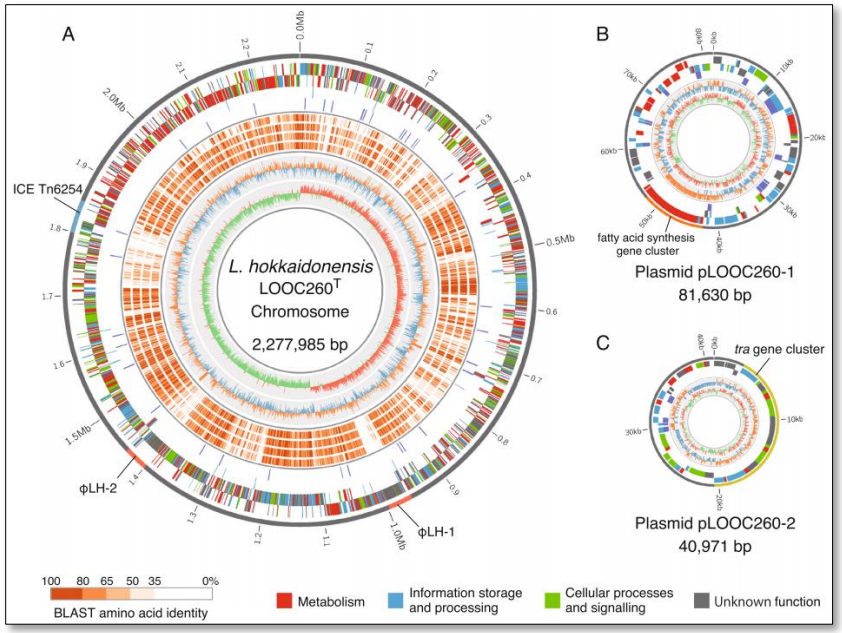
The microbial world: from single bacterial genome to microbiota population characterization

Déborah Moine, Field Application Scientist II Europe

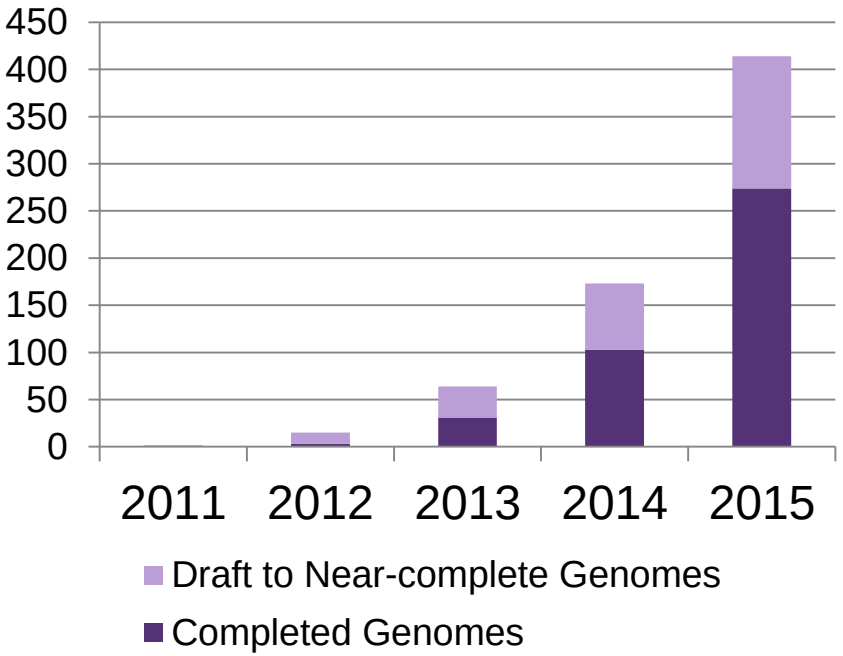
AGENDA OF THE SESSION

- Multiplexing bacterial genomes
- 16S Metagenetic
- Shotgun Metagenomic

PACBIO IS THE NEW GOLD STANDARD FOR BACTERIAL GENOMICS



Growth in published microbial genomes completed with SMRT Sequencing



Tanizawa et al. (2015) Complete genome sequence and analysis of Lactobacillus hokkaidonensis LOOC260T, a psychrotrophic lactic acid bacterium isolated from silage. BMC Genomics 16: 240.

STREAMLINED SOLUTION FOR MULTIPLEXED *DE NOVO* MICROBIAL GENOME ASSEMBLIES

Achieve high base accuracy and closed microbial genome assemblies with the leader in long-read sequencing

- Two verified barcoded adapter kits for up to 16 samples
- Up to 30 Mb of multiplexed microbial genomes
- Library preparation in ~8.5 hours
- Microbial Multiplexing Calculator to assist with equimolar pooling
- Sequencing in 10 hour movie runs
- Automatic demultiplexing
- Optimized assembly parameters for microbial assemblies
- Detection of R-M system motifs

MULTIPLEX MICROBIAL GENOMES FOR COST EFFICIENCY

Two Examples of 30 Mb Microbial Genome Pooling Strategies

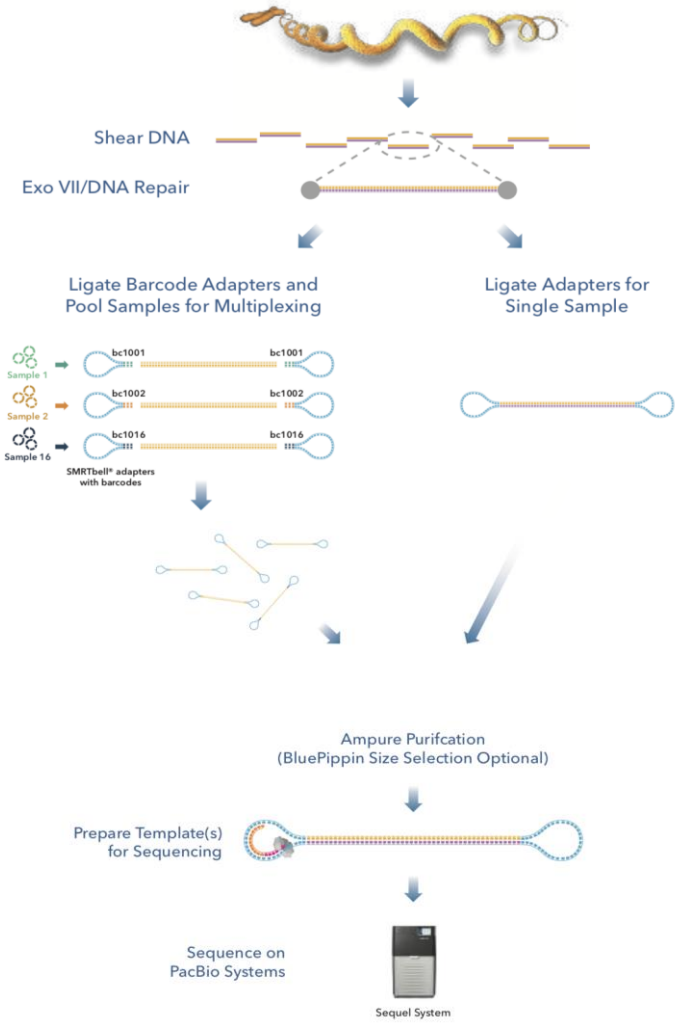
Samples	Sample ID	Genome Size (bp)
1	E. coli 1	4,642,523
2	E. coli 2	4,642,523
3	B. subtilis 1	4,045,593
4	B. subtilis 2	4,045,593
5	S. sonnei 1	4,813,450
6	S. sonnei 2	4,813,450
7	L. monocytogenes 1	3,032,269
8	L. monocytogenes 2	3,032,269

