Optimizing Single-Cell Long-Read Sequencing for Enhanced Isoform Detection in Pancreatic Islets

Maria S. Hansen, Christopher J. Hill, Lori Sussel, and Kristen L. Wells

Diabetes 2026;75:606–616 | https://doi.org/10.2337/db25-0424

TECHNOLOGICAL ADVANCES

Alternative splicing is an essential mechanism for generating protein diversity by producing distinct isoforms from a single gene. Dysregulation of splicing that affects pancreatic function and immune tolerance has been linked to both types 1 and 2 diabetes. Next-generation sequencing technologies, with their short read lengths, are limited in their ability to accurately detect splice variants. Long-read sequencing technologies offer the potential to overcome these limitations by providing full-length transcript information; however, their application in single-cell RNA sequencing has been hindered by technical challenges, including insufficient read lengths and higher error rates. Furthermore, cell types that produce high levels of a single transcript, such as islet endocrine cells, can obscure identification of lower-abundance transcripts. In this study, we optimized a protocol for single-cell long-read sequencing in pancreatic islets to improve read length and transcript detection. Our findings demonstrate that 5′ library preparation protocols outperform 3′ protocols, resulting in better transcript identification. Furthermore, we show that targeted depletion of insulin transcripts enhances the detection of informative reads, highlighting the utility of transcript-depletion strategies. This optimized protocol enables isoform-specific gene expression analysis and reveals differential transcript usage across the various cell types in pancreatic islets. By leveraging this approach, we gain deeper insights into the transcriptomic complexity and cellular heterogeneity within pancreatic islets.

ARTICLE HIGHLIGHTS

- This study addresses the limitations of current single-cell long-read RNA sequencing technologies in detecting full-length transcripts and isoform diversity, particularly in pancreatic islets.
- We demonstrate that optimizing single-cell library preparation protocols reproducibly enhances read length and transcript identification in pancreatic islets.
- Combined with targeted insulin depletion and extended reverse transcription, 5′ capture methods significantly improved read length and isoform detection compared with standard protocols, while maximizing the number of informative reads.
- These improvements yield longer reads in single-cell experiments, substantially enhancing transcript identification and enabling more accurate analysis of isoform diversity.

Alternative splicing plays a critical role in generating protein diversity from the ∼22,000 known protein-coding genes, leading to the production of >140,000 distinct transcripts (1). This process allows for the generation of proteins with different amino acid sequences, affecting their functions and localization within the cell and allowing them to respond readily to changes in the environment (2,3). Splicing dysregulation is a key factor in many diseases, including diabetes, either as a result of inherent mutations in splice sites or RNA binding proteins or in response to changes in environmental conditions, such as inflammatory stress or hyperglycemia (4,5). In the context of type 1 diabetes, diversity in isoform expression has significant implications for pancreatic function and immune tolerance. For example, differential isoform expression of autoantigens IA-2 and G6pc2 between the

Barbara Davis Center, University of Colorado Anschutz Medical Campus, Aurora, CO

Corresponding author: Kristen L. Wells, kristen.wells-wrasman@cuanschutz.edu

Received 30 April 2025 and accepted 5 January 2026

This article contains supplementary material online at https://doi.org/10.2337/figshare.31008040.

pancreas and thymus has been suggested as contributing to the generation of autoreactive T cells in type 1 diabetes (6,7). Furthermore, dysregulated splicing events have been observed in islets from individuals with type 2 diabetes, underscoring the importance of splicing regulation in maintaining proper cellular function and immune homeostasis (5). As one specific example, SNAP-25, a component of the SNARE complex responsible for vesicle fusion and exocytosis, exists in two isoforms (SNAP-25a and SNAP-25b). In SNAP-25b–deficient mice, Ca^{2+} elevations are prematurely activated and delayed in termination, and insulin secretion is increased (8).

Despite the critical need to detect splice variants in the context of diabetes, next-generation sequencing (NGS) technologies remain insufficient for this task. Identifying isoform-specific gene expression requires sequencing reads that span multiple exons of the mRNA transcript. In the human genome, transcript lengths are estimated to average between 1,800 and 4,900 bp, with the mode of distribution ∼2,000 bp (9). NGS technologies have read lengths of 150 bp, making it difficult to identify isoforms. In contrast, long-read sequencing technologies, such as PacBio and Oxford Nanopore Technologies, offer the generation of full-length reads that can capture the full RNA molecule, thereby providing a clearer picture of isoform diversity. Long-read sequencing has uncovered thousands of previously unannotated mRNA isoforms, many of which encode proteins with distinct functions, highlighting alternative isoforms as a major and still-expanding source of proteomic and regulatory diversity (10). Published single-cell long-read RNA sequencing (sclrRNA-seq) libraries often report shorter read lengths than expected, which may limit transcript coverage and isoform detection. For instance, a recent study reported a median read length of 900 bp for sclrRNA-seq of two cancer cell lines (11).

Advances in sequencing technologies, especially single-cell approaches, have revealed the complex heterogeneity within the pancreas, uncovering distinct functional and transcriptomic subpopulations across different cell types. In pancreatic islets, single-cell genomics and Patch-seq have identified transcriptionally and functionally distinct β-cell subpopulations directly linking gene expression to key physiological processes, such as vesicle exocytosis (12). This heterogeneity underscores the importance of characterizing splicing events and their resulting isoforms at the single-cell level. However, sclrRNA-seq technologies come with inherent limitations. Nanopore flow cells produce fewer reads than Illumina, with ∼20,000 reads per cell for a 5,000-cell experiment, well below the typical 30,000–50,000 reads per cell common in NGS. Moreover, Nanopore's higher error rate (1%) increases the likelihood of incorrect barcode and unique molecular identifier (UMI) assignments. To overcome these challenges, we have optimized a protocol for pancreatic islets that improves read length, advancing the utility of long-read sequencing in single-cell transcriptomics.

RESEARCH DESIGN AND METHODS

Dissociation of Pancreatic Islets

Female C57BL/6 mice aged 10 weeks were obtained from The Jackson Laboratory. Pancreatic islets were isolated from mice under ketamine–xylazine–acepromazine anesthesia by collagenase delivery into the pancreas via injection into the bile duct. The collagenase-inflated pancreas was surgically removed and digested. After isolation, islets were dissociated using Accutase in a 37°C bead bath for 25–30 min. Single-cell suspension was filtered through a 40-mm filter and quenched in RPMI medium plus 10% FBS. Cells were washed again with RPMI plus 10% FBS and with PBS plus 0.1% BSA. Single-cell suspensions were loaded into a Genomics Chromium instrument targeting 4,000 cells per sample.

