Genome Sequencing & Assembly

Michael Schatz

March 31, 2014
CSHL Genome Access
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
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
Dickens accidentally shreds the first printing of *A Tale of Two Cities* – Text printed on 5 long spools

- 5 copies × 138,656 words / 5 words per fragment = 138k fragments
- The short fragments from every copy are mixed together
- Some fragments are identical

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

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

Original Fragment  Directed Edge

```
It was the best of
```

```
It was the best
```

```
was the best of
```

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

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.
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.
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 ...
Assembly Complexity
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
GGATGCGCGACACGTCGCATATCCGGTTTGTC AACCTCGGACGGAC
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
Illumina Sequencing by Synthesis

1. Prepare
2. Attach
3. Amplify
4. Image
5. Basecall

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

Imagine raindrops on a sidewalk
Balls in Bins

Total balls: 1000
Balls in Bins 2x

Balls in Bins
Total balls: 2000

Histogram of balls in each bin
Total balls: 2000 Empty bins: 142
Balls in Bins 4x

Balls in Bins
Total balls: 4000

Histogram of balls in each bin
Total balls: 4000 Empty bins: 17
Balls in Bins 8x

Histogram of balls in each bin
Total balls: 8000 Empty bins: 1

Balls in Bins
Total balls: 8000
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, and (3) repeats
Errors in the graph

(Chaisson, 2009)

<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>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>was the worst of times,</td>
</tr>
<tr>
<td></td>
<td>times, it was the age</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>worst of times, it</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) CACACACACACACACACACACACACACACA</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

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

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

300bp

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

10kbp

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
Break
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
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
Translocation  Inversion  Insertion
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
Seed-and-extend with MUMmer

How can quickly align two genomes?

1. Find maximal-unique-matches (MUMs)
   - Match: exact match of a minimum length
   - Maximal: cannot be extended in either direction without a mismatch
   - Unique
     - occurs only once in both sequences (MUM)
     - occurs only once in a single sequence (MAM)
     - occurs one or more times in either sequence (MEM)

2. Cluster MUMs
   - using size, gap and distance parameters

3. Extend clusters
   - using modified Smith-Waterman algorithm
nucmer --maxmatch CO92.fasta KIM.fasta
  --maxmatch Find maximal exact matches (MEMs)

delta-filter -m out.delta > out.filter.m
  -m Many-to-many mapping

show-coords -r out.delta.m > out.coords
  -r Sort alignments by reference position

dnadiff out.delta.m
Construct catalog of sequence variations

mummerplot --large --layout out.delta.m
  --large Large plot
  --layout Nice layout for multi-fasta files
  --x11 Default, draw using x11 (--postscript, --png)
  *requires gnuplot

See manual at 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 PacBio projects
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
ALLPATHS-LG sequencing model

<table>
<thead>
<tr>
<th>Libraries (insert types)</th>
<th>Fragment size (bp)</th>
<th>Read length (bases)</th>
<th>Sequence coverage (x)</th>
<th>Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragment</td>
<td>180*</td>
<td>≥ 100</td>
<td>45</td>
<td>yes</td>
</tr>
<tr>
<td>Short jump</td>
<td>3,000</td>
<td>≥ 100 preferable</td>
<td>45</td>
<td>yes</td>
</tr>
<tr>
<td>Long jump</td>
<td>6,000</td>
<td>≥ 100 preferable</td>
<td>5</td>
<td>no**</td>
</tr>
<tr>
<td>Fosmid jump</td>
<td>40,000</td>
<td>≥ 26</td>
<td>1</td>
<td>no**</td>
</tr>
</tbody>
</table>

*See next slide.

