Techniques for Genome Mapping & Sequencing

Eric D. Green, M.D., Ph.D.
National Human Genome Research Institute
Phone: 301-402-2023
FAX: 301-402-2040
E-Mail: egreen@nhgri.nih.gov
Outline

I. Fundamentals of Genome Mapping

II. Fundamentals of Genome Sequencing

III. Mapping & Sequencing in the Human Genome Project... and Beyond

IV. Comparative Sequencing

Genome Sizes

Human Genome
Mouse Genome

~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

~100,000,000 bp

Yeast Genome

~15,000,000 bp

E. coli Genome

~5,000,000 bp
The Human Cytogenetic Map

Genetic Map

Cytogenetic Map

Physical Map

RH Map  Clone-Based Map  Sequence Map

...GATCTGCTA
TACTACCGC
ATTATTCG...

25 50 75 100 125 150 Mb

20 30 30 20 25 cM
Sequence-Tagged Sites (STSs)

100 kb

STS1  STS2  STS3  STS4  STS5

//  //  //  //  //

PCR Primer 1

GGATTCCGACTAGTGGTCTT... // ...GGATTAGCTAGGTATTGGCTAT
CCTAAGGCTGATCCAGCCAGAA... // ...CCTAATCGATCCATAACCGATA

PCR Primer 2

~60-1000 bp

Physical Mapping: General Principles

- Importance of Physical Maps:
  Localization and Isolation of Genes (e.g., Positional Cloning)
  Study of Genome Organization and Evolution
  Framework for Genome Sequencing

- Physical Mapping Involves Ordering Clones and/or Landmarks

- General Types of Physical Maps:
  Landmark Only (e.g., Radiation Hybrid Maps)
  Clone-Based
  Sequence
Clone-Based Physical Mapping

Construction of YACs and BACs

Green et al. (1998)  
Birren et al. (1998)

High-Molecular Weight DNA  
Partial Restriction Digestion

YAC Insert: ~100-1000 kb  
BAC Insert: ~100-200 kb
Cloning Capacity

- **YAC**
  - ~1,000,000 bp
- **BAC**
  - ~100,000 bp
- **Cosmid**
  - ~45,000 bp
- **Bacteriophage**
  - ~25,000 bp

Genome Sizes

- **Human**
  - ~3,000,000,000 bp
- **Mouse**
  - ~160,000,000 bp
- **Fruit Fly**
  - ~100,000,000 bp
- **Nematode**
  - ~25,000,000 bp
- **Yeast**
  - ~10,000,000 bp
- **E. coli**
  - ~5,000,000 bp

**Bacterial Artificial Chromosomes (BACs)**

- Bacterial-Based Cloning System Developed by Shizuya et al. (1992)
- Based on the *E. coli* F Factor (Fertility Plasmid): Replication Control
- Cloned Inserts: 100-200 kb, Circular DNA
- Low Copy Number
  - Low Yields of DNA by Standard Methods
  - Reasonably Stable
- Relatively Non-Chimeric
- BAC Libraries from Many Different Species now Available
  (e.g., www.chori.org/bacpac)
- See Birren et al. (1998)
Eric D. Green, M.D., Ph.D.
Genome Mapping & Sequencing

Chromosome (~130 Mb)

YAC (~0.5-1.0 Mb)

BAC (~0.1-0.2 Mb)

Genome (~3000 Mb)

Sequence-Ready Contig Map

Marra et al. (1997) and Gregory et al. (1997)
Physical Mapping: Future Prospects

- Strategies for Physical Mapping are Radically Changing in the Sequence-Based Era

- Will Now See a Closer Interplay of Mapping and Sequencing in the Exploration of New Genomes

- Construction of New BAC Libraries will Allow Physical Mapping Studies of More Species’ Genomes

DNA Sequencing
History of DNA Sequencing

- 1870: Miescher discovers DNA
- 1940: Avery proposes DNA as 'Genetic Material'
- 1953: Watson & Crick: Double Helix Structure of DNA
- 1965: Holley: Sequences Yeast tRNA_{Ala}
- 1977: Wu: Sequences λ Cohesive End DNA
- 1980: Sanger: Dideoxy Chain Termination
- 1986: Gilbert: Chemical Degradation
- 1980: Messing: M13 Cloning
- 1990: Hood et al.: Partial Automation
- 2005: • Cycle Sequencing
- • Improved Sequencing Enzymes
- • Improved Fluorescent Detection Schemes

