T cell fate and clonality inference from single-cell transcriptomes

Michael J T Stubbington[1,3], Tapio Lönnberg[1,3], Valentina Proserpio[1], Simon Clare[2], Anneliese O Speak[2], Gordon Dougan[2] & Sarah A Teichmann[1,2]

We developed TraCeR, a computational method to reconstruct full-length, paired T cell receptor (TCR) sequences from T lymphocyte single-cell RNA sequence data. TraCeR links T cell specificity with functional response by revealing clonal relationships between cells alongside their transcriptional profiles. We found that T cell clonotypes in a mouse *Salmonella* infection model span early activated CD4$^+$ T cells as well as mature effector and memory cells.

T lymphocytes recognize specific peptide–major histocompatibility complex (pMHC) combinations presented on the surface of antigen-presenting cells. This highly specific recognition is mediated by the TCR, an extremely diverse heterodimeric cell-surface protein comprising an α- and a β-chain encoded by genes produced by recombination of V(D)J loci during T cell development. The DNA sequence diversity of mouse TCRs has been estimated as 5×10^{21} different paired combinations[1], allowing us to assume that cells with identical paired TCR genes arose from the same T cell clone.

The diversity of individual TCR chains has been used as a proxy for overall clonal diversity within bulk populations of T lymphocytes[2,3], but these studies cannot determine the paired chains within each cell. This limits their ability to perform high-resolution determination of clonal relationships between cells and also to draw conclusions about the antigenic specificities of the cells.

Paired TCR analysis can better discern the 'grammar' of TCR recognition and aid in the design of therapeutic TCR molecules. This has been performed in individual cells using specific amplification or capture of TCR genes[4–6], but these methods do not provide additional information about queried cells. In addition, biases in PCR primer efficiency prevent the accurate determination of TCR expression levels. One method amplifies a small set of 'phenotyping marker' genes[7] in addition to paired TCR profiling, providing limited information about cellular phenotype. However, it requires *a priori* knowledge of informative genes as well as the design and optimization of multiplexed PCR primers.

To connect transcriptional status with antigen specificity, model the dynamics of clonal expansion within T cell populations and investigate T cell phenotypic plasticity, it will be critical to link TCR sequence with transcriptional profiles in individual cells. Single-cell RNA-seq (scRNA-seq) has already proven valuable in investigating the transcriptional heterogeneity and differentiation processes of cell populations[8–10] and has revealed a novel T lymphocyte subset[11]. However, recombined TCR sequences have yet to be reported from T cell scRNA-seq data sets.

Existing computational tools for TCR analysis are designed for experiments that use bulk cell populations and require the targeted amplification of TCR loci[12–15]. Here, we present a novel method that enables full-length, paired TCR sequences to be reconstructed from single-cell RNA-seq data with high accuracy and sensitivity. Importantly, this method requires no alterations to standard scRNA-seq protocols and so can be easily applied to any species and sample for which scRNA-seq is possible.

Our TCR reconstruction tool, TraCeR (**Supplementary Fig. 1a**, **Supplementary Software** and https://www.github.com/teichlab/tracer), extracts TCR-derived sequencing reads for each cell by alignment against 'combinatorial recombinomes' comprising all possible combinations of V and J segments (**Supplementary Fig. 1b**). Reads are then assembled into contiguous sequences that are analyzed to find full-length, recombined TCR sequences. Importantly, the reconstructed recombinant sequences typically contain nearly the complete length of the TCR V(D)J region (**Fig. 1**) and so allow high-confidence discrimination between closely

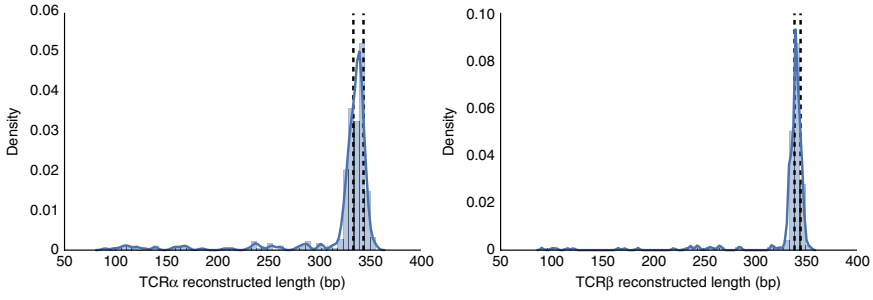

Figure 1 | Length distributions of reconstructed TCR sequences. Lengths of reconstructed sequences trimmed to include the V gene, junction and J gene are plotted as histograms and kernel density estimates. Dashed lines represent the interquartile range of lengths of full-length sequences derived from the combinatorial recombinome files.

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [2]Wellcome Trust Sanger Institute, Cambridge, UK. [3]These authors contributed equally to this work. Correspondence should be addressed to S.A.T. (saraht@ebi.ac.uk).

