Microbial genomics

張傳雄
國立陽明大學 遺傳學研究所
01-06-2003
Why study microbial genomes?

• studies on whole genome sequences give us a complete genomic blueprint for an organism. We can examine how all the parts operate cooperatively to influence the activities and behavior of an entire organism – a complete understanding of the biology of an organism. Microbes provide an excellent starting point for studies of this type as they have a relatively simple genomic structure compared to higher, multi-cellular organisms.

• studies on microbial genomes may provide crucial starting points for the understanding of the genomics of higher organisms.

• analysis of whole microbial genomes also provides insight into microbial evolution and diversity beyond single protein or gene phylogenies.

• analysis of whole microbial genomes is also a powerful tool in identifying new applications for biotechnology and new approaches to the treatment and control of pathogenic organisms.
Phylogeny of the living world

(Data of Cart R. Woese)

YM-Genetics
Genome size variation

From McAllister, UTA
**Early microbial genome sequencing**

<table>
<thead>
<tr>
<th>Organism/organelle</th>
<th>Genome Size</th>
<th>Date</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteriophage MS2</td>
<td>5.6 knt</td>
<td>1976</td>
<td>First organism (ss-RNA)</td>
</tr>
<tr>
<td>SV40</td>
<td>5.8 kbp</td>
<td>1979</td>
<td>First virus</td>
</tr>
<tr>
<td>CaMV</td>
<td>8.0 kbp</td>
<td>1980</td>
<td>Cauliflower mosaic virus</td>
</tr>
<tr>
<td>TMV</td>
<td>6.3 knt</td>
<td>1982</td>
<td>Tobacco mosaic virus</td>
</tr>
<tr>
<td>Bacteriophage λ</td>
<td>48.5 kbp</td>
<td>1982</td>
<td></td>
</tr>
<tr>
<td>chloroplast</td>
<td>121 kbp</td>
<td>1986</td>
<td>Tobacco</td>
</tr>
<tr>
<td>Vaccinia virus</td>
<td>192 kbp</td>
<td>1990</td>
<td></td>
</tr>
<tr>
<td>CMV</td>
<td>229 kbp</td>
<td>1991</td>
<td>Cytomegalovirus</td>
</tr>
<tr>
<td>Marchantia polymorpha</td>
<td>187/121 kbp</td>
<td>1992</td>
<td>liverwort mitochondria &amp; chloroplast</td>
</tr>
<tr>
<td>Variola</td>
<td>186 kbp</td>
<td>1993</td>
<td>Smallpox virus (automated sequencing)</td>
</tr>
</tbody>
</table>
Genome types

• **RNA Genome**
  i) ssRNA
    - *e.g.*, Poliovirus, Rabies virus, HIV retrovirus
  ii) dsRNA
    - *e.g.*, Reovirus
    → This virus’ genome consists of 10-12 linear pieces of ssRNA

• **DNA Genome**
  • ssDNA
    - linear → paroviruses
    - circular → M13 phage
ii) dsDNA genome

- linear - T4 phage

- circular - Herpes viruses
- Simian virus 40 (SV40)

- sealed ends/closed - Poxvirus

- terminated protein - Adenovirus
Viral genomes encode:

- genes for their own replication

- genes for taking over the hosts metabolism and/or integrating into the host’s genome
  - *(retrovirus → integrase)* genes for their own replication

- genes for capsid proteins/viral coat proteins
118 Published Complete Genomes:
- Archaeal: 16 species
- Bacterial: 87 species
- Eukaryal: 15 (Homo sapiens, plants, insects, nematodes, protozoa, fungi, …)

352 Prokaryotic Ongoing Genomes:
- Archaeal: 23 species
- Bacterial: 329 species
The principal steps involved in generating a complete bacterial genome sequence (the only really automatable step)
Microbial genome sequencing strategy

• “Shotgun” sequencing
  – shear DNA into small fragments
  – insert into vector
  – sequence in from vector
  – computer aligns & assembles sequences based on overlap
  – ordering of contigs
  – primer walking to complete sequence

• Working with sequence data
  – open reading frames identified
  – databases searched for similar sequences – genes identified & annotated
  – comparison of genetic complements of different organisms
Whole genome shotgun assembly

1. Find overlapping reads

2. Merge good pairs of reads into longer contigs

3. Link contigs to form supercontigs

4. Derive consensus sequence

http://www-genome.wi.mit.edu/wga/
Laboratory tools for studying whole genomes

Pulsed Field Gel Electrophoresis (PFGE)

- the most important factor in PFGE resolution is switching time, longer switching times generally lead to increased size of DNA fragments which can be resolved
- switching times are optimised for the expected size of the DNA being run on the PFGE gel
- switch time ramping increases the region of the gel in which DNA separation is linear with respect to size
- a number of different apparatus have been developed in order to generate this switching in electric fields however most commonly used in modern laboratories are FIGE (Field Inversion Gel Electrophoresis) and CHEF (Contour-Clamped Homogenous Electrophoresis)
Separating large fragments

Pulsed field gel electrophoresis (PFGE)

Alternating electric fields

FIGE

field inversion

CHEF

transverse alternating
Large insert cloning vectors – BAC & PAC

• conventional plasmid derived cloning vectors are only able to reliably maintain inserts less than 20 kb in size
• there are a number of approaches to generating clones with inserts in an intermediate size range (20 – 80 kb) such as cosmids, etc.
• the most commonly used vectors for cloning extremely large DNA inserts are BACs (Bacterial Artificial Chromosomes) and PACs (P1-derived Artificial Chromosomes)
• both BAC and PAC vectors are plasmid derived vectors distinguished from conventional vectors by extremely tightly controlled low copy numbers
• BAC and PAC vectors both utilise *E. coli* as the host organism
• BAC vectors are based on the *E. coli* single copy F-factor plasmid – the F-factor origin of replication is very tightly controlled
• PAC vectors are based on an identical principle but instead use a single copy origin of replication derived from P1 phage
Example - *Haemophilus influenzae*

- first complete genome sequence of a free living organism (1995)
- important pathogen
- genome is around 1.83 Mb in size
- random sequencing was done for both small insert and large insert (lambda) libraries
- sequencing reactions performed by eight individuals using fourteen ABI 377 DNA sequencers per day over a three month period
- in total around 33000 sequencing reactions were performed on 20000 templates
- plasmid extraction performed in a 96 well format
- 11 Mb of sequence was initially used to generate 140 contigs
- gaps were closed by lambda linking clones (23), peptide links (2), Southern analysis (37) and PCR (42)
Genes are interspersed along DNA molecules, being separated by DNA sequence of unknown function (intergenic regions)
(a) *Escherichia coli*

(b) *Saccharomyces cerevisiae*

(c) Human

(d) Maize

kb

0 10 20 30 40 50

- **Gene**
- **Pseudogene**
- **Repetitive DNA**
Bacterial chromosomes (1)

double stranded DNA, range of sizes:

~750 kb in *Mycoplasma* species

5000 kb (5 Mb) in *Escherichia coli*

10 Mb in *Streptomyces* species
Bacterial chromosomes (2)

- Usually a single chromosome,

- but the following species
  have two chromosomes:

  *Vibrio cholerae* (3.0 and 1.1 Mb)

  *Burkholderia pseudomallei* (3.6 and 2.4 Mb)

  *Rhodobacter spheroides* (3 and 1 Mb)
Bacterial chromosomes (3)

- Usually circular, supercoiled DNA
  (comparing to linear eukaryotic chromosomes)

  e.g., *E. coli*, *Bacillus subtilis*, *Streptomyces coelicolor*,
  *Salmonella typhimurium*, *Streptococcus pneumoniae*,
  *Haemophilus influenzae*

- But linear chromosomes in some spirochaetes [e.g., *Borrelia burgdorferi* (~911 kb)]
Bacterial chromosomes (4)

- Chromosome-associated proteins, ‘histone-like’ proteins or ‘nucleoid-associated’ proteins

  e.g., HU of *Escherichia coli*, 40,000 molecules/cell (approx. 1 molecule/100 bp of DNA)

  involved in packaging the chromosome, regulating transcription
Bacterial plasmids (1)

- Stable, independently replicating ‘additional’ genetic elements
- Mostly circular, supercoiled, double stranded DNA molecules
- But.... linear plasmids in *Streptomyces*, *Borrelia burgdorferi*
  single-stranded DNA plasmids in *Myxococcus xanthus*
Bacterial plasmids (2)

