TITLE

Single cell characterization of CRISPR-modified transcript isoforms with nanopore sequencing

Authors:

Heon Seok Kim¹, Susan M. Grimes¹, Anna C. Hooker¹, Billy T. Lau¹, Hanlee P. Ji¹

Institutions:

¹ Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

Corresponding author:

Hanlee P. Ji

Email: genomics_ji@stanford.edu

CCSR 1115, 269 Campus Drive, Stanford, CA-94305, USA

ABSTRACT

Transcript isoforms are mRNAs that arise from alternative splicing events. During RNA processing, different combinations of a gene's exons lead to a diverse set of isoforms. Polymorphisms or mutations at splice junctions can generate alternative splicing events. Various splicing factors also impact the representation of a gene's transcript isoforms. To assess how these two features contribute to alternative splicing, we developed a single cell approach to introduce CRISPR edits that modify mRNA transcript structure. Our method combines (1) long-read sequencing to characterize the expressed transcripts and identify the edit at single cell resolution; (2) short-read sequencing to match the single cell gene expression profiles of the cells with the altered isoform. First, we modify target exon-intron segments with CRISPR-Cas9. Second, using cDNAs with cell barcodes, we use long read sequencing to directly identify the changes in transcript isoforms from the targeted CRISPR edits. As a variation on this approach, we also determined how modifying specific splicing factors influence isoform expression and structure. Overall, we demonstrate how the integration of single cell long read analysis and CRISPR engineering can be used to directly confirm transcript isoform and target genomic edits at single cell resolution. This approach will improve our understanding of the role of alternative splicing in transcriptional regulation.

INTRODUCTION

For any gene's messenger RNA (**mRNA**), alternative splicing events lead to a diverse set of transcripts that have different combinations of exons. Also referred to as transcript isoforms, these different mRNA species result from genetic variants at intron-exon junctions and the functional contribution of different splicing factors. This diversity of transcripts provides cells with an added level of transcriptional regulation. Frequently, different isoforms have significantly different functions depending on the cell and tissue type. To date, most studies examining the structure and function of isoform variation have relied on CRISPR or other genome engineering methods. These experiments involve introducing a single genetic variant within an exon/intron junction or knocking out a splicing factor. Subsequently, the impact on transcript structure is assessed using conventional short read RNA-seq – this bulk sequencing method averages the changes in transcript expression across many cells.^{1, 2} Importantly, conventional short read approaches may not resolve important transcript isoform features that are present in only a subset of cells and require complex bioinformatic methods to resolve transcript structures. Thus, it remains a challenge to determine how these factors influence the generation of different mRNA species.

Single-cell RNA-sequencing (**scRNA-seq**) has been used for profiling transcript isoforms across individual cells.^{3, 4} Some studies have adapted long-read sequencing for single-cell transcript isoform analysis.^{5, 6} These approaches rely on instruments such as the Oxford Nanopore or Pacific Biosciences sequencers that generate long reads, typically over hundreds of bases if not much longer. Intact cDNAs are generated from single cells and used for library preparation without the fragmentation step required for short read analysis. Sequencing the intact cDNA across the length of a single read can identify transcript structure and expressed genetic variants. However this method can only assess the transcripts existing in the sequenced

sample, without consideration of possible isoform alterations such as exon-intron junction modifications or splicing factor knock-outs.

Figure 1. (a) Overview of single-cell short/long-read integration strategy. (b) Structure of *RACK1* transcript. The red arrows point to CRISPR target sites that disrupt splicing acceptor sequences. (c, d) Heatmaps showing the proportion of each transcript isoform (x-axis) with each cell (y-axis) and clustering based on transcript isoform proportion for *RACK1* exon5 treated cells (c) and multiple exon treated cells (d). (e) Expression level of *RACK1* in each cell cluster in (d). P values were calculated in comparison with cluster 3 ($P = 5.7 \times 10^{-12}$, 2.6×10^{-11} , 9.9×10^{-11} , 1.6×10^{-10} , 1.1×10^{-07} , 9.5×10^{-06} , 2.5×10^{-07} , 1.1×10^{-08} ; two-sided t-test).

Several studies have integrated single cell RNA-seq with CRISPR screening to evaluate the transcriptional phenotypes of cellular networks.⁷⁻¹⁰ These single-cell CRISPR studies require sequencing an individual cell's expressed guide RNA; this barcode functions as an indirect indicator of which cells have a CRISPR edit. However, these methods do not directly verify the genome edit for a given cell. In addition, those approaches rely on short-read sequencing of transcripts where only a small portion of the ends of the mRNA are available. As a result, short read sequencing methods for single cell analysis do not identify changes in transcript structure nor the quantitative expression of individual isoforms.