RECEIVED 28 AUGUST 2015; ACCEPTED 25 JANUARY 2016; PUBLISHED ONLINE 7 MARCH 2016; DOI:10.1038/NMETH.3800

Table 1 | TCR reconstruction statistics

Mouse	TCRα reconstruction(%)	TCRβ reconstruction(%)	Paired productive chains(%)
Uninfected, day 0	39/50 (78)	46/50 (92)	37/50 (74)
Day 14, mouse 1	68/71 (96)	68/71 (96)	66/71 (93)
Day 14, mouse 2	29/39 (74)	35/39 (90)	28/39 (72)
Day 49	87/112 (78)	98/112 (88)	78/112 (70)

related gene segments. We present an analysis of data generated using the SMART-seq protocol[16] with the Fluidigm C1 microfluidics system, though our method works with any scRNA-seq data derived from full-length cDNA.

We analyzed scRNA-seq data from 272 FACS-sorted CD4$^+$ T cells isolated from spleens of C57BL/6N mice (**Supplementary Table 1** and **Supplementary Fig. 2**). We detected at least one productive α-chain in 74–96% of cells, a productive β-chain in 88–96% and paired productive α-β–chains in 70–93% (**Table 1** and **Supplementary Table 2**). This compares favorably with previous PCR-based TCR sequencing approaches that detected productive pairs in 50–82% of cells[4,6,7].

Our method detected two α-chain recombinants in 42% of cells and two β-chain recombinants in 22% (**Supplementary Table 2**). We detected two productive α-chains in 35 of 188 (19%) cells with at least one productive α-chain, and two productive β-chains in 16 of 247 cells with at least one productive β-chain (6%). These data are in line with previous observations[17]. The best-performing PCR-based method did not detect multiple β-recombinants in any of the 1,268 cells studied[7] because it filtered data to remove any TCRβ chains that were represented by <85% of reads. In only one cell (0.3%) did we detect two apparently nonproductive sequences for a locus, and both of those sequences were validated by the PCR-based approach described below.

We compared the TCR sequences reconstructed by our method with those detected by a multiplex PCR-based approach[7] that we adapted for use with mouse cells (**Supplementary Note 1**, **Supplementary Fig. 3** and **Supplementary Table 3**). We also determined the effects of sequencing depth, read pairing and read length upon the performance of our method (**Supplementary Note 2** and **Supplementary Figs. 4** and **5**). Our approach provides sequences in good concordance with those generated by the PCR-based method and is able to successfully reconstruct TCR sequences from a variety of sequencing depths and read types.

We also applied TraCeR successfully to over 700 single-cell transcriptomes from a recently published study[18] (**Supplementary Note 3**). As a negative control, we applied our method to 192 scRNA-seq data sets generated from mouse embryonic stem cells[19]. No TCR sequences were reconstructed from these highly transcriptionally active and promiscuous cells.

Taken together, these data indicate that our method accurately and sensitively determines the sequences of recombined and expressed TCR loci within individual T cells from single-cell RNA-seq data.