Dissociation of Spleens

Spleens were isolated from mice under ketamine–xylazine–acepromazine anesthesia. Spleens were dissociated through a 70-μm strainer in complete Iscove's modified Dulbecco's medium (cIMDM) using a 3-mL syringe plunger. Cells were washed, centrifuged, and treated with 1 mL ammonium–chloride–potassium lysis buffer for 30 s, followed by dilution in cIMDM and a second spin. After one additional cIMDM wash, cells were resuspended in PBS plus 0.1% BSA. Single-cell suspensions were loaded into a Genomics Chromium instrument targeting 4,000 cells per sample.

scRNA-seq Library Preparation and Insulin Depletion

Single-cell libraries were prepared using either the Chromium Next GEM Single Cell 3′ Kit (version 3.1) or the Chromium Next GEM Single Cell 5′ Kit (version 2) following the protocol up to and including step 2.4, stopping just before fragmentation. For the optimized libraries, the following modifications were applied to the 5′ library preparation: 1 μL 10 mmol/L deoxynucleotide triphosphate (dNTP) solution (cat. no. FERR0191; Thermo Fisher Scientific) was added to the reaction in step 1.1; the extension time was increased from 45 min to 2 h in step 1.5; and 1 μL 10 mmol/L dNTP solution was added to the reaction in step 2.2, and the extension time was increased from 1 to 3 min.

Insulin depletion was performed on cDNA from step 2.4 of the 10x Genomics Chromium library preparation using the DepleteX RNA Depletion Panel (Insulin) Kit from Jumpcode Genomics. We followed the PacBio MAS-IsoSeq protocol (December 2022; version 1.0), with the following modifications: during ribonucleoprotein complex formation (step A), we used 0.9 μL Cas9 instead of 2.3 μL and 1.6 μL insulin guide RNA instead of 4.0 μL single-cell boost guide RNA, and during bead cleanup (step D), we used 50 μL (1×) AMPure XP beads instead of 75 μL 1.5× SMRTbell cleanup beads.

After insulin depletion, long-read libraries were prepared from the cDNA using the Ligation Sequencing Kit (version 14; cat. no. SQK-LSK114; Nanopore) and the

PCR Expansion (cat. no. EXP-PCA001; Nanopore). For 3′ libraries, the "Ligation Sequencing V14—Single-Cell Transcriptomics With 3′ cDNA Prepared Using 10x Genomics on PromethION (SQK-LSK114)" protocol was used. For 5′ libraries, the "Ligation Sequencing V14—Single-Cell Transcriptomics With 5′ cDNA Prepared Using 10x Genomics on PromethION (SQK-LSK114)" protocol was used. Short fragment buffer was used for library preparation instead of long fragment buffer. Library beads were used for the flow cell priming mix instead of library solution. Libraries were sequenced on R (version 10.4.1) flow cells on either the PromethION 2 Solo or PromethION 2 Integrated device.

Bulk RNA-seq Library Preparation

Islets were isolated from two male C57BL/6J mice as described above. RNA was purified using the Qiagen RNeasy Micro Kit (cat. no. 74004). Bulk RNA-seq libraries were prepared using the KAPA mRNA HyperPrep Kit (cat. no. 8098123702; Roche), following the KAPA mRNA HyperPrep Kit protocol (version 7.21; cat. no. KR1352; KAPA Biosystems). Insulin depletion was applied between steps 8 and 9 in the KAPA library preparation protocol.

scRNA-seq Preprocessing

Single-cell libraries were processed using the EPI2ME single-cell workflow (version 1.1.0) from Oxford Nanopore Technologies for the identification of cell and UMI barcodes (https://github.com/epi2me-labs/wf-single-cell) using the mm10 2020-A mouse reference provided by 10x Genomics. The kit name was either 3′ or 5′ depending on the 10x library preparation, and we provided kit versions 3 and 1 for the 3′ and 5′ samples we prepared, respectively. We provided the kit version that was indicated in the publications associated with each published data set. All parameters can be found on GitHub (https://github.com/CUAnschutzBDC/sc-islet-longread-analysis/blob/61d52b6ef94ff974a394a9335dd65d28b9e6360e/01_wf_single_cell/trial_parameters.tsv). The pipeline was altered to bypass the StringTie step to use a consistent GTF file from the mm10 genome. These jobs were executed using Singularity and Nextflow. Only the reads that had a sequence quality higher than seven were input into this pipeline.

We manually created a gene-count matrix from the BAM outputs that were produced by the EPI2ME pipeline and then used the emptyDrops function from the DropletUtils package (13) to identify cell barcodes. Using these barcodes, we subset the gene- and transcript-count matrices from the single-cell workflow and calculated the read lengths of cells that were identified using the barcodes. Read lengths were calculated from the BAM files by looking at the length of the query sequences for all primary reads and were subset further by matching barcodes, whether the read was tagged with a gene, and whether the read was tagged with a transcript. Finally, we examined the BAM files and calculated the number of reads that were tagged with a gene and a transcript, before creating the final output plots.

Detailed Analysis of the 5′ Modified Islet Data Set

The filtered data for the 5′ modified islet data set was input into Seurat (version 4.1.3) (14–18) using R (version 4.2.3). Quality control metrics were calculated using the perCellQCMetrics function from the Scuttle package (version 1.8.4) (19). We used the DoubletFinder package (version 2.0.3) (20) to remove cells that were identified as doublets, before running the principal component analysis. We tested several resolutions of uniform manifold approximation and projection before generating final clustering.

Cell types were initially identified using a reference data set (21) and further refined based on the expression of known islet cell type markers. Dimensionality reduction and clustering were first performed using gene-level expression data, followed by a second round of clustering based on transcript-level expression. To evaluate the consistency between gene- and transcript-based cell type annotations, we calculated concordance between the two methods and visualized the results as confusion matrices using the Matrix (version 1.5-4) (22) and pheatmap (version 1.0.12) packages. Clusters identified by transcript-level expression were used for downstream analysis. To assess differential splicing, we conducted pseudobulk differential transcript usage (DTU) analysis using DTUrtle (version 1.0.2) (23), alongside pseudobulk differential gene expression (DGE) analysis. Pseudobulk count matrices were generated by aggregating counts per replicate and cell type using Seurat. Differential expression was tested at the gene level using DESeq2 (version 1.38.3), and DTU was tested at the transcript level using DTUrtle with sparseDRIMSeq (version 0.1.2) for model fitting. Batch effects across replicates were corrected using ComBat-seq from the sva package (version 3.50.0). Most figures were generated using the dplyr (version 1.1.0), ggplot2 (version 3.4.1), readr (version 2.1.4), tidyverse (version 2.0.0), MetBrewer (version 0.2.0), and scAnalysisR (version 0.0.0.9000) (24) packages. Additional details of this pipeline can be found at our code repository (https://github.com/CUAnschutzBDC/sc-islet-longread-analysis). Docker images for RNA-seq preprocessing (STAR, Fastqc, cutadapt, and featurecounts; https://hub.docker.com/r/kwellswrasman/rnaseq_general) and scRNA-seq (DTurtle, scAnalysisR, and Seurat; https://hub.docker.com/r/kwellswrasman/sclrseq_methods_r_docker) are also available.

