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

Tomonori Hosoya[1,3], Hongyang Li[2,3], Chia-Jui Ku[1], Qingqing Wu[1], Yuanfang Guan[2] and James Douglas Engel[1]*

[1]Department of Cell and Developmental Biology

[2]Department of Computational Medicine and Bioinformatics

University of Michigan

3035 BSRB

109 Zina Pitcher Place

Ann Arbor, Michigan 48109-2200

[3]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

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 *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(Tonegawa 1983). 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(Mostoslavsky et al. 2004), while at the level of protein localization on the cell surface for TCRα(Alam and Gascoigne 1998). Loss of allelic exclusion results in dual-TCR expression, which can lead to autoimmunity(Hinz et al. 2001; Auger et al. 2012).

A high-throughput method to study the diversity of lymphocytes at the single-cell level has been reported recently(Han et al. 2014), 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(Weischenfeldt et al. 2008; Dash et al. 2011), 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, which would alter the interpretation of TCR locus activity. Additionally, no reference genome is available for merging multiple sequencing reads, due to the many millions of possible sequences that can be generated from VDJ rearrangement(Robins et al. 2009). Thus, *de novo* sequence assembly or a long-read approach is required to retrieve the original genomic sequence of both alleles in single cells. To significantly improve the traditional Sanger sequencing method employed by us and others(Aifantis et al. 1997; Aifantis et al. 1999; Ku et al. 2017), we developed a high-throughput method that enables analysis of *Trb* allelic exclusion status by sequencing both alleles of the genome in single cells, enabling us to determine whether each allele in those cells underwent either no, unproductive or productive rearrangement.

RESULTS

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

To sequence genomic DNA in the *Trb* gene VDJ region, we employed a multiplex strategy (Fig. 1a; details are described in Methods). Single cells were directly sorted into lysis buffer in individual wells of 96-well PCR plates. Multiplex, nested PCR reactions followed by barcoding PCR amplified the germline (GL: unrearranged configuration) genomic DNA, D-to-J rearranged (DJ) DNA, and all potential V-to-DJ rearranged (VDJ) DNA in the *Trb* locus(Bosc and Lefranc 2000) (Figs. 1b and 1c). The PCR products were mixed and then sequenced in a PacBio RS II long-read sequencer (www.pacb.com). Each circular consensus specification (CCS) read of an insert generated by the PacBio system

