Sierra: Discovery of differential transcript usage from polyA-captured single-cell RNA-seq data

Ralph Patrick[1,2,#], David Humphreys[1,#], Alicia Oshlack[3,4], Joshua W.K. Ho[1,2,5], Richard P. Harvey[1,2,6,*], Kitty K. Lo[7,*]

1 Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010 Australia

2 St. Vincent's Clinical School, UNSW Sydney, Kensington, Australia

3 Murdoch Children's Research Institute, Parkville, Victoria 3052 Australia

4 Peter MacCallum Cancer Centre, Research Division, 305 Grattan Street, Melbourne Vic, 3000, Australia

5 School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

6 School of Biotechnology and Biomolecular Science, UNSW Sydney, Kensington, Australia

7 School of Mathematics and Statistics, Faculty of Science, The University of Sydney, Camperdown, 2006 Australia

# These authors have contributed equally to this manuscript

* Co-corresponding authors: kitty.lo@sydney.edu.au; r.harvey@victorchang.edu.au

Abstract

Background: High-throughput single-cell RNA-seq (scRNA-seq) is a powerful technology for studying gene expression variability in single cells; however, standard analysis approaches only consider the overall expression of each gene, masking additional heterogeneity that could exist through cell type-specific expression of alternative mRNA transcripts.

Results: Here we show that differential transcript usage (DTU) can be readily detected in data-sets generated from commonly used polyA-captured nanodroplet scRNA-seq technology. Our computational pipeline, Sierra, detects and quantifies polyadenylation sites in scRNA-seq data-sets, which are utilised to evaluate DTU between single-cell populations. We validate our approach by comparing cardiac cell populations derived from scRNA-seq to bulk RNA-seq of matched populations obtained through fluorescent activated cell sorting (FACS), demonstrating that we detect a significant overlap in cell-type DTU between the congruous data-sets. We further illustrate the utility of our method by detecting alternative transcript usage in human peripheral blood mononuclear cells (PBMCs), 3'UTR shortening in activated and proliferating cardiac fibroblasts from injured mouse hearts and finally by building an initial atlas of cell type-specific transcript usage across 12 mouse tissues.

Conclusions: We anticipate that Sierra will enable new avenues of transcriptional complexity and regulation to be explored in single-cell transcriptomic experiments. Sierra is available at `https://github.com/VCCRI/Sierra`.

Keywords

scRNA-seq, polyadenylation, isoforms, differential transcript usage

Background

Regulation of cellular transcriptional activity includes changes in the total expression level of genes as well as alternative usage of gene architecture. Alternative, or differential, usage can be thought of as changes in the relative use of gene sub-components leading to expression of alternative mRNA transcript isoforms, which we refer to here as differential transcript usage (DTU). Forms of DTU can include alternative splicing (AS), such as through exon skipping or retained introns, alternative 5' promoter usage or 3 end use, or changes in 3'UTR length through the use of alternative polyadenylation (APA) sites. RNA sequencing (RNA-seq) studies have revealed high levels of alternative transcript usage between tissues, with 95% of multi-exon genes estimated to undergo AS among human tissues [1]. Similarly, APA is widespread among the genome, and is estimated to occur in most genes [2]. While the extensive use of AS and APA among tissues is documented, DTU between the diverse sub-tissue cell-types revealed in recent years by single-cell RNA-seq (scRNA-seq) is relatively unexplored.

There is a strong impetus to develop strategies for studying DTU at the single-cell level, given that studies to date have found the different forms of DTU to play significant functional roles across many biological contexts. Functional consequences of DTU can include changes to mRNA stability, localisation, protein translation and nuclear export as determined by AS [3] or APA in 3'UTRs [4]. DTU plays roles in differentiation and organ development through AS [5, 6], including intron retention (IR) [7, 8], a form of AS that appears widespread in mammals [9, 10]. DTU is also relevant in disease contexts. As examples, AS is linked to disease through miss-splicing caused by genomic variants [11] while 3'UTR shortening has been reported as widespread in cancer cells [12] and has been found to repress tumor-suppressor genes [13]. Alternative usage of introns is also linked to cancer both through intronic polyadenylation [14] and IR [15, 16]. Overall, studies to date reveal a pattern of

widespread alternative mRNA transcript usage of different forms and functional consequences across tissues, development and disease contexts.

The advance of scRNA-seq technologies opens up new avenues for deeper exploration of DTU at the level of single cells [17]; however, there are technical features of the high-throughput nanodroplet technologies like the 10x Genomics Chromium platform, including low depth and limited gene coverage, that makes detecting DTU a non-trivial task. Some scRNA-seq protocols such as Smart-seq provide read coverage across the gene and therefore enable the analysis of alternative isoform expression [17]. For example, analysis of neural Smart-seq scRNA-seq data has demonstrated alternative splicing events at the single-cell level in the brain [18] and methods have been specifically developed for analysing alternative isoform expression in scRNA-seq experiments that have reads spanning the transcript [19–21]. The above studies utilise increased transcript read coverage at the expense of lower-throughput profiling of cells; however, high-throughput nanodroplet technologies like 10x Chromium have become the technology of choice for many scRNA-seq experiments due to the capacity to profile thousands of single-cell transcriptomes at low cost. Currently, only gene-level expression data is routinely utilised in analysing 10x Chromium data; however, the enrichment of 3' ends in barcoded and polyA-captured scRNA-seq means that study of APA and alternative 3'-end usage can be explored, with the potential to unveil additional levels of information among cell types currently masked when only considering an aggregate of each gene. Despite polyA-captured nanodroplet scRNA-seq experiments now routine, to the best of our knowledge there is no computational pipeline described that can leverage such data-sets to identify DTU between cell types.

We present here a novel computational pipeline for unbiased identification of potential polyadenylation (polyA) sites in barcoded polyA-captured scRNA-seq experiments, and evaluation of DTU between cell populations. We demonstrate that we can identify DTU – which we define as any change in relative usage across a gene, including differential exon usage,

alternative 3'UTR usage and changes in 3'UTR length – between cell-types across different species, tissues and disease contexts. These include 10x Chromium scRNA-seq experiments on human PBMCs, murine cardiac interstitial and enriched fibroblast cells from injured and uninjured hearts and a multi-tissue atlas from the Tabula Muris. We validate our approach by comparing DTU calls from cardiac scRNA-seq data to bulk ribo$^-$ RNA-seq of matched cell populations derived from FACS and find a significant overlap in DTU genes from both the scRNA-seq and bulk RNA-seq experiments. We demonstrate that we can detect multiple types of DTU among cardiac interstitial cells types, including alternative 3' end use, APA and even alternative 5' start sites that are validated from the matched bulk RNA-seq samples. We apply our method to detect 3'UTR shortening in proliferating cardiac fibroblasts from injured mouse hearts and show not only that we can detect 3'UTR shortening in proliferating cells as observed previously, but can detect shortening in related activated populations that are not proliferating – a granularity that is not possible to observe in bulk RNA-seq studies. Finally, we apply our approach to 12 tissues from the Tabula Muris, presenting an initial atlas of cell type-specific DTU across mouse tissues. Our analysis pipeline in implemented as an R package, named Sierra, available at `https://github.com/VCCRI/Sierra`.

Results

The Sierra R package contains a start-to-end pipeline for identification of polyadenylated gene regions in scRNA-seq data, differential usage analysis and visualization (Methods). Briefly, the Sierra pipeline starts with a BAM file, such as produced by the 10x Genomics CellRanger software, and the reference GTF file used for mapping (Figure 1). Based on the observation that aligned reads from 10x Chromium scRNA-seq experiments fall into Gaussian-like distributions, peak calling is run to identify sub-regions, or 'gene peaks', within genes that correspond to potential polyA sites. The peak coordinates are utilised to construct

a new reference file of genomic regions, enabling a unique molecular identifier (UMI) matrix of peak coordinates to be built for a supplied list of cell barcodes. Each of the gene peaks is annotated according to the genomic feature it falls on (3'UTR, exon, intron or 5'UTR) and proximity to sequence features including A-rich regions and the canonical polyA motif – enabling discrimination of likely cases of internal priming to A-rich sequences from true polyA sites.

For down-stream analysis, Sierra can utilise either the Bioconductor [22] SingleCellExperiment class [23] or the widely-used Seurat R package [24] to create objects containing the peak counts and annotation information required for DTU testing, in addition to predetermined cell identities/clusters and dimensionality reduction coordinates (t-SNE or UMAP) for visualisation purposes. For Seurat users, Sierra can directly import cluster identities and t-SNE/UMAP coordinates over from a pre-existing Seurat object, allowing for straight-forward analysis of pre-defined cell types.

To test for DTU between cell groups we utilise the differential exon usage method developed for bulk RNA-seq, DEXSeq [25], but applied to peak coordinates, whereby a gene will be called as DTU if it shows significant change in the relative usage of peaks. Cell populations to be tested are first aggregated into a small number of pseudo-bulk profiles, which define replicates, enabling computational efficiency of DTU testing. Sierra also contains several functions for visualisation of peak expression from DTU genes (Figure 1). These include the plotting of relative expression between two or more gene peaks, where the expression of a set of gene peaks is transformed according to their relative usage within each cell population (Figure 1; Methods). Relative expression can then be plotted on, for example, t-SNE or UMAP coordinates as with gene expression data. Finally, in order to aid in interpretation of DTU between cell types, we provide functionality for gene-level plotting of read coverage across defined single-cell populations. Single-cell populations (e.g. clusters) are first extracted from an aggregate BAM file according to cell barcode, generating population-level BAMs.

Read coverage for DTU genes can then be visualised at a single cell-population level (Figure 1). The coverage plots, which also show a gene model, allow for deeper interpretation of the nature of the DTU detected.

Features of Sierra data

We applied Sierra to 19 data-sets representing four main systems or studies: two human PBMC data-sets, (a 7k cell and 4k cell) from 10x Genomics, total non-cardiomyocyte cells (total interstitial population [TIP]) from uninjured (sham) hearts or hearts at 3- or 7-days following myocardial infarction (MI) surgery [26], enriched cardiac fibroblasts (Pdgfra-GFP^{+}) also from MI and sham mouse hearts [26] and the Tabula Muris [27], with twelve tissues from Tabula Muris analysed in this study (Table S1).

Sierra typically detected 30,000 – 50,000 peaks covering 10,000 - 15,000 genes across the data-sets tested (Table S1). Most genes had a small number number of peaks, typically with a median of 2 peaks per gene (Table S1). In the PBMC 7k data-set for example, over 4000 genes were called with 1 peak, followed by 2000 and 1000 genes with 2 and 3 peaks, respectively, while a minority of genes were called with 20 - 100+ peaks (Figure 2A). Considering the genomic features associated with these peaks, we found that for genes with 1-2 peaks, the majority of these fell in 3'UTRs (Figure 2B), while genes with larger numbers of peaks tended to show more intronic peaks. For genes with over 20 peaks, on average 75% of these were intronic (Figure 2B). These metrics were consistent in both the PBMC 4k and TIP data-sets (Figure S1A-D). We next stratified the peaks according to genomic feature type, and examined how increasing stringency of cellular detection rates (i.e. only considering peaks expressed in some x% of cells) affected the feature-type composition of peaks. With no filtering, we found that the largest number of called peaks were intronic, followed by 3'UTRs ($\geq$ 0% detection rate; Figure 2C and Figure S1E,F). Progressive filtering

of peaks according to cell detection rates showed that intronic peaks tended to be detected in a smaller number of cells (Figure 2C and Figure S1E,F). The substantial presence of intronic peaks is in agreement with previous observations made about RNA molecules containing intronic sequence in 10x Genomics Chromium data [28], and likely corresponds to pre-spliced mRNA.

We compared the expression characteristics of the peaks with gene-level expression data from CellRanger (Figure S2A-D) and found a strong correlation between gene expression and expression of peaks in 3'UTRs as expected, with weaker correlations in intronic peaks for both 7k PBMCs (Figure S2A) and the cardiac TIP data-set (Figure S2C). We also compared gene and peak expression using mean expression/dispersion plots. We noticed a wider range of dispersion values in peaks compared to genes for both data-sets, though intronic peaks partially explain this, with a higher dispersion range among more lowly expressed genes (Figure S2B,D). Finally, we annotated each peak according to whether it was proximal to an A-rich region or the canonical polyA motif (Table S1). We found 3'UTR peaks had the highest percentage of proximity to the polyA motif (on average 47%), while 5'UTRs had the lowest (average of 5%). Intronic and exonic peaks also had low levels of polyA motif proximity (average of 9% and 10%, respectively). Conversely, 3'UTR peaks had the lowest proximity to A-rich regions (average of 10%), while intronic peaks had the highest (50%), with exonic and 5'UTR peaks showing an average of 28% and 18%, respectively (Table S1).

