Immunology and Cell Biology (2016), 1–8

www.nature.com/icb

ORIGINAL ARTICLE

Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells

Auda A Eltahla[1,4], Simone Rizzetto[1,4], Mehdi R Pirozyan[1,4], Brigid D Betz-Stablein[1], Vanessa Venturi[2], Katherine Kedzierska[3], Andrew R Lloyd[1], Rowena A Bull[1] and Fabio Luciani[1]

Heterogeneity of T cells is a hallmark of a successful adaptive immune response, harnessing the vast diversity of antigen-specific T cells into a coordinated evolution of effector and memory outcomes. The T cell receptor (TCR) repertoire is highly diverse to account for the highly heterogeneous antigenic world. During the response to a virus multiple individual clones of antigen specific CD8+ (Ag-specific) T cells can be identified against a single epitope and multiple epitopes are recognised. Advances in single-cell technologies have provided the potential to study Ag-specific T cell heterogeneity at both surface phenotype and transcriptome levels, thereby allowing investigation of the diversity within the same apparent sub-population. We propose a new method (VDJPuzzle) to reconstruct the native TCRαβ from single cell RNA-seq data of Ag-specific T cells and then to link these with the gene expression profile of individual cells. We applied this method using rare Ag-specific T cells isolated from peripheral blood of a subject who cleared hepatitis C virus infection. We successfully reconstructed productive TCRαβ in 56 of a total of 63 cells (89%), with double α and double β in 18, and 7% respectively, and double TCRαβ in 2 cells. The method was validated via standard single cell PCR sequencing of the TCR. We demonstrate that single-cell transcriptome analysis can successfully distinguish Ag-specific T cell populations sorted directly from resting memory cells in peripheral blood and sorted after *ex vivo* stimulation. This approach allows a detailed analysis of the TCR diversity and its relationship with the transcriptional profile of different clones.
Immunology and Cell Biology advance online publication, 8 March 2016; doi:10.1038/icb.2016.16

Antigen-specific T cells can provide long-lasting immunity against infections. Long-term memory cells are the result of the clonal expansion of a very small population of naïve cells specific for an epitope, which expand both vigorously and rapidly upon antigen stimulation. During primary infection, T cell clones undergo several phenotypic changes from naïve to memory phenotype, with a complex pattern of diversification of T cell sub-populations, each characterised by distinct surface molecule profiles, and dynamic changes in the transcriptome.[1] These elaborate dynamics make T cells the most heterogeneous cell type in humans.

This heterogeneity is dictated firstly by the vast diversity of the T cell receptor (TCR), which underpins antigen specificity. Each cell carries a specific TCR, with greater than 10^{14} different combinations potentially available, and with a broad spectrum of peptide antigens recognised by each receptor. Upon recognition, the specificity of the TCR for an antigen ultimately drives clonal expansion, and the establishment of a functional memory T cell repertoire.[2] With this vast diversity, it has remained difficult to analyse the relationships between TCR diversity and the phenotypic outcomes of Ag-specific T responses.[3] This challenge is further complicated in the case of an immune response against a rapidly mutating virus, such as hepatitis C virus (HCV), HIV or influenza (Flu), where multiple Ag-specific responses may be activated, each with a distinct HLA-restricted epitope, each inducing a distinct TCR repertoire, and each facing an evolving viral population undergoing mutation events to escape those responses.[4] Therefore, understanding the diversity of each T cell clone which contributes to the Ag-specific response inform the determinants of successful immune protection.[5]

Several factors have limited our knowledge of these complex dynamics, notably including the lack of technologies to study single cells. Recent advances in single cell technologies offer the opportunity to identify and characterize rare sub-populations and their heterogeneity,[3,6] as well as their fate.[7,8] Combinations of flow cytometry, cell sorting, and single cell transcriptomic technologies (scRNA-seq) are rapidly arising as useful methods to study single cell diversity, including of CD8+ T cells.[3] With these approaches human samples have recently been analysed to identify subpopulations of haemopoietic cells and to describe heterogeneous responses to *in vitro* stimulation.[9]

[1]Systems Medicine in Infectious Diseases, Inflammation and Infection Research Centre, School of Medical Sciences, University of New South Wales, Sydney, Australia; [2]Kirby Institute for Infection and Immunity, University of New South Wales, Sydney, Australia and [3]Department of Microbiology and Immunology, University of Melbourne, Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
[4]These authors contributed equally to this work.
Correspondence: Dr F Luciani, Systems Medicine in Infectious Diseases, Inflammation and Infection Research Centre, School of Medical Sciences, University of New South Wales, Sydney 2052, Australia.
E-mail: luciani@unsw.edu.au
Received 31 December 2015; revised 4 February 2016; accepted 5 February 2016; accepted article preview online 10 February 2016

While these advances have allowed the characterisation of transcriptomic heterogeneity at the single cell level, the study of TCR diversity in scRNA-seq data and thereby the link to the transcriptome has been challenging. Single cell analyses of the TCR have generally been performed using gene-specific PCR methods.[10] Such methods allow the identification of the native αβ chain combination for T cells, however this approach is constrained by the number of cells, that can be analysed and does not allow the simultaneous analysis of other transcripts. These methods have been adapted for high-throughput TCR sequencing for scale up of the analysis to large numbers of cells.[11] Directly linking TCR sequence with the transcriptome at the single cell level can substantially increase our understanding of Ag-specific T cell diversity and how heterogeneity contributes to the ultimate outcome of an immune response. Recently, this approach came to the fore with the description of high-throughput methods to link a selected number of genes to the native TCRαβ at the single cell level.[12] Here we describe a new computational workflow to analyse the full transcriptome of flow-sorted Ag-specific CD8+ T cells at the single cell level, and to accurately identify the native TCRαβ. This method allows simultaneous analysis of gene expression and diversity of circulating Ag-specific T cell clones. We applied the method to human HCV-specific CD8+ T cells and demonstrated that scRNA-seq can distinguish T cell sub-populations along with TCRαβ chains.

