BMC Genomics

RESEARCH Open Access

Identification and characterization of the T cell receptor (TCR) repertoire of the cynomolgus macaque (*Macaca Fascicularis*)

Swati Jaiswal[1], Sarah K. Nyquist[2,3,4,5,6,7], Shayla Boyce[1], Tasneem Jivanjee[2,3,4], Samira Ibrahim[2,3,4], Joshua D. Bromley[2,3,4,8], G. James Gatter[2,3,4], Hannah Gideon[9], Kush Patel[9], Sharie Keanne Ganchua[9], Bonnie Berger[5,6], Sarah M. Fortune[10,11], JoAnne L. Flynn[9], Alex K. Shalek[2,3,4,10,11,12] and Samuel M. Behar[1*]

Abstract

Background: Cynomolgus macaque (*Macaca fascicularis*) is an attractive animal model for the study of human disease and is extensively used in biomedical research. Cynomolgus macaques share behavioral, physiological, and genomic traits with humans and recapitulate human disease manifestations not observed in other animal species. To improve the use of the cynomolgus macaque model to investigate immune responses, we defined and characterized the T cell receptor (TCR) repertoire.

Result: We identified and analyzed the alpha (TRA), beta (TRB), gamma (TRG), and delta (TRD) TCR loci of the cynomolgus macaque. The expressed repertoire was determined using 22 unique lung samples from *Mycobacterium tuberculosis* infected cynomolgus macaques by single cell RNA sequencing. Expressed TCR alpha (TRAV) and beta (TRBV) variable region genes were enriched and identified using gene specific primers, which allowed their functional status to be determined. Analysis of the primers used for cynomolgus macaque TCR variable region gene enrichment showed they could also be used to amplify rhesus macaque (*M. mulatta*) variable region genes.

Conclusion: The genomic organization of the cynomolgus macaque has great similarity with the rhesus macaque and they shared > 90% sequence similarity with the human TCR repertoire. The identification of the TCR repertoire facilitates analysis of T cell immunity in cynomolgus macaques.

Keywords: T cell receptor, Cynomolgus macaque, Locus map, NHP, Variable region genes

Background

Experimental animal models are an essential tool in our pursuit of understanding human physiology. The mouse has been incredibly useful in elucidating the major concepts of immunology, including defining the genetic and molecular basis of immunoglobulin and TCR formation and diversity. As part of this effort, the murine TCR repertoire have been extensively characterized and its knowledge is being used to develop new approaches to facilitate antigen discovery and novel treatments for human disease. However, it is not surprising that many human diseases are inadequately modelled in mice. This has been repeatedly emphasized for cancer and is also true for many infectious diseases. Two important examples are acquired immunodeficiency syndrome (AIDS), which is caused by the Human Immunodeficiency Virus-1 (HIV-1), and COVID-19, which is caused by the SARS-CoV2 coronavirus [1–5]. Mice are naturally resistant to both infections. For HIV research, the field largely

*Correspondence: samuel.behar@umassmed.edu

[1] Department of Microbiology and Physiological Systems, Universityof Massachusetts Chan Medical School, Worcester, MA, USA
Full list of author information is available at the end of the article

turned to nonhuman primates (NHP) as a better alternative because they could be infected with a highly related virus, Simian Immunodeficiency Virus (SIV). Consequently, the rhesus macaque's TCR locus was among the first NHP TCR locus to be characterized [6]. Cynomolgus macaques have been increasingly used for biomedical research, especially in the fields of neurology, cardiology, and for drug development [7, 8]. Importantly, they are increasingly used for infectious disease research, including as a model for human HIV [9] and SARS-CoV2 infection [5]. Most NHP species, including rhesus macaques, whether in captivity or in the wild, rapidly succumb to *Mycobacterium tuberculosis* infection [10, 11]. However, Flynn's group finds that following challenge with low dose *M. tuberculosis*, nearly half of infected cynomolgus macaques develop a form of disease that resembles latent TB in people [12–15]. Indeed, the pathology observed among *M. tuberculosis*-infected cynomolgus macaques *recapitulates* the spectrum of human TB pathology [16]. Thus, the cynomolgus macaque is providing insights into human disease not possible with other small animal models.

The tremendous capacity of T cells to recognize diverse antigens has a genetic basis that is inherent in the genomic organization of the T cell receptor (TCR) loci [17]. TCR repertoire diversity arises through genetic mechanisms that minimize the number of genetic elements encoded by the genome while maximizing the potential breadth of expressed TCRs. The germline configuration of TCR genes is not functional. Instead, the TCR loci encode families of variable (V), diversity (D), and joining (J) segments, which undergo rearrangement early during T cell development [17]. Recombination of V, D, and J segments leads to a gene fragment that encodes the V-region domain, which becomes the N-terminus of the TCR protein and determines its antigen specificity. Downstream of the V, D, and J genes are constant (C) region exons, which encode the C-terminus of all TCRs and couples the TCR to the Cluster of differentiation 3 (CD3) protein complex to mediate signal transduction into the T cell. The primary diversity of TCRs arises from the nearly random rearrangement of V, D, and J gene segments, as well as additional diversity that occurs at the V-D and D-J junctions by imprecise recombination and the insertion of non-germline encoded nucleotides (N-regions). TCRs are heterodimers formed by TCRα and TCRβ chains, which are encoded by distinct loci (TRA and TRB, respectively) [18]. The TCRα is encoded when Vα and Jα gene segments recombine; the TCRβ is formed from the recombination of Vβ, D and Jβ gene segments. Additional diversity is created by the random pairing of the TCRα and TCRβ chains. Unlike immunoglobulin genes, somatic mutation does not occur in TCR genes. The potential TCR repertoire varies between animal species and is driven in large part by the number of functional members of V, D, or J segments. In humans, there is the potential to generate 10^{15} unique TCRs.

A second subset of T cells are known as gamma-delta (γδ) T cells, express an alternative TCR, which is encoded by distinct gene segments found in the TRG and TRD loci. The γδ-TCR is structurally similar to the αβ-TCR. Like the TRA and TRB loci, the TRG and TRD loci contain sets of Vγ and Jγ, and Vδ, Dδ and Jδ gene segments, respectively. In general, there are fewer gene segments in the TRG and TRD loci, although the potential diversity is still great because of longer CDR3 regions [19]. γδ T cells remain enigmatic because the antigens they recognize and the antigen presenting molecules that restrict their recognition of antigen are incompletely characterized. Nevertheless, they are identified in the circulation and in the tissues of all mammals, and play important roles in autoimmune disease, and in immunity to infection and cancer [20, 21].

Here we identified the TRA, TRB, TRG and TRD loci of the cynomolgus macaque. Based initially on the homology with human TCR gene segments, and subsequently using the identified gene segments from rhesus macaque and cynomolgus macaque, we systematically identified all the V, D, J, and C gene segments belonging to all four T cell receptor loci. Finally, using the genomic sequences, we designed specific primers for the amplification of the Vα and Vβ regions, and determined which of the V gene segments are expressed in individual subjects. To validate our annotations, we investigated the expressed TCR repertoire in cynomolgus macaques infected with *Mycobacterium tuberculosis*. To minimize the possibility of active infection skewing the TCR repertoire, only samples taken from lung areas where there was no active inflammation or gross infection (i.e., uninvolved lung tissue), were used in the present study. The TCR V-regions used by T cells located in uninvolved regions of lung tissue were analyzed by single cell RNA sequencing. These data will allow the detailed analysis of the T cell responses in cynomolgus macaques as well as comparative immunogenetics studies, comparing different species of macaques, and the evolution of TCR genes among primates.

Results

Identification of the Macaca fascicularis (Macfas) TCR loci

The Macfas genome assembly Macaca_fascicularis_5.0 (GCF_000364345.1) was used to annotate the different TCR loci. Later, we also used the Assembly MFA1912RKSv2 assembly [22]. Based on nucleic acid sequence homology with the human Cα, Cδ, Cβ, and Cγ gene segments, the TRA and TRD loci were identified on Chr.7, and TRB and TRG loci were identified on

Chr.3 (Fig. 1). Subsequently, each human V, D, J, and C gene segment was used to blast the Macfas Chr.7 and 3, to identify homologous gene segments. Similarly, *Macaca mulatta* (Macmul) gene segments were also used to identify homologous genes unique to the macaca genus. Using this approach, we were able to annotate and assemble a map of the Macfas TRA, TRB, TRG, and TRD loci as described in detail below (Fig. 1).

The Macfas TRA locus

The structure of the Macfas TRA locus is like the human locus in that it overlaps the TRD locus on Chromosome 7 (Fig. 1A) [23]. We identified 64 TRAV genes in Macfas genome, three more than the 61 human genes but less than the 67 Macmul genes. The two human gene families TRAV7 and TRAV28, each contain a single member and are absent from the Macfas and Macmul TRA locus (Table 1, Fig. 2). Conversely, the TRA loci of Macfas and Macmul have additional genes in the TRAV11, TRAV22, TRAV23, TRAV24, TRAV25, and TRAV26 families. The greater number of Macmul TRAV genes compared to Macfas results from an expansion of the TRAV22 and TRAV23 families (Table 1). Of the 64 Macfas TRAV genes, 15 are pseudogenes and 2 are

