Novogene | Advancing Genomics, Improving Life

Characterisation of Single Cell full-length transcriptomics in human and mouse samples with short and long read sequencing

Tianran SHI, Man Chun LEONG[2], Yan LIANG, Yiran SUN[1]
[1]Novogene Co. Ltd . [2]NovogeneAIT Genomics Singapore Pte Ltd

INTRODUCTION

The application of long-read sequencing in single cell analysis offers several advantages over short-read sequencing. These advantages encompass the potential for mapping certainty, transcript isoform identification, and detection of structural variants. Here, we evaluate single cell data output, data quality (including cell number and number of genes detected) and concordance of results using illumina and Nanopore sequencing methods.

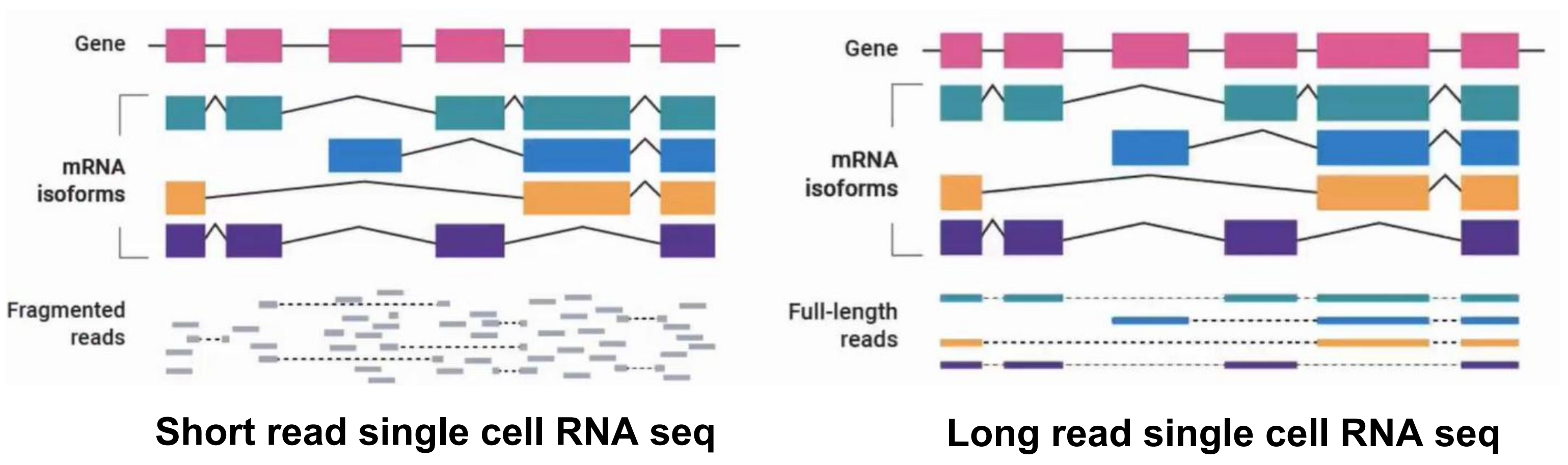
To construct single cell libraries, we utilised Human and Mouse samples, and followed 10x Chromium single cell 3' protocol and the respective manufacturer's protocols — illumina short read and Nanopore long read. The short-read sequencing was performed on Illumina NovaSeq 6000 system using paired-end reads at a depth of ~35,000. For long read sequencing, the library was sequenced on an individual Oxford Nanopore PromethION Cell.

CONCLUSION

We achieved a data output of more than 150M total reads for each library. The data quality assessment, including cell number and median gene number, yield a high level of concordances between short read and long read data. Furthermore, we demonstrated that the data output from a single Nanopore's PromethION cell can yield sufficient reads for isoform detection, structural variation detection and cell clustering for single cell analysis.

AIM

1. To establish a workflow amenable for both single cell short reads and full-length transcriptomics
2. To evaluate the concordance of scRNA-seq data using illumina short read and Nanopore long reads



- ➢ 3' to 5' short reads
- ➢ Gene expression information only
- ➢ Unable to analyze the differences in the isoform between cells

- ➢ Full-length long reads
- ➢ Get the full-length information of mRNA
- ➢ Analyze different mRNA isoform, alternative splicing, fusion genes, etc.

METHODS

Sample Type: human ecchymosis samples (mm10) and mouse bone hematopoietic stem cell samples (hsa).

