Genome Sequencing & Assembly
Michael Schatz

July 6, 2014
Frontiers of techniques in plant sciences
Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats

2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer

3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for PacBio projects

4. Summary & Recommendations
Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats

2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer

3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for PacBio projects

4. Summary and Recommendations
**Shredded Book Reconstruction**

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

<table>
<thead>
<tr>
<th>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
<tr>
<td>It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...</td>
</tr>
</tbody>
</table>

- 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 | Directed Edge
---|---
It was the best of | It was the best
| was the best of

- de Bruijn, 1946
- Idury and Waterman, 1995
- Pevzner, Tang, Waterman, 2001
It was the best of times, it was the age of wisdom, it was the worst of times, it was the age of foolishness.

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

the age of foolishness

it was the age of

the age of wisdom, it was the

After graph construction, try to simplify the graph as much as possible
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 winder of despair …
Assembly Complexity
Milestones in Genome Assembly

1977. Sanger et al.
1st Complete Organism
5375 bp

1995. Fleischmann et al.
1st Free Living Organism
TIGR Assembler. 1.8Mbp

1998. C.elegans SC
1st Multicellular Organism
BAC-by-BAC Phrap. 97Mbp

2000. Myers et al.
1st Large WGS Assembly.
Celera Assembler. 116 Mbp

2001. Venter et al., IHGSC
Human Genome
Celera Assembler/GigaAssembler. 2.9 Gbp

2010. Li et al.
1st Large SGS Assembly.
SOAPdenovo 2.2 Gbp

Like Dickens, we must computationally reconstruct a genome from short fragments.
Assembly Applications

- Novel genomes
- Metagenomes
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - ...
Assembling a Genome

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads

   ...AGCCTAGGGATGCGCGACACGT
   GGATGCGCGACACGTGCAATATCGGTTTGGTCACACCTCGGACGGAC
   CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

4. Detangle graph with long reads, mates, and other links
Why are genomes hard to assemble?

1. **Biological:**
   - (Very) High ploidy, heterozygosity, repeat content

2. **Sequencing:**
   - (Very) large genomes, imperfect sequencing

3. **Computational:**
   - (Very) Large genomes, complex structure

4. **Accuracy:**
   - (Very) Hard to assess correctness
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
Massively Parallel Sequencing

1. Attach
2. Amplify
3. Image

Illumina HiSeq 2000
Sequencing by Synthesis

>60 Gbp / day

http://www.youtube.com/watch?v=l99aKKHcxC4
Typical contig coverage

Imagine raindrops on a sidewalk
How many rain drops should we collect?
1x sequencing
2x sequencing

- Balls in Bins
  - Total balls: 2000

- Histogram of balls in each bin
  - Total balls: 2000
  - Empty bins: 142
4x sequencing
8x sequencing
Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

Formulation comes from the limit of the binomial equation

Resembles a normal distribution, but over the positive values, and with only a single parameter.

**Key property:** The standard deviation is the square root of the mean.
Coverage and Read Length

Idealized Lander-Waterman model

- Reads start at perfectly random positions

- Contig length is a function of coverage and read length
  - Short reads require much higher coverage to reach same expected contig length

- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
  - Recommend 100x coverage

Assembly of Large Genomes using Second Generation Sequencing
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, and (3) repeats
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 ty mes,</td>
<td>was the worst of ty mes,</td>
</tr>
<tr>
<td>the worst of times, it</td>
<td>times, it was the age</td>
</tr>
<tr>
<td></td>
<td>ty mes, it was the age</td>
</tr>
</tbody>
</table>

