Utility Analyses of AVITI Sequencing Chemistry

Silvia Liu[1,2,3,*], Caroline Obert[4], Yan-Ping Yu[1,2,3], Junhua Zhao[4], Bao-Guo Ren[1,2], Kelly Wiseman[4], Benjamin J. Krajacich[4], Wenjia Wang[5], Kyle Metcalfe[4], Mat Smith[4], Tuval Ben-Yehezkel[4] and Jian-Hua Luo[1,2,3,*]

Department of Pathology[1], High Throughput Genome Center[2], Pittsburgh Liver Research Center[3], University of Pittsburgh School of Medicine; Element Biosciences Inc[4], 10055 Barnes Canyon Road, Suite 100, San Diego, CA 92121; Department of Biostatistics[5], University of Pittsburgh School of Public Health.

***- Co-corresponding authors: Jian-Hua Luo, 3550 Terrace Street, Scaife S-728, Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261; E-mail: luoj@upmc.edu; Silvia Liu, S-406 Biomedical Science Tower, 203 Lothrop Street, Pittsburgh, PA 15261; E-mail: shl96@pitt.edu.**

Keywords: nucleic acid sequencing, AVITI, ILLUMINA

Abstract:

Background: DNA sequencing is a critical tool in modern biology. Over the last two decades, it has been revolutionized by the advent of massively parallel sequencing, leading to significant advances in the genome and transcriptome sequencing of various organisms. Nevertheless, challenges with accuracy, lack of competitive options and prohibitive costs associated with high throughput parallel short-read sequencing persist. **Results:** Here, we conduct a comparative analysis using matched DNA and RNA short-reads assays between Element Biosciences AVITI chemistry and Illumina NextSeq 550. Similar comparisons were evaluated for synthetic long-read sequencing for RNA and targeted single-cell transcripts between the AVITI and Illumina NovaSeq 6000. For both DNA and RNA short-read applications, the study found that the AVITI produced significantly higher per sequence quality scores. For PCR-free DNA libraries, we observed up to a 10-fold lower experimentally determined error rate for using the AVITI chemistry compared to the NextSeq 550. For short-read RNA quantification, both AVITI and the NextSeq 550 demonstrated comparable accuracy. With regards to synthetic long-read mRNA and targeted synthetic long read single cell mRNA sequencing, both platforms respective chemistries performed comparably in quantification of genes and isoforms. The AVITI displayed a marginally lower error rate for long reads, with fewer chemistry-specific errors and a higher mutation detection rate. **Conclusion:** These results point to the potential of the AVITI platform as a competitive candidate in high-throughput short read sequencing analyses when juxtaposed with the Illumina NextSeq 550.

Introduction:

DNA and RNA sequencing are essential applications for deeper understanding of modern biology. In the last 20 years, high-throughput massively parallel short-read sequencing platforms have formed the backbone in deciphering the genomes and transcriptomes of numerous organisms [1-3]. As short-read sequencing platforms entered the market, they offered an ever-evolving variety of sequencing mechanisms, throughput capacities and accuracies[4-6]. From that plethora of options, Illumina emerged as the dominant manufacturer of short-read sequencing platforms. One of the most widely used sequencing platforms from the Illumina portfolio is the NextSeq. Although historically, there was a lack of an alternative platform which could offer similar capacity, versatility, and sequencing accuracy, 2022 saw the entry of several new short-read sequencer manufacturers to the market, including Singular Genomics, Ultima Genomics, and MGI[7].

Another recent entry to the short-read sequencing space was that by Element Biosciences, whose AVITI sequencer uses a novel technology known as avidity sequencing. Avidity sequencing utilizes a highly specific multivalent binding between fluorescence-labeled nucleotide polymers and DNA templates within the enzymatic pocket of the engineered DNA polymerase to identify each nucleotide in the DNA templates[8]. The high affinity binding between the nucleotide polymers and DNA templates lead to a reduction of the dissociation constant, decrease in the amount of required sequencing reagents, and an increase in sequencing accuracy. However, the utility of such sequencing improvements remains to be established. In this report, we performed multiple control sample analyses using whole genome, transcriptome, synthetic long-read transcriptome, and targeted synthetic long-read single-cell transcriptome assays on both Illumina and Element Biosciences platforms. The results indicate that while comparable consensus-based long-read metrics were observed between NovaSeq 6000 and AVITI sequencing chemistries, avidity sequencing resulted in substantially better short-read sequencing quality

and accuracy in comparison to that obtained when using Illumina's NextSeq 550, while retaining similar quantification robustness.

Results:

The AVITI benchtop platform from Element Biosciences currently has a capacity of 1 billion polonies per flow cell. This output capacity is more than double that of Illumina's similarly sized NextSeq 550 and NextSeq 1000, which have maximal outputs of 400 million clusters per run and is comparable to that of the newer released NextSeq 2000 which has a maximal output of 1.2 billion clusters. It should be noted, however, that the AVITI is a dual flow cell platform, so the total output is roughly double that of the NextSeq 2000. Although the latter two Illumina platforms were optimized for cluster brightness, reduction in channel cross talk, improved signal to noise ratios and data yield as compared to the NextSeq 550, all three platforms provide equivalent data quality with regards to key applications (https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/nextseq-1000-2000-data-concordance-app-note-970-2020-001.pdf). We utilized Illumina's NextSeq 550 as the control platform for our short-read analyses, to investigate the utility of the AVITI chemistry, normalizing datasets sizes where appropriate. To achieve a robust comparison, we performed short-read sequencing to evaluate a PCR-Free whole genome library preparation of *Escherichia coli* and an mRNA library preparation of an ERCC standard for the evaluation of transcriptome sequencing.

To make direct comparisons of synthetic long-read sequencing from an ERCC standard and single-cell samples from a liver cancer patient (Figure 1), samples starting from the same initial library preparations were utilized in Illumina and Element Biosciences platform specific sequencing. As a NextSeq 2000 was not available for our evaluations, in order to generate a comparable number of short reads from a single sequencing run, we opted to assess the data from a single AVITI flow cell to that generated by a single

lane on a S4 NovaSeq 6000 flow cell. The NovaSeq 6000 is a dual flow cell, production-scale platform, which has a maximum output of 10 billion clusters (or roughly 2.5 billion clusters per lane) per S4 flow cell. Although the platform is significantly larger than the various NextSeq models, similar base quality metrics have been observed (https://q.omnomics.com/ords/f?p=118:34:::::P34_S1,P34_S2:NextSeq%20500,NovaSeq%206000; https://q.omnomics.com/ords/f?p=118:34:::::P34_S1,P34_S2:NextSeq%202000,NovaSeq%206000) (Senabouth et al., 2020). It should be noted though that the percent error rate is lower on the NovaSeq 6000 than the NextSeq 550 (Stoler & Nekrutenko, 2021).

Short-Read Whole Genome Sequencing of the *E. coli* Genome

In this study, which was conducted prior to the September 2022 update for 1 billion polonies per AVITI flow cell, a PCR-free library preparation of the genome of *E.coli* strain B [CIP 103914, NCIB 11595, NRC 745] (~4.6 Mb) was analyzed with 150 cycles of paired-end sequencing. The sequencing outputs from a single run on AVITI and NextSeq 550 were 781,171,133 and 392,056,161 paired end reads, respectively. As shown in Figure 2A, using Element Biosciences v1 chemistry, over 91% of the reads from AVITI in read 1 obtained a per sequence quality score of 40 (equivalent to 0.01% error rate) using FASTQC[9]. The number of AVITI reads with quality score of 40 or higher dropped to 77.8% in read 2, suggesting that the majority of the reads had an accuracy of greater than 99.99%. On the other hand, as the NextSeq platforms only evaluate per base quality scores to Q30, no reads from the NextSeq 550 displayed a similar level of per sequence quality to that of the AVITI. When comparing the number of reads from both platforms that were able to achieve 99.9% accuracy (equivalent to per sequence quality scores of 30), the AVITI was able to ascribe 99.5% reads from read 1 and 97.8% reads from read 2. These results were favorable in comparison with 92.5% reads from read 1 and 89.9% reads from read 2 off the NextSeq 550.

To examine whether the quality differences between the two platforms were related to time-dependent decay in read quality, the Q scores were analyzed cycle by cycle. As shown in Figure 2B, most of the median base Q scores in read 1 of AVITI sequences hovered above 40. The trend of the Q score started to drop after 100 cycles, with the lowest Q score slightly above 37 (99.98% accuracy) at the 150th cycle. On the other hand, most of the average Q scores in read 1 of the NextSeq 550 sequences hovered in mid 30s with a wider range of quality variation. The trend line of Q scores on the NextSeq 550 began to drop after 80 cycles. A significant number of reads fell into abnormal (yellow) range after 120 cycles. Interestingly, there was no overlap of percentile read quality between the two platforms in read 1. For read 2, the Q scores were generally lower for both platforms. The Q scores from AVITI platform appeared to maintain in the normal range (green) for 134 cycles, while some bases fell into abnormal range after 135 cycles. On the other hand, the Q scores from the NextSeq 550 platform maintained most of the reads in normal range up to 64 cycles. The downward trend started at 65 cycles. A significant number of base Q scores fell into abnormal and bad read (red) ranges after 65 cycles. Thus, there was a significant quality gap between the chemistries from the two platforms.