Bulk RNA-seq Analysis

Raw reads were trimmed using cutadapt (version 4.8) and python (version 3.10.14) (25) and aligned to the mouse GRCm38 genome using STAR (version 2.7.11b) (26). Gene expression was quantified with featureCounts (version

2.0.6) (27), and differential expression analysis was conducted using DESeq2 (version 1.44.0) (28) with R (version 4.4.1). All plots were made using ggplot2 (version 3.5.1) in R.

NGS Coverage

The 10k Human DTC Melanoma, Chromium GEM-X Single Cell 5′, and 10k Human DTC Melanoma, Chromium GEM-X Single Cell 3′, aligned BAM files were downloaded from 10x Genomics (https://www.10xgenomics.com/datasets). Coverage plots were generated using NGSplot (version 2.63) (29).

Transcript Coverage

To visualize the transcript coverage for each of our samples, we used the coverage function from the GenomicFeatures (version 1.60.0) package. The reads output from our single-cell postprocessing pipeline were aligned to a transcriptome reference generated from the GRCm38 Ensembl GTF and subset based on matching barcodes and UMIs from our processed Seurat objects. For each sample, a coverage matrix of counts per transcript position was created for every transcript in the GTF file, with each transcript normalized to a length of 100. The counts were then normalized by read depth and then again by applying a softmax transformation to generate matrix values that summed to 1 across each transcript, before a column mean provided a value per each 1–100 position per sample. These coverages were subset based on the number of exons per transcript, supplied by the GTF file, as well as the length of the transcripts, with a prerequisite that the transcript have more than one exon.

Gviz Plots

To visualize our transcriptome-aligned reads on the genome, we use Gviz (version 1.52.0) and the BAM files output from our single-cell postprocessing pipeline. Briefly, reads for each sample were aligned to a transcriptome reference generated using the GRCm38 Ensembl GTF file and subsequently subset based on matching barcodes and UMIs from our processed Seurat objects. The final output BAM files were generated by subsetting the genome-aligned BAM files to only those reads that uniquely mapped to one transcript in our transcriptome alignment. These BAMs were then merged by replicate. For each of these merged BAM files, an alignments track was created using Gviz and displayed alongside the GRCm38 transcripts.

Overrepresentation Analysis

DTU genes were subjected to KEGG pathway overrepresentation analysis using gprofiler2 (version 0.2.1) (30) with a custom background of all expressed genes. KEGG pathways were extracted from the enrichment results and ranked by *P* value. The top 25 pathways were visualized as bubble plots.

3′ Versus 5′ Bias Plot

Isoform-defining splice junctions were identified across data sets. Transcripts were classified based on whether their unique junctions fell in the 3′ region (last 25% of the transcript), 5′ region (first 25% of the transcript), central (middle 50% of the transcript), or both 5′ and 3′ and binned accordingly.

Data and Resource Availability

All data are available on Gene Expression Omnibus: islet scRNA-seq (GSE295353), spleen scRNA-seq (GSE295352), and islet bulk RNA-seq (GSE295351). All source data supporting the findings of this study are provided. Custom analysis pipelines are available on GitHub (https://github.com/CUAnschutzBDC/sc-islet-longread-analysis).

RESULTS

Evaluation of Read Lengths and Isoform Detection in Published Single-Cell Long-Read Data Sets

We aimed to identify isoform differences between islet cell types and subtypes using scRNA-seq. To evaluate the ability of single-cell sequencing technologies to generate full-length reads, we reanalyzed previously published sclrRNA-seq data sets generated using 10x Genomics and Oxford Nanopore Technologies, focusing on their ability to capture full-length transcripts and detect isoform-specific transcript expression. Our analysis included seven sclrRNA-seq libraries from five different studies (11,31–34). The reanalysis revealed an average read length of 794 bp and an average mode of 582 bp, compared with the expected mode distribution of ∼2,000 bp in the human genome (9) (Fig. 1*A*). This discrepancy between the average read length and the expected transcript length underscores the ongoing challenge of capturing full-length transcripts. This shortfall in read length is important because it limits the transcript detection ability. Where gene detection ranges from 60% to 75% of total reads, transcript detection ranges from 30% to 60% of total reads (Fig. 1*B*). These findings highlight the limitations of current sclrRNA-seq technologies in achieving comprehensive transcript-level resolution.

Efficient and Specific Depletion of Insulin From Islet Sequencing Libraries Generates Enhanced Read Diversity

Analyzing transcript expression requires a higher overall read depth than gene expression analysis, because each gene is associated with multiple transcripts. Initial analysis of our sclrRNA-seq libraries of mouse pancreatic islets led to the discovery that the two mouse insulin genes, *Ins1* and *Ins2*, made up 25% of the total reads, impeding our ability to achieve optimal read depth (Fig. 1*C*). To overcome this issue, we incorporated an insulin-depletion step into the protocol and validated the specificity and efficiency of the depletion in a bulk short-read RNA-seq library of mouse pancreatic islets. The depletion was remarkably efficient and highly specific; insulin transcripts

