Sequencing of Microbial Genomes: Profiling Complex Populations on the Sequel System

Cheryl Heiner
Complete genome of *Lactobacillus hokkaidonensis* from psychrotrophic silage

Yasuhiro Tanizawa1, Masanori Toh1

Abstract

**Background:** *Lactobacillus hokkaidonensis* was isolated from Timothy grass silage in Hokkaido, Japan. It is a psychrotrophic silage starter culture in cold regions such as Hokkaido with a growth temperature of as low as 4°C. To elucidate its genetic features, we constructed the complete genome sequence of the organism using PacBio sequencing technology.
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.\(^1\)

Long, single-molecule reads generated high-resolution metagenomic profiles to the species and strain level for pre- and post-fecal-microbiota-transplant (FMT) samples from an individual suffering from chronic C. difficile infection.\(^2\)

---

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](http://molecularcasestudies.cshlp.org/content/2/1/a000448.full.pdf+html)
16S VS WGS – WGS PLUSES & MINUSES

Advantage: whole genome
- Provides functional information, in addition to classification
- Assembly of large contigs
  o 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:
  o Requires ~100 ng reasonable quality DNA
  o Sensitive to host contamination
- Requires more reads for characterization
- Assembly may requires higher coverage, depending on community
- Complex analysis
Full-Length 16S Sequencing
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

16S SEQUENCING AND CCS RESULTS WITH NEW CHEMISTRY

**P3 - C3 chemistry**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>1 Cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total yield</td>
<td>28 Gbases</td>
</tr>
<tr>
<td>Polymerase RL</td>
<td>50,562</td>
</tr>
<tr>
<td>Subread RL</td>
<td>2,152</td>
</tr>
<tr>
<td>Primary (P1) Reads</td>
<td>573,354</td>
</tr>
<tr>
<td>CCS yield (99%)</td>
<td>348,730</td>
</tr>
<tr>
<td>CCS yield (99.9%)</td>
<td>305,316</td>
</tr>
<tr>
<td>CCS yield (99.99%)</td>
<td>232,828</td>
</tr>
</tbody>
</table>

![Comparison of 16S CCS yields](image)

<table>
<thead>
<tr>
<th></th>
<th>P2/C2</th>
<th>P3/C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection</td>
<td>6 hours</td>
<td>15 hours</td>
</tr>
<tr>
<td>Pre-extension</td>
<td>0 hours</td>
<td>2 hours</td>
</tr>
<tr>
<td>P1 reads</td>
<td>454,534</td>
<td>573,354</td>
</tr>
<tr>
<td>Full version</td>
<td>S/P2-C2/5.0</td>
<td>S/P3-C3/5.0</td>
</tr>
</tbody>
</table>
- 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

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

- 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

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

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/10.1101/392333
RE-ANALYSIS REVEALED PREVIOUSLY REPORTED ‘SYSTEMATIC ERRORS’ WERE ACTUALLY GROUND TRUTH

- Authors reanalyzed sequencing data from a *Staphylococcus aureus* isolate used to describe systematic errors in Wagner et al. 2016

- “We recovered 5 ASVs, all of which *exactly matched previously sequenced 16S rRNA genes* from *S. aureus*.”

- “The differences between these intragenomic variants may have been misinterpreted as systematic errors, perhaps because *the short-read genome assembly that was used as the ground truth contained only one of the five rRNA operons in the *S. aureus* 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/10.1101/392333
3rd Party Providers for PacBio 16S Analysis

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 taxonomic assignments based on the differences in the 16S gene sequence.
Gene Profiling and Assembly
### WGS LIBRARY SIZE DEPENDS ON SAMPLE AND PROJECT

<table>
<thead>
<tr>
<th>Project Goal</th>
<th>Library Size</th>
<th>Input Sample</th>
<th>Community Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification, functional profiling</td>
<td>3 kb</td>
<td>100 ng, somewhat fragmented ok</td>
<td>All - complex OK</td>
</tr>
<tr>
<td>Find and co-localize genes, classification</td>
<td>6 kb – 10 kb</td>
<td>0.5 - 1 μg, fairly high molecular weight</td>
<td>All - complex OK</td>
</tr>
<tr>
<td>Assembly</td>
<td>10 kb</td>
<td>1 μg, high molecular weight</td>
<td>Simple, or a few predominant members</td>
</tr>
</tbody>
</table>
LONGER CCS LIBRARIES WITH NEW CHEMISTRY

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

<table>
<thead>
<tr>
<th># Bases</th>
<th>Polymerase RL mean</th>
<th># P1 reads</th>
<th># CCS reads &gt;99% predicted accuracy</th>
<th>Mean CCS read length</th>
<th>Mean # passes</th>
<th>Mean CCS predicted accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>38 Gb</td>
<td>77,846</td>
<td>496,590</td>
<td>273,261</td>
<td>8,345</td>
<td>16</td>
<td>99.9%</td>
</tr>
<tr>
<td>37 Gb</td>
<td>76,318</td>
<td>501,462</td>
<td>271,184</td>
<td>7,705</td>
<td>18</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

CCS Read Length

CCS Read Score

CCS results cell 1, 99% predicted accuracy
ANALYSIS PIPELINE – GENE FINDING AND LOCALIZATION

CCS Read

Gene

find genes

assemble

CCS reads or contigs

dereplicate genes

locate genes on contigs

8 kb mean

~900 bp
FUNCTIONAL CLASSIFICATION OF GENES FOUND

CD-hit clustering
- Gene finding with fragGeneScan

<table>
<thead>
<tr>
<th></th>
<th>2 Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td># Genes</td>
<td>2,691,726*</td>
</tr>
<tr>
<td>100%</td>
<td>695,181</td>
</tr>
<tr>
<td>99%</td>
<td>451,483</td>
</tr>
<tr>
<td>90%</td>
<td>193,897</td>
</tr>
<tr>
<td>Mean size (aa)</td>
<td>317</td>
</tr>
</tbody>
</table>

*Known redundancy

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

Visualized in MEGAN http://megan.informatik.uni-tuebingen.de/
SPIKE-IN REFERENCE ALIGNMENT TO PACBIO CONTIGS
3rd Party Providers for WGS Data Analysis

- CosmosID®
- Era7 bioinformatics
- computomics®
- DNA Nexus
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 hydorpriming and suppression of fungal pathogens Crop Protection


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


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


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


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, Livve 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

Gesud, 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


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


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


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

Breecr, 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/