To investigate whether the difference in the quality of sequencing observed in this PCR-free preparation of the *E. coli* genome translated into differences in error rate in sequencing, we analyzed the error rate of the trimmed and filtered mapped reads (AVITI: 781,056,538 paired-end reads, NextSeq 550: 391,975,818 paired-end reads). Among 162.5 billion bp sequenced by AVITI, a little more than 52 million errors were identified. Thus, the error rate of AVITI sequencing was 0.032%. This represented a more than 10-fold decrease of error rate by comparison with 0.37% from NextSeq 550 (343.2 million errors for 92.7 billion bp sequenced) (Figure 2C) when comparing data from matched samples sequenced on each platform. The full-length perfect read rate of AVITI sequencing, however, only improved slightly in

comparison with NextSeq 550 (96.3% vs 96.2%, Figure 2D). To detect whether the errors from AVITI overlapped with those from NextSeq 550, the position of the errors were plotted across the *E. coli* genome (Figure 2E). As shown in Figure 2E, the concordant errors between AVITI and NextSeq550 were rare, suggesting that most of these errors were likely generated from the sequencing chemistries rather than the DNA sample preparation since these libraries shared the early stages of sample preparation.

Short-Read Transcriptome Sequencing

To investigate the utility of AVITI for transcriptome sequencing, we performed short-read RNA sequencing on a sample of an external RNA control consortium (ERCC). ERCC is a standard RNA cocktail of 92 synthetic RNA transcripts ranging from 250 to 2000 nt in length used for quality control analyses. It provides the underlying truth of both RNA quantities as well as the exact sequence of each RNA transcript[10]. AVITI sequencing generated 54.9 million filtered paired-end reads, while NextSeq 550 had 18.5 million. As shown in Figure 3A, both AVITI and NextSeq 550 produced excellent correlations with the published quantities of the RNA, with Pearson correlation coefficients of 0.983. Of note, the correlation between AVITI and NextSeq 550 was even better: reaching a correlation coefficient of 0.999. The same results were obtained when the output of AVITI sequencing was subsampled and filtered to 19.7 million reads for more direct comparison to the NextSeq 550 output (18.5 million reads), suggesting that the quantification by AVITI and NextSeq550 was extremely similar and reproducible. We then analyzed the quality of the ERCC sequencing by AVITI and NextSeq 550. As shown in Figure 3B, 91% of read 1 and 81% of read 2 from AVITI sequencing reached a per sequence quality score of 40 or better (equivalent to 0.01% error rate or lower). In contrast, no reads from NextSeq 550 were evaluated at this confidence level of accuracy, for reasons previously discussed. While more than 95% of all reads from AVITI reached a per sequence quality score of 30 or better, by comparison, only 80% of reads from NextSeq 550 achieved the same values. Similarly, with regards to base sequencing quality, the AVITI

platform also showed fewer errors for both base substitution (0.046% versus 0.13%) and insertion/deletion (0.00064% versus 0.0027%) analyses in comparison with NextSeq 550 (Figure 3C). When analyses were performed on error rate per transcript, the results showed that the AVITI had an average error rate of 0.047% per transcript versus an error rate of 0.14% from NextSeq 550 (Figure 3D). When specific transcripts were analyzed, 86 of 92 transcripts showed lower error rates on the AVITI platform vs NextSeq 550 (Figure 3E and Supplemental Table 1).

Synthetic Long-Read RNA sequencing

The inclusion of long read sequencing has become more commonplace since its introduction in 2009[11], and subsequent technological advancements. There are two fundamental approaches to long-read sequencing: native (also referred to as third generation sequencing) and synthetic (SLR). Native long read platforms, such as those from PacBio and Oxford Nanopore, can improve *de novo* assembly, mapping uncertainty and detection of structural variants[12], however, they are also prone to lower per read accuracies, while requiring higher input amounts and yielding less data[13-15]. Synthetic long reads overcome those constraints by harnessing the accuracy and throughput provided by short-read sequencers. Comparisons of LoopSeq synthetic long reads to native long read platforms have previously been performed, and as such are not included in this paper[16-18].

Long-read RNA sequencing is also essential to detect isoform expression and to quantify isoform mutation expression. Previously, we showed that the combination of intramolecular barcoding with short-read sequencing of RNA resulted in highly accurate synthetic long-read sequencing data[17]. To investigate the ability to generate similarly accurate long read sequences on the AVITI platform, ERCC transcripts were intramolecularly barcoded, fragmented, and sequenced. As shown in Figure 4A, long-read sequencing using AVITI and NovaSeq 6000 achieved similar quantification results in relation to the

reference standard for all detected transcripts: 0.9617 for AVITI and 0.9588 for NovaSeq 6000. The correlations were slightly lower for both platforms (0.9561 for AVITI and 0.9542 for NovaSeq 6000), if only the full-length transcripts were considered. The base substitution error rate of long-read sequencing from AVITI was 0.051%, while insertion/deletion rate was 0.0075% (Figure 4B). On the other hand, NovaSeq 6000 had a base substitution error rate of 0.054% and insertion/deletion rate of 0.0099%. AVITI also had a slightly lower error rate per transcript than NovaSeq 6000: AVITI had a base substitution error per transcript at 0.054% versus 0.067% for NovaSeq 6000, and insertion/deletion error per transcript at 0.018% versus 0.021% for NovaSeq 6000 (Figure 4C). Fifteen transcripts showed higher error rates on the NextSeq platform while 6 transcripts showed higher error rates in AVITI platform (Figure 4D and Supplemental Table 2). Error position analyses showed higher error rates in the average long transcripts (Figure 4E).

Targeted Single-cell Synthetic Long-Read Sequencing

Long-read single-cell sequencing is an important approach to identify mutation allele expression and isoform switching at individual cell level[19]. To investigate the utility of AVITI with long-read single-cell transcriptome analysis, cDNA samples from a liver cancer patient barcoded with 10x Genomics single-cell 3' index were ligated to LoopSeq unique molecular index adapters, underwent probe capture, and were subsequently intramolecularly barcoded (Figure 1). These targeted cDNA were then sequenced in AVITI platform. A more detailed description of the library preparation for this assay is described in Liu et al[19]. As shown in Figure 5A, the AVITI sequencing generated an average of 2510 long-read transcripts per benign liver cell and 2426 per hepatocellular carcinoma (HCC) cell (Figure 5A). These were similar to the results generated on an Illumina NovaSeq 6000 platform with the same library preparation: 2418 long transcripts per benign liver cell and 2304 per HCC cell. Similar isoforms and genes per cell were also identified on the AVITI and NovaSeq platform (Figure 5A). The single-cell gene expression correlation

between AVITI and NovaSeq platforms were high: Pearson's correlation coefficients were 0.9549 for benign liver and 0.9776 for HCC (Figure 5B). The cross-platform correlation for isoforms, however, was a bit lower: 0.9192 for benign liver and 0.958 for HCC.

Next, we performed single nucleotide polymorphism analysis on long-read transcripts generated by AVITI. As shown in Figure 5C, 53775 out of 368794 single nucleotide polymorphisms (or possible real mutations at the DNA level) expressed in the RNA level were validated by whole exome sequencing of HCC samples on the same platform (Figure 5C). In the long-read transcript data generated using short-reads from a NovaSeq 6000, we were able to validate 24368 out of 375785 SNPs using whole exome sequencing. Interestingly, similarly large numbers of SNPs (AVITI: 315019 and NovaSeq 6000: 351417) present in the RNA level were not present in the cancer genome, suggesting significant RNA editing activity or sequencing errors by the procedure. To investigate which of these non-validated SNPs were platform specific, the non-validated SNPs from AVITI and NovaSeq 6000 were overlapped. The results suggested that 252185 SNPs were common between platforms, indicating their existence before the library preparation. In contrast, 99232 SNPs were NovaSeq 6000 specific while 62834 SNPs were AVITI specific. These platform specific SNPs were generated after the library preparation, and thus, were platform errors. The smaller number of platform-specific errors by AVITI suggests a higher accuracy for the sequencing chemistry. To investigate mutation expression, whole exome sequencing was performed on the genome DNA of gallbladder from the same patient to eliminate the physiological SNP expressions. As shown in Figure 5D, 4868 validated mutations were identified in RNA transcripts using the AVITI, while 3789 mutations were identified with the NovaSeq 6000.

