Computational Biology 2

Pawan Dhar
BII
Lecture 2

Human Genome Sequencing

An overview
PCR

NEXT CYCLE

MELT DNA
94°C

EXTENSION
72°C

PRIMER BINDING
53°C

REVERSE PRIMER
FORWARD PRIMER

TAQ
Restriction enzyme digestion
Partial digestion
Food for thought

To create a genomic library, typically one does not use a 4 base cutter or a rare base cutter restriction enzyme. Why?
Gel electrophoresis apparatus
Equipment used for gel imaging
How to read the DNA gels - 1
How to read the DNA gels - 2
DNA electrophoresis

Recommended agarose concentrations for resolution of linear dsDNA

<table>
<thead>
<tr>
<th>Agarose (%)</th>
<th>Size of fragments separated (kb)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1–30</td>
</tr>
<tr>
<td>0.7</td>
<td>0.8–12</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5–10</td>
</tr>
<tr>
<td>1.2</td>
<td>0.4–7</td>
</tr>
<tr>
<td>1.5</td>
<td>0.2–3</td>
</tr>
<tr>
<td>2.0</td>
<td>0.05–2</td>
</tr>
</tbody>
</table>

*DNA marker (Marker A) migrates approximately at 300 kb and allele typical at 4 kb, independent of agarose concentration between 0.5% and 1.4% in TAE buffer.

DNA markers for gel electrophoresis (bp)

<table>
<thead>
<tr>
<th>λ HindIII</th>
<th>λ HindIII–EcoRI</th>
<th>λ EcoRI</th>
<th>OX174–HindIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>23,130</td>
<td>21,226</td>
<td>21,226</td>
<td>1,353</td>
</tr>
<tr>
<td>9,416</td>
<td>7,421</td>
<td>1,078</td>
<td></td>
</tr>
<tr>
<td>6,557</td>
<td>5,804</td>
<td>4,973</td>
<td>872</td>
</tr>
<tr>
<td>4,361</td>
<td>5,643</td>
<td>4,785</td>
<td>603</td>
</tr>
<tr>
<td>2,322</td>
<td>3,550</td>
<td>2,027</td>
<td>281</td>
</tr>
<tr>
<td>2,027</td>
<td>2,027</td>
<td>2,027</td>
<td></td>
</tr>
<tr>
<td>566</td>
<td>1,904</td>
<td>1,904</td>
<td>271</td>
</tr>
<tr>
<td>125</td>
<td>1,584</td>
<td>1,584</td>
<td>234</td>
</tr>
<tr>
<td>125</td>
<td>1,584</td>
<td>1,584</td>
<td>234</td>
</tr>
<tr>
<td>1,375</td>
<td>1,375</td>
<td>1,375</td>
<td>194</td>
</tr>
<tr>
<td>947</td>
<td>947</td>
<td>947</td>
<td>118</td>
</tr>
<tr>
<td>831</td>
<td>831</td>
<td>831</td>
<td>72</td>
</tr>
<tr>
<td>566</td>
<td>566</td>
<td>566</td>
<td>566</td>
</tr>
<tr>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
</tr>
</tbody>
</table>
1. Why does an uncut plasmid give 3 bands on the gel?
2. What is the molecular weight of these bands?

- Relaxed circle
- Full length linear
- Supercoiled

5.2 kb
DNA Gels under uv light

Ethidium bromide stained gels

Gel photograph on thermal paper
DNA sequencing

Manual Automatic

Final output in human / machine readable format

A = Adenine, T = Thymine
G = Guanine, C = Cytosine
Shotgun sequencing

Hierarchical

Whole genome
1. Shear genome into 1 Kb long fragments
2. Subclone fragments into universal cloning vector
3. Sample library of fragments
4. Use bioinformatics approaches to align the fragments
Whole Genome Shotgun Sequencing Method

