Introduction to Microbial Genomics

Dave Ussery
Comparative Microbial Genomics Workshop
Pathumthani, Thailand
7 March, 2005
Outline

• Introduction
• Comparison of Genome Sizes
• The Human Genome Project
• DNA repeats
• A few words about speed of sequencing
What is a genome?

**genome** dʒiː.nəm. Biol. Formerly also genom -nom. [a. *G. genom* (H. Winkler Verbreitung u. Ursache d. Parthenogenesis (1920) iv. 165), irreg. f. gene gene¹ + chromosom chromosome.] **A haploid set of chromosomes; the sum-total of the genes in such a set.**

1930 *Cytologia* I. 14 Chromosomes from different sets (or genoms) of *Triticum vulgare* show affinity toward each other.

1930 [see **allopolyploidy**].

1932 *Proc. 6th Int. Congr. Genetics* I. 275 The inviability of deficient genomes in the haploid generation serves to some extent as an alternative distinction between mutation and deficiency.

1932 *Proc. 6th Int. Congr. Genetics* II. 5 There are two species having genoms resembling *C. neglecta*.

1952 C. P. Blacker *Eugenics* x. 243 The appearance of such terms as gene-complex and genome (denoting a set of chromosomes as a working unity) testify to the movement towards holism in genetics.

1965 A. M. Srb et al. *Gen. Genetics* (ed. 2) vii. 190 Among organisms with chromosomes, each species has a characteristic set of genes, or genome. In diploids a genome is found in each normal gamete. It consists of a full set of the different kinds of chromosomes.

1970 *Sci. Amer.* Oct. 19/1 The human genome consists of perhaps as many as 10 million genes.
The Central Dogma of Molecular Biology

- **Replication**
  - DNA passes coded information with replication

- **Transcription**
  - RNA synthesis: coded information passed into RNA during transcription

- **Translation**
  - Messenger RNA carries coded information to ribosome during protein synthesis

The buck stops here. Proteins refuse to give away any information.

Proteins provide structure and help carry out almost all biological activity.
What is Biological Information?

DNA → RNA → protein

Genome -> Transcriptome -> Proteome

retroviruses  RNA viruses
<table>
<thead>
<tr>
<th>Organism</th>
<th># bp</th>
<th># genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Φ-X174</td>
<td>5386</td>
<td>9</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>4,600,000</td>
<td>4288</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>13,000,000</td>
<td>5885</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>~100,000,000</td>
<td>~14,000</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>~120,000,000</td>
<td>~10,000</td>
</tr>
<tr>
<td><em>Drosophila melanogastor</em></td>
<td>~180,000,000</td>
<td>~12,000</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>~3,400,000,000</td>
<td>~25,000</td>
</tr>
</tbody>
</table>
Phage λ
50 kb
2 pages

*Escherichia coli*
(bacteria)
4.7 Mb
200 pages

*Saccharomyces cerevisiae*
(yeast)
12.5 Mb
500 pages

*Caenorhabditis elegans*
(nematode)
100 Mb
3 volumes

Human being
3000 Mb
80 volumes
The Human Genome Project

Started ~1985

The U.S. government agreed to invest $200,000,000 U.S. per year for 20 years.

~3,400,000,000 bp per haploid genome
~6,800,000,000 bp per haploid genome

One base per second = 216 years!
<table>
<thead>
<tr>
<th>year</th>
<th># human genes mapped</th>
<th>#years to sequence human genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>none</td>
<td>not possible</td>
</tr>
<tr>
<td>1980</td>
<td>3</td>
<td>~4,000,000 years</td>
</tr>
<tr>
<td>1990</td>
<td>12</td>
<td>~1000 years</td>
</tr>
<tr>
<td>2000</td>
<td>~25,000</td>
<td>draft</td>
</tr>
<tr>
<td>2005</td>
<td>~30,000</td>
<td>new draft + chimp, chicken, dog, mouse, pig, rat...</td>
</tr>
</tbody>
</table>
THE GENOME WAR
HOW CRAIG VENTER TRIED TO CAPTURE THE CODE OF LIFE AND SAVE THE WORLD
JAMES SHREEVE
THE GENE MASTERS

How a New Breed of Scientific Entrepreneurs Raced for the Biggest Prize in Biology

INGRID WICKELGREN
Discussion questions:

Why is the human genome not finished yet?