To perform gene-based single-cell analysis, genes with similar expressions across all cells were removed, while genes with expression standard deviations of at least 1 were retained to perform uniform

manifold approximation and projection (UMAP) analysis[20, 21]. As shown in Figure 5E, the clustering based on 72 genes from the AVITI platform produced three groups of cells: mostly HCC cells, mostly benign liver cells and a mixture of benign liver and HCC cells (transition). The 47-gene model from Illumina produced similar cell segregation. The 63-isoform UMAP analysis obtained from the AVITI platform produced 3 groups of segregation, albeit a little more dispersed. Interestingly, the 42-isoform model from Illumina platforms produced segregation with larger distance. Overall, both AVITI and the NovaSeq 6000 were comparable in terms of quantifying the transcripts for cell type characterization.

Discussion:

Short-read massively parallel sequencing is fundamental for modern genomics[6]. Prior to 2022, Illumina held the majority of short-read sequencers in the commercial space without an heir apparent[5]. Their platforms utilize sequencing by synthesis (SBS) technology to support numerous applications including whole genome sequencing, RNA-Seq and synthetic long read sequencing. Here, we examine whether the AVITI platform described may present a viable alternative to Illumina sequencing. AVITI's affinity-based chemistry produces lower requirements for sequencing reagents, higher confidence of base calling, and lower error rate, while maintaining high throughput capacity.

The lower error rates observed with the AVITI platform are the result of separating the sequencing process into two parts: extension of DNA template by synthesis and affinity identification of the nucleotide in the DNA template one at a time, allowing for optimization of nucleotide identification in a separate procedure. Avidite ligands have higher affinity than the nucleotide monomer, resulting in less sequencing reagents employed in the sequencing process. Dissociation constants of the avidite-polymerase-DNA template complex are negligible and minimal background can be achieved, enabling improved imaging without significant decay. These features may enable a large magnitude of

improvement in quality over the traditional sequencing methods. For a more in depth description of Avidity sequencing, see Arlsan et al[8].

High quality sequencing usually translates into fewer errors detected in analysis. For short-read *E. coli* sequencing, there was up to a 10-fold decrease of error rate detected in reads from the AVITI platform versus the NextSeq 550. However, such improvement was reduced to around 3-fold when transcriptome sequencing was analyzed. There was an improvement in error rate for the NextSeq 550 platform with regards to transcriptome sequencing which saw a reduction in the error rate from 0.37% in *E. coli* genome sequencing to 0.14% in ERCC transcriptome sequencing. The AVITI performed similarly for both applications: *E.coli* (0.032%) and ERCC sequencing (0.047%). Such improvement could be related to less homopolymer sequence in the ERCC RNA. The mild increase of error rate of ERCC sequencing in AVITI platform could be due to errors generated by reverse transcriptase. For synthetic long-read sequencing, the error rate differences between Illumina and Element platforms were statistically insignificant. We speculate that the high coverage per nucleotide may correct most of the errors generated by either platform.

High quality sequencing translates into high confidence of base-calling and lower errors. Low error rate is important in mutation calling and polymorphism analyses. Due to the comparatively lower confidence of base-calling on current sequencing instruments, the mutation calling threshold has been usually set at a large number of reads. This condition requires the increase of sequencing depth which results in a higher cost of sequencing. In human oncology, mutation calling has formed the backbone of modern diagnostic and therapeutic bases[22, 23]. However, when significant heterogeneity and benign sample dilution are present, ambiguity will be generated. Ambiguity of mutation calling may negatively impact the diagnosis and treatment of cancers. The higher confidence of base calling and lower error rate in

avidity sequencing may offer an excellent solution in such situations: a platform with lower error rates may, in the future, help calling correct mutations with fewer reads, and might decrease the cost of sequencing by increasing the signal to noise ratio in mutation detection. It might also increase the number of mutation calls by discovering the mutations previously masked by higher error rates. As a result, Avidity sequencing on the AVITI platform may hold promise to become an important role in increasing the fidelity of modern genomic analyses.

Methods:

PCR-Free Whole Genome Short-Read Library Preparation

Illumina: PCR-Free libraries were prepared with 1 µg/reaction input of *E. coli* (ATCC 11303) using the KAPA HyperPlus (Cat#: KK8514) reagents with KAPA UDI Adapters (Cat#: KK8726). Instead of utilizing the enzymatic fragmentation reagents provided in the kit, acoustic shearing on a Covaris ME220 platform using 8 microTUBE -50 AFA Fiber H Slit Strip v2 tubes (Cat#: 520240.2) was performed the following settings to produce an average insert size of 350 bp: Sample volume: 55 µL; Waveguide: ME220 Waveguide 8 Place (PN 500526); Temperature Set Point: 12 °C; Repeat/Iterations: 7; Repeat Process Treatment Duration (s): 10; Peak Power (W): 50; Duty Factor (%): 20; Cycles Per Burst: 1000; Avg Power: 10; Total Treatment Time per sample (s): 70. Fifty microliters per reaction of the mechanically sheared gDNA was transferred from the respective AFA tubes into corresponding wells of a clean PCR plate. After adding 10 µl of the prepared master mix, end repair and A-tailing was performed on a Bio-Rad C1000 PCR machine with the following parameters: 30 min at 20°C, 30 min at 65°C. Adapter ligation was conducted by adding 5 µl of the respective KAPA Adapter and 45 µL of the prepared master mix to each well. Reactions were incubated in a PCR machine for 15 min at 20°C with the lid temperature off. Samples were then subjected to a 0.8x SPRI cleanup and a dual SPRI size selection (0.5x/0.66x). All reactions were evaluated for concentration in triplicate using Qubit 1x dsDNA HS assay reagents (Cat#:

Q33231) on an Agilent NEO2 Plate Reader and for insert size using HS NGS Reagents (DNF-473-1000) reagents on an Agilent Fragment Analyzer 5300 automated CE platform. Individual reactions of the prepared libraries were normalized for equimolar pooling. The final library pool was evaluated using an Agilent Bioanalyzer 2100 using High Sensitivity DNA Assay reagents (Cat#: 5067-4626). Samples were denatured and diluted for loading on the NextSeq 550 platform per manufacturer's recommendations.

AVITI: The equimolar pooled library was concentrated using an Amicon Ultra-30K concentrator column and centrifuging at 14,000 x g for 10 minutes per manufacturer's guidelines. It should be noted that updates to the Element Biosciences circularization chemistry no longer require the higher input concentration utilized during this preparation. The concentrated library pool was then subjected to circularization with the Element Biosciences Adept Library Compatibility Kit v1.1 (Cat#: 830-00007) according to manufacturer instructions. The circularized library was quantified using qPCR with the provided standard before being sequenced on the AVITI system with 2 x 150 length paired end reads.

mRNA Transcriptome Short-Read Library Preparation

Illumina: For transcriptome sequencing, similar procedures were described previously[24-27]. Briefly, a 1% dilution of ERCC RNA Spike-In Mix (4456740, ThermoFisher) was processed into a short-read library using the TruSeq Stranded mRNA Sample Prep Kit from Illumina, Inc (Cat#: 20020594). The fragmented mRNA was reverse transcribed to cDNA using random primers. This was followed by second stranded cDNA synthesis. The library preparation processes of mono-adenylation, adapter ligation and amplification were performed following the manual provided by the manufacturer. The quantity and quality of the libraries were assessed in a manner similar to that described for whole genome DNA library preparation. The procedure of 150 cycle paired-end sequencing on the Illumina NextSeq550 followed per the manufacturer's guidelines.

AVITI: The linear library pool prepared above was amplified with the KAPA HiFi HotStart Library Amplification kit with Primer Mix (#KK2621) to ensure blunt-end products prior to circularization (90s at 98°C; 5 cycles of 30s at 98°C, 30s at 60°C, 30s at 72°C; and 60s at 72°C). Circularization was performed using the Element Biosciences Adept Library Compatibility Kit v1.1 (Cat#: 830-00007) according to manufacturer instructions. The circularized library was quantified using qPCR with the provided standard before being sequenced on the AVITI system with 2 x 150 length paired end reads.

Synthetic Long-Read mRNA Transcriptome Library Preparation

Illumina: LoopSeq preparation of ThermoFisher ERCC RNA Spike-in Mix (Cat#: 4456740) was executed as follows. First, reverse transcription of 11ul ERCC RNA of a 1% diluted template was performed, using a prepared mix of Maxima H Minus RT Buffer and Reverse Transcriptase 200U/µl (Cat#: EP0752, ThermoFisher), RiboLock RNAse inhibitor 40U/µl (Cat#: EO0381, ThermoFisher), 10mM dNTP (Cat#: N0447L, New England Biolabs), LoopSeq specific 100uM Template Switching Oligo (IDT), and a LoopSeq specific 10uM polyTVN UMI-tagging primer oligo (IDT). Incubation conditions were 50°C for 90 min followed by 85°C for 5 min.

