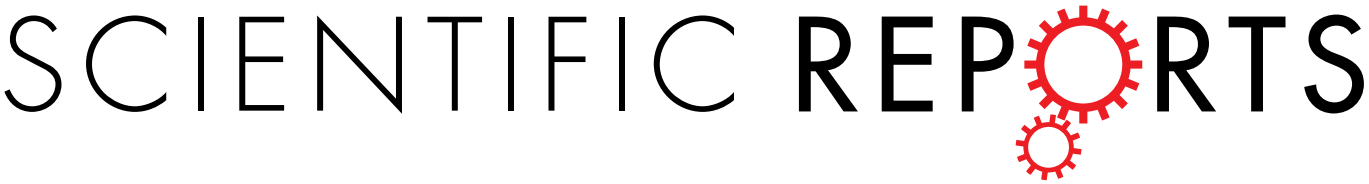

OPEN

Impact of sequencing depth and read length on single cell RNA sequencing data of T cells

Received: 17 February 2017
Accepted: 14 September 2017
Published online: 06 October 2017

Simone Rizzetto[1,2], Auda A. Eltahla[1,2], Peijie Lin[3,4], Rowena Bull[1,2], Andrew R. Lloyd[1,2], Joshua W. K. Ho[3,4], Vanessa Venturi[5] & Fabio Luciani[1,2]

Single cell RNA sequencing (scRNA-seq) provides great potential in measuring the gene expression profiles of heterogeneous cell populations. In immunology, scRNA-seq allowed the characterisation of transcript sequence diversity of functionally relevant T cell subsets, and the identification of the full length T cell receptor (TCRαβ), which defines the specificity against cognate antigens. Several factors, e.g. RNA library capture, cell quality, and sequencing output affect the quality of scRNA-seq data. We studied the effects of read length and sequencing depth on the quality of gene expression profiles, cell type identification, and TCRαβ reconstruction, utilising 1,305 single cells from 8 publically available scRNA-seq datasets, and simulation-based analyses. Gene expression was characterised by an increased number of unique genes identified with short read lengths (<50 bp), but these featured higher technical variability compared to profiles from longer reads. Successful TCRαβ reconstruction was achieved for 6 datasets (81% − 100%) with at least 0.25 millions (PE) reads of length >50 bp, while it failed for datasets with <30 bp reads. Sufficient read length and sequencing depth can control technical noise to enable accurate identification of TCRαβ and gene expression profiles from scRNA-seq data of T cells.

Single cell RNA sequencing (scRNA-seq) has vastly improved our ability to determine gene expression and transcript isoform diversity at a genome-wide scale in different populations of cells. scRNA-seq is becoming a powerful technology for the analysis of heterogeneous immune cells subsets[1,2] and studying how cell-to-cell variations affect biological processes[3,4]. Despite its potential, scRNA-seq data are often noisy, which are caused by a combination of experimental factors, such as the limited efficiency in RNA capture from single cells, and also by analytical factors, such as the challenges in separating true variation from technical noise[5–7]. The quality of scRNA-seq data depends on mRNA capture efficiency[8], the protocol utilised to obtain libraries, as well as sequence coverage and length[3,4]. Bioinformatics tools for the analyses of scRNA-seq data have been rapidly evolving, whereby various algorithms have been proposed to resolve the issues related to scRNA-seq compared to classical bulk transcriptomic analysis[9–11]. However, the lack of a consensus in the data analyses further contributes to difficulties in assessing the quality of the data analysed so far.

One important consideration in designing scRNA-seq experiments is to decide on the desired sequencing depth (*i.e.*, the expected number of reads per cell) and read length[3,6]. These are two important experimental parameters that can be controlled, and which need to be often predetermined before sequencing. For bulk RNA-seq data, sequencing depth and read length are known to affect the quality of the analysis[12]. For scRNA-seq it has been shown that half a million reads per cell are sufficient to detect most of the genes expressed, and that one million reads are sufficient to estimate the mean and variance of gene expression[13]. Low coverage scRNA-seq has also been utilised to show that 50,000 reads per cell are sufficient to classify a cell type in a sample of 301 cells[14]. Nevertheless, this may not be sufficient when more homogenous populations are involved, for example T cell subsets, such as central memory and effector memory cells. In these scenarios, deep sequencing of single cell library may be required for improving detection of genes with low expression[3,6]. Indeed, an important issue for scRNA-seq data is the very large number of genes with no detectable expression in a cell[6]. This overrepresentation

[1]School of Medical Sciences, UNSW, Sydney, Australia. [2]Viral Immunology Systems Program, Kirby Institute for Infection and Immunity, UNSW, Sydney, Australia. [3]Victor Chang Cardiac Research Institute, Sydney, NSW, Australia. [4]St. Vincent's Clinical School, UNSW, Sydney, Australia. [5]Infection Analytics Program, Kirby Institute for Infection and Immunity, UNSW, Sydney, Australia. Correspondence and requests for materials should be addressed to F.L. (email: Luciani@unsw.edu.au)

Dataset	Dataset	Reference	Accession number	Organism	Number of cells	Average reads length (bp)	Average number of PE reads (x10^6 reads)	scRNA-seq protocol	Number of genes expressed (FPKM >1) per cell
1	HCV specific CD8 + T cells	Elthala[20]	E-MTAB-4850	human	54	145	8.4	Smart-Seq. 2	2,563
2	HCV specific CD8 + T cells	Elthala[20]	E-MTAB-4850	human	12	215	3.5	Smart-Seq. 2	3,289
3	Th17 cells (A)	Gaublomme[24]	GSE74833	mouse	399	125	2.5	Smart-Seq	5,128
4	Th17 cells (B)	Gaublomme[24]	GSE74833	mouse	269	100	3.7	Smart-Seq	6,540
5	Th17 cells (C)	Gaublomme[24]	GSE74833	mouse	100	25	1.5	Smart-Seq	4,146
6	CD4 + T cells	Stubbington[21]	E-MTAB-3857	mouse	272	100	4.3	Smart-Seq	2,354
7	CD8 + T cells	Kimmerling[25]	GSE74923	mouse	106	32	1.2	Smart-Seq. 2	6,796
8	Th2	Mahata[26]	E-MTAB-2512	mouse	93	75	16.3	Smarter-Seq	6,401

Table 1. scRNA-seq data sets analysed in this study. List of dataset used for the analysis.

of zeros in scRNA-seq datasets makes it difficult to distinguish technical dropout of transcripts from true biological variation between cells[3] and statistical methods have been developed to at least control this issue (e.g.[15]).

Nonetheless, there has not been any systematic evaluation of the effect of sequencing depth and read length on scRNA-seq data analysis. In designing an scRNA-seq experiment it is optimal to generate data by maximising sequencing depth and utilising the longest read length. This approach would improve the quality of the reads alignment and also maximise the chance of detecting low abundant transcripts. In reality, we are often constrained by the cost of sequencing. Therefore a more practical question is to ask what is the minimum sequencing depth and read length that allows users to obtain adequate information for their desired downstream analyses.