Differential transcript usage among human PBMCs

We next considered the extent to which we could call DTU between human PBMC cell populations as defined by gene-level clustering. Seurat clustering of the 7k PBMCs yielded 16 clusters including several populations of monocytes (Mo), $CD4^+$ T-cells, $CD8^+$ T-cells, B-cells, two natural killer cell (NKC) populations and several minor populations. As DTU

testing only applies to genes with multiple peaks, we evaluated how many genes had multiple 3'UTR/exonic peaks expressed across the PBMC 7k cell populations at increasing cell detection rates. When filtering for peaks with expression in at least 10% of cells, there were over 3000 genes with multiple peaks. This dropped to just over 1000 genes when requiring peaks to be detected in 25% of cells (Figure 2D). For the below analyses we used a 10% detection-rate cutoff. We applied DEXSeq [25] to call peaks exhibiting differential usage (DU) between cell types after aggregating cells within cell types into a small number of pseudo-bulk profiles to create pseudo-replicates (Methods). We restricted our testing to peaks falling on 3'UTRs or exons.

We performed DU analyses both between clusters and aggregated groups of cells (e.g. all $CD4^+$ T-cells vs monocytes). We readily detected significant DTU genes using our pseudo-bulk replicates with DEXSeq between cell types ($P_{adj} < 0.01$; $LFC > 0.5$). We detected the largest numbers of DTU genes when comparing lineages; for example, Mo against lymphoid populations including B-cells (BC), T-cells (TC) or NKCs (see Table 1 for representative examples). When comparing populations Mo1 and $CD4^+$ TC1, 825 distinct peaks were called by DEXSeq as DU, representing 492 DTU genes (i.e. a gene containing at least one DU peak is classified as a DTU gene). In contrast, DEXSeq detected 83 DU peaks corresponding to 63 DTU genes when comparing populations NKC1 to $CD4^+$ TC1 (Table 1). Of the DU peaks, on average 20% were tagged as near an A-rich region, and potentially due to internal priming, while an average of 32% flanked the canonical polyA motif. Among the DTU genes, there were known examples of alternatively spliced genes in immune cells, including *IKZF1* (Ikaros) and *PTPRC* (CD45) [29]. Although *PTPRC* gene expression is ubiquitous among immune cell types, we found we could distinguish clear patterns of alternative peak usage in monocytes compared to other cell types (Figure 2E) according to t-SNE visualisations of relative peak expression, demonstrating that we are able to detect cell type gene expression activity masked when only considering an aggregate of the reads across a gene.

We compared DTU genes to differentially expressed (DE) genes from the same population comparisons using the Seurat *FindMarkers* program with MAST testing [30]. For the PBMC 7k data-set, we found on average that 50% of the DTU genes overlapped with DE genes. We also found a strong positive correlation between the number of DTU and DE genes (Spearman's correlation test; $\rho = 0.79$; $P = 3.2e-51$). Finally, we evaluated the reproducibility of calling DTU genes by comparing the PBMC 7k DTU genes with genes identified by the same cell-type comparisons in the 4k PBMC data-set. For example, comparing $CD14^+$ Mo to $CD4^+$ TCs, there were 653 DTU genes in the PBMC 7k data-set and 247 in the lower-depth PBMC 4k data-set, with 202 overlapping (Figure 2F), representing 82% of the DTU genes from the PBMC 4k analysis. For comparison, we performed the same analysis with DE genes and found a similar level of overlap (Figure 2G). Across all comparisons, we found that on average 60% of DTU genes were matched between PBMC 4k and PBMC 7k data-sets, while for DE genes an average of 80% were matched (Table S2). Thus, the majority of DTU genes can be found in a replicate experiment, and at levels not far below DE testing, indicating that our method of detecting DTU genes is reproducible.

Patterns of differential transcript usage in the mouse heart

We validated Sierra DTU calls by comparing single-cell populations from the heart [26] to a bulk RNA-seq data-set of matched cardiac populations isolated with FACS [31]. Importantly, the bulk RNA-seq experiment used a $ribo^-$ protocol instead of $polyA^+$, which makes this a unique validation resource that does not have the 3' bias of the scRNA-seq. The cardiac scRNA-seq experiment was performed on the total non-cardiomyocyte compartment of the heart (total interstitial population [TIP]) and reported several interstitial cells types including fibroblasts, endothelial cells (ECs) and numerous sub-populations of leukocytes [26] (Figure 3A). In addition to cardiomyocytes, the bulk RNA-seq data-set contains sorted ECs

(CD31$^+$), fibroblasts (CD90$^+$) and leukocytes (CD45$^+$). These populations were isolated from adult mouse hearts at 3 days post-sham or myocardial infarction (MI) surgery; the TIP scRNA-seq experiment contains a sham, and two MI time-points: MI-day 3 and MI-day 7. Both data-sets therefore contain comparable populations and conditions (sham and MI-day 3). To compare with the scRNA-seq we first applied DEXSeq to calculate DTU genes between an aggregate of the scRNA-seq fibroblast cells, ECs and leukocytes from the sham hearts (i.e. pairwise comparisons between these cell lineages) and the sham populations again MI leukocytes (Table S3). We did not include MI fibroblasts or ECs as these populations are diluted in the scRNA-seq due to an overwhelming influx of monocytes and macrophages at MI-day 3 [26].

We found clear examples of DTU masked when considering an aggregate of the gene. As examples, top DTU genes from fibroblast to EC and fibroblast to MI leukocyte comparisons were *Cxcl12* and *Igf1*, respectively. *Cxcl12* was observed to be expressed in fibroblasts, ECS and mural cells (Figure 3B), while *Igf1* was expressed in fibroblasts and macrophage populations (Figure 3C). When we compared relative expression of DU peaks, we found we could distinguish clear patterns of specificity between fibroblasts and ECs for *Cxcl12* (Figure 3D) and between fibroblasts and macrophages for *Igf1* (Figure 3E). We plotted read coverage across *Cxcl12* for the fibroblast and EC populations and compared them to the bulk RNA-seq of sorted fibroblasts and ECs. This showed that the peak upregulated in fibroblasts corresponded to a 3'UTR of a short transcript isoform of *Cxcl12*, while the top EC peak corresponded to a longer isoform (Figure 3F). The difference in transcript isoform expression between fibroblasts and ECs was also observed in the bulk RNA-seq, confirming our observations made in the single-cell data (Figure 3F). *Cxcl12* splice variants have been observed in the heart previously [32]. In the case of *Igf1*, the two DU peaks corresponded to proximal and distal sites on the same 3'UTR, relative to the terminating exon, and are annotated in current RefSeq gene models. The distal peak (peak 2) was specifically expressed

in fibroblasts, and from the bulk RNA-seq, we could observe that fibroblasts preferentially expressed a longer 3'UTR in *Igf1* (Figure 3G), demonstrating that we can detect APA corresponding to changes in 3'UTR length. We could observe other examples of APA from DTU genes between fibroblasts and leukocytes, including *Tfpi* and *Tm9sf3* (Figure S3A,B). We also detected examples of alternative 5' start sites in *Lsp1* and *Plek* (Figure S3C,D), demonstrating that Sierra detects DTU corresponding to a variety of alternative transcript expression events.

We next asked how many DTU genes detected from the single-cell comparisons could also be detected in the bulk RNA-seq. We used the peak coordinates mapped to 3'UTRs and exons as a reference to generate counts from the bulk RNA-seq and again applied DEXSeq to determine DTU between the same cell-types and conditions as with the scRNA-seq data. For all comparisons we found that there was an overlap between the scRNA-seq and bulk RNA-seq DTU genes greater than expected by chance (Figure 3H; Fisher's exact test; $P < 0.05$), after using the set of genes with multiple 3'UTR/exon peaks expressed in the relevant scRNA-seq populations as a random-expectation background. The comparison with the largest and most significant overlap was between sham fibroblasts and MI leukocytes, with 343 out of 531 comparable DTU genes (65%) from the scRNA-seq analysis also observed as DU in the bulk RNA-seq (Figure 3H,I). We next compared the fold-change direction of the peaks called as DU in both the scRNA-seq and bulk RNA-seq experiments. For all comparisons we found a significant positive correlation in fold-change (Spearman's correlation test; $P < 0.05$). The strongest correlation was found in the fibroblast vs EC comparison (Figure 3J; $r = 0.82$). We also considered whether filtering out peaks annotated as A-rich prior to the DTU analysis would improve the correlation with the bulk RNA-seq. We recalculated DU peaks from the single-cell populations, first filtering out peaks proximal to A-rich regions. For 5/6 of the comparisons there was no major change in the metrics used for the comparisons; however, for the sham EC vs sham leukocyte comparison we noticed the overlap increased

from 51% (with Fisher's exact test $P = 0.004$) to 56% ($P = 0.001$) and the Spearman correlation coefficient increased from 0.31 (with Spearman's correlation test $P = 0.005$) to 0.4 ($P = 9e - 04$). Together, these results show that Sierra can detect multiple types of alternative mRNA isoform usage with a significant number of these corroborated by an independent bulk RNA-seq experiment.

3'UTR shortening in activated and proliferating cardiac fibroblasts

Proliferating cells have previously been observed to have shortened 3'UTRs [33, 34]. We sought to determine whether we could apply Sierra to infer 3'UTR shortening at the single-cell level. In the heart, fibroblasts become activated and proliferate following MI, with the peak of proliferation occurring within days 2 to 4 post-MI [35]. scRNA-seq of enriched (Pdgfra-GFP^{+}) murine cardiac fibroblasts at day 3 post-sham or MI has revealed several sub-types of cardiac fibroblasts [26]. In the uninjured heart, predominant fibroblast populations can be distinguished on the basis of the expression of *Ly6a* (Sca1) – referred to as Fibroblast Sca1-low (F-SL) and Sca1-high (F-SH) (Figure 4A) [26, 36]. Following MI, there is the expansion of a pool of activated fibroblasts (F-Act) leading to a population of actively cycling fibroblasts (F-Cyc) (Figure 4B). In between F-Act and F-Cyc in pseudo-time is an intermediary activated population, F-CI (cycling intermediate), that does not express markers of actively proliferating cells.

We investigated whether we could infer 3'UTR shortening in the F-Cyc population compared to the main resting populations, F-SL and F-SH. We applied DEXSeq to find examples of DU peaks falling on 3'UTRs, filtering out peaks tagged as near A-rich regions prior to DU testing in order to enrich for real polyA sites (Table S4). For the DU 3'UTR peaks, we considered all expressed and non A-rich peaks that occurred on the same 3'UTR and ranked them according to their relative location to the terminating exon. Each peak was given a score

between 0 (most proximal to the terminating exon) and 1 (most distal), and we determined if there was a difference in relative location for upregulated and downregulated peaks. We reasoned that an increased number of upregulated peaks proximal to the terminating exon should imply a preference of shortened 3'UTRs, as well as a downregulation of distal peaks.

Comparing F-Cyc to F-SL/F-SH, we found 598 DU 3'UTR peaks representing 424 DTU genes ($LFC > 0.5$; $P_{adj} < 0.01$). Comparing their relative peak locations, we found a strong shift towards proximal peak upregulation in F-Cyc, and a corresponding propensity for distal peaks to be downregulated (Figure 4C; Wilcoxon Rank-sum test; $P = 1.1e-68$). We next considered whether the remaining activated populations also showed evidence of 3'UTR shortening (Table S4). Interestingly, we found that the intermediate F-CI population had a larger number of DU 3'UTR peaks (785; 545 DTU genes) than F-Cyc and also showed a strong pattern of proximal peak upregulation (Figure 4D; Wilcoxon Rank-sum test; $P = 6.5e-98$). We directly compared F-Cyc to F-CI but found only 8 3'UTR peaks showing DU, suggesting the 3'UTR shortening is occurring in the F-CI population prior to cell cycle entry. We also evaluated F-Act relative to F-SL/F-SH and found a smaller number of DU peaks (106; 94 DTU genes), which showed a reduced pattern of proximal peak upregulation, though the shift remained statistically significant (Figure 4E; Wilcoxon Rank-sum test; $P = 5.5e-08$).