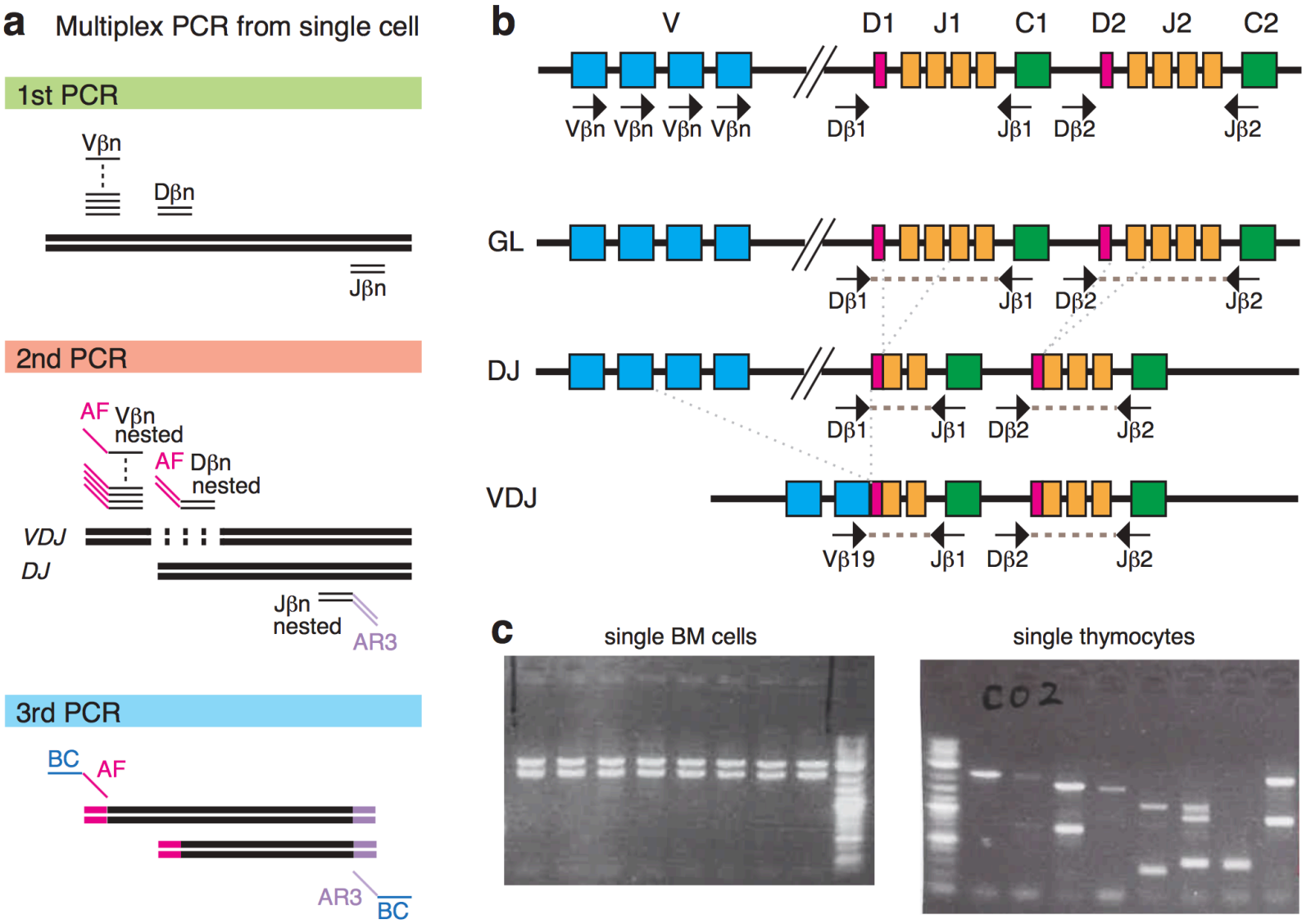
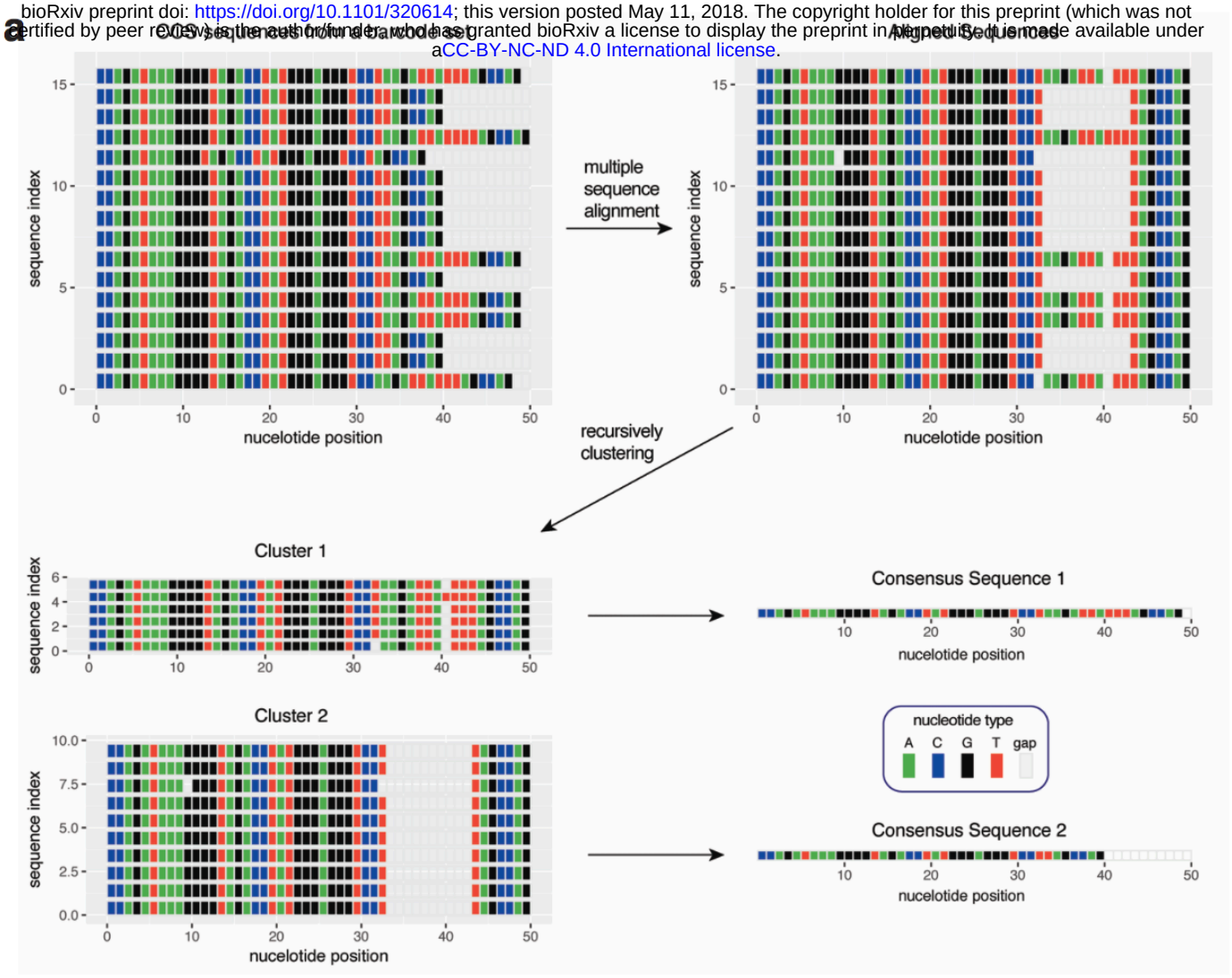
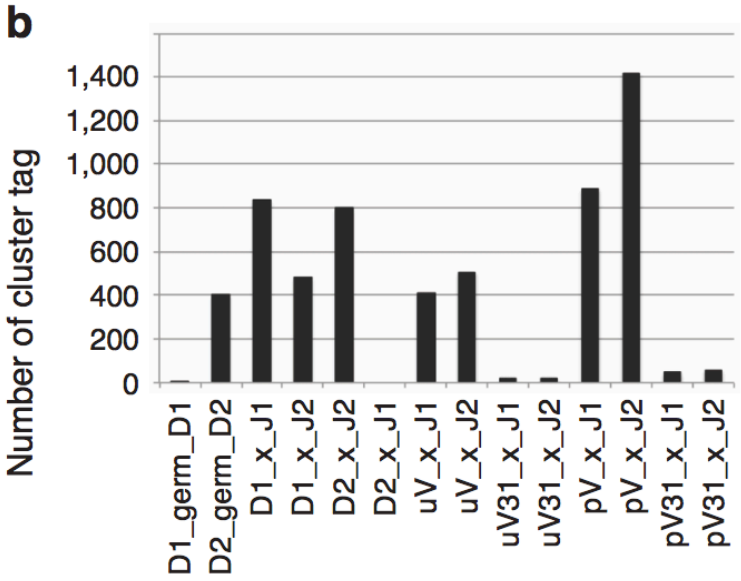


