De Novo Assembly of High-throughput Short Read Sequences

Chuming Chen

Center for Bioinformatics and Computational Biology (CBCB)
University of Delaware

NECC Third Skate Genome Annotation Workshop
May 23, 2011
Outline

• Genome assembly primer
• High-throughput short read sequencing (NGS) assembly pipeline
• Case study
Genome sequencing and assembly

• Genome
  – Long stretches of contiguous DNA sequences (base pairs)
  – Different genome sizes (i.e. virus: 3.5k, Human: 3.3 billion)

• Genome sequencers (NHGRI, Feb. 4, 2011)
  – Sanger-based sequencing (500-600 bases)
  – 454 sequencing (300-400 bases)
  – Illumina and SOLiD sequencing (50-100 bases)

• Sequencing and assembly
  – A genome must be fragmented, sequenced piece by piece and then re-assembled to obtain the full contiguous sequence
Assembly approaches

• Hierarchical assembly
  – Mapping the genome to a set of large insert clones
  – Reduce the assembly of the sequencing reads from the entire genome to a single clone, typically 40 - 200 Kb
  – The genome sequence is then assembled by aligning sequences of adjacent clones

• Whole genome shot-gun assembly
  – The entire genome is fragmented
  – The shotgun process takes reads from random positions along the chromosomes that make up one genome
  – The assembler then reconstructs the reads up to the chromosome length
  – Assembly is possible because the target is over-sampled by the shotgun reads, such that reads overlap
Whole genome shot-gun sequencing and assembly

Genome

Short Reads

Overlapping

Contigs

Scaffolding

Scaffold/SuperContigs

Pairing-end/Mate pair reads

Finishing/Gap closing

Assembled Genome

(modified from Lin, 2010)
Paired-end vs. mate-pair reads

(Fass, 2010)

Paired-end reads:
- Short fragments: 200 ~ 500bp

Mate-pair reads:
- Long fragments: 2 ~ 5 kbp
- Circularized fragment: F/R, short insert length (PE)
- Cut the circularized fragments, sequence both ends, and you end up with a mixed population

(a) Correct assembly
(b) Mis-assembly
(c) Correct assembly
(d) Mis-assembly

(Phillippy, 2008)
Assembly pipeline

Sample DNA

Prepare Insert Libraries

Sequencer Run

Sequence file (.fastq)

Check Quality of Sequences (Filtering)

De Novo Assembly

Consensus sequence file (.fasta)

Check Quality of Assembly/Map reads back
HiSeq2000 sequence data processing pipeline

On-instrument Real Time Analysis

Image Analysis (Firecrest)
- .tiff

Base calling (Bustard)
- .bcl

OLB
- .bcl to .qseq
- .qseq

CASAVA
Sequence Analysis

FASTQ format sequence files:
Single-end reads:
- s_?_1_sequence.txt

Paired-end/Mate-pair reads:
- s_?_1_sequence.txt and s_?_2_sequence.txt

Other files ....
FASTQ format sequence files

s_?_1_sequence.txt

@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAATACCATTTAGTTCAAGGGAAAGCCAGCAAAGTTCGATCAAGGATAGACCGAGG
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffffdfddddffgggfafffcdfffdffdddeaeegfggafaffW\]^a`b^d[dbdbcccccb`dW\a`\caascfaffafaccfcd\cfcc\cb

s_?_2_sequence.txt

@HWI-ST741_0085:1:1101:1444:1939#0/2
CCCAGCTTATCCTTGCAACTCTTCTTAAATAGAGGCACAACATTAATCACCTCCCTCTATGGACTCCTTATACCTACCCCTCTCCCTCCCTCGGCAGGCCAG
+HWI-ST741_0085:1:1101:1444:1939#0/2
edeadffdfaffcdaaeffdfdecefaceccfcdffdfddfffdfffdcafafffcffgagggggg`ggggggfggggggggd]eeda_^BB

@HWI-ST741_0085:1:1101:1444:1939#0/2
ACCATTGCCTTTCTTTACTGCTGCACCTGCTACCTTTAAATAGACTGCTGACTACCATGACACCAGGTCTGCTGCATCTCCCCCTCCATAC
+HWI-ST741_0085:1:1101:1444:1939#0/2
hhhhhhhhhhhhhhhhhhhhhhhhhhhhfghhhhhfhddhdhahhhhffefhhffdfdfedhhehhhhhehhehccccfffd...
A closer look at sequence file