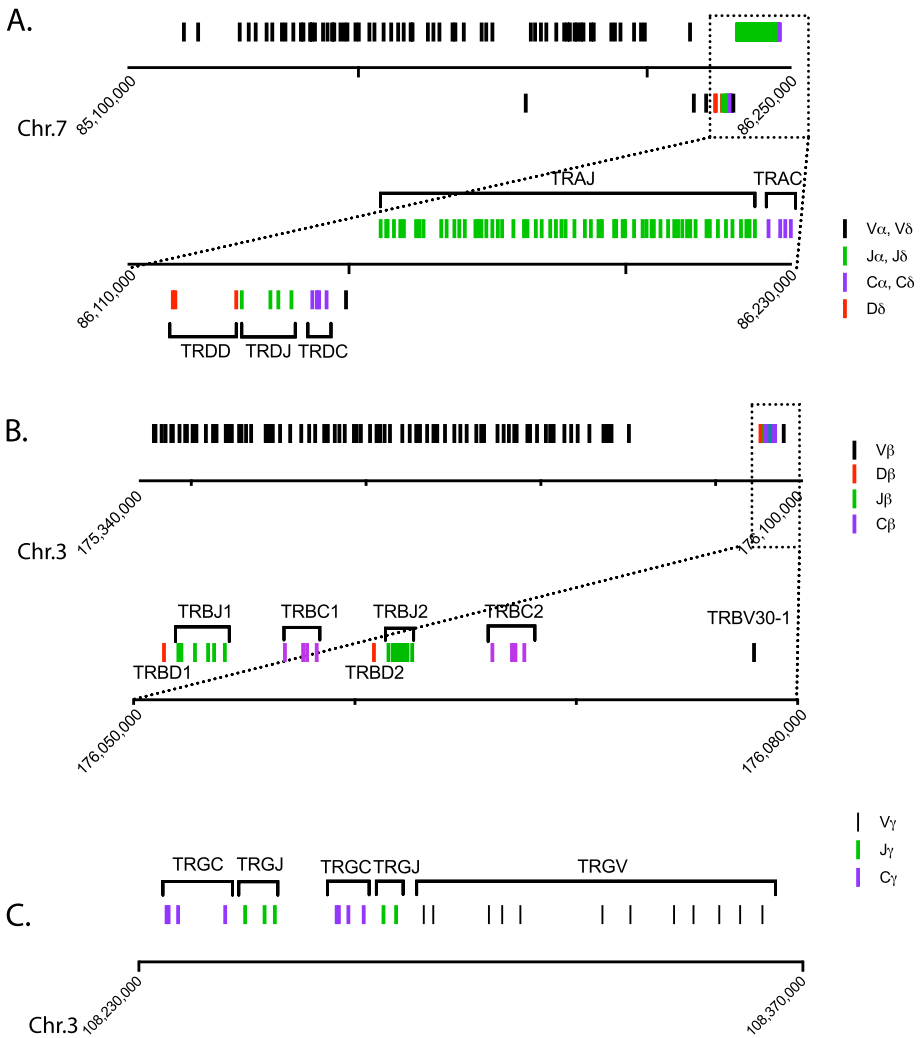

Fig. 1 The Macfas TCR loci. Structure of the TCR loci (**A**) TRA/TRD (**B**) TRB and (**C**) TRG loci based on Macaca_fascicularis_5.0. **A** The TRA and TRD loci are interspersed on Chr. 7. The genes above the x-axis belong to the TRA locus; those below the axis belong to the TRD locus. The boxed region is expanded to show greater detail. **B** The TRB locus is located on Chr. 3. The boxed region is expanded to show greater detail. **C** The TRG locus is located on Chr. 3. Each line represents a gene and the distance between them is proportional to their spacing on Chr.7 and Chr.3. The blue boxes represent the 3′ region of the AMPH gene and exon 10 of STARD3NL, which are boundaries of the TRG locus. The black lines represent V gene segments; green lines are J gene segments; purple lines represent C region exons, and the red lines represent D gene segments

Table 1 Comparison of Macfas, Macmul and human TRAV, TRBV, TRDV, and TRGV genes

TRAV gene segments			
Subgroup	Macfas	Macmul	Homsap
TRAV1	2	2	2
TRAV2	1	1	1
TRAV3	1	1	1
TRAV4	1	1	1
TRAV5	1	1	1
TRAV6	1	1	1
TRAV7	0	0	1
TRAV8	7	7	8
TRAV9	2	2	2
TRAV10	1	1	1
TRAV11	3	3	2
TRAV12	3	3	3
TRAV13	2	2	2
TRAV14	2	2	2
TRAV15	1	1	1
TRAV16	1	1	1
TRAV17	1	1	1
TRAV18	1	1	1
TRAV19	1	1	1
TRAV20	1	1	1
TRAV21	1	1	1
TRAV22	2	3	1
TRAV23	2	4	1
TRAV24	2	2	1
TRAV25	2	2	1
TRAV26	3	3	2
TRAV27	1	1	1
TRAV28	0	0	1
TRAV29	1	1	1
TRAV30	1	1	1
TRAV31	1	1	1
TRAV32	1	1	1
TRAV33	1	1	1
TRAV34	1	1	1
TRAV35	1	1	1
TRAV36	1	1	1
TRAV37	1	1	1
TRAV38	2	2	2
TRAV39	1	1	1
TRAV40	1	1	1
TRAV41	1	1	1
TRAV46	1	1	1
TRAVA	1	1	1
TRAVB	1	1	1
TRAVC	1	1	1
Total	**64**	**67**	**61**

TRBV gene segments			
Subgroup	Macfas	Macmul	Homsap
TRBV1	3	3	1
TRBV2	3	3	1
TRBV3	4	4	2
TRBV4	3	3	3
TRBV5	10	10	8
TRBV6	10	10	9
TRBV7	11	11	9
TRBV8	2	2	2
TRBV9	1	1	1
TRBV10	3	3	3
TRBV11	3	3	3
TRBV12	4	4	5
TRBV13	1	1	1
TRBV14	1	1	1
TRBV15	1	1	1
TRBV16	1	1	1
TRBV17	1	0	1
TRBV18	1	1	1
TRBV19	1	1	1
TRBV20	1	1	1
TRBV21	1	1	1
TRBV22	1	1	1
TRBV23	1	1	1
TRBV24	1	1	1
TRBV25	1	1	1
TRBV26	1	1	1
TRBV27	1	1	1
TRBV28	1	1	1
TRBV29	1	1	1
TRBV30	1	1	1
TRBVA	1	1	1
TRBVB	1	1	1
TRBVC	1	1	1
Total	**78**	**77**	**68**

TRGV gene segments				
Subgroup	GENE	Macfas	Macmul	Homsap
	TRGV1	1	ORF	ORF
	TRGV2	1	1	1
	TRGV3	1	1	1
	TRGV4	0	0***	1
TRGV1	TRGV5P	0	NR	P
	TRGV5	0	NR	1
	TRGV6	P	P	P
	TRGV7	0	NR	P
	TRGV8	1	1***	1
TRGV2	TRGV9	1	1	1
TRGV3	TRGV10	1	ORF	ORF
TRGV4	TRGV11	P	P	ORF
TRGVA	TRGVA	P	P	P
TRGVB	TRGVB	P	P	P
TRGVC	TRGVC	1	P	NR
TRGVD	TRGVD	1	P	NR
Total		**12**	**12**	**14**

TRDV gene segments				
Subgroup	GENE	Macfas	Macmul	Homsap
TRDV1	TRDV1	1	1	1
TRDV1	TRDV1-1	1	1	0
TRDV2	TRDV2	1	1	1
TRDV3	TRDV3	1	1	1
TRDV4	TRDV4	1	1	0
Total		**5**	**5**	**3**

P	pseudogene
ORF	open reading frame
NR	not reported
0	no homologous gene identified
***	nomenclature discrepancy

The numerical value for every subgroup represents the number of genes identified

ORF Open reading frame, *NR* Not reported, *P* Pseudogene, *F* Functional

*** Nomenclature discrepancy

possible pseudogenes (Table 1, Table S1). There might have been a duplication of a section of the TRA locus. A stretch of six genes (TRAV22, TRAV23, TRDV1, TRAV24, TRAV25, and TRAV26) is repeated, and differentiates the human TRAV locus from the macaque locus (Fig. 2B). The sequences of the affected TRAV genes are not identical, indicating continued evolution over time. It is unknown whether other NHP have such duplications. Second, there are three additional TRAV genes in the Macmul genome assembly that are absent in the Macfas genome. These are Macmul TRAV22-2, TRAV23-2, or TRAV23-3 (Fig. 2B). In searching two different assemblies, we found that six Macmul homologs are missing from the Macfas 5.0 assembly, and four genes are missing from the MFA1912RKSv2 assembly. As both assemblies contain multiple gaps in the TRA/TRD loci, the difference in the number of V-genes in the Macfas and Macmul TRA/TRD locus is likely to be a consequence of limitations in the genome assemblies. A difference in the genomic structure of the Macfas and Macmul TRA/TRD cannot be ruled out but based on the high degree of conservation at the gene level, we believe that such a scenario is unlikely. We identified 61 TRAJ genes, which is the same number as rhesus macaque and human TRAJ genes. There is a high degree of conservation between Macfas and *Homo sapiens* (human) TRAJ gene segments (Table S2). Finally, we compared the TRAC exons from all three species. The Macfas and Macmul TRAC genes have identical amino acid sequences (Figure S1).

