APPLICATION NOTE

Microbial *de novo* assembly with linked-read technology

TELL-Seq using Illumina platforms enables rapid, cost-effective, and highly accurate linked-read data

Linked-read sequencing powered by

UNIVERSAL SEQUENCING
innovation for all

illumina®

Introduction

Next-generation sequencing (NGS) is an important tool for analyzing small genomes (≤ 10 Mb), including those of bacteria, viruses, and other microbes. Microbial NGS, including whole-genome sequencing (WGS) and targeted resequencing, enables mapping and *de novo* assembly of novel organisms, completing genomes of known organisms, and comparing genomes across samples. Short-read sequencing (≤ 600 bp per read) is highly accurate and cost-effective for many applications. For *de novo* sequencing, where a reference genome is not available for alignment and reads are assembled as contigs, Universal Sequencing Technology Corporation offers Transposase Enzyme Linked Long-read Sequencing (TELL-Seq), a simple, scalable library prep solution. TELL-Seq uses linked-read sequencing to apply the advantages of short-read Illumina NGS for generating highly accurate and cost-effective long-range sequencing information for assembly of highly polished reference genomes.[1]

This application note demonstrates the exceptional performance of TELL-Seq as part of a comprehensive workflow for microbial *de novo* genome assembly using Illumina NGS Systems (Figure 1).

Methods

Sample preparation

Eight bacterial species with differing GC content (Table 1) were obtained from the American Type Culture Collection (ATCC). Genomic DNA (gDNA) was extracted directly from freeze-dried material using the MagAttract HMW DNA Kit (QIAGEN, Catalog no. 67563).

Library preparation

Libraries were prepared from 0.5 ng of input gDNA using the TELL-Seq WGS Library Prep Kit (Universal Sequencing, Catalog no. 100001). TELL-Bead input into barcoding was consistent across samples and input for PCR was based on genome size; 1.5 µl beads per 5 Mb genome is recommended.

Sequencing

Libraries were sequenced on a MiSeq™ System with a run configuration of 146 × 18 × 8 × 146 bp. Libraries can also be sequenced on a MiniSeq™ System.

Data analysis

Sequencing data was streamed directly from the instrument into the cloud ecosystem for analysis using the BaseSpace™ TELL-Seq Data Analysis App for linked-read analysis and *de novo* assembly.

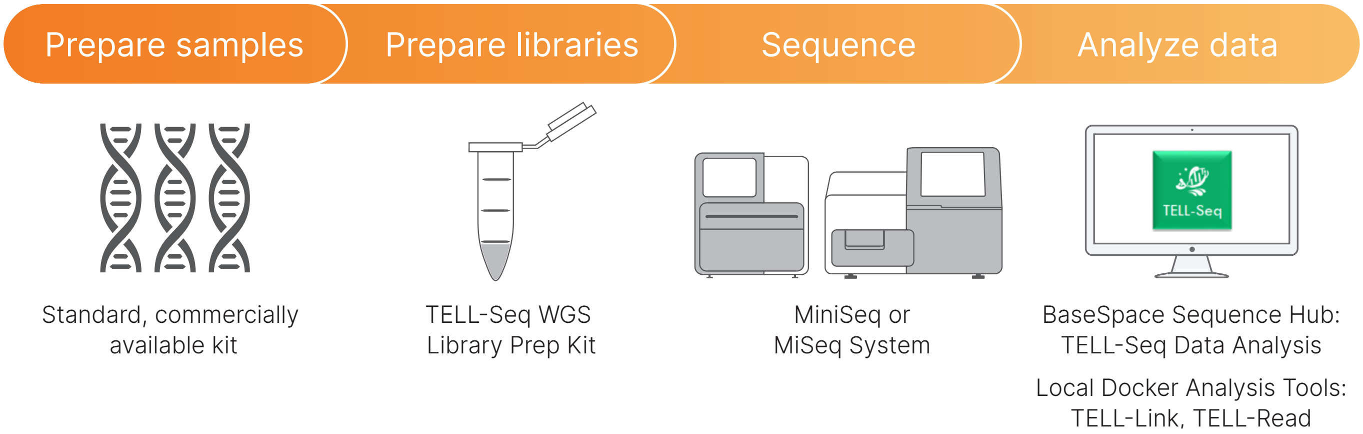
Prepare samples
Prepare libraries
Sequence
Analyze data
TELL-Seq
Standard, commercially available kit
TELL-Seq WGS Library Prep Kit
MiniSeq or MiSeq System
BaseSpace Sequence Hub: TELL-Seq Data Analysis
Local Docker Analysis Tools: TELL-Link, TELL-Read

Figure 1: TELL-Seq workflow—Linked-read sequencing is an integrated, DNA-to-data workflow that includes TELL-Seq library preparation, sequencing on a MiniSeq or MiSeq System, and data analysis with the TELL-Seq BaseSpace App or TELL-Link and Tell-Read Local Docker analysis tools.