Regulated copy number:

- small plasmids (5-10 kb) 50-100 copies/cell
  (1000 in some *Streptomyces* plasmids)

- large plasmids (50-200 kb) 1-10 copies/cell

Limited host range – but some ‘promiscuous’,
  broad host range plasmids
Bacterial plasmids (3)

- Accessory genetic elements encoding adaptive functions:
  - conjugation
  - antibiotic resistance:
    - enzymic degradation (e.g. penicillin)
    - enzymic modification (e.g. chloramphenicol)
    - altered membrane permeability (e.g. tetracycline)
    - alteration of drug target (e.g. streptomycin)
    - alternative metabolic activity (e.g. sulphonamide)
  - virulence (invasion, toxin production)
  - symbiosis
  - substrate degradation
Bacterial viruses (1)

- a.k.a. bacteriophages (eaters of bacteria) or ‘phages’

  **Virulent** phages lyse
  (break open) infected cells

  **Temperate** phages form
  prophages in infected cells
  (undergo lysogeny)
Bacterial viruses (2)

- Phage (or lysogenic) conversion:
  - *Streptococcus pyogenes* (throat infections, scarlet fever)
  - *Corynebacterium diphtheriae* (diphtheria)

  Toxin genes on prophages, so only lysogens are virulent

  Non-virulent (i.e. non-lysogenic) strains ‘converted’ to virulence by phage infection
Mobile genetic elements (1)

a.k.a. transposable genetic elements

Insertion sequences (IS)

Transposons

Pathogenicity islands (PAIs)

(PAIs) are stretches of ORFs on a bacterial chromosome which contain in clustered form main determinants of the bacterium’s Pathogenicity Islands pathogenic potential.
Mobile genetic elements (2)

- Insertion sequences (IS)
  - typically 1300-1500 bp
  - encode ‘transposase’ (Tnp) and other proteins
  - terminal inverted repeat (IR, →) sequences (8-40 bp)
  - duplicated ‘target sequence’ (0-12 bp)

IS1 of *Escherichia coli*
  - unusually small, 768 bp
  - 8 genes (including Tpn)
  - IRs are 23 bp
  - target sequence is 9 bp
# Mobile genetic elements (3)

- Class II (simple) transposons - similar to ISs

<table>
<thead>
<tr>
<th>Transposon</th>
<th>Size (kb)</th>
<th>Gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tn3</td>
<td>3.5</td>
<td>bla</td>
</tr>
<tr>
<td>Tn2513</td>
<td>7</td>
<td>mer</td>
</tr>
<tr>
<td>Tn21</td>
<td>19</td>
<td>mer sul str</td>
</tr>
<tr>
<td>Tn4</td>
<td>22</td>
<td>bla* sul str  (*Tn3)</td>
</tr>
</tbody>
</table>
Mobile genetic elements (4)

- Class I (composite or compound) transposons

  Antibiotic resistance genes flanked by whole ISs

  Tn10 9 kb  *tet* flanked by IS10 (1300 bp, 23 bp IRs)

  Tn9 2.5 kb  *cat* flanked by IS1 (768 bp, 23 bp IRs)
Mobile genetic elements (5)

- Pathogenicity islands

  large regions with different G+C composition from most of the chromosome

  pathogen-specific (loss results in loss of virulence)

  often flanked by direct repeats (c.f. ISs and Tns)

  often targeted to tRNA genes and/or IS elements
Bacterial genome sizes

• Smallest: Mycoplasma genitalium 580 kb (0.58 Mb)
  Largest: Myxococcus xanthus 9200 kb (9.2 Mb)
  Median: ~2000 kb (2.0 Mb)

• Average gene size: 0.9-1.0 kb

• ~90% of genome encodes protein and stable RNA

• The larger the bacterial genome,
  the more genes the genome contains

• Bacterial gene number reflects bacterial lifestyle:
  small genomes = obligate parasites
  large genomes = metabolically flexible and/or development
Bacterial chromosome numbers

- Most bacteria contain a single chromosome
  (± extrachromosomal elements)

- Some bacteria have been found also to contain 2-3 replicons which can be considered either megaplasmids or minichromosomes
  (e.g., 3.0 Mb and 0.9 Mb replicons in *Rhodobacter sphaeroides*)

- A few bacterial genera contain >1 chromosome
  (e.g., 2.1 Mb and 1.2 Mb chromosomes in *Brucella*)

- Some bacteria harbour large replicons essential for survival in a specific ecological niche but not under laboratory conditions
  (e.g., 1.4 Mb and 1.7 Mb replicons in *Rhizobium meliloti* are required for plant symbiosis)
Gene order & orientation

• Gene order in bacteria is NOT constant over evolutionary time, even among bacteria within the same phylum

• No obvious rationale for gene order although genes near the replication origin may be present at increased numbers

• Gene orientation is often more regular: replication and transcription often proceed in the same direction

• The order of genes within operons is commonly conserved:

  *Bacillus subtilis*
  
  trpE  trpD  trpC  trpF  trpB  trpA

  *Escherichia coli*
  
  trpE  trpD  trpC  trpB  trpA

  *Methanobacterium sp.*
  
  trpE  trpG  trpC  trpF  trpB  trpA  trpD
The role of accessory elements in chromosome dynamics

- Most bacterial genomes typically contain many integrated accessory elements: transposons, plasmids, prophage, and pathogenicity islands among others
- Are usually recognizable by their sequence but not always
- Contribute to the variability in genome structures between even closely-related species
- May be functional or nonfunctional
- May be valuable or selfish or both

- *B. subtilis* does not have transposable elements!
Summary

• A number of bacterial genomes have been sequenced; even more are in progress

• Both sequencing and physical analyses give valuable information about genome structure and organization

• Bacterial genomes vary in size; more DNA = more genes

• Chromosomes are mainly circular, but may be linear

• Some bacteria contain >1 chromosome, or >1 copy of an individual chromosome

• Most of the genome is composed of coding sequences

• Gene order is not constant

• Operons are conserved

• Genome structure may be conserved over long evolutionary periods or may undergo rearrangement

• Accessory elements contribute to macromolecular rearrangements
Genome Size (prokaryotes)

- Bacterial genome: $6 \times 10^5$–more than $10^7$
  Smallest known: *Mycoplasma genitalium* (470 protein coding genes, 3 rRNA genes, 33 tRNA genes)

- Prokaryotes genome sizes are roughly proportional to gene numbers.

- Processes affect bacterial genome size:
  Gene duplication, small-scale deletions and insertions, transpositions, horizontal transfer, loss of genes in parasitic lines, etc.
## Microbial Genome Sequencing Projects