Total **33,067,670**

Samples	Sample ID	Genome Size (bp)
1	E. coli 1	4,642,523
2	E. coli 2	4,642,523
3	B. subtilis 1	4,045,593
4	B. subtilis 2	4,045,593
5	R. palustris	5,459,213
	R. palustris - plasmid 1	8,427
6	S. aureus	2,872,915
	S. aureus - plasmid 1	27,041
	S. aureus - plasmid 2	3,125
7	L. monocytogenes	3,032,269
8	S. sonnei 1	4,813,450

Total **33,592,672**

AFFORDABLY CHARACTERIZE COMPLETE MICROBIAL GENOMES ON THE SEQUEL PLATFORM



- Multiplex up to 30 Mb of microbial genomes / 16 samples on one SMRT Cell 1M
- Adjust planned multiplexing depth to balance cost constraints with your requirements for genome completeness
- Use our Microbial Multiplexing Calculator to simplify equimolar pooling
- Assemble most bacterial chromosomes into 5 contigs or fewer

OTHER EXPERIMENTAL DESIGN CONSIDERATIONS

Genome complexity also impacts ability to close microbial genome: (Koren, et, al.)

- **Class I** genomes have few repeats except for the rDNA operon sized 5 to 7 kb. These will likely assemble to <5 contigs with multiplexing up to 30 Mb total genomes.

- **Class II** genomes have many repeats, such as insertion sequence elements, but none greater than 7 kb. These will need higher coverage to close, which can be achieved with lower multiplex for additional coverage across the genome per SMRT Cell. Size selection to enrich larger fragments may benefit assembly.

- **Class III** genomes contain large, often phage-related, repeats >7 kb, including tandem repeats and segmental duplications. These will be difficult to close with 10 kb insert libraries. If a closed genome is required, non-multiplexed sequencing with larger insert libraries can be explored. We offer >15 and >30 kb library protocols.

Koren S, et. Al. *Reducing assembly complexity of microbial genomes with single-molecule sequencing*. Genome Biology 2013;14(9):R101.

ACHIEVE HIGHER COST EFFICIENCY AFTER BUILDING INITIAL EXPERIENCE WITH 30 MB POOLED LIBRARIES

Barcode ID	Sample ID	Pre-Assembly Yield (%)	GC Content (%)	Genome Size (bp)	Contigs (#)	Concordance w. Reference (QV)
BC1002	E. coli control 1	77	50.08	4,642,523	1	55.21
BC1015	E. coli control 2	77.9	50.08	4,642,523	1	56.25
BC1004	B. sub control 1	76.4	43.94	4,045,593	1	51.16
BC1016	B. sub control 2	74.9	43.94	4,045,593	1	51.45
BC1009	E. coli 1	77.4	50.08	4,642,523	1	54.63
BC1018	E. coli 2	77.7	50.08	4,642,523	1	54.91
BC1012	S. sonnei 1	76.4	51.03	4,813,450	1	59.83
BC1020	S. sonnei 2	78.1	51.03	4,813,450	1	54.27
BC1014	L. monocytogenes 1	71.8	37.94	3,032,269	1	53.06
BC1022	L. monocytogenes 2	73.2	37.94	3,032,269	1	49.63

- 10-plex run of 42.4 MB total microbial genomes shown
- Combined use of Barcoded Adapter Kit 8A & Barcoded Adapter Kit 8B
- Less than 5 contig assemblies for main chromosomal genomes can be achieved with Advanced HGAP parameters
- QV50 is roughly 99.999% base accuracy

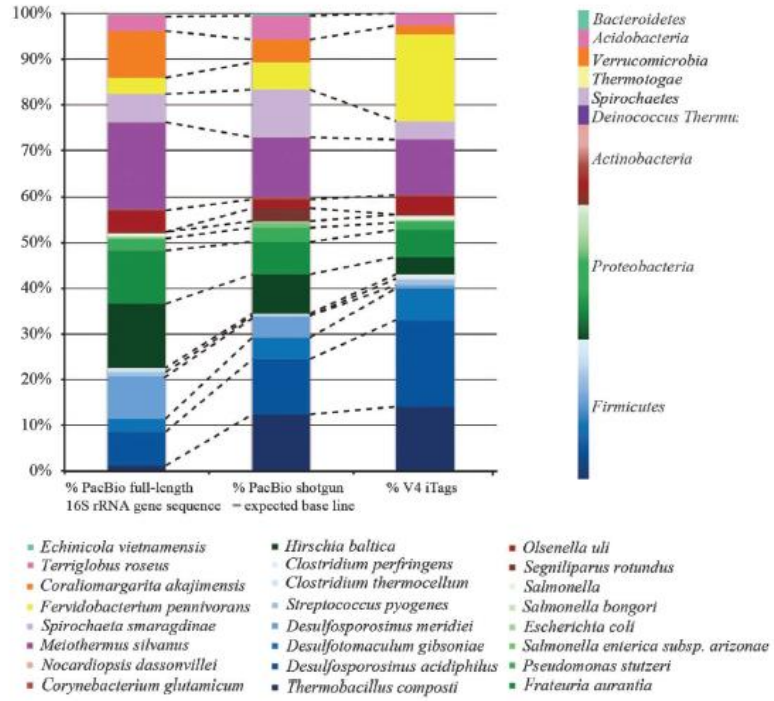
PACBIO®

Full-Length 16S Sequencing

Microbiota identification

CHARACTERIZE COMPLEX POPULATIONS

FULL-LENGTH 16S SEQUENCING: SPECIES-LEVEL PHYLOGENTIC RESOLUTION



Full-length 16S rRNA sequences from a mock microbial community provided a more accurate community profile compared with short-read 16S and shotgun assembly results[1]

