nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-023-01815-7

High-throughput RNA isoform sequencing using programmed cDNA concatenation

Received: 1 October 2021

Accepted: 2 May 2023

Published online: 08 June 2023

Check for updates

Aziz M. Al'Khafaji[1,11] ✉, Jonathan T. Smith[1,11], Kiran V. Garimella[1,11] ✉, Mehrtash Babadi[1,11] ✉, Victoria Popic[1,11] ✉, Moshe Sade-Feldman[1,2], Michael Gatzen[1], Siranush Sarkizova[1], Marc A. Schwartz[1,3,4,5], Emily M. Blaum[1,2], Allyson Day[1], Maura Costello[1], Tera Bowers[1], Stacey Gabriel[1], Eric Banks[1], Anthony A. Philippakis[1], Genevieve M. Boland[6], Paul C. Blainey[1,7,8] ✉ & Nir Hacohen[1,2,9,10] ✉

Full-length RNA-sequencing methods using long-read technologies can capture complete transcript isoforms, but their throughput is limited. We introduce multiplexed arrays isoform sequencing (MAS-ISO-seq), a technique for programmably concatenating complementary DNAs (cDNAs) into molecules optimal for long-read sequencing, increasing the throughput >15-fold to nearly 40 million cDNA reads per run on the Sequel IIe sequencer. When applied to single-cell RNA sequencing of tumor-infiltrating T cells, MAS-ISO-seq demonstrated a 12- to 32-fold increase in the discovery of differentially spliced genes.

Although RNA sequencing has accelerated our understanding of biology, accurate quantification and discovery of RNA isoforms remain a challenge[1]. Alternative splicing is a core regulatory process that modulates the coding sequence, translation efficiency, stability and localization of mRNAs through differential splicing (DS) of exons during transcript maturation. Beyond being an integral component of cellular/organismal development and homeostasis, alternative splicing is implicated in a wide range of pathologies with hallmark isoforms being linked to cardiovascular, neurological and immunological diseases[2,3]. Additionally, mutated and/or dysregulated splicing factors make up a major class of phenotypic alterations associated with tumor progression and therapeutic resistance[4].

High-throughput full-length RNA isoform identification and quantification remain challenging for single-cell and bulk studies as the necessary read lengths (>5 kb) and depths ($>2 \times 10^7$ reads) are not easily attainable by existing sequencing platforms. For example, short-read sequencing platforms (for example, Illumina) achieve more than sufficient throughput ($>1 \times 10^9$ reads) but are hindered by limited read lengths (50–600 bp) that are inadequate to span the majority of human transcripts (~1.6 ± 1.1 kb; Supplementary Fig. 1). As a result, individual short-reads often fail to span successive splice sites, impairing efforts to correctly identify alternative transcript isoforms[5]. A recently developed short-read sequencing approach, Smart-seq3, enhances isoform detection by enabling single-molecule reconstruction via integration of reads from products with the same 5′ unique molecular identifier (UMI)[6]. However, due to the 5′ coverage bias of Smart-seq3, most transcript molecules are only partially reconstructed, resulting in poor isoform identification and discovery. Conversely, the long-read platforms from Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) enable the full-length RNA isoform sequencing needed for robust isoform identification and discovery but suffer from comparatively low read throughput at high costs, limiting the scope of their application. Early limitations in raw base calling accuracy on long-read platforms (error rates of 10–15%) have been mitigated by improvements

[1] Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2] Department of Medicine, Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA. [3] Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [4] Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA. [5] Department of Pediatric Oncology, Dana Farber Cancer Institute, Boston, MA, USA. [6] Division of Surgical Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [7] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [8] Koch Institute for Integrative Cancer Research at the Massachusetts Institute of Technology, Cambridge, MA, USA. [9] Harvard Medical School, Boston, MA, USA. [10] Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, MA, USA. [11] These authors contributed equally: Aziz M. Al'Khafaji, Jonathan T. Smith, Kiran V Garimella, Mehrtash Babadi, Victoria Popic. ✉e-mail: aalkhafa@broadinstitute.org; kiran@broadinstitute.org; mehrtash@broadinstitute.org; vpopic@broadinstitute.org; pblainey@broadinstitute.org; nhacohen@broadinstitute.org

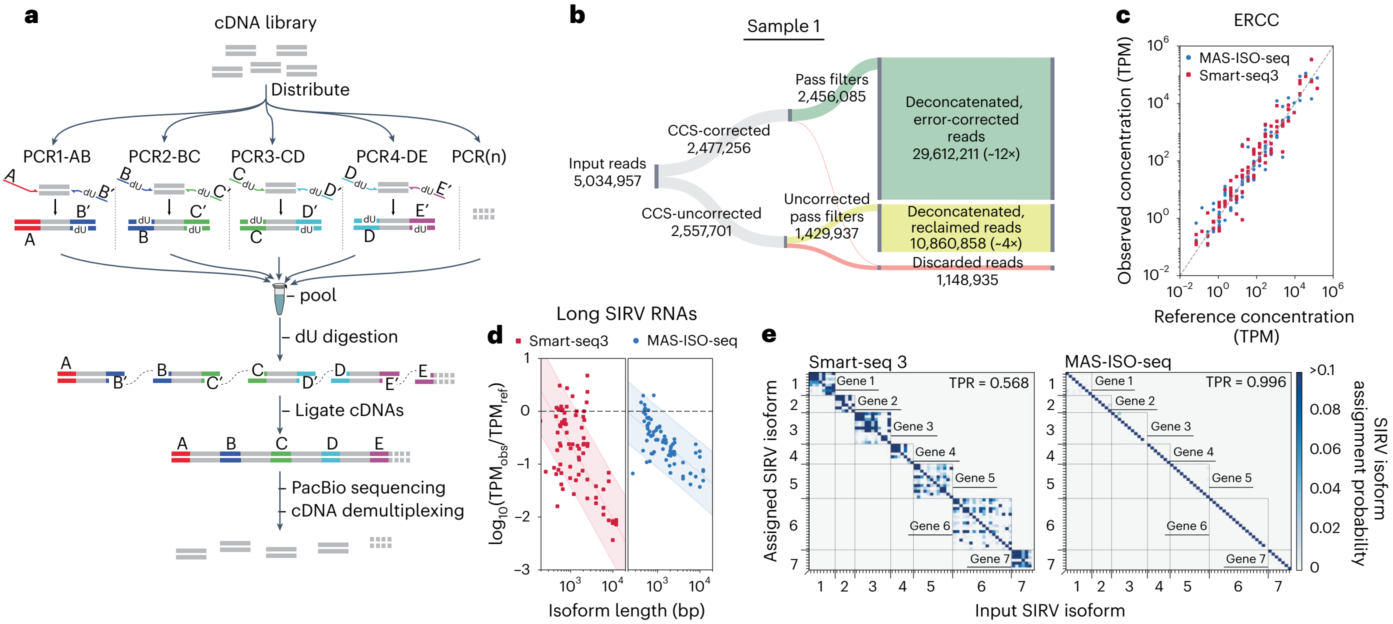

Fig. 1 | MAS-ISO-seq workflow and experimental validation using synthetic RNA isoforms. a, Schematic representation of the MAS-ISO-seq intramolecular cDNA multiplexing workflow. **b**, Sankey diagram reporting MAS-ISO-seq run yield of sample 1 at various stages of processing. **c**, Observed ERCC concentrations as measured in MAS-ISO-seq and Smart-seq3 experiments versus reference concentrations ($R^2 > 0.95$ for both). **d**, Log ratio of observed to reference concentrations of short and long SIRV isoforms in SIRV-Set 4 versus transcript length for Smart-seq3 and MAS-ISO-seq. **e**, Isoform identification confusion matrix for SIRV isoforms as measured by Smart-seq3 reconstructions and MAS-ISO-seq observations.

in pore-based nucleotide reading, circularized consensus sequencing (CCS, or HiFi) and consensus generation strategies for individual library molecules[7–9]. On the PacBio Sequel IIe platform, consensus base quality reaches the Phred-scale quality of Q30 at ~10 circular passes, with marginal quality improvement on additional passes. For the current Sequel IIe instrument and single molecule, real-time (SMRT) Cell 8M chemistry, the optimal library size for reaching ~10 circular passes is 15–20 kb. As transcript lengths typically range substantially shorter (200 bp–5 kb), CCS of individually circularized complementary DNA (cDNA) molecules using the standard Iso-Seq protocol (PacBio) yields an excessive number of circular passes (50–60) and ineffectively uses the available sequencing potential of the platform (Supplementary Fig. 2).

To maximize the sequencing throughput on the PacBio platform, we developed a method for the programmable concatenation of DNA fragments into long composite sequence library molecules, multiplexed arrays sequencing (MAS-seq; Fig. 1a). When MAS-seq is used for sequencing transcript isoforms, we term the approach MAS-isoform-seq (MAS-ISO-seq). The protocol begins by depleting TSO (template switching oligo) priming artifacts via streptavidin/biotin selection of molecules containing the oligo-dT adapter from the input cDNA library. The purified cDNA library is then split across parallel PCRs, which serve to both increase cDNA yield and append reaction-specific deoxy-uracil (dU) containing barcode adapters. Using dU digestion followed by barcode-directed ligation of cDNAs, MAS-ISO-seq generates long concatenated cDNA arrays assembled deterministically with a narrow length distribution that allows for both accurate consensus sequencing and more optimal capacity utilization of the PacBio long-read platform. To drive accurate and specific hybridization, we designed 15 bp ligation barcode adapters with each having a Hamming distance of 11 from all other barcodes[10]. In combination with upstream depletion of TSO priming artifacts via streptavidin/biotin selection, MAS-ISO-seq boosts the sequencing throughput to ~40 million full-length transcripts per SMRT Cell 8M flow cell, a >15-fold increase over CCS-corrected read counts (Fig. 1b).

To demonstrate MAS-ISO-seq's performance, we carried out a 15-member cDNA ligation from two 5′ single-cell gene expression cDNA libraries (10× Genomics) of tumor-infiltrating CD8$^+$ T cells. As expected, we observed a ~15-fold increase in cDNA library length after ligation (Supplementary Fig. 3). MAS-ISO-seq libraries underwent standard CCS library preparation and were sequenced on the PacBio Sequel IIe. Sequenced libraries exhibited corrected read length and circular pass count distributions more comparable to whole-genome CCS data than the standard isoform sequencing method, Iso-Seq, as expected due to longer concatenated library lengths (Supplementary Fig. 4).