To answer these questions, we have focussed on assessing the quality of available scRNA-seq data from T cells, which form a highly heterogeneous population of lymphocytes that play a vital role in mounting successful adaptive immune responses against intracellular pathogens and tumours[16]. T cells are also characterised by a highly diverse repertoire of T cell receptors (TCRs), which identify the specific recognition of the cognate antigen. TCRs are heterodimer proteins composed of two chains, α and β, and a subset of those expressing the $\gamma\delta$ chains, which result from genetic recombination of the V(D)J genes. The diversity of TCR$\alpha\beta$ repertoire has been associated with successful control of many pathogens[17], and more recently with outcome of checkpoint inhibitor immunotherapy for patients with metastatic melanoma[18]. The third complementarity-determining region (CDR3) of the TCR α and β chains forms loops that engage amino acid residues of peptides in complex with MHC. Detection of the CDR3 region is a crucial step to accurately identify the clonality of a T cell repertoire, for instance responding to a viral infection. The highly polymorphic nature of the TCR genes has made their identification very difficult in bulk population sequencing datasets. In the last decade, deep sequencing approaches of bulk TCRs focussed on either α or β chains[19]. The advent of scRNA-seq allowed the identification of the full length TCR of both α and β chains (referred to hereafter as TCR$\alpha\beta$) from T cells[20,21]. This has now led to the capacity to simultaneously detect TCR$\alpha\beta$ and full gene expression profiles in one experiment, thereby allowing direct study of TCR diversity and its interaction with the T cell functions reflected in gene expression profiles.

In this study we performed a comprehensive analysis of the impact of sequencing depth and read length on the detection of full length TCR$\alpha\beta$ sequences, as well as estimation of gene expression and its effect on cell-type identification. Our study aims to fill this gap through performing a re-analysis of eight published scRNA-seq data that have a wide range of read length and sequencing depth, and analysis of simulated datasets that were subsampled from a deeply sequenced human T-cell scRNA-seq dataset. The analysis suggests important precautionary steps for researchers seeking to maximise throughput of single cell experiments without compromising the quality of the results.

Results

To assess the effects of sequencing depth and read length on accurate reconstruction of full length TCR$\alpha\beta$ and gene expression profile from scRNA-seq data, we manually reviewed NCBI's Gene Expression Omnibus[22] and ArrayExpress[23] to identify relevant T-cell scRNA-seq data published prior to April 2016. Eight datasets were identified with accessible data, collectively profiling 1,305 single cells (Table 1). The datasets were generated from mouse[21,23–26] and human-derived cells[20], utilising one of the available versions of the Smart-Seq protocol[27], and had a wide range of sequencing depth (1.2–8.4 million paired-end (PE) reads per cell) and read length (25–215 bp) (Table 1). The mean number of expressed genes in each data set ranged between 2,354 and 6,795 (Table 1).

The effect of sequencing depth and read length on reconstruction of full-length T-cell receptors. We analysed whether sequencing depth and read length affect the detection and reconstruction of TCR$\alpha\beta$. Two recently developed bioinformatics methods for reconstruction of full-length TCR$\alpha\beta$ from scRNA-seq data were used, TraCeR[21] and VDJPuzzle[20]. The analysis performed with VDJPuzzle revealed successful TCR$\alpha\beta$ reconstruction in 1027 cells (79%) (Table 2). This result was consistent with the results from TraCeR, with successful TCR$\alpha\beta$ reconstruction from 953 cells (73%) (Table S1). Six of the eight datasets had a success rate >80% in detection of TCR$\alpha\beta$, and up to 100% for the scRNA-seq dataset with an average read length of 215 bp. The two datasets with lowest detection rate of TCR$\alpha\beta$ had 25 and 32 bp long reads, where only 0% and 1.89% of the cells successfully generated TCR$\alpha\beta$ sequences, respectively (Table 2 and Fig. 1A). In terms of sequencing

Dataset	Number of cells	Average reads length (bp)	Average number of PE reads (x10^6 reads)	TCRα success rate (%) VDJPuzzle	TCRβ success rate (%) VDJPuzzle	TCRαβ success rate (%) VDJPuzzle
1	54	145	8.4	81.48	85.19	81.48
2	12	215	3.5	100.00	100.00	100.00
3	399	125	2.5	99.25	98.75	98.50
4	269	100	3.7	98.51	98.88	97.77
5	100	25	1.5	0.00	0.00	0.00
6	272	100	4.3	89.71	93.38	85.66
7	106	32	1.2	1.89	7.55	1.89
8	93	75	16.3	89.25	93.55	86.02

Table 2. The success rate of reconstructing full-length T-cell receptors (TCR) using VDJPuzzle for the eight scRNA-seq data sets. Success rates for TCRαβ detection in each dataset.

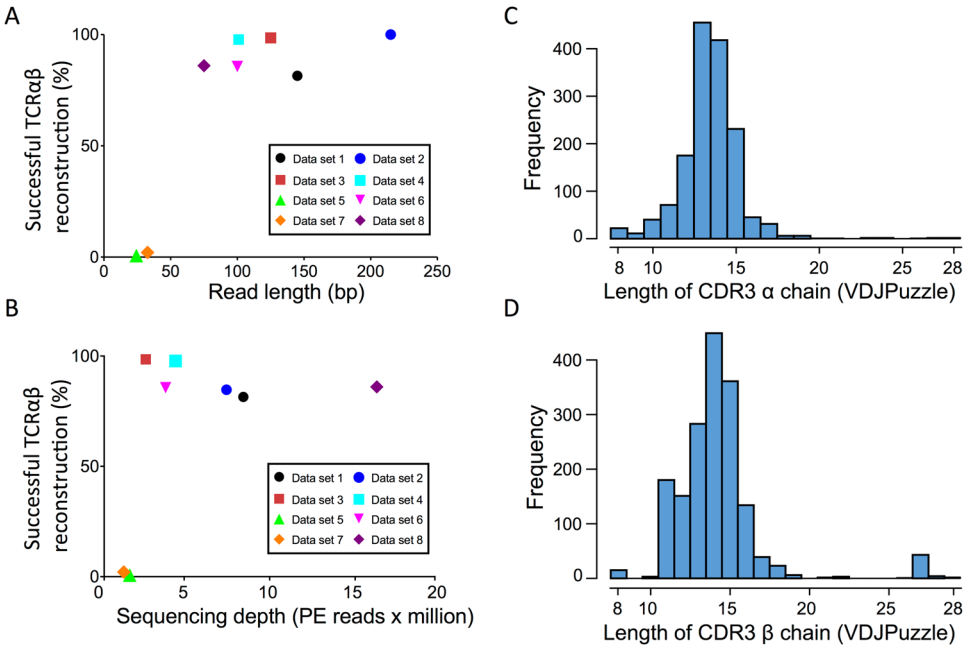

Figure 1. Success rates of TCRαβ reconstruction as a function of read length (**A**) and sequencing depth (**B**) using VDJPuzzle. Panels C and D show the distributions of the length of the reconstructed CDR3α and CDR3β regions, respectively.

depth, in datasets with less than 1.5 million PE reads TCRαβ were successfully detected in less than 1% of the cells, and this success increased rapidly to >80% for depths >0.25 million PE reads (Fig. 1B).