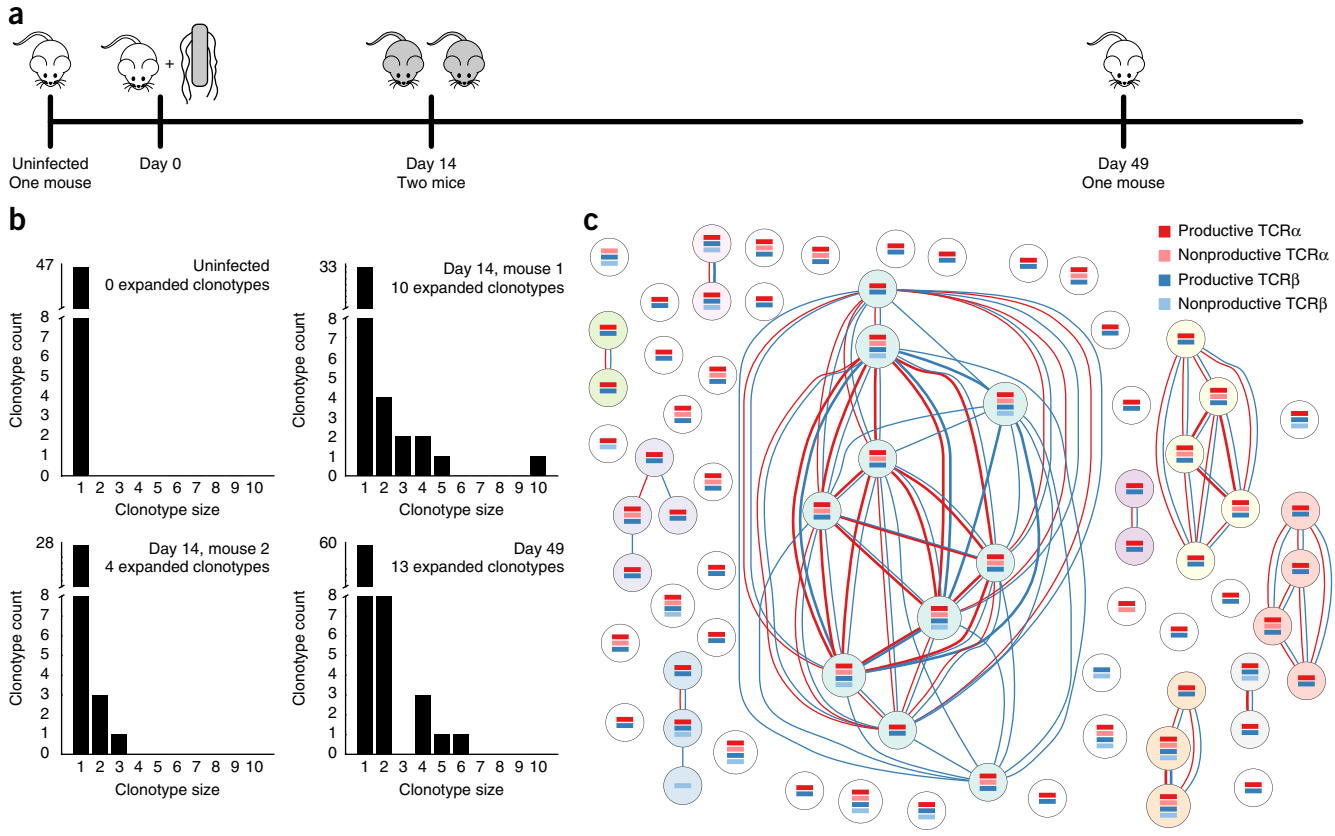

Figure 2 | Clonal CD4$^+$ T cell expansion during *Salmonella* infection. (**a**) Timeline for *Salmonella* infection experiment. (**b**) Distribution of expanded clonotypes within splenic CD4$^+$ T cell populations analyzed by single-cell RNA-seq. Clonotype size indicates the number of cells within the expanded clonotypes. (**c**) Clonotype network graph from day 14, mouse 1. Each node represents a splenic CD4$^+$ T lymphocyte; colored bars indicate reconstructed TCR sequences detected for each cell. Dark colors are productive, light colors are nonproductive. Red edges indicate shared TCRα sequences, blue edges indicate shared TCRβ sequences. Edge thickness is proportional to the number of shared sequences.

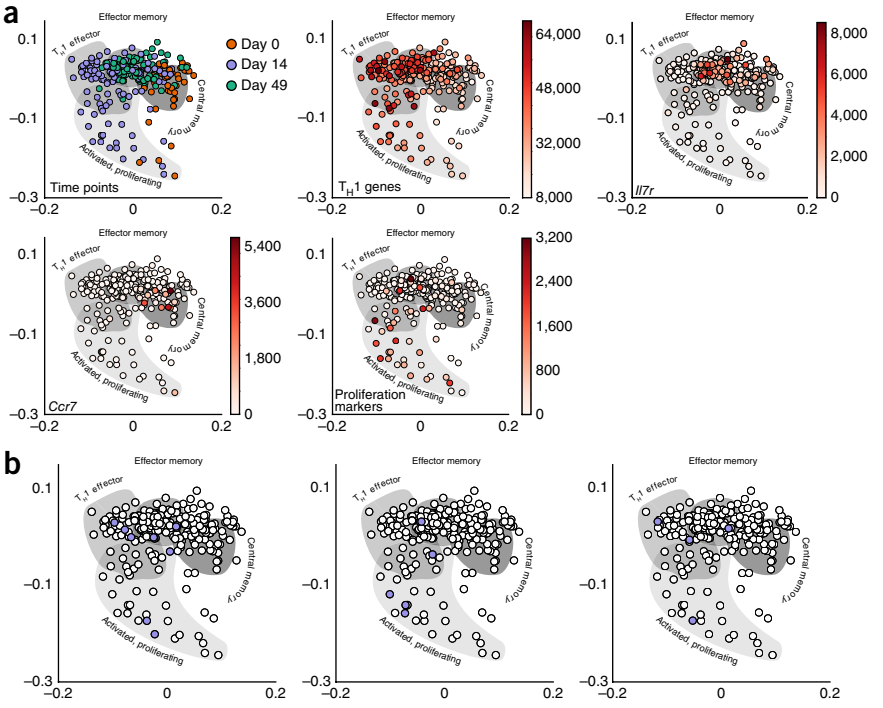

Figure 3 | Distribution of expanded clonotypes throughout the T_H1 response to *Salmonella* infection. ICA was used for dimensionality reduction of single-cell gene expression data. (**a**) Each point represents a $CD4^+$ T cell and is colored according to sampling time point or marker gene expression indicative of phenotype. T_H1 and proliferation marker gene sets are plotted as the sum of transcripts per million (TPM) values for genes within the set. (**b**) Clonotype distribution in gene-expression space. Three representative expanded clonotypes from day 14 mouse 1 are shown as purple points on top of all other cells within the gene-expression space.