(Chaisson, 2009)
## Repetitive regions

<table>
<thead>
<tr>
<th>Repeat Type</th>
<th>Definition / Example</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-complexity DNA / Microsatellites</td>
<td>((b_1b_2\ldots b_k)^N) where (1 \leq k \leq 6) CACACACACACACACACA</td>
<td>2%</td>
</tr>
<tr>
<td>SINEs (Short Interspersed Nuclear Elements)</td>
<td>Alu sequence (~280 bp)</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>Mariner elements (~80 bp)</td>
<td></td>
</tr>
<tr>
<td>LINEs (Long Interspersed Nuclear Elements)</td>
<td>~500 – 5,000 bp</td>
<td>21%</td>
</tr>
<tr>
<td>LTR (long terminal repeat) retrotransposons</td>
<td>Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)</td>
<td>8%</td>
</tr>
<tr>
<td>Other DNA transposons</td>
<td></td>
<td>3%</td>
</tr>
<tr>
<td>Gene families &amp; segmental duplications</td>
<td></td>
<td>4%</td>
</tr>
</tbody>
</table>

- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - Wheat: 16 Gbp; Pine: 24 Gbp
Repeats and Coverage Statistics

- If \( n \) reads are a uniform random sample of the genome of length \( G \), we expect \( k = n \Delta / G \) reads to start in a region of length \( \Delta \).
  - If we see many more reads than \( k \) (if the arrival rate is > \( A \)), it is likely to be a collapsed repeat.
  - Requires an accurate genome size estimate.

\[
\Pr(X - \text{copy}) = \left(\frac{n}{k}\right)\left(\frac{X\Delta}{G}\right)^k \left(\frac{G - X\Delta}{G}\right)^{n-k}
\]

\[
A(\Delta, k) = \ln \left( \frac{\Pr(1 - \text{copy})}{\Pr(2 - \text{copy})} \right) = \ln \left( \frac{(\Delta n / G)^k e^{\frac{-\Delta n}{G}}}{k!} \right) = \frac{n\Delta}{G} - k \ln 2
\]
**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

10kbp circle

2x100 @ ~10kbp (outies)

2x100 @ 300bp (innies)
Scaffolding

• Initial contigs (*aka* unipaths, unitigs) terminate at
  – Coverage gaps: especially extreme GC regions
  – Conflicts: sequencing errors, repeat boundaries

• Iteratively resolve longest, ‘most unique’ contigs
  – Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  – Uniqueness measured by a statistical test on coverage
N50 size

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

Example: 1 Mbp genome

\[
\text{N50 size} = 30 \text{ kbp} \\
(300k + 100k + 45k + 45k + 30k = 520k \geq 500kbp)
\]

Note:
N50 values are only meaningful to compare when base genome size is the same in all cases
Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats

2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer

3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for PacBio projects

4. Summary and Recommendations
Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
University of Maryland
Goal of WGA

- For two genomes, $A$ and $B$, find a mapping from each position in $A$ to its corresponding position in $B$
Not so fast...

- Genome $A$ may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to $B$ (sometimes all of the above)
WGA visualization

• How can we visualize whole genome alignments?

• With an alignment dot plot
  – $N \times M$ matrix
    • Let $i =$ position in genome $A$
    • Let $j =$ position in genome $B$
    • Fill cell $(i,j)$ if $A_i$ shows similarity to $B_j$
  – A perfect alignment between $A$ and $B$ would completely fill the positive diagonal
SV Types

- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

http://mummer.sf.net/manual/AlignmentTypes.pdf
Alignment of 2 strains of *Y. pestis*

http://mummer.sourceforge.net/manual/
Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats

2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer

3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for long read projects

4. Summary and Recommendations
Genome assembly with ALLPATHS-LG
Iain MacCallum
How ALLPATHS-LG works

reads
  └── corrected reads
    └── doubled reads

unipaths
  └── localized data
    └── local graph assemblies
        └── global graph assembly

assembly
*For best results. Normally not used for small genomes. However essential to assemble long repeats or duplications.

Cutting coverage in half still works, with some reduction in quality of results.