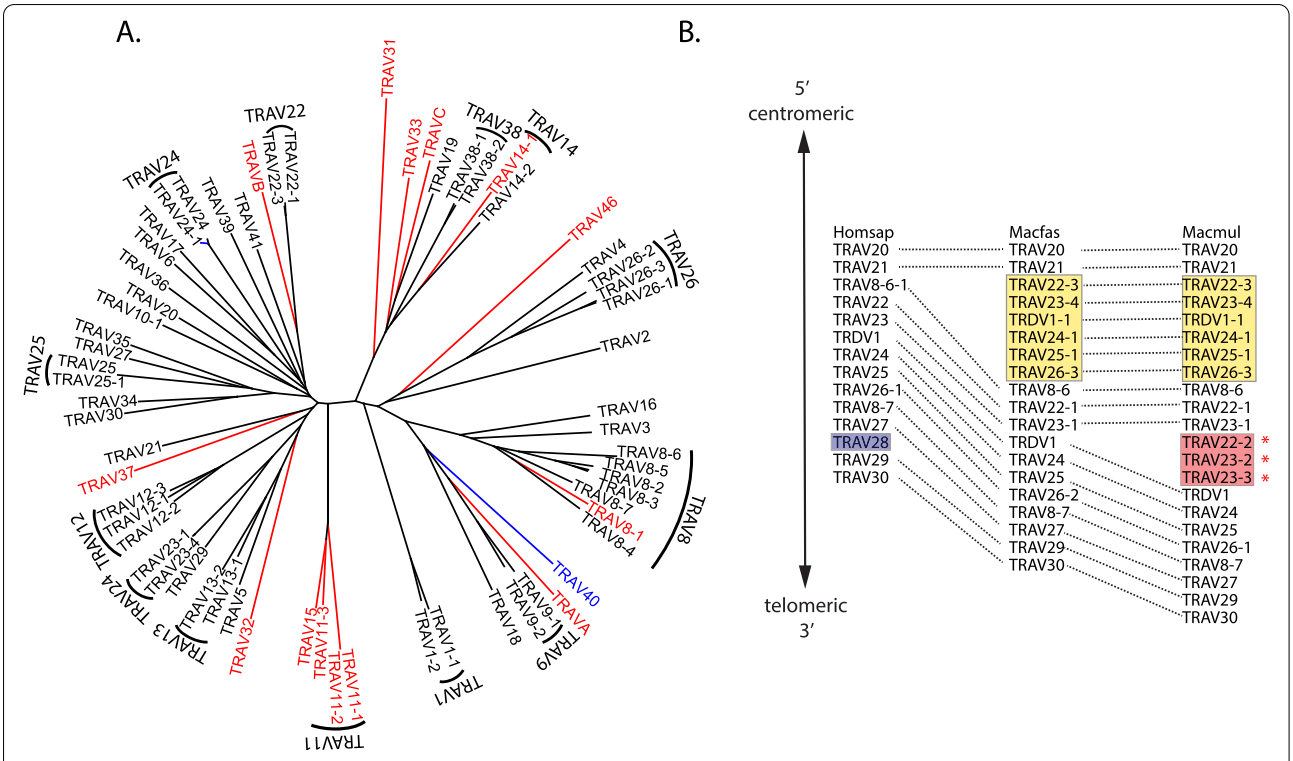

Fig. 2 TRAV families. Phylogenetic tree illustrating (**A**) functional genes (black), pseudogenes (red) and pseudogenes in some haplotypes (blue) of the Macfas TRAV locus. The genes clustered together belong to the same family. **B** Comparison of the TRAV/TRDV locus of human, Macfas, and Macmul. The genes that are exclusive to humans are highlighted in purple. Those TRAV genes found in Macfas and Macmul but not in human are in yellow, and the genes are present only in Macmul but absent in Macmul are in red. *, The absence of Macfas orthologs of TRAV22-2, TRAV-23–2, and TRAV23-3, might be a consequence of gaps in the Macfas genome assembly and should not be construed as reflecting evolutionary differences

The Macfas TRB locus

The Macfas TRB locus (Fig. 1B) is similar in structure to the Macmul TRB locus. We identified 78 TRBV genes, compared to 77 annotated Macmul TRBV genes (Table 1 & Table S3). Both are expanded compared to the human species, for which there exists 68 distinct genes. The overall TRBV family structure is similar, with some variation in the number of members and the number of pseudogenes ($n = 17$) and possible pseudogenes ($n = 8$) (Table 1, Fig. 3). The organization of the TRBJ and TRBC genes is similar in all three species, characterized by a duplication of the TRBJ and TRBC genes (Fig. 1B). Comparing the Macfas and Macmul TRBJ gene segments, four (including the TRBJ2.2P ORF) differ by a single nucleotide; the other 10 genes are 100% conserved (Figure S2, Table S4). The TRBD1 and TRBD2 are also 100% conserved between Macfas and Macmul (Table S4). Similarly, there is a high degree of conservation between Macfas and human TRBJ gene segments (Figure S2). Finally, we compared the TRBC exons from all three species. As noted, there are two TRBC genes, TRBC1 and TRBC2, which are 97% identical. The Macfas and Macmul TRBC1 differ by only two bp and the translated sequence is 100% identical; for TRBC2, there is a single amino acid difference (Figure S1).

The Macfas TRG locus

The Macfas TRG locus is located on chromosome 3 (Fig. 1C). We identified 12 TRGV genes of which 6 are predicted to be functional and an additional 4 are pseudogenes (Fig. 4, Table 1, Table S5). These genes were compared to the homologous genes in human and rhesus (Fig. 4). The same 12 genes were found in the Macmul TRG locus. In general, the Macfas and Macmul orthologs had between 0–2 mismatches (i.e., > 99% identity), while the homology between Macfas and human TRGV genes was 88–95%. The two NHP species lacked TRGV4, TRGV5, TRGV5P, and TRGV7, and Macmul had two additional V genes, TRGVC and TRGVD. The human TRG locus has two clusters of J segments and C-region genes [23, 24]; IMGT/LIGM-DB: IMGT000011 (582,960 bp), human (Homo sapiens) TRG locus), and the Macfas and Macmul loci have a similar structure (IMGT/LIGM-DB: IMGT000059 (197,016 bp), rhesus monkey (Macaca mulatta) TRG locus). The five Macfas TRGJ

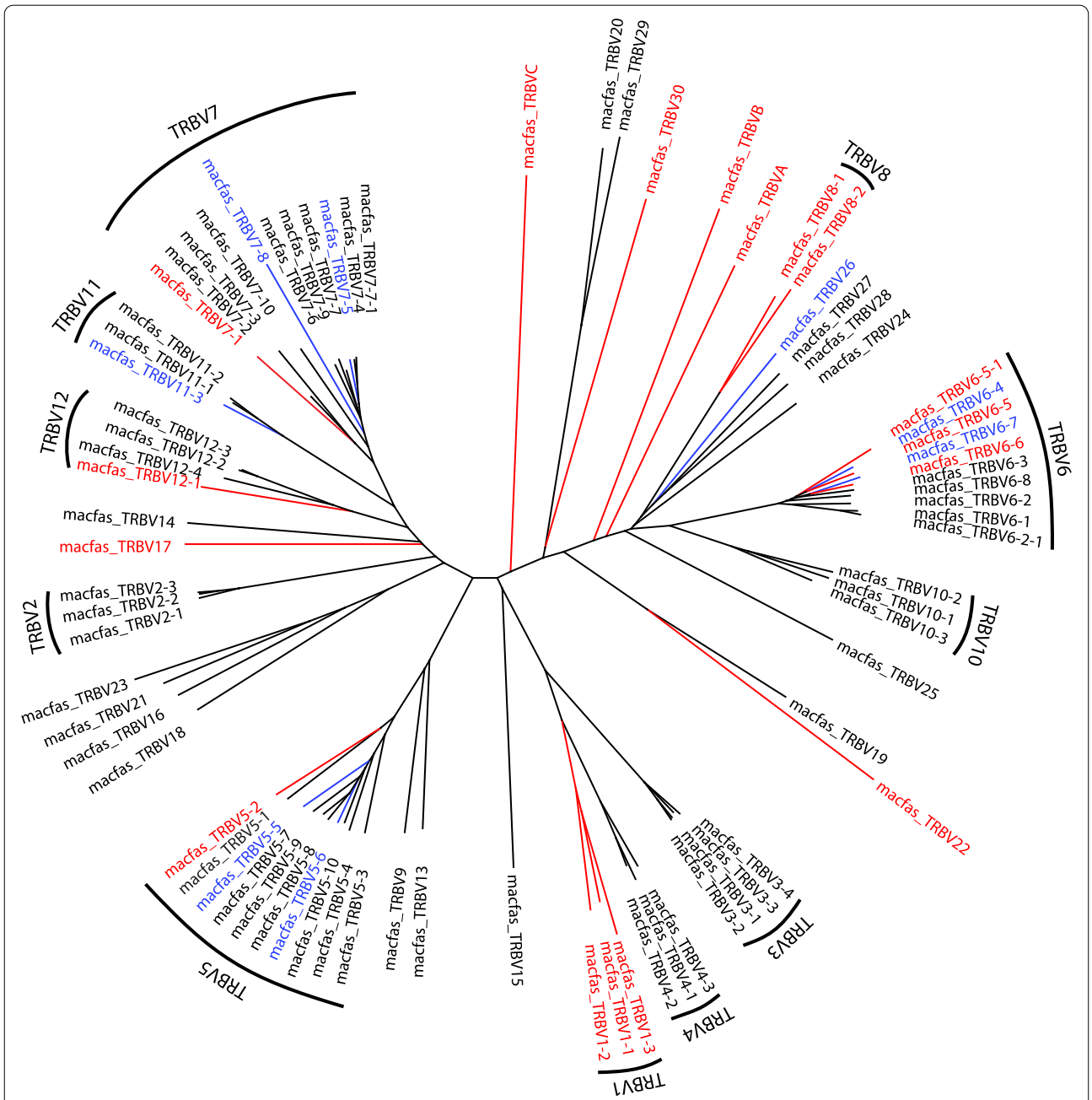

Fig. 3 TRBV families. Phylogenetic tree illustrating functional genes (black), pseudogenes (red) and pseudogenes in some haplotypes (blue) of the Macfas TRBV locus. The genes clustered together belongs to the same family

gene segments are very similar to their Macmul counterparts, with between 0–1 bp differences (Fig. 4B). Similarly, there are two Macfas TRGC regions, each encoded by three exons (Table S5). These are highly similar to their Macmul orthologs. Comparing Macfas and Macmul TRGC2 exon 1, 2, and 3, there are 1, 0 and 2 mismatches, respectively, with an overall amino acid sequence identify of 96.5%.