We demonstrated an application of our approach by investigating the $CD4^+$ T lymphocyte clonotypes present within the spleens of mice before, during or after a nonlethal infection with *Salmonella typhimurium* (**Fig. 2a** and **Supplementary Table 1**), a bacterium that elicits a strong type 1 $CD4^+$ T cell response. We analyzed effector cells ($CD4^+CD8^-TCR\beta^+NK1.1^-CD44^{hi}CD62L^{lo}$) at day 14, when their relative abundance is close to its maximum, and memory cells ($CD4^+CD8^-TCR\beta^+NK1.1^-CD44^{hi}CD62L^{lo}CD127^{hi}$) at day 49, when the infection has been resolved[20] (**Supplementary Fig. 2**).

Analysis of the TCR sequences present in the splenic $CD4^+$ T cells enabled us to identify 12 invariant natural killer T (iNKT) cells[21] that were excluded from further analyses (**Supplementary Note 4**).

We compared recombinant identifiers between cells to find clonally related cells that expressed TCR genes with exactly the same nucleotide sequence. We found no TCR sharing either between cells from different mice or between cells from the uninfected mouse (**Fig. 2b**, **Supplementary Fig. 6** and **Supplementary Table 2**). This is to be expected given the huge potential diversity of TCR nucleotide sequences.

We saw evidence of clonotype expansion within activated $CD4^+$ T lymphocytes from each mouse at day 14 of *Salmonella* infection as well as from the mouse at week 7 after infection (**Fig. 2b**,**c**, **Supplementary Figs. 7–9** and **Supplementary Table 2**). TCR sequences within expanded clonotypes from these mice are likely to be specific for *Salmonella* antigens. Importantly, we observed multiple cells that shared all their detected recombinant sequences, including those that were nonproductive, indicating that our method detected the correct combinations of TCR recombinants within the cells. Furthermore, observations of cells that share multiple TCR sequences provide increased confidence that those cells are genuinely clonally related owing to the extremely small likelihood that identical recombination events would occur in two independent cells during development in the thymus.

Developing T lymphocytes in the thymus first recombine the TCRβ locus and undergo proliferation before recombining their TCRα loci. Cells generated from a single progenitor by this proliferative expansion will all have the same TCRβ recombinant but will each randomly generate a different α-recombinant before continuing maturation and entering the periphery. It is therefore possible that a particular single TCRβ chain could be found with multiple partners, and we detected cells that shared TCRβ sequences but had different TCRα sequences (**Supplementary Fig. 10**). This illustrates the value of paired TCRαβ sequences when inferring accurate clonal relationships between T cells. Note that we found no evidence of contamination across microfluidics chip capture or harvest sites or adjacent wells in the 96-well plates used (**Supplementary Fig. 11**).

Single-cell RNA-seq allows cells to be classified according to their gene expression profiles. We quantified gene expression within each cell and performed independent component analysis (ICA) to reduce the expression space to two dimensions (**Fig. 3a**). The 14,889 informative genes for ICA give much more information than the 17 phenotyping genes used in a previous PCR-based approach[7].

We analyzed the expression of 259 genes that indicate a T_H1 cell fate[22], as well as *Il7r* (CD127), which is indicative of effector memory T cells[23], *Ccr7* (a marker of central memory T cells)[24] and a set of seven genes that are expressed in proliferating cells[25] (**Fig. 3a**). Expression of these genes allowed us to separate the cells into four populations: activated proliferating cells that are differentiating to the T_H1 fate, mature differentiated T_H1 effector cells, effector memory–like cells and central memory–like cells. Cells from the uninfected mouse are mostly central memory–like, cells from the mouse at day 14 have an activated or T_H1 effector phenotype, and cells from day 49 (sorted as $CD127^{hi}$, a marker of effector memory fate) are found in the effector memory region of the ICA gene expression space.

We then determined the distributions of expanded clonotypes within the reduced gene expression space (**Fig. 3b** and **Supplementary Figs. 12–14**), excluding those that shared a TCRβ sequence but had different TCRα recombinants. Cells derived from the same progenitor could be seen throughout the activated differentiating, T_H1 effector and effector memory populations. This suggests that after activation by binding to a *Salmonella* antigen–MHC complex, the progeny of a particular $CD4^+$ T cell

differentiate asynchronously. Members of one clonotype exist across the full spectrum of proliferation and differentiation states that occur during the *Salmonella* response.

Our method is sensitive, accurate and easy to adapt to any species for which annotated TCR gene sequences are available. We also fully expect our method to be easily extended to study B cell receptor or antibody sequences in B lymphocytes, although considerations of clonality would need to take into account the process of somatic hypermutation.

Combining TCR reconstruction with scRNA-seq allows us to assess cellular phenotypes using orders of magnitude more genes than PCR-based approaches while obviating the need for *a priori* knowledge of phenotyping genes of interest. This will permit the discovery of novel or poorly characterized phenotypic subtypes in conjunction with the analysis of their TCR sequences.

Currently, the cost of scRNA-seq is prohibitive for surveys of the entire immune repertoire within an organism. However, it is practical for the analysis of smaller, selected lymphocyte subsets. In our illustrative example we were able to draw meaningful immunological insights from just 272 cells. A recent study sequenced 722 single T lymphocytes[19], and our method found expanded clonotypes likely to provide additional biological insights. Throughput of scRNA-seq methods is increasing while costs decrease, and we expect ever larger data sets to become standard.