All: protocols are either available, or in progress.
To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:

More than one closure allowed (but rare).
Localization

I. Find ‘seed’ unipaths, evenly spaced across genome (ideally long, of copy number CN = 1)

II. Form neighborhood around each seed

reaches to other unipaths (CN = 1) directly and indirectly

read pairs reach into repeats and are extended by other unipaths
19+ vertebrates assembled with ALLPATHS-LG

- spotted gar: N50 = 69 kb
- male ferret: N50 = 67 kb
- female ferret
- bushbaby
- squirrel monkey: N50 = 19 Mb
- NA12878
- chinchilla
- N. brichardi
- A. burtoni
- M. zebra
- P. nyererei
- tenrec
- ground squirrel
- tilapia
- Shrew
- B6
- stickleback
- coelacanth
- N. brichardi
- NA12878

scaffold N50 (Mb)

contig N50 (kb)
Population structure of *Oryza sativa*

New whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*

Strain specific regions

Conclusions

• Very high quality representation of the “gene-space”
  • Overall identity ~99.9%
  • Less than 1% of exonic bases missing

• Genome-specific genes enriched for disease resistance
  • Reflects their geographic and environmental diversity
  • Detailed analysis of agriculturally important loci

• Assemblies fragmented at (high copy) repeats
  • Missing regions have mean k-mer coverage >10,000x
  • Difficult to identify full length gene models and regulatory features
Genome assembly with the Celera Assembler
1. Pre-overlap
   – Consistency checks

2. Trimming
   – Quality trimming & partial overlaps

3. Compute Overlaps
   – Find high quality overlaps

4. Error Correction
   – Evaluate difference in context of overlapping reads

5. Unitigging
   – Merge consistent reads

6. Scaffolding
   – Bundle mates, Order & Orient

7. Finalize Data
   – Build final consensus sequences

http://wgs-assembler.sf.net
Long Read Sequencing Technology

PacBio RS II

Moleculo

Oxford Nanopore

CSHL/PacBio

(Voskoboynik et al. 2013)

Broad/OxNano @ AGBT ***
Sample of 100k reads aligned with BLASR requiring >100bp alignment

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>83.7%</td>
</tr>
<tr>
<td>Insertions</td>
<td>11.5%</td>
</tr>
<tr>
<td>Deletions</td>
<td>3.4%</td>
</tr>
<tr>
<td>Mismatch</td>
<td>1.4%</td>
</tr>
</tbody>
</table>
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

PacBioToCA & ECTools

Hybrid/PB-only Error Correction

HGAP & Quiver

PB-only Correction & Polishing

< 5x PacBio Coverage > 50x

Quiver Performance Results
Comparison to Reference Genome
(M. ruber; 3.1 MB; SMRT® Cells)

<table>
<thead>
<tr>
<th></th>
<th>Initial Assembly</th>
<th>Quiver Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>QV</td>
<td>43.4</td>
<td>54.5</td>
</tr>
<tr>
<td>Accuracy</td>
<td>99.99540%</td>
<td>99.99964%</td>
</tr>
<tr>
<td>Differences</td>
<td>141</td>
<td>11</td>
</tr>
</tbody>
</table>
S. pombe dg21

PacBio RS II sequencing at CSHL
• Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science

Over 275x coverage in 5 SMRTcells using P5-C3

Mean: 5170
103x over 10kbp
7.6x over 20kb
Max: 35,415bp
S. pombe dg21

ASM294 Reference sequence
• 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp

PacBio assembly using HGAP + Celera Assembler
• 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id

Near perfect assembly:
- Chr1: 1 contig
- Chr2: 2 contigs
- Chr3: 2 contigs
- MT: 1 contig
S. pombe dg21

ASM294 Reference sequence
- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp

PacBio assembly using HGAP + Celera Assembler
- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id
A. thaliana Ler-0 sequenced at PacBio

- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

<table>
<thead>
<tr>
<th>Genome size: 124.6 Mbp</th>
<th>Sum of Contig Lengths: 149.5Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome N50: 23.0 Mbp</td>
<td>N50 Contig Length: 8.4 Mb</td>
</tr>
<tr>
<td>Corrected coverage: 20x over 10kb</td>
<td>Number of Contigs: 1788</td>
</tr>
</tbody>
</table>