To further assess the quality of the reconstruction of TCRαβ sequence, we analysed the distribution of CDR3 amino acid sequences across both α and β chains, and the distribution of single cells carrying double α chains. The average CDR3 length of the reconstructed TCRαβ sequences with VDJPuzzle was 14 amino acids for both α and β chains (Fig. 1C and D), with similar results using TraCeR (Fig. S1). This result showed a distribution of CDR3 lengths consistent with those previously estimated with other methods, such as 5′-Race for single cell TCR analysis[28].

One of the major advantages of using scRNA-seq to reconstruct TCR sequences is the possibility to detect double α chains within a single T cell. Overall, 30% (n = 395) of the cells analysed here presented more than one α but not double β. In a single study (datasets 3 and 4 in Table 1), 43% (n = 333) of the cells sequenced presented more than one α, and 44% (n = 337) had more than one β sequence detected. Notably, 29% (n = 225) of these cells had both more than two unique α and two unique β chain sequences, thus suggesting that in this study multiple cell could have been sorted in a single well. In support of this conclusion, the plot of the number of unique genes identified in those cells with two or more than unique α and two unique β chain sequences showed a significantly higher number of gene counts when compared to the remaining cells (Fig. S2). By filtering out cells with more than one α and one β, a total of 309 unique TCRαβ sequences were identified across all datasets. There was no clonotype (defined as cells bearing identical TCRαβ) overlapping between datasets.

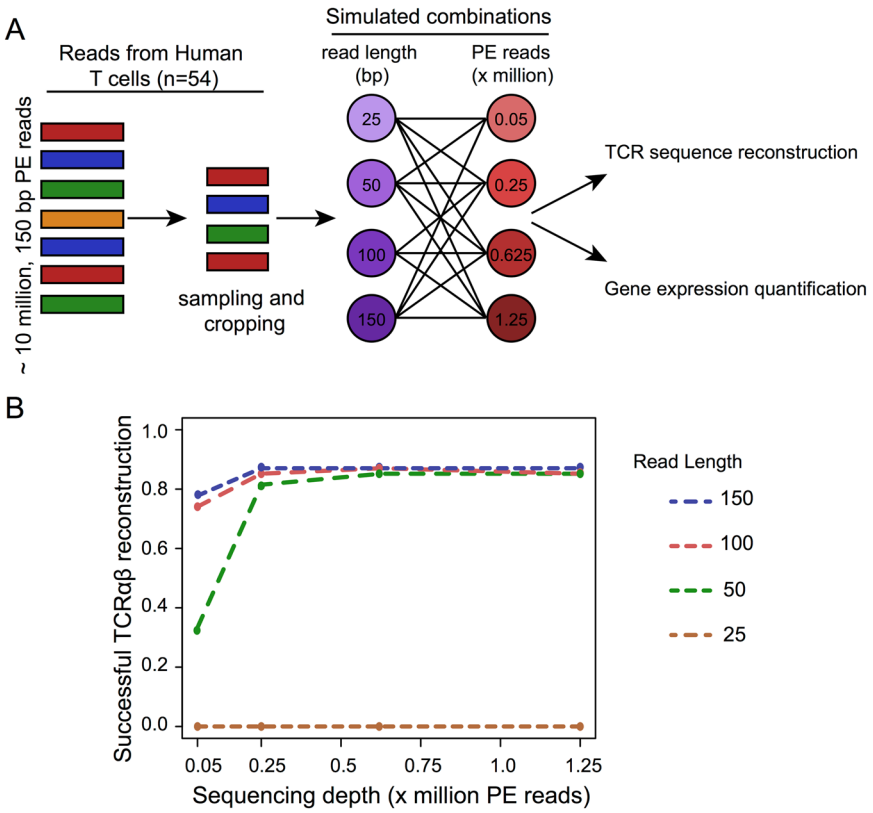

Figure 2. (**A**) Generation of the simulated datasets from real scRNA-seq data 1. (**B**) Success rate for TCRαβ reconstruction as a function of read length and sequencing depth from the simulated datasets.

Effect of sequencing depth and read length on the TCRαβ detection using simulated datasets. To systematically investigate the effect of sequencing depth and read length, we generated simulated datasets with different sequencing depth and read length to assess the success rate of TCRαβ reconstruction. Simulated datasets were all derived from the original datasets 1, which had deep coverage (~8.4 million PE reads per cell), and long read length (150 bp) (Table 1). The original datasets consisted of a total of 54 single cells originated from HCV specific CD8+ T cells from a single subject that previously cleared HCV. Of these cells, 18 were directly sorted from peripheral blood mononuclear cells (PBMC-derived T cells) and the remaining 36 were sorted after *in vitro* expansion following stimulation with cognate antigen. Of these 36, 18 were sorted after a second antigen restimulation 24 hours prior to sorting[20]). From each of the original single cell data (n = 54), we generated 16 randomly subsampled scRNA-seq datasets with all combinations of four different sequencing depths (0.05, 0.25, 0.625 and 1.25 million PE reads) and four different read lengths (25, 50, 100 and 150 bp) (Fig. 2A). For each of the 16 subsampled datasets, the TCRαβ sequence was reconstructed using VDJPuzzle[20], and the success rate was calculated (Figs 2B and S3). Only TCRαβ sequences with a complete CDR3 recognised by the international ImMunoGeneTics information system (IMGT,[29]) were considered as an exact TCRαβ reconstruction.

Success rate of paired α and β was above 80% for datasets which had a minimum read length of 50 bp and a depth of at least 0.25 million reads. This rate was substantially diminished up to 0% for datasets with a number of PE reads per cell below 0.25 million PE reads (Fig. 2B). Finally, the proportion of cells with double α detected was also proportional to both read length and sequencing depth, with the highest success rate corresponding to a depth of 1.25 million PE reads and a read length above 100 bp (Fig. S4). The relationship between the success rate of TCRαβ reconstruction and both sequencing depth and read length was fitted with a sigmoidal function (Fig. S3). The success rate in TCRαβ reconstruction from the experimental datasets (the real dataset) closely followed this specific relationship ($r = 0.97$, $p < 0.01$).