The programmed sequential pattern of MAS-ISO-seq adapters provides landmarks for effective cDNA segmentation and constraints for detecting malformed or otherwise defective array structures. MAS-ISO-seq adapters also enable the utilization of CCS-uncorrected reads that are otherwise discarded using standard methods. To exploit these signals, we developed a composite profile hidden Markov model, Longbow, for the probabilistic annotation and optimal segmentation of each MAS-ISO-seq read via maximum *a* posteriori state path (Methods). Across both single-cell MAS-ISO-seq libraries, 99.01–99.15% of CCS-corrected reads and 54.27–60.72% of CCS-uncorrected reads were found to segment consistently. To maximize precision, segmentation results inconsistent with our expected array structure (that is, off-subdiagonal elements of the matrices in Supplementary Fig. 5a,b) were filtered out (Supplementary Fig. 5c,d). A plurality of filtered reads (sample 1, 29.54%; sample 2, 35.61%) were found to contain fully formed 15-element arrays. Arrays with fewer than 15 cDNAs were more prevalent in CCS-uncorrected reads than in CCS-corrected (Supplementary Figs. 6 and 7). Across both libraries, this process yielded 37–40 million cDNA reads for downstream analysis (a gain of 16.34–22.90× compared to the CCS-corrected read yield; Fig. 1b and Supplementary Fig. 8).

The segmented reads were then filtered again to remove reads that failed to conform to the library structure at the individual cDNA level (Longbow sift command; Methods). The vast majority of sifted, segmented reads from these partial arrays still contained consecutive adapter sequences, a poly(A) tail and had a high

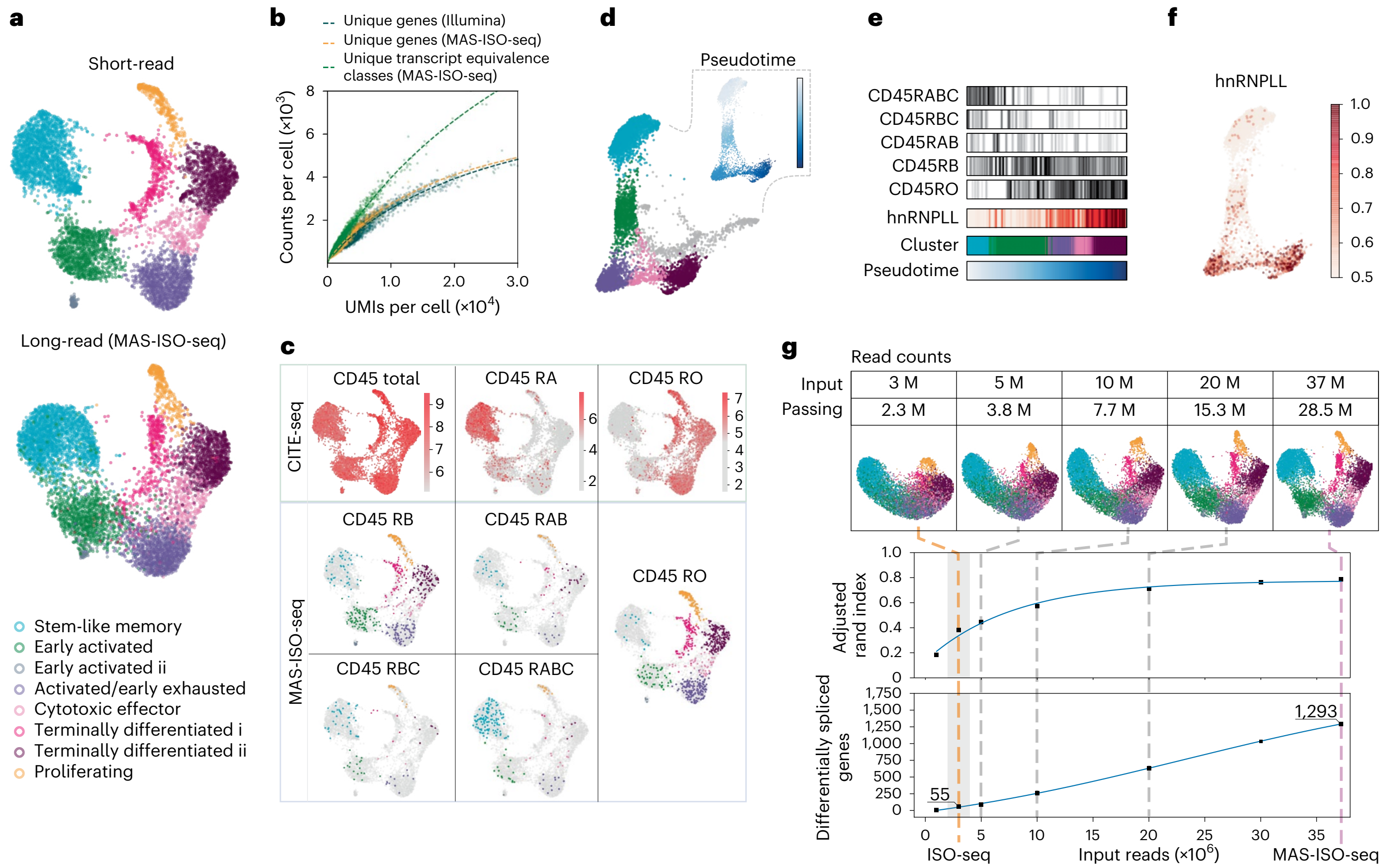

Fig. 2 | Single-cell isoform-resolved sequencing of primary human CD8$^+$ T cells with MAS-ISO-seq. a, UMAP embedding of single-cell gene expression of 5,270 CD8$^+$ T cells from short- or long-read analyses; the long-read UMAP is annotated with the cell identities determined from the short-read data. **b**, Scatter plot of unique gene or transcript counts in cells versus UMI counts per cell for short-read (Illumina) and long-read (MAS-ISO-seq). **c**, CD45 (*PTPRC*) isoform analysis using either CITE-seq or MAS-ISO-seq (natural log raw counts). **d**, Force-directed graph of CD8$^+$ T cells with insets depicting pseudotime progression and differential CD45 isoform expression along the pseudotime axis. **e**, Expression levels of encoded CD45 isoforms and hnRNPLL along pseudotime and in each cluster. **f**, Expression of hnRNPLL along the pseudotime progression (log-normalized counts). **g**, Downsampling analysis of MAS-ISO-seq reads; (top) evolution of UMAP embedding versus depth; (middle) ARI between short-reads reference annotations and downsampled long reads versus depth; (bottom) number of statistically significant differentially spliced genes versus depth. Typical Iso-seq read depths shaded in gray.

mapping quality to the genome (96.90%; Supplementary Figs. 6 and 7). After final filtering across both samples, we obtained ~21 M to 28 M quantification-ready CCS-corrected transcripts (~11- to 13-fold yield increase over the number of CCS-corrected reads) and ~6 M to 8 M quantification-ready CCS-uncorrected transcripts (an additional ~2- to 5-fold increase) for a total 14- to 18-fold increase as compared to raw CCS-corrected reads.

To validate the ability of MAS-ISO-seq to faithfully identify RNA isoforms, we performed full-length RNA sequencing of the Lexogen SIRV-Set 4, a synthetic mixture of spike-in RNA variants (SIRVs) containing 69 RNA isoforms of varying lengths and equal molarity across seven 'genes', 15 long (4–12 kb) SIRVs and 92 ERCC RNA standards with concentration spanning six orders of magnitude[11]. Smart-seq3 short-read sequencing of the SIRV-Set 4 library was performed in parallel to compare short-read isoform reconstructions to our high-throughput long-read sequencing approach. Although quantification of ERCC standards was broadly similar overall between both protocols (Fig. 1c), long isoforms showed markedly reduced length bias in MAS-ISO-seq and Iso-Seq versus Smart-seq3 (Fig. 1d and Supplementary Fig. 9). Smart-seq3 isoform reconstructions exhibited substantial ambiguity in assigning reconstructed transcripts to a specific known isoform (~43% error rate; Fig. 1e). In contrast, MAS-ISO-seq allows direct identification of transcript isoforms without the need for in silico reconstruction, and hence leads to virtually unambiguous isoform assignment (~0.4% error rate; Fig. 1e).

To characterize the performance of MAS-ISO-seq for single-cell RNA sequencing, we performed 10× Genomics 5′ single-cell gene expression on tumor-infiltrating CD8$^+$ T cells. Using the standard 5′ single-cell gene expression protocol, we generated both standard short-read and MAS-ISO-seq long-read libraries from the same full-length cDNA library. To overcome challenges associated with the incompleteness of the available isoform annotations and cDNA truncation artifacts, we developed a graph-based algorithm that assigns each read to an isoform equivalence class based on the junction-level relationship between the read, GENCODE reference annotations and de novo annotations discovered from all reads using StringTie2 (ref. 12; Methods; Supplementary Figs. 10 and 11). After applying conventional QC filtering steps and separating primary tumor cells (Methods), we obtained 5,270 CD8$^+$ T cells containing a median of 4,041 UMIs/cell (short-read data) and 1,701 UMIs/cell (long-read data). Sequencing saturation was higher for the short-read run, 1.98 reads/UMI (short) versus 1.22 reads/UMI (long). We leveraged the presence of a small number of primary tumor cells in our sample and the mutually exclusive expression of several immune and tumor genes to estimate the accuracy of MAS-ISO-seq cell barcode (CBC) assignments to be in the range of 99.0–99.7% (Methods; Supplementary Fig. 12). Despite large discrepancies in sequencing depth between short- and long-read approaches and quantification methodologies (Methods), cell clustering and gene expression were highly concordant (Fig. 2a, adjusted Rand index (ARI) = 0.79, Fig. 2b, concordant gene count saturation curves, and Supplementary Fig. 13, $R^2 = 0.91$). A common set of T-cell transcriptional states, ranging from stem cell-like to terminally differentiated, were observed in both datasets.