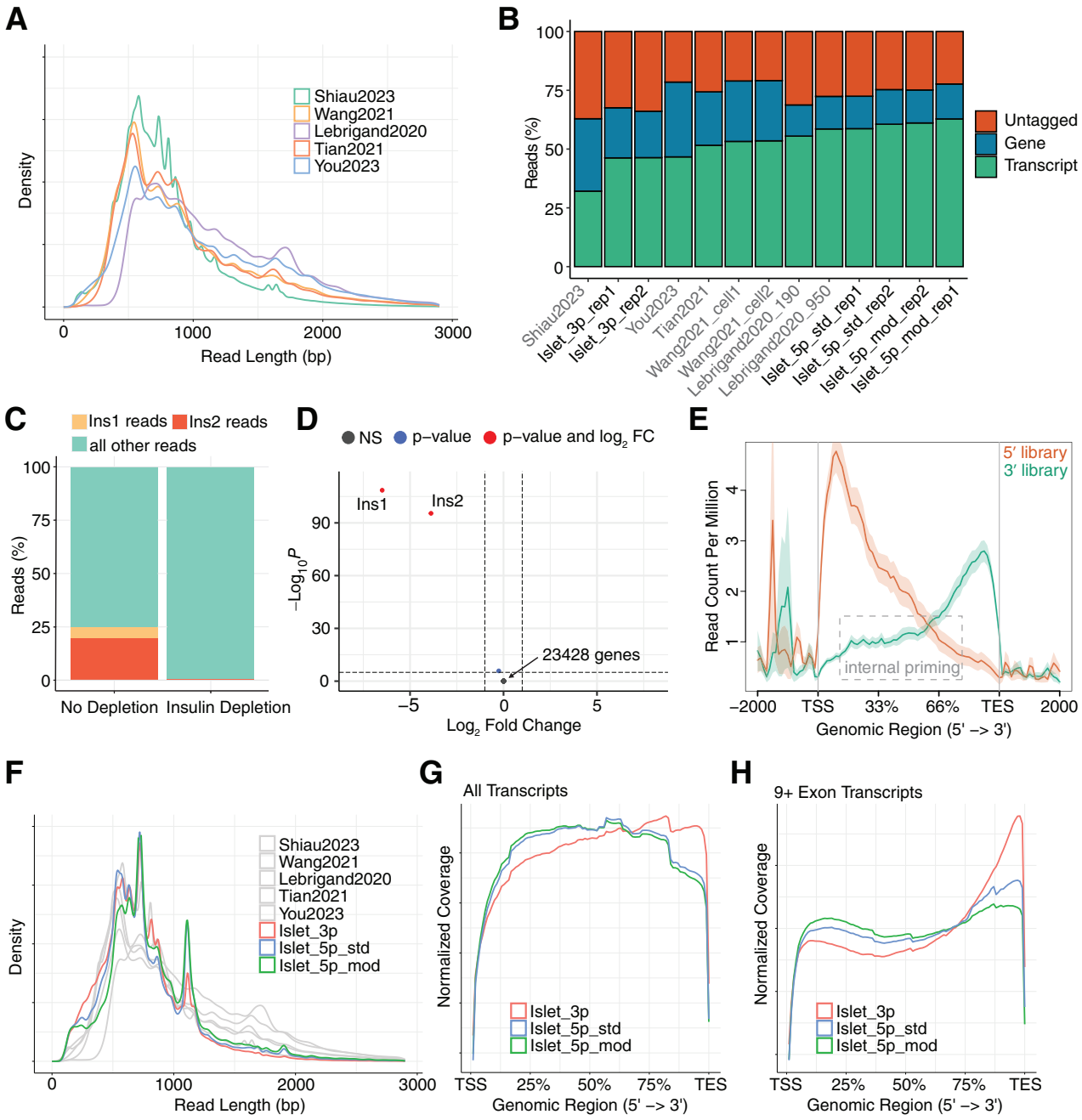

Figure 1—Read length and transcript identification comparison between sclrRNA-seq libraries. *A*: Read length distribution of published sclrRNA-seq libraries prepared with 10x Genomics and Nanopore technologies. Biological replicates are included for data sets from Lebrigand et al. (31) and Wang et al. (33); other data sets are shown as single samples. *B*: Proportion of reads across data sets where the gene is identified, the transcript is identified, or neither is identified. Published reanalyzed data sets and six mouse pancreatic islet samples: two replicates prepared with 3′ 10x Genomics technology, two with 5′ 10x Genomics technology, and two with 5′ modified 10x Genomics technology (incorporating library preparation optimizations). *C*: Proportion of reads aligned to *Ins1* or *Ins2* in scRNA-seq analysis of mouse pancreatic islets pre- and post–insulin depletion. *D*: Volcano plot depicting DGE between nondepleted and insulin-depleted bulk RNA-seq libraries from mouse pancreatic islets. Total of 23,431 genes are shown in the plot. *E*: NGS coverage plot indicating read start sites across the genomic region. Libraries shown are 3′ and 5′ single-cell 10x Genomics preparations derived from human DTC melanoma cells. *F*: Read length distribution comparison across mouse pancreatic islet sclrRNA-seq libraries, including two replicates prepared with 3′ 10x Genomics technology, two with 5′ 10x Genomics technology, and two with 5′ modified 10x Genomics technology. *G*: Coverage plot showing Softmax-normalized read coverage across relative transcript positions for all multiexon transcripts in mouse pancreatic islet scRNA-seq libraries prepared with 3′, 5′, or 5′ modified 10x Genomics technology. *H*: Coverage plot showing Softmax-normalized read coverage across relative transcript positions for all transcripts with nine or more exons in mouse pancreatic islet scRNA-seq libraries prepared with 3′, 5′, or 5′ modified 10x Genomics technology. NS, not significant; TES, transcription end site; TSS, transcription start site.

were uniquely depleted, whereas all other genes remained unaffected (Fig. 1*D* and Supplementary Table 1). The same insulin depletion was then applied to a single-cell pancreatic islet library followed by long-read Nanopore sequencing. Importantly, the insulin depletion was as efficient as in the bulk sample (Fig. 1*C*). This strategy was applied to all subsequent pancreatic islet libraries generated for this study. Notably, the classification of β-cells does not rely on the presence of insulin transcripts. We demonstrated this by computationally removing insulin reads from an scRNA-seq library and repeating cell clustering (Supplementary Fig. 1*A–D*).

Protocol Modifications Enhance Read Length and Transcript Identification in Islet Single-Cell Long-Read Libraries

Most high-throughput sclrRNA-seq methods rely on 10x Genomics single-cell capture and library preparation, which were originally optimized to generate and amplify shorter sequences, raising the question of whether they can effectively amplify full-length transcripts. 10x Genomics offers two types of transcriptomic profiling for scRNA-seq: one that captures the 3′ end of transcripts and another that captures the 5′ end. Studies have shown that 3′ RNA libraries frequently contain internal priming artifacts (35) that would prevent the amplification of full-length reads. To test for internal priming in 3′ versus 5′ libraries, we downloaded libraries generated using each technology in human melanoma samples from data sets created by 10x Genomics and analyzed the genomic coverage (29). A notably higher degree of internal priming was exhibited by 3′ libraries compared with 5′ libraries, as evidenced by an increased number of reads mapping to the central regions of transcripts in genomic coverage plots (Fig. 1*E*). Because reads generated through internal priming cannot span the full length of a transcript, this phenomenon likely contributed to the shorter read lengths observed in these libraries. To test this, sclrRNA-seq libraries were prepared from mouse pancreatic islets in parallel using both 3′ and 5′ capture technologies ($n = 2$ independent biological replicates per capture method). The libraries prepared with 5′ technology resulted in a subtle increase in longer reads, compared with the 3′ libraries ($P < 2 \times 10^{-16}$ by one-sided Wilcoxon test) (Fig. 1*F* and Supplementary Fig. 2*B*). Furthermore, the 5′ library provided substantially improved transcript identification, increasing from 46.4% ± 0.1% to 59.7% ± 1.3% (Fig. 1*B*).