UMI diversity in the RT product was then quantified by qPCR. As the RT polyTVN primer oligo includes a UMI sequence, each first strand generated by RT was uniquely UMI-tagged, allowing quantification by qPCR using LoopSeq Amp Mix S and LoopSeq Amp Additive from the Amplicon LoopSeq kit (# 840-10002, Element Biosciences). A master mix of Amp Mix S and Amp Additive was generated following manufacturer instructions and was spiked with 50X SYBR Green (# S7585, Invitrogen) to final concentration of 0.5X SYBR. A dilution series of the RT product was created and quantified by qPCR to determine the dilution needed for an optimized concentration of 15,000 UMIs/µl. PCR was performed

for 22 cycles using PCR cycling conditions of 95°C for 3 min, 95°C 30s, 60°C 45s, 72°C 8 min, and a 10°C hold.

After the appropriate dilution was determined by qPCR, twelve replicate 20µl PCR reactions were prepared using LoopSeq Amp Mix S, LoopSeq Amp Additive, and RT product diluted to 15,000 UMIs/µl as template. PCR was performed as described above. 0.6x SPRISelect (Cat#: B23319, Beckman-Coulter) was used for cleanup and generation of a final pool of 20µl. Pooled PCR product was used as input into the Extension LoopSeq kit (Cat#: 840-10003, Element Biosciences).

AVITI: Circularization of the synthetic long read transcriptome library prepared above was performed using the Adept Library Compatibility Kit v1.1 (Cat#: 830-00007, Element Biosciences) according to manufacturer instructions. The circularized library was quantified using qPCR with the provided standard before being sequenced on the AVITI system with 2 x 150 length paired end reads.

Synthetic Long Read Single-Cell Transcriptome Library Preparation

Illumina: The Synthetic Long Read Single Cell Transcriptome library was prepared and sequenced as described in Liu et al[19].

AVITI: Circularization of the synthetic long read single cell transcriptome library prepared above was performed using the Adept Library Compatibility Kit v1.1 (#830-00007, Element Biosciences) according to manufacturer instructions. The circularized library was quantified using qPCR with the provided standard before being sequenced on the AVITI system with 2 x 150 length paired end reads.

Bioinformatics Analysis for Short-Read Whole Genome Sequencing Data

E. coli DNA sequences were measured by both Illumina and Element Bioscience platforms. For the raw sequencing data, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)[9] was first applied to check the sequencing quality over read cycle. Then, Trimmomatic[28] was employed to trim the adapter sequences and remove the low-quality reads. After the pre-processing, the surviving reads were aligned to the *E. coli* reference genome by Burrows-Wheeler Alignment Tool[29]. BCFtools[30] was then applied to call the variant sites. Sequencing errors (mismatches and small indels) were calculated from the variant sites, excluding the known SNP sites from the *E. coli* strain.

Bioinformatics Analysis for Short-Read RNA-Seq Data

ERCC spike-in samples were measured by bulk RNA-seq for both platforms. Similar to WGS, the raw RNA-seq data were pre-processed by FastQC[9] and Trimmomatic[28] for quality control and trimming. Then the reads were aligned to the ERCC reference genome by STAR aligner[31]. Transcript counts were collected based on the alignment and the count per million (CPM) was calculated for transcript quantification. BCFtools[30] were applied on the aligned file to call the variant sites. Average error rate was calculated for all the sequencing base-pairs and for each transcript.

Bioinformatics Analysis for Long-Read RNA-Seq Data

ERCC samples measured by the Loop Seq technology were collected from both platforms. Short-read data produced by each platform were processed using a pipeline to generate LoopSeq synthetic long reads. In this pipeline, adapters were trimmed from the short reads using Trimmomatic[28], short reads were demultiplexed by LoopSeq sample index and UMI, and finally assembled usingSPAdes[32]. Synthetic long reads were then aligned to the ERCC reference genome by Minimap2 aligner[33]. The downstream transcript quantification and error rate calculation follow the same pipeline as the ERCC short-read data.

Bioinformatics Analysis for Targeted Long-Read Single-Cell RNA-Seq Data

Single-cell long-read RNA-seq was performed on a paired tumor – benign sample from an HCC patient by both Illumina and Element Bioscience platforms. Based on short-read sequencing, full length transcripts were assembled into long reads by tool SPAdes[32]. In total, four runs were performed on each library and were pooled together for analysis. Long reads were first annotated by SQANTI[34] based on human genome reference hg38, where each long read was assigned to a known gene and transcript, or novel one otherwise. Then the long reads were demultiplexed by their cell barcode (10X index) and molecule barcodes (LoopSeq index). Eventually, based on the long-read annotation and cell assignment, UMIs were quantified at both gene and isoform level. Valid cells were defined as cells with at least 1000 long-read transcripts per cell. Gene counts per cell and isoform counts per cell were normalized to count per million and compared between the two platforms.

For gene/ isoform expression analysis, R/Bioconductor package Seurat[35] was applied to integrate both benign and tumor libraries. Top genes and isoforms were selected by their expression standard deviation across all the valid cells. Cell clustering was performed based on these selected genes and isoforms, and the cells were visualized by UMAP dimension reduction algorithm[20, 21].

For variant calling, single-cell long reads were aligned to human reference genome hg38 by Minimap2 aligner[33]. Variant calling was performed by BCFtools[30]. For the same sample, whole exome sequencing was measured, and variant calling was performed to detect variant events at DNA level. Variants unique to RNA data are potential RNA editing sites or sequencing errors. Variants detected by both RNA and DNA data are potential mutations.

Statistical data analysis

All the statistical analyses were performed by R programming. Pearson correlation was calculated comparing pairwise expression levels.

Data availability

The raw and processed RNA sequencing data has been deposited to GEO database with accession ID GSE246838 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE246838). The ERCC bulk short-read RNA-seq by Element Biosciences AVITI platform can be downloaded from GSE246830, the ERCC bulk short-read RNA-seq by Illumina NextSeq platform can be downloaded from GSE246832; the ERCC bulk Loop-seq long-read RNA-seq by Element Biosciences AVITI platform can be downloaded from GSE246834; the ERCC bulk Loop-seq long-read RNA-seq by Illumina NVSQ600 platform can be downloaded from GSE246835; the HCC single-cell Loop-seq long-read RNA-seq by Element Biosciences AVITI platform can be downloaded from GSE246836; The HCC single-cell Loop-seq long-read RNA-seq by Illumina Novaseq platform has been deposited into GEO with accession ID GSE223743 in our previous research. The raw DNA sequencing data has been deposited to SRA BioProject PRJNA1034988 (http://www.ncbi.nlm.nih.gov/bioproject/1034988). The submission includes DNA-seq on *E. coli* by both Element Biosciences AVITI and Illumina NextSeq platforms, and whole exome sequencing on HCC, hepatocyte and gallbladder samples using Element Biosciences AVITI platform.

Acknowledgment: This work is in part supported by grants from the National Cancer Institute (1R56CA229262-01 to JHL), the National Institute of Digestive Diseases and Kidney (P30- DK120531-01), the National Institutes of Health (UL1TR001857 and S10OD028483) and The University of Pittsburgh Clinical and Translational Science Institute.

Conflict of interest: SL, YPY, BGR, WW, and JHL declare no conflict of interest, while CO, JZ, KB, BJK, KM, MS, and TBY are employees of Element Biosciences, Inc.

Contribution: JHL, SL and TBY conceived the idea. YPY, BGR, TBY, CO, and JHL contributed the experiment design. BGR, CO, JZ, KB, BJK, KM performed the experiments. SL, CO, YPY, BJK, WW, TBY, and JHL performed the analyses.