RESULTS

We have developed a novel bioinformatics pipeline for the linkage of scRNA-seq transcriptomics with TCRαβ at the single cell level. The workflow allows transcript quantification, noise reduction, gene variability analysis, differential expression, co-expression analysis, and finally reconstruction of the full TCRαβ transcripts (Figure 1).

Experimental design

In order to test the hypothesis that the transcriptomic profile of single cells can be utilised to separate sub-populations of Ag-specific CD8+ T cells and to identify accurately the native TCRαβ, three experimental conditions utilising human Ag-specific CD8+ T cells directed against a single epitope were studied. First, we isolated HCV-specific CD8+ T cells from a subject who had spontaneously cleared primary HCV infection (Figure 1). CD8+ T cells specific for the HLA-I A*0201 restricted epitope (CINGVCWTV) were identified via dextramer staining and sorted directly from peripheral blood mononuclear cells (PBMCs). Second, from the same patient and from the same sample time point an epitope-specific (CINGVCWTV) CD8+ T cell line was generated using *ex vivo* peptide stimulation and sorted via dextramer staining. Finally, *ex vivo* generated Ag-specific T cells from the line were briefly re-stimulated with the peptide immediately prior to dextramer staining and sorting. The cells analysed from the first condition (CINGVCWTV specific CD8+ T cells isolated from PBMC) are referred to here as PB-T; and CL-Unstim and CL-Stim to the CINGVCWTV CD8+ T cells isolated from the cell line before and after epitope re-stimulation respectively. Following dextramer staining, single cells were sorted into 96-well plates and full-length cDNA was generated using the Smart-seq2 protocol[13] before sequencing the transcriptome using the Illumina sequencing platform (San Diego, California, USA).

Transcriptomic analysis of single Ag-specific T cells

A total of 63 single cells sampled across the three conditions ($n=21$ cells for each) were analysed, along with three bulk-sorted cell populations (approximately 10 cells per sample) as controls. We quantified gene expression in all 63 single cells. We performed an unsupervised clustering analysis to detect differences among the three populations and to identify variations within the same group of cells. Principal component analysis (PCA) based on Fragments Per Kilobase of transcript per Million (FPKM) values of gene expression showed that PB-T cells clustered separately from the cell line T cells, while there was no clear difference between CL-Unstim and CL-Stim (Supplementary Figure 1). In the PCA analysis four cells (3 CL-Unstim and 1 CL-Stim) were characterised by different profiles compared to other cells and had a low number of genes expressed ($<=1200$); these cells were excluded from downstream analyses (Supplementary Figure 1). For the remaining cells, we detected between 1342 and 2255 genes expressed in PB-T cells, 2433 and 4934 genes for CL-Unstim, and between 2534 and 5220 genes for CL-Stim (Figure 2a).

In order to minimise the confounding effects of varied cell-cycle stage in the transcriptional analyses, the contribution of cell cycle to gene expression was filtered out using scLVM.[14] Following this step,

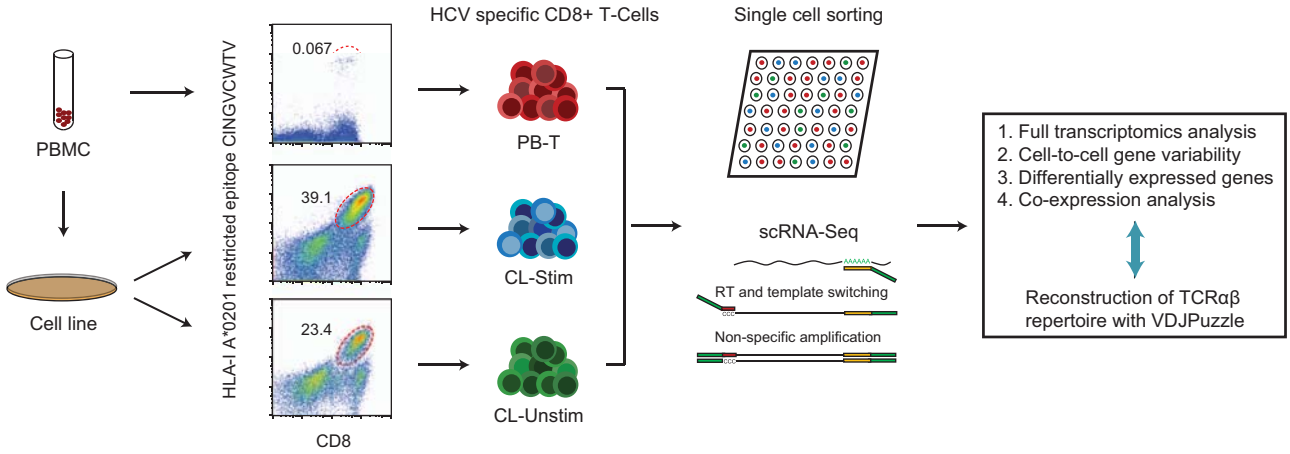

Figure 1 Experiment design. Epitope-specific T cells were identified and sorted via flow-cytometry from PBMCs, and from a T cell line obtained from the same subject who spontaneously cleared HCV infection. Single cell RNA-seq was performed following Smart-Seq2 protocol. A new pipeline was developed for simultaneous analysis of both TCRαβ (VDJPuzzle) and full transcriptome.