Leveraging the distinct splicing patterns of CD45 (*PTPRC*) over the course of T-cell differentiation, we performed orthogonal validations of CD45 isoform expression at the protein level using CITE-seq and compared them to the mRNA levels measured with MAS-ISO-seq[13]. CD45 isoform expression between these two modalities was highly concordant (Fig. 2c). Notably, mRNA measurements were more granular in their ability to resolve the multiple encoded CD45 isoforms present (RO, RA, RAB, RB and RBC) as compared to the antibody-based CITE-seq approach. This is due to the single-epitope specificity of antibodies that limits or does not enable discrimination between closely related isoforms[14]. For example, the CD45 RA antibody cannot distinguish between CD45 RA and RAB. Pseudotime analysis revealed a continuum of T-cell states leading from stem cell-like to activated to terminally differentiated. Canonical CD45 isoform expression and its associated splicing factor, hnRNPLL[13], were tracked clearly along this differentiation trajectory (Fig. 2d–f).

To quantify the impact of the sequencing depth gained by MAS-ISO-seq on cell typing and identification of differentially spliced (DS) genes, we performed an in silico downsampling analysis from a single MAS-ISO-seq run. We processed each dataset identically using the same pipeline and computed the ARI between the cell clustering of the subsampled long-read dataset and the full short-read dataset as a reference. We also determined the number of DS genes across the T-cell subtypes for each downsampling run (Methods). Compared to the read-depth expected from an Iso-Seq run (2–4 M HiFi reads passing filters), the throughput gain afforded by MAS-ISO-seq translates to a 34–47% increase and saturation of ARI between short-read and long-read single-cell clustering and a 12–32-fold gain in identifying DS genes (multiple hypothesis testing correction with false discovery rate (FDR) < 0.05; Fig. 2g and cluster-resolved results given in Supplementary Fig. 14). Notably, a plurality of the DS genes was distinct from the set of differentially expressed (DE) genes (Supplementary Fig. 15).

In this work, we detailed and validated MAS-ISO-seq, a programmable cDNA concatemerization method that boosts the throughput of the PacBio long-read sequencing platform >15-fold to ~40 million deconcatenated reads per run. Using synthetic RNA isoforms as a ground truth library, we demonstrate that MAS-ISO-seq is far superior in confidently identifying RNA isoforms as compared to short-read approaches. Furthermore, we leveraged MAS-ISO-seq to perform single-cell RNA isoform sequencing on human tumor-infiltrating CD8$^+$ T cells. We validated our ability to accurately identify isoforms by resolving canonical CD45 isoform expression differences across the range of observed cell states and orthogonal protein isoform-based measurements. Through downsampling analyses, we demonstrate that the additional throughput afforded by MAS-ISO-seq is sufficient to enable robust cell clustering into known T-cell differentiation states and substantially boosts the identification of DS genes. As adequate sequencing depth is, in part, a function of cellular RNA content, deeper sequencing may be necessary to provide adequate power for downstream single-cell analyses. A related approach, HIT-scISOseq, leverages palindromic adapter sequences to drive ligation of an indeterminate number of cDNAs, enabling ~10 million transcript reads[15]. While producing a fourfold lower yield as compared to MAS-ISO-seq, HIT-scISOseq additionally lacks the sequential array structure that MAS-ISO-seq exploits for accurate segmentation and identification of malformed arrays. Other concatenation approaches for targeted DNA sequencing use Gibson Assembly or Golden Gate Assembly for array formation. These methods also demonstrate considerably lower throughput and lack the error robustness of MAS-ISO-seq arrays[16,17].

Challenges impacting the RNA isoform sequencing field as a whole include cDNA synthesis artifacts, incomplete transcriptome references and transcriptome assembly software with limited performance. We believe that the read throughput afforded by approaches such as MAS-ISO-seq will lower barriers to data generation and catalyze progress to surmount these challenges. The compatibility of MAS-ISO-seq with archived single-cell cDNA libraries generated in cell atlasing studies poises the field to immediately advance isoform discovery and generate cell type-specific isoform-resolved transcriptome references at scale. Furthermore, MAS-ISO-seq will augment a broad range of efforts, including gene fusion identification, proteogenomic resolution, neoantigen discovery and TCR/BCR repertoire sequencing. To date, PacBio and ONT have driven transformative advancements in long-read sequencing, releasing new platforms and chemistries with increased base-level accuracy and throughput (for example, Revio and Q20+). Given the modular and scalable nature of MAS-ISO-seq, the workflow is positioned to co-evolve with compatible long-read sequencing platforms, enabling even greater throughput as read lengths, yield and per-base accuracy increase.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01815-7.

References

1. Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A. & Tilgner, H. U. Getting the entire message: progress in isoform sequencing. *Front. Genet.* **10**, 709 (2019).
2. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
3. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2015).
4. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).
5. Kanitz, A. et al. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**, 150 (2015).
6. Hagemann-Jensen, M. et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
7. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
8. Volden, R. 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. USA* **115**, 9726–9731 (2018).
9. Baid, G. et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2022).
10. Buschmann, T. & Bystrykh, L. V. Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* **14**, 272 (2013).
11. Paul, L. et al. SIRVs: spike-in RNA variants as external isoform controls in RNA-sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/080747 (2016).
12. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).

13. Oberdoerffer, S. et al. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science* **321**, 686–691 (2008).
14. Bio-Rad. Mini-review: CD45 characterization and isoforms. https://www.bio-rad-antibodies.com/cd45-characterization-isoforms-structure-function-antibodies-minireview.html (2023).
15. Shi, ZX., Chen, ZC. & Zhong, JY. High-throughput and high-accuracy single-cell RNA isoform analysis using PacBio circular consensus sequencing. *Nat Commun* **14**, 2631 (2023).
16. Schlecht, U., Mok, J., Dallett, C. & Berka, J. ConcatSeq: a method for increasing throughput of single molecule sequencing by concatenating short DNA fragments. *Sci. Rep.* **7**, 5252 (2017).
17. Kanwar, N., Blanco, C., Chen, I. A. & Seelig, B. PacBio sequencing output increased through uniform and directional fivefold concatenation. *Sci. Rep.* **11**, 1–13 (2021).

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

Methods

Patients consent and sample collection

Patients' $CD8^+$ T cells analyzed in this study were collected under the Dana–Farber/Harvard Cancer Center Institutional Review Board (protocol 11-181) and provided written informed consent before tissue collection.

Single-cell and SIRV cDNA library preparation

Sample dissociation and fluorescence-activated cell sorting (FACS) of $CD3^{\pm}CD8^{\pm}$ T cells. Using the human tumor dissociation kit (Miltenyi Biotec, 130-095-929), freshly isolated tumors were digested to obtain a single-cell suspension. Tissue was placed into a 1.5 ml Eppendorf tube containing 420 μl of Dulbecco's modified Eagle medium (DMEM) with 10% fetal calf serum (FCS), 42 μl of enzyme H, 21 μl of enzyme R and 5 μl enzyme A (provided with the kit). The tissue was minced using surgical scissors, and an additional 512 μl of DMEM with 10% FCS was added to the tube (total volume of 1 ml). Next, the tissue was incubated for 15 min at 37 °C, 350 r.p.m. in a thermomixer (Eppendorf; F1.5). After incubation, the tissue was further digested using a 1 ml syringe plunger over a 50 μm filter (Sysmex, 04-004-2327), making sure to wash the filter with media. Using ACK (ammonium-chloride-potassium) buffer (Gibco, A1049201), RBC lysis was performed and the sample was finally resuspended in DMEM with 10% FCS to count and determine the viability of the cells using a manual hemocytometer (Bright-line, 1492). Cells were then washed twice with cold 1× PBS, and the cells were incubated with live/dead Zombie Violet Dye (Biolegend, 423114) for 15 min at room temperature as suggested by the manufacturer. The cells were then washed and resuspended with 1× PBS containing 1.5% FCS for cell-surface labeling using a standard protocol for 30 min at 4 °C. An antibody panel was used to identify and sort the $CD3^+CD8^+$ T-cell population—human TrueStain FcX (Biolegend, 422302), PE (phycoerythrin) antihuman CD45 (Biolegend, 304008), FITC (fluorescein isothiocyanate) antihuman CD3 (Biolegend, 317306), APC (allophycocyanin)/Cyanine7 antihuman CD235a (Biolegend, 349116) and APC antihuman CD8a (Biolegend, 300912). Sorting of single live $CD3^+CD8^+$ T cells (gating on $Zombie^{low}$, $hCD235a^-$, $hCD45^+$, $hCD3^+$, $hCD8^+$) was performed using a Sony MA900 cell sorter. Cells were sorted into a 15 ml tube containing DMEM with 10% FCS. After sorting, tubes with sorted cells were vortexed briefly, spun down at 1500 r.p.m., 4 °C for 5 min, resuspended and counted for yield (Supplementary Fig. 16).

TotalSeq-C staining and single-cell RNA-sequencing procedure. Sorted $CD3^+CD8^+$ T cells were washed and resuspended with staining buffer (1× PBS + FCS 2.5% + 2 mM EDTA). Next, TruStain FcX (FC blocker; Biolegend, 422301) was added and the sample was incubated for 10 min at 4 °C. After incubation with FcX blocker, the cells were washed with staining buffer once and spun down at 1,500 r.p.m., 4 °C for 5 min. The cells were then incubated for 20 min at 4 °C with the following TotalSeq-C antibody mix: TotalSeq-C0048 antihuman CD45 antibody (Biolegend, 368545), TotalSeq-C0103 antimouse/antihuman CD45R/B220 (Biolegend, 103273), TotalSeq-C0087 antihuman CD45RO (Biolegend, 304259) and TotalSeq-C0063 antihuman CD45 RA (Biolegend, 304163). Before adding the surface antibody mix, equal volumes of each antibody were combined and the mix was spun at 14,000 r.p.m. for 5 min to remove aggregates. After staining, the cells were washed twice with staining buffer, and a final wash was completed in DMEM with 10% FCS before counting. Single-cell RNA libraries were generated using the 10× Genomics Chromium Single Cell V(D)J Reagent Kit using 5′ v1 chemistry with Feature Barcode technology for Cell-Surface Protein (10× Genomics, 1000080). After each step, cDNA generation, gene expression libraries and cell-surface protein libraries samples quality was assessed using the Qubit dsDNA high-sensitivity kit (Invitrogen, Q32854) and the high-sensitivity BioA DNA kit (Agilent, 5067-4626). Samples that passed quality control were sequenced on a NextSeq 500 sequencer (Illumina), using pair-end reads, with 26 reads for read 1 and 55 reads for read 2.