### Complete

- 16 archaea
- 87 bacteria
- 15 eucaryotes

### In progress

- 23 archaea
- 329 eubacteria
- 236 eucaryotes

<table>
<thead>
<tr>
<th>Year</th>
<th>Group</th>
<th>Species</th>
<th>Strain</th>
<th>Genome Size (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>eubacteria</td>
<td>Haemophilus influenzae Rd</td>
<td>KW20</td>
<td>1.83</td>
</tr>
<tr>
<td>1995</td>
<td>eubacteria</td>
<td>Mycoplasma genitalium</td>
<td>G37</td>
<td>0.58</td>
</tr>
<tr>
<td>1996</td>
<td>archaea</td>
<td>Methanococcus jannaschii</td>
<td>DSM2661</td>
<td>1.66</td>
</tr>
<tr>
<td>1996</td>
<td>eubacteria</td>
<td>Synechocystis sp.</td>
<td>PCC6803</td>
<td>3.57</td>
</tr>
<tr>
<td>1996</td>
<td>eubacteria</td>
<td>Mycoplasma pneumoniae</td>
<td>M129</td>
<td>0.81</td>
</tr>
<tr>
<td>1996</td>
<td>eucaryote</td>
<td>Saccharomyces cerevisiae</td>
<td>S288C</td>
<td>13.00</td>
</tr>
<tr>
<td>1997</td>
<td>eubacteria</td>
<td>Escherichia coli</td>
<td>K12</td>
<td>4.60</td>
</tr>
<tr>
<td>1997</td>
<td>eubacteria</td>
<td>Helicobacter pylori</td>
<td>26695</td>
<td>1.66</td>
</tr>
<tr>
<td>1997</td>
<td>eubacteria</td>
<td>Bacillus subtilis</td>
<td>168</td>
<td>4.20</td>
</tr>
<tr>
<td>1997</td>
<td>eubacteria</td>
<td>Borrelia burgdorferi</td>
<td>B31</td>
<td>1.44</td>
</tr>
<tr>
<td>1997</td>
<td>archaea</td>
<td>Methanobacterium thermoautotrophicum delta H</td>
<td>OT3</td>
<td>1.75</td>
</tr>
<tr>
<td>1997</td>
<td>archaea</td>
<td>Archaeoglobus fulgidus</td>
<td>DSM4304</td>
<td>2.18</td>
</tr>
<tr>
<td>1998</td>
<td>eubacteria</td>
<td>Aquifex</td>
<td>VF5</td>
<td>1.50</td>
</tr>
<tr>
<td>1998</td>
<td>eubacteria</td>
<td>Mycobacterium tuberculosis</td>
<td>H37Rv</td>
<td>4.40</td>
</tr>
<tr>
<td>1998</td>
<td>eubacteria</td>
<td>Treponema pallidum</td>
<td>Nichols</td>
<td>1.14</td>
</tr>
<tr>
<td>1998</td>
<td>eubacteria</td>
<td>Chlamydia trachomatis</td>
<td>serovar D</td>
<td>1.05</td>
</tr>
<tr>
<td>1998</td>
<td>eubacteria</td>
<td>Rickettsia prowazekii</td>
<td>Madrid E</td>
<td>1.10</td>
</tr>
<tr>
<td>1999</td>
<td>archaea</td>
<td>Pyrococcus horikoshii</td>
<td>OT3</td>
<td>1.80</td>
</tr>
<tr>
<td>1999</td>
<td>eubacteria</td>
<td>Helicobacter pylori</td>
<td>J99</td>
<td>1.64</td>
</tr>
<tr>
<td>1999</td>
<td>eubacteria</td>
<td>Chlamydia pneumoniae</td>
<td>CWL029</td>
<td>1.23</td>
</tr>
<tr>
<td>1999</td>
<td>eubacteria</td>
<td>Thermotoga maritima</td>
<td>MSB8</td>
<td>1.80</td>
</tr>
<tr>
<td>1999</td>
<td>eubacteria</td>
<td>Lactococcus lactis</td>
<td>IL1403</td>
<td>2.36</td>
</tr>
<tr>
<td>1999</td>
<td>eubacteria</td>
<td>Deinococcus radiodurans</td>
<td>R1</td>
<td>3.28</td>
</tr>
<tr>
<td>1999</td>
<td>archaea</td>
<td>Aeropyrum pernix</td>
<td>K12</td>
<td>1.67</td>
</tr>
<tr>
<td>1999</td>
<td>archaea</td>
<td>Pyrococcus abyssi</td>
<td>GE5</td>
<td>1.76</td>
</tr>
<tr>
<td>2000</td>
<td>eubacteria</td>
<td>Ureaplasma urealyticum</td>
<td>serovar 3</td>
<td>0.75</td>
</tr>
<tr>
<td>2000</td>
<td>eubacteria</td>
<td>Campylobacter jejuni</td>
<td>NCTC11168</td>
<td>1.64</td>
</tr>
<tr>
<td>2000</td>
<td>eubacteria</td>
<td>Chlamydia pneumoniae</td>
<td>AR39</td>
<td>1.23</td>
</tr>
<tr>
<td>2000</td>
<td>eubacteria</td>
<td>Chlamydia trachomatis</td>
<td>MoPn</td>
<td>1.07</td>
</tr>
<tr>
<td>2000</td>
<td>eubacteria</td>
<td>Neisseria meningitidis</td>
<td>MC58</td>
<td>2.18</td>
</tr>
<tr>
<td>2000</td>
<td>eubacteria</td>
<td>Neisseria meningitidis</td>
<td>Z2491</td>
<td>2.27</td>
</tr>
</tbody>
</table>
29 bacterial genome sequences finished in 2001

<table>
<thead>
<tr>
<th>Organism</th>
<th>Taxonomy</th>
<th>Accession</th>
<th>Genomic Links</th>
<th>Institute/University</th>
<th>Project</th>
<th>Publication Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Agrobacterium tumefaciens</em> C58-Dufour</td>
<td>Dickeyville Pea Pathogens</td>
<td>1085 K + 4600</td>
<td>N/A</td>
<td>University of Wisconsin-Madison</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Agrobacterium tumefaciens</em> C58-Caro</td>
<td>Dickeyville Pea Pathogens</td>
<td>1085 K + 4600</td>
<td>N/A</td>
<td>University of Wisconsin-Madison</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td></td>
<td>2586 K</td>
<td>N/A</td>
<td>National Institutes of Health</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td></td>
<td>4813 K</td>
<td>N/A</td>
<td>National Institutes of Health</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Listeria monocytogenes</em></td>
<td></td>
<td>3011 K</td>
<td>N/A</td>
<td>National Institutes of Health</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Salmonella typhimurium</em></td>
<td></td>
<td>4857 K</td>
<td>N/A</td>
<td>National Institutes of Health</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Salmonella typhi</em></td>
<td></td>
<td>4800 K</td>
<td>N/A</td>
<td>National Institutes of Health</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Streptococcus pneumoniae</em></td>
<td></td>
<td>2043 K</td>
<td>N/A</td>
<td>National Institutes of Health</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Cracking the code of microbes

1995: First complete genome sequence for a free living organism (*Haemophilus influenzae*) – cited more than 2,100 times!

2002: ~ 50 bacterial genomes completed

- 10 archaea (red squares) & 34 bacteria (blue squares).
- Obligate bacterial parasites are denoted by triangles.  

*Doolittle, R. Nature 416:697-700, 2002*
# Partial List of Completely Sequenced Genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size (MM base pairs)</th>
<th>Est. Genes*</th>
<th>Completed</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Archaea</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aeropyrum pernix K1</td>
<td>1.67</td>
<td>2,694</td>
<td>1999</td>
<td>Potential source of novel enzymes, etc.</td>
</tr>
<tr>
<td>Archaeoglobus fulgidus</td>
<td>2.18</td>
<td>2,407</td>
<td>1997</td>
<td>Potential source of novel enzymes, etc.</td>
</tr>
<tr>
<td>Methanobacterium</td>
<td>1.75</td>
<td>1,869</td>
<td>1997</td>
<td>Potential source of novel enzymes, etc.</td>
</tr>
<tr>
<td>thermoautotrophicum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pyrococcus abyssi</td>
<td>1.77</td>
<td>1,765</td>
<td>1999</td>
<td>Potential source of novel enzymes, etc.</td>
</tr>
<tr>
<td>Pyrococcus horikoshii</td>
<td>1.74</td>
<td>2,064</td>
<td>1998</td>
<td>Potential source of novel enzymes, etc.</td>
</tr>
<tr>
<td><strong>Bacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aquiflex aeolicus</td>
<td>1.55</td>
<td>1,522</td>
<td>1997</td>
<td>Potential source of novel enzymes, etc.</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>4.21</td>
<td>4,100</td>
<td>1997</td>
<td>Represents sporulating Gram-positive bacteria</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>1.64</td>
<td>1.654</td>
<td>2000</td>
<td>Food-borne pathogen</td>
</tr>
<tr>
<td>Chlamydia trachomatis</td>
<td>1.04</td>
<td>894</td>
<td>1998</td>
<td>Human pathogen</td>
</tr>
<tr>
<td>Chlamydia pneumoniae</td>
<td>1.23</td>
<td>1,052</td>
<td>1998</td>
<td>Human pathogen</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4.64</td>
<td>4,289</td>
<td>1998</td>
<td>Key model organism; human pathogen</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1.83</td>
<td>1,709</td>
<td>1995</td>
<td>Human pathogen; first free-living organism to have genome completely sequenced</td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>1.67</td>
<td>1,553</td>
<td>1997</td>
<td>Major cause of stomach ulcers</td>
</tr>
<tr>
<td>Helicobacter pylori J99</td>
<td>1.64</td>
<td>1,491</td>
<td>1999</td>
<td>Another H. pylori strain</td>
</tr>
<tr>
<td>Mycobacterium tuberculosi</td>
<td>4.41</td>
<td>3,918</td>
<td>1998</td>
<td>Causes tuberculosis</td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>0.58</td>
<td>480</td>
<td>1995</td>
<td>Genome is interesting because it is very small</td>
</tr>
<tr>
<td>Mycoplasma pneumoniae</td>
<td>0.82</td>
<td>677</td>
<td>1996</td>
<td>Leading cause of “walking pneumonia”</td>
</tr>
<tr>
<td>Rickettsia prowazekii</td>
<td>1.11</td>
<td>834</td>
<td>1998</td>
<td>Causes epidemic typhus</td>
</tr>
<tr>
<td>Synechocystis PCC6803</td>
<td>3.57</td>
<td>3,169</td>
<td>1996</td>
<td>Should help us understand photosynthesis</td>
</tr>
<tr>
<td>Treponema pallidum</td>
<td>1.14</td>
<td>1,031</td>
<td>1998</td>
<td>Causes venereal syphilis</td>
</tr>
<tr>
<td>Thermotoga maritima</td>
<td>1.86</td>
<td>1,846</td>
<td>1999</td>
<td>Potential source of novel enzymes, etc.</td>
</tr>
<tr>
<td>Ureaplasma urealyticum</td>
<td>0.75</td>
<td>611</td>
<td>2000</td>
<td>Sexually transmitted pathogen</td>
</tr>
<tr>
<td><strong>Eukaryota</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>~97.0</td>
<td>~19,000</td>
<td>1998</td>
<td>Worm – a key model organism</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>12.07</td>
<td>5,885</td>
<td>1996</td>
<td>Yeast – a key model organism</td>
</tr>
<tr>
<td>Human Chromosome 22**</td>
<td>33.46</td>
<td>600+</td>
<td>1999</td>
<td>First human chromosome to be fully sequenced</td>
</tr>
</tbody>
</table>

