Supplementary Figures: Single cell transcriptome sequencing on the Nanopore platform with ScNapBar

Qi Wang[1], Sven Bönigk[1], Volker Böhm[2,3], Niels Gehring[2,3], Janine Altmüller[4], and Christoph Dieterich*[1,5,6]

[1]Klaus Tschira Institute for Integrative Computational Cardiology, University Hospital Heidelberg, 69120 Heidelberg, Germany
[2]Institute for Genetics, University of Cologne, 50674 Köln, Germany
[3]Center for Molecular Medicine Cologne (CMMC), University of Cologne, 50937 Köln, Germany
[4]Cologne Center for Genomics (CCG), University of Cologne, Weyertal 115b, 50931 Köln, Germany
[5]Department of Internal Medicine III (Cardiology, Angiology, and Pneumology), University Hospital Heidelberg, 69120 Heidelberg, Germany
[6]German Centre for Cardiovascular Research (DZHK)-Partner Site Heidelberg/Mannheim, 69120 Heidelberg, Germany

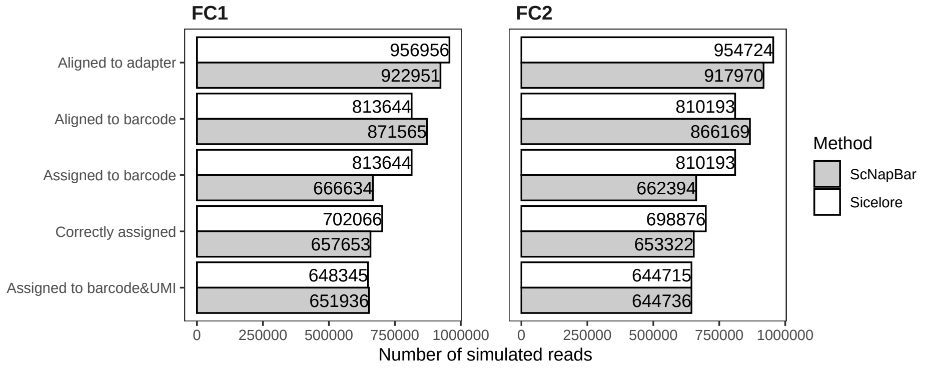

Figure S1: **Number of reads at each processing step in the workflows.** We simulated one million Nanopore reads and processed with ScNapBar and Sicelore, respectively (see “Methods” section). The ScNapBar score cut-off was set to 50.

*christoph.dieterich@uni-heidelberg.de

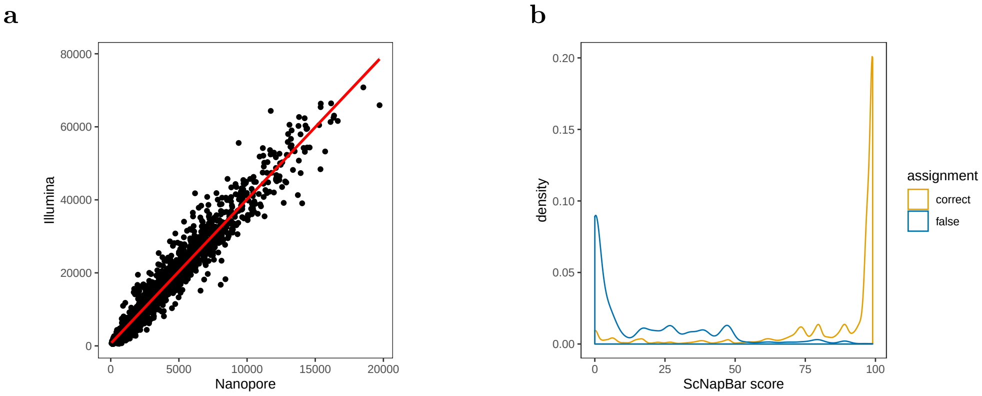

Figure S2: **Statistics of the barcode assignments.** (a) Number of reads per cell between Illumina and Nanopore from the real data. (b) Comparing ScNapBar score of the correct and false assignments from the simulation.

