University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Tian, L;Jabbari, JS;Thijssen, R;Gouil, Q;Amarasinghe, SL;Voogd, O;Kariyawasam, H;Du, MRM;Schuster, J;Wang, C;Su, S;Dong, X;Law, CW;Lucattini, A;Prawer, YDJ;Collar-Fernández, C;Chung, JD;Naim, T;Chan, A;Ly, CH;Lynch, GS;Ryall, JG;Anttila, CJA;Peng, H;Anderson, MA;Flensburg, C;Majewski, I;Roberts, AW;Huang, DCS;Clark, MB;Ritchie, ME

Title:
Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing

Date:
2021-12-01

Citation:
Tian, L., Jabbari, J. S., Thijssen, R., Gouil, Q., Amarasinghe, S. L., Voogd, O., Kariyawasam, H., Du, M. R. M., Schuster, J., Wang, C., Su, S., Dong, X., Law, C. W., Lucattini, A., Prawer, Y. D. J., Collar-Fernández, C., Chung, J. D., Naim, T., Chan, A. ,... Ritchie, M. E. (2021). Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biology, 22 (1), https://doi.org/10.1186/s13059-021-02525-6.

Persistent Link:
https://hdl.handle.net/11343/292011

License:
CC BY

Tian et al. Genome Biology (2021) 22:310
https://doi.org/10.1186/s13059-021-02525-6

Genome Biology

METHOD Open Access

Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing

Luyi Tian[1,2*], Jafar S. Jabbari[1,3], Rachel Thijssen[1,2], Quentin Gouil[1,2], Shanika L. Amarasinghe[1,2], Oliver Voogd[1], Hasaru Kariyawasam[1], Mei R. M. Du[1], Jakob Schuster[1], Changqing Wang[1], Shian Su[1,2], Xueyi Dong[1,2], Charity W. Law[1,2], Alexis Lucattini[3], Yair David Joseph Prawer[4], Coralina Collar-Fernández[5], Jin D. Chung[6], Timur Naim[6], Audrey Chan[6], Chi Hai Ly[6,7], Gordon S. Lynch[6], James G. Ryall[6,8], Casey J. A. Anttila[1], Hongke Peng[1,2], Mary Ann Anderson[1,2,9], Christoffer Flensburg[1,2], Ian Majewski[1,2], Andrew W. Roberts[1,2,9,10,11], David C. S. Huang[1,2], Michael B. Clark[4] and Matthew E. Ritchie[1,2*]

* Correspondence: tian.l@wehi.edu.au; mritchie@wehi.edu.au
[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
Full list of author information is available at the end of the article

Abstract

A modified Chromium 10x droplet-based protocol that subsamples cells for both short-read and long-read (nanopore) sequencing together with a new computational pipeline (*FLAMES*) is developed to enable isoform discovery, splicing analysis, and mutation detection in single cells. We identify thousands of unannotated isoforms and find conserved functional modules that are enriched for alternative transcript usage in different cell types and species, including ribosome biogenesis and mRNA splicing. Analysis at the transcript level allows data integration with scATAC-seq on individual promoters, improved correlation with protein expression data, and linked mutations known to confer drug resistance to transcriptome heterogeneity.

Keywords: Single-cell gene expression, Long-read sequencing, Splicing, Single-cell multi-omics

Background

Single-cell RNA sequencing (scRNA-seq) is a widely adopted method for profiling transcriptomic heterogeneity in health and disease [1]. However, assessing transcript-level changes between cell types using current scRNA-seq protocols is challenging due to their reliance on short-read sequencing. Previous studies using plate-based methods [2, 3] have focused on individual alternative splicing events such as exon skipping, due to the fundamental limitation of short-read sequencing in linking distal splicing outcomes belonging to the same transcript. The Smartseq3 protocol [4] can achieve full-length transcript coverage but is still unable to assemble the complete transcript sequence and is heavily reliant on the reference annotation. Droplet-based methods [5] such as 10x only sequence the 3′ or 5′ end of transcripts which largely precludes

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

isoform identification. Long-read sequencing can overcome this limitation and generate full-length transcript information in single cells, as illustrated in several recent studies [6–9]. However, the throughput of current long-read sequencing platforms is still not comparable to short-read platforms (especially for Pacific Biosciences' Sequel platform) and the per-base accuracy may also be lower (particularly for Oxford Nanopore Technologies), which together create many issues. Limited sequencing throughput introduces a trade-off between the per-cell sequencing depth and the number of cells or genes processed. Protocols such as ScISOr-Seq [6] sequence all processed cells necessitating either shallow depth per cell or high cost, while RAGE-Seq [7] focuses on specific transcripts rather than the whole transcriptome. The recently published ScNaUmi-seq [9] protocol requires saturated sequencing to error-correct UMI sequences, which increases the sequencing cost. On top of the current protocol limitations, another pressing issue is the lack of data analysis pipelines for long-read transcriptome data, especially for single cells. Tools such as *ScNapBar* [10] and SiCeLoRe [9] focus on cell barcode and UMI assignment, while others such as *FLAIR* [11] and *TALON* [12] lack the ability to process single-cell data. New protocols and computational pipelines to overcome these limitations are needed.

To this end, we adapted the popular Chromium 10x protocol and the data analysis platform *FLAMES* to perform single-cell isoform sequencing and data analysis. Adapting the 10x Chromium platform creates a cost-effective approach to discover and quantify isoforms in single cells by integrating data from short- and long-read sequencing technologies. Subsampling single cells from a full 10x run and applying nanopore long-read sequencing can achieve a comparable number of sequencing reads per cell to that obtained from short-read platforms. For data analysis, we developed a computational framework to perform single-cell full-length analysis of mutations and splicing (*FLAMES*), which includes cell barcode and UMI assignment from nanopore reads as well as semi-supervised isoform discovery and quantification. We applied this methodology to human and mouse samples containing different cell types and highlight shared splicing patterns in human cancer cells and mouse quiescent muscle stem cells. Differential transcript usage analysis pinpointed common functional modules and genes across samples. We also demonstrate that these protocols are a promising tool for detecting coding variants of clinical relevance. Taken together, our modified protocol and data analysis pipeline enable comprehensive characterization of the full-length isoforms present in single cells that are currently overlooked in short-read sequencing datasets.

Results

High-throughput single-cell full-length transcriptome sequencing with Chromium 10x

We modified the standard Chromium scRNA-seq protocol (10x Genomics) to better amplify the full-length cDNA. Since the throughput of long-read platforms is still limited compared to Illumina sequencing platforms, we subsample 10–20% of the 10x Chromium generated Gel Bead-in-Emulsions (GEMs) after reverse transcription (Fig. 1A), similar to the method recently described by Lebrigand et al. [9]. This is equivalent to sampling 10–20% of the cells as the cDNA of each cell is still within the GEMs. After subsampling, the GEMs are pooled separately for library preparation. Part of the amplified cDNA from the 10–20% subsample is used for Oxford Nanopore Technologies long-read library preparation and sequencing on a

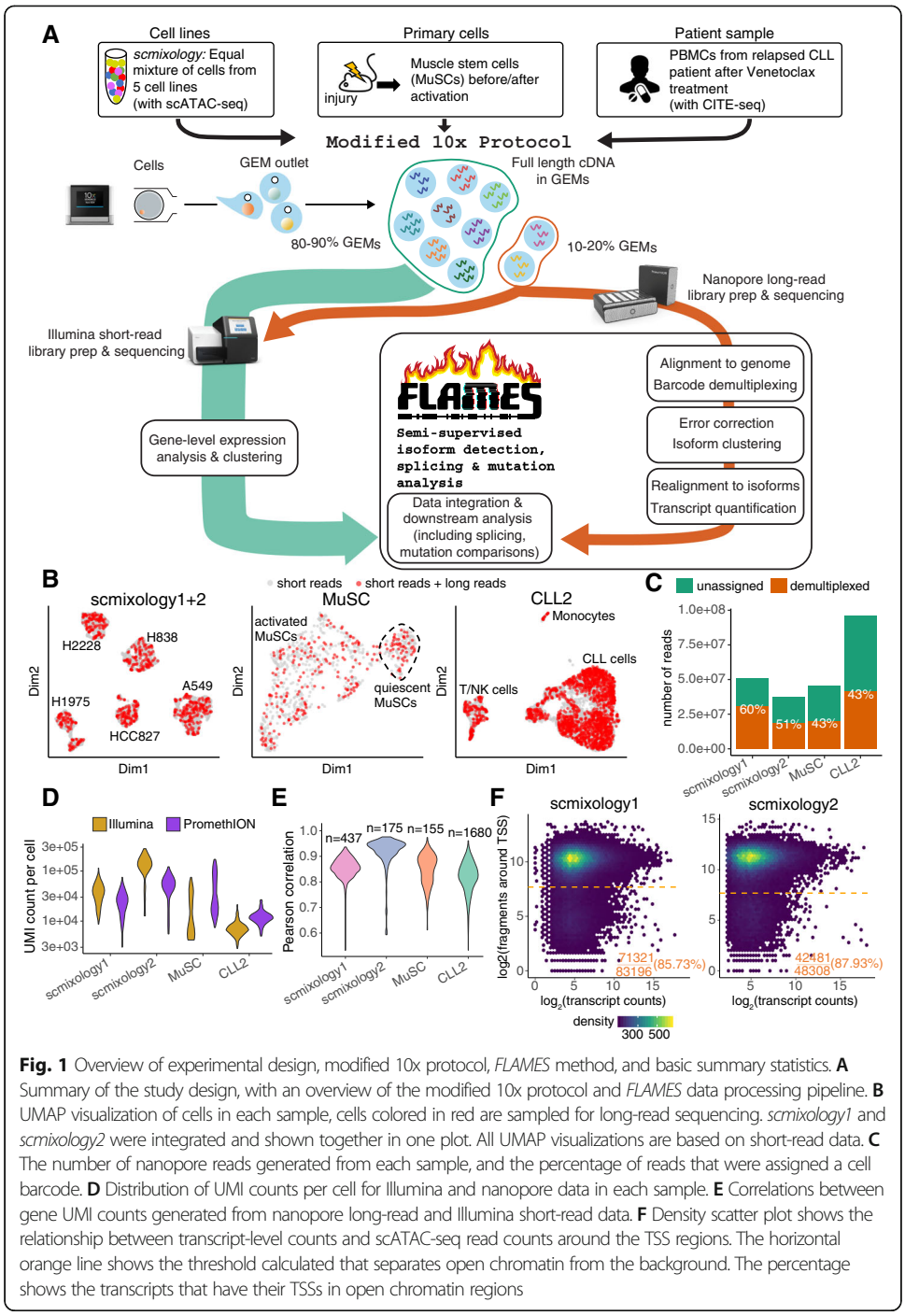


Fig. 1 Overview of experimental design, modified 10x protocol, *FLAMES* method, and basic summary statistics. **A** Summary of the study design, with an overview of the modified 10x protocol and *FLAMES* data processing pipeline. **B** UMAP visualization of cells in each sample, cells colored in red are sampled for long-read sequencing. *scmixology1* and *scmixology2* were integrated and shown together in one plot. All UMAP visualizations are based on short-read data. **C** The number of nanopore reads generated from each sample, and the percentage of reads that were assigned a cell barcode. **D** Distribution of UMI counts per cell for Illumina and nanopore data in each sample. **E** Correlations between gene UMI counts generated from nanopore long-read and Illumina short-read data. **F** Density scatter plot shows the relationship between transcript-level counts and scATAC-seq read counts around the TSS regions. The horizontal orange line shows the threshold calculated that separates open chromatin from the background. The percentage shows the transcripts that have their TSSs in open chromatin regions

PromethION. The remainder of the cDNA from the 10–20% sample together with the GEMs from the 80–90% sample are used for regular 10x library preparation and Illumina sequencing in parallel. In the end, long-read data from the 10–20% subsample of cells and Illumina short-read data for all cells are generated by this protocol.