Fig. 1. High-throughput *Trb* genomic DNA sequencing from single-cells. (**a**) Strategy for multiplex PCR amplification of genomic DNA sequences at the *Trb* gene VDJ loci from single cells. Final PCR products were mixed and then sequenced by PacBio RS II sequencer. Vβn: V region primers. Dβn: D region primers. Jβn: J region primers. AF: adapter forward. AR: adapter reverse. BC: barcode. Details are described in Methods. (**b**) Schematic presentation of VDJ rearrangement at the *Trb* locus with simplified illustration of multiplex primer location. The genomic DNA recombination events are initiated by recombining Dβ (diversity) and Jβ (joining) segments on both chromosomes. Subsequently, on only one chromosome, one of the Vβ (variable) segments is joined to the previously rearranged DJ recombinant. Not only the selection of each VDJ segment, but also multiple lengths of spacer sequence between V, D and J, generates an incalculable number of VDJ sequence possibilities. mRNA splicing joins Cβ (constant) segments to the rearranged VDJ recombinant to generate the final TCRβ protein. VDJ rearrangement that generates a stop codon or elongated transcript results in a predictedly unproductive TCRβ. (**c**) Successful amplification of rearranged VDJ, DJ and germline D-to-J regions were confirmed by agarose gel electrophoresis with NEB 2-log DNA ladder in the last (left panel) and first (right panel) lanes.

corresponds to a single molecule of the original PCR product (Supplementary Fig. S1). Based on barcoding and sequence similarity, multiple CCSs were grouped into clusters (Fig. 2a). For each cluster, the consensus sequence was retrieved, which corresponds to an original PCR product. The consensus sequences were subsequently submitted to IMGT HighV-Quest analysis(Li et al. 2013) (http://www.imgt.org/) to analyze whether the VDJ sequences in any cluster predicted that a functional TCRβ protein was generated. From one *Trb* allele, two (D1-to-J1 and D2-to-J2 related) or one (D1-to-J2 rearrangement related) PCR products are amplified (Supplementary Fig. S2 and S3). Therefore, two to four PCR products (clusters) are recovered from two alleles in each cell (Supplementary Fig. S4). Details for the computational processing of the PacBio output data to calculate allele rearrangement status are described in Methods.