1. Singer, E. et al. (2016) High-resolution phylogenetic microbial community profiling. The ISME Journal. doi:10.1038/ismej.2015.249

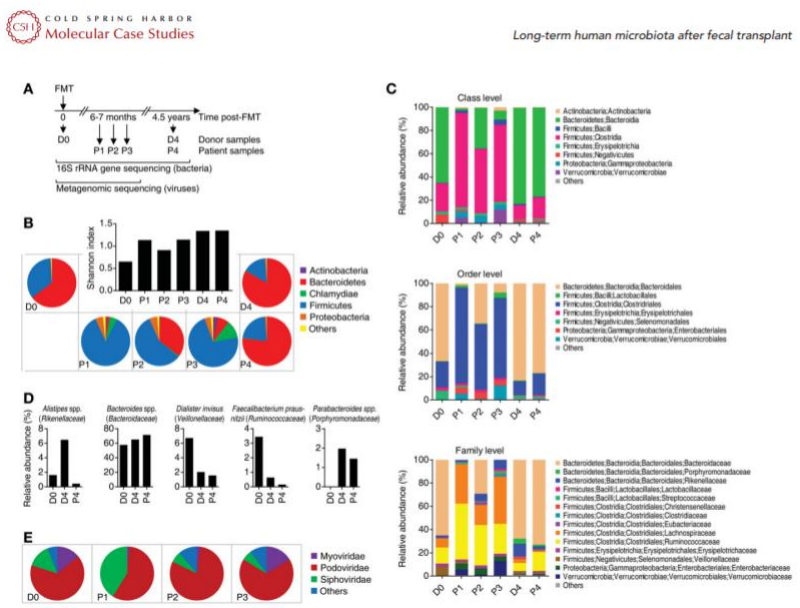
16S VS WGS – 16S PLUSES & MINUSES

Advantage: targeted approach

- Inexpensive
- High multiplex potential
- Robust to sample issues
  - Fragmented, low input samples OK
  - Samples with host contamination OK
- Classification to species level

Disadvantage: targeted approach

- May have limited resolution
  - Dependent on data base representation, completeness and accuracy
- No functional information
- PCR issues
  - Chimeras
  - Potential for biased representation
  - Dependent on having appropriate primer sequence – may exclude some (unknown) organisms



Long-term changes of bacterial and viral compositions in the intestine of a recovered Clostridium difficile patient after fecal microbiota transplantation

Broecker et al. (2016) Cold Spring Harb Mol Case Stud 2: a000448: http://molecularcasestudies.cshlp.org/content/2/1/a000448.full.pdf+html

16S WORKFLOW – LIBRARY PREP

Protocol with optimized PCR conditions

- Includes full-length 16S primer sequences
- Multiplex with barcoded universal primers or barcoded 16S primers
- Input and cycling conditions to limit chimeras

Key factors in limiting chimeras

- Input template amount (keep low)
- PCR cycle number (as low as possible)
  - Different # of cycles for different input amounts
- Extension time (longer may be better)
- High fidelity polymerase to minimize PCR errors

PACBIO

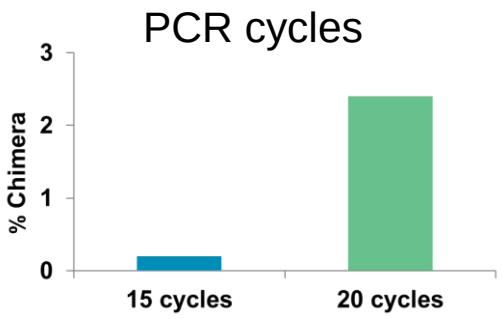
Procedure & Checklist - Full-Length 16S Amplification, SMRTbell® Library Preparation and Sequencing

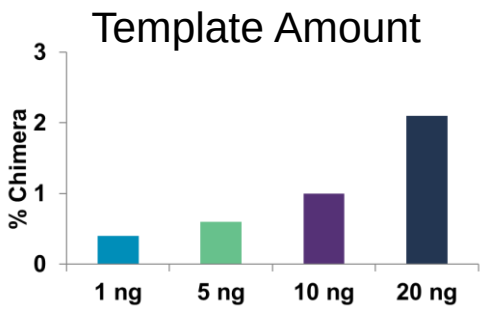
This document contains instructions for amplifying and sequencing a full-length 16S gene from bacterial DNA isolated from metagenomic samples. Tests with mock community samples produced discrete 16S amplicons with adequate yield for library preparation and SMRT sequencing. Data analysis showed good representation of community members in the samples, with low rates of chimerism.

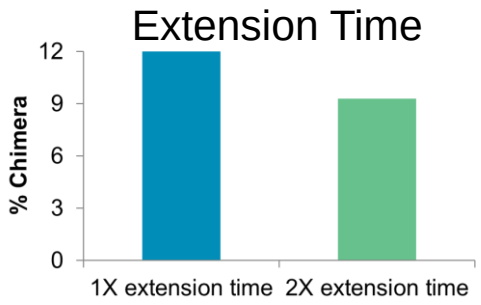
The workflow employs 2 rounds of PCR, the first with universal primer-tailed 16S primers and the second with PacBio Barcoded Universal Primers.

Materials and Kits Needed

Item	Vendor
Amplification	
KAPA HiFi HotStart PCR Kit	KAPA Biosciences
27F Primer tailed with universal sequence*	Any Oligo Vendor
1492R Primer tailed with universal sequence**	Any Oligo Vendor
Barcoded Universal Primers	PacBio
Library Prep	
SMRTbell Template Prep Kit	Pacific Biosciences
AMPure® PB beads	
QC Tools	
Qubit	Invitrogen
BioAnalyzer	Agilent