Multiplexed array assembly of cDNA libraries. cDNA libraries were amplified using the following reaction conditions: 34 μl of H_2O, 25 μl of Kapa HiFi Uracil+ ReadyMix (2×; Roche, 7959079001), 5 μl of primer AAO272 (10 μM, Integrated DNA Technologies (IDT)), 5 μl of primer AAO273 (10 μM, IDT) and 6 μl 10 × 5′ cDNA library (~3 ng μl^{-1}), and the following cycling conditions: 98 °C for 3 min, followed by 5 cycles of 98 °C for 20 s, 65 °C for 30 s and 72 °C for 8 min, followed by a final 72 °C extension for 10 min. Amplified libraries were purified using 0.7× SPRIselect (Beckman Coulter B23318) cleanup and quantified using Qubit (Thermo Fisher Scientific, Q32851). Libraries were further purified using 10 μl (100 μg) Dynabeads kilobaseBINDER (Thermo Fisher Scientific, 60101) with final bead reconstitution in 40 μl TE (tris & EDTA) buffer (Thermo Fisher Scientific, AM9849) after binding/washing. After streptavidin purification, 2 μl of USER (Uracil-Specific Excision Reagent) enzyme (M5505S) was added and incubated at 37 °C for 2 h to uncouple the bound cDNAs from the beads. Following USER digestion, the reaction was placed on a magnet for 5 min, separating the beads and supernatant containing the cDNAs. The cDNA fraction was moved to a fresh tube and purified using 0.7× SPRIselect (Beckman Coulter, B23318) cleanup. After cDNA purification, the following PCR master mix was assembled: 580 μl of H_2O, 750 μl of Kapa HiFi Uracil+ ReadyMix (2×; Roche, 7959079001) and 20 μl 10 × 5′ cDNA library (~6 ng μl^{-1}). In total, 90 μl of the master mix was distributed in 15 PCR tubes, each containing 10 μl of 5 μM MAS-ISO-seq primer pair mix (Supplementary Table 1). The 15 reactions were then thermocycled with the following cycling conditions: 98 °C for 3 min, followed by 8 cycles of 98 °C for 20 s, 65 °C for 30 s and 72 °C for 8 min, followed by a final 72 °C extension for 10 min (optimal cycling number was identified using scaled-down qPCR reaction). Reactions were then pooled in a 5 ml tube and purified using a 0.7× SPRIselect (Beckman Coulter, B23318) cleanup and eluted in 450 μl of TE. In a subsequent reaction, 15 μl of USER Enzyme (M5505S) was added to 435 μl of the pooled product and set to incubate at 37 °C for 2 h. Following USER digestion, 15 μl HiFi Taq DNA Ligase (M0647S) and 51 μl of HiFi Taq DNA Ligase buffer were added to the reaction and incubated in a thermocycler at 42 °C for 2 h. Following ligation, the reaction was purified using a 0.7× AMPure PB Bead (PacBio, 100-265-90) cleanup and eluted in 180 μl of H_2O. Multiplexed array libraries were quantified using Qubit (Thermo Fisher Scientific, Q32851) and Genomic DNA ScreenTape (Agilent, 5067-5365).

SIRV-Set 4 cDNA generation. SIRV-Set 4 (Lexogen, 141.01) was thawed and aliquoted 1 μl into each of the nine PCR tubes on ice. Following primary aliquoting, 2 μl of Tris–EDTA (pH 7.0) was added to each tube and mixed. SIRV stocks were then frozen at −80 °C. For first strand synthesis, the following primary master mix was set up: 15.5 μl of H_2O, 3.2 μl of polyethylene glycol 8,000 50% (wt/vol; VWR, 25322-68-3), 0.24 μl Triton X-100 (10%) solution (Thermo Fisher Scientific, 9002-93-1), 0.32 μl SUPERase·In RNase Inhibitor (Thermo Fisher Scientific, AM2696), 1.6 μl of dNTP (deoxynucleotide triphosphates) mix (10 mM; NEB, N0447S), 0.16 μl OligodT primer (100 μM; IDT; SS3_OligodTVN for Smart-seq3 and MAS_OligodTVN for Iso-seq and MAS-ISO-seq) and 3 μl of SIRV-Set 4 aliquot. Additionally, the following RT master mix was assembled: 1.2 μl of H_2O, 0.8 μl of Tris–HCl (pH 8.5; 1 M), 0.96 μl of NaCl (1 M), 0.8 μl of $MgCl_2$ (100 mM), 0.32 μl of GTP (guanosine triphosphate) (100 mM), 2.56 μl of DTT (dithiothreitol) (100 mM), 0.4 μl of SUPERase•In RNase Inhibitor (Thermo Fisher Scientific, AM2696), 0.64 μl of TSO 100 μM (IDT; SS3_OligodTVN for Smart-seq3 and MAS_OligodTVN for Iso-seq and MAS-ISO-seq) and 0.32 μl of Maxima H-minus RT enzyme 200 U μl^{-1} (Thermo Fisher Scientific, EP0751). Both primary and RT master mixes were added to the thermocycler with the following conditions: 42 °C for 90 min, followed by 10 cycles of 50 °C for 2 min and 42 °C for 2 min, followed by a final 85 °C for 5 min.

Smart-seq3 of SIRV-Set 4. To amplify the cDNA, the cDNA generation reaction was added straight into the following PCR mix: 26.5 µl of H_2O, 16 µl of Kapa HiFi HotStart buffer (5×), 2.4 µl of dNTP mix 10 mM (NEB, N0447S), 0.4 µl of $MgCl_2$ (100 mM), 0.4 µl of fwd_primer 100 µM (IDT), 0.8 µl of rev_primer 10 µM (IDT) and 1.6 µl of Kapa HiFi DNA polymerase (KK2103). The reaction was amplified using the following conditions: 98 °C for 3 min, followed by 13 cycles of 98 °C for 20 s, 65 °C for 30 s and 72 °C for 8 min, followed by a final 72 °C extension for 10 min. Amplified cDNA libraries were purified using 0.7× SPRIselect (Beckman Coulter, B23318) cleanup and quantified using Qubit (Thermo Fisher Scientific, Q32851). Libraries were normalized to 0.1 ng μl^{-1} and tagmented using the following reaction conditions: 7.56 µl of H_2O, 9 µl of tagmentation buffer 4× (Tris–HCl pH 7.5 (40 mM), $MgCl_2$ (20 mM) and DMF (N,N-dimethylmethanamide) (20%)), 1.44 µl amplicon tagmentation mix (Illumina, FC-131-1024) and 4 µl of normalized cDNA libraries. Tagmentation reaction was mixed, spun down and then added to a thermocycler at 55 °C for 10 min. After tagmentation, 2 µl of 2% SDS (sodium lauryl sulfate) was immediately added and incubated for 5 min to halt the reaction. To the tagmented cDNA reactions, 6 µl of Nextera primer pair mixes (0.5 µM) were added. Following the addition of primers, the following PCR was assembled: 25.38 µl of H_2O, 25.2 µl of Phusion buffer 5× (Thermo Fisher Scientific, F530L), 2.7 µl of dNTP mix 10 mM (NEB, N0447S), 0.72 µl of Phusion high-fidelity DNA polymerase 2 U μl^{-1} and added to the thermocycler with the following conditions: 72 °C for 3 min, 98 °C for 3 min, followed by 12 cycles of 98 °C for 10 s, 55 °C for 30 s and 72 °C for 30 s, followed by a final 72 °C extension for 5 min. Amplified final libraries were purified using 0.7× SPRIselect (Beckman Coulter, B23318) cleanup and quantified using Qubit (Thermo Fisher Scientific, Q32851) and Agilent high-sensitivity DNA kit for BioAnalyzer (Agilent, 5067-4626). Libraries were sequenced on an Illumina NovaSeq 6000, using paired-end 150 read lengths.

Smart-seq3 short-read processing workflow

Aligning and stitching UMI-containing reads for SIRV isoform reconstruction. We process Smart-seq3 SIRV Illumina paired-end reads closely following the procedure outlined in ref. 6. We processed raw nondemultiplexed FASTQ files using zUMIs v2.9.4g and STAR v2.5.4b to generate expression profiles for both the 5′ UMI-containing and internal reads. To extract and identify the UMI-containing reads in zUMIs, we specified find_pattern: ATTGCGCAATG for the 5′ read together with base_definition: cDNA (23–150), UMI (12–19) in the configuration YAML file and collapsed UMIs within a Hamming distance of 1. In total, we obtained 3.1×10^8 UMI-containing and 5.6×10^7 internal reads. Next, we proceeded to stitch UMI-containing reads together using stitcher.py[18] starting from the <prefix>.filtered.Aligned.GeneTagged.UBcorrected.sorted.bam output from zUMIs. To avoid UMI collision, we downsampled the aligned reads down to the 20% level before read stitching. We inferred the transcript compatibility set for each 5′ UMI-containing read from the CT tag in the produced BAM file. The most abundant transcript compatibility set was SIRV201, SIRV202 and SIRV205, which contained 7,161 unique UMIs, which is still substantially below the UMI space size $4^8 = 65{,}536$, justifying our chosen read downsampling level (Supplementary Fig. 17). In total, stitcher.py reconstructed 1.35×10^6 molecules. The median and interquartile range for reads/molecules were 8 and 24, respectively (Supplementary Fig. 18). Finally, we generated the transcript identification confusion matrix by iterating over all stitched 5′ reads, assuming a flat prior for both source and target transcripts, and accordingly dividing the assignment probability weight equally to all compatible source and target transcripts.

Quantification of SIRV isoforms. Following the recommendation discussed in ref. 6, we do not use UMIs to quantify isoform abundances. Instead, we used both 5′ UMI-containing and internal reads for quantification. To this end, we ran salmon v1.5.1 in quantification mode with additional arguments '--minAssignedFrags 1 -l IU' on the previously obtained <prefix>.filtered.tagged.Aligned.toTranscriptome.out.bam transcriptome alignments from zUMIs without any downsampling. We read the transcript per million (TPM)-normalized abundances from the salmon_quant/quant.sf output table.

MAS-ISO-seq processing workflow

Error correction. Error correction was performed on-board the PacBio Sequel IIe with the vendor's ccs software v5.0.0 (ref. 7) and settings '--all --subread-fallback --num-threads 232 --streamed <movie_name>.consensusreadset.xml --bam <movie_name>.reads.bam'. With these settings, all reads from the instrument (including those failing CCS correction) are presented in a single BAM[19] file for downstream analysis. Each read is affixed with an auxiliary BAM tag 'rq' indicating overall read quality ranging from $0 < rq < 0.99$ for CCS-corrected reads with predicted accuracy <Q20, $rq \geq 0.99$ for CCS-corrected reads with predicted accuracy ≥Q20, and $rq = -1$ for CCS-uncorrected reads[20].