To calculate the PCR error rate under our conditions, we took advantage of the unrearranged (germline configuration) genomic DNA corresponding to parts of *Trbd1* to *Trbj1-7* and *Trbd2* to *Trbj2-7* (Fig. 1b) in single bone marrow (BM) cells. From single BM cells we obtained two PCR bands (Fig. 1c), which correspond to the lengths of germline DNA amplified using Dβ1 forward and Jβ1 reverse nested primers (2,672 bp) as well as Dβ2 forward and Jβ2 reverse nested primers (1,896 bp). The accuracy of the resultant sequences was 99.768% for each individual CCS. When a consensus was generated from 3 randomly chosen synonymous CCSs, the accuracy increased to 99.991%. These data demonstrate that our method using PCR with Q5® High-Fidelity DNA polymerase (NEB) followed by calculating the consensus sequence from multiple CCSs yields >99.99% accuracy, even after a total of 80 PCR cycles.





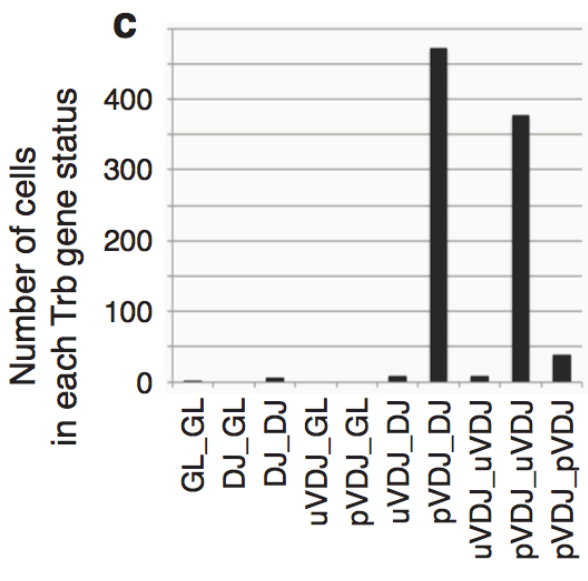


Fig. 2. *Trb* VDJ rearrangement status in single thymocytes. (**a**) A schematic outline of the PacBio raw data processing. Details are described in Methods. (**b**) Cluster tags in a total of 5,877 clusters recovered from single wild type thymocytes at DN4, DP and late DP (post β-selection) stages. Each cluster consensus sequence was categorized into one tag shown on the x-axis (see Supplementary Fig. S3 for detail). (**c**) Calculated *Trb* allele status for 909 single wild type thymocytes at DN4, DP and late DP (post b-selection) stages. Legend: GL: germline allele status; DJ: D-to-J recombined allele; pVDJ: V-to-DJ recombined allele encoding a productive TCRβ protein; uVDJ: V-to-DJ recombined allele encoding a predicted unproductive TCRβ. Possible combination of cluster tags per cell for each VDJ rearrangement status (total 252 different tag combinations) is shown in Supplementary Fig. S4. The data in (**b**) and (**c**) represent a summary of 4 mice for DN4, 5 mice for DP and 4 mice for late DP stages, and details are shown in Supplementary Figs. S5 and S6.