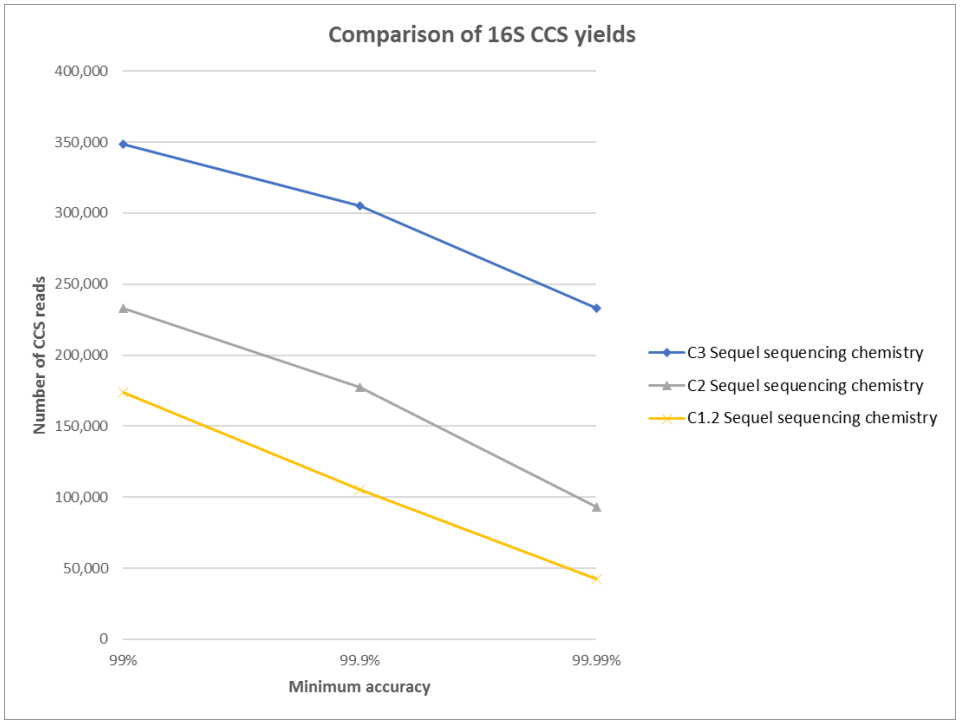


https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Full-Length-16S-Amplification-SMRTbell-Library-Preparation-and-Sequencing.pdf

16S SEQUENCING AND CCS RESULTS WITH NEW CHEMISTRY

P3 - C3 chemistry

Parameter	1 Cell
Total yield	**28 Gbases**
Polymerase RL	**50,562**
Subread RL	2,152
Primary (P1) Reads	573,354
CCS yield (99%)	**348,730**
CCS yield (99.9%)	**305,316**
CCS yield (99.99%)	**232,828**

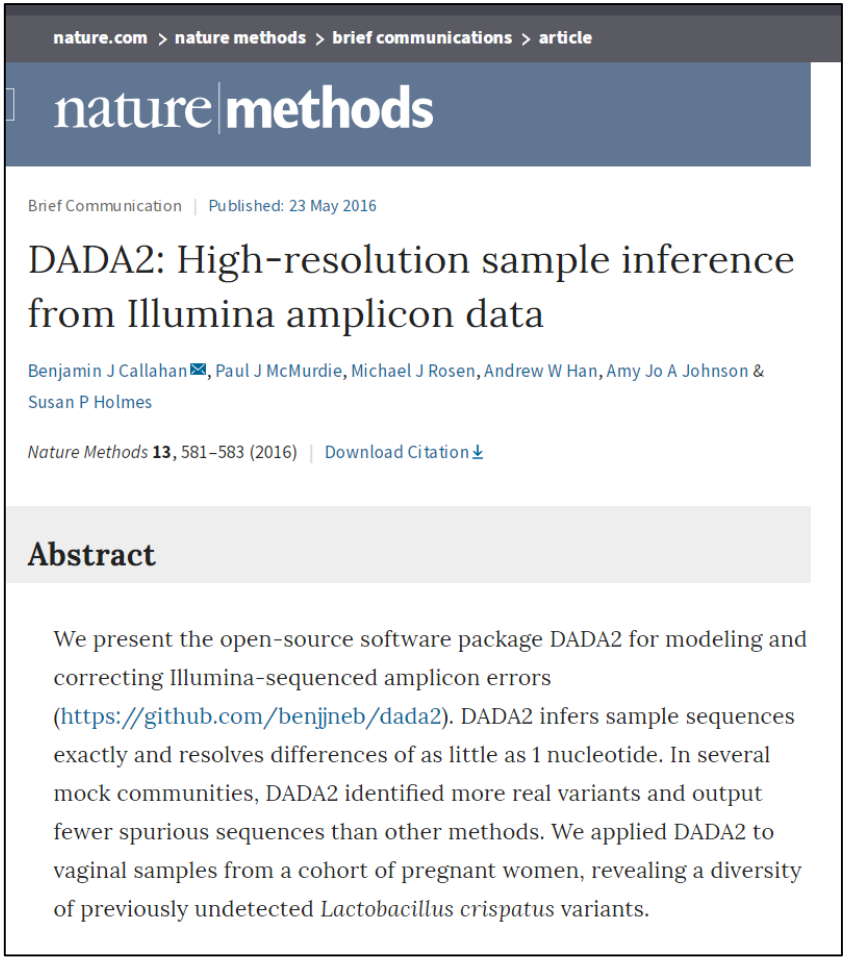


	P2/C2	P3/C3
Data collection	6 hours	15 hours
Pre-extension	0 hours	2 hours
P1 reads	454,534	573,354
Full version	S/P2-C2/5.0	S/P3-C3/5.0

DIVISIVE AMPLICON DENOISING ALGORITHM

Amplicon Sequencing. **Exactly.** *Version 1.8*

- Created by Benjamin Callahan
- Reference-free
- Quality-aware, models errors
- **Resolution down to single-nt differences**
- Fewer false positive sequence variants (no OTUs)
- Steps
  - Quality filtering
  - Dereplication
  - Error modeling
  - **Amplicon Sequence Variant** inference
  - **Chimera removal**
  - **Taxonomic assignment**
- Latest version supports PacBio reads



