**Graduate Computational Genomics**

02-710 / 10-810 & MSCBIO2070

---

**Genome Sequencing**

Takis Benos

Lecture #10, February 7, 2008

Reading: handouts & papers

---

**Overview**

- **Genome sequencing strategies**
  - Hierarchical sequencing ("clone-by-clone")
  - Whole genome sequencing (WGS)

- **Assembly algorithms**
  - Phred, Phrap
  - ARACHNE
  - Celera assembler
Genome sequencing

- Cloning
- Mapping/Sequencing
- Assembly
- Gap filling

Genome sequencing - cloning

- Cloning
- Mapping/Sequencing
- Assembly
- Gap filling

Insert size (kb)
- Plasmids: 1-20
- Cosmids: 30-50
- BACs: 100-250
- YACs: 100-3,000
Genome sequencing - sequencing

- Cloning
- Mapping/Sequencing
- Assembly
- Gap filling

Sequencing methods: hierarchical sequencing

Genome

Subcloning

BAC library (100-200 kb inserts)

Mapping
**Sequencing methods:**

**hierarchical sequencing (cntd)**

- genome
- BAC library *(100-200 kb inserts)*
- plasmid library *(1-2 kb inserts)*

---

**Sequencing methods:**

**whole genome sequencing (WGS)**

- genome
- BAC library *(100-200 kb inserts)*
- plasmid library *(1-2 kb inserts)*

---
**Sequencing methods:**
whole genome sequencing (WGS)

- **genome**
- **BAC library** (100-200 kb inserts)
- **plasmid library** (1-2 kb inserts)
- **contigs**
- **scaffolds**

**History of WGA**

- 1982: \(\lambda\)-virus, 48,502 bp
- 1995: h-influenzae, 1 Mbp
- 2000: fly, 100 Mbp
- 2001 - present
  - human (3Gbp), mouse (2.5Gbp), rat*, chicken, dog, chimpanzee, several fungal genomes
Deliverables (on human)

<table>
<thead>
<tr>
<th></th>
<th>NIHGR</th>
<th>WGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sequence</td>
<td>2.4 Gb</td>
<td>2.9 Gb</td>
</tr>
<tr>
<td>Coverage (sequence)</td>
<td></td>
<td>5.1X Celera</td>
</tr>
<tr>
<td>- pre-draft</td>
<td>2.5X</td>
<td>3X public (*)</td>
</tr>
<tr>
<td>- draft</td>
<td>12X</td>
<td></td>
</tr>
<tr>
<td>- finished</td>
<td>20-25X</td>
<td></td>
</tr>
<tr>
<td>No. of genes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- protein coding</td>
<td>~32,000</td>
<td>~35,500</td>
</tr>
<tr>
<td>- total</td>
<td>~32,000</td>
<td>~38,600 (**)</td>
</tr>
</tbody>
</table>

(*) shredded
(**) no mention of RNA genes

Venter et al. (2001) *Science* **291**: 1304-1351

Hierarchical sequencing (2) - chromosome walking
Advantages & Disadvantages of Sequencing Philosophies

Hierarchical Sequencing
- Starting time overhead (clone mapping step)
- Easier assembly
- Fewer gaps (and relatively easier to close)
- Sequence redundancy (due to inaccurate mapping)

Whole Genome Sequencing (WGS)
- No mapping
- More difficult assembly
- Generally more gaps (and more difficult to close)

Chromosome walking
- The advantages of clone-by-clone without the mapping step

Drosophila, human (Celera), mouse, dog
Yeast, worm, human (NIH-funded)
Rice

Sequencing: base calling
In the (not so) old days...