We demonstrate the use of this protocol by profiling 16,660 cells from diverse biological systems, 2737 of which were sequenced by both long-read and short-read technologies (Fig. 1A, Additional file 1). Firstly, we used our previously published

scmixology design [13], which involved an equal mixture of cells from five cell lines (H2228, H838, H1975, HCC827, A549). Two biological replicates were profiled (*scmixology1* and *scmixology2*) with this protocol together with 10x scATAC-seq for the second replicate. In addition to the cell-line mixtures, we sequenced freshly isolated quiescent and activated muscle stem cells (*MuSC*s) from mouse. Lastly, we applied this protocol to a cryogenically preserved peripheral blood mononuclear cell (PBMC) sample from a patient (*CLL2*) whose chronic lymphocytic leukemia (CLL) had progressed on venetoclax treatment after a durable response. The sample was prepared together with the 10x CITE-seq assay with 17 antibody markers. This demonstrated the broad utility of this approach, which is compatible with different 10x transcriptomic assays and can be applied to both fresh and frozen samples. The Uniform Manifold Approximation and Projection (UMAP) visualization presented the cell populations as expected (Additional file 2: Fig. S1) and revealed no obvious bias in the GEM sampling process (Fig. 1B). Collectively, we generated 230 million long reads across all samples, together with scATAC-seq and CITE-seq data for the *scmixology2* and *CLL2* samples, respectively.

Single-cell isoform detection and quantification with *FLAMES*

We developed a flexible computational framework called *FLAMES* (*F*ull-*L*ength *A*nalysis of *M*utations and *S*plicing in long-read RNA-seq data) to detect and quantify isoforms for both single-cell and bulk long-read data (Fig. 1A, Additional file 2: Fig. S2 and Methods). Input to *FLAMES* are fastq files generated from the long-read platform. Using the cell barcode annotation obtained from short-read data as the reference, it identifies and trims cell barcodes/UMI sequences from the long reads. After barcode assignment, all reads were aligned to the relevant genome to obtain a draft read alignment. The draft alignment is then polished and grouped to generate a consensus transcript assembly. All reads are aligned again using the transcript assembly as the reference and quantified. Benchmarking of the isoform detection component of *FLAMES* on bulk synthetic spike-in control samples showed good performance in terms of recovery of known transcripts while identifying fewer false-positive transcripts compared to other leading methods *FLAIR*, *TALON*, and *StringTie2*. Next, we benchmarked *FLAMES* on a bulk SIRV spike-in dataset [14] for which the isoform structure and abundances are known a priori. Our results clearly show that *FLAMES* outperforms *FLAIR*, *TALON*, and *StringTie2* [15] both in terms of the isoform detection (Fig. S3A, Additional file 3) and quantification (Additional file 2: Fig. S3B). Additional benchmarking of *FLAMES* on a bulk sequins spike-in dataset showed similar results [16].

We used *FLAMES* to preprocess and analyze the four datasets generated. Forty to 60% of the long reads could be assigned to an expected cell barcode and were kept for further analysis (Fig. 1C). The transcript coverage of reads realigned to assembled transcripts showed that the percentages of full-length reads decreased for longer transcripts (Fig. S4), which has also been shown in another study [17]. Reads that are not full-length and cannot be uniquely assigned to transcripts are discarded during data processing. The average UMI count per cell ranges from 10,000 to 60,000, varied by cell type, and was comparable to the short-read counts from the same cells (Fig. 1D). Gene-level UMI counts between the matched nanopore and Illumina data were also

found to be highly correlated (Fig. 1E). The data processed by *FLAMES* showed that our modified 10x protocol generated high-quality long-read data that was comparable to the short-read Illumina data when analyzed at the gene level.

To validate the transcription start sites (TSSs) from the isoforms generated by *FLAMES*, we compared them to the FANTOM5 TSS peaks [18] and found that around 75% of the TSSs are within the FANTOM5 annotation (Additional file 2: Fig. S5A). For the *scmixology* data where scATAC-seq data from the same populations were also available, we aggregated scATAC-seq signals around the TSSs as an indicator of open promoters. The result showed *scmixology1* and *scmixology2* have similar open promoters and more than 85% of the TSSs are within active promoters, supporting the existence of these transcripts (Fig. 1F). In contrast, when we process the *scmixology* data using *TALON*, *FLAIR*, and *StringTie2*, the results were less optimal. The majority of transcripts generated by *FLAIR* and *TALON* did not match the reference annotations (Additional file 2: Fig. S5B), and *FLAIR*, *TALON*, and *StringTie2* had fewer TSSs overlapping the FANTOM5 annotations compared to *FLAMES* (Fig. S5C). We found similar results when comparing the scATAC-seq signals around the TSS regions from transcripts generated by different methods (Fig. S5D), with *TALON* and *FLAIR* having around 40% and 50% of their TSSs in open chromatin regions, respectively. In summary, our comparisons yield similar results between a spike-in dataset and the *scmixology* dataset, with *FLAMES* outperforming *StringTie2*, *FLAIR*, and *TALON* with the latter two likely generating many false transcripts.

Characterization of isoforms reveals the distinct splicing landscapes of different cell populations

We compared the transcripts generated by *FLAMES* to the reference annotation and classified them using the scheme introduced in *SQANTI* [19], including transcripts with all splice junctions matching to reference transcripts (full splice match, FSM) or partially matching to consecutive splice junctions for a reference transcript (incomplete splice match, ISM) and transcripts with novel splice junctions with new (Novel, not in catalog, NNC) or known (Novel, in catalog, NIC) donor and acceptor sites (Fig. 2A). We observed that the number of transcripts detected varies between samples (Fig. 2B) and is correlated with the sequencing throughput as shown in Fig. 1C. While around half of the transcripts detected were novel, the majority of reads aligned to known transcripts with novel transcripts having lower abundance in general. In addition to the comparison with reference annotations, we also compared the transcripts generated from the three human samples to each other (Fig. 2C) and found many transcripts unique to each sample (22%, 27%, and 68% for *scmixology1*, *scmixology2*, and *CLL2*, respectively). A majority of transcripts (~60%) were shared between biological replicates and among these 30% were novel, suggesting that many conserved alternative splicing events in these cell lines have not been annotated. The results from *FLAMES* were also compared to short-read bulk RNA-seq profiled from matched samples analyzed using *STAR* (Additional file 2: Fig. S6). The degree of overlap of known junctions between the long- and short-read results was relatively high across all samples (Additional file 2: Fig. S6A). For novel junctions, the overlap was considerably less (Additional file 2: Fig. S6B), with relatively fewer junctions detected in the short-read data across the different

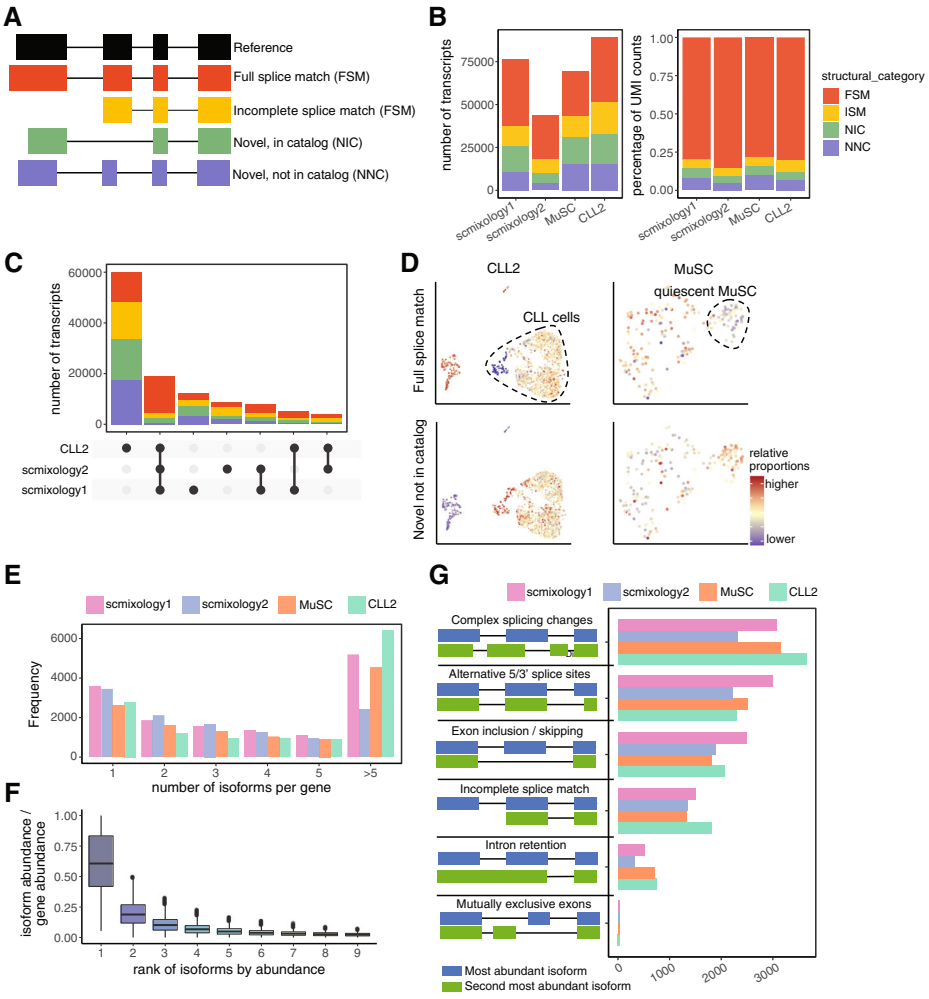


Fig. 2 Overview of the single-cell isoform-level analysis from *FLAMES*. **A** Classification of transcripts according to their splice sites when compared to reference annotations. **B** Summary of transcripts in different categories in **A** both in numbers (left) and in the percentage of UMI counts (right). **C** UpSet plot showing overlap of transcripts in human datasets, where the number of transcripts shared by different sets of samples is indicated in the top bar chart, colored by categories specified in **A**. **D** UMAP visualization of *CLL2* and *MuSC* dataset on the cells sampled for long-read sequencing. Colored by percentage of UMI counts of transcripts in FSM (top) and NNC (bottom) categories. CLL cells and quiescent MuSCs are annotated on the plot. **E** Bar plot of the number of distinct transcripts expressed per gene. Genes with more than five distinct transcripts are merged. **F** Box plot showing the percentage of transcript abundance relative to gene abundance for genes express multiple transcripts. Transcripts are ranked by abundance, shown on the *x*-axis. **G** Summary of the type of alternative splicing between the two most abundant transcripts of each gene. The "Complex splicing changes" category represents transcripts with more than one type of constitutive alternatively spliced event

samples compared to the long-read data, which identified between 3- and 14-fold more novel junctions.