Herein, we present a new approach which characterizes CRISPR-induced alterations of full-length transcript isoforms from single cells. We design CRISPR edits targeting specific exon-intron junctions that potentially alter transcript structure. As another experimental manipulation of transcript isoforms, we use CRISPR to alter specific splicing factor genes and assess how these changes influence a downstream gene's expression of transcript isoforms. Our method relies on single cell nanopore sequencing – an individual read has sufficient length to cover the entire transcript. We use this added information to extrapolate the transcript structure. Subsequently, we match each individual cell's long read information with its corresponding short read transcriptome information to identify general transcriptional profile changes (**Figure 1A**). Likewise, we show that CRISPR edits introducing short insertion deletions (**indels**) are detected within sequences distal from the cDNA's 5' or 3' end that could not be detected with short reads. This allows for validation of CRISPR edits even if they do not affect isoform sequence.

As a proof-of-concept, we designed and conducted an experiment targeting the splicing acceptor of exon 5 within the *RACK1* gene (**Figure 1B**). This gene encodes a receptor for activated C kinase 1. It plays a role in intracellular protein shuffling and anchoring. This gene was chosen because it is one the most highly expressed genes in the human embryonic kidney 293T (**HEK293T**) cell line as determined from single-cell short read gene expression data. In addition, it has eight exons to target with CRISPR. We selected a sequence for this target and generated a plasmid construct expressing Cas9 and target gRNA. These plasmids were transfected into HEK293T cells. After six days, we prepared single cell cDNAs from both the wild-type and transfected cells with the gRNAs (**Methods**). From the single-cell cDNAs, the *RACK1* transcript was amplified using two primers that included sequence from the 5' adaptor and the last exon of *RACK1*. The amplified target cDNA underwent long read sequencing

(Oxford Nanopore). We performed base calling using the program guppy and aligned the long reads to the reference genome, GrCh38 using minimap2.^{11, 12}

Utilizing the cell barcodes at the 5' end of the cDNA enables assignment of a long read covering a transcript to its cell of origin. Since these barcodes will not align to the genome, they should fall into the soft-clipped region of the aligned read. Accordingly, the soft-clipped portion of each read was extracted and compared to the whitelist of cell barcodes identified from the short-read single-cell sequencing. The comparison was performed using vectorization (with token length of 8) and a cosine similarity metric as described for assessment of text document similarity and referred to as tf-idf cosine.¹³ The cosine-similarity score evaluates the frequency of 8-mer occurrence between the whitelisted barcodes and the soft-clipped sequence, with higher scores indicating higher correlation. The edit distance between each of the 5 top scoring barcodes and each 16bp region of the soft-clipped sequence was calculated with the barcode having the minimum edit distance considered a valid match if the edit distance was less than 3.

For the *RACK1* amplicon from HEK293T cells the mean edit distance of selected barcodes was 0.51. The cell barcode's position within the soft-clipped portion of the read was an average of 0.54 bases from the expected position based on the known single cell adapter/barcode construct. Overall, the data integration between single cell long and short read data enabled us to assess those cells with an altered target isoform and their accompanying full transcriptome features.

Using long read sequencing of target cDNAs, we characterized the transcript isoform diversity in each individual cell. Prior to this quantification step the 10bp UMI sequence immediately

following each cell barcode was reviewed. If a set of reads had the same barcode and UMI, these duplicate sequences were discarded. Reads per cell barcode and their associated isoforms were then aggregated and hierarchical clustering was used to determine the distribution of different isoforms among cell subpopulations. Among the wildtype cells, we determined that nearly all of the cells expressed the full length *RACK1* isoform. Specifically, 5030 out of 5074 cells had more than 90% of isoforms with full length *RACK1* transcripts. However, a small subset of wildtype cells (Cluster 2), 44 out of 5074 cells demonstrated transcript isoform heterogeneity, averaging 23% of isoforms skipping one or more exons (**Supplementary Figure 1**). This result demonstrated how long-read analysis combined with scRNA-seq characterizes the underlying transcript isoform diversity present within individual cells.