**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.
Error correction

Given a crystal ball, we could stack reads on the chromosomes they came from (with homologous chromosomes separate), then let each column ‘vote’:

But we don’t have a crystal ball....
ALLPATHS-LG. For every K-mer, examine the stack of all reads containing the K-mer. Individual reads may be edited if they differ from the overwhelming consensus of the stack. If a given base on a read receives conflicting votes (arising from membership of the read in multiple stacks), it is not changed. (K=24)

columns inside the kmer are homogeneous

columns outside the kmer may be mixed

change to C

Two calls at Q20 or better are enough to protect a base
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).
Unipath: unbranched part of genome – squeeze together perfect repeats of size ≥ K

Adjacent unipaths overlap by K-1 bases
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
Create assembly from global assembly graph

- Flatten
- Scaffold
- Patch
- Fix
Large genome recipe: **ALLPATHS-LG vs capillary**

<table>
<thead>
<tr>
<th>Cost</th>
<th>Continuity</th>
<th>Completeness</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>contig N50 (kb)</td>
<td>scaffold N50 (Mb)</td>
<td>genome (%)</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>25</td>
<td>88.7</td>
</tr>
<tr>
<td></td>
<td>17.5</td>
<td>16.9</td>
<td></td>
</tr>
</tbody>
</table>

Mouse Genome
19+ vertebrates assembled with ALLPATHS-LG

- female ferret
- male ferret
- N. brichardi
- spotted gar
- B6
- 129
- N. brichardi
- ground squirrel
- tilapia
- shrew
- P. nyererei
- tenrec
- coelacanth
- chinchilla
- NA12878
- male ferret
- 69 kb
- 67 kb
- 19 Mb
- 19 Mb
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
Hybrid Sequencing

**Illumina**

*Sequencing by Synthesis*

- High throughput (60Gbp/day)
- High accuracy (~99%)
- Short reads (~100bp)

**Pacific Biosciences**

*SMRT Sequencing*

- Lower throughput (600Mbp/day)
- Lower accuracy (~85%)
- Long reads (2-5kbp+)
SMRT Sequencing

Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).

SMRT Sequencing Data

TTGTAAGCAGTTGAAAACTATGTGTCATTAGAATAAAAGACATGAAG
TTGTAAGCAGTTGAAAACTATGTG-TATTAGAATAAAAGACATGAAG
ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGCGGCCTAGG
A-TATAAA-TCAGTTGATCCATTAGAA-AGAAACGC-AAAGGC-GCTAGG
CAACCTTGAATGTAATCGCACTTGAGAAACAAGATTATTATTCGGCGCCG
C-ACCTTG-ATGT-AT--CAGTTGAGAAACAAGATTATTATTCGGCGCCG
TAACGAATCAGATTCTGAAAACAT-ATAACAACCTCCAAA-CACAA
T-ACGAATC-AGATTCTGAAAACAT-ACCTCCAAGACACAAACGACAA
-AGGAGGGGAAGGGGAATATCT-ATAAAAGATTACAATATGATAG
GAGAGG--AA------GAATATCTGAT-AAAAAGATTAAATGTAG
ACT-AATTCAACTAATATACACTTTTA--ACAGAATTGAT-GGAA-GTT
ACTAAATTCAACAATAATACACTTTTA--ACAGAATTGAT-GGAA-GTT
TCGGAGAGATCCAAAACAATGGCC-ATCGCCCTTGA-GTATAC-AATCAA
TC-GAGAGATTCC-AACAAAT--GGGATCG-CTTTGACGTACCACAAATCAA
ATCCAGTGAAATATAATTTATCGAATCCAGAACTTTATCCAAATTAG
ATCCAGT-GAAATATA-CTTAGC-ATCCA-GAATTATATCCAAATTAG

Sample of 100k reads aligned with BLASR requiring >100bp alignment

| Match  | 83.7% |
| Insertions | 11.5% |
| Deletions | 3.4% |
| Mismatch | 1.4% |
Consensus Accuracy and Coverage

Coverage can overcome random errors

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

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

Nature Biotechnology. 30:693–700
PacBio Assembly Algorithms

- **PBJelly**
  - Gap Filling and Assembly Upgrade
  - English *et al* (2012)
  - *PLOS One*. 7(11): e47768

