frontiers

ScNaST - a computational workflow for spatial long Nanopore read transcriptomics

Etienne Boileau [1,2,3], **Xue Li** [4], **Ramona Casper**[5], **Janine Altmüller**[6], **Florian Leuschner** [4,7], **Christoph Dieterich** [1,2,3]

[1] *Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany*
[2] *Department of Internal Medicine III (Cardiology, Angiology, and Pneumology), University Hospital Heidelberg, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany*
[3] *DZHK (German Centre for Cardiovascular Research) Partner Site Heidelberg/Mannheim*
[4] *Department of Cardiology, University Hospital Heidelberg, Im Neuenheimer Feld 410, 69120 Heidelberg, Germany*
[5] *Cologne Center for Genomics (CCG), University of Cologne, Weyertal 115b, 50931 Cologne, Germany*
[6] *Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Max Delbrück Center for Molecular Medicine*
[7] *Center for Systems Biology, Massachusetts General Hospital and Harvard Medical School, Simches Research Building, 185 Cambridge Street, Boston, Massachusetts 02114, USA*

Correspondence*:
Christoph Dieterich
christoph.dieterich@uni-heidelberg.de

ABSTRACT

We introduce Single-cell Nanopore Spatial Transcriptomics (scNaST), a set of tools to facilitate the analysis of spatial gene expression from second- and third-generation sequencing, allowing to generate a full-length single-cell transcriptional landscape of the tissue microenvironment. Taking advantage of the Visium Spatial platform, we adapted a strategy recently developed to assign barcodes to long-read single-cell sequencing data for spatial capture technology. Here, we demonstrate our workflow using four short axis sections of the mouse heart following myocardial infarction. We show that results ... Molecular signatures involved in cardiac remodeling integrated with morphological context may support the development of new therapeutics towards the treatment of heart failure and the reduction of cardiac complications.

Keywords: Spatial transcriptomics, Single-cell RNA sequencing, Oxford Nanopore Technology, Myocardial Infarction

INTRODUCTION

Cell type heterogeneity has recently emerged as a major aspect in redrawing the cellular picture of the mammalian heart (Wang et al., 2020; Tucker et al., 2020; Litviňuková et al., 2020). Single-cell RNA-seq

(scRNA-seq) technology has enabled to explore crosstalk of different cardiac cell populations to identify response signatures involved in remodeling after myocardial infarction (MI) and ischemic injury (Cui et al., 2020; Forte et al., 2020; Ruiz-Villalba et al., 2020; Vafadarnejad et al., 2020; Molenaar et al., 2021; Gladka et al., 2021; Tombor et al., 2021; Heinrichs et al., 2021). These and other data provide a valuable compendium of information to better understand transcriptional changes occuring in cardiomyocyte and non-cardiomyocyte sub-populations in healthy, injured, and regenerating hearts. However, in all these studies, the original tissue architecture is destroyed and, in general, the morphological context is lost, including the relationship of cells to infarct, border and remote zones (van Duijvenboden et al., 2019).

Recent development in spatial transcriptomics addresses this challenge, but few studies only have provided spatially resolved insights into the cardiac transcriptome. Techniques such as microdissection (**??**) or, in particular, spatially barcoded arrays and *in situ* capturing, have enabled to retain the spatial information while profiling the whole transcriptome at near single-cell resolution, allowing to shed light on localized tissue neighbourhoods (**??**), but a detailed characterization of cellular zones of injury, repair and/or remodeling is lacking. MI is a complex spatio-temporal heterogeneous disease involving the whole heart, and unbiased spatial transcriptomics holds a promise to add tissue context to molecular profiling in the search of novel therapeutics.

One of the most established spatial transcriptomics methods is now widely available as the Visium platform by 10x Genomics. After the tissue section is fixed on the spatial slide, stained, and imaged, it is permeabilized to release RNA to bind to capture probes for on-slide cDNA synthesis. Library preparation is performed off-slide, and spatial barcodes and tissue image are used to overlay transcriptomics data with tissue context. Although several new such methods have recently been published (**??**), they are for the most part relying on short-read library preparation. Thus they are subject to amplification biases and fail to generate sufficient overlaps to reconstruct transcriptomes *de novo*.

