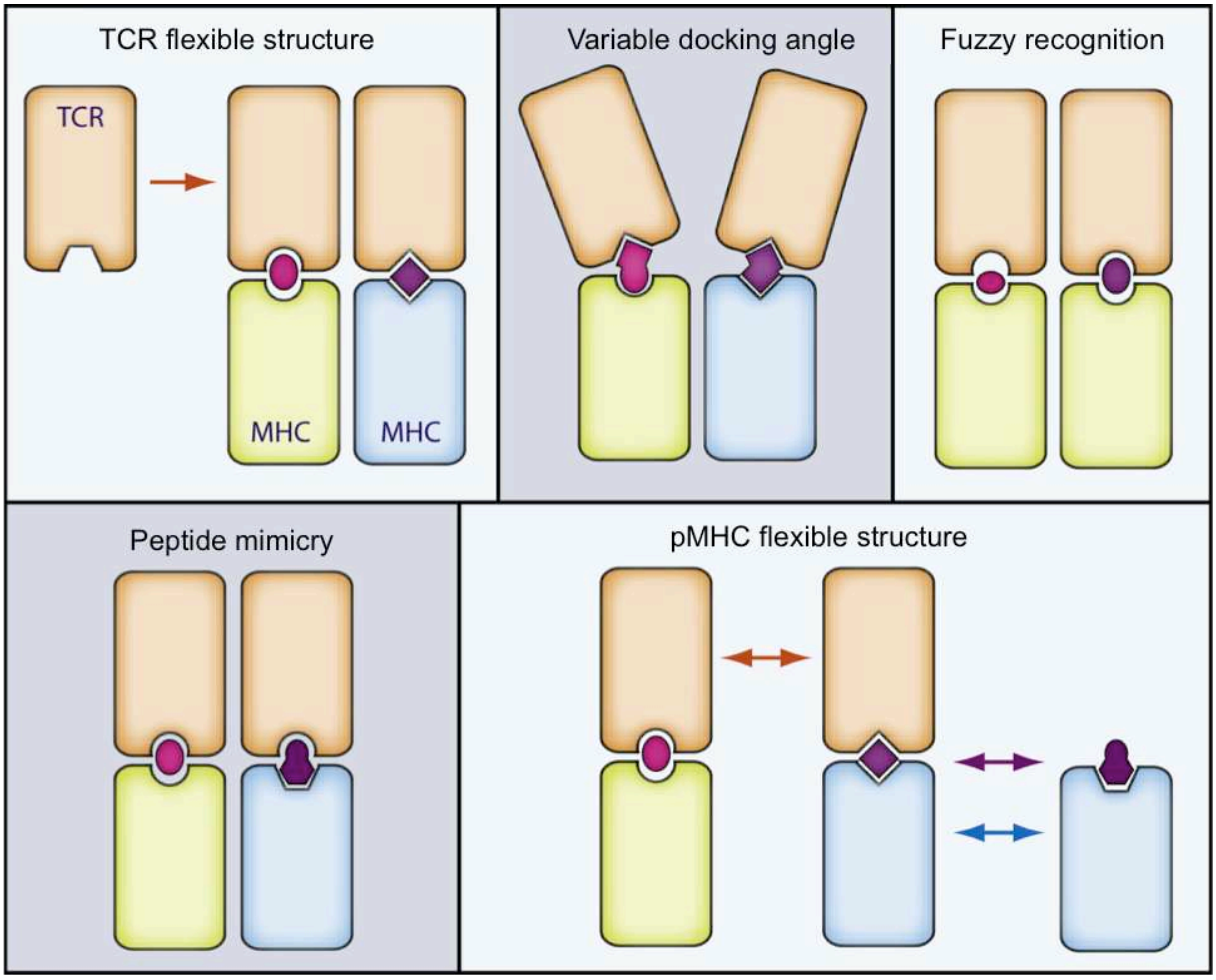


Figure 8. TCR cross-reactivity mechanisms.

TCR is in orange; peptides are red or purple; MHC are green or blue. First row. TCR flexible structure: induced binding site allow TCR to fit with different pMHC complex. Variable docking angle: TCR binds different pMHC using different docking orientations. Fuzzy recognition: structural degeneracy in the peptide can lead to suboptimal but sufficient complementarity between pMHC and TCR, allowed by variations of the kind of binding (i.e. Van der Walls). Second row. Peptide mimicry: different pMHC ligands can form similar interfaces with the cross-reactive TCR if the peptides have close structure. pMHC flexible structure: induced conformation of the pMHC complex allow recognition by TCR. [261].

Cross-reactive repertoire is necessary for the individual, considering the large number of potential pathogenic antigens to which one is exposed over a lifetime. It could compensate a situation with a limited TCR repertoire and still allow a normal immune response. For example, mice deficient for the TdT have impaired CDR3 diversification with a TCR repertoire that is only 5–10% of that calculated for wild type mice[16], but these mice have a relatively normal immune response when infected by LCMV[262]. These TdT-deficient mice are reported to have T-cells with a highly cross-reactive profile when compared to wild type mice. In human, a very restricted repertoire of 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[263].

An essential feature of the cross-reactive T-cell repertoire is that a T-cell can be activated by several foreign peptides, but also by self-antigens. Consequently, the high degree of cross-reactivity among T-cells, which ensures that there is a high frequency of T-cells that respond to any foreign pMHC complex, could potentially lead to autoimmunity. It seems obvious that a high degree of cross-reactivity is an intrinsic and necessary property of antigen recognition by T-cells, and it suggests that the cross-reactive selected repertoire has evolved to have an optimum level of cross-reactivity, which represents the best compromise between the advantage of having a high frequency of T-cells that respond to an antigen and the disadvantage of peripheral autoimmune injuries. One has to keep in mind that the chronic autoimmune diseases that could affect individuals seem less dangerous from a species point of view, than lethal infectious pathogens. For all these reasons, we can be disturbed by the so-called "autoreactive TCR". A TCR is not autoreactive for a simple reason: it is because every TCR are autoreactive, depending of the context of activation via their TCR. Fortunately, TCR activation is not the only one mechanism implicated in the activation of lymphocytes. Interactions with co-receptors, concentration of cytokines and composition of the microenvironment have also their importance. Another way to investigate cross-reactivity is to analyse evolution of the peptide repertoire selected by AIRE complex, which is highly conserved across individuals and could be composed of peptides that are the more prompt to select the highly cross-reactive TCR.

Considering these observations at the molecular level, cross-reactivity will undoubtedly lead to particular and related phenotype in the periphery. In fact, there are numerous papers that are overestimate and that found the explanation of their results in cross-reactivity.

3.3. Cross-reactivity leads to heterologous immunity

The heterologous immunity is defined as an immunity that can develop to one pathogen after exposure to non-identical pathogens. It is easy to think heterologous immunity as a common phenomenon among closely related pathogens, for example, different strains of influenza or DENV, or among different members of the same virus group, because of conserve molecular pattern. However, it also occurs among unrelated pathogens, including parasites, protozoa, bacteria, and viruses[264].

3.3.1. Mechanisms of T-cell dependant heterologous immunity

The mechanism of heterologous immunity is mainly driven by TCR, as it can be conferred, for example, by cross-reactivity between different viruses due notably to mimicry. Alternatively, chronically stimulated T or activated ones that are more reactive to cytokines may be much more sensitive to activation. Under these conditions, low affinity binding to an antigen can trigger activation.

One of the most studied cases of heterologous immunity is heterologous immunity conferred by LCMV (lymphocytic choriomeningitidis virus) against other virus such as Pichnide virus (PV) or Vaccinia virus (VV). LCMV infection in mice elicit a strong T-cells response where LCMV specific memory T-cells could constitute more 20% of the total pool of the CD8 memory cells and the specificity of the repertoire against the different epitope of the virus is relatively stable among genetically identical mice[265].

LCMV and PV are composed of cross-reactive nucleoprotein epitope NP_{205} that have 6 AA in common (LCMV epitope sequence: YTVKYPNL, PV epitope sequence: YTVKFPNM). Upon LCMV infection, staining of splenocytes with LCMV or PV tetramer NP show an induction of the same number of activated CD8 T-cells producing IFN-γ[266]. T-cells activated in response to one of those viruses are also able to recognize the alternative peptide[267]. By looking at the TCR repertoire level, both PV and LCMV induce an expansion of the Vβ16 germline encoded TCR NP specific T-cells. However, the PV induced repertoire is also composed of Vβ5 germline encoded TCR specific T-cells. It indicates the presence of a private repertoire in response to pathogens[266].

When LCMV immunized mice are subsequently infected with PV, they observed a similarly increase in specific NP_{205} cells response but also with a high degree of variability in the TCR repertoire usage. Indeed, 2 mice present a responding repertoire composed of 75% by the Vβ5. The result is consistent with CD8 T-cells that

preferentially expand on PV infection but the Vβ5 was a minor part of the LCMV specific T-cell repertoire.

When PV immunized mice are subsequently infected with LCMV, the responding repertoire to the second virus is composed of a subdominant NP_{205} response due to a proliferation of PV NP_{205}-specific memory CD8 T-cells[267]. Interestingly, analysis of the repertoire revealed differences in TCR usage between mice. Upon the 12 mice tested, (i) 4 had a repertoire like that seen in the previous PV infected mice with Vβ5 (from 30% to 60%), (ii) 6 have a Vβ16 dominant usage with a Vβ5<5%, (iii) 2 other mice had a repertoire with a specific Vβ7 dominancy (89% of the repertoire) and Vβ12 dominancy (69%) even if these Vβ has never been seen as dominant in primary infection. These results suggest that only a part of the selected repertoire during primary infection is recruited during heterologous immunity response.

To investigate the private part of the heterologous immunity, the same author investigated the composition of the repertoire during adoptive transfer. The PV induced repertoire is composed of Vβ5 germline encoded TCR specific T-cells, whereas the LCMV induced repertoire is mostly composed of Vβ16 germline encoded TCR specific T-cells. Experiment of splenocytes infusion from immune mice in recipient then infected with the cross-reactive virus highlight bias in the activation of the NP_{205} specific repertoire. Even if recipients of a single donor generated similar T-cell repertoire following infection, recipients from different donors generated different set of responding TCR repertoire indicating that a part of the cross-reactive expansion was depending of each individual. The general dynamic of the TCR repertoire for this example of heterologous immunity is summarized in Figure 9:

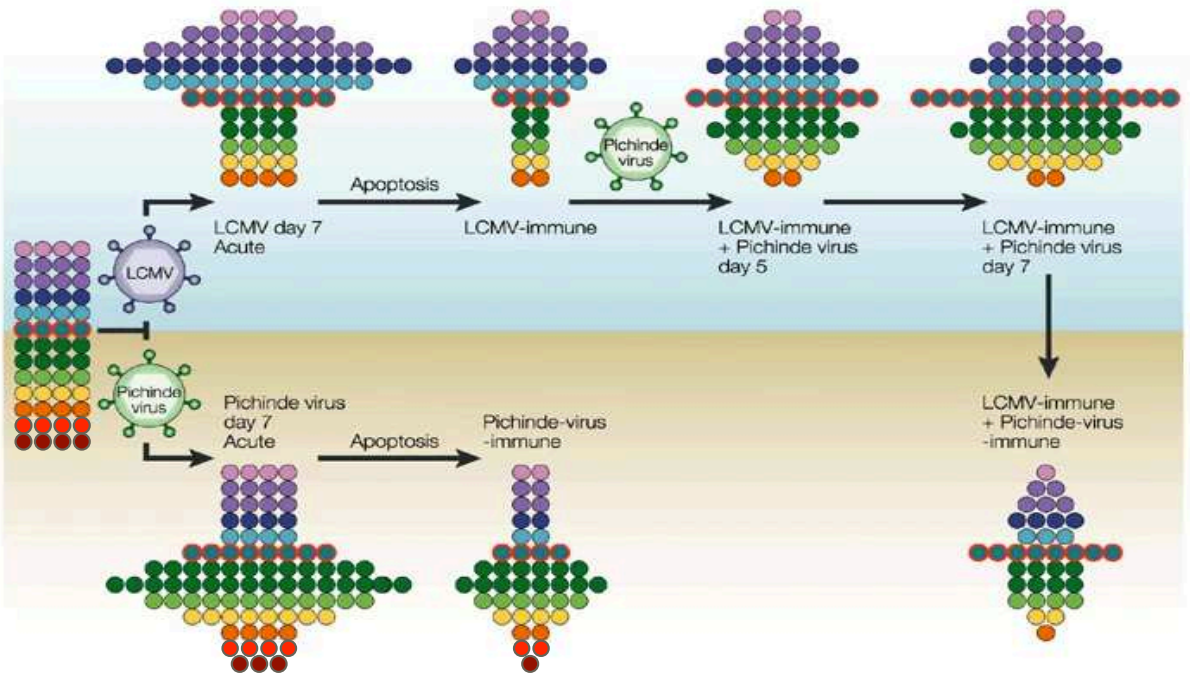


Figure 9. Dynamic of TCR repertoire during heterologous immunity.

The coloured dots represent different T-cell clonotypes that have different specificities. On the left, a naive immune repertoire is challenged with either of two heterologous viruses LCMV or PV. Clonotypes stimulated by viral antigen expand and then undergo apoptosis, and some of them form a memory T-cell pool. The immune system conditioned by the first viral infection (LCMV) is exposed to another virus (PV), T-cell cross-reactive clonotypes (red outline) will expand preferentially and dominate the response. After the response, memory cross-reactive T-cells are preserved and enriched in the resting memory pool(from [268]).

In the other widely studied example of heterologous immunity conferred by LCMV against VV, the immune response is driven by specific CD8 through IFN-γ secretion[264]. In this respiratory mucosal model of infection with VV, memory TLs specific for LCMV were recruited in the lung[269]. Adoptive transfer of LCMV specific splenocytes from immunized mice protect against VV infection, and this protection is reduced if the CD4 or CD8 are depleted in the donor splenocytes[270]. As LCMV immunized mice treated with anti-IFNγ are not protected against VV, the mechanism seems to be IFN-γ dependant.

In this example of LCMV and VV, the cross-reactivity of TCR is involved in heterologous immunity. TCR specific for the LCMV epitopes NP_{205}, GP_{34}, and GP_{118} proliferate after VV challenge. TLs specific for each of LCMV epitopes present cross-reactivity for the VV $A11R_{198}$ epitope (AIVNYANL). The A11R-specific TLs from LCMV immunized mice can bind to both LCMV GP_{34}, and GP_{118} tetramers[271]. Several VV antigens sequences harbour

partial homology with NP_{205}: VV A11R antigen has 4/8 AA in common with NP_{205} and 3/8 AA with GP_{34}, and GP_{118}.

In this second example, mechanism leading to heterologous immunity has also a private part depending of each individual. Indeed, there are variations in the specificity of CD8 memory cells responding to VV in LCMV immunized mice. The study of A11R specific CD8 cells before and after heterologous virus infection shows that low levels of these cells are present in LCMV immune mice (0,2-0,5%). At 12 days post VV infections, only 6/19 mice showed significant increase of T-cells specific for the cross-reactive VV epitope. Upon the nineteen, four mice respond to the GP_{34-41} epitope, three to the NP_{205} and one to the GP_{118}. Adoptive transfer also demonstrated the private specificity of heterologous immunity. Some mice preferentially use cross-reactive responses against the A11R whereas other not, and sometimes cross-reactivity is not seen against any of those epitopes, thereby demonstrating the complexity of heterologous immunity[222].

Even if the heterologous immunity is based on cross-reactivity, it is important to note that there is not necessarily reciprocity. One of the possible explanations to this phenomenon for the example above is that VV is a virus that encodes more than two hundred proteins and a number of T-cell epitopes evaluated to more than thousand with the capacity to trigger memory T-cells. In contrast, LCMV encodes only four proteins. The explanation depending on this number of viral proteins is that the thousand epitopes encoded by VV could more likely stimulate TLs in an LCMV-specific memory subset than would the much more limited number of LCMV epitopes stimulating a VV specific cells[272].

In human, same kind of mechanistic has been reported for heterologous immunity. The mostly studied example is the TCR repertoire during EBV infection. Indeed, an increased level of cross-reactive CTL was observed during this infection[273,274]. This result was first attributed to non-specific bystander but limiting-dilution clonal assays have shown that this multi-specific activity was due to T-cell clones that are cross-reactive[274]. It has been previously related that cross-reactive T-cells exist with specificity for two immunodominant HLA-A2 presented epitopes, Influenza matrix epitope M_{158} (GILGFVFTL) and EBV $BMLF_{1280}$ (GLCTLVAML), which have relatively little sequence similarity. CD8 from patients previously immunized against Flu and with acute EBV

infection were cultured and pulsed with Flu-M_{158} or $BMLF_{1280}$ peptides. CD8 cells recognized both peptides with the same avidity even if there is few AA in common[275].
Another study looked for whether unrelated virus specific memory CD8 T-cells were activated in the immune response to acute infection with heterologous pathogens. Authors investigated antigen-specific CD8 CTL in PBMC from patients at the onset of acute HBV infection. They were detectable from: 43 to 89% for HBV and for 5.5 to 20% for CMV, in CD8 T-cell compartment as determined by pentamer binding. These results suggest that CD8 T-cells specific for CMV may be activated during acute HBV infections. They also found that IL-15, a cytokine important for maintenance of memory T-cells and often produced during acute viral infection, selectively activates CMV specific CTL cells, for spontaneous IFN-γ production, and enhances anti-viral mechanisms[276].

There are few papers on the mechanism of heterologous immunity in human. Most of the studies focused on the results of heterologous autoimmunity. Indeed, it has been suspected for a while that cross-reactivity could lead to the apparition of autoimmune diseases but there are fewer papers on the benefit of cross-reactivity leading to positive heterologous immunity. In the following section, I described a non-exhaustive list of heterologous phenotypes in mice and human.

3.3.2. Protective heterologous immunity

Historically Jenner made one of the most important discoveries in the field of immunology. He noted that milkmaids were generally immunized against smallpox. Jenner postulated that the pus in the blisters that milkmaids received from cowpox protected them from smallpox, which is a similar disease. It is commonly accepted that Jenner was one of the fathers of vaccination. However, what Jenner discovered in this day of 4 May 1796, is more than vaccination. It was the protective action of heterologous immunity[277].
It is the first example suggesting that vaccination may have not only disease-specific effects but also effects on other diseases. Moreover, since Vaccinia was introduced, it was noticed that recipients are less susceptible to diverse infections such as atopic diseases, measles, scarlet fever, and syphilis[278].
Several examples of protective heterologous immunity have been described in mice specially with LCMV that provides protection with mouse cytomegalovirus (MCMV) or

VV. Infection with LCMV, PV or MCMV conferred a considerable level of protection against infection with VV, as shown by reductions in viral load and increased survival in response to lethal doses of VV in systemic and respiratory-mucosal models of infection. Protected mice had a 2 to 200 fold reduced viral load 3 days after infection compared with the challenge of seronegative mice, and the heterologous protection continued for as long as a year[270]. Against, the absence of reciprocal protection has to be noted.

Measles virus

In human, it is noteworthy that in developing countries, live measles virus vaccine, seems to protect against mortality that is not attributed to this infection[279]. The studies on standard measles vaccine suggest that this vaccine has beneficial effects beyond protection against measles. Its implementation significantly improved child survival in Africa and may have enhanced resistance to unrelated infections. Comparative studies showed that two doses of measles vaccine at 4.5 and 9 months of age reduced the mortality to infections other than measles by 30% in children as compared to one dose of measles vaccine at 9 months[280,281].

Cytomegalovirus

CMV is a beta herpes common virus that infects most of the population worldwide, with seroprevalence increasing up to 90% with age[282]. It is a classic example of virus establishing persistent infections. In most of the cases, CMV infections are subclinical and well tolerated, but they cause a significant perturbation of T-cell repertoire as 20% of the lymphocytes can be specific for this pathogens in infected subjects[283,284]. Why such a large population of CMV-specific memory CTL cells would be maintained over time? Only for CMV protection? Considering heterologous immunity, CMV-specific memory CTL presence might contribute to the immunological response against other pathogens. There are several aspects of correlation between immunity against CMV and others pathogens. In a recent paper published by the team of Mark Davis, the positive contribution of CMV in the serological response to influenza vaccination has been observed in human. Indeed, subjects that have been previously infected by CMV have a different immune compared with negative subjects for influenza vaccine responses. Young CMV+ subjects have higher levels in TH1 and TH2 cytokines level as well as a stronger CD8+ responses to IL-6 compared to CMV− subjects. They exhibited elevated

serum levels of IFN-γ, and elevated antibody responses to the influenza vaccine. The comparisons of the antibodies response against influenza are represented in Figure 10:

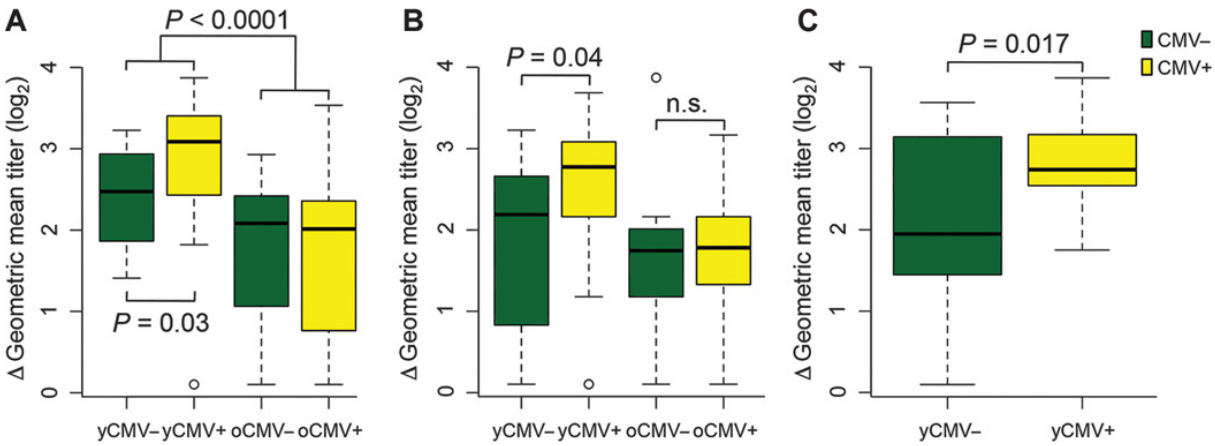


Figure 10. Protective heterologous immunity between CMV and Influenza in human.

Green bars, CMV−; yellow bars, CMV+. To estimate vaccine response in young and older CMV− and CMV+, the Δ geometric mean titer was calculated (ΔGMT). To calculate ΔGMT, post-vaccination GMT was subtracted from pre-vaccination GMT. The GMT is calculated for all three strains in the vaccine, for each individual. The GMT is based on the standardized HAI assay[285]. A higher response is observed in young CMV seropositive individuals (yCMV+) compared to seronegative (yCMV−) in the first (A) and second (B) year following vaccination, as well as in an independent validation study conducted during the 2010 to 2011 influenza season (C). No significant differences (n.s.) were observed between old individuals (oCMV− and oCMV+).

These data indicate that CMV could be a major driver of natural immune response against influenza.

Influenza A virus

The Influenza virus presents many viral strains, which impose annual vaccination against the pathogen. An unresolved question is how these cross-reactive T-cells are working in immunity against IAV infections[286]. Epidemiological study have indicated that exposure to one strain of IAV (H1N1) apparently provided some level of protection against another strain of IAV (H2N2)[287]. Once could expect these results as different strains of Influenza are in fact molecularly very close. On the other hand, once may wonder why is it necessary to vaccine our self each year against Influenza? A clue for these responses is described below.

A more unexpected example of heterologous immunity, as it between unrelated human viruses, occurs between IAV and HCV. Indeed, T-cell specific for an immunodominant HLA-A2-restricted T-cell epitope that is encoded by HCV, cross-react with an Influenza epitope that have 7/9 AA in common (HCV $NS3_{1073}$: CVNGVCWTV, Influenza NA_{231}: CVNGSCFTV). This result defines a strong cross-reactivity between hepatitis C virus and influenza virus dominant epitopes[288]. It can be a clue to explain why some patient clear HCV, but others develop persistent infections[289].

Hepatitis G virus

The same kind of phenomenon could be implied in a relative protection against HIV. Several HIV-exposed individuals persistently remain seronegative. They harbour HIV-specific CTL responses to epitopes that are different from those that are recognized by HIV seropositive individuals[290,291]. The heterologous immunity could be involve as recent observation that seropositive patients that are co-infected with the Hepatitis G virus are less susceptible to progress to AIDS[292,293]. In these two studies published in the *New England Journal of Medicine,* a significant link was found between the HGV chronic infection and the control of HIV infection. Among the HIV seropositive patients, the mortality rate was significantly decreased among those with HGV. These results are summarizing in Figure 11:

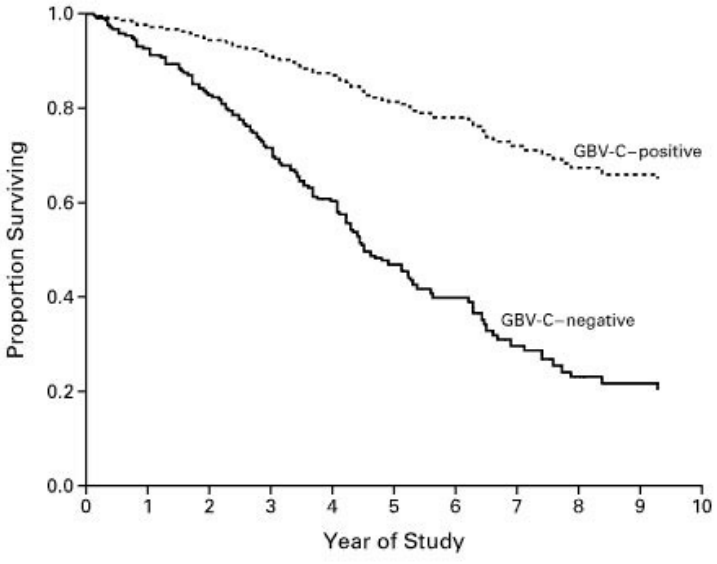


Figure 11. Survival curves for co-infected HIV and HGV patients.

HIV seropositive patients with HGV coinfection survived significantly longer than HIV seropositive without HGV coinfection (P<0.001)(From [293]).

This result was independent of classical factors as, base-line CD4+ cell count, age, race, sex, or mode of HIV transmission. These results have to be moderate if we are involving the heterologous immunity. In fact, they do not give the proof that the control of HIV infection is due to immune system. HGV could interfere with some basic and necessary metabolic pathway that is necessary HIV replication[294].

Mycobacterium tuberculosis and Bacillus Calmette-Guérin

Heterologous immunity is not restricted to virus. BCG immunization seems also to provide positive heterologous immunity effects. Trials in neonates show that BCG vaccination at birth reduces almost half neonatal mortality. Infants have lower rates of neonatal sepsis and respiratory infection[295]. Another epidemiological analysis performed on over 150,000 children under five years old performed in more than thirty countries suggested that BCG vaccination reduce acute lower respiratory infection incidence by 17 to 37%[296].

More specifically, BCG vaccine has been noted to provide protection against *Mycobacterium leprae*. This effect was first demonstrated in 1936 by a positive Mitsuda reaction, a marker for improved cell-mediated immunity against leprosy was observed after BCG vaccination. In a recent meta-analysis, author included nineteen observational and seven experimental analyses. The results of the different observations indicate that the protection conferred by the BCG vaccination increase the protection against *Mycobacterium leprae* of more than 25% but with significant heterogeneity between the different experiments. The protective effect conferred by vaccination estimated by the observational studies was more than 60% also with significant heterogeneity between the different experiments[297].

There is a much more surprising heterologous immunity that have been notice thanks to *M. tuberculosis*. Indeed, it has been reported that BCG induced could induce protection and T-cells immunity against poxvirus[264]. Moreover, it has been observed that BCG vaccinated HIV seropositive adults had lower risk of intestinal nematode infection than unvaccinated patients[298].

The positive effects of heterologous immunity have been reported for several times but unfortunately, monitoring of patients is often done when they are ill. The protective heterologous immunity seems to be a widespread effect especially when you keep in mind that the vast majority of the population was in their childhood, infected by CMV or

EBV. All together, these observations suggest that previous encountered pathogens, and moreover every immune response that lead to a memory repertoire, play a role in shaping our response to subsequent immune challenge. This is not taken in account for most of the clinical studies while it could give a clue to provide insight into the variance between vaccination efficacies in the laboratory versus in the clinic. It can also help to prevent some injuries link to heterologous immunity. Indeed, different papers indicate that heterologous immunity can result in autoimmunity, altered immunopathology and/or changes in the TH1/TH2 balance (immune deviation) that can sometimes lead lethal side effects.

3.3.3. Heterologous autoimmunity and mimicry

As autoimmunity aetiology was for a while suspected to have infectious origin and to be associated with viral infections, once have to question the implication of the T-cells that are specific for pathogens and present unconventional characteristics. Indeed, it has been proposed that viral determinants that mimic host antigens could directly stimulate self-antigen specific T-cells. We have described above the mimicry as a molecular similarity between 2 different antigens. On an autoimmune point of view, this definition can be enlarged to described pathologies. First of all, mimicry can be defined as the detection of TL or antibodies that both react with same antigens. Second, based on epidemiological point of view, mimicry could be described as a link between infection to pathological agent or microbe and development of AD. Third, mimicry can define the autoimmune phenotype obtain in animals model following sensitization with specific epitope either infection with the pathogens[299]. Here, we review some examples of direct cross-reactivity that could be involved in autoimmunity due to unconventional TL.

Multiple sclerosis

Multiple sclerosis (MS) is an inflammatory and autoimmune disease of the central nervous system. It is characterized by motor and sensitive disturbance associated with cognitive impairments[300]. MS has been considered to be TL associated disease. First, the myelin basic protein (MBP), the myelin oligodendrocyte glycoprotein (MOG) and the proteolipid protein, three host antigens, are the main targets of autoreactive CD4 T-cells[301]. Second, the inflammation and the tissue destruction are driven principally by CD8 T-cells[124].

Considering the mimicry as a driver of the pathogenesis, several pathogens such as varicella zoster virus[302] or Chlamydia pneumonia[303] have been suspected to be implicated in MS. However, EBV cumulates incriminating evidences linked to mimicry for a possible driver role in this disease. First, in a recent epidemiological analysis, EBV seropositive, HLA-II predisposing allele (HLA-DRB1*15:01), or both were strongly associated with the development of MS. Second, in patients with MS, TL with specificity for the MBP are also able to react with the EBV nuclear antigen 1[304,305]. Third, molecular homology between MBP and EBV peptides restricted to DRB1*15:01 has been reported[259]. This molecular mimicry is detailed in Figure 12:

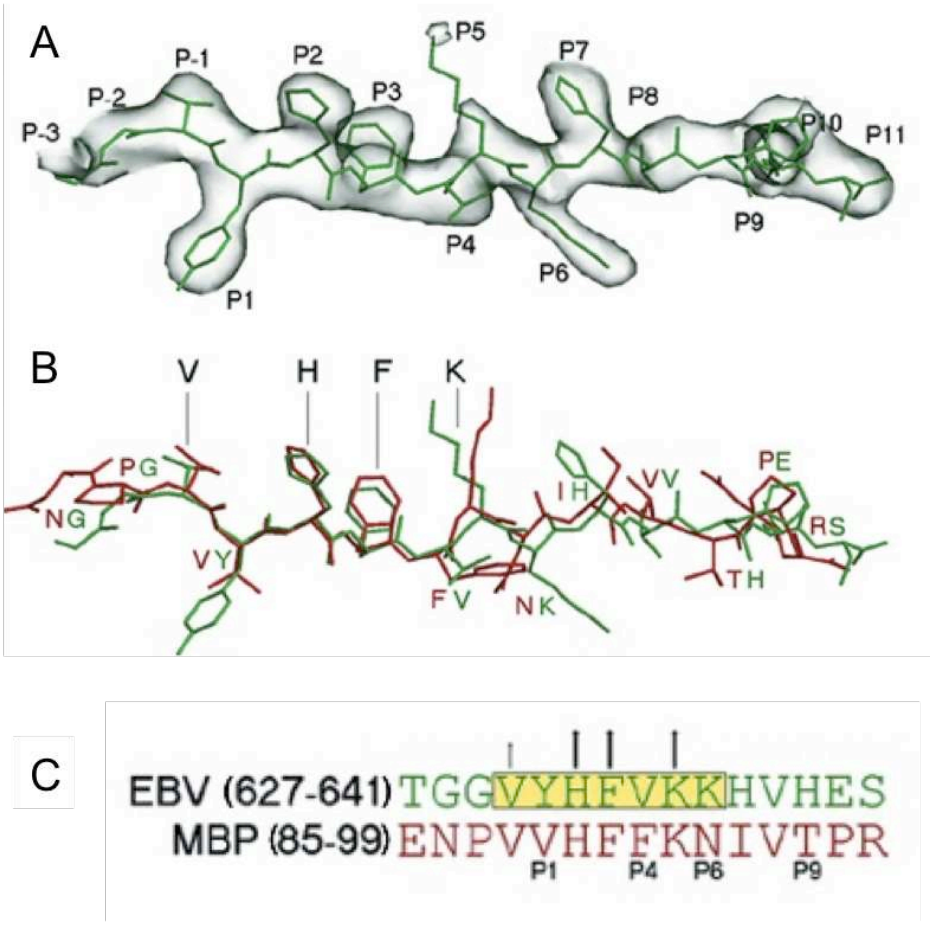


Figure 12. Molecular mimicry between EBV and MBP.

A. The EBV HLA-DRB1*15 restricted peptide is drawn in green. B. Superposition of EBV (green) and MBP (red) peptides based on HLA-II structures. Residues critical for specific clones TCR recognition are labelled in black. C. Alignment of EBV and MBP sequences. Anchors residues required for specific TL stimulation are outlined in yellow, and vertical bars indicate importance for T-cell recognition. (From [259]).

Considering these results, it is reasonable to think that the TCR of the autoreactive TL could share some affinity for EBV peptides and MBP. However, the affinity of these autoreactive TCR is lower for the EBV peptides than for the MBP[306]. One of the hypotheses is that the generation of the immune memory after initial activation in response to EBV, involving autoreactive T-cells with low affinity TCR, leads to the generation of higher relative affinity TL clones that will damage the CNS[307].

Type 1 narcolepsy

Narcolepsy is a rare sleep disorder caused by hypocretin neuronal loss. It is characterized by an excessive daytime sleepiness accompanied by impaired nocturnal sleep. The autoimmune cause was suspected due to strong association with the presence of HLA-II alleles as HLA-II DQB1*06:02 [308–310] and epidemiological studies [311]. Indeed, this epidemiological study was the first to report that H1N1 vaccination has the potential for the development of narcolepsy. They proved a significant increase in narcolepsy cases after vaccination in the Chinese population.

Molecular mimicry has been identified between Influenza NP and the extracellular domain of the human hypocretin receptor 2, which is considered as a target implied in the development of the pathology[312]. Moreover, antibodies from Influenza vaccinated patients were able to cross-react with these two proteins. In addition, the specific implication of TCR was highlight by the involvement of the TCR alpha[313]. A recent studies have definitely proven the implication of the immune system in the physiopathology of the disease as repertoire composed of specific TCR with cross reactivity for hypocretin and influenza A were found in patient [314]. Thus, mimicry appears as a major factor in the development of post-infectious narcolepsy.

Type 1 diabetes

It is largely accepted that T1D is a polygenic disease but environmental factors like viral infections have also been shown to play a decisive role in the appearance of the pathology[315,316]. Among the virus that are suspected to be associated with the appearance of T1D are those that belong to enterovirus. Epidemiological investigations have reported an increase incidence of insulin dependent diabetes after Coxsackie virus epidemics. The infection was 10 times more common in T1D compared with controls[317,318].

Suspicion of mimicry in T1D pathology also comes from the finding of numerous similarities between epitopes of pancreatic β cells and viral components. Cross-reactivity has been noted between the VP-1 protein of enterovirus and the tyrosine phosphatase of beta cells and enterovirus infection can induce cross-reactive immune responses[319]. Second, the presence of the HLA-DR3 is known to be a factor of predisposition for the development of the disease as it is more prompt to complex autoantigen like the glutamic acid decarboxylase. HLA-DR3 restricted CD4 T-cell has been found to cross-react against epitopes with sequences similarity between this autoantigen and a CMV-encoded antigen[320]. Third example, it was found that molecular mimicry with rotavirus could promote autoimmunity to IA2 islet antigens[321].

Interestingly, the inoculation of NOD with gammaherpes virus delayed the onset of T1D[322]. Moreover, the inoculation of NOD with Coxsackie virus was followed by of long-term protection from T1D[323]. These results feed the debate about implication of viral infection in T1D and further studies are needed to clarify the role of heterologous immunity in the physiopathology of T1D.

Guillain-Barré syndrome

Guillain-Barré syndrome (GBS) is an immunologically mediated polyneuropathies characterized by an acute inflammatory polyradiculoneuropathy, progressive weakness, autonomic dysfunction and pain. The role of TL in GBS has been extensively investigated as a central mechanism for autoimmunity because of nerve infiltration by TL, parallel to demyelination[324]. This syndrome is considered as a post-infectious autoimmune disease[325]. Thus, several infections such as viral[326] and bacterial[327] have been reported and suspected to be linked with this syndrome.

Considering the different phenotypes of the disease, the acute motor axonal neuropathy (AMAN) that targets the axon membranes, is the only one clearly associated BGS and with *C. jejuni* mimicry[328]. The first epidemiological evidence of *C. jejuni* in the development of GBS was a case-control study reported that a recent *C.jejuni* infection was more common in patients with GBS[329]. On a immunological level, patients with *C.jejuni* infections present significant titer of IgG that cross-react with GD1 and GM1 gangliosides, two key targets of the disease[330,331]. In addition to these two levels of mimicry, molecular mimicry between the bacterial LPS and the GM1 has been identified[332]. Finally, the disease has been reproduced also in animal model.

Experimental sensitization of rabbits with the bacterial LPS leads to the increase of anti-GM1 IgG and the development of symptoms as limb weakness, resembling to GBS[333,334]. Many results orient aetiology of GBS to infection and the mimicry seems to be involved when *C.jejuni* is implicated is the apparition of the disease. It is one of the many ways that the disease seems to begin and others investigations have to level how *C.jejuni* infection is an important cause in GBS aetiology.

Systemic lupus erythematosus

Systemic lupus erythematosus (SLE) is an autoimmune disease with multiple molecular and heterologous clinical phenotypes. SLE may involve all organs but the most common clinical features are mucocutaneous lesions, arthritis, renal involvement, and haematological disorders. The serology of SLE is characterised by the positivity of many autoantibodies among which the most specific are anti-dsDNA and anti-Sm. B cell dysregulation is usually linked to the development of SLE as they mediates the production of autoantibodies that are fundamental factors in SLE[335].

Considering the mimicry as a driver of the SLE, different pathogens such as parvovirus, CMV or HCV have been suspected to be implicated but one more time, EBV cumulates incriminating evidences linked to the mimicry with this disease[335,336]. Studies have highlighted a higher incidence of EBV infection among SLE patients as well as higher titers of antibodies against EBV[337,338].

The immune response against EBV for SLE patients is also different from healthy individuals. Indeed, the humoral response generates cross-reactive antibodies in genetically susceptible individuals[336]. Different candidate antigens have emerged. For example, the molecular mimicry of PPPGRRP EBV antigen with the PPPGMRPP peptide from Sm of the human spliceosome is consistent with the possibility that EBV infection is implicated in SLE[339].

From these data and observations, we can suspect evidence for implication of viral mimicry in human autoimmunity in some cases. The identification of sequence motifs feed the hypothesis that heterologous autoimmunity could contribute to pathogenesis of autoimmune diseases. However, much of the time the correlation is done without any causality proof and more functional evidences and experiments have to be provided. These cases of pathogenic heterologous immunity are relatively simple to understand,

as mimicry is easy to predict and observe. But, in some case, the problem linked with heterologous immunity are much more insidious.

3.3.4. The original antigenic sin

In the late 1950s, Thomas Francis Jr notices that the antibodies generated during second infection by a second serotype of Influenza had a higher specificity for the first serotype. He used a theological analogy to name this phenomenon: the Original Antigenic Sin (OAS)[340]. Just few years after, the phenomenon was also described and named "the Hoskins effect". It has been suspected to be involved in weak immune responses against Influenza. From an analysis of a vaccination campaign in adolescent pupils, in the 70's, it was concluded that annually repeated vaccinations would not confer protection against epidemic influenza in the long-term[341,342]. Finally, a few years ago, an epidemiologic study has implicated the OAS in the explanation of the variation of immunity level against different serotypes of Influenza. Indeed, key observations about the human immune response against repeated exposure to influenza A is that the first strain infecting an individual apparently produces the strongest specific response. Using epidemiological data, the authors listed the capacity of each birth cohort to respond to different strain of influenza. Indeed, individuals born before 1968 had their first infection with a group 1 virus are protected against virus of the same group. Conversely, after 1968, individuals had their first infection with a group 2 virus are protected against virus of the group 2. The author highlighted that susceptibility profiles of infection mirror the age of reported cases, supporting the life-long effect of an original imprinting of the immune system[343]. These results are summarized in Figure 13:

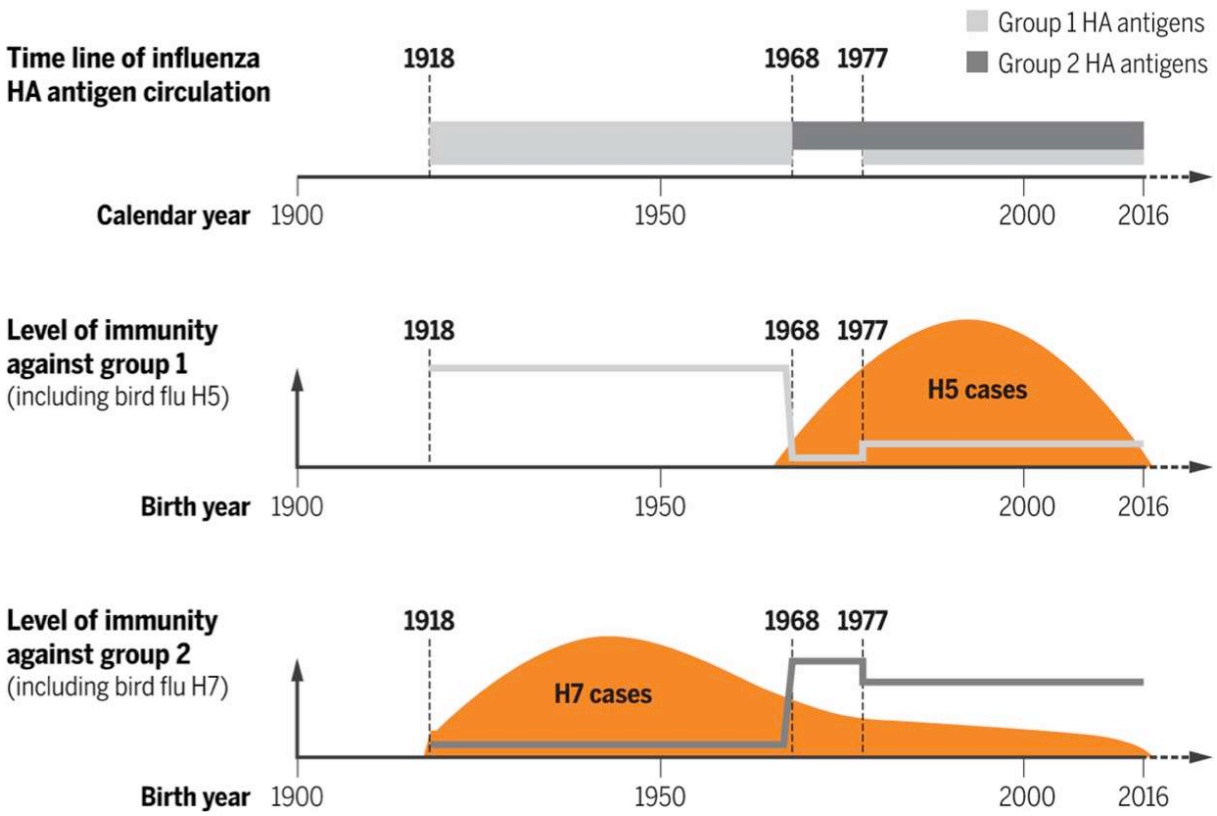


Figure 13. First flu is forever.

A major change in the circulation of influenza strain append in 1968 with a variation in the immunodominant antigen hemagglutinin. This modified the strain of influenza that new birth individuals first encountered in life. As a consequence, the protective immunity level differs by birth year and is reduced against the discordant strain. [344].

The mechanism has been described as selective activation of CTL specific for previously encountered epitopes of a virus, following infection with a heterologous strain presenting mutant variants of those epitopes that would otherwise invoke a different T-cell repertoire[345]. Due to prior exposures to the first antigen, naive lymphocytes do not respond to the variant antigen as for a primary contact but instead memory lymphocytes, specific of the first antigen, are recruited. They interpret the second antigen as the original one and proceeds with a memory response to the variant antigen. At first glance, this seems like a favourable phenomenon like the immune memory. However, it can be problematic when the variant antigen is sufficiently different from the original, which leads to an ineffective clearance of the mutated pathogen. In a more extreme example, it leads to a complete evasion of the mutated pathogen from the immune system, a situation that clearly could have deadly implications. Finally, the inefficacy of the immune system to clear the virus can also lead to an uncontrolled

activation of the immune system and immune related pathologies. The theoretical mechanism of original antigenic sin is summarized in Figure 14:

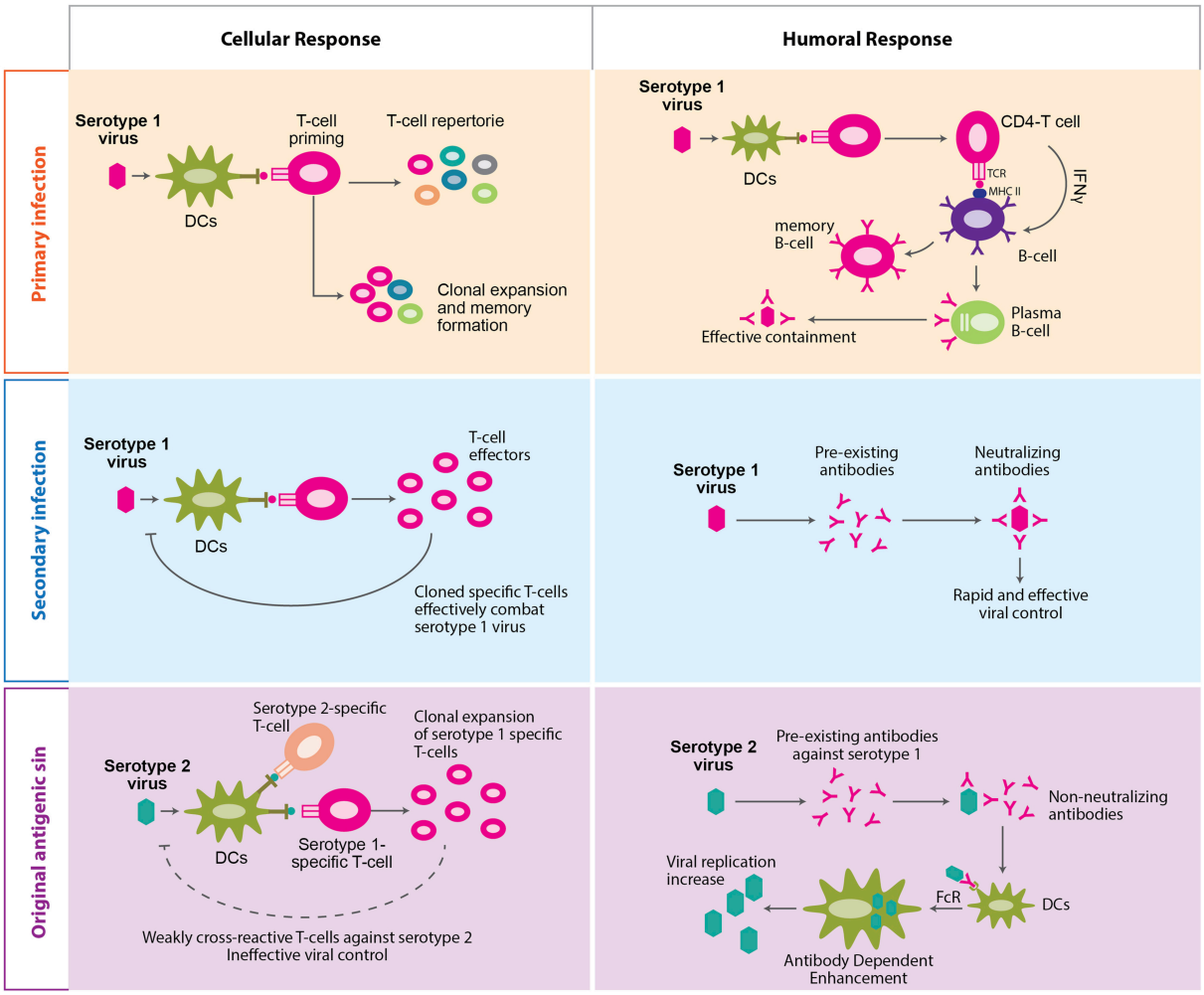


Figure 14. Theoretical mechanism of the original antigenic sin.

During primary infection, serotype 1 viral antigen is processed and presented by the APC, leading to the activation and expansion of specific B and T-cells. The pathogen is cleared and a memory pool specific of the virus serotype 1 is generated. Following the second infection by serotype 2 with variant antigens, the memory pool developed during the primary infection is recruited, because of cross-reactivity. It renders an ineffective response against serotype 2 and a possible immune escape and/or immunopathology(from [340]).

The original antigenic sin seems also to be implicated in the difficulties to design efficient vaccines due to the mutations of viruses, inducing sub-optimal memory response[346–348]. Moreover, in HIV, the high rate of mutations implicated in viral escape could be due to original antigenic sin[345,349]. There are different consequences of the original antigenic sin that have been reported in mice and human. Here is a non-exhaustive list of the phenomena.

Narrow repertoire

The proportion of memory CTL increase with age, but homeostasis controlled their number, as there is an upper limit in the total number of memory cells. Either additional memory T-cells, which could emerge from following infections, are precluded. Either, there must be a deletion of pre-existing memory populations when the host responds to a new pathogen. So CD8 T-cell memory can be compromised by subsequent viral or bacterial infections.

LCMV and PV encode epitopes that induce different but highly cross-reactive diverse TCR repertoires. Homologous viral challenge of immune mice only slightly skewed the repertoire and enriched for predictable TCR motifs. However, striking differences in TCR repertoire evolution were noted when mice received a homologous versus heterologous viral challenge. Heterologous challenge in either virus sequence led to profound narrowing of the repertoire.[266]

The same mechanism is observed in Influenza immune response. T-cells repertoire can cross-react with different Influenza epitope but when viral strains present mutations on these antigens, a more narrow T-cells repertoire emerge[350]. Interestingly, all influenza A viruses that have infected humans in the past 80 years have expressed the conserved HLA-A2 restricted M_{158} immunodominant epitope[351]. During the different infection of influenza during life, the diverse repertoire of the first infection, composed of different Vβ and TCR, derive to a narrow Vβ 17 repertoire with an AA motif in the CDR3 region of the TCR[352]. This immunodominant response does not seem to be adequate and may be a cause of viral escape, as individuals remain susceptible to infection with cross-reactive influenza virus strains[353].

Altered immunity and Immunopathology

The original antigenic sin can also lead to inflammatory deleterious reactions. It gives an explanation that some diseases, like chicken pox, measles or polio are more severe in adults than in children[354,355]. Indeed, heterologous viruses have the potential to over stimulate memory T-cells that are cross-specific for other viruses because their replication would not be prevented by highly specific T-cell action and neutralizing antibodies. These cells are in fact much more specific against the original virus, and are not able to clear the variant strain or the heterologous one. So, the phenomenon can lead to an uncontrolled stimulation of T-cell, inflammation and immune related diseases.

The majority of human being acquires a persistent infection with EBV, whose first contact with the virus is during childhood and is essentially asymptomatic. However, young adults acquire EBV for the first time display symptoms of acute infectious mononucleosis (AIM) [356]. The AIM is characterized by massive expansion of CD8 CTL and symptom can be influenza like illness to severe syndrome with splenomegaly, hepatic injuries and extreme tired[357]. The difference of severity seems to be related to the amplitude of the CD8+ T-cell response rather than the virus load[358]. The OAS could be an explanation of this phenomenon as the activation of cross-reactive memory TLs, which have been generated in response to a previously encountered virus, can be over-activated by cross-reactivity against EBV, but are not sufficiently specific to clear it easily and quickly. One observation supports this hypothesis, as T-cells that are cross-reactive between the influenza M_{158} epitope and the main immunodominant $BMLF_{1280}$ epitope of EBV are present at a high frequency during infectious mononucleosis[275] and the severity of the AIV is correlated with the presence of cross-reactive M1-BMLF-specific TCR repertoire[359].

In human, the original antigenic sin seems to be implicated in the Dengue Haemorrhagic Fever (DHF) or Dengue Shock Syndrome (DSS). It is a severe complication of dengue disease associated with a high rate of mortality. DSS patients die of progressive worsening shock and multi-organ failure. The classical mechanism of shock is linked to increased vascular permeability due to dysfunction of vascular endothelial cells induced by cytokines[360].

Two theories have been proposed to explain the physiopathology of DSS[361]. The first one stipulates that DSS is caused by more virulent strains of the dengue virus. The second one suggests that DSS results from abnormal and exaggerated host immune responses. This second mechanism seems to be due to sequential infections with different strains of dengue virus as epidemiological studies indicating that it occurs when a dengue seropositive individual is secondarily infected with a DENV of a different strain[362,363]. During the second infection, the virus-specific CD8+ T-cell response is dominated by the proliferation of cross-reactive memory T-cells generated during the primary response. These memory T-cells have a lower activation threshold, so they respond faster and are more present compared to naïve precursors, even if they have a lower avidity for the secondary-infecting virus and a worse antiviral function. Consequently, these cross-

reactive CTL will not be able to efficiently clear DENV while the cytokines they over-produced may contribute to increased vascular permeability involved in DSS.

DENV is composed of 2 surface glycoproteins (envelope (E); pre-membrane (M)) that allow fixation on the host cell, the capsid protein (C), and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) that regulate viral replication. NS3, NS4B and NS5 are the most frequently recognized proteins[364] but the level of immunodominance between proteins depends of the DENV strain. DENV 3 gives rise to similar proportions of CD8+ T-cells specific for the structural and non-structural proteins while DENV 1, 2 and 4 elicit a CD8+ T-cell responses are much more specific for the non-structural NS3, NS4B and NS5[365].

For this reason, the sequence of the DENV infection influence disease severity, with DENV 3 followed by DENV 2 appearing to be the worst combinations of subsequent infection[366]. A detailed study focusing on the specificity of T-cells responses in dengue-infected Thai children presenting a DSS concluded that CTL show a low affinity for the vaccine strain and a higher affinity for another strain, presumably from a previous infection[367]. During the second infection, CD8 T-cells specific for the NS3 antigens are increased in DSS patients and there is an association between the level of these T-cell response and the severity of the disease [368]. These CD8 cells display different cytokines production profile upon stimulation with the heterologous challenge, alike CD4 that produce higher TNF-α upon heterologous stimulation, which is a major driver of vascular shock[369,370]. These altered functions of the T-cell repertoire response could result from the original antigenic sin. The scenario of DSS due to the original antigenic sin is depicted in Figure 15:

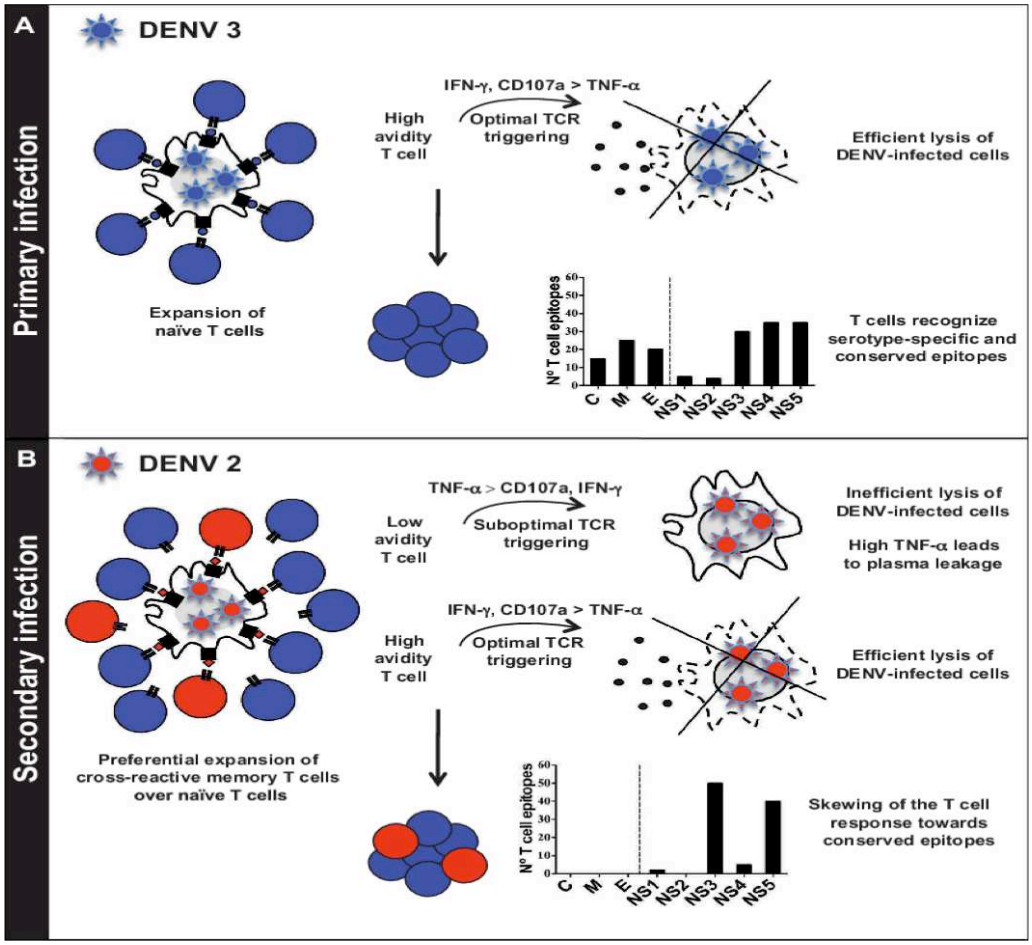


Figure 15. Original Antigenic Sin and the Dengue Shock Syndrome.

A. During a primary infection with DENV 3, there is an expansion of a range of naïve T-cell precursors with high avidity for serotype-specific and conserved DENV 3 epitopes. T-cells that have high specificity for their cognate antigen are activated through TCR signalling which mainly leads to production of cytokines resulting in the clearance of the pathogen. A proportion of the cells became memory cells (blue cells) with specificities for the different epitopes of the virus. B. Dengue Shock Syndrome physiopathology. In the case of a subsequent infection of the same individual occurring with a heterologous DENV 2 serotype, cross-reactive memory T-cells specific for the primary infecting virus will dominate the response due to their increased numbers and their lower activation threshold as compared to naïve T-cell precursors (red cells). Some of these cells may have a lower avidity for DENV 2 antigens, which results in suboptimal TCR activation leading to an inefficient clearance of DENV (From [371]) .

The low avidity has been pointed out to be the major driver of T-cell pathological implication in original antigenic sin as T-cell responses after secondary infections are highly serotype cross-reactive but less efficient in virus clearance[372,373]. A shift in the classical antiviral Th1 to a Th2 cytokines profile is also suspected to correlates with DSS apparition but it have to be confirmed [374]. In fact, this skewing of the immune has been

observed in another manifestation of the original antigenic sin involving the Respiratory Syncytial Virus.

Respiratory syncytial virus (RSV) is the leading cause of hospitalization in infants and young children[375]. However, there is no vaccine available. The reason for this may be the deadly failure of a clinical trial in the 1960's with a formalin-inactivated RSV (FI-RSV) vaccine[376]. The vaccine do not only failed to induce immunity against RSV, but it also resulted in an increased rate of disease severity after a subsequent natural RSV infection in the majority of the volunteers including hospitalization and two cases of fatal disease[377].

These patients presented unusual clinical features with lung pathology associated eosinophilia. Indeed, the two children that died revealed a significant increase in the number of eosinophils present in the lung parenchyma. The presence of eosinophils indicates a Th2 pathway that is not ordinary for anti-RSV response, as Th1 response generally favours virus clearance, whereas Th2 cell responses can prolonged virus replication and, in some cases, can lead to enhanced immunopathology[378].

Experimental mouse models of RSV have reproduced some of the features of this disease. Mice that were immunized with the FI-RSV[379] or in the form of a VV recombinant vaccine (VV-G)[380], developed a severe eosinophilia and a Th2 response after intranasal inoculation with infectious RSV. This contrasted with the Th1 response that was mounted by non-vaccinated RSV-challenged control mice, which easily controlled the infection.

The explanation leads in the composition of the RSV G-specific memory effector $CD4^+$ T-cells repertoire. RSV specific cells present in the lung following infection exhibit the Vβ14 region TCR with limited diversity within the CDR3 region. Remarkably, mice depleted of Vβ14 cells fail to develop pulmonary eosinophilia following subsequent RSV infection. These results demonstrate that immunopathology due to RSV vaccine seems to be link to an oligoclonal $CD4^+$ T-cell population and suggests that the engagement of these RSV G-specific $CD4^+$ T to Th2 differentiation may be dictated by TCR usage[381]. It is an example of heterologous immunity leading to immune deviation. It is interesting to note that mice previously infected with Flu before being vaccinated with VV-G vaccine did not develop eosinophilia subsequently to the RSV challenge[382].

4. Objectives and experiments

In the following sections, I will first detail the experimental design and methodological approach, before summarizing the obtained results.

4.1. The TRiPoD project:

First of all, this work is a part of the European Research Council advanced grant "Tripod: Treg repertoire in Physiology or Diseases" directed by Professor David Klatzmann[383]. The general objective of this work was to decipher the structure of the T cell repertoire in human. In particular, we aimed at portraying the composition of the TCR repertoire in the body and eventually to create a human TCR map in healthy condition. As far as I know, majority of the studies on the TCR repertoire has been performed on blood which contains less than 3% of the total T-cells of the body[384] and on unsorted cells. Moreover, the vascular compartment is not the place of antigenic stimulation. Collection of human tissues is thought to be an essential resource for biomedical research and the anatomical localization and tissue compartmentalization of immune cells is emerging as a key factor in their functions[385]. During my PhD, I have started and managed a research protocol in collaboration with the "Agence de biomedicine", "Coordination des prélèvements d'organes et de tissus" and the cardiac surgery team of Pitié-Salpêtrière hospital. The purpose was to obtain different lymphoid tissues from the same individual.

Adaptive immune system is a dynamic one. It is made of different kind of subpopulations with their own role in the defence of the host and control of immune reaction. The second major aim of this work was to sort cell populations from these organs to obtain an overview of all key actors of the adaptive immune response.

4.2. Original biobanking of sorted cells for TCR repertoire analysis

The original protocol allowed us to harvest lymphoid samples from 36 deceased organ donors. Donors ranged in age from 19-67 years old. We obtained a gender ratio of 1,6 H/F. All patients were free of cancer, hepatitis, HIV. Serological data indicated the status for persistent viruses EBV: 27 seropositive donors; and CMV: 16 seropositive donors. The clinical datas were of great importance for the performed analysis. To have a large mapping of the human immune system, we harvested (i) the thymus as it represents the central compartment of the immune system, (ii) blood and spleen that represent the circulating compartment and (iii) different organs draining lymph nodes (like pancreatic

or mesenteric) as it is the place for antigens presentation. We wanted to collect these samples to map the composition of the TCR repertoire in the body. We aimed to follow the TCR repertoire structure from the early selected one in the thymus to the finally activated one in the lymphoid organs. By looking at TCR repertoire in different localisations of the body, the first objective of this work was to highlight organ TCR repertoire signature.

To obtain the more exhaustive representation of the immune repertoire, we sorted different types of cells based on key phenotypes[386]. From the thymus, five thymocytes populations were sorted. DPCD3$^-$ ($CD3^-CD4^+CD8^+$) that are the most naïve thymocytes and represent the naïve TCR repertoire, DPCD3$^+$ ($CD3^+CD4^+CD8^+$) that expressed a full TCR and that are undergoing beta selection, ThyCD8 ($CD3^+CD4^-CD8^+$), ThyTeff ($CD3^+CD4^+CD8^-CD25^-$) and ThyTreg ($CD3^+CD4^+CD8^-CD25^+$) that are the naïve cells that have passed the thymic selection. In the periphery, including blood, spleen and lymph nodes, Teff ($CD3^+CD4^+CD25^-CD127^+$) and Treg ($CD3^+CD4^+CD25^+CD127^-$) cell subsets were sorted depending of their naïve or activated phenotype based on the respective expression of CD45RA$^+$ or CD45RO$^+$. CD8 T-cells were sorted in four subsets: naïve (nCD8): $CD3^+CD8^+CCR7^+CD45RA^+$, effector memory (emCD8): $CD3^+CD8^+CCR7^-CD45RA^-$, effector memory RA (emraCD8): $CD3^+CD8^+CCR7^-CD45RA^+$), central memory (cmCD8):$CD3^+CD8^+CCR7^+CD45RA^-$. The general protocol to obtain cells is represented in Figure 16:

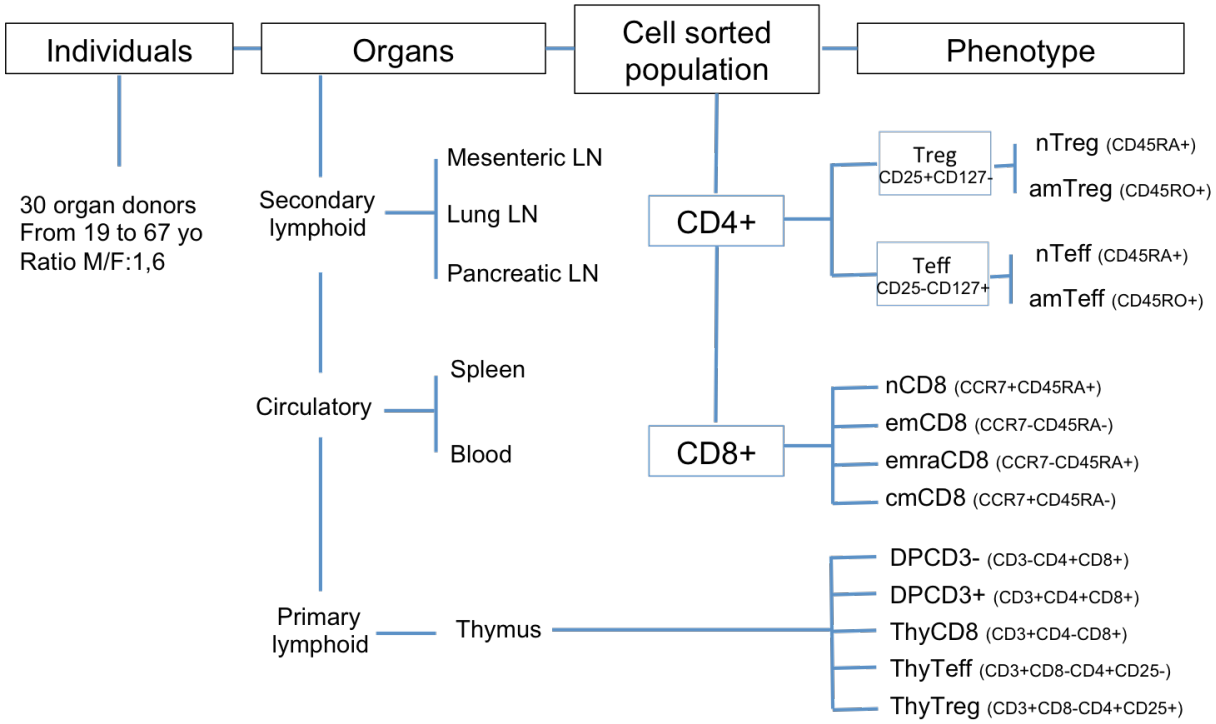


Figure 16. Flow chart of cell sorting strategy.

Overview of the different T-cell phenotypes sorted for this study from 30 donors, with blood, spleen, lymph nodes and thymus., We sorted 8 distinct T-cell subsets population for the periphery and 5 distinct T-cell subsets for the primary lymphoid organ. The cells were sorted according to their different phenotypes.

More than 900 samples were sorted and bio banked. It is, in our knowledge, the largest biobank dedicated to the analysis of human T-cells subsets repertoire in lymphoid tissues obtained from individual organ donor.

4.3. Library preparation and next generation sequencing:

The previous TCR sequencing technologies were restricted in their capacities to capture the diversity of the TCR repertoire. There are different reasons such as the requirement of extensive primer multiplexing or the limitation to a particular size or sequence. When we started this project, we hypothesised that the functionally relevant TCR repertoire is enough transduced in T lymphocytes and present in mRNA sequences. These molecules are composed of one known C region and an adjacent unknown flanking region representing the rearranged V(D)J region. So, we decided to use the mRNA to study the TCR repertoire. We tested different protocols of RNA extraction to finally use the RNAqueous-Kit from Invitrogen®. It allows to biobank the RNA in lysis buffer for a long and it was an easy to use and cost-effective technic for all the samples that we had. The

RNA concentration and sample integrity were systematically determined on NanoDrop (Thermo Fisher®).

When I started this project, I first designed an in-house protocol for the amplification of these genetic materials, adapted from [387]. The sequencing was performed under the 454 protocol (Roche®) based on pyrosequencing technology. It took at least 3 days for the sequencing step just for one sample and the results were not always reproducible. Given the number of samples that we had to perform, we realized that this technology was no longer adapted to our project. At the same the Illumina® HiSeq technology was emerging. It allowed sequencing much more samples in a reduced time. In parallel, Clonetech® developped a standardize protocol of library preparation that was compatible with Illumina® barcoding. This technology allowed us to performed 96 samples in the same batch. We assess that these technologies were adapted to our project in a multi-centric comparison of TCR repertoire library preparation. This comparison was the subject of a paper published in Nature Biotechnology. (See results)

With this established TCR sequencing technology, we were able to provide an in-depth characterisation of the TCR repertoire analysis.

4.4. TCR repertoire analysis

Deep sequencing raw data generated on Hiseq Illumina® were first evaluated for quality using the sequencer-specific quality programs provided by the manufacturer in order to ensure qualification of datasets. Thus, each generated dataset was subject to a standard of first-layer analysis comprising sequence quality control, barcode classification, and initial annotation. Each generated TCR sequence were annotated for V(D)J gene segment usage and boundaries, CDR1, CDR2 & CDR3 identification. This annotation step was performed with MiXCR[95] in accordance with the standards developed at ImMunoGeneTics[13]. For normalisation, we analysed the first 18,000 most expressed β or α CDR3s from each sample. 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 TCRβ-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.

There are three different ways that we used to analyse the thymopoiesis: the network, the Pgen and the specificity of the TCR.

First, 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[101,102,108,111,388]. These results are represented below in Figure 17:

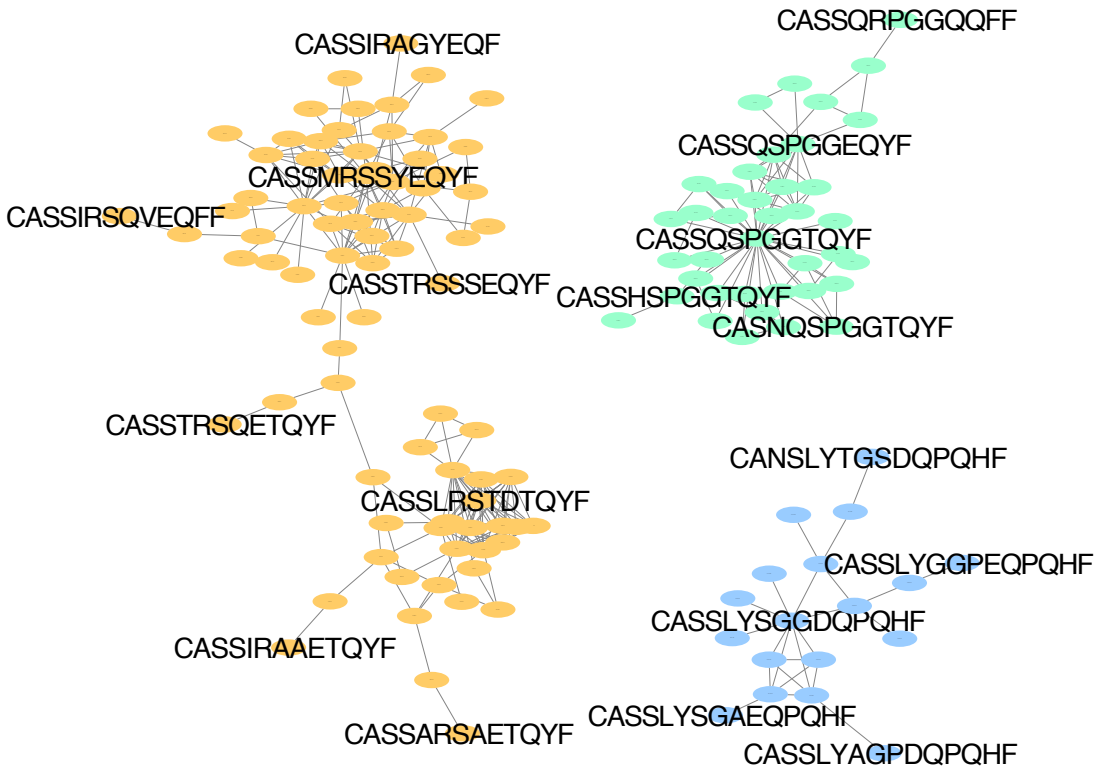


Figure 17. Networks of βCDR3s specific for antigen.

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 loaded with the indicated peptides. Each node represents a clonotype. Two different clonotypes are connected if their βCDR3s differ by at most one amino acid (LD≤1).

The second original way that we used to analyse the thymopoiesis is the generation probability (Pgen). The Pgen of a sequence is inferred using the Olga[389] algorithm, which is inferred by IGoR[230]. IGoR uses out-of-frame sequence information to infer patient-dependent models of VDJ recombination, effectively bypassing selection. From these models, the probability of a given recombination scenario can be computed. The generation probability of a sequence is then obtained by summing over all the scenarios that are compatible with it.

The third original way that we analysed the thymopoiesis was the use of public databases that provides the specificity of the TCR. The virus-associated CDR3 databases used for the search of specificity was compiled from the most complete previously published McPAS-TCR[390] and VDJdb[391] databases. However, public databases indicate TCRs as "specific" for a pathogen or a disease without explicitly specifying how this specificity has been demonstrated. Actually, for most TCRs, they are only "associated" with a condition, meaning found in certain experimental circumstances. As it was an important point, I manually curated the public databases, after checking in the referenced article to only retain TCRs for which specificity has been defined with tetramers/dextramers. In my work, all the TCRs referred to as "virus-specific" have thus been tested for binding a given viral peptide. Virus-specific CDR3s were selected from the original datasets only when derived from a TCR of sorted CD8 T cells that were bound by a specific tetramer. A total of 5,437 such unique tetramer-associated βCDR3s were identified and used. Peptides used for tetramer sorting were from cytomegalovirus (CMV), Epstein-Barr virus (EBV), hepatitis C virus (HCV), herpes simplex virus 2 (HSV2), human immunodeficiency virus (HIV), influenza and yellow fever virus (YFV). Moreover, we also used an original dataset of virus-specific CDR3 single-cell barcoded dextramers. This dataset contains single-cell αβ TCRs from 160 914 $CD8^+$ T cells isolated from peripheral blood mononuclear cells (PBMCs) from 4 healthy donors. A new technology, named dCODE™ Dextramer® developed by Immudex® as used in this experiment. The dCODE™ Dextramer® are DNA barcoded MHC dextramers designed for used in single-cell sequencing analysis. Each of these dextramers has a unique barcode encoding the MHC-peptide specificity displayed. We performed the analysis on 30 dCODE™ dextramers with antigenic peptides derived from infectious diseases (9 from CMV, 12 from EBV, 1 for influenza, 1 for HTLV, 2 for HPV and 5 for HIV). They were simultaneously used to mark cells. We used this dataset to study the presence of multiple specificities in TCR and CDR3. We identified 15,195 unique virus-specific TCRs with at least one binding. We also used single-cell sequencing datasets of TCRs from bronchoalveolar lavages from COVID-19 patients. This dataset contains single-cell alpha/beta TCRs from T cells isolated from bronchoalveolar lavage (BAL) of 9 patients infected by COVID-19[392]. We considered these TCR to be specific of COVID-19 as they were isolated in a primary site of infections.

To our knowledge, it was the first time that such biological samples were analysed these ways. In the main paper of my work, entitled: "Human thymopoiesis selects unconventional $CD8^+$ T cells that respond to multiple viruses", we report novel findings based on human thymocyte TCR sequencing that indeed fit with recent concepts in the field. For example, in a recent review article, Thomas & Crawford hypothesized that "particular TCR sequences and sequence clusters are highly favoured and recurrently generated across humans, and these same clusters tend to dominate responses to relevant pathogens"[393]. Here, we actually report for the first time the generation, thymic selection and characteristics of such sequences. Moreover, we also report totally new findings, showing that clustered/public/$Pgen^{high}$ TCRs have unconventionnal properties, responding to multiple different viruses, far beyond the classic concept of cross-reactivity.

These results are in line with a very recent review by Hayday & Vantourout entitled "The innate biologies of adaptive antigen receptors"[394] which ends with the following statement "our minds should be open to the possibility that various antigen receptor V-region-mediated innate interactions with endogenous and/or microbial moieties may regulate T and B lymphocytes independently of, or in concert with, clonotypic responsiveness. In particular, such interactions may ensure that appropriate lymphocyte repertoires are placed into a state of preparedness for mounting efficacious immune responsiveness." However, in this article, they mostly consider super-antigen-like activation of T cells, not a fuzzy MHC/peptide recognition, which we reveal. Classical thinking about cross-reactivity is that of totally unrelated sequences, which generate a similar spatial configuration (mimotopes) when presented by the MHC. In contrast, we propose here that a whole set of $\alpha\beta$ CD8+ T cells harbour TCRs that have an innate-like sensing of peptide/MHC complexes, i.e. more like pattern recognition receptors, a concept far beyond cross-reactivity.

5. Results and publications

5.1. Human thymopoiesis is not private and stochastic but public and pleiospecific.

Article 1: "Human thymopoiesis selects unconventional CD8⁺ α/β T cells that respond to multiple viruses." (*Quiniou et al., Nature, in review*)

How the immune system shapes a functional repertoire able to respond to not yet encounter antigens is one of the major questions of immunology. Hypothesis admitted is that a sufficiently diverse set of receptors makes it possible to generate a repertoire able to respond efficiently to not yet encounter antigens. However, considering the wide number of antigens and the limited number of lymphocytes in an organism, this theory could not be totally correct. We have notice numerous examples before that go against this theory.

In this paper, we demonstrate for the first time that unconventional CD8⁺ αβ TCR with innate-like recognition properties are preferentially selected in the thymus and respond to vaccination and infection.

First, we analysed the organisation of the TCR repertoire focusing on the CDR3 AA sequence which is independent of the MHC and directly interact with the antigen. We constructed network of the TCR repertoire, using the Levenshtein distance: CDR3 sequences are connected based on their level of similarity allowing a single amino acid difference (insertion, deletion, substitution). We observed an increase of clustered CDR3 and connections in ThyCD8 *vs* DPCD3- that reflect the selection of CDR3 with same capacities to recognize antigens used for thymic selection. These results highlight for the first times the oriented thymic selection process through CDR3 with shared specificities and invalidate the dogma of the largest possible diversity as a necessity to be selected.

Secondly, during this process of selection, a significant part of the selected TCR repertoire is public and sometimes shared by all individual. We found a significant increase of public CDR3 sequences in CD8 versus DPCD3⁺. In CD8, public CDR3 represents 8,6% while it is only 3,7% in DPCD3⁺. The public sequences were mostly those with the highest probability of generation and with shared specificities of

recognition as highlighted by the presence of a public CDR3 network that is selected in the thymus. We also investigated the number of connected nodes between patients. We sampled 1500 CDR3 in each individual for DPCD3$^+$ and CD8$^+$ thymocytes and compared the level of clustering and connection. We observe that the connectivity significantly increased between individual from the beginning of CDR3 recombination to thymic post selection. The number of connections is the higher for CD8$^+$ thymocytes, arguing for the selection of public specificities in the human thymus.

Thirdly, we observed an increase of CDR3 specific for pathogens during the process of selection. The proportion of these specific sequences is higher in CD8$^+$ versus DPCD3$^+$ arguing for an important selection of CDR3s specific of pathogens. We estimated that the viral antigen coverage in human CD8$^+$ is at least about 12%. Moreover, these TCR are reported to bind several unrelated antigens. we investigate at the single cell level the capacity of these cells to bind different kind of pHLA.

Fourthly, we investigated these properties and performed *in vitro* and *in vivo* experiments to outline the functional relevance of our observations. We highlighted that these cells are able not only to bind multiple pHLA, but also to be activated by these complexes and that these cells are ubiquitously recruited during vaccination or infection, including infection with COVID-19.

These results contradict the admitted rule of the clonal selection theory proposed by Burnet, hypothesised that a single TCR specifically recognize only one antigen and unlikely the others[2]. These public TCR repertoires present the specific characteristic to recognize multiple viral antigens.

According to these results, the stochastic mechanism of selection is no longer an acceptable process to explain the selection of the TCR repertoire, especially if we consider that the primary function of the repertoire is the protection against pathogens. These specific protections are related to the capacities of the adaptive system to identify pathogens antigens at the time of infection. So, changing the specificity of the repertoire over time will be ineffective. This paper calls into question the clonal selection theory of Burnet and also the admitted dogma of the necessity to select the most diverse TCR

repertoire to protect against all the possible pathogens that the individual will encounter. However, it provides evidences of a selection of public and unconventional $CD8^+$ T cells with TCR repertoire that is able to interact with multiple and unrelated human viruses.

5.2. Methods comparison for TCR repertoire data generation

Article 2: "Benchmarking of T-cell receptor repertoire profiling reveals large systematic biases" (*Barennes et al., Nature Biotechnology*)

From the biobanking to the generation of RepSeq data results, we have to be able to produce high quality data that can be investigated with all kind of analysis process. Establishing a technological pipeline that can provide an in-depth characterisation of TCR with reproducibility and accuracy was one of the major goals of this project. Indeed, we have seen that TCR repertoire analysis is determinant as it is a key in understanding mechanism of the immune adaptive system. But as every new domain, it lacks gold standard for investigation. To assess if this protocol was reproducible and adapted to catch the different facets of the TCR repertoire, we started a comparison with others technologies from commercial providers or from academic laboratories. We wanted to investigate the possible differences between methods for intra- and inter-method reproducibility and the accuracy of TCR determination, for both, alpha and beta chains. By using next generation sequencing and standardized input samples, we compared different methods used for TCR data generation. We also spiked the different samples with a known clonotype (Jurkat T-cell) at different concentration to estimate the capacity of the different method to detect low to high concentration clones.

Our study revealed that method RACE methods based on RNA performs better than multiplex PCR. DNA based commercial methods provide high quality data, however the absence of corresponding methods for the alpha TCR repertoire limits the interest of such approaches. Our method was less impacted by quantitative limitation of input material, which is an advantage compare to UMI-based protocol. Accuracy provided by UMIs was at the expense of some reduction in TCR diversity capture. It is evident that UMI-based methods require deeper sequencing to identify clonotypes.

All methods relatively accurately detected the Jurkat clonotype spiked into samples at pre-determined frequencies down to 0.1% but we observe intra- and inter method

differences. Results from RACE methods were consistent among themselves, while those from multiplex differed from each other and from RACE method results.

This study highlights the advantages and limitations of different TCR data generation methods and their potential impact on studies. Importantly for my project, these analyses allowed to identify and validate technique protocol that is now being used for all the i3 laboratory TCR repertoire projects.

5.3. Heterologous immunity between phage intestinal bacteria and cancer cells:

Article 3: "Cross-reactivity between MHC class I-restricted antigens from cancer cells and intestinal bacteria." (*Fluckiger et al., Science*)

The intestinal microbiota is a field of intense research. Recently, several studies indicate that the intestinal microbiota could influence autoimmune pathologies or cancer. In this collaborative project, the principal investigators identified a peptide epitope from a phage of an *Enterobacteraceae* that seems to have an anti-tumour effect. Indeed, mice bearing this bacterium harbouring this specific phage mounted a cytotoxic immune response specific for the phage peptide. The phage-peptide CTL also recognized an oncogenic driver. The administration of bacterial strains engineered to express the phage epitope improved the outcome of cancer treatment, and tumours bearing knock-in mutations for the oncogenic driver, that abolish cross-reactivity with phage specific CTL, became immunotherapy-resistant. We have highlighted based on tetramer cell sorting experiments that parts of TCR repertoire specific for the oncogenic driver and the phage were shared. These results highlight for the first time the implication of a phage as a part of heterologous immunity for antitumor activity.

The results support (i) the concept of intestinal microbiota anticancer immunosurveillance[395], (ii) the importance of heterologous immunity[272] and (iii) the concept of infectious symbiosis[396].

5.4. Oligoclonal repertoire in Takayasu lesions:

Article 4: "Specific T follicular helper gene signature discriminates large vessel vasculitis patients." (*Desbois et al., JCI insight, in review*)

Immune response profiles of Takayasu's arteritis (TAK) and giant cell arteritis (GCA), the two most common types of large vessel vasculitis (LVV), are currently poorly known.

The lesions in LVV are characterized by granulomatous inflammatory infiltrate and the presence of unexpected tertiary lymphoid organs in the media of the vessels.
First, we demonstrated for the first time a specific gene signature of circulating $CD4^+$ T-cells that discriminates large vessel vasculitis patients. We were able to collect large blood vessels from TAK patient under surgery and to sort cells from the inflammatory lesions. First, in arterial inflammatory lesions, we found higher proportion of tertiary lymphoid structures composed of $CD4^+$, $CXCR5^+$, $PD\text{-}1^+$ and $CD\text{-}20^+$ cells in TA compared to GCA. Second, TCR sequencing of $CD4^+$ $CXCR5^+$ T-cells located within aorta inflammatory lesions of TAK patients showed oligoclonal populations. The restricted repertory of $CD4^+$ $CXCR5^+$ T-cells within the aorta strongly suggests specific antigenic selection of $CD4^+$ $CXCR5^+$ T-cells. Interestingly, the TCR sequences found in this study have already been reported in diseases with autoimmune mechanisms and B cells abnormalities such as Sjögren syndrome. These results provided important breakthrough in the physiopathology of these diseases and give clues for potential therapeutics.

Altogether, these four studies pointed out the importance of the analysis of the TCR repertoire in physiopathology. The standardization process of this analysis is ongoing and further improvements, such as the normalisation and the generation of gold standard repertoire, needs to be achieved. However, within the accuracy of the method that we have developed, we were able to extract important results on the analysis of the TCR repertoire, on the mechanism of immune response and on physiopathology of diseases. The issues raised a lot of other questions, further discussed in the following section of this manuscript, and many other investigations have to be performed based on these results.

6. Discussions and Perspectives:

6.1. Unconventional TCR repertoire bridge Innate and Adaptive immunity:

It has been suggested that public TCR represent a primordial and necessary germline-encoded repertoire. The specificities of this repertoire that are unconventional in their peptide-binding properties, have higher affinity for MHC or are somehow different from other T-cell responses[397,398]. A parallel could be done with the B cell repertoire looking at the differences between natural IgM and IgG. IgM are low affinity bindings but their capacities of pentamerisation increase their affinity for antigens. There are able to bind low affinity antigens and initiate easily the immune response. The final steps of the immune response involving IgG are much more specific. The public-T-cell response, composed of TCR with high Pgen, may represent a subset of the TCR repertoire with low affinity or "degenerate affinity" but, as they are often present, they can act as "whistle-blower", to initiate an immune response that will be much more specific with highly specific TCR. The action of these public-T-cell gives time to much more specific TCRs, that are less present, to be recruited and to interact with their cognate antigens, with a higher affinity. Thus, IgM and high Pgen TCR are able to quickly initiate a "not specific" immune response.

These high Pgen TCRs share a lot of common features with constitutive receptors of the innate immune system. This sows seeds of doubt in the strict discrimination of innate and adaptive immunity. In the text-book « Immunobiology: The Immune System in Health and Disease »[399], the major characteristics of the innate and adaptive receptors are compared.

First of all, the distinction between TCR and Pattern recognition receptor (PRR), the major components of the innate immune system, is made on the inherited character of the receptor. Entire genes inherited through the germline encode the innate receptors, whereas genes that are combined from individual gene segments during lymphocyte development encode the adaptive receptors. However, we have seen that some TCR with high Pgen are germline encoded. It is also the case for the natural antibodies IgM, which are encoded by unmutated germ-line genes[400]. This part of the adaptive immune

receptor shared this characteristic of germline-encoded proteins, with the slight difference of recombination.

Secondly, the distinction between innate and adaptive receptor is conditioned to the receptor characteristic, i.e. all cells of a class express identical receptors. Here, we are dealing with the structural identity of the receptor. For example, the expression of TLR1 by macrophages. They are not clonal. It is true that TCR with different structures are present on T-cell. However, it has been suggested that subgroup of T-cell could harbour the degenerate TCR[398]. In the textbook cited above, the adaptive immunity is defined by the fact that, all cells of a class, e.g. $CD8^+$ T cells, express a single type of receptor with unique specificity. They are clonal. We have demonstrated that unconventional $CD8^+$ T cells selected during thymopoiesis harboured some identical receptors. This identity is even much more pronounced at the amino acid level. Furthermore, the specificity of these receptors is no longer unique. These receptors are able to bind multiple kind of pHLA complex which keep them away still a little more from the classical definition of adaptive immunity.

Thirdly, innate receptors trigger a very rapid response without the delay imposed by the clonal expansion of cells needed in the adaptive immune response. If we stand in the classical concept of adaptive immunity, the immune response starts as soon as the specific TCR has encountered its cognate antigen. Pretty sure, it could take a long time… The presence of pleiospecific TCR in high frequency allows a reduction of this delay. We have demonstrated that they are able to bind a multitude of pHLA complexes. The high frequency of this cells and their capacities to bind multiple pHLA ligands allow them to trigger a rapid immune response.