Reference list

1. Nyren P, Pettersson B, Uhlen M: **Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay.** *Anal Biochem* 1993, **208:**171-175.
2. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P: **Real-time DNA sequencing using detection of pyrophosphate release.** *Anal Biochem* 1996, **242:**84-89.
3. Ronaghi M, Uhlen M, Nyren P: **A sequencing method based on real-time pyrophosphate.** *Science* 1998, **281:**363, 365.
4. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems.** *J Biomed Biotechnol* 2012, **2012:**251364.
5. McCombie WR, McPherson JD, Mardis ER: **Next-Generation Sequencing Technologies.** *Cold Spring Harb Perspect Med* 2019, **9**.
6. Pervez MT, Hasnain MJU, Abbas SH, Moustafa MF, Aslam N, Shah SSM: **A Comprehensive Review of Performance of Next-Generation Sequencing Platforms.** *Biomed Res Int* 2022, **2022:**3457806.
7. Pollie R: **Genomic Sequencing Costs Set to Head Down Again.** *Engineering* 2023, **23:**3-6.
8. Arslan S, Garcia FJ, Guo M, Kellinger MW, Kruglyak S, LeVieux JA, Mah AH, Wang H, Zhao J, Zhou C, et al: **Sequencing by avidity enables high accuracy with low reagent consumption.** *Nat Biotechnol* 2023.
9. Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, Waite D, Davey RP: **Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics.** *Front Genet* 2013, **4:**288.
10. Finnis M, Dayan S, Hobson L, Chenevix-Trench G, Friend K, Ried K, Venter D, Woollatt E, Baker E, Richards RI: **Common chromosomal fragile site FRA16D mutation in cancer cells.** *Hum Mol Genet* 2005, **14:**1341-1349.
11. Pushkarev D, Neff NF, Quake SR: **Single-molecule sequencing of an individual human genome.** *Nat Biotechnol* 2009, **27:**847-850.
12. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q: **Opportunities and challenges in long-read sequencing data analysis.** *Genome Biol* 2020, **21:**30.
13. Logsdon GA, Vollger MR, Eichler EE: **Long-read human genome sequencing and its applications.** *Nat Rev Genet* 2020, **21:**597-614.
14. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C: **The Third Revolution in Sequencing Technology.** *Trends Genet* 2018, **34:**666-681.
15. Warburton PE, Sebra RP: **Long-Read DNA Sequencing: Recent Advances and Remaining Challenges.** *Annu Rev Genomics Hum Genet* 2023, **24:**109-132.
16. Callahan BJ, Grinevich D, Thakur S, Balamotis MA, Yehezkel TB: **Ultra-accurate microbial amplicon sequencing with synthetic long reads.** *Microbiome* 2021, **9:**130.
17. Liu S, Wu I, Yu YP, Balamotis M, Ren B, Ben Yehezkel T, Luo JH: **Targeted transcriptome analysis using synthetic long read sequencing uncovers isoform reprograming in the progression of colon cancer.** *Commun Biol* 2021, **4:**506.
18. Yu T, Cheng L, Liu Q, Wang S, Zhou Y, Zhong H, Tang M, Nian H, Lian T: **Effects of Waterlogging on Soybean Rhizosphere Bacterial Community Using V4, LoopSeq, and PacBio 16S rRNA Sequence.** *Microbiol Spectr* 2022, **10:**e0201121.
19. Silvia Liu Y-PY, Bao-Guo Ren, Tuval Ben-Yehezkel, Caroline Obert, Mat Smith, Wenjia Wang, Alina Ostrowska, Alejandro Soto-Gutierrez, Jian-Hua Luo: **Long-read single-cell sequencing reveals expressions of hypermutation clusters of isoforms in human liver cancer cells.** *bioRxiv* 2023.

20. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW: **Dimensionality reduction for visualizing single-cell data using UMAP.** *Nat Biotechnol* 2018.
21. Lim HS, Qiu P: **Quantifying Cell-Type-Specific Differences of Single-Cell Datasets Using Uniform Manifold Approximation and Projection for Dimension Reduction and Shapley Additive exPlanations.** *J Comput Biol* 2023, **30:**738-750.
22. Koboldt DC: **Best practices for variant calling in clinical sequencing.** *Genome Med* 2020, **12:**91.
23. Singer J, Irmisch A, Ruscheweyh HJ, Singer F, Toussaint NC, Levesque MP, Stekhoven DJ, Beerenwinkel N: **Bioinformatics for precision oncology.** *Brief Bioinform* 2019, **20:**778-788.
24. Liu S, Nalesnik MA, Singhi A, Wood-Trageser MA, Randhawa P, Ren BG, Humar A, Liu P, Yu YP, Tseng GC, et al: **Transcriptome and Exome Analyses of Hepatocellular Carcinoma Reveal Patterns to Predict Cancer Recurrence in Liver Transplant Patients.** *Hepatol Commun* 2022, **6:**710-727.
25. Luo JH, Liu S, Tao J, Ren BG, Luo K, Chen ZH, Nalesnik M, Cieply K, Ma T, Cheng SY, et al: **Pten-NOLC1 fusion promotes cancers involving MET and EGFR signalings.** *Oncogene* 2021, **40:**1064-1076.
26. Yu YP, Ding Y, Chen Z, Liu S, Michalopoulos A, Chen R, Gulzar ZG, Yang B, Cieply KM, Luvison A, et al: **Novel fusion transcripts associate with progressive prostate cancer.** *Am J Pathol* 2014, **184:**2840-2849.
27. Yu YP, Ding Y, Chen R, Liao SG, Ren BG, Michalopoulos A, Michalopoulos G, Nelson J, Tseng GC, Luo JH: **Whole-Genome Methylation Sequencing Reveals Distinct Impact of Differential Methylations on Gene Transcription in Prostate Cancer.** *Am J Pathol* 2013.
28. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
30. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27:**2987-2993.
31. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
32. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** *J Comput Biol* 2012, **19:**455-477.
33. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-3100.
34. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al: **SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification.** *Genome Res* 2018.
35. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al: **Integrated analysis of multimodal single-cell data.** *Cell* 2021, **184:**3573-3587 e3529.

Figure 1. Schema of AVITI sequencing flowchart. PCR-free whole genome sequencing (WGS), transcriptome sequencing, long-read transcriptome sequencing, and targeted single-cell long-read sequencing processes are outlined. After linear amplification and circularization of the single stranded sequencing template, the process enters the cycle of sequencing process: avidite binding, wash, one base nucleotide extension, reversal of extension block.

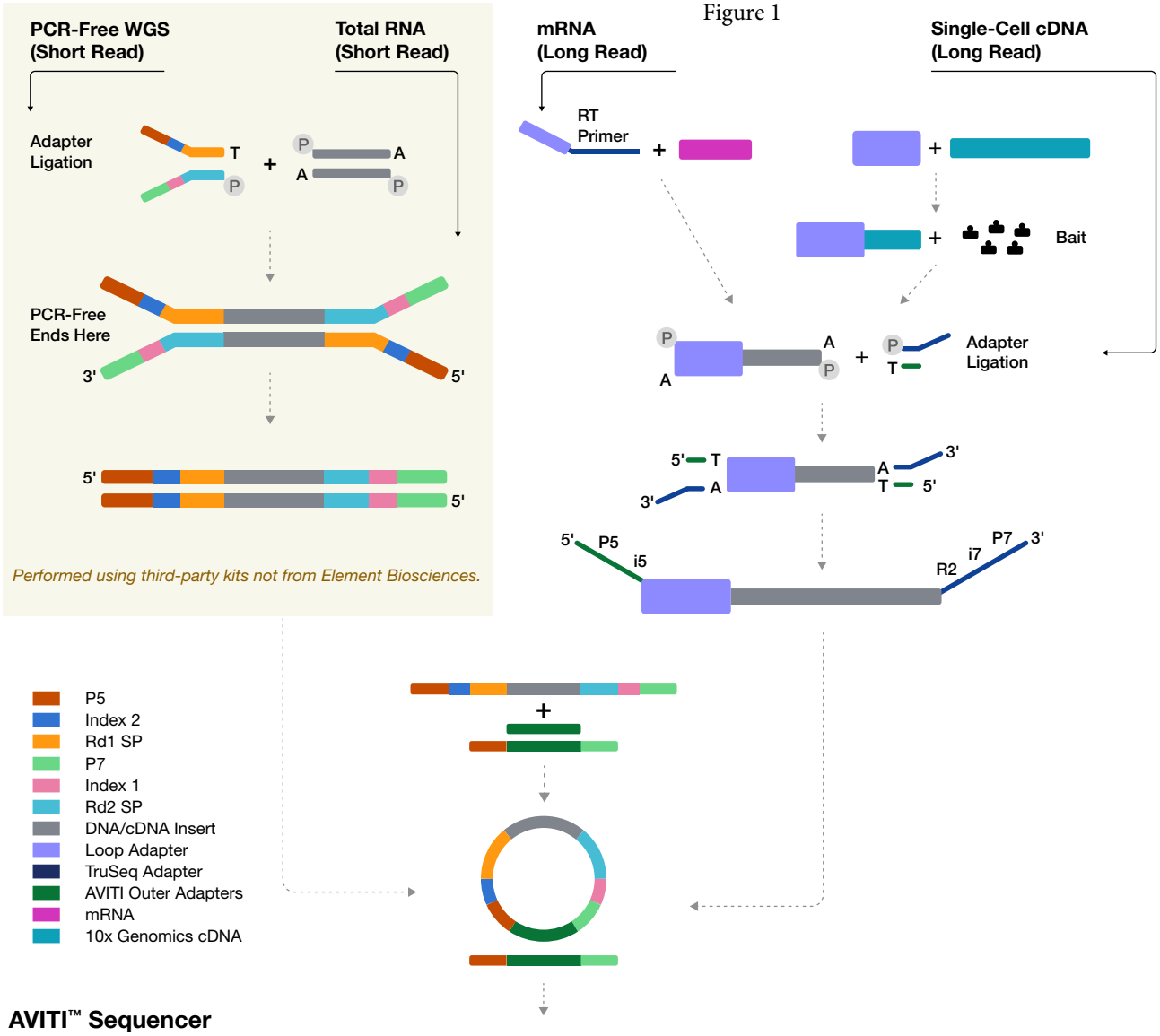
Figure 2. Sequencing of *E. coli* Genome by AVITI and NextSeq550. (A) Per sequence quality score distributions of AVITI and NextSeq550 sequencing of *E. coli* genome. Percentage of reads with scores more than 40, 30, and 20 scores are indicated. (B) Phred score distribution cycle by cycle of AVITI and NextSeq550. Whisker box plots represent 10, 25, 50, 75 and 90 percentile distributions of reads in the indicated cycles. Both read 1 and read 2 were analyzed. (C) Percent error rate of the mapped reads from AVITI and NextSeq550. (D) Percentage of the mapped reads from AVITI and NextSeq550. (E) Histogram of 2000 loci error rate in the *E. coli* genome by AVITI and NextSeq550.