The effect of read length and sequencing depth on the quantification of the gene expression profile. Next, we used the 16 subsampled scRNA-seq datasets to investigate the effect of sequencing depth and read length on read alignment and gene expression quantification. Surprisingly, we observed a slight increase in the total number of aligned PE reads in datasets with shorter read length, especially when the read length was below 100 bp (Fig. 3). This higher level of total read alignment at short read length can be attributed to an increased proportion of reads with multiple alignments, and more discordant alignment of PE reads (Fig. 3). Notably, this relationship with read length was also observed for the proportion of concordant pairs aligned, but with a lower proportion for reads of 25 bp long compared to 50 bp. The reason for such a trend is largely due to the increased number of reads that in general align when read length is <100 bp (first column of Fig. 3). This is due to the fact that shorter reads are more likely to be aligned anywhere in the genome compared to longer ones. Indeed,

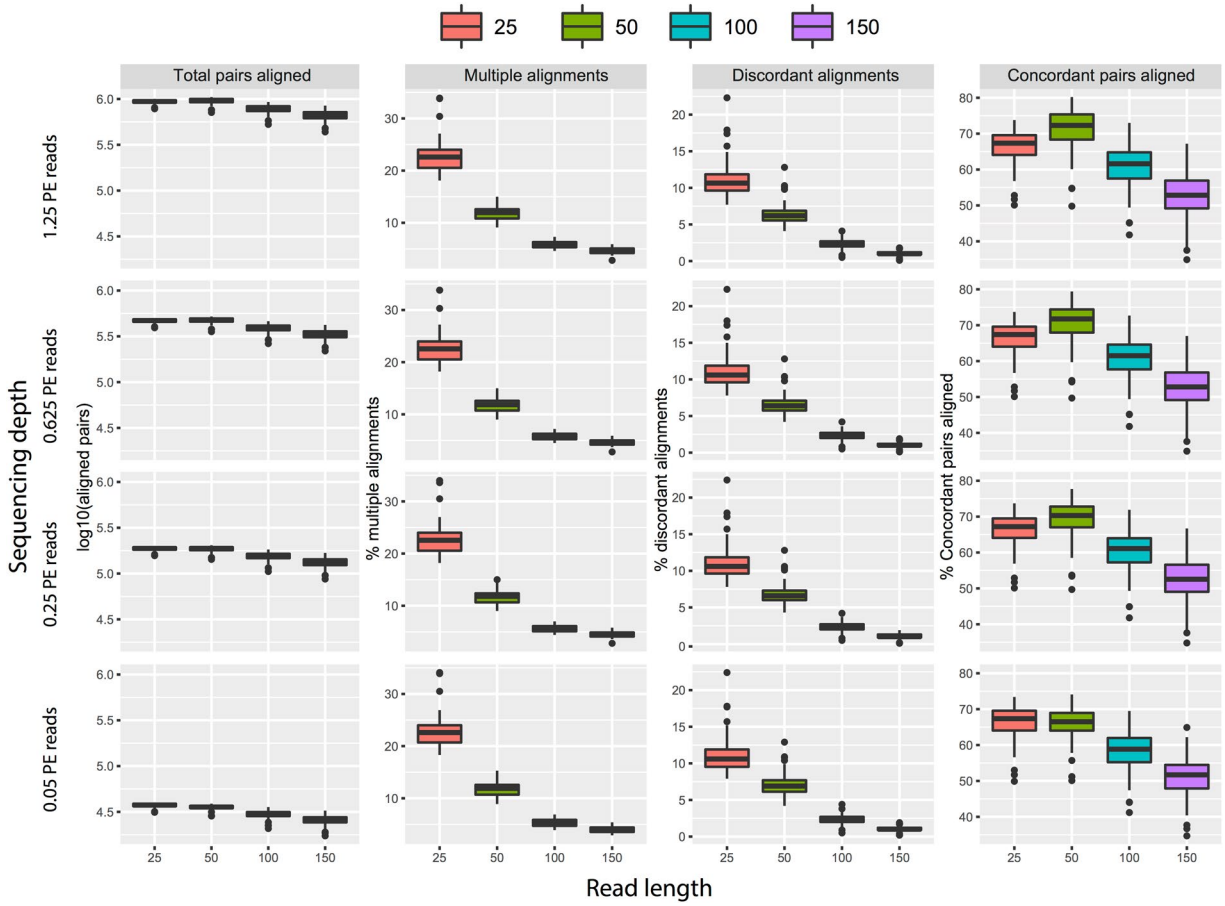

Figure 3. Analysis of the alignment of the simulated datasets as a function of sequencing depth and read length. Shown is the number of paired-end reads aligned (in log10 scale), along with the proportion of concordant and discordant pairs, and of multiple alignment instances.

for read length of 25 bp and 50 bp the proportion of multiple, discordant, and paired concordant reads aligned are all increased compared to long read. In regard of sequencing depth, we observed an increase in the number of aligned reads of 50 bp compared to 25 bp for high coverage data (top rows in Fig. 3). This phenomenon is likely due to the fact that shorter reads (25 bp) have lower mapping quality caused by the very large number of multiple alignments (Fig. 3). Indeed, the aligner (bowtie) assigns a low score to reads that can be aligned multiple times as their correct position is uncertain. This phenomenon is less evident for low coverage where reads aligning to the genome are the limiting factor and are more influenced by sampling bias.

To assess the effect of this trend on the quantification of genes, fragments per kilo base per million (FPKM) were calculated allowing only one alignment per read, hence eliminating a potential confounding factor of multiple alignments. We found that the number of detectable expressed genes (those with FPKM >1) was positively correlated with sequencing depth (Pearson correlation $=0.89$) but negatively correlated with read length (Pearson correlation $=-0.93$). The number of genes that were expressed in at least 10% of the cells showed a similar correlation with sequencing depth and read length (Table 3, Fig. 4A). Notably, there was a positive relationship between number of genes expressed among cells within the same dataset and read length for sequencing depth smaller than 0.625 million PE reads, while there was no variation at higher sequencing depths (Fig. 4B).

In order to quantify the reliability of the gene expression profile as a function of read length and sequencing depth, two simulated datasets with a sequencing depth of 0.05 million PE reads were generated, with read length of 25 bp and 150 bp, respectively. Two replicates for each dataset were simulated. This analysis showed a significantly higher correlation between the gene expression profiles of paired cells from the two replicates with read length 150 bp when compared to the two replicates with read length 25 bp (Fig. 4C). This result suggested that gene expression profiles from short read length dataset have higher levels of technical noise.

To further assess how the technical variation generated by shorter read length and lower sequencing depth affects the identification of the three cell sub-populations available from the experimental scRNA-seq data of HCV specific T cells[20], a clustering algorithm was applied on all the simulated datasets. A newly developed bioinformatics tool CIDR[30] was used to perform dimensionality reduction, Principal Coordinates Analysis (PCoA) and hierarchical clustering on the scRNA-seq gene expression profiles. When cutting the hierarchical clustering to form three clusters based on the experimentally validated cell subsets in the original data (i.e., the ground truth defined by the known cell types in the data set[20]), CIDR achieved the tightest clustering when the dataset has $>=100$ bp long PE reads (Figs 5A,B and S5). This was evident by the higher misclassification rate calculated

Sequencing depth (PE reads x million)				
Read length (bp)	0.05	0.25	0.625	1.25
25	6,081	8,497	9,184	10,849
50	5,665	7,801	8,255	8,240
100	5,141	6,879	7,333	7,440
150	4,836	6,458	6,824	6,936

Table 3. Number of genes expressed in at least 10% of the cells in the simulated data sets, comprised of subsamples of the scRNA-seq data set 1, with various sequencing depths (columns) and read lengths (rows). Analysis of empirical drop out rate on simulated datasets.