Annotation/MAS-ISO-seq array filtration/segmentation/demultiplexing. We developed a composite hidden Markov model toolkit ('Longbow') to enable the per-read labeling of all subsequences of interest (annotation), allowing for insertions, deletions and mismatches in both low- and high-error rate data. This toolkit is based on the open-source hidden Markov model library, pomegranate[21]. Our hidden Markov model formulation considers a MAS-ISO-seq read to be a mosaic of imperfect (but complete) copies of the various known adapter sequences among which the unknown cDNA sequences of interest are present. Given a predefined array and cDNA structure, we combined several instances of the following two probabilistic models for pairwise sequence alignment: the Needleman–Wunsch and random alignment models[22]. Needleman–Wunsch model sections support annotation of sequences known a priori (for example, MAS-ISO-seq adapters; 10× Genomics single-cell 5′ and 3′ adapters). Two instances of the Needleman–Wunsch models were modified to account for expected sequence length (using duration modeling[22]) and used to model poly-A tails and sequences of known length but unknown content (that is, CBCs and UMIs), respectively. Random alignment model sections support annotation of unknown interstitial sequences (that is, cDNA sequences and unexpected nucleotide sequences resulting from sequencing or library construction errors/artifacts). All submodel termini are bi-directionally connected to a secondary random model, which may transition to any other Needleman–Wunsch model. This construction permits the hidden Markov model annotation to skip adapters erroneously absent from a read due to errors in array or cDNA synthesis for downstream filtering or examination.

The state transition diagram and default values for transmission and emission probabilities (used for all MAS-ISO-seq processing performed in this work) are provided in Supplementary Fig. 19. These defaults can optionally be refined using Longbow's train command, which will estimate the parameters of the model using Baum–Welch learning.

Data processing proceeds as follows: Longbow annotations are generated for both the forward- and reverse-complement orientations, retaining the result from the model with higher log-likelihood. Given the design expectation that MAS-ISO-seq adapters should be found in sequence along the length of the read, we verify that each read conforms to this expectation and filter out (via Longbow filter) any read with mis-ordered MAS-ISO-seq adapters. We then segment (via Longbow segment) each read between MAS-ISO-seq adapters and the 10× Genomics single-cell 5′ adapter. Finally, we filter (via Longbow sift) individual segmented reads by whether they conform to the structure of the expected library preparation (that is, the cDNA library itself). Longbow sift enforces that all expected regions in a segmented read are present (that is, a 10× Genomics single-cell 5′ adapter, a CBC, a UMI, the switch oligo leader sequence ('SLS'), cDNA, a poly(A) tail and a 10× Genomics single-cell 3′ adapter). We apply this model to each segmented read and retain those that match this model (Longbow's sift command; Supplementary Fig. 19).

For multiplexed libraries (for example, libraries with different array configurations and run on the same flow cell), the demultiplexing workflow proceeds similarly to the procedure described above with one notable change—annotations are generated for both the forward and reverse-complement read orientations and over each user-specified array design. The annotations from the read orientation and array design that maximize the overall log-likelihood are propagated to subsequent steps.

SIRV isoform alignment. To assign SIRV isoform to MAS-ISO-seq reads, we took reads (both CCS-corrected and CCS-uncorrected) that had been filtered, annotated and segmented by Longbow and annotated their UMIs. We then removed the adapter sequences and poly-A tails from these reads. The resulting reads were aligned to the SIRV-Set 4 transcriptome using minimap2 v2.17-r941 (ref. 23) with the HiFi read preset (minimap2 -ayYL --MD --eqx -x asm20).

SIRV confusion matrix construction. To generate the SIRV confusion matrix, we first followed the steps for SIRV isoform alignment. We then generated the transcript identification confusion matrix by iterating over all read alignments, assuming a flat prior for both source and target transcripts, and accordingly dividing the assignment probability weight equally to all compatible source and target transcripts.

Quantification of SIRV isoforms. To quantify SIRV isoforms from MAS-ISO-seq data, we first followed the steps for SIRV isoform alignment. We then took the primary alignments and removed any in which we could not detect a UMI as a quality control measure. Following the workflow discussed in ref. 6, we do not use UMIs to quantify isoform abundances, and instead, we use salmon v1.5.1 in the long-read quantification mode with arguments '--minAssignedFrags 1 --dumpEqWeights -l U --ont'. The motivation for this choice is twofold, which are as follows: (1) here our goal is to compare MAS-ISO-seq SIRV quantification with the matching Smart-seq3 short-read protocol (Fig. 1c,d). The authors of Smart-seq3 recommend using salmon for quantification, utilizing both 5′ UMI-containing and internal reads. Indeed we found that salmon quantification, compared to UMI-based quantification of stitched 5′ reads, substantially improved Smart-seq3 results. This is likely associated with the utilization of reliable sequencing bias models in salmon and the usage of internal reads; (2) the Smart-seq3 protocol uses a UMI length of 8 bp, which is long enough to avoid collisions when reads are stratified by CBCs in single-cell libraries. Our SIRV library, however, is too complex to allow avoiding UMI collision for several abundant ERCC transcripts, diminishing the utility of UMIs for quantifying the SIRV-Set 4 data.

CBC and UMI annotation. CBC (16 bp) and UMI (10 bp) sequence boundaries are approximately determined during read annotation with Longbow in accordance with the MAS-ISO-seq array design (Supplementary Fig. 19e). To ensure accurate boundary annotation for CBC error correction and UMI-based deduplication, additional post-processing considerations were applied as follows: first, putative CBC sequences were error-corrected against a list of expected barcodes (described below). Next, we aligned the error-corrected CBC to either the 80 bp (in the case of CCS-corrected reads) or 120 bp (in the case of CCS-uncorrected reads) on either end of each read using an accelerated Smith–Waterman algorithm, SSW (v1.2.4)[24], to determine the 5′ boundary between the CBC and UMI. We then aligned the 13 bp sequence between the UMI and the cDNA, the SLS (TTTCTTATATGGG), to the 46 bp (2 × (UMI length + SLS length)) beyond the end of the CBC alignment read using SSW. The UMI was then identified as the sequence between the end of the CBC and the start of the SLS, and each read was tagged accordingly. Note that the length of the resulting UMI sequences can deviate from the expected 10 bp due to indel sequencing errors, errors in oligo synthesis or a missing SLS. To handle the latter, we filtered out reads with SLS Smith–Waterman alignment scores below 10 and UMI lengths deviating from 10 bp by more than 3 bp for CCS-corrected reads and 4 bp for CCS-uncorrected reads.

In the case of SIRV data, no CBC was present in the library and, therefore, it was not annotated. The SIRV UMIs were similarly identified leveraging the structure of the array design. We first annotated each SIRV read with Longbow and then counted bases from the end of the forward adapter to annotate each read with the UMI.

CBC error correction. Correcting for potential CBC errors is a key step in single-cell data analysis, which we performed as follows. We first annotated each long read with a raw CBC as described earlier. We then padded the sequence of this raw CBC to include the adjacent 3 bp on either end. Next, we used a python implementation[25] of the SymSpell[26] symmetric delete spelling correction algorithm to correct all padded long read CBC sequences to a CBC whitelist identified from short-read data (sample 1, ~695,000 entries; sample 2, ~645,000 entries). We did so by sliding a 16 bp window across the padded CBC sequences and performing a lookup in the 10× CBC whitelist within a Levenshtein distance threshold of 2 for CCS-corrected reads and 3 for CCS-uncorrected reads for each such window. We then corrected the CBC to the 10× CBC sequence that had the lowest Levenshtein distance. In the event that no 10× CBC could be found within that Levenshtein distance or if multiple different 10× CBCs were found with the same minimum Levenshtein distance, the long read CBC was not corrected and the containing read was removed from further processing. We found that 97.2% and 96.3% of CCS-corrected reads and 72.2% and 71.12% of CCS-uncorrected reads (for sample 1 and sample 2, respectively) could be unambiguously corrected to a whitelisted CBC sequence. The lower CBC correction rate for CCS-uncorrected reads is expected given the conservative parameters deliberately chosen to minimize misassignment. We implemented this correction mechanism as the correct subcommand in Longbow.

Evaluating the accuracy of CBC identification and error correction. Assigning CBC to reads and correcting for potential sequencing or segmentation errors is a multistage process involving several parameter choices, as described earlier. The overall 'end-to-end' accuracy of CBC assignment can be effectively evaluated using species-mixing experiments[27,28]. Inspired by such experiments, we leveraged the presence of a small number of primary tumor cells in our sample (attributed to $CD3^{\pm}CD8^{+}$ FACS false positives) to evaluate the overall accuracy of MAS-ISO-seq CBC assignment as follows. First, we used short-read sequencing to identify high-purity tumor and immune CBCs. After removing doublets and potentially contaminated cells, we could identify 3,336 high-purity immune and 101 high-purity tumor CBCs, along with a set of genes exhibiting mutually exclusive expression patterns across immune and tumor cells. Our criterion for mutual exclusivity was TPM < 1 in tumor cells and TPM > 100 in immune cells, or vice versa. We could identify 121 immune-specific and 100 tumor-specific such genes. Our criterion for barcode purity was the sum total of total off-target UMIs to be ≤1. The median UMI per cell in our short-read data was ~4,000, so the on-target gene expression purity in our selected CBCs was >99.97%. Next, we studied the expression of the same genes in the same CBCs but in the MAS-ISO-seq data obtained from the same cDNA library. CBC misidentification, sequencing errors and inaccurate barcode error correction lead to the random shuffling of reads between tumor and immune cells. Therefore, off-target counts of tumor genes in immune cells and vice versa can be used to estimate the rate of CBC misassignment. We note that this strategy is practically similar to the 'capture–mark–recapture' method for estimating wildlife population sizes, where 'capturing' and 'marking' steps are done using high-fidelity short-read data, followed by 'recapturing' in MAS-ISO-seq data. Supplementary Fig. 12 shows a scatter plot of total tumor gene expression versus total immune gene expression in MAS-ISO-seq data for the predetermined set of high-purity barcodes using short-read data. Overall, we found 99.82% and 99.65% of reads assigned to tumor and immune cells to be on-target. These accuracy figures are slightly higher for CCS-corrected reads (99.86% and 99.80%, respectively) and only slightly lower for CCS-uncorrected reads (99.62% and 98.99%, respectively). Assuming that CBC errors occur at random with probability p_e (per read) and independently of transcript identity, the odds of CBC misassignment can be straightforwardly estimated as follows:

$$\frac{p_\mathrm{e}}{1-p_\mathrm{e}} \approx \frac{R_{\mathrm{T}\to\mathrm{I}*}}{R_\mathrm{T}}\frac{R_\mathrm{total}}{R_{\mathrm{I}*}} \approx \frac{R_{\mathrm{I}\to\mathrm{T}*}}{R_\mathrm{I}}\frac{R_\mathrm{total}}{R_{\mathrm{T}*}},$$

where $R_{\mathrm{T}\to\mathrm{I}*}$ and $R_{\mathrm{I}\to\mathrm{T}*}$ denote the total number of reads mapping to tumor and immune genes but misassigned to (a predetermined set of) immune and tumor CBCs, I^* and T^*, respectively; R_T and R_I denote the total number of reads in the library mapping to tumor and immune genes, respectively; $R_{\mathrm{T}*}$ and $R_{\mathrm{I}*}$ denote the total number of reads assigned to CBC sets T^* and I^*, respectively; R_total denotes the total number of sequenced reads. Using this formula, we obtain a CBC misassignment rate of 0.3–1.0% (or correct CBC assignment rate of 99.0–99.7%) using either off-target tumor or immune genes as the error estimator.

