TELL-Seq™ Application Note:

Microbial Metagenome Assembly, Taxonomy, and Abundance Estimation

UNIVERSAL SEQUENCING
innovation for all

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

May 2022
Document # 100040, version 1.0

Introduction

Understanding microbial diversity at the strain level has become critical to studying human health and disease. It is similarly crucial for some segments in the food industry and the analysis of wastewater and soil samples in urban and rural areas to cite some examples. Metagenomics—the study of genomic material in a mixed community of organisms—enables the characterization of microbial diversity in clinical, food, and environmental samples (Riesenfeld et al., 2004). However, the difficulty of assembling a genome without using a reference becomes only more complex when recovering and annotating genomes from metagenomic sequencing data. The presence of multiple organisms in a microbial mixture with widely different levels of abundance, relatedness with each other, and repetitive content can vastly complicate the genome assembly process (Ghurye et al., 2016; Breitwieser et al., 2019). In this context, Universal Sequencing Technology (UST) has developed Tell-TaxContigs—a computational pipeline that enables taxonomic identification of bacterial and archaeal species and estimation of their relative abundance. Tell-TaxContigs uses pre-assembled metagenomic data generated with the Transposase Enzyme-Linked Long-read-Sequencing (TELL-Seq™) WGS Library Prep kit and Tell-Link software, developed by Universal Sequencing Technology.

The TELL-Seq library captures long-range information in short sequencing reads—known as linked reads (Chen et al., 2020) — and can be sequenced on a short-read NGS platform, such as an Illumina® instrument. With the long-range resolution that TELL-seq linked reads provide, *de novo* assembly with Tell-Link enables highly contiguous (often complete) genome assemblies of microbial isolates. TELL-Seq-scaffolded fragments are based on Illumina's short reads, which ensures robust sequence fidelity compared to long-read NGS platforms. While, in contrast, long-read technology may require coupling with short-read technology to ensure enough sequence accuracy. Contigs available for genome binning from Tell-Link-powered metagenomic assembly are expected to be superior to other methods in terms of contiguity and fidelity, as well as in higher completion rates of metagenome-assembled genomes with the potential to identify novel microbial species and improve upon existing genome references.

In this application note, we use two popular ATCC® Microbiome Standards as benchmarks for genome assembly of mixed microbial populations, taxonomic characterization and relative abundance estimation. One standard is a mock microbial community that mimics a mixed sample with an even abundance of bacterial strains ('Even'). The other standard is a mock microbial community that mimics a mixed sample with a diversity of abundance of bacterial strains ('Staggered'). We show how TELL-seq in combination with Tell-Link and Tell-TaxContigs can assemble metagenomic sequencing reads and recover fully completed genomes from a microbial mixture sample.

Methods

Sample and library preparations

Genomic DNA from the 20 Strain Even Mix Genomic Material (MSA-1002™, ATCC) and the 20 Strain Staggered Mix Genomic Material (MSA-1003™, ATCC) was used. The Even mix has identical relative abundances for 20 bacterial strains (5%), whereas the Staggered mix has relative abundances for the same 20 bacterial strains ranging from 0.02% to 18%. These 20 bacterial strains belong to the genera *Acinetobacter*, *Bacillus*, *Bacteroides*, *Bifidobacterium*, *Clostridium*, *Cutibacterium*, *Deinococcus*, *Enterococcus*, *Escherichia*, *Helicobacter*, *Lactobacillus*, *Neisseria*, *Porphyromonas*, *Pseudomonas*, *Rhodobacter*, *Actinomyces/Schaalia*, *Staphylococcus*, and *Streptococcus*. One nanogram of DNA from each mixture was processed with the TELL-Seq WGS Library Prep kit and approximately 2.4 million (M) TELL beads in a 22ul reaction. Half of the TELL beads were used for PCR (10 cycles). Moreover, a second prep was generated with the Staggered mix using 5 ng of DNA and approximately 9.5M TELL beads in a 66ul reaction (referred to as 'Staggered-deep'). In this second case, all TELL beads were used for PCR in this case (9 cycles). This second prep was used for deep sequencing (see next section).

Sequencing

Using the TELL-Seq Illumina Sequencing Primer kit, the Even and Staggered TELL-Seq libraries were sequenced on the NextSeq™ instrument (Illumina; mid output kit), while the Staggered-deep TELL-Seq library was sequenced on the NovaSeq™ 6000 instrument (Illumina; S4 flow cell). The 2 × 146 bp PE sequencing protocol was followed, generating 59 million (M) reads for the Even mix, 98M reads for the Staggered mix, and 700M reads for the Staggered-deep mix.