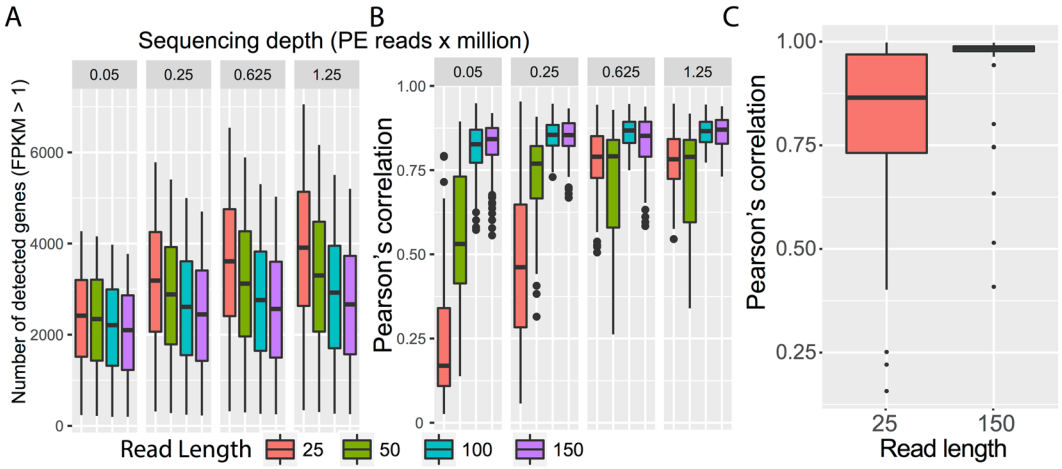

Figure 4. The effect of read length and sequencing depth on the technical error variability using simulated scRNA-seq datasets. A: Number of identified expressed genes (Fragment per Kilobase per Million reads; FPKM >1) as a function of read length and sequencing depth (**A**). Error bars (box plot, mean and 5–95% interval) represent variability across individual cells. (**B**) Mean pairwise cell-to-cell Pearson correlation of gene expression values as a function of sequencing depth and read length. (**C**) The distribution of pairwise cell-to-cell Pearson correlation of gene expression values using subsets of different read length drawn from the original dataset. Original dataset had a read length of 145 bp with depth >8 millions PE reads, two samples drawn from this dataset were taken, with length 25 bp and same depth.

from the clustering analysis with shorter read length: 28% (15/54 of cells were misclassified for read length 25 and 50 bp, and 9% (5/54 for read length 100 and 150 bp (Figs 5C and S5). Sequencing depth did not affect the misclassification rate. To investigate whether the 'tightness' of the clustering is affected by sequencing depth and read length, the within-cluster-sum-of-squares of each cell type was computed. Consistent with the misclassification analysis, longer reads led to tighter clusters, reflected by a substantial decrease in within-class-sum-of-squares for PBMC derived Ag CD8+ T cells (Figs 5D and S5). The effect of read length was less pronounced for the other two *in vitro* expanded subpopulations, as these are biologically more close to each others when compared to the blood derived original population.

To analyse the effect of read length and sequencing depth on specific gene categories, the distribution of gene expression levels (in terms of log(FPKM)) was analysed for highly expressed genes (average FPKM >100), lowly expressed genes (average FPKM <100), housekeeping genes, and transcription factors in all the subsampled simulated datasets. Independent of the gene category, there was a reduction in the number of genes identified with an expression level below 100 FPKM in datasets with a low sequencing depth (<0.05 PE reads x million, Fig. S6). This effect was more evident among the transcription factors, where a combination or short read length and low depth led to a complete loss of lowly expressed genes. There was an increase in the frequency of highly abundant genes with the decrease of read length. To illustrate these trends, six individual genes were considered: three housekeeping genes (*GAPDH, RPL7A*, and *RPL34*), two genes constitutively expressed in CD8+ T cells (*CD8B* and *TRAC*), and one transcription factor (*GAS5*, which is associated with T cell proliferation[31]). The analysis showed that, contrary to the expectation, the gene expression profile of the selected housekeeping genes varied significantly for low depth and short reads (Fig. 6). Notably, the housekeeping gene RPL34 did not vary as much as GAPDH and RPL7A for low depth and short reads (Fig. 6). *GAPDH* and *CD8B* expressions were positively correlated with the read length, while a significant variability was detected for *GAS5*, independent of sequencing depth and read length. *TRAC* did not show any substantial variation.

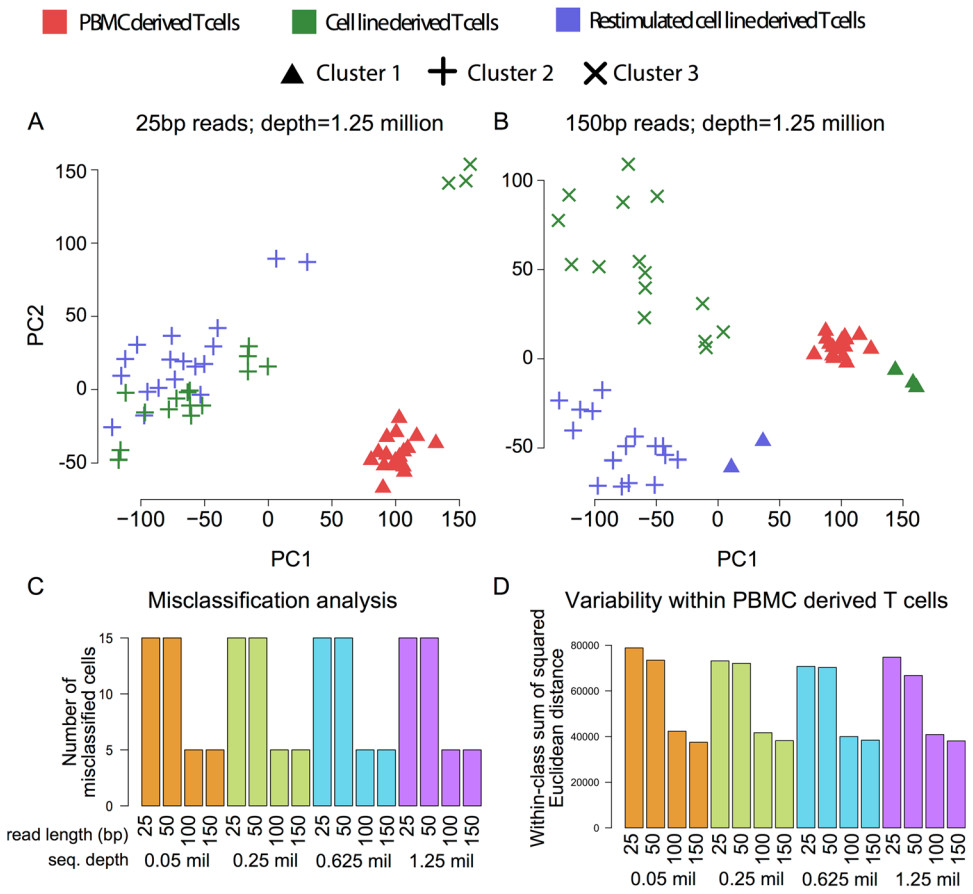

Figure 5. Clustering analysis for the three populations of HCV specific CD8+ T cells. Panels A and B display Principle Coordinate Analysis of the three subsets of cells by varying read length (25 to 150 bp). Coverage for each dataset was set to 1.25 millions of PE reads per cell. The point colours correspond to the 'ground truth' cell type labels (see legend), while the three point styles correspond to the three identified clusters (circle, triangle and cross). Clustering analysis was performed using CIDR, and forcing the number of clusters to be n = 3. Panels C and D display the misclassification and the variability within the same cell type (within-class sum of squares) as a function of read length and sequencing depth, respectively. Panel D displays only results from PBMC-derived T cells.

