High throughput single cell sequencing of both *T-cell-receptor-beta* alleles

Tomonori Hosoya[*,‡], Hongyang Li[†,‡], Chia-Jui Ku[*], Qingqing Wu[*], Yuanfang Guan[†] and James Douglas Engel[*,§]

[*]Department of Cell and Developmental Biology

[†]Department of Computational Medicine and Bioinformatics

University of Michigan

3035 BSRB

109 Zina Pitcher Place

Ann Arbor, Michigan 48109-2200

[‡]These authors contributed equally to this work.

§Corresponding author

James Douglas Engel

3035 BSRB, 109 Zina Pitcher Place

Ann Arbor, MI 48109

Email: engel@umich.edu

Telephone: 734-647-0803

Running title: Validated sequencing of both *Trb* alleles in single cells

ABSTRACT

Allelic exclusion is a vital mechanism for the generation of monospecificity to foreign antigens in B- and T-lymphocytes. Here we developed a high-throughput barcoded method to simultaneously analyze the VDJ recombination status of both mouse *T cell receptor beta* alleles in hundreds of single cells using Next Generation Sequencing.

INTRODUCTION

Vertebrates have evolved both innate and adaptive immune systems to protect individuals against infection, cancer and invasion by parasites. B and T lymphocytes comprise the central components of adaptive immunity and every individual lymphocyte harbors specific reactivity to a single antigen that is conferred by individually unique antigen receptors expressed on the cell surface. The diversity of antigen receptors is generated by DNA recombination, called VDJ rearrangement in jawed vertebrates(1). To maintain the required monospecificity of mature lymphocytes, only one of the two autosomal alleles is allowed to express a functionally rearranged beta- and alpha- chain T cell receptors (TCRβ and TCRα) in T cells, or immunoglobulin heavy and light chain receptor (IgH and IgL) in B cells: both are controlled by a historically opaque mechanism referred to as "allelic exclusion". Allelic exclusion occurs at the genetic level for IgH, IgL and TCRβ(2), while at the level of protein localization on the cell surface for TCRα(3). Loss of allelic exclusion results in dual-TCR expression, which can lead to autoimmunity(4, 5).

A high-throughput method to study the diversity of lymphocytes at the single-cell level has been reported recently(6), yet how one might analyze their mono-specificity remains unclear. Since the abundance of transcripts from a non-functional allele is lower than that of the functional allele(7, 8), traditional RNA-based methods cannot detect the mechanisms underlying allelic exclusion. One major challenge in analyzing the genes themselves is that there is only one copy of DNA representing each allele, unlike the existence of multiple transcribed RNA species. Therefore, extremely accurate DNA sequencing is required to avoid erroneously mis-assigning nucleotide-level mutations,

52 which would alter the interpretation of TCR locus activity. Additionally, no reference
53 genome is available for merging multiple sequencing reads, due to the many millions of
54 possible sequences that can be generated from VDJ rearrangement(9). Thus, *de novo*
55 sequence assembly or a long-read approach is required to retrieve the original genomic
56 sequence of both alleles in single cells. To significantly improve the traditional Sanger
57 sequencing method employed by us and others(10-12), we developed a high-throughput
58 method that enables analysis of *Trb* allelic exclusion status by sequencing both alleles of
59 the genome in single cells, enabling us to determine whether each allele in those cells
60 underwent either no, unproductive or productive rearrangement.
61
62

63 **MATERIALS AND METHODS**

64 **Single cell isolation.** Staged thymocytes were first isolated from mice (C57BL/6J,
65 Jackson Laboratories) between 5 and 8 weeks old using a cell sorter (BD FACSAria III).
66 $Lin^-CD4^-CD8a^-$ $Thy1.2^+cKit^-CD25^+CD28^-$ (DN3a stage), $Lin^-CD3^-TCR\beta^-$ $CD8a^-$
67 $Thy1.2^+cKit^-CD25^-$ (DN4 stage), $CD4^+CD8^+$ (DP stage) and $TCR\beta^+CD4^+CD8^+$ (late DP
68 stage) thymocytes (10,000 to 100,000 cells at each stage) were isolated as described
69 previously(10). Next, single cells were directly sorted into 20 μl of lysis buffer
70 [containing 1x Q5® Reaction Buffer (NEB), 4 μg Proteinase K and 0.1% Triton X100] in
71 one well of a 96 well PCR plate using a Synergy cell sorter (Sony iCyt SY3200). The cell
72 sorter setting was carefully aligned so that sorted cells were precisely deposited into the
73 center of each well. Sorted single cells in the lysis solution were kept on ice, and then
74 digested at 55ºC for 60 minutes followed by 95ºC for 15 minutes (to inactivate Proteinase
75 K) using a PCR thermal cycler within 6 hours after sorting.
76
77 **Multiplex nested PCR.** For the first round of PCR, primers were selected to amplify all
78 potential VDJ rearrangement at the *Trb* locus: 31 V region primers covering all 35 *Trbv*
79 genes(13), 2 D region primers, 2 J region primers and 2 control primers used to detect
80 sequences 3' of the *Actb* gene (Figs. 1a and 1b). The full list of primers used are
81 deposited at (https://umich.box.com/s/3f5q64u2i68dn2i7hlneucxph9oetpvb). Since this
82 method was designed to recover not only rearranged but also germ line configured
83 genomic DNA, the two J primers were designed to be 3' of J1-7 and J2-7. The following
84 PCR condition amplifies the entire J1 (Jβ1 primer coupled with a V or D region primer)
85 as well as J2 genomic DNA region (Jβ2 primer coupled with a V or D segment primer).

86 Since VDJ will be spliced to a C region at the RNA level, the selected Jβ primer
87 sequences remain in genomic DNA after recombination. D region primers were designed
88 5' of D1 and D2. The D segment primers amplify only a D-to-J rearranged genome, but
89 not a V-to-DJ rearranged genome. After V-to-D1J rearrangement, the D1 primer
90 sequence is removed from the genome. Similarly after V-to-D2J2 rearrangement, the D2
91 primer sequence is removed from the genome. The first round of PCR was performed in a
92 60 μl final reaction volume containing of 50 nM of each primer, 1x Q5® Reaction
93 Buffer, 200 μM each dNTPs, 0.4 unit Q5® High-Fidelity DNA Polymerase (NEB) and
94 20 μl of the lysed single cell solution. The PCR condition was 30 sec at 98 ºC followed
95 by 30 cycles of 5 sec at 98 ºC, 10 sec at 66-58 ºC and 2 min at 72 ºC, and then final
96 extension for 2 min at 72 ºC. During the first 5 cycles, the annealing temperature was
97 reduced 2 ºC per cycle from 66 ºC to 58 ºC, and then performed at 56ºC for the last 25
98 cycles.
99 For the second round of PCR, nested primers were selected: 32 nested V region
100 primers containing forward adapter sequences (AF-Vβn), 2 nested D region forward
101 adapter primers (AF-Dβn) and 2 nested J region primers in a reverse adapter orientation
102 (AR3-Jβn). The second round of PCR was performed in a 20 μl final reaction volume
103 containing 50 nM of each primer (Fig. 1a), 1x Q5® buffer, 200 μM of each dNTP, 0.4
104 units Q5® High-Fidelity DNA Polymerase (NEB) and 1 μl of the first round PCR
105 product. The 2nd PCR condition was same as for the first round PCR but for only 25
106 cycles.
107 Finally, in the third round of PCR, unique barcoded-AF and barcoded-AR3
108 combinations of primers were selected for each single cell (Fig. 1a). Barcodes were

109 adopted from a published report(6). The 3rd PCR condition was same as for the second
110 round PCR (5 + 20 cycles). Hot start PCR was performed for all PCR reactions using
111 either a BioRad T100™ or Applied Biosystems 2720 Thermal Cycler. All primers were
112 purchased from Integrated DNA Technologies, Inc. The 37 (for 1st round) or 36 (for 2nd
113 round) PCR primers at 200 μM final concentrations were mixed, aliquoted and frozen for
114 subsequent use. Each PCR mix was prepared immediately before initiating PCR
115 reactions. (The amplification efficiency was reduced if the PCR mix was prepared in
116 advance and repeatedly frozen and thawed. Amplification efficiency was also reduced
117 when 500 nM of each primer was used in the 2nd round PCR reaction). Successful DNA
118 amplification was confirmed by running 2 μl of 3rd round PCR product on agarose gels
119 with 0.4 μg of 2-Log DNA Ladder (NEB) (Fig. 1c): 8 samples were selected from each
120 96 well PCR plate. Of note, the PCR conditions employed here allow amplification of
121 unrearranged (germline configuration) genomic DNA (Fig. 1c left panel).
122
123 **Recovery frequency**
124 To analyze the frequency of recovered wells, a primer pair amplifying the 3' region of the
125 *Actb* gene was added to the 1st round PCR. 1 μl of the 1st round PCR product was used
126 in a PCR reaction with nested *Actb* primers: 5'-CATAGGCTTCACACCTTCCT-3', 5'-
127 CTTTGCCTCCATCTGCATAAC-3' and FAM-labeled probe
128 TGCTAGTCTGAAGCTGCCCTTTCC (ZEN / Iowa Black FQ) purchased from
129 Integrated DNA Technologies, Inc. Luna Universal Probe qPCR Master Mix (NEB,
130 M3004) was used in a 20 μl reaction. The PCR condition was 30 sec at 95 ºC followed
131 by 45 cycles of 1 sec at 95 ºC and 20 sec at 60 ºC using StepOnePlus™ Real-Time PCR