Adequate gene and transcript models are instrumental towards relevant proof of concepts and investigational new drug development in translational cardiac research (Müller et al., 2021). Third-generation or long-read sequencing allow to reconstruct truthful assemblies with fewer gaps and to characterize complete transcript isoforms and chimeric transcripts. This has recently been taken to the single-cell level using either Pacific Biosciences (PacBio) or Oxford Nanopore technology (Nanopore) (**?????**).

Here, taking advantage of the conceptual similarity between spatial and cell barcodes, we introduce Single-cell Nanopore Spatial Transcriptomics (scNaST), a set of tools to facilitate whole transcriptome spatial profiling of full-length transcripts, based on a previously published Bayesian approach for cell barcode assignment (**?**). Our method relies on the commercially available Visium platform by 10x Genomics and Nanopore long-read sequencing, although it can in principle be extended to other technologies, as long as a hybrid sequencing approach is used for spatial barcode assignment.

We demonstrate **??**

MATERIAL AND METHODS

Experimental model of myocardial infarction

A C57BL/6 mouse (female, 8 weeks old, Janvier Labs) was exposed to 5% isoflurane for anesthesia. An intubation cannula was inserted orally into the trachea. The mouse was fixed on a heating plate at 37°C and maintained under anesthesia with 2% isoflurane. An incision was made from the left sternum to the

midclavicular line. Skins and muscle layers were stretched with forceps and sutures. Another incision was made between the third and fourth intercostal space. The heart was exposed and subjected to permanent myocardial infarction with ligation of the left anterior descending (LAD) coronary artery. When the ribs and skins were fully closed, the isoflurane supply was cut off. Oxygen was then supplied until normal breathing was resumed. 24 hours after surgery, 100μl of blood was collected from the retro-orbital sinus. After incubation with heparin for one hour at room temperature, tubes containing blood were centrifuged for 15 min (12,000g at room temperature). 10μl of the supernatant plasma was diluted with 390μl PBS for Cardiac Troponin T (cTnT) analysis. The cTnT level is a reference for cardiac infarction size.

Heart extraction and cryosection

Three days after permanent LAD ligation, the mouse was sacrificed for organ harvest. After washing with cold PBS several times to remove the blood, the heart was transferred into a bath of isopentane (Millipore Sigma) frozen by liquid nitrogen. The freshly obtained heart was kept fully submerged in isopentane for 5 min until fully frozen. Pre-cooled Cryomold on the dry ice was filled with chilled TissueTek OCT compound without introducing bubbles. The frozen tissue was then transferred into the OCT with pre-cooled forceps and placed on the dry ice until the OCT was completely frozen. OCT-embedded tissue blocks were removed from the Cryomold and mounted on the specimen stage. 10μm sections were cut in a cryostat at -10°C and placed within a Capture Area on the pre-equilibrated Visium Spatial Slides (10x Genomics). The slides were later sealed in individual 50 ml Falcon at -80°C ready for further processing.

10X Genomics Visium experiments

Four short axis sections of the heart were processed according to the manufacturer's protocol. The Visium Spatial Tissue Optimization Slide & Reagent kit (10x Genomics) was used to optimize permeabilization conditions. Tissue permeabilization was performed for 24 min. Spatially barcoded full-length cDNA was generated using the Visium Spatial Gene Expression Slide & Reagent kit (10x Genomics). A fraction of each cDNA library was used for nanopore sequencing. cDNA amplification was then conducted for 20 cycles of PCR (identified by qPCR), and 400μl were used in the 10xGenomics Visium library preparation (100μl per section). The libraries were sequenced on a NovaSeq6000 (Illumina), with 29 bases from read 1 and 90 bases from read 2, at a depth of 160M reads per section (640M reads in total). The raw sequencing data was processed with the 10x Genomics Space Ranger 1.1.2 and mapped to the mm10 genome assembly (mm10-2020-A).

Oxford Nanopore sequencing libraries

Libraries for Nanopore sequencing were prepared according to the manufacturer's protocol for direct sequencing of native RNAs (SQK-DCS109 Oxford Nanopore Technologies) with the following minor modifications: 2 ml tubes were used for the HulaMixer, pellets were re-suspended with ABB (room temperature) and centrifugated 2 times, elution was performed in 15μl to have material for TapeStation and Qubit BR, and Flow Cell Priming kit EXP-FLP002 (new version) was used. Starting with 200ng cDNA each, four GridIon flow cells (FLO-MIN106) were loaded with 12ml libraries, with a final cDNA concentration determined by Qubit BR. Base calling was done with Guppy v5.0.12. The High accuracy (HAC) model was selected for base calling (Q-Score cut-off $>$9).