Discussion

This study explored how sequencing depth and read length of scRNA-seq dataset affect various downstream analyses, such as transcript reconstruction, gene expression estimation and cell-type identification. The overall messages of this study can be summarised with two major findings. Firstly, by combining available algorithms for TCRαβ detection, along with simulation-based analysis, this study revealed that accurate detection of full-length TCRαβ is possible and achievable with sequencing depth below 0.25 million PE reads, and with a minimum read length of 50 bp. The detection rate of full length TCRαβ is at least 80% for reads with a sequencing depth >0.25 million PE reads of length at least 50 bp. Notably, the success rate in TCR reconstruction was dramatically reduced for short read length datasets (25 bp). This result can be explained by the poor alignment quality of short reads across the highly variable region of the CDR3 genes. Both methods implemented for TCR reconstruction rely on a *de novo* assembly step, which failed to reconstruct putative TCR contig. Secondly, the poor quality of alignment with short reads (25 or 50 bp) is associated with a higher number of detected genes when compared to datasets with longer reads. This increase in gene expression quantification is also associated to a diminished accuracy and increased misclassification of cell populations. Hence, short read datasets are more prone to technical noise. Future experimental designs should consider the quality of the reads as an important feature to obtain reliable results. Analyses of simulated and real scRNA-seq datasets showed that current methods, such as Smart-seq2 are consistent with a capture efficiency between 3–10% of the total mRNA available[8]. Indeed, the effect of low sequencing depth in the quality of gene expression quantification and TCR reconstruction is likely to be associated to the poor library capture efficiency of mRNA from single cells (<10%)[8], hence it is conceivable that downstream analyses are not affected by large increase in sequencing depth. This study has focussed on T cells, however the results provided here are likely to be valid for other cell types. Notably, in this study we analysed resting memory T cells[20], which are likely to be affected by limited mRNA available compared to other more active cell types such as effector cells. Previous studies have shown that the technical noise strongly depends on

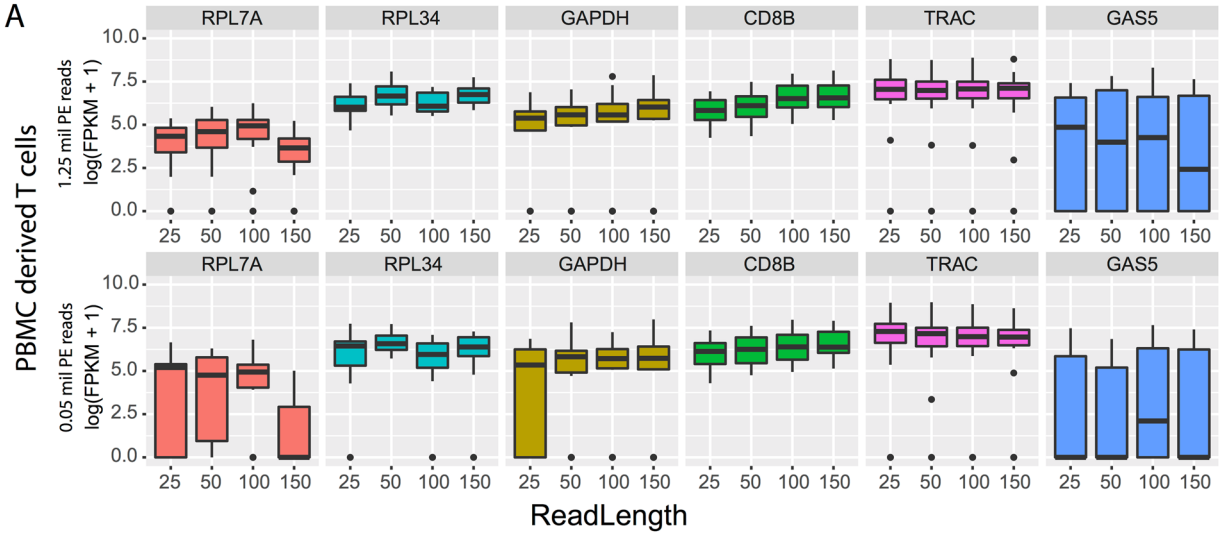

Figure 6. Gene expression profiles of selected genes identified from dataset 1, human HCV- specific CD8+ T cells (Table 1).

the mRNA content of the cell, which is a limiting factor in the detection of biological variation[32,33]. It is therefore likely that these results may be relevant for other cell types with a resting state, such as naïve cells, quiescent cells and dormant cancer cells.

Here we have showed that clustering analysis with scRNA-seq data characterised by 0.25 million PE reads and length >50 bp can accurately distinguish three rare and relatively homogenous cell subsets of antigen specific T cells, thus suggesting that despite technical limitations current scRNA-seq data can successfully be applied to identify differences between rare T cell subsets, such as effector and memory subtypes during an immune response against a viral infection. On the other hand, this may not be sufficient when more homogenous populations are involved, such as central memory and effector memory T cells from the same antigen specific repertoire. Deep sequencing of single cell library may still be required to improve detection of low abundant transcripts. Indeed, an important issue for scRNA-seq data is the very large amount of genes with zero expression[6]. This observation results from real zero expression genes that a single cell may have at the time of RNA extraction, as well as dropout events, which are due to inefficient mRNA capture and library processing.

Full length TCRαβ can be accurately estimated and linked to the gene expression profile of the same cell. The analysis also showed multiple instances of single cells with at least two α and two β sequences detected. These findings are likely explained by the presence of multiple cells per well being sequenced, and the higher detection rate of double chains in the Th17 dataset (datasets 3 and 4) is likely due to the larger sample size compared to the other studies. The high success rate obtained with both available software programs further support the high quality of the scRNA-seq data, which significantly improve the quality of TCR reconstruction with more classical approaches such as bulk sequences and Sanger sequencing. Along with TCRαβ full-length data, the entire transcriptome can be interrogated to identify specific gene profiles associated to T cell subsets, along with the relationship with the TCRαβ clonotypes. Single cell approaches are therefore likely to increase further the accurate identification of novel markers, which could be utilised for detecting novel subpopulations of cells, for instance using flow cytometry. Another improvement is to introduce bar coding of the cell, with approaches such as MARS-seq[34]. These approaches however still lacks the incorporation of full-length transcriptome sequencing, hence affecting the accurate detection of full length TCRαβ. In conclusion, this study showed that future analyses should consider the quality of sequencing output to ensure reliable and accurate single cell transcriptomic profiling.