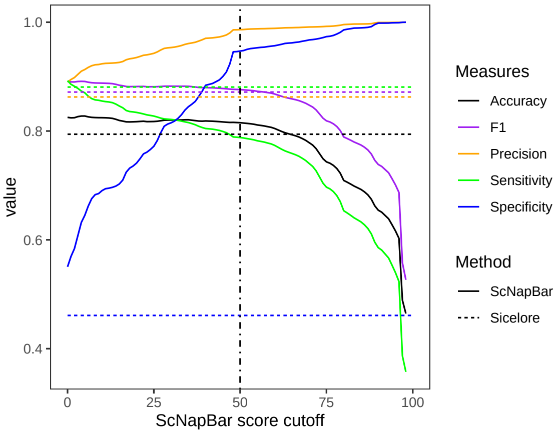

Figure S3: **Benchmarking metrics of barcode assignment from the simulated data between ScNapBar and Sicelore.** We have defined the metrics as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{5}$$

where true positive (TP) is the number of reads which are assigned to the correct barcode; false positive (FP) is the number of reads which are assigned to an incorrect barcode. As we observed ∼ 20% non-cell barcodes from the Illumina library, we further simulated additional 20% reads whose barcode sequences are not within the barcode whitelist but from the other available Illumina barcodes in the protocol and tagged them as true negative (TN).

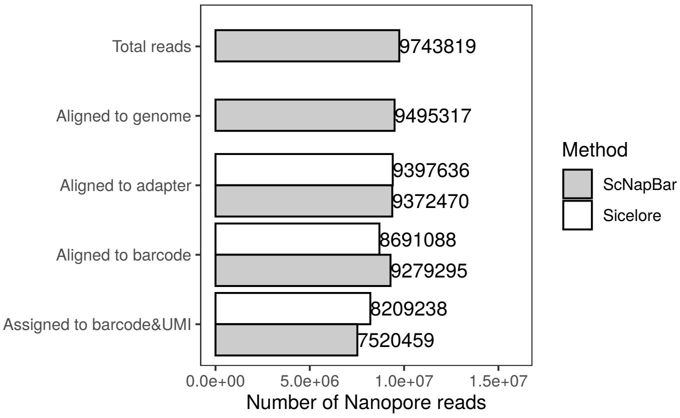

Figure S4: **Number of Nanopore reads assigned to a barcode using ScNapBar and Sicelore.** High Illumina saturation dataset (GEO accession No. GSM3748087) is used.

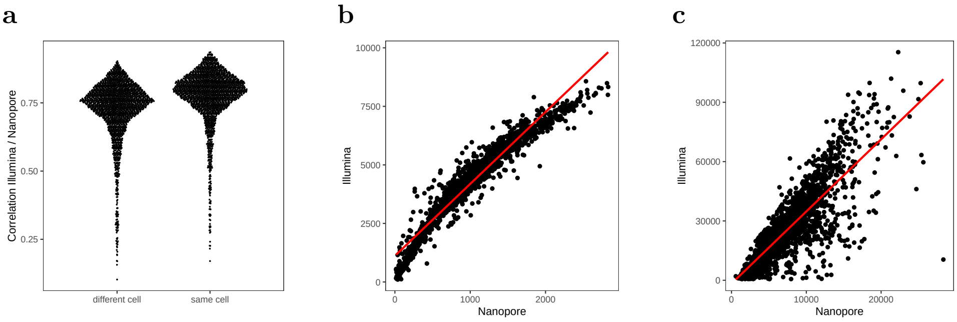

Figure S5: **Correlations between Illumina and Nanopore.** (a) Correlation of gene expression. Pearson's r are calculated based on cellular barcode (same cell: matched barcode, different cell: random unmatched barcode) from Nanopore and Illumina. (b) Number of genes for each cell. (c) Number of unique UMIs for each cell.

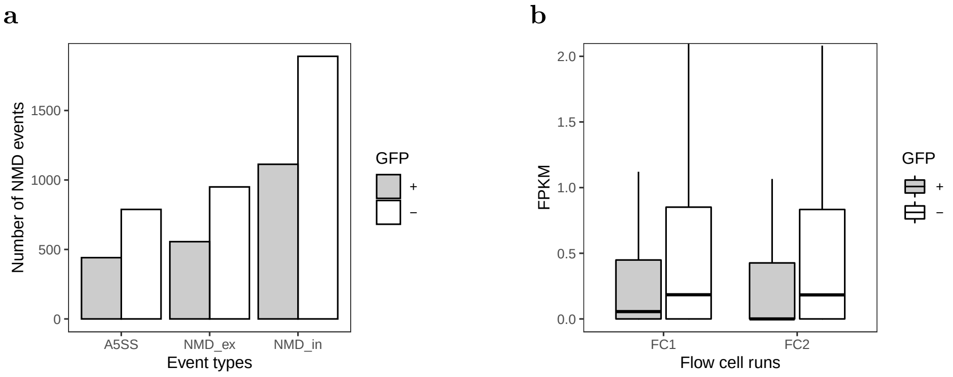

Figure S6: **The NMD events between GFP+/- cells from the pooled FC1 and FC2**. (a) The number of NMD events. NMD_ex: exclusion of an exon; NMD_in: inclusion of an exon; A5SS: the changes occurred at 5' splicing site. (b) The FPKM level of 6,423 known NMD transcripts (t-test between GFP+/- cells p-value <0.01)