Within samples, we found consistent alterations in splicing patterns between cell populations (Fig. 2D). CLL cells had higher proportions of novel transcripts, especially transcripts with novel splice junctions (NNC), compared to non-CLL cells from the same sample, including T cells, NK cells, and monocytes. Similarly, quiescent muscle stem cells also had higher proportions of novel transcripts compared to activated stem cells. Analysis of intronic reads from RNA-seq data [20] has demonstrated that intron retention, which would produce novel transcripts, is increased in quiescent mouse

muscle stem cells and is essential for these cells to exit the quiescent state. This is consistent with our results and suggests that differences in splicing patterns between different cell populations may act as a regulatory mechanism.

After comparing the transcripts identified against reference annotations, we sought to characterize isoforms within the same gene. Around 80% of genes can be expressed as multiple isoforms (Fig. 2E). The average number of isoforms expressed per gene ranges between 3 and 6 and varied between the different samples (Fig. 2E). The distribution of isoform expression is skewed with only a few abundant transcripts dominantly expressed for most genes (Fig. 2F), consistent with previous results [21]. On average, the two most highly expressed isoforms account for 80% of the total counts (median 85%). Next, we categorized the types of alternative splicing between the two most highly expressed isoforms (Fig. 2G). Alternative splicing has mostly been studied based on particular events such as exon skipping or alternative 5′ splice sites using short-read sequencing technology [22]. We found the most common category, comprising around 30% of genes, has more than one type of alternative splicing event between the top two isoforms. This means that the 2 most highly expressed isoforms differ by complex splicing changes involving multiple exons, which may not be accurately characterized by short reads because two isoforms could have a skipped exon near the 5′ end and a different splice site near the 3′ end.

Common classes of genes with differential transcript usage across samples

Following the analysis of isoform abundance, we investigated whether genes with multiple isoforms exhibited differential transcript usage (DTU) between the clusters or cell types shown in Additional file 2: Fig. S1. We focused on changes in the use of internal splice junctions, grouping transcripts with the same intron chain. To mitigate the high dropout rate in single-cell data, we aggregated the transcript counts into pseudo-bulk values per cluster and filtered out transcripts with low abundance. A chi-square test was performed on the pseudo-bulk count matrix for DTU analysis. To assess how sensitive the DTU results are to sequencing level, we down-sampled the number of long-reads from the *scmixology2* dataset (from 20 to 80% of the total number). We observed that although the number of detected isoforms (Additional file 2: Fig. S7A) and the number of genes with DTU (Additional file 2: Fig. S7B) decrease with decreasing sequencing level per cell, 58% of the DTU changes could still be detected with as little as 20% of the starting reads (Additional file 2: Fig. S7B). We performed DTU analysis on all samples and identified between 500 and 1000 genes with differential transcript usage (gDTU) in each sample (Fig. 3A, Additional file 4). Nearly half of the gDTU (244 out of 573) are shared between *scmixology1* and *scmixology2*. Besides this overlap, gDTU were largely unique for each sample (Additional file 2: Fig. S8A). The functional annotation clustering of these genes revealed shared pathways across different samples, especially pathways related to transcription and translation, such as mRNA splicing and ribosome biogenesis (Fig. 3B, Additional file 5). As an example, different gDTU related to mRNA splicing were found in different samples. PQBP1, which is involved in pre-mRNA splicing, is a gene with DTU in the *scmixology* samples (Additional file 2: Fig. S8B), while SRSF2 and SRSF3, which belong to a family that acts both as general splicing factors

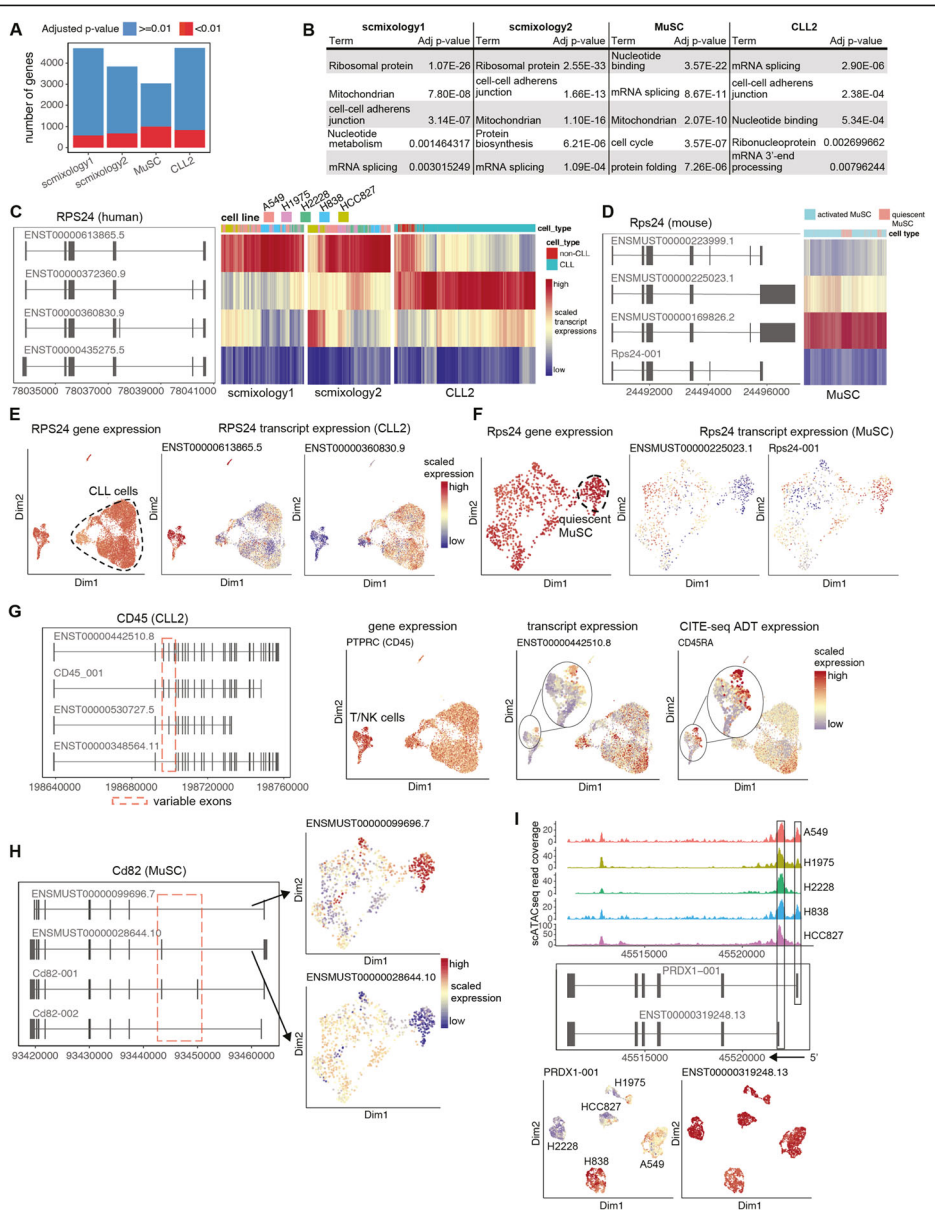


Fig. 3 Summary of differential transcript usage results from *FLAMES*. **A** Summary of results from the statistical testing of DTU detected many significant genes per sample (adjusted *P*-value < 0.01). **B** Table of common functional categories among different samples from the functional enrichment analysis of gDTU. **C** Top 4 most abundant isoforms of RPS24 in human and heatmap of their expression at the single-cell level in the *scmixology1*, *scmixology2*, and *CLL2* samples. **D** Top 4 most abundant isoforms of RPS24 in mouse and heatmap of their expression at the single-cell level in *MuSCs*. **E** UMAP of cells in *CLL2*, colored by RPS24 gene expression and transcript expression. Two transcripts with differential expression on different populations were selected. Transcript expression in each cell is colored by scaled relative expression to highlight the difference between different populations. **F** Similar to **E**, UMAP of cells in *MuSC* sample, colored by RPS24 gene expression and transcript expression. **G** Top 4 most abundant isoforms of CD45 in *CLL2* and UMAP visualizations of the cells colored by (from left to right) gene expression, transcript expression, and corresponding protein expression. H Top 4 most abundant isoforms of CD82 in *MuSC*, with UMAP visualization of cells colored by expression of two isoforms that have differential expression between quiescent and activated *MuSC*. **I** scATAC-seq read coverage for PRDX1 with cells from each cell line aggregated and plotted together. UMAP plots showing isoform expression, with each cell colored by scaled transcript expression

and as regulators of alternative splicing [23], exhibit DTU in the *CLL2* sample (Additional file 2: Fig. S8C).