Trb VDJ rearrangement status in single thymocytes

To precisely determine the genomic status at the *Trb* loci following VDJ rearrangement, we analyzed wild type DN4, DP and late DP stage thymocytes, in which *Trb* VDJ rearrangement had already been completed (i.e. post-β-selection). We obtained a total of 5,877 clusters and recovered the rearrangement status of both alleles from 909 single cells (Fig. 2c, Supplementary Figs. S5 and S6). 52% (472) of the single thymocytes were in a pVDJ (productive VDJ rearranged)/DJ configuration and 41% (377) were in the uVDJ (unproductive VDJ rearranged)/pVDJ configuration, as expected from the 60/40-rule reported previously(Mostoslavsky et al. 2004). 4% (37) of the post-β-selection thymocytes were in a pVDJ/pVDJ arrangement, which was similar to the 3% dual-TCRβ bearing cells as analyzed using cell surface antibodies(Balomenos et al. 1995). The V regions that were most frequently used were V13-2 (Supplementary Fig. S7), which is also the most frequently used in inflamed lung epitope specific CD8+ T cells(Dash et al. 2011). 6,415 V-to-DJ rearranged genomic DNA and complementarity determining region 3 (CDR3) amino acid sequences (4,258 productive and 2,158 unproductive) recovered from single thymocytes in this analysis are available online (see Data Access for detail). Surprisingly we recovered D2-to-J1 rearranged sequences, although only rarely: only 6 out of 7,113 D-to-J rearranged sequences we have recovered were D2-to-J1 recombinants (Supplementary Fig. S8). At the double negative 3a (DN3a) stage, when V-to-DJ rearrangement takes place, only 9% of the cells had recombined to generate a productive (pVDJ) allele [0.5% (pVDJ/GL), 7.2%, (pVDJ/DJ) and 1.5% (uVDJ/pVDJ), Supplementary Fig. S6] probably representing cells predicted to be about to develop to the next (DN3b) stage.

Next, to determine whether the methods developed here can capture an artificial allelic exclusion phenomenon, we analyzed late DP cells isolated from Vb8 transgenic mice(Shinkai et al. 1993) (Tg^{Vb8}), which express a productively rearranged TCRβ protein, which has been shown to repress VDJ rearrangement at the endogenous *Trb* loci(Uematsu et al. 1988). Out of 2,429 clusters generated, 722 clusters bore the DNA sequence of the Tg^{Vb8} transgene (Supplementary Fig. S9) and were removed from the analysis. We then determined the rearrangement status of both alleles from 127 single cells. V-to-DJ rearrangement was not observed at the endogenous locus in 80% of the cells, in agreement with published observations(Uematsu et al. 1988). However, 20% of the cells had either unproductively (13% uVDJ/DJ) or productively (6% pVDJ/DJ) escaped allelic exclusion to allow V-to-DJ rearrangement at one of the endogenous *Trb* loci.

DISCUSSION

In summary, the method developed here provides a high-throughput approach to analyze rearranged *Trb* genomic DNA status at the single cell level. This strategy allowed us to obtain 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(Aifantis et al. 1997; Aifantis et al. 1999; Ku et al. 2017) 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 allelic exclusion at the genetic level(Mostoslavsky et al. 2004).

METHODS

Single cell isolation. Staged thymocytes were isolated using a Synergy cell sorter (Sony iCyt SY3200). $Lin^-CD4^-CD8a^-$ $Thy1.2^+cKit^-CD25^+CD28^-$ (DN3a stage), $Lin^-CD3^-TCR\beta^-CD8a^-Thy1.2^+cKit^-CD25^-$ (DN4 stage), $CD4^+CD8^+$ (DP stage) and $TCR\beta^+CD4^+CD8^+$ (late DP stage) thymocytes were isolated. Single cells were directly sorted into 20 μl of lysis buffer [containing 1x Q5® Reaction Buffer (NEB), 4 μg Proteinase K and 0.1% Triton X100] in one well of a 96 well PCR plate. The cell sorter setting was carefully adjusted so that sorted cells were deposited into the exact middle of each well. Sorted single cells in the lysis solution were kept on ice, and then digested at 55ºC for 60 minutes followed by 95ºC for 15 minutes (to inactivate Proteinase K) using a PCR thermal cycler within 6 hours after sorting.

Multiplex nested PCR. In the first round of PCR, primers were selected to amplify all potential VDJ rearrangement at the *Trb* locus: 31 V region primers covering all 35 *Trbv* genes(Bosc and Lefranc 2000), 2 D region primers, 2 J region primers and 2 control primers used to detect sequences 3' of the *Actb* gene (Fig. 1a, 1b and Supplementary Fig. S10). The first round of PCR was performed in a 60 μl final reaction volume containing of 50 nM of each primer, 1x Q5® Reaction Buffer, 200 μM each dNTPs, 0.4 unit Q5® High-Fidelity DNA Polymerase (NEB) and 20 μl of the lysed single cell solution. The PCR condition was 30 sec at 98 ºC followed by 30 cycles of 5 sec at 98 ºC, 10 sec at 66-58 ºC and 2 min at 72 ºC, and then final extension for 2 min at 72 ºC. During the first 5 cycles, the annealing temperature was reduced 2 ºC per cycle from 66 ºC to 58 ºC, and then performed at 56ºC for the last 25 cycles.