Source: NCBI; *excludes tRNA and rRNA genes; **euchromatic region
Completed and current Brazilian network-based genome projects

Genome sequencing in Brazil was initiated as a means of assimilating genomic technologies into the scientific community of the state of São Paulo, rather than to simply generate the complete genome data of selected organisms. The sequencing was done by a network of existing laboratories in universities and institutes that became known as ONSA — the Organization for Nucleotide Sequencing and Analysis. (This was actually a play on words as ‘onça’, pronounced ‘onsa’, is Portuguese for ‘jaguar’ thus being the Brazilian equivalent of a TIGR.) The strategy has been a success and has blossomed into a major programme in São Paulo, as well as stimulating similar efforts in other parts of Brazil. Complete and current genomics projects in Brazil are listed below.

Complete genome sequence of *citrus* strain of *Xylella fastidiosa*  
(September 1997–March 2000)\(^ {11}\).............. http://watson.fapesp.br/genoma.htm

FAPESP/LICR–Human Cancer Genome Project  

Complete genome sequence of *Xanthomonas citri*  

Complete genome sequence of *Xanthomonas campestris*  
(September 2000–September 2001)

SUCEST Sugar Cane EST Project  

Complete genome sequence of grapevine strain of *Xylella fastidiosa*  
(October 2000–August 2001)

Complete genome sequence of *Leifsonia xyli*  
(January 2001–present)

*Schistosoma mansoni* EST project .................. http://verjo18.iq.usp.br/schisto/

Complete genome sequence of *Chromobacterium violaceum*  
(December 2000–present).............................. http://www.brgene.lncc.br/
“Good” bacteria

- Make yogurt, cheese, sourdough bread
- Actinomycetes: Produce antibiotics (bacteria as factories)
- Plant growth promoting bacteria
- Break down dead matter
- Break down chemicals - bioremediation
- Food for many organisms
- “Good” bacteria in our bodies (trillions!)
Bacteria commonly found on the surfaces of the human body

<table>
<thead>
<tr>
<th>BACTERIUM</th>
<th>Skin</th>
<th>Conjunctiva</th>
<th>Nose</th>
<th>Pharynx</th>
<th>Mouth</th>
<th>Lower</th>
<th>Intestine</th>
<th>Anterior</th>
<th>urethra</th>
<th>Vagina</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Staphylococcus epidermidis</em> (1)</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em> (2)</td>
<td>+</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>+/-</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Streptococcus mitis</em></td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Streptococcus salivarius</em></td>
<td>++</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Streptococcus mutans</em> (3)</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Enterococcus faecalis</em> (4)</td>
<td>+/-</td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Streptococcus pneumoniae</em> (5)</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+/-</td>
</tr>
<tr>
<td><em>Streptococcus pyogenes</em> (6)</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+/-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Neisseria sp.</em> (7)</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Neisseria meningitidis</em> (8)</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Veillonellae sp.</em></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+/-</td>
</tr>
<tr>
<td><em>Enterobacteriaceae</em> (Escherichia coli) (9)</td>
<td>+/-</td>
<td>+/-</td>
<td>+/-</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td><em>Proteus sp.</em></td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em> (10)</td>
<td>+/-</td>
<td>+/-</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+/-</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em> (11)</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Bacteroides sp.</em></td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+/-</td>
</tr>
<tr>
<td><em>Bifidobacterium bifidum</em> (12)</td>
<td>++</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Lactobacillus sp.</em> (13)</td>
<td>+</td>
<td>++</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Clostridium sp.</em> (14)</td>
<td>+/-</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Clostridium tetani</em> (15)</td>
<td>+/-</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Corynebacteria</em> (16)</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Mycobacteria</em></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td>+/-</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Actinomycetes</em></td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Spirochetes</em></td>
<td>+</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Mycoplasmas</em></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td>+/-</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://www.bact.wisc.edu/Bact303/Bact303normalflora
Microbial genomes – the untapped resource

Don A. Cowan

Although the 1990s have ushered in the genome, they have also exposed our limitations for deriving structural and functional information. In parallel, molecular phylogeny has demonstrated that the majority of microbial genomes are currently inaccessible. Key objectives for the next century are the development of techniques for accessing ‘unculturable’ genomes, exploiting their biotechnologically valuable genes and products, and linking genome-sequence data to molecular structure and function.

<table>
<thead>
<tr>
<th>Group</th>
<th>Estimated total species</th>
<th>Known species</th>
<th>Proportion known of total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viruses</td>
<td>130,000⁴</td>
<td>5000</td>
<td>[4]:</td>
</tr>
<tr>
<td>Archaea</td>
<td>≈500⁵</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Bacteria</td>
<td>400,000</td>
<td>4800</td>
<td>[1.2]</td>
</tr>
<tr>
<td>Fungi</td>
<td>1,500,000</td>
<td>69,000</td>
<td>5</td>
</tr>
<tr>
<td>Algae</td>
<td>60,000</td>
<td>40,000</td>
<td>67</td>
</tr>
</tbody>
</table>

These values are substantially underestimated, possibly by 1 to 2 orders of magnitude.

<table>
<thead>
<tr>
<th>Group</th>
<th>Estimated total species</th>
<th>Known species</th>
<th>Proportion of culturability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seawater</td>
<td></td>
<td></td>
<td>0.001–0.100</td>
</tr>
<tr>
<td>Freshwater</td>
<td></td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>Mesotrophic lake</td>
<td></td>
<td></td>
<td>0.1–1.0</td>
</tr>
<tr>
<td>Unpolluted estuarine waters</td>
<td></td>
<td></td>
<td>0.1–3.0</td>
</tr>
<tr>
<td>Activated sludge</td>
<td></td>
<td></td>
<td>1–15</td>
</tr>
<tr>
<td>Sediments</td>
<td></td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>Soil</td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
</tbody>
</table>

These values date from the mid-1990s and will have increased by 10–50%.

[4] Indicates that these values are probably gross underestimates.

[5] 16S rRNA sequence analysis of different biotopes suggests that archael species represent a much higher proportion of in situ diversity than is indicated by microbial culture studies.
Advantages of using microbial genomes

- Procaryotic genomes are much smaller than eucaryotic ones
- No introns
- Little non-coding region between genes
- Most genes & gene functions known
- Comparative genomics can be done with many very similar genomes
- Large numbers of sequenced microbial genomes available
Minimal Genome Size –
Experimental approach


Knock-out 79 randomly selected genes from *Bacillus subtilis*:
Only 6 lethal, 73 are dispensable → 7.5% (6/79) of genome indispensable.

