GENES & GENOME DATABASES

BME 110/BIOL 181
Computational Biology Tools
Prof. Todd Lowe
April 5, 2012
• Discuss Fun Quiz

• Readings:
  • Dummies Chapters 1, 2 (pp. 29-56), Ch 3; NYTimes piece on Jim Kent
  • Assigned Review:
    • "The impact of next-generation sequencing technology on genetics" by E. Mardis
  • Recommended Review:
    • "Application of ‘next-generation’ sequencing technologies to microbial genetics" by MacLean et al.
STUDY SECTIONS

• Study Sections:
  • Tues 6-7:15pm (Physical Sciences Building Rm 305)
  • To be decided (look like it will be Thursday 6-7:15pm)
GENES & GENOMES
THREE DOMAINS OF LIFE

WHAT IS A GENOME?

- A complete set of instructions for life encoded in DNA
- Organized in chromosomes
  - prokaryotes generally have one main circular chromosome
  - eukaryotes have multiple linear chromosomes
- Instructions are generally in the form of genes, and are the “unit” of heredity
WHY SEQUENCE A GENOME?

• We wish to understand how the entire cell / organism works
  • thousands of complex gene interactions
  • complete “parts list” is first step to understanding how parts work together as a whole

• Economy of scale – faster, cheaper to sequence all genes at once, than one at a time by many different researchers
First fully sequenced Organism: *H. Influenza*
The Institute for Genome Research (TIGR) - 1995
HUMAN GENOME

• In 1980’s, initial discussion to sequence human genome (first key meeting here at UCSC!)
  • Began: 1990
  • Planned finish: 2005

• Original Estimates:
  • ~100,000 genes
  • 3 billion nucleotides, projected cost $300 Million
  • less than 5% of genome codes for genes
  • Early opposition to sequencing 95% “junk”, taking money away from basic research

• Final Results:
  • Draft completed June 2000, “Finished” April 2003
  • Final cost ~ $3 billion
  • ~25,000-30,000 genes now
  • 5% of genome is conserved, likely important
PREPARING FOR THE HUMAN SEQUENCING...

- Mapping the human genome
- Practice: sequencing **model** organisms

<table>
<thead>
<tr>
<th>Model Species</th>
<th>Genome Size (Mb)</th>
<th>Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker’s Yeast (S. cerevisiae)</td>
<td>12.5</td>
<td>1996</td>
</tr>
<tr>
<td>E. coli</td>
<td>4.5</td>
<td>1997</td>
</tr>
<tr>
<td>Roundworm (C. elegans)</td>
<td>100</td>
<td>1998</td>
</tr>
<tr>
<td>Fruit fly (D. melanogaster)</td>
<td>160</td>
<td>2000</td>
</tr>
<tr>
<td>Mustard weed (A. thaliana)</td>
<td>120</td>
<td>2000</td>
</tr>
</tbody>
</table>
Sequencing Genomes: Strategy #1

- Original “Top-Down” Strategy
- Deliberate, small chance for major errors
WGS (Whole Genome Sequencing) – “Bottom Up” Strategy
- No ordering of a clone library - straight to sequencing to build “scaffolds”
- Current method used for genome sequencing
A new private company, Celera, announces it would use strategy #2 to beat the public, international effort to sequence the human genome, and planned to patent as many human genes as possible.

Public team effort changes strategy from #1 to mixture of #1 and #2.

Change in strategy meant ordered clone libraries would not be finished, and an alternate method was needed for putting the pieces together (quickly).

UC Santa Cruz to the rescue!

David Haussler & Jim Kent, in just a four weeks, created a clever program (GigAssembler) to run on a network of 100 Dell desktop computers to assemble the genome by Celera’s target date in June.

The whole story:

SEQUENCING TECHNOLOGIES

“Sanger Sequencing”

- Capillary sequencing: ABI 3700 and MegaBASE machines allow first rapid automated sequencing in tiny capillary tubes
- 600-800 bases at a time $\times$ 384 wells $\approx$ ~ 268,000 bases decoded / run
- Bulk of original human genome decoded on these

“NextGen” Technologies (Massively parallel, shorter reads)

- 454 / Roche Sequencers 100-500 bases, 400-600 Million bases / run
- Illumina Genetic Analyzer: ~2x100 bases, 150-200 Billion bases/ run
- ABI SOLiD Sequencing: ~2x50 base each, 80-100 Billion bases / run

Chemical decoding is no longer the major cost (10 Million bp / cent) – re-assembling all the pieces with computer programs and interpretation (bioinformatics) is now the major cost / bottleneck
GENOME CHALLENGES

• **$1,000 Genome Prize** – goal set in 2003 by J. Craig Venter Science Foundation to make personalized medicine available to all ($500,000 prize)

• **X Prize** – build a machine that can sequence 100 people in 10 days or less, for $10,000/genome or less ($10 Million prize): http://genomics.xprize.org/
PLUMMETING COST OF HUMAN GENOME SEQUENCING


Tuesday, April 10, 12