To further improve the read length, several additional optimization steps were introduced into the islet 5′ library preparation protocol (Chromium Next GEM Single Cell 5′ Reagent Kits [version 2]), including increasing the extension time from 45 min to 2 h during GEM-RT incubation and from 1 to 3 min during cDNA amplification, based on the approach outlined by Lebrigand et al. (31), and increasing the amount of dNTP. These modifications resulted in longer reads than those from the 5′ library without modifications ($P < 2 \times 10^{-16}$ by one-sided Wilcoxon test) (Fig. 1*F* and Supplementary Fig. 2*B*) and enabled far better transcript identification than any of the published data sets (62.0% ± 1.2%) (Fig. 1*B*). Additionally, the 5′ modified libraries showed improved transcript coverage (Fig. 1*G* and Supplementary Fig. 2*C*), with the most pronounced improvement observed in longer transcripts (Fig. 1*H* and Supplementary Fig. 2*D*). Comparison of samples showed higher correlation between replicates generated using the same capture method than among samples processed within the same batch, indicating that the data are highly reproducible and that variability is primarily driven by capture technique (Supplementary Fig. 3*A*). Overall, this emphasizes the preference for 5′ over 3′ capture and highlights the necessity for library preparation optimizations to enhance the amplification of full-length reads.

Isolating high-quality RNA from pancreatic islets is notoriously difficult, primarily because of the presence of digestive enzymes, including RNases that are secreted by the exocrine pancreas. To explore whether a different cell type might yield still longer reads, we applied 3′, 5′, and 5′ optimized library preparations, as described above, to lymphocytes isolated from dissociated mouse spleens ($n = 2$ independent biological replicates per capture method). The 5′ lymphocyte samples demonstrated better transcript identification compared with the 3′ pancreatic islet samples (Supplementary Fig. 2*A*). However, the 10x Genomics Chromium library preparation modifications for the 5′ samples did not yield the same improvements in the lymphocyte samples as observed in the pancreatic islet samples (Supplementary Fig. 2*A*, *E*, and *F*). This suggests that the benefits of these optimizations might be tissue specific and highlights the need for additional refinements tailored to different tissue types.

Isoform Variants Identified Between α- and β-Cells and Within β-Cell Subpopulations

With the improved library preparation, the optimized 5′ sclrRNA-seq data sets from mouse pancreatic islets were used to explore whether splicing changes could be detected from different cell types and cell states. Importantly, the 5′ modified sclrRNA-seq data set, merged across replicates, allowed clear identification of all expected cell populations (Fig. 2*A* and *B* and Supplementary Table 2). Importantly, despite depletion of ∼95% of insulin transcripts, enough transcripts remained to identify insulin-expressing cells (Fig. 2*B*). Furthermore, the analysis revealed that cell type identification remains robust whether using gene- or transcript-level expression data for dimensionality reduction and clustering, with >90% concordance between the two approaches (Fig. 2*D–G*). This stability in broad cell type classification aligns with the understanding that major cell types are defined by distinct gene expression patterns. However, when examining substructure within these cell types, substantial differences emerged between gene- and transcript-level analyses, with

