Lawrence Berkeley National Laboratory
Recent Work

Title
Finishing of New Technology Only Microbes and Fungi

Permalink
https://escholarship.org/uc/item/23x9m7dt

Authors
Clum, Alicia
Sun, Hui
LaButti, Kurt
et al.

Publication Date
2009-05-27
Current Workflow:
Receive draft genome containing 454 shotgun titanium and 454 large insert paired end data
Separate gaps into subprojects and utilize unassembled 454 data to close gaps
Add fakes from resolved subprojects and velvet assembly to main project
Perform reactions for remaining subprojects
Repeat performing reactions and data addition until all gaps are closed and all repeats resolved (if possible)
Quality improvement using Illumina and Sanger where necessary
Completed genome ready for annotation

Sanger Data Use:
1. Gap closing reactions-PCR based methods are used to create templates and sequenced with Sanger to close gaps See images below for genome improvements using Sanger.
2. Assembly verification-PCR based methods are used to verify/resolve repeats as necessary.
3. Larger and more complex projects may require fosmid libraries for clone walking and repeat resolution.

Illumina Data Use:
1. Genome quality improvement using in-house tools (polisher.pl and acePolisher).
   - These scripts create files containing corrections and unsupported areas, which are used to make corrections, tag an acefile with various flags (polishTarget, unsupported, slesecaCorrected), and design primers for areas with targeted areas. See image A below for a sample tagged acefile.
2. Alignment to consensus sequence- This can be done to see the Illumina data for areas which need manual inspection. See image B for a Illumina data aligned to a reference using mosaik.
3. De-novo assembly of data- We standardly use Velvet for de-novo assembly of Illumina data. If Velvet contigs hit the main assembly, fakes are created and added to projects to close gaps See image C for a fake generated from a Velvet contig.

Future developments:
Eventually gaps may be pooled and sequenced using Illumina or 454. Other future developments include improvements in repeat resolution software and working with 454 to provide feedback and improvements for Newbler. Future fungal genome improvements may include targeted finishing, a limited number of rounds of reactions, in-silica only improvements, or complete finishing. Improvements to draft genomes will be discussed by Alla Lapidus.

454 Data Use:
1. De-novo assembly and draft creation- 454 standard titanium and a large insert 454 paired end library are assembled by Newbler to create a draft genome.
2. Subproject creation-For each gap a separate project is created (using gapResolution.pl). Unique reads and pairs, as well as unassembled data, are assembled to try and resolve the gap in-silica.
3. Genome quality improvement using in-house tools (polisher.pl)- 454 reads can be aligned and suggest corrections and improvements.

Assemble Illumina data using Velvet
The upper image to the current state Acidovorax delafieldii 2AN and the lower is the draft version. Most gaps have been closed with bubble PCR sequenced by Sanger. Where bubble PCR sequencing failed, attempts were made to close gaps with pairwise primers using the bubble PCR product as a template. To the right is sequencing data for this approach. For more information on bubble PCR see poster by Hope Tice.

Above is an assembly view image of a subproject created by our gapResolution software. It was created by pooling unique areas and mate pairs surrounding a gap. This gap was closed using bubble PCR. The image to the right shows the fake from the subassembly incorporated back into the main assembly. For more information on gapResolution software please see poster by Kurt LaButti and Stephan Trong.

PCR Methods:
New technology only projects have no templates, so we employ different PCR kits depending on the need.

<table>
<thead>
<tr>
<th>PCR methods</th>
<th>Kit</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>bubble PCR</td>
<td>Failsafe PCR System</td>
<td>Epicentre</td>
</tr>
<tr>
<td>long range PCR</td>
<td>LongRange PCR Kit</td>
<td>Qiagen</td>
</tr>
<tr>
<td>multiplex</td>
<td>Proof High Fidelity Kit</td>
<td>Bio-Rad Laboratories</td>
</tr>
<tr>
<td>standard pairwise or combinatorial</td>
<td>Failsafe or iProof</td>
<td>Epicentre/Bio-Rad</td>
</tr>
</tbody>
</table>

Image A: Tags generated from acePolisher. The green tag represents an insertion or deletion. The purple tag is a polishTarget whose area was addressed with a Sanger polishing read. The magenta tags represent “unsupported” bases where there isn’t enough coverage to clearly make a base change.

Image B: Illumina data aligned to a reference sequence using mosaik. These alignments can be used to manually inspect areas where depth of coverage wasn’t enough for automatic correction.

Image C: A fake created from a Velvet assembled Illumina contig is used to close a gap.