Spatial barcode assignment

To account for source-specific quality differences, each heart section (Illumina libraries) was processed separately using Scanpy v1.7.2 (**?**), keeping only spatial barcodes with approximately 150 $<$ counts $<$

18000, 250 < genes < 5000, detected in at least 2 spots, and with less than 40% mitochondrial counts. The resulting datasets were concatenated, normalized, and the union of highly variable genes (per section) were kept for final analysis. Batch balanced KNN (BBKNN) (**?**) with ridge regression (Park et al., 2020) was used for integration and batch correction, starting from a coarse clustering obtained from a BBKNN-corrected graph.

For the Oxford Nanopore libraries, samples were demultiplexed and processed with ScNapBar v1.1.0 using a Naıve Bayes model to assign spatial barcodes (**?**). Briefly, for each heart section, the Space Ranger results (Illumina libraries) were used to parametrize a model of barcode alignment features to discriminate correct versus false barcode assignment in the Nanopore data. FASTQ files were mapped using minimap2 v2.21 (**?**). For transcript isoform quantification, a de novo transcriptome annotation was generated. Alignment files were processed by StringTie v2.1.5 in long read mode with the reference annotation to guide the assembly process (**?**). The annotations were merge into a non-redundant set of transcripts and compared to the reference using GffCompare v0.12.2 (**?**), after removing single-exon transcripts. To generate feature-spatial barcode matrices, alignment files were split into multiple files, one per spatial barcode, based on the barcode assignments, converted to FASTQ, and aligned to the de novo transcriptome with minimap2. Abundances were quantified with Salmon v1.5.2 in alignment-based mode using a long read error model (**?**). Each section was processed separately using Scanpy v1.7.2 and integrated with BBKNN, as described above for the Illumina data. Spatial barcodes were filtered for counts (approximately 50 < counts < 4000), transcripts (approximately 50 < transcripts < 2000, detected in at least 2 spots), and ribosomal genes (¡ 40%).

Identification of anatomical regions

For the Illumina libraries, the neighborhood graph was computed using BBKNN. Spots were clustered with a low resolution (0.3) to identify anatomical regions such as infarct, border and remote zones. Marker genes were identifed using a Wilcoxon rank sum test with Benjamini–Hochberg correction, by comparing the expression of each gene in a given cluster against the rest of the spots. The final clusters were manually annotated.

Labels were then assigned to Nanopore spatial barcodes based on the set of matching Illumina barcodes. However, not all assigned barcodes were labeled due to quality control filtering criteria that were different between Illumina and Nanopore datasets. To assign labels to the remaining Nanopore barcodes, seed labelling was performed with scANVI using the set of assigned labels as groundtruth (**?**). The top expressed transcripts were then identified as described above for the Illumina data.

Spatial spot deconvolution

Spatial spots were deconvoluted using stereoscope (scvi-tools v0.15.0) (**?**). Heart data (Smart-Seq2 and 10x Genomics) from the Tabula Muris (Schaum et al., 2018) were used as reference dataset and highly variable genes were identified. For the Smart-Seq2 data, gene length normalization was applied. The model was trained on the single cell reference dataset on the intersection of genes found in the spatial (Illumina) data, and proportions were inferred for each Visium spot for each cell type in the reference dataset. Labels were then assigned to Nanopore spots are described above.

Spatial gene expression (Illumina)

Differential transcript usage (Nanopore)

To identify changes in relative usage of transcripts/isoforms within genes, differential transcript usage (DTU) tests were performed using satuRn (**?**). Only multi-exon transcripts and genes with more than one isoform were kept for the analyses. The transcript count matrix was further filtered to keep transcripts expressed in a worthwhile number of spots, determined by the design (but greater than at least 10% of the smallet group size), with a CPM count above a threshold of $1(\text{median library size})^{-1}$. In addition, transcripts were kept only if they had a minimum count of 1 across all spots. A quasi-binomial generalized linear model was fit using a design comparing each anatomical region with another, or each anatomical region with the rest of all regions. A two-stage testing procedure was performed using stageR (**?**), with an OFDR of 0.05. Results were reported using a student's t-test statistic, computed with estimated log-odds ratios.