Finally, the discrimination of innate cells vs adaptive cells is based on the ligands recognized. The strategy of PRR is based on the detection of conserved molecular structures produced by microbial pathogens[401]. The classical example of PAMPs recognized by PRR is the lipopolysaccharide (LPS) of gram-negative bacteria. Evolution has tinkered PRRs so that they are able to recognize the common motifs on different pathogens. The classical T cells are defined as able to recognize specific details as proteic epitope. The pleiospecific cells with high *Pgen* TCR have the capacity to respond to a

multiple of epitope from unrelated pathogens. Of course, all the pathogens and antigens have not been tested but we have seen that these TCR were able to respond to CMV, EBV, Influenza and HIV. The first three are classical infectious virus of the human species. HIV is, fortunately, not a common pathogen but it is a retrovirus and they are coexisting with mammalian for millions of years[402]. So, the hypothesis is that evolution has tinkered high *Pgen* TCR so that they are able to recognize antigens of the common infectious pathogens. They are able to recognize the pathogens with the highest probability of infection.

Unconventional $CD8^+$ T cells have elements of both the innate and adaptive immunity. Investigation of gene signature via transcriptome analysis of these T-cell will help to define this subgroup that harbour pleiospecific TCR. The high throughput single cell analysis will provide insight in the differences between the sub- versus supra- specific TCR repertoire.

6.2. Symbiosis with pathogens: are we experienced?

During a lifetime, many different organisms infect human. They can be cleared or persist as chronic infectious pathogens. This latter case is often perceived as a dysfunction of the immune system and, generally, pathogens are defined as more or less negative for wellbeing. However, some human pathogens can protect their hosts from infection by related viruses or from disease caused by completely unrelated pathogens. This symbiosis, defined as a relationship between these two dissimilar entities, is mutualistic if each member benefits from the relationship. However, depending of the circumstances, there is either a benefit or a cost for it[396]. This symbiosis has been presented at the level of each individual, for example we have seen different examples of positive heterologous immunity in human as the CMV that enhance the immune response against influenza and other herpes virus may also increase the immune response to different pathogens. Moreover, this symbiosis has to be evaluated at the human species scale. Those who think it unacceptable that man evolved from apes will be defeated: there is also virus inside us! Indeed, retroviruses have the capabilities to integrate the host genome and become what is known as endogenous retroviruses (ERVs). Over time, retrovirus DNA sequence has been significantly added in mammalian genomes and is estimated to be up to 8% of the human genome [403]. Some of ERV

encoded functional genes involved in major evolutionary steps. In human, two envelope genes of retroviral origin, named syncytin 1 and 2 have a physiological function in the generation of the placenta and are involved in the formation of the syncytiotrophoblast layer[404]. Another example is the expression and activation of amylase in human salivary gland that is controlled by a retroviral insertion[405]. On an immunological point of view, there is a hypothesis about the presence of ERV, which may results in immunity to some lethal pathologies, so only individual with ERV survive[406]. For example, it has been recently highlighted that the role of ERV long terminal repeat domains (LTR) is implicated in the IFN-γ response[407]. Another example is the implication of ERVs protein expression that seems to help the anti-tumour response by serving as tumour-specific antigens. Endogenous retroviral genes coding for peptides recognized by tumour-specific CTL were identified in some human melanoma. These finding suggest that these antigens are targeted by CTLs in melanoma patients but there are for the moment no correlation between the expression of the antigen and survival[408]. In the article 3, the findings reported by Fluckiger et al. seems to go in the same way, with the slightly difference that the anti-tumour effect is not due to an ERV but due to a phage.

These different examples illustrate the importance of infection in evolution, which are a kind of symbiosis between human and pathogens. All humans become infected with multiple herpesviruses during childhood. We have presented numerous example that herpesvirus confers a benefit to the host against virus. There are different major facts to extract from these results. First of all, the classical approach in cellular and molecular immunity is largely based on experiments with pathogen-free mice that have never encountered any pathogens. Previous published papers have highlighted that there are significant immune modulations, with potentially either an alteration or an improvement of the organism response to foreign antigens. I think that an accurate understanding of the immune responses requires to assess the immunity under the presence of these symbiotic agents. Secondly, the level of symbiosis have to be determined concerning these different symbionts, which have been exerting selective pressures on the mammalian immune system for millions of years. We have seen that the presence of some virus can lead to the modification of the TCR repertoire. One major question is to determine the level of action of the symbionts, i.e. if these modifications are due to expansion of the TL via TCR recognition or if the presence of symbionts influences the TCR repertoire generation or selection. It is known that pathogens have

an impact on epigenetic[409]. This influence has to be evaluated especially since it has been established that the epigenetic variations contribute to the heritability of complex characters[410].

Finally, comes the burning question of vaccine. Vaccination is one of the major breakthroughs in medicine. It has saved millions of people. Different vaccines, as those against Hepatitis B, Papilloma virus or meningococcal conjugate vaccine contains only some proteins of the pathogens that act as immunodominant epitopes. The purpose was to strongly focus the immune response against this epitope to improve the efficacy of immunization. Considering previously published results, the ubiquitarian presence of pleiospecific cytotoxic cells and the diminished functionality of a narrow repertoire, this strategy has to be revaluated. In fact, this strategy has been suspected to prevent immune response against virus variant[411,412]. It seems that vaccines with several epitopes, or the inactivated pathogens, might provide better protective immunity based on a broader TCR repertoire diversity. Furthermore, considering the impact of heterologous immunity on health and disease, the impact of vaccination versus natural infection has to be evaluated.

6.3. Heterologous immunity: is the answer blowing in the AIRE?

The thymus is not a chocolate box. Indeed, it seems that the thymus always "knows" what you are gonna get (Gump et al. 1994…). We have pointed out that the thymic selection result of unconventionnal CD8 T cells with a public and highly cross-reactive repertoire, specific for the recognition of pathogens. The thymopeptidome expressed by thymic APC is unique inasmuch as these APCs express AIRE[178] that controls the expression of sets of otherwise tissue-restricted antigens. The thymopeptidome expressed by AIRE in the thymus is known to be fundamental in the process of selection. One hypothesis is that the AIRE complex presents the peptides that are the more likely to contribute to the selection of sufficiently cross-reactive TCR. The thymopeptidome lays in the endogenous protein's sequences of individual. Could the peptidome be a mean of all the peptides that human being has encountered during evolution? This "immune code" inside us has to be deciphered.

As thymocytes positive selection requires both proper interaction with thymic APC presenting self-peptides and thymocyte-thymocyte interaction[413], we plan to evaluated the characteristics of the sequences of peptides presented in the thymus by MCH-I to

understand the mechanisms underlying the selection of a repertoire specific for pathogens recognition. We used two public datasets of peptidome eluted from thymic cells MHC-I[414,415] to look for characteristics that could explain propensity for the selection of repertoires specific for pathogen recognition.

We investigated potential mechanisms for this selection of thymic TCRs with antiviral specificities in the absence of the viral antigens. Recent results show that physicochemical features of peptides can distinguish between non-immunogenic self and pathogen-derived foreign peptides and provide an accurate immunogenicity predictor that can be used to linked epitope immunogenicity and precursor frequency of specific T-cells[416,417]. We have therefore used the underlying classifier dataset, first published by Chowell et al., and from a physico-chemical perspective, to see whether the thymopeptidome is distinct from the overall set of human peptides in terms of immunogenicity. We have used mean values of Kidera factors to map peptides into the physicochemical feature space. Mean values were selected in order to compensate for length difference between eluted peptides and peptides from Chowell dataset. As can be seen from Figure 18.A, thymic peptides are closer to foreign peptides than self-peptides, even if they are also part of the self. Yet they do not have exactly the same feature profile and are situated between those two peptides set. Despite the fact that they are from self, thymic peptides have unique immunogenic features positioning them between foreign and self-peptides, suggesting a mimicry for pathogen-derived peptide properties.

To directly investigate a potential mimicry of the thymopeptidome for viral sequences, we blasted the thymopeptidome onto the public database of viral proteins (ViPR)[418]. We aligned viral- and self-peptide sequences based on the presence at a given position of the same AA or of an AA with the same physico-chemical properties. The hypothesis is that there is sufficiently mimicry between endogenous peptides and viral sequences to select a repertoire that is able to recognize viral proteins. Surprisingly, thymopeptides could be significantly aligned with peptide sequences of viruses known to be infectious for humans (Fig 18.B). We even observed stretch of up to nine identical amino acids (Fig 18.C). Out of 17 families of viruses that infect humans (and available in the database), thymopeptides could be significantly aligned with 15 of them, encompassing altogether 35 different viruses.

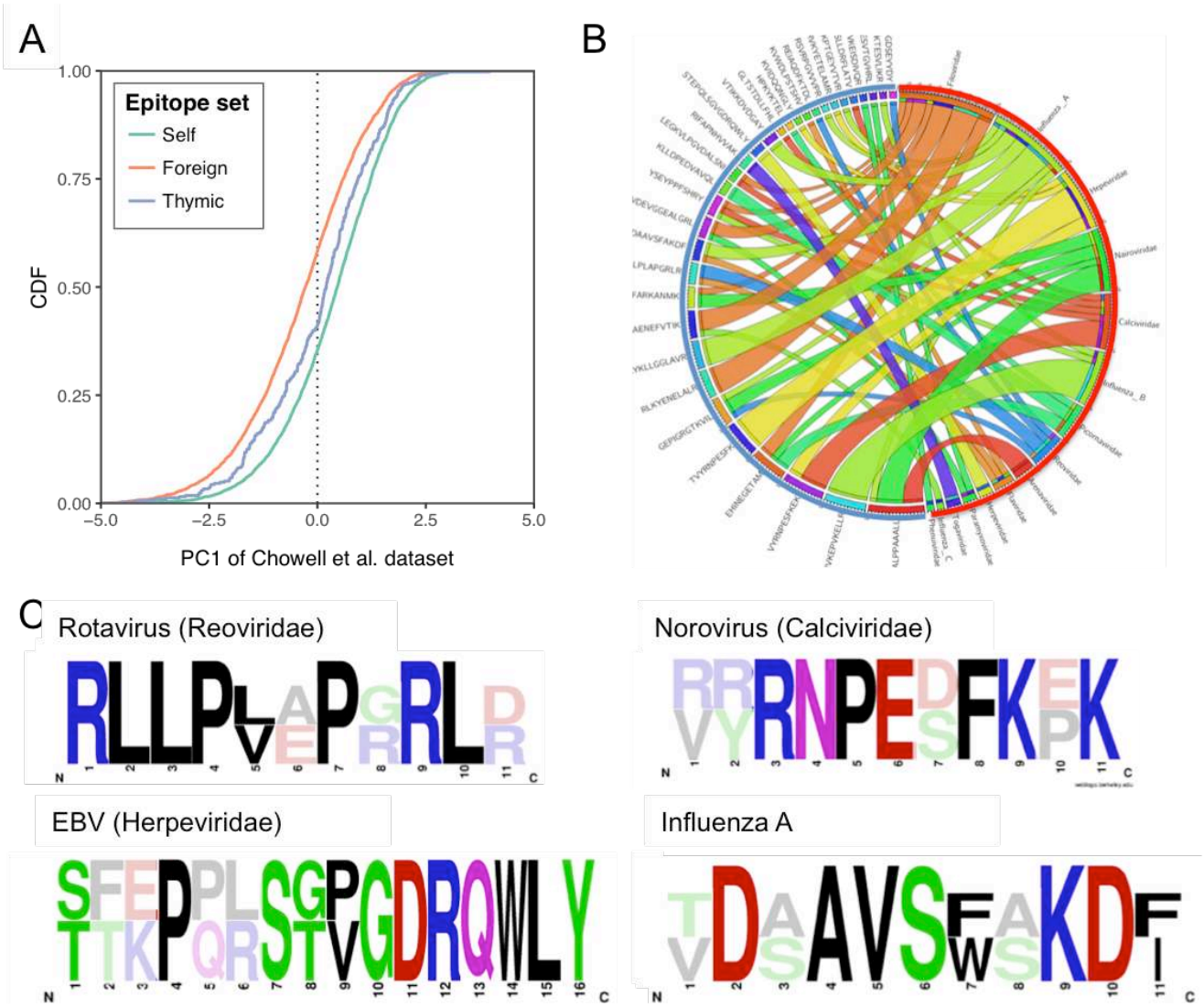


Figure 18. Thymopeptidome immunogenicity and pleiospecificity of CDR3s.

A. Similarity of physico-chemical properties between self-, foreign- and thymo-peptidomes. Cumulative distribution function (CDF) of the PC1 (first principal component explaining most variance[416]) values for self and foreign peptides compared to values for peptides eluted from human thymus tissue. PC1 values for thymic peptides lie in-between values for self and foreign peptide sets and are significantly different for both (D=0.11, P<4.10^{-4} and D= 0.17, P<10^{-9} respectively (Kolmogorov-Smirnov test)). B. Alignment of thymopeptides with viral peptides. Blue half circle represents the thymopeptides and the red half circle represent the virus class. The links on the circos plot represent the presence of an alignment of a thymopeptide with a virus peptide. Link colours represent the viral class. Thickness of the line represents the number of viruses from a viral class. Only thymopeptides with a significant alignment value (E-value<0.05) are represented. C. Sequence logo of thymopeptides alignment with viral peptide. Amino acids are coloured according to their chemical properties: polar (green), basic (blue), acidic (red) and hydrophobic (black).

These results highlight for the unique immunogenicity of the thymopeptidome in human, which appears uniquely suitable for selecting a TCR repertoire affine for, at

least, human viruses. Our endogenous proteins sequences seem to contain an immune code involves in the shaping of the adaptive immune system for cross-reactive recognition of pathogens. Thus, we can hypothesis that some variations in the endogenous proteins sequences that are used in the thymopeptidome leads to variation in the selection of the repertoire and so, in the immune function. More investigations have to evaluate the linked between the composition of the thymopeptidome, which seems to be sufficiently immunogenic to mimic pathogens antigens sequences, and the immune response, including auto-immunity.

Conclusion

Different processes allow the immune system to generate a diverse repertoire, like recombination or selection. In this study, we used an original set of human samples, combined with the latest NGS technology to investigate the mechanism of TCR repertoire selection. By leveraging the specificity of thymic TCR repertoires, we have given a proof that the TCR repertoire generation and selection is not randomly made. Furthermore, we have highlighted a convergence of the repertoire in the human species that allow a highly specific and efficient protection against human pathogens.

The network analysis provides a robust tool to assess the presence of close TCR repertoire that shared specificities. These specificities networks form antigen specific coverage that could act as cross-reactive public repertoire and improve the efficiency of the immune response. Moreover, it helps us to give an explanation to the overlooked heterologous immunity mechanism. Indeed, the shared capacities of recognition of the TCR repertoire appear to be fundamental in the survival of human species, especially in childhood. It seems that the pleiospecific selected repertoire is mainly oriented through pathogens that have been encountered by human being during evolution.

On the other hand, these unconventional CD8 T cells could be deleterious, but the presence of these cells is not totally illogical based on evolutionary perspective. Indeed, purpose of a species is survival via reproduction. So, it is necessary to protect species against acute lethal pathogens even if some individuals suffer from chronic autoreactivity. We anticipate these discoveries to provide unprecedented perspective for development of new types of vaccinations and for personalized therapy based on the TCR repertoire.

Bibliography

1. Alpert, A. et al. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. Nat. Med. **25**, 487–495 (2019).
2. Burnet, S. F. M. The clonal selection theory of acquired immunity. (1959).
3. Schirle, M., Weinschenk, T. & Stevanović, S. Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. J. Immunol. Methods **257**, 1–16 (2001).
4. Haskins, K. et al. The major histocompatibility complex-restricted antigen receptor on T cells. I. Isolation with a monoclonal antibody. J. Exp. Med. **157**, 1149–1169 (1983).
5. Meuer, S. C. et al. Clonotypic structures involved in antigen-specific human T cell function. J Exp Med **157**, 705–719 (1983).
6. Hedrick, S. M., Nielsen, E. A., Kavaler, J., Cohen, D. I. & Davis, M. M. Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins. Nature **308**, 153–158 (1984).
7. Siu, G. et al. The human t cell antigen receptor is encoded by variable, diversity, and joining gene segments that rearrange to generate a complete V gene. Cell **37**, 393–401 (1984).
8. Yanagi, Y. et al. A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. Nature **308**, 145–149 (1984).
9. Brenner, M. B. et al. Pillars Article: Identification of a Putative Second T-cell Receptor. Nature. 1986. 322: 145–149. J. Immunol. **196**, 3509 (2016).
10. Attaf, M., Huseby, E. & Sewell, A. K. αβ T cell receptors as predictors of health and disease. Mol. Immunol. 9 (2015).
11. Chen, H. et al. TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. Nat. Immunol. **13**, 691–700 (2012).

12. 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).
13. Lefranc, M.-P. et al. IMGT(R), the international ImMunoGeneTics information system(R). Nucleic Acids Res. **37**, D1006–D1012 (2009).
14. Scaviner, D. & Lefranc, M.-P. The human T cell receptor alpha variable (TRAV) genes. Exp. Clin. Immunogenet. **17**, 83–96 (2000).
15. Scaviner, D. & Lefranc, M.-P. The human T cell receptor alpha joining (TRAJ) genes. Exp. Clin. Immunogenet. **17**, 97–106 (2000).
16. Cabaniols, J.-P., Fazilleau, N., Casrouge, A., Kourilsky, P. & Kanellopoulos, J. M. Most α/β T Cell Receptor Diversity Is Due to Terminal Deoxynucleotidyl Transferase. J. Exp. Med. **194**, 1385–1390 (2001).
17. Schatz, D. G., Oettinger, M. A. & Baltimore, D. The V (D) J recombination activating gene, Rag-1. Cell **59**, 1035–1048 (1989).
18. Oettinger, M. A., Schatz, D. G., Gorka, C. & Baltimore, D. RAG-1 and RAG-2, adjacent genes that synergistically activate V (D) J recombination. Science **248**, 1517–1523 (1990).
19. Montaudouin, C. et al. Endogenous TCR Recombination in TCR Tg Single RAG-Deficient Mice Uncovered by Robust In Vivo T Cell Activation and Selection. PLoS ONE **5**, e10238 (2010).
20. Notarangelo, L. D., Kim, M.-S., Walter, J. E. & Lee, Y. N. Human RAG mutations: biochemistry and clinical implications. Nat. Rev. Immunol. **16**, 234–246 (2016).
21. Feeney, A. J., Goebel, P. & Espinoza, C. R. Many levels of control of V gene rearrangement frequency. Immunol. Rev. **200**, 44–56 (2004).

22. Livak, F., Burtrum, D. B., Rowen, L., Schatz, D. G. & Petrie, H. T. Genetic Modulation of T Cell Receptor Gene Segment Usage during Somatic Recombination. J. Exp. Med. **192**, 1191–1196 (2000).

23. McMurry, M. T. & Krangel, M. S. A role for histone acetylation in the developmental regulation of V (D) J recombination. Science **287**, 495–498 (2000).

24. Tripathi, R., Jackson, A. & Krangel, M. S. A Change in the Structure of Vβ Chromatin Associated with TCR β Allelic Exclusion. J. Immunol. **168**, 2316–2324 (2002).

25. Ma, Y., Schwarz, K. & Lieber, M. R. The Artemis:DNA-PKcs endonuclease cleaves DNA loops, flaps, and gaps. DNA Repair **4**, 845–851 (2005).

26. Moshous, D. et al. Artemis, a Novel DNA Double-Strand Break Repair/V(D)J Recombination Protein, Is Mutated in Human Severe Combined Immune Deficiency. Cell **105**, 177–186 (2001).

27. Martinez-Valdez, H. & Cohen, A. Coordinate regulation of mRNAs encoding adenosine deaminase, purine nucleoside phosphorylase, and terminal deoxynucleotidyltransferase by phorbol esters in human thymocytes. Proc. Natl. Acad. Sci. **85**, 6900–6903 (1988).

28. Motea, E. A. & Berdis, A. J. Terminal deoxynucleotidyl transferase: The story of a misguided DNA polymerase. Biochim. Biophys. Acta BBA - Proteins Proteomics **1804**, 1151–1166 (2010).

29. Schatz, D. G. V(D)J recombination. Immunol. Rev. **200**, 5–11 (2004).

30. Elhanati, Y., Sethna, Z., Callan, C. G., Mora, T. & Walczak, A. M. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol. Rev. **284**, 167–179 (2018).

31. Nikolich-Žugich, J., Slifka, M. K. & Messaoudi, I. The many important facets of T-cell repertoire diversity. Nat. Rev. Immunol. **4**, 123–132 (2004).

32. Qi, Q. et al. Diversity and clonal selection in the human T-cell repertoire. Proc. Natl. Acad. Sci. **111**, 13139–13144 (2014).

33. Doherty, P. C. & Zinkernagel, R. M. H-2 compatibility is required for T-cell-mediated lysis of target cells infected with lymphocytic choriomeningitis virus. J. Exp. Med. **141**, 502–507 (1975).

34. Zinkernagel, R. M. & Doherty, P. C. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. Nature **248**, 701 (1974).

35. Archbold, J. K. et al. Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition. J. Exp. Med. **206**, 209–219 (2009).

36. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. Nat. Rev. Immunol. **18**, 325–339 (2018).

37. Wieczorek, M. et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Front. Immunol. **8**, (2017).

38. Garboczi, D. N. et al. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature **384**, 134–141 (1996).

39. Rudolph, M. G. & Wilson, I. A. The specificity of TCR/pMHC interaction. Curr. Opin. Immunol. **14**, 52–65 (2002).

40. Buslepp, J., Wang, H., Biddison, W. E., Appella, E. & Collins, E. J. A Correlation between TCR V␣ Docking on MHC and CD8 Dependence: Implications for T Cell Selection. Immunity **Vol. 19, 595–606**, 12 (2003).

41. Garcia, K. C., Teyton, L. & Wilson, I. A. STRUCTURAL BASIS OF T CELL RECOGNITION. Annu. Rev. Immunol. **17**, 369–397 (1999).

42. Werlen, G. & Palmer, E. The T-cell receptor signalosome: a dynamic structure with expanding complexity. Curr. Opin. Immunol. **14**, 299–305 (2002).

43. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. HOW TCRS BIND MHCS, PEPTIDES, AND CORECEPTORS. Annu. Rev. Immunol. **24**, 419–466 (2006).

44. Huppa, J. B. & Davis, M. M. T-cell-antigen recognition and the immunological synapse. Nat. Rev. Immunol. **3**, 973–983 (2003).

45. Kim, S. T. et al. The αβ T Cell Receptor Is an Anisotropic Mechanosensor. J. Biol. Chem. **284**, 31028–31037 (2009).

46. Liu, B., Chen, W., Evavold, B. D. & Zhu, C. Antigen-specific TCR–pMHC catch bonds trigger signaling by fast accumulation of force-prolonged bond lifetimes. Cell (2014).

47. Kersh, G. J., Kersh, E. N., Fremont, D. H. & Allen, P. M. High- and Low-Potency Ligands with Similar Affinities for the TCR: The Importance of Kinetics in TCR Signaling. Immunity 10 (1998).

48. Williams, C. B., Engle, D. L., Kersh, G. J., Michael White, J. & Allen, P. M. A Kinetic Threshold between Negative and Positive Selection Based on the Longevity of the T Cell Receptor–Ligand Complex. J. Exp. Med. **189**, 1531–1544 (1999).

49. Savage, P. A. & Davis, M. M. A Kinetic Window Constricts the T Cell Receptor Repertoire in the Thymus. Immunity **14**, 243–252 (2001).

50. Teague, R. M. et al. Peripheral CD8+ T Cell Tolerance to Self-Proteins Is Regulated Proximally at the T Cell Receptor. Immunity **28**, 662–674 (2008).

51. Rabinowitz, J. D., Beeson, C., Lyons, D. S., Davis, M. M. & McConnell, H. M. Kinetic discrimination in T-cell activation. Proc. Natl. Acad. Sci. **93**, 1401–1405 (1996).

52. Love, P. E. & Hayes, S. M. ITAM-mediated Signaling by the T-Cell Antigen Receptor. Cold Spring Harb. Perspect. Biol. **2**, a002485–a002485 (2010).

53. Kaech, S. M. & Ahmed, R. Memory CD8+ T cell differentiation: initial antigen encounter triggers a developmental program in naïve cells. Nat. Immunol. **2**, 415–422 (2001).

54. Lenardo, M. J. lnterleukin-2 programs mouse ab T lymphocytes for apoptosis. Nature (1991).

55. Wherry, E. J. & Ahmed, R. Memory CD8 T-Cell Differentiation during Viral Infection. J. Virol. **78**, 5535–5545 (2004).

56. Jiang, J., Gross, D., Nogusa, S., Elbaum, P. & Murasko, D. M. Depletion of T Cells by Type I Interferon: Differences between Young and Aged Mice. J. Immunol. **175**, 1820–1826 (2005).

57. Cukalac, T. et al. Reproducible selection of high avidity CD8+ T-cell clones following secondary acute virus infection. Proc. Natl. Acad. Sci. **111**, 1485–1490 (2014).

58. Price, D. A. et al. Avidity for antigen shapes clonal dominance in CD8 $^{+}$ T cell populations specific for persistent DNA viruses. J. Exp. Med. **202**, 1349–1361 (2005).

59. Zehn, D., Lee, S. Y. & Bevan, M. J. Complete but curtailed T-cell response to very low-affinity antigen. Nature **458**, 211–214 (2009).

60. Wensveen, F. M. et al. Apoptosis Threshold Set by Noxa and Mcl-1 after T Cell Activation Regulates Competitive Selection of High-Affinity Clones. Immunity **32**, 754–765 (2010).

61. Pace, L. et al. Regulatory T Cells Increase the Avidity of Primary CD8+ T Cell Responses and Promote Memory. Science **338**, 532–536 (2012).

62. Willinger, T., Freeman, T., Hasegawa, H., McMichael, A. J. & Callan, M. F. C. Molecular Signatures Distinguish Human Central Memory from Effector Memory CD8 T Cell Subsets. J. Immunol. **175**, 5895–5903 (2005).

63. Kim, C. & Williams, M. A. Nature and nurture: T-cell receptor-dependent and T-cell receptor-independent differentiation cues in the selection of the memory T-cell pool: TCR-driven hierarchical differentiation of CD4+ T cells. Immunology **131**, 310–317 (2010).

64. Constant, S., Pfeiffer, C., Woodard, A., Pasqualini, T. & Bottomly, K. Extent of T Cell Receptor Ligation Can Determine the Functional Differentiation of Naive CD4 + T Cells. Journal of Experimental Medicine (1995).

65. Hosken, N., Shibuya, K., Heath, A. W., Murphy, K. M. & O'Garra, A. The Effect of Antigen Dose on CD4 + T Helper Cell Phenotype Development in a T Cell Receptor-alpha/beta-transgenic Model. Journal of Experimental Medicine (1995).

66. Yamane, H. & Paul, W. E. Early signaling events that underlie fate decisions of naive CD4 $^{+}$ T cells toward distinct T-helper cell subsets. Immunol. Rev. **252**, 12–23 (2013).

67. Yamane, H., Zhu, J. & Paul, W. E. Independent roles for IL-2 and GATA-3 in stimulating naive CD4 $^{+}$ T cells to generate a Th2-inducing cytokine environment. J. Exp. Med. **202**, 793–804 (2005).

68. Castelli, C. et al. T-cell recognition of melanoma-associated antigens. J. Cell. Physiol. 9 (2000).

69. Gottschalk, S., Rooney, C. M. & Heslop, H. E. Post-Transplant Lymphoproliferative Disorders. Annu. Rev. Med. **56**, 29–44 (2005).

70. de Pagter, P. J. et al. Human herpesvirus type 6 reactivation after haematopoietic stem cell transplantation. J. Clin. Virol. **43**, 361–366 (2008).

71. Walter, E. A., Finch, R. J., Watanabe, K. S., Thomas, E. D. & Riddell, S. R. Reconstitution of Cellular Immunity against Cytomegalovirus in Recipients of Allogeneic Bone Marrow by Transfer of T-Cell Clones from the Donor. N. Engl. J. Med. 7 (1995).

72. Nagata, S. Fas Ligand-Induced Apoptosis. Annu. Rev. Genet. **33**, 29–55 (1999).

73. Blott, E. J. & Griffiths, G. M. Secretory lysosomes. Nat. Rev. Mol. Cell Biol. **3**, 122–131 (2002).

74. Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. & Toda, M. Immunologic Self-Tolerance Maintained by Activated T Cells Expressing11-2 Receptor a-Chains (CD25). J. Immunol. 15 (1995).

75. Campbell, D. J. & Koch, M. A. Phenotypical and functional specialization of FOXP3+ regulatory T cells. Nat. Rev. Immunol. **11**, 119–130 (2011).

76. Sakaguchi, S., Powrie, F. & Ransohoff, R. M. Re-establishing immunological self-tolerance in autoimmune disease. Nat. Med. **18**, 54–58 (2012).

77. Burchill, M. A. et al. Linked T Cell Receptor and Cytokine Signaling Govern the Development of the Regulatory T Cell Repertoire. Immunity **28**, 112–121 (2008).

78. Hsieh, C.-S., Lee, H.-M. & Lio, C.-W. J. Selection of regulatory T cells in the thymus. Nat. Rev. Immunol. **12**, 157–167 (2012).

79. Lio, C.-W. J. & Hsieh, C.-S. A Two-Step Process for Thymic Regulatory T Cell Development. Immunity **28**, 100–111 (2008).

80. Levine, A. G., Arvey, A., Jin, W. & Rudensky, A. Y. Continuous requirement for the TCR in regulatory T cell function. Nat. Immunol. **15**, 1070–1078 (2014).

81. Zhu, J. & Shevach, E. M. TCR signaling fuels Treg cell suppressor function. Nat. Immunol. **15**, 1002–1003 (2014).

82. Six, A. et al. The Past, Present, and Future of Immune Repertoire Biology – The Rise of Next-Generation Repertoire Analysis. Front. Immunol. **4**, (2013).

83. Thomas-Vaslin, V. et al. Immunodepression and Immunosuppression During Aging. in Immunosuppression - Role in Health and Diseases (ed. Kapur, S.) (InTech, 2012). doi:10.5772/29549.

84. Cochet, M. et al. Molecular detection andin vivo analysis of the specific T cell response to a protein antigen. Eur. J. Immunol. **22**, 2639–2647 (1992).

85. Pannetier, C., Cochet, M., Darche, S., Casrouge, A. & Kourilsky, P. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor f8 chains vary as a function of the recombined germ-line segments. Proc Natl Acad Sci USA 5 (1993).

86. Gorski, J., Piatek, T., Yassai, M., Gorski, J. & Maslanka, K. Improvements in Repertoire Analysis by CDR3 Size Spectratyping.: Bifamily PCR. Ann. N. Y. Acad. Sci. **756**, 99–102 (1995).

87. Bouneaud, C., Kourilsky, P. & Bousso, P. Impact of negative selection on the T cell repertoire reactive to a self-peptide: a large fraction of T cell clones escapes clonal deletion. Immunity **13**, 829–840 (2000).

88. Sourdive, D. J. et al. Conserved T cell receptor repertoire in primary and memory CD8 T cell responses to an acute viral infection. J. Exp. Med. **188**, 71–82 (1998).

89. Musette, P. et al. Expansion of a recurrent V beta 5.3+ T-cell population in newly diagnosed and untreated HLA-DR2 multiple sclerosis patients. Proc. Natl. Acad. Sci. **93**, 12461–12466 (1996).

90. Yamamoto, K. et al. Establishment and application of a novel T cell clonality analysis using single-strand conformation polymorphism of T cell receptor messenger signals. Hum. Immunol. **48**, 23–31 (1996).

91. Sottini, A., Quiros-Roldan, E., Albertini, A., Primi, D. & Imberti, L. Assessment of T-Cell Receptor P-Chain Diversity by Heteroduplex Analysis. Human Immunology (1996).

92. Bolotin, D. A. et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. Eur. J. Immunol. **42**, 3073–3083 (2012).

93. Liu, X. et al. Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire. PLOS ONE **11**, e0152464 (2016).

94. Rosati, E. et al. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnol. **17**, (2017).

95. Bolotin, D. A. et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods **12**, 380–381 (2015).

96. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods **9**, 72–74 (2012).

97. Sepúlveda, N., Paulino, C. D. & Carneiro, J. Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. J. Immunol. Methods **353**, 124–137 (2010).

98. Venturi, V., Kedzierska, K., Turner, S. J., Doherty, P. C. & Davenport, M. P. Methods for comparing the diversity of samples of the T cell receptor repertoire. J. Immunol. Methods **321**, 182–195 (2007).

99. La Gruta, N. L. et al. Epitope-specific TCR repertoire diversity imparts no functional advantage on the CD8+ T cell response to cognate viral peptides. Proc. Natl. Acad. Sci. **105**, 2034–2039 (2008).

100. Yang, X. et al. Structural Basis for Clonal Diversity of the Public T Cell Response to a Dominant Human Cytomegalovirus Epitope. J. Biol. Chem. **290**, 29106–29119 (2015).

101. Dash, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature **547**, 89–93 (2017).

102. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. Nature **547**, 94–98 (2017).

103. Miles, J., Brennan, R. & Burrows, S. T Cell Receptor Bias in Humans. Curr. Immunol. Rev. **5**, 10–21 (2009).

104. Turner, S. J., Doherty, P. C., McCluskey, J. & Rossjohn, J. Structural determinants of T-cell receptor bias in immunity. Nat. Rev. Immunol. **6**, 883–894 (2006).

105. Stewart-Jones, G. B. E., McMichael, A. J., Bell, J. I., Stuart, D. I. & Jones, E. Y. A structural basis for immunodominant human T cell receptor recognition. Nat. Immunol. **4**, 7 (2003).

106. Jerne, N. K. Towards a network theory of the immune system. Ann Immunol **125**, 373–389 (1974).

107. 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).

108. 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).

109. Sun, Y. et al. Specificity, Privacy, and Degeneracy in the CD4 T Cell Receptor Repertoire Following Immunization. Front. Immunol. **8**, (2017).

110. 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).

111. 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).

112. Moss, P. A. H. et al. Extensive conservation of a and f8 chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide. 4 (1991).

113. Serana, F. et al. Identification of a public CDR3 motif and a biased utilization of T-cell receptor V beta and J beta chains in HLA-A2/Melan-A-specific T-cell clonotypes of melanoma patients. J. Transl. Med. **7**, 21 (2009).

114. Zoete, V., Irving, M., Ferber, M., Cuendet, M. A. & Michielin, O. Structure-Based, Rational Design of T Cell Receptors. Front. Immunol. **4**, (2013).

115. Shevach, E. M. Regulatory T Cells in Autoimmmunity. Annu. Rev. Immunol. **18**, 423–449 (2000).

116. Pathiraja, V. et al. Proinsulin-Specific, HLA-DQ8, and HLA-DQ8-Transdimer–Restricted CD4 $^{+}$ T Cells Infiltrate Islets in Type 1 Diabetes. Diabetes **64**, 172–182 (2015).

117. Rodriguez-Calvo, T., Ekwall, O., Amirian, N., Zapardiel-Gonzalo, J. & von Herrath, M. G. Increased Immune Cell Infiltration of the Exocrine Pancreas: A Possible Contribution to the Pathogenesis of Type 1 Diabetes. Diabetes **63**, 3880–3890 (2014).

118. Nejentsev, S. et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature **450**, 887–892 (2007).

119. Erlich, H. et al. HLA DR-DQ Haplotypes and Genotypes and Type 1 Diabetes Risk: Analysis of the Type 1 Diabetes Genetics Consortium Families. Diabetes **57**, 1084–1092 (2008).

120. Bulek, A. M. et al. Structural basis for the killing of human beta cells by CD8+ T cells in type 1 diabetes. Nat. Immunol. **13**, 283–289 (2012).

121. van Lummel, M. et al. Type 1 Diabetes-associated HLA-DQ8 Transdimer Accommodates a Unique Peptide Repertoire. J. Biol. Chem. **287**, 9514–9524 (2012).

122. Öling, V., Reijonen, H., Simell, O., Knip, M. & Ilonen, J. Autoantigen-specific memory CD4+ T cells are prevalent early in progression to Type 1 diabetes. Cell. Immunol. **273**, 133–139 (2012).

123. Seay, H. R. et al. Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. JCI Insight **1**, (2016).

124. Babbe, H. et al. Clonal Expansions of Cd8 $^{+}$ T Cells Dominate the T Cell Infiltrate in Active Multiple Sclerosis Lesions as Shown by Micromanipulation and Single Cell Polymerase Chain Reaction. J. Exp. Med. **192**, 393–404 (2000).

125. Jacobsen, M. et al. Oligoclonal expansion of memory CD8+ T cells in cerebrospinal ¯uid from multiple sclerosis patients. Brain 13 (2002).

126. Skulina, C. et al. Multiple sclerosis: Brain-infiltrating CD8+ T cells persist as clonal expansions in the cerebrospinal fluid and blood. Proc. Natl. Acad. Sci. **101**, 2428–2433 (2004).

127. Oksenberg, J. R. et al. Selection for T-cell receptor Vβ-Dβ-Jβ gene rearrangements with specificity for a myelin basic protein peptide in brain lesions of multiple sclerosis. **362**, 3 (1993).

128. Wucherpfennig, K. W. et al. Shared Human T Cell Receptor Vp, Usage to Immunodominant Regions of Myelin Basic Protein. **248**, 5 (1990).

129. Klarenbeek, P. L. et al. Inflamed target tissue provides a specific niche for highly expanded T-cell clones in early human autoimmune disease. Ann. Rheum. Dis. **71**, 1088–1093 (2012).

130. Liu, X. et al. T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. Ann. Rheum. Dis. **78**, 1070–1078 (2019).

131. Tyndall, A. & Dazzi, F. Chronic GVHD as an autoimmune disease. Best Pract. Res. Clin. Haematol. **21**, 281–289 (2008).

132. Copelan, E. A. Hematopoietic Stem-Cell Transplantation. N. Engl. J. Med. 14 (2006).

133. Meyer, E. H. et al. A distinct evolution of the T-cell repertoire categorizes treatment refractory gastrointestinal acute graft-versus-host disease. Blood **121**, 4955–4962 (2013).

134. Jungbluth, A. et al. Jungbluth AA, Chen Y-T, Stockert E., Busam KJ, Kolb D, Iversen K, Coplan K, Williamson B, Altorki N, Old LJ. Immunohistochemical analysis of NY-ESO-1 antigen expression in normal and malignant human tissues.International Journal of Cancer 2001; 92(6) 856-860. Int. J. Cancer **97**, 878–878 (2002).

135. Kurashige, T. et al. NY-ESO-1 Expression and Immunogenicity Associated with Transitional Cell Carcinoma: Correlation with Tumor Grade. Cancer Res. 5 (2001).

136. Odunsi, K. et al. NY-ESO-1 and LAGE-1 Cancer-Testis Antigens Are Potential Targets for Immunotherapy in Epithelial Ovarian Cancer. Cancer Res. 9 (2003).

137. Dietrich, P.-Y. et al. Prevalent Role of TCR α-Chain in the Selection of the Preimmune Repertoire Specific for a Human Tumor-Associated Self-Antigen. J. Immunol. **170**, 5103–5109 (2003).

138. Le Gal, F.-A. et al. Distinct Structural TCR Repertoires in Naturally Occurring Versus Vaccine-Induced CD8+ T-Cell Responses to the Tumor-Specific Antigen NY-ESO-1: J. Immunother. **28**, 252–257 (2005).

139. Rosenberg, S. A. et al. Durable Complete Responses in Heavily Pretreated Patients with Metastatic Melanoma Using T-Cell Transfer Immunotherapy. Clin. Cancer Res. **17**, 4550–4557 (2011).

140. Garrido, P., Ruiz-Cabello, F., Balanzategui, A., Almeida, J. & Orfao, A. Monoclonal TCR-V␣13.1⁺/CD4⁺/NKa⁺/CD8⁻/⁺dim T-LGL lymphocytosis: evidence for an antigen-driven chronic T-cell stimulation origin. **109**, 10 (2007).

141. Argaet, V. P. et al. Dominant Selection of an Invariant T Cell Antigen Receptor in Response to Persistent Infection by Epstein-Barr Virus. By Victor P. Argaet,* Christopher W. Schmidt,* Scott R. Burrows,* Sharon L. Silins,* Mike G. Kurilla,~ Denise L. Doolan,* Andreas Suhrbier,* Denis J. Moss,. J. Exp. Med. 6 (1994).

142. Kløverpris, H. N. et al. CD8+ TCR Bias and Immunodominance in HIV-1 Infection. J. Immunol. **194**, 5329–5345 (2015).

143. Trautmann, L. et al. Selection of T Cell Clones Expressing High-Affinity Public TCRs within Human Cytomegalovirus-Specific CD8 T Cell Responses. J. Immunol. **175**, 6123–6132 (2005).

144. Wu, L., Antica, M., Johnson, G. R., Scollay, R. & Shortman, K. Developmental Potential of the Earliest Precursor Cells from the Adult Mouse Thymus By Li Wu, MariastefaniaAntica, GregorytL. Johnson,. J. Exp. Med. 11 (1991).

145. Holt, P. G. & Jones, C. A. The development of the immune system during pregnancy and early life. Allergy **55**, 688–697 (2000).

146. Berthault, C. et al. Asynchronous lineage priming determines commitment to T cell and B cell lineages in fetal liver. Nat. Immunol. **18**, 1139–1149 (2017).

147. Carrelha, J. et al. Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. Nature **554**, 106–111 (2018).

148. Godfrey, D. I., Kennedy, J., Suda, T. & Zlotnik, A. A developmental pathway involving four phenotypically and functionally distinct subsets of CD3-CD4-CD8- triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. 10.

149. Kumar, B. V., Connors, T. J. & Farber, D. L. Human T Cell Development, Localization, and Function throughout Life. Immunity **48**, 202–213 (2018).

150. Sinclair, C., Bains, I., Yates, A. J. & Seddon, B. Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. Proc. Natl. Acad. Sci. **110**, E2905–E2914 (2013).

151. Weerkamp, F., Pike-Overzet, K. & Staal, F. J. T. T-sing progenitors to commit. Trends Immunol. **27**, 125–131 (2006).

152. Hori, S., Takashi, N. & Sakaguchi, S. Control of Regulatory T Cell Development by the Transcription Factor Foxp3. Science **299**, 1057–1061 (2003).

153. Watanabe, N. et al. Hassall's corpuscles instruct dendritic cells to induce CD4+CD25+ regulatory T cells in human thymus. Nature **436**, 1181–1185 (2005).

154. Garcia, K. C. et al. An αβ T Cell Receptor Structure at 2.5 A and Its Orientation in the TCR-MHC Complex. Science (1996).

155. Fehling, H. J., Krotkova, A. & Salnt-Ruft, C. Crucial role of the pre-T-cell receptor a gene in development of. **375**, 4 (1995).

156. Yamasaki, S. et al. Mechanistic basis of pre–T cell receptor–mediated autonomous signaling critical for thymocyte development. Nat. Immunol. **7**, 67–75 (2006).

157. Hayes, S. M., Shores, E. W. & Love, P. E. An architectural perspective on signaling by the pre-, alphabeta and gammadelta T cell receptors. Immunol. Rev. **191**, 28–37 (2003).

158. Michie, A. M. & Zúñiga-Pflücker, J. C. Regulation of thymocyte differentiation: pre-TCR signals and β-selection. Semin. Immunol. **14**, 311–323 (2002).

159. Pang, S. S. et al. The structural basis for autonomous dimerization of the pre-T-cell antigen receptor. Nature **467**, 844 (2010).

160. Aifantis, I. & Buer, J. Essential Role of the Pre-T Cell Receptor in Allelic Exclusion of the T Cell Receptor ␣ Locus. Immunity 7 (1997).

161. Thomas-Vaslin, V., Altes, H. K., de Boer, R. J. & Klatzmann, D. Comprehensive Assessment and Mathematical Modeling of T Cell Population Dynamics and Homeostasis. J. Immunol. **180**, 2240–2250 (2008).

162. Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. Nat. Rev. Immunol. **9**, 833 (2009).

163. van Meerwijk, J. P. M. et al. Quantitative Impact of Thymic Clonal Deletion on the T Cell Repertoire. J. Exp. Med. **185**, 377–384 (1997).

164. Zerrahn, J., Held, W. & Raulet, D. H. The MHC Reactivity of the T Cell Repertoire Prior to Positive and Negative Selection. Cell **88**, 627–636 (1997).

165. Sha, W. C. et al. Positive and negative selection of an antigen receptor on T cells in transgenic mice. Nature **336**, 73–76 (1988).

166. Barton, G. M. & Rudensky, A. Y. Evaluating peptide repertoires within the context of thymocyte development. Semin. Immunol. **11**, 417–422 (1999).

167. Zinkernagel, R. M., Callahan, G. N., Klein, J. & Dennert, G. Cytotoxic T cells learn specificity for self H–2 during differentiation in the thymus. Nature **271**, 251–253 (1978).

168. Nikolic-Zugic, J. & Bevan, M. J. Role of self-peptides in positively selecting the T-cell repertoire. Nature **344**, 65–67 (1990).

169. Laufer, T. M., DeKoning, J., Markowitz, J. S., Lo, D. & Glimcher, L. H. Unopposed positive selection and autoreactivity in mice expressing class II MHC only on thymic cortex. Nature **383**, 81–85 (1996).

170. Murata, S. et al. Regulation of CD8+ T Cell Development by Thymus-Specific Proteasomes. Science **316**, 1349–1353 (2007).

171. Nitta, T. et al. Thymoproteasome Shapes Immunocompetent Repertoire of CD8+ T Cells. Immunity **32**, 29–40 (2010).

172. Nakagawa, T. et al. Cathepsin L: Critical Role in Ii Degradation and CD4 T Cell Selection in the Thymus. Science **280**, 450–453 (1998).

173. Viret, C. et al. Thymus-specific serine protease contributes to the diversification of the functional endogenous CD4 T cell receptor repertoire. J. Exp. Med. **208**, 3–11 (2011).

174. Witt, C. M., Raychaudhuri, S., Schaefer, B., Chakraborty, A. K. & Robey, E. A. Directed Migration of Positively Selected Thymocytes Visualized in Real Time. PLoS Biol. **3**, e160 (2005).

175. McCaughtry, T. M., Baldwin, T. A., Wilken, M. S. & Hogquist, K. A. Clonal deletion of thymocytes can occur in the cortex with no involvement of the medulla. J. Exp. Med. **205**, 2575–2584 (2008).

176. Daniels, M. A. et al. Thymic selection threshold defined by compartmentalization of Ras/MAPK signalling. Nature **444**, 724–729 (2006).

177. Naeher, D. et al. A constant affinity threshold for T cell tolerance. J. Exp. Med. **204**, 2553–2559 (2007).

178. Derbinski, J., Schulte, A., Kyewski, B. & Klein, L. Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self. Nat. Immunol. **2**, 1032–1039 (2001).

179. Kyewski, B. & Klein, L. A CENTRAL ROLE FOR CENTRAL TOLERANCE. Annu. Rev. Immunol. **24**, 571–606 (2006).

180. Aaltonen, J. et al. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. Nat. Genet. **17**, 399–403 (1997).

181. Sewell, A. K. Why must T cells be cross-reactive? Nat. Rev. Immunol. **12**, 669–677 (2012).

182. Burroughs, N. J., de Boer, R. J. & Keşmir, C. Discriminating self from nonself with short peptides from large proteomes. Immunogenetics **56**, 311–320 (2004).

183. Müller, V. & Bonhoeffer, S. Quantitative constraints on the scope of negative selection. Trends Immunol. **24**, 132–135 (2003).

184. Stritesky, G. L., Jameson, S. C. & Hogquist, K. A. Selection of Self-Reactive T Cells in the Thymus. Annu. Rev. Immunol. **30**, 95–114 (2012).

185. Apostolou, I., Sarukhan, A., Klein, L. & von Boehmer, H. Origin of regulatory T cells with known specificity for antigen. Nat. Immunol. **3**, 756 (2002).

186. Bains, I., van Santen, H. M., Seddon, B. & Yates, A. J. Models of Self-Peptide Sampling by Developing T Cells Identify Candidate Mechanisms of Thymic Selection. PLoS Comput. Biol. **9**, e1003102 (2013).

187. Arpaia, E., Shahar, M., Dadi, H., Cohen, A. & Rolfman, C. M. Defective T cell receptor signaling and CD8+ thymic selection in humans lacking zap-70 kinase. Cell **76**, 947–958 (1994).

188. Chan, A. C. et al. ZAP-70 deficiency in an autosomal recessive form of severe combined immunodeficiency. Science **264**, 1599–1601 (1994).

189. Hivroz, C. Everything you ever wanted to know about ZAP-70. Med. Sci. MS **21**, 150–155 (2005).

190. Teh, H. S. et al. Thymic major histocompatibility complex antigens and the a{J T-cell receptor determine the CD4/CD8 phenotype ofT cells. Nature 5 (1988).

191. He, X. et al. The zinc finger transcription factor Th-POK regulates CD4 versus CD8 T-cell lineage commitment. Nature **433**, 826–833 (2005).

192. Park, J.-H. et al. Signaling by intrathymic cytokines, not T cell antigen receptors, specifies CD8 lineage choice and promotes the differentiation of cytotoxic-lineage T cells. Nat. Immunol. **11**, 257–264 (2010).

193. Brugnera, E. et al. Coreceptor Reversal in the Thymus: Signaled CD4⁻8⁻ Thymocytes Initially Terminate CD8 Transcription Even When Differentiating into CD8⁻ T Cells. Immunity 13 (2000).

194. Ashton-Rickardt, P. G. et al. Evidence for a differential avidity model of T cell selection in the thymus. Cell **76**, 651–663 (1994).

195. Sebzda, E. et al. Positive and Negative Thymocyte Selection Induced by Different Concentrations of a Single Peptide. Science (1994).

196. Alam, S. M. et al. T-cell-receptor affinity and thymocyte positive selection. **381**, 5 (1996).

197. Elliott, J. I. T cell repertoire formation displays characteristics of qualitative models of thymic selection. Eur. J. Immunol. **27**, 1831–1837 (1997).

198. Williams, O. et al. Interactions with multiple peptide ligands determine the fate of developing thymocytes. Proc. Natl. Acad. Sci. **95**, 5706–5711 (1998).

199. De Magistris, M. T. et al. Antigen analog-major histocompatibility complexes act as antagonists of the T cell receptor. Cell **68**, 625–634 (1992).

200. Hogquist, K. A. et al. T cell receptor antagonist peptides induce positive selection. Cell **76**, 17–27 (1994).

201. Jameson, S. C., Hogquist, K. A. & Bevan, M. J. Specificity and flexibility in thymic selection. Nature **369**, 750–752 (1994).

202. Chau, L. A., Bluestone, J. A. & Madrenas, J. Dissociation of Intracellular Signaling Pathways in Response to Partial Agonist Ligands of the T Cell Receptor. J. Exp. Med. **187**, 1699–1709 (1998).

203. Page, D. M. et al. Negative selection of CD4+ CD8+ thymocytes by T-cell receptor peptide antagonists. Proc. Natl. Acad. Sci. **91**, 4057–4061 (1994).

204. Spain, L. M., Jorgensen, J. L., Davis, M. M. & Berg, L. J. A peptide antigen antagonist prevents the differentiation of T cell receptor transgenic thymocytes. J. Immunol. 10 (2019).

205. Nakano, N., Rooke, R., Benoist, C. & Mathis, D. Positive Selection of T Cells Induced by Viral Delivery of Neopeptides to the Thymus. Science **275**, 678–683 (1997).

206. Smyth, L. A. et al. Altered peptide ligands induce quantitatively but not qualitatively different intracellular signals in primary thymocytes. Proc. Natl. Acad. Sci. **95**, 8193–8198 (1998).

207. Hemmer, B., Stefanova, I., Vergelli, M. & Martin, R. Relationships Among TCR Ligand Potency, Thresholds for Effector Function Elicitation, and the Quality of Early Signaling Events in Human T Cells. 9 (1998).

208. Marodon, G. et al. High diversity of the immune repertoire in humanized NOD.SCID.γc $^{-/-}$ mice: Cellular immune response. Eur. J. Immunol. **39**, 2136–2145 (2009).

209. Pham, H.-P. et al. Half of the T-cell repertoire combinatorial diversity is genetically determined in humans and humanized mice. Eur. J. Immunol. **42**, 760–770 (2012).

210. Shimizu, I., Fudaba, Y., Shimizu, A., Yang, Y.-G. & Sykes, M. Comparison of Human T Cell Repertoire Generated in Xenogeneic Porcine and Human Thymus Grafts: Transplantation **86**, 601–610 (2008).

211. Vandekerckhove, B. A. et al. Thymic selection of the human T cell receptor V beta repertoire in SCID- hu mice. J. Exp. Med. **176**, 1619–1624 (1992).

212. Khosravi-Maharlooei, M. et al. Cross-reactive public TCR sequences undergo positive selection in the human thymic repertoire. J. Clin. Invest. **129**, 2446–2462 (2019).

213. Madi, A. et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res. **24**, 1603–1612 (2014).

214. Jerne, Niels. K. THE NATURAL-SELECTION THEORY OF ANTIBODY FORMATION. Proceedings of the National Academy of Sciences (1955).

215. Jerne, N. K. The somatic generation of immune recognition. Eur. J. Immunol. **1**, 1–9 (1971).

216. Holler, P. & Kranz, D. M. T cell receptors: affinities, cross-reactivities, and a conformer model. Mol. Immunol. **40**, 1027–1031 (2004).

217. Wooldridge, L. et al. A Single Autoimmune T Cell Receptor Recognizes More Than a Million Different Peptides. J. Biol. Chem. **287**, 1168–1177 (2012).

218. Degauque, N., Brouard, S. & Soulillou, J.-P. Cross-Reactivity of TCR Repertoire: Current Concepts, Challenges, and Implication for Allotransplantation. Front. Immunol. **7**, (2016).

219. Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. Immunol. Today **19**, 395–404 (1998).

220. Godfrey, D. I., MacDonald, H. R., Kronenberg, M., Smyth, M. J. & Kaer, L. V. NKT cells: what's in a name? Nat. Rev. Immunol. **4**, 231–237 (2004).

221. Rossjohn, J., Pellicci, D. G., Patel, O., Gapin, L. & Godfrey, D. I. Recognition of CD1d-restricted antigens by natural killer T cells. Nat. Rev. Immunol. **12**, 845–857 (2012).

222. Kim, S.-K. et al. Private specificities of CD8 T cell responses control patterns of heterologous immunity. J. Exp. Med. **201**, 523–533 (2005).

223. Cibotti, R. et al. Public and Private VB T Cell Receptor Repertoires Against Hen Egg White Lysozyme (HEL) in Nontransgenic Versus HEL Transgenic Mice By RicardoCibotti, Jean-PierreCabaniols,* Christophe Pannetier, Christiane Delarbre,IsabelleVergnon,*Jean M. Kanellopoulos,. J. Exp. Med. 12 (1994).

224. Mora, T. & Walczak, A. M. Quantifying lymphocyte receptor diversity. http://biorxiv.org/lookup/doi/10.1101/046870 (2016) doi:10.1101/046870.

225. Arstila, T. P. et al. A Direct Estimate of the Human T Cell Receptor Diversity. Science **286**, 958–961 (1999).

226. Khan, N., Cobbold, M., Keenan, R. & Moss, P. A. H. Comparative Analysis of CD8 $^{+}$ T Cell Responses against Human Cytomegalovirus Proteins pp65 and Immediate Early 1

Shows Similarities in Precursor Frequency, Oligoclonality, and Phenotype. J. Infect. Dis. **185**, 1025–1034 (2002).

227. Argaet, V. P. et al. Dominant Selection of an Invariant T Cell Antigen Receptor in Response to Persistent Infection by Epstein-Barr Virus By Victor P. Argaet,* Christopher W. Schmidt,* Scott R. Burrows,* Sharon L. Silins,* Mike G. Kurilla,~ Denise L. Doolan,* Andreas Suhrbier,* Denis J. Moss,. J. Exp. Med. 6 (1994).
228. Lim, A. et al. Frequent Contribution of T Cell Clonotypes with Public TCR Features to the Chronic Response Against a Dominant EBV-Derived Epitope: Application to Direct Detection of Their Molecular Imprint on the Human Peripheral T Cell Repertoire. J. Immunol. **165**, 2001–2011 (2000).
229. Grandi, N. & Tramontano, E. Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. Front. Immunol. **9**, (2018).
230. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. Nat. Commun. **9**, (2018).
231. 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).
232. Venturi, V. et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. Proc. Natl. Acad. Sci. **103**, 18691–18696 (2006).
233. Fuschiotti, P. et al. Analysis of the TCR α-chain rearrangement profile in human T lymphocytes. Mol. Immunol. **44**, 3380–3388 (2007).
234. Wallace, M. E. et al. Junctional Biases in the Naive TCR Repertoire Control the CTL Response to an Immunodominant Determinant of HSV-1. Immunity **12**, 547–556 (2000).

235. 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).

236. Fazilleau, N. et al. Vα and Vβ Public Repertoires Are Highly Conserved in Terminal Deoxynucleotidyl Transferase-Deficient Mice. J. Immunol. **174**, 345–355 (2005).

237. Venturi, V., Price, D. A., Douek, D. C. & Davenport, M. P. The molecular basis for public T-cell responses? Nat. Rev. Immunol. **8**, 231–238 (2008).

238. Gillespie, G. M. A. et al. Strong TCR Conservation and Altered T Cell Cross-Reactivity Characterize a B*57-Restricted Immune Response in HIV-1 Infection. J. Immunol. **177**, 3893–3902 (2006).

239. Kedzierska, K., Turner, S. J. & Doherty, P. C. Conserved T cell receptor usage in primary and recall responses to an immunodominant influenza virus nucleoprotein epitope. Proc. Natl. Acad. Sci. **101**, 4942–4947 (2004).

240. Price, D. A. et al. T Cell Receptor Recognition Motifs Govern Immune Escape Patterns in Acute SIV Infection. Immunity **21**, 793–803 (2004).

241. Yu, X. G. et al. Mutually Exclusive T-Cell Receptor Induction and Differential Susceptibility to Human Immunodeficiency Virus Type 1 Mutational Escape Associated with a Two-Amino-Acid Difference between HLA Class I Subtypes. J. Virol. **81**, 1619–1631 (2007).

242. Brennan, R. M. et al. Predictable T-Cell Receptor Selection toward an HLA-B*3501-Restricted Human Cytomegalovirus Epitope. J. Virol. **81**, 7269–7273 (2007).

243. Menezes, J. S. et al. A public T cell clonotype within a heterogeneous autoreactive repertoire is dominant in driving EAE. J. Clin. Invest. **117**, 2176–2185 (2007).

244. Zhong, W. & Reinherz, E. L. In vivo selection of a TCR V repertoire directed against an immunodominant influenza virus CTL epitope. Int. Immunol. **16**, 1549–1559 (2004).

245. Lehner, P. J. et al. Human HLA-A0201-restricted cytotoxic T lymphocyte recognition of influenza A is dominated by T cells bearing the Vβ17 gene segment. Journal of Experimental Medicine (1995).

246. Kedzierska, K., La Gruta, N. L., Davenport, M. P., Turner, S. J. & Doherty, P. C. Contribution of T cell receptor affinity to overall avidity for virus-specific CD8+ T cell responses. Proc. Natl. Acad. Sci. **102**, 11432–11437 (2005).

247. Kedzierska, K. et al. Early establishment of diverse T cell receptor profiles for influenza-specific CD8+CD62Lhi memory T cells. Proc. Natl. Acad. Sci. **103**, 9184–9189 (2006).

248. Tynan, F. E. et al. High Resolution Structures of Highly Bulged Viral Epitopes Bound to Major Histocompatibility Complex Class I: IMPLICATIONS FOR T-CELL RECEPTOR ENGAGEMENT AND T-CELL IMMUNODOMINANCE. J. Biol. Chem. **280**, 23900–23909 (2005).

249. Kjer-Nielsen, L. et al. A Structural Basis for the Selection of Dominant alpha beta T Cell Receptors in Antiviral Immunity. Immunity 12 (2003).

250. Tynan, F. E. et al. A T cell receptor flattens a bulged antigenic peptide presented by a major histocompatibility complex class I molecule. Nat. Immunol. **8**, 268–276 (2007).

251. Reiser, J.-B. et al. CDR3 loop flexibility contributes to the degeneracy of TCR recognition. Nat. Immunol. **4**, 241–247 (2003).

252. Armstrong, K. M., Piepenbrink, K. H. & Baker, B. M. Conformational changes and flexibility in T-cell receptor recognition of peptide–MHC complexes. Biochem. J. **415**, 183–196 (2008).

253. Borbulevych, O. Y. et al. T Cell Receptor Cross-reactivity Directed by Antigen-Dependent Tuning of Peptide-MHC Molecular Flexibility. Immunity **31**, 885–896 (2009).

254. James, L. C., Roversi, P. & Tawfik, D. S. Antibody Multispecificity Mediated by Conformational Diversity. **299**, 7 (2003).

255. Hahn, M., Nicholson, M. J., Pyrdol, J. & Wucherpfennig, K. W. Unconventional topology of self peptide–major histocompatibility complex binding by a human autoimmune T cell receptor. Nat. Immunol. **6**, 490–496 (2005).

256. Colf, L. A. et al. How a Single T Cell Receptor Recognizes Both Self and Foreign MHC. Cell **129**, 135–146 (2007).

257. Li, Y. et al. Structure of a human autoimmune TCR bound to a myelin basic protein self-peptide and a multiple sclerosis-associated MHC class II molecule. EMBO J. **24**, 2968–2979 (2005).

258. Hemmer, B. et al. Contribution of Individual Amino Acids Within MHC Molecule or Antigenic Peptide to TCR Ligand Potency. J. Immunol. **164**, 861–871 (2000).

259. Lang, H. L. E. et al. A functional and structural basis for TCR cross-reactivity in multiple sclerosis. Nat. Immunol. **3**, 940–943 (2002).

260. Birnbaum, M. E. et al. Deconstructing the Peptide-MHC Specificity of T Cell Recognition. Cell **157**, 1073–1087 (2014).

261. Yin, Y. & Mariuzza, R. A. The Multiple Mechanisms of T Cell Receptor Cross-reactivity. Immunity **31**, 849–851 (2009).

262. Gilfillan, S. et al. Efficient immune responses in mice lacking N-region diversity. Eur. J. Immunol. **25**, 3115–3122 (1995).

263. 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).

264. Mathurin, K. S., Martens, G. W., Kornfeld, H. & Welsh, R. M. CD4 T-Cell-Mediated Heterologous Immunity between Mycobacteria and Poxviruses. J. Virol. **83**, 3528–3539 (2009).

265. Lin, M. Y., Selin, L. K. & Welsh, R. M. Evolution of the CD8 T-cell repertoire during infections. Microbes Infect. 15 (2000).

266. Cornberg, M. et al. Narrowed TCR repertoire and viral escape as a consequence of heterologous immunity. J. Clin. Invest. **116**, 1443–1456 (2006).

267. Brehm, M. A. et al. T cell immunodominance and maintenance of memory regulated by unexpectedly cross-reactive pathogens. Nat. Immunol. **3**, 627–634 (2002).

268. Welsh, R. M. & Selin, L. K. No one is naive: the significance of heterologous T-cell immunity. Nat. Rev. Immunol. **2**, 417–426 (2002).

269. Chen, H. D. et al. Memory CD8+ T cells in heterologous antiviral immunity and immunopathology in the lung. Nat. Immunol. **2**, 1067–1076 (2001).

270. Selin, L. K., Varga, S. M., Wong, I. C. & Welsh, R. M. Protective Heterologous Antiviral Immunity and Enhanced Immunopathogenesis Mediated by Memory T Cell Populations. J. Exp. Med. **188**, 1705–1715 (1998).

271. Cornberg, M. et al. CD8 T Cell Cross-Reactivity Networks Mediate Heterologous Immunity in Human EBV and Murine Vaccinia Virus Infections. J. Immunol. **184**, 2825–2838 (2010).

272. Welsh, R. M., Che, J. W., Brehm, M. A. & Selin, L. K. Heterologous immunity between viruses: Heterologous immunity between viruses. Immunol. Rev. **235**, 244–266 (2010).

273. Burrows, S. R. et al. Cross-reactive memory T cells for Epstein-Barr virus augment the alloresponse to common human leukocyte antigens: degenerate recognition of major histocompatibility complex-bound peptide by T cells and its role in alloreactivity. Eur. J. Immunol. **27**, 1726–1736 (1997).

274. Burrows, S. R., Khanna, R., Silins, S. L. & Moss, D. J. The influence of antiviral T-cell responses on the alloreactive repertoire. Immunol. Today **20**, 203–207 (1999).

275. Clute, S. C. et al. Cross-reactive influenza virus-specific CD8+ T cells contribute to lymphoproliferation in Epstein-Barr virus-associated infectious mononucleosis. J. Clin. Invest. **115**, 3602–3612 (2005).

276. Sandalova, E. et al. Contribution of herpesvirus specific CD8 T cells to anti-viral T cell response in humans. PLoS Pathog. **6**, e1001051 (2010).

277. Stewart, A. J. & Devlin, P. M. The history of the smallpox vaccine. J. Infect. **52**, 329–334 (2006).

278. Mayr, A. Taking Advantage of the Positive Side-Effects of Smallpox Vaccination. J. Vet. Med. Ser. B **51**, 199–201 (2004).

279. Aaby, P. et al. Non-specific beneficial effect of measles immunisation: analysis of mortality studies from developing countries. BMJ **311**, 481–485 (1995).

280. Aaby, P. et al. The survival benefit of measles immunization may not be explained entirely by the prevention of measles disease: a community study from rural Bangladesh. Int. J. Epidemiol. **32**, 106–115 (2003).

281. Aaby, P. et al. The optimal age of measles immunisation in low-income countries: a secondary analysis of the assumptions underlying the current policy. BMJ Open **2**, e000761 (2012).

282. Bate, S. L., Dollard, S. C. & Cannon, M. J. Cytomegalovirus seroprevalence in the United States: the national health and nutrition examination surveys, 1988–2004. Clin. Infect. Dis. **50**, 1439–1447 (2010).

283. Kuijpers, T. W. et al. Frequencies of circulating cytolytic, CD45RA+ CD27−, CD8+ T lymphocytes depend on infection with CMV. J. Immunol. **170**, 4342–4348 (2003).

284. Sylwester, A. W. et al. Broadly targeted human cytomegalovirus-specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects. J. Exp. Med. **202**, 673–685 (2005).

285. Webster, R. & Stohr, N. C. WHO Manual on Animal Influenza Diagnosis and Surveillance. (2002).

286. Lee, L. Y.-H. et al. Memory T cells established by seasonal human influenza A infection cross-react with avian influenza A (H5N1) in healthy individuals. 13 (2008).

287. Epstein, S. L. Prior H1N1 Influenza Infection and Susceptibility of Cleveland Family Study Participants during the H2N2 Pandemic of 1957: An Experiment of Nature. J. Infect. Dis. **193**, 49–53 (2006).

288. Wedemeyer, H., Mizukoshi, E., Davis, A. R., Bennink, J. R. & Rehermann, B. Cross-Reactivity between Hepatitis C Virus and Influenza A Virus Determinant-Specific Cytotoxic T Cells. J. Virol. **75**, 11392–11400 (2001).

289. Grebely, J. et al. Hepatitis C virus clearance, reinfection, and persistence, with insights from studies of injecting drug users: towards a vaccine. Lancet Infect. Dis. **12**, 408–414 (2012).

290. Kaul, R. et al. CD8+ lymphocytes respond to different HIV epitopes in seronegative and infected subjects. J. Clin. Invest. **107**, 1303–1310 (2001).

291. Shacklett, B. L. Understanding the "Lucky Few": The Conundrum of HIV-exposed, Seronegative Individuals. 2 (2006).

292. Tillmann, H. L., Ockenga, J., Mcmorrow, M. & Stoll, M. Infection with GB Virus C and Reduced Mortality among HIV-Infected Patients. N. Engl. J. Med. 10 (2001).

293. Xiang, J. & Patrick, K. D. Effect of Coinfection with GB Virus C on Survival among Patients with HIV Infection. N. Engl. J. Med. 8 (2001).

294. Xiang, J., McLinden, J. H., Chang, Q., Jordan, E. L. & Stapleton, J. T. Characterization of a Peptide Domain within the GB Virus C NS5A Phosphoprotein that Inhibits HIV Replication. PLoS ONE **3**, e2580 (2008).

295. Aaby, P. et al. Randomized Trial of BCG Vaccination at Birth to Low-Birth-Weight Children: Beneficial Nonspecific Effects in the Neonatal Period? J. Infect. Dis. **204**, 245–252 (2011).

296. Hollm-Delgado, M.-G., Stuart, E. A. & Black, R. E. Acute Lower Respiratory Infection Among Bacille Calmette-Guérin (BCG)–Vaccinated Children. Pediatrics **133**, e73–e81 (2014).

297. Setia, M. S., Steinmaus, C., Ho, C. S. & Rutherford, G. W. The role of BCG in prevention of leprosy: a meta-analysis. Lancet Infect. Dis. **6**, 162–170 (2006).

298. Elliott, A. M. et al. Inverse association between BCG immunisation and intestinal nematode infestation among HIV-1-positive individuals in Uganda. The Lancet **354**, 1000–1001 (1999).

299. Peterson, L. K. & Fujinami, R. S. 3 - MOLECULAR MIMICRY. in Autoantibodies (Second Edition) (eds. Shoenfeld, Y., Gershwin, M. E. & Meroni, P. L.) 13–19 (Elsevier, 2007). doi:10.1016/B978-044452763-9/50007-X.

300. Libbey, J. E., Cusick, M. F. & Fujinami, R. S. Role of pathogens in multiple sclerosis. Int. Rev. Immunol. **33**, 266–283 (2014).

301. Sospedra, M. & Martin, R. Immunology of multiple sclerosis. Annu Rev Immunol **23**, 683–747 (2005).

302. Sotelo, J. On the viral hypothesis of multiple sclerosis: Participation of varicella-zoster virus. J. Neurol. Sci. **262**, 113–116 (2007).

303. Du, C., Yao, S.-Y., Ljunggren-Rose, Å. & Sriram, S. Chlamydia pneumoniae infection of the central nervous system worsens experimental allergic encephalitis. J. Exp. Med. **196**, 1639–1644 (2002).

304. Lünemann, J. D. et al. EBNA1-specific T cells from patients with multiple sclerosis cross react with myelin antigens and co-produce IFN-γ and IL-2. J. Exp. Med. **205**, 1763–1773 (2008).

305. Wucherpfennig, K. W. & Strominger, J. L. Molecular mimicry in T cell-mediated autoimmunity: Viral peptides activate human T cell clones specific for myelin basic protein. Cell **80**, 695–705 (1995).

306. Harkiolaki, M. et al. T cell-mediated autoimmune disease due to low-affinity crossreactivity to common microbial peptides. Immunity **30**, 348–357 (2009).

307. Cao, Y. et al. Functional inflammatory profiles distinguish myelin-reactive T cells from patients with multiple sclerosis. Sci. Transl. Med. **7**, 287ra74-287ra74 (2015).

308. Hor, H. et al. Genome-wide association study identifies new HLA class II haplotypes strongly protective against narcolepsy. Nat. Genet. **42**, 786–789 (2010).

309. Kornum, B. R. et al. Common variants in P2RY11 are associated with narcolepsy. Nat. Genet. **43**, 66–71 (2011).

310. Partinen, M. et al. Narcolepsy as an autoimmune disease: the role of H1N1 infection and vaccination. Lancet Neurol. **13**, 600–613 (2014).

311. Han, F. et al. Narcolepsy onset is seasonal and increased following the 2009 H1N1 pandemic in china. Ann. Neurol. **70**, 410–417 (2011).

312. Ahmed, S. S. et al. Antibodies to influenza nucleoprotein cross-react with human hypocretin receptor 2. Sci. Transl. Med. **7**, 294ra105-294ra105 (2015).

313. Hallmayer, J. et al. Narcolepsy is strongly associated with the T-cell receptor alpha locus. Nat. Genet. **41**, 708–711 (2009).

314. Luo, G. et al. Autoimmunity to hypocretin and molecular mimicry to flu in type 1 narcolepsy. Proc. Natl. Acad. Sci. **115**, E12323–E12332 (2018).

315. Borchers, A. T., Uibo, R. & Gershwin, M. E. The geoepidemiology of type 1 diabetes. Autoimmun. Rev. **9**, A355–A365 (2010).

316. Coppieters, K. T., Wiberg, A. & von Herrath, M. G. Viral infections and molecular mimicry in type 1 diabetes. APMIS **120**, 941–949 (2012).

317. Wagenknecht, L. E., Roseman, J. M. & Herman, W. H. Increased Incidence of Insulin-Dependent Diabetes Mellitus Following an Epidemic of Coxsackievirus B5. Am. J. Epidemiol. **133**, 1024–1031 (1991).

318. Yeung, W.-C. G., Rawlinson, W. D. & Craig, M. E. Enterovirus infection and type 1 diabetes mellitus: systematic review and meta-analysis of observational molecular studies. BMJ **342**, d35–d35 (2011).

319. Härkönen, T., Lankinen, H., Davydova, B., Hovi, T. & Roivainen, M. Enterovirus infection can induce immune responses that cross-react with β-cell autoantigen tyrosine phosphatase IA-2/IAR: Enterovirus Infection and β-Cell Autoantigen Tyrosine Phosphatase. J. Med. Virol. **66**, 340–350 (2002).

320. Hiemstra, H. S. et al. Cytomegalovirus in autoimmunity: T cell crossreactivity to viral antigen and autoantigen glutamic acid decarboxylase. Proc. Natl. Acad. Sci. **98**, 3988–3991 (2001).

321. Honeyman, M. C., Stone, N. L., Falk, B. A., Nepom, G. & Harrison, L. C. Evidence for Molecular Mimicry between Human T Cell Epitopes in Rotavirus and Pancreatic Islet Autoantigens. J. Immunol. **184**, 2204–2210 (2010).

322. Wetzel, J. D. et al. Reovirus Delays Diabetes Onset but Does Not Prevent Insulitis in Nonobese Diabetic Mice. J. Virol. **80**, 3078–3082 (2006).

323. Tracy, S. et al. Toward Testing the Hypothesis that Group B Coxsackieviruses (CVB) Trigger Insulin-Dependent Diabetes: Inoculating Nonobese Diabetic Mice with CVB Markedly Lowers Diabetes Incidence. J. Virol. **76**, 12097–12111 (2002).

324. Ben-Smith, A., Gaston, J. S., Barber, P. C. & Winer, J. B. Isolation and characterisation of T lymphocytes from sural nerve biopsies in patients with Guillain-Barré syndrome and chronic inflammatory demyelinating polyneuropathy. J. Neurol. Neurosurg. Psychiatry **61**, 362–368 (1996).

325. Rodríguez, Y. et al. Guillain–Barré syndrome, transverse myelitis and infectious diseases. Cell. Mol. Immunol. **15**, 547–562 (2018).

326. Steininger, C. et al. Presence of Cytomegalovirus in Cerebrospinal Fluid of Patients with Guillain-Barré Syndrome. J. Infect. Dis. **189**, 984–989 (2004).

327. Kitazawa, K., Tagawa, Y., Honda, A. & Yuki, N. Guillain-Barré syndrome associated with IgG anti-GM1b antibody subsequent to Mycoplasma pneumoniae infection. J. Neurol. Sci. **156**, 99–101 (1998).

328. Shahrizaila, N. & Yuki, N. Guillain-Barré Syndrome Animal Model: The First Proof of Molecular Mimicry in Human Autoimmune Disorder. J. Biomed. Biotechnol. **2011**, 1–5 (2011).

329. Rees, J. H. & Hughes, R. A. C. Campylobacter jejuni Infection and Guillain–Barré Syndrome. N. Engl. J. Med. **333**, 6 (1995).

330. Ho, T. W. et al. Anti-GD1a antibody is associated with axonal but not demyelinating forms of Guillain-Barré syndrome. Ann. Neurol. Off. J. Am. Neurol. Assoc. Child Neurol. Soc. **45**, 168–173 (1999).

331. Yuki, N., Yoshino, H., Sato, S. & Miyatake, T. Acute axonal polyneuropathy associated with anti-GM1 antibodies following Campylobacter enteritis. Neurology **40**, 1900–1900 (1990).

332. Koga, M., Gilbert, M., Li, J. & Yuki, N. Complex of GM1-and GD1a-like lipo-oligosaccharide mimics GM1b, inducing anti-GM1b antibodies. PloS One **10**, e0124004 (2015).

333. Komagamine, T. & Yuki, N. Ganglioside mimicry as a cause of Guillain-Barre syndrome. CNS Neurol. Disord.-Drug Targets Former. Curr. Drug Targets-CNS Neurol. Disord. **5**, 391–400 (2006).

334. Moyano, A. L. et al. Validation of a rabbit model of neuropathy induced by immunization with gangliosides. J. Neurol. Sci. **272**, 110–114 (2008).

335. Tsokos, G. C., Lo, M. S., Reis, P. C. & Sullivan, K. E. New insights into the immunopathogenesis of systemic lupus erythematosus. Nat. Rev. Rheumatol. **12**, 716 (2016).

336. Barzilai, O., Ram, M. & Shoenfeld, Y. Viral infection can induce the production of autoantibodies. Curr. Opin. Rheumatol. **19**, 636–643 (2007).

337. James, J. A. et al. An increased prevalence of Epstein-Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. J. Clin. Invest. **100**, 3019–3026 (1997).

338. Poole, B. D., Scofield, R. H., Harley, J. B. & James, J. A. Epstein-Barr virus and molecular mimicry in systemic lupus erythematosus. Autoimmunity **39**, 63–70 (2006).

339. James, J. A. et al. Systemic lupus erythematosus in adults is associated with previous Epstein-Barr virus exposure. Arthritis Rheum. **44**, 1122–1126 (2001).

340. Vatti, A. et al. Original antigenic sin: A comprehensive review. J. Autoimmun. **83**, 12–21 (2017).

341. Hoskins, T. W. et al. INFLUENZA AT CHRIST'S HOSPITAL: MARCH, 1974. The Lancet **307**, 105–108 (1976).

342. Hoskins, T. W., Davies, JoanR., Smith, A. J., Miller, ChristineL. & Allchin, A. ASSESSMENT OF INACTIVATED INFLUENZA-A VACCINE AFTER THREE OUTBREAKS OF INFLUENZA A AT CHRIST'S HOSPITAL. The Lancet **313**, 33–35 (1979).

343. Gostic, K. M., Ambrose, M., Worobey, M. & Lloyd-Smith, J. O. Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. 6 (2016).

344. Viboud, C. & Epstein, S. L. First flu is forever. Science **354**, 706–707 (2016).

345. Klenerman, P. & Zinkernagel, R. M. Original antigenic sin impairs cytotoxic T lymphocyte responses to viruses bearing variant epitopes. Nature **394**, 482–485 (1998).

346. Adalja et al. (2010). Original Antigenic Sin and Pandemic (H1N1) 2009. Emerging infectious diseases, 16(6), 1028.

347. Kim, J. H., Skountzou, I., Compans, R. & Jacob, J. Original Antigenic Sin Responses to Influenza Viruses. J. Immunol. **183**, 3294–3301 (2009).

348. Nachbagauer, R. et al. Defining the antibody cross-reactome directed against the influenza virus surface glycoproteins. Nat. Immunol. **18**, 464–473 (2017).

349. Park, M. S., Kim, J. I., Park, S., Lee, I. & Park, M.-S. Original Antigenic Sin Response to RNA Viruses and Antiviral Immunity. Immune Netw. **16**, 261 (2016).

350. Haanen, J. B. A. G., Wolkers, M. C., Kruisbeek, A. M. & Schumacher, T. N. M. Selective Expansion of Cross-reactive CD8+ Memory T Cells by Viral Variants. J. exp. Med (1999).

351. Park, K. Y., Lee, M. G., Ryu, J. C. & Park, Y. K. Evolutionary stasis of M1 gene of human influenza A viruses and the possibility of their subtyping by restriction analysis of M1 gene polymerase chain reaction product. Acta Virol. **41**, 231–239 (1997).

352. Lawson, T. M. et al. Influenza A antigen exposure selects dominant Vβ17+ TCR in human CD8+ cytotoxic T cell responses. Int. Immunol. **13**, 1373–1381 (2001).

353. Cobey, S. & Hensley, S. E. Immune history and influenza virus susceptibility. Curr. Opin. Virol. **22**, 105–111 (2017).

354. Gershon, A. A. Is chickenpox so bad, what do we know about immunity to varicella zoster virus, and what does it tell us about the future? J. Infect. **74**, S27–S33 (2017).

355. Rickinson, A. B. & Kieff, E. Fields virology. Epstein-Barr Virus (1996).

356. Balfour, H. H. et al. Behavioral, Virologic, and Immunologic Factors Associated With Acquisition and Severity of Primary Epstein–Barr Virus Infection in University Students. J. Infect. Dis. **207**, 80–88 (2013).

357. Taylor, G. S., Long, H. M., Brooks, J. M., Rickinson, A. B. & Hislop, A. D. The Immunology of Epstein-Barr Virus–Induced Disease. Annu. Rev. Immunol. **33**, 787–821 (2015).

358. Silins, S. L. et al. Asymptomatic primary Epstein-Barr virus infection occurs in the absence of blood T-cell repertoire perturbations despite high levels of systemic viral load. Blood **98**, 3739–3744 (2001).

359. Aslan, N. et al. Severity of Acute Infectious Mononucleosis Correlates with Cross-Reactive Influenza CD8 T-Cell Receptor Repertoires. mBio **8**, (2017).

360. Basu, A. & Chaturvedi, U. C. Vascular endothelium: the battlefield of dengue viruses. FEMS Immunol. Med. Microbiol. **53**, 287–299 (2008).

361. Kurane, I. Dengue hemorrhagic fever with special emphasis on immunopathogenesis. Comp. Immunol. Microbiol. Infect. Dis. **30**, 329–340 (2007).

362. Guzmán, M. G. et al. Epidemiologic studies on Dengue in Santiago de Cuba, 1997. Am. J. Epidemiol. **152**, 793–799 (2000).

363. Sangkawibha, N. et al. Risk factors in dengue shock syndrome: a prospective epidemiologic study in Rayong, Thailand: I. The 1980 outbreak. Am. J. Epidemiol. **120**, 653–669 (1984).

364. Simmons, C. P. et al. Early T-Cell Responses to Dengue Virus Epitopes in Vietnamese Adults with Secondary Dengue Virus Infections. J. Virol. **79**, 5665–5675 (2005).

365. Weiskopf, D. et al. Human CD8 + T-Cell Responses Against the 4 Dengue Virus Serotypes Are Associated With Distinct Patterns of Protein Targets. J. Infect. Dis. **212**, 1743–1751 (2015).

366. Guzman, M. G., Alvarez, M. & Halstead, S. B. Secondary infection as a risk factor for dengue hemorrhagic fever/dengue shock syndrome: an historical perspective and role of antibody-dependent enhancement of infection. Arch. Virol. **158**, 1445–1459 (2013).

367. Mongkolsapaya, J. et al. Original antigenic sin and apoptosis in the pathogenesis of dengue hemorrhagic fever. Nat. Med. **9**, 921–927 (2003).

368. Zivna, I. et al. T cell responses to an HLA-B* 07-restricted epitope on the dengue NS3 protein correlate with disease severity. J. Immunol. **168**, 5959–5965 (2002).

369. Imrie, A. et al. Differential Functional Avidity of Dengue Virus-Specific T-Cell Clones for Variant Peptides Representing Heterologous and Previously Encountered Serotypes. J. Virol. **81**, 10081–10091 (2007).

370. Mangada, M. M. & Rothman, A. L. Altered Cytokine Responses of Dengue-Specific CD4 + T Cells to Heterologous Serotypes. J. Immunol. **175**, 2676–2683 (2005).

371. Rivino, L. T cell immunity to dengue virus and implications for vaccine design. Expert Rev. Vaccines **15**, 443–453 (2016).

372. Mathew, A., Kurane, I., Green, S. & Stephens, H. A. F. Vaughn. DW Kalayanarooj Suntayakorn Ennis FA Rothman A 3999–4004 (1998).

373. Rothman, A. L. Immunity to dengue virus: a tale of original antigenic sin and tropical cytokine storms. Nat. Rev. Immunol. **11**, 532–543 (2011).

374. Chaturvedi, U. C., Agarwal, R., Elbishbishi, E. A. & Mustafa, A. S. Cytokine cascade in dengue hemorrhagic fever: implications for pathogenesis. FEMS Immunol. Med. Microbiol. **28**, 183–188 (2000).

375. Heilman, C. A. From the National Institute of Allergy and Infectious Diseases and the World Health Organization: Respiratory Syncytial and Parainfluenza Viruses. J. Infect. Dis. **161**, 402–406 (1990).

376. Kapikian, A. Z., Mitchell, R. H., Chanock, R. M., Shvedoff, R. A. & Stewart, C. E. AN EPIDEMIOLOGIC STUDY OF ALTERED CLINICAL REACTIVITY TO RESPIRATORY SYNCYTIAL (RS) VIRUS INFECTION IN CHILDREN PREVIOUSLY VACCINATED WITH AN INACTIVATED RS VIRUS VACCINE. Am. J. Epidemiol. **89**, 405–421 (1969).

377. Kim, H. W. et al. RESPIRATORY SYNCYTIAL VIRUS DISEASE IN INFANTS DESPITE PRIOR ADMINISTRATION OF ANTIGENIC INACTIVATED VACCINE12. Am. J. Epidemiol. **89**, 422–434 (1969).

378. Maloy, K. J. et al. Cd4 $^{+}$ T Cell Subsets during Virus Infection: Protective Capacity Depends on Effector Cytokine Secretion and on Migratory Capability. J. Exp. Med. **191**, 2159–2170 (2000).

379. Connors, M. et al. Pulmonary Histopathology Induced by Respiratory Syncytial Virus (RSV) Challenge of Formalin-Inactivated RSV-Immunized BALB/c Mice Is Abrogated by Depletion of CD4+ T Cells. J VIROL **66**, 8 (1992).

380. Openshaw, P. J. M., Clarke, S. L. & Record, F. M. Pulmonary eosinophilic response to respiratory syncytial virus infection in mice sensitized to the major surface glycoprotein G. Int. Immunol. **4**, 493–500 (1992).

381. Varga, S. M., Wang, X., Welsh, R. M. & Braciale, T. J. Immunopathology in RSV Infection Is Mediated by a Discrete Oligoclonal Subset of Antigen-Specific CD4+ T Cells. Immunity **15**, 637–646 (2001).

382. Walzl, G., Tafuro, S., Moss, P., Openshaw, P. J. M. & Hussell, T. Influenza Virus Lung Infection Protects from Respiratory Syncytial Virus–Induced Immunopathology. J. Exp. Med. **192**, 1317–1326 (2000).

383. Klatzmann, D. ERC TRIPOD. https://cordis.europa.eu/project/id/322856/fr (2013).

384. Ganusov, V. V. & De Boer, R. J. Do most lymphocytes in humans really reside in the gut? Trends Immunol. **28**, 514–518 (2007).

385. Schenkel, J. M. & Masopust, D. Tissue-Resident Memory T Cells. Immunity **41**, 886–897 (2014).

386. Maecker, H. T., McCoy, J. P. & Nussenblatt, R. Standardizing immunophenotyping for the Human Immunology Project. Nat. Rev. Immunol. (2012) doi:10.1038/nri3158.

387. Mamedov, I. Z. et al. Preparing Unbiased T-Cell Receptor and Antibody cDNA Libraries for the Deep Next Generation Sequencing Profiling. Front. Immunol. **4**, (2013).

388. Meysman, P. et al. The workings and failings of clustering T-cell receptor beta-chain sequences without a known epitope preference. http://biorxiv.org/lookup/doi/10.1101/318360 (2018) doi:10.1101/318360.

389. 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.

390. 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).

391. Shugay, M. et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res. **46**, D419–D427 (2018).

392. 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.

393. 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).

394. Hayday, A. C. & Vantourout, P. The Innate Biologies of Adaptive Antigen Receptors. Annu. Rev. Immunol. **38**, annurev-immunol-102819-023144 (2020).

395. Routy, B. et al. The gut microbiota influences anticancer immunosurveillance and general health. Nat. Rev. Clin. Oncol. **15**, 382–396 (2018).

396. Roossinck, M. J. The good viruses: viral mutualistic symbioses. Nat. Rev. Microbiol. **9**, 99–108 (2011).

397. Gavin, M. A. & Bevan, M. J. Increased peptide promiscuity provides a rationale for the lack of N regions in the neonatal T cell repertoire. Immunity **3**, 793–800 (1995).

398. Huseby, E. S. et al. How the T Cell Repertoire Becomes Peptide and MHC Specific. Cell **122**, 247–260 (2005).

399. Murphy, K., & Weaver, C. Janeway's immunobiology. Garland Science.(2016). Q. Rev. Biol. **87**, 266–267 (2012).

400. Avrameas, S. Natural autoantibodies: from 'horror autotoxicus' to 'gnothi seauton'. 6 (1991).

401. Mogensen, T. H. Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses. Clin. Microbiol. Rev. **22**, 240–273 (2009).

402. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. Nat. Commun. **8**, 13954 (2017).

403. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature **409**, 860–921 (2001).

404. Mi, S. et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature **403**, 785–789 (2000).

405. Ting, C. N., Rosenberg, M. P., Snow, C. M., Samuelson, L. C. & Meisler, M. H. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev. **6**, 1457–1465 (1992).

406. Ryan, F. Virolution. (Collins, 2009).

407. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science **351**, 1083–1087 (2016).

408. Schiavetti, F., Thonnard, J., Colau, D., Boon, T. & Coulie, P. G. A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. Cancer Res. **62**, 5510–5516 (2002).

409. Silmon de Monerri, N. C. & Kim, K. Pathogens Hijack the Epigenome. Am. J. Pathol. **184**, 897–911 (2014).

410. Johannes, F., Colot, V. & Jansen, R. C. Epigenome dynamics: a quantitative genetics perspective. Nat. Rev. Genet. **9**, 883–890 (2008).

411. Gaschen, B. et al. Diversity considerations in HIV-1 vaccine selection. Science **296**, 2354–2360 (2002).

412. Pircher, H. et al. Viral escape by selection of cytotoxic T cell-resistant virus variants in vivo. Nature **346**, 629 (1990).

413. Choi, E. Y. et al. Thymocyte-Thymocyte Interaction for Efficient Positive Selection and Maturation of CD4 T Cells. Immunity **23**, 387–396 (2005).

414. Adamopoulou, E. et al. Exploring the MHC-peptide matrix of central tolerance in the human thymus. Nat. Commun. **4**, (2013).

415. Espinosa, G. et al. Peptides presented by HLA class I molecules in the human thymus. J. Proteomics **94**, 23–36 (2013).