Why is the human genome so large?

~3,400,000 bp per haploid genome
~6,800,000 bp per haploid genome

One base per second = 216 years!

~2% or less codes for proteins.
Aristotle's ladder of complexity
The “C-value paradox”

The genome size of an organism is defined as the amount of haploid DNA in a genomic set (e.g., an egg or sperm nucleus). This is also referred to as the "C-value"; the "C" means "constant" or "characteristic", since the size of a genome is usually constant for a given species.

The large difference in genome sizes without any seeming relation to an organism’s complexity, is called the C-value paradox.
What does all this DNA do?

90%
50%
2%
<1%
DNA repeats

The approximate size and characteristics of genomes was characterised in the 1960s, in a classic study of the kinetics of DNA reassociation by Britten and Kohne (1968).

They found that the DNA could be divided into four fractions:

1. foldback DNA
2. highly repetitive DNA
3. middle-repetitive DNA
4. single-copy DNA

The repetitive DNA can either be localised to discrete regions, or dispersed.

Highly repetitive DNA

**Dispersed - e.g., Alu family**
- about 300 bp long
- 500,000 copies in humans
- (about 5% of the human genome)
- dispersed throughout the chromosomes

**Localised highly repetitive sequences**
- about 2-10 bp long
- present in millions of copies, often in large blocks
- (about 6% of the human genome)
- associated with heterochromatin
- usually very high A+T content
Localised repetitive DNA

Often, satellite DNA consists of long tandem arrays of repeated sequences, all localised to one or a few discrete regions in the chromosomes. For example, in the kangaroo rat (*Dipodomys ordii*), more than 50% of the genome consists of three families of repeated sequences:

\[(AAG)_n, \text{ where } n = \sim 2.24 \times 10^9\]
\[(TTAGGG)_n, \text{ where } n = \sim 2.2 \times 10^9\]
\[(ACACAGCGGG)_n, \text{ where } n = \sim 1.2 \times 10^9\]
Middle repetitive DNA

- makes up more than 40% of the human genome
- position varies due to transposable elements
- Includes the following types of sequences:
  - Dinucleotide repeats
  - microsatellite DNA
  - TRInucleotide repeats
  - associated with many diseases
  - (e.g., Fragile X, muscular distrophy)
What about bacteria?
<table>
<thead>
<tr>
<th>Organism</th>
<th>%AT</th>
<th>Size (bp)</th>
<th>Atlas</th>
<th>Number of genes</th>
<th>Coding density</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli&lt;br&gt;Strain: K-12, isolate W3110&lt;br&gt;DDBJ NCBI tax</td>
<td>49</td>
<td>4,636,552</td>
<td>Genome Atlas</td>
<td>4085</td>
<td>79% 1135 bp/gene</td>
<td>GenoBase (Japan) unpublished</td>
</tr>
<tr>
<td>Escherichia coli&lt;br&gt;Strain: E2348/69</td>
<td>49</td>
<td>~4,997,000 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Sanger Center Unpublished</td>
</tr>
<tr>
<td>Escherichia coli&lt;br&gt;Strain: 042</td>
<td>49</td>
<td>~5,184,000 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Sanger Center Unpublished</td>
</tr>
<tr>
<td>Escherichia coli&lt;br&gt;Strain: RS218</td>
<td>49</td>
<td>~5,200,000 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>U. Wisconsin Unpublished</td>
</tr>
<tr>
<td>Escherichia coli&lt;br&gt;Strain: O157:H7 (substrain RIMD 0509952)&lt;br&gt;DDBJ NCBI tax</td>
<td>49</td>
<td>5,498,450</td>
<td>Genome Atlas</td>
<td>5361</td>
<td>88% 1026 bp/gene</td>
<td>DNA Res. 8:11-22 February, 2001 (PubMed)</td>
</tr>
<tr>
<td>Escherichia coli&lt;br&gt;Strain: non K-1</td>
<td>49</td>
<td>~5,800,000 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Sanger Center Unpublished</td>
</tr>
</tbody>
</table>
GENOME ATLAS DATABASE