The Macfas TRD locus

The Macfas TRD locus is located on chromosome 7 and overlaps with the TRA locus (Fig. 1A). Three canonical TRDV genes were identified as Macfas homologs of human TRDV1, TRDV2, and TRDV3, with homologies between 91–97% (Fig. 5, Table S6). The two macaque species have an additional gene, TRDV1-1, which is very homologous to TRDV1 (Figs. 2B and 5). We named the Macfas TRDV1-1

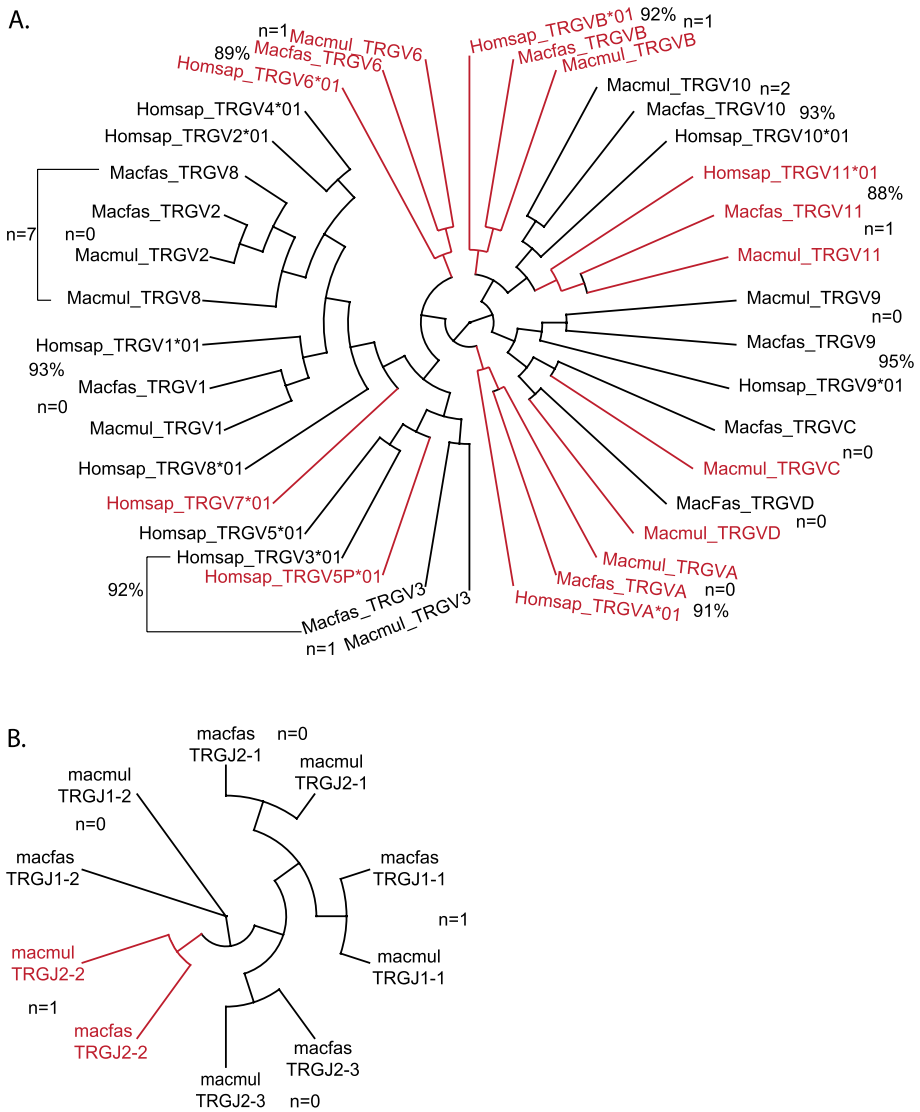

Fig. 4 TRGV and TRGJ gene segment homologies. **A** Phylogenetic tree illustrating functional genes (black) and pseudogenes (red) of the Macfas TRGV locus. The number (i.e., "*n* = 1") is the number of mismatches between the Macfas and Macmul genes. The % is the identity between the Macfas and the human gene. Homologies between other genes of interest are indicated with a dotted line. **B** Phylogenetic tree clustering Macfas and Macmul TRGJ genes

based on its orthologous location although its sequence homology is more similar to Macfas TRDV1. A fifth gene, TRDV4, was identified which was 100% homologous to Macmul TRDV4, for which no human ortholog was identified. Three TRDD and four TRDJ Macfas gene segments were identified, as in the human genome (Table S6). These genes are 100% identical to their Macmul homologs (Fig. 5B). Similarly, the single Macfas TRDC region has 100% DNA sequence identity and predicted amino acid sequence as the Macmul TRDC (Figure S1). There is a two amino acid gap, which we suggest is a consequence of the artificial splicing between exons 2 and exon 3.

The expressed V gene repertoire used by cynomolgus macaque T cells

We determined the expressed TRA and TRB repertoire in cynomolgus macaques infected with *Mycobacterium tuberculosis* by single cell RNA sequencing, The TCR V-regions were amplified using primers as described [25]. Our evaluation of the primers finds that they can be used for analysis of TCRs from rhesus macaque as well (Tables 2 and 3). To determine the functionality of the TRAV and TRBV gene segments we identified, the following criteria were used: (i) Defined L1 exon and L2-V exon, (ii) absence of nonsense or missense mutation, and

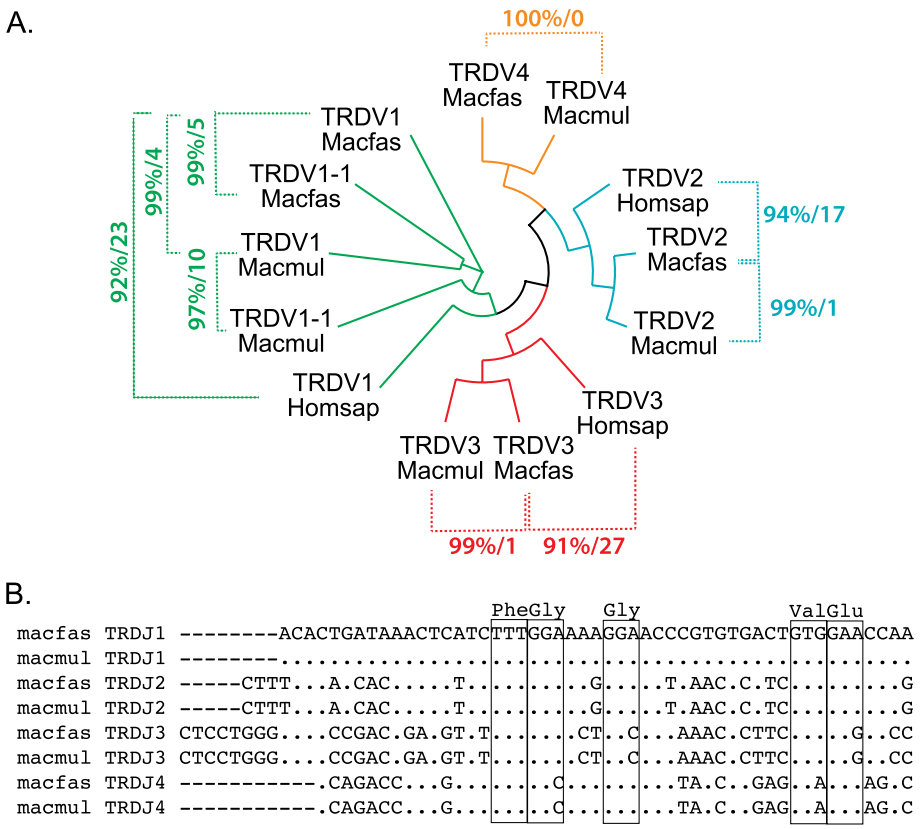

Fig. 5 TRDV and TRDJ gene segment homologies. **A** Phylogenetic tree showing the functional genes human, Macfas and Macmul TRDV genes. Comparisons are indicated with dotted lines and the percent identity is indicated followed by the number of sequence mismatches. Each TRDV gene family is color coded. **B** Alignment of Macmul and Macfas TRDJ showing the conserved amino acids (boxed)

(iii) encodes a cytosine (C) at position 21–23 followed by tryptophan (W) at position 31–33 of the V exon. The terminal amino acids encoded by a functional TRAV gene is usually CAVR, CAL, or CAF. Similarly, the terminal amino acids encoded by a functional TRBV gene is usually CASSQ, CASSL, or CASSE. Based on these criteria, we initially assigned each V gene to be functional if it met these criteria. If the gene had an internal stop codon, or lacked the conserved C or W residue, it was deemed a pseudogene. Finally, if the gene appeared to be functional, but the L1 or L2 parts of the leader sequence could not be identified, or it lacked consensus splice site for intron A, we designated it an open reading frame (ORF) (Tables S1, S3 and S5).

To determine the expressed TRAV and TRBV repertoire, cells obtained from the uninvolved lung tissue of 22 cynomolgus macaques infected with *Mycobacterium tuberculosis* was analyzed by single cell TCR sequencing. The expressed TRAV (Fig. 6A) and TRBV (Fig. 6B) repertoire was determined for each individual macaque. The percentage of individuals that expressed each gene was also calculated. These data allow assignment of each TRAV and TRBV gene as a functional gene or a pseudogene. Overall, there was a good correlation between genes that were predicted to be pseudogenes (based on premature stop codons) and the lack of representation in the transcribed repertoire. However, there were exceptions. For example, TRBV6-4 was predicted to be a pseudogene but was highly represented in the transcribed repertoire. The expected stop codon at position 85 (TAG) was CAG in the transcribed gene, and thus, encoded a functional glutamine (Q). This difference between the germline and the transcribed gene could be the result of a polymorphism or a sequencing error in the genomic reference sequence. Several other genes had similar behavior and were designated as being functional. The status of V genes designated as ORFs, was changed to 'functional' if the V gene was transcribed, or to 'pseudogene' if it was not. To determine whether the macfas homologs of TRAV22-2, TRAV23-2, and TRAV23-3, which are missing from the genome assemblies, were used by T cells, we included the sequences of the macmul V gene orthologs in the reference database. The algorithm did not assign any TCRs to the missing genes.