For the cells transfected with CRISPR, 6028 of the 7548 cells had more than 33% of their *RACK1* isoforms lacking exon 5. This result confirmed the genome edit (**Figure 1C**). We performed hierarchical clustering based on the expression levels of the different *RACK1* transcript isoforms present in each cell. A subset of cells (Cluster 1, 4732 out of 7548 cells) contained CRISPR edits in the splicing acceptor in all three alleles of *RACK1*. A subpopulation of Cluster 2 cells (1234 out of 7548 cells) had predominantly full-length transcripts (82% on average), indicating a lack of CRISPR editing. One factor contributing to this subpopulation not being completely homogeneous may be the introduction of ambient RNA from the highly expressed *RACK1* gene being incorporated into the cell droplets and causing some cross-contamination.

Given the successful proof-of-concept, we designed a multiplexed CRISPR assay that targeted multiple exons in *RACK1*. We selected gRNAs that changed the sequence of splicing acceptor

sites of *RACK1* exons 2-7. We then transfected a pooled plasmid library containing these gRNAs into HEK293T cells with stable Cas9 expression (**Methods**). As previously described, single cell cDNA libraries were generated from these CRISPR-transduced cells. Then, we used long read sequencing of *RACK1* cDNA to determine isoform representation. Our results indicated that the multiplexed CRISPR gRNAs generated edits leading to different exon-skipping events (**Figure 1D**). The cells in Cluster 3 predominantly expressed full length *RACK1* transcripts (83% full length transcripts per cell on average) thus indicating that this subpopulation had not undergone genome editing. Other cells had reads indicating multiplexed CRISPR editing of splicing acceptors. For example, cells in Cluster 1 showed evidence of genome editing in exons 2 to 6 with the disruption of multiple splicing acceptor sites or large deletions (83% exon2 to 6 skipped transcript per cell on average).

We studied the downstream transcriptional changes for the CRISPR-transfected cells with different *RACK1* isoforms. For this analysis, we integrated the single cell data from the long-read sequencing of *RACK1* cDNA with short-read sequencing of the matching single cell transcriptomes by matching cell barcodes as described earlier. Then, we compared short-read gene expression profiles between the *RACK1* isoform clusters for cells which had undergone multiplexed CRISPR editing (**Figure 1D**). First using the short-read single cell data, we compared the *RACK1* expression levels between the non-edited cells versus edited cells. As expected, *RACK1* expression among the non-edited cells (Cluster 3) was significantly higher than CRISPR-edited cells (6.59 fold change, $P = 1.85e-9$) (**Figure 1E**). The lower *RACK1* expression among CRISPR-edited cells was likely the result of mis-splicing and non-sense mediated decay (**NMD**). Next, we compared *RACK1* expression from each cluster of cells having *RACK1* exon 5 CRISPR editing. The cells without edits (Cluster 2) had 1.53 fold higher level of *RACK1* expression ($P = 1.04e-194$) than the cells in other clusters (**Figure 1C and**

Supplementary Figure 2). These experiments demonstrated the characterization of CRISPR transcript isoform changes induced by CRISPR and their transcriptional level at single cell resolution.

Different splicing factor genes impact the generation of alternative splicing events. We leveraged this aspect of isoform regulation to demonstrate another application of our approach. Namely, we determined how different splicing factors affect alternative splicing events of a downstream target gene. This mechanism of isoform generation is particularly crucial for immune cells where changing the regulator state and increased functional flexibility is paramount. For this experiment, we chose to study the *PTPRC* gene. Expressed in T cells, *PTPRC* is a transmembrane phosphatase and its pre-mRNA alternative splicing is critical for changing T cell regulatory states (**Supplementary Figure 3**).¹⁴ *PTPRC* has five highly expressed isoforms. This includes two shorter ones where there is substantial degree of exon loss and several longer isoforms where the majority of exons from the variable region are retained. For example, CD4-CD8- (double negative) T cells and NK precursor cells preferentially express longer isoforms like RABC and RBC and when activated. In addition, T cells and NK cells preferentially express shorter isoforms like RO and RB.

To study how various splicing factors impact the generation of *PTPRC* isoforms, we selected a series of gRNAs targeting a set of 16 splicing factor genes (**Supplementary Table 3**). We chose splicing factors which are expressed and play a critical role pre-mRNA processing in Jurkat cells.^{1, 15, 16} We integrated this multiplexed CRISPR assay with single cell long-read sequencing and determined how each splicing factor contributed to changes in *PTPRC*'s transcript isoform structure. For these experiments, we used the Jurkat human cell lines derived from a T cell leukemia. Our line stably expressed Cas9. We transduced a multiplexed

gRNA lentiviral library targeting 16 splicing factors (two gRNAs per gene) and five non-targeting gRNAs as negative controls (**Supplementary Table 3**). After 14 days, we harvested the cells, generated single cell libraries and conducted sequencing with both short and targeted long reads of *PTPRC*.