Organisms

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Genomes</th>
<th>Species</th>
<th>Total basepairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td>21</td>
<td>21</td>
<td>46,210,894 bp</td>
</tr>
<tr>
<td>Bacteria</td>
<td>219</td>
<td>162</td>
<td>688,041,310 bp</td>
</tr>
<tr>
<td>Eukaryotes</td>
<td>17</td>
<td>17</td>
<td>594,670,914 bp</td>
</tr>
</tbody>
</table>

Mitochondrial DNA

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Genomes</th>
<th>Species</th>
<th>Total basepairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitochondria</td>
<td>464</td>
<td>448</td>
<td>11,742,093 bp</td>
</tr>
</tbody>
</table>
**Microbiology Comment**

**Table 1.** Summary of the published genomes discussed in this update

The accession number for each chromosome is the same for GenBank, EMBL and DDBJ. Note that of the four *Campylobacter* genomes, only *C. jejuni* has been fully sequenced in one piece. *P. berghei* and *P. chabaudi* have been sequenced to 7479 and 10690 pieces, respectively, so we are unable to extract data regarding A+T content and number of tRNAs and rRNAs.

<table>
<thead>
<tr>
<th>Name</th>
<th>Genome length (bp)</th>
<th>A + T (mol%)</th>
<th>No. of genes</th>
<th>tRNAs</th>
<th>rRNAs</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Anaplasma marginale</em> St Maries</td>
<td>1 197 687</td>
<td>50·2</td>
<td>949</td>
<td>37</td>
<td>1</td>
<td>CP000030</td>
</tr>
<tr>
<td><em>Dehalococcoides ethenogenes</em> 195</td>
<td>1 469 720</td>
<td>51·1</td>
<td>1591</td>
<td>46</td>
<td>1</td>
<td>CP000027</td>
</tr>
<tr>
<td><em>Campylobacter jejuni</em> RM1221</td>
<td>1 777 831</td>
<td>69·7</td>
<td>1838</td>
<td>44</td>
<td>3</td>
<td>CP000025</td>
</tr>
<tr>
<td><em>Campylobacter lari</em> RM2100</td>
<td>1 500 000</td>
<td>73·1</td>
<td>1554</td>
<td>42</td>
<td>1</td>
<td>AAFK00000000</td>
</tr>
<tr>
<td><em>Campylobacter upsaliensis</em> RM3195</td>
<td>1 660 000</td>
<td>69·8</td>
<td>1782</td>
<td>43</td>
<td>3</td>
<td>AAFJ00000000</td>
</tr>
<tr>
<td><em>Campylobacter coli</em> RM2228</td>
<td>1 680 000</td>
<td>71·4</td>
<td>1764</td>
<td>43</td>
<td>3</td>
<td>AAFL00000000</td>
</tr>
<tr>
<td><em>Geobacillus kaustophilus</em> HTA426</td>
<td>3 544 776</td>
<td>47·9</td>
<td>3498</td>
<td>87</td>
<td>9</td>
<td>BA000043</td>
</tr>
<tr>
<td><em>Geobacillus kaustophilus</em> pHTA426</td>
<td>47 890</td>
<td>55·8</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>AP006520</td>
</tr>
<tr>
<td><em>Idiomarina loihiensis</em> L2TR</td>
<td>2 839 318</td>
<td>53·0</td>
<td>2628</td>
<td>56</td>
<td>4</td>
<td>AE017340</td>
</tr>
<tr>
<td><em>Salmonella enterica</em> ATCC 9150</td>
<td>4 585 229</td>
<td>47·8</td>
<td>4093</td>
<td>82</td>
<td>7</td>
<td>CP000026</td>
</tr>
<tr>
<td><em>Zymomonas mobilis</em> ZM4</td>
<td>2 056 416</td>
<td>53·7</td>
<td>1998</td>
<td>51</td>
<td>3</td>
<td>AE008692</td>
</tr>
<tr>
<td><em>Plasmodium berghei</em></td>
<td>17 996 878</td>
<td>–</td>
<td>5864</td>
<td>–</td>
<td>–</td>
<td>CAAJ01000000</td>
</tr>
<tr>
<td><em>Plasmodium chabaudi</em></td>
<td>16 866 661</td>
<td>–</td>
<td>5698</td>
<td>–</td>
<td>–</td>
<td>CAAJ01000000</td>
</tr>
</tbody>
</table>
Bacillus cereus, strain ZK          CP000001   GenBank, unpublished
Legionella pneumophila Lens           CR628337 Nature Genetics (2004), In the press.
Kronborg Castle
Growth in GenBank