*B. subtilis* genome: $4.2 \times 10^6 \text{bp} \times 7.5\% = 3.2 \times 10^5 \text{bp}$
Average gene size is 1.25Kb, so the minimal genome size ≈ 254 genes.
Minimal Genome Size – Analytic approach

- Mushegian and Koonin, Trends in Genetics 12(9):334-336, 1996:

By comparison of complete bacterial genomes:

Orthologs among *E. coli*, *H. influenzae*, & *M. genitalium* genes

\[
\text{Overlapping orthologous genes (239)} + \text{[non-orthologous gene displacement]} - \text{[genes specific to parasitic bacteria or of functional redundancy]} = 256 \text{ genes}
\]
Minimal Genome Project

Scientists at TIGR Uncover the Minimal Number of Cellular Genes Needed for Life

ROCKVILLE, Md. - Dec. 8, 1999 - Researchers at The Institute for Genomic Research (TIGR) have uncovered the number of non-essential and essential genes necessary for life in Mycoplasma genitalium, the simplest known cell. The genetic information of Mycoplasma genitalium is 5,000 times smaller than the human genome, but this diminutive genome provides a starting point to define the essential genes required for life.

In the paper, published in the December 10 issue of Science, the minimum number of protein-coding genes required for cellular life in the laboratory is between 265 and 350. Surprisingly, this minimal gene set includes about 100 genes of unknown function. This finding draws into question a prevailing assumption that the basic molecular mechanisms underlying cellular life are understood, at least in broad outline.

"Defining the minimal genome is a very fundamental problem, and no one else seems to be approaching it experimentally," says Hamilton Smith, Nobel laureate and a TIGR investigator at the time this work was initiated.
Global Transposon Mutagenesis
and a Minimal Mycoplasma Genome

Clyde A. Hutchison III,1,2* Scott N. Peterson,1*† Steven R. Gill,1
Robin T. Cline,1 Owen White,1 Claire M. Fraser,1
Hamilton O. Smith,1† J. Craig Venter1‡§

Mycoplasma genitalium with 517 genes has the smallest gene complement of
any independently replicating cell so far identified. Global transposon mu-
tageneis was used to identify nonessential genes in an effort to learn whether
the naturally occurring gene complement is a true minimal genome under
laboratory growth conditions. The positions of 2209 transposon insertions in
the completely sequenced genomes of M. genitalium and its close relative M.
pneumoniae were determined by sequencing across the junction of the trans-
poson and the genomic DNA. These junctions defined 1354 distinct sites of
insertion that were not lethal. The analysis suggests that 265 to 350 of the 480
protein-coding genes of M. genitalium are essential under laboratory growth
conditions, including about 100 genes of unknown function.
Minimum genome size

• How many genes are needed to carry out minimal cell functions for life?
• Some clues from bacteria:
  – *M. genitalium* 467 ORFs
  – *M. pneumoniae* has 677 ORFs
• 250 -350 genes estimated minimum

Mycoplasma mutated
265-350 genes are essential
Bacterial proteomes

- *M. genitalium* with 480 genes has the simplest known genome
- 33% of its genome is expressed during optimal growth conditions
- Remaining proteins are likely expressed during other conditions, in undetectable amounts, or not isolated
Horizontal gene transfer

• before microbial genome sequences became available most of the focus of microbial evolution was on ‘vertical’ transmission of genetic information – mutation recombination and rearrangement within the clonal lineage of a single microbial population
• genome sequences have demonstrated that horizontal transfer of genes (between different types of organisms) are widespread and may occur between phylogenetically diverse organisms
• generally speaking, essential genes (such as 16S rRNA) are unlikely to be transferred because the potential host most likely already contains genes of this type that have co-evolved with the rest of its cellular machinery and and cannot be displaced
• genes encoding non-essential cellular processes of potential benefit to other organisms are far more likely to be transferred (e.g. those involved in catabolic processes)
Scope of Horizontal Gene Transfer (HGT) in bacteria

Length of bars represent amount of coding DNA, native is blue, Foreign due to mobile elements is yellow, other is red. Numbers are the % of foreign DNA

Species and strain specific genetic diversity

- although genome sequencing and analysis is very useful when comparing phylogenetically distant taxa, it is also of interest to examine the genomes of very closely related microorganisms.
- this allows a more quantitative approach for examining the relationships between genotype and phenotype.
- complete genome sequences have been determined for two species of the genus *Chlamydia* (pneumoniiae and trachomatis).
- although the overall genome structure was quite similar, *C. pneumoniae* contained an additional 214 genes most of which have an unknown function.
- two strains of the bacterium *Helicobacter pylori* have been completely sequenced (26695 and J99).
- overall the two strains were very similar genetically with only 6% of genes being specific to each strain.
Case study - *Deinococcus radiodurans*

- discovered in 1956 by Arthur W. Anderson (during experiments in which packaged food was sterilized using radiation instead of heat) at Oregon Agricultural Experiment Station in Corvallis, USA.
- a non-pathogenic, gram-(+), mesophilic, non-spore-forming, non-motile, spherical bacterium (tetrad-forming coccus produces pink to reddish colonies).
- shows remarkable resistance to a range of damage caused by ionizing radiation, desiccation, UV radiation, oxidizing agents, & electrophilic mutagens.
- can endure 1.5 million rads of radiation (i.e., it can withstand radiation 3,000 times what it would take to kill a human).
Deinococcus radiodurans –

the most radiation-resistant organism known

- genome (total of 3.3 Mb) consists of two chromosomes (2.6 and 0.4 Mb) a megaplasmid (177 kb) and a small plasmid (44 kb).
- considerable genetic redundancy was observed in both the chromosomal and plasmid sequences.
- numerous systems for DNA repair, DNA damage export were identified.
- highly efficient DNA repair system: can rapidly repair DNA double strand breaks induced by radiation without rearrangement or increased mutation frequency.
- a significant proportion of the ORFs identified had no database matches - these may be involved in unique cellular adaptations to radiation and stress response.
Genome Sequence of the Radioresistant Bacterium

*Deinococcus radiodurans* R1

Owen White,¹ Jonathan A. Eisen,¹ John F. Heidelberg,¹
Erin K. Hickey,¹ Jeremy D. Peterson,¹ Robert J. Dodson,¹
Daniel H. Haft,¹ Michelle L. Gwinn,¹ William C. Nelson,¹
Delwood L. Richardson,¹ Kelly S. Moffat,¹ Haiying Qin,¹
Lingxia Jiang,¹ Wanda Pamphile,¹ Marie Crosby,¹ Mian Shen,¹
Jessica J. Vamathevan,¹ Peter Lam,¹ Lisa McDonald,¹
Terry Utterback,¹ Celeste Zalewski,¹ Kira S. Makarova,²
L. Aravind,² Michael J. Daly,³ Kenneth W. Minton,³
Robert D. Fleischmann,¹ Karen A. Ketchum,¹ Karen E. Nelson,¹
Steven Salzberg,¹ Hamilton O. Smith,¹* J. Craig Venter,¹*
Claire M. Fraser,¹†

The complete genome sequence of the radiation-resistant bacterium *Deino-
coccus radiodurans* R1 is composed of two chromosomes (2,648,638 and
412,348 base pairs), a megaplasmid (177,466 base pairs), and a small plasmid
(45,704 base pairs), yielding a total genome of 3,284,156 base pairs. Multiple
components distributed on the chromosomes and megaplasmid that contribute
to the ability of *D. radiodurans* to survive under conditions of starvation,
oxidative stress, and high amounts of DNA damage were identified. *Deinococcus
radiodurans* represents an organism in which all systems for DNA repair, DNA
damage export, desiccation and starvation recovery, and genetic redundancy
are present in one cell.
### Table 1. General features of the *D. radiodurans* genome.