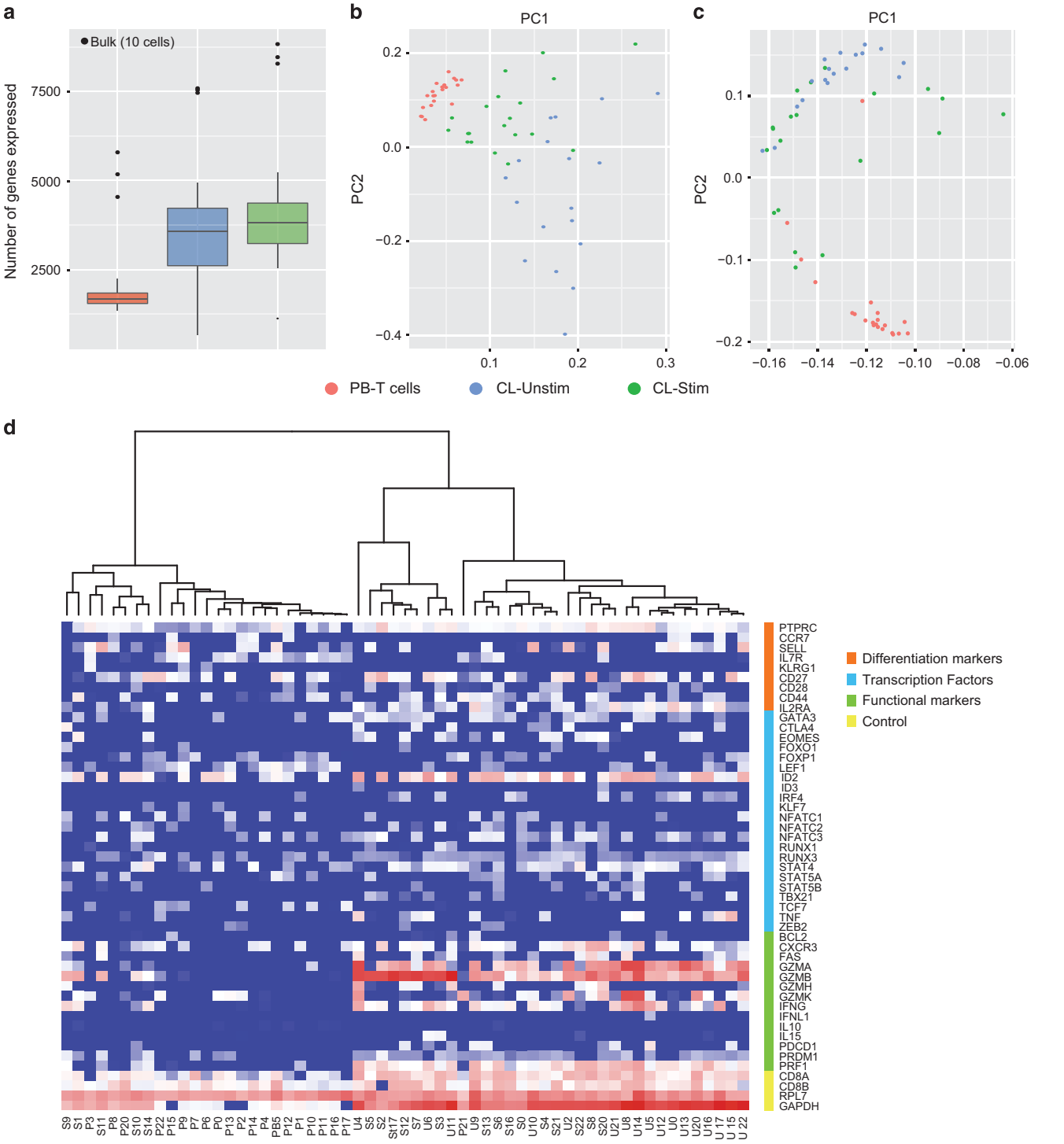

Figure 2 Transcriptomic analysis. (**a**) Number of genes expressed. (**b**) Unsupervised PCA using all the identified genes separates the three T cell populations. (**c**) PCA analysis on a restricted list of CD8 T cell related genes. (**d**) Heatmap of a restricted list of CD8 T cell related genes subdivided by functional categories.

the three populations were clearly separated in the PCA analysis (Figure 2b). This analysis also showed that both CL-Unstim and CL-Stim populations are more scattered than PB-T cells, suggesting a greater degree of heterogeneity in the transcriptome of the T cells derived from the cell line. However, the correlation (Pearson's correlation, R) was lower between cells in CL-Stim sample. The same analysis approach was then applied to a selected list of genes associated with T cell functions. PCA identified PB-T cells as a separate cluster, while CL-Unstim and CL-Stim were clustered together (Figure 2c). Similar clustering analysis based on Pearson's correlations on selected lists of genes categorised according to differentiation, transcription factors, and functional genes showed that PB-T cells were clustered together with a profile broadly separated from CL-Unstim. (Figure 2d).

In support of the good quality of the transcriptional profiling, control genes, such as RPL7, or CD45RO were consistently expressed among all the single T cells, while other immune response related genes were more variably expressed, such as Granzyme genes (A, B, H, K), IFNγ, CXCR4, PDCD1 (PD1), IL2R, SELL (CD62L), CCR7, Eomes, Tbx21 (T-bet), and IL7Rα (Supplementary Figure 2). We further validated the scRNA-seq expression data by qPCR on a selected set of genes, including RPL7, CCR7, GAPDH, and ACTB using a representative set of 10 cells, where these genes were variably expressed in the transcriptome (except RPL7 which was uniform in expression across the cells). Correlation analysis between the log transformed levels of the normalised FPKM values against the mean FPKM of each gene and the corresponding normalised cp values of the qPCR showed consistent trends ($R = 0.88$, P-value < 0.001).