nature.com > nature methods > brief communications > article

nature methods

Brief Communication | Published: 23 May 2016

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes

Nature Methods **13**, 581–583 (2016) | Download Citation

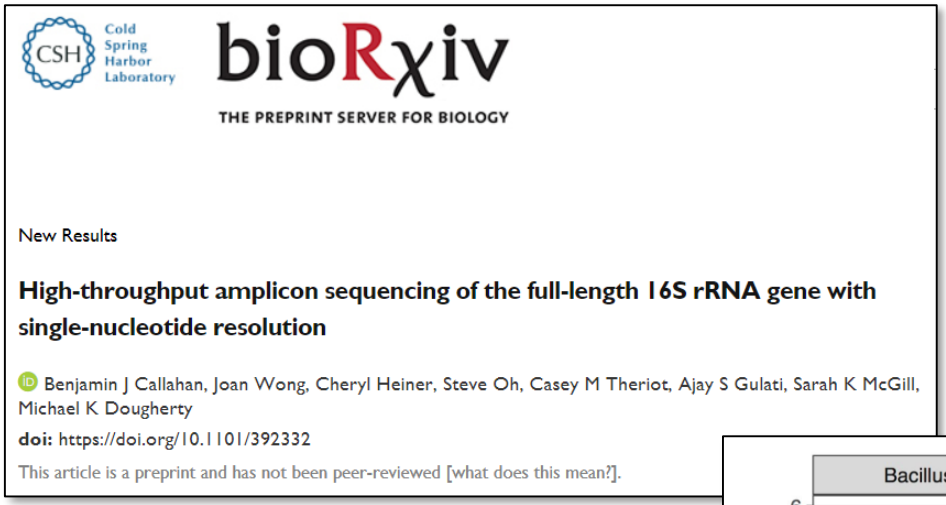
Abstract

We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjjneb/dada2). DADA2 infers sample sequences exactly and resolves differences of as little as 1 nucleotide. In several mock communities, DADA2 identified more real variants and output fewer spurious sequences than other methods. We applied DADA2 to vaginal samples from a cohort of pregnant women, revealing a diversity of previously undetected *Lactobacillus crispatus* variants.

https://benjjneb.github.io/dada2/index.html

Nature Methods volume 13, pages 581–583 (2016)

FULL-LENGTH 16S SEQUENCING



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution

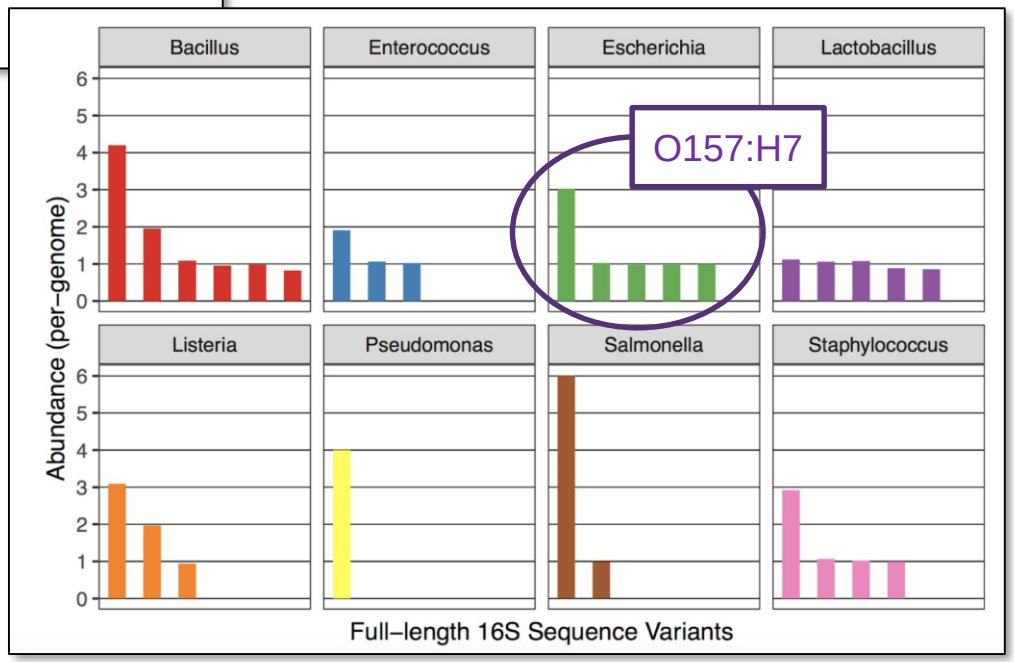
Benjamin J Callahan, Joan Wong, Cheryl Heiner, Steve Oh, Casey M Theriot, Ajay S Gulati, Sarah K McGill, Michael K Dougherty

doi: https://doi.org/10.1101/392332

This article is a preprint and has not been peer-reviewed [what does this mean?].

"The high resolution and accuracy we are reporting derives in part from the exceptional and not-entirely-appreciated accuracy of PacBio CCS sequencing."

- Nearly every bacteria has multiple copies of the 16S housekeeping gene, but in many cases they are not perfect duplicates
- PacBio CCS produced multiple distinct 16S sequences per bacterial genome, and they appear in integer ratios that reflected their copy number in each genome



Callahan, BJ et. al. (2018) High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. bioRxiv doi: http://dx.doi.org/10.1101/39233.

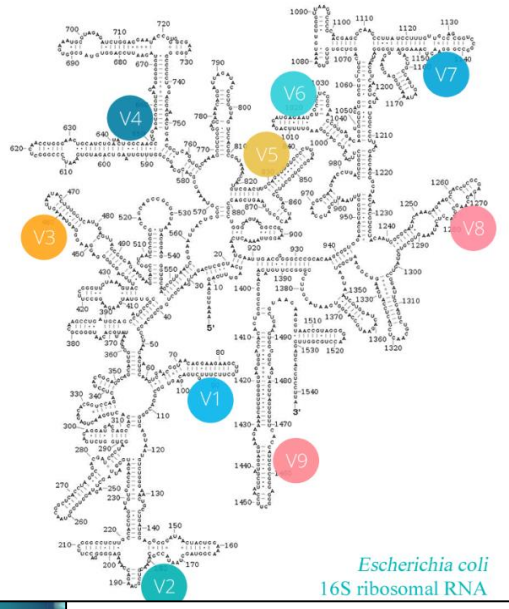
3RD PARTY PROVIDERS FOR PACBIO 16S ANALYSIS