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Length (bp)</th>
<th>Average ORF length (bp)</th>
<th>Protein coding regions</th>
<th>GC content</th>
<th>Repeat content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome I</td>
<td>2,648,638</td>
<td>913</td>
<td>90.8%</td>
<td>57.0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Chromosome II</td>
<td>412,348</td>
<td>1,044</td>
<td>93.5%</td>
<td>66.7%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Megaplasmid</td>
<td>177,466</td>
<td>1,100</td>
<td>90.4%</td>
<td>63.2%</td>
<td>9.2%</td>
</tr>
<tr>
<td>Plasmid</td>
<td>45,704</td>
<td>928</td>
<td>80.9%</td>
<td>56.1%</td>
<td>13.0%</td>
</tr>
<tr>
<td>All</td>
<td>3,284,156</td>
<td>937</td>
<td>90.9%</td>
<td>66.5%</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

**Unusual Characteristics of the Cell Wall**
Table 6. DNA repair genes and pathways encoded by D. radiodurans.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Genes in D. radiodurans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide excision repair</td>
<td>Corresponds to UV endonuclease α, urA = uraAB, urD = urB [31]</td>
</tr>
<tr>
<td>Transcription repair couplings</td>
<td>Experiments suggest that this process may not be present [34]</td>
</tr>
<tr>
<td>UV excision repair</td>
<td>Corresponds to UV endonuclease β (uvrCDE)</td>
</tr>
<tr>
<td>Base excision repair</td>
<td>AKA 3-methyl-guanine glycosylase</td>
</tr>
<tr>
<td></td>
<td>MPG 3-methyl-guanine glycosylase</td>
</tr>
<tr>
<td></td>
<td>Ung uracil DNA glycosylase</td>
</tr>
<tr>
<td></td>
<td>Mug G-U mismatch glycosylase</td>
</tr>
<tr>
<td></td>
<td>Ung2 uracil DNA glycosylase? [35]</td>
</tr>
<tr>
<td></td>
<td>MutS-Mhp 3′pg Fapy and 8-oxo-guanine DNA glycosylase</td>
</tr>
<tr>
<td>MutLV-Nth-1</td>
<td>Likely a G.A glycosylase because most similar to MutLV’s</td>
</tr>
<tr>
<td>MutLV-Nth-2</td>
<td>Thymine glycol glycosylase from [36]</td>
</tr>
<tr>
<td>MutLV-Nth-3</td>
<td>Second Fapy glycosylase from [36]</td>
</tr>
<tr>
<td>MutLV-Nth-4</td>
<td>Unknown</td>
</tr>
<tr>
<td>AP endonuclease</td>
<td>May also be an exonuclease</td>
</tr>
<tr>
<td>Mismatch excision repair</td>
<td>Absence of MutH suggests different strand recognition system than E. coli</td>
</tr>
<tr>
<td>Recombinational repair</td>
<td>Initiations</td>
</tr>
<tr>
<td></td>
<td>Nearly complete RecF pathway (RecG missing)</td>
</tr>
<tr>
<td></td>
<td>Absence of RecF and RecC orthologs suggests that this gene functions differently than in E. coli [27]</td>
</tr>
<tr>
<td></td>
<td>RecA Homology to RecD and RecF11 suggests a role in DSB repair [27]</td>
</tr>
<tr>
<td></td>
<td>Recombinase; may also regulate transcription of other genes</td>
</tr>
<tr>
<td></td>
<td>DNA polymerases</td>
</tr>
<tr>
<td></td>
<td>Repair replication polymerase [39]</td>
</tr>
<tr>
<td></td>
<td>Chromosomal replication polymerase</td>
</tr>
<tr>
<td></td>
<td>DNA polymerase of unknown function (PolX family)</td>
</tr>
<tr>
<td>Ligation</td>
<td>Dnvj Ligation activity is required for all excision and recombinational repair pathways</td>
</tr>
<tr>
<td>dNTP pools, cleanup</td>
<td>dNTP cleanup: more copies than any other prokaryote</td>
</tr>
<tr>
<td>MutS and Nudix family</td>
<td>Ribonucleoside reductase</td>
</tr>
<tr>
<td></td>
<td>Ribonucleoside ribotransferase</td>
</tr>
<tr>
<td>Induction</td>
<td>LexA Transcription repressor, possibly for SOS response</td>
</tr>
<tr>
<td>Other</td>
<td>RadA/SSM5 DNA damage response?</td>
</tr>
<tr>
<td></td>
<td>HepA Likely role in transcription or DNA repair (or both): member of SNF2 family [24]</td>
</tr>
<tr>
<td></td>
<td>MutS2 Possible role in recognizing mismatches but not likely involved in mismatch repair [24]</td>
</tr>
<tr>
<td></td>
<td>XseA Exonuclease VII subunit (but XseB is absent)</td>
</tr>
<tr>
<td></td>
<td>UvrA2 Export of damaged DNA?</td>
</tr>
<tr>
<td>Extracellular nucleases</td>
<td>Degradation of exported DNA?</td>
</tr>
<tr>
<td>SSB</td>
<td>Single-strand DNA binding protein</td>
</tr>
<tr>
<td>CssA</td>
<td>May recruit RecA to cell membrane</td>
</tr>
<tr>
<td>XerC</td>
<td>Site-specific recombinase</td>
</tr>
</tbody>
</table>


Why is *Deinococcus radiodurans* so resistant to ionizing radiation?

John R. Battista, Ashlee M. Earl and Mie-Jung Park

The publication of the fully assembled and annotated sequence of the *Deinococcus radiodurans* R1 genome is expected during 1999 and, if the anecdotal information released to the press is accurate, analysis of the sequence has not revealed much that can be used to explain this organism’s extraordinary capacity to tolerate DNA damage. It appears that most, if not all, of the typical complement of prokaryotic DNA-repair proteins are found in *D. radiodurans*. This observation suggests two equally intriguing possibilities: (1) *D. radiodurans* uses the same DNA-repair strategies as other prokaryotes but does so in a manner that is somehow more effective than in other species or (2) *D. radiodurans* uses a DNA-repair system that has novel components. Nevertheless, the precise DNA damage. Given the scale of the capacity of *D. radiodurans* to survive massive DNA damage, we assume that this organism has evolved specific and distinctive mechanisms to deal with such damage. We have identified several observations that hint at possible mechanisms.
The Effect of High Doses of $\gamma$-Radiation on Physical Matter

To illustrate the magnitude of an irradiation dose from which *Deinococcus radiodurans* recovers easily, we irradiated a glass (Pyrex) beaker (left) to 17,500 Gy. It can be seen that this dose of $\gamma$-rays has turned the glass brown and also has made the glass brittle. Imagine what this dose does to DNA! This radiation dose will break a long DNA double-stranded helix at many places causing fragmentation.

Fig. 1. The ability of *Deinococcus radiodurans* R1 to survive the accumulation of DNA double-strand breaks following exposure to a 3000-Grey ($\text{Gy}$) dose of $\gamma$ radiation. Lane one contains a lambda size standard; lane two contains chromosomal DNA prepared from an untreated culture; lane three contains chromosomal DNA prepared from a culture immediately after irradiation; lanes four to six contain chromosomal DNA prepared from a culture three, six, and nine hours post-irradiation, respectively. Assuming that the size of the *D. radiodurans* genome is 3.2 Mb\textsuperscript{22}, 3000 Gy generates 120 dsbs per genome or, on average, one dsb for every 27 kb\textsuperscript{22}.
Fig. 1. Representative survival curves for Deinococcus radiodurans R1 (squares) and Escherichia coli B/r (diamonds) following exposure to γ radiation. 1E-1 is 1×10⁻¹, or 0.1; each designation on the y-axis therefore represents a reduction in viability by a factor of 10. The D₅₀ dose (i.e., the average dose of ionizing radiation that is required to inactivate a single colony-forming unit) for the E. coli culture is 50 Gray (Gy), approximately 200 times lower than that of D. radiodurans. D. radiodurans has a characteristic shoulder of resistance to approximately 5000 Gy, in which there is no loss of viability. Above 5000 Gy, there is an exponential decline in viability and a D₅₀ dose of between 6000 Gy–7000 Gy for cultures in exponential phase[10].

Fig. 2. Schematic representation of the response of Deinococcus radiodurans to ionizing-radiation-induced DNA damage. As DNA replication, degradation and recombination repair are coordinately regulated, it is proposed that these processes are sensitive to, or responsible for the generation of, intracellular signals. It is believed that the ability of D. radiodurans to survive ionizing-radiation-induced DNA damage involves recombination repair, the regulation of DNA replication and the export of damaged nucleotides. The nature of the inhibitory protein controlling DNA degradation is unknown.
Engineering *Deinococcus radiodurans* for metal remediation in radioactive mixed waste environments

Hassan Brim¹, Sara C. McFarlan², James K. Fredrickson³, Kenneth W. Minton¹, Min Zhai¹, Lawrence P. Wackett⁵, and Michael J. Daly¹*

¹Department of Pathology, Uniformed Services University of the Health Sciences, Bethesda, MD 20814. ²Department of Biochemistry, Biological Process Technology Institute and Center for Biodegradation Research and Informatics, Gertler Laboratory, University of Minnesota, St. Paul, MN 55108. ³Pacific Northwest National Laboratory, Richland, WA 99352. *Corresponding author (mjdaly@pnl.gov).