132 System (Thermo Fisher Scientific). A 90-95% recovery efficiency was routinely obtained
133 from each plate containing 96 single cell sorted wells.
134
135 **PacBio high-throughput sequencing**
136 Since the length of the PCR products described above ranged from 200 to 2700 bp (Fig.
137 1b), a sequencer with the ability to read long DNA fragments is preferred. To this end,
138 Pacific Biosciences single molecule real-time (SMRT) sequencing technology(14) was
139 employed. A PacBio smartbell adapter was ligated to the barcoded PCR product mixture
140 and then sequenced on PacBio RSII sequencer at the University of Michigan Sequencing
141 Facility. Thus each circular consensus sequencing (CCS) read generated from PacBio
142 sequencing corresponds to a single molecule of initial PCR product (Fig. 2a).
143
144 **PacBio raw reads analysis**
145 To obtain DNA sequences amplified from genomic DNA in single cells, we developed an
146 in-house analysis pipeline (Fig. 2a; the full version is available at
147 https://github.com/Hongyang449/scVDJ_seq). To calculate CCS reads of inserts, PacBio
148 ConsensusTools software version 2.0.0 was used with the following settings: --
149 minFullPasses=2, --minPredictedAccuracy=90, --minLength=10 and default for other
150 parameters (https://github.com/PacificBiosciences/SMRT-
151 Analysis/wiki/Documentation).
152 First, to demultiplex CCS reads, we grouped the CCS reads based on the presence of
153 a barcode primer set flanking each end of the CCS. We obtained approximately 35,000
154 CCSs per PacBio RS II SMRTcell on average, and 46% of them started with a barcode-

155 AF primer and ended with a barcode-AR3 primer. The remaining 54% had mutations,
156 deletions, or additions within the barcoded primer sequence or bore truncated barcode
157 primer sequences, and were excluded from the analysis.

158 Next, we generated sub-groups in each single cell (unique barcode set) based on the
159 presence of a V or D primer sequence, and sequences in each sub-group were aligned
160 using multiple sequence alignment software MUSCLE(15)
161 (http://www.drive5.com/muscle/). For each group, hierarchical clustering based on
162 sequence length was performed to create initial clusters, then k-means clustering was
163 recursively performed on the DNA sequences of each initial cluster to generate the final
164 multiple clusters, until every nucleotide position in each cluster achieved >51% identity.
165 Clusters containing only 1 CCS read were excluded. Aberrant CCSs with two PCR
166 products ligated together probably resulted during the smartbell adapter ligation step, and
167 were also excluded from the analysis.

168 To obtain accurate genomic DNA sequence, we calculated a consensus sequence
169 from multiple CCSs in each cluster. Of note, each CCS file that was generated from
170 single zero-mode waveguide (ZMW) corresponds to one copy of PCR product. If a
171 mutation happened during the first PCR cycle, 50% of the final PCR products inherit the
172 mutation at a specific base position in the final PCR product. Similarly, a mutation
173 during the second cycle is inherited by 25% of the final PCR products, 12.5% after the
174 third cycle and so on. After adapter and J region genomic DNA sequences (except the J
175 region used for rearrangement) were removed, the cluster consensus with V primer
176 sequences were submitted to IMGT HighV-Quest analysis(16) (http://www.imgt.org/) to
177 determine whether a specific VDJ sequence in a cluster was predicted to encode a

178 productive or unproductive TCRβ protein. For example, when DNA sequences
179 rearranged with J1-3, J1-4 to J1-7 regions as well as 3' of J1-7 to J1 primer sequences
180 were removed. The removal of the extra J region genomic DNA is essential to accurately
181 determine the productivity of rearrangement, especially when some bases are truncated
182 from J regions.

183 Based on sequence and functionality predicted by IMGT analysis, each cluster
184 consensus sequence was categorized into a tag shown in Supplementary Fig. S1a. Since
185 the status of an allele (GL, DJ, uVDJ and pVDJ) can be determined from possible
186 combinations of 1-2 cluster tags (a total of 24 different patterns are shown in
187 Supplementary Fig. S1b), the rearrangement status from within a single cell (the
188 combination of two separate allelic patterns) was calculated from 2-4 tags recovered from
189 each cell (for a total 253 different combinations of tags). From those, the summary tables
190 were generated.

191 In this study, the rearrangement status of both alleles was obtained from 909 single
192 cells (25% of 3,664 wells) for thymocytes post beta-selection and 193 DN3a stage
193 thymocytes (25% of 784 wells) when V-to-DJ rearrangement takes place. The frequency
194 depends on sequence depth and was 40% when we recovered 29,643 CCSs from 576
195 wells. We obtained 0-4 clusters per single cell (1.62 clusters per cell on average) in this
196 depth of sequencing. PacBio CCS data, the barcode list for each sample, genomic DNA
197 and CDR3-IMGT amino acid sequences for all clusters generated from this study are
198 deposited at
199 (https://umich.box.com/s/3f5q64u2i68dn2i7hlneucxph9oetpvb).
200

201 **RESULTS**

202 **High-throughput *Trb* genomic DNA sequencing from single cells**

203 To sequence both alleles of genomic DNA in the *Trb* gene VDJ region, we employed
204 a multiplex strategy (Fig. 1a; details described in Methods). Since mouse thymocytes
205 were directly sorted into lysis buffer in individual wells of 96-well PCR plates.
206 Multiplex, nested PCR reactions followed by barcoding PCR amplified the germline (GL:
207 unrearranged configuration) genomic DNA, D-to-J rearranged (DJ) DNA, and all
208 potential V-to-DJ rearranged (VDJ) DNA in the *Trb* locus(13) (Figs. 1b and 1c). The
209 PCR products were mixed and then sequenced in a PacBio RS II long-read sequencer
210 (www.pacb.com). Each circular consensus specification (CCS) read of an insert
211 generated by the PacBio system corresponds to a single molecule of the original PCR
212 product (Fig. 2a). Based on barcoding and sequence similarity, multiple CCSs were
213 grouped into clusters (Fig. 2a). For each cluster, the consensus sequence was retrieved,
214 which corresponds to an original PCR product. The consensus sequences were
215 subsequently submitted to IMGT HighV-Quest analysis(16) (http://www.imgt.org/) to
216 determine whether the VDJ sequences in any cluster predicted that a functional TCRβ
217 protein could be generated in that cell. From one *Trb* allele, two [(V-)D1-J1 and (V31-
218)D2-J2 rearrangement] or one [(V-)D-J2] PCR products are amplified (Fig. 1b and
219 Supplementary Fig. S1). Theoretically, between two and four PCR products (clusters) can
220 be recovered from two alleles in each cell. Details for the computational processing of the
221 PacBio output data to calculate allele rearrangement status are described in Materials and
222 Methods.

223 To calculate the PCR error rate under our experimental conditions, we took advantage
224 of the unrearranged (germline configuration) genomic DNA corresponding to parts of
225 *Trbd1* to *Trbj1-7* and *Trbd2* to *Trbj2-7* (Fig. 1b) in single bone marrow (BM) cells. From
226 single BM cells we obtained two PCR bands (Fig. 1c), which correspond to the lengths of
227 germline DNA amplified using Dβ1 forward and Jβ1 reverse nested primers (2,672 bp)
228 as well as Dβ2 forward and Jβ2 reverse nested primers (1,896 bp). The accuracy of the
229 resultant sequences was 99.768% for each individual CCS, including both PCR and
230 PacBio sequencing errors. To increase the accuracy, we employed an approach to obtain
231 the consensus sequence from more than 2 CCSs. If we obtain only 2 CCSs per cluster, an
232 error at a position results in 50% identity at the position (ex. An A in 1 CCS and a G in
233 the other CCS); such a cluster was then removed because it is not >51% identical. If we
234 recover 3 CCSs, an error at any position results in a 66.6% identity at the position (e.g.
235 An A in 2 CCSs and a G in 1 CCS). In this case A is the consensus sequence at the
236 position in this cluster. When a consensus was generated from 3 randomly chosen
237 synonymous CCSs, the accuracy increased to 99.991%. These data demonstrate that this
238 method using PCR with Q5® High-Fidelity DNA polymerase (NEB) followed by
239 calculating the consensus sequence from multiple CCSs yields >99.99% accuracy, even
240 after a total of 80 PCR cycles.
241
242 ***Trb* VDJ rearrangement status in single thymocytes**
243 To precisely determine the genomic status at the *Trb* loci following VDJ
244 rearrangement in single cells, we analyzed wild type DN4, DP and late DP stage
245 thymocytes, in which *Trb* VDJ rearrangement had already been completed (i.e. post-β-