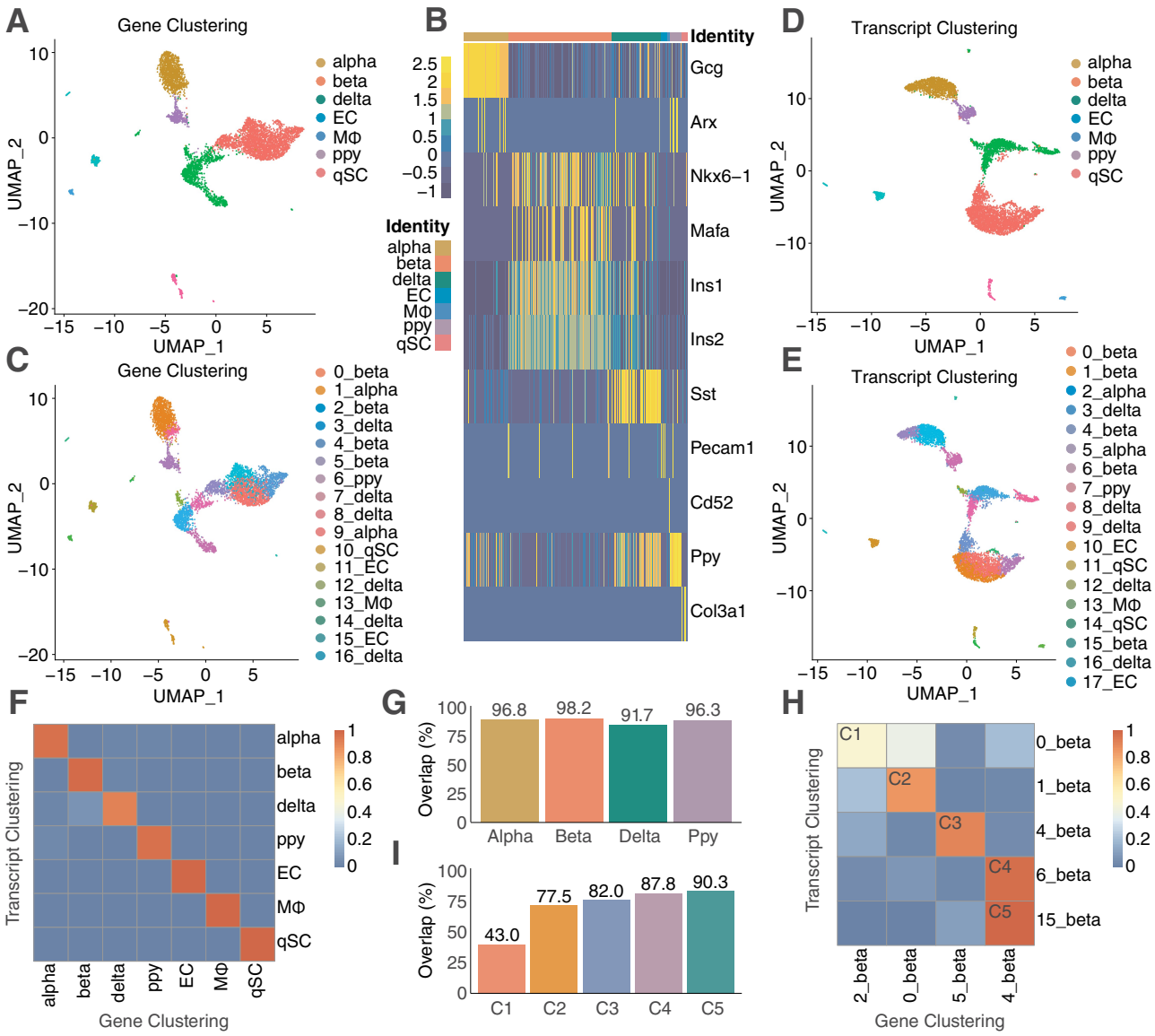

Figure 2—Comparison of single-cell clustering based on gene expression versus transcript-level expression. *A*: Uniform manifold approximation and projection (UMAP) of single cells from insulin-depleted mouse pancreatic islet sclrRNA-seq libraries prepared with 5′ modified 10x Genomics technology (two biological replicates merged), based on gene-level expression. Cells are colored by gene expression profiles reflecting major pancreatic cell types. *B*: Heat map showing expression of cell type–specific markers across single-cell insulin-depleted 5′ modified libraries (two biological replicates merged), based on gene-level clustering. *C*: UMAP of single cells based on gene-level expression, colored by gene expression profiles reflecting cell subpopulations. *D*: UMAP of single cells based on transcript-level (isoform) expression, colored by transcript expression profiles reflecting major pancreatic cell types. *E*: UMAP of single cells based on transcript-level expression, colored by transcript expression profiles reflecting cell subpopulations. *F*: Confusion matrix showing concordance in cell type identification between gene- and transcript-based clustering. *G*: Bar plot quantifying cell type concordance between clustering methods. *H*: Confusion matrix showing low concordance in β-cell subpopulation identification between clustering methods. *I*: Bar plot quantifying β-cell subpopulation concordance between clustering methods. C1–5 indicate comparisons 1 through 5. EC, endothelial cell; ppy, pancreatic polypeptide; qSC, quiescent stellate cell.

consistency ranging from 43% to 90% across subclusters (Fig. 2*I* and Supplementary Fig. 4). These findings suggest that although gene-level expression is sufficient for identifying major cell types, transcript-level analysis provides crucial insights into subtle variations within cell populations. Such variations may reflect different cell states, functions, or responses that are not captured by gene-level analysis alone.

The primary strength of sclrRNA-seq lies in its ability to capture cell-specific isoform expression. To assess differential splicing, we performed pseudobulk DTU analysis on merged 5′ modified replicates, alongside pseudobulk DGE analysis. A pseudobulk approach was chosen because it treats the biological sample, rather than individual cells, as the unit of replication, reducing pseudoreplication and yielding more robust and reproducible results (36,37).

