Tell-TaxContigs: Microbial Metagenome Assembly, Taxonomy, and Abundance Estimation Using UST TELL-seq Long-range Information

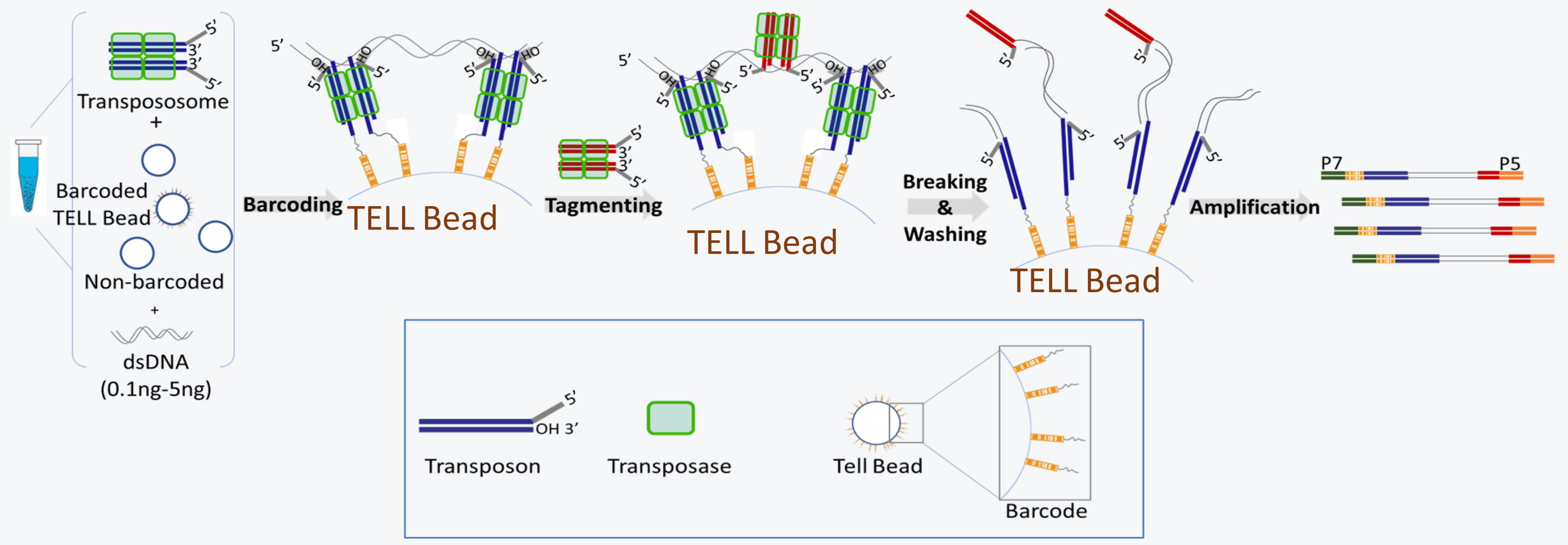
Colin Heberling[1], Long Pham[2], Sree Krishna Chanumolu[4], Hasan H. Otu[3], Yu Xia[2], Peter Chang[2], Andrew Anfora[2], Ivan Garcia-Bassets[2], Tom Chen[2], Yong Wang[1]

[1] Universal Sequencing Technology, Canton, MA (USA); [2] Universal Sequencing Technology, Carlsbad, CA (USA); [3] OTUFY, Lincoln, NE (USA); [4] GeneFront, San Jose, CA (USA)

Abstract. Deconvolving diversity of a metagenome is critical for understanding the role of a given microbial community in human health and disease, small molecule biosynthesis, and other complex ecosystems where more reductive analysis proves to be elusive. Metagenomic assembly is one common method for characterizing a metagenome, especially for identifying novel gene content or novel organisms. However, analyzing sequencing data from microbial mixtures with a high dynamic range of relative abundance of strains, close relatedness, and repetitive genomic content amongst members can vastly complicate the genome assembly process of individual microorganisms and strains. Furthermore, efficient assembly requires high fidelity sequencing reads to avoid ambiguities. We previously developed a method that captures long-range molecular origin information from kilobase-long genomic fragments by a process of DNA barcoding that we called transposase enzyme-linked long read sequencing (TELL-seq) developed by Universal Sequencing Technologies (UST)[1]. TELL-seq barcoded fragments can be sequenced with instruments that process short reads (i.e., high-fidelity sequencing). Here, we show that integration of TELL-seq data with a computational pipeline that combines *de novo* genome assembly (Tell-Link) with taxonomic classification and abundance estimation (Tell-TaxContigs) provides highly accurate metagenomic analyses. We show how the application of Tell-Link and Tell-TaxContigs on sequencing data generated from commercially available microbial mixture standards results in genome assemblies with contiguities larger than 1Mbp (N50), and highly accurate classification and relative abundance estimation for organisms at 0.18% or greater relative abundance, respectively. Therefore, Tell-Link, in combination with genome binning software (e.g. metabat2[2]) provides highly contiguous and high-fidelity genome assemblies of abundant organisms in a metagenomic sample. Tell-TaxContigs classifies contigs and unassembled reads with BLASTn and using a deep learning approach resolves ambiguities, rules out false positive classifications, and accurately estimates relative abundances of classified species. This approach has an average margin of error of lower than 1% in enumerating relative abundance for the microbial mixture standards tested.