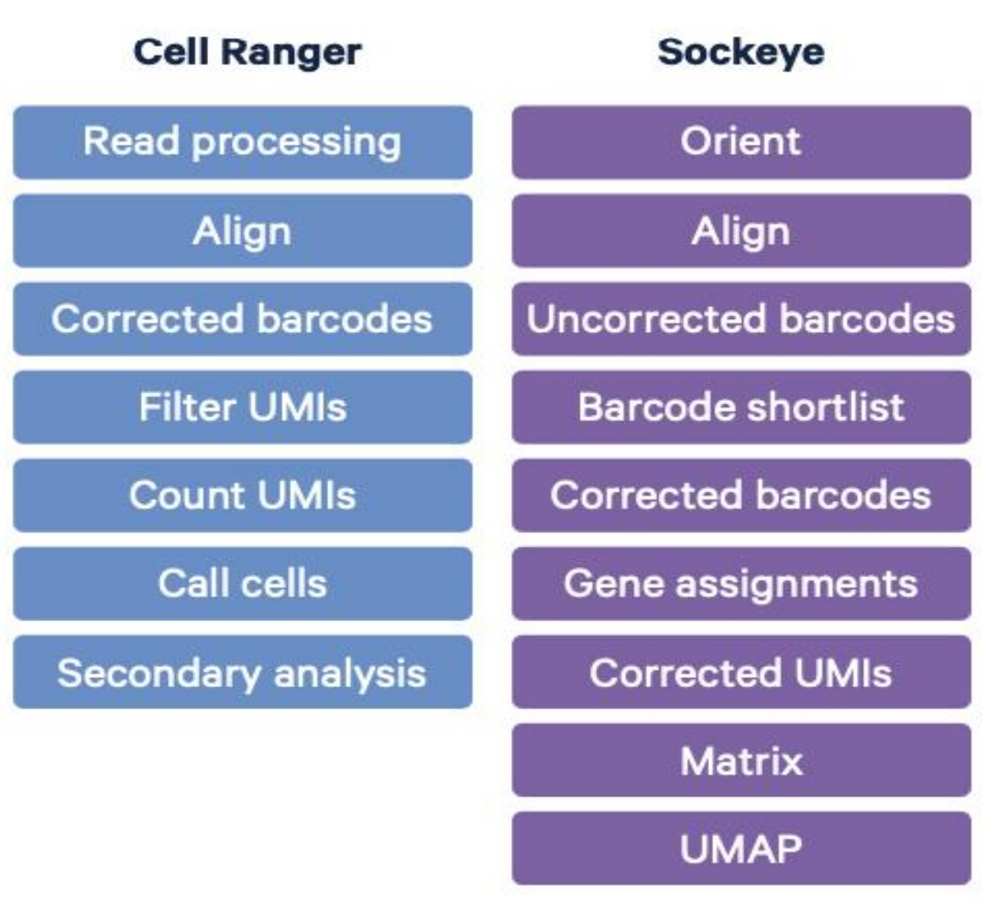
Library prep:

- Gene expression libraries were constructed based on 10× Genomics Next GEM Single Cell 3' Kit v3.1.
- Oxford Nanopore Technologies cDNA libraries were prepared with full-length cDNA generated from Chromium 3' according to the manufacturer's protocol.

Sequencing:

- Short-read sequencing was performed on an Illumina NovaSeq 6000 system to yield 100Gb data output.
- Long-read sequencing library was sequenced on an individual PromethION Flow Cell.

Analysis



Cell Ranger: Read processing; Align; Corrected barcodes; Filter UMIs; Count UMIs; Call cells; Secondary analysis

Sockeye: Orient; Align; Uncorrected barcodes; Barcode shortlist; Corrected barcodes; Gene assignments; Corrected UMIs; Matrix; UMAP

Fig 1. Diagram of Cell Ranger and Sockeye bioinformatics workflows.

RESULTS AND DISCUSSIONS

Outcome:

Sequencing depth and library complexity was comparable across short- and long-read data. Additionally, annotation of cell types and clustering performed similarly between the two sequencing technologies.