Figure 3. Sequencing of ERCC RNA standard by AVITI and NextSeq550. (A) Scatter plot of expression quantification of ERCC RNA between reference standard, AVITI quantification and NextSeq550 quantification. AVITI-sub indicates reduction of sample size from AVITI sequencing to match the mapped read quantity from NextSeq550. (B) Per sequence quality score distributions of AVITI and NextSeq550 sequencing on ERCC standard. Percentage of reads with scores more than 40, 30, and 20 scores are indicated. (C) Percent error rates of AVITI and NextSeq550 on ERCC standard RNA. (D) Error rates per transcript of AVITI and NextSeq550 on ERCC standard RNA. (E) Scatter plot of error rates of individual ERCC RNA between AVITI and NextSeq550 sequencing.

Figure 4. Long-read sequencing of ERCC RNA standard by AVITI and NovaSeq 6000. (A) Scatter plot of expression long-read transcript quantification of ERCC RNA between reference standard, AVITI quantification and NovaSeq 6000 quantification. (B) Percent error rates of AVITI and NovaSeq 6000 on ERCC long-read standard RNA. (C) Percent error rates per long-read transcript of AVITI and NovaSeq 6000 on ERCC standard RNA. (D) Scatter plot of error rates of individual long-read ERCC RNA between AVITI and NovaSeq 6000 sequencing. Blue line indicates linear regression, while the shaded area represents the confidence interval of the linear regression. (E) Distribution of errors by the position of ERCC RNA in AVITI and NovaSeq 6000 sequencing.

Figure 5. Long-read single-cell sequencing by AVITI and NovaSeq platform. (A) Distribution of targeted long-read single-cell sequencing transcripts, gene, or isoforms per cell by Element AVITI and Illumina NovaSeq platforms. (B) Scatter plots of quantification of long-read transcripts by Element AVITI and Illumina NovaSeq. Pearson correlation coefficients are indicated. (C) Distribution of detected nucleotide polymorphism in long-read transcripts and the genome of HCC by Element AVITI and Illumina NovaSeq. Venn diagram shows the common polymorphisms detected by both genome and transcript analysis. The polymorphisms not detected in the genome of cancer were overlapped to identify sequencing errors specific to each platform. (D) Mutation expression in the long-read single-cell sequencing detected by AVITI and NovaSeq based on the screening of the validated polymorphism against the exome sequencing of gallbladder of the same individual. (E) UMAP clustering of HCC and benign liver cells based on standard deviation of gene or isoform expression across all the cells by Element AVITI and Illumina NovaSeq. Groups of cells of benign liver, transition and HCC are indicated.

Figure 1


PCR-Free WGS
(Short Read)
Total RNA
(Short Read)
mRNA
(Long Read)
Single-Cell cDNA
(Long Read)
Adapter
Ligation
T
P
A
PCR-Free
Ends Here
3'
5'
Performed using third-party kits not from Element Biosciences.
RT
Primer
Bait
Adapter
Ligation
P5
i5
R2
i7
P7
P5
Index 2
Rd1 SP
P7
Index 1
Rd2 SP
DNA/cDNA Insert
Loop Adapter
TruSeq Adapter
AVITI Outer Adapters
mRNA
10x Genomics cDNA
AVITI™ Sequencer

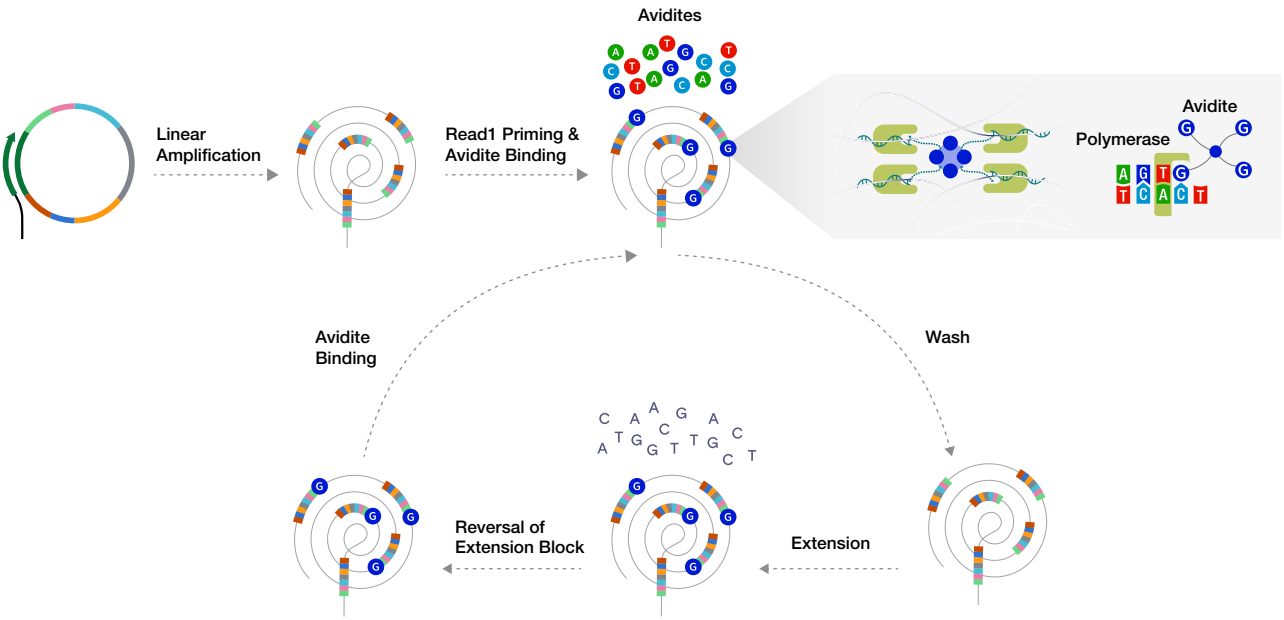

Linear
Amplification
Read1 Priming &
Avidite Binding
Avidites
Avidite
Polymerase
Avidite
Binding
Wash
Reversal of
Extension Block
Extension