RESULTS

Fresh-frozen tissue samples were stained, imaged and fixed on Visium Spatial Gene Expression Slides (10X Genomics) for permeabilization and *in situ* RNA capture. Full-length cDNA libraries were split for the preparation of 3' Illumina short-read and direct long-read Nanopore sequencing libraries. Short-read data were used for the assignment of spatial barcodes to Nanopore reads using the SCNAPBAR workflow, and subsequently used to define anatomical regions within the tissue organization (Fig. 1A). Long-read data were used for transcriptome assembly and transcript abundance quantification, and layered onto the stained images to reveal the spatial organization of isoform expression. The Nanopore data comprises of four heart slices (or samples) with a total of 25,5 million reads, reaching a relatively high sequencing saturation (Fig. 1B), and providing we expect here to have a more uniform coverage... see Figure 1 C, this is a little unexpected. I am checking this... After spatial barcode assignment, libraries had a median of 2.8 million reads per sample, with over 70% assigned reads (Figs. 1D and Supplementary Figure S1A). Despite variations in transcriptome alignments between samples, we observed a good correlation between spots across all samples between Illumina and Nanopore libraries (Fig. Supplementary Figure S1B,C, see also Figs. Supplementary Figure S2 to Supplementary Figure S4).

Clustering based on short-read gene expression defined four broad morphological regions: two remote zones that stem from differences in sequencing depth between heart slices, a border zone, and the infarct. The region classification was then transfered to the Nanopore data (Fig. 2A,B) Remote zones and, to a lesser extent the border zone, are largely associated with cardiomyocyte markers, while the border and infarct zones are characterized by a higher expression of endothelial, myofibroblast, and immune marker genes, as well as with markers of fibrosis and inflammation (Fig. 2C).

We successfully assigned 7616 spatial spots, corresponding to distinct 19794 transcript isoforms in total, encoded by 12474 genes. Transcript isoforms were largely associated with exact matches to the reference annotation, multi-exon transcripts with at least one junction match to the reference (*e.g.* exon skipping and exon extension), transcripts longer than the reference (containment of reference), completely novel transcripts (intergenic), transcripts with exonic overlap, or intronic transcripts (Fig. 2D). Among all genes, we observed 8131 (67,3%) that expressed a single isoform and 3953 (32.7%) that expressed 2 or more isoforms (Fig. 2E). Although predicted by our assembly, genes with many isoforms were expressed at a lower threshold and were not included in our final analyses. We also noticed variations in the number of isoforms per gene across morphological regions of each heart slice, with significant differences between

either of the remote zones and the border and the infarct areas, suggesting a higher transcriptome diversity in the healthy regions (Fig. 2F).

0.1 Regional isoform switching ...

A few interesting cases: Actc1, Crip2, Tnni3, Tmsb4x, Myh7, Sparc, Clu, etc. Need to add deconvolution results: can we identify cellular origin of some of the observed results? We need a way to visualize the different isoforms to see if they all make sense! Snapshots from IGV, etc. Add biological context: MI

DISCUSSION

DATA AVAILABILITY

The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY] [LINK].

CODE AVAILABILITY

AUTHOR CONTRIBUTIONS

TBD

FUNDING

TBD

ACKNOWLEDGMENTS

TBD

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

SUPPLEMENTAL DATA

Supplementary Figure S1

ScNaST methodology.

Supplementary Figure S2

Quality control (Illumina).

Supplementary Figure S3

Quality control (Nanopore).

Supplementary Figure S4

Spatial distribution of UMIs or counts and genes or transcripts.

REFERENCES

Cui, M., Wang, Z., Chen, K., Shah, A. M., Tan, W., Duan, L., et al. (2020). Dynamic transcriptional responses to injury of regenerative and non-regenerative cardiomyocytes revealed by single-nucleus rna sequencing. *Developmental Cell* 53, 102–116.e8. doi:https://doi.org/10.1016/j.devcel.2020.02.019. PMID: 32220304

Forte, E., Skelly, D. A., Chen, M., Daigle, S., Morelli, K. A., Hon, O., et al. (2020). Dynamic interstitial cell response during myocardial infarction predicts resilience to rupture in genetically diverse mice. *Cell Reports* 30, 3149–3163.e6. doi:https://doi.org/10.1016/j.celrep.2020.02.008

Gladka, M. M., Kohela, A., Molenaar, B., Versteeg, D., Kooijman, L., Monshouwer-Kloots, J., et al. (2021). Cardiomyocytes stimulate angiogenesis after ischemic injury in a zeb2-dependent manner. *Nature Communications* 12, 84. doi:10.1038/s41467-020-20361-3. PMID: 33398012