Received 2 September 1999; accepted 12 November 1999

We have developed a radiation resistant bacterium for the treatment of mixed radioactive wastes containing ionic mercury. The high cost of remediating radioactive waste sites from nuclear weapons production has stimulated the development of bioremediation strategies using *Deinococcus radiodurans*, the most radiation resistant organism known. As a frequent constituent of these sites is the highly toxic ionic mercury (Hg) (II), we have generated several *D. radiodurans* strains expressing the cloned Hg (II) resistance gene (*merA*) from *Escherichia coli* strain BL308. We designed four different expression vectors for this purpose, and compared the relative advantages of each. The strains were shown to grow in the presence of both radiation and ionic mercury at concentrations well above those found in radioactive waste sites, and to effectively reduce Hg (II) to the less toxic volatile elemental mercury. We also demonstrated that different gene clusters could be used to engineer *D. radiodurans* for treatment of mixed radioactive wastes by developing a strain to detoxify both mercury and toluene. These expression systems could provide models to guide future *D. radiodurans* engineering efforts aimed at integrating several remediation functions into a single host.
Figure 4. Effect of continuous exposure to γ-radiation and Hg (II) on the growth of strains, containing different copy numbers of the mer operon. Two TGY agar plates (A and B), and two TGY agar plates containing 30 μM Merbromin (C and D) were spotted with 1 x 10^9 cells of each of the indicated strains. Following plate inoculation, plates B and D were placed into the 32ºC irradiator (60 Gy / h) for incubation for 5 days. The control plates (A and C) were incubated at the same temperature in the absence of radiation for the same time.

Figure 6. (A) Mercuric reductase assay. Hg(II)-dependent NADPH oxidation catalyzed by cell extracts prepared from the strains R1 (mer-, tod+; wild type), MD735 (mer+), MD761 (mer+), MD767 (mer+), MD764 (mer-, tod+), MD736 (mer+), and MD737 (mer+) were monitored spectrophotometrically. Decreasing absorbance at 340 nm corresponds to a decreasing NADPH concentration. (B) Mercury volatilization by engineered D. radiodurans.
Case study - *Neisseria meningitidis*

- *N. meningitidis* causes bacterial meningitis and is therefore an important pathogen
- genome is 2.2 megabases in size
- 2121 ORF’s were identified with many having extremely variable G+C% (recently acquired genes)
- many of these recently acquired genes are identified as cell surface proteins
- there is a remarkable abundance and diversity of repetitive DNA sequences
- nearly 700 neisserial intergenic mosaic elements (NIME’s) - 50 to 150 bp repeat elements
- these repeat elements may be involved in enhancing recombinase specific horizontal gene transfer
Case study - *Borellia burgdorferi*


- *B. burgdorferi* is a spirochaete which causes Lyme disease
- it has a 0.91 Mb linear genome and at least 17 linear and circular plasmids which total 0.53 Mb
- 853 predicted ORF’s identified - these encode a basic set of proteins for DNA replication, transcription, translation and energy metabolism
- no genes encoding proteins involved in cellular biosynthetic reactions were identified - appears to have evolved via gene loss from a more metabolically competent precursor
- there is significant amount of genetic redundancy in the plasmid sequences although a biological role has not been determined
- it is possible the these plasmids undergo frequent homologous recombination in order to generate antigenic variation in surface proteins
GC skew Analysis

- Biological Background:
  - Bacterial genomes and some other types of chromosomes exhibit a regional GC skew (strand-specific bias in G:C ratio) related to the direction of replication. In *E. coli*, the skew (calculated as \((G-C)/(G+C)\)) changes sign at the origin and terminus of replication. Differential mutation in the leading and lagging strand of replication has been proposed as a mechanism which could explain this phenomenon.

The genomics revolution

• Genome sequences allow the following questions to be asked:

  – What are the minimal requirements for a “living” organism?
  – How has evolution streamlined microbial genomes?
  – How are microbes related to each other?
  – What are the genomic differences between:
    • obligate parasites and free-living organisms?
    • Phototrophic and chemotrophic organisms?
    • Organotrophic and lithotrophic organisms?
    • Mesophiles and Thermophiles?
    • Pathogenic and non-pathogenic strains?
Applications of microbial genome data

- Pathogen identification in tissue sample
- Virulence gene targets used for diagnosis & prognosis
- Antibiotic resistance genes for determining best treatment
- Identification of genes required for pathogenesis will allow targeted drug/vaccine development
- Determination of gene function in “simple” organisms will help understand function of genes in eukaryotes.
- Discover enzymes which might have industrial applications.
  Develop better “factories” for producing drugs, chemicals & foods, & for biodegradation, bioremediation
- Identify new bacteria & new disease-causing agents
Archaea genomes

*Methanococcus jannaschii*

The first archaeon sequenced

(Science 273:580-586, 1996)

- *Methanococcus jannaschii* lives in ocean thermal vents at 85 °C
- Circular DNA 1.7 Mb with 1738 protein-encoding genes
- Contains 1 large and 2 small chromosomes
- 58% of its genes do not resemble any known gene
Genomic insights

• Prevalence of gene clusters and gene islands (genomic islands). Horizontal gene transfer between microbes, mediated by phage or phage-like elements, appears to be common.

• Closely related bacteria can have significant differences in genome content and structure.

• Intracellular bacterial genomes have reduced genome size.
  e.g., *Buchnera*
  – endosymbiont of aphids
  – 50 million years of genetic isolation
  – only observe gene loss
  e.g., *Rickettsia* – 25% non-coding (vs. 10% for most other bacteria) – evidence of decay
  e.g., *Mycobacterium leprae* – massive decay
**Rickettsia prowazekii genome**

(Nature 396:133-143, 1998)

- The length of complete genome sequence is 1,111,523 bp. This genome contains 834 protein-coding genes.
- Pseudogenes: about 25% of non-coding sequences
  - gene remnants that have been degraded by mutations
  - ‘Reductive evolution’
- Many amino acid and nucleoside biosynthesis genes are absent from *R. prowazekii* and mitochondria.
- More closely related to mitochondrial genomes than any other bacteria
- *Rickettsia* and mitochondria probably share an proteobacterial ancestor and a similar evolutionary history.
The sequencing of *E. coli* genome

- The genome of the non-pathogenic K-12 laboratory strain *E. coli* MG1655 was completely sequenced by Blattner *et al.* (Science 277:1453-1462, 1997)

- The genome sequence of enterohaemorrhagic strain *E. coli* EDL933 (a reference strain for O157:H7) has been completed by Perna *et al.* (Nature 409:529-533, 2001).
E. coli O157:H7

• *E. coli* O157:H7 is a rare but virulent strain of *E. coli*, which lives in the intestinal tracts of mammals and man and causes serious and potentially fatal diseases.

• O157 can survive refrigeration and freezer storage. The major food sources carrying this organism are undercooked hamburger and roast beef, raw milk, improperly processed cider.

• Since 1982, there have been at least 16 major outbreaks in the US. Some 22 deaths have been recorded. CDC experts estimate there may be as many as 20,000 cases per year.
What makes *E. coli* O157:H7 so dangerous?

The pathogenicity (ability to cause damage) and virulence (degree of pathogenicity) of O157:H7 depend on:

1. The genes encoding the so-called Shiga toxin, such as *stx1* and *stx2*;
2. The small, circular DNA molecules that encode “virulence factors”;
3. Pathogenicity island — a section of chromosominal DNA containing many genes that contribute to pathogenicity.
Summary

• Pathogenic *E. coli* O157:H7 shares the common ancestor with non-pathogenic *E. coli* strain based on the genome-scale comparative analysis.

• Lateral gene transfer contribute much more than previously expected to the strain-specific pathogenesis in *E. coli* O157: H7.