ONE CODEX

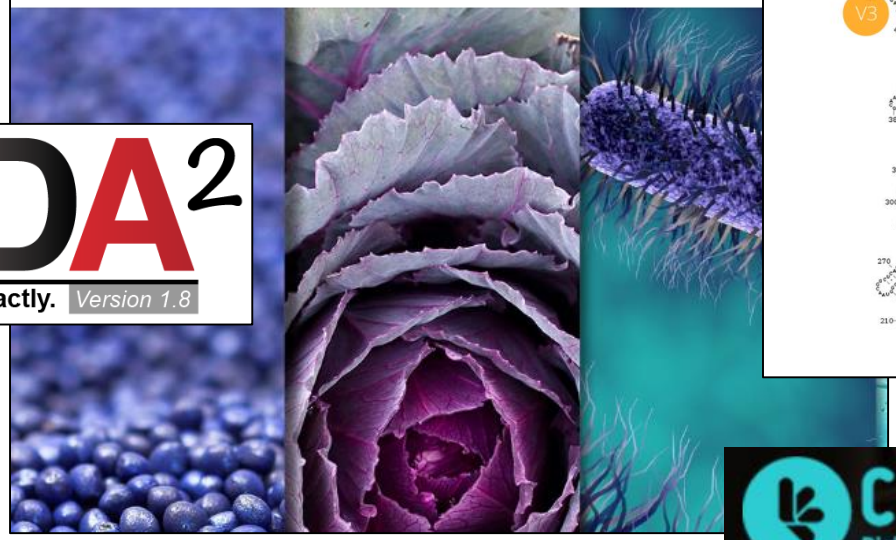
Era7 bioinformatics

100% of the hypervariable regions can be analyzed using PacBio reads

Using long read sequencing technologies we can obtain the sequence of the full gene and, then, have a significantly higher specificity and resolution capacity to do the taxonomic assignments based on the differences in the 16S gene sequence.



computomics®
molecular data analysis



DA2
Amplicon Sequencing. Exactly. Version 1.8

CHUNLAB
Bioinformatics for Microbiology

PACBIO®

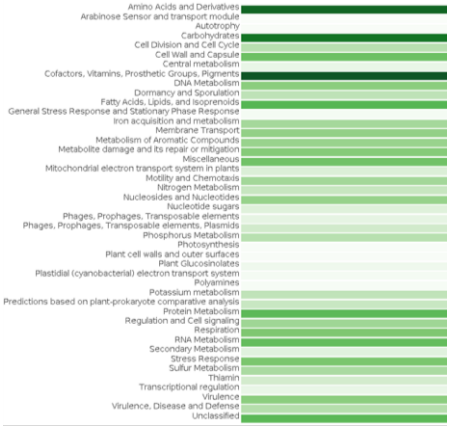
Shotgun Metagenomic (WGS)

Gene Profiling and Assembly

16S VS WGS – WGS PLUSES & MINUSES

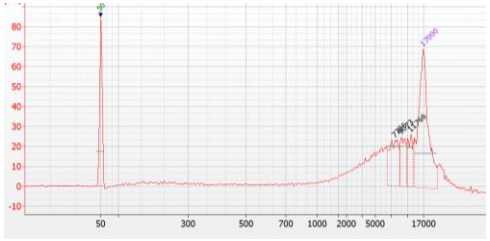
Advantage: whole genome

- Provides functional information, in addition to classification
- Assembly of large contigs
  - Complete genomes possible for simple communities or predominant members
- Epigenetic information with high coverage
- Identification of low abundance community members
- Doesn't require (full-length, correct, full compliment) 16S sequences in data base
- Unbiased representation of what's in the sample (bacteria + other organisms)



Disadvantage: whole genome

- Input sample requirements:
  - Requires ~100 ng reasonable quality DNA
  - Sensitive to host contamination
- Requires more reads for characterization
- Assembly may requires higher coverage, depending on community
- Complex analysis



WGS LIBRARY SIZE DEPENDS ON SAMPLE AND PROJECT

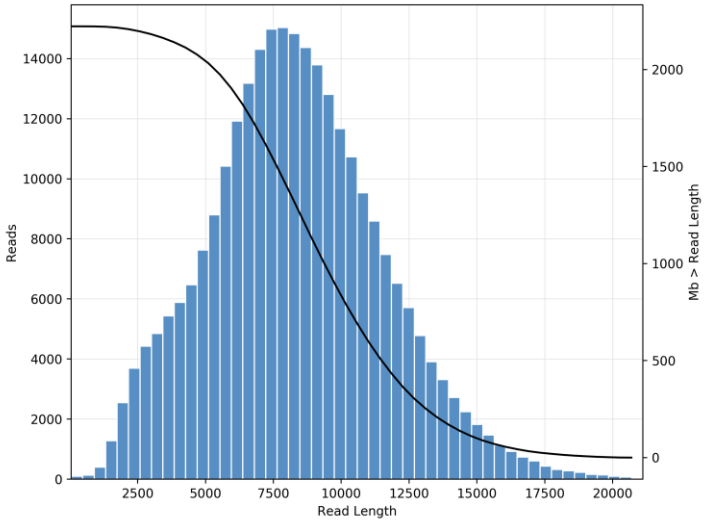
Project Goal	Library Size	Input Sample	Community Complexity
Classification, functional profiling	3 kb	100 ng, somewhat fragmented ok	All - complex OK
Find and co-localize genes, classification	6 kb – 10 kb	0.5 - 1 µg, fairly high molecular weight	All - complex OK
Assembly	10 kb	1 µg, high molecular weight	Simple, or a few predominant members

LONGER CCS LIBRARIES WITH NEW CHEMISTRY

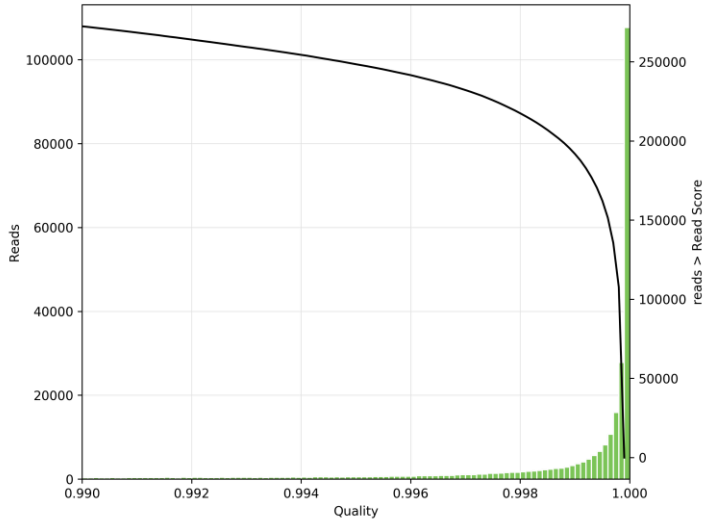
10 kb sheared library - Sequencing and CCS results, 20 hour collection

# Bases	Polymerase RL mean	# P1 reads	# CCS reads >99% predicted accuracy	Mean CCS read length	Mean # passes	Mean CCS predicted accuracy
38 Gb	77,846	496,590	**273,261**	**8,345**	16	**99.9%**
37 Gb	76,318	501,462	**271,184**	**7,705**	18	**99.9%**