Heinrichs, M., Ashour, D., Siegel, J., Büchner, L., Wedekind, G., Heinze, K. G., et al. (2021). The healing myocardium mobilizes a distinct B-cell subset through a CXCL13-CXCR5-dependent mechanism. *Cardiovascular Research* 117, 2664–2676. doi:10.1093/cvr/cvab181. PMID: 34048536

Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C. L., Lindberg, E. L., et al. (2020). Cells of the adult human heart. *Nature* 588, 466–472. doi:10.1038/s41586-020-2797-4. PMID: 32971526

Molenaar, B., Timmer, L. T., Droog, M., Perini, I., Versteeg, D., Kooijman, L., et al. (2021). Single-cell transcriptomics following ischemic injury identifies a role for b2m in cardiac repair. *Communications Biology* 4, 146. doi:10.1038/s42003-020-01636-3. PMID: 33514846

Müller, T., Boileau, E., Talyan, S., Kehr, D., Varadi, K., Busch, M., et al. (2021). Updated and enhanced pig cardiac transcriptome based on long-read rna sequencing and proteomics. *Journal of Molecular and Cellular Cardiology* 150, 23–31. doi:https://doi.org/10.1016/j.yjmcc.2020.10.005. PMID: 33049256

Park, J.-E., Botting, R. A., Conde, C. D., Popescu, D.-M., Lavaert, M., Kunz, D. J., et al. (2020). A cell atlas of human thymic development defines t cell repertoire formation. *Science* 367, eaay3224. doi:10.1126/science.aay3224

Ruiz-Villalba, A., Romero, J. P., Hernández, S. C., Vilas-Zornoza, A., Fortelny, N., Castro-Labrador, L., et al. (2020). Single-cell rna sequencing analysis reveals a crucial role for cthrc1 (collagen triple helix repeat containing 1) cardiac fibroblasts after myocardial infarction. *Circulation* 142, 1831–1847. doi:10.1161/CIRCULATIONAHA.119.044557. PMID: 32972203

Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature* 562, 367–372. doi:10.1038/s41586-018-0590-4. PMID: 30283141

Tombor, L. S., John, D., Glaser, S. F., Luxán, G., Forte, E., Furtado, M., et al. (2021). Single cell sequencing reveals endothelial plasticity with transient mesenchymal activation after myocardial infarction. *Nature Communications* 12, 681. doi:10.1038/s41467-021-20905-1. PMID: 33514719

Tucker, N. R., Chaffin, M., Fleming, S. J., Hall, A. W., Parsons, V. A., Bedi, K. C., et al. (2020). Transcriptional and cellular diversity of the human heart. *Circulation* 142, 466–482. doi:10.1161/CIRCULATIONAHA.119.045401. PMID: 32403949

Vafadarnejad, E., Rizzo, G., Krampert, L., Arampatzi, P., Arias-Loza, A.-P., Nazzal, Y., et al. (2020). Dynamics of cardiac neutrophil diversity in murine myocardial infarction. *Circulation Research* 127, e232–e249. doi:10.1161/CIRCRESAHA.120.317200. PMID: 32811295

van Duijvenboden, K., de Bakker, D. E., Man, J. C., Janssen, R., Günthel, M., Hill, M. C., et al. (2019). Conserved ¡i¿nppb¡/i¿+ border zone switches from mef2- to ap-1–driven gene program. *Circulation* 140, 864–879. doi:10.1161/CIRCULATIONAHA.118.038944. PMID: 31259610

239 Wang, L., Yu, P., Zhou, B., Song, J., Li, Z., Zhang, M., et al. (2020). Single-cell reconstruction of the adult
240 human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function.
241 *Nature Cell Biology* 22, 108–119. doi:10.1038/s41556-019-0446-7. PMID: 31915373