We compared the genes implicated in 3'UTR shortening to the bulk RNA-seq of sorted fibroblasts from sham-day 3 and MI-day 3 hearts described above [31]. While the proliferating cells will comprise a minority within the bulk RNA-seq, the most significant examples of 3'UTR shortening should still be detectable in the bulk RNA-seq. Two of the top significant DTU genes were *Timp2* and *Cd47*, which showed clear patterns of relative higher proximal peak usage in F-Cyc compared to resting fibroblasts (Figure 4F,G). We analysed read distributions in the 3'UTRs, including isolated single-cell populations F-SH/F-SL combined (from sham hearts) and the MI populations (F-Act, F-CI and F-Cyc) and compared these to the bulk RNA-seq of sorted fibroblasts from sham and MI-day 3 hearts (Figure 4H,I). In both

the single-cell MI populations, as well as the MI bulk RNA-seq, we observed a decrease in coverage in the distal regions of the 3'UTR for both *Timp2* (Figure 4H) and *Cd47* (Figure 4I).

Clustering using peak-level expression

Our approach can clearly identify patterns of DTU when analysing single-cell populations defined through gene-level clustering. We next asked whether clustering on peak-level expression data could yield new information. We performed clustering using Seurat on the gene counts, with increasing cluster granularity through modification of the 'resolution' parameter in the Seurat *FindClusters* program, and compared the results through performing clustering on the peak counts, again selecting for peaks falling on 3'UTRs or exons (Methods).

We compared the clustering consistency between gene and peak counts using the Fowlkes and Mallow's (FM) index and the Adjusted Rand Index (ARI). We found in general that the lower clustering resolutions, with fewer clusters, yielded higher consistency as determined by FM index and ARI (Table 2). We also noticed that the effect on cluster numbers returned from peak-level clustering tended to be data-set dependent. For the TIP data-set, there were fewer clusters relative to gene-level clustering (e.g. at a resolution of 0.6, there were 21 vs 25 clusters for peak-level and gene-level clustering, respectively). For the GFP^+ data-set, there were consistently two additional clusters when using peaks, and for the PBMC 7k data-set, the numbers were the same (Table 2). Overall, using peak-level expression in place of gene-level expression does not appear to have a major impact on clustering results, particularly at lower clustering resolutions.

We visually compared the clusters returned by peak and gene-level clustering by imposing the clusters on the same t-SNE coordinates calculated on gene expression in the TIP data-set (Figure S4A-D). We found broad consensus between the clusters returned for most of the cell populations, further indicating that in general, the use of peak expression was not leading to

the identification of many different populations. Of the few differences, we noticed that at resolution 0.6, the peak-level clustering identified a small sub-population of ECs (cluster '17'; Figure S4B) that was not present in the gene-level clustering (Figure S4A). We calculated DTU between cluster '17' and the main EC population (cluster '1') and found only a small number of DU peaks (16) between these clusters; however, differential gene expression testing between clusters '17' and '1' showed that cluster '17' corresponded to a minor population of lymphatic ECs upregulating *Vwf*, *Lyve1* and *Prox1* (Figure S4E). The presence of these lymphatic ECs was observed in the original analysis by marker gene expression [26], but as a subset of cells within a larger cluster. Thus, clustering using peak expression may allow for a finer-resolution identification of some cell-types.

A tissue atlas of cell type-specific differential transcript usage

We applied Sierra to the Tabula Muris [27], a compendium of scRNA-seq data-sets across mouse tissues, to construct an initial tissue atlas of cell type-specific DTU. Applying Sierra's peak calling to each of the 12 tissues, followed by merging of peak coordinates, we obtained a total of 107,425 peaks across the whole data-set. To determine DTU across tissues, we considered tissues that contained at least two cell types with at least 100 cells, leaving 10 tissues for our analysis. We calculated DTU within each tissue by performing pairwise comparisons between each of the cell-types. Stratifying the DU peaks across tissues for each of the cell types, we found the greatest amount of DTU in mammary gland tissue, with most cell types exhibiting over 2000 examples of DU peaks. The smallest amount was observed in the heart tissue, which only contained fibroblasts and ECs after filtering for cell number, and between which there were 106 detected DU peaks (Figure 5A).

As mammary gland tissue contained the greatest number of DTU genes, we focused more on this tissue type (Figure 5B,C). Across the mammary gland cell types, we found the largest

number of DTU genes were called when comparing luminal epithelial cells to T-cells and B-cells. The smallest number of DTU genes was detected in the T-cell vs B-cell comparison. We next considered whether the Sierra DTU calls could be used to define "marker peaks" in an analogous manner to marker genes for cell types. Here we define marker peaks as peaks that are DU between the cell-type of interest and all other cell types. Applying marker DU testing to cell types from the mammary gland tissue, we defined DTU genes with high cell-type relative peak expression, including *Mrc1* in macrophages, *Ctla4* in T-cells and *Dapk1* in stromal cells (Figure 5C). We have made the processed Tabula Muris data available online.

Discussion and conclusion

We have presented Sierra, a computational pipeline for analysis and visualisation of differential transcript usage in scRNA-seq. Sierra is applicable to barcoded droplet-based scRNA-seq experimental data such as produced using the 10x Genomics Chromium system. Our method for detecting genomic regions corresponding to potential polyA sites enables a data-driven approach to detecting novel DTU events, such as alternative 3' usage and APA between single-cell populations. We fist determine the location of potential polyA sites by applying splice-aware peak calling to a scRNA-seq data-set, followed by annotation of the identified peaks and peak-coordinate UMI counting across individual cells. Finally, we use a statistical testing approach to determine genes exhibiting DTU with a novel pseudo-bulk approach to define replicates. While the software and analysis presented here is focused on the analysis of scRNA-seq data from UMI technologies, there is potential for our methodologies to be used in other enrichment based data-sets such as single-cell ChIP-seq or ATAC-seq.

Using the Sierra approach we find thousands of genes that display significant DTU in publicly available data-sets from human and mouse. Importantly, we validate our approach by comparing DTU calls made from scRNA-seq populations in the heart to DTU calls from ribo$^-$

bulk RNA-seq of matched cardiac populations obtained with FACS. The significant overlap in DTU genes for the populations tested points to the DTU detected by Sierra as representing real biology, and not technical artifacts of the scRNA-seq. Despite the significance of the overlap, there were uniquely detected DTU genes in either the bulk RNA-seq or the scRNA-seq. These differences could be due to several reasons, both technical and biological. It is to be expected that more DTU genes will be detected in bulk RNA-seq data due to the increased depth and transcript coverage; however, there could also be examples of DTU genes that can more easily be detected with 3'-end-based scRNA-seq than with bulk RNA-seq, which would indicate a unique benefit of using 3'-end-based scRNA-seq for detecting some forms of alternative isoform expression.

The flexibility of Sierra means that diverse questions can be asked about DTU. While we focus on mature transcripts in this manuscripts, the presence of intronic peaks means that questions about pre-spliced mRNA can be explored as well. The RNA Velocity approach utilises ratios of spliced and unspliced reads from scRNA-seq to estimate rates and direction of cell differentiation [28]. With Sierra, it should be possible to identify specific genes that show differences in relative usage of intronic peaks, as an indicator of changes in the expression-level of pre-spliced mRNA transcripts. Such analyses could be useful in differentiation contexts in identifying what genes are tending to be newly transcribed. There are additional applications that could be used for intronic peaks. One form of AS is intron retention, which has been found to have a role in cancer [15]. Intronic polyadenylation has also been linked to cancer through the inactivation of tumour supressor genes [14]. The majority of intronic peaks from our analysis are annotated as proximal to A-rich regions, indicating that most will be due to internal priming, but by filtering for peaks more likely to represent true polyA sites, analysis of intronic polyadenylation represents a potential application of Sierra.

In conclusion, we have developed a novel computational pipeline for detecting differential transcript usage from 3'-end-based scRNA-seq experiments. Our novel approach to analysing

scRNA-seq yields biological insights unobserved when considering only an aggregate of genes and allows new questions to be asked about the nature of transcriptional regulation between cells.

Methods

Data-sets

The following data-sets are used in this study. The 7k human PBMCs data-set was downloaded from the 10x Genomic website at: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3`. The 4k PBMCs data-set was downloaded from `https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k`. The TIP and GFP$^+$ data-sets are downloadable from `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7376/`. The Tabula Muris dataset was downloaded from the AWS file store provided here. `https://github.com/czbiohub/tabula-muris/blob/master/tabula-muris-on-aws.md` We used cell type annotations provided by the Tabula Muris consortium from here. `https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733`

Previously generated bulk RNA-Seq data sets [31] were downloaded from `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95755`, trimmed using Trimmomatic [37] and aligned using the two-pass STAR alignment method [38].

Sierra pipeline

The Sierra pipeline is implemented as an R package and can be divided into the following main steps, each described in detail below. 1) Splice-aware peak calling is applied to a BAM

file to identify peak coordinates. 2) If multiple data-sets are being analysed, peak coordinates from multiple BAM files are merged together into one set of unified peak coordinates. 3) UMI counting is performed against the set of unified peak coordinates for a provided set of cell barcodes. 4) The peak coordinates are annotated according to the genomic features they fall on and, optionally, according to proximal sequence features corresponding to A-rich regions, T-rich regions or the presence of a canonical polyA motif. 5) Differential transcript usage analysis by applying DEXSeq to pseudo-bulk profiles of cells. 6) Visualisation of relative peak expression and read coverage across genes for select cell populations.

Peak calling

The Sierra peak calling procedure requires as input a BAM file containing the data for the entire experimental run, such as produced by the 10x Genomics CellRanger software. It must include the error corrected cell and UMI barcode tags. Although many peak callers exist in bioinformatics, e.g. for ChIP-seq analysis, it is not appropriate to apply these existing peak callers to scRNAseq data due to the presence of spliced peaks. As the average width of scRNAseq peaks are 500 bp, larger than the size of a typical exon, peaks can contain splice sites and span cross multiple exons. To make our peak caller splice aware, we first extract the splice junctions from a BAM file using ‘regtools‘ [39]. By extracting splice junctions directly from the data, we do not depend on existing transcript annotations. Using these splice junctions, we extract read coverage from the BAM file at all exonic positions. Then for each gene, we find the location of maximum peak coverage and fit a Gaussian curve to the read coverage data at that position. The peak coordinate is set to range from three standard deviations upstream and to three standard deviations downstream from the middle of the peak. We remove read coverage data within this peak location, find the next maximum read coverage location and iterate until the maximum peak coverage is less than a predefined

value. We call these the "exonic" peaks. After finding the exonic peaks, we go back to each splice junction and try to fit a Gaussian curve to the peak coverage within the splice junction. We call these the "intronic" peaks. Sierra ignores peaks in intergenic regions.

Peak merging

Peaks called from multiple independent data-sets are merged into a unified set of peaks as follows. For some n number of data-sets, we perform a pair-wise comparison of data-sets to identify gene-level peaks that should be merged. Given a reference data, for a set of gene-specific peaks, a similarity score is determined between each peak and the peaks from the remaining data-sets. Given a distance, d, between two peaks, calculates as the absolute difference between start and end sites, and a length of the reference peak, l, a similarity score is calculated as $s = 1 - d/l$. Given a large value of d that exceeds the length of the peak, s will become negative – in this situation s is set to 0 to reflect no similarity. Plotting the similarity scores revealed a bimodal distribution with most similarity scores (on average 95% from data-sets shown below) either in a range of 0.75-1 and indicating peaks clearly matching between data-sets, or clearly not matching as indicated by $s = 0$. As examples, the comparison of GFP$^+$ data-sets showed that 97% of peaks were either matched between the sham and MI data-sets with $s = 0$ or $0.75 \leq s \leq 1$ (Figure S5A) and that for the 3 TIP data-sets merged, on average 94% of the comparisons showed $s = 0$ or $0.75 \leq s \leq 1$ (Figure S5B-D). Given the drop-off in similarity distributions observed at $s < 0.75$, a default threshold, $t = 0.75$ is required to match a reference to a comparison, such that $s > t$, and is the parameter used in this study. When evaluating peak matching between a pair of data-sets, we calculate s both from the reference to the comparison (s_r) and from the comparison to the reference (s_c). To ensure the peaks being matched are genuine, we check s_r, but allow for some match variation from t; by default this variation value, v, is 25%. Two peaks are

finally marked as matched if $s_r > t$ and $s_c > t - t * v$. This process is run for all pairwise combinations of data-sets to merge, where given a set of peaks that are matched, the union of the start and end coordinates are taken to create a final merged peak.