Data analysis

Sequencing data was processed according to guidance in the TELL-Seq Data Analysis Roadmap User Guide (Universal Sequencing Technology). Briefly, sequencing outputs were first processed with Tell-Read software (Universal Sequencing Technology) for demultiplexing, linked-read FASTQ data conversion, adapter trimming, barcode error corrections and QC reporting. Then, Tell-Link (UST) was applied to build barcode-aware assembly graphs and assemble contigs.

Tell-TaxContigs was used to analyze Tell-Link pre-assembled reads to output a reliable list of species identifications and their relative abundances. An overview of this pipeline is shown in **Figure 1**. For each contig, Tell-Tax Contigs consolidates their BLAST hits from the NCBI Microbial Genome Database and identifies contigs that align well with a single species and show significantly lower similarity to other species. Using this species list, Tell-Tax Contigs discovers likely community members by finding the species that have a robust percentage of reads uniquely mapping to them. Using deep learning, Tell-Tax Contigs (i) determines potential organisms by eliminating false positives from the likely community members based on evolutionary similarity, contig abundancy, and contig representations; (ii) calculates the abundance of potential organisms based on contig binning and contig abundance; and (iii) adjusts organism abundance using motifs based on GC/AT-rich assessments at organism level. For the last step, a training set is used to identify motifs that result from sequencing known GC/AT-rich genomes and how this relates to errors made in their estimated abundances. Then, the presence of these motifs is assessed in potential organisms and a model is built to adjust their abundancy accounting for the GC/AT bias. Finally, Tell-Tax Contigs reclassifies unused contigs by either mapping them to the identified potential organisms or suggests novel species in the community.

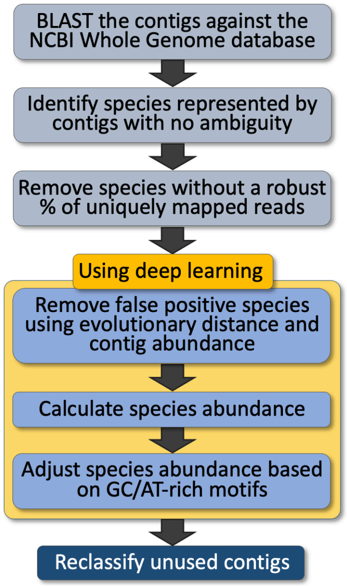
BLAST the contigs against the NCBI Whole Genome database
Identify species represented by contigs with no ambiguity
Remove species without a robust % of uniquely mapped reads
Using deep learning
Remove false positive species using evolutionary distance and contig abundance
Calculate species abundance
Adjust species abundance based on GC/AT-rich motifs
Reclassify unused contigs

Figure 1. Tell-TaxContigs pipeline

To assess the quality of the assembly process, Metagenome-Assembled Genomes (MAGs) were generated using a combination of Tell-Link and Metabat2 to bin contigs (Dongwan et al., 2019). GTDB-Tk software 1.5.0. based on the Genome Taxonomy Database (GTDB) was used for taxonomic classification (Parks et al., 2018; Chaumeil et al., 2019; Parks et al., 2020; Rinke et al., 2021; Parks et al. 2021). Quast was used for standard assembly statistics, including contiguity and assembly length (Gurevich et al., 2013). CheckM was used for providing estimates of genome completeness and contamination based on a set of taxonomic clade-specific marker genes (Parks et al., 2015).

Results

Robust capture of microbial diversity and estimated relative abundance

Even and Staggered DNA mixes—two popular mock microbial mixes used as standards—were processed with the TELL-Seq WGS Library Prep kit. Genome assembly was conducted with Tell-Link, and microbial diversity was captured with TELL-TaxContigs. In **Table 1**, we show a summary of Tell-TaxContigs statistics on these two mock samples.

Microbial Mix	**Even**	**Staggered**	**Staggered-deep**
Total sequencing reads	59M	98M	700M
True positives (out of 20)	20	15	18
False positives	0	0	2
Avg. RA error	0.83%	0.72%	0.91%
Max. RA error	1.91%	3.55%	3.79%

Table 1. Summary statistics for Tell-TaxContigs. RA errors are defined as relative abundance percentage deviations from ground truth.