Table1. Data Output and Data Quality

wf-single-cell	Reads	Cells	Genes (total)	Transcripts (total)
sample mm10	168,648,648	7,710	20,748	47,806
sample hsa	154,825,486	4,917	29,534	83,627

Seurat		sampleC1_5(mm10)	sample1(hsa)
ONT gene	ncells	7,212	4,349
	median	3,195	4,083
	median umi	7,522	13,129
ONT transcript	ncells	7,142	4,687
	median	3,599	4,915
	median umi	6,985	11,643
Illu gene	ncells	6,969	3,977
	median	2,056	1,979
	median umi	6,312	6,605

The data output of the 2 samples was greater than 150M total reads, in line with the official recommendation of ONT;

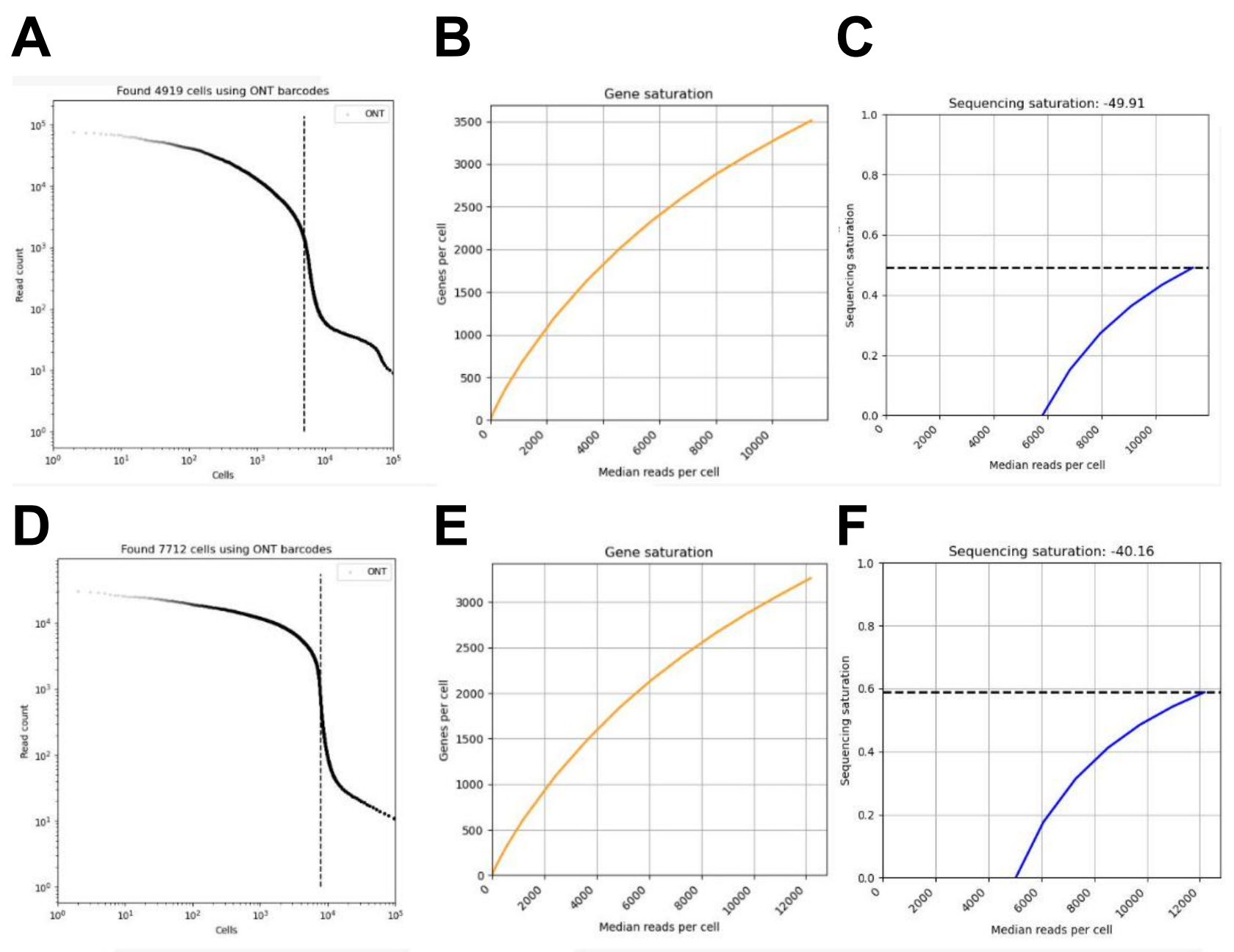


Fig 2. Barcode Rank Plot and Gene Saturation and Sequencing Saturation Plot for mm10 (A-C) and has (D-F). (A, D) The barcode rank plot shows a "steep drop" which means the differentiation between cells and background is obviously. **(B, E) (C, F)** When the target cells is around 6000, 1 ONT Cell produces about 20,000 Median Reads per Cell, the sequencing saturation is 0.5-0.6. It is estimated that at least 2 Cells are required to reach the illumina sequence depth.

