Evaluating Assembly Quality
Michael Schatz

Dec 8, 2014
USDA/ARS Workshop
1. Assembly review
   1. Assembly by analogy
   2. Causes of Mis-assemblies

2. Evaluating Assembly Quality
   1. Assemblathon
   2. Size Statistics
   3. Mate-pair Happiness
   4. CEGMA

3. RNA-seq specific challenges
Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of *A Tale of Two Cities*
  - Text printed on 5 long spools

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

- How can he reconstruct the text?
  - 5 copies x 138,656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical
Greedy Reconstruction

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem
de Bruijn Graph Construction

- $D_k = (V,E)$
  - $V =$ All length-$k$ subfragments ($k < l$)
  - $E =$ Directed edges between consecutive subfragments
    - Nodes overlap by $k-1$ words

- **Locally constructed graph reveals the global sequence structure**
  - Overlaps between sequences implicitly computed

---

**Original Fragment**

```
It was the best of
```

**Directed Edge**

```
It was the best -> was the best of
```

de Bruijn, 1946  
Idury and Waterman, 1995  
Pevzner, Tang, Waterman, 2001
After graph construction, try to simplify the graph as much as possible
de Bruijn Graph Assembly

It was the best of times, it

of times, it was the

it was the worst of times, it

it was the age of

the age of foolishness

the age of wisdom, it was the

After graph construction, try to simplify the graph as much as possible
de Bruijn Graph Assembly

A TALE OF TWO CITIES

BOOK THE FIRST. RECALLED TO LIFE

CHAPTER 1

The Past

It was the best of times, it was the worst of times. It was the age of wisdom, it was the age of foolishness; it was the epoch of belief, it was the epoch of incredulity; it was the season of Light, it was the season of Darkness; it was the spring of hope, it was the winter of despair; we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct to Hell.
Ingredients for a good assembly

Coverage

High coverage is required
- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

Read Length

Reads & mates must be longer than the repeats
- Short reads will have false overlaps forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality

Errors obscure overlaps
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in de novo plant genome sequencing and assembly
Typical sequencing coverage

Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs $1
1x sequencing
2x sequencing
4x sequencing
8x sequencing
Genome Coverage Distribution

Expect Poisson distribution on depth
• Standard Deviation = sqrt(cov)

This is the mathematically model => reality may be much worse
• Double your coverage for diploid genomes
• Can use somewhat lower coverage in a population to find common variants
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness. After graph construction, try to simplify the graph as much as possible.
Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka “unitigs”, “unipaths”
  - Unitigs end because of (1) lack of coverage, (2) errors, (3) heterozygosity/isoform differences, and (4) repeats
Over 50% of mammalian genomes are repetitive
- Large plant genomes tend to be even worse
- Wheat: 16 Gbp; Pine: 24 Gbp
The full tale

... it was the best of times it was the worst of times ...
... it was the age of wisdom it was the age of foolishness ...
... it was the epoch of belief it was the epoch of incredulity ...
... it was the season of light it was the season of darkness ...
... it was the spring of hope it was the winter of despair ...
Errors in the graph

<table>
<thead>
<tr>
<th>Clip Tips</th>
<th>Pop Bubbles</th>
</tr>
</thead>
<tbody>
<tr>
<td>was the worst of times,</td>
<td>was the worst of times,</td>
</tr>
<tr>
<td>was the worst of tymes,</td>
<td>was the worst of tymes,</td>
</tr>
<tr>
<td>the worst of times, it</td>
<td>times, it was the age</td>
</tr>
<tr>
<td></td>
<td>tymes, it was the age</td>
</tr>
</tbody>
</table>

(Chaisson, 2009)
Outline

1. Assembly review
   1. Assembly by analogy
   2. Causes of Mis-assemblies

2. Evaluating Assembly Quality
   1. Assemblathon
   2. Size Statistics
   3. Mate-pairs
   4. CEGMA

3. RNA-seq specific challenges
• Attempt to answer the question:
  “What makes a good assembly?”

• Organizers provided sequence data to assembly experts around the world
  – Assemblathon 1: ~100Mbp simulated genome
  – Assemblathon 2: 3 vertebrate genomes each ~1GB

• Results demonstrate trade-offs assemblers must make

Assemblathon 1: A competitive assessment of de novo short read assembly methods.

Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species
Assembly Results

Scaffolds
- Broad
- DOEJGI
- CSHL

Scaffold Paths
- WTSI-S
- DOEJGI
- Broad

Contig Paths
- BGI
- Broad
- CSHL

Fill Color Key
- Item >= 1
- 1e2
- 1e3
- 1e4
- 1e5
- 1e6
- 1e7
Final Rankings

- ALLPATHS and SOAPdenovo came out neck-and-neck followed closely behind by Celera Assembler, SGA, and ABySS

- My recommendation for “typical” short read assembly is to use ALLPATHS

- Single molecule sequencing becoming extremely attractive if you have access
N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome

50%

N50 size = 30 kbp
(300k + 100k + 45k + 45k + 30k = 520k >= 500kbp)

A greater N50 is indicative of improvement in every dimension:
• Better resolution of genes and flanking regulatory regions
• Better resolution of transposons and other complex sequences
• Better resolution of chromosome organization
• Better sequence for all downstream analysis
19+ vertebrates assembled with ALLPATHS-LG

- Coelacanth
- Stickleback
- Tilapia
- Spotted Gar
- Male ferret
- Female ferret
- Shrew
- P. nyererei
- Tenrec
- Chinchilla
- B6
- NA12878
- Squirrel monkey
- Bushbaby
- Ground squirrel
- A. burtoni
- M. zebra
- N. brichardi
- Female ferret
- Male ferret

scaffold N50 (Mb)
contig N50 (kb)
Ingredients for a good assembly

**Coverage**

- **High coverage is required**
  - Oversample the genome to ensure every base is sequenced with long overlaps between reads
  - Biased coverage will also fragment assembly

**Read Length**

- **Reads & mates must be longer than the repeats**
  - Short reads will have **false overlaps** forming hairball assembly graphs
  - With long enough reads, assemble entire chromosomes into contigs

**Quality**

- **Errors obscure overlaps**
  - Reads are assembled by finding kmers shared in pair of reads
  - High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Estimating coverage with Kmers

Reference:

Reads:

...GATTACA
GATTACAC
TACACGCT...
Estimating coverage with Kmers

Reference:

Reads:

NA12878
QC: Read Coverage

Reference: [Bar graph showing a distribution of reads]

Reads: [Bar graph showing a distribution of reads]

Errors: [Bar graph showing a distribution of errors]

Coverage: [Bar graph showing a distribution of coverage]

Repeats: [Bar graph showing a distribution of repeats]
Wheat Genome
(A. tauschi / CSHL)
Heterozygous Genome

Contact: @mike_schatz
Ingredients for a good assembly

Coverage

High coverage is required
- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

Read Length

Reads & mates must be longer than the repeats
- Short reads will have false overlaps forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality

Errors obscure overlaps
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in de novo plant genome sequencing and assembly
Assembly Validation

Automatically scan an assembly to locate misassembly signatures for further analysis and correction

Assembly-validation pipeline
1. Evaluate Mate Pairs & Libraries
2. Evaluate Read Alignments
3. Evaluate Read Breakpoints
4. Analyze Depth of Coverage

Genome Assembly forensics: finding the elusive mis-assembly.
Paired-end and Mate-pairs

**Paired-end sequencing**
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

**Mate-pair sequencing**
- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads
C/E Statistic

- The presence of individual compressed or expanded mates is rare but expected.

- Do the inserts spanning a given position differ from the rest of the library?
  - Flag large differences as potential misassemblies
  - Even if each individual mate is “happy”

- Compute the statistic at all positions
  - \( \frac{\text{Local Mean} - \text{Global Mean}}{\text{Scaling Factor}} \)

- Introduced by Jim Yorke’s group at UMD
Sampling the Genome

Normal Library
Count=10000, Mean=4000, SD=400

8 inserts: 3kb-6kb
Local Mean: 4048
C/E Stat: \( \frac{(4048-4000)}{(400 / \sqrt{8})} = +0.33 \)