Noise reduction analysis

Gene expression quantification from scRNA-Seq can be affected by different sources of noise.[15] To assess the reliability of our expression data, we compared the values from RNA-seq data of bulk sorted cells with those from single cells. For each gene, we generated three *in-silico* bulk RNA-seq datasets by averaging profiles of 10 single cells. A total of three *in-silico* bulk profiles were generated for each of the three experimental conditions. We then tested the hypothesis that the *in-silico* generated bulk profiles were not statistically different from the observed bulk distributions utilising a standard t-test. For each gene, the test was repeated 1000 times and the mean of the t-values was calculated. This generated a distribution of t-values, which were utilized to identify a threshold (FPKM = 95) above which genes could not be differentiated between the two bulk conditions (Supplementary Figure 3). This threshold was very similar between the T cell populations in the three experimental conditions. The analysis revealed that a subset of genes with an average expression level lower than 95 FPKM were likely to be affected by noise, and hence likely to impair the accuracy of analyses. Between 85 and 95% of the genes were identified as possible noise associated genes (Supplementary Table 1). This consideration was taken into account in downstream analyses of differentially expressed genes.

Analysis of variable genes

One of the advantages of using scRNA-seq is the possibility to assess the variation in gene expression within the same cell population. Brennecke *et al.* proposed a method to identify the set of variable genes utilising a log linear fit model of the squared coefficient of variation over the mean of expression of each control gene within the same population.[16] With this approach, only genes with the squared coefficients of variation (CV) greater than the fitted line were considered variable. Using this method, we found 1083 variable genes in the PB-T cells, 3212 in CL-Unstim, and finally 3363 genes in CL-Stim population. As described above, scLVM correction was applied to remove gene expression variance attributable to cell cycle stage. For each population, we selected the set of most variable genes as the 95th percentile of the distribution of the deviations from the fitted linear model on the squared coefficient of variation. This analysis identified 23, 1222 and 2355 variable genes within PB-T, CL-Unstim, and CL-Stim, respectively (Supplementary Table 2). We identified several variable genes that were expressed in at least two of the three populations (see Venn diagram in the Supplementary Figure 4), of which five were found in all the three experimental populations, including ANXA1, TMED3, UQCRC2, and two genes related to T cell functionality - GZMK, and CD62L. To further decrease the number of false positives, these genes were intersected with those estimated to be non-noisy (as described above), identifying 14, 360, 520 variable genes in the three populations respectively. Interestingly, GZMK and CD62L were still detected as variable genes in all three populations, whereas a control gene (RPL7) demonstrated similar expression levels across cells from all populations (Supplementary Figure 2).

Analysis of differentially expressed genes

Differentially expressed genes between PB-T cells and both CL populations, as well as single cell and bulk-sorted samples were identified using the framework provided by Monocle.[17] We identified 1561 genes, which were differentially expressed (q-value < 0.001) between PB-T cells and CL-Stim, while only 366 genes were differentially expressed between the three bulk experiments. There were 70 genes identified as differentially expressed in both single cell and bulk samples. Thus, 95% of differentially expressed genes were detected only in a single cell setting. Differentially expressed genes were mostly enriched in Gene Ontology (GO) functions designated as respiratory electron transport chain ($n = 47$, q-value < 0.001) and antigen processing and presentation ($n = 54$, q-value < 0.001). Moreover 1459 genes were found to be differentially expressed between CL-Unstim and PB-T cells, but only 307 were found differentially expressed in bulk samples, with 57 genes differentially expressed in both experiments (Supplementary Table 3). These genes were enriched in antigen processing and presentation ($n = 48$, q-value < 0.001) as well as metabolic processes ($n = 482$, q-values < 0.001). When corrected for cell cycle, 320 genes were differentially expressed in PB-T cells compared to CL-Unstim, and 412 compared to CL-Stim, with 202 genes differentially expressed in both comparisons, of which 50 had an expression level greater than 95 FPKM. This final list of genes was enriched in various metabolic processes such as generation of precursor metabolites, and energy and NAD metabolic process (q-value < 0.001).

Co-expression modules analysis

In order to identify patterns of correlated gene expressions within the Ag-specific T cell populations we performed a Weighted Correlated Network Analysis (WGCNA).[18] In total this analysis identified 18 groups of genes with correlated patterns. We selected the five largest modules for GO enrichment analysis, identifying genes involved in viral processes and translational termination (q-value < 0.001). One of the most highly variable genes (GZMK) was clustered with other immunologically relevant genes such as CCL4 ($R = 0.94$), CD44 ($R = 0.70$), NFAT5 ($R = 0.92$) and IL2RG ($R = 0.74$). By contrast, CD62L was clustered with IL7R ($R = 0.65$).

Reconstruction of TCRαβ from scRNA-seq

In order to assess the diversity and clonality of Ag-specific T cells, we developed VDJPuzzle, a pipeline that reconstructs the full TCRαβ from scRNA-seq (Figure 3a). With VDJPuzzle we identified 56 (89%) productive TCRαβ, while at the single chain level, we detected 89% of the αTCRs and 92% of the βTCRs. The final TCR repertoire was very diverse (Table 1), with only 4 cells identified with shared TCRαβ (Supplementary Table 4 for full details of the repertoire). This major TCRαβ clone was found across the three populations (two cells in the PB-T, one in the CL-Stim and one in the CL-Unstim) and carried a shared, but distinct transcriptomic profile (Pearson coefficient 0.677) (Figure 3b). In order to validate the TCR sequences identified by VDJPuzzle, single cell PCR was performed on the same cDNA generated for scRNA-seq for a representative set of 25 cells, of which 11 were sampled from the PB-T cell subset, and five in each of the two