UMI counting

To perform UMI counting, Sierra requires a set of unified peak coordinates, GTF file of gene positions, a scRNAseq BAM file and a white list of cell barcodes. We extract alignments for each gene using the GenomicAlignments package, then count the overlaps between the peak coordinates and the alignments using the *countOverlaps* function in GenomicRanges [40]. Extra filtering is performed to ensure only cells that are in the barcodes white list are counted. The final peak to cell matrix is output in matrix market format.

Detecting differential transcript usage

We test for differential transcript usage using the differential exon usage testing method DEXSeq [25], which was developed for bulk RNA-seq, but here is applied to scRNA-seq after transforming the single-cells to be tested into a small number of pseudo-bulk samples. The use of pseudo-bulk samples allows for computational efficiency of testing. Given two sets of cells to be compared, we first build some n number of pseudo-bulk profiles for each of the cell sets by randomly assigning cells into n groups and summing their peak counts. By default, the value of n is set to 6 (see below). We use the *DEXSeqDataSet* function in DEXSeq to build a DEXSeq object, where *countData* is the raw peak counts, *groupID* is set to gene names, and *featureID* is a set of unique gene-peak numbers. We run the *estimateSizeFactors* function, setting the *locfunc* option to use the *shorth* function from the *genefilter* R package [41], as it is suggested in the DEXSeq documentation that the *shorth* function may provide better results for low counts. After estimating size factors we follow the standard DEXSeq pipeline

for testing differential usage. The function to perform DU testing is implemented in Sierra as *DUTest*, and contains options to filter for genomic feature types and peaks annotated as near A-rich regions prior to DU testing.

The main parameter that is introduced here is the n value determining the number of pseudo-bulk profiles. In order to decide on a default value, we considered the impact of different values of n when running DU testing on the TIP and PBMC 7k data-sets by re-running tests for 10 different seeds (and therefore different random assignments of cells to pseudo-bulk profiles) for three values of n: 3, 6 and 10. We evaluated the number of DTU genes obtained (Figure S6A,B), the consistency of those results across the ten seeds (Figure S6C,D) and finally computational time taken for a DTU test (Figure S6). We found that, on average, there was a small increase in the number of DU genes detected, as well as the consistency, with an increased value of n; however, the computational time increased drastically when increasing n from 6 to 10. Thus, a value of *6* maintains a fast computational time while returning a similar number and consistency of DTU genes to a higher n value of 10, and was therefore selected as the default value for Sierra.

Peak annotation

All reported peaks were annotated to identify overlapping feature information within gene transcripts using the Sierra function *AnnotatePeaksFromGTF*. This function takes peak coordinates (GRanges format) and identifies if the boundaries are within a UTR, exon or intron of known genes (supplied as a GTF formatted file). In cases where a gene has multiple transcripts, an annotation hierarchy is applied such that when peaks overlap multiple features UTR > exon > intron. In cases where a peak overlaps multiple features within a single transcript all features are returned. This same function also assesses genomic sequence downstream from peaks for the existance of poly A motifs (i.e. AAUAAA), or A-rich region

(defined as 13 consecutive A with up to 1 mismatch). Similarly genomic sequence upstream of peaks were assessed for the existance of a T-rich region (13 consecutive T with up to 1 mismatch) as previously identified by [28].

Coverage plots

Cell type BAM files were extracted from the original total single cell population BAM alignment files using the Sierra function *SplitBam*. For bulk RNA-Seq data sets the BAM files had previously been imported into SeqMonk using default RNA-seq settings. Coverage information was extracted from SeqMonk files by exporting wig-like files (via selecting running window generator on specific gene lists). Single-cell split BAM files and bulk RNA-Seq coverage data files were passed to the Sierra *PlotCoverage* function.

Detecting 3'UTR shortening

In order to evaluate 3'UTR shortening, we first calculate DU peaks, selecting for peaks falling on 3'UTRs and filtering out peaks annotated as proximal to an A-rich region to enrich for real polyA sites. This is performed in Sierra using the *DUTest* function with *feature.type = "UTR3"* and *filter.pA.stretch = TRUE*. To ensure that our comparative analysis was performed on the same 3'UTR, we selected exon IDs for DU 3'UTR peaks using the GenomicFeatures [40] *threeUTRsByTranscript* function. For each DU 3'UTR peak, we compared the DU peak to the remaining peaks expressed on the same UTR. The peaks were ordered according to their proximity to the start of the 3'UTR and assigned a score between 0 and 1, with the most proximal peak scored 0 and the most distal scored 1. The ordering scores were compared between up- and downregulated peaks using Wilcoxon rank-sum test to evaluate shifts toward more proximal or distal peak expression.

Plotting relative peak expression

Given two or more peaks from a gene, we calculate relative expression values as follows. First, given a set of i cell population identities (e.g. clusters), and n gene peaks, gene-level population means are calculated from the set of n peaks such that $\mathbf{g} = (g_1, g_2, ..., g_i)$ and peak-level population means are calculated such that,

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1i} \\ \vdots & & & \\ p_{n1} & p_{n2} & \dots & p_{ni} \end{bmatrix}$$

Next, we calculate a population-level relative usage score for each peak, defined as the peak population mean divided by the gene population mean; this produces a matrix of relative usage scores, S, such that $S = P/\mathbf{g}$. Finally, the relative peak expression is calculated for each cell, according to its corresponding population, by dividing the peak expression by the gene mean, plus a pseudo count of 1, and multiplying by the relative usage score. The relative expression values are finally log2 transformed with a pseudo count of 1. Thus, given some cell population 1 containing j cells and an associated $n \times j$ matrix of expression values, E_1, relative expression R is calculated as,

$$R_1 = log2(E/(g_1 + 1) \times (p_{11}...p_{n1}) + 1)$$

R is calculated for all cell populations $1, 2, ...i$ to produce a final combined matrix of relative peak expression. The relative peak expression can then be visualised using similar methods for visualising gene expression. Sierra provides four functions for visualising relative peak expression on 1) t-SNE coordinates, 3) UMAP coordinates, 4) box plots and 5) violin plots. Plots are generated using ggplot2 [42].

Clustering analyses

Gene-level clustering of the 4k and 7k PBMC data-sets was performed using the Seurat R package (version 3.0.2). We applied the following quality control metrics: for both data-sets, cells with $> 10\%$ UMIs mapping to mitochondrial genes were filtered out. We visualised the distribution of expressed genes and UMIs and filtered out cells with outliers. For 7k PBMCs, we filtered out cells with over 15,000 UMIs and 3000 genes and for 4k PBMCs we filtered for 10,000 UMIs and 2000 genes. For both data-sets, the UMIs were then normalized to counts-per-ten-thousand, log-transformed and top 2000 variable genes were selected using the *FindVariableFeatures* program. The variable genes were used for principal component (PC) analysis, with the top 20 PCs input to the *FindNeighbors* program. A range of resolutions (*res* parameter) were applied to the *FindClusters* program with a resolution of 0.6 chosen for both the 7k and 4k PBMCs for the DTU analysis presented in this manuscript.

For the comparisons between gene-level and peak-level clustering, we used three data-sets: the PBMC 7k, Pdgfra-GFP$^+$ and TIP. For the GFP$^+$ and TIP data-sets, which were originally clustered in Seurat version 2, we reclustered in version 3 for the purpose of comparisons, though retaining the same PCs used for the original clustering [26]. For all data-sets, the pipeline for clustering on peaks remained identical to that for the relevant gene-level clustering, such as the same number of PCs used for clustering; however, we experimented with increasing the number of highly variable features included for PC analysis, due the peaks in effect splitting genes into multiple features. For the more heterogeneous TIP data we selected 3500 features, for 7k PBMCs we used 2500 and 2000 were retained for GFP$^+$. We compared the returned clusters across 3 *res* parameters: 0.6, 0.8 and 1.0. To compare gene- and peak-level clusters we used the *adjustedRand* function from the clues R package [43].

Declarations

Availability of data and materials

The data-sets analysed during the current study are available at the following locations:

The 7k human PBMCs data-set: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3`.

The 4k PBMCs data-set: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k`.

The TIP and GFP$^+$ data-sets: `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7376/`.

The Tabula Muris data-set:. `https://github.com/czbiohub/tabula-muris/blob/master/tabula-muris-on-aws.md`

Cardiac bulk RNA-seq data: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95755`

Processed peak count files generated in this study are available on xxxxxx under ID xxxxxx.

Funding

RP acknowledges research support from the National Health and Medical Research Council of Australia (NHMRC; APP1118576, 1074386), the Australian Research Council (ARC) Special Research Initiative in Stem Cell Science (SR110001002), Foundation Leducq Transatlantic Networks of Excellence in Cardiovascular Research (15 CVD 03; 13 CVD 01) and the New South Wales Government Department of Health. JWKH is supported by a Career Development Fellowship by the National Health and Medical Research Council (1105271) and a Future Leader Fellowship by the National Heart Foundation of Australia (100848), and

the HKU-USydney Strategic Partnership Fund.

Author contributions

The study was conceived by RP, DH, AO, JWKH and KKL. The experiments were carried out by RP, DH and KKL with advice from AO, JWKH and RPH. RP wrote the manuscript with input from DH and KKL. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Acknowledgments

Not applicable

Tables.

Comparison	Peaks	% Motif	% A-rich	DTU	DEG	Overlap
Mo1 vs CD4+ TC1	825	32	17	492	2042	251
Mo1 vs BC1	576	33	17	347	2059	189
CD8+ TC1 vs FCGR3A+ Mo	339	35	17	214	1817	87
Mo2 vs NKC2	201	32	19	130	1471	57
NKC1 vs CD4+ TC1	83	29	33	63	1506	35
CD4+ TC1 vs CD8+ TC3	66	26	24	50	786	22
NKC1 vs BC1	39	18	36	26	1339	17
BC1 vs CD8+ TC1	28	36	21	21	957	9
CD4+ TC1 vs CD4+ TC2	33	27	27	28	386	10

Table 1. Differential transcript usage (DTU) examples on the PBMC 7k data-set. Shown are the specific cluster comparisons performed, number of peaks called as differentially used (DU), the % identified as proximal to the canonical polyadenylation motif, the % proximal to an A-rich region, the number of unique called DTU genes, the number of genes called as differentially expressed (DEG), and the number of overlapping genes from the DTU and DEG lists.

Data-set	Resolution	Clusters (gene)	Clusters (peaks)	FM index	ARI
TIP	0.6	25	21	0.92	0.91
TIP	0.8	29	25	0.86	0.85
TIP	1.0	30	27	0.79	0.76
PBMC 7k	0.6	17	16	0.84	0.82
PBMC 7k	0.8	18	18	0.82	0.80
PBMC 7k	1.0	20	20	0.81	0.79
GFP^+	0.6	10	12	0.75	0.70
GFP^+	0.8	12	14	0.75	0.71
GFP^+	1.0	14	16	0.66	0.62

Table 2. Comparisons for clustering results based on peak counts relative to gene counts. Columns from left to right show the data-set tested, the resolution parameter (*res*) for the Seurat *FindClusters* program, number of clusters returned from gene-level clustering, number of clusters from peak-level clustering, Fowlkes and Mallow's (FM) index and Adjusted Rand Index (ARI).

Figure legends

Figure 1. Sierra workflow. Sierra starts with a BAM file produced by an alignment program such as CellRanger. Standard gene-level work-flow (top row) involves using a gene model to produce a matrix of gene-level counts used for clustering. The Sierra pipeline performs peak calling to identify subregions corresponding to potential polyadenylation sites. Peak coordinates are used to build an annotated UMI count matrix for each gene peak. This new data can be used to identify genes showing differential peak usage, with visualisation options for plotting relative peak expression and read coverage across selected cell populations.