Efficiency (bp/person/year)

- 1
- 15
- 150
- 1,500
- 15,000
- 25,000
- 50,000
- 200,000
- >100,000,000

Adapted from Messing & Llaca, PNAS (1998)

DNA Tagged with Radioactivity

G: G Reaction
A: A Reaction
T: T Reaction
C: C Reaction
Radioactive Sequencing

Fluorescent DNA Sequencing

Wilson & Mardis (1997)
Detection of Fluorescently Tagged DNA

DNA Fragments Separated by Electrophoresis

Optical Detection System

Output to Computer

Laser Excites Fluorescent Dyes

Analyzing Fluorescent DNA Sequencing Data

Computer Analysis

A G T A C T G G A T C
Fluorescent DNA Sequencing Results

Slab Gel-Based DNA Sequencing Instruments
Capillary-Based DNA Sequencing Instruments

Large-Scale cDNA Sequencing

• ESTs: Expressed-Sequence Tags

• SAGE: Serial Analysis of Gene Expression

• Full-Insert (Full-Length) cDNA Sequencing

mgc.nci.nih.gov
Large-Scale Genomic Sequencing

Shotgun Sequencing

Wilson & Mardis (1997)
Green (2001)

Subclone Construction

BAC DNA

Prepare Multiple Copies

Randomly Fragment

Subclone Fragments
Shotgun Sequencing Strategy

Poisson Calculations

The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution.

The probability a base is not sequenced is given by:

\[ P_0 = e^{-c} \]

Where:
- \( c \) = fold sequence coverage \((c=LN/G)\),
- \( LN \) = # bases sequenced, i.e. \( L \) = average sequencing read length and \( N \) = # reads,
- \( G \) = target sequence length,
- \( e \approx 2.718 \) (where \( e \approx 2.718281828459 \))

<table>
<thead>
<tr>
<th>Fold Coverage</th>
<th>( P_0 \times 10^{-c} )</th>
<th>% not sequenced</th>
<th>% sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.37</td>
<td>37%</td>
<td>63%</td>
</tr>
<tr>
<td>2</td>
<td>0.135</td>
<td>13.5%</td>
<td>87.5%</td>
</tr>
<tr>
<td>3</td>
<td>0.05</td>
<td>5%</td>
<td>95%</td>
</tr>
<tr>
<td>4</td>
<td>0.018</td>
<td>1.8%</td>
<td>98.2%</td>
</tr>
<tr>
<td>5</td>
<td>0.0067</td>
<td>0.6%</td>
<td>99.4%</td>
</tr>
<tr>
<td>6</td>
<td>0.0025</td>
<td>0.25%</td>
<td>99.75%</td>
</tr>
<tr>
<td>7</td>
<td>0.0009</td>
<td>0.09%</td>
<td>99.91%</td>
</tr>
<tr>
<td>8</td>
<td>0.0003</td>
<td>0.03%</td>
<td>99.97%</td>
</tr>
<tr>
<td>9</td>
<td>0.0001</td>
<td>0.01%</td>
<td>99.99%</td>
</tr>
<tr>
<td>10</td>
<td>0.000045</td>
<td>0.005%</td>
<td>99.995%</td>
</tr>
</tbody>
</table>
Shotgun Sequence Assembly

“Consed” (Gordon et al., 1998)
Shotgun Sequencing Strategy

Sequence Finishing: Resolving Ambiguities

*** Sequence Finishing: Remains Relatively Expensive ***
Historically Significant Genome Sequencing Projects

Bacterial Genome Sequences

Welcome to the Comprehensive Microbial Resource (CMR) Home Page

Funded by: The U.S. Department of Energy and The National Science Foundation

The Comprehensive Microbial Resource (CMR) is a tool that allows the researcher to access all of the bacterial genome sequences completed to date. For each genome not sequenced at TIGR, two kinds of annotation are displayed: the primary annotation taken from the genome sequencing center and the TIGR annotation generated by an automated annotation process at TIGR. Use the CMR to access information on all of the bacterial genomes or any subset of them.