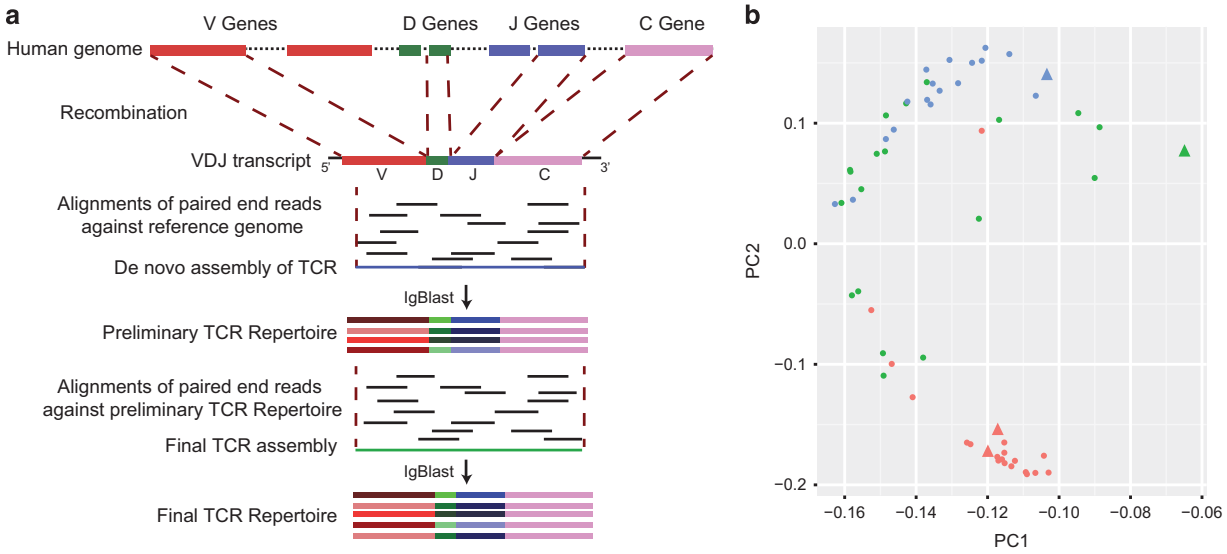

Figure 3 VDJPuzzle workflow and results. (**a**) Schematic of the VDJPuzzle workflow. (**b**) PCA (as Figure 2c) showing the four cells bearing the most common TCRαβ (triangle).

Table 1 TCRαβ repertoire identified from scRNA-seq using VDJPuzzle

CDR3α	*CDR3β*	*Total occurrences*	*Occurrences in PB-T*	*Occurrences in CL-Unstim*	*Occurrences in CL-Stim*
CAVEDTGGFKTIF	CASSSSMESGNTIYF	4	2	1	1
CAMREHTSGTYKYIF	CASSDSLVRGYQETQYF	3	0	1	2
CAFMITGAGSYQLTF	CASSLQEWDPNYGYTF	2	1	0	1
CALSVVNQAGTALIF	CASSLVENTEAFF	2	1	0	1
Unique clones		45	13	16	16
Total number of αβ chains		56	17	18	21

cell line subsets (Supplementary Figure 5, Supplementary Table 5). The comparison was performed on the CDR3 region assuming a positive comparison with 100% match. Of these 25 cells, 5 cells were negative in the PCR – for these cells VDJPuzzle analysis had revealed 3 to be α and β negative, but 2 cells had the TCRβ identified but not the TCRα. The remaining 20 cells all had a positive PCR and TCRαβ reconstruction with VDJPuzzle. Of these 20 cells, 14 had an exact PCR match in α chain, and 17 had a match in the β-chain CDR3 sequences. Only 12 cells (48%) were positive for both αβ via PCR. Notably, our method also detected double α chains in 10 cells (18%), and double β chains in 4 cells (7%), which is in line with the previous reports on frequencies of double α and β T cells.[12,19] In addition, VDJPuzzle also identified 2 cells with two productive α and two productive β sequences.

As expected, the successful reconstruction of full TCRαβ was proportional to the abundance of gene expression (Supplementary Figure 6) – the success rate was 100% when the coverage in the constant region was above 800 FPKM. We also examined whether the length of the Illumina paired-end sequences could affect the successful TCR reconstruction. We found that the subset of cells sequenced with Illumina 300 PE reads ($n = 12$ cells) were not statistically different from those sequenced with NextSeq 150PE reads (data not shown), again suggesting that read length was not affecting the results. Although the sample size is relatively small, a surprisingly high level of heterogeneity was observed (Supplementary Table 4).

VDJPuzzle is freely available (https://github.com/Simo-88/VDJPuzzle).

DISCUSSION

We have developed a new method to reconstruct the human TCRαβ from scRNA-seq in Ag-specific T cells. In doing so we have proposed a workflow to link single cell transcriptomics to TCR by means of a pipeline of established and new computational methods. Our approach has been applied to Ag-specific T cells derived from a subject who cleared HCV infection, and showed a very diverse repertoire of TCRαβ specific for a single HLA-A2 restricted HCV epitope. We have also shown that scRNA-seq can be used to distinguish Ag-specific T cell populations directly from PBMC or from an Ag-specific cell line derived from the same subject by identifying a distinct gene expression signature. Notably our analysis showed successful separation between scRNA-seq profiles from resting Ag-specific T cells sorted from peripheral blood and those with a more active phenotype derived from *ex-vivo* stimulation. We concluded that rare Ag-specific T cells circulating in human blood are characterized by a diverse repertoire at both TCRαβ and gene expression profiles.

The application of SmartSeq2 to single Ag-specific CD8+ T cells from human cells provided good quality data. Previous application of this approach has been focussed on cell lines and in samples from animal models, but has been shown to provide a more uniform coverage across the full transcripts when compared to other scRNA-seq methods,[20] thus increasing the likelihood of successful reconstruction of highly polymorphic transcripts such as TCR. Despite the limited sensitivity and specificity of the current scRNA-seq technologies, the computational analyses described here was sufficient to describe gene expression profiles, including the full reconstruction of TCRαβ.