A combined knowledge of T cell clonal dynamics, TCR specificity and detailed transcriptional phenotype is likely to be of great use in the study of T cell responses to infection, autoantigens or vaccination and will provide insights into both pathogenic mechanisms and therapeutic approaches.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing data from the cells described in this manuscript were deposited at ArrayExpress with accession number E-MTAB-3857.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank V. Svensson, T. Hagai, J. Henriksson and other members of the Teichmann laboratory along with G. Lythe for helpful discussions. We thank the Wellcome Trust Sanger Institute Sequencing Facility for performing Illumina sequencing and the Wellcome Trust Sanger Institute Research Support Facility for care of the mice used in these studies. This work was supported by European Research Council (grant ThSWITCH, number 260507, to S.A.T.) and the Lister Institute for Preventative Medicine (S.A.T.).

AUTHOR CONTRIBUTIONS

M.J.T.S. conceived the project, designed the computational method, wrote the software, designed PCR sequencing primers, analyzed data, generated figures and wrote the manuscript. T.L. and S.C. designed and performed the *Salmonella* experiments. T.L. performed cell collection and purification, generated scRNA-seq libraries, performed gene expression analyses, analyzed data, generated figures and wrote the manuscript. V.P. performed PCR-based TCR-sequencing experiments. A.O.S. designed the cell-sorting strategy, performed the sorting and generated figures. S.A.T. and G.D. supervised work and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Lieber, M.R. *FASEB J.* **5**, 2934–2944 (1991).
2. Becattini, S. *et al. Science* **347**, 400–406 (2015).
3. Mamedov, I.Z. *et al. EMBO Mol. Med.* **3**, 201–207 (2011).
4. Dash, P. *et al. J. Clin. Invest.* **121**, 288–295 (2011).
5. Linnemann, C. *et al. Nat. Med.* **19**, 1534–1541 (2013).
6. Kim, S.-M. *et al. PLoS One* **7**, e37338 (2012).
7. Han, A., Glanville, J., Hansmann, L. & Davis, M.M. *Nat. Biotechnol.* **32**, 684–692 (2014).
8. Buettner, F. *et al. Nat. Biotechnol.* **33**, 155–160 (2015).
9. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
10. Trapnell, C. *et al. Nat. Biotechnol.* **32**, 381–386 (2014).
11. Mahata, B. *et al. Cell Rep.* **7**, 1130–1142 (2014).
12. Bolotin, D.A. *et al. Nat. Methods* **12**, 380–381 (2015).
13. Shugay, M. *et al. Nat. Methods* **11**, 653–655 (2014).
14. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J. & Chain, B. *Bioinformatics* **29**, 542–550 (2013).
15. Kuchenbecker, L. *et al. Bioinformatics* **31**, 2963–2971 (2015).
16. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
17. Brady, B.L., Steinel, N.C. & Bassing, C.H. *J. Immunol.* **185**, 3801–3808 (2010).
18. Gaublomme, J.T. *et al. Cell* **163**, 1400–1412 (2015).
19. Kolodziejczyk, A.A. *et al. Cell Stem Cell* **17**, 471–485 (2015).
20. Mittrücker, H.-W., Köhler, A. & Kaufmann, S.H.E. *Infect. Immun.* **70**, 199–203 (2002).
21. Brennan, P.J., Brigl, M. & Brenner, M.B. *Nat. Rev. Immunol.* **13**, 101–117 (2013).
22. Stubbington, M.J.T. *et al. Biol. Direct* **10**, 14 (2015).
23. Kallies, A. *Immunol. Cell Biol.* **86**, 325–332 (2008).
24. Sallusto, F., Lenig, D., Förster, R., Lipp, M. & Lanzavecchia, A. *Nature* **401**, 708–712 (1999).
25. Whitfield, M.L., George, L.K., Grant, G.D. & Perou, C.M. *Nat. Rev. Cancer* **6**, 99–106 (2006).

ONLINE METHODS

Ethics statement. Mice were maintained under specific pathogen–free conditions at the Wellcome Genome Campus Research Support Facility (Cambridge, UK). These animal facilities are approved by and registered with the UK Home Office. Animals were sacrificed by approved animal technicians in accordance with Schedule 1 of the Animals (Scientific Procedures) Act 1986. Oversight of the arrangements for Schedule 1 killing was performed by the Animal Welfare and Ethical Review Body of the Wellcome Genome Campus.