246 selection). Since DJ joining takes place before the DN3 stage on both alleles, all of the
247 cells analyzed here would be predicted to have undergone at least one productive VDJ
248 recombination event on one allele. We obtained a total of 5,877 clusters and assessed the
249 rearrangement status of both alleles from 909 single cells (Fig. 3). 52% (472) of the
250 single thymocytes were in a [pVDJ (productive VDJ rearranged)/DJ] configuration and
251 41% (377) were in a [uVDJ (unproductively VDJ rearranged)/pVDJ] configuration, as
252 expected from the 60/40-rule(2). 4% (37) of the post-β-selection thymocytes were in a
253 pVDJ/pVDJ arrangement, which was similar to the 3% dual-TCRβ bearing cells as
254 analyzed using cell surface antibodies(17). Stage-specific data is shown in
255 Supplementary Figs. S2 and S3. The V regions that were most frequently utilized for
256 recombination were V13-2 (Fig. 4 and Supplemental Fig. S4), which is also the most
257 frequently used in inflamed lung epitope specific CD8+ T cells(8). 6,415 V-to-DJ
258 rearranged genomic DNA and complementarity determining region 3 (CDR3) amino acid
259 sequences (4,258 productive and 2,158 unproductive) recovered from single thymocytes
260 in this analysis are available online (see Data Access for detail). Surprisingly, we
261 recovered D2-to-J1 rearranged sequences, although only rarely (Supplementary Fig. S2).
262 At the double negative 3a (DN3a) stage, when V-to-DJ rearrangement takes place, only
263 9% of the cells had generated a productive (pVDJ) allele [0.5% (pVDJ/GL), 7.2% had
264 generated a (pVDJ/DJ) recombinant and 1.5% were in a (uVDJ/pVDJ) configuration,
265 Supplementary Fig. S3], all of which probably represent cells that were about to develop
266 to the next (DN3b) stage.
267 Next, to determine whether the methods developed here can capture an artificial
268 allelic exclusion phenomenon, we analyzed late DP cells isolated from Vb8 transgenic

269 mice(18) (Tg^{Vb8}), which express a productively rearranged TCRβ protein; expression of
270 that transgene has been shown to repress endogenous *Trb* VDJ rearrangement(19). Out of
271 2,429 clusters generated, 722 clusters bore the DNA sequence of the Tg^{Vb8} transgene
272 (Supplementary Figs. S4a) and were removed from further analysis. We then determined
273 the rearrangement status of both alleles from 127 single cells (Supplementary Figs. S4b
274 and S4c). V-to-DJ rearrangement was not observed at the endogenous locus in 80% of
275 the thymocytes, in agreement with published observations(19). However, 20% of the
276 cells had either unproductively (13% uVDJ/DJ) or productively (6% pVDJ/DJ) escaped
277 transgene-suppressed allelic exclusion to allow V-to-DJ rearrangement on one of the
278 endogenous *Trb* loci. In the Tg^{Vb8}, D2-to-J2 rearrangement was observed less often than
279 D1-to-J1. This supports the idea that D2-to-J2 rearrangement was more frequently
280 repressed when compared to D1-to-J1 region recombination in the presence of a
281 functional TCRβ transgene, in agreement with published analysis of 9 T cell clones(19).
282

DISCUSSION

In summary, the method developed here provides an extremely accurate, rapid high-throughput approach for the analysis of *Trb* genomic DNA status at the single cell level. This strategy allowed us to analyze the rearrangement status in thousands of single cells, which is more than 10-fold greater than the number from Sanger sequencing methods previously employed by us and others(10-12) and requiring far less time. This strategy would also be directly applicable to the analysis of the *Igl* and *Igh* genes (encoding immunoglobulins in B cells), which are also regulated by recombination and allelic exclusion at the genetic level(2).

306 **REFERENCES**

307 1. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575-
308 581.

309 2. Mostoslavsky, R., F. W. Alt, and K. Rajewsky. 2004. The lingering enigma of the
310 allelic exclusion mechanism. *Cell* 118: 539-544.

311 3. Alam, S. M., and N. R. Gascoigne. 1998. Posttranslational regulation of TCR
312 Valpha allelic exclusion during T cell differentiation. *J Immunol* 160: 3883-3890.

313 4. Hinz, T., E. Weidmann, and D. Kabelitz. 2001. Dual TCR-expressing T
314 lymphocytes in health and disease. *Int Arch Allergy Immunol* 125: 16-20.

315 5. Auger, J. L., S. Haasken, E. M. Steinert, and B. A. Binstadt. 2012. Incomplete
316 TCR-beta allelic exclusion accelerates spontaneous autoimmune arthritis in
317 K/BxN TCR transgenic mice. *Eur J Immunol* 42: 2354-2362.

318 6. Han, A., J. Glanville, L. Hansmann, and M. M. Davis. 2014. Linking T-cell
319 receptor sequence to functional phenotype at the single-cell level. *Nat Biotechnol*
320 32: 684-692.

321 7. Weischenfeldt, J., I. Damgaard, D. Bryder, K. Theilgaard-Monch, L. A. Thoren,
322 F. C. Nielsen, S. E. Jacobsen, C. Nerlov, and B. T. Porse. 2008. NMD is essential
323 for hematopoietic stem and progenitor cells and for eliminating by-products of
324 programmed DNA rearrangements. *Genes Dev* 22: 1381-1396.

325 8. Dash, P., J. L. McClaren, T. H. Oguin, 3rd, W. Rothwell, B. Todd, M. Y. Morris,
326 J. Becksfort, C. Reynolds, S. A. Brown, P. C. Doherty, and P. G. Thomas. 2011.
327 Paired analysis of TCRalpha and TCRbeta chains at the single-cell level in mice.
328 *J Clin Invest* 121: 288-295.

329 9. Robins, H. S., P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O.
330 Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson. 2009. Comprehensive
331 assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:
332 4099-4107.
333 10. Ku, C. J., J. M. Sekiguchi, B. Panwar, Y. Guan, S. Takahashi, K. Yoh, I. Maillard,
334 T. Hosoya, and J. D. Engel. 2017. GATA3 Abundance Is a Critical Determinant
335 of T Cell Receptor beta Allelic Exclusion. *Mol Cell Biol* 37.
336 11. Aifantis, I., J. Buer, H. von Boehmer, and O. Azogui. 1997. Essential role of the
337 pre-T cell receptor in allelic exclusion of the T cell receptor beta locus. *Immunity*
338 7: 601-607.
339 12. Aifantis, I., V. I. Pivniouk, F. Gartner, J. Feinberg, W. Swat, F. W. Alt, H. von
340 Boehmer, and R. S. Geha. 1999. Allelic exclusion of the T cell receptor beta locus
341 requires the SH2 domain-containing leukocyte protein (SLP)-76 adaptor protein. *J*
342 *Exp Med* 190: 1093-1102.
343 13. Bosc, N., and M. P. Lefranc. 2000. The mouse (Mus musculus) T cell receptor
344 beta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. *Exp Clin*
345 *Immunogenet* 17: 216-228.
346 14. Roberts, R. J., M. O. Carneiro, and M. C. Schatz. 2013. The advantages of SMRT
347 sequencing. *Genome Biol* 14: 405.
348 15. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy
349 and high throughput. *Nucleic Acids Res* 32: 1792-1797.
350 16. Li, S., M. P. Lefranc, J. J. Miles, E. Alamyar, V. Giudicelli, P. Duroux, J. D.
351 Freeman, V. D. Corbin, J. P. Scheerlinck, M. A. Frohman, P. U. Cameron, M.

352 Plebanski, B. Loveland, S. R. Burrows, A. T. Papenfuss, and E. J. Gowans. 2013.
353 IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and
354 next generation repertoire immunoprofiling. *Nat Commun* 4: 2333.
355 17. Balomenos, D., R. S. Balderas, K. P. Mulvany, J. Kaye, D. H. Kono, and A. N.
356 Theofilopoulos. 1995. Incomplete T cell receptor V beta allelic exclusion and
357 dual V beta-expressing cells. *J Immunol* 155: 3308-3312.
358 18. Shinkai, Y., S. Koyasu, K. Nakayama, K. M. Murphy, D. Y. Loh, E. L. Reinherz,
359 and F. W. Alt. 1993. Restoration of T cell development in RAG-2-deficient mice
360 by functional TCR transgenes. *Science* 259: 822-825.
361 19. Uematsu, Y., S. Ryser, Z. Dembic, P. Borgulya, P. Krimpenfort, A. Berns, H. von
362 Boehmer, and M. Steinmetz. 1988. In transgenic mice the introduced functional T
363 cell receptor beta gene prevents expression of endogenous beta genes. *Cell* 52:
364 831-841.
365
366
367
368
369
370
371
372
373
374

375 **FOOTNOTES**

376 This work was supported by National Institute of Health Grant AI094642 (to T.H. and
377 J.D.E). The research was also supported in part by the University of Michigan
378 Comprehensive Cancer Center Support Grant (P30 CA046592) for use of the Flow
379 Cytometry and the Sequencing Cores at the University of Michigan.