Apart from the gDTU that are unique to each sample, we also found a few genes (18) that appeared in all samples (Additional file 2: Fig. S8A), including a ribosomal protein RPS24 that had the smallest *P*-value among different samples (Additional file 4). RPS24 is a highly conserved gene between mouse and human, with exons 5 and 6 alternatively spliced to generate different protein coding sequence (Fig. 3C, D). Further analysis showed transcript usage of this gene is altered at multiple levels in different samples. Firstly, the major transcript of RPS24 was distinct between different samples. Transcripts without exons 5 and 6 were most abundant in the cell lines (*scmixology1* and *scmixology2*, Fig. 3C), while the transcript with only exon 5 was highly expressed in the patient-derived PBMC sample (*CLL2*, Fig. 3C) and *MuSC* sample (Fig. 3D). In addition to the major transcript, additional transcripts showed differential transcript usage in different cells within the same sample. The transcript with exons 5 and 6 was more frequently expressed in CLL cells while the transcript without exons 5 and 6 was preferentially expressed in non-CLL cells (Fig. 3E). Besides these known transcripts, we also identified a new isoform in mouse (Rps24-001) that was preferentially expressed in quiescent MuSCs (Fig. 3F). Different RPS24 isoforms have tissue-specific expression [24] and some have been linked to tumor progression [25]; however, the functional differences between the encoded proteins remain unclear. In summary, we have highlighted the heterogeneity of expression of the RPS24 transcript that is missed in a typical gene-level analysis (Fig. 3E, F).

Another category of gDTU of interest is cell-surface proteins. Genes encoding cell-surface proteins produce alternative mRNA isoforms, usually by changing the combinations of consecutive exons corresponding to certain functional domains. Some of the isoforms have been characterized, such as CD45, where alternative splicing of exons 4 to 6 are expressed in different lymphocytes [26]. By analyzing the *CLL2* data where the surface marker expression is available through CITE-seq, we detected multiple isoforms of CD45 (Fig. 3G). We found similar expression patterns between the transcript and the protein it encodes, where the protein was quantified by counting antibody-derived tags (ADT) from CITE-seq data. The result both validated the isoform quantification from the *FLAMES* pipeline and showed that transcript-level analysis can provide a better correlation between mRNA and protein expression that cannot be achieved using gene-level quantification. CD45 was the only DTU gene identified among the CITE-seq panel of 17 antibodies. We also observed other genes encoding important cell surface markers with DTU, such as Cd82 (Fig. 3H), CD47 (Additional file 2: Fig. S8D), and CD44 (Additional file 2: Fig. S8E), with different combinations of consecutive exons. Notably, some of the novel transcripts in CD47 can be observed in different samples (Additional file 2: Fig. S8D), suggesting conserved alternative splicing patterns that are missing from the reference annotation. In sum, we observed previously overlooked isoform diversity of cell-surface proteins, which could introduce functional diversity and contribute to development and diseases [27, 28].

Alternative promoters have been shown to regulate cancer-specific transcription [29]. Here, we found transcripts with different TSSs expressed in different cancer cell lines. We have shown that open chromatin captured by scATAC-seq is correlated with the promoter region indicated by the TSS (Fig. 1F). As an example, we found multiple

isoforms of PRDX1, including a new isoform PRDX1-001 with a different TSS. The new isoform expressed in H838 and A549 contains a novel first exon that is not found in the other cell lines (Fig. 3I). Additional open promoter regions are observed in the scATAC-seq data from H838 and A549 that coincide with the new exon (Fig. 3I). This result suggests that the long-read scRNA-seq data can be integrated with scATAC-seq at the promoter level to further enhance the resolution of integration and reveal promoter heterogeneity that cannot be found via gene-level integration [30].

Long-read scRNA-seq links coding sequence variation to transcriptome heterogeneity

Nanopore long-read sequencing allows full transcript coverage compared to 3′ or 5′-end short-read sequencing, which provides us with a better chance of identifying coding variations. However, the high error rate in nanopore sequencing presents an obstacle [31]. To overcome this challenge, we first exclude homopolymer regions since they have higher sequence error rates [32, 33]. After excluding these regions, ~95% of regions have a reference allele frequency >90% (Additional file 2: Fig. S9A). The remaining single-nucleotide variants (SNVs) were filtered again using a statistical test in *FLAMES* to enrich for true positives and clonal variants, based on the assumption that the occurrence of sequencing errors is independent of cell barcode and such errors will occur randomly in all cells (Fig. 4A and the "Methods" section). We tested our approach on the *scmixology* samples where the five cell lines carry distinct variants. The principal component analysis (PCA) on the filtered allele count matrix successfully recapitulated the expected population structure (Fig. 4B). Louvain clustering on leading principal components generated similar results to the cell type assignment obtained from running *Demuxlet* [34] on the short-read data (98.7% and 99.4% concordance for *scmixology1* and *scmixology2*, respectively), which indicated that we can successfully capture variants in these different cell lines using long-read data. We then performed

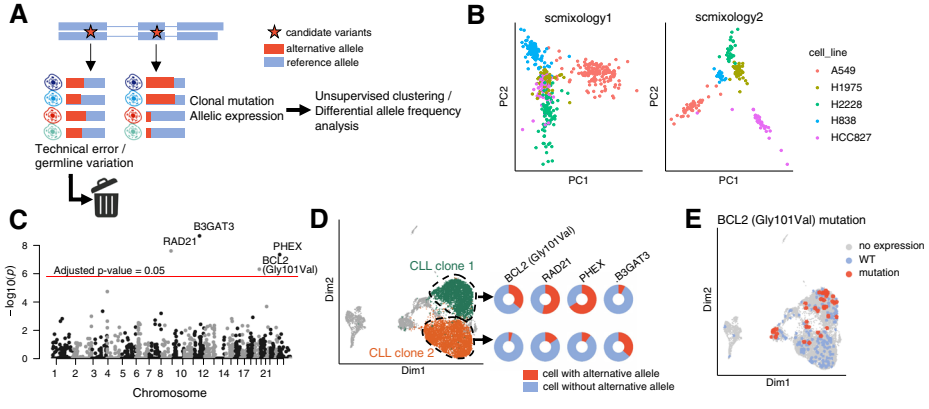


Fig. 4 Summary of differential allele frequency analysis to detect coding mutations. **A** Summary of variation filtering and analysis pipeline implemented in *FLAMES*. Candidate variants are filtered based on allele frequency first, then based on per cell allele frequency to remove technical artifacts. The remaining variants are used for differential allele frequency analysis. **B** PCA on alternative allele matrix using the variants after filtering, colored by unsupervised clustering results using top PCs and annotated with cell lines. **C** Manhattan plot of *P*-values from a differential allele frequency analysis with Benjamini–Hochberg adjustment. Genes with significant variants are labeled. **D** UMAP visualization highlighting two CLL populations that have differential allele frequency for the significant variants. **E** UMAP visualization of cells colored by Gly101Val mutation status

differential allele frequency analysis to find SNVs that were specific to each cell line. The results showed a high precision (80.1%) with the SNVs called from bulk RNA-seq data from individual cell lines.

After we confirmed the approach using the cell lines, a similar analysis was performed on the *CLL2* dataset to examine the relationship between transcriptome heterogeneity and sequence variations. We searched for SNVs that only existed in the cancer cells and had a differential allele frequency across different CLL transcriptional clusters. We found four significant variants (adjusted *P*-value < 0.05, chi-square test, Benjamini–Hochberg correction) associated with different CLL clusters (Fig. 4C, D). Material from the same sample was further analyzed using bulk RNA-seq and whole exome sequencing. All four variants were detected in the bulk RNA-seq, and 3 out of 4 were present in the bulk whole exome sequencing (the RAD21 variant was outside the capture region). The variant in PHEX is a known polymorphism and the variant allele frequency was consistent with it being a germline SNV, so the difference of its allele frequency could be a result of allele-specific expression. By investigating the allele frequency of these variants across clusters (Additional file 2: Fig. S9B), we identified two subclones, where SNVs in the genes BCL2, RAD21, and PHEX are enriched in *subclone1* and the variant located in the gene B3GAT3 is enriched in *subclone2* (Fig. 4D). Although the analysis itself does not require genomic sequence data as a reference, it is important to validate the variants detected since the changes in allele frequency may also come from allele-specific expression or RNA editing. The Gly101Val mutation has been confirmed to promote resistance to venetoclax treatment by reducing the affinity of BCL2 for venetoclax [35], and patient *CLL2* was known to carry ~25% Gly101Val mutations. Our modified protocol shows that the mutation is not just subclonal, but also linked to specific transcriptional clusters. The distribution of the Gly101Val mutation was further confirmed by capture enrichment of the BCL2 gene (Additional file 2: Fig S9C), which greatly increased the coverage of the mutation site (Additional file 2: Fig S9D). Taken together, this approach provides unbiased high-throughput linking of single-cell variants and transcriptomic heterogeneity.

Discussion

Transcriptional activity is typically summarized at the gene level due to the limitations of short-read technology, especially in scRNA-seq studies. The recent development of long-read sequencing technology promises sequencing of the full-length transcript, but its application to single cells has been slowed by a lack of protocols and data analysis pipelines. Using a modified 10x Genomics scRNA-seq protocol that includes subsampling of cells, we demonstrate that a similar per-cell sequencing level can be achieved to short-read technologies at a much lower cost than other methods [6] because only a subset of the cell population undergoes nanopore sequencing. As library preparation artifacts can cause a substantial fraction of reads to lack barcodes (40–60% in this study), enriching for full-length cDNAs that contain barcodes as proposed in Lebrigand et al. [9] has the potential to further enhance the efficiency of sequencing. The hybrid approach taken in our study combines the advantages of throughput and accuracy from short-read sequencing to identify cell barcodes and perform clustering with the strengths of long-read sequencing that enables isoform discovery and quantification, the results of which are combined in an integrated analysis. This modified protocol is

compatible with a variety of 10x transcriptomic assays and could be potentially applied to any single-cell transcriptomic protocol with cell barcoding [36]. It could also be combined with other single-cell long-read sequencing approaches such as R2C2 [37] or other sequencing platforms including PacBio Sequel II. As the throughput and accuracy of long-read sequencing platforms increases, it may be possible to skip the sampling step and short-read sequencing altogether and apply long-read sequencing alone rather than the current hybrid approach to achieve similar aims and further reduce sequencing costs.

To analyze noisy long-read data, we developed the *FLAMES* pipeline that can detect and quantify novel and known isoforms in single-cell and bulk samples [16]. *FLAMES* implements a strict filter on consensus read clustering and filters out spurious reads that are mostly truncated reads with incomplete 3′ or 5′ ends. It also has two rounds of alignment using the genome and the recovered isoforms as reference to reduce alignment artifacts and distinguish reads aligned to similar isoforms. The recently published *SiCeLoRe* method can also detect and quantify isoforms, although it requires deep sequencing to perform UMI correction and has not been as extensively validated. *FLAMES* also identifies cell-type-specific splicing patterns and variants through comparative analysis. Further validation of our approach is available in the recent demonstration that gene fusions can be detected using the *scmixology* data and the output of *FLAMES* together with the *JAFFAL* [38] long-read fusion finding tool. Areas for further improvement of *FLAMES* include using more sophisticated methods for assessing differential transcript usage and allele frequency between clusters (a simple chi-square test is currently applied), extending the framework to accommodate multi-sample analyses, and applying an optimized nanopore-specific variant caller to improve the mutation analysis. There is also a pressing need for further systematic benchmarking of long-read isoform analysis methods, which is currently being pursued by the Long-read RNA-seq Genome Annotation Assessment Project (LR-GASP, https://www.gencodegenes.org/pages/LRGASP/).