Cell preparation. Female C57BL6/N mice aged 6–8 weeks were infected intravenously with 0.2 ml *Salmonella typhimurium* M525 containing 5×10^5 CFU of bacteria in sterile phosphate buffered saline (PBS, Sigma-Aldrich). At day 14 or 49 post infection (p.i.) mice were sacrificed with spleens and livers being harvested. An uninfected mouse (day 0) was also sacrificed. The sample size of four mice was chosen to provide sufficient example data for application of our method. Mice were randomly chosen to receive infection and randomly assigned to sacrifice at day 14 or day 49. No blinding was performed. Bacteria were enumerated from the livers by serial dilution and plating onto agar plates (Oxoid) to confirm levels of infection. Single-cell suspensions were prepared by homogenizing spleens through 70 µm strainers and lysing erythrocytes. Following incubation with CD16/CD32 blocking antibody, the cells from the uninfected mouse and from day 14 p.i. were stained with titrated amounts of fluorochrome conjugated antibodies for CD44(FITC), CD25(PE), CD62L(PE-CF594), TCRβ(PerCP-Cy5.5), CD8α(APC-H7), NK1.1(BV421), and CD4(BV510). The cells from day 49 p.i. were stained with antibodies for CD44(FITC), CD127(PE), CD62L(PE-CF594), TCRβ(PerCP-Cy5.5), NK1.1(APC), CD8α(APC-H7), CD4 (BV510), and Sytox Blue viability stain. Antibody details can be found in **Supplementary Table 4**. Cell sorting was performed using a BD FACSAria II instrument using the 100-µm nozzle at 20 psi using the single cell sort precision mode. The cytometer was set up using Cytometer Setup and Tracking beads and compensation was calculated using compensation beads (for antibodies, eBioscience UltraComp) and cells (for Sytox Blue) using automated software (FACSDiva v6).

Single-cell RNA-sequencing and gene expression quantification. Capture and processing of single $CD4^+$ T cells was performed using the Fluidigm C1 autoprep system. Cells were loaded at a concentration of 1,700 cells μl^{-1} onto C1 capture chips for 5–10 µm cells. We used microscopic inspection of the C1 capture sites to determine which contained only a single cell so as to exclude empty wells or those containing multiple cells. From the five chips used in this work we captured 329 single cells out of a possible 480 (68.5%). ERCC (External RNA Controls Consortium) spike-in RNAs (Ambion, Life Technologies) were added to the lysis mix. Reverse transcription and cDNA preamplification were performed using the SMARTer Ultra Low RNA kit (Clontech). Sequencing libraries were prepared using Nextera XT DNA Sample Preparation kit with 96 indices (Illumina), according to the protocol supplied by Fluidigm. Libraries from 303 single cells were pooled and sequenced on Illumina HiSeq2500 using paired-end 100-base reads.

Reads were mapped to the *Mus musculus* genome (Ensembl version 38.70) concatenated with the ERCC sequences, using GSNAP[26] with default parameters. Gene-specific read counts were calculated using HTSeq[27]. Thirty-one cells (out of 303) with detected transcripts for fewer than 2,000 genes, or with more than 10% of measured exonic reads corresponding to genes coded by the mitochondrial genome, were excluded from further analyses. The 272 cells that passed these quality control steps were used in the analyses presented here.

Reconstruction and analysis of TCR sequences from RNA-seq data. Combinatorial recombinome files were separately created for the TCRα and TCRβ chains. To generate these fasta files, nucleotide sequences for all mouse V and J genes were downloaded from The International ImMunoGeneTics information system[28] (IMGT, http://www.imgt.org). Every possible combination of V and J genes was generated for each TCR locus such that each combination was a separate sequence entry in the appropriate recombinome file. Within the recombinome files we did not attempt to encompass all possible sequences that could be generated by junctional diversity during V(D)J recombination. Instead, ambiguous 'N' nucleotide sequence characters (not to be confused with 'N nucleotides' added by terminal deoxynucleotidyl transferase during recombination within the cell) were introduced into the junction between V and J genes in each sequence entry to improve alignments of reads that spanned diverse junctional sequences (**Supplementary Fig. 1b**). Seven N nucleotides were used in TCRβ combinations whilst one N nucleotide was used in the TCRα combinations. V gene leader sequences are not well annotated within IMGT and so 20 N nucleotides were added at the 5′ end of the V sequence to permit alignment of sequencing reads that included the leader sequence.

TCRα or TCRβ constant region cDNA sequences were downloaded from ENSEMBL and appended to the 3′ end of each combined sequence to permit alignment of reads that ran into the constant region. The full-length TCRα constant region was used whilst the only the first 259 nucleotides of the TCRβ constant gene were used since these are identical between both *Trbc* homologs that are found within the mouse genome. The combinatorial recombinomes used in this work can be found alongside the other tools at https://www.github.com/teichlab/tracer.

RNA-seq reads from each cell were aligned against each combinatorial recombinome independently using the Bowtie 2 aligner[29]. Bowtie 2 is ideal for alignment against the recombinomes because it can align against ambiguous N nucleotides within a reference and also introduce gaps into both the reference and read sequences. This allows it to align reads against the variable junctional regions. We used the following Bowtie 2 parameters with low penalties for introducing gaps into either the read or the reference sequence or for aligning against N nucleotides: -no-unal -k 1-np 0-rdg 1,1-rfg 1,1.