UMI error correction. Reads were first partitioned into groups, such that reads with the same CBC and transcript equivalence class (TEC; described in the following section "Quantification of 10× Genomics 5′ CD8+ T-cell isoform expression") were grouped together. UMI correction was then performed separately on each resulting read group. We formulated UMI correction as a minimum vertex cover problem on a bipartite graph $G = (T, S, E)$, where T and S are two disjoint and independent sets of nodes and E is the set of edges, constructed as follows (Supplementary Fig. 20a). Let R be the set of reads in a given group, we defined the set of target nodes T to consist of all unique three-tuple (UMI, cDNA length and GC content) combinations generated from the reads in R and the set of source nodes S to consist of all the reads in R. We then added an edge $(s, t) \in E$ between a source node s and a target node t if the following three conditions held: (1) the Levenshtein distance between the UMI of s and t was no greater than 2 for CCS-corrected reads and 3 for CCS-uncorrected reads, (2) the difference in cDNA length between s and t was no greater than 50 bp for CCS-corrected reads and 100 bp for CCS-uncorrected reads, and (3) the difference in GC content between s and t was no greater than 0.05 for CCS-corrected reads and 0.15 for CCS-uncorrected reads. The constraint parameters were selected to reflect the rate of indel sequencing errors and the empirical distributions of cDNA lengths and GC content of intragroup reads with identical UMIs. Under these constraints, an edge between a read and a target encoded the possibility that they represent the same molecule. Given the resulting graph, we applied an iterative greedy strategy to select the minimum subset of targets in T that cover all the read nodes in S. In particular, starting with the initial assignment, we iteratively chose the target in T with the highest degree (that is, the greatest number of supporting reads). The UMIs of the reads assigned to each selected target were then corrected to the UMI with the maximal support in the group (in the case of ties, priority was given to UMIs closer in length to the expected 10 bp). Post correction, reads with UMIs deviating from the expected length of 10 bp by more than 3 bp were filtered out. Such reads were found to be primarily missing either the UMI itself or the subsequent 13 bp switch leader sequence (TTTCTTATATGGG). Note this filtering criterion further restricted the admissible UMI lengths as compared to the precorrection UMI-based filtering. Supplementary Fig. 20b shows the reduction in the number of UMIs before and after correction at each locus.

Quantification of 10× Genomics 5′ CD8$^+$ T-cell isoform expression. To cross-annotate MAS-ISO-seq reads against a reference transcriptome (for example, GENCODE) and to obtain a single-cell isoform count matrix, we took reads (both CCS-corrected and CCS-uncorrected) that had been annotated, filtered and segmented by Longbow with CBCs and UMIs properly identified and error-corrected. We then extracted the cDNA bases from each read, thereby removing the library structure and poly-A sequences (implemented in longbow extract). These resulting extracted reads were aligned to a version of the GRCh38 human reference genome with alternate contigs removed (GCA_000001405.15_GRCh38_no_alt_analysis_set.fa) using minimap2 v2.24-r1122 with the splicing preset (for CCS-corrected reads: minimap2 -ayYL --MD --eqx -x splice:hq, for CCS-uncorrected reads: minimap2 -ayYL --MD --eqx -x splice).

We then filtered these aligned reads, removing unmapped reads, reads with secondary or supplementary alignments, reads with mapping quality of 0, reads with length >15 kb and reads with clipping on either end of length >1 kb. We then processed the resulting reads with StringTie2 v2.2.1 (ref. 12) using GENCODE v37 (ref. 29) as baseline transcript annotations to create new transcriptome annotations specific to each of our samples (stringtie -Lv -G gencode.v37.primary_assembly.annotation.gtf -o annotations.gtf -A gene_abund.out).

We developed a graph-based algorithm to accurately characterize and quantify isoform expression. First, we converted the aligned reads to genome interval annotations in gff format using spliced_bam2gff v1.3 (https://github.com/nanoporetech/spliced_bam2gff; spliced_bam2gff -S -M aligned_reads.bam > aligned_reads.gff). We then performed comparisons between the GENCODE transcriptome annotations, the new transcriptome annotations and the aligned read annotations using gffcompare v0.12.6 (ref. 30). These comparisons were as follows (base versus query): new transcriptome versus GENCODE, GENCODE versus new transcriptome, aligned reads versus GENCODE, aligned reads versus new transcriptome (gffcompare -V -r base.gff -s base.fasta query_gff_name). These comparisons resulted in relationships between each query interval and the intervals in the given base file.

We assembled a directed multigraph using the output of these comparisons where each node is a transcript or read. The edges represent the relationships produced by gffcompare between each node (including the gffcompare classification codes), and edge direction is the query node to the base node. We first assigned gene names to the new transcriptome reads by traversing edges between the new transcript nodes and the GENCODE nodes to create a set of gene name/classification code pairs. If this set of pairs contained a single pair with a classification code of '=', this was used as the gene name for the new transcript node. Otherwise, we assign a gene equivalence class to the new transcript node, which is composed of all accumulated gene name/classification code pairs (Algorithm 3 in Supplementary Note). Once these gene names/equivalence classes were identified, the new transcript nodes in the graph were updated with their new gene assignments. To assign gene names to each read, we followed similar steps but traversed edges from read nodes to both new transcript nodes and GENCODE nodes to create the pair set. To assign transcripts to each read, we performed a traversal similar to the gene assignment process for each read, but created transcript ID/classification code pairs. This resulted in TECs for each read. These equivalence classes either contain no transcript assignment, one transcript assignment (with classification code), or multiple transcript assignments (with classification codes; see Algorithm 4 in Supplementary Note). For additional details on the graph construction method, see Supplementary Note and Supplementary Fig. 10.

Following the assignment of a TEC to each MAS-ISO-seq read, we created a count matrix by tallying for each TEC and CBC combination how many unique UMI occurrences there were. Using this equivalence class formulation enabled improved and automated processing of isoform expression data compared to direct alignment to either GENCODE or StringTie2-derived transcriptomes (Supplementary Fig. 12).

Single-cell analysis

Short-read 10× Genomics 5′ gene expression and antibody capture preprocessing. We quantified the produced 5′ RNA capture and TotalSeq-C antibody capture libraries using Cell Ranger v3.1.0 count workflow. We imported the count data into AnnData format using scanpy v1.7.2 read_10x_h5 command. Our preliminary investigations indicated that the Cell Ranger automatic cell identification algorithm had used an excessively conservative cutoff, leading to the loss of 30–50% viable nonempty droplets (primarily of stem-like memory T cells origin, a cell type that exhibits relatively lower transcriptional complexity). As a countermeasure, we loaded the raw count data from Cell Ranger count output raw_feature_bc_matrix.h5 and kept every droplet expressing >500 unique genes and >80% nonmitochondrial genes. We performed a preliminary round of clustering and differential gene expression analysis using scanpy standard workflow[31]. We identified and removed nonimmune cell clusters of likely primary tumor origin. We additionally identified and removed doublets using scrublet v0.2.3. The estimated doublet rate was 14%, which is the expected figure for loading ~10,000 cells. Finally, we log-transformed the antibody capture counts and treated them as cell-level annotations for the rest of the analysis.

Long-read single-cell MAS-ISO-seq isoform expression preprocessing. As a first step, we converted the TEC-level UMI count matrix produced by the MAS-ISO-seq workflow to an AnnData object. During this conversion, additional metadata was added to the counts. Many of the new transcripts and genes discovered by StringTie2 could be unambiguously assigned back to a GENCODE annotation. In particular, new genes were assigned to known genes in GENCODE v37 if the new genes had transcripts overlapping exactly one unique gene in GENCODE v37. In addition, an interval list containing T-cell receptor genes[32] was cross-referenced, and transcripts found overlapping these intervals were accordingly marked. To harmonize the long- and short-read AnnData objects for joint analyses, we only kept the mutual CBCs between the two datasets. We could identify 100% of T-CBCs identified from the short-read dataset in the MAS-ISO-seq long-read dataset, indicating the high fidelity of our CBC error correction algorithm.

Normalization, clustering and embedding. We imported the harmonized short- and long-read AnnData objects to seurat v4.0.3 using SeuratData v0.2.1 and SeuratDisk v0.0.0.9015 helper packages[33]. We performed a negative binomial (NB) variance-stabilizing transformation (VST) on each count dataset separately using sctransform v0.3.2. We treated TEC counts similarly to gene counts, which is justified because TEC counts exhibit the same class of technical noise and statistical dropout as gene counts. Given the much larger number of TECs, we found it necessary to increase the number of TECs used for training the NB model from the default value of 2,000–10,000. We did not notice any substantial change in the downstream results by increasing this figure any further. The Pearson residuals for all cells and genes were exported to AnnData. We performed clustering and embedding separately for short- and long-read datasets using the same workflow as follows. We selected the top 5,000 genes (or isoforms) sorted in the descending order of total Pearson residual as highly variable features (HVFs). The HVFs were z-scored independently to equalize the role of each gene (or isoform). We reduced the feature set down to 30 using PCA and calculated the $k = 100$ nearest neighbor graph for each cell in the PCA space based on the Euclidean distance. The resultant neighbor graph was used for obtaining a 2D embedding using UMAP and clustering using the Leiden algorithm with a resolution parameter set to 1.1. We performed differential gene expression (DE) analysis on the short-read dataset based on t-test, which is an appropriate statistical test for VST counts, as implemented in scanpy rank_genes_groups method. The DE genes were used for annotating the clusters shown in Fig. 2 using known T-cell subtype markers.