Figure 2. (a) Overview of single-cell CRISPR screen integrated with long-read sequencing. (b) Boxplot showing the ratio of short *PTPRC* transcript isoform (RO and RB) for cells with gRNAs targeting indicated genes. P values are calculated in comparison with Nontarget. Genes which have less than three cells with target gRNAs are not shown. (c) Heatmap showing proportion of each transcript isoform (x-axis) with each cell (y-axis) and clustering based on transcript isoform proportion for cells having indicated gRNA sequence.

First, we determined which gRNAs were expressed within a given individual cell (**Figure 2A**).

This method relies on using a primer to polymerase extend over the gRNA adjacent to a given cell barcode and then sequencing the product to obtain both sequences (**Supplementary Table**

4).¹⁷ With the paired gRNA and cell barcode sequence, one determines the distribution of

expressed gRNAs across individual cells. To assess the reliability of our long-read cell barcode matching process, we amplified our 16 targeted genes, conducted nanopore sequencing and then identified the CRISPR induced mutations at the gRNA target-site. In parallel, we conducted the single cell short-read sequencing, determine the gRNA for each cell and matched the short with the long read cell barcodes (**Supplementary Figure 4A**). Based on this comparative analysis, CRISPR mutations were introduced at the target site at a significantly increased rate in cells with the appropriate target gRNA versus cells versus those without the guide (**Supplementary Table 5**). For example, we confirmed that cells expressing the *SRSF5-1* gRNA had a variety of indels at the targeted Cas9 cleavage site in the transcript. This was in contrast to cells with *SRSF5-2* gRNA or other gRNAs which lacked indels in the target (**Supplementary Figure 4B**). Our quantitative analysis showed that the average mutation rate per cell with *SRSF5-1* gRNA was 64.32% compared to 12.73% of cells with gRNAs targeting other gene at *SRSF5-1* target site (**Supplementary Figure 4C**). The difference in mutation frequency was clearly evident and highly significant ($P = 5.85e-11$) despite the high rate of error in nanopore sequencing. Overall, this result indicated that targeted long read sequencing identified the CRISPR genotypes and that the short/long-read barcode matching was reliable.

Next, we used long read sequencing of full-length cDNA from individual cells to determine the structure of *PTPRC* transcripts. We clustered individual cells based on the expressed gRNA sequences targeting the different splicing factor genes. We determined the expression and relative ratios of the five most abundant *PTPRC* transcript isoforms for these cells using long-read sequencing. For each long-read, the isoform structure was identified by determining aligned bases within each exon. The cell barcode was identified using the sequence matching methodology described earlier. Among the different cells with the CRISPR gene targets versus the negative control, we compared the average ratio of short *PTPRC* isoform category and the

fold change differences (**Figure 2B and Supplementary Figure 5**). The knock-out of *HNRNPLL* and *SRSF5* reduced *PTPRC* exon skipping events, resulting in lower RO and RB abundance (2.01- and 1.18-fold change respectively). In comparison, the knock-out of *PCBP2* and *HNRNPD* increased exon skipping events resulting in higher RO and RB abundance (1.15- and 1.16-fold change respectively). Therefore, *HNRNPLL* and *SRSF5* induce exon skipping of *PTPRC*. In contrast, *PCBP2* and *HNRNPD* inhibited exon skipping. Although *HNRNPLL* and *SRSF5* knock-outs inhibited *PTPRC* exon skipping, their isoform expression patterns were significantly different (**Supplementary Figure 6 and 7**). The ratio of RBC was comparable between the two knock-outs but the ratio of RABC was much higher for the *HNRNPLL* knock-out (5.60-fold change, $P = 5.90e-37$). The consequences of *PCBP2* and *HNRNPD* knock-outs were nearly equivalent with respect to the expression of the RB isoform ($P = 0.1$). However the *PCBP2* knock-out had greater expression of the RO isoform (1.20-fold change, $P = 0.017$) (**Supplementary Figure 8**).