Near 0 indicates overall happiness
C/E-Statistic: Expansion

Normal Library
Count=10000, Mean=4000, SD=400

8 inserts: 3.2kb-6kb

Local Mean: 4461
C/E Stat: \( \frac{(4461-4000)}{(400 / \sqrt{8})} = +3.26 \)

C/E Stat ≥ 3.0 indicates Expansion
C/E-Statistic: Compression

8 inserts: 3.2 kb-4.8kb
Local Mean: 3488
C/E Stat: \( \frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62 \)
C/E Stat ≤ -3.0 indicates Compression
Hawkeye & AMOS: Visualizing and assessing the quality of genome assemblies
Long Read Sequencing Technology

PacBio RS II

Moleculo

Oxford Nanopore

CSHL/PacBio

(Voskoboynik et al. 2013)

CSHL/ONT
O. sativa pv Indica (IR64)

Genome size:  ~370 Mb
Chromosome N50:  ~29.7 Mbp

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Contig NG50</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiSeq Fragments</td>
<td>19 kbp</td>
</tr>
<tr>
<td>25x 456bp (3 runs 2x300 @ 450 FLASH)</td>
<td></td>
</tr>
<tr>
<td>“ALLPATHS-recipe”</td>
<td>18 kbp</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td>36x 2x50bp @ 2100</td>
<td></td>
</tr>
<tr>
<td>51x 2x50bp @ 4800</td>
<td></td>
</tr>
<tr>
<td>HGAP</td>
<td>4.0 Mbp</td>
</tr>
<tr>
<td>22.7x @ 10kbp</td>
<td></td>
</tr>
<tr>
<td>Nipponbare BAC-by-BAC Assembly</td>
<td>5.1 Mbp</td>
</tr>
</tbody>
</table>

HGAP Read Lengths
Max: 53,652bp
22.7x over 10kbp
(discarded reads below 8500bp)
Ingredients for a good assembly

**Coverage**

*High coverage is required*
- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

**Read Length**

*Reads & mates must be longer than the repeats*
- Short reads will have false overlaps forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

**Quality**

*Errors obscure overlaps*
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in de novo plant genome sequencing and assembly
Detection and Correction with Quake

1. Count all “Q-mers” in reads
   - Fit coverage distribution to mixture model of errors and regular coverage
   - Automatically decide threshold for trusted k-mers

2. Correction Algorithm
   - Consider editing erroneous kmers into trusted k-mers in decreasing likelihood
   - Includes quality values, nucleotide/nucleotide substitution rate

Quake: quality-aware detection and correction of sequencing reads.
Gene Analysis with CEGMA

- Defined a set of 248 “core eukaryotic genes” (CEGs)
  - Highly conserved and in low copy numbers across all known eukaryotic species
  - House keeping genes and other basic functions

- Developed a robust alignment-based search tool to seek out those genes in your new assembly
  - Your ability to discover these 248 CEGs is highly correlated with finding the rest of the genes in the genome

Assessing the gene space in draft genomes
1. Assembly review
   1. Assembly by analogy
   2. Causes of Mis-assemblies

2. Evaluating Assembly Quality
   1. Assemblathon
   2. Size Statistics
   3. Mate-pairs
   4. CEGMA

3. RNA-seq specific challenges
RNA-seq Overview
RNA-seq Challenges

**Challenge 1: Eukaryotic genes are spliced**
Solution: Use a spliced aligner, and assemble isoforms

*TopHat: discovering spliced junctions with RNA-Seq.*

**Challenge 2: Read Count != Transcript abundance**
Solution: Infer underlying abundances (e.g. FPKM)

*Transcript assembly and quantification by RNA-seq*

**Challenge 3: Transcript abundances are stochastic**
Solution: Replicates, replicates, and more replicates

*RNA-seq differential expression studies: more sequence or more replication?*
Assembly Summary

Assembly quality depends on

1. **Coverage**: low coverage is mathematically hopeless
2. **Repeat composition**: high repeat content is challenging
3. **Read length**: longer reads help resolve repeats
4. **Error rate**: errors reduce coverage, obscure true overlaps

• Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  – Extensive error correction is the key to getting the best assembly possible from a given data set

• Watch out for collapsed repeats & other misassemblies
  – Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together
Questions?

http://schatzlab.cshl.edu/
@mike_schatz