Our new method for reconstructing the full TCRαβ from scRNA-seq, VDJPuzzle, has been validated via single cell PCR amplification showing a high level of sensitivity. Our findings are in line with those obtained using a recently developed method, which was proposed to reconstruct TCRαβ from scRNA-seq from total CD4 T cells in a mouse model of Salmonella infection.[19] Despite the different approaches and cell types between ours, and the recently published algorithm, both methods yielded similar success rates in reconstructing TCRαβ, thus suggesting that scRNA-seq has sufficient quality to be utilised for accurate identification of highly polymorphic transcripts in both humans and other samples. The TCR repertoire diversity identified here in the single Ag-specific T cells was very high, which is in line with previous findings of high TCR diversity in both HCV and also other infections.[21,22] The role of this highly diverse TCR repertoire specific for a single antigen remains unclear.[23] As single cell analysis of these highly diverse Ag-specific TCR repertoires at high sensitivity is now achievable delineation of the importance of such diversity will become possible.[3] The method proposed here provides a powerful tool to unravel the mechanisms driving a successful T cell response.

The method described here may be affected by low specificity and sensitivity.[20] Such limitations are known to arise, for example, from sequence amplification bias[15] and the presence of background noise, which could significantly impair identification of low expression transcripts. In order to reduce the number of falsely identified variable genes detected, we applied a series of statistical approaches, including noise reduction techniques and correction for cell cycle stage that allowed us to determine the set of variable genes with higher sensitivity in the cells studied here. From both PCA and variable gene analyses we observed a broader heterogeneity in PBMC-derived T cells compared to cell line-derived T cells. This is likely to be due to a higher level of noise in PBMC derived Ag-specific T cells that masks the biological variation. Nevertheless, the higher level of heterogeneity and the lower level of gene expression in PB-T cells are perhaps expected as these cells are likely to include only small numbers of residual effectors, but largely resting memory cells (as the subject had cleared the infection more than 6 months prior to the sampling time point analysed here). On the other hand, cell line derived scRNA-seq profiles are affected by active stimulations and further proliferation of cells, which are all factors that are likely to increase the RNA production and the overall physiological activity of the cell.

This work also led to the identification of double α and β chains within individual cells, which has been rarely observed. Indeed, the standard approach for TCR sequencing has been nested PCR followed by Sanger sequencing. This method presents significant issues in detecting two distinct sequences from the same chromatogram, thus limiting the possibility of detecting two dissimilar TCR sequences from a single cell. On the other hand, the possibility of contamination resulting in these double α and β sequences was minimised. Indeed, sorting technologies, based on microfluidics or flow-cytometry, can affect the precision in sorting individual cells. In this work, flow cytometry was employed for single cell sorting (Influx, BD Biosciences, USA) using a drop-single mode, which optimises precision but results in sub-optimal recovery. Similar results showing evidence of double αβ chains were reported in a recent study employing scRNA-seq,[19] thus suggesting that the advance of single cell technologies and sequencing can unravel previously unknown complexity in TCR distribution within individual cells.

As a proof of concept, we have applied this methodology to samples derived from only one subject, but the approach can be readily expanded to larger data sets and more complex scenarios. For instance, longitudinally collected human samples may be utilized to study the relationship of T cell differentiation pathways with the diverse TCR repertoire. An important application of this method is the possibility to link the surface phenotype to the clonotype of T cells using their TCR. For instance, this method can be combined via index sorting to link surface phenotype of Ag-specific T cells with the transcriptomic profiles as well as the TCR clonotype, thus significantly improving the current knowledge on the extent of phenotypic and genotypic diversity that characterise human Ag-specific T cell responses and the role that these play in disease progression. Finally, co-expression analysis of single T cells can be applied to build regulatory networks linked to successful T cell responses to pathogens.

METHODS

Blood sample and processing

Stored PBMC samples from a subject who had previously cleared HCV was made available from the Hepatitis C Incidence and Transmission Study in prisons cohort (HITS-p).[24] Ethical approvals were obtained from Human Research Ethics Committees of Justice Health (reference number GEN 31/05), New South Wales Department of Corrective Services (reference number 05/0884), and the University of New South Wales (reference numbers 05094, 08081) - all located in Sydney, Australia.

Generation of the Ag-specific T cell line

PBMC were separated from freshly collected whole blood by density gradient centrifugation, before 4×10^5 cells per well were cultured in medium (AIM-V with 1% Glutamax, 1% Penicillin-Streptomycin) supplemented with 10% heat inactivated human AB serum (Sigma, USA) in 96-well U-bottomed plates in presence of 10 μg ml^{-1} of the CINGVCWTV peptide (aa 1073–1081 of HCV non-structural 3 protein). For the initial culture, the cytokines, IL-7 (10 ng ml^{-1}, Miltenyi Biotec), IL-12 (300 ρg ml^{-1}, Miltenyi Biotec) and IL-15 (10 ng ml^{-1}, Miltenyi Biotec) were added. Fresh AIM-V media and IL-2 (20 U ml^{-1}, Miltenyi Biotec) was added to the wells on days 4 and 7 post-stimulation. On day 10 post-stimulation, cells were re-stimulated with irradiated autologous monocytes supplemented with IL-2, IL-7, IL-12 and IL-15 in same concentrations as stated for day 1. On day 14 and 17 fresh AIM-V media containing IL-2 (U ml^{-1}) was added to the wells. The established cell line was harvested at day 20 and cryopreserved in RPMI with 20% DMSO. Re-stimulation of the cell line was performed with 10 μg ml^{-1} of the CINGVCWTV peptide per well (2 10^5 cells in each well) for 24 h.