Figure 2

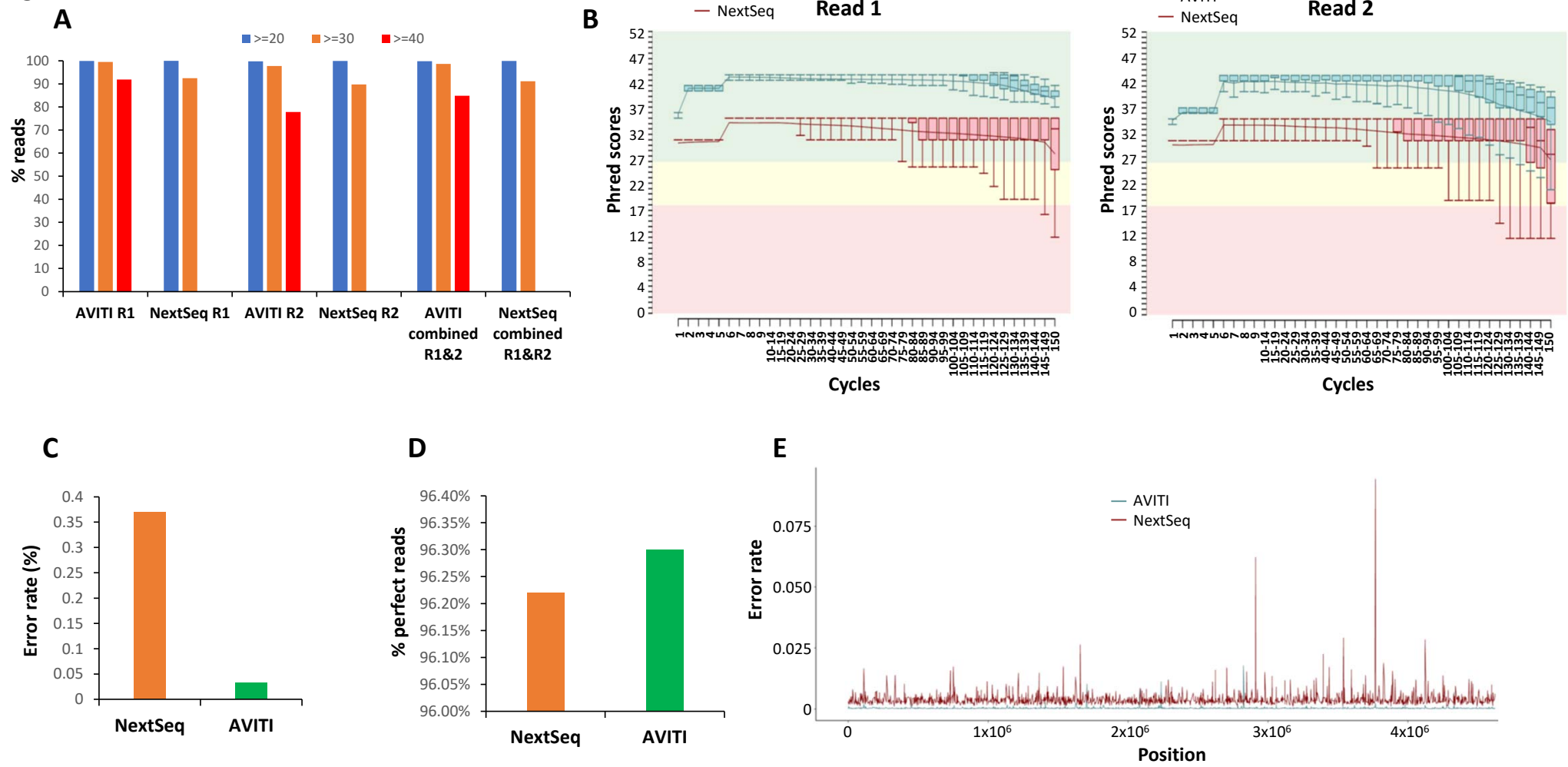
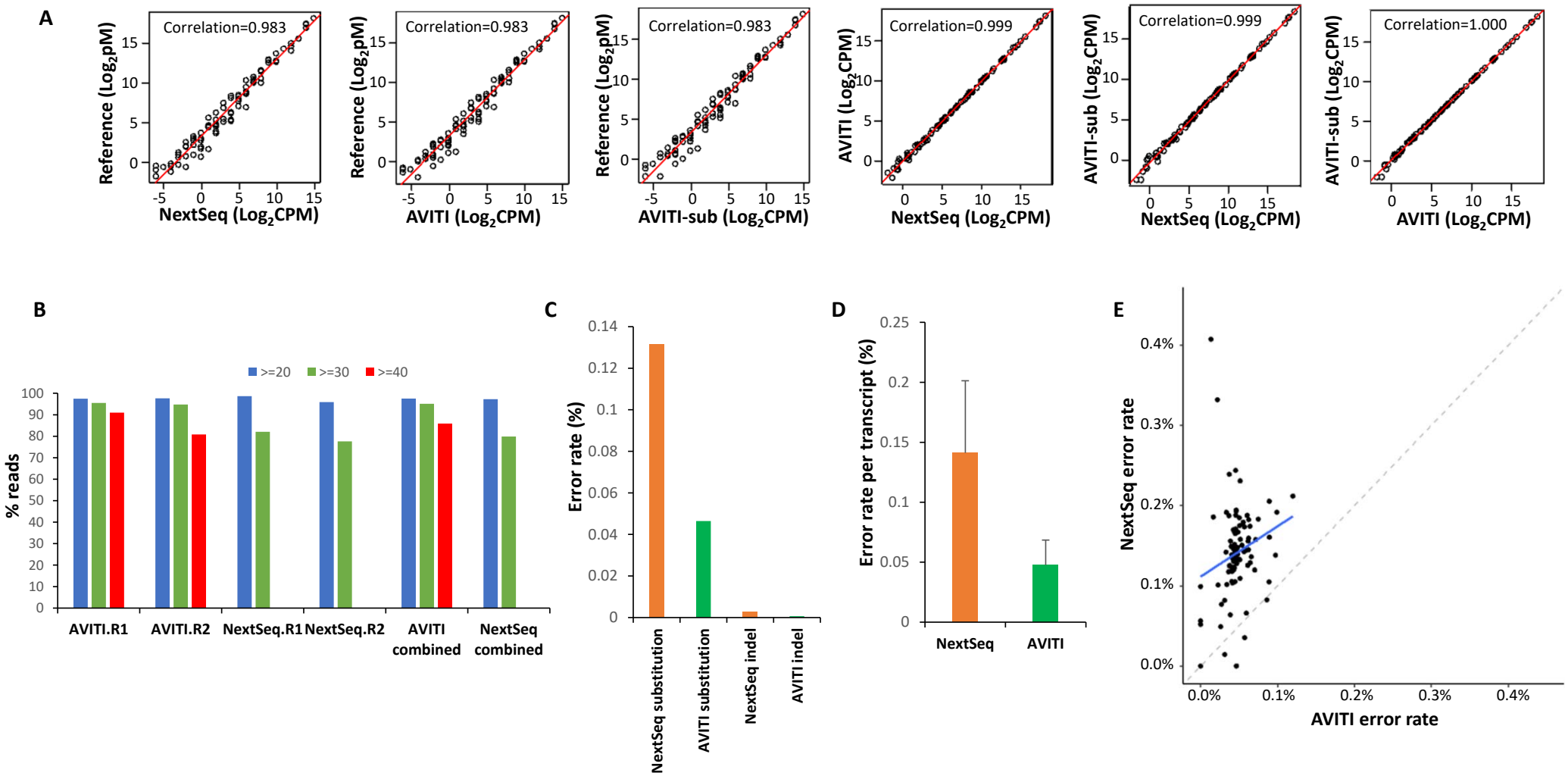

A
>=20
>=30
>=40
% reads
AVITI R1
NextSeq R1
AVITI R2
NextSeq R2
AVITI combined R1&2
NextSeq combined R1&R2
B
AVITI
NextSeq
Read 1
Phred scores
Cycles
Read 2
C
Error rate (%)
NextSeq
AVITI
D
% perfect reads
NextSeq
AVITI
E
AVITI
NextSeq
Error rate
Position

Figure 3


A
Correlation=0.983
Reference (Log2pM)
NextSeq (Log2CPM)
Correlation=0.983
Reference (Log2pM)
AVITI (Log2CPM)
Correlation=0.983
Reference (Log2pM)
AVITI-sub (Log2CPM)
Correlation=0.999
AVITI (Log2CPM)
NextSeq (Log2CPM)
Correlation=0.999
AVITI-sub (Log2CPM)
NextSeq (Log2CPM)
Correlation=1.000
AVITI-sub (Log2CPM)
AVITI (Log2CPM)
B
>=20
>=30
>=40
% reads
AVITI.R1
AVITI.R2
NextSeq.R1
NextSeq.R2
AVITI combined
NextSeq combined
C
Error rate (%)
NextSeq substitution
AVITI substitution
NextSeq indel
AVITI indel
D
Error rate per transcript (%)
NextSeq
AVITI
E
NextSeq error rate
AVITI error rate