Fred Sanger method
1. Start at primer
2. Extend DNA (insert)
3. Include dideoxy-nucleotide to stop DNA synthesis
4. Separate products by length, using gel electrophoresis
5. Read manually

Nowadays...
1. Start at primer
2. Extend DNA (insert)
3. Include dideoxy-nucleotide with fluorochrome
4. Separate products by length, using capillary transfer
5. Read automatically as they run
Calling the bases

- Base calling: the first obstacle
  - First and last bases of low quality (noisy, unevenly spaced)
  - Low complexity regions: difficult to distinguish between bases

![Base calling example](image)

Phred: Phil's Read Editor

- Four phase procedure
  - Phase-1: idealized evenly spaced peak locations are predicted with simple Fourier analysis methods that examine the four traces in the regions surrounding each point
  - Phase-2: actual peak locations are located in the trace; this is done independently for the four traces, so there may be base overlap
  - Phase-3: observed and predicted peaks matches using dynamic programming
  - Phase-4: evaluation of the trace surrounding each called base using various 4-5 criteria (i.e., peak has the largest signal, greater than minsize, flanked by resolved peaks, etc)

- Base quality score = $-10 \times \log_{10}(P(\text{error}))$
Typical contig coverage

Lander-Waterman statistics

$L$ : read length  
$T$ : minimum detectable overlap  
$\sigma : 1 - T/L$

$G$ : genome size  
$N$ : number of reads  
$c$ : coverage ($NL / G$)

$E(#islands) = Ne^{-c\sigma}$

$E(island size) = L(e^{c\sigma} - 1) / c + 1 - \sigma$
Example - “experimental” data

<table>
<thead>
<tr>
<th>X coverage</th>
<th># ctgs</th>
<th>% &gt; 2X</th>
<th>avg ctg size (L-W)</th>
<th>max ctg size</th>
<th># ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>284</td>
<td>54</td>
<td>1,234 (1,138)</td>
<td>3,337</td>
<td>526</td>
</tr>
<tr>
<td>3</td>
<td>597</td>
<td>67</td>
<td>1,794 (4,429)</td>
<td>9,589</td>
<td>1,092</td>
</tr>
<tr>
<td>5</td>
<td>548</td>
<td>79</td>
<td>2,495 (21,791)</td>
<td>17,977</td>
<td>1,398</td>
</tr>
<tr>
<td>8</td>
<td>495</td>
<td>85</td>
<td>3,294 (302,545)</td>
<td>64,307</td>
<td>1,762</td>
</tr>
<tr>
<td>complete</td>
<td>1</td>
<td>100</td>
<td>1.26 M</td>
<td>1.26 M</td>
<td>1,329</td>
</tr>
</tbody>
</table>

Caveat: numbers based on artificially chopping up the genome of Wolbachia pipientis dMel
  • Physical gaps not considered
  • Fewer repeats

“Experimental” data (cntd)

sequence represented by continuous library

random reads

coverage

gap

sequence represented by discontinuous library

random reads

coverage

gap

sequence gap

library gap

Wendl, Barbazuk, BMC Bioinf, 2005, 6: 245
Genome sequencing - assembly

- Cloning
- Mapping/Sequencing
- Assembly
- Gap filling

Assembly paradigms

- Overlap-layout-consensus
  - greedy (TIGR Assembler, phrap, CAP3...)
  - graph-based (Celera Assembler, Arachne)

- Eulerian path (Sequencing by hybridization)
**Phred: Phil’s Revised Assembly Program**

- **Greedy algorithm**
  - Repeated bases at the begin/end of a sequence are removed
  - All pairwise matches of length > minlength are identified
  - Pick the largest scoring overlap and merge the two fragments
  - Repeat until no more merges can be done

- Uses the base quality scores from Phred

---

**Repeats**

- **Low-Complexity DNA** (e.g. ATATATATACATA...)