Flow cytometric sorting of HCV-specific CD8+ T cells

HCV-specific CD8+ T cells were stained with NS31073-Dextramer (Immudex; Copenhagen, Denmark). Live/Dead fixable blue dye (Invitrogen) was used to exclude non-viable cells from the analysis. Subsequently, cells were washed and surface stained with the following monoclonal antibodies (mAbs): CD3-APC-Cy7, CD8-Alexa Fluor700 and CD19-PE-Cy5 (BD Biosciences). After the final wash the cells were resuspended in 500 μl of PBS, 2 mM EDTA and 0.5% BSA (Sigma-Aldrich, Germany) solution kept in dark at 4 °C until data acquisition. After exclusion of non-viable/CD19+ cells, CD3+ CD8+ dextramer+ cells were sorted directly into 96-well PCR plates (Eppendorf) using a BD Influx cell sorter (BD Influx; BD Biosciences, Franklin Lakes, NJ) and stored at − 80 °C for subsequent TCRαβ analysis.

Single cell RNA sequencing

Single cell RNA sequencing was performed as described by Picelli *et al.*[13,25] with the following modifications. Cell lysis buffer was prepared by adding 1 μl RNase inhibitor (Clontech, Mountain View, CA) to 19 μl Triton X-100 solution (0.2% v/v). Cells were sorted directly into tube containing 2 μl lysis buffer, 1 μl dNTP mix (10 mM) and 1 μl oligo-dT primer at 5 μM. Tubes were stored at -80 °C until needed. Reverse-transcription and PCR amplification were performed as described[13] with the exception of reducing the IS PCR primer to a 50 nM final concentration and increasing the number of cycles to 28. Sequencing libraries were prepared using the Nextera XT Library Preparation

Kit (Illumina; San Diego, CA, USA) and sequencing was performed on Illumina MiSeq and NextSeq sequencing platforms.

RNA-seq bioinformatics and computational analyses

For standard RNA-seq analysis, the quality of Illumina reads was assessed with FastQC, next reads have been trimmed to 100 nt using Trimmomatic.[26] Reads were then aligned to the UCSC hg18 reference genome using TopHat2.[27] FPKMs were estimated with cuffquant and cuffnorm (http://cole-trapnell-lab.github.io/cufflinks/). Further data analysis on transcriptomics was performed with in-house R scripts. Cell cycle noise analysis was performed with scLVM.[14] Variable genes were obtained by fitting a linear model of the coefficients of variation as a function of the mean of expression of each gene;[16] this was repeated for each experimental cell set. Variable genes were selected with a FDR of 0.05 (*P*-value of deviation from the fitted model <0.05). Differentially expressed genes were detected with Monocle. More detail and source. GO enrichment analyses were performed with GOrilla[28] using the list of genes expressed in at least one cell as background.

VDJPuzzle

Reconstruction of the TCR repertoire from scRNA-Seq data was undertaken using VDJPuzzle. This software implements a method based on selecting reads that either overlap with the constant segment, or with the V(D)J region, of a reference set of TCR gene sequences. The selected reads are then used to reconstruct a *de novo* assembly (using Trinity[29]) of both the α and β chains. The assembled sequences are filtered with IgBlast[30] with the respect to the IMGT database, and then used to build a new reference TCR. For each cell, all of the paired end reads are re-aligned against the preliminary TCR sequences (resulting from the *de novo* assembly as above) using Bowtie2,[31] allowing for alignment against ambiguous nucleotides within a reference and for gaps into both the reference and read sequences. The matching reads are then extracted and assembled again with Trinity[29] to generate the final TCRαβ repertoire. The generation of a second assembly (using again Trinity) is performed to increase the likelihood of detection of reads overlapping with the TCRαβ transcripts which are highly diverse compared to the reference genome. Clonotype and CDR3 are identified using MigMap (https://github.com/mikessh/migmap). MigMap is a wrapper to IgBlast that extracts the CDR3 region directly from the reads aligned against the reference genome. The output provides an identification string for the V, D and J segments and of the CDR3. It reports the position of the Cys starting residue of variable segment, and the Phe/Trp residue of J segment that marks the end of CDR3. MigMap was utilized to identify productive TCRs from the complete reconstructed TCR sequence set by checking for in frame amino acid translations and the absence of stop codons. Positive identification of TCR from the reconstructed transcript was accepted only if their reported V and J alignments had Evalues below 0.01. The details of this software are available at https://github.com/Simo-88/VDJPuzzle.

TCR sequencing via RT-PCR

Full length (oligodT primed) cDNA produced from single cells was used as template in two PCR reactions, one for each TCR locus. This was performed using a method first explained by,[32,33] which has been adapted with some modifications. Full-length (oligodT primed) cDNA produced from single cells was used as template. TCR transcripts from each cell were amplified by multiplex nested PCR as previously reported,[32] with some modifications. Reactions for TCRα and β were separated into two separate 96 well PCR plates, internal forward TRAV and TRBV were used at 2.5 pmol concentration, and internal TRAC and TRBC primers were used at 10 pmol, with 10x coralload PCR buffer (Qiagen, Germany). The PCR conditions were 95 °C for 2 min, followed by 35 cycles od 95 °C for 20 s, 52 °C for 20 s, and 72 °C for 45 s, the final extension was performed at 72 °C for 7 min. PCR cleanup was performed by illustra ExoProStar (GE healthcare life Sciences), briefly, 1 μl of ExoProStar was added to 5 μl of positive PCR product and the mixture was incubated at 37 °C for 15 min followed by a 80 °C inactivation for 15 min. Sequencing reactions were performed using the 5 pmol of TRAC internal of TRBC internal primers for TCRα and β respectively. The sequencing reactions were performed in 20 μl reaction mix using BigDye Terminator v3.1 cycle sequencing kit (Life technologies) containing, 1 μl of DMSO, 4 μl 5x sequencing buffer, 1 μl of big dye terminator v3.1 and 6 μl of PCR product. Purified product was submitted for Sanger sequencing at the Ramaciotti Centre for Functional Genomics at the University of New South Wales, Australia