Table 2 Enrichment primers for TRAV in Macfas and Macmul. The code, name, and sequence of the primers are from [25]. For purposes of this paper, the sequence of each primer is divided into two regions: (i) the 5 'handle (in red) which is common to all primers); and (ii) the TRAV-gene specific sequence (in blue). The last column shows the specificity of the primer. Bolded TRAV-genes are specific for macfas; TRAV genes in italics are specific for macmul. P, pseudogene; ORF, open reading frame

CODE	NAME	Oligonucleotide sequence (common/TRAV specific)		Macfas & Macmul TRAV (**Macfas only**; *Macmul only*)
A01	macfas_TRA_1	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGT CTG TGA CCC AGC TTG ACA GCC A	TRAV8-2 | TRAV8-3 | TRAV8-5
A02	macfas_TRA_2	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAA GAA CAG CTC CCT GCA CAT CAC A	TRAV25 | TRAV25-1
A03	macfas_TRA_3	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAA GGA TCC CAG CCT GCA GAC TCA G	TRAV24-1 | TRAV24
A04	macfas_TRA_4	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAA TAC TTT CAA GAA CTG CTT GGA A	**TRAV11-1** | TRAV11-2 ORF
A05	macfas_TRA_5	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAA TCC TTC AGT CTC AAG ATC TCA G	TRAV38-1 | TRAV38-2
A06	macfas_TRA_6	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAC TGC ACT TAC AGC AAC AGT GCT T	TRAV12-1 | TRAV12-3
A07	macfas_TRA_7	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAG CCA CAT ACC GTA AAG AAA CCA C	**TRAV9-1** | TRAV9-2
A08	macfas_TRA_8	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGC TCC AGA TGA AAG ACT CTG CCT C	TRAV1-1 | TRAV1-2
A09	macfas_TRA_9	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	GAT GCT GCT GAG TAC TTC TGT GCT G	TRAV8-4 | TRAV8-7
A10	macfas_TRA_10	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CAG CAG CTG GAG CAG AGT CCT CGG T	TRAV27
A11	macfas_TRA_11	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	GAC TCA GCA ACG TAT TTC TGT GCA A	TRAV14-2
A12	macfas_TRA_12	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	TGG ACT CAG CAG TAT ACT TCT GTG C	TRAV19
B01	macfas_TRA_13	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	GTT CAA ATA TGG CTA AGA GGC AAG G	TRAV13-2
B02	macfas_TRA_14	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CAT ATG GAG CAG GGA GAC AAA CAT T	TRAV15 P
B03	macfas_TRA_15	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAT CTC TTC TGG TAT GTC CAG TAC C	TRAV16
B04	macfas_TRA_16	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CTG TAA GCC AGC ATA ACC ACC ATG T	TRAV8-1 P
B05	macfas_TRA_17	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	TGT AAC TCT CAA CTG CAG TTA TGA A	TRAV36
B06	macfas_TRA_18	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	GAA GCG CCC ATC TTC CTG ATG ATA T	TRAV30
B07	macfas_TRA_19	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAA GCG ACA TCA CAC TGA ACT G	TRAV37 ORF
B08	macfas_TRA_20	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGA ATG GCC TCT CTG ACA ATC ACT G	**TRAV26-2**
B09	macfas_TRA_21	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CAA ATG GAG CAG TGA AGC AGG AGG G	TRAV39
B10	macfas_TRA_22	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGA CAT TAT ATG GCT TAC ACT GGT A	TRAV34
B11	macfas_TRA_23	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	ATC AGG GTT ATT CTA AGT CAA ATG C	TRAV33 P
B12	macfas_TRA_24	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	GGA GAC AAC TTG GTT CTC AAC TGC A	TRAV21
C01	macfas_TRA_25	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CCT CTC TGA TCA TCA CAG AAG ACA G	TRAV26-1 | TRAV26-3
C02	macfas_TRA_26	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGT CCA GAA AGG AGG GAT TTC AAT T	TRAV23-1 | *TRAV23-2* | *TRAV23-3* | *TRAV23-4*
C03	macfas_TRA_27	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	ACT TCT TCA AGC ACA TTT AAC ACC T	TRAV35
C04	macfas_TRA_28	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CTG GCT GCA ACA GCA TCC AGG AGG A	TRAV41
C05	macfas_TRA_29	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAT GCT AAA CAT GTC TCC CTG CAT A	TRAV32 P
C06	macfas_TRA_30	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	ACA TAT ATT CTT TCA AAT ACG GAC C	**TRAV5**
C07	macfas_TRA_31	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AAA CCA TCT GCC CTT GTG AGC GAC T	TRAV3
C08	macfas_TRA_32	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	TCA ACA AGG AGA AGA GGA TCC TCA G	TRAV17
C09	macfas_TRA_33	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	ACA AGA CAG CCA AAC ATT TCT CTC T	TRAV13-1
C10	macfas_TRA_34	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	ATC ATG TGG TAC CAA CAG TTT CCC A	TRAV4
C11	macfas_TRA_35	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CCC TGA GAG GGC AGC TCT AAC ATT A	TRAV18
C12	macfas_TRA_36	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGC ACA GCT CGA TAG AGC CAG CCA G	**TRAV12-2**
D01	macfas_TRA_37	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	ACT CTT AAA CAG AGT TTG TTT CAT A	TRAV6
D02	macfas_TRA_38	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	TCT CAG CAG GGA CGA TAC AAC ATG A	TRAV2
D03	macfas_TRA_39	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CTG CTG AAG GTC CTA CAT TCC TGA T	TRAV29
D04	macfas_TRA_40	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	TCA TGA CTT TCA GTG AGA ACA CAA A	TRAV10
D05	macfas_TRA_41	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CAC ATT TCC TCC TCC CAG ACC ACA G	**TRAV22-1** | *TRAV22-3*
D06	macfas_TRA_42	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	CAG GTA TCA GAC TCA GCC GTG TAC T	TRAV40
D07	macfas_TRA_43	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGG TGA CGC AGA GTC CTG AGG CCC T	TRAV20
D08	macfas_TRA_44	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	AGA GTC AAT ATT CAG TGA GCT TCC A	**TRAV31 P**
D09	macfas_TRAV11_3	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	GTT TGG AAT ATC CCG ACC TCT CAT C	TRAV11-3 P
D10	macfas_TRAV14-1	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	TCT GGT ACA AGT AGC CCA GGA GTG G	TRAV14-1 P
D11	macfas_TRAV46	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	TCC TTA AGC ATG TGC AGC CAA GAG A	TRAV46 P
D12	macfas_TRAV8-6	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G	GAA CGA CAC GCT GAG TAC TTC TGT G	TRAV8-6 P

Table 3 Enrichment primers for TRBV in Macfas and Macmul. The code, name, and sequence of the primers are from [25]. For purposes of this paper, the sequence of each primer is divided into two regions: (i) the 5 ' handle (in red) which is common to all primers); and (ii) the TRBV-gene specific sequence (in blue). The last column shows the specificity of the primer. Bolded TRBV-genes are specific for macfas. P, pseudogene; ORF, open reading frame