For Research Use Only. Not for use in diagnostic procedures.

Table 1: Overview of sequenced microbial genomes

Bacteria	ATCC Catalog no.	Gram	Genome size (bp)	% GC content	No. of chromosomes/ plasmids
Clostridium perfringens	ATCC 13124	Positive	3,256,676	28.38%	1/0
Campylobacter jejuni	ATCC 700819	Negative	1,637,699	30.56%	1/0
Bacillus cereus	ATCC 14579	Positive	5,430,163	35.29%	1/1
Bacillus subtilis	ATCC 6051	Positive	4,295,427	43.35%	1/1
Escherichia coli MG1655	ATCC 700926	Negative	4,642,497	50.79%	1/0
Rhodopseudomonas palustris	ATCC 17001	Negative	5,262,262	65.15%	1/0
Bordetella pertussis	ATCC 9797	Negative	4,045,794	67.68%	1/0
Rhodobacter sphaeroides	ATCC BAA-808	Negative	4,628,173	68.77%	2/5

Results

The TELL-Seq LIbrary Prep Kit was evaluated for microbial WGS and *de novo* assembly.

Quality control (QC) of gDNA extraction

To maximize performance with the TELL-Seq WGS Library Prep Kit, input gDNA fragment length is recommended to be greater than 20 kb, with fragments smaller than 10 kb removed before library preparation. For this evaluation, purified gDNA from the various bacterial species was assayed by pulsed field gel electrophoresis. The average fragment size was ~ 30 kb with smearing below 10 kb (Figure 2). While not ideal, these samples were carried through library prep and sequencing without removal of any small DNA fragments.

Robust library yield with low input

Prepared TELL-Seq libraries displayed a broad size distribution of 300-1000 bp (Figure 3). Distribution of fragment sizes was not impacted by genome size or % GC content. TELL-Seq library prep resulted in moderate yield for various microbial species with low input (Figure 4). Microbial species with high GC content (> 60%) consistently resulted in lower yields compared to other samples; however, this did not impact sequencing metrics.

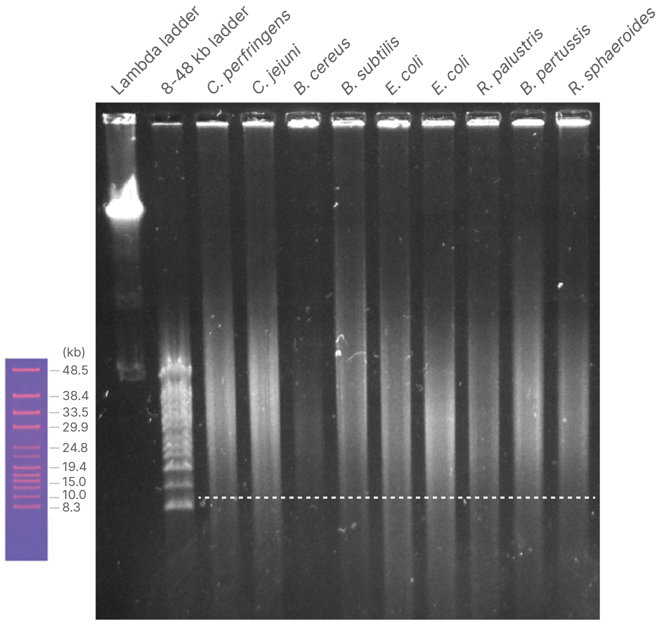

Figure 2: Extracted gDNA sizing and QC—Sizing of extracted gDNA from the various bacterial species was determined by agarose gel electrophoresis. Fragments were ~ 30 kb on average with smearing below 10 kb (dashed line).