High quality assembly of chromosome arms
Assembly Performance: $8.4\text{Mbp}/23\text{Mbp} = 36\%$
MiSeq assembly: $63\text{kbp}/23\text{Mbp} = .2\%$
**Human CHM1**

http://blog.pacificbiosciences.com/2014/02/data-release-54x-long-read-coverage-for.html

**CHM1hert** sequenced at PacBio

- Sequenced using the P5 enzyme and C3 chemistry
- Size selection using an 20kb elution window on a BluePippin™ device from Sage Science
- Total coverage: 54x

<table>
<thead>
<tr>
<th>Genome size:</th>
<th>3.0 Gb</th>
<th>Sum of Contig Lengths:</th>
<th>3.2 Gb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome N50:</td>
<td>90.5 Mbp</td>
<td>N50 Contig Length:</td>
<td>4.38 Mbp</td>
</tr>
<tr>
<td>Average read length:</td>
<td>7,680 bp</td>
<td>Max Contig:</td>
<td>44 Mbp</td>
</tr>
</tbody>
</table>

High quality draft assembly
Assembly Performance: 4.38Mbp/90.5Mbp = 4.5%
Sanger HuRef assembly: 107kbp / 90.5Mbp = .1%
Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

\[
CNS\ Error = \sum_{i=[c/2]}^{c} \left( \binom{c}{i} e^i (1-e)^{n-i} \right)
\]

*Nature Biotechnology.* 30:693–700
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
What should we expect from an assembly?

Analysis of dozens of genomes from across the tree of life with real and simulated data

Summary & Recommendations

< 100 Mbp:  HGAP/PacBio2CA @ 100x PB C3-P5
expect near perfect chromosome arms

< 1GB:  HGAP/PacBio2CA @ 100x PB C3-P5
high quality assembly: contig N50 over 1Mbp

> 1GB:  hybrid/gap filling
expect contig N50 to be 100kbp – 1Mbp

> 5GB:  Email mschatz@cshl.edu

Error correction and assembly complexity of single molecule sequencing reads.

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395
## Acknowledgements

<table>
<thead>
<tr>
<th>Schatz Lab</th>
<th>CSHL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giuseppe Narzisi</td>
<td>Hannon Lab</td>
</tr>
<tr>
<td>Shoshana Marcus</td>
<td>Gingeras Lab</td>
</tr>
<tr>
<td>James Gurtowski</td>
<td>Iossifov Lab</td>
</tr>
<tr>
<td>Alejandro Wences</td>
<td>Levy Lab</td>
</tr>
<tr>
<td>Hayan Lee</td>
<td>Lippman Lab</td>
</tr>
<tr>
<td>Rob Aboukhalil</td>
<td>Lyon Lab</td>
</tr>
<tr>
<td>Mitch Bekritsky</td>
<td>Martienssen Lab</td>
</tr>
<tr>
<td>Charles Underwood</td>
<td>McCombie Lab</td>
</tr>
<tr>
<td>Rushil Gupta</td>
<td>Ware Lab</td>
</tr>
<tr>
<td>Avijit Gupta</td>
<td>Wigler Lab</td>
</tr>
<tr>
<td>Shishir Horane</td>
<td>IT Department</td>
</tr>
<tr>
<td>Deepak Nettem</td>
<td></td>
</tr>
<tr>
<td>Varrun Ramani</td>
<td></td>
</tr>
<tr>
<td>Piyush Kansal</td>
<td>NBACC</td>
</tr>
<tr>
<td>Eric Biggers</td>
<td>Adam Phillippy</td>
</tr>
<tr>
<td>Aspyn Palatnick</td>
<td>Sergey Koren</td>
</tr>
</tbody>
</table>
Thank you!

http://schatzlab.cshl.edu

@mike_schatz