This study focussed only on T cells and scRNA-seq protocols available that allow full-length sequencing of transcripts. Barcoded methods, such as those with microfluidics based technologies and unique molecular identifiers (UMIs) are restricted to sequencing of a small fragment of the mRNA transcript, thus limiting the detection of full TCR sequences. At the time of this analysis there were no studies that performed UMI-based protocols for T cell subsets. A recent study has utilised a cell line to benchmark four UMI-based protocols (CEL-seq. 2, Drop-seq, MARS-seq, SCRB- seq) against Smart-seq. 1, and Smart-seq. 2[35]. The analysis tested the sensitivity of UMI and Smart-seq methods on the same samples by assessing the number of genes detected as a function of sequencing depth. This analysis showed that Smart-seq. 2 had the highest sensitivity of detection, and 0.625 Million PE reads were sufficient to obtain an optimum number of genes, which is in line with our simulations. UMI-methods however quantified mRNA levels with less amplification noise due to the use of UMI. A comprehensive analysis of 15 experimental protocols utilizing ERCC spike-ins on the effect of sequencing depth and on the limit of detection (lower molecular-detection limit, for a given sequencing depth), confirmed high sensitivity of Smart-seq protocols and comparable but variable accuracy when compared to UMI-based methods[36]. Consistent with our findings this analysis also showed that sensitivity is highly dependent on sequencing depth, experimentally confirming our value for the optimal sequencing depth (>0.625 million PE reads, i.e. 1.25 million

reads in total) whereby the increase in read depth from 1 million reads to 4.5 million reads per sample results in marginally increased sensitivity. Finally, UMI-based methods can be modified to obtained VDJ region of the TCR at the single cell level. For instance 10x Genomics has recently released a new protocol to sequence VDJ sequences from bar coded T cells (https://www.10xgenomics.com/vdj/). This approach however does not provide simultaneous analysis of gene expression and TCR repertoire from the same cell.

In conclusion, our results based on T cell data are consistent with benchmark studies on other cells, showing that the full-length scRNA-seq methods provide good accuracy, high sensitivity and larger potential for applications in T cell biology. Future development of barcoded technologies that allow full-length transcriptomic sequencing will allow the application of these technologies to a larger number of cells, enabling comprehensive study of the role of T cells in experimental models and human disease states.

Material and Methods

ScRNA-seq data. Raw data were downloaded from NCBI's Gene Expression Omnibus and ArrayExpress (Table 1).

Generation of simulated datasets. Simulated data were obtained by generating subset of reads from Dataset 1 in Table 1 by randomly reducing read length and sequencing depth using an in-house python scripts. Original dataset consisted of 54 scRNA-seq data all with read length of 145 bp and sequencing depth of about 8.4 million PE reads. Sixteen combinations of read length and sequencing depth were considered: read length of 25, 50, 100 and 150 bp; and sequencing depth of 0.1, 0.5, 1.25 and 2.5 million. A new set of paired fastq file for each combination was then generated. For each dataset, reduced depth was obtained by randomly subsampling the original set of PE reads while the shorter read length was obtained by randomly cropping original PE reads.

Gene expression quantification. PE reads were analysed for quality control using FastQC, and reads were trimmed using Trimmomatic[37]. For Trimmomatic we used the following parameters: Nextera adapters, leading = 3, trailing = 3, window length = 4, window quality = 15, average quality = 20, minimum length was chosen according to the read length of the dataset. Alignment of PE reads was performed with TopHat2. For the alignment, the default option was used, (https://ccb.jhu.edu/software/tophat/manual.shtml).

Gene expression was estimated with the pipeline available in Cufflinks 2.2.1, utilising CuffQuant with parameter–max-frag-multihits equal to 1, which allows maximum one alignment per fragment. Gene expression quantification (in FPKM) was normalised with CuffNorm using default parameters. Resulting FPKM values were manipulated in R using the package Monocle (version 2)[38], using the detectGenes function to count and filter genes by FPKM value. Downstream analysis, which included Pearson's correlation analysis, number of genes expressed, and gene expression analysis by gene categories, was performed with an in-house R script. Transcription factors and housekeeping genes have been selected from available list in the literature[39,40].

Dropout rate and clustering analysis. Dropout analysis, principal coordinates analysis and clustering were performed using CIDR[30], which requires raw read counts as input data. The tool featureCounts was used to obtain the read counts, with the–*primary* option to allow only primary alignments. The *dropoutCandidates* Boolean matrix output by the *determinDropoutoutCandidates* method of CIDR is used to calculate the figures in Table 3 and Fig. 5 – a gene is considered 'expressed' in a sample if the corresponding entry in the *dropoutCandidates* matrix has a value of *FALSE*.

For clustering, the CIDR parameters *nCluster* and *wThreshold* were set to be 3 and 6 respectively, while the other CIDR parameters were left as defaults. Within each cluster, the first two CIDR principal coordinates were used to calculate Euclidean distances between all pairs of samples, the squares of which sum to the within-class sum of squares.

Misclassification rate was used to evaluate the accuracy of clustering, which is defined as the number of misclassified cells divided by the total number of cells. To define misclassified cells, each CIDR cluster is associated with the ground truth cluster, which gives the biggest intersection, and those cells that are not in the intersection are counted as misclassified cells.

Reconstruction of TCRαβ. TCRαβ of the downloaded dataset were reconstructed using VDJPuzzle[41]. A second method was used to validate the VDJPuzzle result using the program TraCeR[21] (see Supplementary Text). VDJPuzzle algorithm is briefly outlined in the following steps (see[41] for more details): for each single cell and for each chain: i) it aligns the reads to the full reference genome; ii) it extracts the reads that align to one known VDJ gene or constant region and assemble these reads using a de novo assembly algorithm (Trinity[42]); iii) all reconstructed sequences with a match to the IMGT database are collected in a preliminary TCR sequence identification; iv) the original reads are re-aligned (using bowtie 2[43]) against the putative TCR sequences, to include additional reads that could have been lost in the first alignment;.v) resulting aligned reads aligning to the preliminary repertoire are assembled with Trinity and interrogated to IMGT with MigMap, a smart wrapper for IgBlast (https://github.com/mikessh/migmap).

Successful reconstruction of a TCR from a single cell was defined as at least one complete and in-frame TCR sequence (α, β, or both) identified in the IMGT database. The exact procedure was performed as previously reported[20].

The fit of the proportion of cells with successful TCRαβ reconstruction as a function of read length and sequencing depth was performed using a two-dimensional sigmoidal function implemented in the scipy package in python (the "curve_fit")

$$SR_{\alpha,\beta} = \frac{q1}{1 + e^{-a(x-x_0)}} \times \frac{q2}{1 + e^{-b(y-y_0)}} \quad (1)$$

Where x represents the read length and y represents the sequencing depth. The obtained fitting values are $q1 = 0.94$, $q2 = 0.95$, $a = 0.52$, $b = 98.67$, $x_0 = 24.2$, $y_0 = 8.85$.

Availability of data and materials. The scRNA-seq data analysed are freely available and accession numbers are provided in Table 1.

