The Human Genome Project

Doug Brutlag
Professor Emeritus
Biochemistry & Medicine (by courtesy)
Maeve’s Office Hours

• Monday and Friday 3:30 to 5:00 PM
• Beckman Center B403A
• Please Email Maeve to set up an appointment if you want to meet with her.
  - maeveo@stanford.edu
• Take the elevator to the fourth floor of Beckman and turn left. The hallway leads directly to the lab B403 and my office is inside the lab at B403A.
Current Topics in Genome Analysis 2010
A lecture series covering contemporary areas in genomics and bioinformatics
January 12 - March 23, 2010

Course Syllabus and Handouts
All lectures are on Tuesday mornings from 10:00 am to 11:30 am. Lectures are held in the Lipsett Amphitheatre, NIH Clinical Center (Building 10).

All handouts are in PDF format. To view, download the free Adobe Acrobat Reader.

January 12
The Genomic Landscape circa 2010
Eric Green, NHGRI

January 19
Biological Sequence Analysis I
Andy Baxevanis, NHGRI

January 26
Biological Sequence Analysis II
Andy Baxevanis, NHGRI

February 2
Mining Data from Genome Browsers
Tyra Wolfsberg, NHGRI

February 9
Next-Generation Sequencing Technologies
Elliott Margulies, NHGRI

February 16
Large-Scale Expression Analysis
Paul Meltzer, NCI

February 23
Regulatory and Epigenetic Landscapes of Mammalian Genomes
Laura Elnitski, NHGRI

March 2
Introduction to Population Genetics
Lynn Jorde, University of Utah

March 9
Genome-Wide Association Studies
Karen Mohlke, University of North Carolina

March 16
Genomics of Microbes and Microbiomes
Julie Segre, NHGRI

March 23
Pharmacogenomics
Howard McLeod, University of North Carolina
Hierarchical Sequencing vs. Whole Genome Shotgun Sequencing

from Gibson & Muse, A Primer of Genome Science
http://www.sinauer.com/genomics/
The Human Genome Project: How should we do it?

  - Use multiple length clones 2 kb, 10 kb and 50 kb
  - Use clone end sequencing generating mate-pairs
  - Able to use long clones to leap over repeated regions
  - Clone length permits one to measure length of repeated regions.
  - Will find more polymorphisms (SNPs)
  - Costs less
  - BAC clone artifacts
    - Differential amplification
    - BACs not stable in bacteria will be lost.
    - Repeated regions will recombine and be lost

  - Preferred clone-by-clone BAC sequencing
  - Distributed versus monolithic organization
  - BACs linked to genetic maps
  - Costs less (sequence 4x human genome)
  - Finishing simplified and fewer gaps
  - Haplotyping automatic
  - Longer repeat regions lengths measured
History of Whole Genome Assembly

1997

Gene Myers

Let's sequence the human genome with the shotgun strategy

That is impossible, and a bad idea anyway

Phil Green

Thanks to Serafim Batzoglou
Public Human Genome Project Strategy
http://www.genome.gov/
Contig Formation as Mapping Progresses
Lander & Waterman 1988


**MATHEMATICAL ANALYSIS OF RANDOM CLONE FINGERPRINTING**

![Graph showing expected number of islands](image)

- **Expected Number of Islands (in units of GL)**
- **Genome Equivalents Fingerprinted**

Graph labels:
- $\theta = 75\%$
- $\theta = 50\%$
- $\theta = 25\%$
- $\theta = 0\%$

Mathematical equations:

\[ G = \text{haploid genome length in bp}; \]
\[ L = \text{length of clone insert in bp}; \]
\[ N = \text{number of clones fingerprinted}; \]
\[ \alpha = \frac{N}{G} = \text{probability per base of starting a new clone}; \]
\[ T = \text{amount of overlap in base pairs needed to detect overlap}; \]
\[ \theta = \frac{T}{L}; \]
\[ c = \text{redundancy of coverage} = \frac{LN}{G}. \]
Public Genome Assembly Process