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAAT</td>
<td>CCCAGCTTATCCTTGCAACTCTTCTTAAATAGAGGCACAACATTAATCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+HWI-ST741_0085:1:1101:1444:1939#0/1</td>
<td>+HWI-ST741_0085:1:1101:1444:1939#0/2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. the unique instrument name
2. flowcell lane
3. tile number within the flowcell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. index number for a multiplexed sample (0 for no multiplexing)
7. the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

(Wikipedia.org)
Quality scores

- Phred quality scores $Q$ are defined as a property which is logarithmically related to the base-calling error probabilities $P$
- A Phred score of a base is: $Q_{\text{phred}} = -10 \log_{10} P$, where $P$ is the estimated probability of a base being wrong

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90 %</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99 %</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9 %</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99 %</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999 %</td>
</tr>
</tbody>
</table>

- Phred scoring scheme, encoded as an ASCII character by adding 64 to the Phred value

@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTTGCAAGATACGATACATGATACGGGAAT
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffffdffffffgggaffffcdffbfdddeaegf[gaaffW^a]  

HiSeq score

Phred score

(Wikipedia.org)
Sequence reads quality assessment

- **FastQC** (Baraham Bioinformatics, UK)
  - Basic Statistics
  - Per Base Sequence Quality
  - Per Sequence Quality Scores
  - Per Base Sequence Content
  - Per Base GC Content
  - Per Sequence GC Content
  - Per Base N Content
  - Sequence Length Distribution
  - Duplicate Sequences
  - Overrepresented Sequences
  - Overrepresented Kmers
Per base sequence quality

Yellow box: 25-75 quartile; Black whisker: min-max; Red line: median; Blue line: mean
Trim Sequences

- Quality trimming
  - Based on quality scores
- Ambiguity trimming
  - Remove stretches of Ns
- Adapter sequence trimming
  - Remove sequence adapters
- Base trim
  - Remove a specified number of bases at either 3' or 5' end of the reads
- Length trimming
  - Remove reads shorter or longer than a specified threshold
De Novo Assembly algorithms

• Overlap/Layout/Consensus Graph
• de Bruijin Graph
Overlap/Layout/Consensus graphs

- A node corresponds to a read, an edge denotes an overlap between two reads.
- The overlap graph is used to compute a layout of reads and consensus sequence of contigs by pair-wise sequence alignment.
- Good for sequences with limited number of reads but significant overlap. Computational intensive for short reads (short and high error rate).
- Example assemblers: Celera Assembler, Arachne, CAP and PCAP
de Bruijin graphs

- No need for all against all overlap discovery.
- Break reads into smaller sequences of DNA (K-mers, K denotes the length in bases of these sequences).
- Captures overlaps of length K-1 between these K-mers.
- More sensitive to repeats and sequencing errors.
- By construction, the graph contains a path corresponding to the original sequence.
- Example assemblers: Euler, Velvet, ABySS, AllPaths, SOAPdenovo, CLC Bio

(modified from Schatz, 2010)
CLC Bio De Novo assembly

- Make a table of the words (K-mers) seen in the reads.
- Build de Bruijn graph from the word table.
- Use the reads to resolve the repeats.
- Use the information from paired reads to resolve larger repeats.
- Output resulting contigs based on the paths.

(modified from CLC Bio, 2011)
Word (K-mer) size

To strike a balance, CLC bio’s de novo assembler chooses a word length based on the amount of input data: the more data, the longer the word length. It is based on the following:

word size 12: 0 bp - 30000 bp
word size 13: 30001 bp - 90002 bp
word size 14: 90003 bp - 270008 bp
word size 15: 270009 bp - 810026 bp
word size 16: 810027 bp - 2430080 bp
word size 17: 2430081 bp - 7290242 bp
word size 18: 7290243 bp - 21870728 bp
word size 19: 21870729 bp - 65612186 bp
word size 20: 65612187 bp - 196836560 bp
word size 21: 196836561 bp - 590509682 bp
word size 22: 590509683 bp - 1771529048 bp
word size 23: 1771529049 bp - 5314587146 bp
word size 24: 5314587147 bp - 15943761440 bp
word size 25: 15943761441 bp - 47831284322 bp
word size 26: 47831284323 bp - 143493852968 bp
word size 27: 143493852969 bp - 430481558906 bp
word size 28: 430481558907 bp - 1291444676720 bp
word size 29: 1291444676721 bp - 3874334030162 bp
word size 30: 3874334030163 bp - 11623002090488 bp
word size 31: 11623002090489 bp and up

