4 - Genome Assembly and Validation (Concepts)

Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia

Wednesday afternoon
Assembly project workflow | Prior Knowledge

- Karyotype: Genome size, Ploidy
- Heterozygocity
- GC content
- Contaminants / Symbionts
- Data Sets:
  - Close relatives
  - Genes / ESTs / RNAseq / Markers
- Mitocondria
- Chloroplast
Experiment design (you choose the data!)

• Know your biological question.

• Plan your data processing (from an information perspective).

• Decide on conditions and biological/technical replicas.

• Decide on technologies and coverages:
  • How will the typical bias affect your experiment?
  • Is the coverage enough? Significant results?
The genome assembly problem (WGS)

Original DNA

Fragments

Sequenced ends

Contigs

Scaffold
Planning and "informed guesses"

Whole genome information (unknown?)
+ Bias Information (unknown)

"known error profile"
short sequences \(10^{2-3}\)
large data \(10^X\)

\(Y \ll X\)

small data \(10^Y\)
long sequences \(10^{10}\)
"unknown error profile"
Whole genome information (known?)
+ Bias Information (unknown, reduced?)

Assemble and Scaffold

Scaffolds & Contigs

Validate and release
The assembly is just a probabilistic model of a genome, condensing the information from the experimental evidence.

All the information is already present in the experimental results.
A correct assembly has:

The right *motifs*,
the correct number of times,
in correct order and position.

None of which is assessed by length stats.
A modern assembler

Using SOAPdenovo2 as an example
Genomic DNA

Fragment and paired-end sequencing of libraries with variant insert sizes.

A

150~500 bp

2~10 Kb

Represent read sequence overlap using de Bruijn graph

Remove erroneous connections on the graph

B

C

(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles

D

Break at repeat boundaries and output contigs

Scaffold construction

E

F

Gap closure
Fragment and paired-end sequencing of libraries with variant insert sizes.

A

150~500 bp
2~10 Kb

B

Represent read sequence overlap using de Bruijn graph

Remove erroneous connections on the graph

C

(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles

D

Break at repeat boundaries and output contigs
C
(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles

D
Break at repeat boundaries and output contigs
Scaffold construction

E

F
Gap closure
Assembly validation

Using biological knowledge to figure out what are…

The right *motifs,*
the correct number of times,
in correct order and position.
Direct experimental evidence: the reads

ACTGACTGCCTGTGTGTGTGTGTGTGTGTGTGTGGACTGTTAAA

ACTGACTGC

GACTGTAAAA

structure sequence

The right *motifs*, the correct number of times, in correct order and position.
<table>
<thead>
<tr>
<th>Error type</th>
<th>Transcripts</th>
<th>Assembly</th>
<th>Read evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family collapse</td>
<td>geneAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>geneAB</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>geneAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chimerism</td>
<td>geneC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>geneB</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsupported insertion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incompleteness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fragmentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Local misassembly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redundancy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Read evidence**

- **bases in reads**
- **agreement**
- **coverage**
- **no reads align to insertion**
- **read pairs align off end of contig**
- **bridging read pairs**
- **read pairs in wrong orientation**
- **all reads assign to best contig**
Direct experimental evidence: other evidence

- Genome size, ploidy
- GC content
- Symbionts
- Plastids
- ESTs, cDNAs, peptides, genome walking

The right *motifs*,
the correct number of times,
in correct order and position.
Indirect experimental evidence: genomes in general

- Genes! They have structure
- Repeats
- Chromosome macrostructure
  - (circular?, number, telomeres, …)

The right *motifs*,
the correct number of times,
in correct order and position.
Indirect experimental evidence: other species

- Close relatives: proteins, transcripts, genomes
- Distant relatives: single-copy genes, phylogeny, HGT

The right **motifs**, the correct number of times, in correct order and position.
Questions?