380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397

398 **AUTHOR CONTRIBUTIONS**

399 T.H. designed the study, performed experiments, wrote computer script, analyzed the
400 data, and wrote the manuscript. H.L. designed the study, wrote computer script and
401 edited the paper. C.K. designed the study, performed experiments, analyzed the data and
402 edited the manuscript. Q.W performed experiments and edited the paper. Y.G. designed
403 the study and edited the paper. J.D.E. designed the study and edited the paper.

404

405 **DISCLOSURE DECLARATION**

406 The authors declare no competing interests.

407

408 **FIGURE LEGENDS**

409 **Fig. 1. High-throughput *Trb* genomic DNA sequencing from single-cells.** (**a**) Strategy
410 for multiplex PCR amplification of genomic DNA sequences at the *Trb* gene VDJ loci
411 from single cells. Final PCR products were mixed and then sequenced by PacBio RS II
412 sequencer. Vβn: V region primers. Dβn: D region primers. Jβn: J region primers. AF:
413 adapter forward. AR: adapter reverse. BC: barcode. Details are described in Methods. (**b**)
414 Schematic presentation of VDJ rearrangement at the *Trb* locus with simplified illustration
415 of multiplex primer location. Genomic DNA recombination events are initiated by
416 recombining Dβ (diversity) and Jβ (joining) segments on both chromosomes.
417 Subsequently, on only one chromosome, one of the Vβ (variable) segments is joined to
418 the previously rearranged DJ recombinant. Not only the selection of each VDJ segment,
419 but also multiple lengths of spacer sequence between V, D and J, generates an
420 incalculable number of VDJ sequence possibilities. mRNA splicing joins Cβ (constant)

421 segments to the rearranged VDJ DNA recombinants to generate a final TCRβ protein.
422 VDJ rearrangement that generates a stop codon or elongated transcript results in a
423 predictedly unproductive TCRβ. (**c**) Successful amplification of rearranged VDJ, DJ and
424 germline D-to-J regions were confirmed by agarose gel electrophoresis with NEB 2-log
425 DNA ladder in the last (left panel) and first (right panel) lanes.
426
427 **Fig. 2. A schematic outline of the PacBio raw data processing.** Details are described
428 in Methods.
429
430 **Fig. 3. *Trb* VDJ rearrangement status in single thymocytes.** (**a**) Cluster tags in a total
431 of 5,877 clusters recovered from single wild type thymocytes at DN4, DP and late DP
432 (post β-selection) stages. Each cluster consensus sequence was categorized into one tag
433 shown on the x-axis (see Supplementary Fig. S1 for detail). (**b**) Calculated *Trb* allele
434 status for 909 single wild type thymocytes at DN4, DP and late DP (post β-selection)
435 stages. Legend: GL: germline allele status; DJ: D-to-J recombined allele; pVDJ: V-to-DJ
436 recombined allele encoding a productive TCRβ protein; uVDJ: V-to-DJ recombined
437 allele encoding a predicted unproductive TCRβ. The data in (**b**) and (**c**) represent a
438 summary of 4 mice for DN4, 5 mice for DP and 4 mice for late DP stages, and details are
439 shown in Supplementary Figs. S2 and S3.
440
441 **Fig. 4. V and J region usage.** Frequency of each V region (a) and J region (b) utilization
442 was calculated out of 3,356 VDJ clusters recovered from wild type single thymocytes at
443 DN4, DP and late DP stages. Stage-specific data are shown in Supplementary Fig. S4.

444

445 **Fig. 5. Endogenous *Trb* rearranged status in thymocytes of Vb8 transgenic mice.**

446 Number of tags (a) for recovered PCR fragments and rearrangement status (b) calculated

447 from late DP stage single cells isolated from wild type (black) and Vb8 transgenic (Tg,

448 white) mice. Data represent the summary from 4 mice.

449

Figure 1

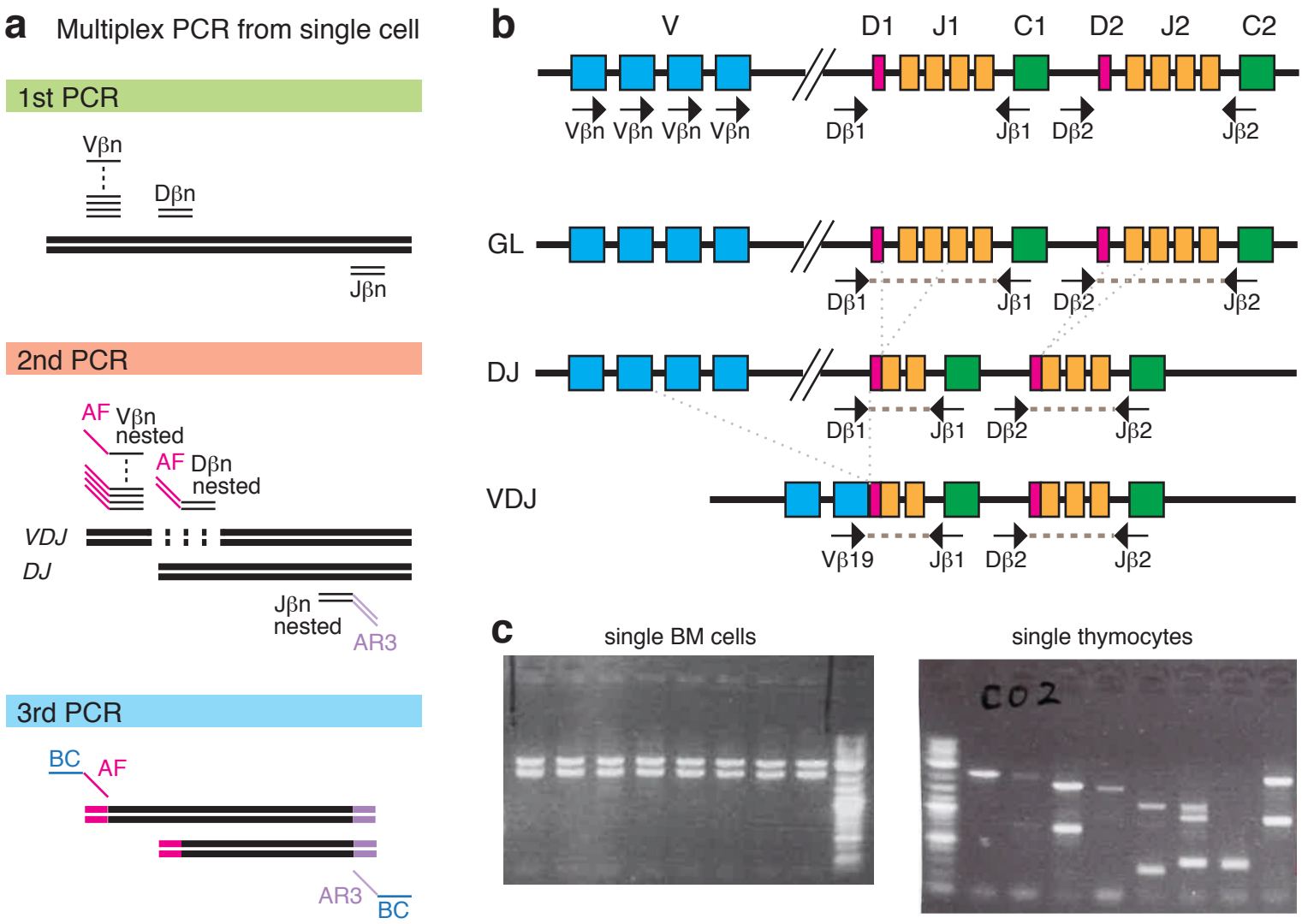

a
Multiplex PCR from single cell
1st PCR
Vβn
Dβn
Jβn
2nd PCR
AF Vβn nested
AF Dβn nested
VDJ
DJ
Jβn nested
AR3
3rd PCR
BC
AF
AR3
BC
b
V
D1
J1
C1
D2
J2
C2
Vβn Vβn Vβn Vβn
Dβ1
Jβ1
Dβ2
Jβ2
GL
DJ
VDJ
Vβ19
c
single BM cells
single thymocytes