Diffusion pseudotime (DPT) analysis. We performed DPT analysis closely following the scanpy hematopoiesis trajectory analysis workflow[34] with one notable modification. We noticed that using scaled highly variable Pearson residuals in place of log-transformed counts resulted in cleaner force-directed graphs. The latter is expected given that Pearson residuals are more Gaussian-like compared to log-transformed counts and, thus, better suited to the assumptions of the DPT model. Accordingly, we substituted the standard preprocessing and normalization step with the sctransform workflow.

Annotating CD45 isoforms. The GENCODE v37 human transcriptome reference contains a rather extensive set of isoform annotations for CD45, including the RO, RABC, RB, RBC and RAB. These annotations, however, are frequently incomplete and miss a large portion of the coding sequences. For instance, out of the available annotations for *PTPRC* (CD45), only two (ENST00000348564: CD45RO and ENST00000442510: CD45RABC) extend all the way to the 3′ UTR (Supplementary Fig. 11a). Given the primarily short-reads origin of currently available transcriptome annotations, we expect this caveat to prevail among most other genes, as also indicated by other authors[35,36]. Incompleteness and truncation of reference isoform annotations can turn into a source of quantification bias. For instance, when we attempted to directly align MAS-ISO-seq reads to the GENCODE v37 reference, we noticed that the aligner (minimap2) preferred the more complete and longer annotations, ENST00000348564 (CD45RO) and ENST00000442510 (CDRABC), as the primary alignment target for the vast majority of CD45 reads, irrespective of the differential inclusion or exclusion of shorter but biologically relevant exons such as the A, B and C. This strong alignment bias masks the alternative splicing pattern of CD45 expected in different T-cell subtypes (Supplementary Fig. 21). We found utility in refining GENCODE annotations using StringTie2, which resulted in the extension of several incomplete GENCODE annotations and improved the specificity of isoform assignments to different CD8$^+$ T-cell subtypes (Supplementary Fig. 11b).

Our proposed transcript annotation and quantification workflow combines the higher fidelity of StringTie2 transcript definitions with the diversity, naming convention and community consensus of transcript definitions published by the GENCODE consortium. As detailed earlier, we align MAS-ISO-seq reads to the reference genome using the splice-aware minimap2 aligner, cross-reference every read against both GENCODE and StringTie2 transcript definitions and assign a TEC to each read. Truncated reads, for example, due to strand invasion or internal poly-A priming, are typically assigned to richer TECs involving many compatible annotations whereas full-length reads are assigned to narrower classes (Supplementary Fig. 10). Depending on the nature of the desired downstream analysis, TECs involving unambiguous splice junction patterns can be readily identified and ambiguous TECs can be neglected. We applied this strategy to quantify CD45 isoforms and achieved substantially higher isoform assignment specificity compared to direct alignment to either GENCODE or StringTie2-derived annotations (Supplementary Fig. 11c). The results shown in Fig. 2b,c are based on this automated workflow.

Finally, the highest specificity in isoform assignment can be achieved by processing the genomic alignment of each read using a decision tree to bucket reads according to the presence/absence of manually specified landmark exons, for example, A, B and C in CD45 (Supplementary Fig. 11d). This quantification strategy, while achieving slightly higher specificity compared to our automated workflow, is not scalable or suitable for genome-wide studies. In summary, leveraging the increased read-depth afforded by MAS-ISO-seq, we find improving algorithms for de novo isoform identification and clustering, benchmarking the available isoform quantification pipelines (for example, FLAIR[36] and TALON[37]) and producing more complete transcriptome annotation references to be crucial areas of future method and resource development.

Quantifying the impact of the sequencing depth via downsampling. We quantify the impact of the sequencing depth gained by MAS-ISO-seq on cell type clustering and identification of DE and spliced genes by performing a series of in silico downsampling experiments from a single MAS-ISO-seq run. More explicitly, we took the set of all deconcatenated MAS-ISO-seq transcripts from 'sample 2' (37,164,708 full-length transcripts, comprising 22,613,229 CCS-corrected and 14,551,479 CCS-uncorrected reads; Supplementary Fig. 8) and randomly subsampled this set to obtain 1 M, 3 M, 5 M, 10 M, 20 M and 30 M transcripts. The standard Iso-Seq protocol yields 2–4 M HiFi reads (CCS-corrected and rq ≥ 0.99). This read-depth range is indicated in Fig. 2g with gray shading for reference. We processed each subsampled dataset identically using the same quantification, normalization, clustering and embedding workflow. We computed the ARI between the cell clustering of each subsampled long-read dataset and the full short-read dataset as follows. Because the appropriate cell clustering resolution is practically chosen in relation to biological considerations and varies with sequencing depth, we determined the cell clustering resolution for each downsampled long-read dataset by sweeping a Leiden clustering resolution range (0.5–2.0) and identifying the resolution that maximized ARI concordance with the fixed reference short-read cell clustering. We also determined the number of DE and DS genes across the T-cell subtypes for each downsampling run. The results of this analysis are shown in Fig. 2g and Supplementary Fig. 15.

Identification of DS genes. We consider two types of DS statistical tests for every expressed gene. In the global DS test, first, we wish to determine whether the isoforms of a given gene are DE in different cell clusters. To this end, we produce a contingency table with TEC counts and cell clusters as rows and columns, and with the aggregated isoform expression counts as entries. A nontrivial global DS pattern is equivalent to having a statistical dependence between the columns and rows of this contingency table. The latter can be canonically assessed using Fisher's exact test generalized to arbitrary $m \times n$ contingency tables with $m, n \geq 2$. Notably, we found the requirements for fast Chi-squared asymptotic approximation to be out of reach for the majority of cases. Therefore, we use the fisher.test as implemented in R v4.1.1 to perform the test using 1*e*6 permutations. In cluster-resolved DS test, we perform a cluster-resolved DS test for every gene, whereby we wish to know whether a gene exhibits differential isoform usage in each of the clusters versus the rest. Like before, we form a contingency table with two columns (the cluster of interest and the rest), with aggregated TEC counts in rows. We similarly obtained a *P* value by performing a permutation-based Fisher's test for every gene and every cluster. Finally, for both tests, we treat the obtained *P* values as a collection of independent hypotheses and adjust the *P* values for FDR at level $\alpha = 0.05$ using the Benjamini–Hochberg step-up procedure.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Links to the datasets used in this study can be found at https://github.com/broadinstitute/mas-seq-paper-data.
Human tumor-infiltrating CD8$^+$ T cells single-cell RNA-sequencing data are available from dbGAP with accession number phs003200.v1.p1.

Code availability

An online repository of code for the Longbow tool used in this study can be found at https://github.com/broadinstitute/longbow.

References

18. Larsson, A. J. & Sandberg, R. stitcher.py. Zenodo. https://doi.org/10.5281/zenodo.3765223 (2020).
19. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).
20. CCS Docs. What is in the reads.bam? https://ccs.how/faq/reads-bam.html (2022)
21. Schreiber, J. Pomegranate: fast and flexible probabilistic modeling in python. *J Mach Learn Res* **18**, 1–6 (2018).
22. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis* (Cambridge University Press, 1998).
23. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
24. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS ONE* **8**, e82138 (2013).
25. Garbe, W. (2012). SymSpell [Computer software]. https://github.com/wolfgarbe/SymSpell
26. Garbe, W. 1000x Faster spelling correction algorithm. https://gist.github.com/SebastiaanLubbers/8402454 (2012).
27. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
28. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
29. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
30. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Res* **9**, 304 (2020).
31. Wolf, A., Ramirez, F. & Rybakov, S. Preprocessing and clustering 3k PBMCs. Scanpy documentation. https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html. (2022)
32. HGNC. Gene group: T cell receptors (TR). https://www.genenames.org/data/genegroup/#!/group/370 (2023).
33. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
34. Wolf, A., Ramirez, F. & Rybakov, S. Trajectory inference for hematopoiesis in mouse. Scanpy documentation. https://scanpy-tutorials.readthedocs.io/en/latest/paga-paul15.html. (2022)
35. Glinos, D.A., Garborcauskas, G. & Hoffman, P. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359 (2022).
36. Tang, A. D. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1438 (2020).
37. Seki, M., Oka, M., Xu, L., Suzuki, A. & Suzuki, Y. Transcript identification through long-read sequencing. *Methods Mol. Biol.* **2284**, 531–541 (2021).

Acknowledgements

We thank W. Kretzschmar for the helpful discussions. This work was supported by Broad Institute SPARC awards 800353 (to A.M.A., K.V.G., N.H. and P.C.B.) and 800307 (to K.V.G.); National Institutes of Health grant (U19 AI082630 to N.H.), Adelson Medical Research Foundation (to N.H.), National Human Genome Research Institute grant (RM1HG006193 to N.H. and P.C.B.), with additional support from the Center for Cell Circuits at the Broad Institute (HG006193). M.A.S. is a Cancer Research Institute Irvington Fellow supported by the Cancer Research Institute (CRI award 4071).

Author contributions

A.M.A. conceived and developed the molecular workflow and designed and performed the experiments. K.V.G. developed the statistical annotation software with contributions from J.S. J.S. developed the data processing pipeline with contributions from V.P. and M.G. and performed bioinformatic analyses. M.B. performed Smart-seq3 and single-cell RNA-seq data analysis and statistical modeling and devised the isoform identification algorithm with contributions from J.S. V.P. developed the UMI and CBC error correction algorithms and conducted the bioinformatic analysis with contributions from A.M.A. S.S. aided through discussions and analysis. G.M.B., E.M.B. and M.S.F. consented patients, collected samples and processed and generated the 10× Genomics scRNA-seq data. M.A.S assisted with T-cell data analysis. M.C., A.D., T.B. and S.G. aided in the data generation and helped troubleshoot early iterations of the protocol. A.A.P. and E.B. provided access to cloud computing and other resources to facilitate data processing and analysis. A.M.A., K.V.G., J.S., M.B., V.P., P.C.B. and N.H. cowrote the manuscript.