FIGURE CAPTIONS

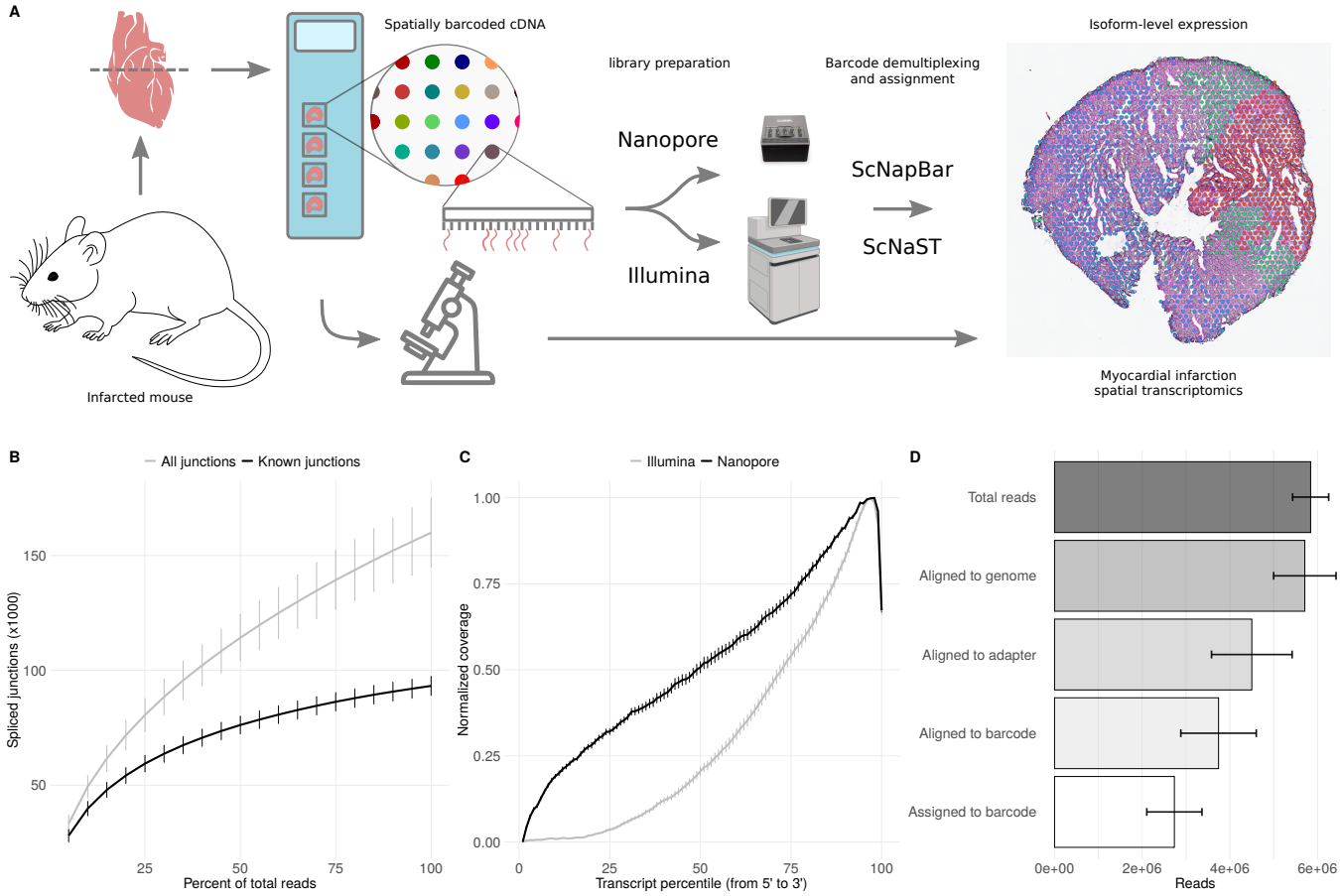

Figure 1. scNaST methodology. A Schematic of the scNaST workflow using a hybrid sequencing approach on Nanopore and Illumina platforms to assign spatial barcodes to long-read sequencing. B Nanopore sequencing saturation showing the number of splice sites detected at various levels of subsampling. A curve that reaches a plateau before getting to 100% data suggest that all known junctions in the library have been detected. The curve shows the mean ± SE of four samples. C Normalized transcript coverage for Nanopore and Illumina. The curves show the mean ± SE of four samples. D Reads assigned by scNapBar at each step of the workflow. The bars show the mean ± SE of four samples.