Genomic DNA

Sequence Each Fragment with Shotgun Approach

Align Contiguous Sequences

Generate Finished Sequence
Special Human Genome issues

Feb 2001
Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

Eric Lander
the lead author
of Nature paper
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>one of the two (or more) forms of a gene</td>
</tr>
<tr>
<td>BAC</td>
<td>an artificial chromosome vector that carries gene to be cloned</td>
</tr>
<tr>
<td>Cloning</td>
<td>the process of generating sufficient copies of DNA</td>
</tr>
<tr>
<td>cDNA</td>
<td>a DNA sequence made from mRNA</td>
</tr>
<tr>
<td>Conserved</td>
<td>Similarity of the two sequences at RNA, DNA and protein level in two distinct organisms</td>
</tr>
<tr>
<td><strong>Genome</strong></td>
<td>The complete DNA sequence of an organism</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td><strong>Genotype</strong></td>
<td>total set of genes, usually refers to a pair of alleles that a person has at a given locus</td>
</tr>
<tr>
<td><strong>Haplotype</strong></td>
<td>A particular combination of alleles or sequence variations that are closely linked and therefore inherited together</td>
</tr>
<tr>
<td><strong>Heterochromatin</strong></td>
<td>Compact gene poor regions of chromosomes full of repeat sequences, very difficult to clone</td>
</tr>
<tr>
<td><strong>Poly-morphism</strong></td>
<td>A region of the genome that varies between individual members of a population, present in significant number of people</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Polymorphism</td>
<td>A region of the genome that varies between individual members of the population. To be called polymorphic the variant should be present in significant numbers of people in the population.</td>
</tr>
<tr>
<td>Proteome</td>
<td>The complete set of proteins produced by the cell.</td>
</tr>
<tr>
<td>Splicing</td>
<td>Removal of introns from transcribed RNAs. May also remove exons giving rise to different proteins (splice variants).</td>
</tr>
<tr>
<td>Transcriptome</td>
<td>The complete set of RNAs transcribed from a genome.</td>
</tr>
<tr>
<td><strong>Raw Sequence</strong></td>
<td>Individual unassembled sequence</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td><strong>Coverage (depth)</strong></td>
<td>The average number of times a nucleotide is represented by a high quality base in a collection of raw sequence. Corresponding to a PHRED score of at least 20.</td>
</tr>
<tr>
<td><strong>Finished clone</strong></td>
<td>A large insert clone for which full shotgun sequence has been produced.</td>
</tr>
<tr>
<td><strong>Contig</strong></td>
<td>The result of joining an overlapping collection of sequences or clones.</td>
</tr>
<tr>
<td><strong>Genome glossary (5/5)</strong></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Scaffold</strong></td>
<td>The result of connecting contigs by linking information from paired-end reads from plasmids / paired end reads from BACs / known m RNA etc. The contigs in the scaffold are ordered and oriented with respect to each other.</td>
</tr>
<tr>
<td><strong>GigAssembler</strong></td>
<td>A computer program that merges individual sequenced clones into a draft genomic sequence.</td>
</tr>
<tr>
<td><strong>Public Sequence Database</strong></td>
<td>Three international sequence databases: GenBank, EMBL and DDBJ.</td>
</tr>
</tbody>
</table>
Genome gurus

* Fred Sanger: Basic sequencing technique

* Leroy Hood: Automated sequencers

* Michael Hunkapiller: ABI PRISM sequence Analyser

* Philip Green: Computer programs for rapid assembly of fragments into complete genome sequence
The Human Genome timeline

1985: Charles DeLisi begins initiates a discussion on a mammoth project to sequence complete human genome

1988: NIH establishes Office of Human Genome Research, renamed later as NCHGR, headed by James Watson

Early 1990s: Sequencing project very expensive, adopts map first sequence later approach, strategy

1992: Francis Collins replaces Watson as Head of NCHGR
The Human Genome timeline

1992: Craig Venter sets up TIGR. First one to sequence M.genitalium

1996: Bermuda conference. Scientists agree to formalize the conditions of data access

1998: Craig Venter launches Celera to sequence human genome with the shotgun method, did not follow Bermuda principles

1999: Scientists respond to Venter’s challenge, launch G5!!

G5 - Whitehead Institute for Biomedical Research, Sanger Centre, Baylor College of Medicine, Washington University, Joint Genome Institute
The Human Genome timeline

1999 - 2000: First Complete genome sequence of #22 published in Nature (German - Japanese collaboration)

2000: Craig Venter and Francis Collins together announce completion of the ‘working draft’ of human genome project
Human genome centres worldwide

Leaders of G16 centres