Fingerprint clone contig

Pick clones for sequencing
### Table 1 Key large-insert genome-wide libraries

<table>
<thead>
<tr>
<th>Library name*</th>
<th>GenBank abbreviation</th>
<th>Vector type</th>
<th>Source DNA</th>
<th>Library segment or plate numbers</th>
<th>Enzyme digest</th>
<th>Average insert size (kb)</th>
<th>Total number of clones in library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caltech B</td>
<td>CTB</td>
<td>BAC</td>
<td>987SK cells</td>
<td>All</td>
<td>HindIII</td>
<td>120</td>
<td>74,496</td>
</tr>
<tr>
<td>Caltech C</td>
<td>CTC</td>
<td>BAC</td>
<td>Human sperm</td>
<td>All</td>
<td>HindIII</td>
<td>125</td>
<td>263,040</td>
</tr>
<tr>
<td>Caltech D1 (CITB-H1)</td>
<td>CTD</td>
<td>BAC</td>
<td>Human sperm</td>
<td>All</td>
<td>HindIII</td>
<td>129</td>
<td>162,432</td>
</tr>
<tr>
<td>Caltech D2 (CITB-E1)</td>
<td></td>
<td>BAC</td>
<td>Human sperm</td>
<td>All</td>
<td>EcoRI</td>
<td>202</td>
<td>24,960</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2,501–2,565</td>
<td>EcoRI</td>
<td>182</td>
<td>46,326</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2,566–2,671</td>
<td>EcoRI</td>
<td>142</td>
<td>97,536</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3,000–3,253</td>
<td>EcoRI</td>
<td>142</td>
<td>97,536</td>
</tr>
<tr>
<td>RPCI-1</td>
<td>RP1</td>
<td>PAC</td>
<td>Male, blood</td>
<td>All</td>
<td>Mbol</td>
<td>110</td>
<td>115,200</td>
</tr>
<tr>
<td>RPCI-3</td>
<td>RP3</td>
<td>PAC</td>
<td>Male, blood</td>
<td>All</td>
<td>Mbol</td>
<td>115</td>
<td>75,513</td>
</tr>
<tr>
<td>RPCI-4</td>
<td>RP4</td>
<td>PAC</td>
<td>Male, blood</td>
<td>All</td>
<td>Mbol</td>
<td>116</td>
<td>105,251</td>
</tr>
<tr>
<td>RPCI-5</td>
<td>RP5</td>
<td>PAC</td>
<td>Male, blood</td>
<td>All</td>
<td>Mbol</td>
<td>115</td>
<td>142,773</td>
</tr>
<tr>
<td>RPCI-11</td>
<td>RP11</td>
<td>BAC</td>
<td>Male, blood</td>
<td>All</td>
<td></td>
<td>178</td>
<td>543,797</td>
</tr>
</tbody>
</table>

Total of top 5: 1,482,502
**Table 2 Total genome sequence from the collection of sequenced clones, by sequence status**

<table>
<thead>
<tr>
<th>Sequence status</th>
<th>Number of clones</th>
<th>Total clone length (Mb)</th>
<th>Average number of sequence reads per kb*</th>
<th>Average sequence depth†</th>
<th>Total amount of raw sequence (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finished</td>
<td>8,277</td>
<td>897</td>
<td>20–25</td>
<td>8–12</td>
<td>9,085</td>
</tr>
<tr>
<td>Draft</td>
<td>18,969</td>
<td>3,097</td>
<td>12</td>
<td>4.5</td>
<td>13,395</td>
</tr>
<tr>
<td>Predraft</td>
<td>2,052</td>
<td>267</td>
<td>6</td>
<td>2.5</td>
<td>667</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>23,147</strong></td>
</tr>
</tbody>
</table>

*The average number of reads per kb was estimated based on information provided by each sequencing centre. This number differed among sequencing centres, based on the actual protocols used.*

†The average depth in high quality bases (≥99% accuracy) was estimated from information provided by each sequencing centre. The average varies among the centres, and the number may vary considerably for clones with the same sequencing status. For draft clones in the public databases (keyword: HTGS_draft), the number can be computed from the quality scores listed in the database entry.
Comparing Chromosome 2 Sequence Versus Genetic Maps


Figure 5 Positions of markers on previous maps of the genome (the Genethon\textsuperscript{101} genetic map and Marshfield genetic map (http://research.marshfieldclinic.org/genetics/genotyping_service/mgsver2.htm), the GeneMap\textsuperscript{99} radiation hybrid map\textsuperscript{100}, and the Whitehead YAC and radiation hybrid map\textsuperscript{26}) plotted against their derived position on the draft sequence for chromosome 2. The horizontal units are Mb but the vertical units of
Synteny Between Human and Mouse