Figure 2. Representative feature of Sierra data from a 7k cells PBMC data-set. (A) Counts of genes according to number of detected peaks. Dotted red line indicates median number of peaks. (B) Average composition of genomic feature types that peaks fall on, according to number of peaks per-gene. (C) Percentage of cells expressing each genomic feature type with increasing stringency of cellular detection rates for peaks. (D) Number of genes expressing multiple (≥ 2) 3'UTR or exonic peaks with increasing stringency of cellular detection rates. (E) Comparison of *PTPRC* gene expression across cell populations on t-SNE coordinates with peaks identified as DU in monocytes. (F,G) Overlapping genes from a $CD14^+$ monocyte vs $CD4^+$ T-cell comparisons for the PBMC 7k and PBMC 4k data-sets for (F) DTU genes and (G) DE genes, visualised with [44].

Figure 3. Comparison of differential transcript usage between cardiac scRNA-seq and bulk RNA-seq populations. (A) t-SNE plot of the cardiac TIP cell lineages. (B,C) Gene expression visualsed on t-SNE for (B) *Cxcl12* and (C) *Igf1*. (D,E) Relative peak expression visualised on t-SNE for example DU peaks between (D) sham fibroblasts and sham ECs: *Cxcl12*; and (E) sham fibroblasts and MI leukocytes: *Igf1*. (F,G) Read coverage plots across the *Cxcl12* and *Igf1* genes for (F) single-cell and bulk fibroblast and EC populations (*Cxcl12*) and (G) single-cell and bulk sham fibroblast and MI leukocyte populations (*Igf1*). (H) Fisher's exact tests on the number of overlapping DTU genes detected from scRNA-seq and bulk for different cell-type/condition comparisons. Shown are the -log10 P-values and the percent of single-cell DTU genes overlapping the bulk. Red line indicates the significance (0.05) threshold. (I) Overlapping genes between single-cell and bulk RNA-seq from the sham fibroblast and MI leukocyte comparison. (J) Log fold-change comparisons for DU peaks identified in both the single-cell and bulk RNA-seq for the sham fibroblast vs EC analysis.

Figure 4. 3'UTR shortening in activated and proliferating cardiac fibroblasts following MI. (A,B) UMAP visualisation of fibroblast populations from Pdgfra-GFP^{+}/Cd31^{-} mouse cardiac cells at 3 days post-sham or MI surgery showing (A) an aggregate of all cells and (B) the UMAP plot separated according to condition. (C-E) Counts of 3'UTR peaks showing differential usage according to their relative location to the terminating exon. Location of 0 indicates the peak most proximal to the terminating exon, with 1 representing the most distal. Comparisons performed are for (C) F-Cyc against F-SL and F-SH combined, (D) F-CI against F-SL and F-SH combined and (E) F-Act against F-SL and F-SH combined. (F,G) Relative expression of peaks most distal and proximal (to terminating exon) for (F) *Timp2* and (G) *Cd47* as visualised on UMAP coordinates. (H,I) Read coverage across 3'UTR for select single-cell fibroblast populations from sham (F-SL/F-SH combined) and MI (F-Act, F-CI, F-Cyc) data-sets compared to bulk RNA-seq of FACS-sorted fibroblasts from sham and MI conditions for (H) *Timp2* and (I) *Cd47*.

Figure 5. Detecting differential transcript usage across the Tabula Muris data-set. (A) Comparison of the number of DU peaks across cell types within each tissue. Only cell types with more than 100 cells are included in the analysis. (B) and (C) Mammary gland tissue results. (B) Number of DU peaks between cell types. (C) Relative expression plot of DTU genes between a cell type and all remaining cell types in the tissue.

Supplementary figure legends

Figure S1. Peak metrics for 4k PBMC and heart heart total interstitial population (TIP) data-sets. (A,B) Counts of genes according to number of detected peaks where dotted red line indicates median number of peaks for the (A) PBMC 4k data-set and (B) TIP data-set. (C,D) Average composition of genomic feature types that peaks fall on, according to number of peaks per-gene for the (C) PBMC 4k data-set and (D) TIP data-set. (E,F) Percentage of cells expressing each genomic feature type with increasing stringency of cellular detection rates for peaks for the (E) PBMC 4k data-set and (F) TIP data-set.

Figure S2. Expression features of peaks comparing Sierra output to the output from the 10x Genomics CellRanger program. (A) Correlation of CellRanger-derived gene expression to mean peak expression (across a gene) for the PBMC 7k data comparing peaks in each genomic feature type – 3'UTRs, exons, introns and 5'UTRs. (B) Mean expression vs dispersion plots for PBMC 7k data comparing CellRanger gene counts to all peaks and peaks according to genomic feature type. (C) Expression correlation for heart TIP data. (D) Mean expression vs dispersion comparisons for heart TIP data.

Figure S3. Gene read distribution coverage plots across scRNA-seq sham fibroblast and MI-day 3 leukocytes from TIP data compared to representative bulk RNA-seq sham fibroblast and MI-day 3 leukocyte samples for (A) Tfpi, (B) Tm9sf3, (C) Lsp1 and (D) Plek.

Figure S4. Comparison of Seurat clustering between using gene and peak expression with different resolution (*res*) parameters. (A-D) t-SNE visualisation of TIP cell clusters for (A) gene-level clustering with $res = 0.6$, (B) Peak-level clustering with $res = 0.6$, (C) gene-level clustering with $res = 1.0$ and (D) gene-level clustering with $res = 1.0$.

Figure S5. Distribution of peak similarity scores between pairs of data-sets for (A) sham vs MI GFP^+ cells, (B) the TIP sham and MI-day 3 data-sets, (C) TIP sham and MI-day 7 data-sets and (D) TIP MI-day 3 and MI-day 7 data-sets.

Figure S6. Comparison of metrics from differential transcript usage testing with increasing numbers (3, 6 and 10) of pseudo-bulk profiles. Evaluated of 6 cell type comparisons for the TIP and PBMC 7k data-sets across 10 randomised pseudo-bulk splits. Shown are mean and standard deviation for each of the following metrics: (A,B) Number of DU peaks detected per-comparison for (A) TIP and (B) 7k PBMCs. (C,D) Average consistency of DTU genes detected, where consistency is defined as the frequency of a DTU gene being detected ($P_{adj} < 0.05$) across the 10 splits for (C) TIP and (D) 7k PBMCs. (E,F) DU peak calculation time (in seconds) for (E) TIP and (F) 7k PBMCs.

Supplementary tables legends

Table S1. Metrics from peak calling, counting and annotation for the 19 individual data-sets analysed is this study. Shown are 1) the number of peaks, 2) number of genes, 3) median peaks per-gene, 4-7) % of peaks falling on 3'UTRs, exons, introns or 5'UTRs, 8-11) % of peaks with a poly(A) motif according to 3'UTRs, exons, introns or 5'UTRs and 12-5) % of peaks down-stream from an A-rich stretch according to 3'UTRs, exons, introns or 5'UTRs.

Table S2. Comparison of overlapping detected differential transcript usage genes and differentially expressed genes for different cell-type comparisons between the PBMC 7k and PBMC 4k data-sets.

Table S3. Output for DTU testing between cell-types from the TIP data-set.

Table S4. Output for DTU testing between activated/proliferating fibroblast populations and the resting populations. DTU testing based on 3'UTR peaks, after filtering out peaks tagged as proximal to an A-rich region.

References

1. Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008.

2. Ruijia Wang, Dinghai Zheng, Ghassan Yehia, and Bin Tian. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Research*, 28(10):1427–1441, 2018.

3. Francisco E. Baralle and Jimena Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18:437 EP –, 05 2017.

4. Bin Tian and James L. Manley. Alternative polyadenylation of mrna precursors. *Nature Reviews Molecular Cell Biology*, 18:18 EP –, 09 2016.

5. Eric T. Wang, Amanda J. Ward, Jennifer Cherone, Thomas T. Wang, Jimena Giudice, Daniel Treacy, Peter Freese, Nicole J. Lambert, Tanvi Saxena, Thomas A. Cooper, and Christopher B. Burge. Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins. *Genome Research*, 2015.

6. Allissa A Dillman, David N Hauser, J Raphael Gibbs, Michael A Nalls, Melissa K McCoy, Iakov N Rudenko, Dagmar Galter, and Mark R Cookson. mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nature Neuroscience*, 16:499 EP –, 02 2013.

7. Karen Yap, Zhao Qin Lim, Piyush Khandelia, Brad Friedman, and Eugene V. Makeyev. Coordinated regulation of neuronal mrna steady-state levels through developmentally controlled intron retention. *Genes & Development*, 26(11):1209–1223, 2012.

8. Justin J.-L. Wong, William Ritchie, Olivia A. Ebner, Matthias Selbach, Jason W.H. Wong, Yizhou Huang, Dadi Gao, Natalia Pinello, Maria Gonzalez, Kinsha Baidya, Annora Thoeng, Teh-Liane Khoo, Charles G. Bailey, Jeff Holst, and John E.J. Rasko. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, 154(3):583 – 595, 2013.

9. Darya P. Vanichkina, Ulf Schmitz, Justin J.-L. Wong, and John E.J. Rasko. Challenges in defining the role of intron retention in normal biology and disease. *Seminars in Cell & Developmental Biology*, 75:40 – 49, 2018. Diversity of transcripts emanating from protein-coding genes.

10. Robert Middleton, Dadi Gao, Aubin Thomas, Babita Singh, Amy Au, Justin J-L Wong, Alexandra Bomane, Bertrand Cosson, Eduardo Eyras, John E. J. Rasko, and William Ritchie. Irfinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biology*, 18(1):51, 2017.

11. Marina M. Scotti and Maurice S. Swanson. RNA mis-splicing in disease. *Nature Reviews Genetics*, 17:19 EP –, 11 2015.

12. Christine Mayr and David P. Bartel. Widespread shortening of 3'utrs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673 – 684, 2009.

13. Hyun Jung Park, Ping Ji, Soyeon Kim, Zheng Xia, Benjamin Rodriguez, Lei Li, Jianzhong Su, Kaifu Chen, Chioniso P. Masamha, David Baillat, Camila R. Fontes-Garfias, Ann-Bin Shyu, Joel R. Neilson, Eric J. Wagner, and Wei Li. 3'UTR shortening represses tumor-suppressor genes in trans by disrupting cerna crosstalk. *Nature Genetics*, 50(6):783–789, 2018.

14. Shih-Han Lee, Irtisha Singh, Sarah Tisdale, Omar Abdel-Wahab, Christina S. Leslie, and Christine Mayr. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, 561(7721):127–131, 2018.

15. Heidi Dvinge and Robert K. Bradley. Widespread intron retention diversifies most cancer transcriptomes. *Genome Medicine*, 7(1):45, 2015.

16. Hyunchul Jung, Donghoon Lee, Jongkeun Lee, Donghyun Park, Yeon Jeong Kim, Woong-Yang Park, Dongwan Hong, Peter J Park, and Eunjung Lee. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nature Genetics*, 47:1242 EP –, 10 2015.

17. Ángeles Arzalluz-Luque and Ana Conesa. Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biology*, 19(1):110, 2018.

18. David Lukacsovich, Jochen Winterer, Lin Que, Wenshu Luo, Tamas Lukacsovich, and Csaba Földy. Single-cell rna-seq reveals developmental origins and ontogenetic stability of neurexin alternative splicing profiles. *Cell Reports*, 27(13):3752 – 3759.e4, 2019.

19. Yan Song, Olga B. Botvinnik, Michael T. Lovci, Boyko Kakaradov, Patrick Liu, Jia L. Xu, and Gene W. Yeo. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular Cell*, 67(1):148 – 161.e5, 2017.

20. Joshua D. Welch, Yin Hu, and Jan F. Prins. Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, 44(8):e73–e73, 01 2016.

21. Yuanhua Huang and Guido Sanguinetti. Brie: transcriptome-wide splicing quantification in single cells. *Genome Biology*, 18(1):123, 2017.

22. Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

23. Aaron Lun and Davide Risso. *SingleCellExperiment: S4 Classes for Single Cell Data*, 2017. R package version 1.0.0.

24. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, III Mauck, William M., Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019/09/11 2019.

25. Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. *Genome Research*, 22(10):2008–2017, 2012.