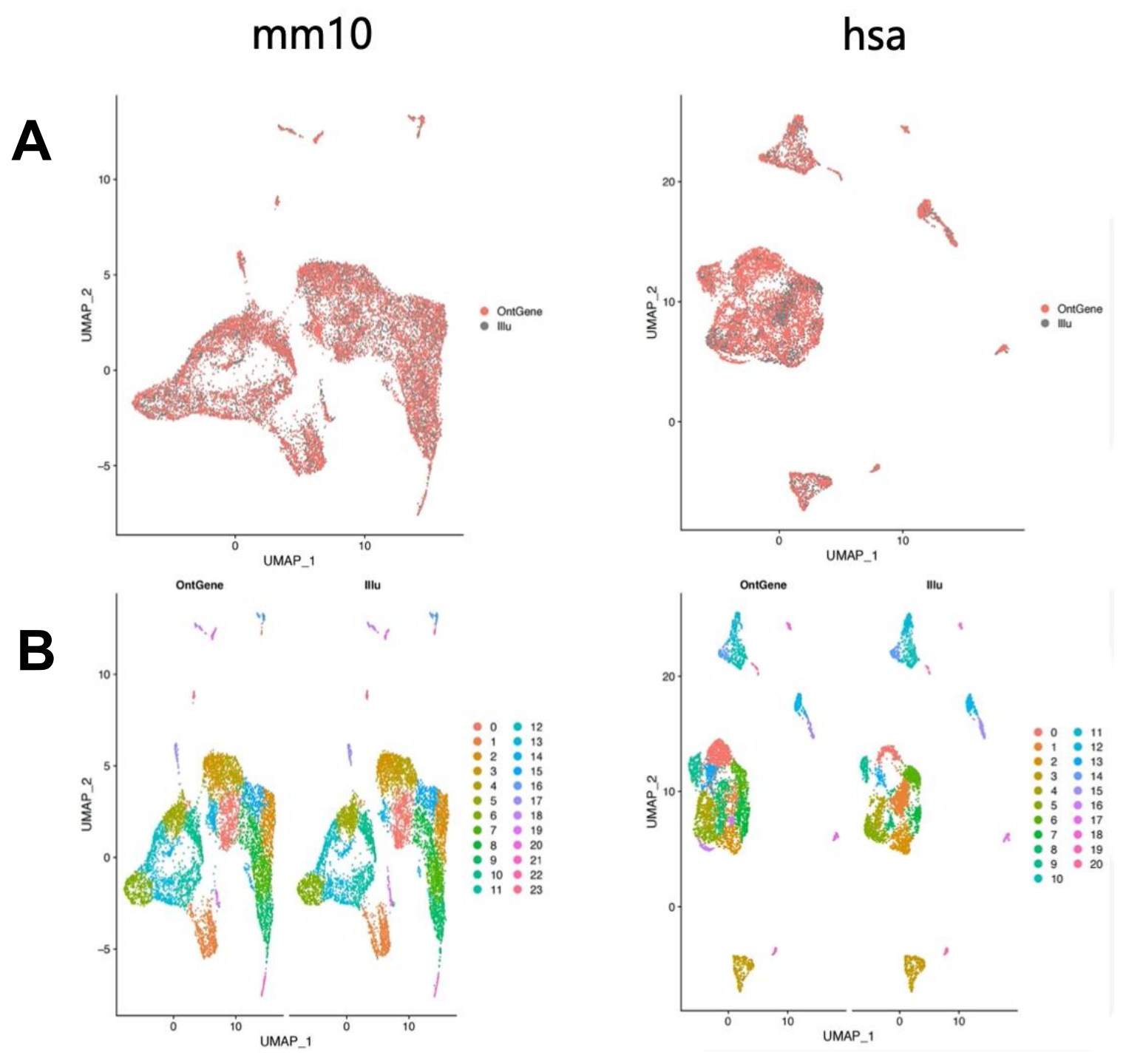


Fig 3. (A) Integration grouping (B) Individual grouping. UMAP plots show high consistency of the cell annotation grouping results in both short reads and long reads sequencing data.