(CLC Bio, 2011)
Repeats or sequencing errors

Graph Reduction

SNP or Sequencing Error

Repeat Sequence

(modified from CLC Bio, 2011)
Assembly quality assessment

• Continuity
  – Lengths distribution of contigs/scaffolds.
  – Average length, minimum and maximum lengths, combined total lengths.
  – N50 captures how much of the assembly is covered by relatively large contigs.
  – The N50 is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly.
  – Compute N50
    • first ordering all contigs (or scaffolds) by length,
    • then summing up their lengths (starting with the longest) until the sum exceeds 50% of the total length of all contigs.
    • the corresponding contig length is N50.

• Accuracy or “Correctness”
  – Base accuracy – the frequency of calling the correct nucleotide at a given position in the assembly.
  – Mis-assembly rate – the frequency of rearrangements, significant insertions, deletions and inversions.
Case study

• Show the basic steps involved in De Novo assembly of high-throughput short read sequences

• Data
  – 3 lanes of Illumina HiSeq 2000 short read sequences for the little skate (a couple weeks ago)

• Assembler
  – CLC Bio Genomics Workbench 4.6
Import sequence data

qseq files specifies whether a read has passed a quality filter or not. If checked, these reads will be ignored during import.

The reads are automatically trimmed when a quality score B is encountered in the input file.
Quality trimming

1. Convert quality score (Q) to error probability: \( P_{\text{error}} = 10^{Q/-10} \), low values are high quality bases.
2. For every base a new value is calculated: \( \text{Limit} - P_{\text{error}} \), negative for low quality bases.
3. For every base, the running sum of this value is calculated. If the sum drops below zero, it is set to zero.
4. The region retained is between the first positive value of the running sum and the highest value of the running sum. Everything before and after this region will be trimmed off.

If this maximum is set to e.g. 2, the algorithm finds the maximum length region containing 2 or fewer ambiguities and then trims away the ends not included in this region.
Adapter trimming

1. Remove adapter
2. Discard when not found
3. Discard when found

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Strand</th>
<th>Alignment score</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' adapter small RNA</td>
<td>ATCGTAGGCACCT...</td>
<td>Minus</td>
<td>3, 5, 15, 2</td>
<td>Remove adapter</td>
</tr>
<tr>
<td>454 Sequence Prim...</td>
<td>GCCTCCCTCGGC...</td>
<td>Plus</td>
<td>3, 2, 15, 2</td>
<td>Remove adapter</td>
</tr>
<tr>
<td>454 Sequence Prim...</td>
<td>GCCTTGCCAGGCC...</td>
<td>Minus</td>
<td>3, 2, 15, 2</td>
<td>Remove adapter</td>
</tr>
<tr>
<td>454 miRNA forwar...</td>
<td>GCCTCCCTCGGC...</td>
<td>Plus</td>
<td>3, 2, 15, 2</td>
<td>Discard when not f...</td>
</tr>
<tr>
<td>454 miRNA revers...</td>
<td>GCCTTGCCAGGCC...</td>
<td>Minus</td>
<td>3, 2, 15, 2</td>
<td>Discard when not f...</td>
</tr>
<tr>
<td>Illumina HiSeq 200...</td>
<td>ACACTCTTCCCTA...</td>
<td>Plus</td>
<td>2, 3, 4</td>
<td>Remove adapter</td>
</tr>
</tbody>
</table>

Number of reads: 1,000
Number of nucleotides: 9

<table>
<thead>
<tr>
<th>Name</th>
<th>Matches found</th>
<th>Reads discarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina HiSeq 200...</td>
<td>19</td>
<td>0</td>
</tr>
</tbody>
</table>