In **Table 2**, we provide a detailed breakdown of TELL-TaxContigs performance. In the Even dataset, the 20 bacterial strains were correctly identified, and no false positives were detected. In the Staggered dataset, Tell-TaxContigs identified fifteen out of twenty species correctly, missing the five very-low-abundant species (i.e., missing those with 0.02% relative abundance). In the Staggered-deep dataset (sequenced ~7-fold deeper compared to the first Staggered dataset), Tell-TaxContigs identified three of the five very-low-abundant species (i.e., 0.02% relative abundance) in addition to the fifteen species with abundance equal or larger than 0.18% (**Table 2**). In this case, two species were falsely identified as potential organisms. These two species, however, had a combined estimated abundance of less than 0.1%. Overall, the abundance calls were accurate with an average relative abundance error of less than 1% for the true species.

Bacterial strains	**True Abundance**		**Tell-TaxContigs Estimated Abundance**			**Estimated Read Coverage**		
Sample	**Even**	**Staggered**	**Even**	**Staggered**	**Staggered-deep**	**Even**	**Staggered**	**Staggered-deep**
Staphylococcus epidermidis	5%	18.00%	5.13%	17.38%	15.63%	340.96x	2,038.82x	14,563.00x
Streptococcus mutans	5%	18.00%	6.23%	14.45%	14.68%	430.77x	2,575.86x	18,399.01x
Porphyromonas gingivalis	5%	18.00%	5.44%	16.12%	15.82%	365.79x	2,187.32x	15,623.69x
Escherichia coli	5%	18.00%	4.47%	20.49%	21.79%	171.09x	1,023.05x	7,307.49x
Rhodobacter sphaeroides	5%	18.00%	4.91%	18.32%	16.12%	182.84x	1,093.34x	7,809.58x
Staphylococcus aureus	5%	1.80%	5.62%	1.58%	1.49%	293.54x	175.53x	1,253.77x
Streptococcus agalactiae	5%	1.80%	5.68%	1.57%	2.14%	376.50x	225.14x	1608.12x
Bacillus cereus	5%	1.80%	4.33%	2.19%	2.18%	158.56x	94.81x	677.24x
Clostridium beijerinckii	5%	1.80%	5.67%	1.58%	1.33%	144.69x	86.52x	618.01x
Pseudomonas aeruginosa	5%	1.80%	5.49%	1.66%	0.91%	135.65x	81.12x	579.40x
Lactobacillus gasseri	5%	0.18%	4.15%	0.24%	0.11%	454.72x	27.19x	194.22x
Helicobacter pylori	5%	0.18%	4.08%	0.25%	0.10%	512.49x	30.64x	218.89x
Acinetobacter baumannii	5%	0.18%	4.46%	0.20%	0.15%	214.11x	12.80x	91.45x
Neisseria meningitidis	5%	0.18%	3.09%	0.67%	0.10%	380.71x	22.77x	162.61x
Cutibacterium acnes	5%	0.18%	3.67%	0.28%	0.10%	346.13x	20.70x	147.84x
Enterococcus faecalis	5%	0.02%	6.08%	NA	0.02%	282.60x	1.88x	13.41x
Bacteroides vulgatus	5%	0.02%	6.79%	NA	0.05%	166.83x	1.11x	7.92x
Deinococcus radiodurans	5%	0.02%	4.55%	NA	0.04%	267.61x	1.78x	12.70x
Actinomyces odontolyticus	5%	0.02%	3.68%	NA	NA	271.64x	1.80x	12.89x
Bifidobacterium adolescentis	5%	0.02%	5.87%	NA	NA	412.22x	2.74x	19.56x

Table 2. Performance of Tell-TaxContigs on the ATCC Mock 20 Even and 20 Staggered samples (including a deep-sequencing sample)

Generation of highly contiguous metagenome-assembled genomes

In **Table 3**, we show results from Tell-Link/Metabat2 MAGs analyses (the genomes were classified with GTDB-tk 1.5.0. and assembly statistics were calculated with Quast and CheckM). Tell-Link was able to attain quality drafts and excellent contiguity scores with few unclassified MAGs. Metrics for describing 'quality' draft genome assemblies has been described by Bowers et al. (2017). MAGs, especially, often suffer in quality, and so the authors described MIMAG (minimum information of MAGs) as a set of guidelines to evaluate an assembly's quality, including genome contiguity and sequence fidelity, with the latter evaluated mostly by the detection of common conserved genes. Quality draft in Table 3 is denoted as medium quality draft