416. Chowell, D. et al. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8 $^{+}$ T cell epitopes. Proc. Natl. Acad. Sci. **112**, E1754–E1762 (2015).

417. Pogorelyy, M. V. et al. Exploring the pre-immune landscape of antigen-specific T cells. Genome Med. **10**, (2018).

418. Pickett, B. E. et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. **40**, D593–D598 (2012).

Aucune source spécifiée dans le document actif.

Annexes

Article 1: "Human thymopoiesis selects unconventional CD8$^+$ α/β T cells that respond to multiple viruses." (Quiniou et al., Nature, in review).

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.

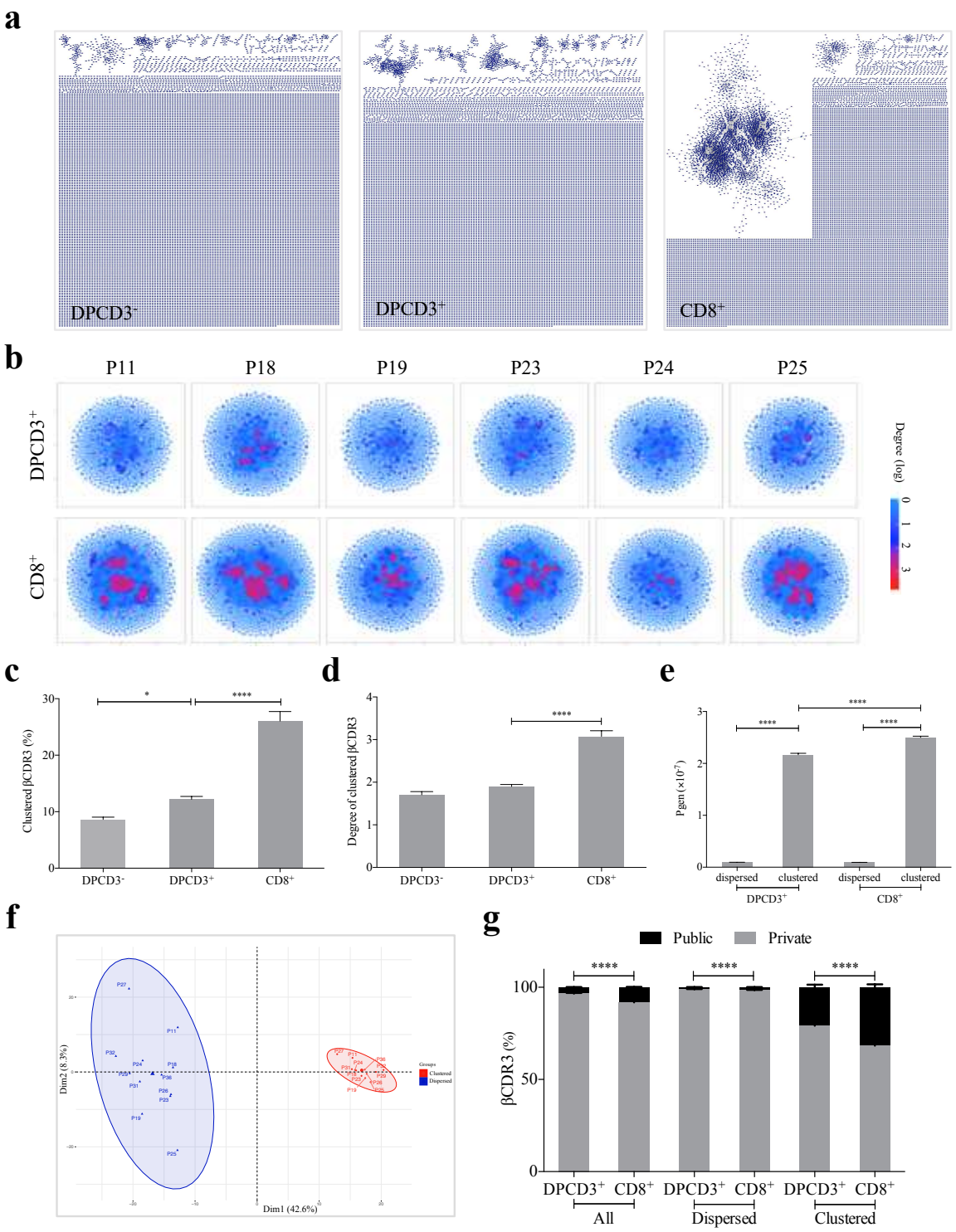


Figure 19. Thymocyte differentiation produces clustered and public CDR3s with high generation probability.

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.

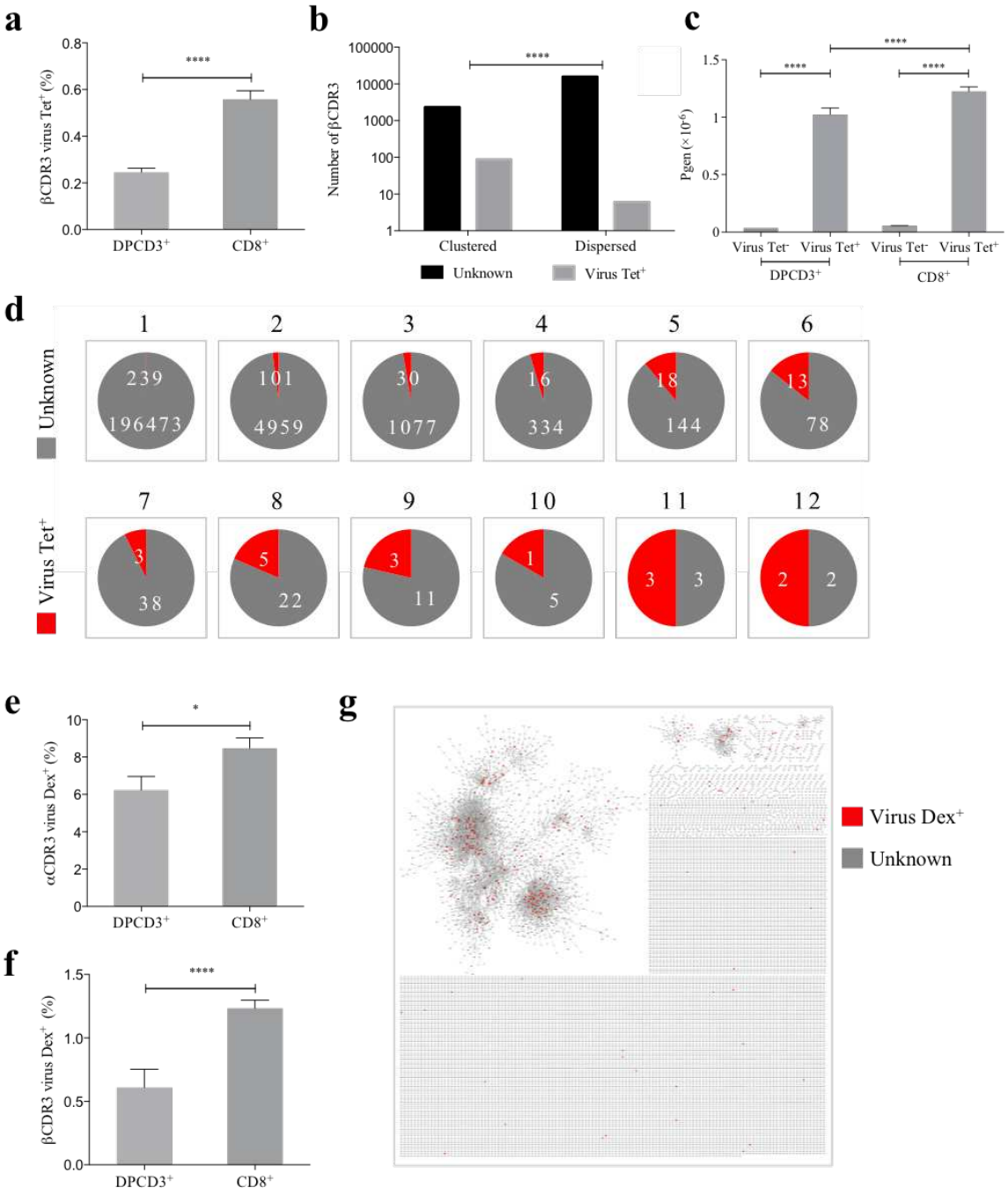


Figure 2. Clustered public TCRs are enriched for virus-specific TCRs.

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

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

We also analysed the binding of HLA-mismatched dextramers. When HLA-mismatch EBV Dex^+ cells are overlaid on a TSNE representation based on single-cell specificity (Fig. 3e), some of them corresponded to cells also labelled by HLA-matched EBV dextramers (Fig. 3a), while some others did not (Fig. 3e, red arrow). As for the HLA-matched dextramers, single TCRs were found to bind HLA-mismatched dextramers loaded with distinct unrelated peptides from EBV (Fig. 3f) and some TCRs even bound HLA-mismatched dextramers loaded with epitopes from different viruses such as CMV, HIV, HPV, HTLV-1 or influenza (Fig. 3g). There were also mostly positive correlations for the binding to different HLA-mismatched

dextramer specificities (Fig. 3h). Altogether, within a dataset of >160,000 single $CD8^+$ T cells, among the 66,191 that did bind dextramers, 24,083 could bind more than one viral-derived peptide from either the same or different viruses, and presented by HLA-matched or even HLA-mismatched 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].

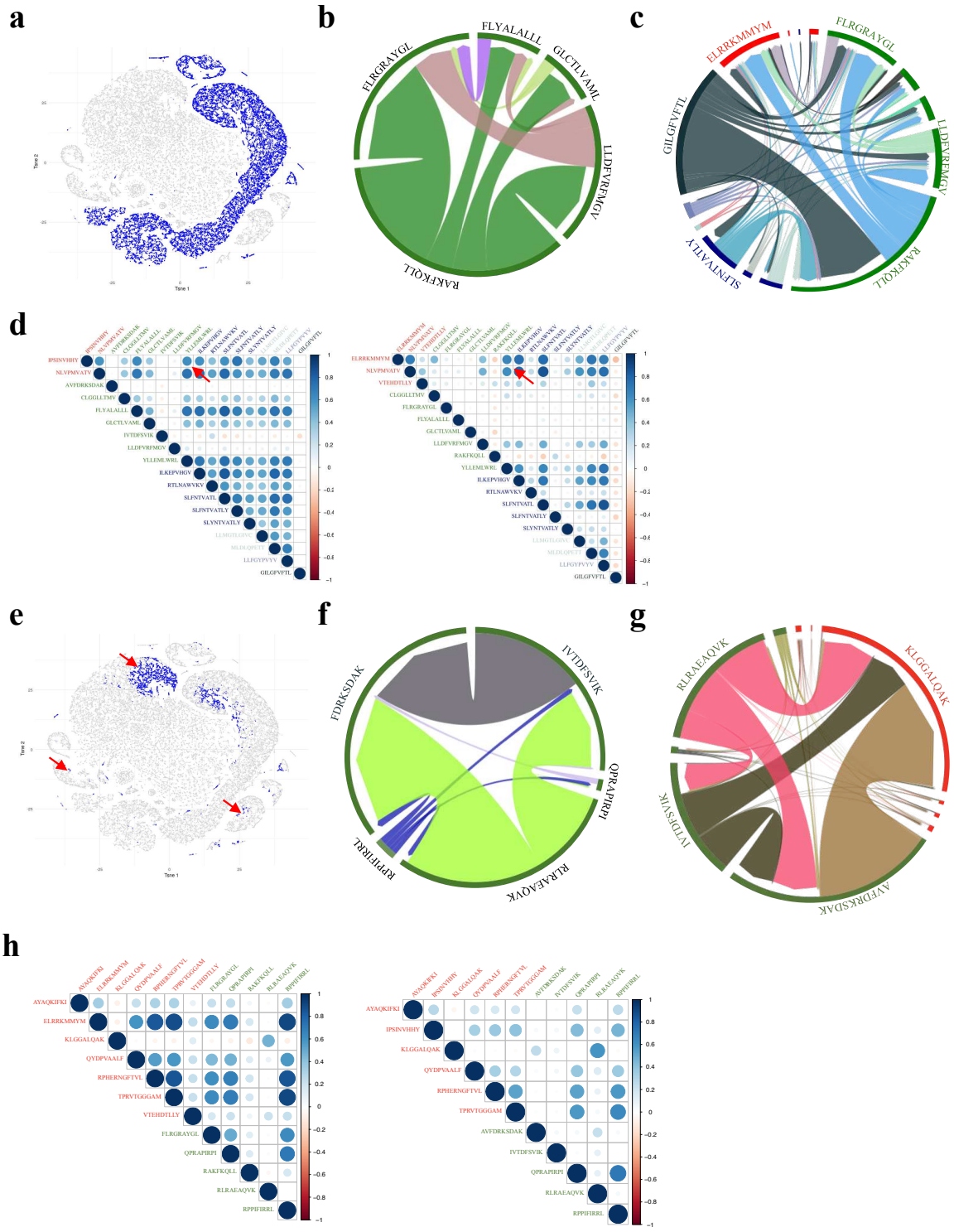


Figure 3. Innate-like TCR binding properties.

a. TSNE representation of the single-cell TCR specificities from one individual. HLA-matched EBV-Dex$^+$ cells are in blue. **b.** Chord diagram showing TCR binding to multiple HLA-matched EBV Dextramers loaded with different peptides. Each segment represents TCR binding to the peptides marked above. The size of the segments corresponds to the number of TCRs binding to these peptides. The link

between segments identifies multiple TCR binding to different peptides. **c.** Chord diagram showing TCR binding to HLA-matched dextramers loaded with peptides from unrelated viruses. The colours of the segments represent the different viruses: CMV (red), EBV (green), HIV (dark blue), HPV (light green), HTLV (purple) and influenza (dark grey). The same colour code is used in d, g and h. The full list of the different peptides is in Supplementary table 4. **d.** Correlation between the binding scores for the different HLA-matched virus-specific dextramers from 2 patients. Significant correlations (p-value < 0.01, Pearson correlation) are represented by coloured circles. The intensity of the colours and the size of the circles are proportional to the correlation coefficients. Positive correlations are displayed in blue and negative correlations in red. **e.** TSNE representation of the single-cell TCR specificities from one individual. HLA-mismatched EBV-Dex$^+$ cells are in blue. Red arrows highlight cells only labelled by HLA-mismatched EBV dextramers. **f.** Chord diagram showing TCR binding to HLA-mismatched dextramers loaded with peptides from EBV. **g.** Chord diagram showing TCR binding to HLA-mismatched dextramers loaded with peptides from unrelated viruses. The full list of the different peptides is in Supplementary table 5. **h.** Correlation between the binding scores for the different HLA-mismatched virus-specific dextramers from 2 patients.

These peculiar properties led us to assess the functional relevance of the innate-like recognition of psT cells' TCRs. We first evaluated the in vitro cross-activation of T cells with different peptides. Human effector memory CD8$^+$ T cells (emCD8$^+$) were purified according to their binding of CMV or EBV fluorescent dextramers; the sorted cells were then stimulated by either the peptide that was used to purify them, or by different ones, and their activation was measured by their IFNγ production (Fig. 4a). All T cells were efficiently non-specifically activated by PMA/ionomycin. T cells that did not bind any dextramers could not be stimulated by any peptide. In contrast, dextramer-sorted cells could be activated by their cognate peptide and almost as well by the other (Fig. 4b). Thus, the binding of multiple dextramers appears functionally relevant, translating into proper psT cells activation by multiple unrelated peptides.

We then investigated whether we could detect the involvement of psT cells during *in vivo* immune responses. We first analysed individuals vaccinated against yellow fever (YFV) (Fig. 4c)[20] or influenza (Flu) (Fig. 4d), identifying their TCR repertoires responding to YFV and Flu using the ALICE[21] and TCRNET algorithms[22,23], respectively. Within these repertoires, we looked for known viral-specific TCR sequences. Besides YFV- and Flu-specific β and α CDR3s (Fig. 4c & d), we could also detect CDR3s of psT cells, i.e. assigned to one or even multiple

other viral specificities, notably for CMV and EBV. We then also analysed the TCR repertoire of T cells from bronchoalveolar lavages (BAL) from patients with COVID-19 pulmonary infections[24], i.e. cells responding to the local infection. We observed that many of these BAL T cells have the characteristics of psT cells: their CDR3s (i) could be detected within the clusters of psT cells responding to YFV and Flu infection (Fig. 4 c & d); (ii) they are highly connected to virus-specific sequences from databases (Fig. 4 e & f) and (iii) many harbour a TCR assigned to at least two specificities from CMV, EBV or Flu. Thus, psT cells migrate to the sites of primary antiviral immune responses.

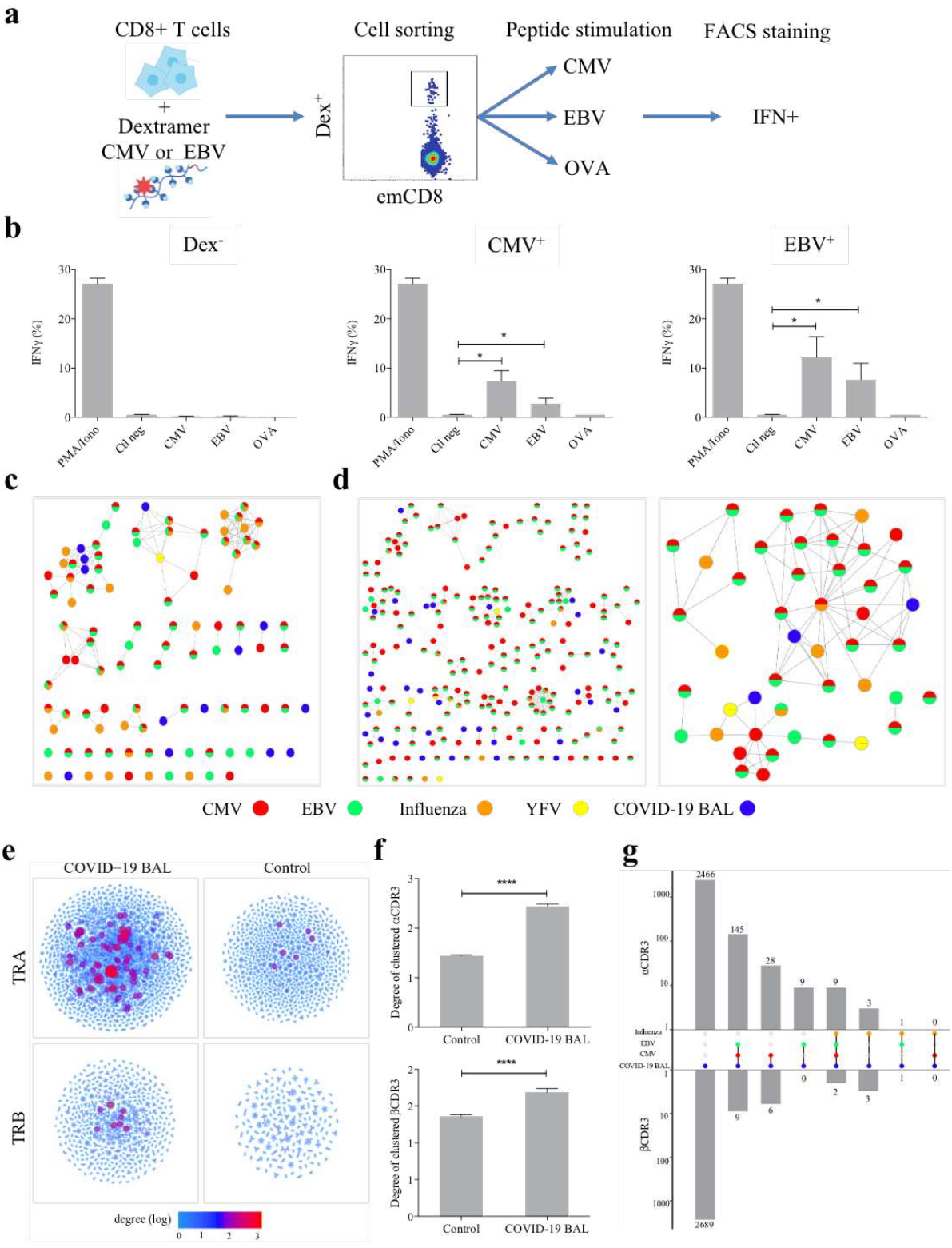


Figure 4. In vitro activation and in vivo recruitment of psT cells. a. Schematic representation of the *in vitro* cross-activation experiment. **b.** In vitro activation of innate-like T cells. Percentage of IFNγ producing emCD8^{+} cells after activation with PMA/ionomycin (positive control), or CMV, 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 another indication that T cell recognition could have been selected to be more fuzzy than stringent. Further studies will have to evaluate the contribution of fuzzy immune responses to the efficacy but also the immunopathology of antimicrobial responses and to autoimmunity.

Materials and methods:

Human samples:

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

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

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

Isolation of thymocytes and extraction of RNA:

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

TCR repertoire library preparation and sequencing

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

as well as a forward primer hybridising to the SMART-Seqv4 sequence added by template-switching and a second PCR targeting the PCR1 amplicons with reverse and forward primers including Illumina Indexes allowing for sample barcoding. PCR2 were then purified using AMPure beads (Beckman-Coulter®). The cDNA samples were quantified and their integrity was checked using DNA electrophoresis performed on an Agilent 2100 Bioanalyzer System in combination with the Agilent DNA 1000 kit, according to the manufacturer's protocol. Sequencing was performed with Hiseq 2500 (Illumina®) SR-300 protocols using the LIGAN-PM Genomics platform (Lille, France).

TCR deep sequencing data processing

FASTQ raw data files were processed for TRB sequence annotation using MiXCR[37] software (v2.1.10) with RNA-Seq parameters. MiXCR extracts TRBs and provides corrections of PCR and sequencing errors.

Network generation and representation

To construct a network, we computed a distance matrix of pairwise Levenshtein distances between CDR3s using the "stringdist"[38] R package. When two sequences were similar under the defined threshold, LD>1 (i.e., at most one amino acid difference), they were connected and designated as "clustered" nodes. CDR3s with more than one amino acid difference from any other sequences are not connected and were designated as "dispersed" nodes.

Layout of networks for Fig. 1b and Supplementary Fig. 9a were obtained by using the graphopt algorithm of the "Igraph"[39] R package and plotted in 2D with "ggplot2" to generate figures[40]. Only clustered nodes are represented, edges are not shown and colours represent the node degree (log scale).

Layouts of detailed networks in Fig. 1a, Fig. 2g, Fig. 4c, Fig. 4d & Supplementary Fig. 1a, Supplementary Fig. 3 & Supplementary Fig. 12 were done with Cytoscape[41].

Statistical analysis and visualisation

Normalisation was performed by sampling on the top α or β 18,000 CDR3s based on their frequency in each sample. The repertoires with less than 18,000 α or β CDR3s were not included in the statistical analysis. The numbers of samples included in the statistical analysis for the β repertoire were: two for $DPCD3^-$, ten for $DPCD3^+$ and twelve for $CD8^+$. The numbers

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

Probability of generation calculation

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

CDR3 connections between individuals

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

Virus-specific βCDR3 tetramer public databases

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

Virus-specific CDR3 single-cell dextramer public dataset

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

threshold defining positive binding at UMI counts greater than 10 for any given dextramer. This identified 15,195 unique virus-specific TCRs with at least one binding.

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

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

Cross-activation experiment

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

Influenza vaccination protocol

Peripheral blood was obtained before vaccination and on day 14 after vaccination. PBMCs were stained with antibodies anti-CD3 (AF700), anti-CD8 (FITC), CD45RA (PeCy7), CCR7 (BV421). Cells were sorted by fluorescent activated cell sorting (Becton Dickinson™ FACSAria

II) with purity >95% to collect populations based on the following labelling: naïve CD8 were gated as $CD3^+CD8^-CD45RA^+CCR7^+$ and effector memory CD8 were gated as $CD3^+CD8^-CD45RA^-CCR7^-$. RNA extraction, library preparation, sequencing and raw data processing were performed as described above.

Identification of TCR enrichment after vaccination

Homologous groups of TCRs that are specifically recruited during an antigen-specific response following yellow fever or influenza vaccination were sought using with the previously published algorithms ALICE[21] and TCRNET[22,23], respectively. The difference between the two algorithms is that ALICE uses the VDJ rearrangement model as a control[45] while TCRNET uses real control samples as background.

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).

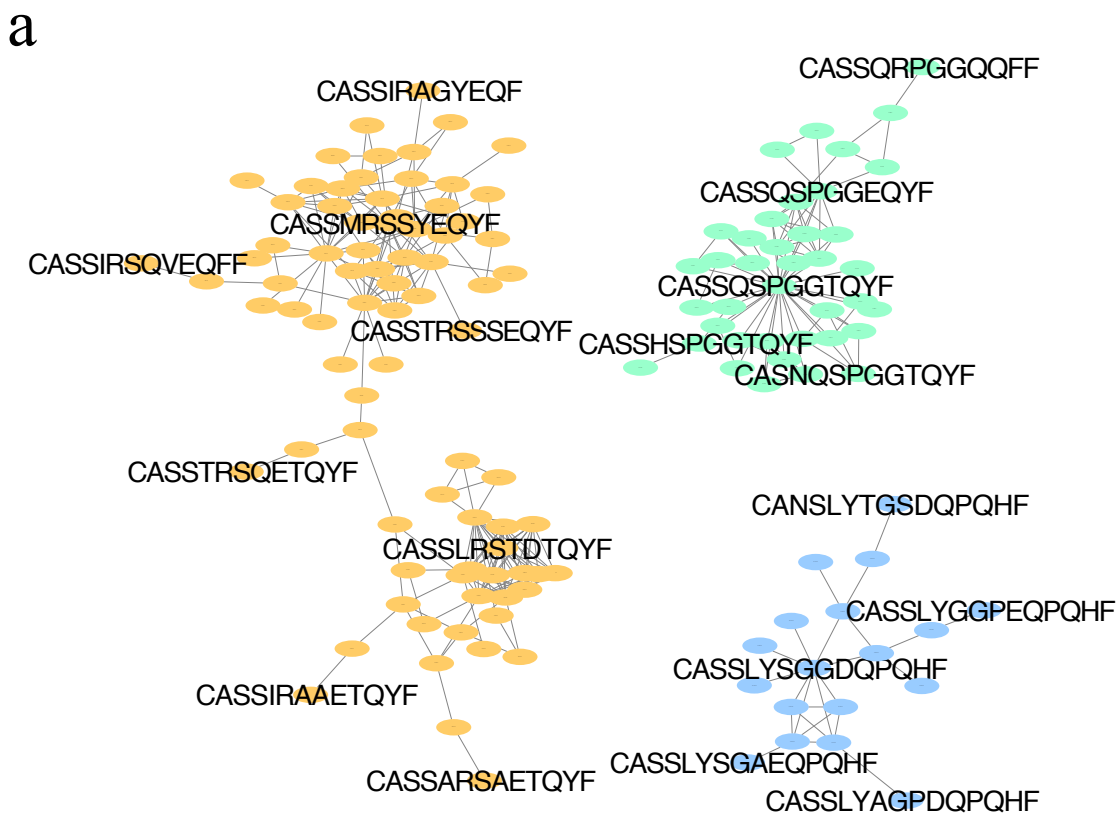
41. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

42. Krijthe, J. H. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R Package Version 013 URL Httpsgithub ComjkrijtheRtsne* (2015).

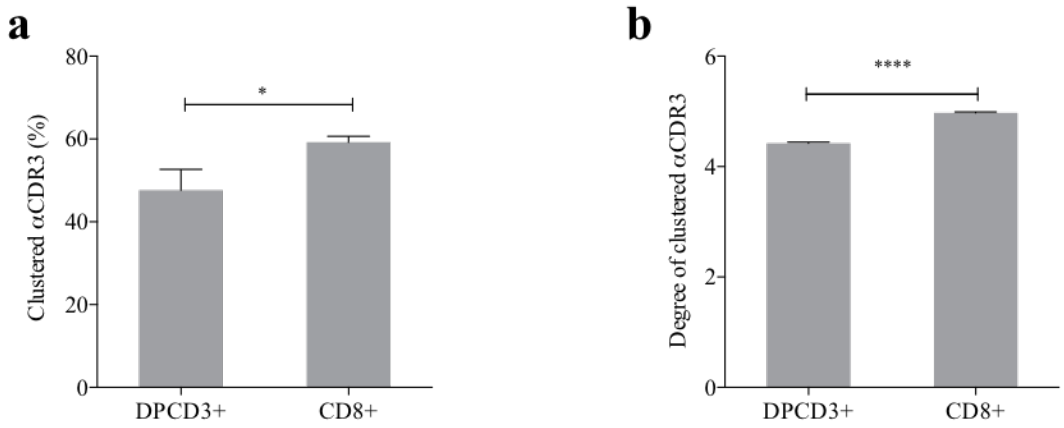
43. Sethna, Z., Elhanati, Y., Jr, C. G. C. & Mora, T. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. 8.

44. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, (2018).

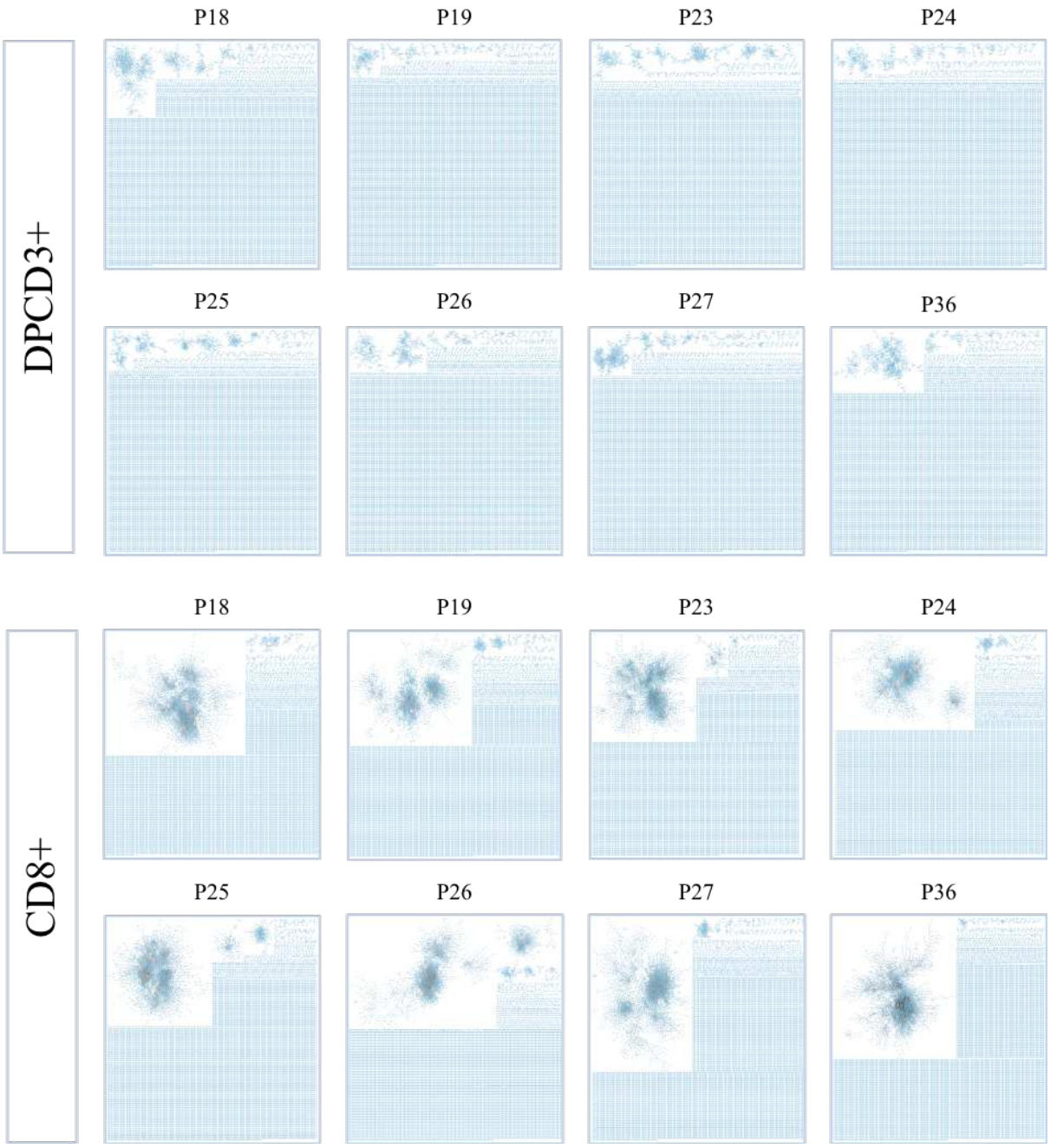
45. Murugan, A., Mora, T., Walczak, A. M. & Callan, C. G. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci.* **109**, 16161–16166 (2012).



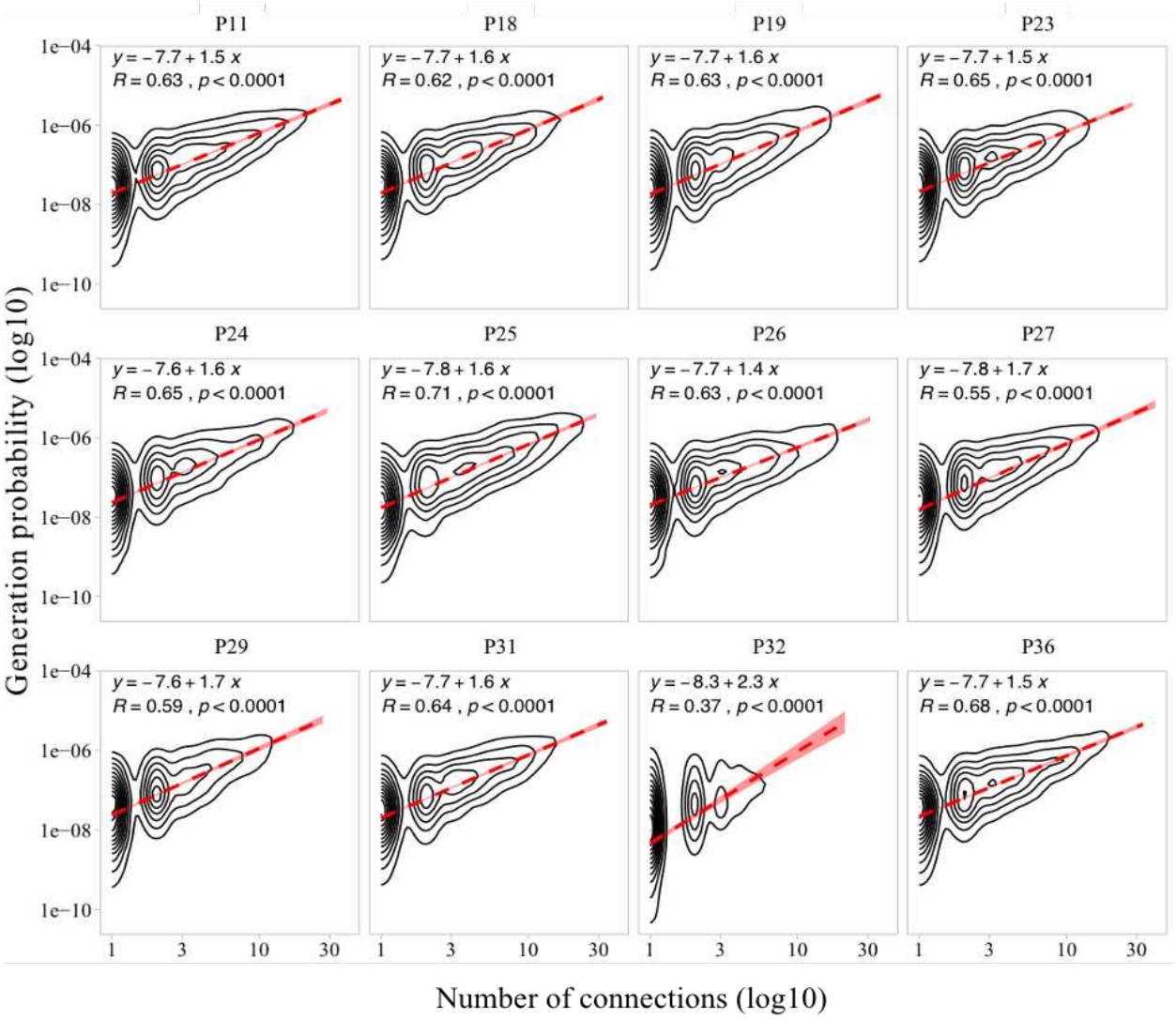
Supplementary Figure 1. TCRs with the same specificity form clusters. a. Networks of βCDR3s specific for GILGFVFTL from influenza (orange), GLCTLVAML from Epstein-Barr virus (green) and FPRPWLHGL from human immunodeficiency virus (blue) are shown. These βCDR3s are from TCRs identified on CD8 T lymphocytes 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).



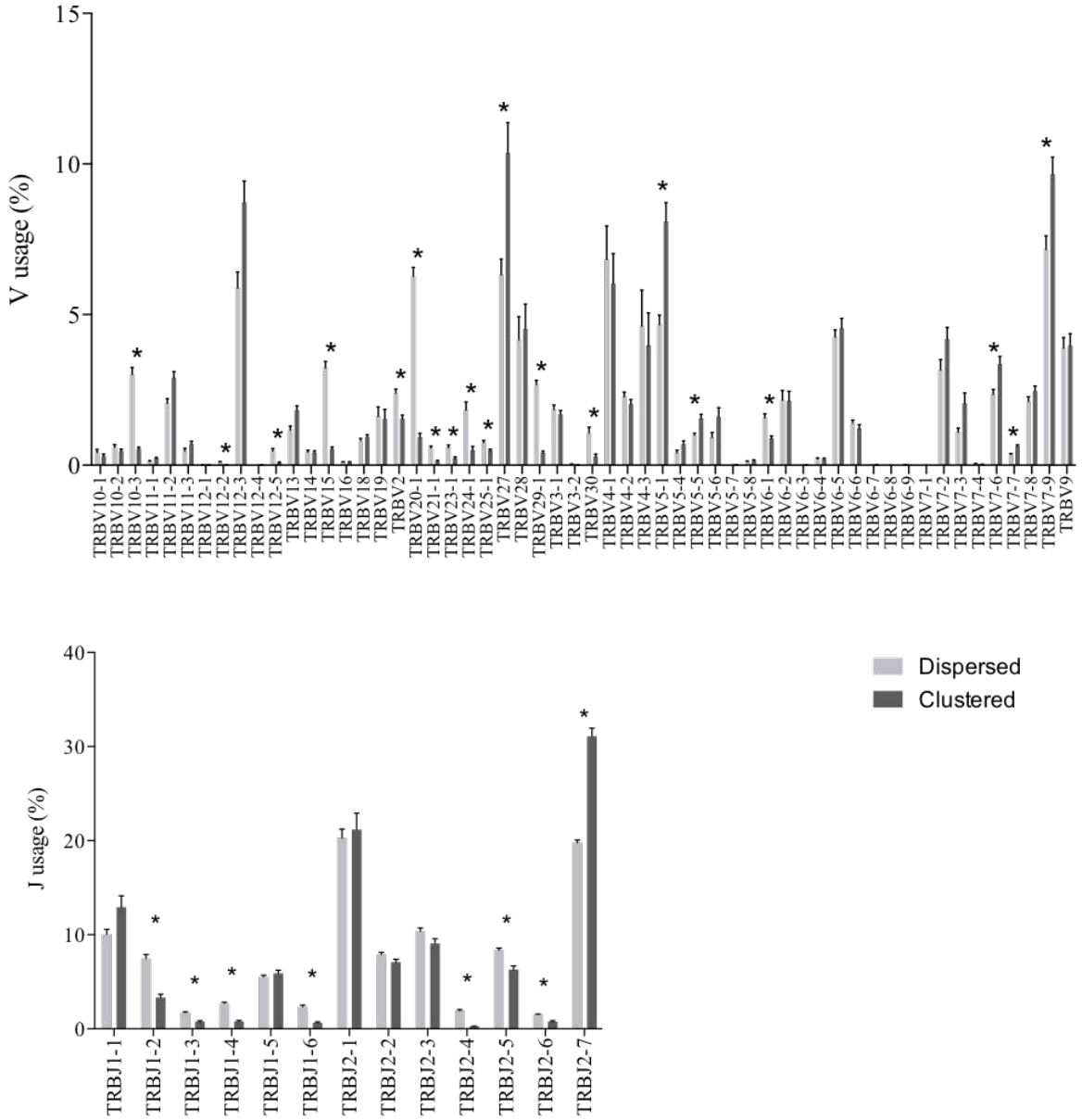
Supplementary Figure 2. Clustered αCDR3s from DPCD3$^+$ and CD8$^+$ thymocytes. Analyses were performed on the first 18,000 most frequent αCDR3s per sample (n=6 for DPCD3$^+$ and n=10 for CD8$^+$ thymocytes). **a.** Percentage of clustered αCDR3s. (mean ± s.e.m., *p=0.016, Mann-Whitney test). **b.**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



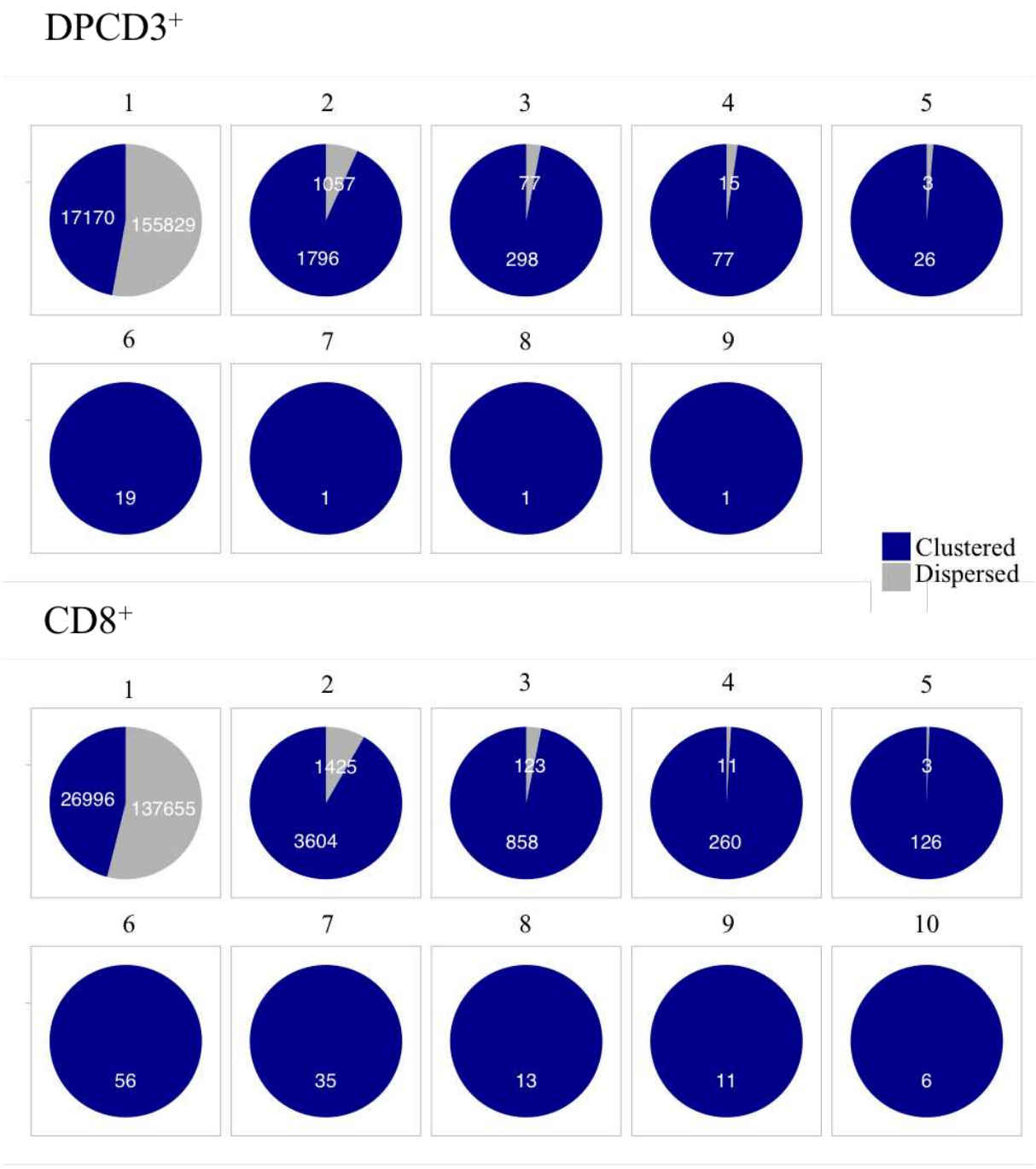
Supplementary Figure 3. βCDR3 network during thymopoiesis. Representation of the 18,000 most frequent βCDR3 networks from DPCD3^{+} and CD8^{+} thymocytes of eight donors (Pn).



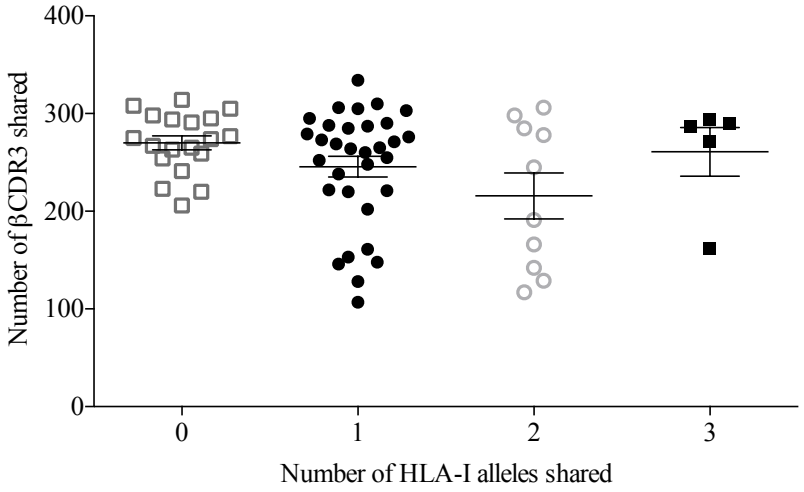
Supplementary Figure 4. Correlation between Pgen and βCDR3 number of connections in the CD8⁺ thymocyte repertoire. The contour plots represent the generation probability as a function of βCDR3 connections in the CD8⁺ thymocytes for donors P11 to P36. Linear regression curves between Pgen and number of 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.



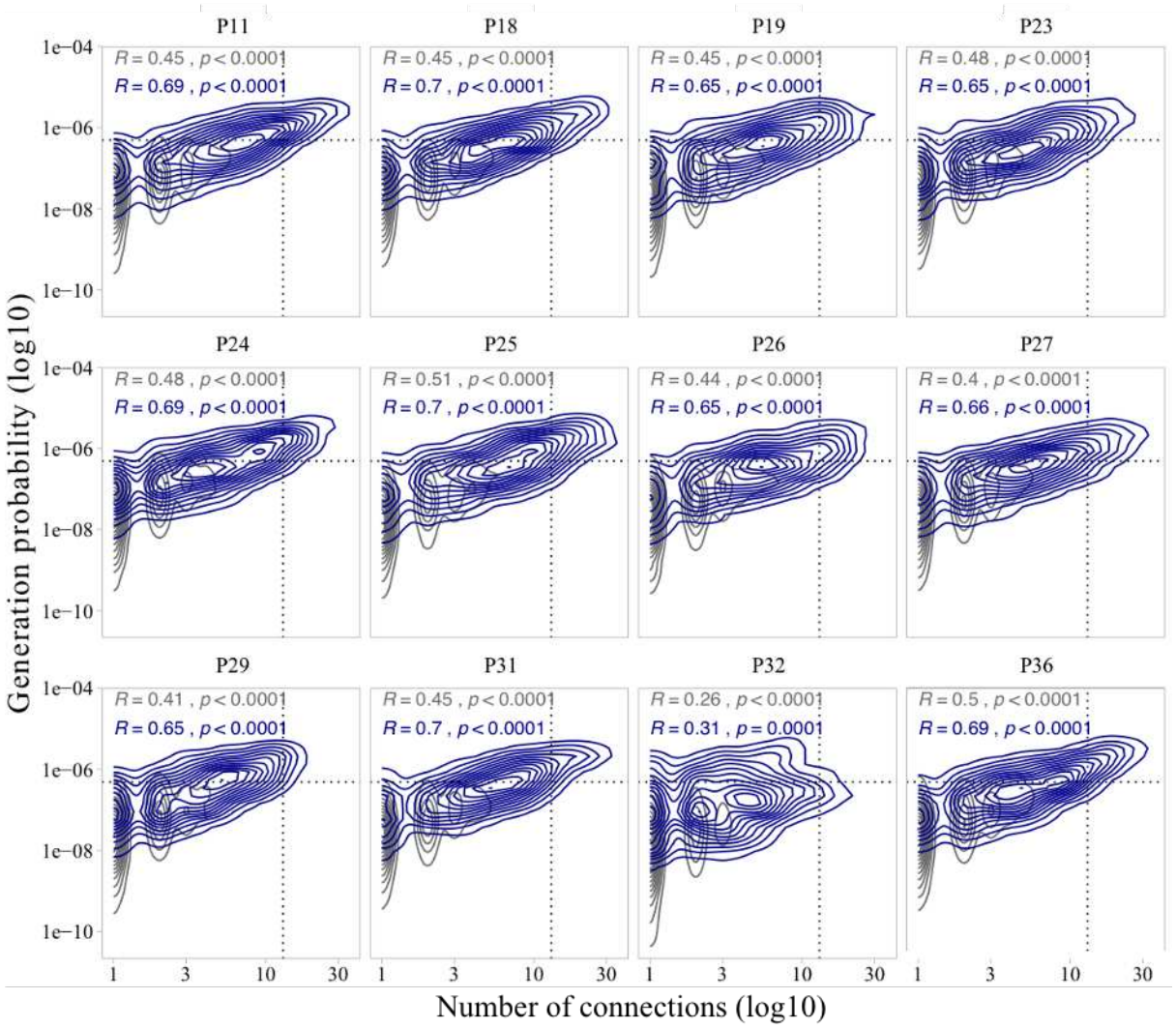
Supplementary Figure 5. Clonogram representation of TCR Vβ and Jβ usage in clustered versus dispersed CD8⁺ thymocytes. The bar plots represent the mean percentage of TCR Vβ (up) and Jβ (down) in dispersed (light grey) versus clustered (dark grey). (* p<0.01, multiple t-test).



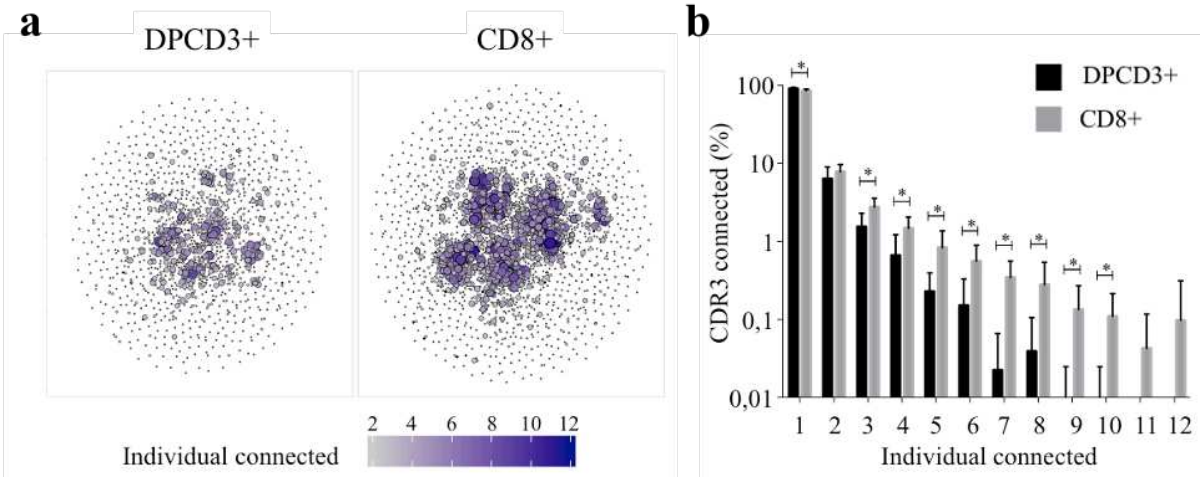
Supplementary Figure 6. βCDR3 sharing between individuals. Pie charts represent the sharing between individuals before (DPCD3$^+$) and after thymic selection (CD8$^+$). Colours represent the dispersed (grey) or clustered (blue) CDR3s. Sharing was analyzed within the 10 donors for which there were at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



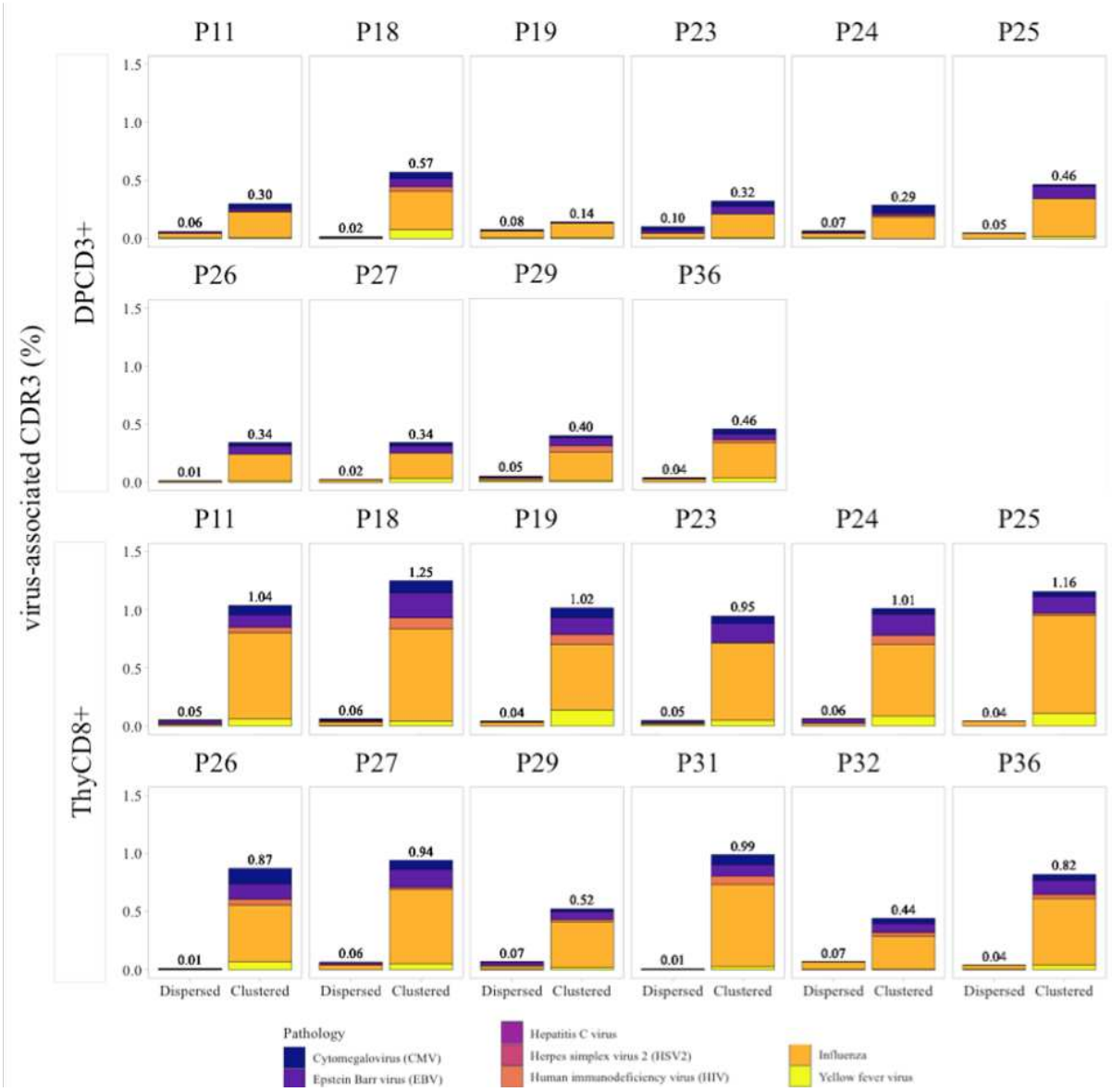
Supplementary Figure 7. The number of public βCDR3s in CD8⁺ thymocytes is independent of the number of HLA-I alleles shared. Each dot represents the number of βCDR3s shared between two donors in the first 18,000 CD8⁺ thymocytes. There is no significant difference in the number of public βCDR3s according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



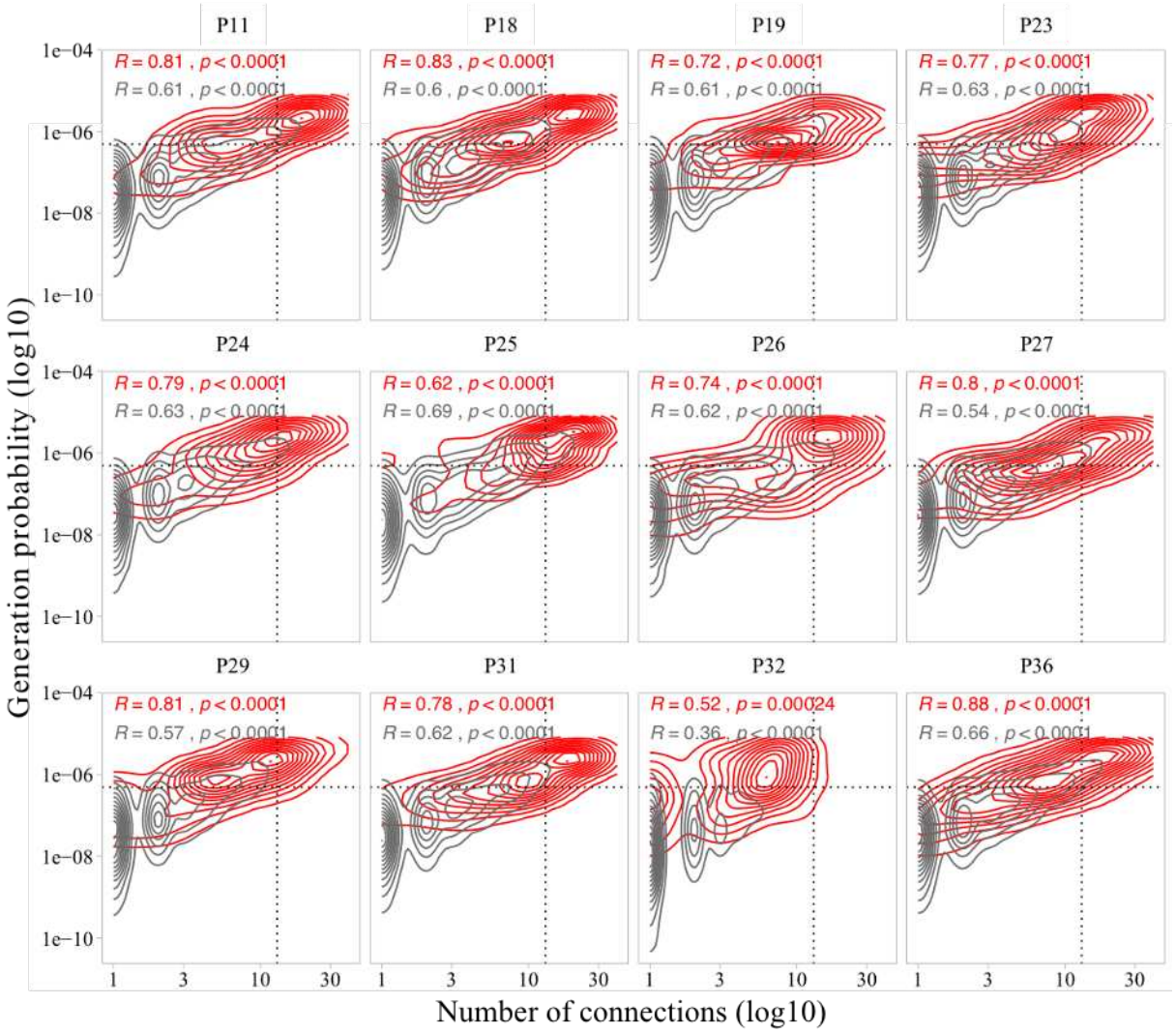
Supplementary Figure 8. Enrichment of public βCDR3s in the CD8⁺ thymocyte repertoires. Representation of the generation probability as a function of βCDR3 connections in individuals (Pn). The contour plots represent shared (blue) or private (grey) βCDR3s. The Pearson correlation coefficient "R" and p-value "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).



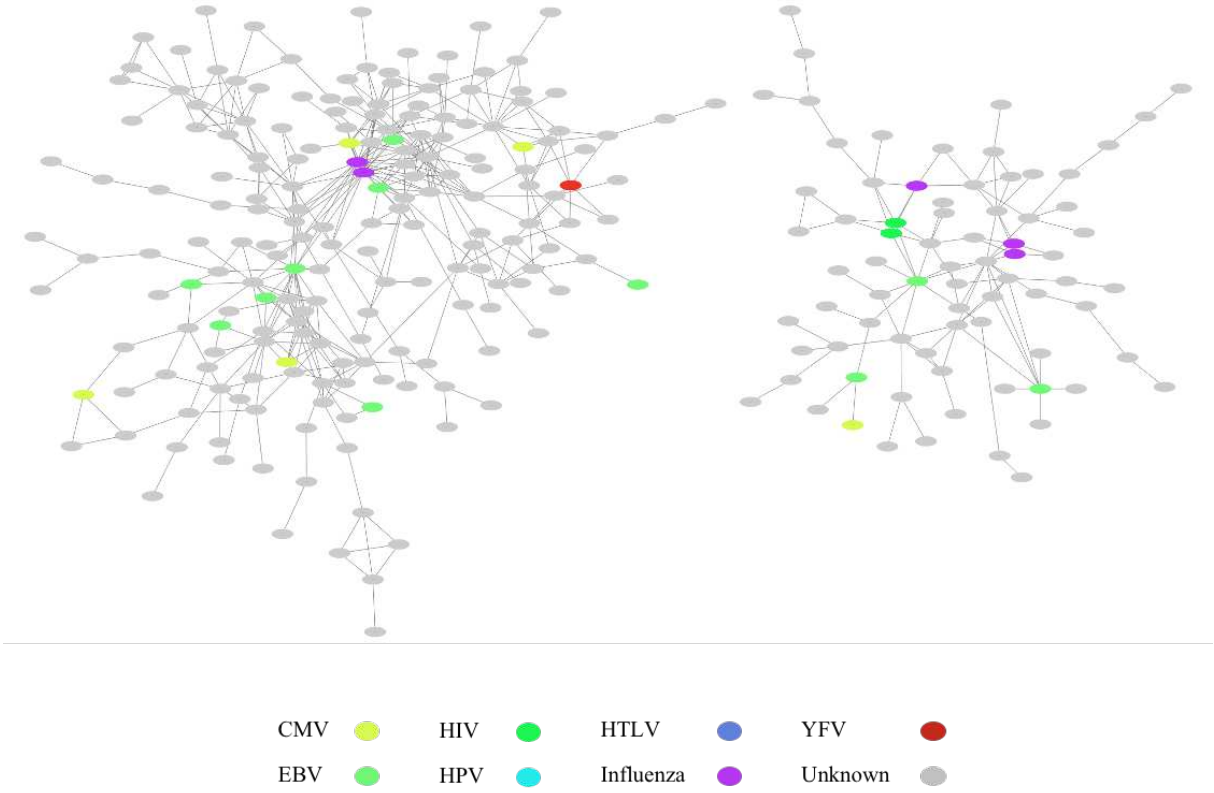
Supplementary Figure 9. Convergence of public βCDR3 specificities during thymopoiesis. a. CDR3 connections between individuals. The top 1,500 βCDR3s were sampled from DPCD3$^+$ (left) and CD8$^+$ (right) cells from each individual and pooled. The CDR3s are clustered based on LD≤1 with colour and size 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).



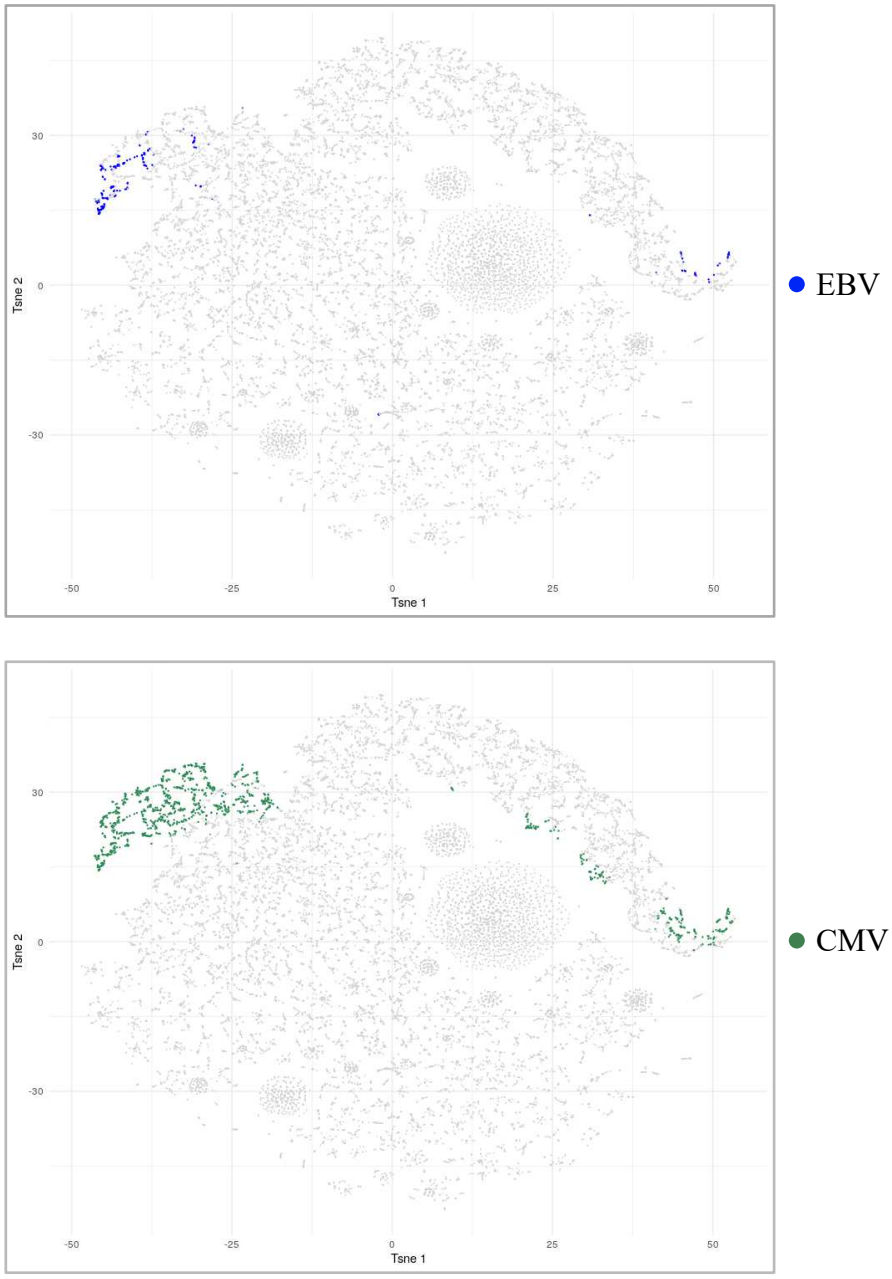
Supplementary Figure 10. Virus-specific CDR3s among DPCD3^{+} and CD8^{+} thymocytes. Bar plots represent the percentage, within the top 18,000 βCDR3s of each donors, of βCDR3s from TCRs identified as virus-specific based on tetramer identification[14,15]. For each panel, the percentage is calculated 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.



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

		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.

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

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

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

This work was primarily funded by the TRiPoD ERC-Advanced EU (322856) grant to DK, and by the LabEx Transimmunom (ANR-11-IDEX-0004-02) and RHU iMAP (ANR-16-RHUS-0001) grants.

Corresponding author

Correspondence to Prof. David Klatzmann

Immunology-Immunopathology-Immunotherapy Laboratory (i3) and Clinical Investigation Center for Biotherapies (CIC-BTi)

Pitié-Salpêtrière Hospital, 83 boulevard de l'Hôpital, F-75013, Paris, France.

E-mail: david.klatzmann@sorbonne-universite.fr

Article 2: "Benchmarking of T-cell receptor repertoire profiling reveals large systematic biases." (Barennes et al., Nature Biotechnology)

Benchmarking of T-cell receptor repertoire profiling methods reveals large systematic biases

Pierre Barennes[1,2], Valentin Quiniou[1,2], Mikhail Shugay[3,4,5,6], Evgeniy S. Egorov[4], Alexey N. Davydov[6], Dmitriy M. Chudakov[3,4,5,6], Imran Uddin[7], Mazlina Ismail[7], Theres Oakes[7], Benny Chain[7], Anne Eugster[8], Karl Kashofer[9], Peter P. Rainer[10], Samuel Darko[11], Amy Ransier[11], Daniel C. Douek[11], David Klatzmann[1,2], Encarnita Mariotti-Ferrandiz[1]*

1. Sorbonne Université, INSERM, Immunology-Immunopathology-Immunotherapy (i3), Paris, France
2. AP-HP, Hôpital Pitié-Salpêtrière, Biotherapy (CIC-BTi) and Inflammation-Immunopathology-Biotherapy Department (i2B), Paris, France
3. Center of Life Sciences, Skoltech, Moscow, Russia
4. Genomics of Adaptive Immunity Department, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia
5. Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Russian National Research Medical University, Moscow, Russia
6. Adaptive Immunity Group, Central European Institute of Technology, Brno, Czechia
7. Division of Infection and Immunity, University College London, United Kingdom
8. DFG-Centre for Regenerative Therapies Dresden, Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Fetscherstrasse 105, 01307 Dresden, Germany
9. Diagnostic and Research Institute of Pathology, Medical University of Graz, Graz, Austria
10. Division of Cardiology, Medical University of Graz, Graz, Austria
11. Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, United States

*Corresponding author:

Encarnita Mariotti-Ferrandiz, PhD

encarnita.mariotti@sorbonne-universite.fr

Monitoring the T-cell receptor (TCR) repertoire in health and disease can provide key insights into adaptive immune responses, but the accuracy of current TCR sequencing (TCRseq) methods is unclear. Here, we systematically compared the results of nine commercial and academic TCRseq methods, including six RACE-PCR and three multiplex-PCR approaches, when applied to the same T-cell sample. We found marked differences in accuracy and intra- and inter-method reproducibility for alpha (TRA) and beta (TRB) TCR

chains. Most methods showed a lower ability to capture TRA than TRB diversity. Low RNA input generated non-representative repertoires. Results from the 5'RACE-PCR methods were consistent among themselves, but differed from the RNA-based multiplex-PCR results. Using an in silico meta-repertoire generated from 108 replicates, we found that one gDNA-based method and two non-UMI RNA-based methods were more sensitive than UMI methods in detecting rare clonotypes, despite the better clonotype quantification accuracy of the latter.

Main

Sequencing of the TCR repertoire is increasingly used to measure lymphocyte dynamics in health, in pathological contexts such as autoimmune disease, infections and cancer[1–6], and following interventions such as vaccination[4,7–9]and immunotherapy[10–12], with the goal of identifying TCR biomarkers of disease or of clinical response to treatment and to stratify patients for precision medicine[13]. These diverse applications have different requirements in terms of sensitivity, specificity and depth. Accurately capturing the TCR repertoire therefore presents great challenges. . Challenges are reproducibility, replicability and sensitivity since the diversity of the repertoire is by essence unknown when studying a new sample.

T-cell receptors, which drive T-cell activation by antigenic peptide recognition, are heterodimers formed by an α and a β chain[14] produced by somatic V(D)J rearrangements during thymopoiesis[15] of 47V and 61J functional TRA genes and 48V, 2D, 12J functional TRB genes[16]. The stochastic V(D)J recombination generates a combinatorial diversity that is further increased by random nucleotide excision and addition at the V(D)J junctions. The independent recombination and subsequent pairing of TRA and TRB chains add an additional level of combinatorial diversity. Recently, computational chain pairing experiments suggested that the potential diversity of the paired repertoire is $\sim 2 \times 10^{19}$ TCRs[17], while the

number of different TRB clonotypes in an individual has been estimated to range from 10^6 to 10^8 [18–20]. The TCR repertoire is dynamic, as lymphocytes are continuously generated, die and expand in response to stimulation, and reflects both an individual's immune potential and history.

A large number of TCRseq methods have been developed. They are all complex multistep protocols, and each step may have a profound impact on the TCRseq data and hence on their interpretation[21]. Methods can be broadly classified as DNA- or RNA-based, and the latter can be categorized as using multiplex PCR (mPCR) with panels of V and J primers[18,22,23] or using rapid amplification of cDNA-ends by PCR (RACE-PCR)[6,24–26] optionally incorporating unique molecular identifiers (UMI) to limit PCR amplification bias and sequencing errors[6,26–28]. Each method has potential advantages and limitations[24,29–32]. Specifically, DNA-based methods are believed to be more quantitative and can be used on samples where RNA quality may not be guaranteed. In contrast, RNA-based methods are believed to be more sensitive because of the presence of multiple mRNA copies per cell, and also are more amenable to UMI incorporation[33]. In addition, although TCR germ line gene variability between individuals has not been extensively explored, the requirement of efficient primers for mPCR might also be an additional limitation. However, the relative robustness and accuracy of the different approaches have not been systematically compared. Here, we compared 9 different TCRseq library preparation protocols by analyzing the TCR repertoire of aliquots of the same T-cell sample. Nine protocols for TCR library preparation were selected according to at least one the following criteria: used by groups other than the one who developed it [6,18,22,24–26], (ii) their association with well-known analysis tools[23,27,34] and (iii) commercially available [18,22,26,35]. Our results clearly show that (i) RACE-PCR and mPCR

perform differently in TCR clonotype detection, (ii) few of the evaluated methods accurately capture the TRA diversity and (iii) UMI-based methods are likely appropriate to study major clonotypes with a high accuracy while robust non-UMI based methods allows to capture a larger repertoire.

RESULTS

Experimental design to evaluate the robustness of human T-cell receptor repertoire analysis

We set out to compare 9 different academic or commercial protocols for library preparation and sequencing (**Supplementary material and methods**; **Supplementary Table 1**) based either on RACE-PCR (RACE-1 to RACE-6) or on multiplex-PCR (mPCR-1 to 3). We sequenced nucleic acids from $CD4^+CD25^-CD127^+$ effector T-cells (**Supplementary Fig.1a**) sorted from two healthy donors (experiments A&B). In experiment A, we evaluated the accuracy and sensitivity of the different methods by spiking donor A T-cell RNA (RACE-1 to RACE-6 and mPCR-3) or DNA (mPCR-1 and mPCR-2) aliquots with different amounts of RNA or DNA from Jurkat cells (**Supplementary Fig.1b**). In experiment B, we analyzed the impact of decreasing amounts of the input material quantity by processing donor B RNA aliquots of 100 ng and 10 ng (**Supplementary Fig.1c**)**.** In both experiments, the $CD4^+CD25^-CD127^+$ T-cells were sorted, and the RNA and DNA were extracted and aliquoted in a single laboratory. Triplicates of aliquots were distributed to service providers and academic laboratories. Raw and/or pre-filtered sequences data were all processed using MiXCR[36].

We obtained from 5.10^5 to 2.10^6 reads per aliquot depending on the method (**Supplementary Fig.2a-b**). Numbers of unique V, J and VJ sequences as well as UMI distribution for RACE-1 and RACE-2 (**Supplementary Fig.2a-c**) were comparable between all

the methods. Numbers of TCR sequences and clonotypes were correlated in a method-dependent manner, but not globally, suggesting that the sequencing depth required for a given number of clonotypes is method-dependent (**Supplementary Fig.2d**).

Replicability and reproducibility differ among methods

For each method, we first analyzed the proportion of reads that were identified as TCRs (**Fig.1a and Supplementary Fig.2**). For 7/9 methods, we observed 20 to 60% of non-aligned reads, which were mainly explained by no V and/or J sequence identification. TCR sequences had a high-quality score (phred > 30, **Fig.1b**) and contained less than 1% PCR errors (**Fig.1c**), except for RACE-2, RACE-6, mPCR-2 and mPCR-3. Note that these parameters could not be assessed for one of the commercialized mPCR-1 for which undisclosed proprietary pre-processing of the data is performed.

Using a VDJ rearrangement model (Methods), we computed 17 rearrangement parameters for TRA and TRB sequences from experiments A&B (**Supplementary Fig.3**) and calculated Jensen-Shannon Divergence (JSD) distances between samples per parameter. Multi-Dimensional Scaling (MDS, **Fig.1d**) showed that, within each experiment, samples obtained with the same method clustered together, suggesting that each method imposed its methodological imprint on the repertoire profile.

We further compared the different library methods' replicability (i.e. the similarity among data obtained with the same method) and reproducibility (i.e. the similarity among data obtained with different methods) using JSD as a measure of the distance between datasets[37]. **Figure 1e** showed that for TRB, both the replicability and reproducibility of RACE-6 and mPCR-2 are lower than for all the other methods tested. However, when considering TRA, replicability is higher for RACE-3 and RACE-5 and reproducibility is higher for RACE-3,

RACE-5 and RACE-2 (with and without UMI). Since RACE-6 showed extremely low replicability for TRB samples and was not reproduced by any other methods, we excluded it from further analysis. Altogether, our results showed that many fundamental parameters of the TCR repertoire, as well as inter-sample replicability and reproducibility, vary between the different methods tested.

The observed TRBV gene usage varies between RACE- and multiplex-PCR RNA-based methods.

We compared the TRBV usage obtained from the sequencing data with the percentage of TRBV protein expression quantified by flow cytometry (FC) (**Fig.2a and Supplementary Figs.4a-b**). mPCR-1 data were highly correlated with FC data (**Fig.2b**, R^2>0.9, $P < 5.10^{-12}$), which likely reflects the undisclosed proprietary filtering by the provider. All other methods also showed a significant R^2 Pearson correlation score ranging from 0.4 to 0.8, $P < 0.05$) with TRBV protein expression (**Fig.2a-b)**, except for mPCR-3 (R^2<0.2, $P > 0.05$). The Pearson correlation of TRBV gene usage within replicates prepared with the same method (**Fig.2c**) was high (R^2>0.9). However, clustering showed that mPCR-3 formed a distinct cluster with a low correlation score (R^2<0.5) with other methods. The RACE methods data were highly correlated between each other (R^2>0.8), except RACE-1 and RACE-1_U, which had a lower correlation (0.6<R^2<0.7). mPCR-1 and mPCR-2 formed an independent "DNA cluster" with an R^2>0.6 when compared to RACE replicates and a low correlation with mPCR-3 (R^2<0.4). This low correlation with mPCR-3 could in part be explained by a skewed TRBV9, TRBV29-1 and TRBV20-1 usage (**Supplementary Fig.4c**). Spearman correlation scores were higher between FC data and mPCR-3 as well as RACE-1, and globally between the methods (**Supplementary**

Fig.4d-e). In summary, RACE-PCR methods and gDNA-based mPCR methods showed comparable TRBV usage results, in contrast with the mPCR-3 RNA based method.

Robustness of TRA and TRB detection is method-dependent

We compared the similarity and composition of the 1% most predominant clonotypes (1%_MPC) detected by each method. The Morisita-Horn similarity index (MH) was calculated for each replicate across all the methods for both TRA (**Fig**.**3a-left**) and TRB sequences (**Fig**.**3a-right**). TRA repertoires from RACE-3 and RACE-5 clustered together, inter- and intra-replicates having a high degree of similarity (MH≈0.8). RACE-1, RACE-2 and RACE-4 have a lower inter- and intra-method similarity (0.2<MH<0.5), but a higher similarity with RACE-3 and RACE-5. Comparable clustering was obtained with the Jaccard similarity index (JSI), a measure independent of clonotype frequency (**Supplementary Fig.5a**). For the TRB repertoires, MH scores were low when comparing RACE and mPCR protocols (MH≈0.36), but high within the RACE cluster (0.6>MH>0.9). There was poor similarity between the results of the three mPCR methods, regardless of the template. Differences between RACE and mPCR methods disappeared when calculating the JSI, suggesting a bias in clonotype frequency, as expected when comparing RNA- with DNA-based methods, but less when comparing RNA-based methods. Similar results were obtained by iteratively increasing the percentage of clonotypes (**Supplementary Fig.5b**). Rényi diversity profiles (**Supplementary Fig.5c**) showed comparable results for TRB with all the methods, but the diversity of TRA varied depending on the method. However, the potential diversity estimated using Chao extrapolation was variable between methods (**Supplementary Fig.5d**).

To test a possible bias in capturing the TRA diversity for some methods, we pooled and compared the three spiking replicates per method from experiment A, as suggested by Greiff

et al.[21]. The MH similarity significantly increased for all the RACE-based methods for TRA (**Fig.3b-top**) (except RACE-3) and for TRB (**Fig.3b-bottom**), with the TRA MH similarity remaining lower than that of TRB. Similar observations were made for mPCR replicates. This suggests that for a given depth of sequencing, the TRB diversity is better captured than that of TRA.

Detection sensitivity of rare TCRs depends on the method

To determine the accuracy of the different library amplifications for different clonotype frequencies, we compared the observed frequencies of the TCR from the Jurkat spike-in to their theoretical frequencies of 1/10, 1/100 and 1/1000. (**Supplementary Fig.1b**). TRA observed frequencies were on average 3 times lower than expected (**Fig.4a-top**; **Supplementary Table 2 and Supplementary Fig.6a**). In contrast, TRB frequencies were on average 3 times higher than the theoretical percentage, except for mPCR-1 (**Fig.4a-bottom**; **Supplementary Table 2 and Supplementary Fig.6a**). For most of the methods, except RACE-1_U, RACE-4 and mPCR-3, the ratio between the different dilutions was maintained, as shown by the mean slope values close to 1 (**Fig.4b**).

We then compared the inter-sample variation in clonal frequency for those TCR sequences shared between all replicates of an individual method (excluding the Jurkat clone). **Figure 4c** represents the standard deviation of the frequency of each shared clonotype (dots) per method (see details in **Supplementary Fig.6b-d**). For TRA, RACE-3 and RACE-5 had the highest number of clonotypes shared between the 9 replicates and the lowest standard deviation. For TRB, all the methods captured a high number of shared clonotypes, and mPCR-1 and RACE-3 had the lowest standard deviation. Finally, pooling all the clonotypes from all the replicates, we identified 9 TRA and 31 TRB clonotypes shared by all the

replicates of all methods, corresponding to the most predominant clonotypes (**Supplementary Fig.7**). RACE-3, RACE-5 (both RNA-based) and mPCR-1 (DNA_based) showed the lowest inter-sample variability in TCR frequency.

The quantity of starting material impacts TCR diversity capture

One major limitation when analyzing TCR repertoire is the number of T-cells that can be analyzed. Focusing on 4 RNA-based methods, we analyzed the influence of input RNA quantity on TRA and TRB repertoires (**Supplementary Fig.1c**). We compared two sets of samples, one containing 10 ng or 100 ng (corresponding to 10^4 and 10^5 cells, respectively). For all the methods, the richness was higher with large (100 ng) than small (10 ng) samples (**Supplementary Fig.8a**). Rényi diversity profiles (**Supplementary Fig.8b**) showed that when alpha < 2 (i.e. when the diversity metric is influenced by rare clones), the diversity of small samples is less than that of larger ones. In contrast, at alpha = 2 (Simpson index) or above, diversity profiles of both samples overlap. Thus, a low RNA input influences the number of rare TCR sequences detected, but not the distribution of the more abundant TCRs.

Finally, we evaluated the inter-sample similarity as a function of RNA input quantity by calculating the MH index with either the TRVJ combination usage (VJ_usage), all clonotype frequencies (Overall), or with the frequencies of the 1% most predominant clonotype (1%_MPC) (**Supplementary Fig.8c-middle**). For TRA, the similarity between 10 ng replicates was lower at the level of VJ usage and of all clonotypes compared with that between 100 ng replicates (**Supplementary Fig.8c-top&bottom**). For TRB, the results were comparable regardless of the quantity (MH>0.5). When focusing on the 1% MPC, the similarity was comparable regardless of the quantity for both TRA and TRB. These results indicated that RNA quantity impacts rare clonotype detection.

Reliability and sensitivity of each method highlighted using an in silico meta-repertoire

One unavoidable issue when aiming at capturing the diversity of a repertoire is sampling, i.e. only a fraction of the cells are analyzed and then a fraction of their nucleic acids[21]. To better assess the ability of each method robustly to capture rare and frequent clonotypes, we took advantage of the fact that altogether we generated 45 TRA and 63 TRB replicates of the same cell sample. We aggregated these results to generate an in silico meta-repertoire. To ensure the accuracy of the TCR sequences composing this meta-repertoire, we removed singletons and kept clonotypes found by at least 3 methods.

We first analyzed how many of the clonotypes present in this meta-repertoire were detected by each method. For TRA (**Fig.5a-left**), RACE-3 and RACE-5 datasets included up to 50% of the meta-repertoire clonotypes (MRC) compared to 10 to 20% for the other RACE method datasets. Similar results were found for TRB (**Fig.5a-right**). We then computed for each method the fraction of MRC found in 0, 1, 2, 3 etc. up to 9 replicates. The dot-heatmaps (**Fig.5b**) showed that for TRA, RACE-3 and RACE-5 clearly outperformed the other methods, capturing up to 40% of the MRC in all 9 replicates (**Fig.5b-left**; Replicate number=9) and missing (i.e. never captured in any of the 9 replicates) less than 1% of the MRC (**Fig.5b-left**; Replicate number=0). The other RACE protocols detected only 1% of MRC in all 9 replicates and missed 15 to 20% of the MRC (**Fig.5b-left**). In contrast, there was much less difference between the methods for TRB (**Fig.5b-right).**

Finally, we analyzed the frequency of MRC TCRs that were detected or not by each method (**Fig.5c and Supplementary Fig.9**). For TRA (**Fig.5c-left**), the frequency of MRC found in 9 replicates (red boxplots) ranged from 1% to 0.001% for RACE-3 and RACE-5 and from 1% to 0.05% for the other methods. In contrast, clonotypes not detected in any replicates (black

boxplots) were present at 10- to 100-fold lower abundance. A similar overall pattern was seen for TRB, although the frequencies were shifted to a lower range. This analysis suggested that RACE-3 and RACE-5 had increased sensitivity, and hence were able to detect a larger proportion of clonotypes at lower abundances. These differences were more evident for TRA than for TRB (**Fig.5c-right)**. The other methods compared behaved very similarly to each other. Importantly, those results were independent of sample size (**Supplementary Fig.10**).

DISCUSSION

Interpreting the TCR repertoire is an increasingly important tool in understanding the underlying causes of immune-mediated diseases and in assisting the development of new immunotherapeutic strategies. However, despite hundreds of TCRseq studies in the last decade using a variety of different methodologies, there has been no systematic study comparing them.

In this work, we compared methods developed by academics, at a time when there was little or no reliable commercial service provision, with some currently available commercial methods. Both RNA- and gDNA-based methods were included. To avoid mis-implementation of protocols, each method (including appropriate pre-processing of sequence data) was performed by the laboratory or commercial provider (except for kit providers) that developed them.

Unexpectedly, some consistent differences were observed in TRBV usage when compared to FC measurement of TRBV-encoded proteins, especially for RNA-based profiling. This might reflect bias in amplification of RNA transcripts according to their expression levels, more

efficient transcription of some V genes, or differences in nonsense-mediated decay[38]. Further studies, using single-cell RNAseq may shed light on this phenomenon.

Working with human samples often imposes limits on the number of available T-cells. Notably, lymphopenia is a common feature in people undergoing treatment (transplantation, immunosuppressive therapy) or with autoimmune disease[39] and infections. Additionally, T-cell subsets of interest, as well as available counts of tumor-infiltrating T-cells, may be limited. Therefore, it is important to identify which methods provide reliable TCRseq profiles for small numbers of T-cells. In this context, we observed that, regardless of the method, starting from a highly polyclonal population, the initial amount of material is critical to obtaining representative results, notably in terms of diversity and rare clone detection.

Although our study focused on polyclonal CD4 T-cells from healthy repertoires, we analyzed a wide range of global and sequence-specific repertoire parameters, including V(D)J gene usage, junctional diversity, repertoire diversity and sequence sharing. These parameters are all relevant to any other alpha/beta T-cell populations, as indeed are all parameters routinely used to analyze repertoires of samples from pathological and clinical human samples[40].

Because our study incorporated multiple replicates tested with each method, we were able to explore method replicability, i.e. the ability of each method to reproduce the same repertoire from different sub-samples from the same individual. Our results showed that, except mPCR-3, all the methods provided consistent results among replicates. We also evaluated the reproducibility, i.e. the extent to which different methods record the same results when applied to the same sample. We observed a low degree of TRB clonotype overlap between repertoires amplified from gDNA and RNA (cDNA), perhaps reflecting differences in gDNA and RNA copy numbers. The four RACE methods produced relatively

similar repertoires as revealed by the Morisita-Horn index. The mPCR on gDNA showed low reproducibility between methods, suggesting that the choice of multiplexing primers might bias the amplification of some clonotypes, as suggested previously[31]. However, most RACE methods (not tested for mPCR) had a lower efficiency in capturing TRA rather than TRB diversity, which may reflect the 2- to 3-fold lower number of TRA transcripts than TRB transcripts[28].

Finally, sensitivity is important for the study of circulating blood T-cells, especially when the goal is to track a few expanded clones associated with infection or autoimmunity, or in response to treatment. However, assessing sensitivity based on sample overlap is a complex performance metric, since it is impacted by experimental variability, but also by sampling. In order to tackle this problem directly, we generated an in silico meta-repertoire which provided a more robust platform with which to directly compare the sensitivity performance of the different methods. Interestingly, using this standard, we found that two non-UMI methods (RACE-3 and RACE-5) had greater sensitivity than UMI-based methods (RACE-1 and RACE-2) and were able to detect clonotypes at a 10-fold lower frequency. In part, this results from the reads-per-UMI cutoff, which may lead to a decrease in observed TCR diversity if sequencing coverage is not sufficient. For example, introducing a hard cutoff which discards all UMIs with less than 5 reads leads to a decrease in observed TCR diversity. UMI-based methods may be more accurate for assessing clonotype frequency, in line with their use to quantify and correct for PCR errors and bias[41]. Furthermore, a threshold of 2-4 reads per UMI efficiently protects against artefacts and cross-sample contamination[42], which becomes critical with tighter cluster density on modern Illumina machines. UMI-based methods may require several replicates or higher sequencing coverage to consistently and unambiguously

identify rare TCR sequence clonotypes. Noteworthy, both RACE-1 and RACE-2 methods performed better after UMI correction (see **Table 1**).