In the second round of PCR, nested primers were selected: 32 nested V region primers containing forward adapter sequences (AF-Vβn), 2 nested D region forward adapter primers (AF-Dβn) and 2 nested J region primers in a reverse adapter orientation (AR3-Jβn). The second round of PCR was performed in a 20 μl final reaction volume containing 50 nM in each primer (Fig. 1a and Supplementary Fig. S10), 1x Q5® buffer, 200 μM of each dNTP, 0.4 units Q5® High-Fidelity DNA Polymerase (NEB) and 1 μl of the first round PCR product. The 2nd PCR condition was same as for the first round PCR but for only 25 cycles.

Finally, in the third round of PCR, unique barcoded-AF and barcoded-AR3 combinations of primers were selected for each single cell (Fig. 1a and Supplementary Fig. S11). Barcodes were adopted from a published report(Han et al. 2014). The 3rd PCR condition was same as for the second round PCR (5 + 20 cycles). Hot start PCR was performed for all PCR reactions using either a BioRad T100™ or Applied Biosystems 2720 Thermal Cycler. All primers were purchased from Integrated DNA Technologies, Inc. The 37 (for 1st round) or 36 (for 2nd round) PCR primers at 200 μM final concentrations were mixed, aliquoted and frozen for subsequent use. Each PCR mix was prepared immediately before initiating PCR reactions. (The amplification efficiency was reduced if the PCR mix was prepared in advance and repeatedly frozen and thawed. Amplification efficiency was also reduced when 500 nM of each primer was used in the 2nd round PCR reaction). Successful DNA amplification was confirmed by running 2 μl of 3rd round PCR product on agarose gels with 0.4 μg of 2-Log DNA Ladder (NEB) (Fig. 1c): 8 samples were selected from each 96 well PCR plate. Of note, the PCR

conditions employed here allow amplification of unrearranged (germline configuration) genomic DNA (Fig. 1c left panel).

Recovery frequency

To analyze the frequency of recovered wells, a primer pair amplifying the 3' region of the *Actb* gene was added to the 1st round of PCR (). 1 μl of the 1st round PCR product was used in a PCR reaction with nested *Actb* primers: 5'-CATAGGCTTCACACCTTCCT-3', 5'-CTTTGCCTCCATCTGCATAAC-3' and FAM-labeled probe TGCTAGTCTGAAGCTGCCCTTTCC (ZEN / Iowa Black FQ) purchased from Integrated DNA Technologies, Inc. Luna Universal Probe qPCR Master Mix (NEB, M3004) was used in a 20 μl reaction. The PCR condition was 30 sec at 95 ºC followed by 45 cycles of 1 sec at 95 ºC and 20 sec at 60 ºC using StepOnePlus™ Real-Time PCR System (Thermo Fisher Scientific). A 90-95% recovery efficiency was routinely obtained from each plate containing 96 single cell sorted wells.

PacBio high-throughput sequencing

Since the length of the PCR products described above ranged from 200 to 2700 bp (Fig. 1b), a sequencer with the ability to read long DNA fragments is preferred. To this end, Pacific Biosciences single molecule real-time (SMRT) sequencing technology(Roberts et al. 2013) was employed. A PacBio smartbell adapter was ligated to the barcoded PCR product mixture and then sequenced on PacBio RSII sequencer at the University of Michigan Sequencing Core. Each circular consensus sequencing (CCS) read generated

from PacBio sequencing corresponds to a single molecule of initial PCR product (Supplementary Fig. S1).

PacBio raw reads analysis

To obtain DNA sequences amplified from genomic DNA in single cells, we developed an in-house analysis pipeline (Fig. 2a, Supplementary Figs. S12 and 13; the full version is available at https://github.com/Hongyang449/scVDJ_seq). To calculate CCS reads of inserts, PacBio ConsensusTools software version 2.0.0 was used with the following settings: --minFullPasses=2, --minPredictedAccuracy=90, --minLength=10 and default for other parameters (https://github.com/PacificBiosciences/SMRT-Analysis/wiki/Documentation).

First, to demultiplex CCS reads, we grouped the CCS reads based on the presence of a barcode primer set flanking each end of the CCS. We obtained approximately 35,000 CCSs per PacBio RS II SMRTcell on average (Supplementary Fig. S1b), and 46% of them started with a barcode-AF primer and ended with a barcode-AR3 primer. The remaining 54% had mutations, deletions, or additions within the barcoded primer sequence or bore truncated barcode primer sequences, and were excluded from the analysis.