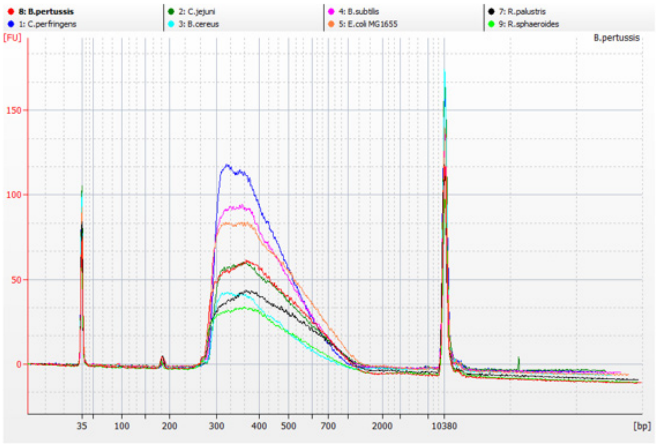

Figure 3: Comparison of library size distribution across microbial genomes of varying GC content—TELL-Seq libraries exhibit a broad fragment size range of 300-1000 bp, regardless of genome size or GC content.

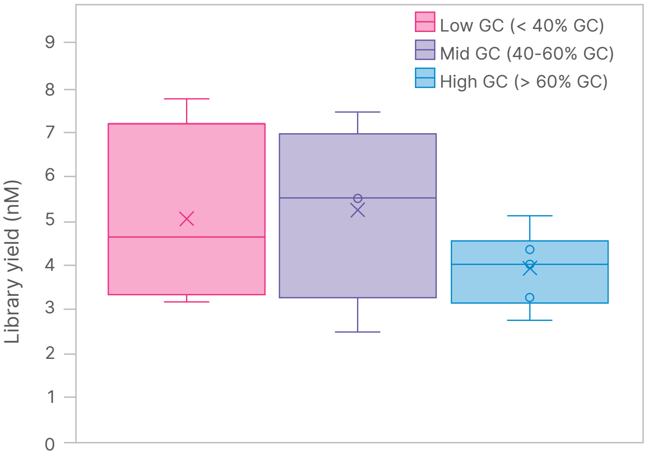

Figure 4: Comparison of library yields across microbial genomes of varying GC content—Species with high GC content (> 60%) consistently showed reduced yield compared to species with low (< 40%) and mid (40-60%) GC content.

Consistent assembly metrics across microbial species

Sequencing data was downsampled to 1-3M read pairs and 100× coverage per microbial sample. Analysis with the TELL-Seq BaseSpace App showed fairly consistent metrics across samples with varying GC content, including mean and median SLF size (Table 2), an indicator of the original molecule size going into library preparation.

Uniform coverage across genomes with varying GC content

Analysis with the TELL-Seq BaseSpace App showed that most assemblies were high quality as measured by NG50 values, number of contigs, number of misassemblies, and genome coverage percentages (Table 2 and Figure 5). Importantly, the quality of genome assembly was not significantly affected by genome size or GC content (Figure 5). It should be noted that the relatively high number of misassemblies observed with *B. pertussis* is likely due to the challenging nature of that genome (Table 2).

Near complete assembly of chromosomes and plasmids

Plasmids are a common component of bacterial genomes, yet can be difficult to study by microbial WGS and left out of *de novo* assembly.[2] Analysis with the TELL-Seq BaseSpace App showed near complete assembly of both chromosomes and plasmids in the same bacterial species (Table 3).

Table 2: *De novo* assembly metrics

	C. perfringens	*C. jejuni*	*B. cereus*	*B. subtilis*	*E. coli MG1655*	*R. palustris*	*B. pertussis*	*R. sphaeroides*
Genome size	~3.3 Mb	~1.6 Mb	~5.4 Mb	~4.3 Mb	~4.6 Mb	~5.3 Mb	~4.0 Mb	~4.6 Mb
GC content	28%	31%	35%	43%	51%	65%	68%	69%
Chromosome no./ plasmids no.	1/0	1/0	1/1	1/1	1/0	1/0	1/0	2/5
Reference sequence source	ATCC	ATCC	ATCC	ATCC	ATCC	ATCC	ATCC	NCBI[a]
TELL-Read analysis metrics								
Read pairs	1,650,000	800,000	3,040,000	2,230,000	2,410,000	2,720,000	2,130,000	2,450,000
Mean coverage	100×	99×	100×	100×	100×	100×	100×	100×
Mean SLF size (bp)	25,416	24,903	15,219	23,533	21,226	20,761	27,676	20,132
Median SLF size (bp)	18,761	17,680	10,239	17,542	15,653	14,713	21,044	15,351
TELL-Link analysis metrics								
No. of contigs (≥ 10 kb)	1	1	2	2	1	1	2	8
Contig NG50	3,237,188	1,615,713	5,315,730	4,184,919	4,618,375	5,242,132	3,577,459	3,155,294
Misassemblies	2	1	2	0	0	3	20	3
Genome fraction	99.11%	99.51%	99.03%	99.44%	99.71%	99.96%	96.03%	98.91%

a. Results indicated that the ATCC reference genome for *R. sphaeroides* may be incomplete; the NCBI reference genome was used for analysis.