Costs:
- Mismatch cost: 2
- Gap cost: 3

Match thresholds:
- Minimum score: 10
- Allow internal matches: ✔
- Allow end matches: ✔
- Minimum score at end: 4
Sequence filtering

1. Select sequencing data
2. Quality trimming
3. Adapter trimming
4. Sequence filtering

Trim bases
- Remove 5' terminal nucleotides
- Remove 3' terminal nucleotides

Filter on length
- Discard reads below length: 15
- Discard reads above length: 1,000
Trimming result handling

1. Select sequencing data
2. Quality trimming
3. Adapter trimming
4. Sequence filtering
5. Result handling

Output options
- Create list of discarded sequences
- Create report

Result handling
- Open
- Save

Log handling
- Make log
Trimming report

Summary

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of reads</th>
<th>Avg. length</th>
<th>Number of reads after trim</th>
<th>Percentage retained</th>
<th>Avg. length after trim</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1_1_sequence (paired)</td>
<td>190,790,624</td>
<td>93.4</td>
<td>190,784,343</td>
<td>~100%</td>
<td>93.1</td>
</tr>
<tr>
<td>S_3_1_sequence (paired)</td>
<td>1,655,955</td>
<td>63.9</td>
<td>1,654,977</td>
<td>99.94%</td>
<td>63.0</td>
</tr>
<tr>
<td>S_5_1_sequence (paired)</td>
<td>209,140,424</td>
<td>92.9</td>
<td>209,131,515</td>
<td>~100%</td>
<td>92.6</td>
</tr>
<tr>
<td>S_3_1_sequence</td>
<td>2,016,294</td>
<td>62.4</td>
<td>2,015,199</td>
<td>99.95%</td>
<td>61.8</td>
</tr>
<tr>
<td>S_5_1_sequence</td>
<td>223,212,034</td>
<td>92.4</td>
<td>223,201,524</td>
<td>~100%</td>
<td>92.1</td>
</tr>
<tr>
<td>S_5_1_sequence</td>
<td>2,373,091</td>
<td>62.2</td>
<td>2,371,693</td>
<td>99.98%</td>
<td>61.5</td>
</tr>
</tbody>
</table>

Trim settings

- Removal of low quality sequence. (limit = 0.05).
- Removal of ambiguous nucleotides: maximal 2 nucleotides allowed.
- Removal of adapter sequences, using the following adaptors:
  # Illumina HiSeq 2000 PE Adapter (ACACTCTTTCCCTACACGACGCTCTTCCGATCT), strand = Plus, action = Remove adapter, score = [2, 3, 10, 4]

Detailed results

<table>
<thead>
<tr>
<th>Trim</th>
<th>Input reads</th>
<th>No trim</th>
<th>Trimmed</th>
<th>Nothing left or Discarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trim on quality</td>
<td>629,187,422</td>
<td>618,707,269</td>
<td>10,474,178</td>
<td>5,975</td>
</tr>
<tr>
<td>Ambiguity trim</td>
<td>629,181,447</td>
<td>628,910,632</td>
<td>256,018</td>
<td>14,797</td>
</tr>
<tr>
<td>Adapter trimming</td>
<td>629,166,650</td>
<td>617,589,864</td>
<td>11,589,357</td>
<td>7,396</td>
</tr>
</tbody>
</table>
Assembly parameters

Minimum length fraction of reads must match the reference
Minimum fraction of similarity between read and the reference
If checked, end gaps will be treated as mismatches
General assembly options

1. Select sequencing reads
2. Set parameters
3. Set general assembly options

Options
- Add conflict annotations

Conflict resolution
- Unknown nucleotide (N)
- Ambiguity nucleotides (R, Y, etc)
- Vote (A, C, G, T)
- **Non-specific matches**
  - Random
  - Ignore

Contig length
- Minimum contig length: 200
Assembly result handling

1. Select sequencing reads
2. Set parameters
3. Set general assembly options
4. Result handling

**Output options**
- Create simple contig sequences (fast)
- Map reads back to contigs (slow)
  - Update contigs based on mapped reads
  - Create summary report
  - Create list of un-mapped reads