Microbial Mix	Even	Staggered	Staggered-deep
Sequencing depth	59M	98M	700M
True positives (out of 20)	16	7	8
Unclassified	1	0	3
Quality draft	15	5	7
Avg. contiguity	1.52	2.73	1.64

Table 3. Summary statistics for MAGs analyses from TELL-Link/Metabat2 metagenomic assembly and contig binning. 'Unclassified' denotes binned genome assemblies that could not be classified with GTDB-tk, using GTDB's sets of core marker genes. See text for definitions on quality draft. Contiguity is defined as assembly length divided by N50 score, with the ideal case being a contiguity score of 1.0.

genome assemblies as defined in Bowers et al. (2017). The genome assemblies are not high quality draft genome assemblies because rRNA genes cannot be detected; this has been demonstrated to be an issue with assembling and/or binning highly conserved genomic regions with Tell-Link assembler, Metabat2, or a combination of both. However, the genomes otherwise meet all other requirements for high quality draft genome assemblies: all listed have >90% completion and <5% contamination, metrics based on taxonomic clade-specific marker gene detection. Therefore, even without identification of rRNA genes, the marker gene identifications by GTDB-tk and CheckM indicate high sequence fidelity, complementing the high sequence contiguity. Average contiguity scores (defined as assembly length divided by N50 score) greater than 1 represented reductions in genome contiguity (i.e. more fragmented). However, as most N50 scores were on a scale of millions of base pairs, MAGs showed high contiguity. For comparison, a typical Illumina-based draft genome assembly of length 5 Mbp and N50 of 100 kbp (and passing all minimum quality checks) would have a contiguity score of 50. Individual genome assembly statistics can be found in **Table 4**.

Table 4 also shows that low abundance organisms are difficult to assemble, and there may be diminishing returns as sequencing depth increases. In the Even mix, the two *Streptococcus* and *Staphylococcus* species were missed, suggesting that it is difficult to

Sample (color-coded)	Even / Staggered / Staggered-deep																				
Properties	Contigs			Length (Mbp)			N50 (Mbp)			Length/N50			Completeness			Contamination			Quality draft		
Staphylococcus epidermidis	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Streptococcus mutans	NA	6	NA	NA	1.87	NA	NA	1.33	NA	NA	1.40	NA	NA	92.77	NA	NA	0.00	NA	NA	Yes	NA
Porphyromonas gingivalis	10	15	10	2.12	2.07	2.13	1.23	2.02	1.17	1.72	1.02	1.82	98.82	98.79	99.29	0.00	0.00	0.00	Yes	Yes	Yes
Escherichia coli	3	11	4	4.51	4.41	4.52	4.50	4.38	4.50	1.00	1.01	1.00	99.87	98.32	99.97	0.04	0.04	0.04	Yes	Yes	Yes
Rhodobacter sphaeroides	23	11	5	4.37	4.39	4.41	3.03	3.11	3.10	1.44	1.41	1.42	94.11	97.36	99.24	0.15	0.15	0.15	Yes	Yes	Yes
Staphylococcus aureus	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Na	NA	NA	NA	NA
Streptococcus agalactiae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bacillus cereus	7	173	NA	5.33	1.19	NA	2.72	2.07	NA	1.96	5.75	NA	97.61	95.57	NA	0.33	193.23	NA	Yes	No	NA
Clostridium beijerinckii	42	110	NA	5.92	5.68	NA	2.64	1.37	NA	2.24	4.16	NA	99.19	96.77	NA	2.42	2.42	NA	Yes	Yes	NA
Pseudomonas aeruginosa	6	262	3	6.29	5.53	6.29	6.27	1.26	6.28	1.00	4.39	1.00	98.40	86.96	99.68	0.45	0.92	0.45	Yes	No	Yes
Lactobacillus gasseri	2	NA	NA	1.80	NA	NA	1.79	NA	NA	1.00	NA	NA	98.36	NA	NA	0.00	NA	NA	Yes	NA	NA
Helicobacter pylori	1	NA	14	1.59	NA	1.59	1.59	NA	1.55	1.00	NA	1.03	99.09	NA	98.20	0.00	NA	0.00	Yes	NA	Yes
Acinetobacter baumannii	15	NA	5	7.93	NA	3.83	2.05	NA	3.81	3.86	NA	1.00	100.00	NA	99.45	175.51	NA	0.27	No	NA	Yes
Neisseria meningitidis	8	NA	48	2.02	NA	1.84	1.99	NA	1.69	1.01	NA	1.09	98.23	NA	92.14	0.19	NA	0.19	Yes	NA	Yes
Cutibacterium acnes	1	NA	75	2.48	NA	2.25	2.48	NA	2.01	1.00	NA	1.12	98.90	NA	79.07	0.00	NA	0.00	Yes	NA	No
Enterococcus faecalis	3	NA	NA	2.68	NA	NA	2.67	NA	NA	1.00	NA	NA	98.89	NA	NA	0.00	NA	NA	Yes	NA	NA
Bacteroides vulgatus	7	NA	NA	4.82	NA	NA	3.12	NA	NA	1.55	NA	NA	98.45	NA	NA	0.19	NA	NA	Yes	NA	NA
Deinococcus radiodurans	5	NA	NA	3.03	NA	NA	2.61	NA	NA	1.16	NA	NA	99.04	NA	NA	0.21	NA	NA	Yes	NA	NA
Actinomyces odontolyticus	2	NA	NA	2.36	NA	NA	2.36	NA	NA	1.00	NA	NA	97.17	NA	NA	0.47	NA	NA	Yes	NA	NA
Bifidobacterium adolescentis	1	NA	NA	1.99	NA	NA	1.99	NA	NA	1.00	NA	NA	97.15	NA	NA	0.00	NA	NA	Yes	NA	NA