Overall, *HNRNPLL* had the most significant effect on *PTPRC* isoform regulation. We compared the cells expressing the nontarget-1 gRNA versus the cells expressing the *HNRNPLL*-1 gRNA (**Figure 2C**). In comparing the cells with nontarget-1 gRNA, the cells expressing the *HNRNPLL*-1 gRNA had a significantly different expression patterns and ratios of *PTPRC* isoforms. These cells had a relatively lower expression of RO and RB isoforms (2.45-fold, $P = 1.16e-32$).

To verify this result, we selected the *HNRNPLL* gene for an independent knock-out experiment (**Figure 3**). We used an electroporation approach to introduce the Cas9 ribonucleoprotein (RNP) into Jurkat cells (**Methods**). Six days after electroporation, we prepared single cell cDNA libraries from the wild-type Jurkat cells and KO pool cells. Subsequently, we performed single cell short-read sequencing to enumerate RNA expression and long-read sequencing to

determine the *PTPRC* isoforms per each cell. Similar to the previous single-cell CRISPR screen result, wild-type cells demonstrated abbreviated transcripts (i.e. RO, RB) while *HNRNPLL* RNP-treated cells expressed the longer transcript isoforms (i.e. RABC, RBC) (7.35-fold, $P < 1.0e-5$, **Figure 3B and 3C**). When comparing the RNP versus lentivirus based *HNRNPLL* gRNAs, we observed similar *PTPRC* transcript structures (**Figure 2C and 3C**).

Figure 3. (a) Overview of splicing factors affect alternative splicing. (b) Quantification of short transcript isoform per target gene. For each sample (x-axis) the ratio of transcript isoform RO and RB among all *PTPRC* isoforms for cells are shown as box plot. (c) Heatmap showing proportion of each transcript isoform (x-axis) with each cell (y-axis) and clustering based on transcript isoform proportion for each sample.

This experimental system had an additional functional readout to assess *PTPRC* isoform expression. Exposing Jurkat cells to phorbol 12-myristate 13-acetate (**PMA**) and ionomycin induces the expression of shorter isoforms of *PTPRC*.¹⁸ Taking advantage of this cellular property, we challenged both CRISPR-treated and wildtype cells with this small molecule combination then subsequently performed single-cell sequencing. Overall, our results showed

that PMA and ionomycin stimulation increased the differences in isoform expression between the wild-type cells and *HNRNPLL* RNP-treated cells. Most of the stimulated wild-type cells had RO and RB transcript isoforms. However, the stimulated *HNRNPLL* RNP-treated cells had less RO and RB transcript isoforms (10.32-fold, $P < 1.0e-5$, **Figure 3B and 3C**).

In addition to *PTPRC*, we analyzed the impact of splicing factors on myosin light chain 6 (*MYL6*) transcript isoforms. Exon6 skipping of *MYL6* is known to be regulated by various splicing factors (**Supplementary Figure 9A**).^{1, 19} We found that *CELF2* targeted cells had a higher percentage of full-length *MYL6* transcript isoforms compared to cells targeted by other splicing factors (1.44-fold, $P = 1.9e-14$, **Supplementary Figure 9B**), indicating that disrupting *CELF2* reduced the occurrence of exon-skipping.

As conducted with short reads, single-cell CRISPR screens has proven useful but does not verify the edit genotype and misses transcript features because of the limitation on sequence length. Addressing these limitations, our new methodology leverages long-read sequencing technology for single cell CRISPR analysis. As a result, CRISPR edit genotypes are directly confirmed in the target transcript. Isoform variation and levels of expression becomes evident at the resolution of individual cells. In addition to the effect of knock-out of splicing factors on transcript isoform generations, we can assess the transcriptomic result of known exon skipping or intron retention events as we shown in **Figure 1**. Beyond its application on identifying transcript isoforms, we demonstrated that this approach also enables one to directly assess CRISPR genetic alterations such as indels or single nucleotide variants present in the transcript even if they do not change the isoform structure. One limitation is that this direct evaluation requires expression of an adequate number of transcripts – future work will test the range of sensitivity for different levels of gene expression.

For this study, we demonstrated that our method can be applied across different CRISPR transduction approaches including plasmid, lentiviral and RNP delivery. In the case of the RNP method, one can use this method to assess individual or sets of cells directly for the presence of the target alteration. This may obviate the need for maintaining the expression of a guide RNA within an individual cell for sequencing, as other single cell CRISPR methods require.