CODE	NAME	Oligonucleotide sequence (common/TRBV specific)		Macfas and Macmul TRBV-GENE (**Macfas only**)
A01	macfas_TRB_1	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CAA ACA GAA AGC AAA GAT GGA T	TRBV21
A02	macfas_TRB_2	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CAA AGA TGG ATT GTA CCC CCG A	TRBV23 | TRDV4
A03	macfas_TRB_3	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CAA GAG GAC AGC AAG TGA CAC T	**TRBV5-6 P**
A04	macfas_TRB_4	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CAG CCA CTC TGC AAT GCT ATC C	TRBV13
A05	macfas_TRB_5	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CAT GCA CTG AGT CAA CAC AGT C	TRBV6 P
A06	macfas_TRB_6	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CCC CAA AGC ACC TGA TCA CAG C	TRBV9
A07	macfas_TRB_7	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CCG ATA TTC AGA GAG GAG ACC T	TRBV22 P
A08	macfas_TRB_8	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA CCT GAG CAT CTG GGA CAT GAT T	TRBV1-1 P| TRBV1-2 P
A09	macfas_TRB_9	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GAA GAG AAA GGA TGT AGC TCT C	TRBV7-1 P
A10	macfas_TRB_10	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GAA GAT CAC TCT GGA ATG TTC T	TRBV25
A11	macfas_TRB_11	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GAA TGT TCT GGT ATC GAC AAG A	TRBV28
A12	macfas_TRB_12	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GAG TTC AAG TTC TTG ATT TCC T	TRBV16
B01	macfas_TRB_13	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GAT TTT AAC AAT GAA GCA GAC A	TRBV15
B02	macfas_TRB_14	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GCC AAG CAG GGA TGT CTG TCA A	TRBV29
B03	macfas_TRB_15	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GGA AGG ACA GAA TGT GAC CCT G	TRBV19
B04	macfas_TRB_16	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GGA ATT CAT TCT ACA GTG TTC C	TRBV26 P
B05	macfas_TRB_17	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GGA CAC AGT CAT GTT TAT TGG T	TRBV18
B06	macfas_TRB_18	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA GGA GTT CTC AGG ATA TGT GCC A	TRBVA P
B07	macfas_TRB_19	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA TCA GTC TCA GCC CCA TGC CTG G	TRBV6-8 ORF
B08	macfas_TRB_20	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA TCA GAG GAC AGC AAG CGA CAT T	TRBV5-1
B09	macfas_TRB_21	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA TGG GAC AAG AAG TAA CTA TCA G	TRBV12-4
B10	macfas_TRB_22	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA TGG GAC AAG CAG TGA CTC TCA G	TRBV12-1 ORF
B11	macfas_TRB_23	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA TTC CAG GTC CTG AAG ACA GCA C	TRBV6-2 | TRBV6-4 | TRBV6-6 | TRBV6-7 | TRBV6-8
B12	macfas_TRB_24	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAA TTT CTG ATA TAC TTT CTG AGA G	TRBV14
C01	macfas_TRB_25	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAC AAC AAG TCT CCG ATA GAT GAT T	TRBV12-2 | TRBV12-3
C02	macfas_TRB_26	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAC ACC AAT GGA GTA GAA GAG CAG C	TRBV30
C03	macfas_TRB_27	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAC CTC TGT GAA GAT CGA GTG CCG T	TRBV20
C04	macfas_TRB_28	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAC CTC TGT GAA GAT CGA GTG CCG G	TRBV2-1 | TRBV2-2 | TRBV2-3
C05	macfas_TRB_29	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAG AAA TTG CTG AAG ATA ATG TTT A	TRBV3-1 | TRBV3-2 | TRBV3-3 | TRBV3-4
C06	macfas_TRB_30	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAG AAG TCT TTG AAA TGT GAA CAA C	TRBV4-1 | TRBV4-2 | TRBV4-3
C07	macfas_TRB_31	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AAT CAC CCA GAG CCC AAG ACA CAA G	TRBV10-1 | TRBV10-3
C08	macfas_TRB_32	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	ACC TTG ACG TGT CAC CAG ACT TGG A	TRBV10-2 | TRBV10-3
C09	macfas_TRB_33	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AGA TGA TTC ACA GTT GCC TAA GGA T	TRBV11-1 | TRBV11-2 | TRBV11-3
C10	macfas_TRB_34	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AGA TGC TCT CCT ATC TCT GGG CAC A	TRBV5-3 | TRBV5-4 | TRBV5-5 | TRBV5-6 | TRBV5-7 | TRBV5-8 | TRBV5-9 | TRBV5-10
C11	macfas_TRB_35	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AGG GCC CAG AGT TTC TGA CTT ACT T	TRBV7-7-1 | TRBV7-9 | TRBV7-10
C12	macfas_TRB_36	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	AGG TGT GAT CCA ATT TCG GGT CAT G	TRBV7-4 | TRBV7-5 | TRBV7-6 | TRBV7-7 | TRBV7-8 | TRBV7-9 | **(TRBV7-7-1)**
D01	macfas_TRB_37	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	ATG TAC TGG TAT CGA CAA GAC CCA G	TRBV6-1 | TRBV6-2 | TRBV6-2-1 | TRBV6-3 | TRBV6-5 | TRBV6-7 | TRBV24
D02	macfas_TRB_38	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	ATG TCC TGG TAT CGA CAA GAC CCA G	TRBV6-6 | TRBV27
D03	macfas_TRB_39	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	CAG GTG TGA TCC AAT TTC AGG TCA T	TRBV7-2 | TRBV7-3
D04	macfas_TRB_40	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	CCA GAC AAA TCA GGG CTG CCC AGT G	TRBV7-2 | TRBV7-5 | TRBV7-8
D05	macfas_TRB_41	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	CTG GTA TCG ACA AGA CCC AGG CAT G	TRBV6-1 | TRBV6-2 | TRBV6-2-1 | TRBV6-3 | TRBV6-5 | TRBV6-6 | TRBV6-7 | TRBV6-8
D06	macfas_TRB_42	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	TTA GGA ACC ATT ATT CAA TGT TCT G	TRBV8-1 P | TRBV8-2 P
D07	macfas_TRBV1_3	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	CTC TCG CTT ATG CCT TCA TGT GCT C	TRBV1-3 P
D08	macfas_TRBV17_1	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	GCC ATA TAT CTC CGC AGT AGC GAT G	**TRBV17 P**
D09	macfas_TRBV_6_5_1	TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GAG	TCC CCA ATG GCT ACA ATG TCT TTA C	**TRBV6-5-1 P**

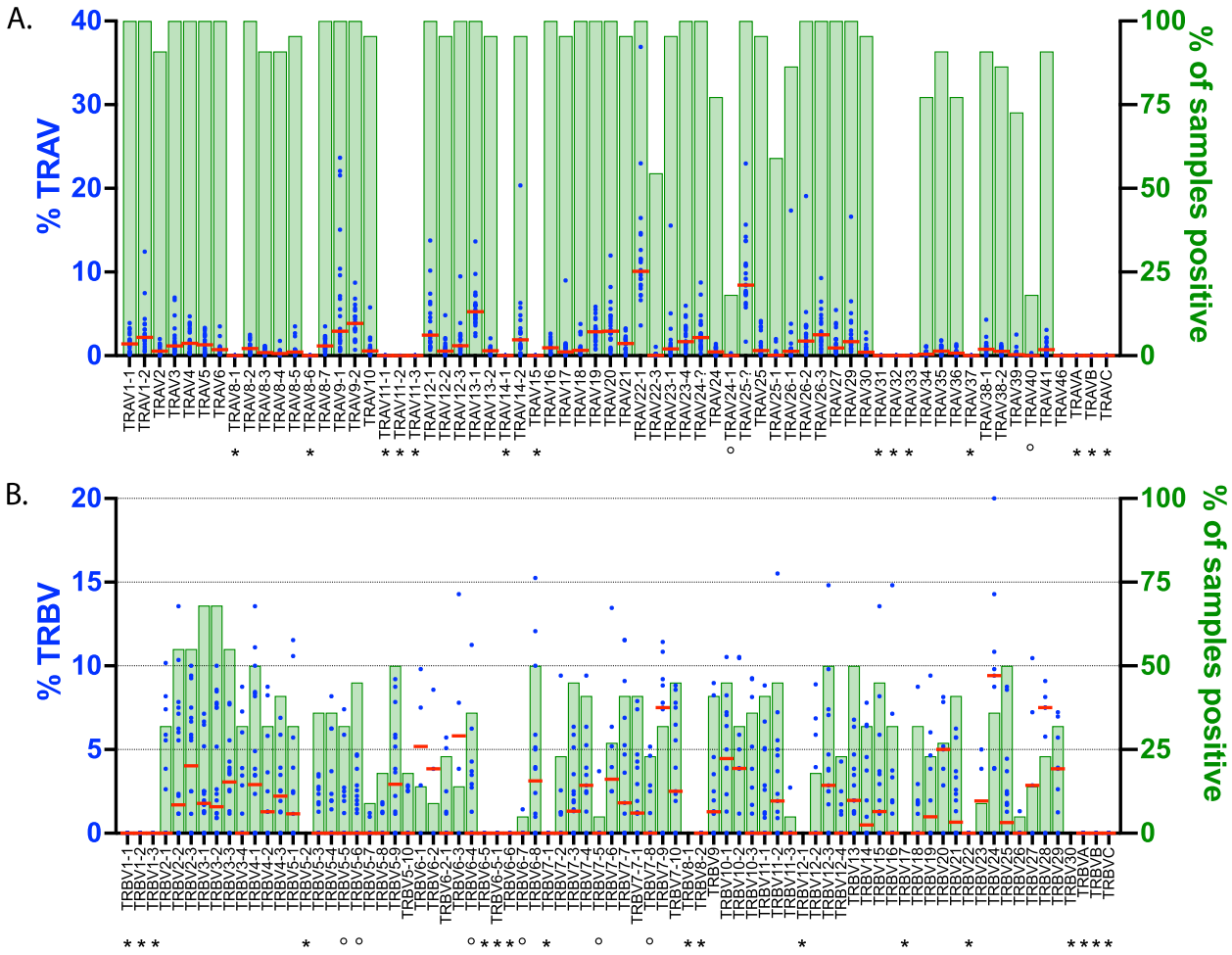

Fig. 6 The expressed TRAV and TRBV repertoire. Single cell analysis of lung mononuclear cells from cynomolgus macaques reveals their functionally expressed TRAV and TRBV repertoire. Each dot represents a different subject ($n = 22$). All samples are from uninvolved lung tissue (i.e., uninfected tissue as observed at autopsy). TRAVs were determined for 22 subjects with a median of 968 cells (interquartile range 496–1833). TRBVs were determined for 21 subjects, with a median of 855 cells (interquartile range 375–1587). The distribution of TRAV (**A**) and TRBV (**B**) V regions segments used by the T cells from each individual was calculated. The values within averaged. Red bar, median (left axis). The number of individuals expressing a given TRAV or TRBV gene was also calculated (right axis). *, pseudogene; °, pseudogene in some haplotypes

Discussion

The nucleic acid sequence of recombined V, D, and J gene segments encodes the protein structure of the TCR and contains immunological information about T cell responses. The complementarity determining region 3 (CDR3), defined as the V-D-J or V-J recombination site, is unique to each unique T cell clone, sometimes referred to as a clonotype. Analytical approaches are beginning to predict the antigen specificity based on the primary sequence of the TCR. In the absence of the antigen specificity, the TCR sequence can be used as a surrogate of antigen specificity. As T cells undergo clonal expansion after encountering antigens, TCR sequences are being used to track T cells, monitor immune responses, and identify new antigens for human tumors and pathogens [26–28]. Advances in T cell therapy are being driven by our ability to clone and recombinantly express TCRs, as exemplified by adoptive cell therapy (ACT) [29, 30]. Thus, defining the V, D, and J gene segments is an important step in the analysis of T cell immunity.

We identified and annotated the TRA, TRB, TRG, and TRD loci of the cynomolgus macaque. There is generally more than 90% identity between the different V, D, and J gene segments in the human, rhesus and cynomolgus macaque's TCR repertoire. We also find that there is expansion of TCR beta locus of Macfas and Macmul compared to human. These differences, which are likely to have occurred by gene duplication [31, 32], may have occurred

in response to changes in selective pressure during evolution of the TCR loci [33, 34]. As one might expect, the structure of the different TCR loci is highly conserved between rhesus and cynomolgus macaque. The genomic differences we detected (e.g., Fig. 2A) are more likely to be due to ascertainment bias arising from problems with genomic sequencing and assembly, than true evolutionary events. The TCR conservation between cynomolgus and rhesus macaques can be leveraged in the analysis of the expressed TCR repertoire. We used a set of TRA and TRB primers to enrich expressed TCR genes from *M. tuberculosis* infected cynomolgus macaques. Our analysis of these primers shows them to be suitable for enrichment of TRA and TRB genes from rhesus macaques, and therefore this set of primers can be used for both species of macaques [25]. Similarly, nested primers for rhesus TCR enrichment using the 10X Genomics platform can also be used for cynomolgus macaques as the regions to which they anneal are 100% conserved [35, 36].