CCS Read Length

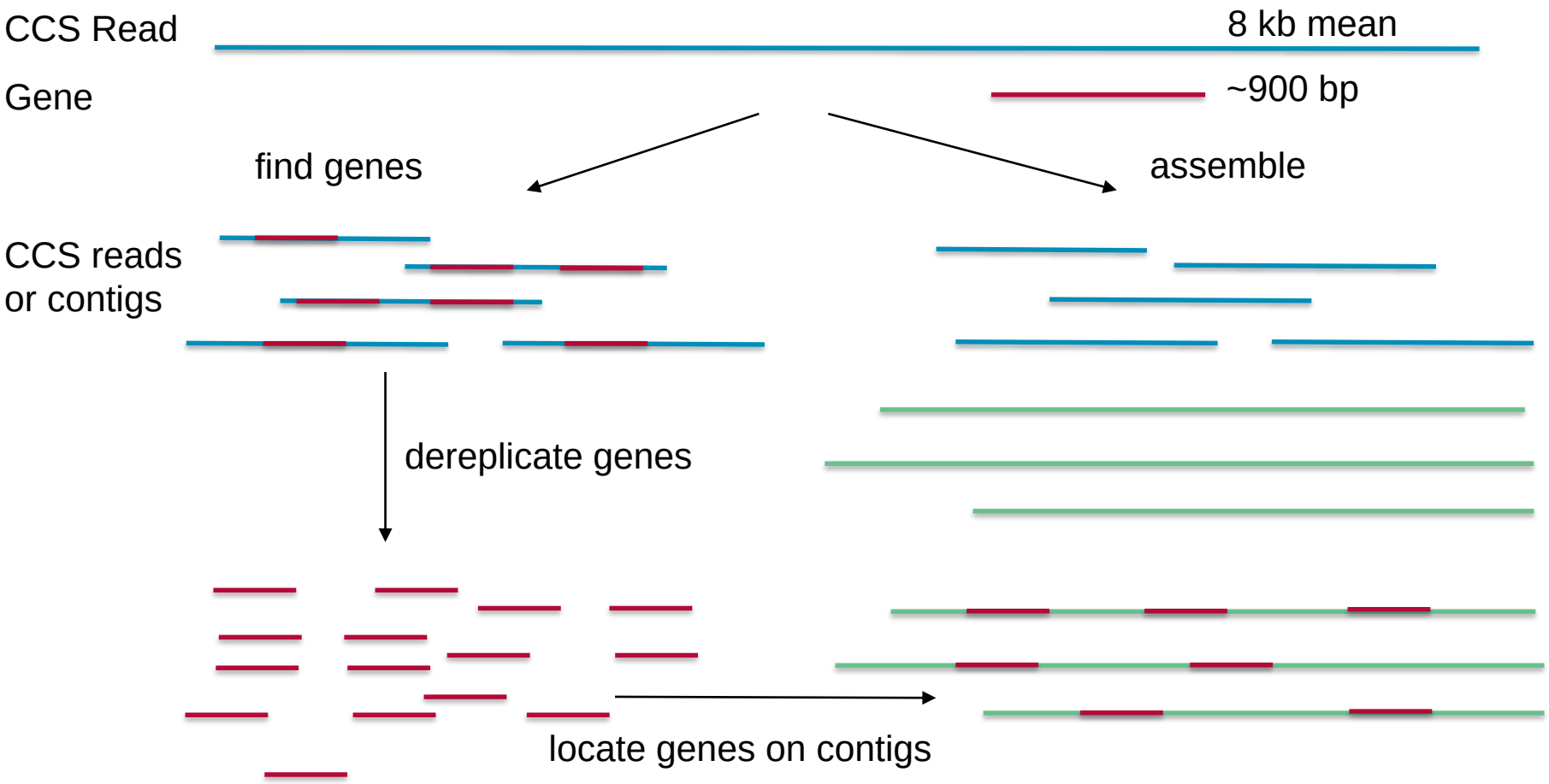


CCS Read Score



CCS results cell 1, 99% predicted accuracy

ANALYSIS PIPELINE – GENE FINDING AND LOCALIZATION



FUNCTIONAL CLASSIFICATION OF GENES FOUND

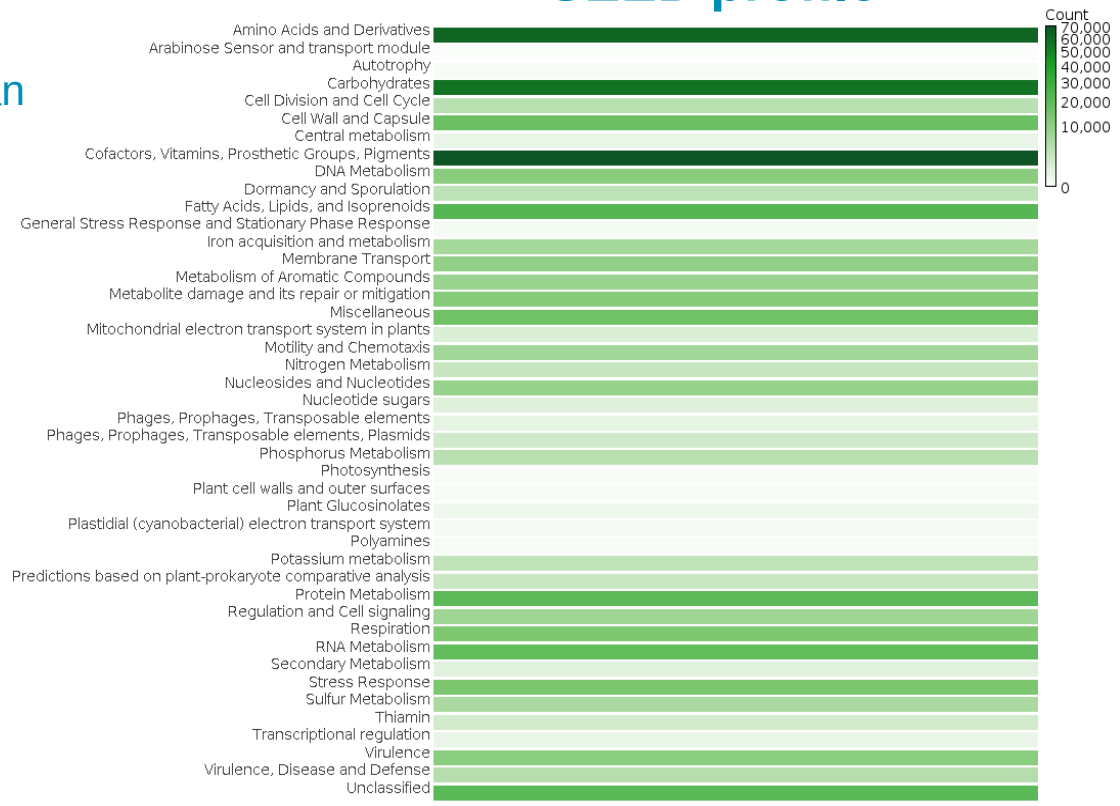
CD-hit clustering

- Gene finding with fragGeneScan

	2 Cells
# Genes	**2,691,726***
100%	695,181
99%	**451,483**
90%	193,897
Mean size (aa)	317

*Known redundancy

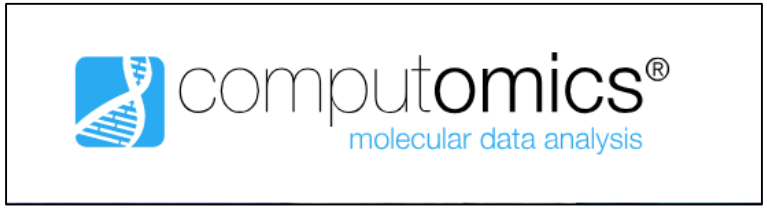
SEED profile



https://sourceforge.net/projects/fraggenescan/

diamond blastp
Visualized in MEGAN http://megan.informatik.uni-tuebingen.de/

3RD PARTY PROVIDERS FOR WGS DATA ANALYSIS



RECENT PUBLICATIONS – CHARACTERIZING COMMUNITIES

P.E. Zida et al. (2018) Increasing sorghum yields by seed treatment with an aqueous extract of the plant Eclipta alba may involve a dual mechanism of hydropriming and suppression of fungal pathogens *Crop Protection*

Heeger, Felix et al. (2018) Long-read DNA metabarcoding of ribosomal rRNA in the analysis of fungi from aquatic environments *BioRxiv*

Beaulaurier, John et al. (2017) Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nature Biotechnology*

Perraudeau, Fanny et al. (2017) Accurate determination of bacterial abundances in human metagenomes using full-length 16S sequencing reads *BioRxiv*

Wang, Yuan et al. (2017) Profiling of oral microbiota in early childhood caries using Single-Molecule Real-Time Sequencing *Frontiers in Microbiology*

Edwards, Joan E et al. (2017) PCR and omics based techniques to study the diversity, ecology and biology of anaerobic fungi: Insights, challenges andopportunities. *Frontiers in Microbiology*

Tedersoo, Leho et al. (2017) PacBio metabarcoding of fungi and other eukaryotes: errors, biases and perspectives. *The New Phytologist*

Zhao, J et al. (2017) Reduction in fecal microbiota diversity and short-chain fatty acid producers in Methicillin-resistant Staphylococcus aureus infected individuals as revealed by PacBio single molecule, real-time sequencing technology. *European Journal of Clinical Microbiology & Infectious Diseases*

Nakano, Kazuma et al. (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Human Cell*

Tai, Phillip WL et al. (2018) Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras *Molecular Therapy*

Xu, Haiyan et al. (2017) The effects of probiotics administration on the milk production, milk components and fecal bacteria microbiota of dairy cows *Science Bulletin*

Pootakham, Wirulda et al. (2017) High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Scientific Reports*

Sadowsky, Michael J et al. (2017) Analysis of gut microbiota – An ever changing landscape. *Gut Microbes*

McRose, Darcy L et al. (2017) Diversity and activity of alternative nitrogenases in sequenced genomes and coastal environments. *Frontiers in Microbiology*

Meng, Xiangli et al. (2017) Metataxonomics reveal vultures as a reservoir for Clostridium perfringens. *Emerging Microbes and Infections*