Our method can be used with CRISPR base editing methodology. Unlike Cas9 nuclease which induces a complete gene KO, base editors introduce base substitution. However, base editors can induce a range of different base substitutions outside of what was originally planned. This lack of specificity leads to unexpected amino acid substitutions unless there is only one target nucleotide at the site of interest. Combining our long-read methodology with base editors, we can evaluate in a highly scalable and rapid fashion those cells with the intended substitution versus those with an unintentional base change. This feature allows one to determine the phenotypic effects of single base substitution within the target of interest.

METHODS

Cell culture conditions

HEK293T (ATCC CRL-11268) cells and Cas9-stable HEK293T (T3251; Applied Biological Materials Inc., Richmond, BC, CANADA) were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS). Jurkat (ATCC TIB-152) and Cas9-stable Jurkat (SL555, GeneCopoeia, Inc., Rockville, MD, USA) cells were maintained in Roswell Park Memorial Institute (RPMI) 1640 Medium supplemented with 10% FBS.

Transfection and electroporation condition

We used 1.2×10^6 HEK293T cells to transfect the Cas9 expression plasmid (1000ng) and sgRNA plasmids (1000ng) using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. For electroporation, 2×10^5 Jurkat cells were used. We used 1250ng of TrueCut™ Cas9 Protein v2 (Invitrogen) and then we used 585 ng of sgRNA were incubated at room temperature for 10 minutes before the electroporation to form the RNP complex. RNP was added to Jurkat cells, resuspended in R buffer and electroporated using Neon electroporation system (1700 V / 20 ms / 1 pulse, Invitrogen). Electroporated cells were transferred to a 24-well plate containing the culture medium. After six days of transfection or electroporation, cells were subjected to single-cell library preparation or stimulation.

Single-cell library preparation

Single-cell cDNA and gene expression libraries are prepared using Chromium Next GEM Single Cell 5' Library & Gel Bead Kit v1.1 or v2 (10X Genomics, Pleasanton, CA, USA) as per the manufacturer's protocol. The cDNA and gene expression libraries are amplified with 16 and 14 cycles of PCR respectively. The quality of gene expression libraries is confirmed using 2% E-Gel (ThermoFisher Scientific, Waltham, MA, USA). The libraries are then quantified using Qubit (Invitrogen) and sequenced on Illumina sequencers (Illumina, San Diego, CA, USA).

Long-read sequencing

From the single-cell full length cDNA, 25 ng were used to amplify transcripts. Primer sequences: Partial_R1 – CTACACGACGCTCTTCCGATCT, RACK1_ex8 – ACACTCGCACCAAGGTTGTCCG, PTPRC_ex7 – CCAGAAGGGCTCAGAGTGGT, MYL6_ex7 - ACACAGGGAAAGGCACGGACTCTGG. KAPA HiFi HotStart ReadyMix (Roche, Basel,

Switzerland) was used for amplification. Libraries were prepared with 900fmol of each amplicon for MinION flow cell FLO-MIN106D (Oxford Nanopore Technologies, Oxford, UK) and Promethion flow cell FLO-PRO002 (Oxford Nanopore Technologies) or 80fmol for Flongle flow cell FLO-FLG001 (Oxford Nanopore Technologies) using Native Barcoding Expansion and Ligation Sequencing Kit (Oxford Nanopore Technologies) as per the manufacturer's protocol. Libraries were sequenced on MinION or Promethion over 48h.

Lentiviral gRNA library production

The oligonucleotide pool for gRNA library cloning was ordered using IDT oPools Oligo Pools (Coralville, Iowa, USA). Amplified gRNA cassettes were cloned to lentiGuide-Puro (Addgene plasmid #52963) using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, Ipswich, MA, USA). Cloned plasmids were purified and electroporated to ElectroMAX Stbl4 Competent Cells (New England Biolabs).

Lentivirus production

2.0×10^6 HEK293T cells were plated 24h prior to transfection. Cells were transfected with lentiviral sgRNA library (2000ng), psPAX2 (1500ng, Addgene plasmid #12260) and pMD2.G (500ng, Addgene plasmid #12259) using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. The viral supernatant was collected after 48hr of transfection, filtered through a 0.45µm filter, and used.

Lentivirus transduction

To 1.0×10^5 Cas9-stable Jurkat, the lentiviral supernatant and 8µg of polybrene (Sigma-Aldrich, MO, USA) were added and the mixture was centrifuged at 800g for 30 minutes at 32 Celsius

degree. After that, cell pellets were resuspended to fresh media and plated in a 6-well plate. After 72 hours, transduced cells were selected by puromycin (Life technologies, CA, USA).