Conclusions

We identified and annotated the TRA/D, TRB and TRG loci of the cynomolgus macaque. The TRA and TRB genomic sequences were used to design primers, and as reference sequences, to amplify and identify TCR sequences expressed by single cells from the lungs of cynomolgus macaques. By using these data to analyze the αβTCRs expressed by mature T cells, we were able to discern which V genes were functional based on their RNA expression. This allowed us to refine and validate our predictions based on the genomic sequences. Altogether, these data show the utility of these TCR reference sequences, and we expect that they will be useful for the study of T cell immunity in cynomolgus macaques.

Methods

Source of genomic sequence

The genome of the cynomolgus macaque (NCBI: taxid 9541), also known as the crab eating macaque, has been sequenced and we used assembly Macaca_fascicularis_5.0 (GenBank assembly accession: GCA_000364345.1) [37, 38]. The RefSeq numbers for Chromosomes 3 and 7 are NC_052257.1 and NC_052261.1, respectively. Additional gene sequences were obtained from Assembly MFA1912RKSv2 for Macaca fascicularis (crab-eating macaque) (GenBank assembly accession: GCA_012559485.3) [22]. The formal genus and species name is *Macaca fascicularis*, which we abbreviate as Macfas. The rhesus macaque (i.e., *Macaca mulatta*; Macmul) TCR sequences were obtained initially from the literature [39] and later from IMGT (http://www.imgt.org) [40]. The human (i.e., *Homo sapiens*) TCR sequences were obtained from IMGT. In cases where more than one allele was available, the first allele was used for sequence comparisons.

Annotation and analysis of Macfas TCR repertoire

To identify the location of the Macfas TCR loci, the human TRAC, TRBC, TRGC, and TRDC were blasted against the Macfas genome. Subsequently, all human gene segments were individually blasted against the Macfas genome. As Macfas gene segments were defined, they were also used to look for other homologous genes. At the beginning of this study, the sequences of the Macmul TCBV genes were available and were used to look for homologous genes [39]. The names of the genes were assigned based on the homology with the human genes, and the location in the genome. The leader sequence (L1 & L2), TRV region, D region and J chain were identified for each gene. The annotation was done by following standard IMGT rules (http://www.imgt.org). Clustal Omega was used for multiple sequence alignments (https://www.ebi.ac.uk/Tools/msa/clustalo/) [41] and visualized using Archaeopteryx for Figs. 2, 3, 4 and 5 [42]. Sequences were entered and tracked in Snap Gene (version 5.0).

Expressed TCR repertoire of cynomolgus macaques

Cells from bronchoalveolar lavage (BAL), single cell suspensions of lung, or lung tissue, were obtained from cynomolgus macaques infected with *Mycobacterium tuberculosis* and single cell RNAseq libraries were created [43]. Primers were synthesized that were specific for the different TRAV, TRBV, TRAC, and TRBC gene segments based on the genomic sequences described herein and used to enrich and amplify the TCR sequences from T cells in scRNA-Seq libraries generated using 3' barcoded Seq-Well [25, 44]. Primers were not designed for pseudogenes that had internal stop codons, or for some V genes that were not initially identified. The libraries were sequenced and then aligned to the TCR reference sequences. The samples were analyzed for 48 TRAV and 73 TRBV genes. The V region and J region sequences were mapped using BOWTIE 2 as part of the TCRGO algorithm (https://github.com/ShalekLab/tcrgo/tree/master/tcrgo) [25]. Briefly, reads are aligned with the V and J regions in the reference TCR database, containing the sequences annotated in this report (see Results, below). Each read from a Seq-Well library includes nucleic acid tags that identify the cell of origin (cell barcode) and the transcript of origin (unique molecular identifier, UMI). Reads with matching cell barcode and UMI are merged, and a consensus V and J region mapping is determined based on sequence similarity identified among the majority of reads. A consensus CDR3 sequence is identified from reads with shared mappings.

Abbreviations
ACT: Adoptive cell therapy; AIDS: Acquired immunodeficiency syndrome; C: Constant; CD3: Cluster of differentiation 3; CDR3: Complementary determining region 3; D: Diversity; HIV-1: Human Immunodeficiency Virus-1; J: Joining; γδ: Gamma-delta; Macfas: *Macaca fascicularis*; Macmul: *Macaca mulatta*; *Mtb*: *Mycobacterium tuberculosis*; NHP: Non-human primates; ORF: Open Reading Frame; SIV: Simian Immunodeficiency Virus; TCR: T cell receptor; V: Variable.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08867-0.

Additional file 1: Figure S1. Constant region homology. Alignment of the amino acid sequence of the TCR constant regions, derived from the in silico splicing of the human, Macfas and Macmul TRAC, TRBC, TRGC, and TRDC exons. Dots represent identity. Amino acids are represented by the 1-letter code. X is undetermined.

Additional file 2: Figure S2. TRBJ gene segment homology. Alignment of the nucleic acid sequences of the human, Macfas and Macmul TRBJ genes. Dots represent identity.

Additional file 3: Table S1. Macfas TRAV genes. The table includes the genomic order of the Macfas TRAV genes, their genomic coordinates, NCBI accession number, the functional status of each TRAV gene, whether the L1, L2, and V regions were identified, and their amino acid and nucleic acid sequences. F, functional; P, pseudogene; ORF, open reading frame; *, expression not analyzed; I, identified; NI, not identified; N/A, not available. Amino acids are represented by the 1-letter code. X is undetermined; dots represent identity; *, stop codon.

Additional file 4: Table S2. Macfas TRAJ sequence. The table includes the genomic order of the Macfas TRAJ genes including their genomic coordinates, their NCBI accession number, the % identity to the nearest human homology, their length, and their nucleic acid sequence. The nucleic acid sequence is arranged to show the triplets encoding the conserved amino acid motif "F/W G X G".

Additional file 5: Table S3. Macfas TRBV genes. The table includes the genomic order of the Macfas TRBV genes, their genomic coordinates, NCBI accession number, the functional status of each TRBV gene, whether the L1, L2, and V regions were identified, and their amino acid and nucleic acid sequences. F, functional; P, pseudogene; ORF, open reading frame; *, expression not analyzed; I, identified; NI, not identified; N/A, not available. Amino acids are represented by the 1-letter code. X is undetermined; dots represent identity; *, stop codon. TRBV6-4 is a pseudogene in the genomic sequence but the expressed gene is function.

Additional file 6: Table S4. Macfas TRBJ and TRBD sequences. The different tabs of this spreadsheet list the Macfas TRBJ genes including their genomic coordinates, their NCBI accession number, the % identity to the nearest human homology, their length, and their nucleic acid sequence. The nucleic acid sequence is arranged to show the triplets encoding the conserved amino acid motif "F G X G". The nucleic acid sequence of Macfas TRBD1 and TRBD2 is shown and compared to their orthologs in Macmul and Homsap.

Additional file 7: Table S5. Macfas TRGV, TRGJ and TRGC sequences. The different tabs of this spreadsheet list the Macfas TRGV, TRGJ and TRGC genes, their genomic coordinates, NCBI accession number, the functional status of each TRBV gene, the L1, L2, and V region amino acid and nucleic acid sequences. F, functional; P, pseudogene; ORF, open reading frame; N/A, not available. Amino acids are represented by the 1-letter code. X is undetermined; dots represent identity; *, stop codon.

Additional file 8: Table S6. Macfas TRDV, TRDJ, TRDD and TRDC sequences. The different tabs of this spreadsheet list the TRDV, TRDJ, TRDD and TRDC genes, their genomic coordinates, NCBI accession number, the functional status of each TRBV gene, the L1, L2, and V region amino acid and nucleic acid sequences. F, functional; P, pseudogene; ORF, open reading frame; N/A, not available. Amino acids are represented by the 1-letter code. X is undetermined; dots represent identity; *, stop codon.

Acknowledgements
Roisin Floyd, Marc Wadsworth and Travis Hughes have performed original sequence libraries for depletion experiment that were helpful in analyzing the expressed repertoire, Jake Rosenburg and Andy Tu helped in the development of TCR pipeline and interpretation of results.

Third party material
All of the data herein are owned by the authors. Sequences were obtained from publicly available data bases (i.e., IMGT and NCBI) and no permissions are required.

Authors' contributions
Conceptualization, S.J. and S.M.B.; Methodology, S.J., S.K.N., T.J., S.I., J.B., J.G. and S.M.B.; Investigation, S.J., S.B., S.K.N, S.M.B.; NHP Sample facilitation, S.K.G., K.P., H.G.; Resource and Supervision, A.K.S. and B.B.; Formal analysis, S.J. and S.M.B.; Writing & Editing, S.J. and S.M.B.; Funding Acquisition, J.F., S.M.F., S.M.B. The author(s) read and approved the final manuscript.

Funding
This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93019C00071, and additional support from The Bill and Melinda Gates Foundation.

Availability of data and materials
All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Declarations

Ethics approval and consent to participate
Chinese cynomolgus macaques were purchased from Valley Biosystems (West Sacramento, CA). All experimental manipulations, protocols, and care of the animals were approved by the University of Pittsburgh School of Medicine Institutional Animal Care and Use Committee (IACUC). The protocol assurance number for our IACUC is D16-00118. Our specific protocol approval numbers for this project are 18124275 and IM-18124275–1. The IACUC adheres to national guidelines established in the Animal Welfare Act (7 U.S.C. Sections 2131—2159) and the Guide for the Care and Use of Laboratory Animals (8th Edition) as mandated by the U.S. Public Health Service Policy. Sedated animals were humanely euthanatized using sodium pentobarbital and phenytoin. All methods are reported in accordance with ARRIVE guidelines (for complete details about the animals used in this study, see Gideon, Hughes, Tzounas, et al. [43].

Consent for publication
Not applicable.