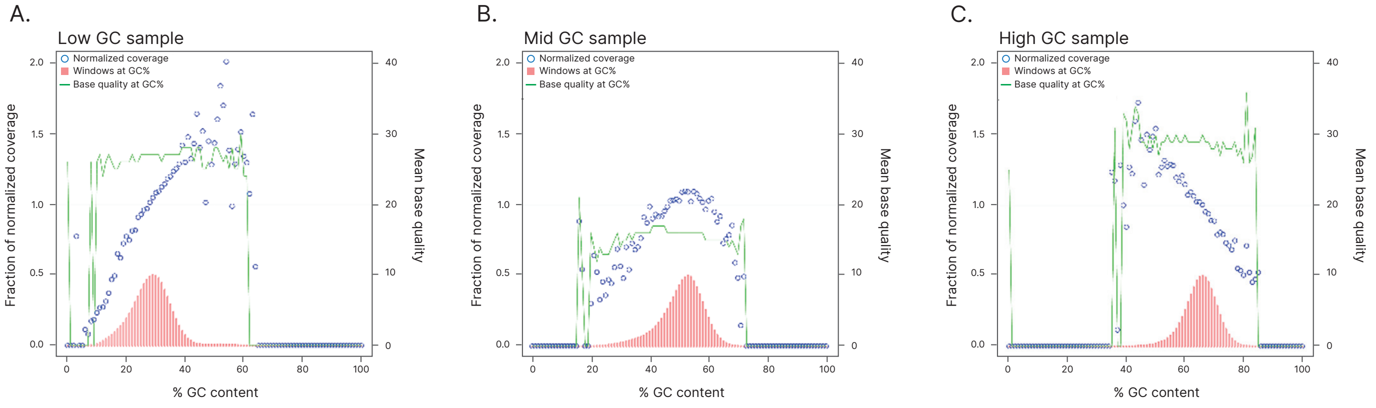

Figure 5: Comparison of read coverage across microbial genomes of varying GC content—TELL-Seq provides consistent and comparable read coverage across microbial genomes of varying GC content. Representative coverage plots are shown for samples with (A) low, (B) mid, and (C) high GC content. Normalized coverage is shown with blue circles, actual GC composition with red bars, and base quality as a function of GC % with a green line.

Table 3: Assembly of chromosomes and plasmids

	Reference		TELL-Seq Assembly	
	Chromosome (bp)	Plasmid (bp)	Chromosome (bp)	Plasmid (bp)
B. cereus	5,414,965	15,198	5,315,730	15,265
B. subtilis	4,211,212	84,215	4,184,919	84,280
R. sphaeroides	3,188,521	124,310	3,155,294	185,204
	942,929	114,178	935,061	106,877
		105,281		47,532
		100,819		20,948
		52,135		11,871
				10,448

Only contigs ≥ 10 kb are listed

Summary

TELL-Seq technology enables Illumina NGS systems to generate highly accurate data while reducing costs, turnaround time, and DNA input requirements, making TELL-Seq an ideal solution for applications such as highly polished small genome *de novo* assembly. This application note demonstrates the exceptional performance of TELL-Seq library preparation combined with Illumina NGS for microbial WGS, even for samples with challenging genomic regions with high GC content or suboptimal input DNA.

Learn more

Microbial WGS,
illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/microbial-whole-genome-sequencing.html

MiSeq Sequencing System,
illumina.com/systems/sequencing-platforms/miseq.html

TELL-Seq technology,
universalsequencing.com/technology

References

1. Chen Z, Pham L, Wu TC, et al. Ultra-low input single tube linked-read library method enables short-read second-generation sequencing systems to generate highly accurate and economical long-range sequencing information routinely. *Genome Res*. 2020;30(6):898-909. doi: 10.1101/gr.260380.119.
2. Williams LE, Detter C, Barry K, et al. Facile recovery of individual high-molecular-weight, low-copy number natural plasmids for genomic sequencing. *Appl Environ Microbiol*. 2006;72(7):4899-4906.

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com