For each chain, separately, we used the reads that aligned to the appropriate recombinome as input to the Trinity RNA-seq assembly software[30] using its default parameters.

V, D and J gene sequences downloaded from IMGT were used to generate appropriate databases for use with IgBLAST[31]. Contigs assembled by Trinity were used as input to IgBlast and the resulting output text files were processed with a custom parsing script. Contigs were classed as representing TCR sequences if they contained gene segments from the correct locus (i.e., TCRα genes for TCRα contigs) and if their reported V and J alignments had

E-values below 5×10^{-3}. If multiple contigs within the same cell represented the same recombinant sequence, these were collapsed so that the sequence was only represented once in the cell for subsequent analyses. In some cases where two contigs derived from the same original sequence but one was shorter than the other, IgBLAST assigned different V sequences if the shorter sequence did not provide sufficient information to distinguish between highly similar genes. This typically occurred with V genes that were part of the evolutionary expansion events that caused gene duplication and triplication within the TCRα locus[32]. In these cases, the sequences were collapsed into a single assignment that used the results from the longest contig. The IgBLAST results for the TCR sequences within each cell were then reduced to an identifying string (**Supplementary Fig. 1a**) consisting of the V gene name, the junctional nucleotide sequence and the J gene name (for example, TRBV31_AGTCTTGACACAAGA_TRBJ2-5), which was used for comparisons between sequences within other cells.

It is important to determine whether a particular TCR mRNA sequence is productive and therefore able to be translated to produce a full-length TCR polypeptide chain. To do this for the reconstructed TCR sequences, we first converted them to entirely full-length sequences by using full-length V and J gene sequences from IMGT appropriate to the gene segments assigned by IgBLAST. Because sequences from laboratory mouse strains are well-characterized and TCRs do not undergo somatic hypermutation, we can make the assumption that variations between the RNA-seq derived sequences and the reference sequences outside the junctional region are due to PCR and/or sequencing errors and so can be ignored. We check that these full-length sequences are in the correct reading frame from the start of the V gene to the start of the constant gene and that they lack stop codons. If this is the case, the sequence is classed as productive. For analysis of CDR3 amino-acid sequences we translate the productive recombinants and define the CDR3 as the region flanked by the final cysteine residue of the V gene and the conserved FGXG motif in the J gene as previously described[14]. Although this approach is likely to be most accurate when analyzing data from inbred, well-characterized mouse strains, it does risk losing useful sequence information that represents genuine germline polymorphism that may be present when analyzing cells from humans or other outbred populations. To address this, it is possible to instruct TraCeR to omit the step that replaces the assembled sequences with reference sequences from IMGT. In this case, productivity is calculated solely from the sequence as assembled by Trinity. For the data presented here, both methods give almost identical assessments of productivity (**Supplementary Fig. 15**).

Expression levels of the TCR genes found within a cell were quantified by appending that cell's full-length recombinant sequences to a file containing the entire mouse transcriptome (downloaded from http://bio.math.berkeley.edu/kallisto/transcriptomes/) and then using this file for the generation of an index suitable for use with the pseudoalignment-based Kallisto algorithm[33]. This index was then used with the RNA-seq reads for the cell as input for Kallisto in quantification mode to calculate transcripts per million (TPM) values for each TCR sequence. If a cell was assigned more than two recombinant sequences for a particular locus (5/272, 1.8% of cells in this study), the sequences were ranked by their TPM values and the two most highly expressed were used for further analyses. Kallisto's speed in constructing indices and performing expression quantification makes it ideal for this task.

After assignment of TCR sequences to each cell within an experiment, we used custom Python scripts to compare the recombinant identifiers present in each cell to find cases where multiple cells contained the same identifier. These analyses were used to generate network graphs where each node in the graph represents a single cell and edges between the nodes represent shared TCR sequences.

Code availability. The analyses described above are performed by our tool, TraCeR which is freely available at https://www.github.com/teichlab/tracer and as **Supplementary Software**.

PCR-based sequencing of TCR sequences. Primers were designed to amplify all possible recombined TCR sequences from both the TCRα and TCRβ loci (all sequences can be found in **Supplementary Table 5**). Two constant region primers were designed to be complementary to the *Trac* or *Trbc* genes close to their 5′ ends. Sets of primers complementary to all TCRα and -β V gene sequences downloaded from IMGT were also designed. Primers were designed to regions of homology between V genes and included degeneracy where appropriate so as to minimize the number of primers required. In total, 34 TCRα and 31 TCRβ primers were used. All primers were designed with a T_m of 71–73 °C. All V gene primers were designed with the sequence ACACTCTTTCCCTACACGACGCTCTTCCGATCT at their 5′ end to allow amplification by the Illumina PE 1.0 primer (AA TGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT), whereas the constant region primers were designed with the sequence TCGGCATTCCTGCTGA ACCGCTCTTCCGATCT at their 5′ end so that they could be amplified by barcoding primers containing a unique 11-nt index sequence (**Supplementary Table 5**). The barcoding primers also contain the Illumina PE 2.0 sequence.