26. Nona Farbehi, Ralph Patrick, Aude Dorison, Munira Xaymardan, Vaibhao Janbandhu, Katharina Wystub-Lis, Joshua WK Ho, Robert E Nordon, and Richard P Harvey. Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *eLife*, 8:e43882, 2019.

27. Nicholas Schaum, Jim Karkanias, Norma F. Neff, Andrew P. May, Stephen R. Quake, Tony Wyss-Coray, Spyros Darmanis, Joshua Batson, Olga Botvinnik, Michelle B. Chen, Steven Chen, Foad Green, Robert C. Jones, Ashley Maynard, Lolita Penland, Angela Oliveira Pisco, Rene V. Sit, Geoffrey M. Stanley, James T. Webber, Fabio Zanini, Ankit S. Baghel, Isaac Bakerman, Ishita Bansal, Daniela Berdnik, Biter Bilen, Douglas

Brownfield, Corey Cain, Michelle B. Chen, Min Cho, Giana Cirolia, Stephanie D. Conley, Aaron Demers, Kubilay Demir, Antoine de Morree, Tessa Divita, Haley du Bois, Laughing Bear Torrez Dulgeroff, Hamid Ebadi, F. Hernán Espinoza, Matt Fish, Qiang Gan, Benson M. George, Astrid Gillich, Geraldine Genetiano, Xueying Gu, Gunsagar S. Gulati, Yan Hang, Shayan Hosseinzadeh, Albin Huang, Tal Iram, Taichi Isobe, Feather Ives, Robert C. Jones, Kevin S. Kao, Guruswamy Karnam, Aaron M. Kershner, Bernhard M. Kiss, William Kong, Maya E. Kumar, Jonathan Y. Lam, Davis P. Lee, Song E. Lee, Guang Li, Qingyun Li, Ling Liu, Annie Lo, Wan-Jin Lu, Anoop Manjunath, Andrew P. May, Kaia L. May, Oliver L. May, Marina McKay, Ross J. Metzger, Marco Mignardi, Dullei Min, Ahmad N. Nabhan, Norma F. Neff, Katharine M. Ng, Joseph Noh, Rasika Patkar, Weng Chuan Peng, Robert Puccinelli, Eric J. Rulifson, Shaheen S. Sikandar, Rahul Sinha, Rene V. Sit, Krzysztof Szade, Weilun Tan, Cristina Tato, Krissie Tellez, Kyle J. Travaglini, Carolina Tropini, Lucas Waldburger, Linda J. van Weele, Michael N. Wosczyna, Jinyi Xiang, Soso Xue, Justin Youngyunpipatkul, Macy E. Zardeneta, Fan Zhang, Lu Zhou, Andrew P. May, Norma F. Neff, Rene V. Sit, Paola Castro, Derek Croote, Joseph L. DeRisi, Geoffrey M. Stanley, James T. Webber, Ankit S. Baghel, Michelle B. Chen, F. Hernán Espinoza, Benson M. George, Gunsagar S. Gulati, Aaron M. Kershner, Bernhard M. Kiss, Christin S. Kuo, Jonathan Y. Lam, Benoit Lehallier, Ahmad N. Nabhan, Katharine M. Ng, Patricia K. Nguyen, Eric J. Rulifson, Shaheen S. Sikandar, Serena Y. Tan, Kyle J. Travaglini, Linda J. van Weele, Bruce M. Wang, Michael N. Wosczyna, Hanadie Yousef, Andrew P. May, Stephen R. Quake, Geoffrey M. Stanley, James T. Webber, Philip A. Beachy, Charles K. F. Chan, Benson M. George, Gunsagar S. Gulati, Kerwyn Casey Huang, Aaron M. Kershner, Bernhard M. Kiss, Ahmad N. Nabhan, Katharine M. Ng, Patricia K. Nguyen, Eric J. Rulifson, Shaheen S. Sikandar, Kyle J. Travaglini, Bruce M. Wang, Kenneth Weinberg, Michael N. Wosczyna, Sean M. Wu, Ben A.

Barres, Philip A. Beachy, Charles K. F. Chan, Michael F. Clarke, Seung K. Kim, Mark A. Krasnow, Maya E. Kumar, Christin S. Kuo, Andrew P. May, Ross J. Metzger, Norma F. Neff, Roel Nusse, Patricia K. Nguyen, Thomas A. Rando, Justin Sonnenburg, Bruce M. Wang, Irving L. Weissman, Sean M. Wu, Stephen R. Quake, The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection, processing, Library preparation, sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.

28. Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018.

29. Annalisa Schaub and Elke Glasmacher. Splicing in immune cells—mechanistic insights and emerging topics. *International Immunology*, 29(4):173–181, 06 2017.

30. Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, 16(1):278, 2015.

31. Gregory A. Quaife-Ryan, Choon Boon Sim, Mark Ziemann, Antony Kaspi, Haloom Rafehi, Mirana Ramialison, Assam El-Osta, James E. Hudson, and Enzo R. Porrello.

Multicellular transcriptional analysis of mammalian heart regeneration. *Circulation*, 136(12):1123–1139, 2017.

32. Raul Torres and Juan C. Ramirez. A chemokine targets the nucleus: Cxcl12-gamma isoform localizes to the nucleolus in adult mouse heart. *PLOS ONE*, 4(10):1–10, 10 2009.

33. Rickard Sandberg, Joel R. Neilson, Arup Sarma, Phillip A. Sharp, and Christopher B. Burge. Proliferating cells express mrnas with shortened 3' untranslated regions and fewer microrna target sites. *Science*, 320(5883):1643–1647, 2008.

34. Ran Elkon, Jarno Drost, Gijs van Haaften, Mathias Jenal, Mariette Schrier, Joachim AF Oude Vrielink, and Reuven Agami. E2f mediates enhanced alternative polyadenylation in proliferation. *Genome Biology*, 13(7):R59, 2012.

35. Xing Fu, Hadi Khalil, Onur Kanisicak, Justin G. Boyer, Ronald J. Vagnozzi, Bryan D. Maliken, Michelle A. Sargent, Vikram Prasad, Iñigo Valiente-Alandi, Burns C. Blaxall, and Jeffery D. Molkentin. Specialized fibroblast differentiated states underlie scar formation in the infarcted mouse heart. *The Journal of Clinical Investigation*, 128(5):2127–2143, 5 2018.

36. James J.H. Chong, Vashe Chandrakanthan, Munira Xaymardan, Naisana S. Asli, Joan Li, Ishtiaq Ahmed, Corey Heffernan, Mary K. Menon, Christopher J. Scarlett, Amirsalar Rashidianfar, Christine Biben, Hans Zoellner, Emily K. Colvin, John E. Pimanda, Andrew V. Biankin, Bin Zhou, William T. Pu, Owen W.J. Prall, and Richard P. Harvey. Adult cardiac-resident msc-like stem cells with a proepicardial origin. *Cell Stem Cell*, 9(6):527 – 540, 2011.

37. Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 04 2014.

38. Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 10 2012.

39. Yang-Yang Feng, Avinash Ramu, Kelsy C. Cotto, Zachary L. Skidmore, Jason Kunisaki, Donald F. Conrad, Yiing Lin, William C. Chapman, Ravindra Uppaluri, Ramaswamy Govindan, Obi L. Griffith, and Malachi Griffith. Regtools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. *bioRxiv*, 2018.

40. Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for computing and annotating genomic ranges. *PLOS Computational Biology*, 9(8):1–10, 08 2013.

41. R. Gentleman, V. Carey, W. Huber, and F. Hahne. *genefilter: genefilter: methods for filtering genes from high-throughput experiments*, 2017. R package version 1.60.0.

42. Hadley Wickham. ggplot2: Elegant graphics for data analysis. *Springer-Verlag New York*, 2009.

43. Fang Chang, Weiliang Qiu, Ruben Zamar, Ross Lazarus, and Xiaogang Wang. clues: An r package for nonparametric clustering based on local shrinking. *Journal of Statistical Software, Articles*, 33(4):1–16, 2010.

44. Hanbo Chen and Paul C. Boutros. Venndiagram: a package for the generation of highly-customizable venn and euler diagrams in r. *BMC Bioinformatics*, 12(1):35, 2011.

Figure 1

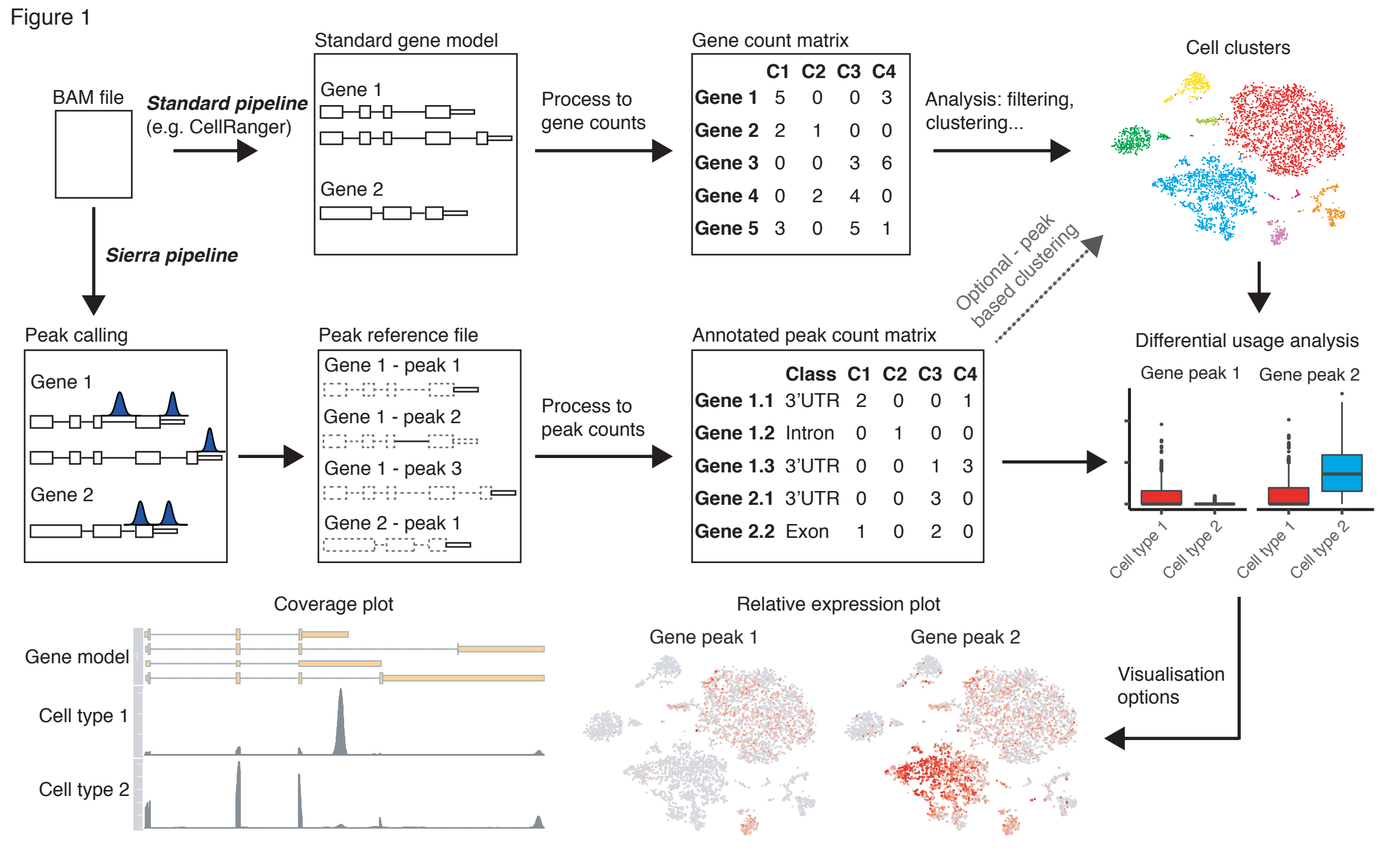

BAM file
Standard pipeline
(e.g. CellRanger)
Standard gene model
Gene 1
Gene 2
Process to
gene counts
Gene count matrix
C1 C2 C3 C4
Gene 1 5 0 0 3
Gene 2 2 1 0 0
Gene 3 0 0 3 6
Gene 4 0 2 4 0
Gene 5 3 0 5 1
Analysis: filtering,
clustering...
Cell clusters
Sierra pipeline
Optional - peak
based clustering
Peak calling
Gene 1
Gene 2
Peak reference file
Gene 1 - peak 1
Gene 1 - peak 2
Gene 1 - peak 3
Gene 2 - peak 1
Process to
peak counts
Annotated peak count matrix
Class C1 C2 C3 C4
Gene 1.1 3'UTR 2 0 0 1
Gene 1.2 Intron 0 1 0 0
Gene 1.3 3'UTR 0 0 1 3
Gene 2.1 3'UTR 0 0 3 0
Gene 2.2 Exon 1 0 2 0
Differential usage analysis
Gene peak 1
Gene peak 2
Cell type 1
Cell type 2
Cell type 1
Cell type 2
Coverage plot
Gene model
Cell type 1
Cell type 2
Relative expression plot
Gene peak 1
Gene peak 2
Visualisation
options