Result handling
- Open
- Save

Log handling
- Make log
Assembly report

Summary statistics

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Average length</th>
<th>Total bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads</td>
<td>629,173,407</td>
<td>92.43</td>
<td>58,156,992,907</td>
</tr>
<tr>
<td>Matched</td>
<td>599,285,257</td>
<td>92.94</td>
<td>55,697,623,043</td>
</tr>
<tr>
<td>Not matched</td>
<td>29,888,150</td>
<td>82.29</td>
<td>2,459,369,864</td>
</tr>
<tr>
<td>Contigs</td>
<td>2,494,829</td>
<td>610</td>
<td>1,523,965,030</td>
</tr>
<tr>
<td>Reads in pairs</td>
<td>162,778,034</td>
<td>362.64</td>
<td></td>
</tr>
<tr>
<td>Broken paired reads</td>
<td>431,137,809</td>
<td>91.92</td>
<td></td>
</tr>
</tbody>
</table>

Quality assessment

Total length of sequences (bp): 1,523,965,030
Total number of contigs: 2,494,829
Max contig length (bp): 22,049
Mean contig length (bp): 610.85
Median contig length (bp): 371
Min contig length (bp): 200
N25: 1720
N50: 891
N75: 435
N90: 251
Total GC count (bp): 650,659,197
GC (%): 42.70
View assembled contig
Assembled contigs

>Consensus from Contig 1
Average coverage: 25.94
ACCATCATCAGTGATAGGGTTTACCTTTAGCAAATTTTAGGCTAGCAGAAACTTTGAAGCT
AAACCTCTGAAAAATTGGCAAATCTTTTTGTATCTTTTGCAGCTGAGTTTAGAATTGTAC
AAAGCACTCTATCTTCTTTGGCAGATTGATAAAAATCTGATGCTTCTGTAGCACGTGGAGTA
AACACTTCCCTCTGAAATTAGCTCTGTTGAATGGTACAGAAACACTGCCAGAAACAC
TTGGCACCCTCAATCACAATTCCTTTGTAGCACCTTGGTGGATTTTA

>Consensus from Contig 2
Average coverage: 13.80
TCCCCCTCCCCCAAAAAATCAGCAACTCAAATGGGGGTGGAGTTGTCTGTCTTCAATTGCGAG
GAAATACGGTATAGCGATCTAAAGACAATCTGATCTTTTGGAGTCATATATTCAGGGTCTT
GATTACCTTGAGTTCTGCTTTGCAATGCATATCTGATATGCACCTTTAATTCACA
ATTGCCATTTAAAAACTATAGAGCCTTATATTCTACCCCTAAACCATTTTGGGGTAAA
AGCAATGAGCAGCTTTCTGTCTTTTCAGGTATAAATCTCAGTGAGAGCTGCATAAGAAAATGT
GGAGACGTCAAACATTTTT

>Consensus from Contig 3
Average coverage: 24.54
CACCCCCACACACACAGCCACACTGTCCCGACCCCTCCCTTCACCTACGTAAGGGCGCA
GCAGCAGCTCAGCTTTTGACATCATTTTGCTTGGCAGCTCGTCGTCAGAGTGGAAGT
CCTCCTGTCCATCGACCAACAAATCTCCTGTAGTGATTCGGAGGATGTCGAAGAGCCTC
CGTAATCCTCTCTGTCGGCAGAGACACTGTCTCCTCTCCTCCCACACACACCTGGTCTC
CACACCCACCTCATGTTGGAG
Summary

• Genome sequencing and assembly problem
• Short read sequence assembly pipeline
  – Sequence data format (FASTQ)
  – Read quality assessment
  – Sequence trimming
  – De Novo assembly algorithms and tools
  – Assembly quality assessment
• Case study
  – Little Skate Illumina HiSeq 2000 short read sequences
  – CLC Bio Genomics Workbench
References

• Dawei Lin, Short Read Assembly, UC Davis 2010 Bioinformatics Short Course, September 13, 2010, UC Davis
• Joe Fass, Genome Sequence Assembly in Action!, UC Davis 2010 Bioinformatics Short Course, September 13, 2010, UC Davis
• CLC Bio. Manual for CLC Genomics Workbench 4.6 Windows, Mac OS X and Linux April 1, 2011
Thank You!

Questions???