Competing interests
A.K.S. reports compensation for consulting and/or SAB membership from Merck, Honeycomb Biotechnologies, Cellarity, Repertoire Immune Medicines, Ochre Bio, Third Rock Ventures, Hovione, Relation Therapeutics, FL82, and Dahlia Biosciences. All other authors report no conflict of interest.

Author details
[1]Department of Microbiology and Physiological Systems, Universityof Massachusetts Chan Medical School, Worcester, MA, USA. [2]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Microbiology Graduate Program, Massachusetts Institute of Technology, Cambridge, MA, USA. [9]Department of Microbiology and Molecular Genetics and Center for Vaccine Research, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [10]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [11]Ragon Institute of MGH, MIT and Harvard, Boston, MA, USA. [12]Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA.

Received: 22 April 2022 Accepted: 1 September 2022
Published online: 12 September 2022

References

1. Urano E, Okamura T, Ono C, Ueno S, Nagata S, Kamada H, Higuchi M, Furukawa M, Kamitani W, Matsuura Y, et al. COVID-19 cynomolgus macaque model reflecting human COVID-19 pathological conditions. Proc Natl Acad Sci U S A. 2021;118(43):e2104847118.
2. Shiina T, Suzuki S, Congy-Jolivet N, Aarninк A, Garchon HJ, Dereuddre-Bosquet N, Vaslin B, Tchitchek N, Desjardins D, Autran B, et al. Cynomolgus macaque IL37 polymorphism and control of SIV infection. Sci Rep. 2019;9(1):7981.
3. Rodgers MA, Ameel C, Ellis-Connell AL, Balgeman AJ, Maiello P, Barry GL, Friedrich TC, Klein E, O'Connor SL, Scanga CA. preexisting simian immunodeficiency virus infection increases susceptibility to tuberculosis in Mauritian Cynomolgus Macaques. Infect Immun. 2018;86(12):e00565.
4. Mohns MS, Greene JM, Cain BT, Pham NH, Gostick E, Price DA, O'Connor DH. Expansion of simian immunodeficiency virus (SIV)-specific CD8 T cell lines from SIV-Naive Mauritian Cynomolgus Macaques for adoptive transfer. J Virol. 2015;89(19):9748–57.
5. Salguero FJ, White AD, Slack GS, Fotheringham SA, Bewley KR, Gooch KE, Longet S, Humphries HE, Watson RJ, Hunter L, et al. Comparison of rhesus and cynomolgus macaques as an infection model for COVID-19. Nat Commun. 2021;12(1):1260.
6. Greenaway HY, Ng B, Price DA, Douek DC, Davenport MP, Venturi V. NKT and MAIT invariant TCRα sequences can be produced efficiently by VJ gene recombination. Immunobiology. 2013;218(2):213–24.
7. Carlsson HE, Schapiro SJ, Farah I, Hau J. Use of primates in research: a global overview. Am J Primatol. 2004;63(4):225–37.
8. Ebeling M, Kung E, See A, Broger C, Steiner G, Berrera M, Heckel T, Iniguez L, Albert T, Schmucki R, et al. Genome-based analysis of the nonhuman primate Macaca fascicularis as a model for drug safety assessment. Genome Res. 2011;21(10):1746–56.
9. Van Rompay KKA. Tackling HIV and AIDS: contributions by non-human primate models. Lab Anim (NY). 2017;46(6):259–70.
10. Matz-Rensing K, Hartmann T, Wendel GM, Frick JS, Homolka S, Richter E, Munk MH, Kaup FJ. Outbreak of tuberculosis in a colony of rhesus monkeys (Macaca mulatta) after possible indirect contact with a human TB patient. J Comp Pathol. 2015;153(2–3):81–91.
11. Sapolsky RM, Share LJ. A pacific culture among wild baboons: its emergence and transmission. PLoS Biol. 2004;2(4): e106.
12. Cadena AM, Hopkins FF, Maiello P, Carey AF, Wong EA, Martin CJ, Gideon HP, DiFazio RM, Andersen P, Lin PL, et al. Concurrent infection with Mycobacterium tuberculosis confers robust protection against secondary infection in macaques. PLoS Pathog. 2018;14(10): e1007305.
13. Maiello P, DiFazio RM, Cadena AM, Rodgers MA, Lin PL, Scanga CA, Flynn JL. Rhesus macaques are more susceptible to progressive tuberculosis than Cynomolgus macaques: a quantitative comparison. Infect Immun. 2018;86(2):e00505.
14. Flynn JL, Gideon HP, Mattila JT, Lin PL. Immunology studies in non-human primate models of tuberculosis. Immunol Rev. 2015;264(1):60–73.
15. Lin PL, Ford CB, Coleman MT, Myers AJ, Gawande R, Ioerger T, Sacchettini J, Fortune SM, Flynn JL. Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. Nat Med. 2014;20(1):75–9.
16. Scanga CA, Flynn JL. Modeling tuberculosis in nonhuman primates. Cold Spring Harb Perspect Med. 2014;4(12): a018564.
17. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, Olshen RA, Weyand CM, Boyd SD, Goronzy JJ. Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci U S A. 2014;111(36):13139–44.
18. Gouaillard C, Huchenq-Champagne A, Arnaud J, Chen CI CL, Rubin B. Evolution of T cell receptor (TCR) alpha beta heterodimer assembly with the CD3 complex. Eur J Immunol. 2001;31(12):3798–805.
19. Murphy K, Weaver C. Janeway's Immunobiology. 2017.
20. Kim DS, Lee KY, Yang WI, Han SJ, Hwang EH. Gamma/delta T lymphocytes in the BCG granulomatous lesions. Yonsei Med J. 1996;37(5):319–24.
21. Zhao Y, Niu C, Cui J. Gamma-delta (gammadelta) T cells: friend or foe in cancer development? J Transl Med. 2018;16(1):3.
22. Jayakumar V, Nishimura O, Kadota M, Hirose N, Sano H, Murakawa Y, Yamamoto Y, Nakaya M, Tsukiyama T, Seita Y, et al. Chromosomal-scale de novo genome assemblies of Cynomolgus macaque and common marmoset. Sci Data. 2021;8(1):159.
23. Lefranc MP, Lefranc G. The T Cell Receptor FactsBook. London: Academic Press; 2001.
24. Lefranc MP, Chuchana P, Dariavach P, Nguyen C, Huck S, Brockly F, Jordan B, Lefranc G. Molecular mapping of the human T cell receptor gamma (TRG) genes and linkage of the variable and constant regions. Eur J Immunol. 1989;19(6):989–94.
25. Jivanjee T, Ibrahim S, Nyquist SK, Gatter GJ, Bromley JD, Jaiswal S, Berger B, Behar SM, Love JC, Shalek AK: Enriching and Characterizing T-Cell Repertoires from 3' Barcoded Single-Cell Whole Transcriptome Amplification Products. 2022. https://doi.org/10.48550/arXiv220311266.
26. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol. 2020;38(10):1194–202.
27. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, et al. Identifying specificity groups in the T cell receptor repertoire. Nature. 2017;547:94.
28. Munson DJ, Egelston CA, Chiotti KE, Parra ZE, Bruno TC, Moore BL, Nakano TA, Simons DL, Jimenez G, Yim JH, et al. Identification of shared TCR sequences from T cells in human breast cancer using emulsion RT-PCR. Proc Natl Acad Sci U S A. 2016;113(29):8272–7.
29. Feldman SA, Assadipour Y, Kriley I, Goff SL, Rosenberg SA. Adoptive cell therapy–tumor-infiltrating lymphocytes, t-cell receptors, and chimeric antigen receptors. Semin Oncol. 2015;42(4):626–39.
30. Tang L, Zheng Y, Melo MB, Mabardi L, Castano AP, Xie YQ, Li N, Kudchodkar SB, Wong HC, Jeng EK, et al. Enhancing T cell therapy through TCR-signaling-responsive nanoparticle drug delivery. Nat Biotechnol. 2018;36(8):707–16.
31. Ohno S. Evolution by Gene Duplication. Heidelberg, Germany: Springer-Verlag; 1970.
32. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.
33. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. The evolution of mammalian gene families. PLoS ONE. 2006;1: e85.
34. Olson MV. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. 1999;64(1):18–23.
35. Walsh ES, Tollison TS, Brochu HN, Shaw BI, Diveley KR, Chou H, Law L, Kirk AD, Gale M Jr, Peng X. Single-cell-based high-throughput Ig and TCR repertoire sequencing analysis in rhesus macaques. J Immunol. 2022;208(3):762–71.
36. Abdulhaqq S, Ventura AB, Reed JS, Bashirova AA, Bateman KB, McDonald E, Wu HL, Greene JM, Schell JB, Morrow D, et al. Identification and Characterization of antigen-specific CD8(+) T cells using surface-trapped TNF-alpha and single-cell sequencing. J Immunol. 2021;207:2913.
37. Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol. 2011;29(11):1019–23.
38. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, McVeigh R, O'Neill K, Robbertse B, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020;2020:baaa062.
39. Greenaway HY, Kurniawan M, Price DA, Douek DC, Davenport MP, Venturi V. Extraction and characterization of the rhesus macaque T-cell receptor β-chain genes. Immunol Cell Biol. 2009;87(7):546–53.
40. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, et al. IMGT(R), the international ImMunoGeneTics information system (R) 25 years on. Nucleic Acids Res. 2015;43(Database issue):D413-422.
41. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47(W1):W636–41.
42. Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics. 2009;10:356.

43. Gideon HP, Hughes TK, Tzouanas CN, Wadsworth MH 2nd, Tu AA, Gierahn TM, Peters JM, Hopkins FF, Wei JR, Kummerlowe C, et al. Multimodal profiling of lung granulomas in macaques reveals cellular correlates of tuberculosis control. Immunity. 2022;55(5):827-846 E810.
44. Tu AA, Gierahn TM, Monian B, Morgan DM, Mehta NK, Ruiter B, Shreffler WG, Shalek AK, Love JC. TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures. Nat Immunol. 2019;20(12):1692–9.

Publisher's Note