- **Microsatellite repeats** (e.g. CAGCAGTAGCAGCACCAG)
  - \((a_1...a_n)^y\) where \(k \sim 3-6\)

- **Transposons**
  - SINE (Short Interspersed Nuclear Elements)
    - e.g., ALU: ~300bp-long, \(10^6\) copies
  - LINE (Long Interspersed Nuclear Elements)
    - ~4000-long, 200,000 copies
  - LTR retroposons (Long Terminal Repeats (~700 bp) at each end)
    - Cousins of HIV

- **Gene Families**
  - Genes duplicate & then diverge (paralogs)

- **Recent duplications**
  - ~100,000bp-long, very similar copies

- **Genome content**:
  - Bacterial genomes: 5%
  - Mammals: 50%
Repeats, errors, and contig lengths

- Repeats shorter than read length are easily resolved
  - Read that spans across a repeat disambiguates order of flanking regions

- Repeats with bp differences > sequencing error rate are OK
  - Throw out overlaps between two reads in different copies of the repeat

- To make the genome appear less repetitive:
  - Increase read length
  - Decrease sequencing error rate

Role of error correction:
Discards up to 98% of single-base sequencing errors decreases error rate
⇒ decreases effective repeat content
⇒ increases contig length

Assembly algorithms

- Overlap-layout-consensus
  - greedy (phrap, TIGR Assembler, CAP3...)
  - graph-based (Celera Assembler, Arachne)

- Eulerian path (Sequencing by hybridization)
Methods

ARACHNE: A Whole-Genome Shotgun Assembler

Serafim Batzoglou,1,2,3 David B. Jaffe,2,3,4 Ken Stanley,2 Jonathan Butler,2
Sante Gnerre,2 Evan Mauceli,2 Bonnie Berger,1,5 Jill P. Mesirov,2 and Eric S.
Lander2,6,7

1Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;
2Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts 02141, USA; 4Department of
Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588, USA; 5Department of Mathematics,
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; 6Department of Biology, Massachusetts
Institute of Technology, Cambridge, Massachusetts 02139, USA

Genome Research 12:1 (2002), 177-189

ARACHNE - algorithm outline

- Input and trimming
- Overlap detection
- Error correction
- Evaluation of alignments
- Identification of paired pairs
- Contig assembly
- Identification of repeat contigs
- Creation of scaffolds
- Filling gaps in scaffolds
- Consensus computation
**ARACHNE - Trimming**

- find longest contiguous sequence with error less than 5% (use quality values)
- trim further if any base with Q<10 is within 12 bases of either end
- throw away read if length < 50 after trimming
- identify vector by aligning with *E. coli* and known cloning vector sequences
- remove vector from beginning and/or end of read

---

**ARACHNE - Overlap detection**

- 24-mer indexing
- index only 1/2 of all k-mers
  - for (x1,x2) where x1 is the reverse compl of x2, store whichever k-mer is alphabetically first
- exclude high-copy k-mers
- create read pairs for all reads that share one or more k-mers
Correcting errors in reads

Note: polymorphisms may confuse this scheme

Evaluation of alignments (pairs)

- Penalty (P) for each discrepant base is minimum of:
  - quality scores of the two aligned bases
  - quality scores of the two bases on their left and on their right

- Penalty score is then $10^{P/10}$
  - Related to the Phred score

- Discard pairs with penalty score > 100
**Merge paired reads**

- **Overlap graph:**
  - Nodes: reads $r_1$,...,$r_n$
  - Edges: overlaps ($r_i$, $r_j$, shift, orientation, score)

**Merge paired reads: implementation**

Reads that come from two regions of the genome (blue and red) that contain the same repeat.
**Contig assembly**

- Paired pairs form the initial contigs
- Next, mark repeat boundaries before doing further merging
- Only merge read pairs when they do not cross a repeat boundary

**Repeat detection**

![Diagram A](image1)

![Diagram B](image2)
Repeat detection (cntd)

We want to merge reads up to potential repeat boundaries

Scaffold formation and gap filling

these are usually repetitive contigs
ARACHNE2: improvements

- Scaffold breaking and re-joining introduced
- Gaps can now be filled by individual reads, not just by contigs
  - This is equivalent to the "stones" method in the Celera Assembler
- Memory usage was reduced fourfold