TR chain	Method	Replicability	Reliability	Sensitivity	Cost per sample	Controls & standards	Format type	fastq data availability
TRA	**RACE-1**	7	4	4	~230	-	lab protocol	YES
	RACE-1_U	4	5	4	~230	UMI	lab protocol	YES
	RACE-2	5	4	5	230-280	-	service or kit	YES
	RACE-2_U	4	5	5	230-280	UMI	service or kit	YES
	RACE-3	3	2	3	~150	-	kit	YES
	RACE-4	5	6	4	~150	-	lab protocol	YES
	RACE-5	2	3	3	~300	-	lab protocol	YES
TRB	**mPCR-1**	3	3	3	~350-550*	synthetic TCRs	service or kit	NO
	mPCR-2	6	7	7	~25	-	lab protocol	YES
	mPCR-3	5	5	3	~350-550*	-	service or kit	YES
	RACE-1	6	5	4	~230	-	lab protocol	YES
	RACE-1_U	4	6	5	~230	UMI	lab protocol	YES
	RACE-2	6	6	6	230-280	-	service or kit	YES
	RACE-2_U	6	6	7	230-280	UMI	service or kit	YES
	RACE-3	2	2	3	~150	-	kit	YES
	RACE-4	3	5	4	~150	-	lab protocol	YES

Table 1: Comparative performance of the nine TCRseq molecular methods. For each method, an average rank score for TRA (top) and TRB (bottom) sequencing has been calculated for Replicability, Reliability, and Sensitivity (three first column) and practical information have been summarized (4 last columns). Ranks have been calculated as the average of the ranks for results from Fig. 1e, 2c, 3b, 4c for "Replicability"; Fig. 1e, 2b, 4b, 5a, 5b for "Reliability"; Fig. 4c, 5b & Supplementary Fig. 2a, 5c for "Sensitivity". Rank values are comprised between 2 (best) and 7 (worst) and represented as bars with their values. Details are provided as Supplementary information. Cost per sample" is expressed in USD as per current prices for a depth of 1 million TCR sequences per sample on a 25 million reads sequencing format. The costs cover reagents for library preparation to sequencing. *mPCR1 and mPCR3 price ranges correspond to the cost for either purchasing kits (lowest price) or service up to sequencing and basic data analyses from the provider.

Such in silico standards may be of value in further comparative TCRseq method evaluation, although ideally synthetic repertoires recapitulating at least the extent of the TRAVJ and TRBVJ combinations and distributions may provide an even more robust alternative. Two such approaches have been proposed for specific clone detection in Minimal Residual Diseases[43,44] as well as for the BCR, but not TCR, repertoire[45], still at a very low diversity level. The construction of such gold standard repertoires is currently very costly and remains a major challenge that the Adaptive Immune Receptor Repertoire Community (AIRR-C)[46],

engaged in AIRR-seq standardization[47,48], may tackle in the future. Finally, in this study some data were pre-processed using proprietary (mPCR-1, mPCR-3) or published[27,34] (RACE-1_U and RACE-2_U) tools and then aligned and error-corrected using MiXCR (v2.1.10)[36]. To further optimize TCR data accuracy, it would also be interesting to benchmark available software analysis tools, especially regarding UMI analysis and sequence alignment. Our datasets generated using different methods should be a valuable complement to using datasets generated purely in vitro[49,50].

In conclusion, the take-home messages from this work are the following. Firstly, there are satisfactory TCRseq methods based on either DNA or RNA input, and in both cases the amount of material impacts both diversity and the detection of rare clones. Secondly, various methods are optimal for detecting maximal diversity, while others most accurately quantify the abundance of specific clonotypes. For the latter, UMI-based methods are potentially more accurate, although they could miss relevant but rare clones. In contrast, non-UMI RACE methods are more sensitive in capturing rare clones, especially for TRA. Thirdly, the availability of raw data is crucial in allowing reliable and reproducible in-depth analyses of TCR repertoires; the mPCR-1 service provider does not provide access to raw sequence data, while mPCR-1 and mPCR-3 do not disclose the proprietary pre-processing filters. In contrast, the RACE-2 provider provides raw data and all preprocessing algorithms. We summarized our results as well as practical aspects in Table 1. Regarding the results, we calculated for each method a rank value for Replicability, reliability and sensitivity based on various measures (**Table 1** and **Supplementary file**). We also summarized cost per sample, presence of controls or standards, format of the method and raw data availability. The Table 1 highlight the advantages and disadvantages of the different methods which could serve as guidance for end-users. Improved and more sophisticated data analyses are essential to

extract the full power of TCR repertoire data. We anticipate that now that TCR sequencing has come of age, the next key developments in the field will come from novel methods of data analysis, as has been the case in the related field of global transcriptomics.

Acknowledgments: We are grateful to M. Barbie for providing the human samples. This work benefited from equipment and services from the iGenSeq core facility, at ICM. This work was supported the ERC-Advanced TRiPoD (322856), LabEx Transimmunom (ANR-11-IDEX-0004-02) and RHU iMAP (ANR-16-RHUS-0001) grants to DK. EMF is funded by European Research Area Network – Cardiovascular Diseases (ERA-CVD, JCT2018, ANR-18-ECVD-0001) and iReceptorPlus (H2020 Research and Innovation Programme 825821) grants. MS and DMC were supported by a grant from the Ministry of Science and Higher Education of the Russian Federation (075-15-2019-1789). This work was funded in part by the intramural program of the National Institute of Allergy and Infectious Diseases (DCD). BC was supported by the National Institute for Health Research UCL Hospitals Biomedical Research. AE was supported by DFG CRTD (FZ 111). AND was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 (LQ1601). KK and PPR were supported by the European Research Area Network – Cardiovascular Diseases (ERA-CVD, JCT2018, AIR-MI Consortium) program.

Author contributions: PB, VQ, ESE, AND, IU, MI, TO, AE, SD, AR, KK, PR performed the experiments and raw data pre-processing. PB, VQ, MS and MI analyzed the data. EMF, DMC, BC, DCD and DK designed the experiments. PB, DK and EMF wrote the manuscript with input from all authors. DK and EMF conceived the study, which was supervised by EMF. DK, BC, AE, DCD, DMC, KK and MS obtained funding for the study.

Competing Interests statement: DMC and MS are cofounders of MiLaboratory LLC. AE, AND, AR, BC, DD, DK, EMF, ESE, IU, KK, MI, PB, PR, SD, TO, VQ declare no conflict of interest.

REFERENCES

1. Cui, J.-H. *et al.* TCR Repertoire as a Novel Indicator for Immune Monitoring and Prognosis Assessment of Patients With Cervical Cancer. *Front. Immunol.* **9**, (2018).
2. Davis, M. M. The αβ T Cell Repertoire Comes into Focus. *Immunity* **27**, 179–180 (2007).
3. Lindau, P. & Robins, H. S. Advances and applications of immune receptor sequencing in systems immunology. *Curr. Opin. Syst. Biol.* **1**, 62–68 (2017).
4. Miles, J. J., Douek, D. C. & Price, D. A. Bias in the [alpha][beta] T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol. Cell Biol.* **89**, 375 (2011).
5. Schrama, D., Ritter, C. & Becker, J. C. T cell receptor repertoire usage in cancer as a surrogate marker for immune responses. *Semin. Immunopathol.* **39**, 255–268 (2017).
6. Heather, J. M. *et al.* Dynamic Perturbations of the T-Cell Receptor Repertoire in Chronic HIV Infection and following Antiretroviral Therapy. *Front. Immunol.* **6**, (2016).
7. Howson, L. J. *et al.* MAIT cell clonal expansion and TCR repertoire shaping in human volunteers challenged with Salmonella Paratyphi A. *Nat. Commun.* **9**, 253 (2018).
8. 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).
9. Sycheva, A. L. *et al.* Quantitative profiling reveals minor changes of T cell receptor repertoire in response to subunit inactivated influenza vaccine. *Vaccine* **36**, 1599–1605 (2018).
10. Hogan, S. A. *et al.* Peripheral Blood TCR Repertoire Profiling May Facilitate Patient Stratification for Immunotherapy against Melanoma. *Cancer Immunol. Res.* **7**, 77–85 (2019).
11. Jin, Y. *et al.* TCR repertoire profiling of tumors, adjacent normal tissues, and peripheral blood predicts survival in nasopharyngeal carcinoma. *Cancer Immunol. Immunother.* **67**, 1719–1730 (2018).
12. Wieland, A. *et al.* T cell receptor sequencing of activated CD8 T cells in the blood identifies tumor-infiltrating clones that expand after PD-1 therapy and radiation in a melanoma patient. *Cancer Immunol. Immunother.* **67**, 1767–1776 (2018).
13. Six, A. *et al.* The Past, Present, and Future of Immune Repertoire Biology – The Rise of Next-Generation Repertoire Analysis. *Front. Immunol.* **4**, (2013).
14. Chien, Y. H., Gascoigne, N. R., Kavaler, J., Lee, N. E. & Davis, M. M. Somatic recombination in a murine T-cell receptor gene. *Nature* **309**, 322–326 (1984).
15. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
16. Lefranc, M.-P. Nomenclature of the Human T Cell Receptor Genes. *Curr. Protoc. Immunol.* **40**, A.1O.1-A.1O.23 (2000).
17. Dupic, T., Marcou, Q., Walczak, A. M. & Mora, T. Genesis of the αβ T-cell receptor. *PLOS Comput. Biol.* **15**, e1006874 (2019).
18. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107 (2009).

19. Warren, R. L. *et al.* Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* **21**, 790–797 (2011).
20. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci.* **111**, 13139–13144 (2014).
21. Greiff, V., Miho, E., Menzel, U. & Reddy, S. T. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends Immunol.* **36**, 738–749 (2015).
22. Wang, C. *et al.* High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl. Acad. Sci.* **107**, 1518–1523 (2010).
23. Zhang, W. *et al.* IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis. *Genetics* **201**, 459–472 (2015).
24. Douek, D. C. *et al.* A Novel Approach to the Analysis of Specificity, Clonality, and Frequency of HIV-Specific T Cell Responses Reveals a Potential Mechanism for Control of Viral Escape. *J. Immunol.* **168**, 3099–3104 (2002).
25. Eugster, A. *et al.* Measuring T cell receptor and T cell gene expression diversity in antigen-responsive human CD4+ T cells. *J. Immunol. Methods* **400–401**, 13–22 (2013).
26. Mamedov, I. Z. *et al.* Preparing Unbiased T-Cell Receptor and Antibody cDNA Libraries for the Deep Next Generation Sequencing Profiling. *Front. Immunol.* **4**, (2013).
27. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* **11**, 653–655 (2014).
28. Oakes, T. *et al.* Quantitative Characterization of the T Cell Receptor Repertoire of Naïve and Memory Subsets Using an Integrated Experimental and Computational Pipeline Which Is Robust, Economical, and Versatile. *Front. Immunol.* **8**, (2017).
29. Liu, X. *et al.* Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire. *PLOS ONE* **11**, e0152464 (2016).
30. Rosati, E. *et al.* Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, (2017).
31. Dunn-Walters, D., Townsend, C., Sinclair, E. & Stewart, A. Immunoglobulin gene analysis as a tool for investigating human immune responses. *Immunol. Rev.* **284**, 132–147 (2018).
32. Doenecke, A., Winnacker, E.-L. & Hallek, M. Rapid amplification of cDNA ends (RACE) improves the PCR-based isolation of immunoglobulin variable region genes from murine and human lymphoma cells and cell lines. *Leukemia* **11**, 1787–1792 (1997).
33. Nielsen, S. C. A. & Boyd, S. D. Human adaptive immune receptor repertoire analysis—Past, present, and future. *Immunol. Rev.* **284**, 9–23.
34. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J. & Chain, B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* **29**, 542–550 (2013).
35. Taylor, S., Yasuyama, N. & Farmer, A. A SMARTer approach to profiling the human T-cell receptor repertoire. *J. Immunol.* **196**, 209.5-209.5 (2016).
36. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
37. Yokota, R., Kaminaga, Y. & Kobayashi, T. J. Quantification of Inter-Sample Differences in T-Cell Receptor Repertoires Using Sequence-Based Information. *Front. Immunol.* **8**, (2017).
38. Gudikote, J. P. & Wilkinson, M. F. T-cell receptor sequences that elicit strong down-regulation of premature termination codon-bearing transcripts. *EMBO J.* **21**, 125–134 (2002).
39. Schulze-Koops, H. Lymphopenia and autoimmune diseases. *Arthritis Res. Ther.* **6**, 178–180 (2004).

40. Miho, E. *et al.* Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Front. Immunol.* **9**, (2018).
41. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
42. Britanova, O. V. *et al.* Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *J. Immunol.* **196**, 5005–5013 (2016).
43. Brüggemann, M. *et al.* Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. *Leukemia* (2019) doi:10.1038/s41375-019-0496-7.
44. Knecht, H. *et al.* Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. *Leukemia* (2019) doi:10.1038/s41375-019-0499-4.
45. Friedensohn, S. *et al.* Synthetic Standards Combined With Error and Bias Correction Improve the Accuracy and Quantitative Resolution of Antibody Repertoire Sequencing in Human Naïve and Memory B Cells. *Front. Immunol.* **9**, (2018).
46. Breden, F. *et al.* Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front. Immunol.* **8**, 1418 (2017).
47. Rubelt, F. *et al.* Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nature Immunology* https://www.nature.com/articles/ni.3873 (2017) doi:10.1038/ni.3873.
48. Vander Heiden, J. A. *et al.* AIRR Community Standardized Representations for Annotated Immune Repertoires. *Front. Immunol.* **9**, 2206 (2018).
49. Zhang, Y. *et al.* Tools for fundamental analysis functions of TCR repertoires: a systematic comparison. *Brief. Bioinform.* doi:10.1093/bib/bbz092.
50. Weber, C. R. *et al.* immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics* doi:10.1093/bioinformatics/btaa158.

METHODS

Blood effector T cell isolation

Peripheral blood mononuclear cells (PBMC) from two healthy blood donors (Etablissement Français du sang; French Blood Center) were obtained with written informed consent for biomedical research. The experiments carried out were in conformity with the Helsinki Declaration on Biomedical Research. Donors A (experiment A) and B (experiment B) were both men, 36 and 54 years old, respectively. $CD3^+CD4^+CD127^+CD25^-$ cells ($CD4^+$ T effector cells) were sorted at the Sorbonne Université laboratory as follows: $CD4^+$ cells were isolated by Lymphoprep (Stemcell®) density gradient and positive selection using the Dynabeads™ CD4 Positive Isolation Kit (Invitrogen®). Enriched $CD4^+$ T-cells were then labeled with anti-$CD3^+$, $CD4^+$, $CD127^+$ and $CD25^+$ antibodies and effector T-cells were sorted on a FACS ARIA II with a purity > 95% (**Supplementary Fig.1a**).

Jurkat cell culture

The Jurkat cell line with a known TCR (TRAV8-4-CAVSDLEPNSSASKIIF-TRAJ3; TRBV12-3-CASSFSTCSANYGYTF-TRBJ1-2) (clone E6-1), from ATCC, was grown in 5% CO_2, in RPMI 1640 medium, supplemented with 10% (v/v) fetal bovine serum (FBS), 2 mM L-glutamine, 50 U/mL penicillin, and 50 µg/mL streptomycin at the Sorbonne Université laboratory.

RNA and DNA extraction

In experiment A, DNA and RNA were both extracted using TRIzol Reagent (Invitrogen®) from 5 million Jurkat cells and 20 million $CD4^+$ T effector cells and, in experiment B, only RNA was extracted using the RNAqueous-Kit (Invitrogen®) from 7.2 million $CD4^+$ T effector cells following the manufacturer's recommendations. DNA concentration and RNA concentration

were measured on a NanoDrop1000 (Thermo Scientific™) and RNA integrity was determined on a Bioanalyzer (Agilent®) with measurements higher than 8. RNA and DNA extraction and validation were performed at the Sorbonne Université laboratory.

Aliquot preparation for method comparison

In experiment A, 100 ng of RNA or DNA from the $CD4^+$ effector T-cells sorted from donor A was split into 3 aliquots that were spiked with different amounts of RNA or DNA from the Jurkat cell line, at ratios of 1/10, 1/100 and 1/1000. Each spiked aliquot was further split into 3 and all replicates were processed by all methods tested (7 for RNA and 2 for DNA; **Supplementary Fig.1b**). With experiment B, we analyzed the impact of the input material quantity. RNA from sorted $CD4^+$ effector T-cells of donor B was extracted, split into 15 aliquots of 100 ng each and 15 aliquots of 10 ng each and processed in triplicate using 5 of the RNA-based methods (**Supplementary Fig.1c**). Aliquots were prepared at the Sorbonne Université laboratory and sent to the partners.

Flow Cytometry

Vβ identification was performed on enriched $CD4^+$ effector T-cells from experiment A (see *Blood effector T cell isolation* for enrichment procedure) stained with the IOTest Beta Mark TR Repertoire Kit (Beckman Coulter®) according to the manufacturer's protocol as well as with CD4-APC, CD127-BV421, CD25-PECy7. Data acquisition was performed on a Cytoflex® (Beckman Coulter®) using CytExpert® software. FlowJo® was used for data analysis. Vb frequencies were calculated on $CD4^+CD25^-CD127^+$ gated cells (**Supplementary Fig.4a-b**). Staining was performed at the Sorbonne Université laboratory.

TCR library preparation and sequencing

The nine protocols for TCR library preparation compared in this study were selected according to at least one the following criteria: published use by groups other than the one who developed it (mPCR-1, mPCR-3, RACE-1, RACE-2, RACE-4 and RACE-5), (ii) their association with well-known analysis tools (RACE-1, RACE-2, mPCR-2) and (iii) commercially available (RACE-2, RACE-3, mPCR-1, mPCR-3). Sequencing protocols were harmonized taking into account published recommendations or recommendations provided by the manufacturer of commercial kits or by the owner or users of the protocol. All protocols are detailed in **Supplementary material and methods**.

TCR deep sequencing data processing

FASTQ raw data files were obtained from each method, except for Multiplex-1 & 2, for which we obtained, respectively, FASTA file and FASTQ files following proprietary pre-processing. For RACE-1 and RACE-2, UMI pre-processing was performed following protocols published elsewhere[27,28,34]. FASTQ and FASTA files were then processed for TRB and TRA sequence annotation using the MiXCR software[36] (v2.1.10) with RNA-Seq default parameters (*-p rna-seq –s hsa*) as available online. MiXCR extracts TRA and TRB repertoire providing correction of PCR and sequencing errors.

Data analysis

Statistical comparisons and multivariate analyses were performed using R software version 3.5.0 (www.r-project.org). We used the ggplot2 package to generate figures[51], except heatmaps. More complex analyses are detailed in the next section.

Comparing VDJ rearrangement statistics

An empirical VDJ rearrangement model for each method was built as follows. We analyzed clonotype tables to obtain comprehensive statistics of VDJ rearrangements including the frequencies of V/D/J segment usage, number of added N Bases (namely “insert profile”, i.e. the probability distribution of having A/T/G/C inserted in the N-region of CDR3 given that we observe a certain base inserted before it) and V/J segment trimming bases, with the IGoR package[52]. This model is built in a 'greedy' way in the sense that it uses best alignments provided by MiXCR rather than running expectation maximization procedures as described in Murugan et al.[53]. We utilized the Jensen-Shannon divergence (JSD) between distributions of VDJ usage to define the following two statistics that we use for comparative analysis of different TCRseq methods: 1) *replicability* measured as the distance between different samples produced by the same protocol and 2) *reproducibility* measured as the distance between samples produced by two different protocols. MDS used for sample mapping was performed on rank-transformed distances to avoid the distorting effect of outliers. All the analyses involve VDJ usage inferred from weighted data (TCR clonotype is weighted by its frequency in the sample) to account for TCRseq method amplification biases.

Similarity analysis

Pearson and Spearman correlations, the Morisita-Horn index[54] (MH) and the Jaccard similarity index[55] (JSI) were used to assess the similarity between samples. The MH index takes into account the relative abundance of species in the sample, while the JSI is a measure of the intersection between two populations relative to the size of their union, and is independent of relative abundances. Both indices vary between 0 (no overlap) and 1 (perfect overlap). JSI and MH were calculated using the DIVO package[56] on R. In order to

discriminate indices represented by a heatmap with the pheatmap package[57], we used a different set of colors. The Pearson and Spearman correlations are presented as yellow/white/orange (**Fig.2c and Supplementary Fig.4e**), MH is presented as blue/white/red (**Fig.3a**) and JSI is presented as purple/yellow/green (**Supplementary Fig.5a**).

Diversity profiling

The diversity was analyzed using two indices. Rényi entropy[58] is a generalization of Shannon entropy, which increases when both species richness and evenness are high. Rényi entropy is a function of a parameter α spanning from (i) the species richness ($\alpha = 0$), which corresponds to the number of clonotypes regardless of their abundance, to (ii) the clonal dominance ($\alpha \rightarrow +\infty$), corresponding to the frequency of the most predominant clonotype. For $\alpha = 1$, the Shannon diversity index is computed. The exponential of the Rényi entropy corresponds to the actual number of clonotypes in the datasets[59] and is used to build a diversity profile[60]. It was computed using the entropy package[61] on R. ChaoE[62] index was calculated with the iNEXT package[63] as a measure of extrapolation of the possible number of clonotypes based on the observed clonotypes. Rarefaction curves were interpolated from 0 to the current sample size and then extrapolated to the size of the largest of samples, allowing comparison of diversity estimates. Interpolation and extrapolation were based on ChaoE multinomial models[64].

Meta-repertoire construction

We generated an in silico meta-repertoire from the sequences obtained from the 108 replicates (45 for TRA and 63 for TRB). This meta-repertoire, for each chain, was designed to minimize biases by (i) pooling all clonotypes from the 9 datasets and removed singletons to

avoid introducing noise due to PCR errors, (ii) Selecting non-reprocessed datasets, meaning before UMI, (iii) keeping only clonotypes found by at least 3 different methods to avoid bias toward one particular method. The threshold was defined to reach a dataset size as close as possible to the original datasets to avoid additional sampling, (iv) normalizing the size of each dataset to the lowest dataset to ensure the same weighting for each method. Completion of the representative meta-repertoire was achieved by pooling all the datasets. This generated a pooled dataset of 14 458 TRA and 18 735 TRB clonotypes.

Methods-only references

51. Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis*. (2016).
52. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, (2018).
53. 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).
54. Horn, H. S. Measurement of 'Overlap' in Comparative Ecological Studies. *Am. Nat.* **100**, 419–424 (1966).
55. Jaccard, P. The distribution of the flora in the Alpine zone. *New Phytol.* **11**, 37–50 (1912).
56. Sadee, C., Pietrzak, M., Seweryn, M. & Rempala, G. *Tools for analysis of diversity and similarity in biological system (Diversity and Overlap Analysis Package)*. (2017).
57. Kolde, R. *Package 'pheatmap'*. (2019).
58. Renyi, A. On measures of information and entropy. *Proc. 4th Berkeley Symp. Math. Stat. Probab.* 547–561 (1961).
59. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
60. Chaara, W. *et al.* RepSeq Data Representativeness and Robustness Assessment by Shannon Entropy. *Front. Immunol.* **9**, 1038 (2018).
61. Hausser, J. & Strimmer, K. *Estimation of Entropy, Mutual Information and Related Quantities*. (2014).
62. Colwell, R. K. *et al.* Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J. Plant Ecol.* **5**, 3–21 (2012).
63. Hsieh, T. C., Ma, K. H. & Chao, A. *Package iNEXT: Interpolation and Extrapolation for Species Diversity*. (2019).
64. Chao, A. et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monogr. 84, 45–67 (2014).
65. Corrie, B. D. *et al.* iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* **284**, 24–41 (2018).

Data Availability

All the fastq data obtained in this study, including the Jurkat Clone E6-1 (ATCC®TIB-152™) cell line TCR alpha and beta sequences, were deposited in the NCBI Sequence Read Archive repository following MiAIRR standard recommendations[47] under the BioProject ID PRJNA548335. The aligned sequence data will be stored in an iReceptor Repository at Sorbonne Université as a repository in the AIRR Data Commons and can be explored and downloaded through the iReceptor Gateway[65] (https://gateway.ireceptor.org). Source data for TCRVb flow cytometry data are provided as **Supplementary Fig.4a-b**. All other data are available from the corresponding author upon request.

Code Availability

All software packages and programs are publicly available and open source. Scripts used to analyze the data with MiXCR are available from https://mixcr.milaboratory.com ; Decombinator from https://github.com/innate2adaptive/Decombinator; MiGEC from https://github.com/mikessh/migec; detailed VDJ rearrangement statistics scripts are available from https://github.com/antigenomics/repseq-protocol-comparison. There is no restriction on the use of the code or data.

FIGURE

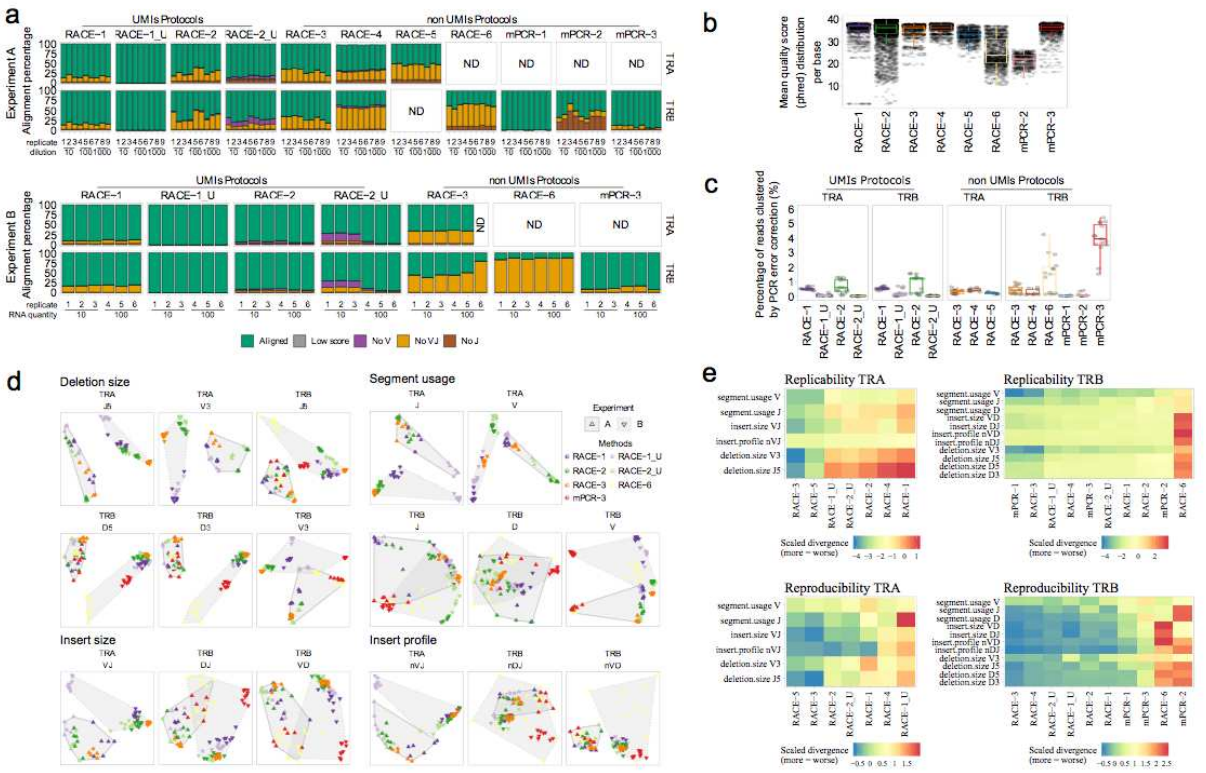


Fig. 1: Performance statistics and VDJ rearrangement model of each method for experiments A and B.

a, The proportion of sequence reads aligned for TRA or TRB genes for each TCRseq replicate per experiment (Experiment A, top, Experiment B, bottom). The bars represent the percentage of TRA and TRB alignment, and the reason for alignment failure is color coded. **b**, Distribution of the reads quality control per base (QC) for each method over all datasets (n = 15 (9 technical replicates from experiment A + 6 technical replicates from experiment B)), computed with fastQC software (www.bioinformatics.babraham.ac.uk/projects/fastqc). **c**, Percentage of reads collapsed after PCR error correction for all samples in the study (n=15 as in Figure 1b). For each method, the MiXCR clustering strategy was applied to correct for PCR errors and collapse reads. Each box-plot represents the percentage of clustered reads. **d**, Multi-dimensional scaling (MDS) of V(D)J recombination parameters. MDS was performed

based on the Jensen-Shannon Divergence (JSD) calculated between replicates on weighted VDJ segment usage (Segment usage), non-template nucleotide insert size distributions (Insert size), V/D/J segment trimming distributions (Deletion size), and nucleotide frequencies in N-inserts (Insert profile). JSD values were transformed to rank for better visualization. Solid and dotted polygons outline samples from experiments A and B, respectively. Colors represents the different methods as in B (only methods used in both experiments are presented). **e**, Replicability and reproducibility of the TRA and TRB repertoires for each method. The average JSD calculated in D (rows) for TRA (left) and TRB (right) measured between replicates produced by the same method (Replicability, top) or replicates of a given method and all other protocols (Reproducibility, bottom) was used as distance metric to compare different protocols (columns). Columns are sorted according to the mean scaled distance (averaged over all rows) from the lowest (best replicability/reproducibility) to the highest (worst replicability/reproducibility). Distance values are shown using a color scale. Jurkat TCR sequences were removed from datasets for this analysis. Boxplots visualize five summary statistics: the median, two hinges (the lower and upper hinges correspond to the first and third quartiles) and two whiskers (the upper whisker extends from the hinge to the largest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge).

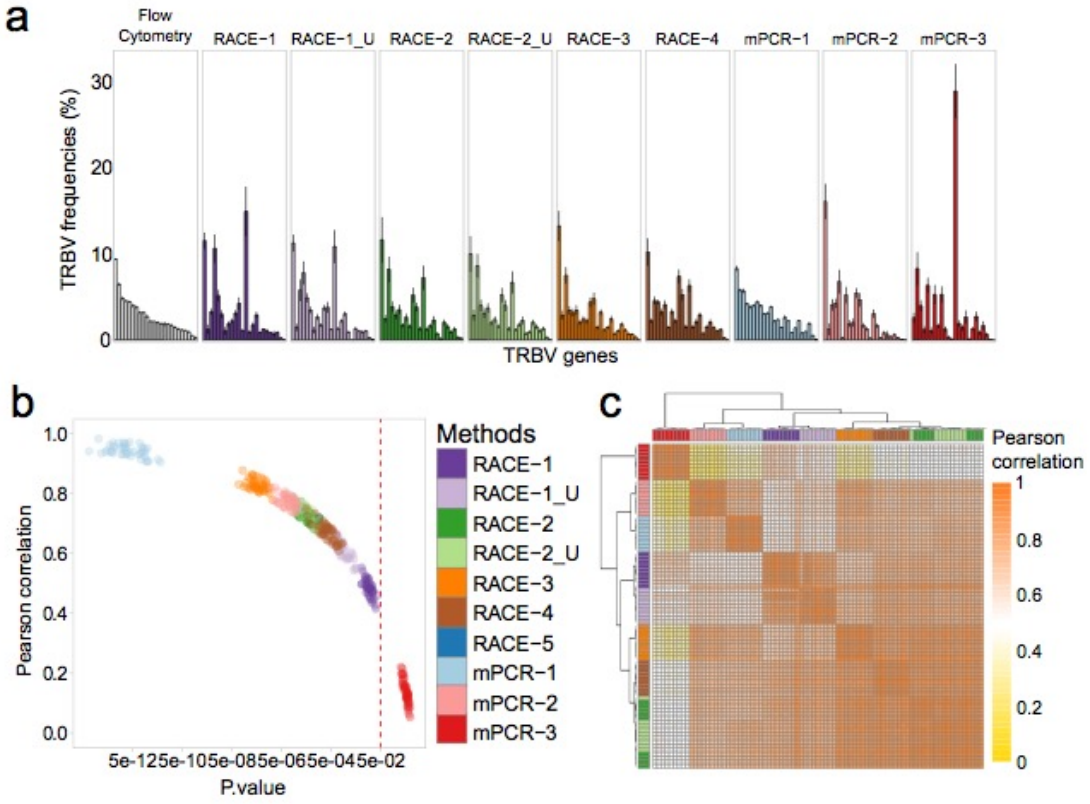


Fig. 2: TRBV usage comparison between flow cytometry and TCRseq.

a, Flow cytometry and TCRseq TRBV frequencies. Bar plots represent the TRBV frequencies calculated from flow cytometry stained $CD4^+$ T effector cells for the 24 TRBV for which antibodies are available and from the TCRseq data (n = 9 technical replicates from experiment A), considering only clonotypes annotated for the same 24 TRBV (original TRBV frequencies were used accordingly). Histograms of the 24 TRBV frequencies are presented as mean values +/- SD and organized by decreasing order using frequencies obtained by flow cytometry as a reference (TRBV20-1, TRBV19, TRBV12-3/4, TRBV28, TRBV2, TRBV3-1, TRBV30, TRBV6-5/9, TRBV9, TRBV5-1, TRBV4-1/2, TRBV27, TRBV29-1, TRBV6-6, TRBV11-2, TRBV10-3, TRBV25-1, TRBV6-2, TRBV18, TRBV5-5, TRBV14, TRBV5-6, TRBV13, TRBV4-3). **b**, TRBV usage correlation between flow cytometry and TCRseq. Pearson's correlation of the TRBV frequencies between the flow cytometry datasets (n=5) and the TCRseq replicates (n=9

from experiment A) was calculated for each method. The plot is represented by the correlation score (y-axis) and the *P*-value (x-axis) of the correlation, allowing the classification of the methods. **c**, Heatmap of the Pearson correlations between each replicate for the distribution of TRBV gene usage (n=62 unique TRBV genes). The Euclidean distance was used for hierarchical clustering as a color-coded matrix ranging from 0 (yellow, maximum dissimilarity) to 1 (orange, maximum similarity). Jurkat TCR sequences were removed from datasets for this analysis.

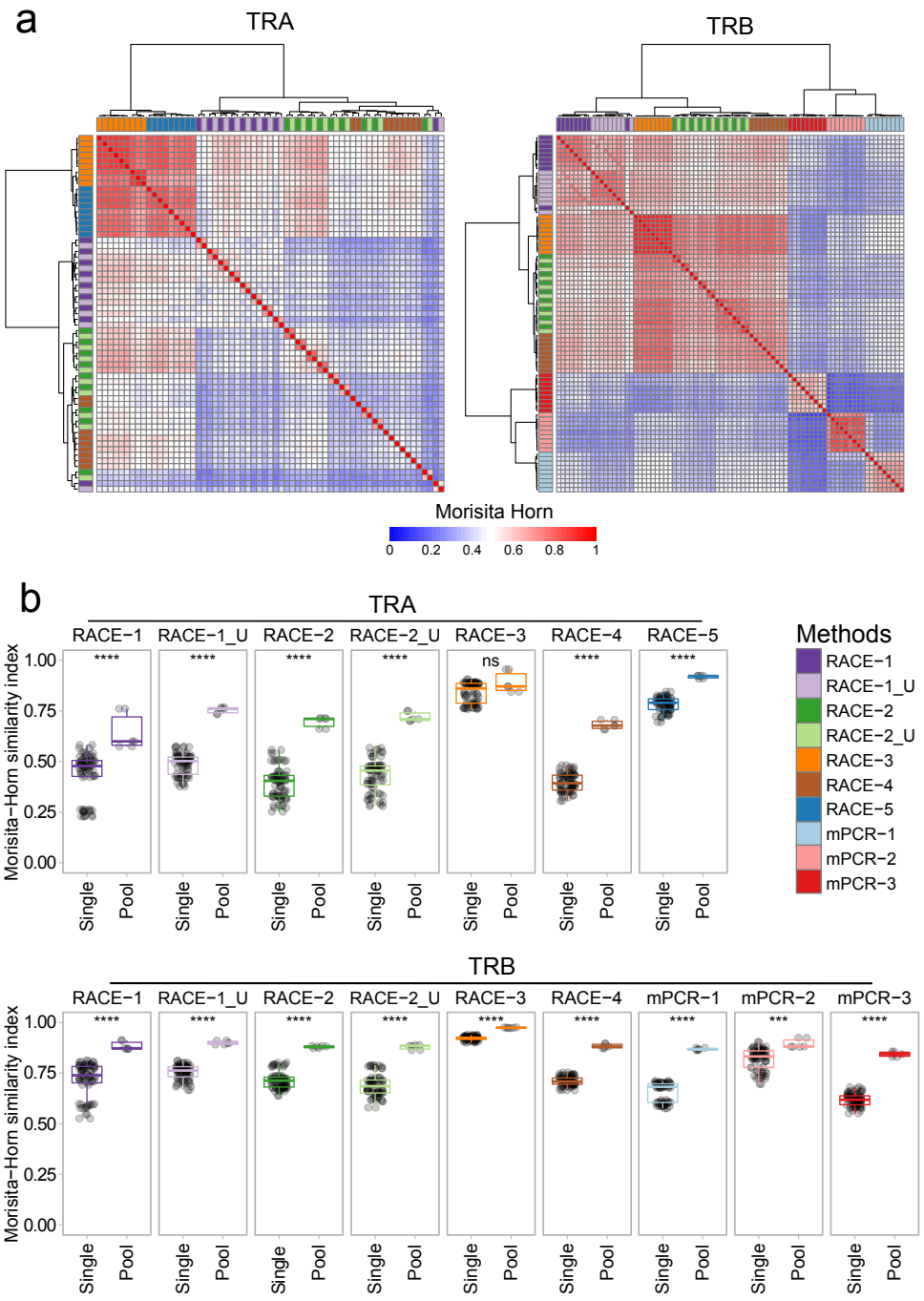


Fig. 3: The reproducibility of detection of major TCR clonotypes by different methods.

a, Heatmaps of the Morisita-Horn similarity index (MH). MH scores were calculated between each replicate across all methods for the top 1% of most predominant clonotypes (MPC) for TRA (left) and TRB (right). The Euclidean distance was used for hierarchical clustering as a color-coded matrix ranging from 0 (blue, maximum dissimilarity) to 1 (red, maximum

similarity). **b**, Comparison between individual replicates (Single, n = 9 from experiment A) and pooled replicates (Pool, n = 3) by the MH similarity index. Datasets from replicates of the same dilution were pooled for each method to get 1 pooled sample per dilution. Singletons (count=1) were removed; MH similarity scores were calculated for the top 1% of most predominant clonotypes for TRA (left) and TRB (right). Boxplots visualize five summary statistics: the median, two hinges (the lower and upper hinges correspond to the first and third quartiles) and two whiskers (the upper whisker extends from the hinge to the largest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge). Mann-Whitney statistical test was performed to compare Single vs Pool results by method and by chain. ns, not significant; *P<0.05; **P<0.01; ***P<0.001; ****P<0.0001. Jurkat TCR sequences were removed from datasets for this analysis.

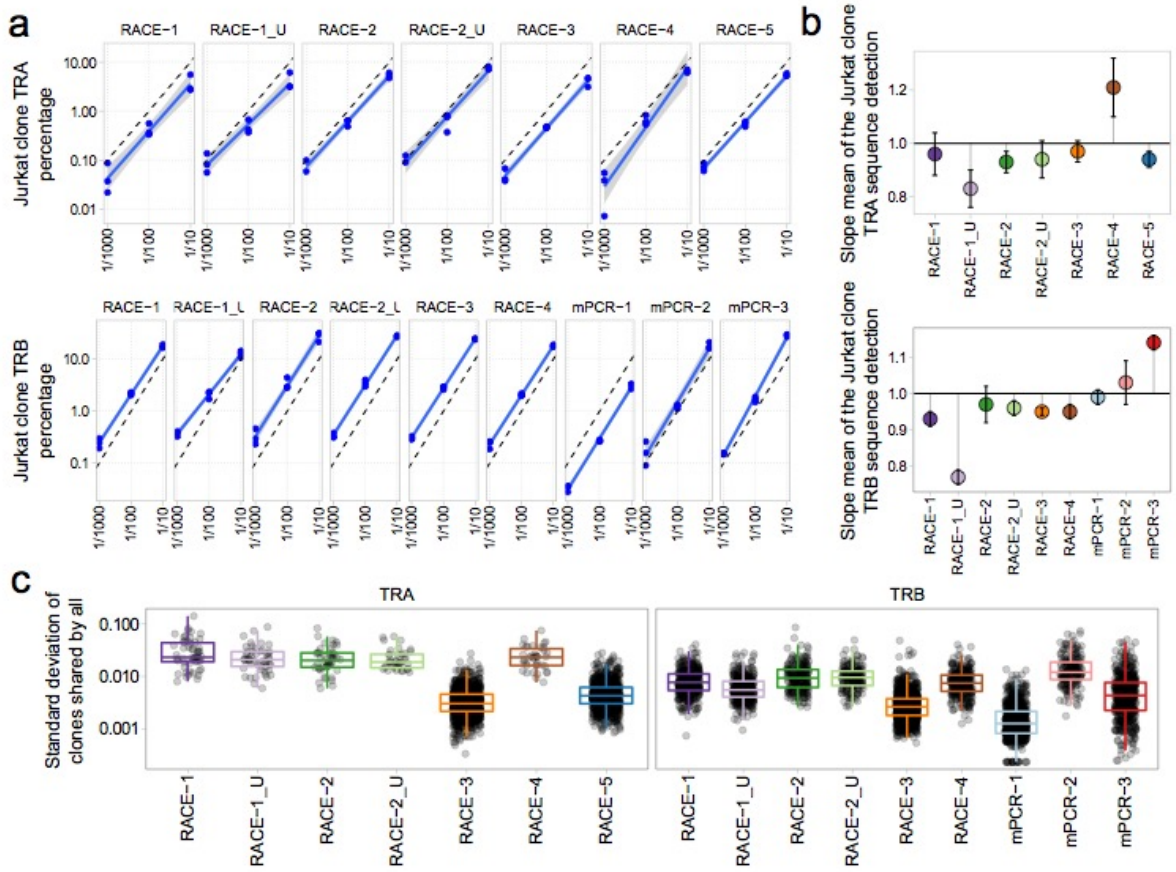


Fig. 4: Sensitivity of TCR sequence detection by different methods.

a, Jurkat clone percentage. Jurkat TRA (top) and TRB (bottom) clonotype percentages were calculated for each experiment per dilution (1/10, 1/100 and 1/1000 spike-in) and are represented by the blue dots. The blue line represents linear regression and the black dashed line represents the theoretically expected percentage. Errors bands in light grey represent the 95% confidence interval. **b**, Slope of the Jurkat tracking linear regression. Slope was computed between dilutions. Data are represented as mean values of the slope +/- SD by method (n = 9 technical replicates from experiment A) for TRA (top) and TRB (bottom). **c**, Standard deviation of the clonotypes shared among all the replicates (n=9 from experiment A), after excluding the Jurkat clone, calculated per method, for TRA (left) and TRB (right). Boxplots visualize five summary statistics: the median, two hinges (the lower and upper hinges correspond to the first and third quartiles) and two whiskers (the upper

whisker extends from the hinge to the largest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge).

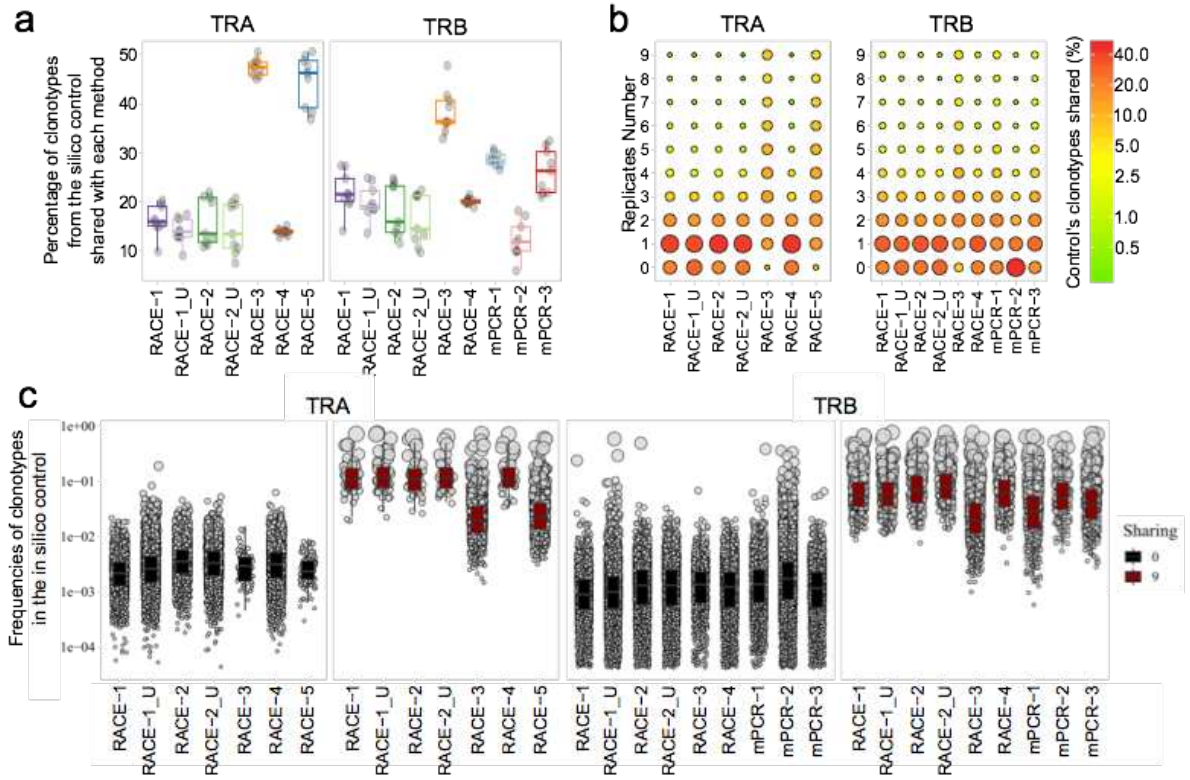


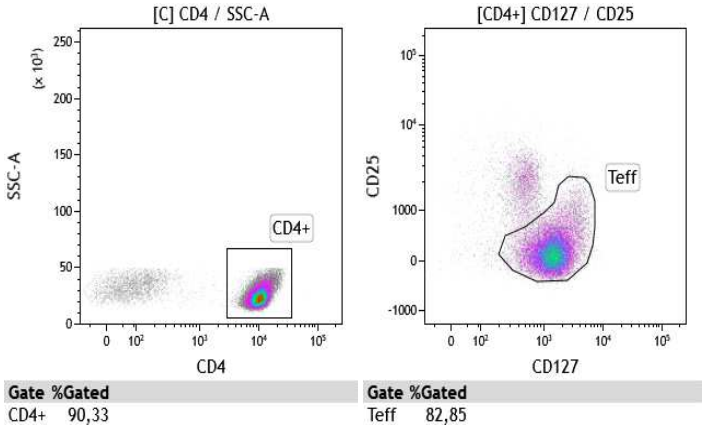
Fig. 5: Sharing with robust and representative meta-repertoire.

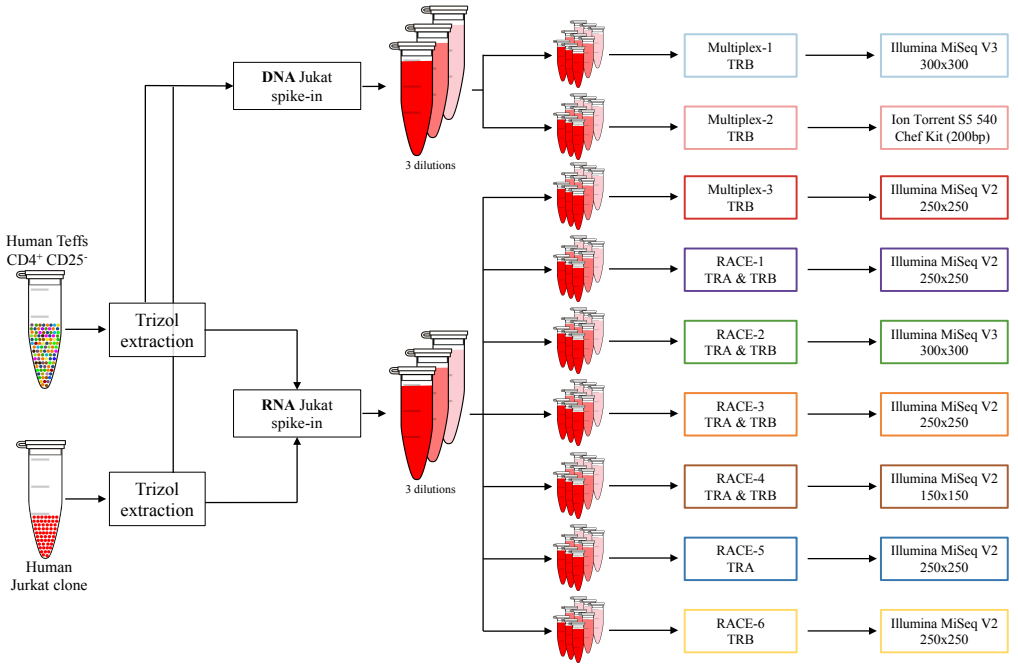
a, Replicate sharing fraction in meta-repertoire repertoire (focus on meta-repertoire clonotypes) for TRA (left) and TRB (right). The values represented correspond to the percentage of clonotypes from each replicate (n = 9 technical replicates from experiment A) per method found in the meta-repertoire, median and the 1st and 3rd quartiles are shown. **b**, Replicability of replicate methods with meta-repertoire for TRA (left) and TRB (right). By chain, heatmaps on the left represent the fraction, which corresponds to the percentage of meta-repertoire clonotypes found in 1 to 9 replicates per method (0: unseen in any of the replicates). **c**, Distribution of meta-repertoire clonotypes in the replicates (n=9 from experiment A) by methods for TRA (left) and TRB (right). Each dot represents a meta-repertoire clonotype and the boxplot represents the average frequencies. Black boxplots with corresponding gray dots represent the unseen clonotypes (seen in n=0 replicates) and red boxplots with corresponding gray dots represent the clonotypes found by all the

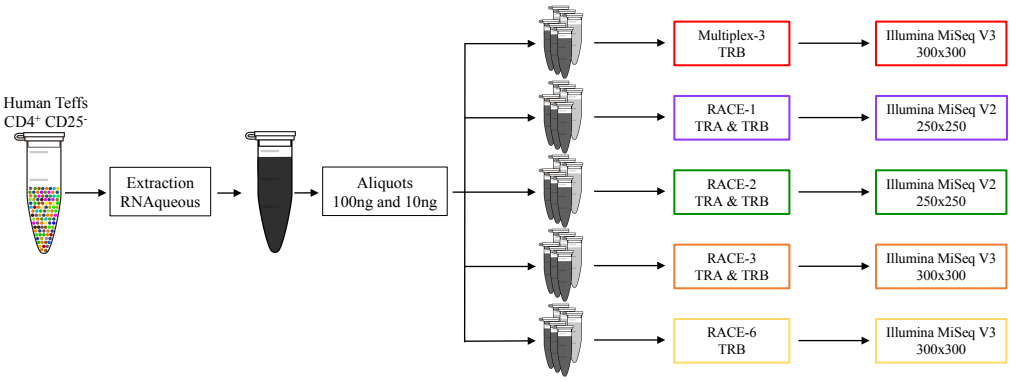
replicates (seen in n=9 replicates). Each method is represented independently. Jurkat TCR sequences were removed from datasets for this analysis. Boxplots visualize five summary statistics: the median, two hinges (the lower and upper hinges correspond to the first and third quartiles) and two whiskers (the upper whisker extends from the hinge to the largest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge).

Supplementary material and methods

Supplementary Figure legends

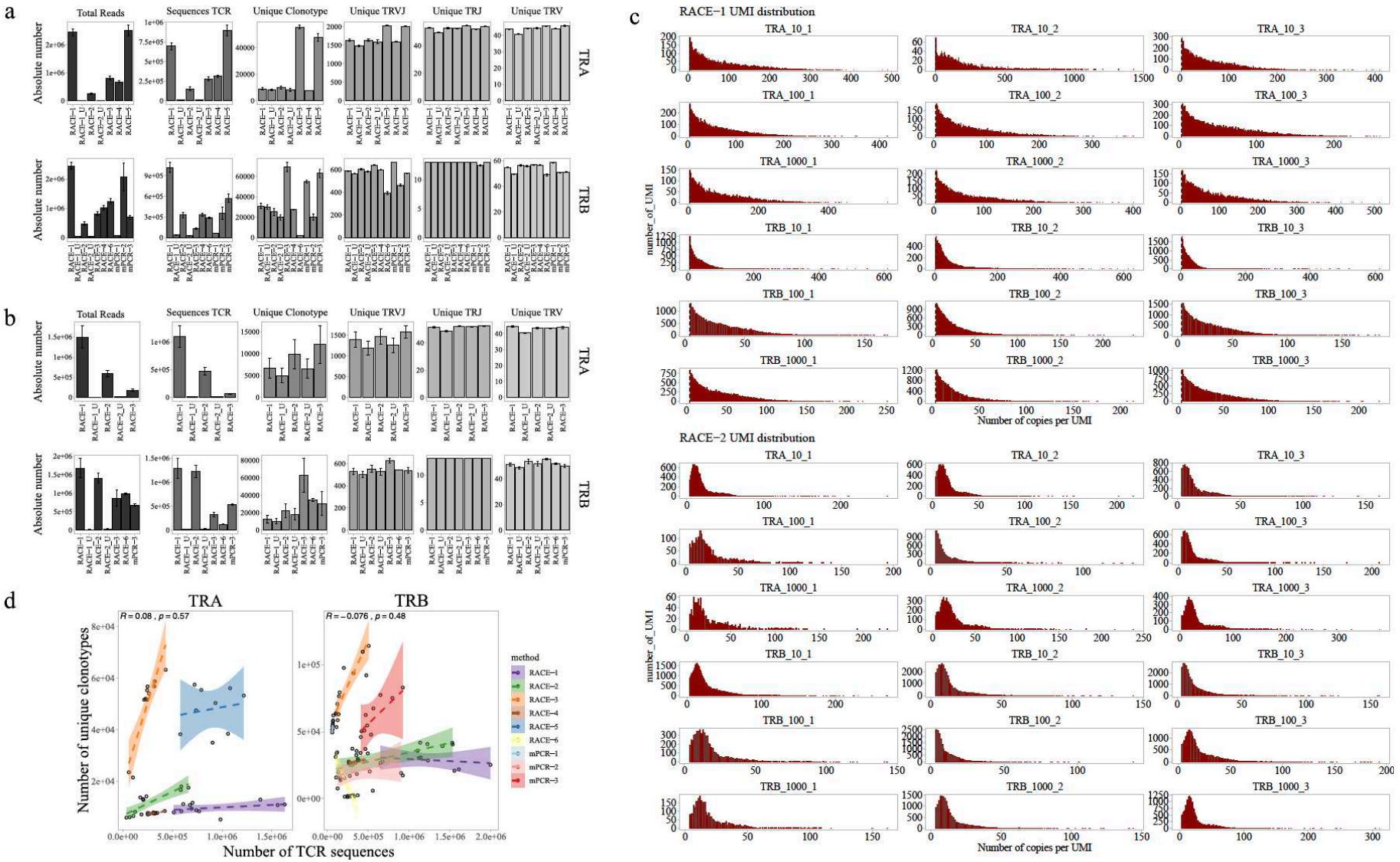






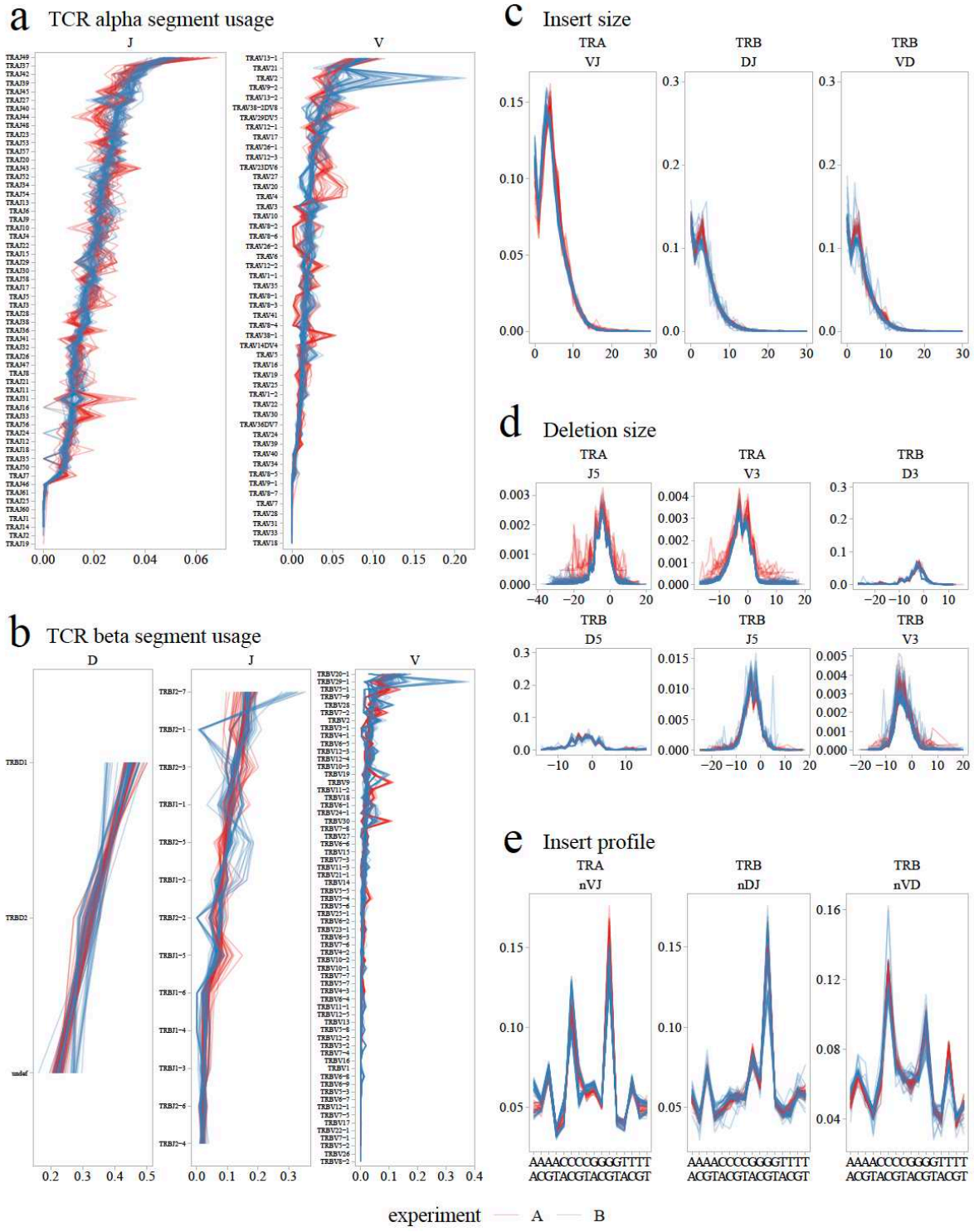
Supplementary Fig. 1: Sorting cell strategy and experimental design.

a, Gating strategy for effector T cell sorting. $CD4^+CD25^-$ T cells were gated on previously enriched lymphocyte (left panel). Effector T cells were gated and sorted based on $CD127^+CD25^-$ expression (right panel). **b**, Experiment A. DNA and RNA of FACS-sorted effector T cells ($CD4^+CD25^-CD127^+$) from donor 1 and cultured Jurkat leukemic T-cells were extracted. Jurkat DNA or RNA were then added to effector T cells DNA or RNA, respectively, at the following ratios: 1/1000, 1/100, 1/10 to a final quantity of 100ng. Triplicate of each ratio condition were processed by 9 TCR library preparation methods (7 with RNA and 2 with DNA). **c**, Experiment B. RNA of FACS-sorted effector T cells ($CD4^+CD25^-CD127^+$) from donor 2 was extracted. Triplicate of aliquots prepared from two quantity of RNA (10ng and 100ng) were processed by 5 TCR library preparation methods.


a
Total Reads
Sequences TCR
Unique Clonotype
Unique TRVJ
Unique TRJ
Unique TRV
Absolute number
TRA
TRB
b
Total Reads
Sequences TCR
Unique Clonotype
Unique TRVJ
Unique TRJ
Unique TRV
Absolute number
TRA
TRB
c
RACE−1 UMI distribution
TRA_10_1
TRA_10_2
TRA_10_3
TRA_100_1
TRA_100_2
TRA_100_3
TRA_1000_1
TRA_1000_2
TRA_1000_3
TRB_10_1
TRB_10_2
TRB_10_3
TRB_100_1
TRB_100_2
TRB_100_3
TRB_1000_1
TRB_1000_2
TRB_1000_3
number_of_UMI
Number of copies per UMI
RACE−2 UMI distribution
TRA_10_1
TRA_10_2
TRA_10_3
TRA_100_1
TRA_100_2
TRA_100_3
TRA_1000_1
TRA_1000_2
TRA_1000_3
TRB_10_1
TRB_10_2
TRB_10_3
TRB_100_1
TRB_100_2
TRB_100_3
TRB_1000_1
TRB_1000_2
TRB_1000_3
number_of_UMI
Number of copies per UMI
d
TRA
TRB
R = 0.08 , p = 0.57
R = −0.076 , p = 0.48
Number of unique clonotypes
Number of TCR sequences
method
RACE−1
RACE−2
RACE−3
RACE−4
RACE−5
RACE−6
mPCR−1
mPCR−2
mPCR−3

Supplementary Fig. 2: Sequencing statistics summary per method and experiment.

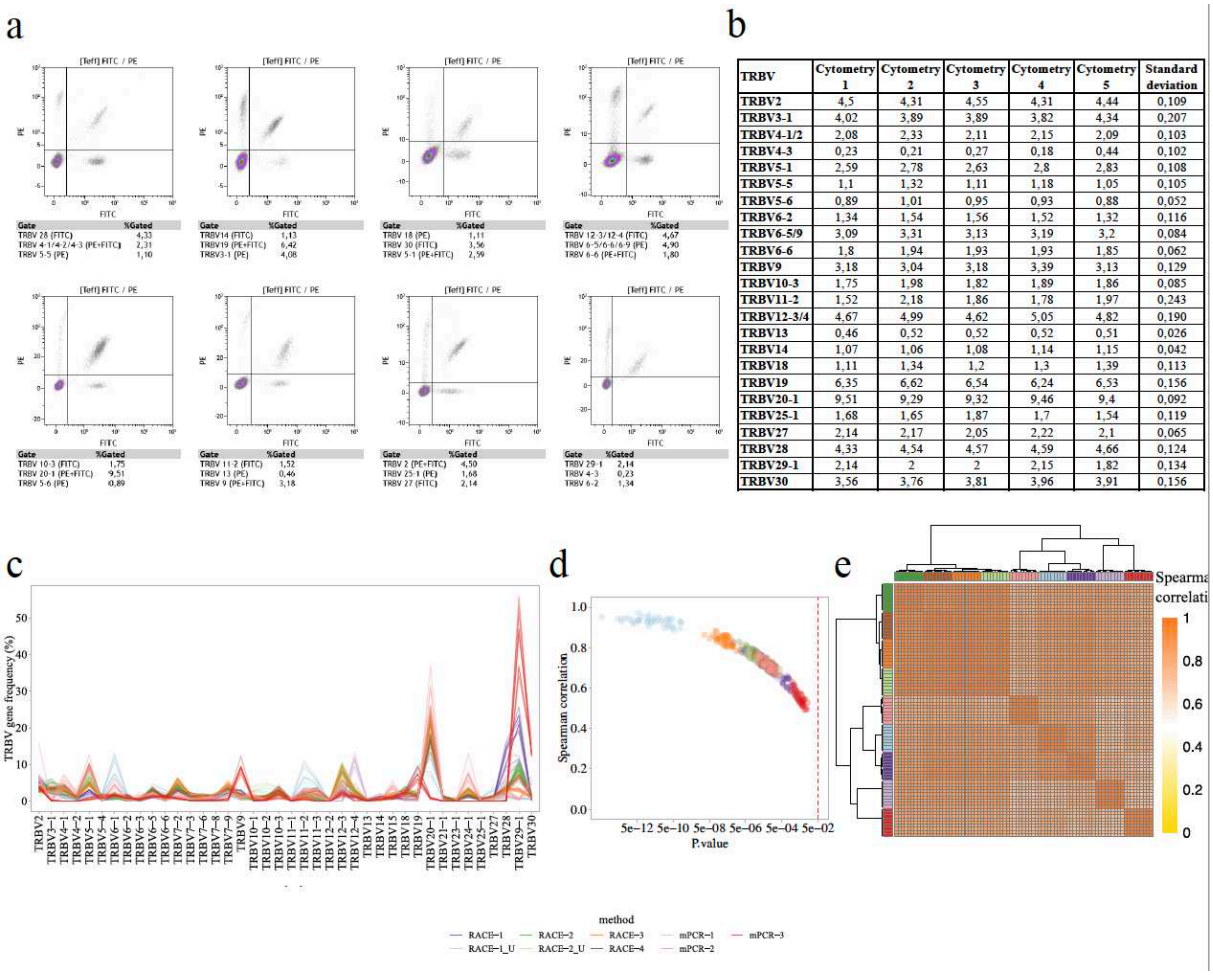
a-b, For experiment A (**a**, n = 9) and experiment B (**b**, n = 6), histogram plots represent from left to right the mean values +/- SEM of raw reads (Total Reads), TCR Sequences, unique clonotype (V-CDR3aa-J), V-J combination (Unique TRVJ), J gene (Unique TRJ) and V gene (Unique TRV) obtained by each method for both TRA (top) and TRB (bottom) chains. **c**, UMI distribution per sample for RACE-1 and RACE-2 methods from experiment A. X-Axis represents the number of copies of each UMI and y-axis represents the number of UMI for each copy-number. **d**, Correlation between the number of unique clonotype and the number of TCR sequences per method and chain from experiment A. Error bands represent the 95% confidence interval.



Supplementary Fig. 3: VDJ rearrangement statistic distributions.

VDJ rearrangement parameter distribution have been assessed as described in material and methods using "weighted" data, i.e. the distributions are scaled by the number of reads associated with each unique clonotype. Results are represented for experiment A (red) and

experiment B (blue) data as follows: **a**-**b**, Variable (V), Diversity (D) and Joining (J) segment usage distributions for TCR alpha (**a**) and beta (**b**) chains. **c**, Insert size distributions for VJ junctions (TRA) and VD/DJ junctions (TRB). **d**, Number of nucleotides trimmed from 5' and 3' ends of V, D and J segments for TRA and TRB chains. **e**, Non-template nucleotide frequencies for VJ (nVJ), VD/DJ (nVD and nDJ) junctions. Insertion probabilities of an observed base (top letter) given previous base (bottom letter) are given, 5'->3' direction is considered for nVJ and nVD, 3'->5' is considered for nDJ. Jurkat TCR sequences were removed from datasets for this analysis.

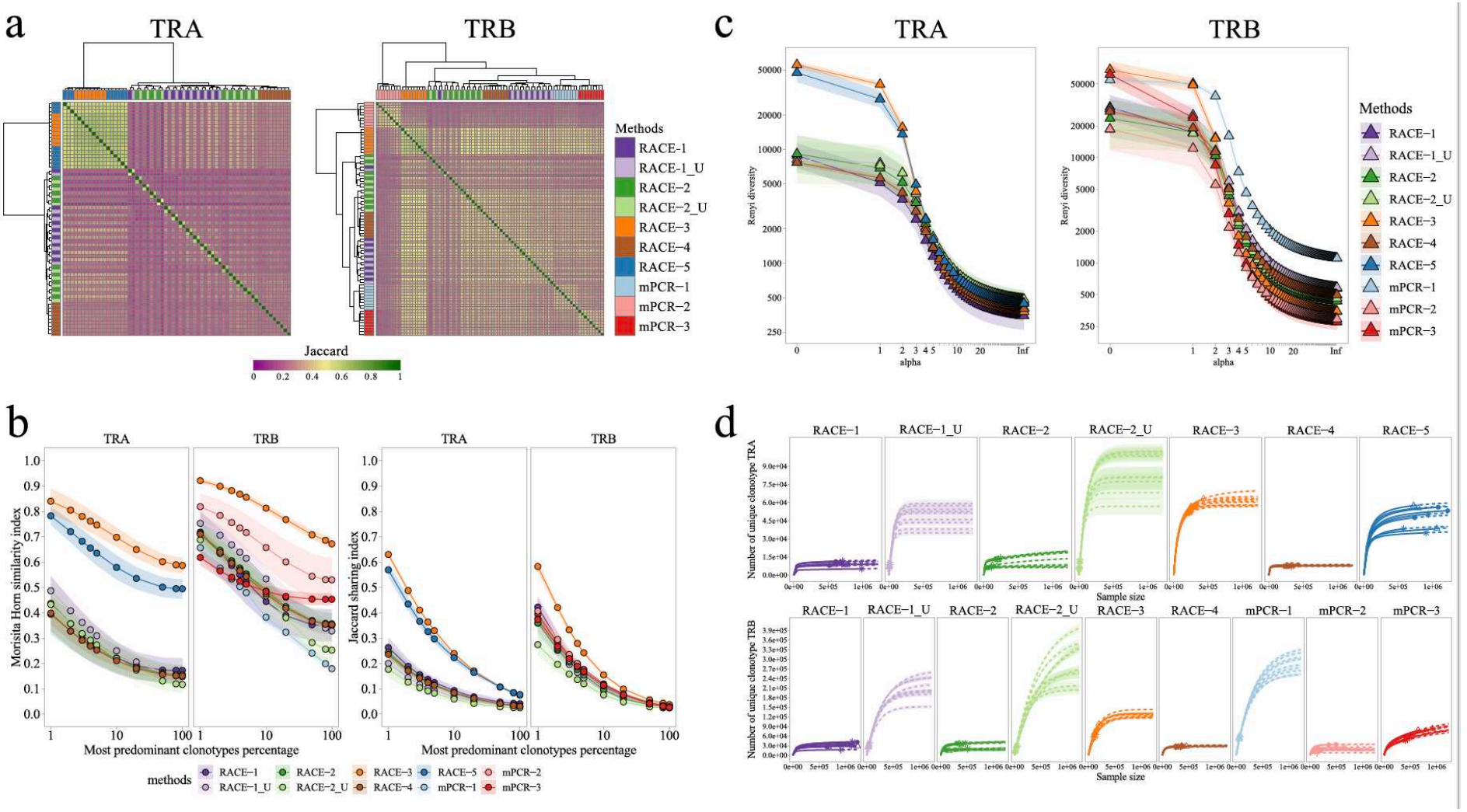


TRBV	Cytometry 1	Cytometry 2	Cytometry 3	Cytometry 4	Cytometry 5	Standard deviation
TRBV2	4,5	4,31	4,55	4,31	4,44	0,109
TRBV3-1	4,02	3,89	3,89	3,82	4,34	0,207
TRBV4-1/2	2,08	2,33	2,11	2,15	2,09	0,103
TRBV4-3	0,23	0,21	0,27	0,18	0,44	0,102
TRBV5-1	2,59	2,78	2,63	2,8	2,83	0,108
TRBV5-5	1,1	1,32	1,11	1,18	1,05	0,105
TRBV5-6	0,89	1,01	0,95	0,93	0,88	0,052
TRBV6-2	1,34	1,54	1,56	1,52	1,32	0,116
TRBV6-5/9	3,09	3,31	3,13	3,19	3,2	0,084
TRBV6-6	1,8	1,94	1,93	1,93	1,85	0,062
TRBV9	3,18	3,04	3,18	3,39	3,13	0,129
TRBV10-3	1,75	1,98	1,82	1,89	1,86	0,085
TRBV11-2	1,52	2,18	1,86	1,78	1,97	0,243
TRBV12-3/4	4,67	4,99	4,62	5,05	4,82	0,190
TRBV13	0,46	0,52	0,52	0,52	0,51	0,026
TRBV14	1,07	1,06	1,08	1,14	1,15	0,042
TRBV18	1,11	1,34	1,2	1,3	1,39	0,113
TRBV19	6,35	6,62	6,54	6,24	6,53	0,156
TRBV20-1	9,51	9,29	9,32	9,46	9,4	0,092
TRBV25-1	1,68	1,65	1,87	1,7	1,54	0,119
TRBV27	2,14	2,17	2,05	2,22	2,1	0,065
TRBV28	4,33	4,54	4,57	4,59	4,66	0,124
TRBV29-1	2,14	2	2	2,15	1,82	0,134
TRBV30	3,56	3,76	3,81	3,96	3,91	0,156

Supplementary Fig. 4: Correlation assessment between flow cytometry and TCRseq TRBV expression level.

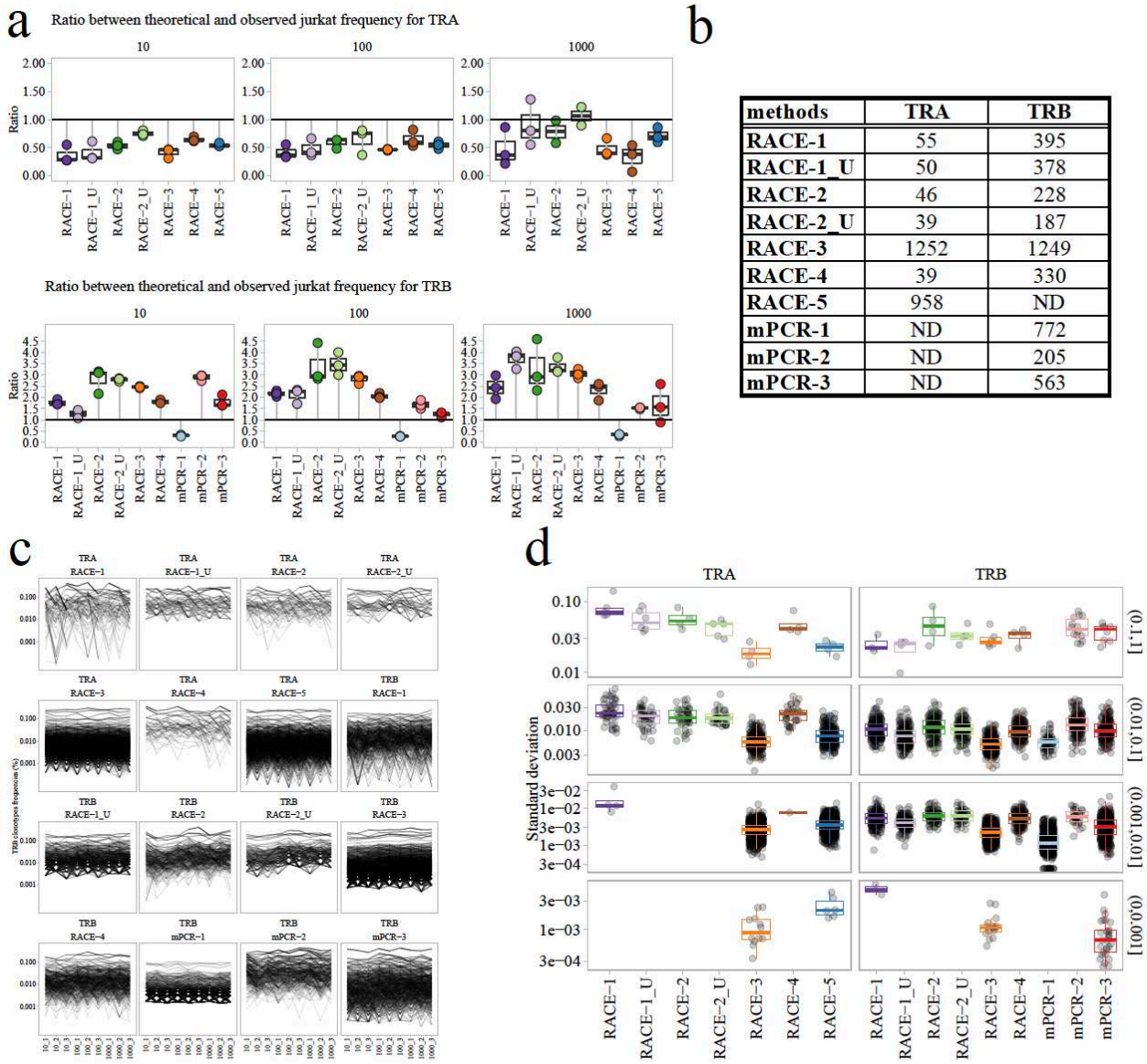
a, TRBV gene repertoire of effector T cells by flow cytometry analysis. Each dotplot corresponds to 3 TRBVs: one TRBV is conjugated to FITC, another one with PE, and the third to both FITC and PE. Percentage of each TRBV subset are shown below the dotplot. Total $CD4^+$ lymphocyte TRBV coverage is 64,03%. **b**, Summary table of the TRBV gene usage and standard deviation analyzed by flow-cytometry for the 5 cellular replicas, each replica is composed of 50 000 cells from the previously harvested effector T cells. **c**, TRBV distribution using experiment A (after Jurkat sequences removal) TCRseq datasets. TRBV frequencies were re-calculated on the top 5% MPC for each method. Only TRBV with significant differences between at least two methods are represented (Two way ANOVA, P-values <

0.05). **d**, Spearman's correlation of the TRBV frequencies between the 5 flow cytometry datasets and the 9 TCRseq replicates was calculated for each method. The plot is represented by the correlation score (y-axis) and the *P*-value (x-axis) of the correlation, allowing the classification of the methods. **e**, Heatmap of the Spearman's correlations between each replicate for the distribution of TRBV gene usage (n=62 unique TRBV genes). The Euclidean distance was used for hierarchical clustering as a color-coded matrix ranging from 0 (yellow, maximum dissimilarity) to 1 (orange, maximum similarity). Jurkat TCR sequences were removed from datasets for this analysis. Jurkat TCR sequences were removed from datasets for this analysis.


a
TRA
TRB
Methods
RACE-1
RACE-1_U
RACE-2
RACE-2_U
RACE-3
RACE-4
RACE-5
mPCR-1
mPCR-2
mPCR-3
Jaccard
0 0.2 0.4 0.6 0.8 1
b
TRA
TRB
TRA
TRB
Morisita Horn similarity index
Jaccard sharing index
Most predominant clonotypes percentage
Most predominant clonotypes percentage
methods
RACE-1
RACE-1_U
RACE-2
RACE-2_U
RACE-3
RACE-4
RACE-5
mPCR-1
mPCR-2
mPCR-3
c
TRA
TRB
Renyi diversity
alpha
Methods
d
RACE-1
RACE-1_U
RACE-2
RACE-2_U
RACE-3
RACE-4
RACE-5
mPCR-1
mPCR-2
mPCR-3
Number of unique clonotype TRA
Number of unique clonotype TRB
Sample size

Supplementary Fig. 5: Most predominant clonotype similarity indices and clonotype diversity.

a, Heatmaps of Jaccard similarity index (JSI). JSI scores were calculated between each replicate across all methods on the top 1% most predominant clonotypes (MPC) for TRA (left) and TRB (right). Euclidean distance was used for hierarchical clustering as a color-coded matrix ranging from 0 (violet, maximal dissimilarity) to 1 (green, maximal similarity). **b**, Morisita-Horn and Jaccard similarity profiles. Similarity profiles computed on increasing percentages of most predominant clonotypes (MPC): from 1% MPC to 100% MPC (the latter corresponding to elsewhere mentioned as "Overall"). For each method, mean similarity metrics were calculated between replicates (n = 9 technical replicates). Shaded areas represent the standard deviation of the mean. **c**, Rényi diversity profile. For each dataset (n = 9 technical replicates), diversity metrics using clonotype frequencies were calculated for increasing values of Rényi order α until stabilization of the resulting diversity (see material and methods) and displayed as a Rényi diversity profile. α varies from 0 (Richness) to ∞ (Berger-Parker); for $\alpha = 1$, the Shannon entropy was computed as described in Rényi[1]. Shaded areas represent the standard deviation of the mean. **d**, Rarefaction curves for TRA (top) and TRB (bottom), representing the number unique of clonotype as a function of the number of TCR sequences (Sample size), are displayed for each method. Solid and dashed lines represent interpolated and extrapolated values, respectively. Dots indicate the observed sample size and richness (number of unique TRA or TRB clonotype). Shaded areas represent 95% confidence intervals. Jurkat TCR sequences were removed from datasets for this analysis.



methods	TRA	TRB
RACE-1	55	395
RACE-1_U	50	378
RACE-2	46	228
RACE-2_U	39	187
RACE-3	1252	1249
RACE-4	39	330
RACE-5	958	ND
mPCR-1	ND	772
mPCR-2	ND	205
mPCR-3	ND	563

Supplementary Fig. 6: Clonotypes shared by all replicates per method details.

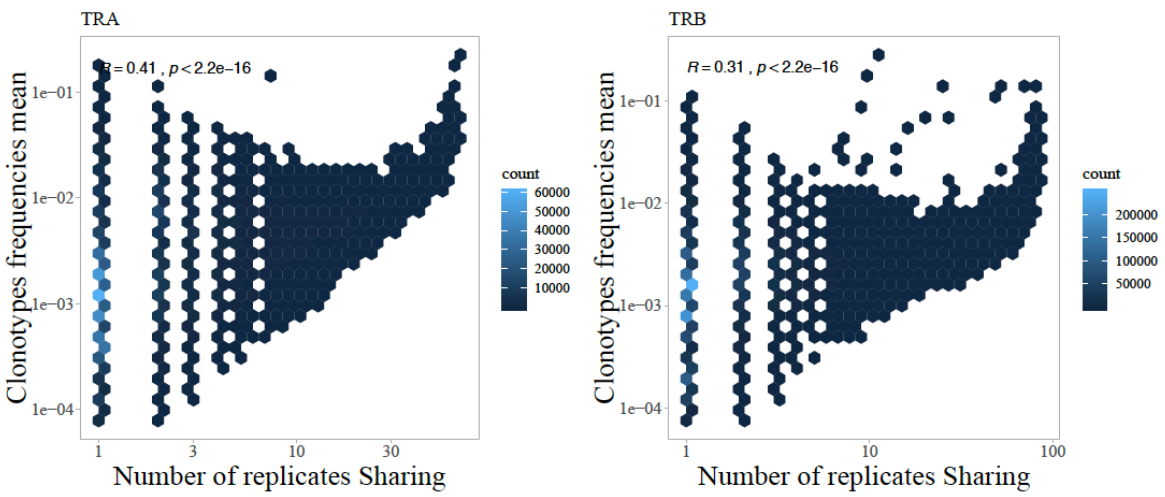
a, Quantification of Jurkat TRA and TRB clonotype detection per method. The plots represent for each method and each dilution (n = 3 technical replicates per dilution), the ratio between the observed and the expected frequencies of TRA (top) and TRB (bottom) clonotypes from the Jurkat spike-in. **b**, Number of unique clonotypes shared by all replicates per method, for TRA and TRB. **c**, Tracking frequencies distribution of these shared clonotypes. **d**, Standard deviations of frequencies of clonotypes shared between all the replicates (n=9) per method, split into frequency ranks, for TRA (left) and TRB (right). Jurkat

TCR sequences were removed from datasets for this analysis. Boxplots visualize five summary statistics: the median, two hinges (the lower and upper hinges correspond to the first and third quartiles) and two whiskers (the upper whisker extends from the hinge to the largest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge).

a

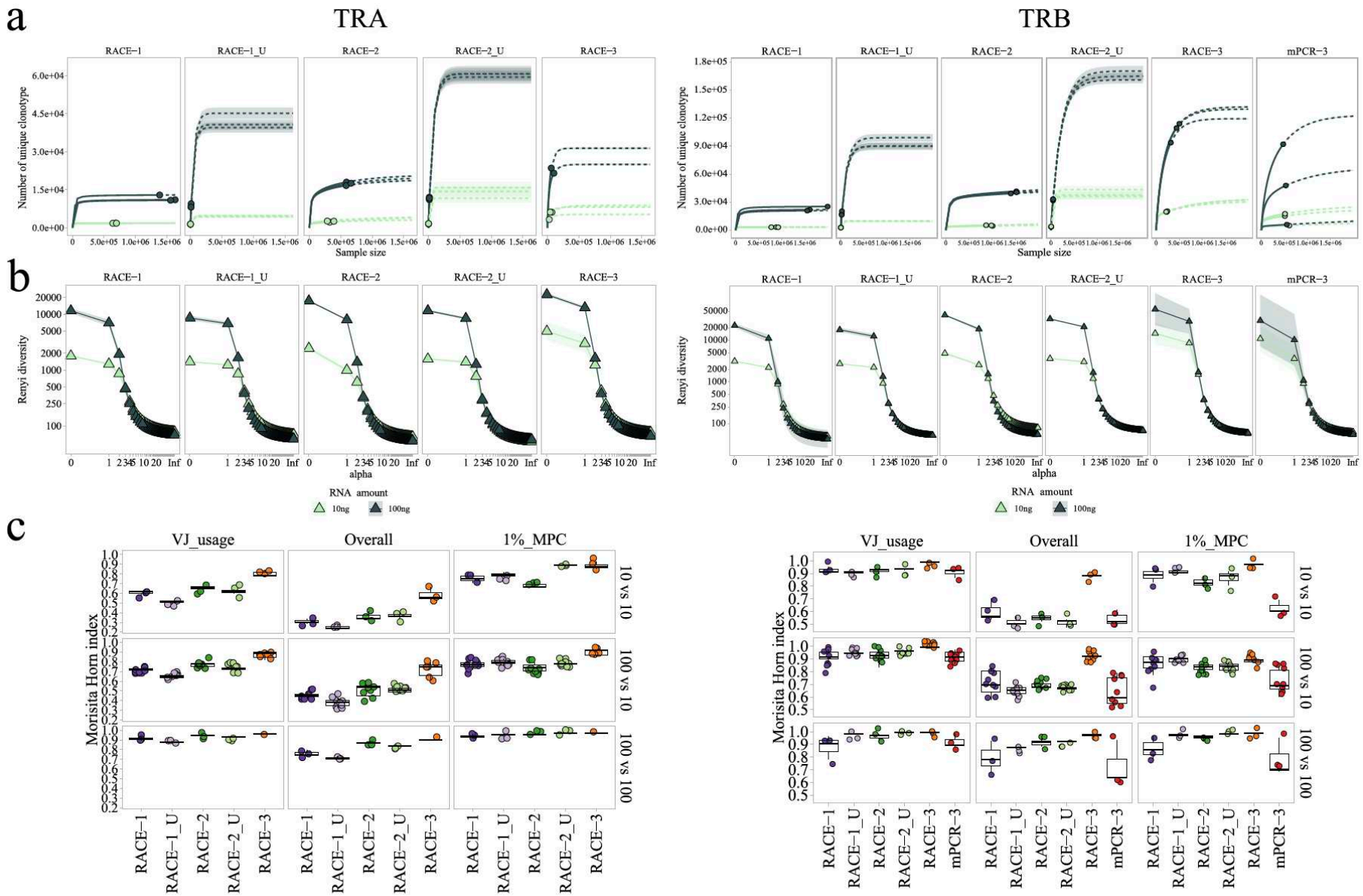
Number of Replicates Sharing	TRA		TRB	
	Number of clonotypes	Percentage of clonotypes	Number of clonotypes	Percentage of clonotypes
1	473758	68,2571	1476606	74,0494
2	118976	17,1416	358103	17,9583
3	38420	5,5354	81067	4,0654
4	19641	2,8298	31626	1,5860
5	10951	1,5778	13765	0,6903
6	7348	1,0587	7790	0,3907
7	5061	0,7292	4907	0,2461
8	3726	0,5368	3395	0,1703
9	2783	0,4010	2435	0,1221
10	2138	0,3080	1887	0,0946
11	1712	0,2467	1500	0,0752
12	1444	0,2080	1128	0,0566
13	1056	0,1521	998	0,0500
14	899	0,1295	792	0,0397
15	776	0,1118	668	0,0335
16	610	0,0879	588	0,0295
17	592	0,0853	524	0,0263
18	496	0,0715	482	0,0242
19	416	0,0599	430	0,0216
20	381	0,0549	382	0,0192
21	306	0,0441	297	0,0149
22	251	0,0362	303	0,0152
23	244	0,0352	255	0,0128
24	203	0,0292	245	0,0123
25	197	0,0284	228	0,0114
26	169	0,0243	212	0,0106
27	159	0,0229	183	0,0092
28	141	0,0203	184	0,0092
29	109	0,0157	177	0,0089
30	107	0,0154	152	0,0076
31	93	0,0134	143	0,0072
32	89	0,0128	138	0,0069
33	83	0,0120	120	0,0060
34	60	0,0086	117	0,0059
35	76	0,0109	120	0,0060
36	60	0,0086	116	0,0058
37	55	0,0079	100	0,0050
38	48	0,0069	106	0,0053
39	43	0,0062	73	0,0037
40	29	0,0042	98	0,0049
41	37	0,0053	89	0,0045
42	35	0,0050	69	0,0035
43	30	0,0043	69	0,0035
44	29	0,0042	55	0,0028
45	21	0,0030	59	0,0030
46	23	0,0033	62	0,0031
47	24	0,0035	70	0,0035
48	14	0,0020	59	0,0030
49	15	0,0022	53	0,0027
50	18	0,0026	48	0,0024
51	23	0,0033	46	0,0023
52	9	0,0013	44	0,0022
53	13	0,0019	50	0,0025
54	7	0,0010	33	0,0017
55	6	0,0009	49	0,0025
56	13	0,0019	42	0,0021
57	12	0,0017	43	0,0022
58	8	0,0012	37	0,0019
59	8	0,0012	38	0,0019
60	7	0,0010	40	0,0020
61	5	0,0007	32	0,0016
62	7	0,0010	37	0,0019
63	9	0,0013	38	0,0019
64	X	X	27	0,0014
65	X	X	29	0,0015
66	X	X	37	0,0019
67	X	X	36	0,0018
68	X	X	32	0,0016
69	X	X	28	0,0014
70	X	X	41	0,0021
71	X	X	23	0,0012
72	X	X	41	0,0021
73	X	X	19	0,0010
74	X	X	24	0,0012
75	X	X	15	0,0008
76	X	X	13	0,0007
77	X	X	22	0,0011
78	X	X	18	0,0009
79	X	X	19	0,0010
80	X	X	24	0,0012
81	X	X	31	0,0016
Total	694079	100	1994081	100

b



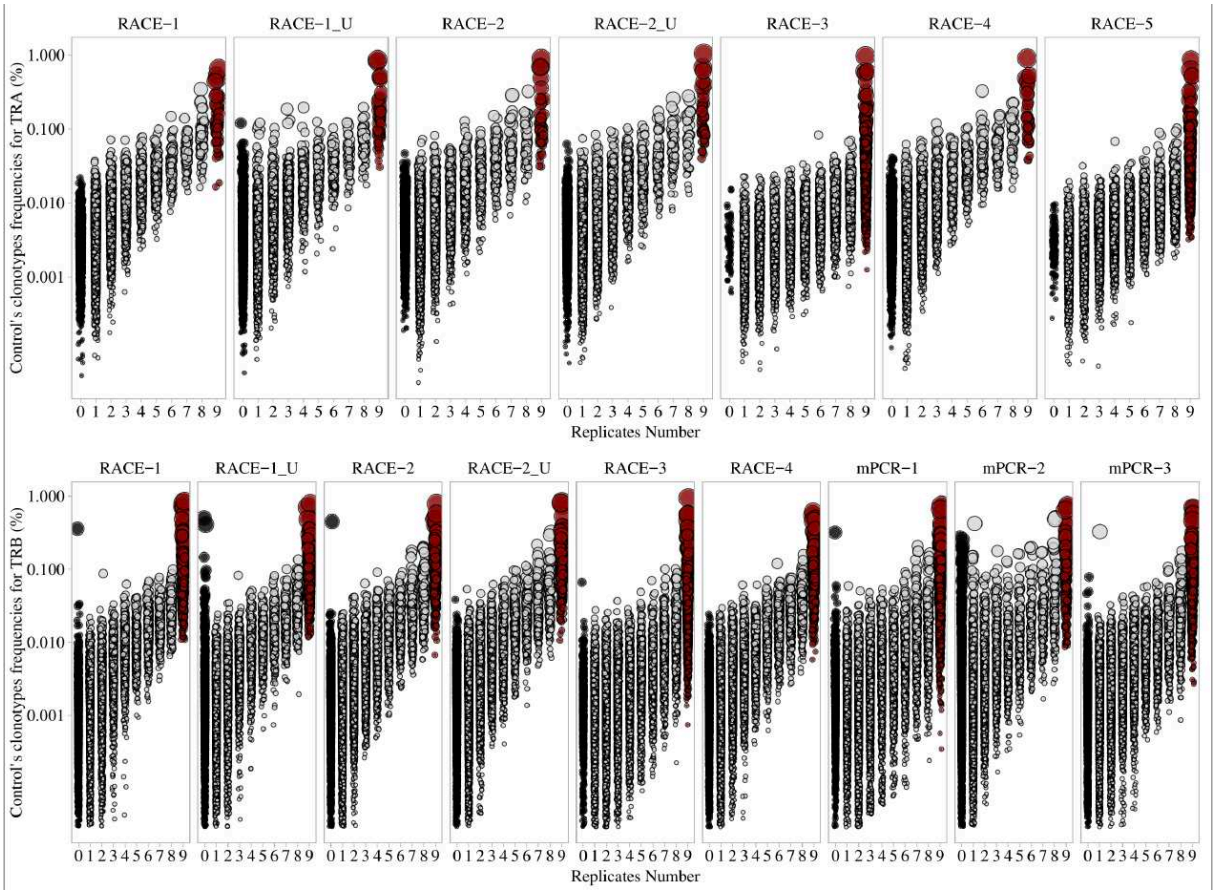
Supplementary Fig. 7: Clonotypes sharing by all replicates and abundance correlation with clonotype sharing.

a, Table of the clonotypes sharing distribution. From all the clonotypes identified across all the replicates for all the methods without any filter on clone size (694 079 TRA clonotypes and 1 994 081 TRB clonotypes), the number and percentage of TRA and TRB clonotypes shared between increasing number of replicates are shown. **b**, Correlation statistics between the average frequency of clonotypes among all replicates and their sharing. To avoid minimizing clonotype frequencies found in single replicates, for each clonotype, the average frequency was calculated as a function of the number of replicates in which it was identified (For example, if clonotype A is found at a frequency of 10% in 1 replicate, the average frequency is 10; while if B is found at 10% in replicate 1 and 7% in replicate 2, the average is 8.5%).


a
TRA
TRB
RACE−1
RACE−1_U
RACE−2
RACE−2_U
RACE−3
mPCR−3
Number of unique clonotype
Sample size
b
Renyi diversity
alpha
RNA amount
10ng
100ng
c
VJ_usage
Overall
1%_MPC
Morisita Horn index
10 vs 10
100 vs 10
100 vs 100

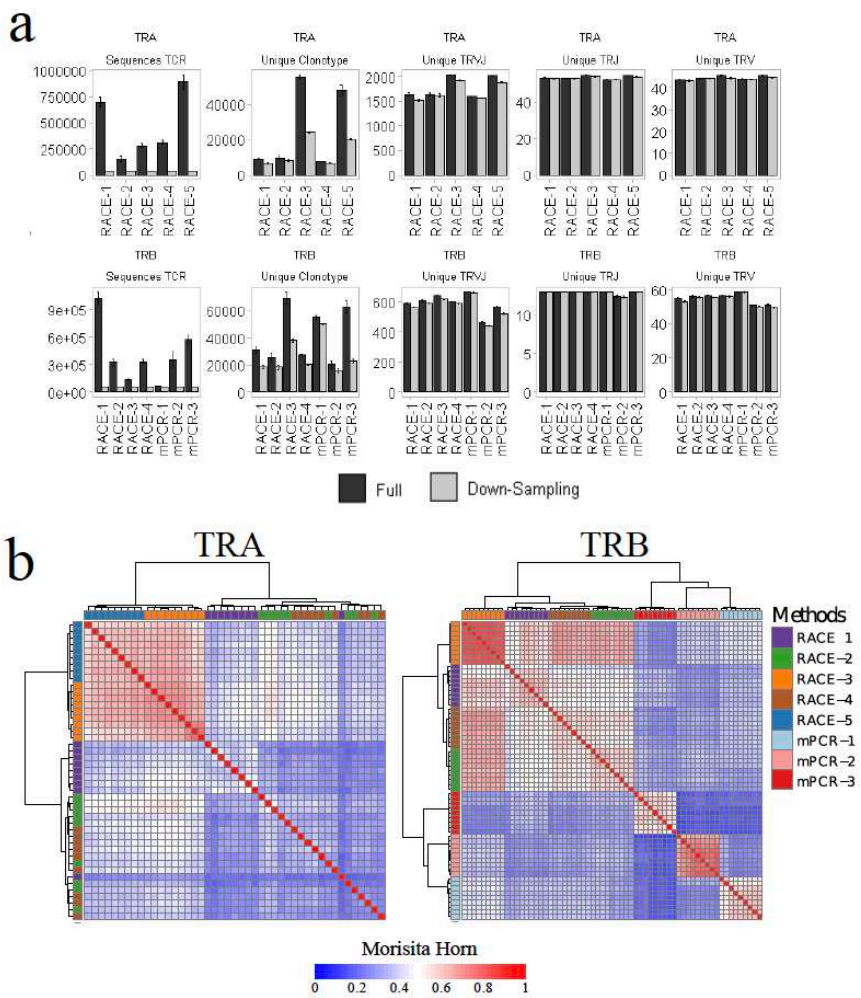
Supplementary Fig. 8: The impact of RNA quantity on TRA and TRB repertoires.

a, Rarefaction curves, representing the number of unique clonotype as a function of the number of TCR sequences (Sample size), are displayed for each method, with 100ng (dark green, n=3 replicates from experiment B) or 10ng (light green, n=3 replicates from experiment B) as input RNA. Solid and dashed lines represent interpolated and extrapolated values, respectively. Dots indicate the observed sample size and richness (number of unique clonotype). Shaded areas represent the 95% confidence interval. **b**, Rényi diversity profile. For each dataset, diversity metrics using clonotype frequencies were calculated for increasing values of Rényi order α until stabilization of the resulting diversity (see material and methods) and displayed as a Rényi diversity profile. α varies from 0 (Richness) to ∞ (Berger-Parker); for α = 1, the Shannon entropy was computed as described in Rényi[1]. Each graph represents the profiles obtained from data generated from 100ng (dark green, n=3 replicates from experiment B) or 10ng (light green, n=3 replicates from experiment B) of input RNA. Shaded areas represent the standard deviation of the mean. **c**, Morisita-Horn scores were calculated between 10ng (top, n=3 replicates from experiment B), 100 ng (middle, n=3 replicates from experiment B) and 10 vs. 100ng (bottom, n=6 replicates from experiment B) replicates on V-J usage, overall clonotypes and the 1% MPC. Results are presented for TRA in the left panel and for TRB in right panel. Boxplots visualize five summary statistics: the median, two hinges (the lower and upper hinges correspond to the first and third quartiles) and two whiskers (the upper whisker extends from the hinge to the largest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge).