Table 4. Detailed genome assembly statistics for Mock 20 Even MAGs from Tell-Link/Metabat2.

disambiguate similar species from the same genera for either Tell-Link, Metabat2, or both. However, organisms that are assembled and correctly classified typically have far superior assemblies than purely Illumina-based MAGs and in many cases are superior to Illumina-based isolate assemblies as well.

Summary

We have shown that TELL-Seq technology can provide highly accurate standard metagenomic analyses in the form of taxonomic classifications and relative abundance profiles. We have also shown that assembly of metagenomic reads using TELL-Seq technology results in highly contiguous and highly accurate binned genome assemblies of the most abundant organisms in a metagenomic sample, depending on sequencing depth. We believe that the combination of Tell-Link and Tell-TaxContigs will aid future discovery efforts in microbiome research, and the higher completion rates of MAGs will enable more comprehensive annotation of uncultured microbes.

Learn more (hyperlinks)

TELL-Seq technology

TELL-Seq technology video

TELL-Seq guides

TELL-seq software guides: Tell-Read, Tell-Sort, and IGV visualization of TELL-Seq data

GTDB-Tk software

Quast software

CheckM software

Additional TELL-seq applications: microbial (Illumina)

Link to raw data (hyperlinks)

All sequencing data is available at NCBI under the BioProject PRJNA831850: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA831850

Ordering information (hyperlinks)

Reagent boxes	
TELL-Seq WGS Library Reagent Box 1	#100001
TELL-Seq WGS Library Reagent Box 2	#100002
Primer boxes	
TELL-Seq Library Index Primer Kit	#100003*
TELL-Seq Illumina Sequencing Primer Kit	#100004

* 100009 and 100010 contain additional indexes

TELL-Seq safety data sheets

Acknowledgements

Hasan H. Otu and Sree Krishna Chanumolu, OTUFY, LLC (Lincoln, NE)

References (hyperlinks)

Bowers et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol. 35:725–731 (2017)*

Breitwieser et al. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform. 20:1125-1126 (2019)*

Chaumeil et al. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics, 36:1925–1927 (2019)*

Chen 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.* 30:898-909 (2020)

Dongwan et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ, 7:e7359 (2019)*

Ghurye et al. Metagenomic assembly: overview, challenges and applications. *Yale J. Biol. Med. 89:353-362 (2016)*

Gurevich et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29:1072-1075 (2013)*

Parks et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res., 25:1043-1055 (2015)*

Parks et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotech., 36:996-1004 (2018)*

Parks et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotech., 38:1079-1086 (2020)*

Parks et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucl. Acids Res. gkab776 (2021)*

Riesenfeld et al., Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet., 38:525-52 (2004)*

Rinke et al. A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat. Microbiol. 6:946–959 (2021)*

For research use only. Not for use in diagnostic procedures