Figure 2

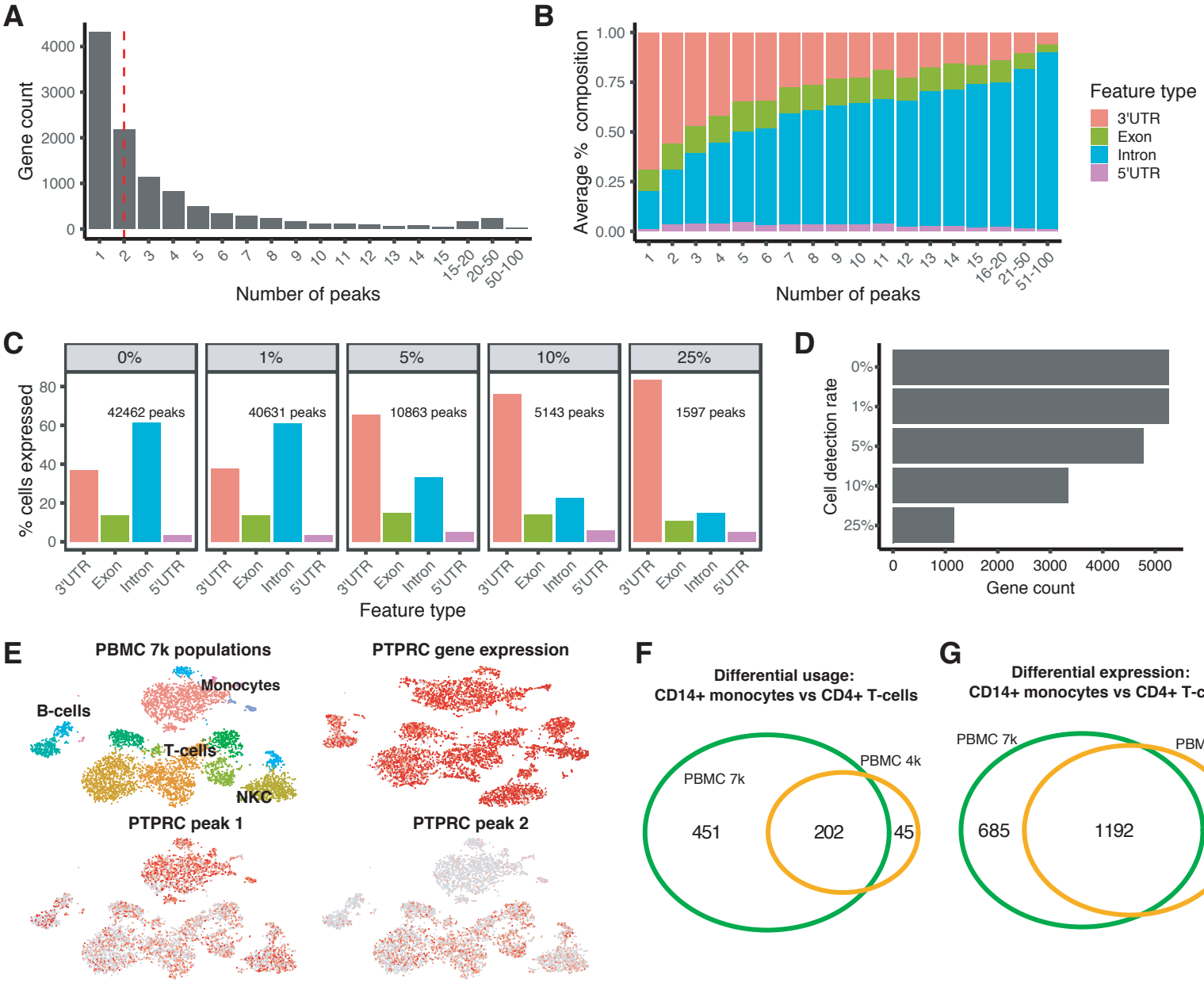

A
Gene count
Number of peaks
B
Average % composition
Number of peaks
Feature type
3'UTR
Exon
Intron
5'UTR
C
% cells expressed
0%
1%
5%
10%
25%
42462 peaks
40631 peaks
10863 peaks
5143 peaks
1597 peaks
3'UTR
Exon
Intron
5'UTR
Feature type
D
Cell detection rate
Gene count
E
PBMC 7k populations
Monocytes
B-cells
T-cells
NKC
PTPRC gene expression
PTPRC peak 1
PTPRC peak 2
F
Differential usage:
CD14+ monocytes vs CD4+ T-cells
PBMC 7k
PBMC 4k
451
202
45
G
Differential expression:
CD14+ monocytes vs CD4+ T-cells
PBMC 7k
PBMC 4k
685
1192
250

Figure 3

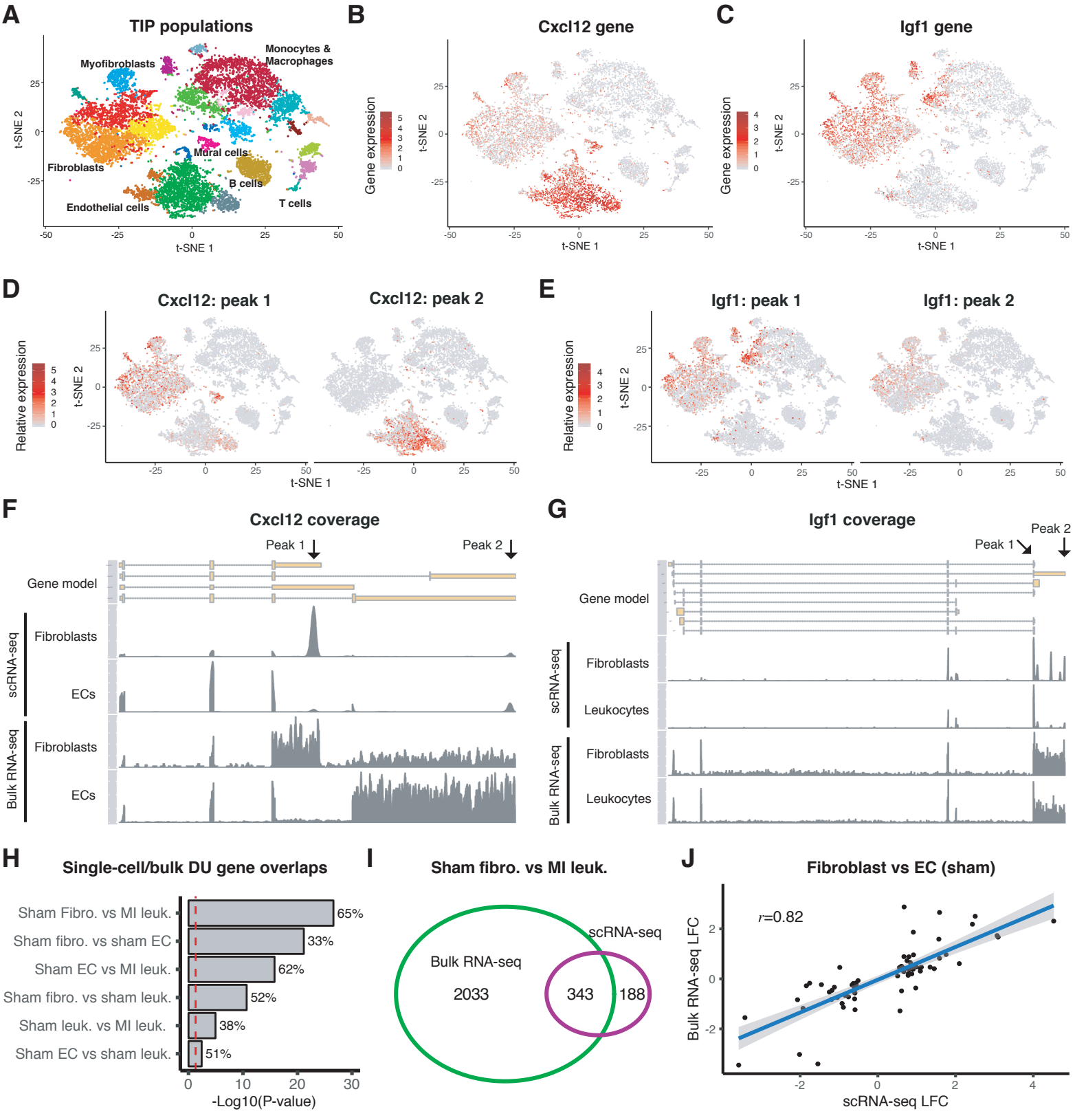

A
TIP populations
Myofibroblasts
Monocytes & Macrophages
Mural cells
Fibroblasts
B cells
T cells
Endothelial cells
t-SNE 1
t-SNE 2
B
Cxcl12 gene
Gene expression
C
Igf1 gene
D
Cxcl12: peak 1
Cxcl12: peak 2
Relative expression
E
Igf1: peak 1
Igf1: peak 2
F
Cxcl12 coverage
Peak 1
Peak 2
Gene model
scRNA-seq
Fibroblasts
ECs
Bulk RNA-seq
G
Igf1 coverage
Leukocytes
H
Single-cell/bulk DU gene overlaps
Sham Fibro. vs MI leuk.
Sham fibro. vs sham EC
Sham EC vs MI leuk.
Sham fibro. vs sham leuk.
Sham leuk. vs MI leuk.
Sham EC vs sham leuk.
65%
33%
62%
52%
38%
51%
-Log10(P-value)
I
Sham fibro. vs MI leuk.
Bulk RNA-seq
scRNA-seq
2033
343
188
J
Fibroblast vs EC (sham)
r=0.82
Bulk RNA-seq LFC
scRNA-seq LFC