Supplementary Fig. 9: Detailed distribution of meta-repertoire clonotypes in the replicates by method for TRA and TRB.

Distribution of meta-repertoire clonotypes by replicates sharing number for TRA (top) and TRB (bottom). Black dots (0) represent the unseen clonotypes and red dots (9) represent the clonotypes found by the 9 replicates. Each method is represented independently. Jurkat TCR sequences were removed from datasets for this analysis.


a
TRA
Sequences TCR
Unique Clonotype
Unique TRVJ
Unique TRJ
Unique TRV
TRB
RACE-1
RACE-2
RACE-3
RACE-4
RACE-5
mPCR-1
mPCR-2
mPCR-3
Full
Down-Sampling
b
TRA
TRB
Methods
RACE 1
RACE–2
RACE–3
RACE–4
RACE–5
mPCR–1
mPCR–2
mPCR–3
Morisita Horn
0
0.2
0.4
0.6
0.8
1
c
TRA
TRB
Morisita Horn similarity index
Jaccard sharing index
Most predominant clonotypes percentage
methods
RACE-1
RACE-2
RACE-3
RACE-4
RACE-5
mPCR-1
mPCR-2
mPCR-3
d
TRA
TRB
Control's clonotypes sharing (%)
RACE-1
RACE-2
RACE-3
RACE-4
RACE-5
mPCR-1
mPCR-2
mPCR-3

Supplementary Fig. 10: Down-sampling analysis.

We performed such a down-sampling based on the smallest sample (except UMI-processed methods) in terms of TCR sequences for TRA and TRB separately (37599 TRA and 57468 TRB). **a**, Histogram plots represent from left to right the mean values +/- SD of TCR sequences, unique clonotype (V-CDR3aa-J), V-J combination (Unique TRVJ), J gene (Unique TRJ) and V gene (Unique TRV) obtained by each method for both TRA (top) and TRB (bottom) chains, for the full datasets (n = 9) in dark and the down-sampled datasets (n = 9) in grey. **b**, Heatmaps of the Morisita-Horn similarity index (MH). MH scores were calculated between each replicate across all methods for the top 1% of most predominant clonotypes (MPC) for TRA (left) and TRB (right). The Euclidean distance was used for hierarchical clustering as a color-coded matrix ranging from 0 (blue, maximum dissimilarity) to 1 (red, maximum similarity). **c**, Morisita-Horn and Jaccard similarity profiles. Similarity profiles computed on increasing percentages of most predominant clonotypes (MPC): from 1% MPC to 100% MPC (the latter corresponding to elsewhere mentioned as “Overall”). For each method, similarity metrics were calculated between replicates (n = 9). Shaded areas represent the standard deviation of the mean. **d**, We also generated a meta-repertoire using the sampled repertoires. Replicate (n = 9) sharing fraction in meta-repertoire (focus on meta-repertoire clonotypes) for TRA (left) and TRB (right). The values represented correspond to the percentage of clonotypes from each replicate per method found in the meta-repertoire, median and the 1st and 3rd quartiles are shown. As you can see, despite this drastic downsampling for some of the methods, the results are very much comparable with those obtained on the full datasets. Boxplots visualize five summary statistics: the median, two hinges (the lower and upper hinges correspond to the first and third quartiles) and two whiskers (the upper whisker extends from the hinge to the largest value no further than

1.5*IQR from the hinge (where IQR is the inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge).

Supplementary table legends

Protocol	Multiplex-1	Multiplex-2	Multiplex-3	RACE-1	RACE-2	RACE-3	RACE-4	RACE-5	RACE-6
Methods identification	Adaptive Technology, Robins, H.S. et al. Blood (2009)	iMonitor Zhang, W. et al. Genetics (2015)	iRepertoire Wang, C. et al. PNAS (2010)	Chiou, B team Oakes, T. et al. Front. Immunol. (2017) Uddin, I. et al. Methods Mol. Biol. (2018)	MGLaboratory LLC Mamedov, I.Z. et al. Front. Immunol. (2013)	Takara Taylor, S. et al. J. Immunol. (2016)	Douek D team Douek, D.C. et al. J. Immunol. (2002)	Eugster, A. team Eugster, A. et al. J. Immunol. Methods (2013)	Invitrogen
TCR chain	Beta	Beta	Beta	Alpha & Beta (independent reactions)	Alpha & Beta (independent reactions)	Alpha & Beta (Same reaction)	Alpha & Beta (independent reactions)	Alpha	Beta
library_layout	paired	single	paired	paired	paired	single	paired	paired	paired
Starting material	gDNA	gDNA	RNA	RNA	RNA	RNA	RNA	RNA	RNA
PCR method	m-PCR	m-PCR	amx-PCR	TRAB RACE	TRAB RACE	TRAB RACE	TRAB RACE	TRA RACE	TRB RACE
Forward Primer targets	TRBV genes n=24	TRBV genes n=24	TRBV genes n=24	Anchor primer (target added by ligation)	NNV Oligo (target added by template switching) - MGLaboratory LLC	SMART-Seq v4 Oligonucleotide (target added by template switching) - Takara Bio	SPIA anchor primer (target added by template switching)	Template Switch Oligo (target added by template switching)	Abridged Anchor Primer (target added by nucleotidyltransferase (TdT enzyme))
TSO+-seq length (nt)	Not applicable	Not applicable	Not applicable	58	Proprietary (Not disclosed)	Proprietary (Not disclosed)	23	37	30
Reverse primer target	TRBJ n=12	TRBJ n=13	TRBC	TRAC & TRBC	TRAC & TRBC	TRAC & TRBC	TRAC & TRBC	TRAC	TRBC
UMI (yes/no)	No	No	No	Yes	Yes	No	No	No	No
UMI-size (nt)	NA	NA	NA	12	12	NA	NA	NA	NA
RT (Yes/no)	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reverse Transcriptase	Not applicable	Not applicable	Proprietary (Not disclosed)	Superscript IV (Invitrogen)	SMARTScribe (Takara Bio)	SMARTScribe (Takara Bio)	SuperScript II (Invitrogen)	SuperScript II (Invitrogen)	SuperScript II (Invitrogen)
Number of PCR	2	1	2	3	2	2	3	3	1
PCR1 cycle number	30	30	25	4	18	20	20	15	35
PCR2 cycle number	7	NA	30	6	10	20	5	11	NA
PCR3 cycle number	NA	NA	NA	10-15 (real time PCR)	NA	NA	3	25	NA
Polymerase type	HotStarTaq Plus DNA Polymerase, (Qiagen)	Ion AmpliSeq HiFi Mix, NEBNext Fast DNA Library Prep Set	Proprietary (Not disclosed)	Phusion High Fidelity (New England Biolabs)	Q5 HotStart High-Fidelity DNA Polymerase (New England Biolabs)	SeqAmp DNA polymerase (Takara Bio)	KAPA HiFi HotStart (Roche)	PrimeSTAR HS DNA Polymerase (Takara Bio)	*Taq* DNA Polymerase (invitrogen)
Purification method	PCR Cleanup Beads (Adaptive Biotechnologies)	AMPure XP beads	AMPure XP beads + gel migration	AMPure XP beads	AMPure XP beads	AMPure XP beads	AMPure XP beads	QIAquick PCR Purification Kit	AMPure XP beads
Amplicons length	400bp	100-2000bp	520bp	670-710 bp	670-710 bp	700bp	550	670bP	620bp
Sequencing chemistry used	MiSeq v3 2x300	Ion Torrent S5 540 Chef Kit (2000bp)	MiSeq v2 2x250* and v3 2x300**	MiSeq v2 2x250	MiSeq v3 2x300** and v2 2x250***	MiSeq v2 2x250* and v3 2x300**	MiSeq v2 2x150	MiSeq v2 2x250	MiSeq v2 2x250* and v3 2x300**
PhiX percentage	5%	NA	10%	15-20%	20%	10%	20%	30%	10%
Pre-processing method (Software)	Proprietary (Not disclosed)	none	Proprietary (Not disclosed)	Decombinator for UMI	MiGEC for UMI	NA	NA	NA	NA
UMI threshold (Meaning UMI cut-off (reads/UMI))	NA	NA	NA	NONE	4	NA	NA	NA	NA
Alignment/annotation tool	MiXCR (our-seq alignment parameter)								

*TSO : Template Switching Oligo
** Sequencing chemistry used in Experiment A
*** Sequencing chemistry used in Experiment B

Supplementary Table 1: Methods summary table.

Table summarizing each protocol by method.

Dilution	Replicates	TRA							TRB								
		RACE-1	RACE-1_U	RACE-2	RACE-2_U	RACE-3	RACE-4	RACE-5	mPCR-1	mPCR-2	mPCR-3	RACE-1	RACE-1_U	RACE-2	RACE-2_U	RACE-3	RACE-4
1/1000	1	0,0369	0,0554	0,0583	0,0896	0,0376	0,0382	0,0598	0,0278	0,1479	0,2594	0,1931	0,4038	0,2315	0,3203	0,3262	0,2477
	2	0,0217	0,0802	0,0982	0,1224	0,0402	0,0543	0,0687	0,0352	0,1548	0,0897	0,2977	0,3264	0,2923	0,3140	0,2861	0,1867
	3	0,0864	0,1359	0,0000	0,0000	0,0668	0,0071	0,0861	0,0358	0,1545	0,1562	0,2437	0,3829	0,4592	0,3774	0,3021	0,2597
1/100	1	0,3600	0,3673	0,6380	0,8051	0,4493	0,8192	0,4801	0,2798	1,5019	1,2813	2,0399	2,3331	2,8327	2,9997	2,6046	2,1898
	2	0,5575	0,6650	0,4841	0,3696	0,4783	0,5336	0,6064	0,2652	1,6500	1,1181	2,2980	1,7128	2,9392	3,4259	2,9580	2,0251
	3	0,3328	0,4149	0,6372	0,7570	0,4683	0,5995	0,5498	0,2616	1,8794	1,3268	2,1735	2,2762	4,4213	4,0048	2,9174	1,9785
1/10	1	2,7366	3,1731	6,0191	8,0655	3,1161	6,1469	5,3301	3,2976	27,3554	21,1914	17,1010	12,9059	21,5958	27,0760	24,1411	17,9631
	2	5,5243	6,1062	4,7304	7,0884	4,6755	6,9603	5,2839	3,3055	29,7710	16,4809	19,1637	14,3869	31,5247	28,4652	25,0016	18,9881
	3	2,8056	3,1257	5,3871	7,3826	4,6158	6,2865	5,8529	2,6962	29,6532	16,5424	16,9448	10,8251	30,9180	28,1424	24,5969	17,4689

Supplementary Table 2: Jurkat clone percentages per method.

Each value represents the percentage of the Jurkat TRA (left) and TRB (right) clonotype observed in each replicate from the 1/1000 (top), 1/100 (middle) and 1/10 (bottom) dilution condition for each method (column).

Supplementary Material and methods

- **Multiplex-1**

DNA was sent to Adaptive technology® (USA) for TCR amplification following their Survey protocol[2,3]. Briefly, from genomic DNA (gDNA), each TCR rearrangement is amplified using a set of 24 forward primers, each targeting one TRBV human gene, and 12 reverse TRBJ primers targeting each TRBJ gene to ensure a complete coverage of the CDR3 sequence. A 30 cycle PCR1 leads to amplicons further purified. 2µL of such purified amplicons are amplified in a second round of PCR (7 cycles) where sample barcodes are added. PCR2 amplicons are further purified. The sequencing is then carried out on a MiSeq V3 Illumina sequencer using a 2x300bp read length protocol by the provider.

- **Multiplex-2**

TCRB sequencing was performed using primers from the IMonitor publication[4] and a protocol adapted for Ion Torrent sequencing. Primer stocks were made by adding 2µl of each 100µM TCR-J primer to 176µl of PCR grade water and 2µl of each 100µM TCR-V Primer to 136µl of PCR grade water creating a stock solution 1µM for each primer. The PCR reactions contained 2µl Ion Ampliseq 5x HiFi PCR Mastermix (Thermo Fisher Scientific), 2µl of each primer stock and 40ng DNA with water to a final volume of 10µl. PCR cycling conditions were 99°C for 2 min, then 25 cycles of 99°C for 15 sec and 65°C for 60sec, followed by a hold at 10°C. PCR products were purified by AMPure XP beads using a 1,8x volume of beads and elution in 25µl of water. Subsequently the PCR products were end-repaired, phosphorylated and ligated to IonXpress adapters using the NEBNext Fast DNA Library Prep Set for Ion torrent (NEB CatNr E6270L) according to manufacturer's recommendations. A post-amplification was applied with 8 cycles followed by another round of AMPure cleanup. Molarity of the library was determined by Ion Library TaqMan

Quantitation Kit (Thermo Fisher CatNr 4468802). The library was sequenced on Ion Torrent S5XL using the 530 Chip Kit and the Ion 510 & Ion 520 & Ion 530 Kit – Chef (Thermo Fisher CatNr A34019) to read sequences up to 400bp length and to obtain 0.5-1.5x106 reads per sample.

- **Multiplex-3**

RNA has been processed with the iProfile-human TRB kit (iRepertoire®). Briefly, each TRB is reverse-transcribed using a set of 24 forward primers, each targeting one TRBV human gene, and a reverse primer located in the TRBC gene to ensure a complete coverage of the TRB sequence. PCR1 primers include barcodes to allow sample identification after the sequencing. PCR1 products are then purified on magnetic beads and a second round of PCR is performed. Amplified libraries are excised from agarose gel and purified[5]. The sequencing is then carried out on a MiSeq V2 Illumina sequencer using a 2x250bp read length protocol for the experiment A and a MiSeq V3 Illumina sequencer using a 2x300bp read length protocol for the experiment B at "Institut du Cerveau et de Moelle epinière" (ICM) to obtain 0.5-1.5x106 reads per sample. Details can be found on the iRepertoire resources document available online (https://docs.wixstatic.com/ugd/c9f231_a02ccf9aa124475e86a7c384c810c077.pdf)

- **RACE-1**

The alpha and beta chains of the TCR repertoire were sequenced by amplifying cDNA generated from the 10ng and 100ng RNA samples provided. Details of the method are published[6,7]. Briefly, the RNA is reverse transcribed from a TCR constant region primer. A oligonucleotide containing an anchor primer, together with a 12 base pair UMI is ligated onto the 3' end of the cDNA using the single strand T4 RNA ligase 1 (NEB, #MO204). Subsequent PCR reactions are then used to amplify the TCR cDNA (from the 3' end of the

constant region until the 3' end of the RNA), and to introduce adaptors for Illumina sequencing and indices for sample multiplexing. The sequencing is then carried out on a MiSeq V3 Illumina sequencer using a 2x300bp read length protocol for both experiment to obtain 0.5-1.5x10^6 reads per sample.

- **RACE-2**

TCR alpha and beta libraries were prepared on 10 or 100ng of RNA from each sample with Human TCR kit (MiLaboratory® LLC) following provider protocol. Briefly, it is a 5'RACE protocol with template switch oligo that carries 12 random "N" nucleotides (UMI). Reverse transcription starts from the primers located in the TRBC and TRAC genes, in one tube. 1st PCR amplifies cDNA with a forward primer annealing on the template switch oligo, and reverse nested primers annealing on TRBC and TRAC genes. 2nd PCR is performed separately for TCR alpha and beta chains and introduces Illumina adapters with indexing. PCR product is purified using AMPUre XP beads (Beckman-Coulter®), analyzed with real-time PCR. The sequencing is then carried out on a MiSeq V3 Illumina sequencer using a 2x300bp read length protocol for the experiment A and a MiSeq V2 Illumina sequencer using a 2x250bp read length protocol for the experiment B to obtain 0.5-1.5x10^6 reads per sample.

- **RACE-3**

T cell receptor (TCR) alpha/beta libraries were prepared on 10 or 100ng of RNA from each sample with SMARTer Human TCR a/b Profiling Kit (Takarabio®) following provider protocol[8]. Briefly, the reverse transcription was performed using a mixture of TRBC and TRAC reverse primers and further extended with a template-switching oligonucleotide (SMART-Seq® v4). cDNAs were then amplified following two semi-nested PCR: a first PCR with TRBC and TRAC reverse primers as well as a forward primer hybridizing to the SMART-Seq v4 sequence added by template-switching and a second PCR targeting the PCR1 amplicons with reverse

and forward primer including Illumina Indexes allowing for sample barcoding. PCR2 are then purified using AMPUre XP beads (Beckman-Coulter®). The sequencing is then carried out on a MiSeq V2 Illumina sequencer using a 2x250bp read length protocol for the experiment A and a MiSeq V3 Illumina sequencer using a 2x300bp read length protocol for the experiment B at "Institut du Cerveau et de Moelle epinière" (ICM) to obtain 0.5-1.5x10^6 reads per sample.

- **RACE-4**

T cell receptor genes were sequenced using a method similarly as described in Gros et al. 2014[9]. In brief, extracted RNA was denatured and cDNA was synthesized using an oligo (dT) primer and a template-switching oligo during an incubation at 42°C for 90 minutes then 70°C for 10 minutes. Following purification with AMPure XP beads (Beckman-Coulter®), cDNA was amplified using the 5PIIA primer (5'- AAGCAGTGGTATCAACGCAGAGT-3') and constant region primers TCRb (5'-TGCTTCTGATGGCTCAAACACAGCGACCT-3') or TCRa (5'-TCTCAGCTGGTACACGGCAGGGTCAGGGT-3') with the following cycling conditions: (95°C 5 min, 5 cycles of 98°C 15 sec, 72°C 1 min, 5 cycles of 98°C 15 sec, 70°C 10 sec, 72°C 1 min, 10-15 cycles of 98°C 15 sec, 68°C 10 sec, 72°C 1 min). The amplicons were purified using AMPure XP beads, amplified further to incorporate Illumina sequences. The sequencing is then carried out on a MiSeq V2 Illumina sequencer using a 2x150bp read length protocol for both experiment to obtain 0.5-1.5x10^6 reads per sample.

- **RACE-5**

First-strand cDNA was synthesized by 5' RACE. RNA, 1 mM dNTP and 0.125 □M final of the 3' primer binding to the TCRa C region (5'-CACTGTTGCTCTTGAAGTCC-3') were denatured for 5', 65°C. A mix containing 2.5 □M final of the template switching primer (5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTrGrGrG-3'), 2 mM DTT, 20 U rRNAsin

(Promega®, Madison, WI), 50 mM TrisHCl (pH 8), 187 mM KCl, 3 mM MnCl2, and 0.05% Tween 20 was added to a final volume of 20 µl. After preincubation (2', 42°C), 200 U SuperScript II Reverse Transcriptase (Invitrogen®, Life Technologies) were added, incubation was continued for 90', 42°C, and 15', 70°C. cDNA was purified using the MinElute PCR Purification Kit (Qiagen®). Whole cDNA was amplified over 3 rounds PCR with PrimeSTAR HS DNA Polymerase (TAKARA®, Japan), allowing the addition of barcodes and adaptors for Illumina sequencing. Primers: 1st: 5'-TCGGTGAATAGGCAGACAGA-3' and 5'-GTGACTGGAGTTCAGACGTG-3', 2nd: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNGCAGGGTCAGGGTTCTGGAT-3' and 5'-CAAGCAGAAGACGGCATACGAGATindexGTGACTGGAGTTCAGAC-3' and 3rd: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC-3' and 5'-CAAGCAGAAGACGGCATACGAGATindexGTGACTGGAGTTCAGAC-3'. Amplification was for 15, 12, and 25 cycles with the following conditions: 1st: 50 µl total (98°C 10 min, 56°C 7min, 72°C 1 h 50min), 2nd: 1 µl in 20 µl total (98°C 10 min, 49*°C 7 min, 72°C 1 h 50 min) (asterisk [*] represents 2°C increment per cycle starting at 49°C), and 3rd: 1 µl 1:100 in 50 µl total (98°C 10 min, 58°C 7 min, 72°C 1 h 50 min). The final product was purified using AMPUre XP beads (Beckman-Coulter®). The sequencing is then carried out on a MiSeq V2 Illumina sequencer using a 2x250bp read length protocol for both experiment to obtain 0.5-1.5x10^6 reads per sample.

- **RACE-6**

This method of amplification was performed using the 5'RACE kit, Version 2.0 (Invitrogen®) according to the supplier's recommendations. Briefly, the first strand cDNA was synthesized using a constant region-specific TRBC-RT primer (5'-CACGTGGTCGGGGWAGAAGC-3'). RT and PCR amplification were performed using manufacturer recommendations on 10 or 100ng

RNA. Libraries were purified by using AMPUre XP beads (Beckman-Coulter®). The sequencing is then carried out on a MiSeq V2 Illumina sequencer using a 2x250bp read length protocol for the experiment A and a MiSeq V3 Illumina sequencer using a 2x300bp read length protocol for the experiment B at "Institut du Cerveau et de Moelle epinière" (ICM) to obtain 0.5-1.5x10^6 reads per sample.

Supplementary references

1. Rényi, A. On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Stat. Probab. Univ. Calif. Press* **1**, 547–561 (1961).
2. Carlson, C. S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, (2013).
3. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**, 4099–107 (2009).
4. Zhang, W. *et al.* IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis. *Genetics* **201**, 459–472 (2015).
5. Wang, C. *et al.* High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl. Acad. Sci.* **107**, 1518–1523 (2010).
6. Oakes, T. *et al.* Quantitative Characterization of the T Cell Receptor Repertoire of Naïve and Memory Subsets Using an Integrated Experimental and Computational Pipeline Which Is Robust, Economical, and Versatile. *Front. Immunol.* **8**, (2017).
7. Uddin, I. *et al.* An Economical, Quantitative, and Robust Protocol for High-Throughput T Cell Receptor Sequencing from Tumor or Blood. in *Cancer Immunosurveillance: Methods*

and Protocols (eds. López-Soto, A. & Folgueras, A. R.) 15–42 (Springer New York, 2019). doi:10.1007/978-1-4939-8885-3_2.

8. Taylor, S., Yasuyama, N. & Farmer, A. A SMARTer approach to profiling the human T-cell receptor repertoire. *J. Immunol.* **196**, 209.5-209.5 (2016).
9. Gros, A. *et al.* PD-1 identifies the patient-specific $CD8^+$ tumor-reactive repertoire infiltrating human tumors. *J. Clin. Invest.* **124**, 2246–2259 (2014).

Article 3: “Cross-reactivity between MHC class I-restricted antigens from cancer cells and intestinal bacteria.” (Fluckiger et al., Science)

Title: Crossreactivity between MHC class I-restricted antigens from cancer cells and an enterococcal bacteriophage.

Authors: Aurélie Fluckiger[1-3], Romain Daillère[1-3], Mohamed Sassi[4], Barbara Susanne Sixt[5,6-10], Peng Liu[6-10], Friedemann Loos[6-10], Corentin Richard[11-13], Catherine Rabu[17,18], Maryam Tidjani Alou[1,2,14], Anne-Gaëlle Goubet[1,2], Fabien Lemaitre[1], Gladys Ferrere[1,2] , Lisa Derosa[1,2,14], Connie PM Duong[1,2], Meriem Messaoudene[15], Andréanne Gagné[15], Philippe Joubert[15], Luisa De Sordi[16], Laurent Debarbieux[16], Sylvain Simon[17,18], Clara-Maria Scarlata[19], Maha Ayyoub[19], Belinda Palermo[20], Francesco Facciolo[21], Romain Boidot[22], Richard Wheeler[23], Ivo Gomperts Boneca[23], Zsofia Sztupinszki[24], Krisztian Papp[25], Istvan Csabai[25], Edoardo Pasolli[26], Nicola Segata[27], Carlos Lopez-Otin[7-10,28], Zoltan Szallasi[24,29-31], Fabrice Andre[32,33], Valerio Iebba[34], Valentin Quiniou[35,36], David Klatzmann[35,36], Jacques Boukhalil[37], Saber Khelaifia[37], Didier Raoult[37], Laurence Albiges[1,14,38], Bernard Escudier[1,38,39], Alexander Eggermont[1-14], Fathia Mami- Chouaib[40], Paola Nistico[20], François Ghiringhelli[41], Bertrand Routy[15,42], Nathalie Labarrière[17,18], Vincent Cattoir[4,43,44], Guido Kroemer[6-10,45,46,47]*, and Laurence Zitvogel[1-3,14,47]*

Affiliations:
[1] Gustave Roussy Cancer Campus (GRCC), Villejuif, France.
[2] Institut National de la Santé et de la Recherche Médicale, U1015, Institut Gustave Roussy, Villejuif, France
[3] Center of Clinical Investigations in Biotherapies of Cancer (CICBT) 1428, Villejuif, France.
[4] Université Rennes 1, Laboratoire de Biochimie Pharmaceutique, Inserm U1230 - UPRES EA 2311, Rennes, France.
[5] Laboratory for Molecular Infection Medicine Sweden, Umeå Centre for Microbial Research, Department of Molecular Biology, Umeå University, 90187, Umeå, Sweden.
[6] Cell Biology and Metabolomics Platforms, Gustave Roussy Cancer Campus, Villejuif, France. [7] Equipe 11 labellisée Ligue contre le Cancer, Centre de Recherche des Cordeliers, Paris, France [8] INSERM U1138, Paris, France.
[9] Université de Paris, Paris, France
[10] Sorbonne Université, Paris, France.
[11] Research Platform in Biological Oncology, Dijon, France.
[12] GIMI Genetic and Immunology Medical Institute, Dijon, France.
[13] University of Burgundy-Franche Comté, Dijon, France.
[14] Université Paris-Saclay, Villejuif, F-94805, France.
[15] Centre de recherche du centre hospitalier de l'université de Montréal (CRCHUM), 900 rue Saint-Denis, H2X 3H8 Montréal, Québec, Canada.
[16] Department of Microbiology, Institut Pasteur, F-75015 Paris, France
[17] CRCINA, INSERM, Université d'Angers, Université de Nantes, Nantes, France.
[18] LabEx IGO "Immunotherapy, Graft, Oncology," Nantes, France.
[19] Cancer Research Centre of Toulouse, INSERM UMR 1037, 31037 Toulouse, France; Université Toulouse III Paul Sabatier, 31330 Toulouse, France; Institut Universitaire du Cancer de Toulouse-Oncopole, 31100 Toulouse, France.
[20] Unit of Tumor Immunology and Immunotherapy, Department of Research, Advanced Diagnostics and Technological Innovation, IRCCS Regina Elena National Cancer Institute, Rome,Italy
[21] Thoracic Surgery Unit, Department of Surgical Oncology, IRCCS Regina Elena NationalCancer Institute, Rome, Italy.
[22] Unit of Molecular Biology - Department of Biology and Pathology of Tumors - Georges-

François Leclerc anticancer center - UNICANCER - Dijon – France
[23] Institut Pasteur, Unit Biology and genetics of the bacterial cell wall, Paris, France
[24] Computational Health Informatics Program (CHIP), Boston Children's Hospital, Boston, MA, USA.
[25] Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary.
[26] Department of Agricultural Sciences, University of Naples Federico II, Naples, Italy
[27] Department CIBIO, University of Trento, Trento, Italy.
[28] Dpto. de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, Oviedo, Spain.
[29] Harvard Medical School, Boston, MA, USA.
[30] Danish Cancer Society Research Center, Copenhagen, Denmark.
[31] MTA-SE-NAP, Brain Metastasis Research Group, 2nd Department of Pathology, Semmelweis University, Budapest, Hungary.
[32] Department of Cancer Medicine, Breast Cancer Committee, Gustave Roussy, Villejuif, France.
[33] INSERM Unit 981, Gustave Roussy, Villejuif, France.
[34] Department of Public Health and Infectious Diseases, Section of Microbiology, Sapienza University of Rome, Rome 00185, Italy.
[35] AP-HP, Hôpital Pitié-Salpêtrière, Clinical Investigation Center in Biotherapy (CIC-BTi) and Immunology-Inflammation-Infectiology and Dermatology Department (3iD), F-75651, Paris, France
[36] Sorbonne Université, INSERM, Immunology-Immunopathology-Immunotherapy (i3), F 75651,Paris, France
[37] URMITE, Aix Marseille Université, UM63, CNRS 7278, IRD 198, INSERM 1095, IHU-Méditerranée Infection, 13005 Marseille, France.
[38] Department of Medical Oncology, Gustave Roussy, Villejuif, France.
[39] INSERM U981, GRCC, Villejuif, France.
[40] INSERM UMR 1186, Integrative Tumour Immunology and Genetic Oncology, Gustave Roussy, EPHE, PSL, Faculté de Médecine, Université Paris-Sud, Université Paris-Saclay, Villejuif, France.
[41] Department of Medical Oncology, Center GF Leclerc, Dijon, France.
[42] Division d'hémato-oncologie, département de médicine, centre hospitalier de l'université de Montréal (CHUM), Montréal, Québec, Canada.
[43] CHU de Rennes - Hôpital Ponchaillou, Service de Bactériologie-Hygiène hospitalière, Rennes, France.
[44] CNR de la Résistance aux Antibiotiques (laboratoire associé 'Entérocoques'), Rennes, France.
[45] Pôle de Biologie, Hôpital Européen Georges Pompidou, Assistance Publique–Hôpitaux de Paris, Paris, France
[46] Department of Women's and Children's Health, Karolinska University Hospital, 1 Stockholm, Sweden.
[47] Suzhou Institute for Systems Biology, Chinese Academy of Medical Sciences, Suzhou, China
*Correspondence to: laurence.zitvogel@gustaveroussy.fr; kroemer@orange.fr

One sentence summary:

Cytotoxic T lymphocytes that recognize antigens from a prophage of a commensal enterococcus can mediate anticancer immunosurveillance by recognizing cross-reactive tumor-associated antigens.

Abstract:

It has been speculated that the intestinal microbiota induces commensal-specific memory T cells that then cross-react with tumor-associated antigens. Here, we identified MHC class I-binding epitopes within the tail length tape measure protein (TMP) of a prophage found in the genome of *Enterococcus hirae.* Mice bearing *E. hirae* strains harboring this prophage mounted a TMP-specific H-2K^{b} restricted CD8^{+} T lymphocyte response upon immunotherapy with cyclophosphamide or anti-PD1 antibodies. Such TMP-specific T cells also recognized a 78%-identical H-2K^{b}-binding peptide derived from the proteasome (20S) subunit beta type-4 (PSMB4), allowing them to control mouse tumors expressing this oncogenic driver. Administration of bacterial strains engineered to express the TMP epitope improved the outcome of immunotherapy. Tumors bearing PSMB4 knock-in mutations that abolish crossreactivity with TMP became immunotherapy-resistant. In renal and lung cancer patients, the presence of the enterococcal prophage in stools, as well as the expression of a TMP-cross reactive antigen by tumors, predicted the long-term benefit of PD-1 blockade. In melanoma patients, we detected T cell clones recognizing naturally processed cancer antigens that are cross-reactive with microbial peptides. Altogether, these results support the idea that intestinal microbe-specific T cell responses contribute to anticancer immunosurveillance.

Main Text:

Unleashing immune responses against tumor-associated antigens through chemotherapy, radiotherapy, targeted therapies or immune checkpoint inhibitors has become the mainstay of successful cancer treatments (*1*, *2*). The recent discovery that the gut microbiota determines the cancer-immune set point, thus influencing the clinical outcome of antineoplastic therapies, has rekindled the concept that microbes or their products modulate not only intestinal but also systemic immunity (*3*, *4*). Indeed, memory responses by interferon-□ □IFN□) secreting CD4^{+} and CD8^{+} T cells specific for *Enterococcus hirae, Bacteroides fragilis,* and *Akkermansia*

muciniphila are associated with favorable clinical outcome in cancer patients (*5–8*), suggesting that microbe- specific T lymphocytes may contribute to antitumor immune responses. The mechanisms through which microbes trigger chronic intestinal inflammation and systemic autoimmune disease have not been resolved (*9*). The theory of molecular mimicry (*10–14*) posits that T cells elicited by bacteria or viruses accidentally recognize autoantigens as they 'escape' from self-tolerance inducing mechanisms (such as clonal deletion or inactivation). While MHC class I and class II binding epitopes encoded by bacterial genomes may be immunogenic (*10–14*), very few reports have demonstrated that microbe-specific $CD4^+$ or $CD8^+$ T lymphocytes attack normal or neoplastic tissues (*15–17*).

Cyclophosphamide (CTX) induces the translocation of *E. hirae* from the gut lumen to the mesenteric and splenic immune tissues, thereby eliciting specific $CD4^+$ and $CD8^+$ T lymphocytes producing interleukin-17 (IL17) and interferon-□ (IFN□□□ correlating with therapeutically effective anticancer immune responses (*6*, *18*). Broad-spectrum antibiotics abolished the therapeutic efficacy of CTX unless *E. hirae* was supplied by oral gavage (*6*). When comparing a panel of distinct *E. hirae* strains (Table S1, Figure S1A) for their capacity to restore the antibiotic-perturbed anticancer effects of CTX, we found that only a few *E. hirae* isolates (such as 13144 and IGR11) were efficient (Figure 1A-B, Ref. (*6*)). Given that the therapeutic efficacy of the combination of CTX and *E. hirae* 13144 is abrogated by the depletion of $CD8^+$ T cells or the neutralization of IFN□ □□□, we screened the differential capacity of *E. hirae* strains to elicit memory T cell responses after priming of the host, measured as the *ex vivo* recall response (IFN□ secretion) of splenic $CD8^+$ T cells against various *E. hirae* strains loaded onto dendritic cells(DC) (Figure S2A). While *E. hirae* 13144 triggered specific $CD8^+$ T cell responses (that were not cross-reactive against irrelevant enterococci), *E. hirae* 708 and 13344 (two prototypic inefficient strains) □□□ failed to do so (Figure S2A).

To identify relevant T cell epitopes, we aligned the sequences of bacterial genes encoding putative cell wall and secreted proteins for immunogenic (13144) *versus* non-immunogenic (708 and 13344) *E. hirae* strains, followed by the in silico identification of 13144-specific nonapeptides with strong affinity (<50 nM) for the MHC class I $H\text{-}2K^b$ protein (Table S2). Subsequently, we recovered splenic $CD8^+$ T cells from mice that had been exposed to *E. hirae* 13144 and CTX (Figure 1C), restimulated them *in vitro* with pools of potentially immunogenic

nonapeptides from *E. hirae* 13144 to measure IFN□ production (Table S2, Figure S2B) and finally split the most efficient pool (No. 7) into individual peptides (Figure 1D). This approach led to the identification of one dominant epitope (one-letter amino acid [aa] code: TSLARFANI, abbreviation TMP1) in position 187 to 197 of the aa sequence of the phage tail length tape measure protein (TMP, 1506 aa) from a 39.2 kb prophage of *E. hirae* 13144 (Figure 1D, Figure S3, Tables S2-S3). The 39.2kb prophage encodes 65 genes, including one shared between all 18 *E. hirae* genomes and 38 unique to *E. hirae* 13144 (Figure S1B), encoding capsid, portal and tail structures characteristic of *Siphoviridae* phages. Importantly, the TMP1 epitope of the 39.2kb prophage from *E. hirae* 13144 and the prophage fragment contained in *E. hirae* IGR11 showed 100% sequence identity (Figure S3 and S4A). Accordingly, *E. hirae* IGR11 was as efficient as *E. hirae* 13144 in reducing the growth of MCA205 sarcomas treated with CTX (Figure 1A-B). In contrast, the absence of a *bona fide* TMP1 epitope (observed in *E. hirae* 708 and 13344, Figure S1B) and a mutation in position 3 of the TSLARFANI peptide (L□F observed in *E. hirae* ATCC9790, Figure S4A) correlated with the lack of anticancer effects of these *E. hirae* strains (Figure 1B and Ref. (*6*)). ELIspot assays designed to detect peptide-specific IFN□-producing T cells revealed that mice gavaged with *E. hirae* 13144 or IGR11 mounted a $CD8^+$ T cell response against TMP1 (but not against the control peptides TMP2 and TMP3), while mice receiving *E. hirae* strains lacking TMP1 (strains 708, 13344) or a strain possessing a mutated TMP1 (strain ATCC9790) were unable to do so (Figure 1D). We used a fluorescent H-$2K^b$/TSLARFANI tetrameric complex (and its negative control H-$2K^b$/SIINFEKL binding to ovalbumine (OVA) specific $CD8^+$ T cells) to detect the frequency and distribution of TMP1-specific cytotoxic T lymphocytes (CTLs) in naive and MCA205 sarcoma bearing C57BL/6 mice. We observed a specific increase in splenic $CD8^+$ T cells that recognized the TMP1 peptide (but not the OVA peptide SIINFEKL) at day 7 following treatment with CTX and gavage with *E. hirae* 13144 (Figure 1E), as well as in tumor draining lymph nodes (LN) of tumor bearers at day 14 after treatment with CTX and gavage with *E. hirae* 13144 (Figure S2C-D). Splenic TMP1 (but not OVA)-specific (H-$2K^b$/TSLARFANI tetramer-positive) CTLs also increased in their frequency after gavage with *E. hirae* IGR11 (but not 13344 nor ATCC9790) (Figure 1E). The H-$2K^b$/TSLARFANI tetramer-positive CTLs were specifically enriched in the $CXCR3^+CCR9^+$ fraction of $CD8^+$ T cells from secondary lymphoid organs (Figure S2C). Even in mice colonized with human fecal materials, CTX administration and oral gavage with *E. hirae* 13144 induced an

anticancer effect (Figure S2E) and an expansion of H-2K^b/TSLARFANI tetramer-positive CTL in tumor draining LN at day 7 and in tumor beds at day 17 while vanishing from mesenteric LN (Figure S2F-H). Hence, immunogenic *E. hirae* elicits a H-2K^b restricted CTL response against the TMP-derived peptide TMP1/TSLARFANI.

To explore the capacity of TMP1-specific H-2K^b restricted T cells to control the growth of MCA205 cancers, we subcutaneously (s.c.) immunized naive C57BL/6 mice with dendritic cells (DC) loaded with heat-inactivated *E. hirae* 13144 (positive control), the naturally occurring TMP1/TSLARFANI peptide from 13144 and IGR11, its L□F mutant from *E. hirae* ATCC9790 ('mut3', Figure 2A, Figure S4A) or other non-immunogenic bacterial peptides (group 1, Figure S2B). In this prophylactic setting, DC pulsed with TMP1 (but not mut3) were as efficient as the whole *E. hirae* extract in reducing tumor growth (Figure 2B-C). Next, we explored whether the TMP1 peptide would be able to confer immunogenicity to the usually inefficient bacterium *Escherichia coli* strain DH5α in the therapeutic setting, in which antibiotic treatment is followed by gavage with different bacterial strains and CTX-based chemotherapy (Figure 1A and Ref. (*6*)). *E. coli* engineered to express TMP1 (Figure S5) was as efficient as *E. hirae* 13144 in restraining MCA205 tumor growth (Figure S4B, Figure 2D) and eliciting tetramer binding CTL in the spleen (Figure 2E). In contrast, *E. coli* expressing an irrelevant sequence (encoding mouse EGFP protein), mut3 or or mutant TMP1 bearing a S□A exchange in the anchor position 2 ('mut2') (Figure 2A) failed to induce such a cancer-protective immune response (Figure 2D-E). To explore the mechanism by which TMP1 exerts its anticancer activity against MCA205 tumors in C57BL/6 mice, we investigated whether H-2K^b-restricted mouse tumor antigens with high identity to the TMP1 peptide (TSLARFANI) exist. Using the NCBI BLASTP suite, we found that the peptide (GSLARFRNI) belonging to the proteasome subunit beta type-4 (PSMB4) located at amino acid positions 76-84 shared a strong homology (7 out of 9 amino acids with identical amino acids at the MHC Class I anchoring positions 2 and 9) with TMP1 (Figure 3A). We queried for potential neoepitopes of MCA205 but found no significant homology with TMP1, prompting us to focus on the non-mutated PSMB4 peptide. In fact, some mouse tumors (such as MCA205 sarcomas and TC1 lung cancers) overexpress the PSMB4 antigen compared with their normal tissues of origin, while others (such as MC38 colon cancers) failed to do so (Figure 3B). This correlates with the fact that MCA205 and TC1 tumors respond to the treatment with

CTX+*E. hirae* 13144, while MC38 cancer does not (Figure S6A-B). PSMB4 is an oncogenic driver involved in proliferation and invasion (*19*) in a variety of malignancies such as glioblastoma (*20*), melanoma (*21*) and breast cancers (*22*), associated with dismal prognosis (*19*, *20*, *22*). CRISPR/Cas9-mediated genomic knock-in of the PSMB4 sequence replacing GSLARFRNI by GALARFRNI (with an S→A exchange in position 2) or GSFARFRNI (with an L→F exchange in position 3 equivalent to mut 3 of TSLARFANI) in MCA205 cells (Figure S7) significantly affected tumor growth kinetics (Figure S6C-D), suggesting that this PSMB4 epitope contributes to the oncogenic activity of PSMB4. While these knock-in mutations did not interfere with the efficacy of CTX treatment alone, they drastically blunted the anticancer effects of *E. hirae* 13144 (Figure 3C-D). We extended these findings to a second tumor model where the anticancer effects of the combination of CTX+*E. hirae* 13144 were additive even in the absence of antibiotic-induced dysbiosis. Introducing a knock-in mutation in position 3 of PSMB4 into TC1 lung cancer cells again compromised the antitumor effects of CTX (Figure 3E). Moreover, in the setting of PD-1 blockade, administration of E. *hirae* 13144 without prior conditioning with antibiotics reduced the growth of parental but not PSMB4-mutated MCA205 cancers (Figure S6E). These results support the idea that the TSLARFANI TMP1 peptide encoded by *E. hirae* 13144 indeed induces T cell responses against the PSMB4-derived GSLARFRNI peptide across different tumor types and therapy modalities.

Reenforcing the notion of molecular mimicry between phage-encoded and cancer antigens, flow cytometric analyses using fluorescent-labelled tetramers H-2K^b/TSLARFANI (from TMP1) and H-2K^b/GSLARFRNI (from PSMB4) identified a subset of double-positive CTLs that infiltrate MCA205 tumors from CTX/*E. hirae 13144*-treated mice (Figure S6F) and that was as frequent as CTLs recognizing the PSMB4 peptide only (Figure 4A). We purified the splenic CD8$^+$ T cells using either the TMP1-H-2K^b or PSMB4-H-2K^b specific tetramers and stimulated them with irrelevant (OVA-derived-SIINFEKL) *versus* relevant (TMP-derived TSLARFANI or PSMB4-derived GSLARFRNI) peptides (Figure 4B). CD8$^+$ T cells binding H-2K^b-TMP1 tetramers produced IFN□ not only in response to TMP1 (up to 5-fold increase in IFN□ secreting T cells) but also in response to the PSMB4 epitope (2-fold increase, as much as with heat-killed *E. hirae* 13144 processed by DC) (Figure 4C, Figure S6G). Similarly, CD8$^+$ T cells binding H-2K^b-PSMB4 tetramers functionally recognized TMP1, albeit less efficiently than the PSMB4 epitope

(Figure S6G). We analyzed the T cell receptor (TCR) repertoire of these two tetramer-reactive $CD8^+$ T cell subsets. In accordance with the functional data, half of the $CD8^+$ T cells labelled with PSMB4-H-$2K^b$ tetramers shared clonotypes with the much wider TCR repertoire of T cells labelled with the TMP1-H-$2K^b$ specific tetramers (Figure 4D, Table S4-S5) (but not with the negative fraction, Figure S6H). In sum, T cells recognizing the TMP1 epitope of immunogenic *E. hirae* can crossreact with a peptide contained in the oncogenic driver PSMB4 and *vice versa*.

Temperate bacteriophages are bacterial viruses that can transfer virulence, antimicrobial resistance genes, and immunogenic sequences to new bacterial hosts (*23*). The TMP protein, which contains a variable number of tandem repeats with highly conserved tryptophan and phenylalanine residues at fixed positions is encoded by the genome of *Siphoviridae* phages (*24*, *25*). To investigate the capacity of the *E. hirae 13144* phage to lysogenize other bacterial species *in vivo*, we performed culturomic analyses of the ileal content from C57BL/6 mice subjected to oral gavage with *E. hirae* 13144 and systemic CTX therapy, followed by PCR analyses seeking TMP sequences (Figure S8A-B). We tested 7 to 18 bacterial colonies from each animal and a total of 76 colonies. We only found lysogenic conversion of *E. gallinarum* by the *E. hirae*-temperate phage *in vivo*, as confirmed by sequencing of the phage genome in the second host (Figure 4E, Figure S8B-C). In contrast, none of the 90 colonies (mostly of *E. gallinarum)* isolated from naive mice harbored the TMP sequence (Figure S8A). Similarly, *in vitro* coculture of TMP^+ *E. hirae* 13144 together with TMP^- *E. gallinarum spp*. at a 1:1 ratio uncovered a significant (~15%) rate of lysogenic conversion (Figure S8D). Examination of a preparation admixing *E. hirae* 13144 and *E. gallinarum* at a 1:10 ratio by means of transmission electron microscopy revealed numerous phages with the typical *Siphoviridae* morphology in the medium, whereas control cultures (bacteria separately) were free of such phages (Figure 4F). Altogether, these results indicate that the TMP1 peptide-encoding *Siphoviridae* phage from *E. hirae* 13144 is a virulent phage.

We next explored the possible pathophysiological relevance of these findings. We first screened a total of 3,027 adult and mother-infant metagenomes (*26*), validated by a second independent metagenomic-assembly based screening of 9,428 metagenomes (*27*) (*28*), to assess the breadth of coverage (BOC) of the *E. hira*e genome and its phages (Figure S9A). *E.*

hirae was present with 100% confidence (i.e. BOC > 80%) in less than 150 fecal samples from disparate geography, age and datasets. This phage (and its host) could be vertically transmitted from mothers to infants and then colonizes the neonate. There was an increased prevalence of the phage (57%) in fecal microbiomes from children (representing 16% of all metagenomes, Fisher's test p-value <0.00001). Of note, the *E. hirae* 13144 phage was detectable in many samples lacking the presence of the *E. hirae* core genome, suggesting that other bacteria than *E. hirae* can host this phage. All host genomes belonged to the *Enterococcus* genus (except two assigned to *Coprobacillus*), in particular *E. faecalis* (80 genomes), *E. faecium* (23), and *E. hirae* (15), suggesting that phage 13144 (and its homologues from *E. hirae* 708, and 13344) are genus- specific but not species-specific.

Contrasting with metagenomics that has a low sensitivity to detect poor abundance species, culturomics followed by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF) provides a technology for detecting rare *E. hirae* colonies in the stool of healthy individuals (*29*) or cancer patients (*8*). PCR analyses of each single cultivatable enterococcal colony (up to 5 per species and individual) from 76 cancer patients led to the detection of the TMP sequence encompassing the TMP1 peptide in 34% of the patients, only in *E. faecalis* and *E. hirae* (Figure S9B, Figure S10). Advanced renal and lung cancer patients (cohort described in Ref. (16)) with detectable fecal TMP at diagnosis exhibited prolonged overall survival after therapy with immune checkpoint inhibitors targeting PD-1 (Figure 4G). Therefore, we screened sixteen TMP-derived nonapeptides predicted to bind the human MHC class I HLA-A*0201 with high affinity for their ability to prime naive $CD8^+$ T cells from six healthy volunteers *in vitro*. We found 6 out of 16 epitopes capable of triggering significant peptide-specific IFN release that were located in two distinct regions of the TMP protein (504-708 and 1397-1462, Figure S11A-B, Table S6). Using the NCBI BLASTP suite, we searched the human cancer peptidome (of the TCGA database) for a high degree of homology with these 6 HLA-A*0201 -restricted immunogenic nonapeptides. We found that only the TMP-derived peptide KLAKFASVV (aa 631-639) shared significant homology (7 out of 9 aa, with identical residues at the MHC anchoring positions 2 and 9) with a peptide contained in the protein glycerol-3-phosphate dehydrogenase 1-like (GPD1-L) (Figure S11C). GPD1-L reportedly counteracts the oncogenic HIF1 -dependent adaptation to hypoxia, and its expression is

associated with favorable prognosis in head and neck squamous cell carcinomas (*30–32*). The TCGA transcriptomics database unveiled that high expression of GPD1-L is associated with improved overall survival in lung adenocarcinoma and kidney cancers (Figure S11D). Moreover, high expression of GPD1-L mRNA by tumors at diagnosis was associated with improved progression-free survival in three independent cohorts of non-small cell lung cancer (NSCLC) patients (n=157, Table S7) treated with anti-PD1 Abs (Figure S11E-F). Expression of GPD-1L failed to correlate with that of PD-L1 in NSCLC (Figure S11G). Of note, mutations in or adjacent to the 631-639 amino acid sequence of GPD-1L gene could rarely be identified in several types of neoplasia (Figure S12).

We derived an HLA-A*0201-restricted, phage peptide (KLAKFASVV)-specific T cell line from peripheral blood mononuclear cells of a human volunteer. Clones from this line also recognized the HLA-A*0201 -restricted, GPD-1L epitope (KLQKFASTV) (Figure S13A-C). Moreover, we detected $CD8^+$ T cells binding HLA-A*0201/KLAKFASVV tetramers exhibiting hallmarks of effector functions after *in vitro* stimulation of PBMC with the KLAKFASVV phage epitope in 3 out of 6 NSCLC patients (Fig. S13D-F). In the reverse attempt searching for molecular mimicry between well known and naturally processed non-mutated melanoma differentiation antigens recognized by human T cell clones (such as HLA-A*0201-binding MART-1 or MELOE epitopes) and gut commensal antigens, we found microbial analogs in the public microbiome data bases (Figure S14, Table S8-Table S9, Figure S15, Table S8-S10). Some of these microbial peptides are recognized by the corresponding TCR (Tables S9- S10) with similar affinities as the parental (tumoral) epitope.

Altogether the present results demonstrate that microbial genomes code for MHC class I-restricted antigens that induce a memory $CD8^+$ T cell response, which then crossreacts with cancer antigens. Several lines of evidence plead in favor of this interpretation, as exemplified for the TMP1 epitope found within a phage that infects enterococci. First, naturally occurring (‘mut3’ in *E. hirae* strain ATCC9790) or artificial mutations (‘mut2’ or ‘mut3’ *in E.coli*) introduced into the TMP1 epitope suppressed the tumor-prophylactic and therapeutic potential of bacteria expressing TMP1. Second, transfer of the TMP1-encoding gene into *E. coli* conferred immunogenic capacity to this proteobacterium, which acquired the same antitumor properties as

TMP1-expressing *E. hirae*. Third, when cancer cells were genetically modified to remove the TMP1-crossreactive peptide within the PSMB4 protein, they formed tumors that could no longer be controlled upon oral gavage with TMP1-expressing *E. hirae*. Fourth, cancer patients carrying the TMP phage sequence in fecal enterococci spp. or the GPD1-L tumoral antigen homologous to TMP epitopes exhibited a better response to PD-1 blockade, suggesting that this type of microbe-cancer cross-reactivity might be clinically relevant.

Recent reports point to the pathological relevance of autoantigen-crossreactive, microbiota-derived peptides for autoimmune disorders such as myocarditis, lupus and rheumatoid arthritis (*34–36*). Given the enormous richness of the commensal proteome (*37*), we expect the existence of other microbial antigens mimicking auto- and tumor antigens. In fact, we extended these findings to naturally processed melanoma-specific antigens that have microbial orthologs recognized by the same TCRs. Global phage numbers have been estimated to reach as high as 10^{31} particles with the potential of 10^{25} phage infections occurring every second (*38*, *39*). Thus, the perspective opens that, within the microbiota, bacteriophages may enrich the therapeutic armamentarium for modulating the intestinal flora and for stimulating systemic anticancer immune responses.

References and Notes:

1. P. Sharma, J. P. Allison, Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. *Cell*. **161**, 205–214 (2015).

2. L. Galluzzi, A. Buqué, O. Kepp, L. Zitvogel, G. Kroemer, Immunological Effects of Conventional Chemotherapy and Targeted Anticancer Agents. *Cancer Cell*. **28**, 690–714 (2015).

3. L. Zitvogel, Y. Ma, D. Raoult, G. Kroemer, T. F. Gajewski, The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. *Science*. **359**, 1366–1370 (2018).

4. T. Tanoue, S. Morita, D. R. Plichta, A. N. Skelly, W. Suda, Y. Sugiura, S. Narushima, H. Vlamakis, I. Motoo, K. Sugita, A. Shiota, K. Takeshita, K. Yasuma-Mitobe, D. Riethmacher, T. Kaisho, J. M. Norman, D. Mucida, M. Suematsu, T. Yaguchi, V. Bucci, T. Inoue, Y. Kawakami, B. Olle, B. Roberts, M. Hattori, R. J. Xavier, K. Atarashi, K. Honda, A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature*. **565**, 600–605 (2019).

5. M. Vétizou, J. M. Pitt, R. Daillère, P. Lepage, N. Waldschmitt, C. Flament, S. Rusakiewicz, B. Routy, M. P. Roberti, C. P. M. Duong, V. Poirier-Colame, A. Roux, S. Becharef, S. Formenti, E. Golden, S. Cording, G. Eberl, A. Schlitzer, F. Ginhoux, S. Mani, T. Yamazaki, N. Jacquelot, D. P. Enot, M. Bérard, J. Nigou, P. Opolon, A. Eggermont, P.-L. Woerther, E. Chachaty, N. Chaput, C. Robert, C. Mateus, G. Kroemer, D. Raoult, I. G. Boneca, F. Carbonnel, M. Chamaillard, L. Zitvogel, Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science*. **350**, 1079–1084 (2015).

6. R. Daillère, M. Vétizou, N. Waldschmitt, T. Yamazaki, C. Isnard, V. Poirier-Colame, C. P. M. Duong, C. Flament, P. Lepage, M. P. Roberti, B. Routy, N. Jacquelot, L. Apetoh, S. Becharef, S. Rusakiewicz, P. Langella, H. Sokol, G. Kroemer, D. Enot, A. Roux, A. Eggermont, E. Tartour, L. Johannes, P.-L. Woerther, E. Chachaty, J.-C. Soria, E. Golden, S. Formenti, M. Plebanski, M. Madondo, P. Rosenstiel, D. Raoult, V. Cattoir, I. G. Boneca, M. Chamaillard, L. Zitvogel, Enterococcus hirae and Barnesiella intestinihominis Facilitate Cyclophosphamide-Induced Therapeutic Immunomodulatory Effects. *Immunity*. **45**, 931– 943 (2016).

7. Y. Rong, Z. Dong, Z. Hong, Y. Jin, W. Zhang, B. Zhang, W. Mao, H. Kong, C. Wang, B. Yang, X. Gao, Z. Song, S. E. Green, H. K. Song, H. Wang, Y. Lu, Reactivity toward Bifidobacterium longum and Enterococcus hirae demonstrate robust CD8+ T cell response and better prognosis in HBV-related hepatocellular carcinoma. *Exp. Cell Res.* **358**, 352–359 (2017).

8. B. Routy, E. Le Chatelier, L. Derosa, C. P. M. Duong, M. T. Alou, R. Daillère, A. Fluckiger, M. Messaoudene, C. Rauber, M. P. Roberti, M. Fidelle, C. Flament, V. Poirier- Colame, P. Opolon, C. Klein, K. Iribarren, L. Mondragón, N. Jacquelot, B. Qu, G. Ferrere, C. Clémenson, L. Mezquita, J. R. Masip, C. Naltet, S. Brosseau, C. Kaderbhai, C. Richard, H. Rizvi, F. Levenez, N. Galleron, B. Quinquis, N. Pons, B. Ryffel, V. Minard-Colin, P. Gonin, J.-C. Soria, E. Deutsch, Y. Loriot, F. Ghiringhelli, G. Zalcman, F. Goldwasser, B. Escudier, M. D. Hellmann, A. Eggermont, D. Raoult, L. Albiges, G. Kroemer, L. Zitvogel, Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science*. **359**, 91–97 (2018).

9. N. R. Rose, Negative selection, epitope mimicry and autoimmunity. *Curr. Opin. Immunol.***49**, 51–

55 (2017).

10. V. Rubio-Godoy, V. Dutoit, Y. Zhao, R. Simon, P. Guillaume, R. Houghten, P. Romero, J.-C. Cerottini, C. Pinilla, D. Valmori, Positional scanning-synthetic peptide library-based analysis of self- and pathogen-derived peptide cross-reactivity with tumor-reactive Melan- A-specific CTL. *J. Immunol. Baltim. Md 1950*. **169**, 5696–5707 (2002).

11. L. Vujanovic, M. Mandic, W. C. Olson, J. M. Kirkwood, W. J. Storkus, A mycoplasma peptide elicits heteroclitic CD4+ T cell responses against tumor antigen MAGE-A6. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **13**, 6796–6806 (2007).

12. M. E. Perez-Muñoz, P. Joglekar, Y.-J. Shen, Y.-J. Shen, K. Y. Chang, D. A. Peterson, Identification and Phylogeny of the First T Cell Epitope Identified from a Human Gut Bacteroides Species. *PloS One*. **10**, e0144382 (2015).

13. Y. Yang, M. B. Torchinsky, M. Gobert, H. Xiong, M. Xu, J. L. Linehan, F. Alonzo, C. Ng,A. Chen, X. Lin, A. Sczesnak, J.-J. Liao, V. J. Torres, M. K. Jenkins, J. J. Lafaille, D. R. Littman, Focused specificity of intestinal TH17 cells towards commensal bacterial antigens. *Nature*. **510**, 152–156 (2014).

14. J. N. Chai, Y. Peng, S. Rengarajan, B. D. Solomon, T. L. Ai, Z. Shen, J. S. A. Perry, K. A. Knoop, T. Tanoue, S. Narushima, K. Honda, C. O. Elson, R. D. Newberry, T. S. Stappenbeck, A. L. Kau, D. A. Peterson, J. G. Fox, C.-S. Hsieh, Helicobacter species are potent drivers of colonic T cell responses in homeostasis and inflammation. *Sci. Immunol.* **2** (2017), doi:10.1126/sciimmunol.aal5068.

15. Q. Ji, A. Perchellet, J. M. Goverman, Viral infection triggers central nervous system autoimmunity via activation of CD8+ T cells expressing dual TCRs. *Nat. Immunol.* **11**, 628–634 (2010).

16. V. P. Balachandran, M. Łuksza, J. N. Zhao, V. Makarov, J. A. Moral, R. Remark, B. Herbst,G. Askan, U. Bhanot, Y. Senbabaoglu, D. K. Wells, C. I. O. Cary, O. Grbovic-Huezo, M. Attiyeh, B. Medina, J. Zhang, J. Loo, J. Saglimbeni, M. Abu-Akeel, R. Zappasodi, N. Riaz, M. Smoragiewicz, Z. L. Kelley, O. Basturk, Australian Pancreatic Cancer Genome Initiative, Garvan Institute of Medical Research, Prince of Wales Hospital, Royal North Shore Hospital, University of Glasgow, St Vincent's Hospital, QIMR Berghofer Medical Research Institute, University of Melbourne, Centre for Cancer Research, University of Queensland, Institute for Molecular Bioscience, Bankstown Hospital, Liverpool Hospital, Royal Prince Alfred Hospital, Chris O'Brien Lifehouse, Westmead Hospital, Fremantle Hospital, St John of God Healthcare, Royal Adelaide Hospital, Flinders Medical Centre, Envoi Pathology, Princess Alexandria Hospital, Austin Hospital, Johns Hopkins Medical, Institutes, ARC-Net Centre for Applied Research on Cancer, M. Gönen, A. J. Levine, P. J. Allen, D. T. Fearon, M. Merad, S. Gnjatic, C. A. Iacobuzio-Donahue, J. D. Wolchok, R. P. DeMatteo, T. A. Chan, B. D. Greenbaum, T. Merghoub, S. D. Leach, Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature*. **551**, 512– 516 (2017).

17. C. P. Bradley, F. Teng, K. M. Felix, T. Sano, D. Naskar, K. E. Block, H. Huang, K. S. Knox, D. R. Littman, H.-J. J. Wu, Segmented Filamentous Bacteria Provoke Lung Autoimmunity by

Inducing Gut-Lung Axis Th17 Cells Expressing Dual TCRs. *Cell Host Microbe*. **22**, 697-704.e4 (2017).

18. S. Viaud, F. Saccheri, G. Mignot, T. Yamazaki, R. Daillère, D. Hannani, D. P. Enot, C. Pfirschke, C. Engblom, M. J. Pittet, A. Schlitzer, F. Ginhoux, L. Apetoh, E. Chachaty, P.-L. Woerther, G. Eberl, M. Bérard, C. Ecobichon, D. Clermont, C. Bizet, V. Gaboriau- Routhiau, N. Cerf-Bensussan, P. Opolon, N. Yessaad, E. Vivier, B. Ryffel, C. O. Elson, J. Doré, G. Kroemer, P. Lepage, I. G. Boneca, F. Ghiringhelli, L. Zitvogel, The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide. *Science*. **342**, 971–976 (2013).

19. G. Y. Lee, P. M. Haverty, L. Li, N. M. Kljavin, R. Bourgon, J. Lee, H. Stern, Z. Modrusan, S. Seshagiri, Z. Zhang, D. Davis, D. Stokoe, J. Settleman, F. J. de Sauvage, R. M. Neve, Comparative oncogenomics identifies PSMB4 and SHMT2 as potential cancer driver genes. *Cancer Res.* **74**, 3114–3126 (2014).

20. Y.-C. Cheng, W.-C. Tsai, Y.-C. Sung, H.-H. Chang, Y. Chen, Interference with PSMB4 Expression Exerts an Anti-Tumor Effect by Decreasing the Invasion and Proliferation of Human Glioblastoma Cells. *Cell. Physiol. Biochem. Int. J. Exp. Cell. Physiol. Biochem. Pharmacol.* **45**, 819–831 (2018).

21. X. Zhang, D. Lin, Y. Lin, H. Chen, M. Zou, S. Zhong, X. Yi, S. Han, Proteasome beta-4 subunit contributes to the development of melanoma and is regulated by miR-148b. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **39**, 1010428317705767 (2017).

22. H. Wang, Z. He, L. Xia, W. Zhang, L. Xu, X. Yue, X. Ru, Y. Xu, PSMB4 overexpression enhances the cell growth and viability of breast cancer cells leading to a poor prognosis. *Oncol. Rep.* **40**, 2343–2352 (2018).

23. M. G. Weinbauer, Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).

24. M. Piuri, G. F. Hatfull, A peptidoglycan hydrolase motif within the mycobacteriophage TM4 tape measure protein promotes efficient infection of stationary phase cells. *Mol. Microbiol.* **62**, 1569–1585 (2006).

25. M. Belcaid, A. Bergeron, G. Poisson, The evolution of the tape measure protein: units, duplications and losses. *BMC Bioinformatics*. **12 Suppl 9**, S10 (2011).

26. E. Pasolli, L. Schiffer, P. Manghi, A. Renson, V. Obenchain, D. T. Truong, F. Beghini, F. Malik, M. Ramos, J. B. Dowd, C. Huttenhower, M. Morgan, N. Segata, L. Waldron, Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods*. **14**, 10231024 (2017).

27. P. Ferretti, E. Pasolli, A. Tett, F. Asnicar, V. Gorfer, S. Fedi, F. Armanini, D. T. Truong, S. Manara, M. Zolfo, F. Beghini, R. Bertorelli, V. De Sanctis, I. Bariletti, R. Canto, R. Clementi, M. Cologna, T. Crifò, G. Cusumano, S. Gottardi, C. Innamorati, C. Masè, D. Postai, D. Savoi, S. Duranti, G. A. Lugli, L. Mancabelli, F. Turroni, C. Ferrario, C. Milani, M. Mangifesta, R. Anzalone, A. Viappiani, M. Yassour, H. Vlamakis, R. Xavier, C. M Collado, O. Koren, S. Tateo,

M. Soffiati, A. Pedrotti, M. Ventura, C. Huttenhower, P. Bork, N. Segata, Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe*. **24**, 133-145.e5 (2018).

28. E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, N. Segata, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*. **176**, 649-662.e20 (2019).

29. B. Samb-Ba, C. Mazenot, A. Gassama-Sow, G. Dubourg, H. Richet, P. Hugon, J.-C. Lagier, D. Raoult, F. Fenollar, MALDI-TOF identification of the human Gut microbiome in people with and without diarrhea in Senegal. *PloS One*. **9**, e87419 (2014).

30. T. J. Kelly, A. L. Souza, C. B. Clish, P. Puigserver, A hypoxia-induced positive feedback loop promotes hypoxia-inducible factor 1alpha stability through miR-210 suppression of glycerol-3-phosphate dehydrogenase 1-like. *Mol. Cell. Biol.* **31**, 2696–2706 (2011).

31. Z. Feng, J. N. Li, L. Wang, Y. F. Pu, Y. Wang, C. B. Guo, The prognostic value of glycerol- 3-phosphate dehydrogenase 1-like expression in head and neck squamous cell carcinoma. *Histopathology*. **64**, 348–355 (2014).

32. S.-C. Liu, S.-M. Chuang, C.-J. Hsu, C.-H. Tsai, S.-W. Wang, C.-H. Tang, CTGF increases vascular endothelial growth factor-dependent angiogenesis in human synovial fibroblasts by increasing miR-210 expression. *Cell Death Dis.* **5**, e1485 (2014).

33. S. Simon, Z. Wu, J. Cruard, V. Vignard, A. Fortun, A. Khammari, B. Dreno, F. Lang, S. J. Rulli, N. Labarriere, TCR Analyses of Two Vast and Shared Melanoma Antigen-Specific T Cell Repertoires: Common and Specific Features. *Front. Immunol.* **9**, 1962 (2018).

34. C. Gil-Cruz, C. Perez-Shibayama, A. De Martin, F. Ronchi, K. van der Borght, R. Niederer, L. Onder, M. Lütge, M. Novkovic, V. Nindl, G. Ramos, M. Arnoldini, E. M. C. Slack, V. Boivin-Jahns, R. Jahns, M. Wyss, C. Mooser, B. N. Lambrecht, M. T. Maeder, H. Rickli, L. Flatz, U. Eriksson, M. B. Geuking, K. D. McCoy, B. Ludewig, Microbiota-derived peptide mimics drive lethal inflammatory cardiomyopathy. *Science*. **366**, 881–886 (2019).

35. T. M. Greiling, C. Dehner, X. Chen, K. Hughes, A. J. Iñiguez, M. Boccitto, D. Z. Ruiz, S. C. Renfroe, S. M. Vieira, W. E. Ruff, S. Sim, C. Kriegel, J. Glanternik, X. Chen, M. Girardi, P. Degnan, K. H. Costenbader, A. L. Goodman, S. L. Wolin, M. A. Kriegel, Commensal orthologs of the human autoantigen Ro60 as triggers of autoimmunity in lupus. Sci. Transl. Med. 10 (2018), doi:10.1126/scitranslmed.aan2306.

36. M. F. Konig, L. Abusleme, J. Reinholdt, R. J. Palmer, R. P. Teles, K. Sampson, A. Rosen, P. A. Nigrovic, J. Sokolove, J. T. Giles, N. M. Moutsopoulos, F. Andrade, Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. *Sci. Transl. Med.* **8**, 369ra176 (2016).

37. J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, MetaHIT Consortium, P. Bork, J. Wang, MetaHIT Consortium, An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

38. M. L. Pedulla, M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. 535 Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, R. W. Hendrix, G. F. Hatfull, Origins of highly mosaic mycobacteriophage genomes. *Cell*. **113**, 171–182 (2003).

39. K. E. Wommack, R. R. Colwell, Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev. MMBR*. **64**, 69–114 (2000).

Acknowledgments: We are thankful to the animal facility team of Gustave Roussy and all the technicians from Centre GF Leclerc. We are very endebted to Dr Oliver Kepp, Gustave Roussy for figure design, and to Prof. Hans Georg Rammensee from the Department of Immunology, Institute for Cell Biology, University of Tübingen, Tübingen, Germany for his careful guidance in peptide selection and reading of the paper. LZ and GK were supported by the Ligue contre le Cancer (équipe labelisée); Agence National de la Recherche (ANR) – Projets blancs; ANR under the frame of E-Rare-2, the ERA-Net for Research on Rare Diseases; Association pour la recherche sur le cancer (ARC); Cancéropôle Ile-de-France; Chancelerie des universités de Paris (Legs Poix), Fondation pour la Recherche Médicale (FRM); a donation by Elior; the European Commission (Horizon 2020: Oncobiome); the European Research Council (ERC); Fondation Carrefour; High-end Foreign Expert Program in China (GDW20171100085 and GDW20181100051), Institut National du Cancer (INCa); Inserm (HTE); Institut Universitaire de France; LeDucq Foundation; the LabEx Immuno-Oncology; the RHU Torino Lumière; the Seerave Foundation; the SIRIC Stratified Oncology Cell DNA Repair and Tumor Immune

Elimination (SOCRATE); ONCOBIOME H2020 network, CARE network (directed by Prof. Mariette, Kremlin Bicêtre AP-HP), and the SIRIC Cancer Research and Personalized Medicine (CARPEM); RHU Torino Lumière (ANR-16-RHUS-0008). The results shown here are based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. National Research, Development andInnovation Fund of Hungary Project no. FIEK_16-1-2016-0005. Z.S was supported by the Research and Technology Innovation Fund NAP2-2017-1.2.1-NKP-0002, Breast Cancer Research Foundation (BCRF-17-156). Z.S and I.C were supported by the Novo Nordisk Foundation Interdisciplinary Synergy Programme Grant (NNF15OC0016584). PN was supported by the Italian Association for Cancer Research AIRC IG 19822. Mouse TCR sequencing was performed by the TRiPoD ERC-Advanced EU (322856) grants to Prof. David Klatzmann.

Competing interests statement: RD, DR, LZ and GK are cofounders of everImmune, a biotech company devoted to the use of commensal microbes for the treatment of cancers. RD is a full-time employee of everImmune. RD and LZ hold patents on immunogenic phage sequences.

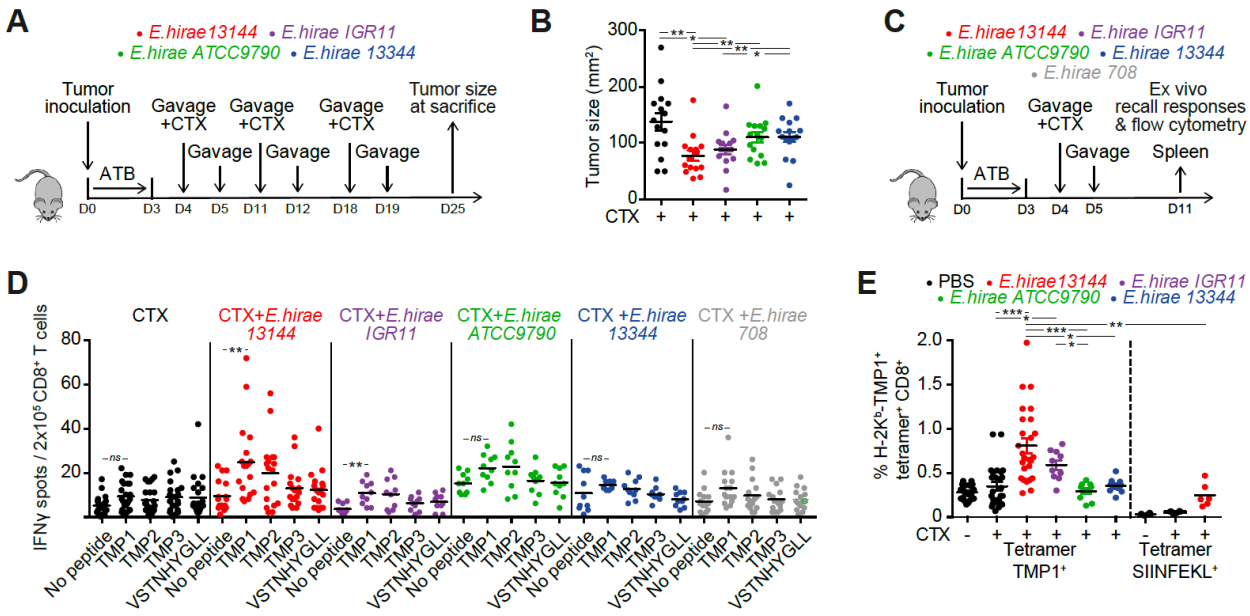

A
• E.hirae13144 • E.hirae IGR11
• E.hirae ATCC9790 • E.hirae 13344
Tumor inoculation
Gavage +CTX
Gavage +CTX
Gavage +CTX
Tumor size at sacrifice
Gavage
Gavage
Gavage
ATB
D0 D3 D4 D5 D11 D12 D18 D19 D25
B
Tumor size (mm²)
300
200
100
0
CTX + + + + +
C
• E.hirae13144 • E.hirae IGR11
• E.hirae ATCC9790 • E.hirae 13344
• E.hirae 708
Tumor inoculation
Gavage +CTX
Ex vivo recall responses & flow cytometry
Gavage
Spleen
ATB
D0 D3 D4 D5 D11
D
IFNγ spots / 2x10⁵ CD8⁺ T cells
CTX
CTX+E.hirae 13144
CTX+E.hirae IGR11
CTX+E.hirae ATCC9790
CTX +E.hirae 13344
CTX +E.hirae 708
No peptide TMP1 TMP2 TMP3 VSTNHYGLL
E
• PBS • E.hirae13144 • E.hirae IGR11
• E.hirae ATCC9790 • E.hirae 13344
% H-2Kᵇ-TMP1⁺ tetramer⁺ CD8⁺
CTX - + + + + + - + +
Tetramer TMP1⁺
Tetramer SIINFEKL⁺

Figure 1

Legends to Figures

Figure 1. Phage Tail Length Tape Measure Protein as the unique antigenic sequence in *E. hirae* 13144.

A, B. C57BL/6 mice bearing MCA205 sarcomas were conditioned with broad spectrum antibiotics (streptomycin, colistin, ampicillin, vancomycin) for 3 days before performing oral gavages with *E. hirae* strain 13144 and i.p. injections of cyclophosphamide (CTX), as indicated (A), and tumor size was recorded for each mouse at sacrifice on day 25 (B). C-E. Naïve C57BL/6 mice were conditioned with antibiotics, gavaged with distinct *E. hirae* strains and treated with CTX (C). Day 11 purified $CD8^+$ T splenocytes were restimulated *ex vivo* in a recall assay with bone marrow-derived dendritic cells loaded with the indicated peptides (Table S2, group 7) to quantify IFNγ-secreting $CD8^+$T cells (D). H-$2K^b$/TMP1 (TSLARFANI) or H-$2K^b$/SIINFEKL tetramer binding $CD8^+$ splenocytes were detected by cytofluorometry at day 11 (E). Also refer to Figure S2. Each graph assembles results from 2-3 independent experiments containing groups of 5-6 mice. ANOVA statistical analyses (Kruskal-wallis test): *$p<0.05$, **$p<0.01$, ***$p<0.001$. Refer to the statistical report.

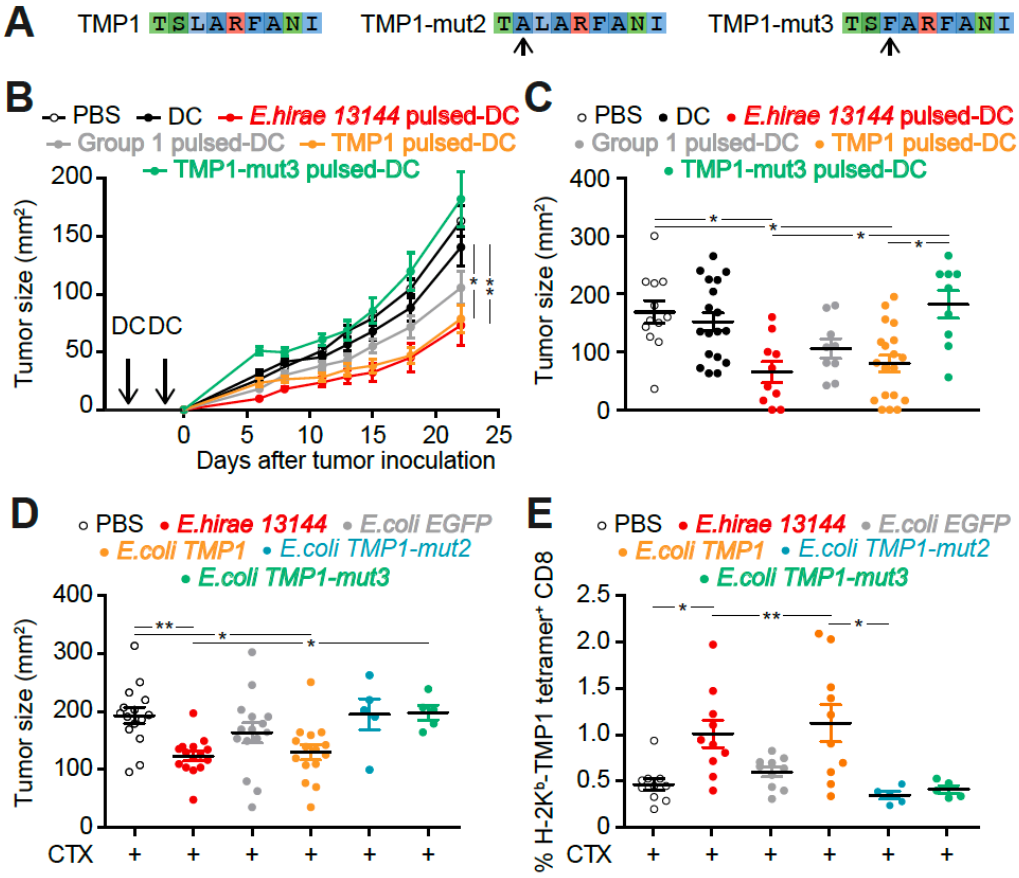

A
TMP1 TSLARFANI
TMP1-mut2 TALARFANI
TMP1-mut3 TSFARFANI
B
PBS
DC
E.hirae 13144 pulsed-DC
Group 1 pulsed-DC
TMP1 pulsed-DC
TMP1-mut3 pulsed-DC
Tumor size (mm²)
DC DC
Days after tumor inoculation
C
PBS
DC
E.hirae 13144 pulsed-DC
Group 1 pulsed-DC
TMP1 pulsed-DC
TMP1-mut3 pulsed-DC
Tumor size (mm²)
D
PBS
E.hirae 13144
E.coli EGFP
E.coli TMP1
E.coli TMP1-mut2
E.coli TMP1-mut3
Tumor size (mm²)
CTX
E
PBS
E.hirae 13144
E.coli EGFP
E.coli TMP1
E.coli TMP1-mut2
E.coli TMP1-mut3
% H-2Kᵇ-TMP1 tetramer⁺ CD8
CTX

Figure 2

Figure 2. Prophylactic and therapeutic immunization using Phage Tail Length Tape Measure Protein (TMP) against sarcomas.

A. Sequence of the immunogenic epitope TMP1 (TSLARFANI) with the artificial and naturally occuring mutations in positions 2 and 3, respectively. B-C. *Prophylactic vaccinations*. TLR3 ligand-exposed dendritic cell (DC) were pulsed with peptides or heat-inactivated bacteria and then s.c. inoculated twice into mice. One month later, MCA205 sarcomas were implanted in the opposite flank, followed by monitoring of tumor size (means±SEM in B, individual results in C). D-E. *Therapeutic settings.* MCA205 tumor bearing mice were treated with cyclophosphamide (CTX) and gavaged with *E. hirae* 13144 or *E. coli* (like in Fig. 1A) that were genetically modified to express the indicated peptides or enhanced green fluorescent protein (EGFP) as a negative control. Tumor growth at sacrifice (D) and the frequency of H-2K^b/TMP1 tetramer binding splenic CD8$^+$ T cells (E) were monitored. Results are shown for 12-18 animals, gathered from 2-3 independent experiments. ANOVA statistical analyses (Kruskal-wallis test): **p<0.05, ** p<0.01, ***p<0.001.* Refer to the statistical report.