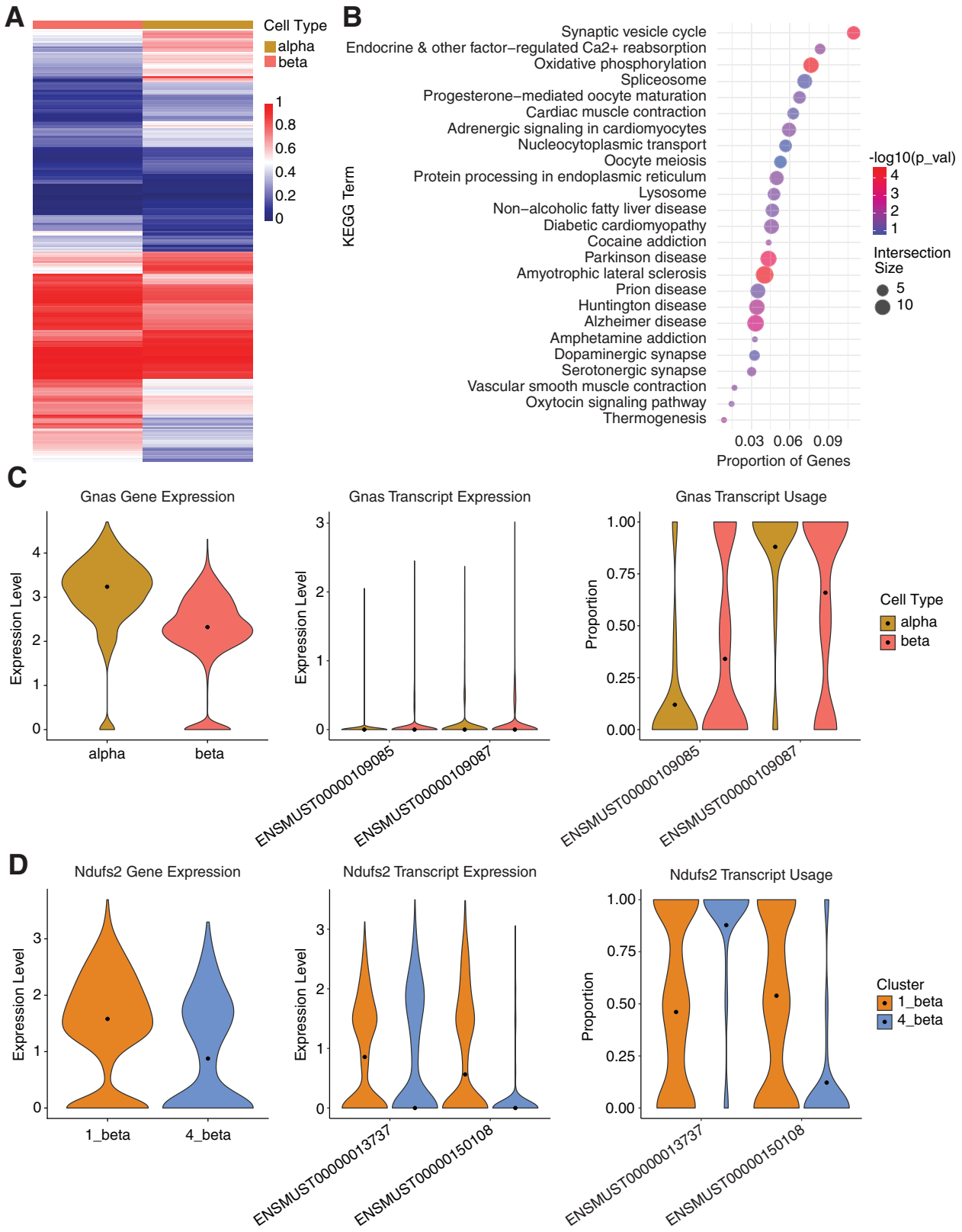

Figure 3—DTU between cell types and cell subpopulations. *A*: Heat map showing DTU between α- and β-cells. Each row represents a transcript (significant at false discovery rate <0.05), and each column represents a cell type. Colors indicate the fraction of a gene's expression contributed by each transcript within each cell type. DTU genes were identified from pseudobulk analysis of two biological replicates of mouse pancreatic islet sclrRNA-seq libraries prepared with 5′ modified 10x Genomics technology. *B*: KEGG pathway overrepresentation analysis of DTU genes between α- and β-cells. Recall shown on *x*-axis (i.e., proportion of functionally annotated genes in the query that are associated with each pathway). Bubble size represents the number of DTU genes in the pathway, and bubble color

DTU analysis (23) identifies proportional differences in the transcript composition of a gene, comparing how much each transcript contributes to total gene expression across conditions. Using this analysis, 218 DTU events were identified between α- and β-cells (Fig. 3*A*). We also detected 2,095 DGE genes, of which 56 overlapped between the DTU and DGE analyses (Supplementary Table 3). As a representative example of β-cell subpopulation analysis, we focused on the comparison between the 1_β and 4_β subpopulations, where we identified 316 DGE genes and nine DTU genes, with two genes overlapping between the two analyses (Supplementary Table 4). Overrepresentation analysis of DTU genes revealed enrichment in pathways such as synaptic vesicle cycle, calcium reabsorption, and oxidative phosphorylation in both α- versus β-cells and 1_β versus 4_β comparisons (Fig. 3*B* and Supplementary Fig. 5*A*). Specifically, when comparing α- and β-cells, we identified isoform-specific differences in *Gnas*, a key regulator of cAMP signaling and insulin secretion (Fig. 3*C* and Supplementary Fig. 5*B*). Gnas encodes Gsα, the stimulatory G-protein α subunit that couples G-protein–coupled receptors to adenylate cyclase (38). Isoforms ENSMUST00000109085 and ENSMUST00000109087 differ in the exon encoding the αA helix of the central α-helical domain, where ENSMUST00000109085 lacks a 14–amino acid segment present in ENSMUST00000109087 and contains a Gly-to-Ser substitution within this region. Alterations in the αA helix could plausibly affect local conformation or interactions with the GTPase domain involved in nucleotide exchange, although the functional consequences of this isoform-specific difference have not been experimentally determined. Additional DTU genes of interest in this comparison include *Pcsk2*, which encodes a prohormone convertase required for insulin processing, as well as *Prdx2* and *Calm2*, which contribute to antioxidant defense and calcium signaling, respectively. Notably, individual transcripts analyzed for *Pcsk2* and *Calm2* each contain a retained intron (ENSMUST00000124751 for *Pcsk2*; ENSMUST00000150137 for *Calm2*) and are predicted to produce a nonfunctional protein. In contrast, the Prdx2 isoforms encode identical protein sequences but differ in their 5′ untranslated regions, potentially affecting RNA stability or translation efficiency. Analysis of two β-cell subpopulations (1_β and 4_β) revealed distinct isoform usage of *Ndufs2*, a central component of oxidative phosphorylation and ATP production (Fig. 3*D* and Supplementary Fig. 5*C*). Among the *Ndufs2* isoforms, only ENSMUST00000013737 is currently predicted to produce a protein, whereas ENSMUST00000150108 has an undefined coding sequence. This comparison also highlighted isoform differences in *Stxbp1*, a gene important for insulin granule exocytosis, where only ENSMUST00000050000 produces protein, whereas the other isoforms (ENSMUST00000113222 and ENSMUST00000192333) contain retained introns and do not encode for a protein product. To assess whether 3′ or 5′ library preparation introduces bias in detecting transcript isoforms, such as preferential capture of variation near the transcript ends, we examined transcript identification based on splice junctions (Supplementary Fig. 5*D*). This analysis showed that there was limited 3′ or 5′ bias, with differences largely driven by a few highly expressed transcripts. Representative gene plots illustrate that 5′ libraries generally provide reads spanning more exon junctions (Supplementary Fig. 6), whereas 3′ libraries better detect certain long transcripts with substantial 3′ end variation (Supplementary Fig. 7) and may disproportionately enhance detection of some transcripts because of internal priming (Supplementary Fig. 8).

DISCUSSION

This study demonstrates how an improved sclrRNA-seq library preparation protocol from isolated islets produces longer reads and increases the proportion of reads that can be confidently assigned to specific transcripts, improving the utility of long-read sequencing data for identifying splice variants and cellular heterogeneity in pancreatic endocrine cell populations. Specifically, this study demonstrates that islet sclrRNA-seq libraries prepared with 5′ protocols outperform those prepared with 3′ protocols. Although neither library type showed a strong overall 3′ or 5′ bias, 3′ libraries are prone to internal priming, whereas 5′ reads tend to cover more exon junctions, providing more complete isoform identification. Enhancements to the 5′ library preparation further improve read length and transcript tagging efficiency in pancreatic islets. Furthermore, depleting insulin transcripts from the pancreatic islet libraries proved to be a highly effective strategy for maximizing informative reads, demonstrating the broader potential of targeted transcript depletion in scRNA-seq experiments.

Although the modified 5′ protocol significantly improved read length in islet samples, lymphocyte samples showed significant improvement only with the unmodified 5′ protocol, with no additional benefit from the modifications. This indicates that individual cell types will require unique modifications and additional optimizations. Despite the significant improvements in read length achieved with the modified protocol, it did not meet expectations for full-length transcript coverage. Achieving this goal will require additional modifications to the 10x chemistry, including adjustments to the master mix and reverse transcriptase.

Although full-length coverage was not achieved for all transcripts, we successfully analyzed transcript expression and identified DTU across cell types and cell subpopulations.

indicates statistical significance (−log10 *P* value). Top 25 enriched KEGG pathways are shown. *C*: DGE, differential transcript expression (DTE), and DTU analyses of *Gnas* between α- and β-cells. DTU shows the relative contribution of each transcript to the gene's overall expression. *D*: DGE, DTE, and DTU analyses of *Ndufs2* between β-cell subpopulations 1_β and 4_β.