Figure 2

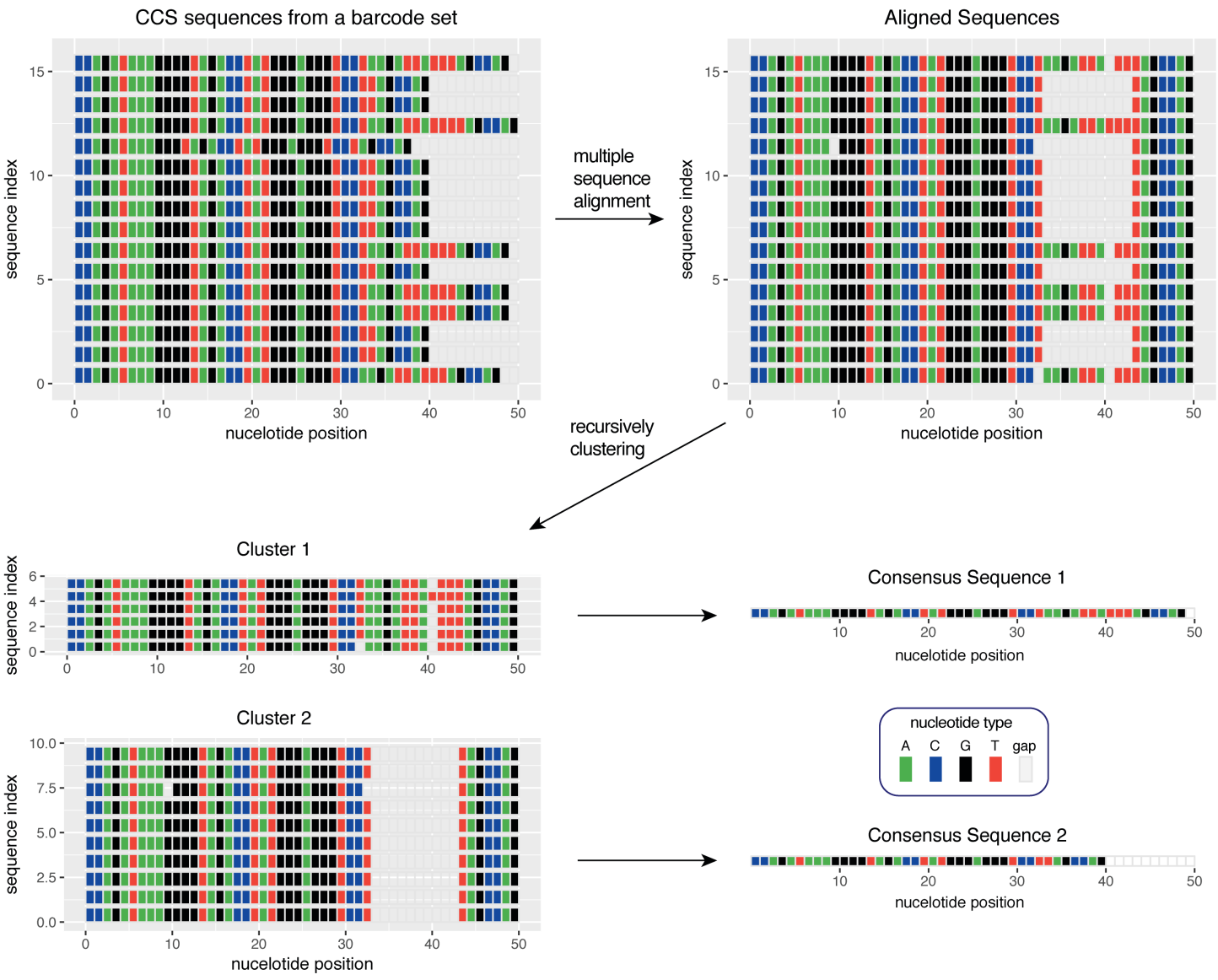

CCS sequences from a barcode set
sequence index
nucelotide position
multiple sequence alignment
Aligned Sequences
sequence index
nucelotide position
recursively clustering
Cluster 1
sequence index
Consensus Sequence 1
nucelotide position
Cluster 2
sequence index
nucelotide position
nucleotide type
A C G T gap
Consensus Sequence 2
nucelotide position

Figure 3

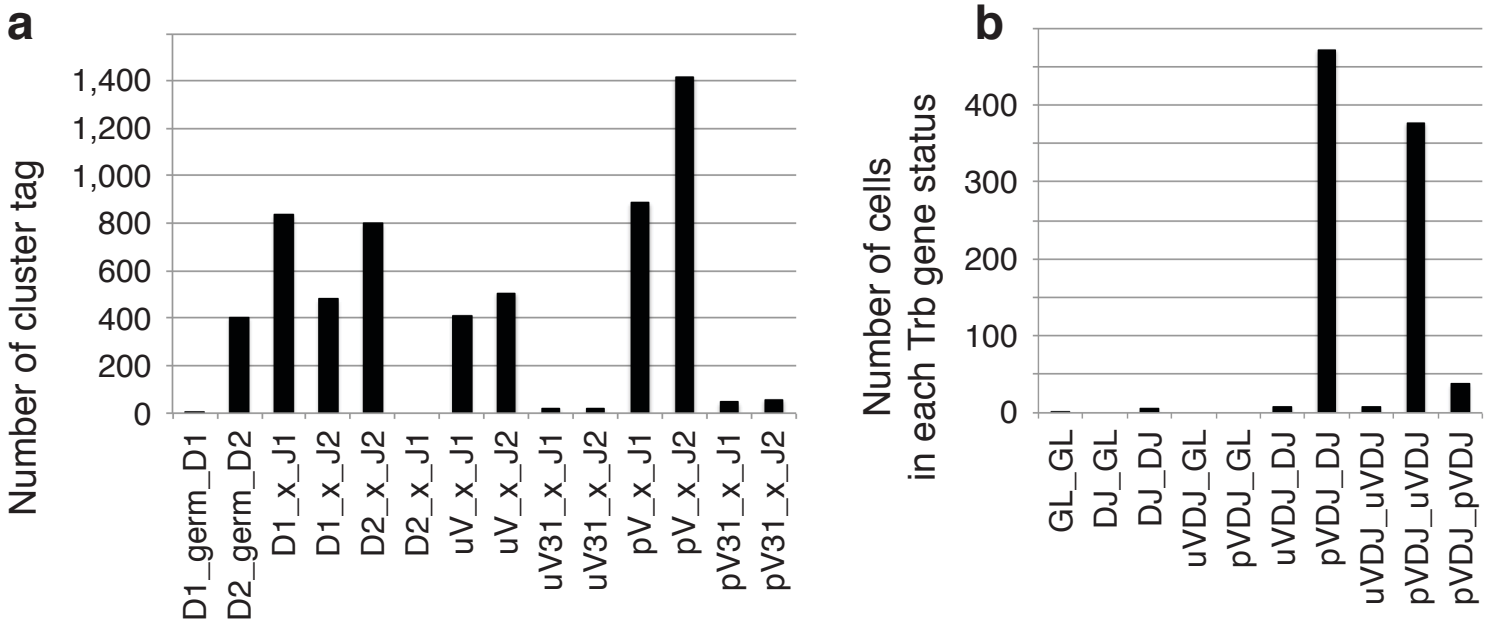
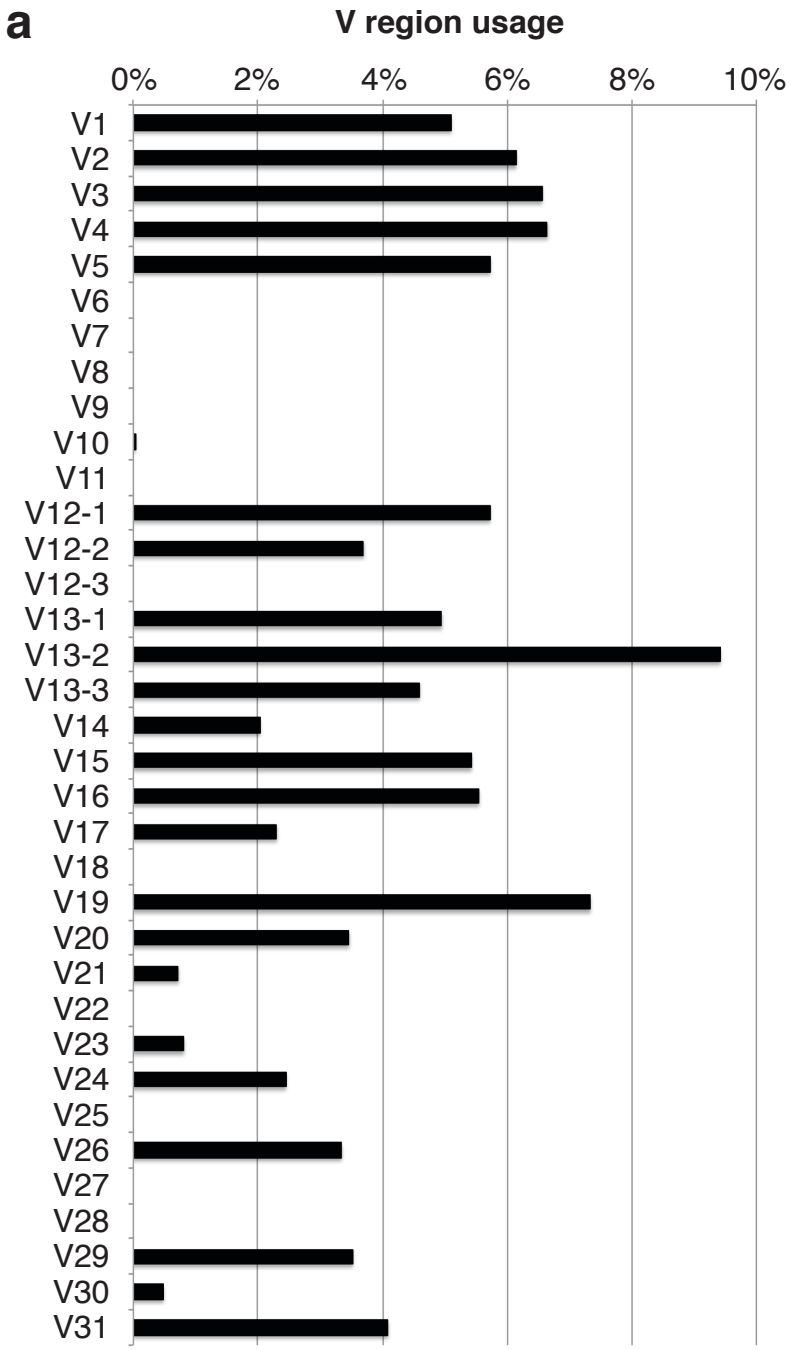

a
Number of cluster tag
1,400
1,200
1,000
800
600
400
200
0
D1_germ_D1
D2_germ_D2
D1_x_J1
D1_x_J2
D2_x_J2
D2_x_J1
uV_x_J1
uV_x_J2
uV31_x_J1
uV31_x_J2
pV_x_J1
pV_x_J2
pV31_x_J1
pV31_x_J2
b
Number of cells
in each Trb gene status
400
300
200
100
0
GL_GL
DJ_GL
DJ_DJ
uVDJ_GL
pVDJ_GL
uVDJ_DJ
pVDJ_DJ
uVDJ_uVDJ
pVDJ_uVDJ
pVDJ_pVDJ