Full-length (oligo-dT primed) cDNA produced from single cells by the C1 system (Fluidigm, USA) was used as template in two PCR reactions, one for each TCR locus. 0.4 µl of cDNA were used in each reaction along with each V primer at 0.06 µM and the constant primer at 0.3 µM. Phusion DNA polymerase (NEB, USA) was used to perform the amplification in 25 µl final volume. The cycling conditions for this step were 98 °C 30 s; 98 °C 10 s, 60 °C 10s, 72 °C 30 s × 16 cycles; 72 °C 5 min. 4 °C. A 1 µl aliquot of the first reaction was used as template in a second PCR amplification, again using Phusion in a 25 µl reaction volume. Here, the Illumina PE 1.0 primer was used with a barcoding primer unique for each cell and each primer was at 0.4 µM. The cycling conditions for this step were 98 °C 30 s; 98 °C 10 s, 58 °C 10s, 72 °C 30 s × 16 cycles; 72 °C 5 min. 4 °C. PCR products of the correct size for sequencing were purified using 0.7 volumes of AMPure XP beads (Beckman Coulter) according to the manufacturer's instructions. Purified products were pooled and submitted to the Wellcome Trust Sanger Institute (WTSI) Sequencing Facility for sequencing using a MiSeq (Illumina) with 250-bp paired-end reads.

Processing PCR data. Reads generated by MiSeq sequencing of PCR products were de-multiplexed by the WTSI Sequencing Facility according to their barcode sequences. Reads were then trimmed to remove low-quality regions and adaptor sequences

doi:10.1038/nmeth.3800

using TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The TCR-targeted PCR primers were designed to provide amplicons short enough such that the forward and reverse paired reads would overlap upon sequencing enabling read pairs to be merged using FLASH[34]. Merged read sequences were then filtered to remove those under 200 nt in length to remove artifactual sequences. Following this step, read sequences for each cell were subsampled where necessary such that there were 50,000 sequences or fewer from each cell. This reduced the computational time and requirements for the next stage whilst still providing sufficient information about the sequences present. As described previously[7], we assumed that sequences from an individual cell that had at least 95% sequence identity were derived from the same original cDNA sequence and so these were combined to generate a consensus sequence. The consensus sequences for each cell were analyzed by IgBLAST to find sequences that represented recombined TCRs and the number of sequencing reads supporting each TCR were used to filter out background sequences that had few reads.

Comparing PCR and RNA-seq data. For each cell, sequences derived from PCR analysis or reconstructed from RNA-seq data were trimmed to only include the regions assigned by IgBlast as containing V, D or J sequences. This removed any leader sequences or constant regions. Trimmed reconstructed RNA-seq sequences were aligned against the trimmed PCR-derived sequences in a set of pairwise comparisons using BLAST. If an alignment was reported, the number of mismatches across the entire alignment were counted, as were the number of mismatches between the nucleotides that encoded the CDR3 region (defined here as the 30 nt following the end of the framework 3 region as annotated by IgBlast). If the CDR3 regions contained any mismatches, the alignment was classed as discordant; otherwise the two sequences were classed as concordant. Sequences from one method (RNA-seq or PCR) that did not align successfully with any sequence from the other method were classed as discordant.

Gene expression quantification and dimensionality reduction. Genes were filtered to remove those expressed (TPM > 1) in fewer than three cells. Dimensionality reduction of the remaining gene expression data was performed by independent component analysis (ICA) using the FastICA Python package.

For plotting gene expression for each cell within ICA space, 259 genes indicating a T_H1-like fate and seven indicators of proliferation (*Mki67*, *Mybl2*, *Bub1*, *Plk1*, *Ccne1*, *Ccnd1* and *Ccnb1*) were taken from previous work[22,25], and their expression levels (in TPM) were summed for each cell.

Clonotype distribution within gene expression space. Cells that did not seem to be derived from the same progenitor (same TCRβ but differing TCRα chains) were removed from the expanded clonotype groups. Cells belonging to a particular expanded clonotype were then plotted within the ICA reduced gene expression space.

26. Wu, T.D. & Nacu, S. *Bioinformatics* **26**, 873–881 (2010).
27. Anders, S., Pyl, P.T. & Huber, W. *Bioinformatics* **31**, 166–169 (2015).
28. Lefranc, M.-P. *et al. Nucleic Acids Res.* **37**, D1006–D1012 (2009).
29. Langmead, B. & Salzberg, S.L. *Nat. Methods* **9**, 357–359 (2012).
30. Grabherr, M.G. *et al. Nat. Biotechnol.* **29**, 644–652 (2011).
31. Ye, J., Ma, N., Madden, T.L. & Ostell, J.M. *Nucleic Acids Res.* **41**, W34–W40 (2013).
32. Bosc, N. & Lefranc, M.-P. *Dev. Comp. Immunol.* **27**, 465–497 (2003).
33. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Preprint at arXiv:1505.02710 (2015).
34. Magoč, T. & Salzberg, S.L. *Bioinformatics* **27**, 2957–2963 (2011).