With *FLAMES*, we characterized full-length isoforms at the single-cell level across species and cell types. We detected thousands of novel transcripts expressed at low levels, most of which were unique to each sample. Novel transcripts also showed distinct enrichment patterns at the single-cell level where CLL cells and quiescent muscle stem cells often have more novel transcripts. We frequently found complex splicing changes between the two most abundant isoforms of a gene, suggesting the inadequacy of studying individual splicing junctions via short reads. We observed cell-type-specific isoform usage for genes that are enriched for common functions such as mRNA splicing and ribosome biogenesis. The expression of ribosomal and spliceosome proteins has low correlation with their gene-level expression [39], indicating some protein expression levels might be regulated by alternative splicing. Genes encoding cell-surface proteins such as CD44 and CD47 often have cell-type-specific splice variants, some of which have not been previously annotated, and may result in different functions [40]. The examples we show indicate that identification of cell-type-specific isoforms may have broad applications in many fields such as immunotherapy, where cell surface proteins play an important role.

Compared to short-read scRNA-seq, our approach provides a better linkage between proteome and genome as it can resolve the full-length transcript. Transcript abundance

can be aligned directly to single-cell protein measurements, which is particularly helpful for proteins with multiple isoforms such as CD45. We also showed that data integration with scATAC-seq is possible by summarizing open chromatin signals at the exact transcription start site rather than at the gene level. Full transcript coverage unlocks the potential to detect mutations and allelic expression for specific isoforms in an unbiased way compared to methods such as GoT-seq [41]. Through unsupervised analysis, we not only identified coding variants that cannot be easily detected by 3′ or 5′-end short-read sequencing, but we could also associate variants to different transcriptional signatures. This can improve our understanding of the correlation between the single-cell transcriptome and genome in cancer. The detection of cluster-specific transcript fusions is also possible using this type of data [38].

Conclusions

Our approach provides new avenues for characterizing single-cell transcriptomic heterogeneity at the transcript level and unveils new questions and challenges. Benchmarking studies that compare the performance of emerging long-read sequencing protocols and computational methods are needed to help our understanding of the strengths and weaknesses of different approaches. Many unanswered questions about transcript expression and alternative splicing in single cells also remain. For example, do the majority of novel transcripts simply reflect stochastic noise in the splicing machinery [42], or are they indicative of a genuine increase in protein diversity? Our methods and analysis provide a starting point for addressing these questions to accelerate isoform-level studies in single cells.

Methods

Human cell lines

The cell culture and sample preparation of the *scmixology* cell lines are as previously described. Briefly, the five cell lines (H2228, H838, H1975, HCC837, A549) were retrieved from the ATCC (https://www.atcc.org/) and cultured in Roswell Park Memorial Institute (RPMI) 1640 medium with 10% fetal calf serum and 1% penicillin–streptomycin. The cells were grown independently at 37°C with 5% carbon dioxide until near 100% confluency. The cells were then counted and mixed in equal numbers. The mix was used for Chromium 10x library preparation. The first batch was derived from the same sample featured in a previous study (*scmixology1*). The cell lines were cultured again using the same protocol to create a second batch (*scmixology2*) processed by both Chromium 10x scRNA-seq and scATAC-seq.

Mouse muscle stem cells

Animals

All procedures were approved by the Animal Ethics Committee of The University of Melbourne and conformed to the Australian code of practice for the care and use of animals for scientific purposes as stipulated by the National Health and Medical Research Council of Australia. Mice were housed in the Biological Research Facility at The University of Melbourne under a 12-h light–dark cycle, with drinking water and standard chow provided ad libitum. $Pax7^{creERT2}$R26-$eYFP^{fl/fl}$ mice were generated from

Pax7^{creERT2} and R26-eYFP$^{fl/fl}$ founder mice strains [43, 44] both on a C57BL6 background.

Muscle injury and stem cell isolation

At 3 months of age, Pax7^{creERT2}; R26R^{YFP} mice received daily 100μL tamoxifen (20mg/ml in corn oil) for 5 days to label MuSCs with YFP. Two weeks after the final injection, animals were killed and hindlimb muscles excised and dissociated as described previously [45]. Cells were sorted on a FACS Aria III (BD Biosciences) with gating based on YFP.

To obtain activated MuSCs, mice were anesthetized with isoflurane and muscles were injured with 1.2% barium chloride (Sigma). Lower hindlimbs received 40μL barium chloride via an intramuscular injection. Activated cells were isolated 72 h postinjury while quiescent MuSCs were isolated from uninjured muscles as described previously.

Human CLL sample (*CLL2*)

After providing written informed consent, the patient sample was collected after progression on venetoclax treatment [46] (Human Research Ethics Committee approvals: Melbourne Health 2011.044, 2016.305, 2012.274; Peter MacCallum Cancer Centre 11/18; Walter and Eliza Hall Institute 05/04, 13/01). Blood was collected in EDTA tubes and processed within 2 h. Peripheral blood mononuclear cells (PBMCs) were isolated using Ficoll-Paque Plus (#17144002, lot:10258101, GE Healthcare) density gradient centrifugation and cells were cryopreserved. PBMCs were thawed, rested for 2 h, and incubated with Fc Receptor blocking solution (Human TruStain FcX, Biolegend) for 10 min prior to staining with TotalSeq C antibodies (Biolegend) at 4 ° for 30 min. PBMCs were washed three times and stained with propidium iodide (PI, Sigma). Viable cells (PI negative) were flow sorted using the FACSAria (BD) and diluted to 1000 cells/μL.

scRNA-seq cell isolation and library preparation

Single-cell capture and cDNA amplification for mouse skeletal muscle stem cell and cell lines was performed using 10x Genomics Chromium Single cell 3′ Library and Gel Bead (v2 for *MuSC* and *scmixology1* and v3 for *scmixology2*) and for the *CLL2* sample using Chromium Single Cell 5′ Library & Gel Bead Kit (v1) according to manufacturer instructions with some modifications. Full-length cDNA generation was carried out as described in detail at protocols.io [47]. Briefly, we followed the standard 10x Genomics user guide, with RT time increased to 2 h in the *CLL2* sample only to potentially increase the reverse transcription of longer transcripts. After GEM-RT, we transferred 10–20% volumes of GEMs into a new tube and performed subsequent steps in parallel for both 10–20% and 80–90% subsample where each subsample is treated as a separate sample according to the 10x user guide for Illumina library preparation for *scmixology* and *MuSC* samples and increasing cDNA PCR extension step to 3 min for the *CLL2* sample. Long-read library preparation is detailed at protocols.io [47]. cDNA generation for hybrid capture input was similar to the main protocol except that cDNA from the remaining sample (i.e., 80–90%, excluding the 10–20% subsample) was used for amplification using primers 10x Partial R1: CTACACGACGCTCTTCCGATCT and T5′ PCR Primer IIA: AAGCAGTGGTATCAACGCAGAG in place of FPSfilA and RPSfilBr.

To capture the BCL2 gene, all annotated isoforms from Gencode and Ensembl data bases and isoforms identified in ~20% subsamples were integrated and used for IDT xGEN Lockdown capture probe design with one probe for each exon up to 1Kb. To design probes for exons shorter than 120 nucleotides, they were concatenated to flanking exons and two probes were designed to include short exons and part of flanking exons. Hybrid capture was done with 2μg of cDNA input using IDT xGEN Hybridisation and Wash Kit. Captured cDNA were pulled down with streptavidin beads and washed with IDT buffers. The targeted cDNA library was amplified in a similar manner to the main protocol for 12 cycles with Takara LA Taq DNA polymerase in place of PrimeSTAR GXL DNA polymerase.

Standard single-cell Illumina libraries were prepared according to 10x protocol. Illumina HiSeq2500 was used for sequencing *scmixology1* (2 × 125 cycles) and *MuSC* (2 × 150 cycles). Other libraries were sequenced on NextSeq 500 (1 × 28/1 × 91 cycles plus 8 base index cycle) using the v2 150 cycle High Output kit (Illumina) as per the manufacturer's instructions. The base calling and quality scoring were determined using Real-Time Analysis on board software v2.4.6, while the FASTQ file generation and de-multiplexing utilized bcl2fastq conversion software v2.15.0.4. Full-length cDNA libraries from subsamples or capture cDNA were prepared using Oxford Nanopore Technologies SQK-LSK109 Ligation Sequencing Kit with the following modifications: incubation times for end-preparation and A-tailing were lengthened to 15 min, and all washes were performed with 1.8X Ampure beads to conserve smaller fragments. SFB was used for the final wash of the libraries. Fifty femtomole per library was sequenced on PromethION FLO-PRO002 R9.4.1 flow cells according to manufacturer protocols.

Single-cell ATAC-seq (scATAC-seq)

The cells from five cell lines were counted and mixed equally. Cell nuclei were isolated and washed according to the Nuclei Isolation for Single Cell ATAC Sequencing (10x Genomics) protocol, with 1 million cells to start with (0.2 million from each cell line) and 3 min lysis on 100μL buffer. Nuclei were then used to generate scATAC-seq libraries according to the Chromium Single Cell ATAC Reagent Kits User Guide (10x Genomics; CG000168 Rev B). Sequencing libraries were loaded on an Illumina sequencer with 2 × 75 paired-end kits using the following read length: 72 bp read 1, 8 bp i7 index, 16 bp i5 index, and 72 bp read 2. In the sequencing reaction, reads 1 and 2 contain the DNA insert, while the index reads, i5 and i7, capture the cell barcodes and sample indices, respectively. Cells were sequenced on Illumina HiSeq2500 with near around 300 million read pairs in total.

scATAC-seq data analysis

scATAC-seq sequencing data was demultiplexed, preprocessed, and aligned with the default settings of the single-cell ATAC *Cell Ranger* platform (1.0). The reference used for alignment through the *Cell Ranger* platform was hg38. Next, *Picard tools* (http://broadinstitute.github.io/picard/) was used to remove the PCR duplicates. *Samtools* [48] (1.7) was used to extract read pairs that have mapping quality (MAPQ)>30, were non-mitochondrial, and not chimerically mapped. *bedtools* [49] (v.2.26.0) was used to identify reads in mate pairs (i.e., fragments) and adjust the start of the paired reads to account for the 9-bp region that the transposase enzyme occupies during transposition

(i.e., +4 bp for + strand and −5 bp for − strand). Next, *Demuxlet* was used to identify the cell lines of each cell barcode using the genotypic information acquired in our previous benchmarking studies [13, 50].