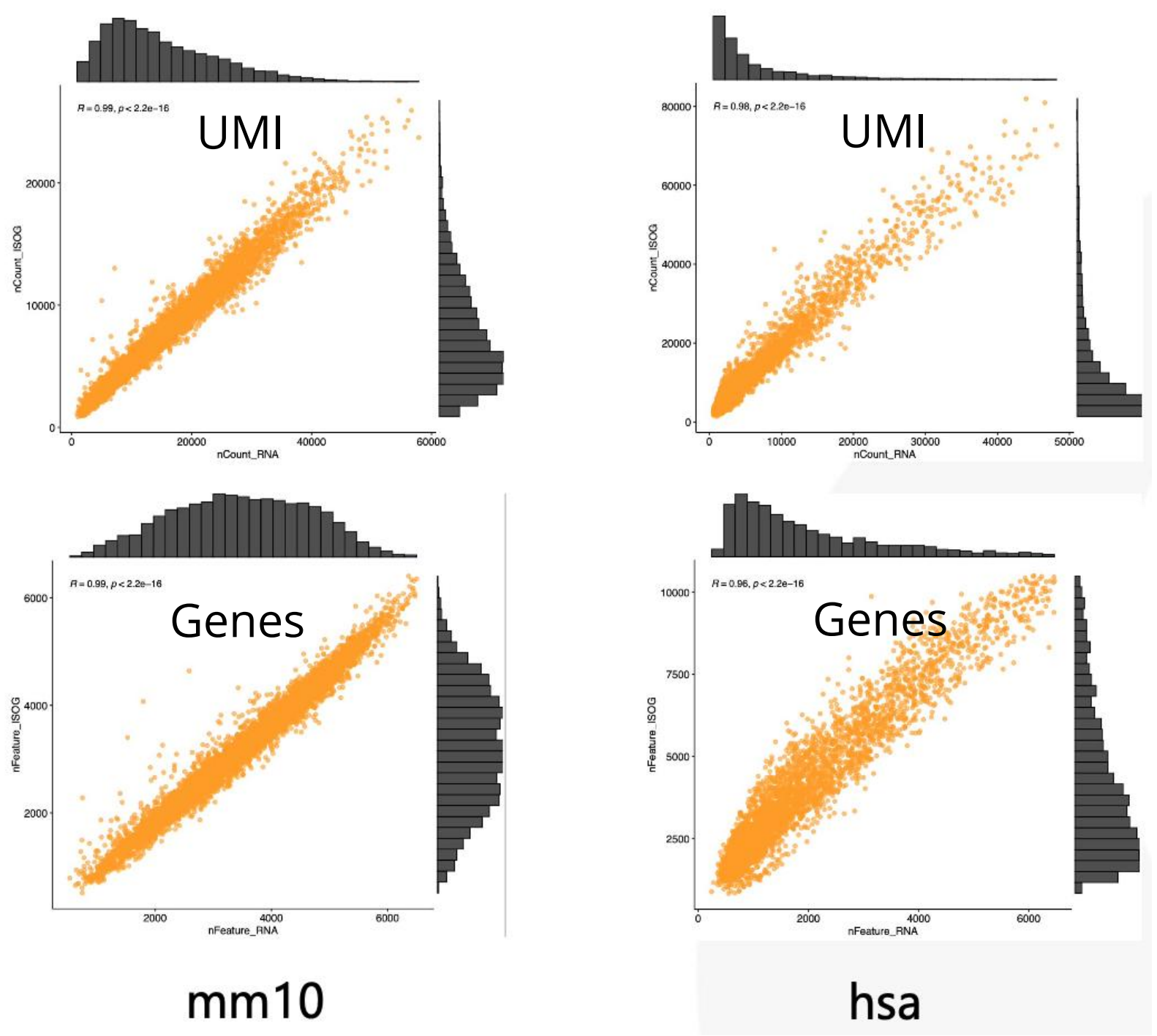
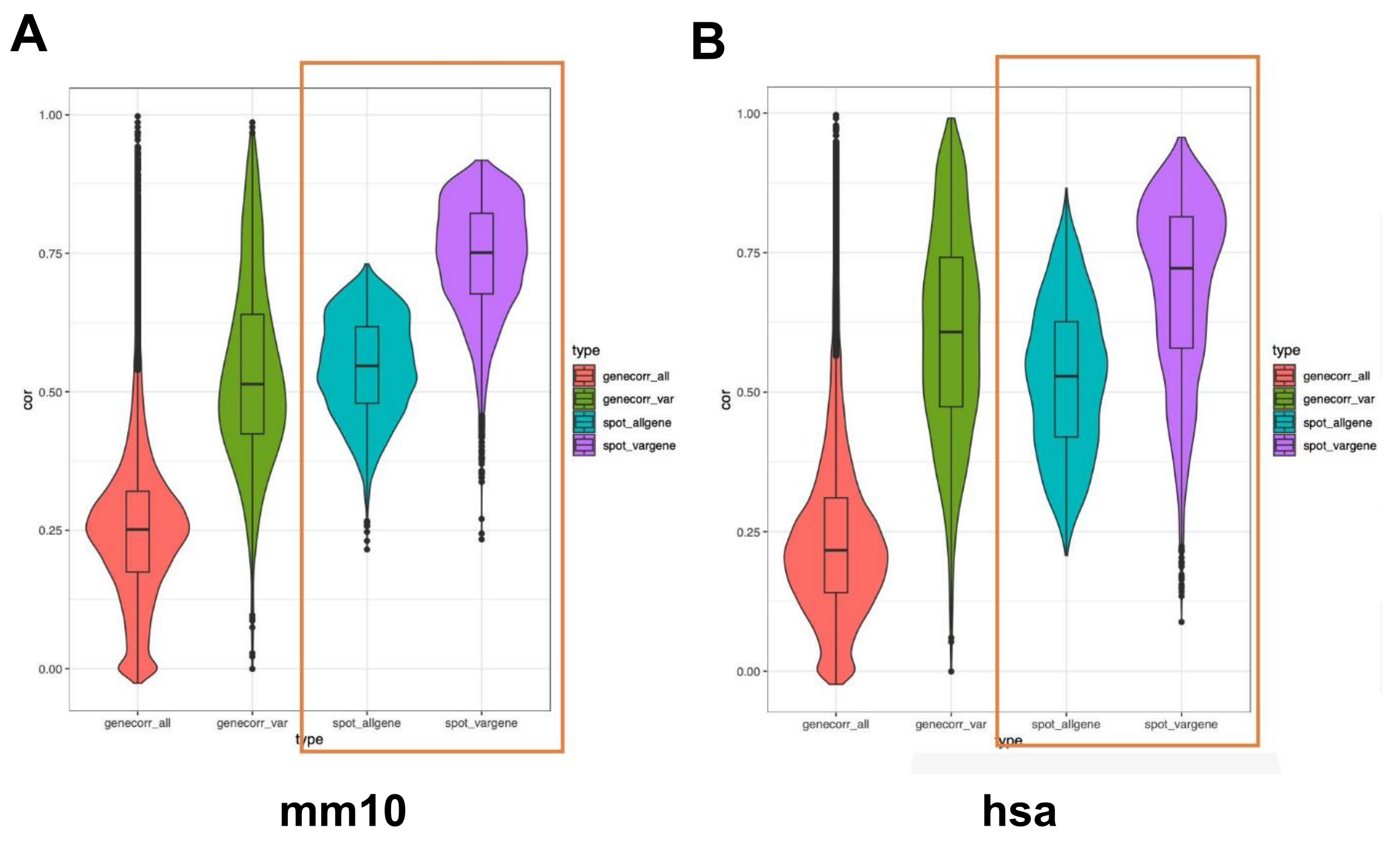


Fig 4. UMI and gene number shows high consistency of > 96%. Each point represents a cell, and the value represents the number of UMIs in the cell (upper figure)/number of genes (lower figure), the x-axis is the short read result, and the y-axis is the Nanopore long read sequencing result.



Legend:

genecorr_all : Correlation analysis of overlap genes of short read and long read
genecorr_var : Correlation analysis of overlap hypervariable genes of short read and long read
spot_allgene : Correlation analysis of genes in overlap cells of short read and long read
Spot_vargene : Correlation analysis of hypervariable genes in overlap cells of the short read and long read

Fig 5. Gene Expression correlation plots. **(A)** Gene expression correlation of mm10 sample. Short read and long read overlap cell num: 6528. all-gene overlap num: 16,049. var_gene overlap num: 1,191; **(B)** Gene expression correlation of hsa sample. Short read and long read overlap cell num: 3512. all_gene overlap num: 15,976. var_gene overlap num: 1,142.

NovogeneAIT Genomics Singapore Pte. Ltd.
(Regional Office for Asia Pacific, Middle-East & Africa)

25 Pandan Crescent #05-15 TIC Tech Centre, Singapore 128577
e: marketing_amea@novogeneait.sg
T: +65 6513 4308
novogene.com/amea-en