Comparison of PCR and RNAseq data

Reconstructed CDR3 regions from the RNA-seq derived sequences were aligned against the Sanger sequences of the CDR3 region derived from the direct TCRαβ PCR products. The CDR3 regions from the Sanger sequences were identified using IgBlast. For each cell, only 100% matched sequences between the two methods were considered concordant.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We acknowledge NHMRC and ACH2 for funding. SR is supported by the University International Postgraduate Award UNSW Australia. ARL is supported by an NHMRC Practitioner Fellowship (No. 1043067), RAB is supported by an NHMRC Career Development Fellowship (No. 1060443), VV is supported by an NHMRC Career Development Fellowship (1067590), and KK is supported by an NHMRC SRFB fellowship (1102792). We are grateful to Simone Picelli for helpful clarifications on the SmartSeq2 protocol, and to Oanh Nguyen for technical assistance with the TCR-PCR protocol.

Author contributions: FL designed the research. AAE performed the scRNA, MR performed the flow cytometry and the PCR TCRαβ experiments. SR developed the computational methods, SR, FL, AAE, MR, RAB analysed the data. VV, BDB-S, KK, ARL contributed to methods and provided reagents and samples. FL, RAB, ARL supervised experiments. SR, FL, AAE, MR wrote the manuscript. All authors reviewed the final manuscript.

1 Kaech SM, Cui W. Transcriptional control of effector and memory CD8+ T cell differentiation. *Nat Rev Immunol* 2012; **12**: 749–761.
2 Kaech SM, Wherry EJ. Heterogeneity and cell-fate decisions in effector and memory CD8+ T cell differentiation during viral infection. *Immunity* 2007; **27**: 393–405.
3 Newell EW, Davis MM. Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. *Nat Biotechnol* 2014; **32**: 149–157.
4 Bull RA, Leung P, Gaudieri S, Deshpande P, Cameron B, Walker M *et al.* Transmitted/Founder Viruses Rapidly Escape from CD8+ T Cell Responses in Acute Hepatitis C Virus Infection. *Journal of virology* 2015; **89**: 5478–5490.
5 Turner SJ, La Gruta NL, Kedzierska K, Thomas PG, Doherty PC. Functional implications of T cell receptor diversity. *Curr Opin Immunol* 2009; **21**: 286–290.
6 Chattopadhyay PK, Gierahn TM, Roederer M, Love JC. Single-cell technologies for monitoring immune systems. *Nature immunology* 2014; **15**: 128–135.
7 Buchholz VR, Flossdorf M, Hensel I, Kretschmer L, Weissbrich B, Graf P *et al.* Disparate individual fates compose robust CD8+ T cell immunity. *Science* 2013; **340**: 630–635.
8 Gerlach C, Rohr JC, Perie L, van Rooij N, van Heijst JW, Velds A *et al.* Heterogeneous differentiation patterns of individual CD8+ T cells. *Science* 2013; **340**: 635–639.
9 Fan HC, Fu GK, Fodor SP. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* 2015; **347**: 1258367.
10 Dash P, Wang GC, Thomas PG. Single-Cell Analysis of T-Cell Receptor alphabeta Repertoire. *Methods Mol Biol* 2015; **1343**: 181–197.
11 Linnemann C, Heemskerk B, Kvistborg P, Kluin RJ, Bolotin DA, Chen X *et al.* High-throughput identification of antigen-specific TCRs by TCR gene capture. *Nat Med* 2013; **19**: 1534–1541.
12 Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat Biotechnol* 2014; **32**: 684–692.
13 Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* 2014; **9**: 171–181.
14 Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* 2015; **33**: 155–160.
15 Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015; **16**: 133–145.
16 Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* 2013; **10**: 1093–1095.
17 Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014; **32**: 381–386.
18 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9**: 559.

19 Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G *et al.* Simultaneously inferring T cell fate and clonality from single cell transcriptomes. *bioRxiv* (e-pub ahead of print 28 August 2015; doi:http://dx.doi.org/10.1101/025676).
20 Grun D, van Oudenaarden A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* 2015; **163**: 799–810.
21 Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* 2004; **4**: 123–132.
22 Attaf M, Huseby E, Sewell AK. Alphabeta T cell receptors as predictors of health and disease. *Cell Mol Immunol* 2015; **12**: 391–399.
23 Miles JJ, Douek DC, Price DA. Bias in the alphabeta T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol* 2011; **89**: 375–387.
24 Luciani F, Bretana NA, Teutsch S, Amin J, Topp L, Dore GJ *et al.* A prospective study of hepatitis C incidence in Australian prisoners. *Addiction* 2014; **109**: 1695–1706.
25 Picelli S, Björklund ÅKK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* 2013; **10**: 1096–1098.
26 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–2120.
27 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 2013; **14**: R36.
28 Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 2009; **10**: 48.
29 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011; **29**: 644–652.
30 Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* 2013; **41** (Web Server issue): W34–W40.
31 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods* 2012; **9**: 357–359.
32 Wang GC, Dash P, McCullers JA, Doherty PC, Thomas PG. T cell receptor alphabeta diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection. *Sci Transl Med* 2012; **4**: 128ra42.
33 Nguyen TH, Rowntree LC, Pellicci DG, Bird NL, Handel A, Kjer-Nielsen L *et al.* Recognition of distinct cross-reactive virus-specific CD8+ T cells reveals a unique TCR signature in a clinical setting. *J Immunol* 2014; **192**: 5039–5049.

The Supplementary Information that accompanies this paper is available on the Immunology and Cell Biology website (http://www.nature.com/icb)