Genome sequencing

Prof. Dr. Hamdy M. El-Aref
Assiut University, Fac. of Agriculture
Genetics Department
Genome sequencing

DNA
the molecule of life

Trillions of cells
Each cell:
- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- 80,000 genes code for proteins that perform all life functions
DNA Cloning Overview

- Animal DNA is cut with BamHI restriction enzyme.
- Animal gene is ligated into the BamHI site of the plasmid.
- The plasmid is injected into the E. coli bacterium.
- Transformation occurs, resulting in a recombinant plasmid inside the bacterium.
- Many copies of the recombinant plasmid are produced within E. coli colonies.
Overview: Dideoxy (Sanger) Method

1. Denaturation (95–96 °C)
2. Annealing (50–55 °C)
3. Extension (60–70 °C)
4. PRODUCTS

- A
- AC
- ACC
- ACCG
- ACCGT
- ACCGTA
- ACCGTA

5. Gel electrophoresis
Sanger Method: Generating Read

1. Start at primer (restriction site)
2. Grow DNA chain
3. Include ddNTPs
4. Stops reaction at all possible points
5. Separate products by length, using gel electrophoresis
Automatic DNA sequencing

(A) ddA, ddC, ddT, ddG, ddNTPs – each with a different fluorescent label

Sequencing reactions, fractionation of products

Fluorescent bands move past the detector

(B) CACCGCAATCGAATAATTTAACTTTCCAAAAGTTAAGCCTTGG
Cloning vectors for genome sequencing

(A) ~1 kb insert
- Phagemid
- T7
- Sp6
- mcs
- Phage origin/packaging
- Amp-resistance
- High copy number plasmid vector

(B) ~2-10 kb insert
- Plasmid
- T7
- Sp6
- mcs
- Amp-resistance
- High copy number plasmid vector

(C) ~100 kb insert
- Packaging initiation
- High copy number plasmid vector
- mcs
- Phage lytic induction
- Kan-resistance
- P1 clone
- Amp-resistance
- cosN/LoxP sites for recombination

(D) < 300 kb insert
- BAC
- T7
- Sp6
- mcs
- Cm-resistance
- Single copy F plasmid
Sequencing vector M13

DNA insert

Recombinant M13 vector (double-stranded DNA)

Transfect E. coli

Protein coat

DNA core

Phages are released

Recombinant M13 phage

Single-stranded DNA
DNA Cloning and Sequencing with **BACs**

- Transformation
- Restriction + ligation
- Plasmid extraction
- Cellular screening
- Cell culture

*Laboratory Techniques*
## Sizes of inserts in sequencing vectors

<table>
<thead>
<tr>
<th>Vector</th>
<th>Size (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>100 Kb</td>
</tr>
<tr>
<td>YAC</td>
<td>300 - 1500 Kb</td>
</tr>
<tr>
<td>BAC</td>
<td>70 - 300 Kb</td>
</tr>
<tr>
<td>Cosmid</td>
<td>~ 40 Kb</td>
</tr>
<tr>
<td>Plasmid</td>
<td>2 - 10 Kb</td>
</tr>
<tr>
<td>M13 or Phagmid</td>
<td>~1 Kb</td>
</tr>
</tbody>
</table>
Genome Sequencing

- Isolate Chromosome
- Shear DNA into Fragments
- Clone into Seq. Vectors
- Sequence
Genome Sequencing Strategies

1. Clone-by-clone or hierarchical sequencing

2. Shotgun sequencing
Clone-by-clone or hierarchical sequencing strategy

- After constructing a complete physical map, clone by clone sequencing can be started in any specific region.
BAC to BAC Sequencing
Or BAC to Plasmid Sequencing

Whole Genome shotgun Sequencing

This step not needed in shotgun sequencing

© 2002 The Center for the Advancement of Genomics (TCAG).
Clone-by-clone or hierarchical sequencing strategy

**Advantages:**
1. Ability to fill gap and re-sequence the regions.
2. Ability to distribute the clones to other labs.
3. Ability to check the produced sequence by restriction enzymes

**Disadvantages:**
1. Expensive and time-consuming for construction of the physical map.
2. Experienced personnel are required.
VENTER’S SHOTGUN
Whole genome shotgun sequencing

• It is a much faster approach, and enabled researchers to speed up the timetable for sequencing enormously.

• The shotgun method was developed by J. Craig Venter et al., 1996.
Shotgun Sequencing

cut many times at random (Shotgun)

Get one or two reads from each segment

~500 bp
**Shotgun Sequencing**

- Used to sequence whole genomes
- Steps:
  - DNA is broken up randomly into smaller fragments
  - Dideoxy method produces reads
  - Look for overlap of reads
Fragment Assembly

Overlap reads and extend to reconstruct the original genomic region
Problems with the shotgun approach.

(A) Problems with tandemly repeated DNA

(Tandem repeat)

-DNA

Fragments

-Sequences

GATTAGATTA

GATTAGATTAGATTA

Incorrect overlap

(B) Problems with genome-wide repeats

(Two genome-wide repeats)

-DNA

Fragments

-Sequences

GCATAGCT...

...GCATAGCT

Incorrect overlap
Shotgun Sequencing

- Very efficient process for small-scale (~10 kb) sequencing (preferred method)
- First applied to whole genome sequencing in 1995 (*H. influenzae*)
- Now standard for all prokaryotic genome sequencing projects
- Successfully applied to *D. melanogaster*
- Moderately successful for *H. sapiens and other genotypes* …
Comparison of two sequencing methods

Hierarchical sequencing
Chromosomes

- Generate and align large BAC or P1 clones
- Fragment and sequence a subset of the clones

Contig assembly and bioinformatics analysis

Shotgun sequencing

- Fragment and sequence entire genome
Sequencing Successes

T7 bacteriophage
completed in 1983
39,937 bp, 59 coded proteins

*Escherichia coli*
completed in 1998
4,639,221 bp, 4293 ORFs

*Saccharomyces cerevisiae*
completed in 1996
12,069,252 bp, 5800 genes
Sequencing Successes

*Drosophila melanogaster*
completed in 2000
116,117,226 bp, 13,601 genes

Homo sapiens
completed in 2003
3,201,762,515 bp, 31,780 genes

Oryza Sativa (Rice)
430 million base
60,000 genes
Genomes to Date

- 8 vertebrates (human, mouse, rat, fugu, zebrafish)
- 2 plants (arabadopsis, rice)
- 2 insects (fruit fly, mosquito)
- 2 nematodes (C. elegans, C. briggsae)
- 1 sea squirt
- 4 parasites (plasmodium, guillardia)
- 4 fungi (S. cerevisae, S. pombe)
- 200 bacteria and archebacteria
- 1900+ viruses
Sequenced Genomes

http://www.genomeneuwsnetwork.org/
RNA Alternative Splicing and Gene Function in Eukaryotes
Amylase = 2
Tropomiosine = 10
Troponin-T = 64

Alternative splicing: Different mRNAs from same pre-mRNA
Thank you

Prof. Dr. Hamdy El-Aref
Sequencing strategies

- Whole genome
Rapid re-sequencing of human Ad1: Time trial

- Have sequence of Ad 1.
- In theory, have a minimally tiled set of PCR primers to cover entire 36,001 base genome.
- In theory, have a minimally tiled set of sequencing primers as well.
- Want draft sequence in a minimal time, including primer delivery from a vendor.
- In practice design two parallel sets of minimally tiled PCR primers and amplify two sets.
- In practice, assume 750 base reads--> 48 primers, one direction.
- Compare with consensus: Determine accuracy, timing and evaluate operation.