Competing interests

The funding that contributed to the subject matter of this manuscript is described as follows: Broad Institute SPARC award, National Institutes of Health grant (U19 AI082630), Adelson Medical Research Foundation, National Human Genome Research Institute grant (RM1HG006193), support from the Center for Cell Circuits at the Broad Institute (HG006193) and Cancer Research Institute award (4071). A.M.A., K.V.G., J.S., M.B., P.C.B. and N.H. are inventors on a licensed, pending international patent application, having serial number PCT/US2021/037226, filed by Broad Institute of MIT and Havard, Massachusetts General Hospital and Massachusetts Institute of Technology, directed to certain subject matter related to the MAS-seq method described in this manuscript. Broad Institute of MIT and Harvard and Pacific Biosciences of California entered into a collaboration agreement relating to this research subsequent to the submission of this manuscript. A.A.P. is a venture partner and employee of GV. He has received funding from Verily, Microsoft, Illumina, Bayer, Pfizer, Biogen, Abbvie, Intel and IBM. M.S.F. receives funding from Bristol-Myers Squibb. G.M.B. has served on SAB and on the steering committee for Nektar Therapeutics. She has SRAs with Olink Proteomics and Palleon Pharmaceuticals. She served on SAB and as a speaker for Novartis. N.H. holds equity in BioNTech and is a founder and equity holder of Danger Bio. P.C.B. is a consultant to and/or holds equity in companies that develop or apply genomic or genome editing technologies: 10× Genomics, General Automation Lab Technologies/Isolation Bio, Celsius Therapeutics, Next Gen Diagnostics LLC, Cache DNA, Concerto Biosciences, Stately Bio, Ramona Optics, Bifrost Biosystems and Amber Bio. P.C.B.'s group receives research funding from industry for unrelated work. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01815-7.

Correspondence and requests for materials should be addressed to Aziz M. Al'Khafaji, Kiran V. Garimella, Mehrtash Babadi, Victoria Popic, Paul C. Blainey or Nir Hacohen.

Peer review information *Nature Biotechnology* thanks Omid Faridani and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Aziz M. Al'Khafaji, Kiran V Garimella, Mehrtash Babadi, Victoria Popic, Paul C. Blainey, Nir Hacohen

Last updated by author(s): March 4th, 2023

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ ☐ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

The software tools used to collect raw data from the sequencer and process it into counts / raw gene and isoform expression information are below. Additional details are located in the Methods section.

Smart-seq3 data processing
- Smart-seq3 Alignment and Stitching: zUMIs v2.9.4g, STAR v2.5.4b, stitcher.py v1.0
- Smart-seq3 Quantification of SIRV isoforms: salmon v1.5.1
- Smart-seq3 Cell Barcode annotation: Cell Ranger v3.1.0

PacBio error correction, MAS-seq data processing and analysis
- PacBio Read Error Correction: PacBio's ccs v5.0.0
- MAS-seq Annotation / Filtration / Segmentation / Demultiplexing / Cell Barcode Pseudocount Calculation / Cell Barcode Correction / UMI Correction: custom software - Longbow v0.6.4 (https://github.com/broadinstitute/longbow/releases/tag/v0.6.4)
- MAS-seq Read Data Processing Pipeline (SIRV and T cell, reads -> count matrix): custom software - https://github.com/broadinstitute/longbow/tree/main/wdl (commit hash: 9844452445cc803f814613e203eaf971814babbc)

Supporting analysis notebooks and scripts
Supporting analysis notebooks and scripts are listed below (custom software - https://github.com/broadinstitute/long-read-pipelines/tree/jts_covid_19_workflow I commit hash: 82bfe460778e56ac642942e62638d594f22bd6bf):
- MAS-seq QC analysis (e.g. ligation heat matrix plotting): MAS-seq_QC_report_template-static.ipynb
- MAS-seq Cell Barcode & UMI Annotation for SIRVs: docker/lr-10x/tag_mas_sirv_umi_positions.py
- MAS-seq Raw Cell Barcode & UMI Annotation for T cells: docker/lr-10x/tool.py, SSW v1.2.4
- Smart-seq3 Cell Barcode extraction: docker/lr-10x/extract_ilmn_bc_conf_scores.py
- Cell Barcode Count Merging: docker/lr-transcript_utils/python/merge_barcode_counts.py
- MAS-seq Isoform Expression Count Matrix Creation: docker/lr-transcript_utils/python/create_count_matrix_anndata_from_tsv.py

- MAS-seq Isoform Expression Preprocessing: docker/lr-transcript_utils/python/create_count_matrix_anndata_from_tsv.py

Cell barcode pseudocount calculation, cell barcode correction, UMI deduplication, transcriptome generation/comparison
- MAS-seq Read Alignment and Primary Read Filtration: minimap2 v2.17-r941, samtools 1.10
- MAS-seq Read UMI deduplication: UMI-tools v1.1.1
- Novel Transcriptome Annotation / Reference Generation: StringTie2 v2.2.1, gffread v0.12.7

Data analysis

The software tools used to process the gene and isoform expression data into additional results are below. Additional details are located in the Methods section.

Supporting analysis notebooks and scripts
Supporting analysis notebooks and scripts are listed below (custom software - https://github.com/broadinstitute/mas-seq-paper-data/tree/mb_revised_notebooks/ | commit hash: 0c166c63ce4d8ee37c835c89c9693bb65becb51c)
- An end-to-end reproducible workflow for the short-reads and long-reads scRNA-seq data analysis: scripts/scrna_seq_analysis
- The in silico downsampling power analysis workflow: scripts/downsampling_analysis

Quantification and single-cell analysis software
- Short-reads 5' 10x Gene Expression Quantification: Cell Ranger v3.1.0, scanpy v1.7.2, scrublet v0.2.3
- Gene & Isoform Expression Normalization / Clustering / Embedding: SeuratData v0.2.1, SeuratDisk v0.0.0.9015, sctransform v0.3.2, scanpy v1.7.2
- Diffusion Pseudotime Analysis: scanpy v1.7.2, sctransform v0.3.2
- CD45 Isoform Annotation Refinement: minimap2 v2.17-r941, StringTie2 v2.2.1
- Differentially Spliced Gene Identification: R v4.1.1 fisher.test

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The SIRV set 4 datasets generated and analyzed during the current study are available via the Broad Institute FTP server at gsapubftp-anonymous@ftp.broadinstitute.org:/MasSeqNatBiotech2021 .

Human peripheral blood monouclear cells (PBMC) / tumor-infiltrating CD8+ T cells single-cell RNA sequencing data are available from dbGAP with accession number phs003200.v1.p1.

All other data supporting the findings of this study are available from the corresponding authors on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No sample size calculation was performed.

Data exclusions

Data were not excluded from analysis except as described in the Methods section, which includes filtering steps for excessive read lengths, malformed MAS-seq arrays, doublet removal, and filtering non-immune cell clusters of likely primary tumor origin.

Replication

We applied the MAS-ISO-seq method independently on three samples (two single-cell samples consisting of tumor-infiltrating CD8+ T cells and one bulk RNA sample consisting of a Spike-in RNA Variant Control Mix) and found similar improvements in output yield in all three cases.

Randomization

Not applicable to this study; does not include subjects that require allocation into experimental groups.

Blinding

Not applicable to this study; does not include subjects that require allocation into blinded experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used	TotalSeq-C antibody mix: TotalSeq - C0048 anti-human CD45 Antibody (Biolegend; Cat# 368545), TotalSeq - C0103 anti-mouse/human CD45R/B220 (Biolegend; Cat# 103272), TotalSeq - C0087 anti-human CD45RO (Biolegend; Cat# 304259), and TotalSeq - C0063 anti-human CD45RA (Biolegend; Cat# 304163).
Validation	Each lot of these antibodies was quality control tested by immunofluorescent staining with flow cytometric analysis and the oligomer sequence was confirmed by sequencing. TotalSeq™-C antibodies are compatible with 10x Genomics Chromium Single Cell Immune Profiling Solution, and were purchased from Biolegend.

Human research participants

Policy information about studies involving human research participants

Population characteristics	The two samples used in this study are from male patients age 31 and 59 with metastatic melanoma treated with checkpoint blockade therapy as part of standard of care treatment. Both samples used in this study are from male patients diagnosed with cutaneous melanoma.
Recruitment	All patients analyzed in this study provided written informed consent for the collection of tissue and matched normal blood samples for research and genomic profiling, prior to sample collection and processing. No compensation was provided to the participants.
Ethics oversight	This study was approved by the Dana-Farber/Harvard Cancer Center Institutional Review Boad (DF/HCC Protocol 11-181)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Using the human tumor dissociation kit (Miltenyi Biotec; Cat# 130-095-929), freshly isolated tumors were digested to obtain a single cell suspension. Tissue was placed into a 1.5mL Eppendorf tube containing 420µL of DMEM with 10% FCS, 42µL of enzyme H, 21µL of enzyme R, and 5µL enzyme A (provided with the kit). The tissue was minced using surgical scissors, and an additional 512µL of DMEM with 10% FCS was added to the tube (total volume of 1ml). Next the tissue was incubated for 15 min at 37°C, 350 rpm in a thermomixer (Eppendorf; F1.5). After incubation, the tissue was further digested using a 1 ml syringe plunger over a 50µm filter (Sysmex; Cat# 04-004-2327), making sure to wash the filter with media. Using ACK buffer (Gibco; Cat# A1049201), RBC lysis was performed and the sample was finally resuspended in DMEM with 10% FCS in order to count and determine the viability of the cells using a manual hemocytometer (Bright-line; Cat# 1492).
Instrument	Cell sorting was done on a Sony MA900 cell sorter, model LE-MA900FP
Software	Cells were collected using the SONY cell sorter software

Cell population abundance	The fraction of CD3+CD8+ T cell prior to sorting was 14-17%. Post sorting, cell in sorted tubes were reanalyzed on the machine using the same gating plots that were used for the actual sorting to evaluate purity. Purity post sorting ranged from 96-98%.
Gating strategy	Sorting of single live CD3+CD8+ T cells (gating on Zombie-low, hCD235a-, hCD45+, hCD3+, hCD8+) was performed using a Sony MA900 cell sorter.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.