Figure 4


a
V region usage
0%
2%
4%
6%
8%
10%
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12-1
V12-2
V12-3
V13-1
V13-2
V13-3
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
V29
V30
V31

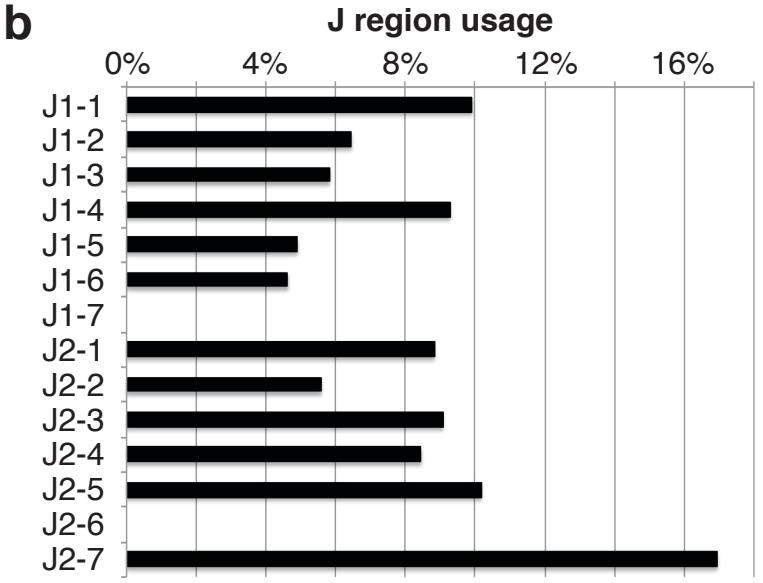
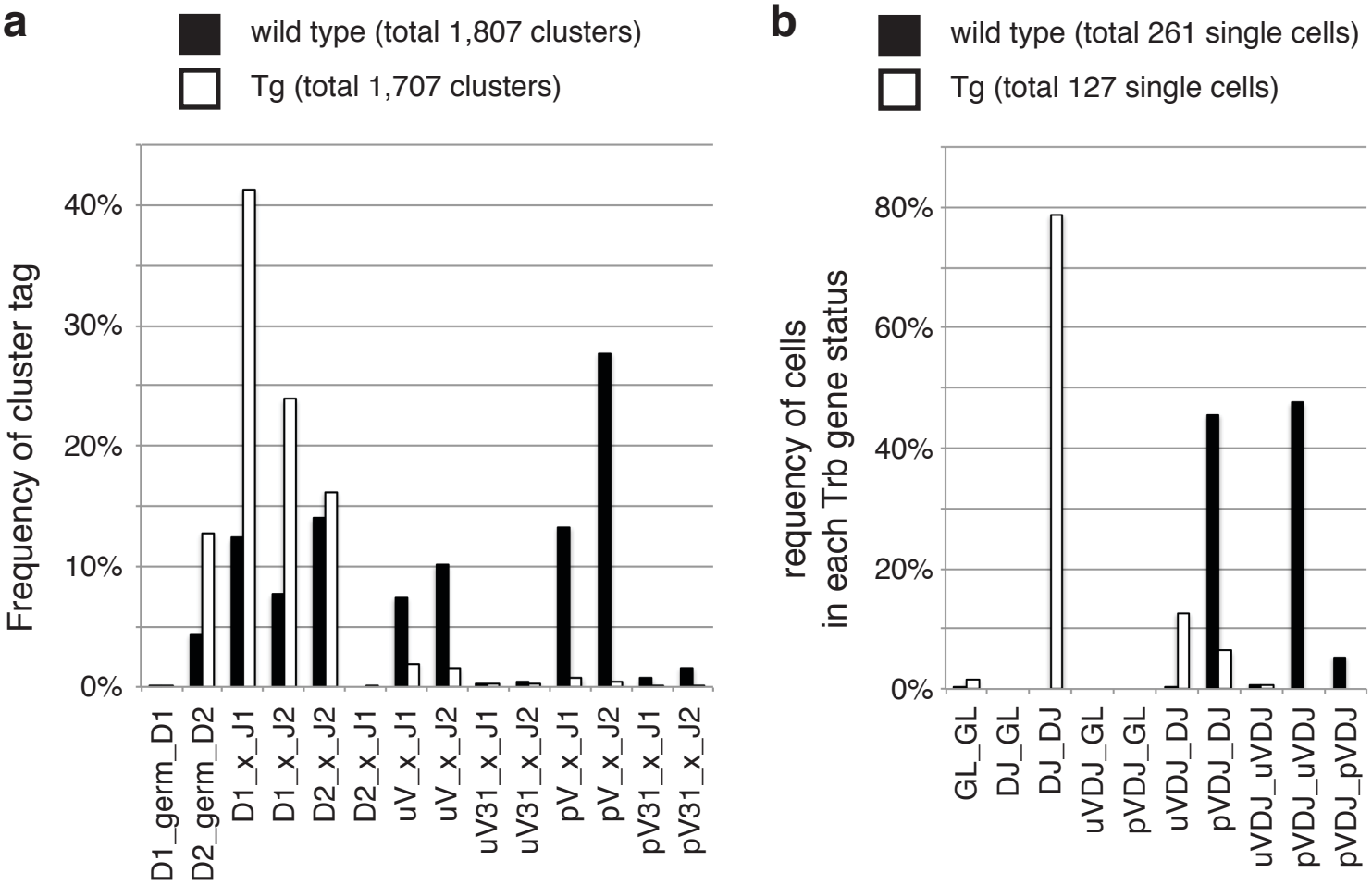

b
J region usage
0%
4%
8%
12%
16%
J1-1
J1-2
J1-3
J1-4
J1-5
J1-6
J1-7
J2-1
J2-2
J2-3
J2-4
J2-5
J2-6
J2-7

Figure 5


a
wild type (total 1,807 clusters)
Tg (total 1,707 clusters)
Frequency of cluster tag
40%
30%
20%
10%
0%
D1_germ_D1
D2_germ_D2
D1_x_J1
D1_x_J2
D2_x_J2
D2_x_J1
uV_x_J1
uV_x_J2
uV31_x_J1
uV31_x_J2
pV_x_J1
pV_x_J2
pV31_x_J1
pV31_x_J2
b
wild type (total 261 single cells)
Tg (total 127 single cells)
requency of cells
in each Trb gene status
80%
60%
40%
20%
0%
GL_GL
DJ_GL
DJ_DJ
uVDJ_GL
pVDJ_GL
uVDJ_DJ
pVDJ_DJ
uVDJ_uVDJ
pVDJ_uVDJ
pVDJ_pVDJ

a

cluster tag	V x J1	D1 x J1	1 x 2	D2 x J2	V x J2
tag1		D1_germ_J1			
tag2				D2_germ_J2	
tag3		D1_x_J1			
tag4			D1_x_J2		
tag5				D2_x_J2	
tag6			D2_x_J1		
tag7 (V1 to V30)	uV_x_J1				
tag8 (V1 to V30)					uV_x_J2
tag9	uV31_x_J1				
tag10					uV31_x_J2
tag11 (V1 to V30)	pV_x_J1				
tag12 (V1 to V30)					pV_x_J2
tag13	pV31_x_J1				
tag14					pV31_x_J2

b

status	pattern	V x J1	D1 x J1	1 x 2	D2 x J2	V x J2
GL	g1	__	D1_germ_J1	__	D2_germ_J2	__
DJ	d1	__	D1_x_J1	__	D2_germ_J2	__
DJ	d2	__	D1_germ_J1	__	D2_x_J2	__
DJ	d3	__	D1_x_J1	__	D2_x_J2	__
DJ	d4	__	__	D1_x_J2	__	__
DJ	d5	__	__	D2_x_J1	__	__
uVDJ	u1	__	__	__	__	uV_x_J2
uVDJ	u2	uV_x_J1	__	__	D2_germ_J2	__
uVDJ	u3	uV_x_J1	__	__	D2_x_J2	__
uVDJ	u4	uV31_x_J1	__	__	__	__
uVDJ	u5	uV31_x_J2	__	__	__	__
uVDJ	u6	__	D1_germ_J1	__	__	uV31_x_J2
uVDJ	u7	__	D1_x_J1	__	__	uV31_x_J2
uVDJ	u8	uV_x_J1	__	__	__	uV31_x_J2
pVDJ	p1	__	__	__	__	pV_x_J2
pVDJ	p2	pV_x_J1	__	__	D2_germ_J2	__
pVDJ	p3	pV_x_J1	__	__	D2_x_J2	__
pVDJ	p4	pV31_x_J1	__	__	__	__
pVDJ	p5	pV31_x_J2	__	__	__	__
pVDJ	p6	__	D1_germ_J1	__	__	pV31_x_J2
pVDJ	p7	__	D1_x_J1	__	__	pV31_x_J2
pVDJ	p8	uV_x_J1	__	__	__	pV31_x_J2
pVDJ	p9	pV_x_J1	__	__	__	uV31_x_J2
pVDJ	p10	pV_x_J1	__	__	__	pV31_x_J2