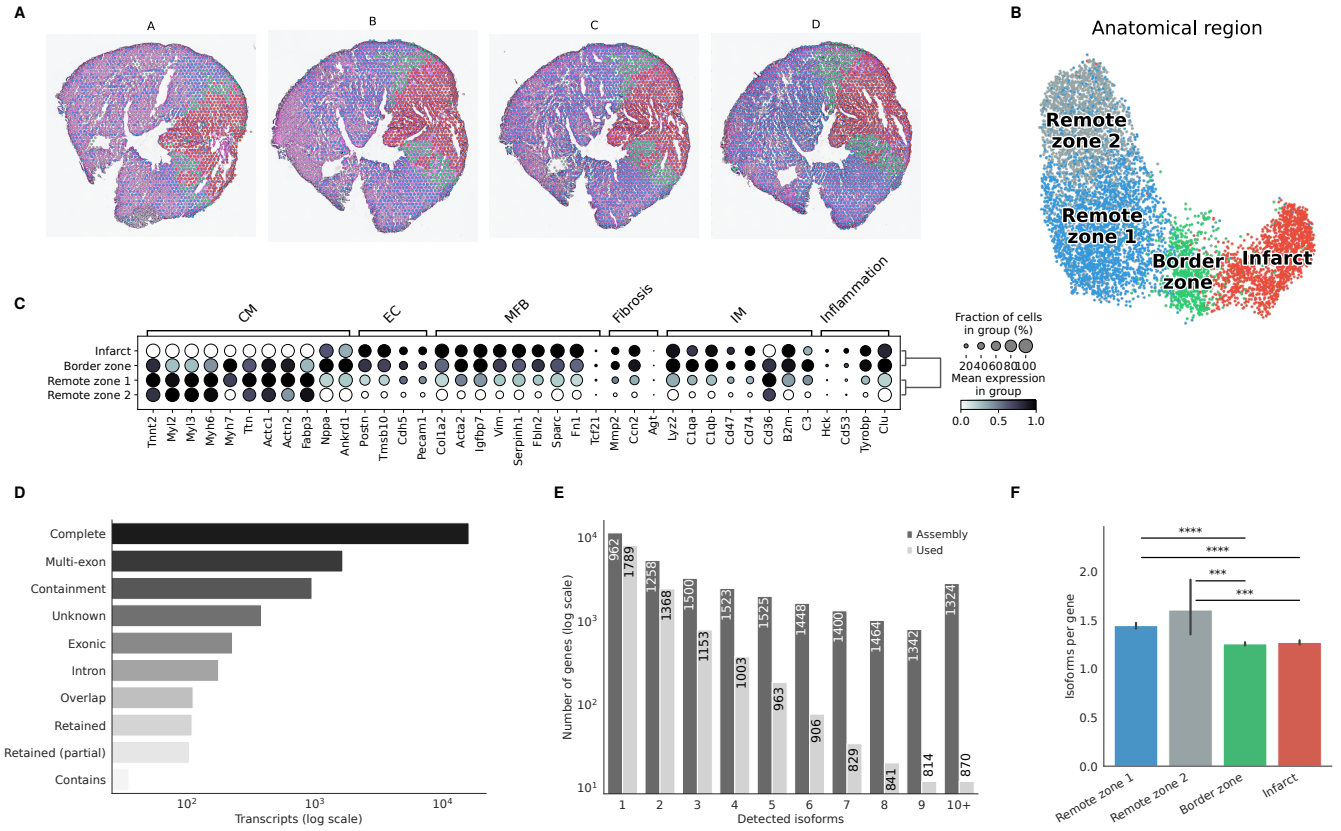

Figure 2. Defining morphological regions after MI. A Annotation of mouse heart regions after MI via short-read clustering, transfered to the Nanopore data. B UMAP representation of the Nanopore data using the region annotation from short-read clustering. C Dot plot showing the expresson of selected markers associated with the expresson of CM=cardiomyocytes, EC=endothelial cells, MFB=myofibroblasts, IM=immune cells, or with fibrosis and inflammation. D Barplot showing how full-length transcripts obtained with ScNaST compare to the existing mouse annotation. Labels Complete (=), Multi-exon (j), Containment (k), Unknown (u), Exonic (x) , Intron(i), Overlap (o), Retained (m, n), and Contains (y) are explained in `https://ccb.jhu.edu/software/stringtie/gffcompare.shtml`. E Barplot showing the frequency distribution of the number of isoforms per gene, either stemming from the assembly, or found in the final data after quality control filtering. The median length of transcripts is indicated in each bar for each category. F Average number of isoform per gene detected for each morphological region. Significance was measured using a Mann-Witney U-test (*** = <0.001, **** = <0.0001)