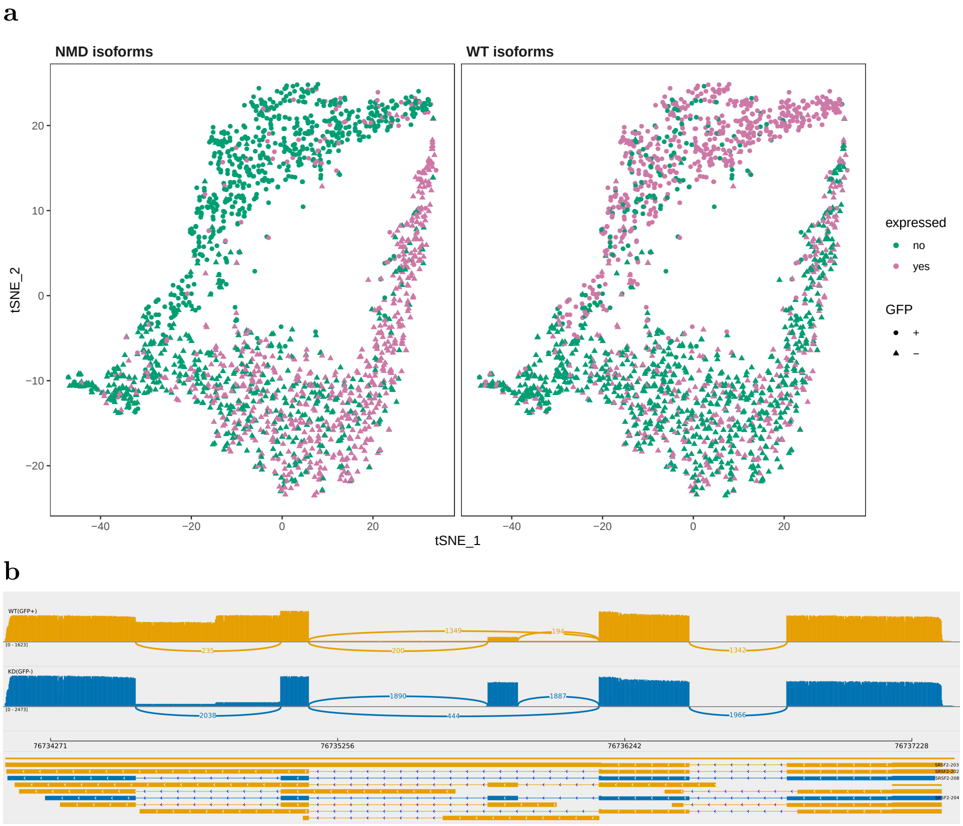

Figure S7: **Isoform expression of the SRSF2 gene.** (a) Isoform expression levels in the GFP+/- cells ("expressed" means number of reads $\geqslant 1$). SRSF2-204 and SRSF2-208 (NMD) are shown on the left. SRSF2-202 and SRSF2-203 (WT) are shown on the right. (b) Sashimi plot of the mapped Nanopore reads shows splice junctions of SRSF2 transcripts. Genomic region of chr17:76,734,108-76,737,393 between GFP+ (WT) and GFP- (NMD). The NMD transcripts are colored blue in the annotation track, while the others are colored yellow. The inclusion level of the exon 3 is clearly higher in NMD cells