The count matrix was generated for each barcode-separated BAM file using the *featureCounts* function in the *Rsubread* package [51] (1.32.4) in the *R* environment (v. 3.5.1). The annotation features were promoter regions (i.e., TSS − 500 bp to TSS + 200 bp) corresponding to isoforms identified by the various long-read analysis methods (i.e., *FLAMES*, *TALON*, *FLAIR*, *StringTie2*). To determine the background of scATAC-seq and identify the open promoter, we performed analysis on randomized TSS, where a random position in each gene body was used as the TSS. The 90% percentile of the fragment count around random TSS (7.64) was used as the threshold to determine the open promoters and is annotated in Fig. 1F and Additional file 2: Fig. S5D.

For Fig. 3I, scATAC-seq coverage was calculated from the aligned BAM files using *compute_coverage* in the *plyranges* package [52] (1.7.14) and visualized using *view_coverage* in the *superintronic* [53] package (0.99.4).

Illumina short-read data analysis

The fastq data were processed by *scPipe* to generate a gene count matrix for all samples except *CLL2*, which was processed by *Cell Ranger* (3.0.0) to generate the antibody and gene count matrix. Each gene count matrix was used as input to the standard *Seurat* pipeline with normalization performed by *SCTransform* [54]. Clustering was performed for the *MuSC* and *CLL2* samples with a resolution equal to 0.6. The cell line annotation for *scmixology* was acquired using *Demuxlet* with the same parameters as our previous benchmark study. Integration of the *scmixology1* and *scmixology2* datasets was performed using *Seurat*. The clustering results and cell line annotation are shown in Additional file 2: Fig. S1.

Nanopore sequencing and data preprocessing

We performed basecalling on the raw fast5 data using *Guppy* (1.8.1 for *MuSC* sample and 3.1.5 for *scmixology* and *CLL2*) from Oxford Nanopore Technologies. For each read, we locate the barcode sequence by searching for the flanking sequence before the cell barcode. The cell barcodes identified from the short-read data provide a reference to search for and trim in the long reads. An edit distance of up to 2 is allowed during cell barcode matching. Reads that failed to match any cell barcode were discarded. Sequences following the cell barcode were used as UMIs and trimmed. For the 3′ end protocol, the polyA tail after the UMI sequence was trimmed and sequences after the polyA tail were kept. The cell barcode and UMI sequence were integrated into the fastq read header as per *scPipe* [55]. The processed fastq was used as input for genome alignment and further analysis.

Detection and quantification of isoforms

Reads were aligned to the genome by *minimap2* [56] (-ax splice --junc-bonus 1 -k14 --secondary=no --junc-bed) using Gencode reference (human hg38.v33, mouse mm10.vM24). *FLAMES* summarizes the alignment for each read by grouping reads with similar splice junctions (<5bp by default) to get a raw isoform annotation. The

raw isoform annotation is compared against the reference annotation to correct potential splice site and transcript start/end errors. Transcripts that have similar splice junctions (<5bp by default) and transcript start/end (<100bp by default) to the reference transcript were merged with the reference. This process will also collapse isoforms that are likely to be truncated transcripts. This is achieved by modeling the possibility of a read to be truncated as a linear function to the isoform length, given that longer isoforms are more likely to have truncated reads with incomplete 5′/3′ ends (depending on the 10x protocol applied). Next, the sequence of each polished transcript was extracted and used as the updated reference. The reads were realigned to this reference by *minimap2*. The transcript coverage of individual reads is summarized in Additional file 2: Fig. S3. We noticed that the *scmixology2* data contained more reads that were not full-length, which might relate to the difference in sample preparation time or the v2 and v3 10x chemistry. The transcripts with fewer than 5 full-length aligned reads (>95% coverage by default) were discarded. The reads were assigned to transcripts based on both alignment score, fractions of reads aligned, and transcript coverage. Reads that cannot be uniquely assigned to transcripts or had low transcript coverage (< 60%) were discarded. The UMI transcript count matrix was generated by collapsing the reads with the same UMI in a similar way to what is done for short-read scRNA-seq data, but allowing for an edit distance of up to 2. The counts of transcripts from the same gene were aggerated to generate the gene-level UMI count and compared to the gene count generated from the short-read data in Fig. 1E. All default parameters mentioned above are encoded in a configuration file (config_sclr_nanopore_default.json) and can be changed to fit different protocols. For example, the 5-bp distance for splice site correction could be smaller for data generated by PacBio which has a lower error rate. *FLAMES* is written in python and R and uses other packages, including *pysam* [48] (0.15.2), *numpy* [57] (1.14.2), and *editdistance* (0.5.3) (pypi.org/project/editdistance).

Comparison of *FLAMES* to other tools

Direct RNA sequencing data from SIRV spike-in E2 mix which contains 69 synthetic isoform transcripts (from 7 SIRV genes) was downloaded from NCBI (SRX3204589). The reference annotation for the spike-ins was provided to all methods to ensure a fair comparison. Alignment was performed using *minimap2* (2.17), with "-ax splice --splice-flank=no -k14 --secondary=no" and "--junc-bed." For *TALON*, the mapped reads were processed using *TranscriptClean* (1.02) to correct for mismatches and microindels. Following correction, long reads were collapsed into a transcript isoform quantification table in *TALON* (4.1), using the SIRV annotation. We ran the *FLAIR* pipeline (1.4) using default parameters with all modules including "flair align," "flair correct," "flair collapse," and "flair quantify." The SIRV annotation was supplied in the "flair correct" step to help correct misaligned splicing sites. We ran *StringTie2* (2.0.4) with "-L -G -c 10" and the SIRV annotation. For *FLAMES*, we used the default parameters except "strand_specific:1" and we filtered transcripts with at least 10 reads. For comparison, transcripts generated by *TALON* and *FLAIR* were also filtered to have a read count of at least 10. The *scmixology1* data was processed in a similar way to the SIRV data. Gencode human hg38.v33 was used as the gene reference annotation for

each method. *TALON* was run in parallel on each chromosome to reduce compute time and the results were aggregated later. Unfortunately, we were unable to install the *SiCeLoRe* software on our system to compare with *FLAMES.*

Isoform classification and splicing analysis

SQANTI2 (4.1, https://github.com/Magdoll/SQANTI2) was used to compare the transcripts identified to the reference with parameter "-g --cage_peak --coverage --force_id_ignore." We used the FANTOM5 cage peak dataset on hg19 and mm9 and lifted these to the hg38 and mm10 reference using UCSC's *liftOver* tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). The isoform classification was extracted from the *SQANTI2* result and plotted in Fig. 2B. The *gffcompare* [58] (0.11.2) program was used with parameter "-T -R -M" to compare isoform annotations generated from different samples (Fig. 2B). It was also used to compare isoform annotations obtained after downsampling, which was achieved by randomly subsampling the long reads in the bam file (20%, 40%, 60%, and 80%) and re-running the pipeline with the same parameters (Additional file 2: Fig. S7A-B). Results from these comparisons were plotted using *UpsetR* [59] (1.4.0). We ranked transcript abundance for each gene that had multiple isoforms and obtained the alternative splicing events from the most expressed transcript and the second most expressed transcript. We used a common model to classify the splicing events [60], where alternative 5/3′ splice site includes alternative promoter and alternative polyadenylation. Transcripts with more than one splicing event were classified as *complex splicing changes* (Fig. 2G).

Differential transcript usage analysis

We filtered genes to have at least two isoforms, each with more than 15 UMI counts. For each gene, the per cell transcript counts were merged by group to generate pseudo-bulk samples. For *scmixology*, the groups are based on cell line identity inferred by known genetic variation, and for the MuSC and *CLL2* data, the groups are based on the clusters acquired from *Seurat* clustering, shown in Additional file 2: Fig. S1. The top 2 highly expressed transcripts for each group were selected and a UMI count matrix where the rows are selected transcripts and columns are groups was used as input to a chi-square test of independence (*chisq.test* in R). *P*-values were adjusted by Benjamini–Hochberg correction [61] and results were summarized in Additional file 4. We performed functional clustering with DAVID [62] using genes with significant DTU (adjusted *P*-value < 0.01) as input (Additional file 5). The transcript structures in all figures were plotted using *geom_alignment* in *ggbio* [63] (1.36.0). We performed imputation of transcript counts on cells that were not sampled for long-read sequencing using the shared nearest neighbor network constructed by *Seurat* [64] (3.1.5). We then scaled the transcript expression matrix for each gene and the results were used in heatmaps and UMAP visualizations (available in *Seurat*) in Fig. 3 and Additional file 2: Fig. S8. The imputed results were used for visualization purposes only.

Variant calling and clonal analysis

First, we identified candidate SNVs using *FLAMES* by excluding homopolymer regions (runs >3 of the same nucleotides), positions with coverage of fewer than 100 reads, and

positions with reference allele frequency less than 10% or greater than 90%. For each candidate SNV, we generated an allele count matrix of the reference and alternative allele. Next, we collect the cells that have reads with the reference allele and the alternative allele and performed a binomial test on the allele counts, assuming that under the null hypothesis each cell has the same probability of having the alternative and reference allele as their allele frequency in the whole sample. *P*-values were adjusted by Benjamini–Hochberg correction and candidate SNVs with an adjusted *P*-value < 0.05 were kept for further analysis. PCA was applied to the filtered alternative allele count matrix from the *scmixology* data and first two PCs were plotted in Fig. 4B. The top 5 PCs were selected to build the shared neighbor network (*scran::buildSNNGraph* [65] $k = 20$, $d = 5$) which is used for Louvain clustering (*igraph::cluster_louvain* [66]) in order to examine whether the allele count matrix captured the variants in the cell lines. Next, we performed differential allele frequency analysis in a similar way to the DTU analysis, but with the allele counts aggregated for each cluster. The candidate SNVs with adjusted *P*-values < 0.05 were selected and compared to the VCF reference generated from a previous study using bulk RNA-seq. After analyzing the *scmixology* dataset, we processed the *CLL2* data in the same way and conducted differential allele frequency analysis on the *CLL2* clusters shown in Additional file 2: Fig. S1. The bar plot of cells with alternative alleles of significant SNVs is shown in Additional file 2: Fig S9B and summarized in Fig. 4D. The capture enrichment data was analyzed in a similar way, with the reads processed by *FLAMES* and an allele count matrix generated for the Gly101Val mutation. All plots were generated using *ggplot2* unless otherwise specified and most of the analysis was performed in *R* [67] (4.0) unless otherwise specified.