Guide RNA sequencing

The sgRNA direct capture was performed using Chromium Next GEM Single Cell 5' Library & Gel Bead Kit v2 (10X Genomics) as previously described.¹⁷ 6pmol of gRNA scaffold binding primer (oJR160) was added to RT master mix directly before droplet generation. After cDNA amplification, we performed 0.6X left-sided SPRI cleanup reaction for cDNAs and 0.6X-1.8X double-sided SPRI selection for gRNA using SPRIselect (Beckman Coulter Life Sciences, CA, USA). The sgRNA fractions were amplified by primers oJR163 and oJR163 (Supplementary table) and sequenced with gene expression library.

Single cell transcript isoform analysis

Short read transcripts: Basecalling for 5' gene expression libraries was performed using cellranger 3.2 (10X Genomics) mkfastq, followed by alignment to reference genome GRCh38, and transcript quantification using cellranger-count. In preparation for integrated analysis, the transcript count matrices generated by cellranger were processed by Seurat 3.0.2²⁰ QC filtering removed cells with fewer than 100 or more than 8000 genes, cells with more than 30% mitochondrial genes and cells predicted to be doublets by DoubletFinder.²¹ Additionally, any genes present in 3 or fewer cells were removed. Dimension reduction was performed using principal component analysis (PCA) and UMAP with 30 principal components and cluster resolution of 0.8.

Long read transcripts: Basecalling was performed using guppy 4.5.4¹¹ and alignment to the reference genome GRCh38 using minimap2 v2.17.¹² A custom python script utilizing pysam and the exon coordinates for the gene of interest (*RACK1*, *PTPRC* or *MYL6*) identified reads which span exon1 through exon8 for *RACK1*, exon3 through exon7 for *PTPRC*, or exon5 through exon7 for *MYL6*. For each read, exons which were present in the transcript were identified and recorded: for example 1234-678 indicated a *RACK1* transcript which skips exon 5.

Integration of long and short reads: Using knowledge of the adapter/cell barcode/UMI structure in the reads as well as the valid single cell barcodes from short-read sequencing, the putative long-read barcode is identified by evaluating the soft clipped portions of the aligned long reads, which are extracted utilizing the pysam module in a custom python script. A second custom python script used a machine-learning approach to identify the barcode. First, the list of valid short-read barcodes was vectorized using the CountVectorizer function from the Scikit-learn python module, with a kmer length of 8 to create a reference list. The 5' soft-clipped region of each read was then vectorized in the same way and compared to the created reference using a cosine similarity metric also from Scikit-learn. Similarly, the 3' soft-clipped region of each read was evaluated by matching the reverse-complement of the soft-clipped sequence to the reference list.

For each read, the barcodes with highest cosine similarity to the reference list were evaluated further. Edit distance to the reference barcode was calculated for each 16bp window across the search sequence for the top five barcodes with non-zero cosine similarity score. The barcode with lowest edit distance (and in cases of a tie, the highest cosine similarity score), was selected for final filtering. If the sequenced cDNA barcode had an edit distance < 3 from the reference short-read barcode it was considered a matching barcode, otherwise the read was not

considered a match to any of the short-read barcodes and was excluded from further integrated analysis. Output from this script was then summarized using awk and bash commands to provide transcript counts per isoform exon pattern identified from the initial long read processing for each barcode. A clustered heatmap was generated in R using heatmap.2, cutree and dendextend, showing the proportion of each isoform per cell barcode. The isoform cluster for each barcode was then added to the Seurat object as a metadata column using AddMetaData, to enable this categorization to be used in subsequent visualization plots.

Guide RNA in-vitro transcription

sgRNA was in-vitro transcribed by T7 RNA polymerase using the MEGAshortscript T7 kit (Invitrogen) according to the manufacturer's protocol. Templates for sgRNA were generated by extension of two complementary oligo nucleotides. Transcribed RNA was purified by column purification. Purified RNA was quantified by Qubit 4 Fluorometer (Invitrogen).

Jurkat cell stimulation

2×10^6 Jurkat cells were stimulated with eBioscience™ Cell Stimulation Cocktail (Invitrogen) using half of the recommended concentration in 6-well plate for 24h. Stimulated cells were subjected to single-cell library preparation.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

None to disclose.