References

1. Proserpio, V. & Lonnberg, T. Single-cell technologies are revolutionizing the approach to rare cells. *Immunol Cell Biol* **94**, 225–229, https://doi.org/10.1038/icb.2015.106 (2016).
2. Vieira Braga, F. A., Teichmann, S. A. & Chen, X. Genetics and immunity in the era of single-cell genomics. *Hum Mol Genet* **25**, R141–R148, https://doi.org/10.1093/hmg/ddw192 (2016).
3. Grun, D. & van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* **163**, 799–810, https://doi.org/10.1016/j.cell.2015.10.039 (2015).
4. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610–620, https://doi.org/10.1016/j.molcel.2015.04.005 (2015).
5. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**, 1145–1160, https://doi.org/10.1038/nbt.3711 (2016).
6. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* **17**, 63, https://doi.org/10.1186/s13059-016-0927-y (2016).
7. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**, https://doi.org/10.12688/f1000research.7223.1 (2016).
8. Marinov, G. K. *et al*. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* **24**, 496–510, https://doi.org/10.1101/gr.161034.113 (2014).
9. Poirion, O. B., Zhu, X., Ching, T. & Garmire, L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. *Front Genet* **7**, 163, https://doi.org/10.3389/fgene.2016.00163 (2016).
10. Conesa, A. *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13, https://doi.org/10.1186/s13059-016-0881-8 (2016).
11. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145, https://doi.org/10.1038/nrg3833 (2015).
12. Chhangawala, S., Rudy, G., Mason, C. E. & Rosenfeld, J. A. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol* **16**, 131, https://doi.org/10.1186/s13059-015-0697-y (2015).
13. Wu, A. R. *et al*. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* **11**, 41–46, https://doi.org/10.1038/nmeth.2694 (2014).
14. Pollen, A. A. *et al*. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* **32**, 1053–1058, https://doi.org/10.1038/nbt.2967 (2014).
15. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* **16**, 241, https://doi.org/10.1186/s13059-015-0805-z (2015).
16. Schober, K. & Busch, D. H. A synergistic combination: using RNAseq to decipher both T-cell receptor sequence and transcriptional profile of individual T cells. *Immunol Cell Biol* **94**, 529–530, https://doi.org/10.1038/icb.2016.32 (2016).
17. Nikolich-Zugich, J., Slifka, M. K. & Messaoudi, I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* **4**, 123–132, https://doi.org/10.1038/nri1292 (2004).
18. Tumeh, P. C. *et al*. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571, https://doi.org/10.1038/nature13954 (2014).
19. Ruggiero, E. *et al*. High-resolution analysis of the human T-cell receptor repertoire. *Nat Commun* **6**, 8081, https://doi.org/10.1038/ncomms9081 (2015).
20. Eltahla, A. A. *et al*. Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. *Immunology and Cell Biology*, https://doi.org/10.1038/icb.2016.16 (2016).
21. Stubbington, M. J. T. *et al*. T cell fate and clonality inference from single-cell transcriptomes. *Nature Methods* **13**, 329–332, https://doi.org/10.1038/nmeth.3800 (2016).
22. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).
23. Brazma, A. *et al*. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **31**, 68–71 (2003).
24. Gaublomme, J. T. *et al*. Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell* **163**, 1400–1412, https://doi.org/10.1016/j.cell.2015.11.009 (2015).
25. Kimmerling, R. J. *et al*. A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nat Commun* **7**, 10220, https://doi.org/10.1038/ncomms10220 (2016).
26. Mahata, B. *et al*. Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep* **7**, 1130–1142, https://doi.org/10.1016/j.celrep.2014.04.011 (2014).
27. Picelli, S. *et al*. Full-length RNA-seq from single cells using Smart-seq. 2. *Nat Protoc* **9**, 171–181, https://doi.org/10.1038/nprot.2014.006 (2014).
28. Ozawa, T., Tajiri, K., Kishi, H. & Muraguchi, A. Comprehensive analysis of the functional TCR repertoire at the single-cell level. *Biochem Biophys Res Commun* **367**, 820–825, https://doi.org/10.1016/j.bbrc.2008.01.011 (2008).
29. Lefranc, M. P. *et al*. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* **43**, D413–422, https://doi.org/10.1093/nar/gku1056 (2015).
30. Lin, P., Troup, M. & Ho, J. W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* **18**, 59, https://doi.org/10.1186/s13059-017-1188-0 (2017).
31. Mourtada-Maarabouni, M., Hasan, A. M., Farzaneh, F. & Williams, G. T. Inhibition of human T-cell proliferation by mammalian target of rapamycin (mTOR) antagonists requires noncoding RNA growth-arrest-specific transcript 5 (GAS5). *Mol Pharmacol* **78**, 19–28, https://doi.org/10.1124/mol.110.064055 (2010).
32. Brennecke, P. *et al*. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**, 1093–1095, https://doi.org/10.1038/nmeth.2645 (2013).
33. Ramskold, D. *et al*. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777–782, https://doi.org/10.1038/nbt.2282 (2012).

34. Jaitin, D. A. *et al*. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779, https://doi.org/10.1126/science.1247651 (2014).
35. Ziegenhain, C. *et al*. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631–643 e634, https://doi.org/10.1016/j.molcel.2017.01.023 (2017).
36. Svensson, V. *et al*. Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* **14**, 381–387, https://doi.org/10.1038/nmeth.4220 (2017).
37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
38. Trapnell, C. *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386, https://doi.org/10.1038/nbt.2859 (2014).
39. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in Genetics* **29**, 569–574, https://doi.org/10.1016/j.tig.2013.05.010 (2013).
40. Zhang, H. M. *et al*. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Research* **40**, D144–D149, https://doi.org/10.1093/nar/gkr965 (2011).
41. Eltahla, A. A. *et al*. Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. *Immunol Cell Biol* **94**, 604–611, https://doi.org/10.1038/icb.2016.16 (2016).
42. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
43. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, https://doi.org/10.1186/gb-2009-10-3-r25 (2009).

Acknowledgements

We acknowledge the National Health and Medical Research Council of Australia (NHMRC) and Australian Centre for HIV and HCV Research (ACH2) for funding. SR is supported by the University International Postgraduate Award UNSW Australia. ARL is supported by an NHMRC Practitioner Fellowship (No. 1043067), AE is supported by an NHMRC Early Career Fellowship (No. 1130128). JWKH is supported by an NHMRC/National Heart Foundation Career Development Fellowship (No. 1105271), RAB, VV and FL are supported by NHMRC Career Development Fellowships (No. 1060443, 1067590, 1128416).

Author Contributions

S.R. analyzed and interpreted the scRNA-seq data. F.L. led the study and designed the experiments. S.R., P.L., J.H., and F.L. contributed softwares and performed bioinformatics analyses the data. S.R., F.L., P.L. and J.H. contributed to the statistical analysis, data collection, bioinformatics analyses. A.R.L., R.B., and A.E. contributed to the design of the study and in the analysis of the data. F.L., S.R., wrote the manuscript. V.V., J.H., and A.E. were major contributors to the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-12989-x.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.