Implementation of the *FLAMES* software

FLAMES is available as both a Python (https://github.com/LuyiTian/FLAMES) and R/Bioconductor package (https://bioconductor.org/packages/FLAMES) (0.99.31). The R package provides functionality of the original Python package using the *basilisk* software (1.2.1) (https://bioconductor.org/packages/release/bioc/html/basilisk.html) to create and manage a "frozen" conda environment, allowing all required Python packages to be downloaded from R. This allows for specific package versions to be accessed, without relying on the user to self-manage the required Python modules. Using this *basilisk* environment, the *reticulate* package (1.18) (https://cran.r-project.org/web/packages/reticulate/index.html) was used to call the *FLAMES* Python functions directly from R. The *FLAMES* package (Additional file 2: Fig S2) thus provides a series of wrapper functions for these reticulate Python calls. All basic data type conversion between R and Python, such as R named list to Python dictionary, is handled by *reticulate*; however, a few more complex data types required manual conversion.

The Python *FLAMES* package provided two pipelines, one for single-cell data and one for bulk long reads which have been re-written in R. Each function called from either pipeline is supplied as an R exported function to facilitate user configuration changes if wishing to manually run the pipeline, without the automated pipeline functions. To assist with interfacing with other Bioconductor packages, *FLAMES* stores the count and annotation data in either a *SingleCellExperiment* (1.12.0) object (https://bioconductor.org/packages/release/bioc/html/SingleCellExperiment.html) or *SummarizedExperiment* (1.20.0) (http://bioconductor.org/

packages/release/bioc/html/SummarizedExperiment.html) for the single-cell and bulk pipeline, respectively, as well as providing a number of additional output files. The package includes vignettes for running the *FLAMES* pipeline without the read alignment and realignment steps, which are platform dependent.

Bulk RNA-seq and exome data

Bulk data from the 5 Human Cell lines used in the *scmixology* experiments were downloaded from GEO accession number GSE86337.

PBMCs from *CLL2* were thawed and stained with CD19-BV510 antibody (BD Biosciences) at 4 ° for 30 min. PBMCs were washed stained with propidium iodide (PI, Sigma). Viable cells (PI negative) were flow sorted into a CD19+ tumor sample using the FACSAria (BD). RNA was extracted using an RNeasy Mini Kit (Qiagen #74104) according to the manufacturer's instructions. Library preparation and sequencing were performed, using the Truseq Stranded Total RNA library prep kit (Illumina) and 100-bp single-end sequencing protocol.

The data was aligned to the human genome (hg38) using *STAR* (2.7.3) with default parameters. Junctions with more than 5 uniquely mapping reads were retained and classified as known or novel based on the Gencode annotation. These were compared to the unique junctions obtained from *FLAMES* for isoforms with at least 1 read in the relevant single-cell sample (*scmixology1* and *CLL2* with non-tumor cells removed) that were classified (as known and novel) and extracted using *SQANTI3*. DNA was extracted from the *CLL2* PBMCs using the QIAamp DNA Mini Kit (Qiagen # 51304) according to the manufacturer's instructions. Library preparation and sequencing were performed at the Australian Genome Research Facility (AGRF), using the TruSeq Nano kit (Illumina), SureSelectXT2 Target Enrichment System, and Human All Exon v6 Capture Library (Agilent Technologies) and 150-bp paired-end sequencing protocol. The bulk sequencing data was aligned with *BWA mem* [68] (0.7.17) (exome) and *STAR* [69] (2.7.3) (RNA) and variants were called by *superFreq* [70] (1.4.3). Select variants detected in the scRNA-seq analysis were visually inspected.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02525-6.

Additional file 1. Summary of the datasets generated in this study.

Additional file 2. Supplementary Figures.

Additional file 3. Summary of SIRV isoform detection results (false negatives, false positives and true positives) for the 4 methods compared (*FLAMES*, *FLAIR*, *TALON* and *StringTie2)*.

Additional file 4. Table of genes with differential transcript usage (gDTU) in each sample.

Additional file 5. Table of functional clustering results from DAVID for gDTU in each sample.

Additional file 6. Review history.

Acknowledgements

We thank Clare Weeden and Marie-Liesse Asselin-Labat for providing the cell lines used in this study, Stephen Wilcox for Illumina sequencing, Kathleen Zeglinski for designing the *FLAMES* logo and creating Fig. 1A, and Joan Heath for providing critical feedback on this manuscript.

Review history

The review history is available as Additional file 6.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

L.T. developed the FLAMES package, performed experiments, conducted data analysis, and wrote the manuscript. J.S.J. developed the modified 10x protocol and performed experiments. R.T., M.A.A, H.P., and C.J.A.A collected patient samples and performed experiments related to *CLL2*. C.H.L., J.D.C., T.N., and A.C. performed experiments related to *MuSC*. Y.D.J.P. and C.C.F performed experiments related to the *scmixology* data. Q.G., H.P., and A.L performed nanopore sequencing. S.L.A., H.K., M.R.M.D., X.D., C.W.L., C.F., and I.M. analyzed data, and O.V, S.S., J.S, and C.W. developed software. G.S.L, J.G.R., A.W.R., and D.C.S.H. facilitated the *MuSC* and *CLL2* studies and provided helpful comments and discussion. J.S.J, Q.G, and M.B.C. planned the research, provided helpful comments and discussion, and wrote the manuscript. M.E.R. planned and supervised the research and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by funding from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (grant numbers 2018-182819 and 2019-002443 to M.E.R.), a Melbourne Research Scholarship to L.T, a Leukemia & Lymphoma Society (LLS) of America Fellowship to R.T. (5467-18) and LLS Specialized Center of Research [SCOR] grant no. 7015-18 to A.W.R. and D.C.S.H., a Cure Cancer and Cancer Australia grant (1186003 to R.T.), fellowships and grants from the Australian National Health and Medical Research Council (NHMRC: Project Grant 1143163 to M.E.R.; Program Grants 1016701 to D.C.S.H. and 1113577 to A.W.R.; Research Fellowships 1079560 to A.W.R., 1156024 to D.C.S.H., 1072662 to M.B.C. and 1104924 to M.E.R.; Investigator Grant 1174902 to A.W.R.), the Australian Research Council (Discovery Project 200102903 to M.E.R.), the Genomics Innovation Hub, Victorian State Government Operational Infrastructure Support, and Australian Government NHMRC IRIISS.

Availability of data and materials

Raw data are available from GEO under accession numbers GSE126906 [55] and GSE154869 (*scmixology1*) [71], GSE154870 [72] and GSE142285 (*scmixology2*) [73], and GSE154868 (*MuSC*) [74], and the patient data (*CLL2*) is available from EGA under accession number EGAS00001005597 [75]. Refer to Additional file 1 for a summary of these datasets. The *FLAMES* [76] source code is available from https://github.com/LuyiTian/FLAMES (Python version) and https://bioconductor.org/packages/FLAMES [77] (R version) under the GNU General Public License (≥ 2). The processed data and scripts used in this study are available from https://github.com/LuyiTian/FLTseq_data [78].

Declarations

Ethics approval and consent to participate

Animal experiments were approved by the Animal Ethics Committee of The University of Melbourne and conformed to the Australian code of practice for the care and use of animals for scientific purposes as stipulated by the National Health and Medical Research Council of Australia.

For the CLL data, informed consent was obtained for sample collection and participation in research. The research was approved by human research ethics committees at Melbourne Health (2011.044, 2016.305, 2012.274), Peter MacCallum Cancer Centre (11/18), and the Walter and Eliza Hall Institute (05/04, 13/01). All experimental methods comply with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia. [2]Department of Medical Biology, The University of Melbourne, Parkville, VIC, Australia. [3]Australian Genome Research Facility, Victorian Comprehensive Cancer Centre, Melbourne, VIC, Australia. [4]Centre for Stem Cell Systems, Department of Anatomy and Neuroscience, The University of Melbourne, Parkville, VIC, Australia. [5]The Florey Institute of Neuroscience and Mental Health, Parkville, VIC, Australia. [6]Centre for Muscle Research, Department of Physiology, The University of Melbourne, Melbourne, VIC, Australia. [7]Present address: Department of Neurology, Stanford University, Stanford, CA, USA. [8]Present address: VOW, North Parramatta, NSW, Australia. [9]Clinical Haematology, Peter MacCallum Cancer Centre and Royal Melbourne Hospital, Melbourne, VIC, Australia. [10]Centre for Cancer Research, University of Melbourne, Melbourne, VIC, Australia. [11]Victorian Comprehensive Cancer Centre, Melbourne, VIC, Australia.

Received: 30 April 2021 Accepted: 21 October 2021
Published online: 11 November 2021

References

1. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581(7808):303–9. Available from: https://doi.org/10.1038/s41586-020-2157-4.
2. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013;498:236–40. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3683364&tool=pmcentrez&rendertype=abstract.
3. Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, et al. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. Mol Cell. 2017;67:148–61.e5. Available from: https://doi.org/10.1016/j.molcel.2017.06.003.

4. Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks G-J, Larsson AJM, et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat Biotechnol. 2020;38(6):708–14. Available from: http://www.nature.com/articles/s41587-020-0497-0. https://doi.org/10.1038/s41587-020-0497-0.
5. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. Available from: http://www.nature.com/articles/ncomms14049.
6. Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat Biotechnol. 2018;36:1197–202. Available from: https://doi.org/10.1101/364950%0A. https://www.biorxiv.org/content/early/2018/07/08/364950.
7. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. 2019;10:3120. Available from: https://www.biorxiv.org/content/early/2018/09/24/424945.full.pdf+html%0A. https://www.biorxiv.org/content/biorxiv/suppl/2018/09/24/424945.DC1/424945-1.pdf.
8. Volden R, Vollmers C. Highly multiplexed single-cell full-length cDNA sequencing of human immune cells with 10X Genomics and R2C2. bioRxiv. 2021. Available from: https://doi.org/10.1101/2020.01.10.902361.
9. Lebrigand K, Magnone V, Barbry P, Waldmann R. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat Commun. 2020;11:4025. Available from: https://doi.org/10.1038/s41467-020-17800-6.
10. Wang Q, Bönigk S, Böhm V, Gehring N, Altmüller J, Dieterich C. Single cell transcriptome sequencing on the Nanopore platform with ScNapBar. RNA. 2021;27(7):763–70. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8208055/.
11. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11(1):1438. Available from: https://doi.org/10.1038/s41467-020-15171-6.
12. Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Zeng W, et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv. 2019. Available from: https://doi.org/10.1101/672931.
13. Tian L, Dong X, Freytag S, Lê Cao K-A, Su S, JalalAbadi A, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods. 2019;16(6):479–87. Available from: http://www.nature.com/articles/s41592-019-0425-8. https://doi.org/10.1038/s41592-019-0425-8.
14. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15(3):201–6. Available from: https://doi.org/10.1038/nmeth.4577.
15. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20(1):278. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1910-1.
16. Dong X, Tian L, Gouil Q, Kariyawasam H, Su S, De Paoli-Iseppi R, et al. The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools. NAR Genom Bioinform. 2021;3(2):lqab028. Available from: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqab028/6253498.
17. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun. 2019;10:3359. Available from: http://www.nature.com/articles/s41467-019-11272-z. https://doi.org/10.1038/s41467-019-11272-z.
18. Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Lassmann T, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24670764.
19. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 2018:28(3):396–411. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29440222%0A. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5848618.
20. Yue L, Wan R, Luan S, Zeng W, Cheung TH. Dek modulates global intron retention during muscle stem cells quiescence exit. Dev Cell. 2020;53:661–76.e6. Available from: https://doi.org/10.1016/j.devcel.2020.05.006.
21. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013;14(7):R70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23815980.
22. Faigenbloom L, Rubinstein ND, Kloog Y, Mayrose I, Pupko T, Stein R. Regulation of alternative splicing at the single-cell level. Mol Syst Biol. 2015;11:845. Available from: https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20156278.
23. Manley JL, Krainer AR. A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev. 2010;24(11):1073–4. Available from: http://genesdev.cshlp.org/cgi/doi/10.1101/gad.1934910.
24. Gazda HT, Grabowska A, Merida-Long LB, Latawiec E, Schneider HE, Lipton JM, et al. Ribosomal protein S24 gene is mutated in Diamond-Blackfan anemia. Am J Hum Genet. 2006;79:1110–8 Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929707634740.
25. Zhang H, Ye J, Weng X, Liu F, He L, Zhou D, et al. Comparative transcriptome analysis reveals that the extracellular matrix receptor interaction contributes to the venous metastases of hepatocellular carcinoma. Cancer Genet. 2015;208:482–91. Available from: https://doi.org/10.1016/j.cancergen.2015.06.002.
26. Dawes R, Petrova S, Liu Z, Wraith D, Beverley PCL, Tchilian EZ. Combinations of CD45 isoforms are crucial for immune function and disease. J Immunol. 2006;176(6):3417–25. Available from: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.176.6.3417.
27. Ray TA, Cochran K, Kozlowski C, Wang J, Alexander G, Cady MA, et al. Comprehensive identification of mRNA isoforms reveals the diversity of neural cell-surface molecules with roles in retinal development and disease. Nat Commun. 2020;11:3328. Available from: https://doi.org/10.1038/s41467-020-17009-7.
28. Chen C, Zhao S, Karnad A, Freeman JW. The biology and role of CD44 in cancer progression: therapeutic implications. J Hematol Oncol. 2018;11:1–23.
29. Demircioğlu D, Cukuroglu E, Kindermans M, Nandi T, Calabrese C, Fonseca NA, et al. A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. Cell. 2019;178(6):1465–77.e17. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419309067.

30. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.e21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598.
31. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19:90. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1462-9.
32. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017;8:1326. Available from: https://doi.org/10.1038/s41467-017-01343-4.
33. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21:30. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1935-5.
34. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018;36:89–94 Available from: http://www.nature.com/doifinder/10.1038/nbt.4042.
35. Blombery P, Anderson MA, Gong J, Thijssen R, Birkinshaw RW, Thompson ER, et al. Acquisition of the recurrent Gly101Val mutation in BCL2 confers resistance to venetoclax in patients with progressive chronic lymphocytic leukemia. Cancer Discov. 2019;9(3):342–53. Available from: http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-18-1119.
36. Lebrigand K, Bergenstråhle J, Thrane K, Mollbrink A, Barbry P. The spatial landscape of gene expression isoforms in tissue sections. bioRxiv. 2020. Available from: https://www.biorxiv.org/content/10.1101/2020.08.24.252296v1.
37. Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, et al. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proc Natl Acad Sci. 2018;115:9726–31 Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1806447115.
38. Davidson NM, Chen Y, Ryland GL, Blombery P, Göke J, Oshlack A. JAFFAL: detecting fusion genes with long read transcriptome sequencing. bioRxiv. 2021. Available from: https://www.biorxiv.org/content/10.1101/2021.04.26.441398v1.
39. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020; Springer US; Available from:. https://doi.org/10.1038/s41576-020-0258-4.
40. Westoby J, Herrera MS, Ferguson-Smith AC, Hemberg M. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. Genome Biol. 2018;19(1):191. Available from: https://www.biorxiv.org/content/early/2018/01/17/248716. https://doi.org/10.1186/s13059-018-1571-5.
41. Nam AS, Kim K-T, Chaligne R, Izzo F, Ang C, Taylor J, et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. Nature. 2019;571(7765):355–60. Available from: http://www.nature.com/articles/s41586-019-1367-0. https://doi.org/10.1038/s41586-019-1367-0.
42. Melamud E, Moult J. Stochastic noise in splicing machinery. Nucleic Acids Res. 2009;37(14):4873–86. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp471.
43. Murphy MM, Lawson JA, Mathew SJ, Hutcheson DA, Kardon G. Satellite cells, connective tissue fibroblasts and their interactions are crucial for muscle regeneration. Development. 2011;138(17):3625–37. Available from: http://dev.biologists.org/cgi/doi/10.1242/dev.064162.
44. Srinivas S, Watanabe T, Lin CS, William CM, Tanabe Y, Jessell TM, et al. Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. BMC Dev Biol. 2001;1(1):1–8. https://doi.org/10.1186/1471-213X-1-4.
45. Liu L, Cheung TH, Charville GW, Rando TA. Isolation of skeletal muscle stem cells by fluorescence-activated cell sorting. Nat Protoc. 2015;10(10):1612–24. Available from: http://www.nature.com/articles/nprot.2015.110. https://doi.org/10.1038/nprot.2015.110.
46. Roberts AW, Davids MS, Pagel JM, Kahl BS, Puvvada SD, Gerecitano JF, et al. Targeting BCL2 with venetoclax in relapsed chronic lymphocytic leukemia. N Engl J Med. 2016;374:311–22. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1513257.
47. Jabbari JS, Tian L. Massively parallel long-read sequencing of single cell RNA isoforms. protocols.io; 2019.
48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352.
49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. Available from:https://doi.org/10.1093/bioinformatics/btq033.
50. Holik AZ, Law CW, Liu R, Wang Z, Wang W, Ahn J, et al. RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. Nucleic Acids Res. 2017;45(5):e30. Available from: https://doi.org/10.1093/nar/gkw1063.
51. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. Available from: https://doi.org/10.1093/bioinformatics/btt656.
52. Lee S, Cook D, Lawrence M. plyranges: a grammar of genomic data transformation. Genome Biol. 2019;20(1):4. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1597-8.
53. Lee S, Zhang AY, Su S, Ng AP, Holik AZ, Asselin-Labat M-L, et al. Covering all your bases: incorporating intron signal from RNA-seq data. NAR Genom Bioinform. 2020;2(3):lqaa073. Available from: https://academic.oup.com/nargab/article/2/3/lqaa073/5910008.
54. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):296. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1.
55. Tian L, Su S, Dong X, Amann-Zalcenstein D, Biben C, Seidi A, et al. scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLOS Comput Biol. 2018;14:e1006361. Available from: http://www.biorxiv.org/content/early/2017/08/14/175927.
56. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics 2018;34:3094–3100. Available from: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778

57. van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng. 2011;13(2):22–30. Available from: http://ieeexplore.ieee.org/document/5725236/. https://doi.org/10.1109/MCSE.2011.37.
58. Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. F1000Research. 2020;9:304. Available from: https://f1000research.com/articles/9-304/v1.
59. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33:2938–40. Available from: https://academic.oup.com/bioinformatics/article/33/18/2938/3884387
60. Frankiw L, Baltimore D, Li G. Alternative mRNA splicing in cancer immunotherapy. Nat Rev Immunol. 2019;19(11):675–87. Available from: http://www.nature.com/articles/s41577-019-0195-7.
61. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300. Available from: http://www.jstor.org/stable/2346101.
62. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. Available from: http://www.nature.com/articles/nprot.2008.211. https://doi.org/10.1038/nprot.2008.211.
63. Yin T, Cook D, Lawrence M. ggbio: an R package for extending the grammar of graphics for genomic data. Genome Biol. 2012;13:R77. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-8-r77.
64. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502. Available from: http://www.nature.com/doifinder/10.1038/nbt.3192.
65. AT LL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17(1):75. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0947-7.
66. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal 2006;Complex Sy:1695. Available from: http://igraph.sf.net
67. R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, Austria; 2019. Available from: https://www.r-project.org/
68. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;00:2013.
69. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635.
70. Flensburg C, Oshlack A, Majewski IJ. Detecting copy number alterations in RNA-Seq using SuperFreq. Bioinformatics. 2021;btab440. Available from: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btab440/6300509.
71. Tian L, Jabbari JS, Ritchie ME. Long and short-read single cell RNA-seq profiling of human lung adenocarcinoma cell lines using 10X version 2 chemistry. [Internet]. Available from: https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/gds/?term=GSE154869
72. Tian L, Gouil Q, Jabbari JS, Ritchie ME. Long and short-read single cell RNA-seq profiling of human lung adenocarcinoma cell lines using 10X V3 chemistry. [Internet]. Available from: https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/gds/?term=GSE154870
73. Amarasinghe SL, Tian L, Jabbari JS, Ritchie ME. Designing a single cell ATAC-Seq dataset to validate long read RNA-Seq isoforms and benchmark scATAC-Seq analysis methods. Available from: https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/gds/?term=GSE142285
74. Ly CH, Tian L, Jabbari JS, Ritchie ME, Lynch GS, Ryall JG. Long and short-read single cell RNA-seq profiling of mouse muscle stem cells. Available from: https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/gds/?term=GSE154868
75. Tian L, Peng H, Thijssen R, Ritchie ME, Huang DCS, Roberts AW. Comprehensive characterization of single cell full-length isoforms in human and mouse with long-read sequencing. Available from: https://ega-archive.org/studies/EGAS00001005597
76. Tian L. FLAMES package [Internet]. 2021. Available from: https://github.com/LuyiTian/FLAMES
77. Tian L, Voogd O, Schuster J, Wang C, Su S, Ritchie ME. Full-length analysis of mutations and splicing in long-read RNA-seq data, version 0.99.31 [Internet]. Available from: https://bioconductor.org/packages/FLAMES
78. Luyi Tian, Ritchie ME. Data and Scripts associated with the FLAMES paper [Internet]. Available from: https://github.com/LuyiTian/FLTseq_data

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