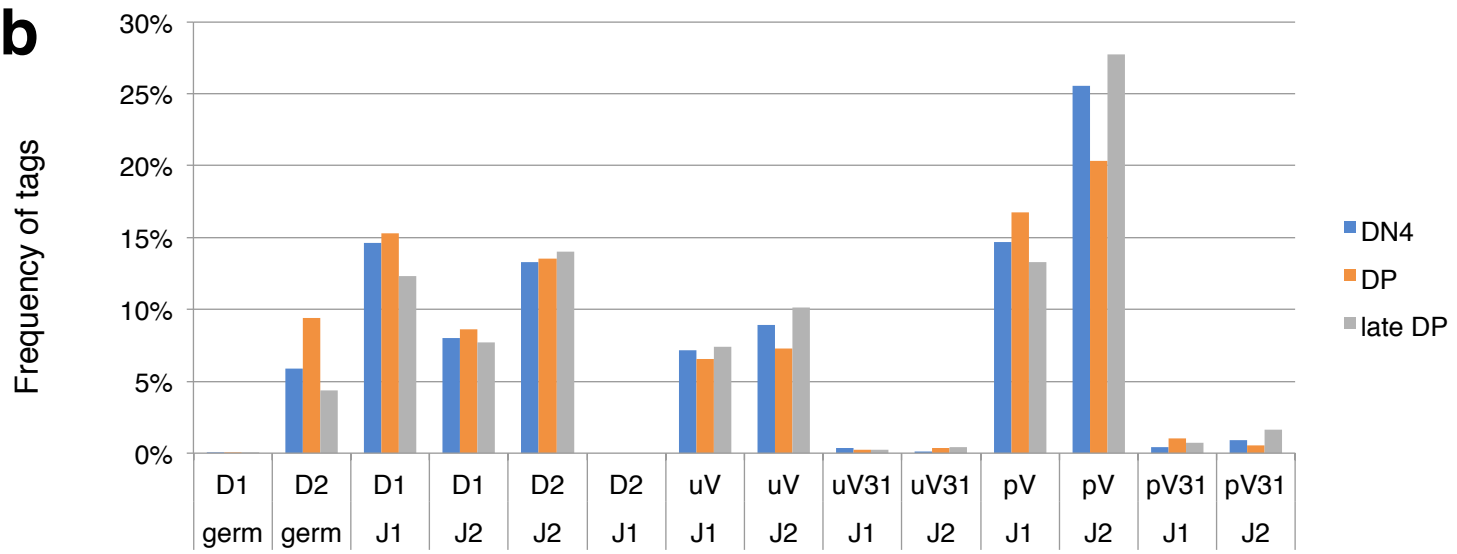
Supplementary Figure S1

(**a**) Full list of cluster tags (PCR products). PCR fragments amplified using Dβ forward (D1 or D2) and Jβ reverse (J1 or J2) nested primers include both germline configuration (germ) and rearranged DNA (x). PCR fragments amplified using Vβ forward (V) and Jβ reverse (J1 or J2) nested primers encode either productive (p, green) or unproductive (u, red) TCRβ protein. (**b**) Full list of *Trb* gene status and patterns. Status of an allele are shown with possible combination of 1-2 PCR fragments. GL: germline configuration (patter g). DJ: D-to-J rearranged (pattern d). uVDJ: V-to-DJ rearranged DNA encoding unproductive TCRβ (pattern u). pVDJ: V-to-DJ rearranged DNA encoding productive TCRβ (pattern p). Rearrangement with *Trbv31* region (V31) increases the complexity. Full list of possible combination of PCR fragment tags per cell is found in tag_status.csv file in code.zip file. Combination of 2-4 tags used to identify allele status of each cell.

a

	D1	D2	D1	D1	D2	D2	uV	uV	uV31	uV31	pV	pV	pV31	pV31		total
sampleID	germ	germ	J1	J2	J2	J1	J1	J2	J1	J2	J1	J2	J1	J2		
DN3a_WT_mouse2	50	108	93	44	21	0	22	10	0	1	4	5	0	0		358
DN3a_WT_mouse3	18	65	93	49	28	1	29	37	1	1	10	2	1	0		335
DN3a_WT_mouse4	21	70	87	52	23	0	26	17	2	0	5	7	0	0		310
DN3a_WT_mouse6	31	78	86	31	24	1	25	16	1	3	12	9	0	0		317
DN4_WT_mouse62	0	16	47	26	48	0	19	28	4	1	59	107	2	4		361
DN4_WT_mouse64	0	28	69	39	58	0	32	36	0	0	66	98	3	3		432
DN4_WT_mouse67	0	18	70	30	63	0	34	44	0	1	50	118	1	3		432
DN4_WT_mouse72	1	36	59	39	53	0	35	41	2	0	71	104	1	5		447
DP_WT_mouseWT	0	91	167	89	163	0	81	99	4	5	169	239	13	5		1125
DP_WT_mouse2	2	34	57	25	37	0	22	19	1	1	59	64	4	2		327
DP_WT_mouse3	0	26	35	34	35	0	22	15	1	0	54	53	4	0		279
DP_WT_mouse4	0	29	60	24	39	0	11	18	0	2	60	57	1	6		307
DP_WT_mouse6	0	46	48	34	50	0	21	24	0	0	59	75	3	0		360
lateDP_WT_mouse20	0	27	67	42	63	0	25	46	0	0	76	111	6	9		472
lateDP_WT_mouse25	1	18	58	29	52	0	41	53	1	2	53	143	1	11		463
lateDP_WT_mouse57	0	16	44	30	57	0	31	36	3	2	44	108	5	3		379
lateDP_WT_mouse77	0	18	54	38	81	0	37	48	0	4	67	139	1	6		493
count	D1	D2	D1	D1	D2	D2	uV	uV	uV31	uV31	pV	pV	pV31	pV31		total
	germ	germ	J1	J2	J2	J1	J1	J2	J1	J2	J1	J2	J1	J2		
DN3a	120	321	359	176	96	2	102	80	4	5	31	23	1	0		1320
DN4	1	98	245	134	222	0	120	149	6	2	246	427	7	15		1672
DP	2	226	367	206	324	0	157	175	6	8	401	488	25	13		2398
late DP	1	79	223	139	253	0	134	183	4	8	240	501	13	29		1807
%	D1	D2	D1	D1	D2	D2	uV	uV	uV31	uV31	pV	pV	pV31	pV31		total
	germ	germ	J1	J2	J2	J1	J1	J2	J1	J2	J1	J2	J1	J2		
DN3a	9%	24%	27%	13%	7%	0%	8%	6%	0%	0%	2%	2%	0%	0%		
DN4	0%	6%	15%	8%	13%	0%	7%	9%	0%	0%	15%	26%	0%	1%		
DP	0%	9%	15%	9%	14%	0%	7%	7%	0%	0%	17%	20%	1%	1%		
late DP	0%	4%	12%	8%	14%	0%	7%	10%	0%	0%	13%	28%	1%	2%		