AUTHORS' CONTRIBUTIONS

HSK and HPJ were involved in conception and design of the study, development of methodology, acquisition of data, analysis and interpretation of data and writing of the manuscript. HSK, SMG and HPJ were involved in analysis and interpretation of data. SMG, BTL and HPJ were involved in the development of methodology and acquisition of data. All authors contributed to the writing of the manuscript. HPJ oversaw all aspects of the study.

ACKNOWLEDGEMENTS

This work was supported by US National Institutes of Health grants P01HG000205 (HPJ, BTL, SMG), R01HG006137 (HPJ, HSK), R35HG011292-01 to BTL and R33 CA247700 (HPJ, HSK, BTL). HPJ also received support from the Clayville Foundation.

DATA AVAILABILITY

The sequencing data have been deposited in the NCBI Sequence Read Archive database under the accession number PRJNA708300.

REFERENCES

1. Gazzara, M.R. et al. Ancient antagonism between CELF and RBFOX families tunes mRNA splicing outcomes. *Genome Res* **27**, 1360-1370 (2017).
2. Bebee, T.W. et al. The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. *Elife* **4** (2015).
3. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
4. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
5. Singh, M. et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat Commun* **10**, 3120 (2019).
6. Gupta, I. et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* (2018).
7. Jaitin, D.A. et al. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896 e1815 (2016).
8. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866 e1817 (2016).
9. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* **14**, 297-301 (2017).
10. Adamson, B. et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882 e1821 (2016).
11. Wick, R.R., Judd, L.M. & Holt, K.E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**, 129 (2019).

12. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
13. Salton, G. & Buckley, C. Global text matching for information retrieval. *Science* **253**, 1012-1015 (1991).
14. Hermiston, M.L., Xu, Z. & Weiss, A. CD45: a critical regulator of signaling thresholds in immune cells. *Annu Rev Immunol* **21**, 107-137 (2003).
15. Oberdoerffer, S. et al. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science* **321**, 686-691 (2008).
16. ten Dam, G.B. et al. Regulation of alternative splicing of CD45 by antagonistic effects of SR protein splicing factors. *J Immunol* **164**, 5287-5295 (2000).
17. Replogle, J.M. et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol* **38**, 954-961 (2020).
18. Koretzky, G.A., Picus, J., Schultz, T. & Weiss, A. Tyrosine phosphatase CD45 is required for T-cell antigen receptor and CD2-mediated activation of a protein tyrosine kinase and interleukin 2 production. *Proc Natl Acad Sci U S A* **88**, 2037-2041 (1991).
19. Li, J. & Yu, P. Genome-wide transcriptome analysis identifies alternative splicing regulatory network and key splicing factors in mouse and human psoriasis. *Sci Rep* **8**, 4124 (2018).
20. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
21. McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324 (2019).

FIGURE LEGENDS

Figure 1. (a) Overview of single-cell short/long-read integration strategy. (b) Structure of *RACK1* transcript. The red arrows point to CRISPR target sites that disrupt splicing acceptor sequences. (c, d) Heatmaps showing the proportion of each transcript isoform (x-axis) with each cell (y-axis) and clustering based on transcript isoform proportion for *RACK1* exon5 treated cells (c) and multiple exon treated cells (d). (e) Expression level of *RACK1* in each cell cluster in (d). P values were calculated in comparison with cluster 3 (P= 5.7e-12, 2.6e-11, 9.9e-11, 1.6e-10, 1.1e-07, 9.5e-06, 2.5e-07, 1.1e-08 ; two-sided t-test).

Figure 2. (a) Overview of single-cell CRISPR screen integrated with long-read sequencing. (b) Boxplot showing the ratio of short *PTPRC* transcript isoform (RO and RB) for cells with gRNAs targeting indicated genes. P values are calculated in comparison with Nontarget. Genes which have less than three cells with target gRNAs are not shown. (c) Heatmap showing proportion of each transcript isoform (x-axis) with each cell (y-axis) and clustering based on transcript isoform proportion for cells having indicated gRNA sequence.

Figure 3. (a) Overview of splicing factors affect alternative splicing. (b) Quantification of short transcript isoform per target gene. For each sample (x-axis) the ratio of transcript isoform RO and RB among all *PTPRC* isoforms for cells are shown as box plot. (c) Heatmap showing proportion of each transcript isoform (x-axis) with each cell (y-axis) and clustering based on transcript isoform proportion for each sample.

b

Ratio of short transcript

$$= \frac{\text{RO read \#} + \text{RB read \#}}{\text{Total } PTPRC \text{ read per cell}}$$

1-3 - - - 7... 1-3 - B - 7...