Number of bp in GenBank

Year

10^12
10^10
10^8
10^6
10^4

#bp in GenBank

Moore’s law
Bioinformatics for Dummies

Find Windows-friendly tools that save time and get results

A Reference for the Rest of Us!

FREE eTips at dummies.com

Jean-Michel Claverie, PhD
Research Director at France’s Centre National de la Recherche Scientifique (CNRS)

Cedric Notredame, PhD
Professor of Bioinformatics at Switzerland’s Lausanne University and the CNRS
What is Bioinformatics?

Bioinformatics is the application of machine learning processes to biological information.
bioinformatics, n.
The science of information and information flow in biological systems, esp. the use of computational methods in genetics and genomics.

1978 P. HOGEWEG in *Simulation* 31 90/1 Since 1970 she has been a staff member at the Subfaculty of Biology of the University of Utrecht, with her main field of research in bioinformatics. 1985 *Jnl. Theoret. Biol.* 113 719 (heading) Tumor escape from immune elimination... R. J. De Beer, Bioinformatics Group, University of Utrecht. 1986 *Philos. Trans. Royal Soc. A.* 317 324 The area of modelling mutants from a known structure has been revolutionized by the latest tools of molecular graphics... This is a key element in the whole technology and has attracted much interest (for example, the recent E.E.C. ‘Bioinformatics’ programme). 1987 *Science* 4 Sept. 1108/3 One of the latest developments [at the European Molecular Biology Laboratory] has been the creation of a new research program in bioinformatics. This is intended to bring together research in computing science, structural biology, and molecular genetics. 1996 *Fast Company* Aug.-Sept. 32/3 A lot of breakthroughs in medicine will come out of the efforts of bioinformatics. 2001 *N.Y. Times* 4 Jan. B6/2 The hope...is to make New York a leader in cutting-edge fields like bioinformatics, in which computers are used to decipher genes and proteins.
The Central Dogma of Molecular Biology

Replication
(DNA passes coded information with replication)

Transcription
RNA synthesis: coded information passed into RNA during transcription

Translation
Messenger RNA carries coded information to ribosome during protein synthesis

The buck stops here. Proteins refuse to give away any information.

Proteins provide structure and help carry out almost all biological activity.
What is Biological Information?

DNA -> RNA -> protein

Genome -> Transcriptome -> Proteome
The DNA sequence contains information.

But what kind of information?
Summary

1. A “genome” is sum of all the DNA sequences (chromosomes) in an organism.

2. The sizes of genomes range about 100,000,000 fold - from small viruses to very large amoebas.

3. Many eukaryotic genomes contain large fractions of repeated DNA sequences, and there is no apparent correlation between the size of an organism and its biological complexity (”C-value Paradox”)

4. Sequence databases are growing faster than computational power!