b



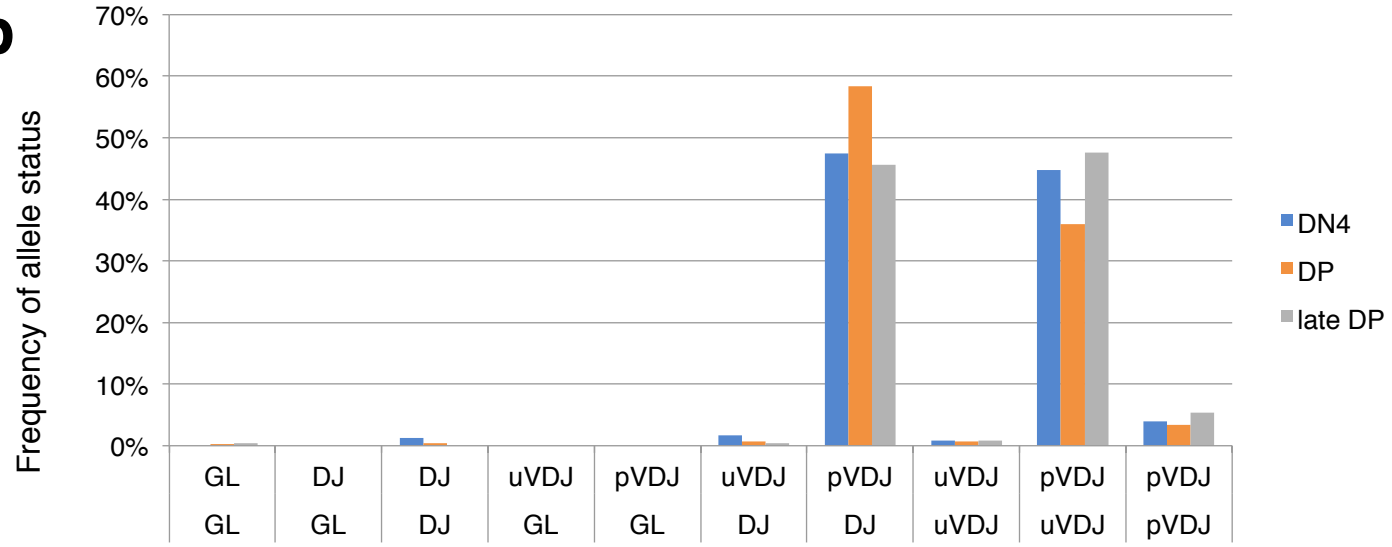
Supplementary Figure S2

Tags for PCR fragments. Number (**a**) and frequency (**b**) of cluster tags recovered from wild type single thymocytes at DN3a, DN4, DP and late DP stages.

a

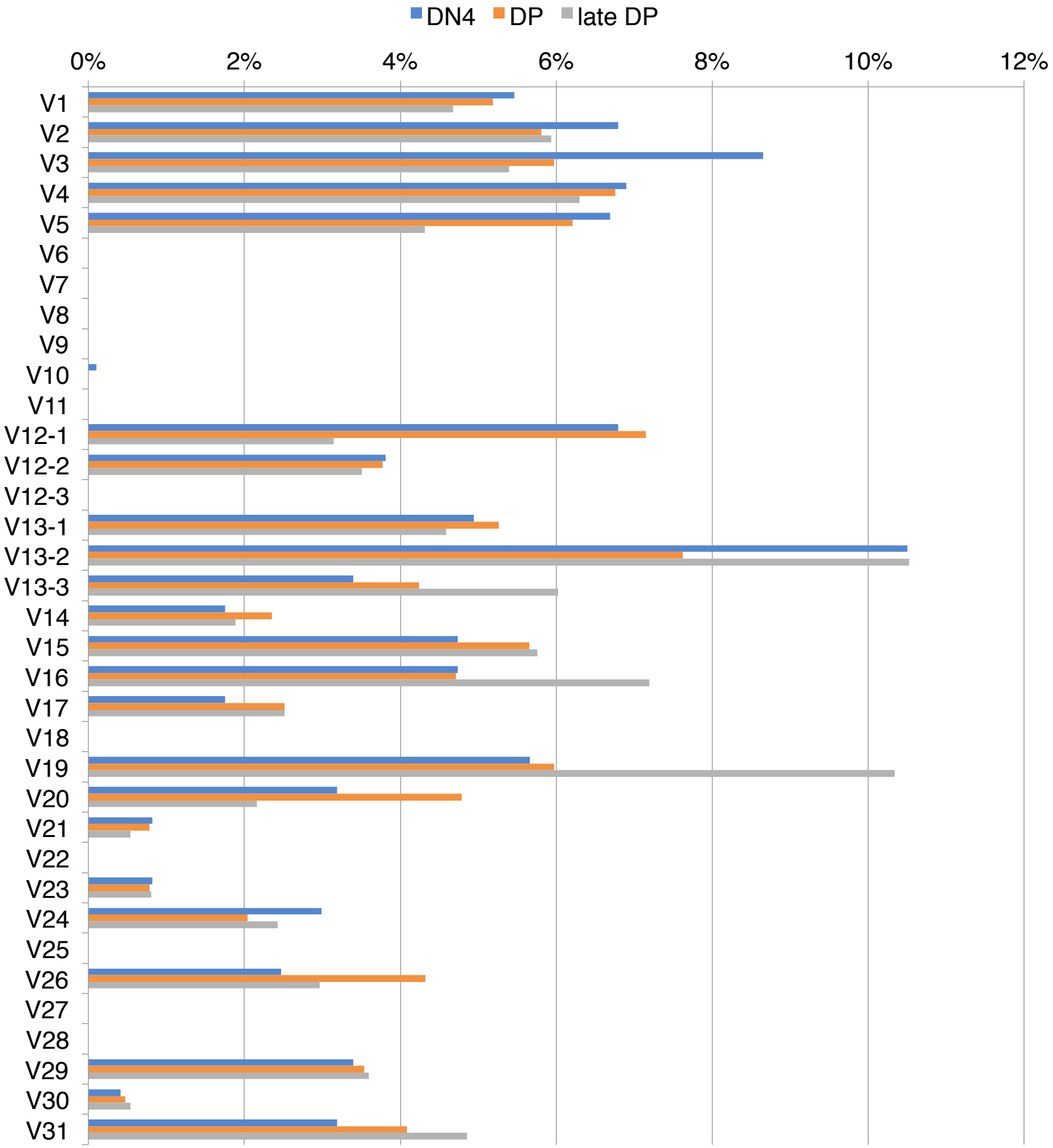
sampleID	GL GL	DJ GL	DJ DJ	uVDJ GL	pVDJ GL	uVDJ DJ	pVDJ DJ	uVDJ uVDJ	pVDJ uVDJ	pVDJ pVDJ	ID12 to 15		total
DN3a_WT_mouse2	33	10	15	0	0	6	4	2	1	0	0		71
DN3a_WT_mouse3	11	1	10	1	0	16	3	3	0	0	0		45
DN3a_WT_mouse4	8	7	12	0	0	7	3	3	0	0	0		40
DN3a_WT_mouse6	16	2	7	0	1	4	4	2	2	0	0		38
DN4_WT_mouse62	0	0	0	0	0	1	20	0	17	3	0		41
DN4_WT_mouse64	0	0	2	0	0	1	26	0	30	3	0		62
DN4_WT_mouse67	0	0	0	0	0	1	28	1	24	2	0		56
DN4_WT_mouse72	0	0	1	0	0	1	34	1	31	1	0		69
DP_WT_mouseWT	0	0	1	0	0	2	122	3	96	9	0		233
DP_WT_mouse2	1	0	1	0	0	0	29	0	15	2	1		49
DP_WT_mouse3	0	0	0	0	0	0	21	0	8	1	0		30
DP_WT_mouse4	0	0	0	0	0	1	36	0	11	0	0		48
DP_WT_mouse6	0	0	0	0	0	0	37	0	21	2	0		60
lateDP_WT_mouse20	0	0	0	0	0	0	38	0	27	1	0		66
lateDP_WT_mouse25	1	0	0	0	0	0	25	1	37	3	0		67
lateDP_WT_mouse57	0	0	0	0	0	0	18	0	27	1	0		46
lateDP_WT_mouse77	0	0	0	0	0	1	38	1	33	9	0		82
count	GL GL	DJ GL	DJ DJ	uVDJ GL	pVDJ GL	uVDJ DJ	pVDJ DJ	uVDJ uVDJ	pVDJ uVDJ	pVDJ pVDJ	ID12 to 15		total
DN3a	68	20	44	1	1	33	14	10	3	0	0		194
DN4	0	0	3	0	0	4	108	2	102	9	0		228
DP	1	0	2	0	0	3	245	3	151	14	1		420
late DP	1	0	0	0	0	1	119	2	124	14	0		261
%	GL GL	DJ GL	DJ DJ	uVDJ GL	pVDJ GL	uVDJ DJ	pVDJ DJ	uVDJ uVDJ	pVDJ uVDJ	pVDJ pVDJ	ID12 to 15		total
DN3a	35%	10%	23%	1%	1%	17%	7%	5%	2%	0%	0%		
DN4	0%	0%	1%	0%	0%	2%	47%	1%	45%	4%	0%		
DP	0%	0%	0%	0%	0%	1%	58%	1%	36%	3%	0%		
late DP	0%	0%	0%	0%	0%	0%	46%	1%	48%	5%	0%		

b



Supplementary Figure S3

VDJ rearrangement status of both *Trb* alleles. *Trb* VDJ rearrangement status for both alleles calculated from wild type single thymocytes at DN3a, DN4, DP and late DP stages. Number (**a**) and frequency (**b**) are shown.



Supplementary Figure S4

V gene segment usage in wild type single thymocytes at DN4, DP and late DP stages. Frequency was calculated from 971 (DN4), 1,273 (DP) and 1,112 (late DP4) VDJ sequence recovered for each stage.