Celera assembler

A Whole-Genome Assembly of Drosophila

We report on the quality of a whole-genome assembly of Drosophila melanogaster and the nature of the computer algorithms that accomplished it. Three independent external data sources essentially agree with and support the assembly's sequence and ordering of contigs across the euchromatic portion of the genome. In addition, there are isolated contigs that we believe represent nonrepetitive packets within the heterochromatic portion of the centromeres. Comparison with a previously sequenced 2.9-megabase region indicates that sequencing accuracy within nonrepetitive segments is greater than 99.99% without manual curation. As such, this initial reconstruction of the Drosophila sequence should be of substantial value to the scientific community.
Celera assembler - overview

Celera assembler - sequence quality requirements

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Requested</th>
<th>Received</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read length / accuracy</td>
<td>500 bp @ 98% (min)</td>
<td>551 bp @ 98%</td>
</tr>
<tr>
<td>Shotgun coverage</td>
<td>10X (min)</td>
<td>14.6 X</td>
</tr>
<tr>
<td>Reads in pairs</td>
<td>70% (min)</td>
<td>72.8%</td>
</tr>
<tr>
<td>Insert length variance</td>
<td>±3% (max)</td>
<td>±10%</td>
</tr>
<tr>
<td>False-positive pairs</td>
<td>1% (max)</td>
<td>0.34%</td>
</tr>
<tr>
<td>BAC map coverage</td>
<td>15 X (min)</td>
<td>13.18 X</td>
</tr>
<tr>
<td>Ratio of 2 kb to 10 kb</td>
<td>4 to 1 (max)</td>
<td>1.32 to 1</td>
</tr>
</tbody>
</table>

(trimmed)
Assembler Principles

- Detect repeats and so avoid being misled by them, leave for the last.
- Make 1st order use of mate-pairs: first to circumnavigate and later to fill in repeats.
- Make all the sure moves first
  - tiered phases that get progressively more aggressive
  - output a complete audit trail of the evidence for assembly.

Assembler pipeline (updated 2006)

- Reads (typically 800bp) are quality-trimmed to 98% accuracy based on the quality score (avg trim length = 500-900bp; 590bp for human in year 2000)
- Remove sequences with *E. coli* matches
- Repeated screening makes run time and graph size reasonable (avoid $10^6$ Alu repeat comparisons)
Assembler pipeline (updated 2006)

- Finds overlaps ≥40 bp allowing 6% mismatch
- 22-mer “seed” matches with linear extension; stops when probability of seeing the observed number of errors is >10^{-6}
- Avoid 22-mers that occur too frequently
- Overlaps are re-evaluated as SNPs are “corrected”
**Assembler pipeline (updated 2006)**

Unique DNA unitig

Arrival Intervals

Repetitive DNA unitig

Discriminator Statistic is log-odds ratio of probability unitig is unique DNA versus 2-copy DNA.

-10 0 +10

Dist. For Repetitive

Dist. For Unique

Definitely Repetitive

Don’t Know

Definitely Unique

**Assembler pipeline (updated 2006)**

Trim & Screen

Overlapper

Unitiger

Scaffolder

Repeat Rez I, II

Mated reads

Benos 02-710/MSCBIO2070 7-FEB-2008

Benos 02-710/MSCBIO2070 7-FEB-2008
Assembler pipeline (updated 2006)

Trim & Screen

Overlapper

Unitiger

Scaffolder

Repeat Rez I, II

Unitig > 0

Genome sequencing - finishing

- Cloning
- Mapping/Sequencing
- Assembly
- Gap filling
For some reading...


Acknowledgements

Some of the slides used in this lecture are copied or modified from lectures of:

- Serafim Batzoglou, Stanford University
- Mihai Pop, University of Maryland
- Arthur L. Delcher, University of Maryland

Pictures from the following: