New technology drafts: production and improvements

Alla Lapidus, Ph.D

DOE Joint Genome Institute,
Walnut Creek, CA
All Genome projects

Major Sequencing Centers
May 2009: 4849 projects

- JGI: 22%
- JCVI: 15%
- BROAD: 9%
- WashU: 7%
- Sanger: 6%
- BMC-HGSC: 5%
- GENOSCOPE: 3%
- World: 33%
- BMC-HGSC: 5%
- Sanger: 6%
- WashU: 7%
- JGI: 22%
- JCVI: 15%
- BROAD: 9%
- GENOSCOPE: 3%
- World: 33%

Advancing Science with DNA Sequence
Advancing Science with DNA Sequence

Genome projects
Archaea + Bacteria only

Sequencing Centers for Archaea & Bacteria
May 2009: 3549 projects

- JGI 23%
- JCVI 18%
- BROAD 9%
- WashU 6%
- BCM 5%
- WORLD 37%

298 Complete Genomes
137 Complete Genomes

http://www.genomesonline.org/
Microbial and Fungal Whole Genome Sequencing Strategy

- 454 standard (15-20x)
- 454 PE (10-20kb; 5x)
- Fosmids ends (for fungi)

**ASSEMBLY**

Next:

- Illumina reads: gaps, quality improvement

Next:

- Sanger finishing reads: gaps, quality improvement

**Finished Genome**

- No gaps
- No misassemblies
- < 1 error/50kb
- ACGT only
Why more than one platform?

- **454** - high quality reliable skeletons of genomes (454 std + 454 PE): correctly assembled contigs; problems with repeats (unassembled or assembled in contigs outside of main scaffolds); homopolymer related frame shifts

- Illumina data is used to help improve the overall consensus quality, correct frameshifts and to close secondary structure related gaps; not ready for de-novo assembly of complex genomes (too many gaps!)

- Sanger – finishing reads; fosmids – larger repeats and templates for primer walk – less cost effective but very useful in many cases
1. Pyrosequence and Sanger to obtain main ordered and oriented part of the assembly – Newbler assembler

2. GapResolution (in house tool) to close some (up to 40%) gaps using unassembled 454 data – PGA or Newbler assemblers

3. Solexa reads to detect and correct errors in consensus – in house created tool (the Polisher) and close gaps (Velvet)

* Fosmids ends not used for microbes
Advancing Science with DNA Sequence

**Gaps: 454 contigs vs. 454 reads aligned to final consensus**

<table>
<thead>
<tr>
<th>Brachybacterium faecium DSM 04810 (72% GC)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>size (bp)</th>
<th>GC%</th>
<th>454 coverage</th>
<th>Newbler contigs</th>
<th>454 reads aligned</th>
<th>Newbler contigs</th>
<th>454 reads aligned</th>
<th>% of genome in gaps</th>
<th>min gap (bp)</th>
<th>max gap (bp)</th>
<th>average gap (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,615,825</td>
<td>71.7</td>
<td>24</td>
<td>250</td>
<td>124</td>
<td>54099</td>
<td>3256</td>
<td>1.5</td>
<td>0.09</td>
<td>2</td>
<td>6008</td>
</tr>
</tbody>
</table>

231 gaps – SG Sanger reads only

19 gaps – SG + finishing reads
What is in 454 gaps?

- Repeats
- High GC secondary structures

16S rRNA operon

Finished reference

rRNA repeats

454 scaffolds
Step 1 For each gap, identify read pairs from contigs found on different scaffolds

Step 2 Assemble reads in contigs adjacent to the gap and reads obtained from contigs outside the scaffold. Sometimes use assembler other than Newbler for sub-assemblies (PGA)
**Step 3** If gap is not closed, tool designs primers for sequencing reactions

**Step 4** Iterate as necessary (in sub-assemblies)

See Poster **FF0070** for details (K.LaButti, S.Trong, etc)
Another way of solving gaps: use Illumina reads

- Velvet assembly
- Blast Velvet contigs against Newbler ends
- Use proper Velvet contigs to close gaps

Velvet contigs close gaps caused by hairpins and secondary structures
Low quality areas – areas of potential frameshifts

Assemblies contain low quality regions (red tags)
Advancing Science with DNA Sequence

Homopolymer related frameshifts

Frameshift 1 (AAAAA, should be AAAA)

Frameshift 2 (CCCC, should be CCC)

Modified from N. Ivanova
**Step 1:** Align Illumina data to 454-only or Sanger/454 hybrid assembly

**Step 2:** Analyze and correct consensus errors

- **Unsupported**
  - Illumina coverage < 10X
  - Illumina coverage >= 10X and <70% of Illumina bases agree with the reference base

- **Corrections**
  - Illumina coverage >= 10X and at least 70% Illumina bases disagrees with the reference base

**Step 3:** Design sequencing reactions for low quality and unsupported Illumina areas
Other approaches (further assembly improvement)

- CONSED and CloneView (in house development) – main visualization (and not only) tools
- bPCR – bubble PCR for clone-free primer walking (Poster 53 – Hope Tice)
Conclusions

• Combination of different sequencing platforms is the best approach for de novo genome sequencing. It allows to combine all the benefits together and produce high quality drafts. None of the platforms are ready to be used for the de novo complex assemblies / finishing by themselves. Bioinformatics tools can significantly improve the situation.

• High quality draft assembly is a great starting point for finishing

• The more problems are resolved at the draft stage, the faster and cheaper will the finishing be

• Approaches, developed for draft improvement, can be used for automated high throughput finishing (we use them for microbial and fungal genomes – see posters FF0070, FF0076, FF0081)

Comment: I think finished genomes are better than any other options and they serve any purposes. Large centers are responsible for the data quality, which will serve people for decades.
Advancing Science with DNA Sequence

JGI (>400 ongoing projects)

Polisher and gapResolution:
Brian Foster
Kurt LaButti
Stephan Trong
Cliff Han
Tom Brettin

bPCR protocol:
Hope Tice
Alicia Clum
Jan-Fang Cheng

Production Dpt:
Susan Lucas – department head

QA/QC and R&D
Matt Nolan
Alex Copeland
Feng Chen