www.tigr.org
First Eukaryotic Genome Sequence


First Animal Genome Sequence

Second Animal Genome Sequence

Science 287:2185-2195, 2000

Revised Timetable for Human Genome Sequencing
February, 2001 Publications

BAC-by-BAC Shotgun Sequencing

Green (2001)
Whole-Genome Shotgun Sequencing

Whole-Genome Shotgun Sequence Assembly
April, 2003 Completion

International Human Genome Sequencing Consortium

- 6 Countries
- 20 Sequencing Centers
- 1000’s of Individuals
- ~1,000 bases per second, 24 hours per day, 7 days per week
S. CON. RES. 10

Designating April 2003 as “Human Genome Month” and April 25 as “DNA Day”;

IN THE SENATE OF THE UNITED STATES
February 27, 2003

Mr. Gregg (for himself, Mr. Kennedy, Ms. Snowe, and Mr. Daschle) submitted the following concurrent resolution; which was considered and agreed to

CONCURRENT RESOLUTION
Designating April 2003 as “Human Genome Month” and April 25 as “DNA Day”.
October, 2004 Publication

Nature 431:931-945, 2004
All of the original goals of the Human Genome Project have been accomplished!

What’s Next?
Mapping the Human Genome  ~1990 to ~2000

Sequencing the Human Genome  ~1998 to ~2003

Interpreting the Human Genome Sequence  ~2003 to ???

Beyond The Human Genome Project
~3,000 bp (0.0001%) of Human Genome Sequence
Comparative Sequence Analysis

Using the Experiments of Evolution to Decode the Human Genome

Comparing Genomes is Like Cryptography
Functional Elements: Coding vs. Non-Coding

- Coding Sequences (i.e., Genes)
  - Relatively EASY to Identify
  - Mostly Know What to Look For
  - Complementary Data Sets Available (ESTs, cDNAs)
  - Ever-Improving Computational Gene Predictions

- Non-Coding Functional Sequences
  - HARD to Identify
  - Very Little Known About What to Look For
  - Virtually No Complementary Data Sets Available
  - Poor Computational Predictions

Major role for comparative sequence analysis will be the identification of functionally important, non-coding sequences

Hybrid Shotgun Sequencing

Green (2001)
What is the optimal mixture???

- BAC-by-BAC Shotgun Sequence Reads
- Whole-Genome Shotgun Sequence Reads
- Redundant Coverage: ~10-fold ~5-fold 0

Nature 428:493-521, 2004
Nature 420:520-562, 2002

Initial sequencing and comparative analysis of the mouse genome
Genome sequence of the Brown Norway rat yields insights into mammalian evolution
Human-Rodent Sequence Comparisons

- ~40% in Alignments
- ~5% Under Selection
- ~1.5% Protein Coding
- ~3.5% Non-Coding

Multi-Species Sequence Comparisons
Multi-Species Comparative Sequence Analysis

• Targeted Genomic Regions
• BAC-Based Sequencing in Multiple Vertebrates
• Identify Highly Conserved Non-Coding Sequences
• Conserved Sequences Correlate with Functional Elements

Thomas et al. (2003)
Additional Vertebrate Genome Sequencing Efforts

- Chimpanzee
- Macaque
- Dog
- Cow
- Monodelphis

Chicken, Xenopus, Zebrafish, Pufferfish

Low-Redundancy Sequencing of Multiple Vertebrate Genomes

Margulies et al., *PNAS*, in press, 2005
Future Genomes to Sequence???

ENCODE Project

- **ENCODE**: ENCyclopedia Of DNA Elements
- Goal: Compile a comprehensive encyclopedia of all functional elements in the human genome
- Initial pilot project: 1% of human genome
- Apply multiple approaches to study and analyze that 1% in a consortium fashion
Current Big Challenges…

- Defining “Saturation Points” in Terms of Information Gained by Comparative Sequence Analyses

- The “$1000 Genome”

- Medical Sequencing (aka, Human Re-Sequencing)
Bibliography