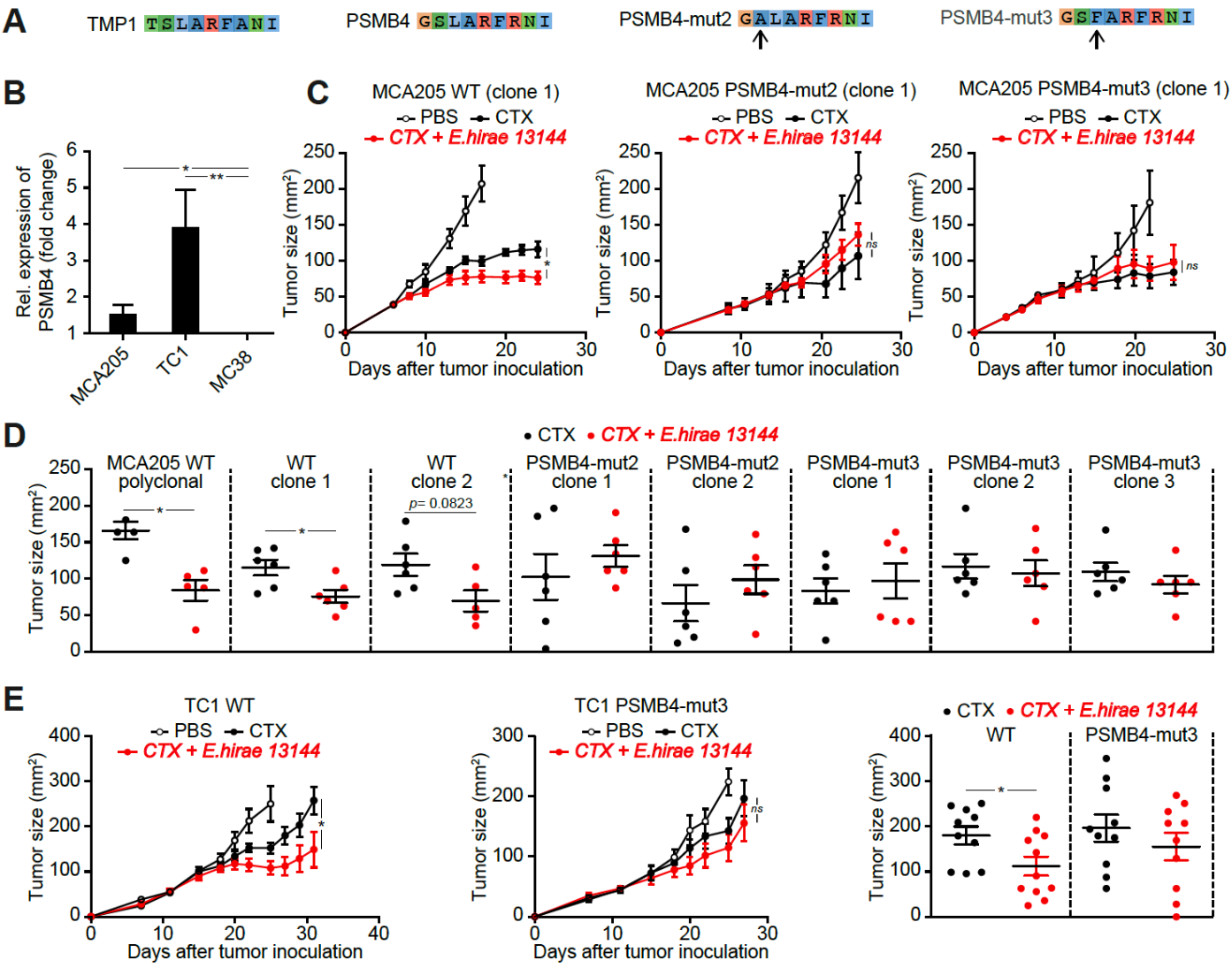

A
TMP1 TSLARFANI
PSMB4 GSLARFRNI
PSMB4-mut2 GALARFRNI
PSMB4-mut3 GSFARFRNI
B
Rel. expression of PSMB4 (fold change)
MCA205
TC1
MC38
C
MCA205 WT (clone 1)
MCA205 PSMB4-mut2 (clone 1)
MCA205 PSMB4-mut3 (clone 1)
PBS
CTX
CTX + E.hirae 13144
Tumor size (mm²)
Days after tumor inoculation
ns
D
CTX
CTX + E.hirae 13144
MCA205 WT polyclonal
WT clone 1
WT clone 2
p= 0.0823
PSMB4-mut2 clone 1
PSMB4-mut2 clone 2
PSMB4-mut3 clone 1
PSMB4-mut3 clone 2
PSMB4-mut3 clone 3
Tumor size (mm²)
E
TC1 WT
TC1 PSMB4-mut3
PBS
CTX
CTX + E.hirae 13144
Tumor size (mm²)
Days after tumor inoculation
WT
PSMB4-mut3
ns

Figure 3

Figure 3. Molecular mimicry between enterophage TMP and the oncogenic driver PSMB4 in two mouse cancers.

A. Sequence alignment of the enterophage TMP1 peptide and a PSMB4 epitope with its two experimental mutants. B. Relative expression of PSMB4 mRNA in MCA205 sarcoma, TC1 lung cancer and MC38 colon carcinomas as compared to their healthy tissue of origin (mean ratio±SEM, n=3). C-D. Therapeutic response of wild type *versus* knock-in mutants of MCA205 to cyclophosphamide (CTX) alone or in combination with immunogenic *E. hirae* strain 13144 (setting as in Fig. 1A). Results are shown as tumor growth kinetics (means±SEM) for selected MCA205 clones (C) or as individual results (one dot corresponds to one mouse) on day 25 (D).

E. Therapeutic response of wild type *versus* mutated TC1 lung cancers to CTX alone or in combination with *E. hirae* 13144 (setting as in Fig. 1A, but without antibiotic preconditioning) reflected by tumor growth kinetics and individual tumor sizes at sacrifice. Results are shown as means ±SEM. Mann Whitney test or ANOVA statistical analyses (Kruskal-wallis test): **p<0.05, ** p<0.01, ***p<0.001.* Refer to the statistical report.

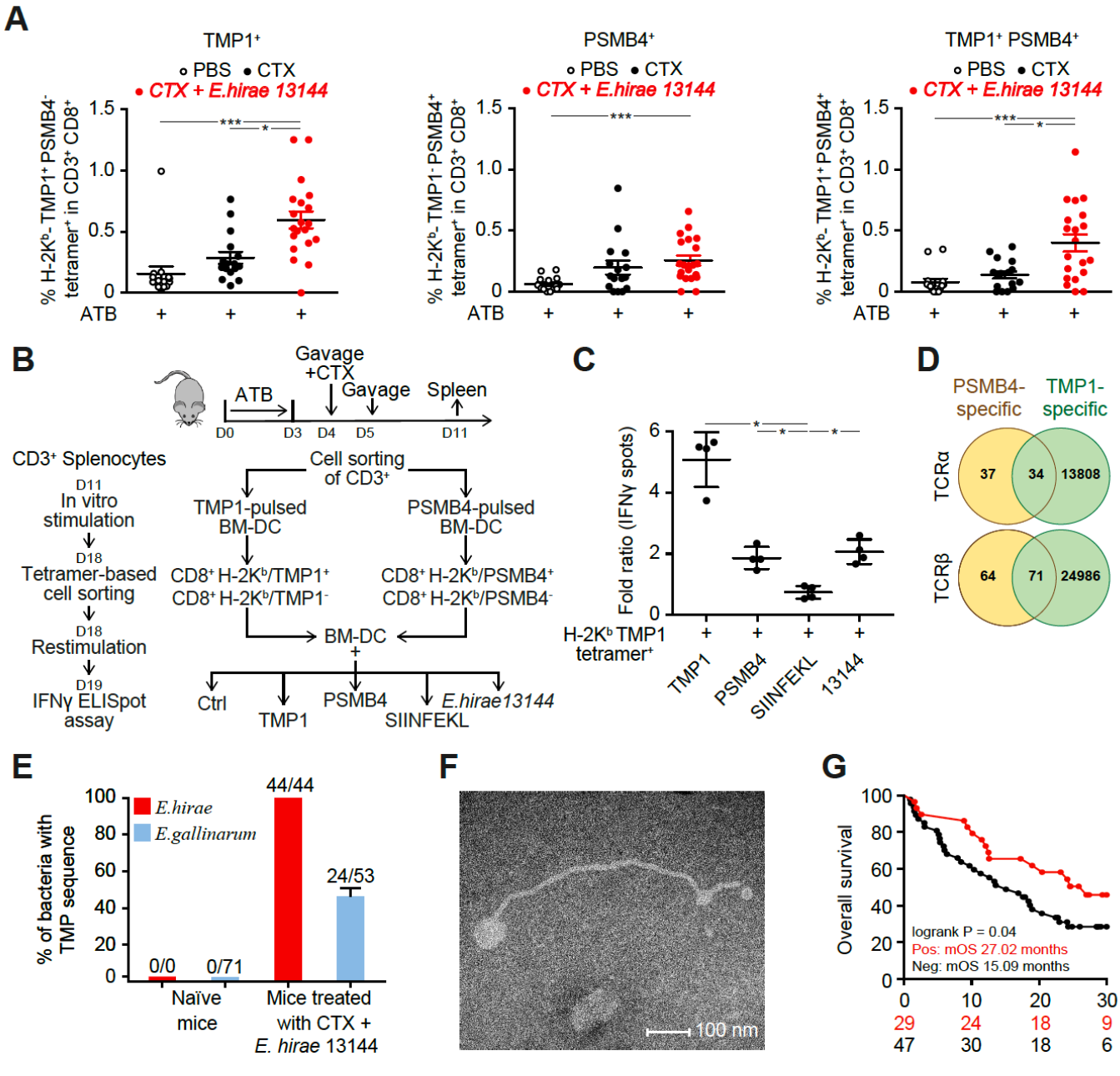

A
TMP1+
PSMB4+
TMP1+ PSMB4+
PBS CTX
CTX + E.hirae 13144
% H-2Kb- TMP1+ PSMB4- tetramer+ in CD3+ CD8+
% H-2Kb- TMP1- PSMB4+ tetramer+ in CD3+ CD8+
% H-2Kb- TMP1+ PSMB4+ tetramer+ in CD3+ CD8+
ATB
B
Gavage +CTX
Gavage
Spleen
D0 D3 D4 D5 D11
CD3+ Splenocytes
D11 In vitro stimulation
D18 Tetramer-based cell sorting
D18 Restimulation
D19 IFNγ ELISpot assay
Cell sorting of CD3+
TMP1-pulsed BM-DC
PSMB4-pulsed BM-DC
CD8+ H-2Kb/TMP1+
CD8+ H-2Kb/TMP1-
CD8+ H-2Kb/PSMB4+
CD8+ H-2Kb/PSMB4-
BM-DC
Ctrl
TMP1
PSMB4
SIINFEKL
E.hirae13144
C
Fold ratio (IFNγ spots)
H-2Kb TMP1 tetramer+
TMP1
PSMB4
SIINFEKL
13144
D
PSMB4-specific
TMP1-specific
TCRα
TCRβ
37
34
13808
64
71
24986
E
% of bacteria with TMP sequence
E.hirae
E.gallinarum
0/0
0/71
44/44
24/53
Naïve mice
Mice treated with CTX + E. hirae 13144
F
100 nm
G
Overall survival
logrank P = 0.04
Pos: mOS 27.02 months
Neg: mOS 15.09 months
0 10 20 30
29 24 18 9
47 30 18 6

Figure 4

Figure 4. TMP crossreacts with the PSMB4 cancer epitope and affects human anticancer immune responses.

A. Flow cytometry analysis of $CD8^+$ tumor-infiltrating lymphocytes (from tumors treated as in Fig. 1A) after co-staining with two different tetramers (H-$2K^b$/TMP1 and H-$2K^b$/PSMB4, sequences in Fig. 3A). Each dot depicts one tumor. The graphs assemble the results of 3 independent experiments with 5 mice/group. B,C. Purified $CD3^+$ T splenocytes from animals treated with CTX and *E. hirae 13144* were restimulated *ex vivo* with bone marrow-derived dendritic cells (DC) loaded with TMP1 or PSMB4 peptide. One week after *ex vivo* restimulation, peptide- specific $CD8^+$ T cells were purified after staining with the corresponding tetramer to measure IFNγ secretion in response to DC loaded with peptides (TMP1, PSMB4, SIINFEKL as negative control) or heat-inactivated *E. hirae* 13144. These results were performed in parallel on the tetramer-binding *versus* non-binding fraction and were normalized to the PBS controls (Ctrl). Each dot represents one culture. Mann Whitney test or ANOVA statistical analyses (Kruskal- wallis test): *$*p<0.05$, ** $p<0.01$, *** $p<0.001$*. D. Venn diagram of TCRα and β chains from tetramer positive $CD8^+$ T cells specific for PSMB4 (yellow) or TMP1 (green). E. Lysogenic conversion of *E. gallinarum* by the *E. hirae* siphoviridae phage *in vivo*. Ileal content was obtained from naïve mice or from mice receiving *E. hirae* together with cyclophosphamide (CTX), followed by cultivation and isolation of bacterial colonies, MALDI-TOF identification and PCR-based detection of TMP. Results are from 5 mice/group. F. Transmission electron microscopy of the phage produced by *E. hirae* 13144. G. Kaplan Meier survival plots of 76 patients with non-small cell lung cancer or renal cell cancer subjected to PD-1-targeting immunotherapy, stratified according to the presence or absence of TMP in at least 5 *E. faecalis* or *E. hirae* colonies/patient. Univariate Log-rank (Mantel-Cox) analysis. Refer to the statistical report.

Science
AAAS

Crossreactivity between MHC class I-restricted antigens from cancer cells and an enterococcal bacteriophage.

Authors: Aurélie Fluckiger, Romain Daillère, Mohamed Sassi, Barbara Susanne Sixt, Peng Liu, Friedemann Loos, Corentin Richard, Catherine Rabu, Maryam Tidjani Alou, Anne-Gaëlle Goubet, Fabien Lemaitre, Gladys Ferrere, Lisa Derosa, Connie PM Duong, Meriem Messaoudene, Andréanne Gagné, Luisa De Sordi, Laurent Debarbieux, Sylvain Simon, Clara-Maria Scarlata, Maha Ayyoub, Belinda Palermo, Francesco Facciolo, Romain Boidot, Richard Wheeler, Ivo Gomperts Boneca, Zsofia Sztupinszki, Krisztian Papp, Istvan Csabai, Edoardo Pasolli, Nicola Segata, Carlos Lopez-Otin, Zoltan Szallasi, Fabrice Andre, Valerio Iebba, Valentin Quiniou, David Klatzmann, Jacques Boukhalil, Saber Khelaifia, Didier Raoult, Laurence Albiges, Bernard Escudier, Alexander Eggermont, Fathia Mami-Chouaib, Paola Nistico, Nathalie Labarrière, François Ghiringhelli, Bertrand Routy, Vincent Cattoir, Guido Kroemer*, and Laurence Zitvogel*.

*Correspondence to: laurence.zitvogel@gustaveroussy.fr; kroemer@orange.fr

This PDF file includes:

Methods:

Cell culture, reagents and tumor cell lines. MC38, TC1, MCA205 (WT or PSMB4-mutated) tumor cell lines or clones were cultured at 37°C with 5% CO_2 in RPMI 1640 medium containing 10% fetal calf serum (FCS), 2 mM L-glutamine, 100 UI/mL penicillin/streptomycin, 1 mM sodium pyruvate and MEM non-essential amino acids (henceforth referred to as complete RPMI 1640). All these reagents were purchased from Gibco-Invitrogen (Carlsbad, CA, USA).

Mice. All animal experiments were carried out in compliance with French and European laws and guidelines and regulations. The local institutional board approved all mouse experiments (permission number: 2016-109-7450). All mouse experiments were performed at the animal facility in Gustave Roussy Cancer Campus where animals were housed in specific pathogen-free conditions. Female C57BL/6 were purchased from Harlan (Gannat, France). Mice were used at an age between 7 and 12 weeks of age.

Antibiotic treatments. Mice were treated during 3 days (biotinylated) an antibiotic (ATB) solution containing ampicillin (1 mg/mL), streptomycin (5 mg/mL), colistin (1 mg/mL) and vancomycin (0.25 mg/mL) (Sigma-Aldrich) added to the sterile drinking water of mice. Antibiotic activity was confirmed by cultivating fecal pellets resuspended in brain heart infusion (BHI) broth + 15% glycerol at 0.1 g/mL on COS (BD Columbia agar with 5% sheep blood, BioMérieux) plates for 48h at 37°C in aerobic and anaerobic conditions. In the context of bacterial or fecal transplantation, mice received 3 days of ATB before undergoing bacterial or fecal transplantation the next day by oral gavage using animal feeding needles. ATB were not used for Figure 3E and Figure S6E.

Tumor challenge and treatment. Syngeneic C57BL/6 mice were inoculated subcutaneously (s.c.) with 1×10^6 MC38 colon cancer cells, 0.8×10^6 MCA205 sarcoma cells or 0.8×10^6 TC1 lung cancer cells. When tumors reached 20 to 35 mm^2 in size, the mice were treated intraperitoneally (i.p.) with cyclophosphamide (CTX, 100mg/kg) (Endoxan Baxter, was provided by Institut de Cancérologie Gustave Roussy, Villejuif, France) or anti-PD-1 mAb (250µg/mouse; clone RMP1-14) or isotype control (clone 2A3) (BioXcell, NH, USA). Depending on the experimental setting, mice were injected with CTX once or 3 times at 1-week intervals. Mice

were injected 4 times at 3-day intervals with anti-PD-1 mAb. Tumor size was routinely monitored every 3 days by means of a caliper.

Gut colonization with dedicated commensal species. *Enterococcus hirae* 13144 were originally isolated from spleens of SPF mice treated with CTX in our laboratory. *E. hirae 708* was provided by INRA (P. Langella), while *E. hirae 13344, ATCC9790* were provided by Prof. Cattoir, CHU de Caen, France. *L. plantarum* was provided by Prof. Ivo Gomperts Boneca from the Institut Pasteur strain repository, France. All *E.hirae* IGR strains were isolated from the stools of NSCLC patients in our laboratory, according to patient informed consent and local IRB approval (ancillary study "Oncobiotics"). All bacteria were grown in COS plates for 24 to 48 hours at 37°C in aerobic conditions. Colonization of ATB pre-treated mice was performed by oral gavage with 100 µl of suspension containing 1×10^9 bacteria. For bacterial gavage, we used suspensions of 10^{10} CFU/mL, monitored using a fluorescence spectrophotometer (Eppendorf) at an optical density of 600 nm in PBS. Depending on the experimental setting, 2 or 6 bacterial gavages were performed for each mouse: the first, the same day as CTX injection, and then 24 hours after the injection of CTX. For anti-PD1 mAb, 5 bacterial oral gavages were performed for each mouse: the first, the same day and 24h before the first anti-PD1 injection, and the same day for the three other injections of anti-PD1 Abs. The efficacy of colonization was confirmed by culturing the feces 48 hours post-gavage. Fecal pellets were harvested and resuspended in BHI+15% glycerol at 0.1 g/mL. Serial dilutions of feces were plated onto COS plates and incubated for 48 hours at 37°C in aerobic and anaerobic conditions. After 48 hours, the identification of specific bacteria was accomplished using a Matrix-Assisted Laser Desorption/Ionisation Time of Flight (MALDI-TOF) mass spectrometer (Andromas, Beckman Coulter, France).

Culture and propagation of bone marrow-derived dendritic cells. Bone marrow-derived dendritic cells (BM-DCs) were generated by flushing bone marrow precursors from the femurs and tibia of female C57Bl/6 WT mice aged between 8 and 12 weeks. Bones were collected in sterile PBS, washed in alcohol and Iscove's medium (IMDM, Sigma-Aldrich) baths, extremities of bones were cut and flushed using a 26G needle. After red blood cell lysis, cells were cultured in IMDM supplemented with 10% of FCS + 2mM L-glutamine + 100 UI/mL penicillin/streptomycin + 50µM 2-mercaptoethanol (Sigma-Aldrich) (referred herein as complete

IMDM medium) at 0.5 × 10^6/mL and treated with 40ng/mL of GM-CSF (supernatant of GM-CSF transfected-cells J558) and 10 ng/mL of recombinant interleukin-4 (IL-4) for BM-DCs (from Peprotech). Cells were split at day 3 and used in experiments on day 7 or 8.

Test of memory TC1 immune response and H-2K^b restricted-peptides on splenic CD8$^+$ T cells. Interferon-γ (IFN-γ) ELISPOT assay were performed in 96-well PVDF bottomed sterile plates (Millipore MSIP S4510) by means of a commercial kit (Cell sciences, Newburyport, US) according to the manufacturer's instructions. After PVDF membrane activation with ethanol 35%, plates were coated overnight with capture antibody to IFN-γ and washed before incubation of blocking buffer during 2 hours. BM-DC (1 × 10^5/well) were exposed to heat-inactivated (2 hours at 65°C) bacterial strains (*E. hirae* 13144, *E.hirae* 708, *E.hirae* 13344 and *L.plantarum* at a multiplicity of infection [MOI] of 1:10) or pulsed with peptides (20µg/mL) and were added to CD8$^+$ T cells (2 x 10^5/well) for 20 hours at 37°C. Cells were then removed and plates were developed with a biotinylated antibody specific for IFN-γ during 1 hour and 30 minutes, followed by streptavidin-alkaline phosphatase during 1 hour. Finally, the substrate of streptavidin (BCIP/NBT buffer) was added for 5-20 min. Spots were counted by means of a CTL Immunospot Analyzer (Cellular Technology Limited, Cleveland, OH).

Vaccination of mice. BM-DCs were activated with poly I:C (10µg/mL, Invivogen) overnight before infection with heat-inactivated (2 hours at 65°C) bacterial strains (MOI 10) or pulsed with peptides (20µg/mL, peptide 2.0). After 6 hours of incubation with bacteria or 1 hour of incubation with peptides, BM-DCs were washed 3 times with PBS before subcutaneous injection in the right flank of mice (1.5 x 10^5 cells per mice). Mice were vaccinated twice at 10 days apart and challenged 4 weeks after the second vaccination with the minimal tumorigenic dose of MCA205 tumor cells in left flank.

Flow cytometry analyses. In experiments without tumor, spleens were harvested 7 days after the injection of CTX. In tumor growth experiments, spleens, tumors and tumor draining lymph node were harvested at different time points, 7, 14 and 21 days after the first injection of CTX into mice bearing MCA205 tumors. Excised tumors were cut into small pieces and digested in RPMI medium containing Liberase^{TM} at 25 µg/mL and DNase1 at 150 UI/mL (Roche) for 30 minutes at

37°C and then crushed and filtered twice using 100 and 40µm cell strainers (Becton & Dickinson, BD). Lymph nodes and spleen were crushed in RPMI medium and subsequently filtered through a 70 µm cell strainer. Two million splenocytes, tumor cells or lymph node cells were pre-incubated with purified antimouse CD16/CD32 (clone 93; eBioscience) for 15 minutes at 4°C, before membrane staining. Dead cells were excluded using the Live/Dead Fixable Yellow dead cell stain kit (Life Technologies). Anti-mouse antibodies for CD3 (145-2C11), CD4 (GK1.5), CD8 (eBioH35-17.2), CXCR3 (CXCR3-173), CCR9 (CW-1.2), and TMP specific tetramer (BD, BioLegend, eBioscience and Cliniscience). Stained samples were acquired on Canto II 7 colors cytometer (BD) and analyses were performed with FlowJo software (Tree Star, Ashland, OR, USA).

Human T cell responses to HLA-A*0201 restricted-TMP epitopes. Cytapheresis cones were collected from healthy volunteers (Etablissement français du sang, EFS) and peripheral blood mononuclear cells (PBMC) were separated using a Ficoll Hypaque (Sigma Aldrich) gradient. We selected only donors with the HLA-A02*01 haplotype determined by immunofluorescence and flow cytometry. PBMC were washed and resuspended in the separation medium (PBS, 1mM ethylenediaminetetraacetic acid, 2% human AB^+ serum) for magnetic bead separation. $CD14^+$ monocytic cells (human CD14 MicroBeads, Miltenyi) were enriched from 75×10^6 peripheral blood mononuclear cells (PBMC) and cultured at 0.5 x 10^6/mL in IMDM supplemented with 10% human AB^+ serum, 1% of 2 mmol/L glutamine (GIBCO Invitrogen), 1000 IU/mL GM-CSF and 1000 IU/mL IL-4 (Miltenyi). Cells were split at day 3 and used in experiments on day 6 or 7. Such (DC-like) cells were seeded in 96-well plates at 1×10^5 cells/well either alone or in the presence of peptides (20µg/mL) for 2 hours at 37°C, 5% CO2. The remaining autologous PBMC fractions were enriched for $CD8^+$ T cells ($CD8^+$ T Cell Isolation Kit, human, Miltenyi). The enriched $CD8^+$ T cells were washed, counted and resuspended at 1×10^5cells/well in RPMI-1640 supplemented with 10% human AB+ serum, 1% 2 mMol/L glutamine, 1% penicillin/streptomycin (GIBCO Invitrogen) and 50 U/mL IL-2 (Proleukin). DC-peptide/ T cell co-cultures were incubated for one week at 37°C, 5% CO_2 (medium was changed every 2 days). Then, the pools of cells were seeded in 96-well ELIspot plates at 2 x 10^5 cells/well and restimulated with or without peptides (20µg/mL) or anti-CD3/anti-CD28 coated beads (1µL/mL, Dynabeads T-Activator, Invitrogen) as a positive control for 20 hours at 37°C. IFN-γ ELISPOT

assays were performed in 96-well PVDF bottomed sterile plates (Millipore MSIP S4510) by using a IFN-γ ELISPOT kit (Cell sciences, Newburyport, Etats-Unis) according to the manufacturer's instructions.

In vitro stimulation of PBMCs from healthy volunteers and cancer patients with HLA-A2-restricted phage and cancer peptides. Cytapheresis cones were collected from healthy volunteers (EFS) and peripheral blood mononuclear cells (PBMC) were separated using a Ficoll Hypaque gradient. We selected only donors with HLA-A02*01 haplotype determined by flow cytometry with anti-HLA-A2 antibody (BB7-2 clone). 4 x 10^7 PBMC were seeded in 96 well/plates at 2 x 10^5 cells/well in RPMI 1640 medium supplemented with 8% human serum (HS), 50 IU/mL of IL-2 (Proleukin, Novartis) and stimulated with 5 µM of KLAKFASVV peptide. After 14 days, each microculture was evaluated for the percentage of specific $CD8^+$ T lymphocytes by double staining with a HLA-A2- KLAKFASVV peptide tetramer and anti-CD8 mAb (Clone RPA-T8, Biolegend) using a FACS Canto HTS. Cross-reactivity of positive microcultures was evaluated by double staining with a HLA-A2 KLQKFASTV peptide tetramer and anti-CD8. HLA-A2 /peptide monomers were produced by the recombinant platform facility P2R, from SFR Santé, as previously described (*1*). Microcultures that contained at least 0.5% of specific T cells were selected, pooled and sorted with the relevant tetramer-coated beads and amplified as previously described (*2*). After the amplification step, purity and cross-reactivity of sorted T cell lines were evaluated by tetramer/CD8 double labeling. CD107A mobilization and TNFα production were evaluated after stimulation of sorted T cells with 5 µM of each peptide (KLAKFASVV or KLQKFASTV). After a 5 hour-stimulation period in the presence of brefeldin A at 10 µg/mL (Sigma, B7651), T cells were labeled with phycoerythrin (PE)-conjugated anti-CD8 antibody (Clone RPA-T8, Biolegend) and fixed with PBS 4% paraformaldehyde (VWR, 100504-858). Lymphocytes were then stained for TNF production using APC conjugated anti-TNFα (clone Mab11, Biolegend). Concerning CD107A labeling, specific T cells were stimulated for 3 hours at 37°C in the presence of Alexa-F647-conjugated mAb specific for CD107A (clone H4A3, Biolegend). T cells were then stained with anti-CD8 antibody (Clone RPA-T8, Biolegend) and analyzed by flow cytometry.

Short-term Ag-specific T cell lines from HLA-A*0201 lung cancer patients. Peripheral blood was collected from non-small cell lung cancer (NSCLC) patients at the time of surgery, after informed consent and PBMC were isolated on a Ficoll Hypaque gradient. Only PBMC from patients bearing the HLA-A*02*01 haplotype were used for *in vitro* short-term Ag-specific stimulation. $CD8^+$ T cells were positively enriched using an anti-CD8-coated magnetic microbeads (Miltenyi Biotec) selection process resulting in more than 95% purity. $CD8^+$ T cells were seeded in 96-well plates at 2 x 10^5 cells/well in RPMI-1640 medium supplemented with 10% human serum. Autologous CD8-depleted PBMC were used as antigen presenting cells (APCs), irradiated, pulsed with TMP epitope 10 (KLAKFASVV) (5 μg/mL) for 2 hours at 37°C in 5% CO_2 and plated with $CD8^+$ T cells at a 1:3 ratio. After 24 hours, human recombinant IL-2 (Miltenyi Biotec) and IL-7 (PeproTech Inc.) (25 U/mL and 5 ng/mL, respectively) were added to the culture wells. After one week, cells were restimulated with the soluble TMP epitope 10 peptide (1 μg/mL) for an additional week before functional analysis.

HLA-A2/peptide tetramer staining. Phycoerythrin (PE)-labeled HLA-A*0201/peptide (KLAKFASVV) tetramers were used. FITC-CD8 monoclonal antibody was purchased from Miltenyi Biotec (BW135/80). Briefly, 2-3 x 10^5 short-term *in-vitro* expanded T cells were first incubated with tetramer (10 μg/mL, 30 minutes, room temperature). After washing, cells were incubated with FITC-CD8 mAb (20 minutes, 4°C). Dead cells were excluded using propidium iodide staining (MP Biomedicals). Cells were immediately acquired on BD FACS Celesta and analyzed using FACS Diva software (BD).

Interferon (IFN)-γ and granzyme B (GrB) production. 3-4 x 10^5 autologous CD8-depleted PBMC isolated from NSCLC patients were pulsed with the relevant TMP epitope 10
(KLAKFASVV) or the irrelevant TMP epitope 14 (KMAALAASA) (1 μg/mL) (Figure S13E). Alternatively, instead of autologous APC, T2 cell lines (purchased from the American Type Culture Collection and routinely checked for mycoplasma using Mycoplasma PCR Reagent, Euroclone, Italy) were used at 1 x 10^5 cells/well and pulsed (or not) with the relevant TMP epitope 10, the irrelevant TMP epitope 14, or with irrelevant MART-1 A27L (ELAGIGILTV) and gp100 209-217 (IMDQVPFSV) peptides (1 μg/mL) (Figure S13F). Whenever indicated, the HLA Class I-blocking antibody (W6/32 mAb) was added (*3*). After 1 hour incubation at 37°C in

5% CO2, autologous 2-3 x 10^5 $CD8^+$ T-cell lines were added to monitor antigen-specific activation markers, IFN-γ and GrB production. Cells were co-cultured for 5 hours at 37°C in 5% CO_2, in the presence of the protein transport inhibitor GolgiStop (BD). After 5 h, T cells were collected, washed in PBS and incubated for 30 minutes at 4°C with the following mAbs from BD Biosciences: PE-CD3 (SP34-2), APC-H7-CD8 (SK1), FITC-CD4 (SK3), BV786-CD137 (4B4-1). After washing, cells were fixed and permeabilized by means of the Cytofix/Cytoperm kit (BD Biosciences), following the manufacturer's instructions. Intracellular staining was performed for 30 minutes at room temperature by the use of following mAbs from BD Biosciences: PE-Cy7-IFN-γ (B27), and Alexa Fluor647-GrB (GB11). Cells were immediately acquired on a FACS Celesta and analyzed using FACS Diva software (BD).

MART-1 and MELOE-1-specific T cell clones and responses to bacterial peptides.

Principles.

We identified microbial analogs of non-mutated tumor-associated antigens relevant to human malignancies (such as the MART-1/Melan-A melanoma differentiation antigen or MELOE-1 abberrantly expressed antigen). In the public microbiome database (metaHIT), 5 and 11 microbial sequences shared more than 78% homology with EAAGIGILTV (from MART-1/Melan-A) (Figure S14A) and TLNDECWPA (from MELOE-1) (Figure S15A), respectively. The cross-reactivity of 11 MART-1/Melan-A -specific T-cell clones to each of the 5 bacterial peptides was measured. All the 11 T cell clones recognized 2 out of the 5 bacterial peptides with EC_{50} values similar to those found for the MART-1/Melan-A-AA27L peptide (Figure S14B). Another bacterial peptide was recognized by 2 of the 11 T cell clones, though with a low affinity (Fig S14B). The cross-reactivity of all the MART-1/Melan-A-specific T-cell clones tested might be linked to the frequent occurrence of TRAV12-2 segments (which are highly flexible) (*33*) within the alpha chains of their TCRs (Table S8). We also evaluated 11 microbial peptides for their capacity to stimulate 10 MELOE-1 specific T cell clones, which exhibited a bias towards another TRAV segment (TRAV19, Table S9). Four out of 10 MELOE-1 specific T cell clones responded to at least 1 bacterial peptide (Figure S15B). One of these peptides, differing from the cognate peptide at positions 6 (P6) and 8 (P8) (predicted as weak binder to the HLA-A2 molecule), was recognized by 3 MELOE-1-specific T-cell clones with EC_{50} values similar to the one observed for the WT MELOE-1 peptide (Figure S15B). The two other analogous peptides with different P6

and P8 residues were also recognized by two T cell clones, with an EC_{50} around 10^{-9} M (Figure S15B), suggesting that these two positions are not essential for TCR recognition.

Calculation of EC50 for bacterial epitopes.

EC50 of MART-1 and MELOE-1-specific T-cell clones were evaluated for each bacterial peptide, by measuring TNFα production after co-culture with TAP-deficient T2 cells loaded with a range of peptides, at an effector/target ratio of 1:2. After a 5h-stimulation period in the presence of brefeldin A at 10 µg/mL (Sigma, B7651), T cells were labeled with PE-conjugated specific anti-CD8 antibody (Clone RPA-T8, BioLegend) and fixed with PBS 4% paraformaldehyde (VWR, 100504-858). Lymphocytes were then stained for cytokine production using APC conjugated anti-TNFα (clone cA2, Miltenyi Biotec).

Stool detection of phage TMP sequence by PCR in human or mouse samples (Figure S10).

We cultivated the stools (from cancer patients) or ileal material (mice) after several dilutions in aerobic conditions and permissive medium to allow for the isolation of enterococci colonies (according to a procedure described in (*4*). We performed a PCR of the TMP sequence in each single cultivatable *Enterococcus* colony. One colony was placed into 100µl of nuclease-free water to release the bacterial DNA and PCR was performed with 5µl of DNA, 12.5µl of PCR master mix (Thermofischer Scientific), 5µl nuclease-free water and 1.25µl of pairs of TMP primers (20 µM) (refer to Figure S10 for the position of the probe sets). PCR products were separated on 1.5% agarose gel containing ethidium bromide and revealed by UV exposure. The sequence of primers are: forward 5'-ACTGCAGCCGTAAAATGGGA-3' and reverse 5'-TCCGTATCGTTTGCCAGCTT-3' (amplicon 1026 bp).

Lysogenic conversion of *E. gallinarum* by the *E.hirae* 13144 phage.

In vivo. To investigate the capacity of the *E. hirae 13144* phage to lysogenize other bacterial species *in vivo*, we performed culturomic analyses of the ileal content from C57BL/6 mice subjected to oral gavage with *E. hirae* 13144 and systemic CTX therapy, followed by PCR analyses seeking TMP sequences (Figure S8A-B). We tested 7 to 18 bacterial colonies from each animal and a total of 76 colonies. We only found lysogenic conversion of *E. gallinarum* by the *E. hirae*-temperate phage *in vivo*, as confirmed by sequencing of the phage genome in the second

host (Figure 4F, Figure S8B-C). In contrast, none of the 90 colonies (mostly of *E. gallinarum)* isolated from naive mice harbored the TMP sequence (Figure S8A).

In vitro. One *E. gallinarum* strain (isolated from naïve mice) were incubated at a ratio of 1:1 (10^7 of each bacteria), in the presence of small intestinal organoids, during one hour before treatment with mafosfomide (25µg/mL). Six and twenty hours post-incubation, organoid supernatants were plated to allow for the isolation of *E.gallinarum* colonies followed by PCR-based detection of the TMP sequence on each *E.gallinarum* colony. For preparation of small intestine organoids, ileal intestinal crypts were isolated and enriched from 8-12 week old C57BL/6 mice as previously described (*5*) with the following modifications. Briefly, pieces of ileum washed in PBS were incubated in PBS containing 2mM EDTA for 30 minutes on ice. Fragments were then rinsed 3 times with PBS containing 10% FCS and filtered through a 70-µm cell strainer. Crypts were pelleted, washed with Advanced DMEM/F12 (ADF) (Invitrogen), resuspended in 1mL of Cultrex PathClear Reduced Growth Factor BME (Bio-Techne, Minnesota, United States) and 50µL drops were pipetted into a 24 well plate. Drops were overlayed with ADF containing the following: 100 U/mL penicillin G sodium, 100 µg/mL streptomycin sulfate, 2 mM L-glutamine, 10 mM HEPES, 1x N2 supplement, 1x B27 supplement, 50 ng/mL mEGF, 100 ng/mL mNoggin (Peprotech, Hamburg, Germany), N-acetylcysteine (Sigma) (reagents from Invitrogen unless otherwise indicated) and 10% conditioned medium of Cultrex® HA-R-Spondin1-Fc 293T Cells (Bio-Techne). Organoids were passaged once per week and utilized in experiments 7 days post splitting.

Transmission electron micrographs of the bacteriophage. The negative staining of particles was realized on supernatant of a co-culture of *E. hirae* 13144 and *E. gallinarum* admixed at a 1:10 ratio for 20 hours. Staining was performed using a 5% solution of ammonium molybdate. Images were acquired using a Tecnai G2, operating at 200 keV. Scale bars are shown on micrograph pictures.

Generation of TMP-expressing *E. coli.* A DNA fragment containing the P23 promoter sequence was generated by annealing two complementary primers (5'-CAATAAAAAATCAGACCTAAGACTGATGACAAAAAGAGCAAATTTTGATAAAATAG TATTAGAATTAAATTAAAAAGGGAGGCCAAATATAG-3' and 5'-

GATCCTATATTTGGCCTCCCTTTTTAATTTAATTCTAATACTATTTTATCAAAATTTGC TCTTTTTGTCATCAGTCTTAGGTCTGATTTTTTATTGCATG-3'). The sequence was then inserted into SphI/BamHI-digested vector pDL278 (Addgene 46882, gift from Gary Dunny) (*6*) to generate vector pDL278-P23. A part of the TMP gene (N-terminal 1185 nucleotides of TMP, including the epitope TSLARFANI, fused to a C-terminal FLAG-tag) was amplified from E. hirae 13144 genomic DNA (5'-TCCGGATCCATGGCACAAAGTAAAACAGTCAAAGCG-3', 5'-CAGGAATTCTTACTTGTCGTCATCGTCTTTGTAGTCACGTAGTAAACTATCACGTAAT CGAACTTC-3') and inserted into BamHI/EcoRI-digested vector pDL278-P23 to generate vector pDL278-P23-TMP-FLAG. Mutations in the epitope were introduced using the QuikChange Lightning Kit (Agilent). Primers 5'-AACGAGCTAAGGCAGTAGCAGCTGTATCTGCAGAC-3' and 5'-GTCTGCAGATACAGCTGCTACTGCCTTAGCTCGTT-3' were used to mutate position 2 (S to A, pDL278-P23-TMP-mut2-FLAG), primers 5'-ATTAGCAAAACGAGCGAAGGAAGTAGCAGCTGTATCTG-3' and 5'-CAGATACAGCTGCTACTTCCTTCGCTCGTTTTGCTAAT-3' were used to mutate position 3 (L to F, pDL278-P23-TMP-mut3-FLAG). To generate the control plasmid pDL278-P23-EGFP, EGFP was amplified from pCIB1(deltaNLS)-pmGFP (Addgene 28240, gift from Chandra Tucker (*7*)) using primers 5'-CTTGGATCCATGGTGAGCAAGGGCGAG-3' and 5'-CAGGAATTCCTACATAATTACACACTTTGTC-3' and inserted into BamHI/EcoRI-digested vector pDL278-P23. Plasmids were transformed into chemically competent E.coli DH5α (NEB) and the presence of plasmids with the correct insert was verified by sequence analysis (5'-CCCAGTCACGACGTTGTAAAACG-3' and 5'-GAGCGGATAACAATTTCACACAGG-3'). Expression of EGFP and TMP-FLAG in E. coli was verified by western blot analysis using antibodies targeting GFP (Cell Signaling, 2956) or FLAG (Sigma-Aldrich, F7425), respectively.

CRISPR/Cas9-mediated mutations of mouse *Psmb4* in MCA205 and TC1 lung cancer cells. Wild type MCA205 cell line was purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA) and was maintained in RPMI-1640 medium (Thermo Fisher Scientifc, Inc., Waltham, MA, USA) supplemented with 10% FBS (Thermo Fisher Scientific, Inc), 100 U/mL penicillin and 100µg/mL streptomycin (Thermo Fisher Scientific, Inc.) at 37˚C. For the CRISPR knock in mutations, we designed the gRNA (sequence AGATATTGCGGAAACGAGCC) by using the CRISPR design tool developed by the Zhang lab (http://crispr.mit.edu/).

Oligonucleotides containing the designed sequence were synthesized (Sigma) and ligated into the pX458 backbone (Addgene #48138, (*8*)) containing the Cas9 gene (human codon-optimised and fused with 2A-GFP allowing for selection) under a CBh promoter and the cloned sgRNA under a U6 promoter. Homology templates (sequence attached) containing the mutation sites were synthesized by Invitrogen GeneArt Gene Synthesis (Thermo Fisher Scientifc, Inc.). The cloned pX458 plasmid and synthesized homology arms were cotransfected into MCA205 cells by means of lipofectamine 3000 (Thermo Fisher Scientific, Inc.) following the manufacturer's protocol. Forty-eight hours after transfection, GFP-positive cells were sorted to 96-well plates as single cells before surviving clones were expanded in duplicated conditions, one for frozen storage at - 80 and the other for genomic DNA extraction. The targeted region in genomic DNA from clones was further amplified by PCR using the Phusion® High-Fidelity PCR Master Mix (New England BioLabs; pswich, MA, USA) and primers 5'CTCAGGGACCCTTTTCACGA 3' and 5'CCCACTCCCTGTTCTACACA 3', and purified with the Monarch® DNA Gel Extraction Kit (New England BioLabs) before being sent to Eurofins Genomics GmbH (BERSBERG GERMANY) for sequencing with the primer 5'GGACCCTTTTCACGATTCAGG 3'. Positive clones were expanded and subjected to DNA extraction for further validating the mutations.

Genome sequencing and analysis. The whole genome sequence of 5 *E. hirae* (13144, 708, 13152, 13344 and EH-17) strains was determined with PacBio technology (GATC Biotech, Konstanz, Germany). Genomic DNA was isolated from 15 other *E. hirae* isolates using the using the Quick-DNA fungal/bacterial miniprep kit (Zymo Research, Irvine, CA) according to the manufacturer's recommendations. After DNA shearing, the DNA libraries were prepared using the NEBNext Ultra DNA library prep kit for Illumina (New England Biolabs, Ipswich, MA) and sequenced as paired-end reads (2 x 300 bp) using an Illumina MiSeq platform and the MiSeq reagent kit version 3. The Illumina reads were trimmed using Trimmomatic (*9*), quality filtered with the Fastx-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and assembled using SPAdes (*10*). Protein sequences were predicted using prokka v1.11 software (*11*). Prophage regions were detected using PHAST software. Predicted proteins were annotated using BLASTp against the National Center for Biotechnology Information (NCBI) non- redundant (NR) database.

Phylogenomic and comparative genomics. Single nucleotide polymorphisms in 20 *E. hirae*
genomes was investigated using the parsnp program (*12*) and using the genome of strain 1314

genome as a reference. Phylogenetic analysis was performed by considering the 47,303 polymorphic sites retained in the core genome of the 20 genomes. Maximum likelihood phylogeny was constructed using Fastree (*13*). Phylogenetic tree was visualized using figtree (http://tree.bio.ed.ac.uk/ software/figtree/). Complete proteome sequences of 20 *E. hirae* strains were compared using by BlastP and pairwise alignments using ClustalW. We clustered the *E. hirae* homologous genes using orthoMCL (*14*) on the translated protein sequences of all predicted genes with a conservative parameter value of 70% amino acid sequence identity and 50% sequence coverage. The determination of the different unique core genomes was based on the homology clusters found by orthoMCL.

TCR sequencing of TMP1- and PSMB4-specific CD8$^+$ T cells. H-2K^b-TMP1 tetramer binding CD8$^+$T cells were isolated from spleen, tumor draining lymph nodes and MCA205 tumor beds after animal exposure to CTX+*E.hirae* 13144, using Facs cell sorting and were pooled into two fractions (positive or negative for the H-2K^b-TMP1 staining, regardless of tissue location). H- 2K^b-PSMB4 tetramer binding CD8$^+$T cells from tumor draining lymph nodes and MCA205 tumor beds were cell sorted by Facs after exposure of the animal to CTX+*E. hirae* 13144 and were pooled into 2 fractions (positive or negative for the H-2K^b-PSMB4 staining, regardless of tissue location). Moreover, H-2K^b-PSMB4 tetramer binding CD8$^+$T cells were harvested (cell sorted by FACS) from vaccine draining lymph nodes after immunization of naive mice with PSMB4 peptides admixed with TLR3 ligands. RNA from those T cell pools (positive and negative fractions) were isolated by means of lysis buffer with the RNAqueous-Kit (Invitrogen®) extraction kit, according to the manufacturer's protocol. The RNA concentration and sample integrity were determined on Nanodrop (ThermoFisher). T cell receptor (TCR) libraries were prepared with the RNA from each sample with SMARTer Human TCR α/β Profiling Kit (Takarabio) following the provider's protocol. Briefly, the reverse transcription was performed using TRBC reverse primers and further extended with a template-switching oligonucleotide (SMART-Seq v4). cDNAs were then amplified following two semi-nested PCR: a first PCR with TRBC and TRAC reverse primers as well as a forward primer hybridizing to the SMART-Seqv4 sequence added by template-switching and a second PCR targeting the PCR1 amplicons with reverse and forward primer including Illumina Indexes allowing for sample barcoding. PCR2 are then purified using AMPure beads (Beckman-Coulter). The quantification and integrity of cDNA

samples was carried out using DNA electrophoresis performed on Agilent 2100 Bioanalyser System in combination with the Agilent DNA 1000 kit, according to the manufacturer's protocol. Sequencing has been performed with Miseq (Illumina) SR-300 protocols at Institut du Cerveau et de la Moelle (Paris, France). FASTQ raw data files were processed for TRAs and TRBs sequences annotation using MiXCR (*15*) software (v2.1.10) with RNA-Seq parameter. MiXCR extracts TRA and TRB providing corrections of PCR and sequencing errors. Generation of datasets was done by concatenating the FASTQ raw data files based on the specificity of the different sorted cell population samples from the different organs. We obtained 4 datasets representing $CD8^+TMP^+$ (meaning CD8 binding to H-$2K^b$-TMP1 tetramer) $CD8^+TMP^-$, $CD8^+PSMB4^+$ and $CD8^+PSMB4^-$ TCR repertoires. These repertoires were respectively composed of 40.734, 416.541, 208 and 532.360 unique clonotypes. Venn diagrams and samples comparisons were performed using R software version 3.5.0 (www.r-project.org) and Prism (GraphPad Software, LaJolla, CA). To compare the TCR sharing of $PSMB4^+$ with TMP^+ vs TMP^- TCRs, a random sampling of 13842 αTCRs and 25057 βTCRs was performed 10 times within the TMP^- repertoire (Fig. 4H).

Statistical analyses. A statistical report has been written for each panel (online material). Data analyses and representations were performed with Prism 6 software (GraphPad, San Diego, CA, USA). Tumor size differences were calculated either using Anova or a dedicated software (https://kroemerlab.shinyapps.io/TumGrowth/). Briefly, tumor growth was subjected to a linear mixed effect modeling applied to log pre-processed tumor surfaces. P-values were calculated by testing jointly whether both tumor growth slopes and intercepts (on a log scale) were different between treatment groups of interests. Survival probabilities were estimated using the Kaplan-Meier method. Cutoffs for continuous variables were chosen using the median value or an optimal cutoff approach. Survival curves were evaluated using the log-rank test. All reported tests are two-tailed and were considered significant at P-values <0.05.

References:

M. Bodinier, M. A. Peyrat, C. Tournay, F. Davodeau, F. Romagne, M. Bonneville, F. Lang, Efficient detection and immunomagnetic sorting of specific T cells using multimers of MHC class I and peptide with reduced CD8 binding. Nat. Med. **6**, 707–710 (2000).

N. Labarriere, A. Fortun, A. Bellec, A. Khammari, B. Dreno, S. Saïagh, F. Lang, A full GMP process to select and amplify epitope-specific T lymphocytes for adoptive immunotherapy of metastatic melanoma. Clin. Dev. Immunol. **2013**, 932318 (2013).

C. J. Barnstable, W. F. Bodmer, G. Brown, G. Galfre, C. Milstein, A. F. Williams, A. Ziegler, Production of monoclonal antibodies to group A erythrocytes, HLA and other human cell surface antigens-new tools for genetic analysis. Cell. **14**, 9–20 (1978).

B. Samb-Ba, C. Mazenot, A. Gassama-Sow, G. Dubourg, H. Richet, P. Hugon, J.-C. Lagier, D. Raoult, F. Fenollar, MALDI-TOF identification of the human Gut microbiome in people with and without diarrhea in Senegal. PloS One. **9**, e87419 (2014).

T. Sato, R. G. Vries, H. J. Snippert, M. van de Wetering, N. Barker, D. E. Stange, J. H. van Es, A. Abo, P. Kujala, P. J. Peters, H. Clevers, Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature. **459**, 262–265 (2009).

D. J. LeBlanc, L. N. Lee, A. Abu-Al-Jaibat, Molecular, genetic, and functional analysis of the basic replicon of pVA380-1, a plasmid of oral streptococcal origin. Plasmid. **28**, 130– 145 (1992).

M. J. Kennedy, R. M. Hughes, L. A. Peteya, J. W. Schwartz, M. D. Ehlers, C. L. Tucker, Rapid blue-light-mediated induction of protein interactions in living cells. Nat. Methods. **7**, 973–975 (2010).

F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, F. Zhang, Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. **8**, 2281–2308 (2013).

A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. **30**, 2114–2120 (2014).

A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, P. A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. J. Comput. Mol. Cell Biol. **19**, 455–477 (2012).

T. Seemann, Prokka: rapid prokaryotic genome annotation. Bioinforma. Oxf. Engl. **30**, 2068–2069 (2014).

T. J. Treangen, B. D. Ondov, S. Koren, A. M. Phillippy, The Harvest suite for rapid core- genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. **15**, 524 (2014).

S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**, 307–321 (2010).

L. Li, C. J. Stoeckert, D. S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic

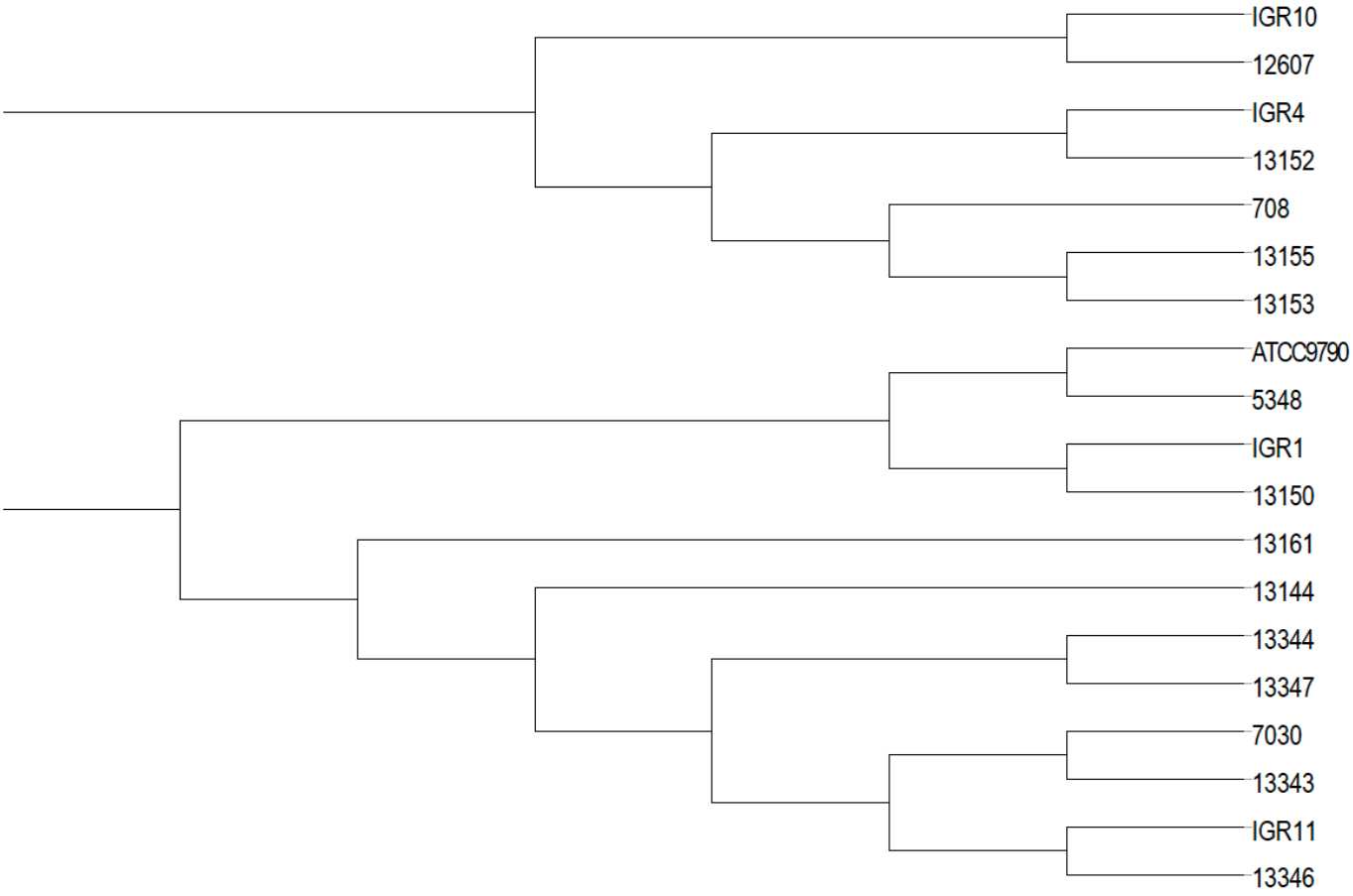
genomes. Genome Res. **13**, 2178–2189 (2003).

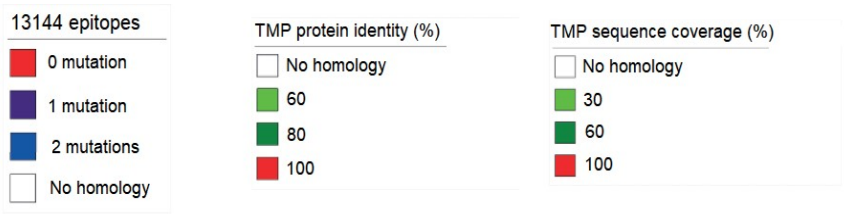
D. A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, M. Chudakov, MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods. **12**, 380–381 (2015).

B. Routy, E. Le Chatelier, L. Derosa, C. P. M. Duong, M. T. Alou, R. Daillère, A. Fluckiger, M. Messaoudene, C. Rauber, M. P. Roberti, M. Fidelle, C. Flament, V. Poirier- Colame, P. Opolon, C. Klein, K. Iribarren, L. Mondragón, N. Jacquelot, B. Qu, G. Ferrere, C. Clémenson, L. Mezquita, J. R. Masip, C. Naltet, S. Brosseau, C. Kaderbhai, C. Richard, H. Rizvi, F. Levenez, N. Galleron, B. Quinquis, N. Pons, B. Ryffel, V. Minard-Colin, P. Gonin, J.-C. Soria, E. Deutsch, Y. Loriot, F. Ghiringhelli, G. Zalcman, F. Goldwasser, B. Escudier, M. D. Hellmann, A. Eggermont, D. Raoult, L. Albiges, G. Kroemer, L. Zitvogel, Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science. **359**, 91–97 (2018).

S. Simon, Z. Wu, J. Cruard, V. Vignard, A. Fortun, A. Khammari, B. Dreno, F. Lang, S. J. Rulli, N. Labarriere, TCR Analyses of Two Vast and Shared Melanoma Antigen-Specific T Cell Repertoires: Common and Specific Features. Front. Immunol. **9**, 1962 (2018).

A


IGR10
12607
IGR4
13152
708
13155
13153
ATCC9790
5348
IGR1
13150
13161
13144
13344
13347
7030
13343
IGR11
13346


13144 epitopes
0 mutation
1 mutation
2 mutations
No homology
TMP protein identity (%)
No homology
60
80
100
TMP sequence coverage (%)
No homology
30
60
100

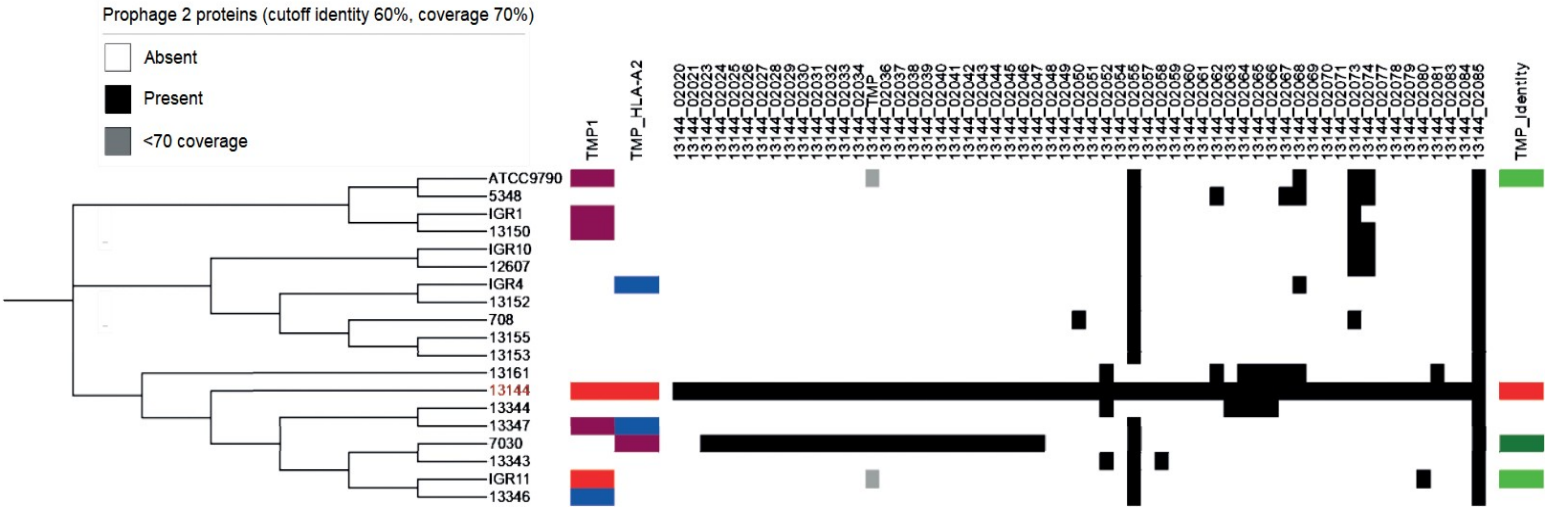

Prophage 2 proteins (cutoff identity 60%, coverage 70%)
Absent
Present
<70 coverage
TMP1
TMP_HLA-A2
TMP_identity
ATCC9790
5348
IGR1
13150
IGR10
12607
IGR4
13152
708
13155
13153
13161
13144
13344
13347
7030
13343
IGR11
13346

Figure S1. Clading and comparative analysis of *E. hirae* 13144 39.2kb-prophage protein sequence with other *E. hirae* strains.

A. Phylogenetic tree of 19 *E.hirae* genomes based on SNP alignments.

B. A particular genomic trait of *E. hirae* 13144 is that it encodes two intact prophage regions (of

40.6 kb and 39.2 kb) showing weak sequence identities with the most common *Enterococcus* phage phiEf11 *vB_EfaS_IME197* (14% and 11% of shared genes, respectively) (Table S3). Comparative analysis of the 39.2kb prophage of *E. hirae* 13144 with 18 other sequenced *E. hirae* genomes showed that this phage was strain-specific, although portions of its genome were detectable in other *E.hirae* strains. Comparative analysis through a "heatmap" clustering based on a matrix of presence (black) and absence (white) of the *E.hirae* 13144 39.2kb-prophage protein sequence or the TMP1 epitope and HLA-A2 TMP epitope 10 without mutation (red), with 1 mutation (violet) or 2 mutations (blue).

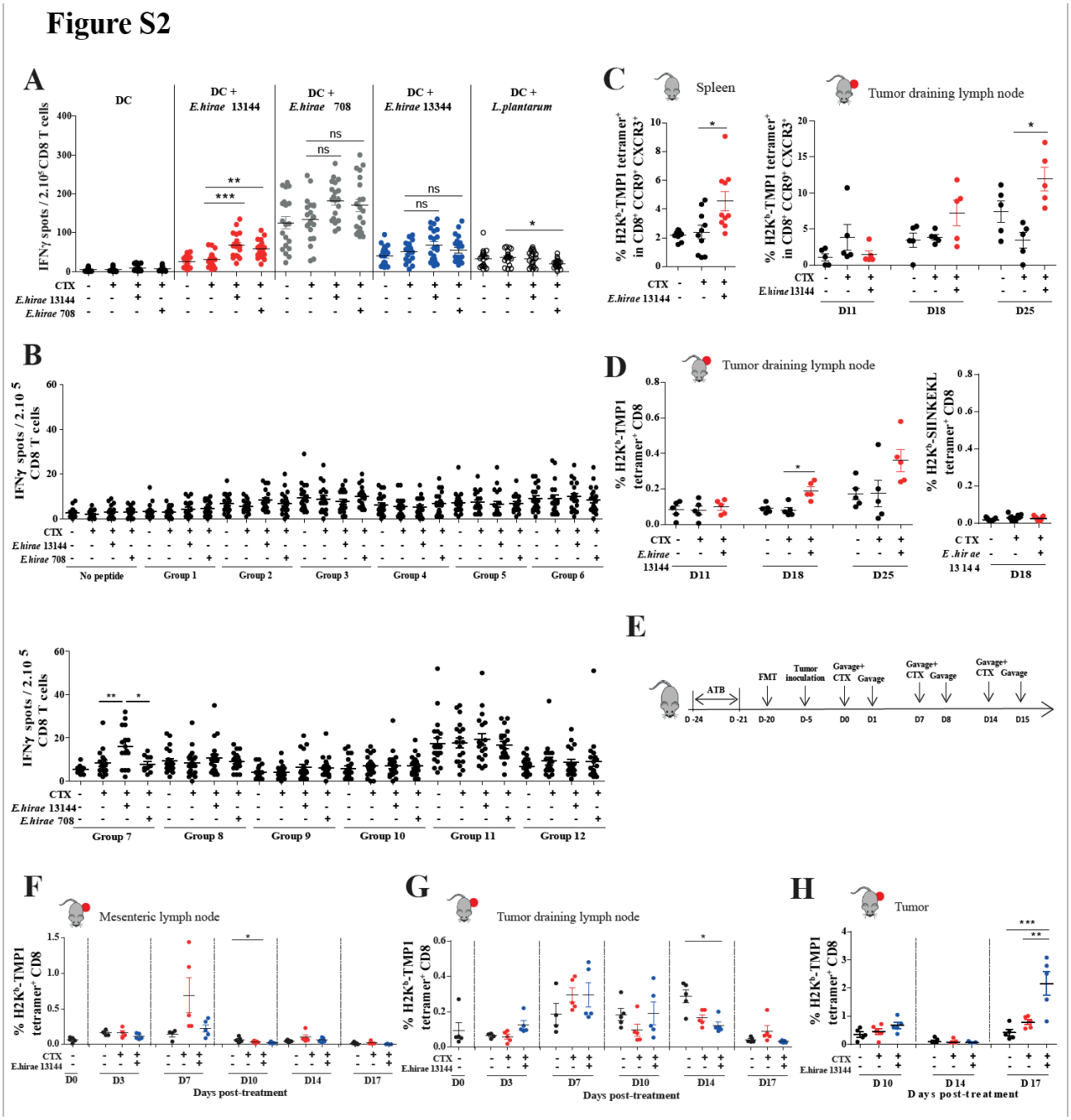


Figure S2. Identification of group 7 as the only group of peptides containing an immunogenic one.

A-B. Naive mice were treated with broad spectrum antibiotics (streptomycin, colistin, ampicillin, vancomycin) for 3 days before oral gavage with *E. hirae* strain 13144 or 708 ($1x10^9$ bacteria) was performed prior to and after systemic administration of cyclophosphamide (ip CTX - 100 mg/kg) or saline solution (NaCl) at day 4 once (like in Fig. 1C). One week later, purified $CD8^+$ T

splenocytes were restimulated *ex vivo* in a recall assay with bone marrow-derived DC loaded with saline or distinct heat killed (65°C during 2 hours) bacterial strains (A) or groups of peptides (Table S2) (B). IFNγ-secreting $CD8^+$T cells (spots) were determined after 24h of coculture. Each dot represents one mouse. The experiment (with 5 mice/group) was performed three times. Statistical analyses revealed that only group 7 of peptides induced a significant response (B). C. Flow cytometric determination of H-$2K^b$/TSLARFANI tetramer-binding $CD8^+$ T cells in the spleen of naive mice (left panel) and tumor draining lymph node (right panel) of tumor bearers in the gate of $CXCR3^+CCR9^+$ double positive cells. D. Flow cytometry analyses of H-$2K^b$/TSLARFANI (left panel) or H-$2K^b$/SIINFEKL (right panel) tetramer binding CTL in tumor draining lymph nodes at various time points in tumor bearing mice treated with CTX and *E. hirae* 13144 (like in Fig. 1A). One representative experiment out of two yielding similar conclusions is shown. Each dot represents one animal, each experiment containing 5-6 mice/group. Data from two independent experiments are depicted. E-H. Avatar mice are SPF C57BL/6 animals treated with 3 days of ATB to allow establishment of a fecal microbial transplant (FMT from a breast cancer patient) 21 days prior to *CTX*+*E. hirae* 13144 therapy (E). Kinetic study of H-$2K^b$/TSLARFANI tetramer-binding $CD8^+$ T cells by flow cytometry analyses in various organs (mesenteric lymph nodes (mLN) (F), tumor draining lymph nodes (tdLN) (G) and tumor beds (MCA205) (H) during a therapy of established MCA205 with the combination of CTX+*E. hirae* 13144 in human fecal material (FMT) subjected avatar mice according to the experimental setting detailed in E. Each dot represents one animal, each experiment containing 5-6 mice/group. Data from one representative experiment are depicted. ANOVA statistical analyses (Kruskal-wallis test): *$*p<0.05, **p<0.01, ***p<0.001$*. Refer to the statistical report.

Figure S3

TMP 39.2-kb prophage (1506 aa)

> *E.hirae*_13144_02029

Figure S3. *E. hirae* 13144 39.2kb-prophage sequence alignment with location of the TMP protein and TMP epitopes. The position of the H-2K^b-restricted TMP1 epitope and HLA- A2*0201 TMP epitope 10 are indicated in red and green, respectively.

igure S4 A

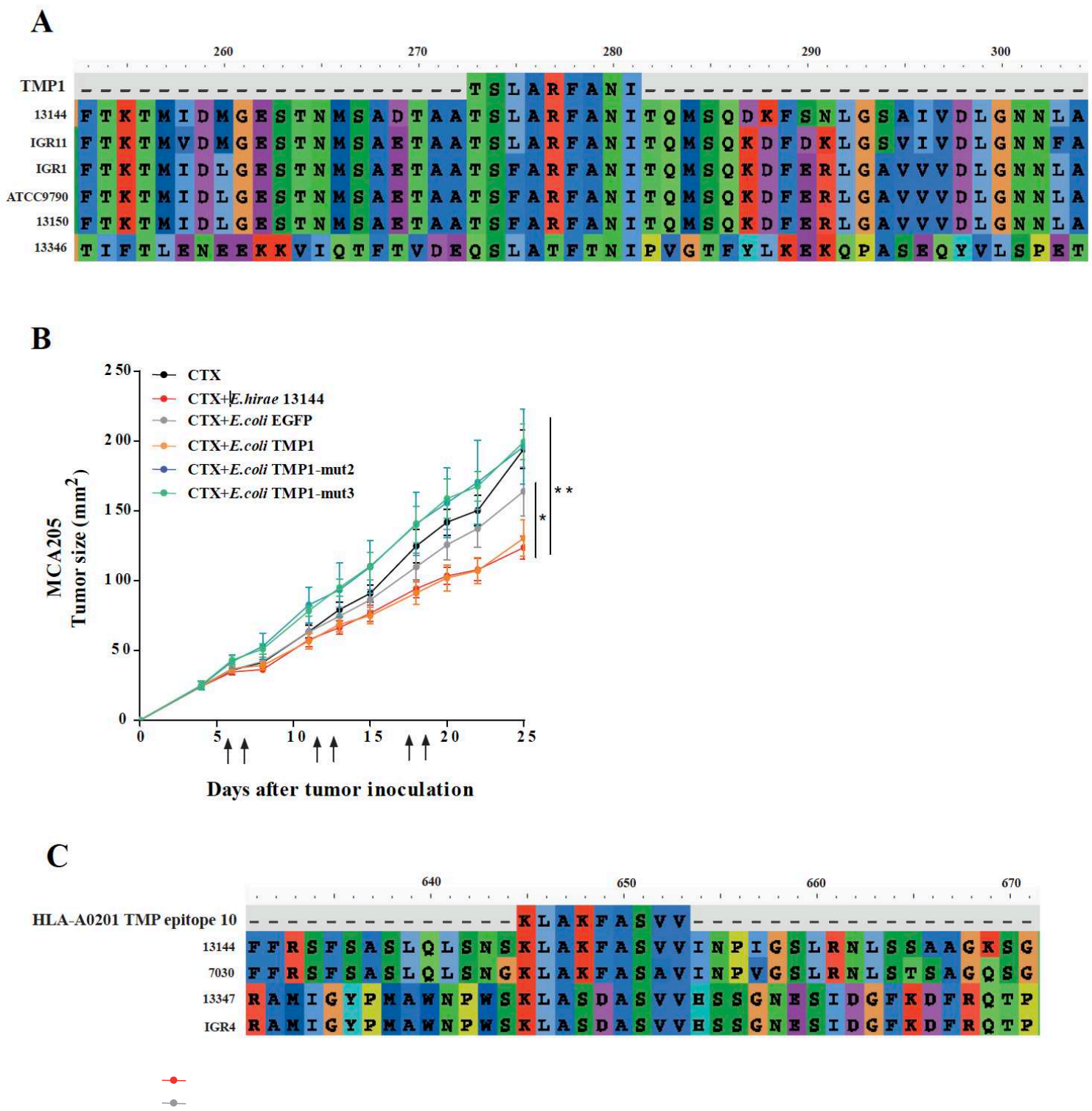

A
260
270
280
290
300
TMP1
13144
IGR11
IGR1
ATCC9790
13150
13346
B
CTX
CTX+E.hirae 13144
CTX+E.coli EGFP
CTX+E.coli TMP1
CTX+E.coli TMP1-mut2
CTX+E.coli TMP1-mut3
MCA205 Tumor size (mm2)
Days after tumor inoculation
C
640
650
660
670
HLA-A0201 TMP epitope 10
13144
7030
13347
IGR4

Figure S4. Sequence alignment of the immunogenic epitope region within 39.2kb-prophage of *E. hirae* 13144. A. The immunogenic peptide TMP1 (TSLARFANI, A) from *E. hirae* 13144 identified in Figure 1 and Figure S2 is aligned to sequences from other *E.hirae* strains tested in this study. B. Complete tumor growth curves of the experiment shown in Fig. 2D. C57BL/6 mice bearing MCA205 sarcomas were treated with CTX and gavaged with *E. hirae* 13144 or *E.coli* genetically modified to express TMP1 (TSLARFANI), TMP1 mut2 (TALARFANI), TMP1 mut3 (TSFARFANI) or EGFP sequence (as control). The means±SEM of tumor sizes at different time points for 12-18 animals, gathered from 2-3 independent experiments are shown. C. The immunogenic peptide HLA-A*0201 TMP epitope 10 (KLAKFASVV) from *E. hirae* 13144 identified is aligned to the sequences from other *E.hirae* strains.

Figure S5

A

TMP-FLAG = TSLARFANI

MAQSKTVKAVLTAIDKGFTQTMGSATSSLKKLSSNASDIPSNLNTVSGAMKSFGDKTASIGQSIEKVGGSMTKGITLPIAGAVGAVTTAA
VKWESAFTGVKKTNDEMVDSNGKVIYSYDDLEKGLRDLAKELPTSHEEIAKVAEAAGQLGIKTDKVVGFTKTMIDMGESTNMSADTAA
TSLARFANITQMSQDKFSNLGSAIVDLGNNLATTESEITEMGLRLAGAGKQIGMTEGDIVGFAAALSSVGIEAEAGGSAFSRLMVQMQ
LATETGVKAFEPLKQAVAIQGVSWEKFVHAVNWGGKELTAVSKQMGVPASELKKLYKEASKASGSLEDFANVTGRTGEEFAELFKSNP
SQAMIEFIQGLKDSEKHGISAIKVLDMGITEVRLRDSLLRDYKDDDDK

TMP-mut2-FLAG (mutation in position 2) = TSLARFANI to TALARFANI

MAQSKTVKAVLTAIDKGFTQTMGSATSSLKKLSSNASDIPSNLNTVSGAMKSFGDKTASIGQSIEKVGGSMTKGITLPIAGAVGAVTTAA
VKWESAFTGVKKTNDEMVDSNGKVIYSYDDLEKGLRDLAKELPTSHEEIAKVAEAAGQLGIKTDKVVGFTKTMIDMGESTNMSADTAA
TALARFANITQMSQDKFSNLGSAIVDLGNNLATTESEITEMGLRLAGAGKQIGMTEGDIVGFAAALSSVGIEAEAGGSAFSRLMVQMQ
LATETGVKAFEPLKQAVAIQGVSWEKFVHAVNWGGKELTAVSKQMGVPASELKKLYKEASKASGSLEDFANVTGRTGEEFAELFKSNP
SQAMIEFIQGLKDSEKHGISAIKVLDMGITEVRLRDSLLRDYKDDDDK

TMP-mut3-FLAG (mutation in position 3) = TSLARFANI to TSFARFANI

MAQSKTVKAVLTAIDKGFTQTMGSATSSLKKLSSNASDIPSNLNTVSGAMKSFGDKTASIGQSIEKVGGSMTKGITLPIAGAVGAVTTAA
VKWESAFTGVKKTNDEMVDSNGKVIYSYDDLEKGLRDLAKELPTSHEEIAKVAEAAGQLGIKTDKVVGFTKTMIDMGESTNMSADTAA
TSFARFANITQMSQDKFSNLGSAIVDLGNNLATTESEITEMGLRLAGAGKQIGMTEGDIVGFAAALSSVGIEAEAGGSAFSRLMVQMQ
LATETGVKAFEPLKQAVAIQGVSWEKFVHAVNWGGKELTAVSKQMGVPASELKKLYKEASKASGSLEDFANVTGRTGEEFAELFKSNP
SQAMIEFIQGLKDSEKHGISAIKVLDMGITEVRLRDSLLRDYKDDDDK

B

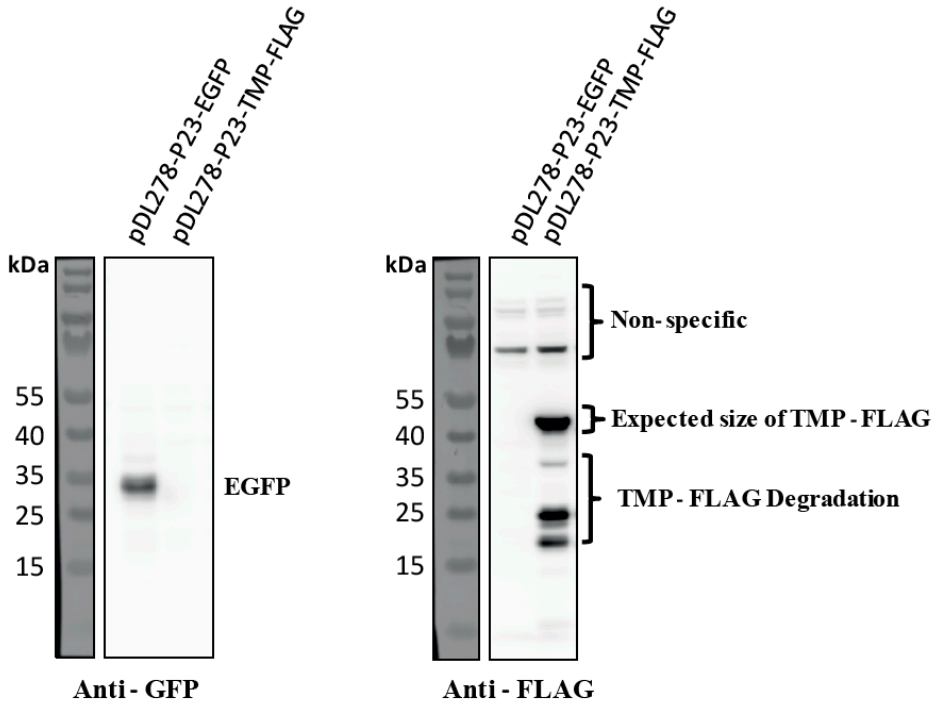


Figure S5. Sub-cloning expression of part of the TMP gene in *E.coli.*

A. Amino acid sequences of TMP-FLAG, TMP-mut2-FLAG and TMP-mut3-FLAG expressed in *E. coli* DH5α. Note that only the N-terminal part of the TMP protein, including the indicated

variants of the epitope (green), was expressed as fusion protein with a C-terminal FLAG tag (blue). B. Western blot analysis demonstrating expression of EGFP and TMP-FLAG in *E. coli* strains transformed with pDL28-P23-EGFP or pDL28-P23-TMP-FLAG, respectively.

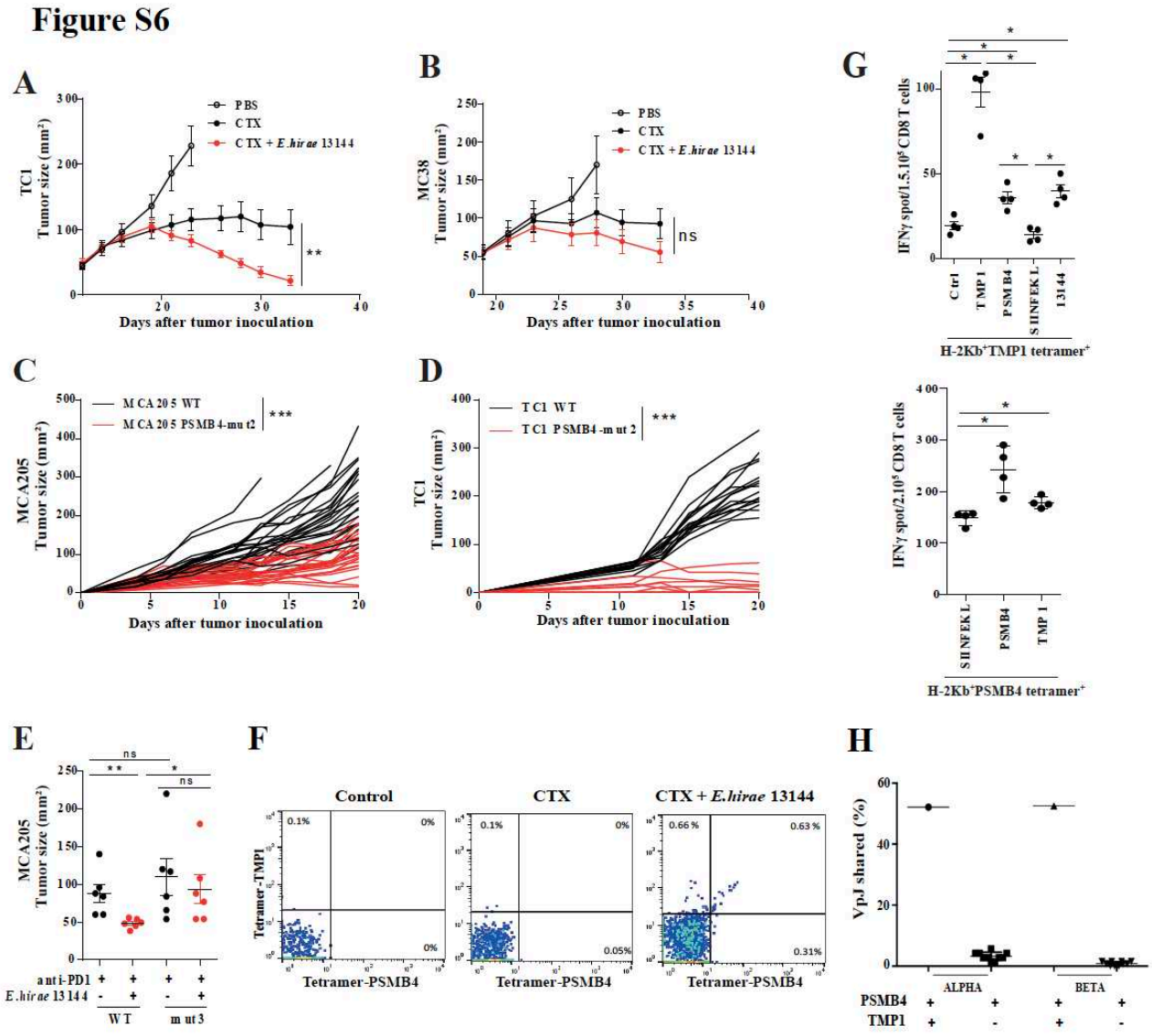


Figure S6. Molecular mimicry between the TMP1 phage and the PSMB4 oncogenic driver.

A-B. Tumor growth of TC1 (A) and MC38 (B) cancers with or without therapy combining CTX ± *E. hirae* 13144 (or saline) following the experimental setting described in Fig1A. C-D. Tumor growth of WT clones or clone harboring a knock-in mutation in position 2 (mut2) of GSLARFRNI for MCA205 (C) and TC1 (D). Each line represents one animal. Two concatenated experiments comprising each 6 mice/group are depicted. E. Tumor sizes at day 9 of SPF mice implanted with WT or mut3 MCA205 clones treated every 3 days three times with anti-PD1 ± *E. hirae* 13144. F. Flow cytometric determination of splenic T cells co-staining with two different tetramers (TMP1 related H-2K^b/TSLARFANI or PSMB4-related GSLARFRNI/H-2K^b complexes). Representative dot plot of CD3$^+$CD8$^+$ splenic T lymphocytes staining with either or both tetramers in one representative tumor bearing animal treated with PBS, or CTX or

CTX+*E.hirae 13144*. G. Cells were prepared following a protocol of *in vitro* expansion (detailed in Figure 4B) in which BM-DC were pulsed with TMP1 or PSMB4 peptide. The number of IFNγ secreting $CD8^+$ GSLARFRNI/H-$2K^{b+}$ (reexpanded after *in vitro* stimulation with PSMB4 peptides, and apostrophed "$CD8^+PSMB4^+$" on the graph) after stimulation with BM-DC pulsed with the three different peptides, and the number of IFNγ secreting $CD8^+$ H-$2K^b$/$TSLARFANI^+$ (reexpanded after *in vitro* stimulation with TMP1 peptides and apostrophed "$CD8^+TMP1^+$" on the graph) after stimulation with BM-DC pulsed with the three different peptides are both depicted in two representative experiments (lower and upper graph respectively). Of note, the binding affinity for H2-K^b of GSLARFRNI peptide is lower than that of TSLARFANI (EC50 :216.85 versus 7.81 nM respectively). We did not observe any IFNγ secretion by the negative fraction $CD8^+$ H-$2K^b$/$TSLARFANI^-$ cells. H. TCRα/β sequences shared between PSMB4-specific TCRs and TCRs from $CD8^+$ T cells that bind H-$2K^b$TMP1-tetramers or fail to do so. Each point represents the result of a random sampling of the same number of clonotypes as for the comparison with TMP1-specific TCRs. Mann Whitney test or ANOVA statistical analyses (Kruskal-wallis test): *$p<0.05$, ** $p<0.01$, ***$p<0.001$. Refer to the statistical report.

Figure S7

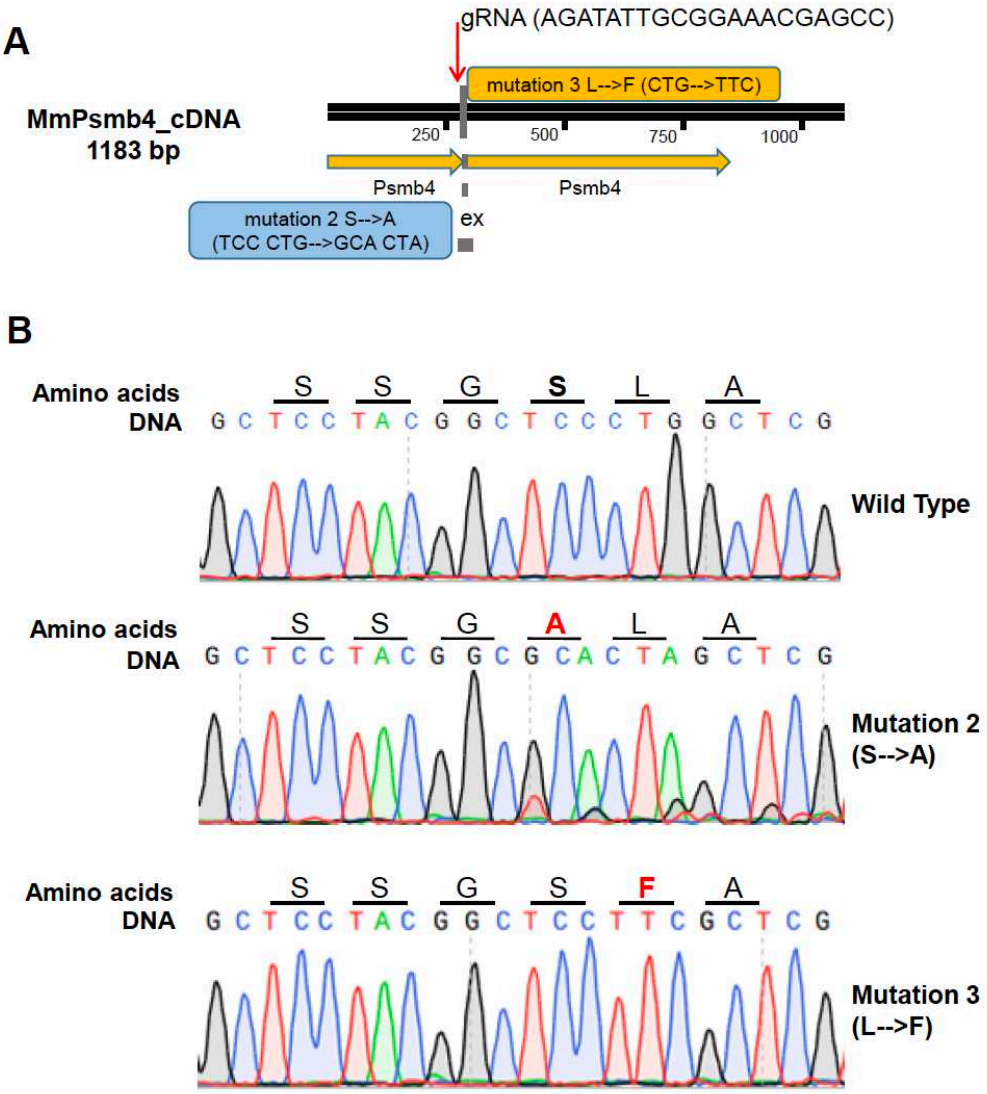


Figure S7. Generation of *pmsb4*-mutated MCA205 cell lines by means of the CRISPR/Cas9 technology. A. Schematic diagrams of *Psmb4* cDNA, and the designed mutation sites. The target site of sgRNA and point mutations are indicated. B. Representative sequence electropherograms for the validation of Pmsb4 mutation 2 and mutation 3 introduced by CRISPR/Cas9. Mutated amino acids are highlighted in red. Similar methods were used for engineering TC1 cells.

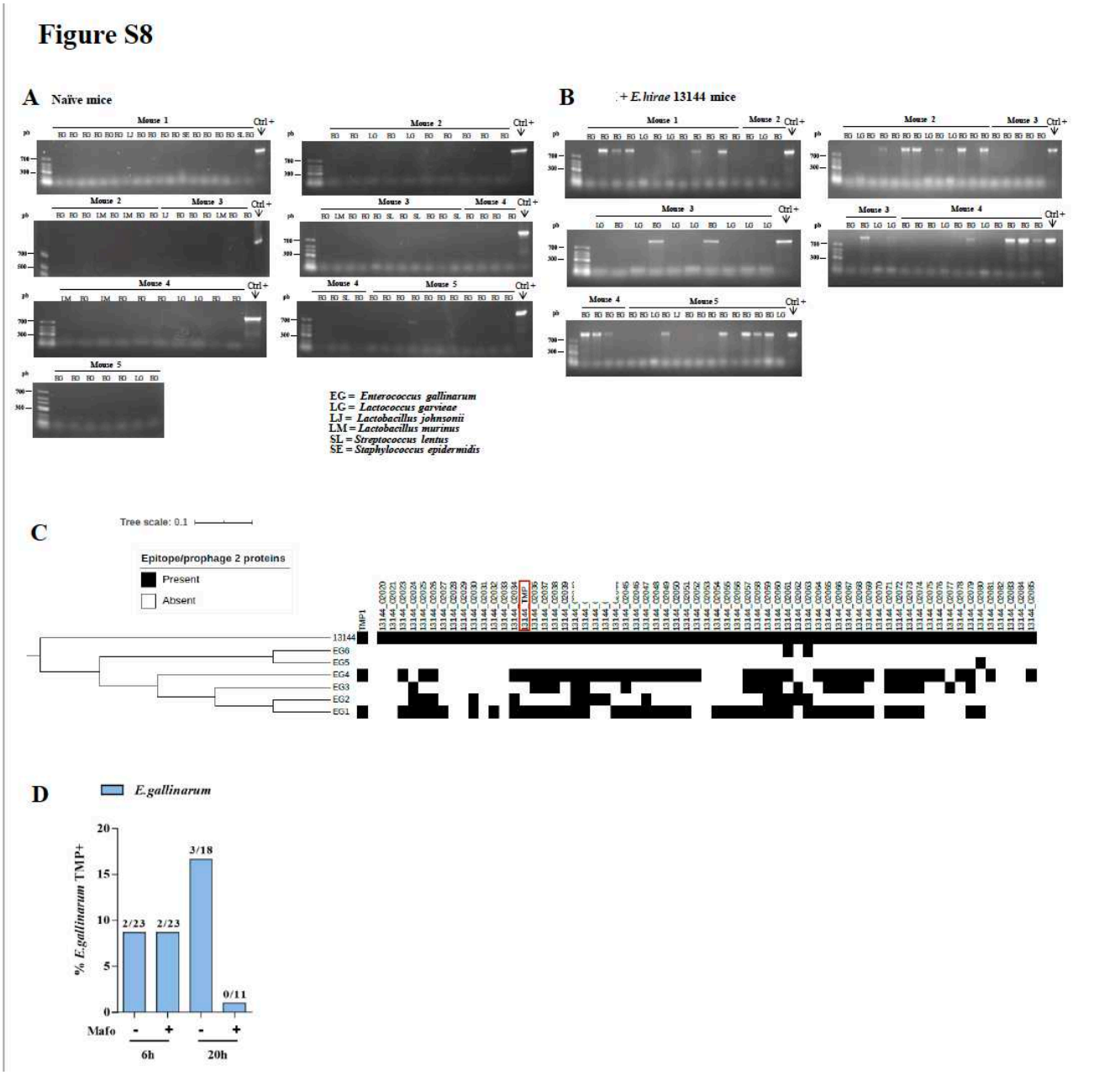


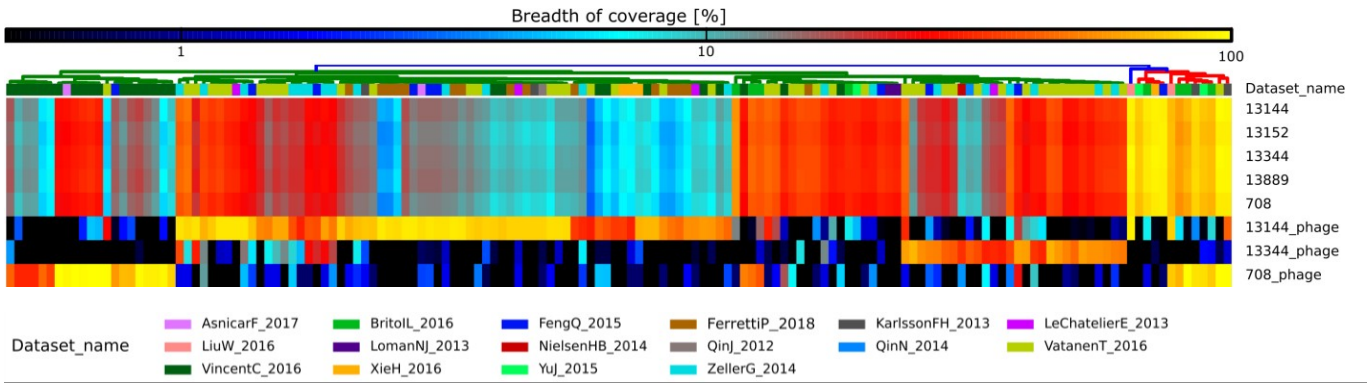
Figure S8. Identification of ileal bacterial colonies after treatment with CTX+oral gavage with *E. hirae* 13144.