Next, we generated sub-groups in each single cell (unique barcode set) based on the presence of a V or D primer sequence, and sequences in each sub-group were aligned using multiple sequence alignment software MUSCLE(Edgar 2004) (http://www.drive5.com/muscle/). For each group, hierarchical clustering based on sequence length was performed to create initial clusters, then k-means clustering was

recursively performed on the DNA sequences of each initial cluster to generate the final multiple clusters, until every nucleotide position in each cluster achieved >51% identity. Clusters containing only 1 CCS read were excluded. Aberrant CCSs with two PCR products ligated together probably resulted during the smartbell adapter ligation step, and were also excluded from the analysis.

To obtain accurate genomic DNA sequences, we calculated a consensus sequence from multiple CCS file alignments in each cluster. Of note, each CCS file that was generated from single zero-mode waveguide (ZMW) corresponds to one copy of PCR product. If a mutation happened during the first PCR cycle, 50% of the final PCR products inherit the mutation at a specific base position in the final PCR product. Similarly, mutation during the second cycle is inherited in 25% of the final PCR product, 12.5% at third cycle and so on. After adapter and extra J region genomic DNA sequences were removed (Supplementary Fig. S14), cluster consensus with V primer sequences were submitted to IMGT HighV-Quest analysis(Li et al. 2013) (http://www.imgt.org/) to determine whether the VDJ sequence in a cluster was predicted to encode a productive or unproductive TCRβ protein. The removal of the extra J region genomic DNA is essential to accurately determine the productivity of rearrangement, especially when some bases are truncated from J regions.

Based on sequence and functionality analyzed by IMGT, each cluster consensus sequence was categorized into a tag shown in Supplementary Fig. S3a. Since the status of an allele (GL, DJ, uVDJ and pVDJ) can be determined from possible combinations of 1-2 cluster tags (total 24 different pattern shown in Supplementary Fig. S3b), the rearrangement status of a single cell (combination of two allele patterns) were calculated

from 2-4 tags recovered from each cell (for a total 253 different combinations of tags shown in Supplementary Fig. S4). From those, the summary tables shown in Supplementary Fig. S13 were generated. Unreliable clusters that had <51% nucleotide identity after removal of extra J regions were excluded from the table.

DATA ACCESS

PacBio CCS data, barcode list for each sample, genomic DNA and CDR3-IMGT amino acid sequences for all clusters generated through this study are deposited at (https://umich.box.com/s/3f5q64u2i68dn2i7hlneucxph9oetpvb).

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

T.H. designed the study, performed experiments, wrote the computer scripts, analyzed the data, and wrote the manuscript. H.L. designed the study, wrote the computer scripts and edited the paper. C.K. designed the study, performed experiments, analyzed the data and edited the manuscript. Q.W performed experiments and edited the paper. Y.G. designed the study and edited the paper. J.D.E. designed the study and edited the paper.

DISCLOSURE DECLARATION

The authors declare no competing interests.

FIGURE LEGENDS

Fig. 1. High-throughput *Trb* genomic DNA sequencing from single-cells. (**a**) Strategy for multiplex PCR amplification of genomic DNA sequences at the *Trb* gene VDJ loci from single cells. Final PCR products were mixed and then sequenced by PacBio RS II sequencer. Vβn: V region primers. Dβn: D region primers. Jβn: J region primers. AF: adapter forward. AR: adapter reverse. BC: barcode. Details are described in Methods. (**b**) Schematic presentation of VDJ rearrangement at the *Trb* locus with simplified illustration of multiplex primer location. The genomic DNA recombination events are initiated by recombining Dβ (diversity) and Jβ (joining) segments on both chromosomes. Subsequently, on only one chromosome, one of the Vβ (variable) segments is joined to the previously rearranged DJ recombinant. Not only the selection of each VDJ segment, but also multiple lengths of spacer sequence between V, D and J, generates an incalculable number of VDJ sequence possibilities. mRNA splicing joins Cβ (constant) segments to the rearranged VDJ recombinant to generate the final TCRβ protein. VDJ rearrangement that generates a stop codon or elongated transcript results in a predictedly unproductive TCRβ. (**c**) Successful amplification of rearranged VDJ, DJ and germline D-to-J regions were confirmed by agarose gel electrophoresis with NEB 2-log DNA ladder in the last (left panel) and first (right panel) lanes.