- **PacBioToCA & ECTools**
  - Hybrid/PB-only Error Correction
  - *Nature Biotechnology*. 30:693–700

- **HGAP & Quiver**
  - PB-only Correction & Polishing
  - Chin *et al* (2013)

---

PacBio Coverage

- < 5x
- > 50x
S. cerevisiae W303

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

Mean: 5910

83x over 10kbp

8.7x over 20kb

Max: 36,861bp

Over 175x coverage in 16 SMRTcells / 2 days using P5-C3
S. cerevisiae W303
S288C Reference sequence
• 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp
PacBio assembly using HGAP + Celera Assembler
• 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id
S. cerevisiae W303

S288C Reference sequence
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler
- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id

Near-perfect assembly:
All but 1 chromosome assembled as a single contig

35kbp repeat cluster
**A. thaliana Ler-0**


*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

**Genome size:** 124.6 Mbp

**Chromosome N50:** 23.0 Mbp

**Corrected coverage:** 20x over 10kb

**Sum of Contig Lengths:** 149.5Mb

**N50 Contig Length:** 8.4 Mb

**Number of Contigs:** 1788

High quality assembly of chromosome arms

Assembly Performance: 8.4Mbp/23Mbp = 36%

MiSeq assembly: 63kbp/23Mbp = .2%
ECTools: Error Correction with pre-assembled reads

https://github.com/jgurtowski/ectools

Short Reads -> Assemble Unitigs -> Align & Select - > Error Correct

Can Help us overcome:
1. Error Dense Regions – Longer sequences have more seeds to match
2. Simple Repeats – Longer sequences easier to resolve

However, cannot overcome Illumina coverage gaps & other biases
A. thaliana Ler-0


![Graph showing coverage and read lengths compared to different assembly tools. The graph compares HGAP and ECTools against coverage levels, with a clear trend indicating increased read lengths with higher coverage.]
O. sativa pv Indica (IR64)

PacBio RS II sequencing at PacBio
- Size selection using an 10 Kb elution window on a BluePippin™ device from Sage Science

Over 14.1x coverage in 47 SMRTcells using P5-C3

Max: 54,288bp

Mean: 10,232bp

12.34x over 10kbp

4.1x over 20kb
**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><strong>MiSeq Fragments</strong></td>
<td>19,078</td>
</tr>
<tr>
<td>25x 456bp</td>
<td></td>
</tr>
<tr>
<td>(3 runs 2x300 @ 450 FLASH)</td>
<td></td>
</tr>
<tr>
<td><strong>“ALLPATHS-recipe”</strong></td>
<td>18,450</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><strong>ECTools</strong></td>
<td>271,885</td>
</tr>
<tr>
<td>10.7x @ 10kbp</td>
<td></td>
</tr>
</tbody>
</table>

ECTools Read Lengths  
Mean: 9,348  
Max: 54,288bp  
10.75x over 10kbp
What should we expect from an assembly?

https://en.wikipedia.org/wiki/Genome_size
Assembly Complexity of Long Reads

Assembly complexity of long read sequencing
Assembly Complexity of Long Reads

Assembly complexity of long read sequencing
Assembly Complexity of Long Reads

Assembly complexity of long read sequencing
Assembly Recommendations

• Long read sequencing of eukaryotic genomes is here

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

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

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

  > 5GB:    Email mschatz@cshl.edu

• **Caveats**
  – Model only as good as the available references (esp. haploid sequences)
  – Technologies are quickly improving, exciting new scaffolding technologies
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
Acknowledgements

Schatz Lab
Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Alejandro Wences
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Rushil Gupta
Avijit Gupta
Shishir Horane
Deepak Nettem
Varrun Ramani
Piyush Kansal
Eric Biggers
Aspyn Palatnick

CSHL
Hannon Lab
Gingeras Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

IT Department

NBACC

Adam Phillippy
Sergey Koren
Thank You!

http://schatzlab.cshl.edu

@mike_schatz