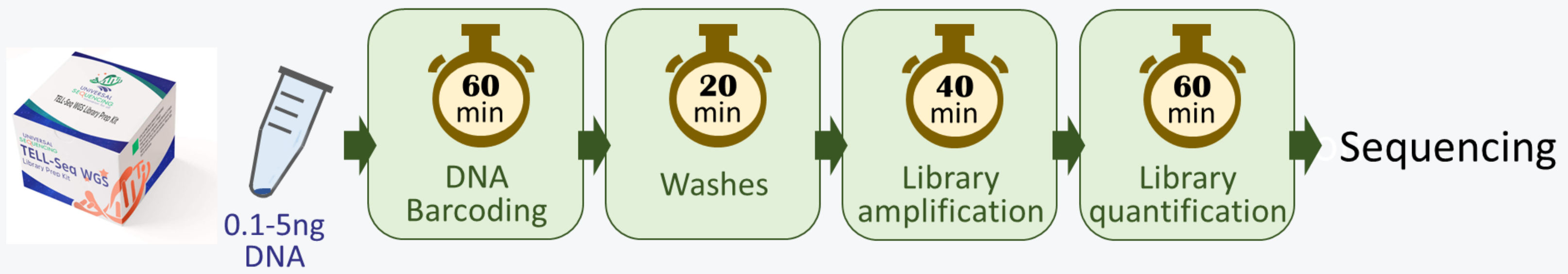
How Does UST TELL-seq Work?