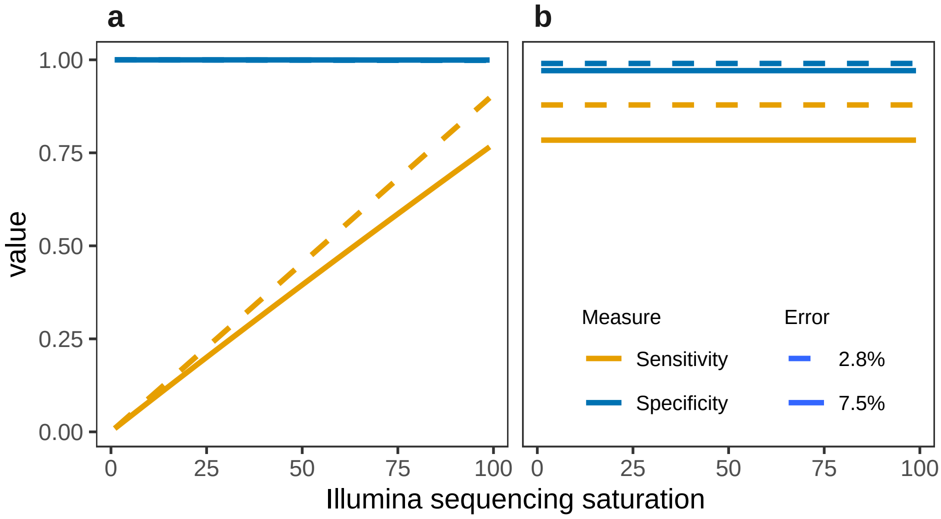

Figure S8: **Sensitivity and specificity of ScNapBar on 100 Illumina libraries with different error rates.** (a) Barcode assignment with UMI matches. (b) Barcode assignment without UMI matches (ScNapBar score >50). 28.1% barcodes contain at least one mismatch or indel from the low error-rate Nanopore reads, as opposed to the 46.4% barcodes from the actual Nanopore reads.

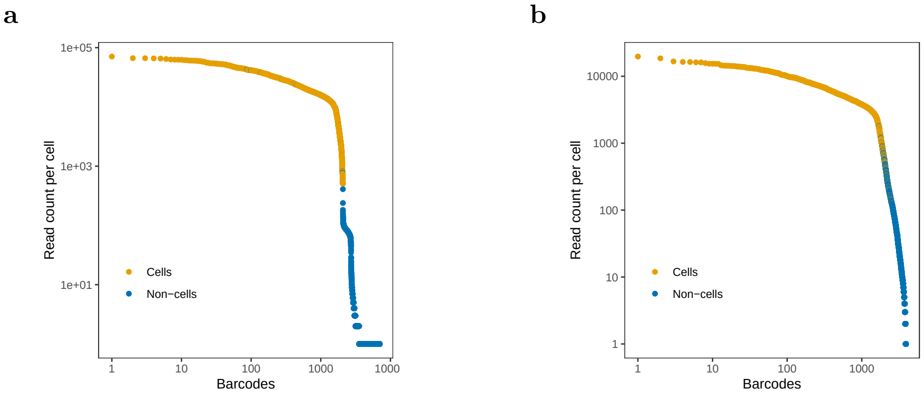

Figure S9: **Barcode rank plots from the real data.** (a) Illumina, (b) Nanopore.