Fig. 2. *Trb* VDJ rearrangement status in single thymocytes. (**a**) A schematic outline of the PacBio raw data processing. Details are described in Methods. (**b**) Cluster tags in a total of 5,877 clusters recovered from single wild type thymocytes at DN4, DP and late DP (post β-selection) stages. Each cluster consensus sequence was categorized into one tag shown on the x-axis (see Supplementary Fig. S3 for detail). (**c**) Calculated *Trb* allele status for 909 single wild type thymocytes at DN4, DP and late DP (post β-selection) stages. Legend: GL: germline allele status; DJ: D-to-J recombined allele; pVDJ: V-to-DJ recombined allele encoding a productive TCRβ protein; uVDJ: V-to-DJ recombined allele encoding a predicted unproductive TCRβ. Possible combination of cluster tags per cell for each VDJ rearrangement status (total 252 different tag combinations) is shown in Supplementary Fig. S4. The data in (**b**) and (**c**) represent a summary of 4 mice for DN4, 5 mice for DP and 4 mice for late DP stages, and details are shown in Supplementary Figs. S5 and S6.

REFERENCES

Aifantis I, Buer J, von Boehmer H, Azogui O. 1997. Essential role of the pre-T cell receptor in allelic exclusion of the T cell receptor beta locus. *Immunity* **7**: 601-607.

Aifantis I, Pivniouk VI, Gartner F, Feinberg J, Swat W, Alt FW, von Boehmer H, Geha RS. 1999. Allelic exclusion of the T cell receptor beta locus requires the SH2 domain-containing leukocyte protein (SLP)-76 adaptor protein. *J Exp Med* **190**: 1093-1102.

Alam SM, Gascoigne NR. 1998. Posttranslational regulation of TCR Valpha allelic exclusion during T cell differentiation. *J Immunol* **160**: 3883-3890.

Auger JL, Haasken S, Steinert EM, Binstadt BA. 2012. Incomplete TCR-beta allelic exclusion accelerates spontaneous autoimmune arthritis in K/BxN TCR transgenic mice. *Eur J Immunol* **42**: 2354-2362.

Balomenos D, Balderas RS, Mulvany KP, Kaye J, Kono DH, Theofilopoulos AN. 1995. Incomplete T cell receptor V beta allelic exclusion and dual V beta-expressing cells. *J Immunol* **155**: 3308-3312.

Bosc N, Lefranc MP. 2000. The mouse (Mus musculus) T cell receptor beta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. *Exp Clin Immunogenet* **17**: 216-228.

Dash P, McClaren JL, Oguin TH, 3rd, Rothwell W, Todd B, Morris MY, Becksfort J, Reynolds C, Brown SA, Doherty PC et al. 2011. Paired analysis of TCRalpha and TCRbeta chains at the single-cell level in mice. *J Clin Invest* **121**: 288-295.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

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

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

Ku CJ, Sekiguchi JM, Panwar B, Guan Y, Takahashi S, Yoh K, Maillard I, Hosoya T, Engel JD. 2017. GATA3 Abundance Is a Critical Determinant of T Cell Receptor beta Allelic Exclusion. *Mol Cell Biol* **37**.

Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P, Freeman JD, Corbin VD, Scheerlinck JP, Frohman MA et al. 2013. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* **4**: 2333.

Mostoslavsky R, Alt FW, Rajewsky K. 2004. The lingering enigma of the allelic exclusion mechanism. *Cell* **118**: 539-544.

Roberts RJ, Carneiro MO, Schatz MC. 2013. The advantages of SMRT sequencing. *Genome Biol* **14**: 405.

Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**: 4099-4107.

Shinkai Y, Koyasu S, Nakayama K, Murphy KM, Loh DY, Reinherz EL, Alt FW. 1993. Restoration of T cell development in RAG-2-deficient mice by functional TCR transgenes. *Science* **259**: 822-825.

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

Uematsu Y, Ryser S, Dembic Z, Borgulya P, Krimpenfort P, Berns A, von Boehmer H, Steinmetz M. 1988. In transgenic mice the introduced functional T cell receptor beta gene prevents expression of endogenous beta genes. *Cell* **52**: 831-841.

Weischenfeldt J, Damgaard I, Bryder D, Theilgaard-Monch K, Thoren LA, Nielsen FC, Jacobsen SE, Nerlov C, Porse BT. 2008. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. *Genes Dev* **22**: 1381-1396.