A-B. PCR amplification of the TMP sequence in each colony growing after seeding of ileal content in aerobic conditions to isolate Gram$^+$ bacteria. A photograph of each agarose electrophoresis gel is shown for each animal. A depicts the results in 5 naive mice (A) and B depicts the findings after CTX+oral gavage with the phage encoding bacterium *E. hirae* 13144 (10^9 cfu) (B). Each vertical lane corresponds to the bacterium identified in MALDI-TOF. Initials are detailed in the lower part of panel A. The positive control (Ctl+) represents the DNA of *E. hirae 13144*. C. Sequence

alignment of the prophage harbored by *E.hirae 13144* in 6 strains of *E. gallinarum* (EG1 to EG6) harvested from ileal material after oral gavage of naive mice with *E.hirae 13144* and therapy with CTX. Comparative analysis through a "heatmap" clustering based on a matrix of presence (black) and absence (white) of the *E.hirae* 13144 39.2kb-prophage protein sequence or the TMP1 epitope and TMP protein. D. Small intestine stem cell crypt derived-organoids were incubated with *E. hirae* 13144 and *E. gallinarum* at a 1:1 ratio for 6 or 20 hrs ± the CTX derivative mafosfamide. Then, live colonies of *E. gallinarum* were harvested and analyzed by PCR. The percentages of *E. gallinarum* which turned positive for TMP detection are indicated in a representative experiment out of two yielding similar results.

Figure 39

A


Breadth of coverage [%]
1
10
100
Dataset_name
13144
13152
13344
13889
708
13144_phage
13344_phage
708_phage
Dataset_name
AsnicarF_2017
LiuW_2016
VincentC_2016
BritoIL_2016
LomanNJ_2013
XieH_2016
FengQ_2015
NielsenHB_2014
YuJ_2015
FerrettiP_2018
QinJ_2012
ZellerG_2014
KarlssonFH_2013
QinN_2014
LeChatelierE_2013
VatanenT_2016

B

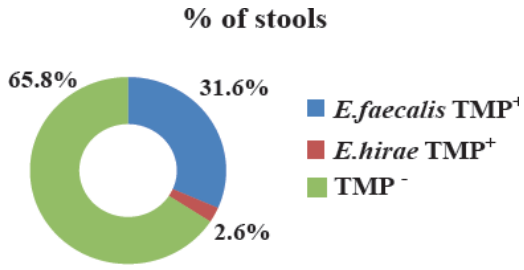

% of stools
65.8%
31.6%
2.6%
E.faecalis TMP+
E.hirae TMP+
TMP -

Figure S9. Breadth of coverage of the *E. hirae* genome and its phage in the MG reference catalog and culturomics analyses of patients stools. A. Breadth of coverage (BOC) of different *E.hirae* sequences *(E.hirae* 13144, 708, ATCC9790) and its phage in 3,027 adult and mother-infant metagenomes (mostly from human stools but also from various mucosae) by reference-based mapping of metagenomic reads from 17 publicly available datasets annotated in curatedMetagenomicData. The BOC measures the fraction of the genome that is covered by the reads in the metagenomes. The color code is indicated with highest BOC towards yellow/red colors and no BOC in black. We first screened a total of 3,027 adult and mother-infant metagenomes (mostly from human stools but also from other mucosae) by reference-based mapping of metagenomic reads from 17 publicly available datasets annotated in curatedMetagenomicData to assess the breadth of coverage (BOC) of the *E. hira*e genome and its phages. *E. hirae* was present with 100% confidence (i.e. BOC > 80%) in 13 samples from disparate geography, age and datasets. In another ~40 cases, the presence of *E. hirae* was very likely but could not be confirmed with high confidence because of insufficient sequencing depth. In 70% of the samples in which *E. hirae* was confidently found, one of the three phage sequences (from *E. hira*e 13144, 708 or 13344) were also detected, though in a partially mutually exclusive fashion (Figure S9A). Of note, the *E. hirae* 13144 phage was detectable in many samples lacking the presence of the *E. hirae* core genome, suggesting that other bacteria than *E. hirae* can host this phage. Analysis of the global prevalence of these phages (irrespective of the presence of *E. hira*e) further confirmed their mutually co-exclusion in the microbiome. We could detect the presence of phage 13144 at 0.66 BOC in three mother-infant paired stool specimens and in the infants at 1, 3, and 7 days after birth suggesting that this phage (and its host) can be vertically transmitted from mothers to infants and then colonize the neonate. We complemented this analysis by a metagenomic-assembly based screening of 9,428 metagenomes, confirming the presence of phage 13144 in humans across the world at a low prevalence (272 positive samples), though possibly with a overrepresentation among a non-Westernized population from Madagascar (19 positive samples). Importantly, this analysis highlighted an increased prevalence of the phage (57%) in fecal microbiomes from children (representing 16% of all metagenomes, Fisher's test p-value <0.00001). We confirmed the integration of the phage into the genome of distinct bacterial species for 128 positive samples (47%), when re-evaluating 154,723 microbial genomes reconstructed from the same 9,428 samples. All host genomes belonged to the

Enterococcus genus (except two assigned to *Coprobacillus*), in particular *E. faecalis* (80 genomes), *E. faecium* (23 genomes), and *E. hirae* (15 genomes), suggesting that phage 13144 (and its homologues from *E. hirae* 708, and 13344) are genus-specific but not species-specific. B. Percentages of stools with detectable TMP sequences in *E. hirae* and/or *E. faecalis* colonies assessed by culturomics followed by PCR in 76 NSCLC and RCC bearing patients (cohort described in (*16*)), the corresponding Kaplan Meier curves indicating overall survival featuring in Figure S11E-F. Up to 5 colonies per species and individual (either colonies from *E. avium, E. casseliflavus, E. durans, E. faecium*, *E. faecalis, E. gallinarum, E. hirae*) from 76 cancer patients led to the detection of the TMP sequence encompassing the TMP1 peptide (aligned in Figure S10) in 34% of the patients. PCR detected TMP sequences only in *E. faecalis, with* 29 colonies positive out of 118 colonies that were tested (not in *E. faecium* nor *E. durans*).

Figure S10

TCAACTACGATACCTCCCATCTTGCTTTTGTTGTTCTGCTAAATTAAGATTCATCGGGCTACCTAGTGCTCCCACTACTTGGCCAGTA
TCCATCACAACAGTTAAATGACGTATTTCTTCTAAGATTTCTACCATTTTCCCCATCGGCGTATTATCTATAGAATGTTTTACCTCAA
TTGCATTTGATCGTTTTATCAAACGACTGCCTGTAATAGACTGGTGAATGCTTGAAATCATATCTTTTGCACTTTGTACGGCAACTA
CGTATCTTCTCGAATACCTGCTGCCACACCCTGTGCAAGGAAAACACCAACGTCATATTTCAATAGGCGTGATGGAGATTTAATCT
TTGCTTTTTTCTGTGCTTCTGCATTAACTGCAGCTACTAAATTTTGCATAGCAGATACTGCTTCTCCTTGGCTAGCTCTAATACCAGA
GCAACACCTTTAGCCATATTAGATCCAACGGATGTCATATTGACTGAACCAGCACCACTTTTTACTGCGTTTCCTAATTCTCTACCA
GCGTTATTAGCTGAACCAACCTGTGATGCCAATCCTTGAACGAAACTACGCCCAAGCTGTTCTCCTACGCTTCGCATAGCAGAAGC
TTTCGTTTTTACTCCTTCAACAGGCGCTGATCCTAATTCTGCACCTTTTTGTTTTGCGTTGCTTTTCTGTTTCGCTAGCCCTGTGTTAT
ACTCTCCAGCATTAGAATTACCGCCGCTATTGTATTCTTTACCTTTACTTTTTGCGCCCTTTGCACCAGATGAAGCTACATCACCAGA
AGTTTTTTCCGCTTCTGCTTTCTTATTTTTTACGCCTGAGTTCATCTGATTCATTAACTCTTGACCGACATTATTGATTTCAACTTTTCC
TGAGTTCAACCCGTCGATTAATTTGTTTTTACCATCTTGACCGTTTTTAAACAAATCAGGCGGTAATGATTGCAAGGTATTCACAAT
GTCAGCTTTTGACATATTCGCCCATAATTTAGGATCGTTGCTTTGCAATCCTTGAACTAGTCCGTTAGAACCATCAATTCCTCGTTG
ACGTAACATTCCAGCTAATAAAGCCATTTGTTGGTCAATGCTAGCACCGTTGTTTACATACGATTGATAAATTCCTAATAACTGTTG
GTCTGTCACTCCTTTTAATTGAGCTAAATTATCAGCCGTCACCGCAATTTTATTTGCACCATTTTGTGAAATAATCGCTAGAAGCTG
CGCTCCTTGTTCTAATTCACTTTGACGTATCTGATCGTTCTGTGTTTGTAATTGCGTAATTTGGTTTTGGAAAGCCGCCTTTTCTGAT
TCAGTTTTCGCTTGGTTCTTTTGTGTTTCCAATTGCTGAATTTGTGCATTATTCTCCTGCACTTGCTGTGCTTGAATTTCTCCAAAAGT
TTTTAAACTTGATAAAGTTTGTTCTTTTTCTTGTTCACTTAACGCTTGGTTGTTATTCAGCTTATTCACACCAGCTTCGACAAACTGTT
TCTGTTGATCCAACAATTGATCACGAATAATATTCGTTTGATTTTGCAAAGTTGCTCTTTGCTGATCGGTTAACTCTTGACCTTCTAC
CGTTTTATTATTCTTCAACTGATTAGAGTAATCTGTGTATACTTTCAACAGATCGCTATTGTTTGATTGAACAGCCTTCATATACTCA
GTTGAAGCATTGGCAAAAATCTTTTGTTTTTCAGCTTCCGATTTACCTTCTGCCGCTTCAATCTGCTTATTATAGGTTTCAACAGCCT
TTTTCTGTTGTTCTTTTAGATTTGTCACTAAATCAAGTGTATGATTGAAATAAGCTTCTACGCCAGCCGTGCTACCGTTTTGCTGTGA
GAAAAGTTCAGTCATTGCCTGTTTAGCTTCATCAAGTTTTGATGAGTAATTTTCAACACTATTTGATACCTCTTCCATATTCATGGAC
ATGGCTTTCGTAGTGTCTTTCGATTTCTTCCCTAATTCTTCTGTGCTTTTAGCTGCTTTTTTTAGAGCAGAATCAGAAAACATGGTAT
CCCAGTCTTTTCCGATATCAGCTAAACTTTTCTTCACATCTTTAAATGATTTATCGGCTCCTTTTGAATCGCCTTTTAATCTTTGCCAA
AGTCCTTTTACTCCGTTAGCAATGGCCATTATTGCATTTACTGCTGTCTTTCCTACAGTAATAATGGCTCGCAATCCATCTACAAAAC
CTGCAATAGCAAAAGTAACTCCGACAAGAGTTCCTGTTCCTAACCATTTAAAAATATTTCCTAATCCTTTTATTGTTTTAGTAACACT
CGCGGAGCTAGGAAGTACACTTTTAAACGATTTTACTATTCCGCTAAAAGCGGTTTTCACGTAGCCTTGAATGTTCATAAAATTGG
ATTTCCAAGCTTGCACTACACCAACTATTGTAGTGGTTATTGCTACTAAAATTGCAGTTATAGGATTGCTCAACATAGCTCCTGTTA
AACTAGCTATAGATCGTATACCCGTTGCTGCAAATGTTCTAAAACCTCCACCTGCTTTTGAGGCGGCTACACCAAGTCCTGATAAA
ACCGTCCCTGATTTACCAGCTGCAGAAGATAAATTCCTTAGCGACCCAATAGGATTAATAACAACGGAGGCGAATTTCGCTAATTT
GCTATTAGATAATTGTAAAGAAGCAGAAAAAGAACGGAAAAAGTTAGTAACTTTATTCCCTTCCCCTAGCATATTTAGCTGTCTTT
GGCTTGCTCGAAGATTTGCTCCAAAAGTGTCCAATGTGGGAAAGAGACCTACAATGGTATCTTTTAGCGTAGTAAAACGGGTAAG
CAGATTTACATTTATCCCCGCACTTTCAAGCCCTGCAAGATTTGATTTATATTTAGAAACAAACCCTTTTACAGCTTGTAATGCGCTA
CTAGAACCGTTTTTGATAGGAGTAACGATAAATTGTTTCCACTTGCTATCTATGTTTCCAGCTTTCTCAAACATTGTTGAAATTGTTT
TGCCAAAAACACTAGTCATTTTCCCAAACACTTTTAATACAGGACCAGCAGAAGCAGCTAATGCAGCCATTTTTAAAATAAATTCTT
GAGTTTTTGGATCAGCTGATGCAAAAGCCTCGGCCATATTTGCTAAAGCTTCAATCATAGGCTTAGCAGCACTTATTGCGCTATTT
AATGCGGCTACTAATGGACCGCCAAACGTAATTGCTACATCGTTTAATTGACCACGTAAAATCTTTAACTGTGATTCTGTAGT**TCCG**
TATCGTTTGCCAGCTTCTTCTGCTAGAGCTGTATTTTCGTTAAACGCTTCGTTACCTCGTTTTACAGCACCTTCAAAGACATCACTCG
CATTAGCCGCACGTAGTAAACTATCACGTAATCGAACTTCGGTAATCCCCATATCATCAAGTACTTTAATAGCTGAGATTCCATGCT
TTTCTGAGTCTTTCAAACCTTGAATAAACTCAATCATAGCTTGAGAAGGATTACTCTTGAATAATTCCGCGAACTCTTCGCCAGTTC
GACCAGTAACATTTGCAAAATCTTCCAAACTTCCAGACGCCTTGCTTGCTTCTTTATATAATTTTTTCAATTCTGAAGCTGGTACTCC
CATTTGTTTAGAAACAGCTGTTAATTCTTTACCACCCCAATTAACAGCATGAACAAATTTTTCCCAAGACACTCCTTGTATAGCTACA
GCTTGTTTTAAAGGTTCAAAAGCTTTAACCCCTGTTTCGGTGGCTAATTGCATTTGTACCATCAACCTAGAAAAAGCTGAACCACCC
GCTTCGGCCTCTATACCAACAGATGATAACGCCGCTGCAAAACCGACAATGTCTCCTTCAGTCATACCAATTTGTTTTCCTGCACCA
GCCAAACGGAGTCCCATTTCTGTGATTTCTGATTCAGTAGTTGCTAAGTTATTCCCTAAGTCAACAATAGCTGAGCCAAGATTGCT
AAATTTATCTTGAGACATTTGAGT**AATATTAGCAAAACGAGCTAAGGAAGT**AGCAGCTGTATCTGCAGACATATTTGTTGATTCG
CCCATATCGATCATTGTTTTAGTAAATCCGACAACTTTATCAGTTTTTATTCCTAACTGTCCAGCTGCTTCTGCTACTTTTGCAATTTC
TTCATGACTAGTAGGTAATTCTTTTGCTAAATCTCTAAGGCCTTTTTCTAAATCATCATAAGAATAAATGACTTTACCGTTAGAATCG
ACCATCTCATCGTTGGTCTTTTTAACACCAGTAAATGCACTT**TCCCATTTTACGGCTGCAGT**TGTGACTGCTCCAACAGCACCCGCA

ATTGGGAGTGTGATACCTTTAGTCATCGAACCGCCGACTTTTTCAATGCTTTGGCCGATACTTGCAGTTTTATCACCAAAACTTTTC
ATCGCACCACTAACTGTGTTCAAATTACTGGGAATATCAGAAGCATTCGAACTAAGTTTTTTTAGCGAAGAGGTAGCACTCCCCAT
TGTCTGAGTAAACCCTTTATCTATTGCTGTAAGTACCGCTTTGACTGTTTTACTTTGTGCCAC

Primers

Sequence of TMP1

Figure S10. Sequence of the Phage Tail Length Tape Measure Protein in *E. hirae*. Nucleotide sequence of the whole TMP protein as well as binding area for PCR primers indicated in green and TMP1 epitope sequence indicated in red.

Figure S11

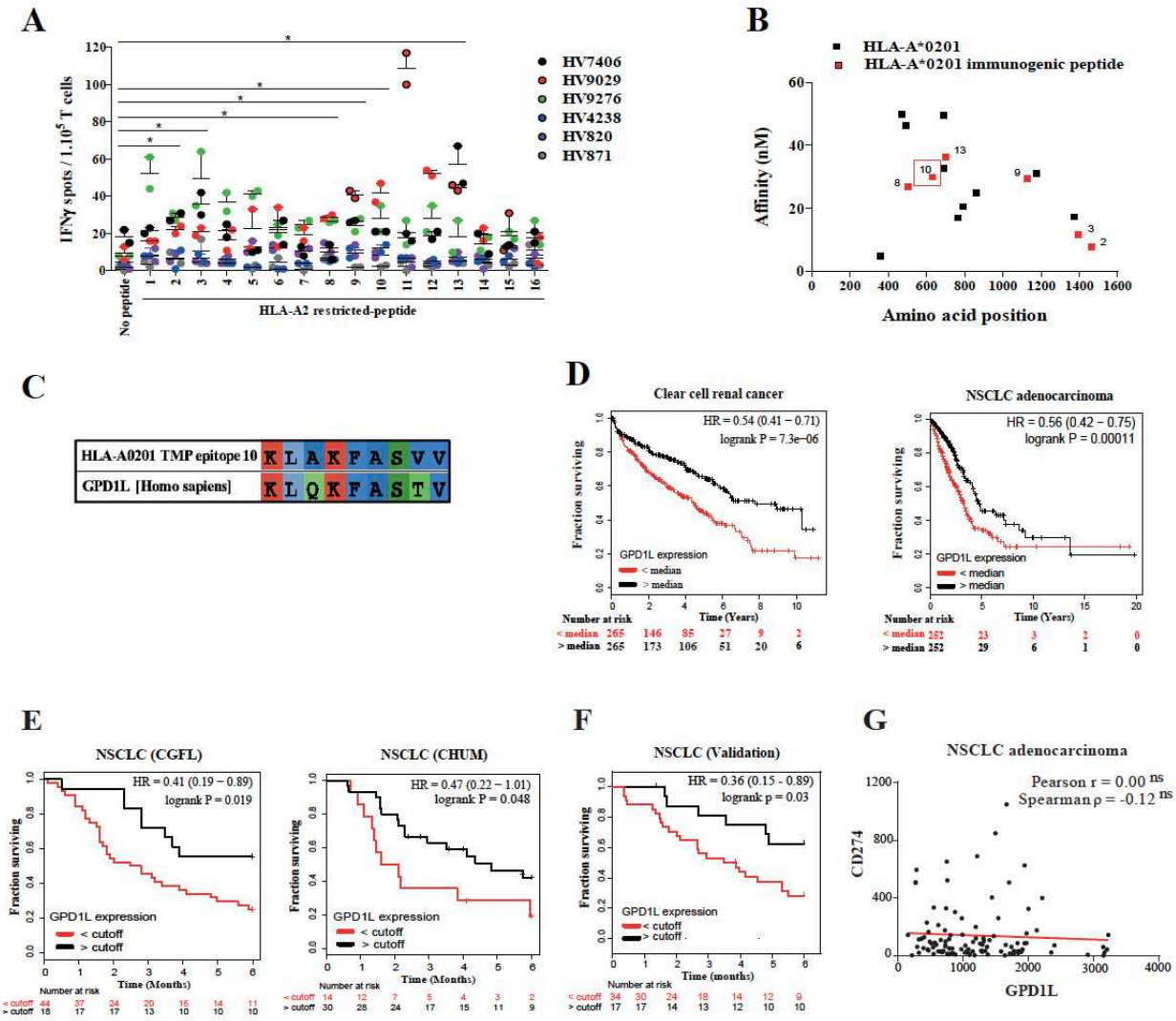


Figure S11. Identification and functional impact of TMP-crossreactive epitopes in the GPD1-L protein .

A. Priming of naive $CD8^+$ T cells from six HLA-A*0201 healthy volunteers with autologous monocyte-derived DC pulsed (or not) with 16 HLA-A*0201 binding TMP epitopes (Table S6). Restimulation at day 7 with each of the 16 TMP peptides for IFNγ ELIspot assays and enumeration of positive spots. ANOVA statistical analyses: *$p<0.05$. B. The HLA-A*0201 binding and immunogenic epitopes (Table S6) are located in defined domains of the TMP protein, as indicated by the color code (red: 6 peptides with significant reactivity in A) and the amino acid sequence position, as a function of their binding affinity to the MHC class I allele (calculated *in silico*). C. Blast sequence alignment of the immunogenic HLA-A*0201-restricted

TMP epitope 10 (KLAKFASVV) with a sequence belonging to the GPD1-L protein (KLQKFASTV). D. Impact of GPD1-L mRNA expression on survival in 530 clear renal cell cancers (left panel) and lung adenocarcinoma (right panel) from the TCGA. Patients were segregated according to the median value of GPD1-L expression, and Kaplan Meier curves of overall survival were compared by Cox regression univariate analysis. E-F. Kaplan Meier curves for time to progression following PD-1 blockade in second line therapy in 44 stage IIIC/IV NSCLC patients (CHUM test cohort, E) corroborated with a second cohort of 62 stage IIIC/IV NSCLC patients (CGFL cohort, E) and then a validation cohort of stage IIIC/IV NSCLC patients (F, n=51) using an optimal cut-off value for GPD1-L tumor expression obtained in RNA sequencing for each cohort (Table S7 for patients description). G. Absence of correlations between GPD1-L and CD274/PD-L1 mRNA expression in lung cancers (TCGA, CHUM and CGFL cohorts together), as determined by Spearman and Pearson calculations.

Figure S12

A **GPD1L**

>sp|Q8N335|GPD1L_HUMAN Glycerol-3-phosphate dehydrogenase 1-like protein

MAAAPLKVCIVGSGNWGSAVAKIIGNNVKKLQKFASTVKMW VFEETVNGRKLTDIINNDHENVKYLPGHKL
PENVVAMSNLSEAVQDADLLVFVIPHQFIHRICDEITGRVPKKALGITLIKGIDEGPEGLKLISDIIREKMGIDISVL
MGANIANEVAAEKFCETTIGSKVMENGLLFKELLQTPNFRITVVDDADTVELCGALKNIVAVGAGFCDGLRCG
DNTKAAVIRLGLMEMIAFARIFCKGQVSTATFLESCGVADLITTCYGGRNRRVAEAFARTGKTIEELEKEMLNGQ
KLQGPQTSAEVYRILKQKGLLDKFPLFTAVYQICYESRPVQEMLSCLQSHPEHT

^{30}KLQKFASTV^{38}

B

Gene Name	Sample Name	AA Mutation	Primary Tissue
GPD1L	TCGA-DZ-6132-01	K30T	Kidney
GPD1L	TCGA-17-Z053-01	L31F	Lung
GPD1L	U343	K33N	Central nervous system

C

Gene Name	Sample Name	AA Mutation	Primary Tissue
GPD1L	TCGA-AA-3667-01	K29N	Large intestine
GPD1L	TCGA-KN-8431-01	K39_W41del	Kidney
GPD1L	TCGA-KN-8431-01	K39_V42delinsI	Kidney
GPD1L	T593	M40V	Large intestine
GPD1L	EGC3	W41R	Stomach

D

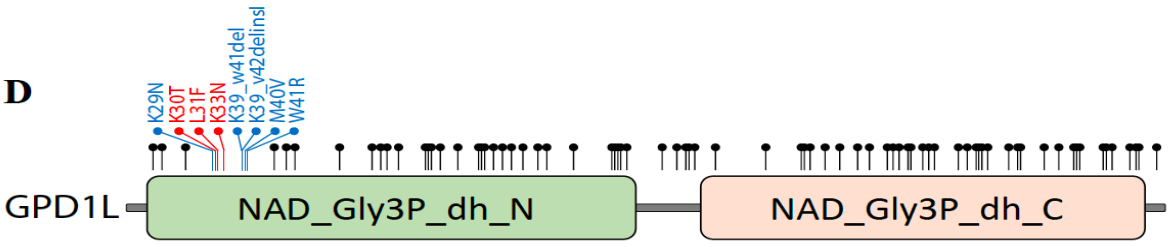


Figure S12. Cancer-associated mutations in GPD1-L from cBIOPORTAL and COSMIC. A. Protein sequence of GPD1-L. B. Mutations annotated in the conserved sequence KLQKFASTV (highlighted in red). C. Mutations annotated in positions adjacent to the conserved sequence KLQKFASTV (highlighted in blue). Gray background indicates two mutations found in the same sample. D. Distribution of all cancer-associated GPD1-L mutations. Mutations in the conserved sequence KLQKFASTV are highlighted in red and adjacent mutations are highlighted in blue.

Figure S13

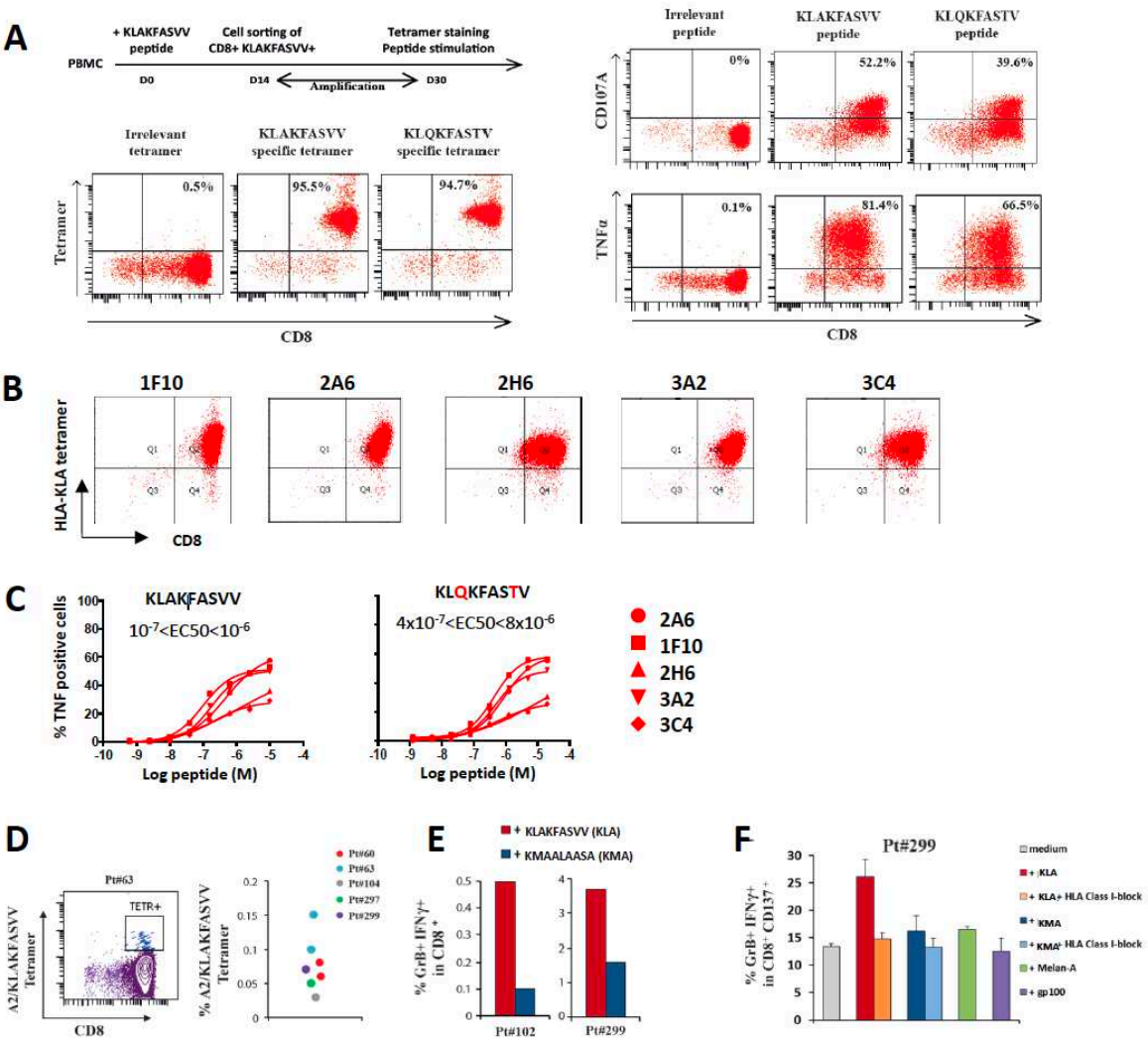

A
+ KLAKFASVV peptide
Cell sorting of CD8+ KLAKFASVV+
Tetramer staining Peptide stimulation
PBMC
D0
D14
Amplification
D30
Irrelevant tetramer
KLAKFASVV specific tetramer
KLQKFASTV specific tetramer
0.5%
95.5%
94.7%
Tetramer
CD8
Irrelevant peptide
KLAKFASVV peptide
KLQKFASTV peptide
CD107A
0%
52.2%
39.6%
TNFα
0.1%
81.4%
66.5%
CD8
B
1F10
2A6
2H6
3A2
3C4
HLA-KLA tetramer
CD8
C
KLAKFASVV
10-7<EC50<10-6
KLQKFASTV
4x10-7<EC50<8x10-6
% TNF positive cells
Log peptide (M)
2A6
1F10
2H6
3A2
3C4
D
Pt#63
TETR+
A2/KLAKFASVV Tetramer
CD8
% A2/KLAKFASVV Tetramer
Pt#60
Pt#63
Pt#104
Pt#297
Pt#299
E
+ KLAKFASVV (KLA)
+ KMAALAASA (KMA)
% GrB+ IFNγ+ in CD8+
Pt#102
Pt#299
F
Pt#299
% GrB+ IFNγ+ in CD8+ CD137+
medium
+ KLA
+ KLA+ HLA Class I-block
+ KMA
+ KMA+ HLA Class I-block
+ Melan-A
+ gp100

Figure S13. Crossreactive T cells recognizing HLA-A*0201-restricted peptides from TMP and GPD1-L proteins.

A-C. In vitro stimulation of T cells from normal volunteers. A. Experimental setting (upper left panel). HLA-A*0201^{+} PBMC extracted from a healthy volunteer were stimulated in microculture assays for 14 days, *ex vivo*, with KLAKFASVV peptide. Each well was screened for costaining with both tetramers (HLA-A*0201/KLAKFASVV from TMP and HLA-A*0201 /KLQKFASTV from GPD1-L). One well out of 192 tested was positive and subjected to cell sorting using HLA- A*0201/KLAKFASVV multimer-coated beads and re-expanded for 15 days. The resulting T cell line was further characterized by flow cytometric analyses for binding to each and both tetramers (A, lower panels) and for its functional cross-reactivity with both epitopes (KLAKFASVV from TMP and KLQKFASTV from GPD1-L), as measured by degranulation (CD107A surface expression) and TNFα release assays. One representative dot plot is depicted for each experimental condition of stimulation (A, upper right panels). B. After cloning of the T cell line by limiting dilution assays, we studied the five KLAKFASVV-specific CD8^{+} CTL clones (1F10, 2A6, 2H6, 3A2 and 3C4) for their capacity to bind the HLA-A*0201/KLAKFASVV tetramer (B) and to secrete TNFα after exposure to increasing concentrations of the two peptides. Three clones exhibited lower or similar affinity for KLQKFASTV (KLQ) compared with KLAKFASVV (KLA) peptides (C). D-F. *In vitro* stimulation of T cells from NSCLC patients. Representative flow cytometry dot plot analyses of PBMC from one of the five HLA-A*0201 NSCLC patients tested, after short term *ex vivo* restimulation with KLAKFASVV peptide followed by tetramer staining using HLA-A*0201/KLAKFASVV tetramers (D, left). Results from 5 different NSCLC patients (D, right panel), each dot representing the percentages of tetramer^{+}CD8^{+} T cells for each patient evaluated. Dots with a similar color represent values from two independent experiments performed using the same patient's PBMC. E-F. Percentages of co- production of GrB and IFN-γ, as measured by multicolor intracellular staining, in CD8^{+} T-cell lines obtained from three HLA-A*0201 NSCLC patients (out of 6 tested) after priming with TMP epitope 10 (KLAKFASVV) and final exposure to the same TMP epitope p10 or irrelevant TMP epitope p14 (KMAALAASA) (E) or irrelevant MART-1/MelanA or pg100 peptides with or without

neutralizing antibodies blocking MHC class I molecules (W6/32) (F). The percentages of effector cells in $CD8^+$ T cells (Pt#102, Pt#299, panel E) or in $CD8^+$ $CD137^+$ T cells (means± SEM of Pt#297 and Pt#299.

Figure S14

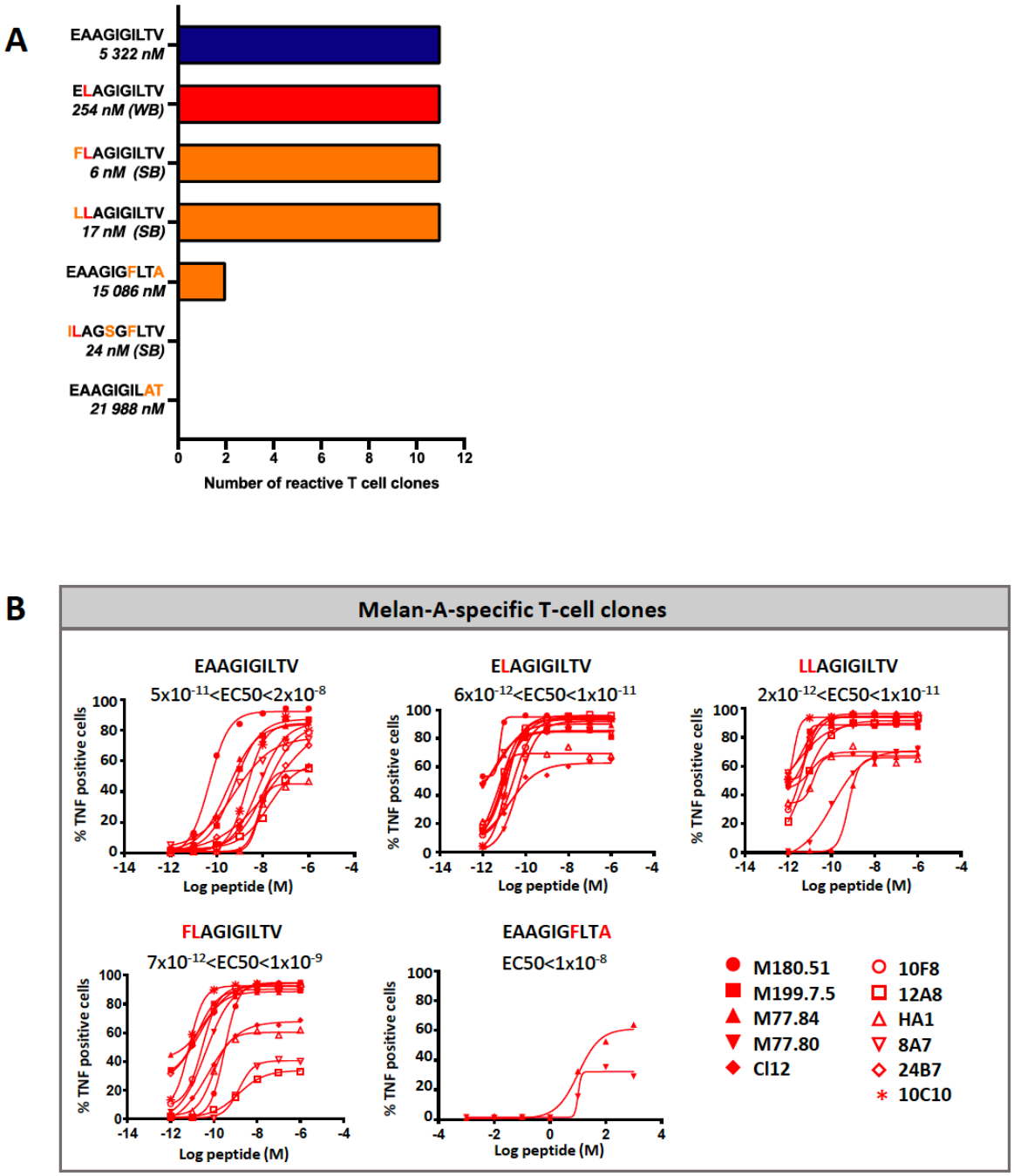


Figure S14. Crossreactivity of T cell clones specific for the MART-1 melanoma peptide with microbial antigens.

A. Numbers of Melan-A specific $CD8^+$T cell clones reactive against bacterial peptides. The blue histogram represents the number of clones reactive against their cognate/naturally

processed epitope (11/11 T cell clones). The red histogram represents the number of Melan-A-specific T cell clones reactive against the high affinity analog peptide Melan-A A27L. Orange histograms represent the number of crossreactive T-cell clones against each bacterial peptide (as selected by an *in silico* approach, Table S8), the differences in the decapeptide amino-acid sequence being highlighted with orange letters. HLA-A*0201-predicted binding affinities (NetMHCprediction) are indicated for each peptide (strong (SB) *versus* weak (WB) binding affinity). B. Functional avidities of MART-1/Melan-A specific $CD8^+$T cell clones in response to naturally processed versus synthetic versus bacterial analogs. Red curves represent T cell clones crossreactive against at least one bacterial peptide. Functional avidities were evaluated by measuring TNFα production in response to T2 cells loaded with a dose range of each indicated peptide, at an E:T ratio of 1:2, by intracellular staining in flow cytometry. Ranges of EC_{50} for each peptide were calculated using PRISM software.

Figure S15

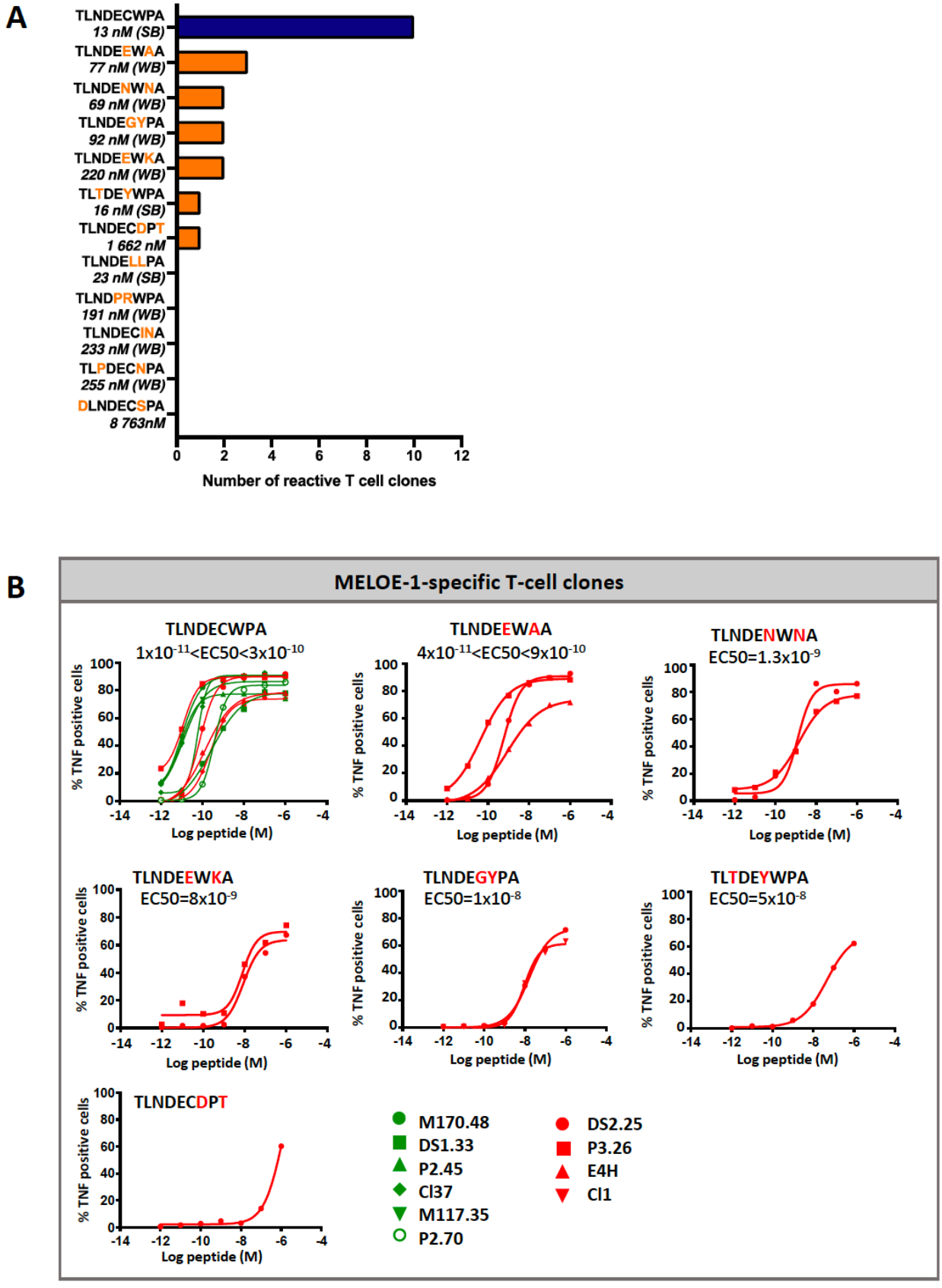

A
TLNDECWPA
13 nM (SB)
TLNDEEWAA
77 nM (WB)
TLNDENWNA
69 nM (WB)
TLNDEGYPA
92 nM (WB)
TLNDEEWKA
220 nM (WB)
TLTDEYWPA
16 nM (SB)
TLNDECDPT
1 662 nM
TLNDELLPA
23 nM (SB)
TLNDPRWPA
191 nM (WB)
TLNDECINA
233 nM (WB)
TLPDECNPA
255 nM (WB)
DLNDECSPA
8 763nM
0 2 4 6 8 10 12
Number of reactive T cell clones
B
MELOE-1-specific T-cell clones
TLNDECWPA
1x10-11<EC50<3x10-10
TLNDEEWAA
4x10-11<EC50<9x10-10
TLNDENWNA
EC50=1.3x10-9
TLNDEEWKA
EC50=8x10-9
TLNDEGYPA
EC50=1x10-8
TLTDEYWPA
EC50=5x10-8
TLNDECDPT
% TNF positive cells
Log peptide (M)
M170.48
DS1.33
P2.45
Cl37
M117.35
P2.70
DS2.25
P3.26
E4H
Cl1

Figure S15. Cross-reactivity of T cell clones specific for the MELOE-1 melanoma peptide with microbial antigens.

A. Numbers of MELOE-1 specific $CD8^+$T cell clones reactive against bacterial peptides. The blue histogram represents the number of clones reactive against their cognate/naturally processed epitope (10 T cell clones). Orange histograms represent the number of crossreactive T-cell clones against each bacterial peptide (designed upon *in silico* selection, Table S8), the differences in the decapeptide amino-acid sequence being highlighted with orange letters. HLA-A*0201-predicted binding affinities (NetMHCprediction) are indicated for each peptide (strong (SB) versus weak (WB) binding affinity). B. Functional avidities of MELOE-1 specific $CD8^+$T cell clones in response to naturally processed versus bacterial analogs. Red curves represent T cell clones cross-reactive against at least one bacterial peptide and green curves represent MELOE-1-specific T-cell clones reactive only to the cognate peptide TLNDECWPA. Functional avidities were evaluated by measuring TNFα production in response to T2 cells loaded with a dose range of each indicated peptide, at an E:T ratio of 1:2, by flow cytometry. Ranges of EC_{50} for each peptide were calculated using PRISM software.

Supplemental Tables:

Table S1. Description of *E. hirae* strains.

Table S2. H-2K^b restricted-*E. hirae* epitopes (<50nM).

We performed sequence alignments of bacterial genes encoding putative cell wall and secreted proteins for immunogenic (13144) *versus* non-immunogenic (708 and 13344) *E. hirae* strains (using the PSORT software), followed by a selection of high affinity epitopes for the MHC class I H-2K^b protein (<50 nM binding affinity) using the NetMHC software.

Table S3. Seeking prophage sequences in *E. hirae* 13144 genomes.

Table S4. List of TRA sequences shared between TMP1 and PSMB4-specific TCRs.

Table S5. List of TRB sequences shared between TMP1 and PSMB4-specific TCRs.

Table S6. List of TMP epitopes selected *in silico* to bind with high affinity (<50nM) HLA-A*0201 molecules.

Table S7. Description of cancer patients treated with anti-PD1 Abs in three independent cohorts (corresponding to Figure S11E-F).

Table S8. Sequence of peptides tested in MART-1 and MELOE-1-specific T cell clones.

Table S9. TCR sequence of MART-1-specific T cell clones. Recurrent motifs already described in the CDR3β of MART-1-specific T-cell clones (*17*) are indicated in bold.

Table S10. TCR sequence of MELOE-1-specific T cell clones. Recurrent motifs already described in the CDR3α of MELOE-1 specific –T-cell clones (*17*) are indicated in bold.

Table S1. Description of *E.hirae* strains

Species	***Origin***	***Cancer***	***Patient outcome***
Enterococcus hirae 13144	*Murine – CTX-treated*		
Enterococcus hirae 708	*Human - Unknown*		
Enterococcus hirae 13344	*Human - Blood*		
Enterococcus hirae ATCC9790	*Type strain CIP 53.48T*		
Enterococcus hirae 5348	*Human - Unknown*		
Enterococcus hirae 7030	*Human – Liver abscess*		
Enterococcus hirae 12607	*Environmental – RiskManche project*		
Enterococcus hirae 13150	*Environmental - Water*		
Enterococcus hirae 13152	*Environmental - Water*		
Enterococcus hirae 13153	*Environmental - Water*		
Enterococcus hirae 13155	*Environmental – RiskManche project*		
Enterococcus hirae 13161	*Environmental - Cockle*		
Enterococcus hirae 13343	*Conservation liquid of kidney*		
Enterococcus hirae 13346	*Human - Urine*		
Enterococcus hirae 13347	*Blood culture*		
Enterococcus hirae IGR1	*Human (stool)*	*Lung*	*Responder*
Enterococcus hirae IGR4	*Human (stool)*	*Lung*	*Complete Responder*
Enterococcus hirae IGR10	*Human (stool)*	*Lung*	*Responder*
Enterococcus hirae IGR11	*Human (stool)*	*Lung*	*Responder*

Table S2. H-2K^b restricted-*E.hirae* epitopes (<50nM binding affinity)

Group	Hirae	sequence	names of the proteins
Group 1	708	INAKFSSQL	Membrane proteins related to metalloendopep□dases
	708	YIYNHYKDM	Membrane proteins related to metalloendopep□dases
	708	YVYGKSRTM	Membrane proteins related to metalloendopep□dases
	708	IAFLSYKLF	cell surface protein precursor
Group 2	708	IMYEYMYPV	hypothe□cal protein
	708	SSMEYFLKV	Phage tail length tape-measure protein
	708	ISFFQENQL	Collagen adhesin
	708	TNLLFMTSL	extracellular protein
Group 3	708	KIFSIFMLL	Phospha□dylinositol-specific phospholipase C
	708	LNIFKFNRF	Chi□nase
	708	MTYDYRGGF	Chi□nase
	708	PSYMFRTSF	Chi□nase
Group 4	708	QSYTYYMTA	cell wall surface anchor family protein
	708	ITFSHYEPT	cell wall surface anchor family protein
	13144	SAFPYEQEL	C3 family ADP-ribosyltransferase
	13144	YNYSKSYPV	hypothe□cal protein
Group 5	13144	VSFSHYRPG	hypothe□cal protein
	13144	VTFLGYNAF	cell surface protein
	13144	TVYTFHVNI	cell surface protein
	13144	TSYSPLFLL	cell surface protein (puta□ve)
Group 6	13144	TNYIYPNIL	2',3'-cyclic-nucleo□de 2'-phosphodiesterase
	13144	VVPILFLGL	FmtB protein
	13144	KNYKAYVEL	hypothe□cal protein
	13144	SAMKYGIPL	hypothe□cal protein
Group 7	13144	TSLARFANI	Phage tail length tape-measure protein → TMP1
	13144	AMIEFIQGL	Phage tail length tape-measure protein → TMP2
	13144	VAITFGGPL	Phage tail length tape-measure protein → TMP3
	13144	VSTNHYGLL	hypothe□cal protein
Group 8	13144	VMFGLFITI	cell surface protein precursor
	13144	TVFSLVSLL	Chi□nase
	13144	SIYNLEKPL	IgA1 protease
	13144	YTIIRYGNL	IgA1 protease
Group 9	13144	SNGLLYTPM	IgA1 protease
	13144	NNYHYVGGL	IgA1 protease
	13144	SMFLNCNNL	hypothe□cal protein
	13144	IAFQGYSSL	hypothe□cal protein
Group 10	13144	QVTNFFNMF	hypothe□cal protein
	13144	IMLGLFMTM	cell surface protein precursor
	EH17	MSFTFFSST	hypothe□cal protein
	EH17	IAFQNFVNL	Chi□nase
Group 11	EH17	SMFIAFQNF	Chi□nase
	EH17	LNYDYGNRI	Chi□nase
	EH17	AGICFFTGV	Pep□doglycan N-acetylglucosamine deacetylase
	EH17	VEYTYFPTL	Membrane proteins related to metalloendopep□dases

Group 12	***EH17***	*AAYVFEMNF*	*Membrane proteins related to metalloendopep□dases*
	EH17	*EMYRKLSTL*	*Membrane proteins related to metalloendopep□dases*
	EH17	*YNYGYKSVL*	*enhancin family protein*
	EH17	*VIHELYNSL*	*bacteriocin immunity protein*

Table S3. Seeking prophage sequence in *E.hirae* 13144 genome

Region	*Region Length*	*Completeness*	*Score*	*# Total Proteins*	*Region Position*	*Most Common Phage*	*GC %*
1	*40.6Kb*	*intact*	*150*	*58*	*481066-521729*	*PHAGE_Entero_phiEf11_NC_013696(9)*	*33.79%*
2	*39.2Kb*	*intact*	*140*	*59*	*2123983-2163272*	*PHAGE_Entero_vB_IME197_NC_028671(6)*	*34.95%*

Table S4. List of TRA sequences shared between TMP1 and PSMB4-specific TCRs

V	J	aaSeqCDR3	CDR3dna	VpJ	VJ	cloneCount
TRAV14D-2	TRAJ22	CAASASSGSWQLIF	TGTGCAGCAAGCGCATCTTCTGGCAGCTGGCAACTCATCTTT	TRAV14D-2 CAASASSGSWQLIF TRAJ22	TRAV14D-2 TRAJ22	57
TRAV14-1	TRAJ26	CAASDNYAQGLTF	TGTGCAGCAAGTGATAACTATGCCCAGGGATTAACCTTC	TRAV14-1 CAASDNYAQGLTF TRAJ26	TRAV14-1 TRAJ26	405
TRAV16D-DV11	TRAJ17	CAMRDLNSAGNKLTF	TGTGCTATGAGAGACCTTAACAGTGCAGGGAACAAGCTAACTTTT	TRAV16D-DV11 CAMRDLNSAGNKLTF TRAJ17	TRAV16D-DV11 TRAJ17	19
TRAV16D-DV11	TRAJ27	CAMREDTNTGKLTF	TGTGCTATGAGAGAGGACACCAATACAGGCAAATTAACCTTT	TRAV16D-DV11 CAMREDTNTGKLTF TRAJ27	TRAV16D-DV11 TRAJ27	123
TRAV3D-3	TRAJ30	CAVSDTNAYKVIF	TGCGCAGTCAGTGACACAAATGCTTACAAAGTCATCTTT	TRAV3D-3 CAVSDTNAYKVIF TRAJ30	TRAV3D-3 TRAJ30	297
TRAV7-3	TRAJ9	CAVSNMGYKLTF	TGTGCAGTGAGCAACATGGGCTACAAACTTACCTTC	TRAV7-3 CAVSNMGYKLTF TRAJ9	TRAV7-3 TRAJ9	208
TRAV8-1	TRAJ18	CATGDRGSALGRLHF	TGTGCTACTGGAGATAGAGGTTCAGCCTTAGGGAGGCTGCATTTT	TRAV8-1 CATGDRGSALGRLHF TRAJ18	TRAV8-1 TRAJ18	19
TRAV6D-7	TRAJ31	CALGGNSNNRIFF	TGTGCTCTGGGGGGGAATAGCAATAACAGAATCTTCTTT	TRAV6D-7 CALGGNSNNRIFF TRAJ31	TRAV6D-7 TRAJ31	17
TRAV14D-2	TRAJ22	CAASASSGSWQLIF	TGTGCAGCCTCTGCATCTTCTGGCAGCTGGCAACTCATCTTT	TRAV14D-2 CAASASSGSWQLIF TRAJ22	TRAV14D-2 TRAJ22	14
TRAV16N	TRAJ52	CAMRENTGANTGKLTF	TGTGCTATGAGAGAGAACACTGGAGCTAACACTGGAAAGCTCACGTTT	TRAV16N CAMRENTGANTGKLTF TRAJ52	TRAV16N TRAJ52	14
TRAV3D-3	TRAJ21	CAVRDLSNYNVLYF	TGCGCAGTCAGGGATTTGTCTAATTACAACGTGCTTTACTTC	TRAV3D-3 CAVRDLSNYNVLYF TRAJ21	TRAV3D-3 TRAJ21	10
TRAV14D-1	TRAJ26	CAARNNYAQGLTF	TGTGCAGCAAGAAATAACTATGCCCAGGGATTAACCTTC	TRAV14D-1 CAARNNYAQGLTF TRAJ26	TRAV14D-1 TRAJ26	156
TRAV12D-3	TRAJ33	CALSNYQLIW	TGTGCTCTGAGCAACTATCAGTTGATCTGG	TRAV12D-3 CALSNYQLIW TRAJ33	TRAV12D-3 TRAJ33	8
TRAV12D-3	TRAJ31	CALSDRDSNNRIFF	TGTGCTCTGAGTGATCGAGATAGCAATAACAGAATCTTCTTT	TRAV12D-3 CALSDRDSNNRIFF TRAJ31	TRAV12D-3 TRAJ31	5
TRAV6D-7	TRAJ31	CALGGNSNNRIFF	TGTGCTCTGGGTGGGAATAGCAATAACAGAATCTTCTTT	TRAV6D-7 CALGGNSNNRIFF TRAJ31	TRAV6D-7 TRAJ31	4
TRAV12D-3	TRAJ31	CALSDRDSNNRIFF	TGTGCTCTGAGTGATCGGGATAGCAATAACAGAATCTTCTTT	TRAV12D-3 CALSDRDSNNRIFF TRAJ31	TRAV12D-3 TRAJ31	255
TRAV14D-3-DV8	TRAJ22	CAASASSGSWQLIF	TGTGCAGCAAGTGCAAGTTCTGGCAGCTGGCAACTCATCTTT	TRAV14D-3-DV8 CAASASSGSWQLIF TRAJ22	TRAV14D-3-DV8 TRAJ22	2
TRAV12D-3	TRAJ31	CALSDRHSNNRIFF	TGTGCTCTGAGTGATCGACATAGCAATAACAGAATCTTCTTT	TRAV12D-3 CALSDRHSNNRIFF TRAJ31	TRAV12D-3 TRAJ31	1
TRAV8-1	TRAJ50	CATDPLASSSFSKLVF	TGTGCTACTGACCCCCTAGCATCCTCCTCCTTCAGCAAGCTGGTGTTT	TRAV8-1 CATDPLASSSFSKLVF TRAJ50	TRAV8-1 TRAJ50	49
TRAV10	TRAJ27	CAASRGTNTGKLTF	TGTGCAGCAAGCAGAGGCACCAATACAGGCAAATTAACCTTT	TRAV10 CAASRGTNTGKLTF TRAJ27	TRAV10 TRAJ27	446
TRAV7-2	TRAJ12	CAAPGTGGYKVVF	TGTGCAGCCCCCGGGACTGGAGGCTATAAAGTGGTCTTT	TRAV7-2 CAAPGTGGYKVVF TRAJ12	TRAV7-2 TRAJ12	360
TRAV6N-6	TRAJ22	CALRAASSGSWQLIF	TGCGCTCTGAGGGCAGCATCTTCTGGCAGCTGGCAACTCATCTTT	TRAV6N-6 CALRAASSGSWQLIF TRAJ22	TRAV6N-6 TRAJ22	332
TRAV13-2	TRAJ26	CAIDQYAQGLTF	TGTGCTATAGACCAATATGCCCAGGGATTAACCTTC	TRAV13-2 CAIDQYAQGLTF TRAJ26	TRAV13-2 TRAJ26	296
TRAV14-2	TRAJ44	CAGTGSGGKLTL	TGTGCAGGGACTGGCAGTGGTGGAAAACTCACTTTG	TRAV14-2 CAGTGSGGKLTL TRAJ44	TRAV14-2 TRAJ44	276
TRAV12D-3	TRAJ23	CALSGENYNQGKLIF	TGTGCTCTGAGTGGGGAGAATTATAACCAGGGGAAGCTTATCTTT	TRAV12D-3 CALSGENYNQGKLIF TRAJ23	TRAV12D-3 TRAJ23	256
TRAV19	TRAJ50	CAAGGVASSSFSKLVF	TGCGCAGCAGGGGGGGTAGCATCCTCCTCCTTCAGCAAGCTGGTGTTT	TRAV19 CAAGGVASSSFSKLVF TRAJ50	TRAV19 TRAJ50	236
TRAV13-1	TRAJ6	CALVLTSGGNYKPTF	TGTGCTTTGGTCCTAACCTCAGGAGGAAACTACAAACCTACGTTT	TRAV13-1 CALVLTSGGNYKPTF TRAJ6	TRAV13-1 TRAJ6	228
TRAV6N-6	TRAJ22	CALSVASSGSWQLIF	TGCGCTCTGAGTGTCGCATCTTCTGGCAGCTGGCAACTCATCTTT	TRAV6N-6 CALSVASSGSWQLIF TRAJ22	TRAV6N-6 TRAJ22	230
TRAV12-2	TRAJ58	CALSDPGTGSKLSF	TGTGCTTTGAGTGATCCAGGCACTGGGTCTAAGCTGTCATTT	TRAV12-2 CALSDPGTGSKLSF TRAJ58	TRAV12-2 TRAJ58	182
TRAV5D-4	TRAJ22	CAASTSSGSWQLIF	TGTGCTGCAAGTACATCTTCTGGCAGCTGGCAACTCATCTTT	TRAV5D-4 CAASTSSGSWQLIF TRAJ22	TRAV5D-4 TRAJ22	128
TRAV14D-2	TRAJ23	CAASEDYNQGKLIF	TGTGCAGCAAGTGAGGATTATAACCAGGGGAAGCTTATCTTT	TRAV14D-2 CAASEDYNQGKLIF TRAJ23	TRAV14D-2 TRAJ23	100
TRAV13N-1	TRAJ27	CAMEPGTNTGKLTF	TGTGCTATGGAACCGGGCACCAATACAGGCAAATTAACCTTT	TRAV13N-1 CAMEPGTNTGKLTF TRAJ27	TRAV13N-1 TRAJ27	80
TRAV12D-3	TRAJ31	CALSDRHSNNRIFF	TGTGCTCTGAGTGATCGACACAGCAATAACAGAATCTTCTTT	TRAV12D-3 CALSDRHSNNRIFF TRAJ31	TRAV12D-3 TRAJ31	58
TRAV13N-4	TRAJ28	CVLSLLPGTGSNRLTF	TGTGTTCTGAGTCTGCTACCAGGCACTGGGAGTAACAGGCTCACTTTT	TRAV13N-4 CVLSLLPGTGSNRLTF TRAJ28	TRAV13N-4 TRAJ28	2
TRAV8-1	TRAJ50	CATDPLASSSFSKLVF	TGTGCTACTGACCCCCTAGCATCCTCCTCCTTCAGCAAGCTGGTGTTT	TRAV8-1 CATDPLASSSFSKLVF TRAJ50	TRAV8-1 TRAJ50	2
TRAV12-2	TRAJ43	CVRNNNNAPRF	TGTGTTCGCAATAACAACAATGCCCCACGATTT	TRAV12-2 CVRNNNNAPRF TRAJ43	TRAV12-2 TRAJ43	346
TRAV16N	TRAJ40	CAMRENTGNYKYVF	TGTGCTATGAGAGAGAATACAGGAAACTACAAATACGTCTTT	TRAV16N CAMRENTGNYKYVF TRAJ40	TRAV16N TRAJ40	234
TRAV8D-2	TRAJ9	CATDVGYKLTF	TGTGCTACAGATGTGGGCTACAAACTTACCTTC	TRAV8D-2 CATDVGYKLTF TRAJ9	TRAV8D-2 TRAJ9	108
TRAV12-2	TRAJ43	CVRNNNNAPRF	TGTGTTCGCAATAACAACAATGCCCCAAGATTT	TRAV12-2 CVRNNNNAPRF TRAJ43	TRAV12-2 TRAJ43	30
TRAV16D-DV11	TRAJ17	CAMRDLNSAGNKLTF	TGTGCTATGAGAGACCTTAACAGTGCAGGGAACAAGCTAACTTTT	TRAV16D-DV11 CAMRDLNSAGNKLTF TRAJ17	TRAV16D-DV11 TRAJ17	2
TRAV3D-3	TRAJ21	CAVRDLSNYNVLYF	TGCGCAGTCAGGGATTTGTCTAATTACAACGTGCTTTACTTC	TRAV3D-3 CAVRDLSNYNVLYF TRAJ21	TRAV3D-3 TRAJ21	2
TRAV4D-3	TRAJ15	CAADQGGRALIF	TGTGCTGCTGACCAGGGAGGCAGAGCTCTGATATTT	TRAV4D-3 CAADQGGRALIF TRAJ15	TRAV4D-3 TRAJ15	2
TRAV14D-2	TRAJ22	CAASASSGSWQLIF	TGTGCAGCAAGTGCCTCTTCTGGCAGCTGGCAACTCATCTTT	TRAV14D-2 CAASASSGSWQLIF TRAJ22	TRAV14D-2 TRAJ22	180
TRAV14D-3-DV8	TRAJ22	CAASASSGSWQLIF	TGTGCAGCAAGTGCATCTTCTGGCAGCTGGCAACTCATCTTT	TRAV14D-3-DV8 CAASASSGSWQLIF TRAJ22	TRAV14D-3-DV8 TRAJ22	180
TRAV12D-3	TRAJ33	CALSNYQLIW	TGTGCTCTCAGCAACTATCAGTTGATCTGG	TRAV12D-3 CALSNYQLIW TRAJ33	TRAV12D-3 TRAJ33	135
TRAV6D-7	TRAJ31	CALGGNSNNRIFF	TGTGCTCTGGGTGGAAATAGCAATAACAGAATCTTCTTT	TRAV6D-7 CALGGNSNNRIFF TRAJ31	TRAV6D-7 TRAJ31	37

Table 33: List of TRB sequences shared between TM-1 and PSMB4-specific TCRs

V	J	aaSeqCDR3	CDR3dna	VpJ	VJ	count
TRBV13-3	TRBJ2-3	CARDGAETLYF	TGTGCCAGGGACGGTGCAGAAACGCTGTATTTT	TRBV13-3 CARDGAETLYF TRBJ2-3	TRBV13-3 TRBJ2-3	318
TRBV12-2	TRBJ2-5	CASAFNQDTQYF	TGTGCCAGCGCTTTTAACCAAGACACCCAGTACTTT	TRBV12-2 CASAFNQDTQYF TRBJ2-5	TRBV12-2 TRBJ2-5	696
TRBV13-2	TRBJ2-7	CASGDFYEQYF	TGTGCCAGCGGGGACTTTTATGAACAGTACTTC	TRBV13-2 CASGDFYEQYF TRBJ2-7	TRBV13-2 TRBJ2-7	16
TRBV13-2	TRBJ2-7	CASGDFYEQYF	TGTGCCAGCGGTGATTTCTATGAACAGTACTTC	TRBV13-2 CASGDFYEQYF TRBJ2-7	TRBV13-2 TRBJ2-7	138
TRBV13-2	TRBJ1-2	CASGDNANSDYPF	TGTGCCAGCGGGGACAATGCAAACTCCGACTACCCCTTC	TRBV13-2 CASGDNANSDYPF TRBJ1-2	TRBV13-2 TRBJ1-2	42
TRBV13-2	TRBJ1-2	CASGDNANSDYTF	TGTGCCAGCGGGGACAATGCAAACTCCGACTACACCTTC	TRBV13-2 CASGDNANSDYTF TRBJ1-2	TRBV13-2 TRBJ1-2	372
TRBV13-2	TRBJ2-4	CASGDRGSQNTLYF	TGTGCCAGCGGTGACAGGGGTAGTCAAAACACCTTGTACTTT	TRBV13-2 CASGDRGSQNTLYF TRBJ2-4	TRBV13-2 TRBJ2-4	1
TRBV13-2	TRBJ2-4	CASGDRGSQNTLYF	TGTGCCAGCGGTGATCGGGGTAGTCAAAACACCTTGTACTTT	TRBV13-2 CASGDRGSQNTLYF TRBJ2-4	TRBV13-2 TRBJ2-4	166
TRBV13-2	TRBJ1-4	CASGDSNERLFF	TGTGCCAGCGGGGATTCCAACGAAAGATTATTTTTC	TRBV13-2 CASGDSNERLFF TRBJ1-4	TRBV13-2 TRBJ1-4	6
TRBV13-2	TRBJ1-4	CASGDSNERLFF	TGTGCCAGCGGTGATTCCAACGAAAGATTATTTTTC	TRBV13-2 CASGDSNERLFF TRBJ1-4	TRBV13-2 TRBJ1-4	1
TRBV13-2	TRBJ1-4	CASGDSNERLFF	TGTGCCAGCGGTGACAGCAACGAAAGATTATTTTTC	TRBV13-2 CASGDSNERLFF TRBJ1-4	TRBV13-2 TRBJ1-4	45
TRBV13-2	TRBJ2-5	CASGGDRGQDTQYF	TGTGCCAGCGGTGGGGACAGGGGGCAAGACACCCAGTACTTT	TRBV13-2 CASGGDRGQDTQYF TRBJ2-5	TRBV13-2 TRBJ2-5	14
TRBV13-2	TRBJ2-1	CASGGTAPIYAEQFF	TGTGCCAGCGGTGGGACAGCTCCTATCTATGCTGAGCAGTTCTTC	TRBV13-2 CASGGTAPIYAEQFF TRBJ2-1	TRBV13-2 TRBJ2-1	18
TRBV17	TRBJ2-5	CASGTGTQDTQYF	TGTGCTAGCGGGACTGGGACCCAAGACACCCAGTACTTT	TRBV17 CASGTGTQDTQYF TRBJ2-5	TRBV17 TRBJ2-5	300
TRBV19	TRBJ1-3	CASRNRGSGNTLYF	TGTGCCAGCAGAAACAGGGGTTCTGGAAATACGCTCTATTTT	TRBV19 CASRNRGSGNTLYF TRBJ1-3	TRBV19 TRBJ1-3	584
TRBV13-1	TRBJ2-7	CASSDAAGQYF	TGTGCCAGCAGTGATGCGGCTGGGCAGTACTTC	TRBV13-1 CASSDAAGQYF TRBJ2-7	TRBV13-1 TRBJ2-7	268
TRBV13-3	TRBJ1-2	CASSDARGQDSDYTF	TGTGCCAGCAGTGATGCCAGGGGGCAGGACTCCGACTACACCTTC	TRBV13-3 CASSDARGQDSDYTF TRBJ1-2	TRBV13-3 TRBJ1-2	618
TRBV13-3	TRBJ2-1	CASSDGAEQFF	TGTGCCAGCAGTGATGGTGCTGAGCAGTTCTTC	TRBV13-3 CASSDGAEQFF TRBJ2-1	TRBV13-3 TRBJ2-1	5
TRBV13-3	TRBJ2-1	CASSDGAEQFF	TGTGCCAGCAGTGATGGAGCTGAGCAGTTCTTC	TRBV13-3 CASSDGAEQFF TRBJ2-1	TRBV13-3 TRBJ2-1	254
TRBV13-1	TRBJ1-4	CASSDGGSNERLFF	TGTGCCAGCAGTGATGGGGGGTCCAACGAAAGATTATTTTTC	TRBV13-1 CASSDGGSNERLFF TRBJ1-4	TRBV13-1 TRBJ1-4	1
TRBV13-1	TRBJ1-4	CASSDGGSNERLFF	TGTGCCAGCAGTGATGGGGGGTCCAACGAAAGATTATTTTTC	TRBV13-1 CASSDGGSNERLFF TRBJ1-4	TRBV13-1 TRBJ1-4	228
TRBV13-3	TRBJ1-2	CASSDHANSDYTF	TGTGCCAGCAGTGACCATGCAAACTCCGACTACACCTTC	TRBV13-3 CASSDHANSDYTF TRBJ1-2	TRBV13-3 TRBJ1-2	37
TRBV13-1	TRBJ2-7	CASSDRDWVYEQYF	TGTGCCAGCAGTGACCGGGACTGGGTCTATGAACAGTACTTC	TRBV13-1 CASSDRDWVYEQYF TRBJ2-7	TRBV13-1 TRBJ2-7	186
TRBV13-1	TRBJ2-4	CASSDRTGGFSQNTLYF	TGTGCCAGCAGTGATCGTACAGGGGGCTTCAGTCAAAACACCTTGTACTTT	TRBV13-1 CASSDRTGGFSQNTLYF TRBJ2-4	TRBV13-1 TRBJ2-4	1474
TRBV13-1	TRBJ2-1	CASSDWGNYAEQFF	TGTGCCAGCAGCGACTGGGGGAACTATGCTGAGCAGTTCTTC	TRBV13-1 CASSDWGNYAEQFF TRBJ2-1	TRBV13-1 TRBJ2-1	8
TRBV13-1	TRBJ2-1	CASSDWGNYAEQFF	TGTGCCAGCAGTGACTGGGGGAACTATGCTGAGCAGTTCTTC	TRBV13-1 CASSDWGNYAEQFF TRBJ2-1	TRBV13-1 TRBJ2-1	128
TRBV13-1	TRBJ2-5	CASSELWGGQDTQYF	TGTGCCAGCAGTGAACTCTGGGGGGGCCAAGACACCCAGTACTTT	TRBV13-1 CASSELWGGQDTQYF TRBJ2-5	TRBV13-1 TRBJ2-5	346
TRBV13-1	TRBJ2-7	CASSEPEYEQYF	TGTGCCAGCAGTGAACCAGAATATGAACAGTACTTC	TRBV13-1 CASSEPEYEQYF TRBJ2-7	TRBV13-1 TRBJ2-7	130
TRBV12-1	TRBJ2-7	CASSFRDISYEQYF	TGTGCCAGCTCTTTCCGGGACATCTCCTATGAACAGTACTTC	TRBV12-1 CASSFRDISYEQYF TRBJ2-7	TRBV12-1 TRBJ2-7	124
TRBV12-1	TRBJ2-7	CASSFRDSSYEQYF	TGTGCCAGCTCTTTCCGGGACAGCTCCTATGAACAGTACTTC	TRBV12-1 CASSFRDSSYEQYF TRBJ2-7	TRBV12-1 TRBJ2-7	38
TRBV14	TRBJ2-7	CASSFRGPYEQYF	TGTGCCAGCAGTTTCAGGGGTCCCTATGAACAGTACTTC	TRBV14 CASSFRGPYEQYF TRBJ2-7	TRBV14 TRBJ2-7	62
TRBV14	TRBJ2-7	CASSFRVPYEQYF	TGTGCCAGCAGTTTCAGGGTTCCCTATGAACAGTACTTC	TRBV14 CASSFRVPYEQYF TRBJ2-7	TRBV14 TRBJ2-7	288
TRBV12-1	TRBJ2-4	CASSGDRDKNTLYC	TGTGCCAGCTCTGGCGACAGGGACAAAAACACCTTGTACTGT	TRBV12-1 CASSGDRDKNTLYC TRBJ2-4	TRBV12-1 TRBJ2-4	2
TRBV12-1	TRBJ2-4	CASSGDRDQNTLYF	TGTGCCAGCTCTGGCGACAGGGACCAAAACACCTTGTACTTT	TRBV12-1 CASSGDRDQNTLYF TRBJ2-4	TRBV12-1 TRBJ2-4	326
TRBV13-3	TRBJ1-2	CASSGTRNSDYPF	TGTGCCAGCAGTGGAACCAGAAACTCCGACTACCCCTTC	TRBV13-3 CASSGTRNSDYPF TRBJ1-2	TRBV13-3 TRBJ1-2	12
TRBV13-3	TRBJ1-2	CASSGTRNSDYTF	TGTGCCAGCAGTGGAACCAGAAACTCCGACTACACCTTC	TRBV13-3 CASSGTRNSDYTF TRBJ1-2	TRBV13-3 TRBJ1-2	288
TRBV12-1	TRBJ2-3	CASSGTTSAETLYF	TGTGCCAGCTCCGGGACAACTAGTGCAGAAACGCTGTATTTT	TRBV12-1 CASSGTTSAETLYF TRBJ2-3	TRBV12-1 TRBJ2-3	2
TRBV19	TRBJ2-3	CASSIGGTSSAETLYF	TGTGCCAGCAGTATAGGGGGGACCTCTAGTGCAGAAACGCTGTATTTT	TRBV19 CASSIGGTSSAETLYF TRBJ2-3	TRBV19 TRBJ2-3	532
TRBV19	TRBJ2-3	CASSIGGTSSAETLYF	TGTGCCAGCAGTATAGGGGGGACCTCTAGTGCAGAAACGCTGTATTTT	TRBV19 CASSIGGTSSAETLYF TRBJ2-3	TRBV19 TRBJ2-3	1
TRBV17	TRBJ2-7	CASSIGTGAYEQYF	TGTGCTAGCAGTATAGGGACAGGGGCCTATGAACAGTACTTC	TRBV17 CASSIGTGAYEQYF TRBJ2-7	TRBV17 TRBJ2-7	228
TRBV12-2	TRBJ2-5	CASSLDKDTQYF	TGTGCCAGCTCTCTCGACAAAGACACCCAGTACTTT	TRBV12-2 CASSLDKDTQYF TRBJ2-5	TRBV12-2 TRBJ2-5	14
TRBV12-2	TRBJ2-5	CASSLDKDTQYF	TGTGCCAGCTCTCTCGACAAAGACACCCAGTACTTT	TRBV12-2 CASSLDKDTQYF TRBJ2-5	TRBV12-2 TRBJ2-5	2
TRBV12-2	TRBJ2-3	CASSLDSSAETLYF	TGTGCCAGCTCTCTCGACTCTAGTGCAGAAACGCTGTATTTT	TRBV12-2 CASSLDSSAETLYF TRBJ2-3	TRBV12-2 TRBJ2-3	386
TRBV16	TRBJ2-4	CASSLERGASQNTLYF	TGTGCAAGCAGCTTAGAAAGGGGAGCTAGTCAAAACACCTTGTACTTT	TRBV16 CASSLERGASQNTLYF TRBJ2-4	TRBV16 TRBJ2-4	454
TRBV16	TRBJ2-4	CASSLETGGARQNTLYF	TGTGCAAGCAGCTTAGAAACTGGGGGGGCGCGTCAAAACACCTTGTACTTT	TRBV16 CASSLETGGARQNTLYF TRBJ2-4	TRBV16 TRBJ2-4	114
TRBV16	TRBJ2-5	CASSLGGQDTQYF	TGTGCAAGCAGCTTGGGGGGCCAAGACACCCAGTACTTT	TRBV16 CASSLGGQDTQYF TRBJ2-5	TRBV16 TRBJ2-5	33
TRBV16	TRBJ2-5	CASSLGGQDTQYF	TGTGCAAGCAGCTTGGGGGGCAAGACACCCAGTACTTT	TRBV16 CASSLGGQDTQYF TRBJ2-5	TRBV16 TRBJ2-5	7
TRBV16	TRBJ2-5	CASSLGGQDTQYF	TGTGCAAGCAGCTTAGGGGGACAAGACACCCAGTACTTT	TRBV16 CASSLGGQDTQYF TRBJ2-5	TRBV16 TRBJ2-5	260
TRBV26	TRBJ1-1	CASSLGQINTEVFF	TGTGCCAGCAGTCTGGGACAAATCAACACAGAAGTCTTCTTT	TRBV26 CASSLGQINTEVFF TRBJ1-1	TRBV26 TRBJ1-1	434
TRBV26	TRBJ1-1	CASSLGQSNTEVFF	TGTGCCAGCAGTCTGGGACAATCAAACACAGAAGTCTTCTTT	TRBV26 CASSLGQSNTEVFF TRBJ1-1	TRBV26 TRBJ1-1	4
TRBV26	TRBJ1-1	CASSLGQSNTEVFF	TGTGCCAGCAGTCTGGGACAAAGCAACACAGAAGTCTTCTTT	TRBV26 CASSLGQSNTEVFF TRBJ1-1	TRBV26 TRBJ1-1	28
TRBV12-2	TRBJ2-5	CASSLMGNQDTQYF	TGTGCCAGCTCTCTCATGGGGAACCAAGACACCCAGTACTTT	TRBV12-2 CASSLMGNQDTQYF TRBJ2-5	TRBV12-2 TRBJ2-5	9
TRBV12-2	TRBJ2-5	CASSLMGNQDTQYF	TGTGCCAGCTCTCTCATGGGGAACCAAGACACCCAGTACTTT	TRBV12-2 CASSLMGNQDTQYF TRBJ2-5	TRBV12-2 TRBJ2-5	2
TRBV29	TRBJ1-4	CASSLSSSNERLFF	TGTGCTAGCAGTTTAAGTAGTTCCAACGAAAGATTATTTTTC	TRBV29 CASSLSSSNERLFF TRBJ1-4	TRBV29 TRBJ1-4	370
TRBV19	TRBJ1-3	CASSMEETSSGNTLYF	TGTGCCAGCAGTATGGAGGAGACATCTTCTGGAAATACGCTCTATTTT	TRBV19 CASSMEETSSGNTLYF TRBJ1-3	TRBV19 TRBJ1-3	2
TRBV19	TRBJ1-1	CASSNRENTEVFF	TGTGCCAGCAGCAACAGGGAAAACACAGAAGTCTTCTTT	TRBV19 CASSNRENTEVFF TRBJ1-1	TRBV19 TRBJ1-1	25
TRBV19	TRBJ1-1	CASSNRENTEVFF	TGTGCCAGCAGCAACAGGGAAAACACAGAAGTCTTCTTT	TRBV19 CASSNRENTEVFF TRBJ1-1	TRBV19 TRBJ1-1	2
TRBV14	TRBJ2-1	CASSPDRGYAEQFF	TGTGCCAGCAGCCCAGACAGGGGGTATGCTGAGCAGTTCTTC	TRBV14 CASSPDRGYAEQFF TRBJ2-1	TRBV14 TRBJ2-1	18
TRBV14	TRBJ2-1	CASSPDRGYAEQFF	TGTGCCAGCAGCCCAGACAGGGGGTATGCTGAGCAGTTCTTC	TRBV14 CASSPDRGYAEQFF TRBJ2-1	TRBV14 TRBJ2-1	2
TRBV21	TRBJ2-4	CASSPGQGASQNTLYF	TGTGCTAGCAGTCCGGGACAGGGGGCCAGTCAAAACACCTTGTACTTT	TRBV21 CASSPGQGASQNTLYF TRBJ2-4	TRBV21 TRBJ2-4	165
TRBV29	TRBJ2-7	CASSPGTGGYEQYF	TGTGCTAGCAGTCCCGGGACAGGGGGCTATGAACAGTACTTC	TRBV29 CASSPGTGGYEQYF TRBJ2-7	TRBV29 TRBJ2-7	566
TRBV29	TRBJ2-5	CASSPGTGNQDTQYF	TGTGCTAGCAGCCCCGGGACAGGGAACCAAGACACCCAGTACTTT	TRBV29 CASSPGTGNQDTQYF TRBJ2-5	TRBV29 TRBJ2-5	47
TRBV29	TRBJ2-5	CASSPGTGNQDTQYF	TGTGCTAGCAGCCCCGGGACAGGGAACCAAGACACCCAGTACTTT	TRBV29 CASSPGTGNQDTQYF TRBJ2-5	TRBV29 TRBJ2-5	2
TRBV29	TRBJ2-3	CASSPGTNSAETLYF	TGTGCTAGCAGCCCCGGGACAAATAGTGCAGAAACGCTGTATTTT	TRBV29 CASSPGTNSAETLYF TRBJ2-3	TRBV29 TRBJ2-3	262
TRBV4	TRBJ1-1	CASSPQDTEVFF	TGTGCCAGCAGCCCCCAGGACACAGAAGTCTTCTTT	TRBV4 CASSPQDTEVFF TRBJ1-1	TRBV4 TRBJ1-1	28
TRBV4	TRBJ1-1	CASSPQDTEVFF	TGTGCCAGCAGCCCCCAGGACACAGAAGTCTTCTTT	TRBV4 CASSPQDTEVFF TRBJ1-1	TRBV4 TRBJ1-1	2
TRBV2	TRBJ2-7	CASSQDLGGRWEQYF	TGTGCCAGCAGCCAAGACCTGGGGGGGCGCTGGGAACAGTACTTC	TRBV2 CASSQDLGGRWEQYF TRBJ2-7	TRBV2 TRBJ2-7	658
TRBV5	TRBJ2-4	CASSQENGGSQNTLYF	TGTGCCAGCAGCCAAGAGAATGGGGGTAGTCAAAACACCTTGTACTTT	TRBV5 CASSQENGGSQNTLYF TRBJ2-4	TRBV5 TRBJ2-4	762
TRBV5	TRBJ1-3	CASSQRDRGSGNTLYF	TGTGCCAGCAGCCAACGGGACAGGGGATCTGGAAATACGCTCTATTTT	TRBV5 CASSQRDRGSGNTLYF TRBJ1-3	TRBV5 TRBJ1-3	59
TRBV5	TRBJ1-3	CASSQRDRGSGNTLYF	TGTGCCAGCAGCCAACGGGACAGGGGATCTGGAAATACGCTCTATTTT	TRBV5 CASSQRDRGSGNTLYF TRBJ1-3	TRBV5 TRBJ1-3	2
TRBV5	TRBJ2-4	CASSQVLGSQNTLYF	TGTGCCAGCAGCCAAGTGCTGGGGAGTCAAAACACCTTGTACTTT	TRBV5 CASSQVLGSQNTLYF TRBJ2-4	TRBV5 TRBJ2-4	282
TRBV17	TRBJ2-7	CASSRDRSYEQYF	TGTGCTAGCAGTAGGGACAGATCCTATGAACAGTACTTC	TRBV17 CASSRDRSYEQYF TRBJ2-7	TRBV17 TRBJ2-7	4
TRBV17	TRBJ2-7	CASSRDRSYEQYF	TGTGCTAGCAGTAGGGACAGGTCCTATGAACAGTACTTC	TRBV17 CASSRDRSYEQYF TRBJ2-7	TRBV17 TRBJ2-7	28
TRBV17	TRBJ2-7	CASSRGGEQYF	TGTGCTAGCAGTAGGGGGGGTGAACAGTACTTC	TRBV17 CASSRGGEQYF TRBJ2-7	TRBV17 TRBJ2-7	130
TRBV12-1	TRBJ2-4	CASSRGLGGRQNTLYF	TGTGCCAGCTCTCGCGGACTGGGGGGGCGGCAAAACACCTTGTACTTT	TRBV12-1 CASSRGLGGRQNTLYF TRBJ2-4	TRBV12-1 TRBJ2-4	162
TRBV12-1	TRBJ1-4	CASSRPNERLFF	TGTGCCAGCTCTCGCCCAAACGAAAGATTATTTTTC	TRBV12-1 CASSRPNERLFF TRBJ1-4	TRBV12-1 TRBJ1-4	572
TRBV15	TRBJ2-4	CASSRRESQNTLYF	TGTGCCAGCAGCCGCCGGGAGAGTCAAAACACCTTGTACTTT	TRBV15 CASSRRESQNTLYF TRBJ2-4	TRBV15 TRBJ2-4	138
TRBV15	TRBJ2-4	CASSRRESQNTLYF	TGTGCCAGCAGCCGCCGGGAGAGTCAAAACACCTTGTACTTT	TRBV15 CASSRRESQNTLYF TRBJ2-4	TRBV15 TRBJ2-4	2
TRBV19	TRBJ2-4	CASSRTGGQNTLYF	TGTGCCAGCAGTAGGACTGGGGGTCAAAACACCTTGTACTTT	TRBV19 CASSRTGGQNTLYF TRBJ2-4	TRBV19 TRBJ2-4	136
TRBV17	TRBJ2-4	CASSSGGGQNTLYF	TGTGCTAGCAGTTCCGGCGGGGGTCAAAACACCTTGTACTTT	TRBV17 CASSSGGGQNTLYF TRBJ2-4	TRBV17 TRBJ2-4	312
TRBV29	TRBJ1-6	CASSSGGNSPLYF	TGTGCTAGCAGTTCAGGGGGAAATTCGCCCCTCTACTTT	TRBV29 CASSSGGNSPLYF TRBJ1-6	TRBV29 TRBJ1-6	9
TRBV29	TRBJ1-6	CASSSGGNSPLYF	TGTGCTAGCAGTTCAGGGGGGAATTCGCCCCTCTACTTT	TRBV29 CASSSGGNSPLYF TRBJ1-6	TRBV29 TRBJ1-6	616
TRBV29	TRBJ1-6	CASSSGGNSTLYF	TGTGCTAGCAGTTCAGGGGGGAATTCGACCCTCTACTTT	TRBV29 CASSSGGNSTLYF TRBJ1-6	TRBV29 TRBJ1-6	58
TRBV19	TRBJ1-4	CASSSGQGSERLFF	TGTGCCAGCAGTTCGGGACAGGGAAGCGAAAGATTATTTTTC	TRBV19 CASSSGQGSERLFF TRBJ1-4	TRBV19 TRBJ1-4	24
TRBV12-1	TRBJ1-2	CASSSGTGGSDYTF	TGTGCCAGCTCTTCCGGGACAGGGGGGTCCGACTACACCTTC	TRBV12-1 CASSSGTGGSDYTF TRBJ1-2	TRBV12-1 TRBJ1-2	312

TRBV17	TRBJ2-4	CASSTGLGQNTLYF	TGTGCTAGCAGTACAGGGTTAGGTCAAAACACCTTGTACTTT	TRBV17 CASSTGLGQNTLYF TRBJ2-4	TRBV17 TRBJ2-4	412
TRBV19	TRBJ1-3	CASSWDSSGNTLYF	TGTGCCAGCAGTTGGGACAGCTCTGGAAATACGCTCTATTTT	TRBV19 CASSWDSSGNTLYF TRBJ1-3	TRBV19 TRBJ1-3	7788
TRBV31	TRBJ1-1	CAWSLRGANTEVFF	TGTGCCTGGAGTCTAAGGGGTGCAAACACAGAAGTCTTCTTT	TRBV31 CAWSLRGANTEVFF TRBJ1-1	TRBV31 TRBJ1-1	1
TRBV31	TRBJ1-1	CAWSLRGANTEVFF	TGTGCCTGGAGTCTAAGGGGTGCAAACACAGAAGTCTTCTTT	TRBV31 CAWSLRGANTEVFF TRBJ1-1	TRBV31 TRBJ1-1	432
TRBV1	TRBJ2-7	CTCSADRAGGYEQYF	TGCACCTGCAGTGCAGATAGGGCAGGGGGCTATGAACAGTACTTC	TRBV1 CTCSADRAGGYEQYF TRBJ2-7	TRBV1 TRBJ2-7	200
TRBV29	TRBJ2-7	CVSSPGTGGYGQYF	TGTGTTAGCAGTCCCGGGACAGGGGGCTATGGACAGTACTTC	TRBV29 CVSSPGTGGYGQYF TRBJ2-7	TRBV29 TRBJ2-7	2
TRBV19	TRBJ1-3	GDSSWESSGNTRYF	GGGGACAGCAGTTGGGAAAGCTCTGGAAATACGCGATATTTT	TRBV19 GDSSWESSGNTRYF TRBJ1-3	TRBV19 TRBJ1-3	2
TRBV21	TRBJ2-1	YAEQFF	TATGCTGAGCAGTTCTTC	TRBV21 YAEQFF TRBJ2-1	TRBV21 TRBJ2-1	366

'able S6. List of TMP epitopes selected in silico to bind HLA-A2 with high affinity (<50nM)

Peptide	Start	Stop	HLA	Sequence	Affinity(nM)
1	357	365	HLA-A0201	AMIEFTQGL	4.88
2	1462	1470	HLA-A0201	KMVETLEEI	7.8
3	1397	1405	HLA-A0201	RLLKYDVGV	11.55
4	765	773	HLA-A0201	TLVGVTFAI	16.94
5	1374	1382	HLA-A0201	AMQNLVAAV	17.32
6	793	801	HLA-A0201	ATMATANGV	20.56
7	862	870	HLA-A0201	AMSMNMEEV	24.83
8	504	512	HLA-A0201	KVFGKMTSV	26.84
9	1130	1138	HLA-A0201	LLGIYQSYV	29.4
10	631	639	HLA-A0201	KLAKFASVV	29.89
11	1176	1184	HLA-A0201	KLWANMSKA	30.99
12	692	700	HLA-A0201	MLSNPITAI	32.68
13	700	708	HLA-A0201	ILVAITTTI	36.32
14	491	499	HLA-A0201	KMAALAASA	46.27
15	691	699	HLA-A0201	AMLSNPTTA	49.58
16	473	481	HLA-A0201	NMAEAFASA	49.85

Table S7. Patient characteristics. Description corresponding to Figure 4E-F

Stage IIIC/IV NSCLC	*Cohort (CHUM)*	*Cohort (CGFL)*	*Cohort (validation)*
***Numbers** (n)*	*44*	*62*	*51*
***Age** (mean, range)*	*65 (45-81)*	*65,5 (46-85)*	*65 (42-83)*
***Gender (n)** Male Female*	*20 (45.5%)* *24 (54.5%)*	*48 (77.4%)* *14 (22.6%)*	*25 (49%)* *26 (51%)*
Smokers (n)			
Yes No NA	*41 (93.2%)* *3 (6.8%)*	*54 (87.1%)* *6 (9.7%)* *2 (3.2%)*	*41 (80.4%)* *10 (19.6%)* *0 (0%)*
***Histology (n)** Adenocarcinoma*			
Squamous cell carcinoma	*32 (72.7%)*	*30 (48,.4%)*	*47 (92.2%)*
Other	*8 (18.2%)* *4 (9.1%)*	*31 5(50%)* *1 (1.6%)*	*4 (7.8%)* *0 (0%)*
Immunotherapy (n)			
Pembrolizumab Nivolumab	*23 (52.3%)*	*0 (0%)*	*20 (39.2%)*
Atezolizumab	*21 (47.7%)* *0 (0%)*	*62 (100%)* *0 (0%)*	*28 (55%)* *3 (5.8%)*
Line of therapy (n)			
1L	*7 (15.9%)*	*0 (0%)*	*8 (15.7%)*
2L	*36 (81.8%)*	*62 (100%)*	*43 (84.3%)*
NA	*1 (2.3%)*	*0 (0%)*	*0 (0%)*
PDL-1 status (n)			
>50%	*20 (45.5%)*	*16 (25.8%)*	*14 (27.5%)*
<50% NA	*11 (25%)* *13 (29.5%)*	*33 (53.2%)* *13 (21%)*	*37 (72.5%)* *0 (0%)*

'able S8. Peptide sequences

Peptide_Name	*Peptide_Sequence*
MART-1_A2_26-35_WT	EAAGIGILIV
MART-1_A2_26-35_Mut	ELAGIGILIV
MART-1_A2_26-35_B1	EAAGIGILAT
MART-1_A2_26-35_B2	EAAGIGFLTA
MART-1_A2_26-35_B3	FLAGIGILTV
MART-1_A2_26-35_B4	ILAGSGILTV
MART-1_A2_26-35_B5	LLAGIGILTV
MELOE-1_A2_36-44_WT	TLNDECWPA
MELOE-1_A2_36-44_B1	DLNDECSPA
MELOE-1_A2_36-44_B2	TLNDECDPT
MELOE-1_A2_36-44_B3	TLNDECINA
MELOE-1_A2_36-44_B4	TLNDEEWAA
MELOE-1_A2_36-44_B5	TLNDEEWKA
MELOE-1_A2_36-44_B6	TLNDEGYPA
MELOE-1_A2_36-44_B7	TLNDELLPA
MELOE-1_A2_36-44_B8	TLNDENWNA
MELOE-1_A2_36-44_B9	TLNDPRWPA
MELOE-1_A2_36-44_B10	TLPDECNPA
MELOE-1_A2_36-44_B11	TLTDEYWPA

able S9. CDR3 alpha and beta sequences of MART-1-specific T-cell clones

MART-1-specific T-cell clones					
T-cell clone	*TRBV*	*CDR3beta*	*TRAV*	*CDR3alpha*	*origin[1]*
10C10	*4-3*	*CASSPGTLSDTQYFG*	*12-2*	*CAVNLEGNNRLAFG*	*Patient PBMC*
12A8	*20-1*	*CSARD**GLG**ELFFG[2]*	*12-2*	*CAVNFDQTGANNLFFG*	
24B7	*4-2*	*CASSQDRGGAETQYFG*	*12-2*	*CAASQGFQKLVFG*	
Cl12	*19*	*CASRWGYLSNQPQHFG*	*35*	*CAGLGAQKLVFG*	
10F8	*20-1*	*CSARD**GLG**ELFFG*	*12-2*	*CAVNLEGNNRLAFG*	
8A7	*5-5*	*CASSSGEGLDTQYFG*	*12-2*	*CAVKAIYFG*	*HV PBMC*
HA1	*25-1*	*CASSEPYKETQYFG*	*12-2*	*CAVGTGTYKYIFG*	
M77-84	*6-1*	*CASSEEVAWGRAETQYFG*	*39*	*CAVDIVPTNDYKLSFG*	*TIL*
M77-80	*28*	*CASTSALLAGGEQYFG*	*29*	*CAASVNARLMFG*	
M199.75	*28*	*CASSLQ**GLG**TEAFFG*	*12-2*	*CALNQAGTALIFG*	
M180.51	*28*	*CASSFE**GLG**TEAFFG*	*19*	*CALSDNTNAGKSTFG*	

MART-1 specific T-cell clones were obtained either from PBMC of melanoma patients or healthy volunteers, fter a step of peptide stimulation, followed by HLA-p/multimer sorting and cloning by limiting dilution, or irectly from TIL spontaneously enriched in Melan-A specific T lymphocytes (Godet et al., Eur J Immunol. 010).