Figure 4

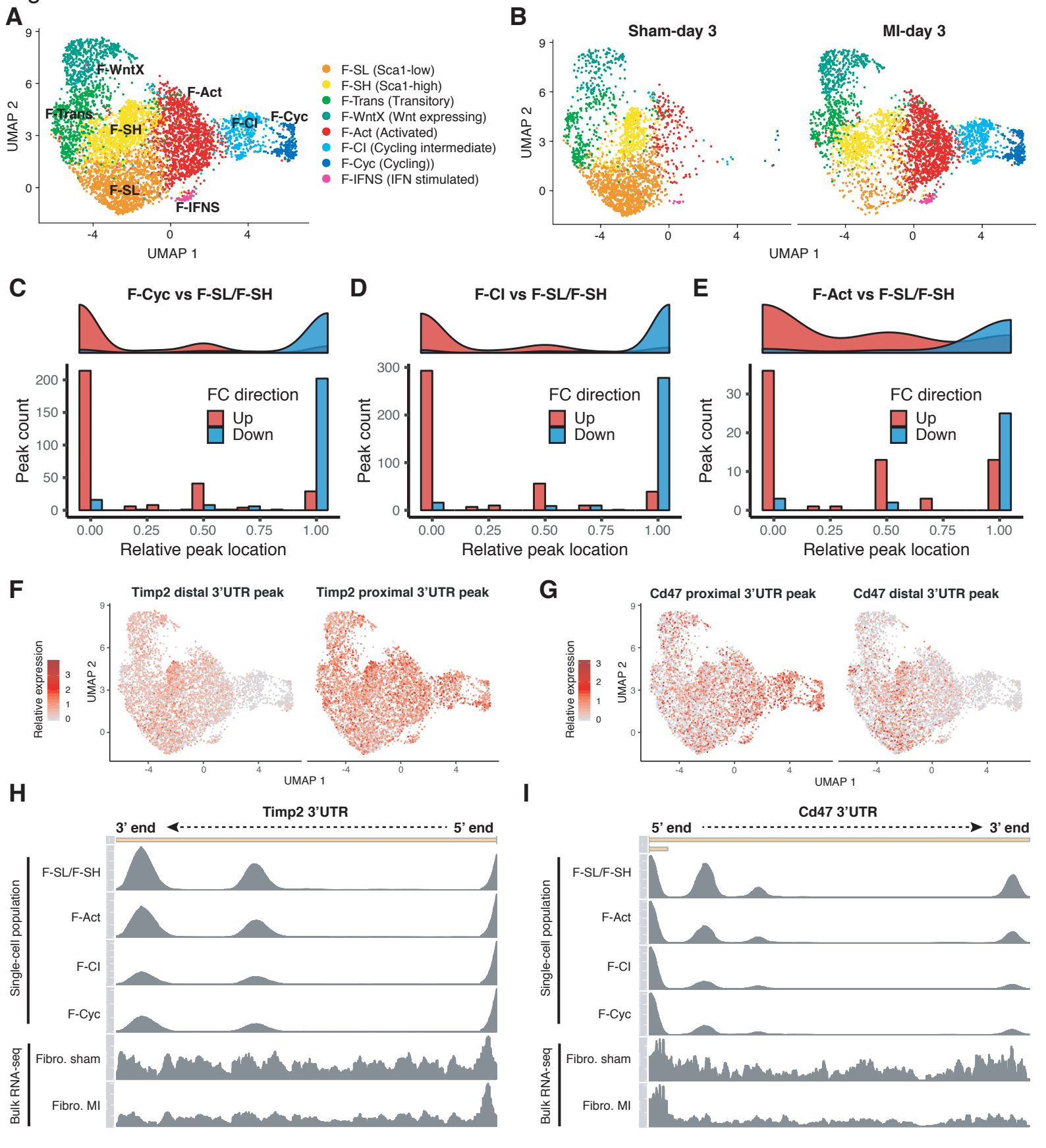
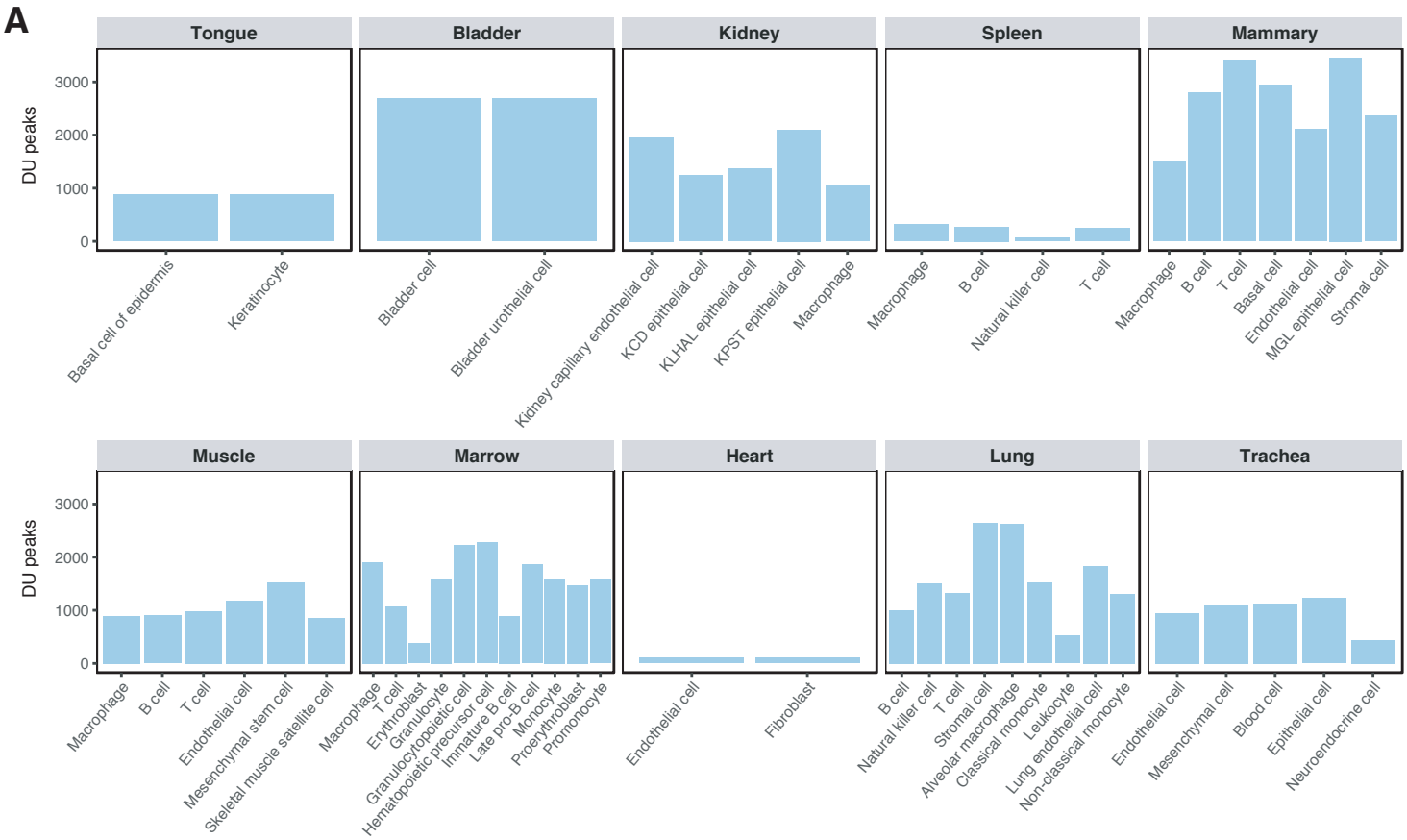
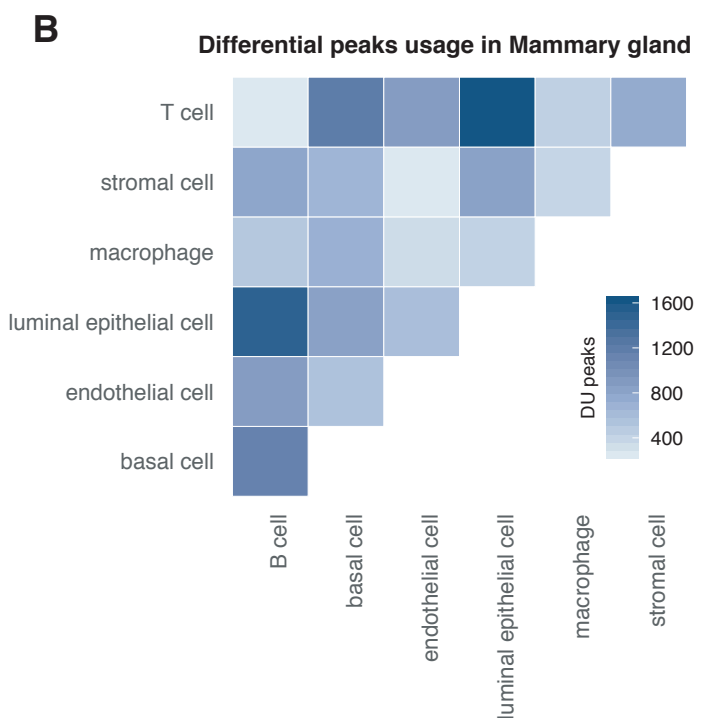
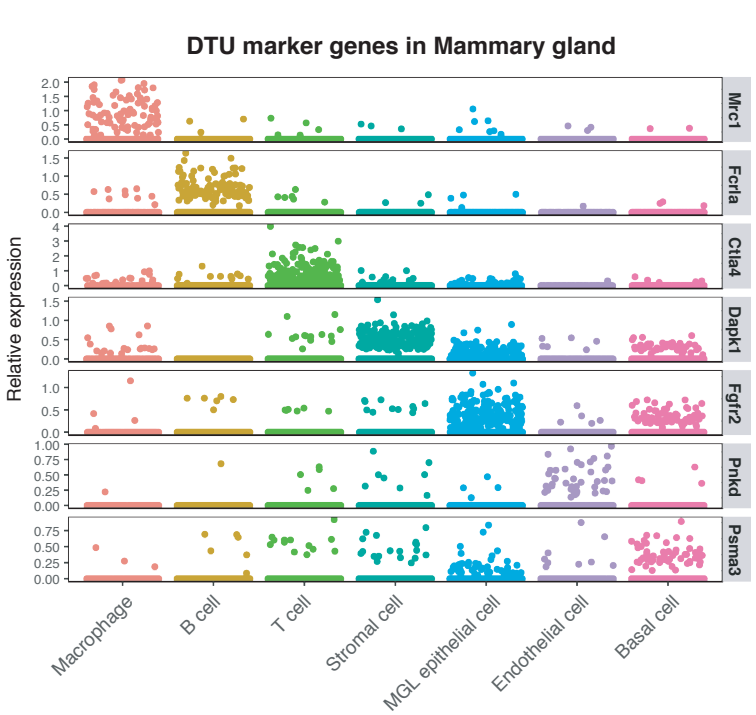

A
F-WntX
F-Act
F-Trans
F-SH
F-Cl
F-Cyc
F-SL
F-IFNS
UMAP 2
UMAP 1
F-SL (Sca1-low)
F-SH (Sca1-high)
F-Trans (Transitory)
F-WntX (Wnt expressing)
F-Act (Activated)
F-Cl (Cycling intermediate)
F-Cyc (Cycling))
F-IFNS (IFN stimulated)
B
Sham-day 3
MI-day 3
C
F-Cyc vs F-SL/F-SH
D
F-Cl vs F-SL/F-SH
E
F-Act vs F-SL/F-SH
Peak count
FC direction
Up
Down
Relative peak location
F
Timp2 distal 3'UTR peak
Timp2 proximal 3'UTR peak
Relative expression
G
Cd47 proximal 3'UTR peak
Cd47 distal 3'UTR peak
H
Timp2 3'UTR
3' end
5' end
Single-cell population
F-SL/F-SH
F-Act
F-Cl
F-Cyc
Bulk RNA-seq
Fibro. sham
Fibro. MI
I
Cd47 3'UTR
5' end
3' end

Figure 5


A
Tongue
Bladder
Kidney
Spleen
Mammary
DU peaks
3000
2000
1000
0
Basal cell of epidermis
Keratinocyte
Bladder cell
Bladder urothelial cell
Kidney capillary endothelial cell
KCD epithelial cell
KLHAL epithelial cell
KPST epithelial cell
Macrophage
Macrophage
B cell
Natural killer cell
T cell
Macrophage
B cell
T cell
Basal cell
Endothelial cell
MGL epithelial cell
Stromal cell
Muscle
Marrow
Heart
Lung
Trachea
Macrophage
B cell
T cell
Endothelial cell
Mesenchymal stem cell
Skeletal muscle satellite cell
Macrophage
T cell
Erythroblast
Granulocyte
Granulocyte/monocyte progenitor cell
Hematopoietic precursor cell
Immature B cell
Late pro-B cell
Monocyte
Proerythroblast
Promonocyte
Endothelial cell
Fibroblast
B cell
Natural killer cell
T cell
Stromal cell
Alveolar macrophage
Classical monocyte
Leukocyte
Lung endothelial cell
Non-classical monocyte
Endothelial cell
Mesenchymal cell
Blood cell
Epithelial cell
Neuroendocrine cell
Key
KCD: Kidney collecting duct
KLHAL: Kidney loop of Henle ascendling limb
KPST: Kidney proximal straight tubule
MGL: Mamary gland luminal


B
Differential peaks usage in Mammary gland
T cell
stromal cell
macrophage
luminal epithelial cell
endothelial cell
basal cell
DU peaks
1600
1200
800
400
B cell
basal cell
endothelial cell
luminal epithelial cell
macrophage
stromal cell


C
DTU marker genes in Mammary gland
Relative expression
Mrc1
Fcrla
Ctla4
Dapk1
Fgfr2
Pnkd
Psma3
Macrophage
B cell
T cell
Stromal cell
MGL epithelial cell
Endothelial cell
Basal cell