Transposase Enzyme Linked Long-Read Sequencing

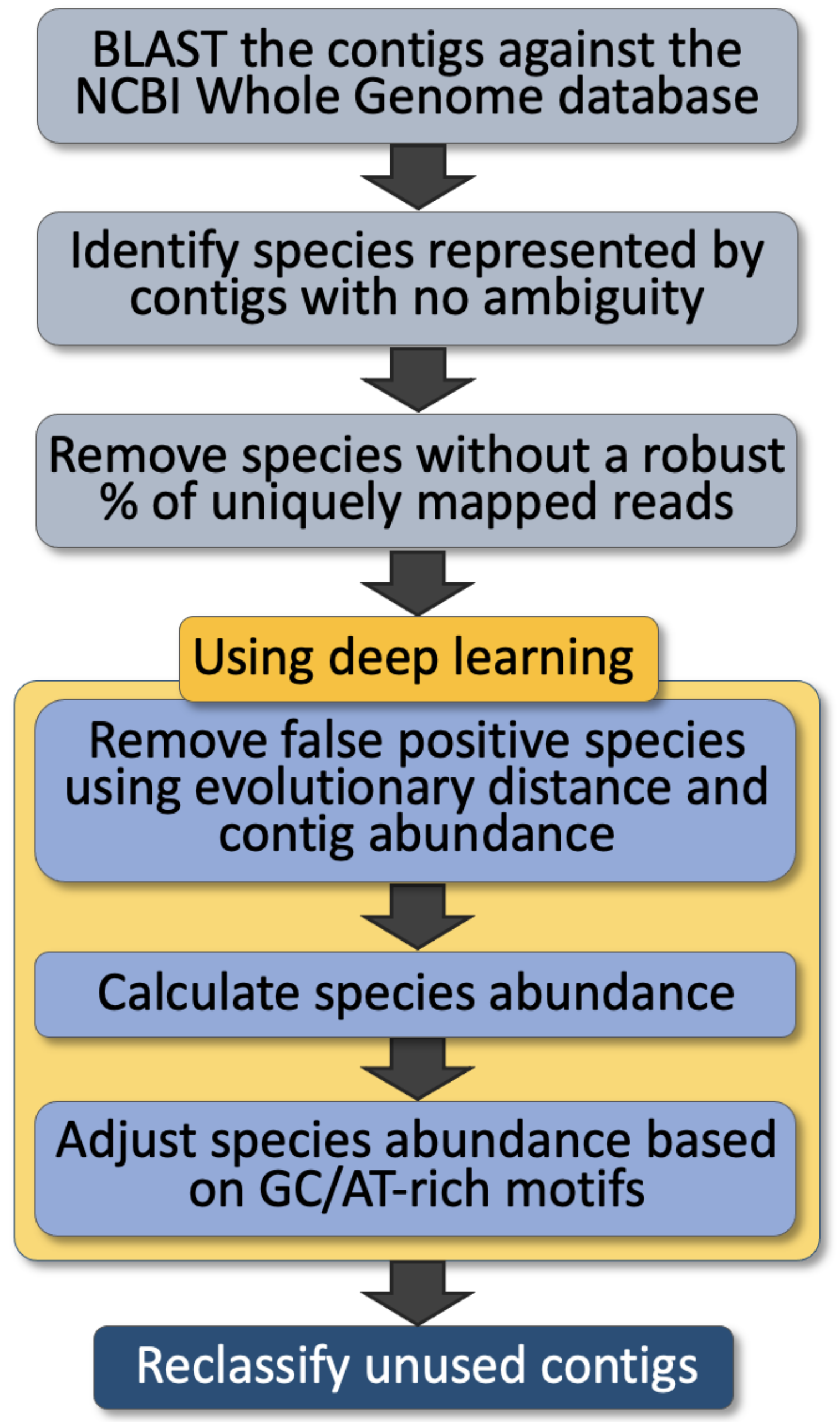


TELL-seq Library Workflow is Simple & Scalable

Total Time: 3 hours



Tell-TaxContigs Overview



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

1. Metagenomic Assembly & Genome Binning

Metagenomic assembly summary

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

Notes on metagenomic assembly:

Contiguity is defined as assembly length divided by N50 score. Quality draft denotes medium quality draft assembly according to Bowers et al. (2017)[3] minimum information of MAGs.

Detailed metagenomic assembly and genome binning results

Sample (color-coded)	Even / Staggered / Staggered-deep																				
Species Name	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

2. Taxonomic Classification & Abundance Estimation

Detailed classification results

Bacterial strains	True Abundance		Tell-TaxContigs Estimated Abundance		
Species Name	Even	Staggered	Even	Staggered	Staggered-deep
Staphylococcus epidermidis	5%	18.00%	5.13%	17.28%	15.63%
Streptococcus mutans	5%	18.00%	6.23%	14.45%	14.68%
Porphyromonas gingivalis	5%	18.00%	5.44%	16.26%	15.82%
Escherichia coli	5%	18.00%	4.47%	20.67%	21.79%
Rhodobacter sphaeroides	5%	18.00%	4.91%	18.18%	16.12%
Staphylococcus aureus	5%	1.80%	5.62%	1.57%	1.49%
Streptococcus agalactiae	5%	1.80%	5.68%	1.56%	2.14%
Bacillus cereus	5%	1.80%	4.33%	2.17%	2.18%
Clostridium beijerinckii	5%	1.80%	5.67%	1.57%	1.33%
Pseudomonas aeruginosa	5%	1.80%	5.49%	1.61%	0.91%
Lactobacillus gasseri	5%	0.18%	4.15%	0.23%	0.11%
Helicobacter pylori	5%	0.18%	4.08%	0.24%	0.10%
Acinetobacter baumannii	5%	0.18%	4.46%	0.21%	0.15%
Neisseria meningitidis	5%	0.18%	3.09%	0.65%	0.10%
Cutibacterium acnes	5%	0.18%	3.67%	0.31%	0.10%
Enterococcus faecalis	5%	0.02%	6.08%	NA	0.02%
Bacteroides vulgatus	5%	0.02%	6.79%	NA	0.05%
Deinococcus radiodurans	5%	0.02%	4.55%	NA	0.04%
Actinomyces odontolyticus	5%	0.02%	3.68%	NA	NA
Bifidobacterium adolescentis	5%	0.02%	5.87%	NA	NA

Classification summary

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%

Conclusions. UST TELL-seq allows for long-range information to be integrated into short-read sequencing projects. This offers several advantages to standard whole metagenome shotgun sequencing projects. We have shown that *de novo* assembly of TELL-seq barcoded Illumina reads results in highly contiguous, high fidelity binned genome assemblies with N50 scores in the Mbp range. Achieving this level of contiguity and fidelity normally requires the hybridization of high-fidelity short reads with error-prone long-read technologies, ideally from the same DNA extraction[4]. Using UST TELL-seq is a simpler process that can achieve similar results, with only a single sequencing source. As a result, the highly contiguous and accurate contigs generated with TELL-seq can be used for various metagenomics downstream analysis steps, such as taxonomic classification and relative abundance estimation, with relatively low margins of error, as we have demonstrated with Tell-TaxContigs. Tell-TaxContigs will be released publicly later this year.

References

1. Chen, Z et al. 2020. *Genome Research* **30**: 898-909.
2. Dongwan et al. 2019. PeerJ, 7:e7359
3. Bowers et al. 2017. Nat. Biotechnol. 35:725–731
4. Wick et al. 2017. PLoS Comput Biol 13(6): e1005595.

UNIVERSAL SEQUENCING
innovation for all