In bold are indicated recurrent motifs already described in the CDR3β of MART-1-specific T-cell clones (Simon t al, Front Immunol., 2018).

Table S10. CDR3 alpha and beta sequences of MELOE-1-specific T-cell clones

MELOE-1-specific T-cell clones					
T-cell clone	*TRBV*	*CDR3beta*	*TRAV*	*CDR3alpha*	*origin*[1]
P2.70	*10-3*	*CAISESWGRDTEAFFG*	*19*	*CALSEAKYNQGGKLIFG*	*Patient PBMC*
P2.45	*14*	*CASSQPSRDRKDNEQFFG*	*19*	***CALSGP****LLGTSYGKLTFG*[2]	
P3.26	*3-1*	*CASSQGSGTSGRRDNEQFFG*	*19*	***CALSGP****ISGGGADGLTFG*	
Cl1	*10-3*	*CAIARTANYGYTFG*	*24*	*CAFIQGNNDMRFG*	
Cl37	*14*	*CASSQERDRGRTNEQFFG*	*19*	***CALSGP****ILTGGGNKLTFG*	
E4H	*11-1*	*CASSVQVSGANVLTFG*	*17*	*CASRGTPLVFG*	*HV PBMC*
DS1.33	*7-2*	*CASSSGLAGTRNYEQYFG*	*19*	*CALRGPMDTGRRALTFG*	
DS2.25	*20-1*	*CSATSLAGIDYGYTFG*	*19*	***CALSGP****FSGGYNKLIFG*	
M170.48	*3-1*	*CASSHKWKREPTDTQYFG*	*19*	***CALSGP****FSDGQKLLFA*	*TIL*
M117.35	*19*	*CASSISEPARRDNEQFFG*	*19*	*CALRGPILTGGGNKLTFG*	

[1] MELOE-1 specific T-cell clones were obtained either from PBMC of melanoma patients or healthy volunteers, after a step of peptide stimulation, followed by HLA-p/multimer sorting and cloning by limiting dilution, or directly from TIL spontaneously enriched in MELOE-1 specific T lymphocytes (Godet et al., Eur J Immunol. 2010).

[2] In bold are indicated recurrent motifs already described in the CDR3α of MELOE-1 specific – T-cell clones (Simon et al, Front Immunol., 2018).

In red are indicated T cell clones cross-reactive against at least one bacterial peptide.

Article 4: “Specific T follicular helper gene signature discriminates large vessel vasculitis patients.” (Desbois et al., JCI insight, in review)

Specific T follicular helper gene signature discriminates large vessel vasculitis patients

AC. Desbois[1,2,3], P. Régnier*[1,2,3], V. Quiniou*[1,2], A. Lejoncour[1,2,3], A. Maciejewski-Duval[1,2], C. Comarmond[1,2,3], H. Vallet[1,2], M. Rosenzwag[1,2], N. Derian[1,2], J Pouchot[4], M Samson[5], T. Manoliu[6], B Bienvenu[7], P. Fouret[8], F. Koskas[9], M Garrido[1,2], D. Sène[10], P. Bruneval[11] , P. Cacoub[1,2,3] , D. Klatzmann[1,2] and D. Saadoun[1,2,3]

*These authors are co-authors

[1] Sorbonne Universités, UPMC Univ Paris 06, INSERM, UMR S 959, Immunology-Immunopathology-Immunotherapy (I3); F-75005, Paris, France;

[2] Biotherapy (CIC-BTi) and Inflammation-Immunopathology-Biotherapy Department (DHU i2B), Hôpital Pitié-Salpêtrière, AP-HP, F-75651, Paris, France;

[3] AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Department of Internal Medicine and Clinical Immunology, F-75013, Paris, France, Centre national de références Maladies Autoimmunes et systémiques rares et Maladies Autoinflammatoires rares
[4]Service de Médecine Interne, Hôpital Européen Georges Pompidou, Paris, France
[5] Service de Médecine Interne, CHU Dijon, Dijon, France
[6]Institut du Cerveau et de la Moelle épinière, Hôpital de la Pitié-Salpétrière, Paris, France
[7]Service de Médecine Interne, CHU Caen, France
[8]Laboratoire d'anatomopathologie ; Groupe Hospitalier Pitié-Salpétrière, Paris, France
[9] Service de Chirurgie vasculaire, Groupe Hospitalier Pitié-Salpétrière, Paris, France
[10] Service de Médecine Interne, Hôpital Lariboisière, Paris, France
[11]Laboratoire d'anatomopathologie, Hôpital Européen Georges Pompidou, Paris, France

Correspondence: David Saadoun, MD, PhD, Department of Internal Medicine and Laboratory I3 « Immunology, immunopathology, immunotherapy » UMR 7211 (CNRS/UPMC) INSERM U959, Hôpital Pitié-Salpétrière, 47-83 boulevard de l'Hôpital, 75013 Paris.

Phone: + (33)(1) 42 17 80 88. Fax: + (33)(1) 42 17 80 33. E Mail: david.saadoun@ aphp.fr

Conflict of interest: The authors have declared that no conflict of interest exists.

Abstract

Background: Takayasu's arteritis (TAK) and giant cell arteritis (GCA), the two most common types of large vessel vasculitis (LVV) exhibit distinct immune response profiles that are currently poorly known.

Objective: To compare transcriptome and phenotype profiles of $CD4^+$ T cells in patients with TA or GCA.

Methods: We performed microarray gene analysis of purified $CD4^+$ T cells and $CD19^+$ B cells of TA and GCA patients. We further investigated their functionality in peripheral blood and arterial lesions in TA and GCA patients and in healthy donors (HD).

Results: Among 730 significantly dysregulated genes in $CD4^+$ T cells of TA compared to GCA patients, we identified the overexpression of CXCR5, CCR6 and CCL20. Circulating $CD4^+$ $CXCR5^+$ T follicular helper (TFH) cells were significantly higher in TA patients as compared to GCA and HD [median of 15.4 % (10; 30.8) versus 5.3 % (1.4; 12.2) and 9.7% (5.6; 12.5) % of $CD4^+$ cells ($p<0.0001$ and $p=0.0001$)], respectively. Among TFH subpopulations, $CD4^+$ $CXCR5^+$ $CCR6^+$ $CXCR3^-$ TFH^-17 cells were specifically increased in TA. Functionally, $CD4^+$ $CXCR5^+$ T cells helped B cells to proliferate, to differentiate into memory cells and to secrete IgG through JAK/STAT pathway. In arterial inflammatory lesions, we found higher proportion of tertiary lymphoid structures composed of $CD4^+$, $CXCR5^+$, PD^-1^+ and CD^-20^+ cells in TA compared to GCA. Sequencing of the α/β TCR repertoire revealed oligoclonal $CD4^+$ $CXCR5^+$ T cells in the aorta of TA patients, suggesting an antigenic stimulation.

Conclusion: We established a TFH^-specific signature in circulating $CD4^+$ T cells that efficiently distinguishes LVV patients. TFH and B cells cooperation might be critical in developing tertiary lymphoid structures in TA aorta.

Introduction:

Takayasu's arteritis (TA) and giant cell arteritis (GCA) are the two most common types of large vessel vasculitis (LVV). Historically, TA and GCA have been considered as distinct diseases on the basis of differences in age at disease onset, ethnic distribution and clinical features including predilection for different arterial territories. All patients with TA have disease involvement of the aorta or its primary branches. In contrast, GCA is traditionally considered as a disease of the cranial arteries. However, with the more frequent use of angio-computed tomography (CT) or 18F-fluorodeoxyglucose positron emission tomography (FDG-PET) recent reports estimated the presence of large vessel involvement in 30 to 70% of patients with GCA [1,2]. Older necropsic study found vascular changes in the large arteries in up to 80% of patients with GCA [3] .

Lesions in LVV are characterized by granulomatous inflammatory infiltrates of the media, the media-intimal junction and the adventitia, often affecting the vasa vasorum. The inner half of the aorta is more often affected than the outer half and adventitia in GCA lesions[4]. Intimal hyperplasia is frequently observed whereas the adventitia is relatively spared in GCA as compared to TA. Scarring can be seen in the later phase, with dense adventitial fibrosis and great fibrous thickening of the intima.

Pathological mechanisms in LVV are not well understood. T cells have been shown to be critical since the secretion of inflammatory cytokines was abolished when T cells were depleted in SCID mice grafted with human inflammatory temporal arteries[5]. Consistently, both diseases are driven by Th1 and Th17 unbalanced immune responses[5,6]. Associations with specific MHC class II molecules (HLA-B52 in TA and HLA-DRB1*04 in GCA) have been demonstrated[7,8]. Increasing evidence also supports a role for B cells in the pathogenesis of LVV. Immuno-histochemical analyses of aortic wall samples in patients with GCA and TA have shown B cells in the inflamed arterial lesions [11,12] and some studies have pinpointed the presence of tertiary lymphoid organs (TLOs) in aorta of LVV patients [11,12]. Altogether, these data suggest the role of T and B cells interactions in pathogenesis of LVV. However, immune activation pathways specifically involved in each disease are poorly known. Herein, we compared microarray gene analysis of purified $CD4^+$ T cells of TA and GCA patients. We further investigated their functionality in peripheral blood and arterial lesions in TA and GCA patients.

Methods

Patients

The study population consisted of 54 TA [median (IQR) age: 32.4 years (27.2; 53.2)] and 52 GCA patients [median age of 74.7 (66.3; 83.2)] (Table 1) and 60 age and sex-matched healthy donors (HD) [64.1% of female, age: 38 years (21; 93)]. All TA patients fulfilled the American College of Rheumatology (ACR) and/or Ishikawa criteria modified by Sharma [13,14]. GCA patients fulfilled the international criteria for GCA [17]. The study was approved by our institutional ethics review board and was performed according to the Helsinki declaration. Patients gave informed consent.

Transcriptome of CD4+ T and CD19+B cells

$CD3^+$ T cells were isolated from PBMC of active TA (n=25) and GCA (n=27) patients by negative isolation using DYNABEADS® untouched™ Human T Cells Kit (Thermofisher Scientific) according to the manufacturer's instructions. Patients should not receive steroids higher than 10mg/day and/or immunosuppressants. CD4+ cells were then isolated by positive selection using DYNABEADS® CD4 isolation kit according to the manufacturer's instructions. $CD19^+$ B cells were isolated from PBMCs of active TA (n = 10) and GCA (n = 8) patients by DYNABEADS® CD19 positive isolation kit according the instructions of the manufacturer. Once isolated, total RNA from CD4, or CD19 positive cells was extracted using the NucleoSpin® RNA kit (Macherey-Nagel), according to the manufacturer's instructions. Total RNA was quantified by a NanoDrop ND-1000 spectrophotometer. Samples with RNA concentration<20ng/µl were excluded. For quality control, RNA dilution was performed using Agilent RNA 6000 Nano Kit and 1µL of the sample was run on the Nano chip using an Agilent 2100 electrophoresis bioanalyser. The quality of total RNA was assessed by the profile of the electropherogram and by the RNA integrity number that was included between 7.3 and 9.3. For Illumina Beadarrays, cRNA samples were prepared using Illumina TotalPre-96 RNA Amp kit (LifeTechnologies) and hybridized to Illumina Human HT-12 v4 Beadarrays.

Data acquisition and normalization

Raw IDAT files were processed using *illuminaio* R package and concatenated into a single text file summarizing all patients and genes. Data were further background-corrected using *limma* R package and inter-chip batch effects were also removed for each

patient group using *ComBat* method from *sva* R package. Additionally, low-quality samples and true outliers were removed, leading finally to the following sample numbers: 25 for $CD4^+$ from TA, 8 for $CD19^+$ from TA, 27 for $CD4^+$ from GCA and 6 for $CD19^+$ from GCA.

Generation of gene signatures

A broad variety of gene signatures specific of known B, T and TFH immune cell populations were generated using the GEO dataset GSE118165. Multiple differential transcriptome analyzes were performed using *limma* R package and crossing of such results were used to extract highly specific gene signatures of interest populations. The final gene signatures may include specifically upregulated or downregulated genes in the target population versus the others, as depicted in each signature file (**Supplementary Table 1**).

Analysis of cell surface markers and intracellular cytokines in PBMC by flow cytometry

PBMC of TA and CGA patients or healthy controls were stained with the following conjugated monoclonal antibodies, at predetermined optimal dilutions, for 20 minutes at 37°C: CD4-APC-Alexa Fluor 750, CD45RA-PB (Beckman Coulter), CXCR3-APC-Alexa Fluor 700, CXCR5-PE dazzle, ICOS-FITC, PD1-PE, CCR6-APC (BD biosciences). IL-17 staining (IL-17 eFluor 660, ebiosciences) was performed using fixation/permeabilization kit (BD cytofix-cytoperm) in accordance with the manufacturer protocol. Data were acquired using a Navios flow cytometer and analysed with the Kaluza analysis software (Beckman Coulter).

Immunohistochemistry

Detection of CXCR5, PD1, CD20, CD38, CD27, BAFF, CXCL-13, IL-21 and CD4 was performed on fixed, paraffin-embedded samples (aorta) from 7 TA and 7 GCA inflammatory aorta lesions and 3 non-inflammatory aorta. After dewaxing in baths of xylene and ethanol, slides were submitted to antigen retrieval by heating in citrate buffer pH 6.0. Blocking of endogenous peroxidase was performed. Before incubation with primary antibodies, Fc receptor was blocked with normal goat serum 5%. Slides were incubated over night with monoclonal mouse anti-human CD4 (dilution 1:50, Abcam)

rabbit polyclonal anti-CXCR5 (dilution 1:20, Abcam), mouse monoclonal anti-PD1 (dilution 1:50, Abcam), monoclonal anti-CD20 (dilution 1:200, DAKO), CD38 (dilution 1:100, Abcam), CD27 (dilution 1:100, Abcam), BAFF (dilution 1:100, Abcam), CXCL13 (dilution 1:100, R&D), IL-21 (dilution 1:100; Merc) or with isotype control: polyclonal Rabbit IgG or monoclonal mouse IgG (Abcam). Slides were then incubated for 30 minutes at room temperature with a biotinylated secondary antibody (1:250). To amplify the signal, Avidin-Biotin complex is then incubated with the tissue section and peroxidase was revealed by diaminobenzidine (DAB) in the presence of H2O2. Finally, slides were mounted in Mowiol, and evaluated under microscopy.

Gene expression quantification at the mRNA level

Quantification of mRNA expression was performed on CD4 positive cells of TA and GCA patients. CD20, PD-1, CXCR5 and CD-45 genes were analyzed. Briefly, total RNA was extracted using the High Pure FFPET RNA Isolation Kit (Roche) for aorta and NucleoSpin® RNA kit (Macherey-Nagel) for CD4 positive cells and reverse-transcribed using SuperScript VILO cDNA Synthesis Kit (Invitrogen) according to the manufacturer's instruction. Gene expression was determined by real-time PCR. Each cDNA sample was amplified in triplicate using SYBR Green (Applied Biosystems) on 7500 FAST Real-time PCR System (Applied Biosystems). The thermal cycling conditions comprised an initial denaturation step at 95°C for 10 minutes, followed by 40 cycles at 95°C for 15 seconds and 65°C for 1 minute. Quantitative values were obtained from the threshold cycle (Ct) number at which the increase in the signal associated with exponential growth of PCR products began to be detected. *RPLPO* gene was used as an endogenous control and each sample was normalized on the basis of its RPLPO content and the expression of CD45. The primers were designed to span introns and are listed in **supplementary Table 2**.

Cultures of B cells and T cells

$CXCR5^+$ or $CXCR5^-$ $CD4^+$ T cells (50 000 cells each/well) of active TA (n=8) patients were cultured with 20 000 naïve B cells (defined as $CD27^-$ IgD^+ $CD19^+$ cells) in the presence of a surperantigen (Cytostim, human, Miltenyi Biotec) (2µl per million of cells) in RPMI1640 complete medium supplemented with 10% heat-inactivated FBS with or without Jak inhibitors (ruxolitininb). Proliferation by CFSE staining was performed at day 3 and differentiation of B cells by flow cytometry at day 7. Cytokines concentrations

were determined in culture supernatants at day 7 by Multiplex® (Merck Millipore). The IgG and IgM concentrations were measured by ELISA (Human IgM, IgA and IgG quantitation Set, Bethyl Laboratories).

Results:

Specific CXCR5, CCR6 and PD1 CD4$^+$ T cells genes signature in TA

We performed a microarray gene analysis of purified CD4$^+$ T cells of active GCA (n=27) and TA (n=25) patients (with steroids ≤10mg/day and no immunosuppressants) to analyse distinct molecular pathways. We found that CD4$^+$ T cells of GCA and TA patients exhibit distinct RNA signatures. A total of 730 genes were significantly dysregulated between these two diseases (with an adjusted p-value threshold of 0.05) of which 419 and 311 of them were up-regulated and down-regulated in TA versus GCA, respectively. Among the most up-regulated genes in TA, we found CXCR5, CCR6 and CCL20 **(Figure 1A).** Interestingly, CXCR5 and CCR6 genes are expressed by T Follicular Helper (TFH cells in blood, in particular by TFH-17. As TA and GCA patients are characterized by different ages, we confirmed that the expression of CXCR5 was not correlated with the age of patients in both diseases (**Figures S1A and B**). We next performed a principal component analysis (PCA) using the signature previously generated. We showed a gene signature leading to a very clear discrimination between the 2 diseases and the genes explaining the most such discrimination included CXCR5, CCR6, PD-1 and CCL20 that were all up-regulated in TA **(Figure 1B).** We confirmed by qPCR the overexpression of CXCR5 mRNA in CD4$^+$ T cells of TA compared to GCA patients (p=0.02) (**Figure S1C**). Next, using publicly available GEO microarray data (GSE118165), we constructed a TFH-specific gene signature and computed a signature score using an in-house algorithm (*see Material and Methods section*) that allows estimating the enrichment of a signature in a given patient group. We found that this TFH-specific gene-signature was significantly more present in the CD4$^+$ T cells of TA compared to GCA patients (**Figure 1C**).

We next confirmed by flow cytometry the over-expression of CXCR5 by CD4$^+$ cells in TA. The proportion of CXCR5$^+$ CD4$^+$ cells [defined as circulating TFH (cTFH)] among CD4$^+$ cells was dramatically higher in TA compared to GCA patients or healthy donors (HD) [median proportion of CXCR5$^+$ CD4$^+$ T cells of 15.4 (10; 30.8)% in TA versus 5.3 (1.4; 12.2)% in GCA (p<0.0001) and 9.7 (5.6; 12.5)% in HD (p=0.0001)] **(Figures 1D and E).** CXCR5$^+$ cells were also higher in active TA patients as compared to inactive one (**Figure 1F**).

Circulating CCR6$^+$ CXCR5$^+$ CD4$^+$ TFH-17 cells are significantly increased in TA

We next studied more deeply TFH differentiation in LVV patients. The same procedure as the one described in **Figure 1C** was performed with a TFH17-specific gene signature obtained

from publicly available GEO dataset (GSE118165). The results show that this signature was also highly enriched in $CD4^+$ T cells of TA compared to GCA patients (**Figure 2A**). Also, the raw expression of some of the most enriched genes composing this $TFH17^-$specific signature were indeed greatly enriched in $CD4^+$ T cells of TA versus GCA patients and allow to efficiently separate both diseases (**Figure 2A**). By flow cytometry, we found that the frequency of total $CXCR6^+$ cells among $CD4^+$ T cells was also higher in TA as compared to GCA patients or HD [18.5 (2.8; 35.5)% in TA vs 4 (0.6; 10)% in GCA patients (p=0.001) and 7 (1.8; 17.5)% in HD (p=0.01)]. The proportion of $CXCR5^+$ $CCR6^+$ cells among $CD4^+$ cells was also increased [4.52 (1.07; 13.35)% in TA vs 0.69 (0.13; 2.16)% in GCA (p=0.0001) and 2.3 (0.3; 4.2)% in HD (p=0.02)] **(Figure 2 B and C).** Circulating $CXCR5^+$ T cells in TA were predominantly composed of TFH^-17 cells (**Figure 2D**). Finally, we studied cytokines involved in TFH17 differentiation. We found that IL^-17 and IL^-6 were increased in culture supernatants of TA compared to GCA patients [IL^-17: 79.2 (±109.9) pg/ml in TA vs 17.3 (±21.6) in GCA, p=0.02 and IL^-6: 69.8 (±72.9) pg/ml in TA and 31.5(±26.8) in GCA, p=0.06] **(Figures 2E and F).**

Specific B cells activation profile in TA

As we have evidenced a TFH gene signature in TA, we next studied the transcriptome of B cells in both diseases. To better characterize the differences between $CD19^+$ B cells of TA and GCA patients, we performed a biological pathways enrichment analysis using GSVA R package and plotted all the obtained results on a network graph, each node representing an enriched gene set coming from Gene Ontology database (**Figure 3A**). Among all the clusters represented, we notably found that a cluster of pathways was specifically enriched in TA versus GCA patients, and that this cluster was itself enriched in pathways closely related to B cells homeostasis : “B cell proliferation”, “B cell activation”, “B cell differentiation” and “Regulation of B cell differentiation”. Moreover, the most important genes associated with the previously described enriched pathways were very clearly enriched in TA versus GCA patients (**Figure 3B**). Together, these data strongly suggest that the homeostasis and the activation of B cells may participate in the physiopathology of TA disease. Consistently, histological analysis of TA aorta revealed major infiltrates of CD20 positive cells with nodular organisation whereas the presence of $CD20^+$ cells was weaker in GCA with less nodular organization (**Figure 3C**). We also observed the expression of markers of B cells differentiation and growth such as CD27, CD38 and BAFF (**Figure 3C and S2A**). We next quantified the expression of CD20 in TA and GCA arteries. The surface area of CD20

positive cells was significantly higher in TA aorta as compared to GCA (p=0.01) (**Figure 3D**). The relative expression of CD20 mRNA by qPCR was also higher in TA aorta as compared to GCA aorta (p=0.05). Consistent with these data, we found that the absolute number of circulating $CD19^+$ cells was significantly increased in TA patients as compared to GCA patients and HD [268.6 (±131) vs 111.9 (±78.8) and 203 (±116.6), p=0.002 and p=0.02, respectively] (**Figure S2B**). The frequency of circulating $CD19^+$ cells was also increased in TA compared to GCA patients and HD [13.8 (±4)% vs 8.8 (±5)% and 11.6 (±4.5)%, p=0.04 and p=0.0004, respectively] (**Figure 3E**). Levels of BAFF in sera of TA tended to be higher compared to GCA patients [498.9 (±141.3) pg/ml vs 424.9(±177.4)] (**Figure S2C**). We did not find significant differences in B cells sub-populations between TA, GCA patients and HD.

$CXCR5^+$ $CD4^+$T cells efficiently help naïve B cells in TA through JAK/STAT pathway

As we have shown a specific TFH differentiation in TA compared to GCA, we next aimed to confirm the functionality their functionality in TA. $CXCR5^+$ or $CXCR5^-$ $CD4^+$ T cells of TA patients were cultured with naïve $CD27^-$ IgD^+ $CD19^+$ B cells of the same patients in the presence of a superantigen (Cytostim®). $CXCR5^+$ $CD4^+$ T cells induced a higher proliferation of B cells at day 3 as compared to $CXCR5^-$ $CD4^+$ T cells (**Figure 4A and B**). Consistently, the proportion of $CD19^+$ cells was significantly increased in cultures including $CXCR5^+$ $CD4^+$ T cells as compared to those with $CXCR5^-$ $CD4^+$ T cells at day 7 [33.8% (±12) versus 24.9% (±6), p=0.008], respectively (**Figure 4C**). Compared to $CXCR5^-$ $CD4^+$ T cells, $CXCR5^+$ $CD4^+$ T cells of TA patients were able to enhance naïve B cells to differentiate into memory B cells [6.7% (±3.4) vs 3.7 % (±2.1), p=0.008], and to secrete IgG (p=0.04) (**Figures 4D and 4E**). We next studied by which cellular pathways CXCR5+ $CD4^+$ T cells were able to help B cells to differentiate. We found that inhibition of JAK/STAT pathway with ruxolitinib led to significant decrease of CD27+ B cells at day 7 [6.7% (±3.4) vs 1.3% (±0.52), p=0.01] (**Figures 4D**) and decrease in IgG/IgM ratio [3.17 (±0.58) vs 1.5 (±0.05), p<0.0001] (**Figures 4E**). Finally, we found that ruxolitinib inhibited significantly IL-6 secretion [11.26 (±10.26) vs 0.63 (±1.07) pg/ml, p=0.02] (**Figure 4F**) although no difference was noted between $CXCR5^+$ and $CXCR5^-$ $CD4^+$ T cells. Altogether, these results showed that $CXCR5^+$ $CD4^+$ T cells that were shown to be increased in TA, were able to help B cells proliferation and differentiation, through JAK/STAT pathway.

Increased tertiary lymphoid structures composed of $CXCR5^+$, $CD4^+$, $PD\text{-}1^+$, $CD20^+$ cells within inflammatory aortic lesions in TA patients

The presence of TLO (Tertiary Lymphoid Organs) has previously been shown in LVV [12]. Thus, we compared the presence of tertiary lymphoid structures in aortic wall of TA (n=12) and GCA (n=15) patients to further study the immune response within aorta inflammatory lesions. All histological aorta samples with haematoxylin and eosin and Masson's trichome colorations were evaluated by the same pathologist. Interestingly, we confirmed that TLO were more frequently observed in aorta of TA patients as compared to GCA patients ($p<0.05$) **(Figures 5A, B and C**). TLO exhibited high expression of CXCR5 within their periphery, as for CD4 staining in TA aorta (**Figure 5B**). PD-1 and BCL-6 were also highly expressed in these structures.

TCR sequencing of CXCR5$^+$ CD4$^+$ T cells in TA aorta

Next, we performed TCR repertoire analysis of FACS-sorted CXCR5$^+$ CD4$^+$ T cells and CXCR5$^-$ CD4$^+$ T in peripheral blood and in aorta in 2 TA patients (**Figure 6A**). We found a very broad repertoire of peripheral CD4$^+$ T cells, whereas it switches to oligoclonality for CD4$^+$ T cells originating from aorta TA lesions (**Figure 6B and S3A**). This trend was dramatically much stronger for CD4$^+$ CXCR5$^+$ T cells than for CD4$^+$ CXCR5$^-$ T cells in arteries of both patients. Indeed, we observed that the TCR repertoire in the aortic cells was much narrow for CXCR5$^+$ CD4$^+$ T cells than in CXCR5$^-$ CD4$^+$ T cells. In TA p# 1, one clonotype represent 97.6% of the TCRα repertoire in CD4$^+$ CXCR5$^+$ aortic T cells. In TA p# 2, two major clonotypes represent 96.1% of the TCRα repertoire founded in CD4$^+$ CXCR5$^+$ aortic T cells. The most frequent clonotypes founded in CD4$^+$ CXCR5$^+$ aortic T cells are reported on **Figure S3B**. Motifs of the major aortic CD4$^+$ CXCR5$^+$ sequences have been previously reported in auto-immune/inflammatory diseases [15,16]and in vascular diseases [17,18] (**Figure S3B**).

Discussion

Specific signature of CD4 T cells in TA disease

We demonstrated for the first time a specific gene signature of circulating $CD4^+$ T cells that discriminates large vessel vasculitis patients. Among 730 differentially expressed genes, CXCR5, CCR6 and CCL20 were shown to be one of the most significantly up-regulated genes in $CD4^+$ T cells of TA compared to GCA patients. Up-regulation of CXCR5 and CCR6 was further confirmed by flow cytometry. Circulating $CD4^+$ $CXCR5^+$ T cells have been found in recent studies to have functional characteristics similar to TFH. Indeed, circulating $CXCR5^+$ $CD4^+$ T cells promote survival, proliferation and differentiation of B cells into plasma cells [19]. However, the phenotype of circulating TFH differs from "conventional" tissue-specific TFH cells that express high levels of PD-1 and ICOS. In peripheral blood, only few $CXCR5^+$ $CD4^+$ T cells express ICOS or PD-1[20]. Circulating TFH can be distinguished according their membrane expression of CCR6 and CXCR3 [21]. In our study, we also demonstrated an increase of $CXCR5^+$ $CCR6^+$ $CXCR3^-$ $CD4^+$ T cells in TA, corresponding to TFH-17 population, that was consistently confirmed by transcriptomic analysis using a TFH17 specific gene signature that was greatly enriched in TA compared to GCA patients.

Specific activation profile of B cells in TA

Our unsupervised biological pathways analysis of $CD19^+$ B cells in LVV patients revealed a significant enrichment of B cells-related pathways, such as “B cell activation”, “B cell differentiation”, “B cell proliferation” and “Regulation of B cell proliferation” in TA compared to GCA. Furthermore, the genes involved in such pathways discriminate very efficiently patients of both diseases, showing a clear up-regulation in TA. Consistently, our study confirmed increased B cells number in blood and also within the arterial lesions in TA as compared to GCA patients. Serum levels of BAFF tended to be higher in TA. Only few data have been published on B cells in TA. A study has reported an increased absolute number and frequency of peripheral blood $CD19^{++}CD20^{-}$/CD27high antibody-secreting cells in patients with active TA[22]. The role of B cells in TA could also be supported by the use of rituximab in case-reports[22]. In GCA, studies reported a decrease in circulating B cells in patients with active GCA [23] that normalized rapidly under treatment.

TLO in TA

Consistent with previous studies, we observed the presence of TLO in aorta lesions of TA and GCA patients[12,24,25]. However, we found significant differences between the two diseases.

First, the proportion of TLO was clearly increased in TA compared to GCA aorta lesions. Moreover, their repartition was dramatically different with a highly ordered nodular organisation of the inflammatory infiltrates in TA. TLO may be observed in tissues affected by non-resolving inflammation as a result of infection, autoimmunity, cancer, and allograft rejection. These highly ordered structures are composed of the cells present in the lymphoid follicles typically associated with the spleen and lymph node compartments [26]. The structural similarities between TLOs and B cells follicles found in secondary lymphoid organs suggest a local recruitment of naive cells via HEV, their activation, as well as the establishment of an immunological humoral memory supported by TFH cells. In the present study, TLOs were mainly detected in the adventitia of TA aortic specimen. Analysis of immune adventitial cells revealed a high percentage of memory and antigen-experienced B cells and also the presence of cells expressing canonical TFH cell markers, such as CXCR5, Bcl6, and PD-1. TCR sequencing of $CD4^+$ $CXCR5^+$ T cells located within aorta inflammatory lesions of TA patients showed oligoclonal populations. The restricted repertory of $CD4^+$ $CXCR5^+$ T cells within the aorta strongly suggests antigenic selection of TFH cells. Interestingly, the TCR sequences found in our study have already been reported in diseases known to be linked to autoimmune mechanisms and B cells abnormalities such as Sjogren syndrome [15] and multiple sclerosis[16], and also in coronary and cardiac injuries [17,27]. Altogether, these results suggest an antigen specific immune response based on cooperation of TFH and B cells occurring in the adventitia of TA vessels.

Cooperation between TFH and B cells in TA was further demonstrated by functional test showing that $CD4^+$ $CXCR5^+$ T cells helped B cells to proliferate and to differentiate. The cooperation between TFH and B cells in TA was mediated by JAK/STAT pathway. In GCA, Zhang et al. have shown that tofacitinib effectively suppressed innate and adaptive immunity in the vessel wall, with reduced proliferation and minimal production of the effector molecules interferon-γ, interleukin-17, and interleukin-21[28]. Taken together, these data suggest that JAK inhibitors might be a promising therapeutic strategy in TA.

In summary, we established a TFH-specific signature in circulating CD4+ T cells that efficiently discriminates LVV patients. In arterial inflammatory lesions of TA, we found high proportion of highly organized tertiary lymphoid structures with oligoclonal $CD4^+$ $CXCR5^+$ T cells expansion. We demonstrated the cooperation between TFH and B cells in TA that was mediated by JAK/STAT pathway. Our results provided important insight into the pathogenesis of LVV and pave the way for interesting therapeutic avenue in TA.

REFERENCES

1. Agard C, Barrier J-H, Dupas B, Ponge T, Mahr A, Fradet G, et al. Aortic involvement in recent-onset giant cell (temporal) arteritis: a case-control prospective study using helical aortic computed tomodensitometric scan. Arthritis Rheum. 2008;59:670–6.
2. Prieto-González S, Arguis P, García-Martínez A, Espígol-Frigolé G, Tavera-Bahillo I, Butjosa M, et al. Large vessel involvement in biopsy-proven giant cell arteritis: prospective study in 40 newly diagnosed patients using CT angiography. Ann Rheum Dis. 2012;71:1170–6.
3. Ostberg G. An arteritis with special reference to polymyalgia arteritica. Acta Pathol Microbiol Scand Suppl. 1973;237:Suppl 237:1-59.
4. Stone JR, Bruneval P, Angelini A, Bartoloni G, Basso C, Batoroeva L, et al. Consensus statement on surgical pathology of the aorta from the Society for Cardiovascular Pathology and the Association for European Cardiovascular Pathology: I. Inflammatory diseases. Cardiovasc Pathol Off J Soc Cardiovasc Pathol. 2015;24:267–78.
5. Deng J, Younge BR, Olshen RA, Goronzy JJ, Weyand CM. Th17 and Th1 T-cell responses in giant cell arteritis. Circulation. 2010;121:906–15.
6. Saadoun D, Garrido M, Comarmond C, Desbois AC, Domont F, Savey L, et al. Th1 and Th17 cytokines drive Takayasu Arteritis inflammation. Arthritis Rheumatol Hoboken NJ. 2015;
7. Weyand CM, Hicok KC, Hunder GG, Goronzy JJ. The HLA-DRB1 locus as a genetic component in giant cell arteritis. Mapping of a disease-linked sequence motif to the antigen binding site of the HLA-DR molecule. J Clin Invest. 1992;90:2355–61.
8. Chen S, Luan H, Li L, Zeng X, Wang T, Li Y, et al. Relationship of HLA-B*51 and HLA-B*52 alleles and TNF-α-308A/G polymorphism with susceptibility to Takayasu arteritis: a meta-analysis. Clin Rheumatol. 2017;36:173–81.
9. Weyand CM, Goronzy JJ. Giant-Cell Arteritis and Polymyalgia Rheumatica. Solomon CG, editor. N Engl J Med. 2014;371:50–7.
10. Watanabe R, Zhang H, Berry G, Goronzy JJ, Weyand CM. Immune checkpoint dysfunction in large and medium vessel vasculitis. Am J Physiol Heart Circ Physiol. 2017;312:H1052–9.
11. Graver JC, Sandovici M, Diepstra A, Boots AMH, Brouwer E. Artery tertiary lymphoid organs in giant cell arteritis are not exclusively located in the media of temporal arteries. Ann Rheum Dis. 2018;77:e16.
12. Clement M, Galy A, Bruneval P, Morvan M, Hyafil F, Benali K, et al. Tertiary Lymphoid Organs in Takayasu Arteritis. Front Immunol. 2016;7:158.
13. de Souza AWS, de Carvalho JF. Diagnostic and classification criteria of Takayasu arteritis. J Autoimmun. 2014;48–49:79–83.
14. Sharma BK, Jain S, Suri S, Numano F. Diagnostic criteria for Takayasu arteritis. Int J Cardiol. 1996;54 Suppl:S141-147.
15. Joachims ML, Leehan KM, Lawrence C, Pelikan RC, Moore JS, Pan Z, et al. Single-cell analysis of glandular T cell receptors in Sjögren's syndrome. JCI Insight [Internet]. 2016 [cited 2018 Sep 4];1. Available from: https://insight.jci.org/articles/view/85609
16. Vandevyver C, Mertens N, van den Elsen P, Medaer R, Raus J, Zhang J. Clonal expansion of myelin basic protein-reactive T cells in patients with multiple sclerosis: Restricted T cell receptor V gene rearrangements and CDR3 sequence. Eur J Immunol. 1995;25:958–68.
17. Slachta CA, Jeevanandam V, Goldman B, Lin WL, Platsoucas CD. Coronary Arteries from Human Cardiac Allografts with Chronic Rejection Contain Oligoclonal T Cells:

Persistence of Identical Clonally Expanded TCR Transcripts from the Early Post-Transplantation Period (Endomyocardial Biopsies) to Chronic Rejection (Coronary Arteries). J Immunol. 2000;165:3469–83.
18. Winchester R, Wiesendanger M, O'Brien W, Zhang H-Z, Maurer MS, Gillam LD, et al. Circulating Activated and Effector Memory T Cells Are Associated with Calcification and Clonal Expansions in Bicuspid and Tricuspid Valves of Calcific Aortic Stenosis. J Immunol. 2011;187:1006–14.
19. Morita R, Schmitt N, Bentebibel S-E, Ranganathan R, Bourdery L, Zurawski G, et al. Human blood CXCR5(+)CD4(+) T cells are counterparts of T follicular cells and contain specific subsets that differentially support antibody secretion. Immunity. 2011;34:108–21.
20. Kim CH, Rott LS, Clark-Lewis I, Campbell DJ, Wu L, Butcher EC. Subspecialization of CXCR5+ T cells: B helper activity is focused in a germinal center-localized subset of CXCR5+ T cells. J Exp Med. 2001;193:1373–81.
21. Ueno H, Banchereau J, Vinuesa CG. Pathophysiology of T follicular helper cells in humans and mice. Nat Immunol. 2015;16:142–52.
22. Hoyer BF, Mumtaz IM, Loddenkemper K, Bruns A, Sengler C, Hermann K-G, et al. Takayasu arteritis is characterised by disturbances of B cell homeostasis and responds to B cell depletion therapy with rituximab. Ann Rheum Dis. 2012;71:75–9.
23. van der Geest KSM, Abdulahad WH, Chalan P, Rutgers A, Horst G, Huitema MG, et al. Disturbed B cell homeostasis in newly diagnosed giant cell arteritis and polymyalgia rheumatica. Arthritis Rheumatol Hoboken NJ. 2014;66:1927–38.
24. Ciccia F, Rizzo A, Maugeri R, Alessandro R, Croci S, Guggino G, et al. Ectopic expression of CXCL13, BAFF, APRIL and LT-β is associated with artery tertiary lymphoid organs in giant cell arteritis. Ann Rheum Dis. 2017;76:235–43.
25. Graver JC, Boots AMH, Haacke EA, Diepstra A, Brouwer E, Sandovici M. Massive B-Cell Infiltration and Organization Into Artery Tertiary Lymphoid Organs in the Aorta of Large Vessel Giant Cell Arteritis. Front Immunol. 2019;10:83.
26. Ruddle NH. High Endothelial Venules and Lymphatic Vessels in Tertiary Lymphoid Organs: Characteristics, Functions, and Regulation. Front Immunol. 2016;7:491.
27. Winchester R, Wiesendanger M, O'Brien W, Zhang H-Z, Maurer MS, Gillam LD, et al. Circulating activated and effector memory T cells are associated with calcification and clonal expansions in bicuspid and tricuspid valves of calcific aortic stenosis. J Immunol Baltim Md 1950. 2011;187:1006–14.
28. Zhang H, Watanabe R, Berry GJ, Tian L, Goronzy JJ, Weyand CM. Inhibition of JAK-STAT Signaling Suppresses Pathogenic Immune Responses in Medium and Large Vessel Vasculitis. Circulation. 2018;137:1934–48.

Legends

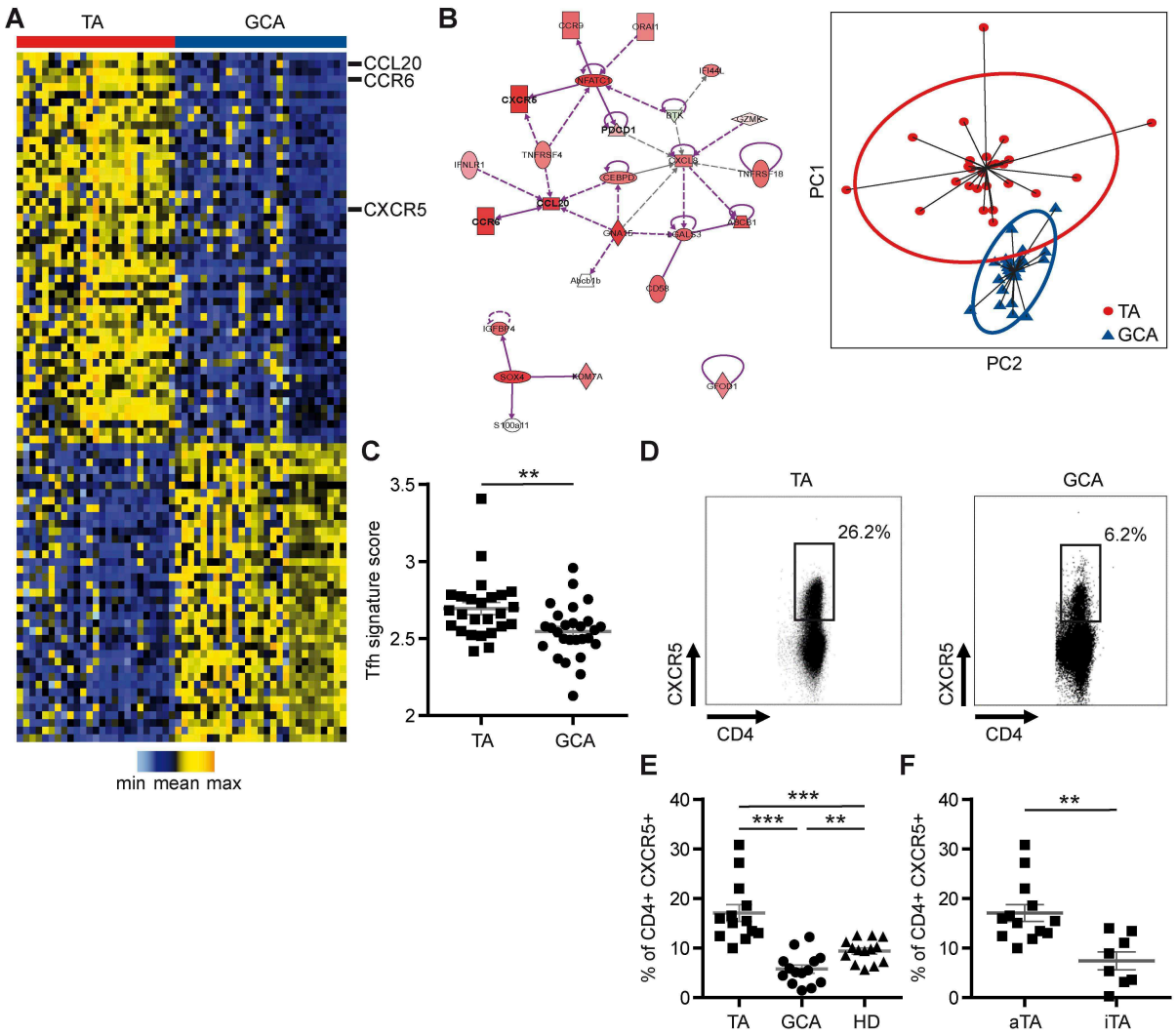


Figure 1: **Specific CXCR5, CCR6 and PD1 CD4$^+$ T cells genes signature in TA**

(**A**) A microarray gene analysis of purified CD4$^+$ T cells of active GCA (n=27) and TA (n=25) patients was performed. Included patients had steroids ≤ 10mg/day and no immunosuppressants. CD4$^+$ T cells of GCA and TA patients exhibited distinct mRNA signatures. CXCR5, CCR6 and CCL20 were significantly up-regulated in TA as compared with GCA patients. (**B**) PCA analyse revealed that the gene signature leading to the best discrimination between GCA and TA included CXCR5, CCR6 and PD-1 genes. (**C**) Using publicly available GEO microarray data (GSE118165), we constructed a TFH-specific gene signature that is significantly more present in the CD4+ T cells of TA versus GCA patients. (**D-E**) Increased frequency of CXCR5+ cells in CD4+ T cells of TA patients in comparison with GCA patients and HD confirmed by flow cytometry. The membrane expression of CXCR-5, CCR-6 and CXCR-3 was analysed among CD4$^+$ T cells of TA (n=13), and GCA (n=14) patients and HD (n=14) by flow cytometry. (**D**) Dot plots represent the proportion of CXCR5$^+$ CD4$^+$ T cells in a TA patient (left panel) and a GCA patient (right panel). (**E**)The whole patient's data are displayed on scatter plots. (**F**) Also, the frequency of CXCR5$^+$ cells was higher in active (n=13) compared to inactive (n=8) TA patients.
**P<0.01. These data are shown as the mean ±SEM. ***P<0.01* ****P<0.001*

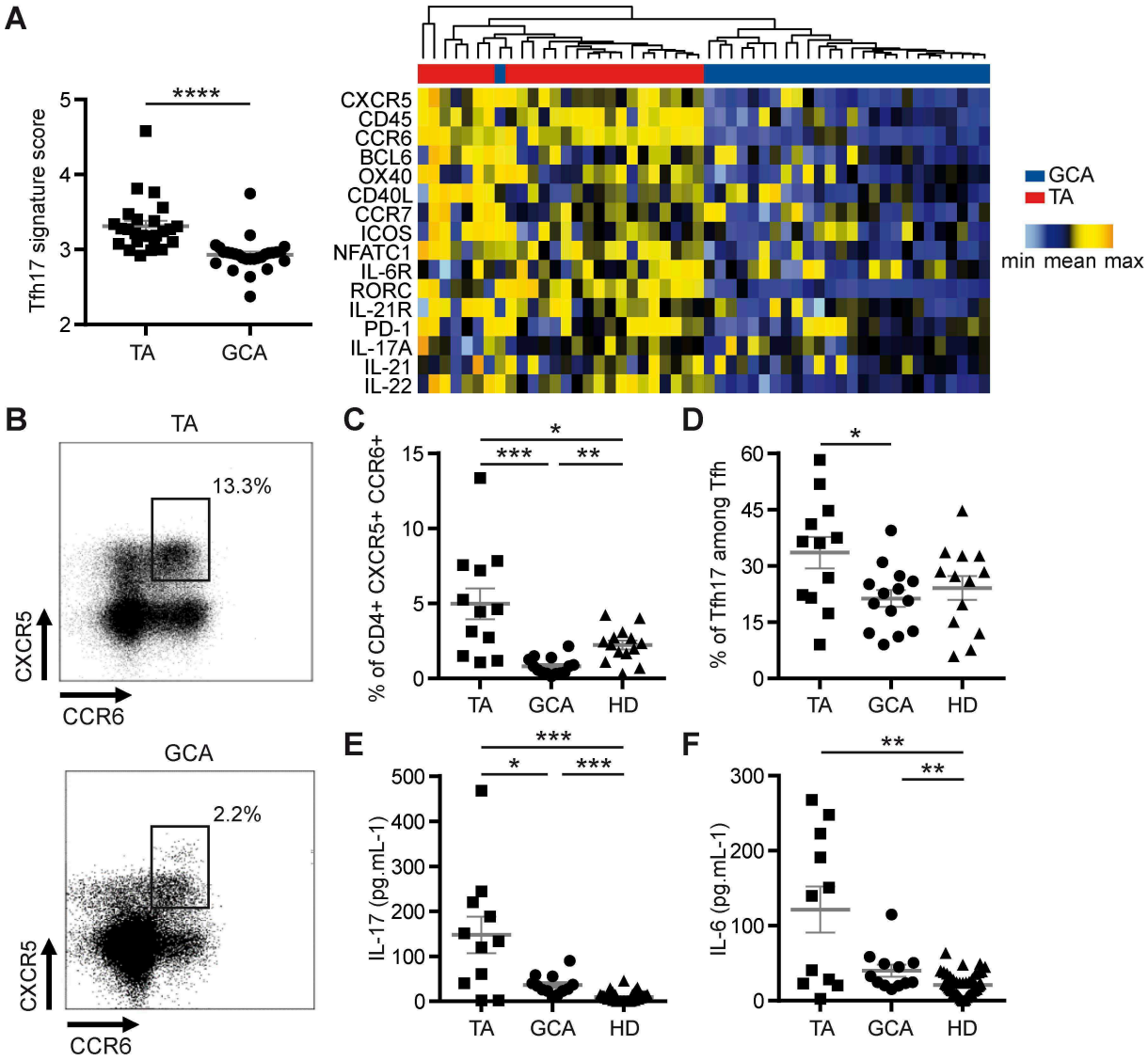


Figure 2: Circulating CCR6^{+} CXCR5^{+} CD4^{+} T cells (TFH^{-}17) are significantly increased in TA

(**A**)TFH17-specific gene signature obtained from publicly available GEO dataset (GSE118165) is highly enriched in CD4+ T cells of TA (n=25) versus GCA (n=27) patients. (**B-C**). The frequency of CXCR5^{+} CXCR6^{+} cells was also higher in CD4^{+} T cells of TA (n=12) patients as compared to GCA (n=14) patients or HD (n=13). (**D**). The proportion of TFH^{-}17 cells was higher in TA as compared to GCA patients and HD. (**E- F**) PBMC of TA (n=11) and GCA (n=12) patients and HD (n=49) were stimulated for 4 hours with 0.05 µg/mL Phorbol 12^{-}myristate 13^{-}acetate (PMA) and 1 mM (1µg/mL) ionomycin. Quantitative determination of IL^{-}17, IL^{-}6, IFNγ, IL^{-}4 and IL^{-}13 was performed in culture supernatants. Levels of IL^{-}17 were significantly higher in TA compared to GCA patients. Levels of IL^{-}6 tended to be higher in TA compared to GCA patients, although not statistically significant.
*P<0.05, ***P<0.001. These data are shown as the mean ±SEM

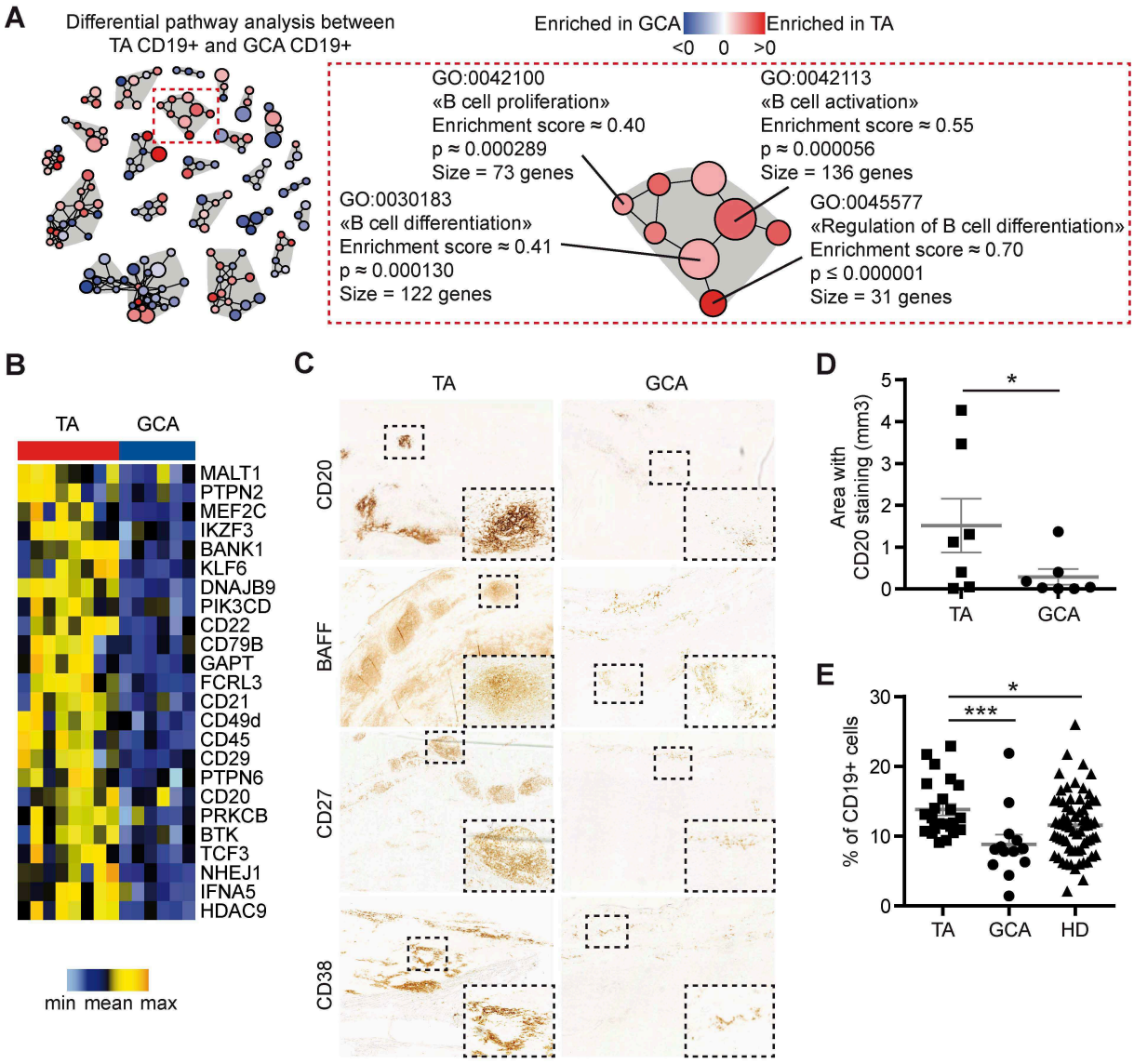


Figure 3: Specific B cells activation profile in TA patients

(**A**) Specific enrichment of cluster of pathways related to B cells homeostasis : "B cell proliferation", "B cell activation", "B cell differentiation" and "Regulation of B cell differentiation" in TA versus GCA patients.(**B**) Heatmap showing the important separation between TA (n=8) and GCA (n=6) patients using genes related to B cells homeostasis.(**C**) Major infiltrates of CD20 positive cells with nodular organisation in TA compared to GCA aorta. Expression of markers of B cells differentiation and growth such as BAFF, CD38 and CD27. (**D**) Surface area staining of CD20 positive cells in TA (n=7) and GCA (n=7) aorta. (**E**). Frequency of $CD19^{+}$ cells was also increased in active and untreated TA patients (n=23) compared to active and untreated GCA patients (n=13) and HD (n=77).
**P<0.05, ***P<0.001*. These data are shown as the mean ±SEM.

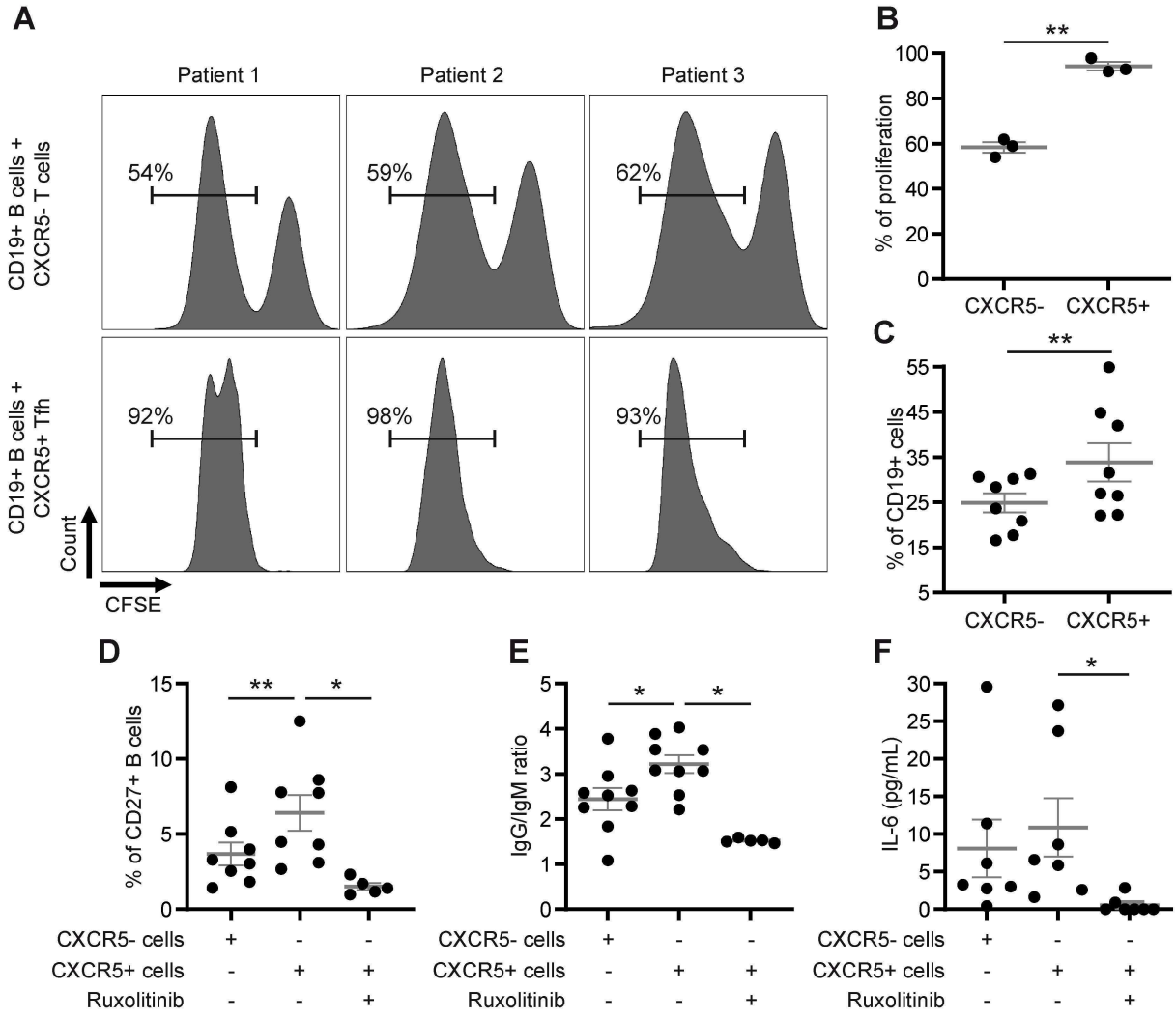


Figure 4: CXCR5$^+$ CD4$^+$ T cells efficiently help naive B cells in TA through JAK/STAT pathway

(**A**) FACS-sorted CXCR5$^+$ or CXCR5$^-$ CD4$^+$ T cells of TA (n=8) and GCA (n=8) patients were cultured with naïve B cells (defined as CD27$^-$ IgD$^+$ CD19$^+$ cells) with stimulation with a surperantigen. Included patients had steroids ≤ 10mg/day and no immunosuppressants Proliferation of B cells was assessed at day 3 by CFSE staining. Histograms of 3 TA patients are represented (P1, P2 and P3) for each culture condition. (**B**)The proportion of proliferative B cells was higher if they are cultured with CXCR5$^+$ CD4$^+$ T cells as compared to culture with CXCR5$^-$ T cells. (**C**)The proportion of CD19$^+$ cells was significantly higher at day 7 in cocultures including CXCR5$^+$ T cells as compared to those with CXCR5$^-$ cells. Results of 8 active TA patients are shown. (**D**) The proportion of CD27$^+$ B cells was significantly higher at day 7 in cocultures including CXCR5$^+$ T cells as compared to those with CXCR5$^-$ cells. Inhibition of the JAK pathway via Ruxolitinib led to suppression of differentiation to CD27+ B cells. Results of 8 active TA patients are shown. (**E**) Measurement of IgG and IgM secretion were performed in culture supernatant (n=9) of naïve B cells cultured with CXCR5$^+$ or CXCR5$^-$ cells. The subsequent secretion IgG/IgM ratio was higher in B cells cultured with CXCR5$^+$ T cells. Ruxolitinib was also associated with a decreased IgG/IgM ratio. (**F**) IL-6 secretion was significantly inhibited when Ruxolitinib was added to cocultures (n=7).
*$P<0.05$,**$P<0.01$. These data are shown as the mean ±SEM

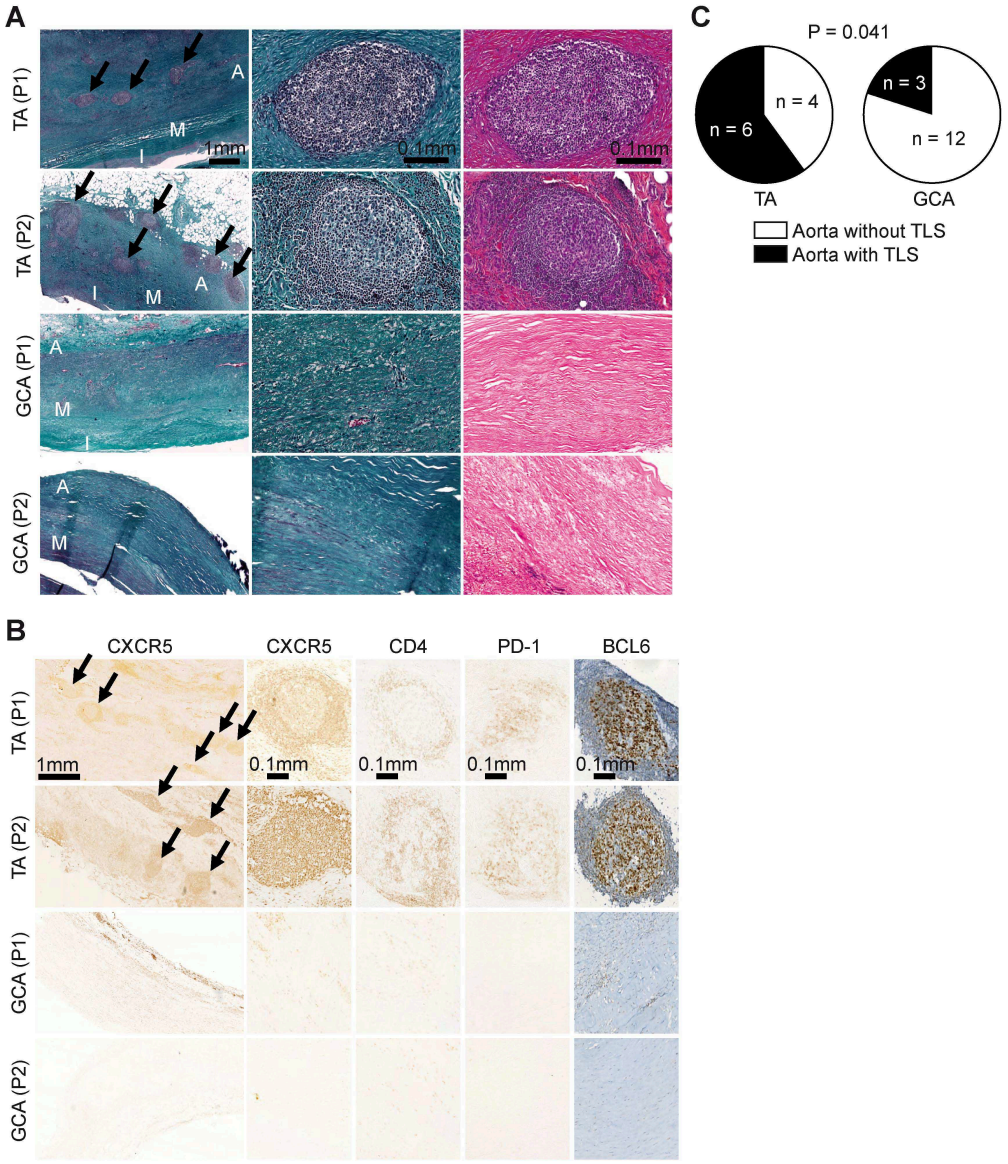


Figure 5: Increase tertiary lymphoid organs within arterial inflammatory lesions of TA

(A) Histological analysis of TA aorta revealed the presence of tertiary lymphoid organs (TLO) in the adventitia (black arrows). Histological analysis of GCA aorta reveals inflammatory infiltrates without nodular and follicular organization. Magnification x 2. A: Adventitia, M: Media, I: Intima (**B**) Analysis of tertiary lymphoid structures (black arrows) in TA aorta by immunohistochemistry revealed a high expression of $CXCR^{-}$ 5. Lymphoid structures also contained $CD4^{+}$ T cells in their periphery and cells with positive staining for $PD^{-}1$ and BCL6. (**C**) The presence of TLO in the aortic wall of TA (n=10) and GCA (n=15) patients was compared. TLOs were more frequently observed in TA compared to GCA patients.
*$P<0.05$. These data are shown as the mean ±SEM

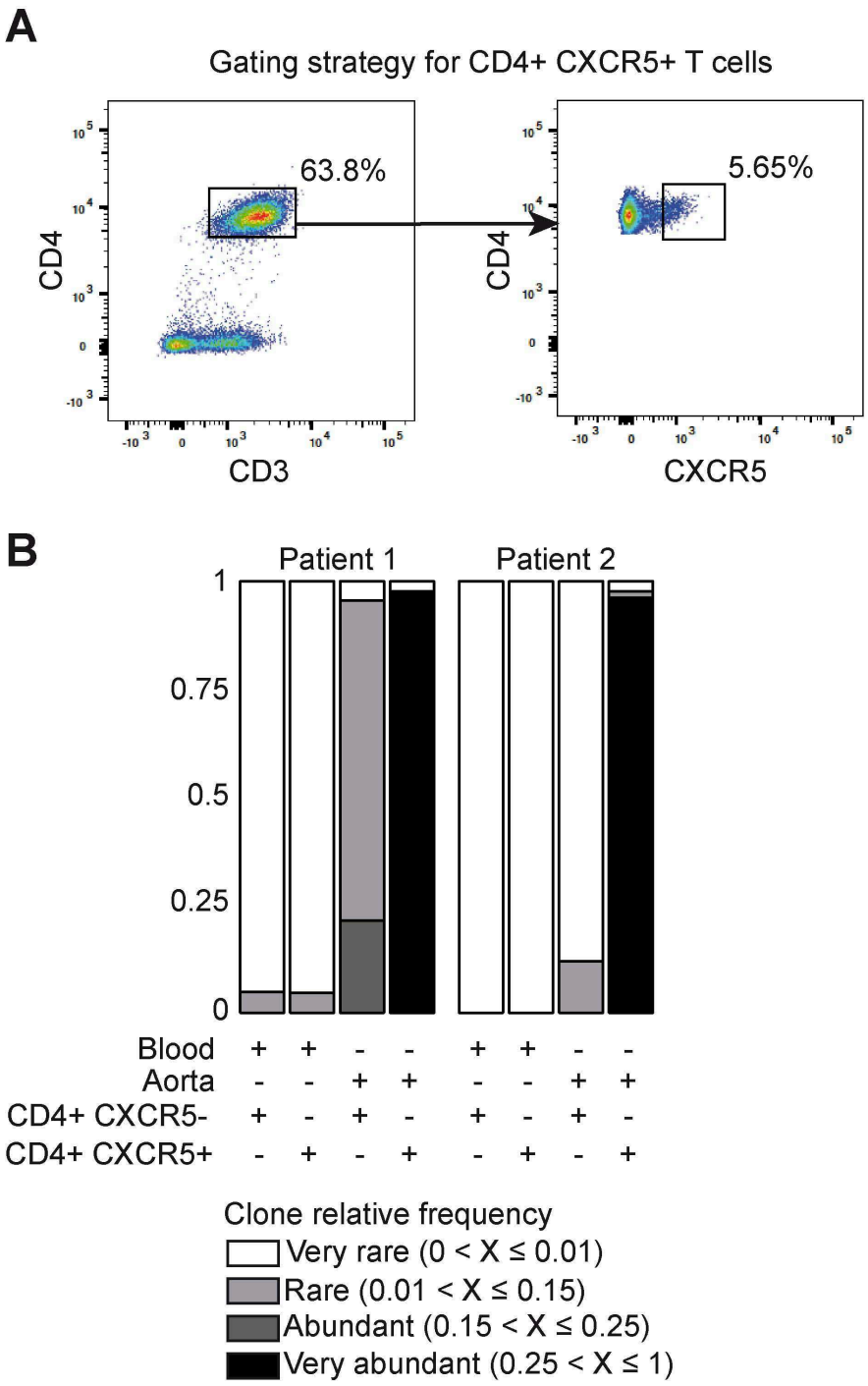


Figure 6: Oligoclonal distribution of arterial TFH cells in TA aorta

(**A**) Gating strategy used to FACS^{-}sorted CD4^{+} CXCR5^{-} and CD4^{+} CXCR5^{+} T cells either coming from blood or aorta samples in TA patients (**n=2**). (**B**) Oligoclonal profile of aorta CD4^{+} CXCR5^{+} T cells compared to blood CD4^{+} CXCR5^{+} T cells and CD4^{+} CXCR5^{-} T cells in TA patients (n=2)

Supplementary Data

Tables:

Table S1: List of genes included in Tfh signature

Table S2.List of primers used for quantification of gene expression

Figures

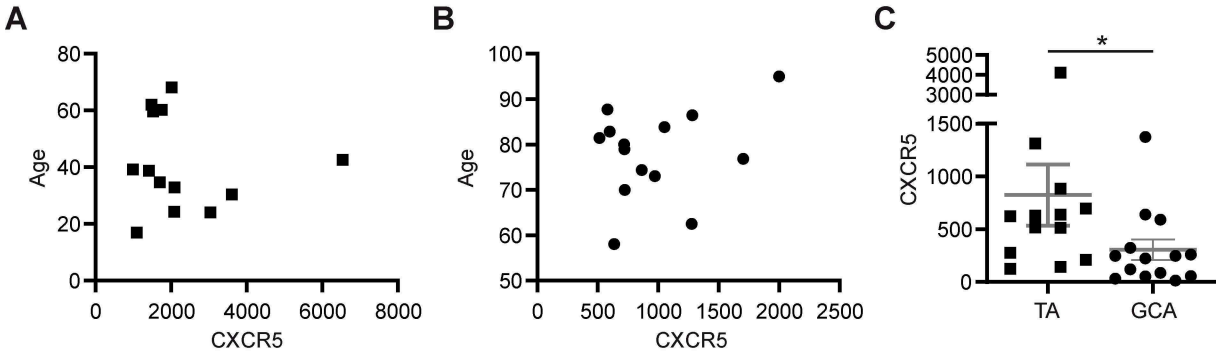


Figure S1: **A and B**. The expression of CXCR5 was not correlated with the age of patients in TA and GCA. **C**. Overexpression of CXCR5 mRNA by qPCR in $CD4^+$ T cells of TA compared to GCA patients; * $p<0.05$.

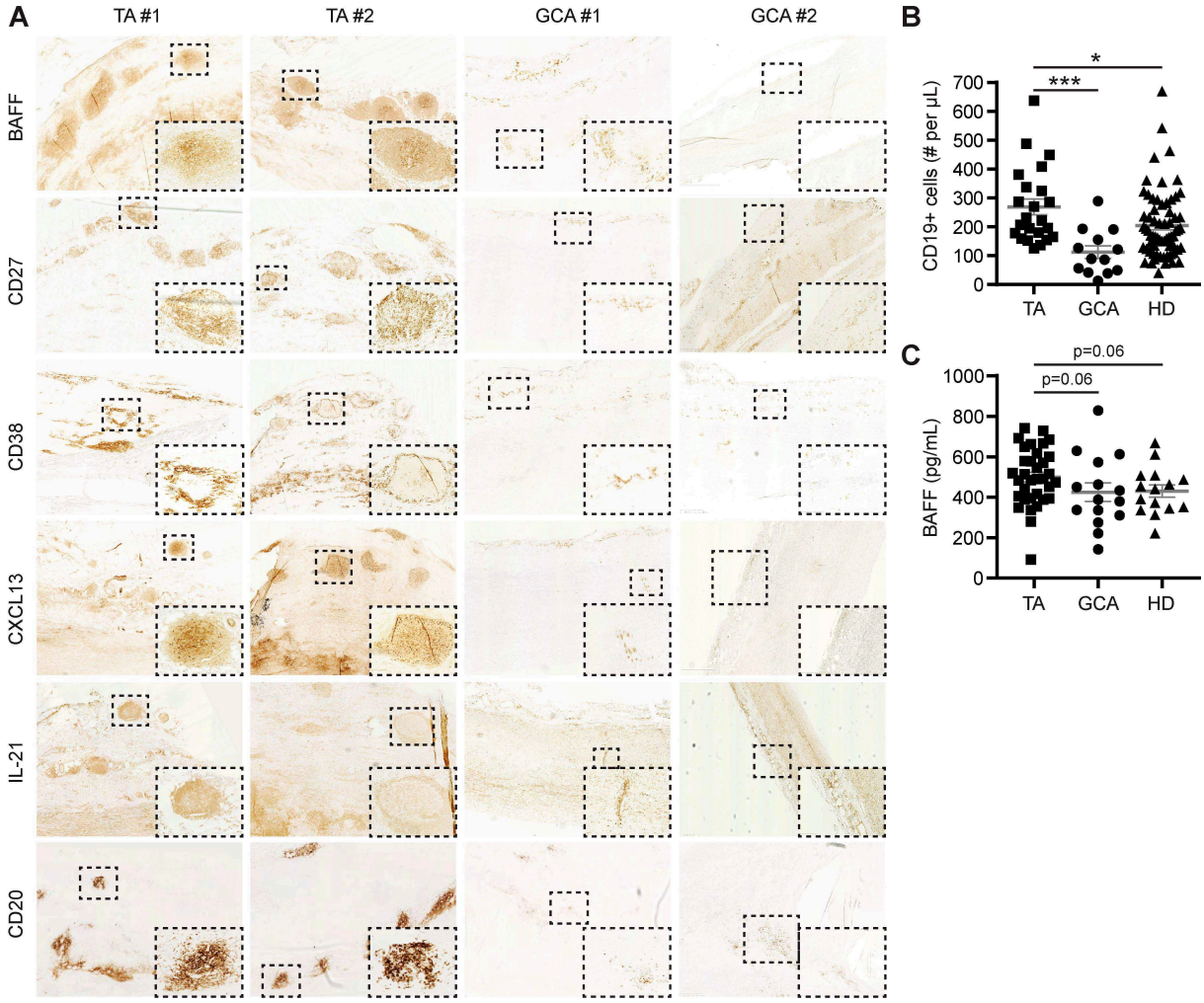
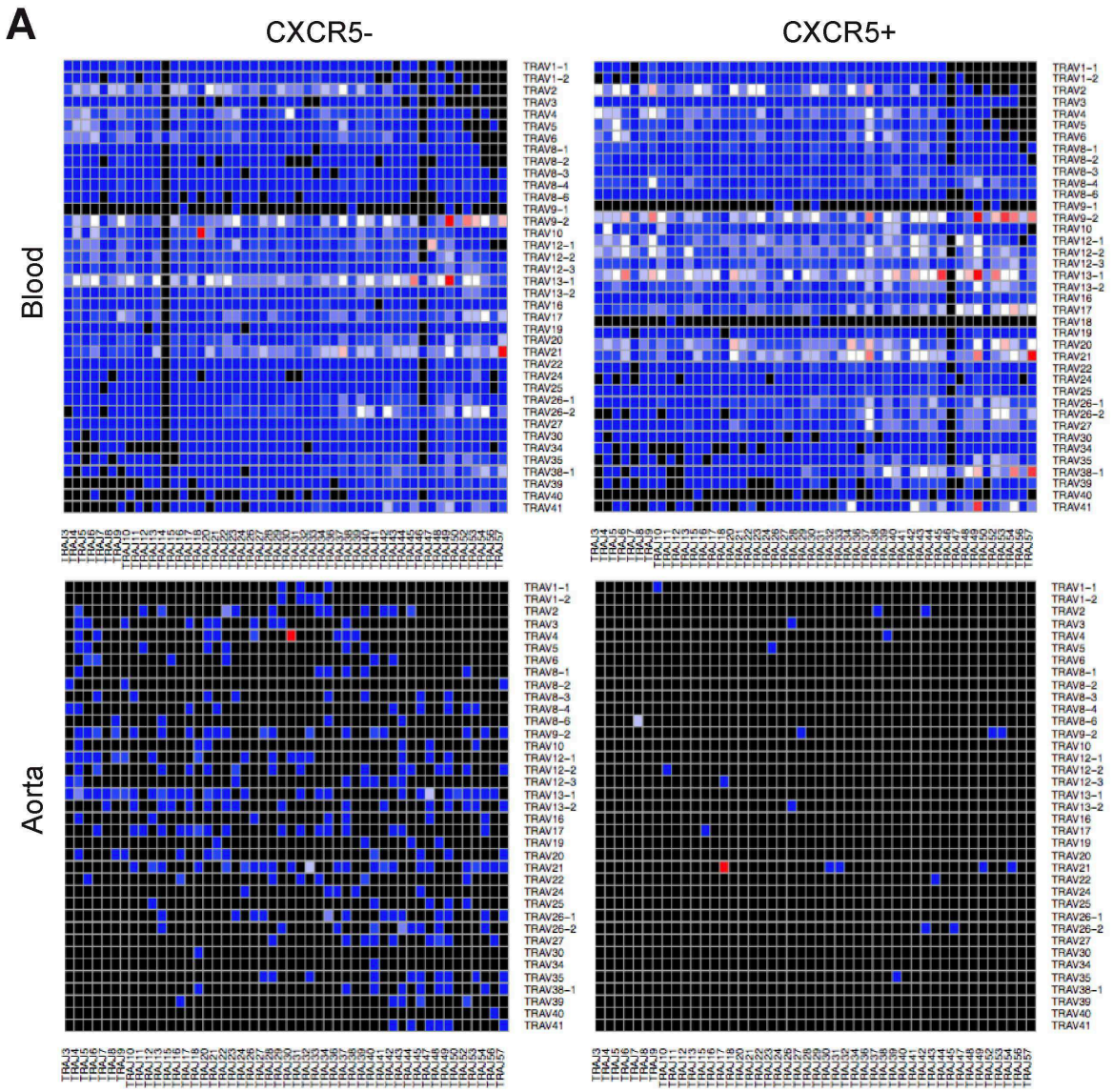


Figure S2: **A**. Expression of markers of B cells differentiation and TLO formation such as IL-21, CXCL13, CD27, CD38 and BAFF in TA and GCA. **B**. The absolute number of circulating CD19+ cells was significantly increased in TA compared to GCA patients and HD. *p<0.05 and ***p<0.001. **C**. Levels of BAFF in sera of TA tended to be higher compared to GCA patients.



B

Patient	Aortic clonotype	TCR gene	Blast sequence	Pathologies	References
1	AFNANTGNQFY	a	NSNTGNQFY	Sjören's syndrome	Joachims et al., 2016
1	ASSRTSGCYNEQF	b	ASSRTSGGHNEOF ASSRTSGVYNEQF ASSLTSGSYNEQF	Sjören's syndrome Calcific aortic stenosis Cardiac graft arteriosclerosis	Joachims et al., 2016 Winchester et al., 2011 Slachta et al., 2000
2	CAVIKAAGNKLTF	a	CAVVPIKAAGNKLTF	Sjören's syndrome	Joachims et al., 2016
2	CAVTSYGNNRLAF	a	CAVVGYGNNRLAF CA-THYGNNRLAF	Sjören's syndrome	Joachims et al., 2016 Vandevyver et al., 1995

Figure S3. **A**. We found a very broad repertoire of peripheral CD4+ T cells, whereas it switches to oligoclonality for CD4+ T cells originating from aorta TA lesions. This trend was dramatically much stronger for CD4+ CXCR5+ T cells than for CD4+ CXCR5- T cells in arteries of both TA patients. **B**. Motifs of the major aortic CD4+ CXCR5+ sequences have been previously reported in auto-immune/inflammatory diseases and in vascular diseases.

Table 1: Main characteristics of GCA and TA patients

Parameters	TA N=54	GCA n-=52
Demographic features		
Median age [IQR]	32.4 [27.2; 53.2]	74.7 [66.3; 83.2]
Female gender	43 (79.6%)	35 (67.4%)
Geographic origin		
Caucasian	20 (36.4%)	50 (96.2%)
African	15 (27.4%)	0 (0%)
North African	15 (27.4%)	2 (3.8%)
Other	5 (9%)	0 (0%)
Clinical features		
Numano Classification		
I	7 (14.3%)	NA
II	8 (16.3%)	NA
III	5 (10.2%)	NA
IV	1 (2%)	NA
V	27 (55%)	NA
Stroke	12 (22.2%)	7 (13.4%)
Aortic aneurysms	18 (33.3%)	4 (7.7%)
Aortitis	54 (100%)	14 (26.9%)
Optic neuritis	NA	5 (5.6%)
Mean C reactive Protein (mg/l) (±SD)	21.4 (±28)	45.3 (±48.3)

Supplementary Tables

Table S1: List of genes included in Tfh signature

Upregulated genes	Downregulated genes
CCR6	CXCR3
PDCD1	CCR7
ICOS	PTPRC
CD3E	CD8A
CD4	CD19
CXCR5	CD14
BCL6	IL1R2
IL6R	
IL21R	
IL21	
TNFRSF4	
CD40LG	
RORC	
IL17A	
IL22	
NFATC1	
IL2RB	
IL2RG	
IL1R1	

Table S2: List of primers used for quantification of gene expression

PD1-for	CGTCTGGGCGGTGCTACAA
PD1-rev	TGACACGGAAGCGGCAGT
CD20-for	CAACTGTGAACCAGTAATCCCT
CD20-rev	GCATCACTGACAAAATGCCCAAG
CXCR5-for	GCTAACGTCGGAAATGGA
CXCR5-rev	GCAGGGCAGAGATGATTT
CD45-for	GTATTTGTGGCTTAAACTCTTGGCAT
CD45-rev	TCCAGTGGGGGAAGGTGTTG

Supplementary Methods

Computation of signature scores

We used specific gene signatures to compute a signature score using an R custom algorithm for each patient involved in the dataset, in order to estimate a given sample orientation in terms of transcriptomic profile. For a given gene signature and patient, the final signature score is the sum of all subscores computed for individual genes. A subscore is defined by the current gene expression value pondered by a given linear weight, depending 1) on the gene direction (up-regulated or down-regulated) and 2) on the intensity of the patient's gene expression compared to all patients' expression of the same gene. In the case of an up-regulated gene, a patient gene expression value close to the maximum of the gene-matching distribution will lead to a weight close to 1, whereas an expression value close to the minimum of the gene-matching distribution will lead to a weight close to 0, and vice versa for a down-regulated gene. In the end, the final score reflects the likeliness of the signature to match the patient's transcriptomic profile: the higher the score is, the more likely the signature describes well the patient's transcriptomic profile. With this technique, we can either compare several signature scores in one patient group, or compare a given signature score between several patient groups.

Analysis of cytokine production in LVV patients

Peripheral blood mononuclear cells (PBMCs) were stimulated for 4 hours with 0,05 µg/mL Phorbol 12-myristate 13-acetate (PMA) and 1 mM (1µg/mL) ionomycin (Sigma-Aldrich). Culture supernatants were harvested and immediately frozen at -80°C. Quantitative determination of IL-17 and IL-6 was performed in culture supernatant using Human Cytokine 25-Plex (Invitrogen, Cergy Pontoise, France) in accordance with the manufacturer protocol.

Counting of B cells

Images where converted from Hamamatsu slide scanner native format ndpi to .tiff format to facilitate the workflow process under Fiji, an open source image processing toolbox. Most of the pre-established parameters of the (Hamamatsu) software were maintained for all the samples. Fiji was downloaded from http://fiji.sc/Fiji (US National Institutes of Health, Bethesda, MD, USA) and used for image segmentation and analysis. For reliable and

automated data analysis, we developed an efficient image analysis software solution as plugin to Fiji. Our plugin quantified the signal of the specific marked cells selectively and separately for the rest of the sample. First, the image brightness and contrast were changed to increase the contrast of the specific stained cells and then were separated from the background using a simple intensity threshold. Subsequently, our plugin generated a binary mask from the images of the respective marker after intensity thresholding. The binary masks were separately applied to the original channel and the specific marked regions were identified. Measurements of area and mean grey value were calculated and then processed under Microsoft Excel.

Valentin Quiniou – Thèse de doctorat d'Immunologie - 2020

Summary:

Specificity is the hallmark of the adaptive immune response. For T cells, specificity is mediated by T cell receptors (TCRs), which interact with peptides presented by major histocompatibility complex molecules through their complementary-determining-region-3 (CDR3). TCRs are formed by rearrangements between hundreds of gene segments at the alpha and beta loci, theoretically generating $>10^{19}$ sequences. TCR selection during thymopoiesis releases for each individual a repertoire of approximately 10^7 to 10^8 unique TCRs. How evolution shaped a process that selects TCRs that would efficiently respond to antigens that are not yet present is a central question of immunology. The paradigm is that the selection of a diverse enough repertoire of TCRs should always provide a proper specificity for any given need. Expansion of such rare T cells would provide enough fighters for an efficacious immune response and later enough antigen-experienced cells for memory. However, this theory is challenged by overlooked data questioning the essence of specificity, including robust experimental and epidemiological data highlighting heterologous immunity, i.e. exposure to one agent protecting against infection by another. We show here that, during human thymopoiesis, there is a marked enrichment of CDR3s with high generation probabilities, shared between individuals and specific for viral peptides. Furthermore, at the single cell level, unique alpha/beta TCRs to which these CDR3s belong can bind multiple peptides from different viruses. Our results support an evolutionary selection that favors polyspecific TCRs for antiviral responses and heterologous immunity.