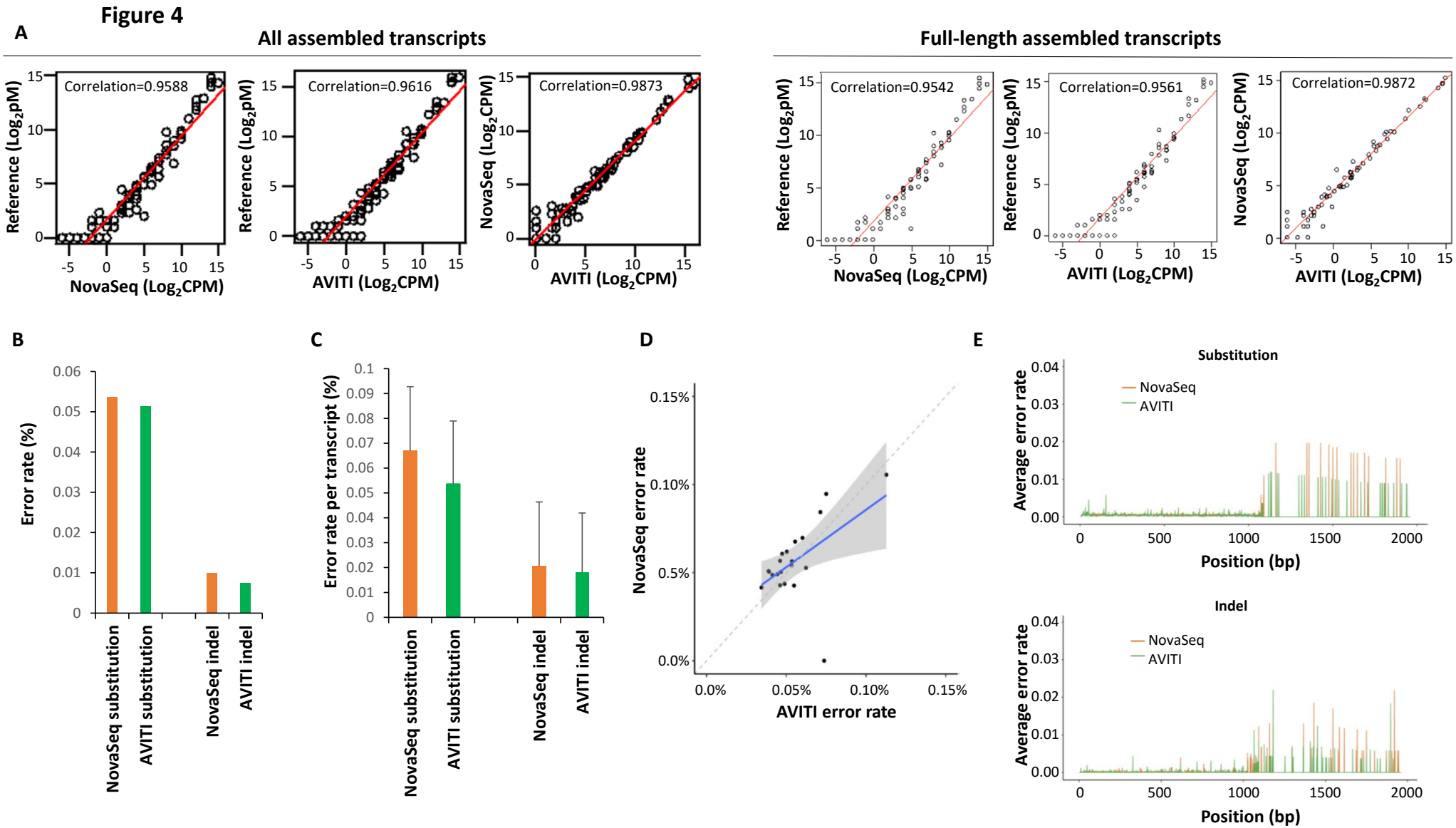
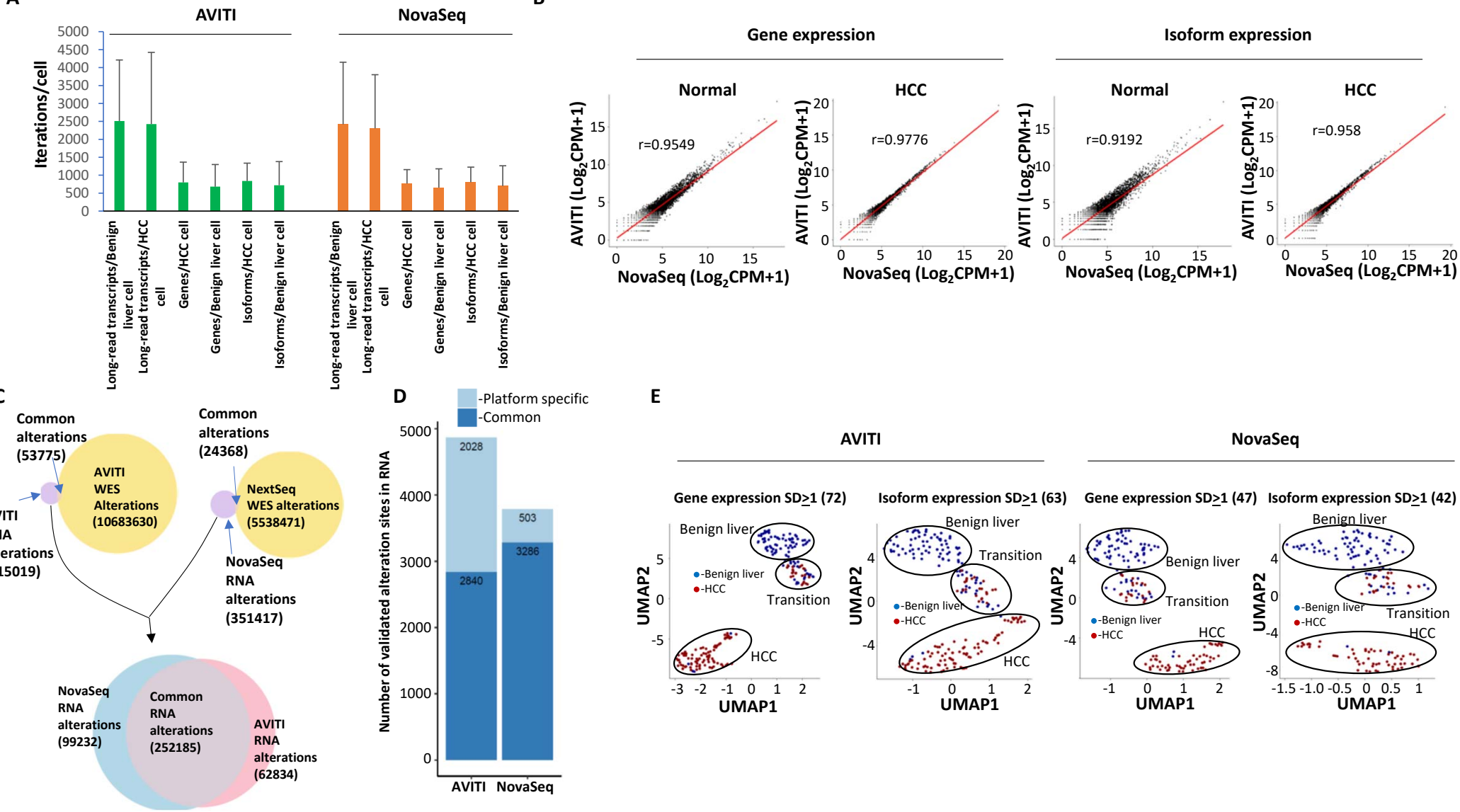


Figure 4
A
All assembled transcripts
Full-length assembled transcripts
Correlation=0.9588
Correlation=0.9616
Correlation=0.9873
Correlation=0.9542
Correlation=0.9561
Correlation=0.9872
Reference (Log2pM)
NovaSeq (Log2CPM)
AVITI (Log2CPM)
B
Error rate (%)
NovaSeq substitution
AVITI substitution
NovaSeq indel
AVITI indel
C
Error rate per transcript (%)
D
NovaSeq error rate
AVITI error rate
E
Substitution
Indel
Average error rate
Position (bp)
NovaSeq
AVITI

Figure 5


A
AVITI
NovaSeq
Iterations/cell
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Long-read transcripts/Benign liver cell
Long-read transcripts/HCC cell
Genes/HCC cell
Genes/Benign liver cell
Isoforms/HCC cell
Isoforms/Benign liver cell
B
Gene expression
Isoform expression
Normal
HCC
r=0.9549
r=0.9776
r=0.9192
r=0.958
AVITI (Log2CPM+1)
NovaSeq (Log2CPM+1)
C
Common alterations (53775)
AVITI WES Alterations (10683630)
AVITI RNA alterations (315019)
Common alterations (24368)
NextSeq WES alterations (5538471)
NovaSeq RNA alterations (351417)
NovaSeq RNA alterations (99232)
Common RNA alterations (252185)
AVITI RNA alterations (62834)
D
-Platform specific
-Common
Number of validated alteration sites in RNA
2028
2840
503
3286
AVITI
NovaSeq
E
AVITI
NovaSeq
Gene expression SD≥1 (72)
Isoform expression SD≥1 (63)
Gene expression SD≥1 (47)
Isoform expression SD≥1 (42)
Benign liver
Transition
HCC
-Benign liver
-HCC
UMAP1
UMAP2