Motooka, Daisuke et al. (2017) Fungal ITS1 deep-sequencing strategies to reconstruct the composition of a 26-species community and evaluation of the gut mycobiota of healthy Japanese individuals. *Frontiers in Microbiology*

Driscoll, Connor B et al. (2017) Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in Genomic Sciences*

Hagen, Live H et al. (2017) Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. *Applied and Environmental Microbiology*

Li, Jing et al. (2017) Bacterial microbiota of Kazakhstan cheese revealed by single molecule real time (SMRT) sequencing and its comparison with Belgian, Kalmykian and Italian artisanal cheeses *BMC Microbiology*

Lam, Ka-Kit et al. (2016) BIGMAC : breaking inaccurate genomes and merging assembled contigs for long read metagenomic assembly. *BMC Bioinformatics*

Gesudu, Qimu et al. (2016) Investigating bacterial population structure and dynamics in traditional koumiss from Inner Mongolia using single molecule real-time sequencing. *Journal of Dairy Science*

Zheng, Yi et al. (2016) Using PacBio long-read high-throughput microbial gene amplicon sequencing to evaluate infant formula safety. *Journal of Agricultural and Food Chemistry*

Singer, Esther et al. (2016) High-resolution phylogenetic microbial community profiling. *The ISME Journal*

Gall, Cory A et al. (2016) The bacterial microbiome of Dermacentor andersoni ticks influences pathogen susceptibility. *The ISME Journal*

Armanhi, Jaderson Silveira Leite et al. (2016) Multiplex amplicon sequencing for microbe identification in community-based culture collections. *Scientific Reports*

Bao, Weichen et al. (2016) Assessing quality of Medicago sativa silage by monitoring bacterial composition with single molecule, real-time sequencing technology and various physiological parameters. *Scientific Reports*

Myer, Phillip R et al. (2016) Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers. *Journal of Microbiological Methods*

Frank, J A et al. (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific Reports*

Ikuta, Tetsuro et al. (2016) Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *The ISME Journal*

Tsai, Yu-Chih et al. (2016) Resolving the complexity of human skin metagenomes using single-molecule sequencing. *mBio*

Broecker, Felix et al. (2016) Long-term changes of bacterial and viral compositions in the intestine of a recovered Clostridium difficile patient after fecal microbiota transplantation *Molecular Case Studies*

Chen, Yuan et al. (2015) Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. *Fungal Genetics and Biology*

https://www.pacb.com/applications/complex-populations/microbial/

TAKE HOME MESSAGE

✓ **Microbial genome multiplexing** → **Simple, efficient workflow** from sample DNA to high-quality microbial genome assemblies including epigenetic modifications detection

✓ **16S profiling** → **Protocol available** for full-length 16S allowing Species level resolution of microbial communities

✓ **Shotgun Metagenomic (WGS)** → **New possibility** of long-read metagenomic profiling and metagenome assembly using last Sequel release (V 6.0)

PACBIO®

www.pacb.com