These advancements are critical for uncovering the full complexity of transcriptomes and hold immense potential for broad application across tissues, enabling deeper insights into cellular heterogeneity, isoform regulation, and functional diversity. Importantly, isoform-level analyses in this study were restricted to previously annotated transcripts, and we did not attempt to identify novel isoforms. Additionally, reads were assigned to an isoform only when they mapped uniquely to a single annotated isoform, excluding ambiguous mappings. This strategy was chosen to prioritize robustness and confidence in isoform assignment. Although scRNA-seq allows the identification of distinct cell populations and subpopulations, the precise number and composition of clusters can be influenced by analytical choices, such as clustering resolution, preprocessing, and dimensionality reduction, supporting previous analyses showing that subpopulations may not be consistently defined across studies (39). In addition, transcript capture and read depth can influence subpopulation detection; although adequate coverage suffices for gene-level conclusions, even greater depth is required to resolve transcript-level differences, because each gene is often associated with multiple transcripts. Despite these limitations, single-cell transcriptomic analyses still provide valuable insights into isoform-level regulation within specific cell types. Investigating these variations at the single-cell level allows us to uncover the intricate heterogeneity within tissues, offering a deeper understanding of the functional and transcriptional diversity that would otherwise go unnoticed. Understanding splicing dysregulation in pancreatic islets is particularly important, because it may reveal how alternative splicing shapes β-cell function, immune tolerance, and β-cell susceptibility in diabetes.

Acknowledgments. The authors thank Laura White and Jay Hesselberth, University of Colorado Anschutz Medical Campus, for their guidance and support with Nanopore long-read sequencing technologies; Mia Smith, University of Colorado Anschutz Medical Campus, for guidance on spleen sample preparation; and Scott Beard, Barbara Davis Center Cytometer Core, University of Colorado Anschutz Medical Campus, for islet and spleen isolations.

Funding. This work was supported by National Institutes of Health grants P30DK116073, R01 DK082590, and U01 DK127505 (L.S.).

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

Author Contributions. M.S.H. was responsible for data acquisition and prepared the original manuscript. M.S.H., C.J.H., and K.L.W. contributed to data analysis and the graphical presentation of results. C.J.H. and K.L.W. developed the computational pipelines. L.S. and K.L.W. reviewed and edited the manuscript. All authors contributed to the study's methodology and conceptualization. K.L.W. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Gonzàlez-Porta M, Frankish A, Rung J, et al. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol 2013;14:R70

2. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 2003;72:291–336

3. Piazzi M, Bavelloni A, Salucci S, et al. Alternative splicing, RNA editing, and the current limits of next generation sequencing. Genes (Basel) 2023;14:1386

4. Juan-Mateu J, Villate O, Eizirik DL. Mechanisms in endocrinology: alternative splicing: the new frontier in diabetes research. Eur J Endocrinol 2016;174:R225–R238

5. Jeffery N, Richardson S, Chambers D, et al. Cellular stressors may alter islet hormone cell proportions by moderation of alternative splicing patterns. Hum Mol Genet 2019;28:2763–2774

6. Diez J, Park Y, Zeller M, et al. Differential splicing of the IA-2 mRNA in pancreas and lymphoid organs as a permissive genetic mechanism for autoimmunity against the IA-2 type 1 diabetes autoantigen. Diabetes 2001;50:895–900

7. Dogra RS, Vaidyanathan P, Prabakar KR, et al. Alternative splicing of G6PC2, the gene coding for the islet-specific glucose-6-phosphatase catalytic subunit-related protein (IGRP), results in differential expression in human thymus and spleen compared with pancreas. Diabetologia 2006;49:953–957

8. Daraio T, Bombek LK, Gosak M, et al. SNAP-25b-deficiency increases insulin secretion and changes spatiotemporal profile of Ca^{2+} oscillations in β cell networks. Sci Rep 2017;7:7744

9. Lopes I, Altab G, Raina P, et al. Gene size matters: an analysis of gene length in the human genome. Front Genet 2021;12:559998

10. Mattioli K, Bulyk ML. Beyond the gene: decoding alternative isoforms. Trends Genet. 11 December 2025 [Epub ahead of print]. DOI: 10.1016/j.tig.2025.11.003

11. Shiau C-K, Lu L, Kieser R, et al. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. Nat Commun 2023;14:4124

12. Camunas-Soler J, Dai X-Q, Hang Y, et al. Patch-seq links single-cell transcriptomes to human islet dysfunction in diabetes. Cell Metab 2020;31:1017–1031.e4

13. Lun ATL, Riesenfeld S, Andrews T, et al.; Participants in the 1st Human Cell Atlas Jamboree. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol 2019;20:63

14. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol 2024;42:293–304

15. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573–3587.e29

16. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–1902.e21

17. Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–420

18. Satija R, Farrell JA, Gennert D, et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33:495–502

19. McCarthy DJ, Campbell KR, Lun ATL, et al. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 2017;33:1179–1186

20. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst 2019;8:329–337.e4

21. Baron M, Veres A, Wolock SL, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst 2016;3:346–360.e4

22. Jagan D. Matrix: sparse and dense matrix classes and methods. Published 4 April 2023. Accessed 23 April 2025. Available from https://CRAN.R-project.org/package=Matrix

23. Tekath T, Dugas M. Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle. Bioinformatics 2021;37:3781–3787

24. Wells KL, Miller CN, Gschwind AR, et al. Combined transient ablation and single-cell RNA-sequencing reveals the development of medullary thymic epithelial cells. Elife 2020;9:e60188

25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet 2011;17:10–12

26. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21
27. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–930
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550
29. Shen L, Shao N, Liu X, et al. ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics 2014;15:284
30. Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–W198
31. Lebrigand K, Magnone V, Barbry P, et al. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat Commun 2020;11:4025
32. Tian L, Jabbari JS, Thijssen R, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol 2021;22:310
33. Wang Q, Bönigk S, Böhm V, et al. Single-cell transcriptome sequencing on the Nanopore platform with ScNapBar. RNA 2021;27:763–770
34. You Y, Prawer YDJ, De Paoli-Iseppi R, et al. Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE. Genome Biol 2023;24:66
35. Svoboda M, Frost HR, Bosco G. Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. NAR Genom Bioinform 2022;4:lqac035
36. Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. Nat Commun 2021;12:738
37. Squair JW, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. Nat Commun 2021;12:5692
38. Weinstein LS, Xie T, Zhang Q-H, et al. Studies of the regulation and function of the Gs alpha gene Gnas using gene targeting technology. Pharmacol Ther 2007;115:271–291
39. Mawla AM, Huising MO. Navigating the depths and avoiding the shallows of pancreatic islet cell transcriptomes. Diabetes 2019;68:1380–1393