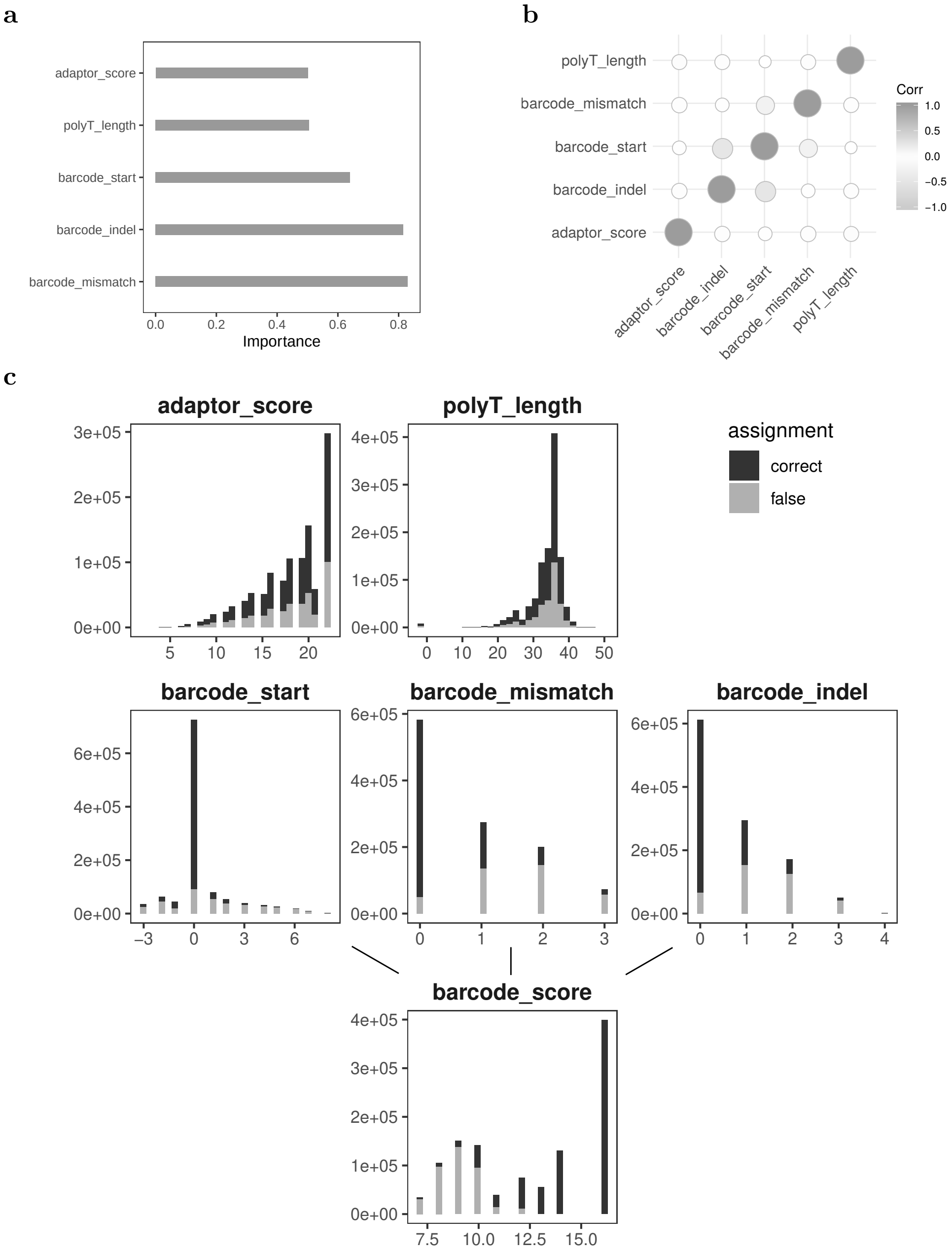

Figure S10: **Combined figure of simulation approach, feature set and feature importance for motivation.** (a) ScNapBar Feature importances towards the assignment correctness. (b) Feature correlations. (c) Histogram shows the frequencies of the features from the matched barcodes in the simulation. There are clear differences of barcode_mismatch, barcode_indel, barcode_start between the distributions of correct and wrong barcode assignments.

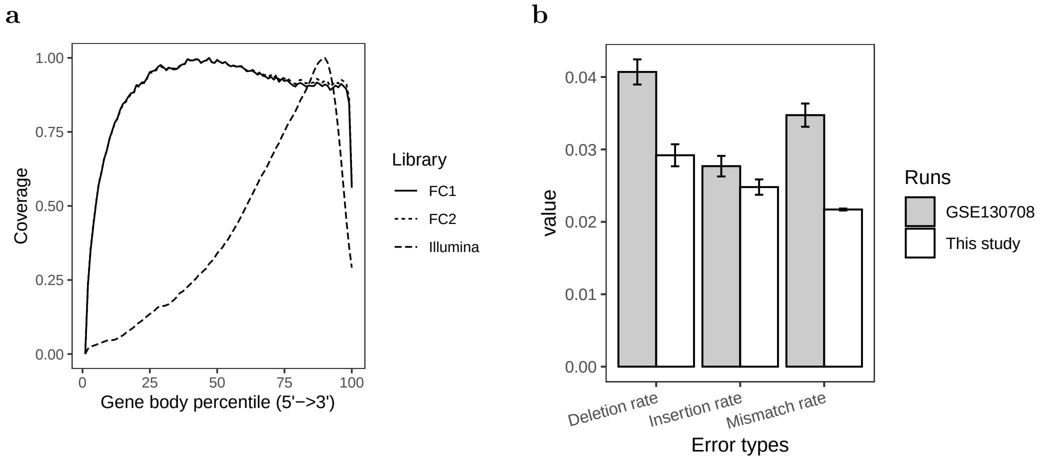

Figure S11: **Sequencing statistics of Nanopore.** (a) Gene body read coverage of Nanopore and Illumina. (b) Nanopore sequencing error rates