Table 1. Celera-generated data input into assembly.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Number of reads for different insert libraries</th>
<th>Total number of base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 kbp</td>
<td>10 kbp</td>
</tr>
<tr>
<td>No. of sequencing reads</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>11,736,757</td>
<td>7,467,755</td>
</tr>
<tr>
<td>C</td>
<td>853,819</td>
<td>881,290</td>
</tr>
<tr>
<td>D</td>
<td>952,523</td>
<td>1,046,815</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>1,498,607</td>
</tr>
<tr>
<td>Total</td>
<td>13,543,099</td>
<td>10,894,467</td>
</tr>
<tr>
<td>Fold sequence coverage (2.9-Gb genome)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>2.20</td>
<td>1.40</td>
</tr>
<tr>
<td>C</td>
<td>0.16</td>
<td>1.17</td>
</tr>
<tr>
<td>D</td>
<td>0.18</td>
<td>0.20</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0.28</td>
</tr>
<tr>
<td>Total</td>
<td>2.54</td>
<td>2.04</td>
</tr>
<tr>
<td>Fold clone coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>2.96</td>
<td>11.26</td>
</tr>
<tr>
<td>C</td>
<td>0.22</td>
<td>1.33</td>
</tr>
<tr>
<td>D</td>
<td>0.24</td>
<td>1.58</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>2.26</td>
</tr>
<tr>
<td>Total</td>
<td>3.42</td>
<td>16.43</td>
</tr>
</tbody>
</table>

Insert size* (mean)

| Average | 1,951 bp | 10,800 bp | 50,715 bp |

Insert size* (SD)

| Average | 6.10% | 8.10% | 14.90% |

% Mates†

| Average | 74.50 | 80.80 | 75.60 |

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.
Celera Scaffolds

http://www.sciencemag.org/cgi/content/full/291/5507/1304

Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.
Celera Assembler

5.11X Celera Reads
39X mate pairs

5.11X Celera Reads
39X mate pairs

Public Bactigs
(from 33,421 BACs)

Shredder

Matcher

Bactigs & Celera pairs
(binned by BAC)

Celera-unique reads

WGA

WGA

WGA

WGA

WGA+Shredder

Components₁

Components₂

Componentsₙ

WGA Assembly

CSA Assembly

Unique Scaffolds

BAC Scaffolds

Combining Assembler

Tiler
Chromosome 21: Public vs Celera Assemblies

http://www.sciencemag.org/cgi/content/full/291/5507/1304
Chromosome 8: Public vs. Celera
http://www.sciencemag.org/cgi/content/full/291/5507/1304
Finishing Strategy for the Public Genome Project

### Table 2: Finished sequence and gaps, HGSC Build 35

<table>
<thead>
<tr>
<th>Chr</th>
<th>Total finished sequence (kb)</th>
<th>Euchromatic gaps†</th>
<th>Heterochromatic gaps‡</th>
<th>Estimate of total gap size§ (kb)</th>
<th>Unfinished clones</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>222,828</td>
<td>32</td>
<td>1,605</td>
<td>2</td>
<td>21,115</td>
<td>17</td>
<td>850</td>
</tr>
<tr>
<td>2</td>
<td>237,503</td>
<td>20</td>
<td>2,512</td>
<td>1</td>
<td>5,412</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>194,636</td>
<td>5</td>
<td>1,935</td>
<td>1</td>
<td>3,435</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>187,161</td>
<td>14</td>
<td>1,250</td>
<td>1</td>
<td>2,900</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>177,703</td>
<td>5</td>
<td>92</td>
<td>1</td>
<td>1,500</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>167,318</td>
<td>10</td>
<td>658</td>
<td>1</td>
<td>3,000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>154,759</td>
<td>11</td>
<td>869</td>
<td>1</td>
<td>1,340</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>142,613</td>
<td>9</td>
<td>662</td>
<td>1</td>
<td>2,300</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>117,781</td>
<td>40</td>
<td>1,955</td>
<td>2</td>
<td>19,955</td>
<td>12</td>
<td>600</td>
</tr>
<tr>
<td>10</td>
<td>131,614</td>
<td>12</td>
<td>1,020</td>
<td>1</td>
<td>19,955</td>
<td>8</td>
<td>400</td>
</tr>
<tr>
<td>11</td>
<td>131,131</td>
<td>7</td>
<td>322</td>
<td>1</td>
<td>5,082</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>130,259</td>
<td>8</td>
<td>796</td>
<td>1</td>
<td>5,095</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>95,560</td>
<td>6</td>
<td>715</td>
<td>2</td>
<td>17,200</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>88,291</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>17,220</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>81,342</td>
<td>10</td>
<td>737</td>
<td>2</td>
<td>18,260</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>78,885</td>
<td>4</td>
<td>143</td>
<td>2</td>
<td>10,000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>77,800</td>
<td>9</td>
<td>875</td>
<td>1</td>
<td>7,500</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>74,656</td>
<td>3</td>
<td>97</td>
<td>1</td>
<td>1,368</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>55,786</td>
<td>5</td>
<td>5,015</td>
<td>1</td>
<td>340</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>59,505</td>
<td>4</td>
<td>1,157</td>
<td>1</td>
<td>1,766</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>34,170</td>
<td>3</td>
<td>53</td>
<td>2</td>
<td>11,620</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>34,765</td>
<td>11</td>
<td>460</td>
<td>2</td>
<td>14,330</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>150,394</td>
<td>12</td>
<td>750</td>
<td>1</td>
<td>3,750</td>
<td>14</td>
<td>700</td>
</tr>
<tr>
<td>Y</td>
<td>24,672</td>
<td>9</td>
<td>1,480</td>
<td>2</td>
<td>33,098</td>
<td>7</td>
<td>350</td>
</tr>
<tr>
<td>Total</td>
<td>2,851,331</td>
<td>250</td>
<td>25,165</td>
<td>30</td>
<td>200,167</td>
<td>58</td>
<td>2,000</td>
</tr>
</tbody>
</table>

*The total length of tilling paths including only finished bases of clones in Build 35. Roughly 2.19 Mb of sequence on chromosome Y was derived directly from the equivalent pseudoautosomal region on chromosome X.
†Defined as gaps in euchromatic regions, including junctions with heterochromatic/centromeric sequences, for which no clone was available (see text).
‡Defined here as gaps in heterochromatic regions (see text and Supplementary Note 2 on heterochromatic sequence). Separate gaps were counted for centromeres and pericentromeric heterochromatin, even when the two were contiguous. Centromere sizes were taken from ref. 62 or in some cases provided directly by the sequencing centres (see Supplementary Note 2). Acrocentric sizes are based on centromere ratios from ref. 63. The sizes of large heterochromatic gaps are typically difficult to estimate accurately owing to their repeat structure and polymorphic nature. Other regions might arguably be called heterochromatin for example, the pericentromeric regions of chromosomes 19 and 3 and a ~400-kb gap on the Y chromosome, but are classified as euchromatin here.
§The sum of lengths for finished sequence, estimated heterochromatic gaps, euchromatic gaps and unfinished clone gaps. The total length is only approximate because of uncertainty in gap sizes, particularly for heterochromatic and centromeres.
||Those in the tilling path but for which it has not been possible to obtain finished sequence. Unfinished sequence from these clones is deposited in public databases. These gaps are all listed at 50 kb, reflecting the approximate average size of the gap.
**Figure 1** Comparison of previous draft sequence with current near-complete sequence of chromosome 7 (ref. 24). At large scale, there was good collinearity between draft and near-complete sequence, although some inversions were present in the draft due to lack of sufficient anchors in some regions. At finer scale, the draft sequence contained some sequence contigs for which order and orientation were not known. The inset shows a region of 500 kb with sequence derived from three overlapping BACs. BACs at each end were finished at the time of draft assembly, whereas the middle BAC was at an early stage of shotgun coverage in which contigs were not yet ordered and oriented.
Substitutions in BAC Overlaps with BACs from Same or Different Libraries
Gaps in BAC Overlaps with BACs from
Duplications and Deletions in the Human Genome

Figure 4 Segmental duplications across the genome. a, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length \( \geq 10 \, \text{kb} \) and sequence identity \( \geq 95\% \)). Large duplications are shown to approximate scale; smaller ones are indicated as ticks. Sequence gaps are indicated above the chromosomes in red. Large gaps (\( >300 \, \text{kb} \)) are shown to approximate scale; smaller gaps are indicated as ticks with those that are 50 kb or smaller shown as shorter ticks. Unfinished clones are indicated as black ticks. b, Percentage of human clones with indicated segmental duplications.
Percentage of Chromosomes Duplicated

Available Next Generation Sequencing Technologies

- **Illumina – Solexa**

- **Illumina Technology**
  - [http://www.illumina.com/technology/sequencing_technology.ilmn](http://www.illumina.com/technology/sequencing_technology.ilmn)

- **454 Life Sciences**

- **454 Life Sciences Technology**

- **Applied Biosystems Inc. (ABI) SOLID Sequencing**
  - [http://solid.appliedbiosystems.com/](http://solid.appliedbiosystems.com/)

- **ABI SOLID Sequencing Technology**
Next Generation Sequencing Technologies in Development

- **Pacific BioSciences**
- **Pacific BioSciences Technology**
- **Helicos**
- **Helicos Technology**
- **Complete Genomics**
- **Complete Genomics Technology**