• Preferential transversions (G↔T) may be attributed to the transcription-coupled repair of damage associated with oxidative stress.
Typical characteristics of fungal genomes

- **Small**
  - *Saccharomyces cerevisiae* 6 MB (5651 genes)
  - *Schizosaccharomyces pombe* 13.8 MB (4824 genes)
  - *Aspergillus nidulans* ~13 MB (~12000 genes)

Compared to:
- *Drosophila melanogaster* 137MB (~15000 genes)
- *Arabidopsis thaliana* 126MB (~15000 genes)
- *Homo sapiens* 1300 MB (~30000 genes)
Typical characteristics of fungal genomes

• Introns
  – few, often none (43% of *S. pombe* genes, total 4730)
  – small: 50-200bp compared to ≥10 kb in mammals
  – *S. pombe* mean 81bp, mold 48bp; range 29-819 bp
Typical characteristics of fungal genomes

• Little repetitive DNA – single copy genes
  – 50-60% of nuclear genome is transcribed into mRNA in *S. cerevisiae*
  – 33% in *Schizophyllum commune* (a basidiomycete fungus), *Bremia lactucae* (oomycete)

Compared to:
  – 1% transcribed in humans
The genome sequence of *Schizosaccharomyces pombe*


We have sequenced and annotated the genome of fission yeast (*Schizosaccharomyces pombe*), which contains the smallest number of protein-coding genes yet recorded for an eukaryote: 4,624. The centromeres are between 35 and 110 kilobases (kb) and contain related repeats including a highly conserved 1.8-kb region. Regions upstream of genes are longer than in budding yeast (*Saccharomyces cerevisiae*), possibly reflecting more extended control regions. Some 43% of the genes contain introns, of which there are 4,730. Fifty genes have significant similarity with human disease genes; half of these are cancer related. We identify highly conserved genes important for eukaryotic cell organization including those required for the cytoskeleton, compartmentation, cell-cycle control, proteolysis, protein phosphorylation and RNA splicing. These genes may have originated with the appearance of eukaryotic life. Fewer similarly conserved genes that are important for multicellular organization were identified, suggesting that the transition from prokaryotes to eukaryotes required more new genes than the transition from unicellular to multicellular organization.

<table>
<thead>
<tr>
<th>Table 1 Genome content for the three chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (bp)</td>
</tr>
<tr>
<td>Chromosome 1</td>
</tr>
<tr>
<td>Chromosome 2</td>
</tr>
<tr>
<td>Chromosome 3</td>
</tr>
<tr>
<td>Whole genome</td>
</tr>
</tbody>
</table>

*Mean gene length excluding introns.

† Gene density, given as average bp per gene.
Figure 2 Intergene regions. Distribution of intergene regions given for all genes and for divergent and convergent pairs of genes, for both S. pombe and S. cerevisiae. A total of 4,890 intergene regions from S. pombe were analysed from a database prepared just before completion of the whole genome, and 5,788 intergene regions from S. cerevisiae were analysed. Histograms show the number of regions in 200-bp bins.
Figure 3 Comparison of proteins in *S. pombe* (Sp.), *S. cerevisiae* (Sc.) and *C. elegans* (C.e.). **a.** Pie chart comparing the homology of proteins of *S. pombe* with those of *S. cerevisiae* and *C. elegans*. **b.** Pie chart comparing the homology of proteins of *S. cerevisiae* with those of *S. pombe* and *C. elegans*. For example, Sp proteins in Sc and C.e. means *S. pombe* proteins with homologues found in *S. cerevisiae* and *C. elegans*. The absolute numbers of proteins are given for both yeasts.
<table>
<thead>
<tr>
<th>Interpro accession no.</th>
<th>S. pombe</th>
<th>S. cerevisiae</th>
<th>H. sapiens</th>
<th>D. melanogaster</th>
<th>C. elegans</th>
<th>A. thaliana</th>
<th>Interpro name</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPR001687</td>
<td>213</td>
<td>1</td>
<td>267</td>
<td>1</td>
<td>436</td>
<td>5</td>
<td>231</td>
</tr>
<tr>
<td>IPR001680</td>
<td>114</td>
<td>2</td>
<td>97</td>
<td>3</td>
<td>277</td>
<td>8</td>
<td>183</td>
</tr>
<tr>
<td>IPR000719</td>
<td>111</td>
<td>3</td>
<td>119</td>
<td>2</td>
<td>579</td>
<td>3</td>
<td>377</td>
</tr>
<tr>
<td>IPR000504</td>
<td>80</td>
<td>4</td>
<td>61</td>
<td>5</td>
<td>307</td>
<td>7</td>
<td>182</td>
</tr>
<tr>
<td>IPR001620</td>
<td>67</td>
<td>5</td>
<td>63</td>
<td>4</td>
<td>155</td>
<td>20</td>
<td>101</td>
</tr>
<tr>
<td>IPR001841</td>
<td>44</td>
<td>6</td>
<td>33</td>
<td>12</td>
<td>215</td>
<td>15</td>
<td>120</td>
</tr>
<tr>
<td>IPR001440</td>
<td>35</td>
<td>7</td>
<td>33</td>
<td>12</td>
<td>150</td>
<td>21</td>
<td>92</td>
</tr>
<tr>
<td>IPR001066</td>
<td>36</td>
<td>8</td>
<td>46</td>
<td>8</td>
<td>44</td>
<td>64</td>
<td>45</td>
</tr>
<tr>
<td>IPR001617</td>
<td>33</td>
<td>9</td>
<td>42</td>
<td>9</td>
<td>75</td>
<td>40</td>
<td>67</td>
</tr>
<tr>
<td>IPR000822</td>
<td>32</td>
<td>10</td>
<td>51</td>
<td>7</td>
<td>712</td>
<td>2</td>
<td>403</td>
</tr>
<tr>
<td>IPR001357</td>
<td>14</td>
<td>23</td>
<td>10</td>
<td>30</td>
<td>24</td>
<td>82</td>
<td>17</td>
</tr>
<tr>
<td>IPR000062</td>
<td>8</td>
<td>29</td>
<td>9</td>
<td>31</td>
<td>8</td>
<td>99</td>
<td>9</td>
</tr>
<tr>
<td>IPR002064</td>
<td>5</td>
<td>32</td>
<td>5</td>
<td>35</td>
<td>4</td>
<td>102</td>
<td>6</td>
</tr>
<tr>
<td>IPR001206</td>
<td>6</td>
<td>31</td>
<td>6</td>
<td>34</td>
<td>12</td>
<td>94</td>
<td>13</td>
</tr>
<tr>
<td>IPR000002</td>
<td>5</td>
<td>32</td>
<td>3</td>
<td>37</td>
<td>3</td>
<td>103</td>
<td>4</td>
</tr>
<tr>
<td>IPR001452</td>
<td>21</td>
<td>16</td>
<td>23</td>
<td>18</td>
<td>220</td>
<td>14</td>
<td>82</td>
</tr>
<tr>
<td>IPR001849</td>
<td>21</td>
<td>16</td>
<td>26</td>
<td>16</td>
<td>253</td>
<td>11</td>
<td>89</td>
</tr>
<tr>
<td>IPR000387</td>
<td>9</td>
<td>28</td>
<td>11</td>
<td>29</td>
<td>112</td>
<td>29</td>
<td>47</td>
</tr>
<tr>
<td>IPR001138</td>
<td>27</td>
<td>13</td>
<td>52</td>
<td>6</td>
<td>0</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>IPR002293</td>
<td>21</td>
<td>16</td>
<td>32</td>
<td>13</td>
<td>43</td>
<td>65</td>
<td>36</td>
</tr>
<tr>
<td>IPR000853</td>
<td>7</td>
<td>30</td>
<td>2</td>
<td>38</td>
<td>26</td>
<td>80</td>
<td>20</td>
</tr>
</tbody>
</table>

Domain identifiers are from Interpro, which integrates PROSITE, PRINTS and Pfam. Only domains within the most frequent 40 found in S. pombe are given. The numbers of proteins with these domains and their ranking is given for S. pombe and the other eukaryotes listed. At the right end of the table is a classification of 1–3; see text for an explanation. NA, not applicable.
The genome of *Neurospora crassa*

- holds a total of about 43 Mb distributed on 7 chromosomes (the genome is not yet completely sequenced). The estimation of protein coding genes is about 13,000.

- GC content:
  - Protein coding regions—59%
  - Noncoding regions—49%

- Biased codon usage
  - Strong preference for C at the 3rd position. Codons ending in A are generally used rarely. Of the stop codons TAA is the most frequent one.

Stephanie Edelmann, SE & Staben, C. Exp. Mycol. 18:70-81 (1994)
The genome of *Neurospora crassa*

- **Introns**
  --- Over 80% of *N. crassa* protein coding genes have introns ranging from 21 to 859 bp. The mean intron size is 101 bp.
  --- Patterns for the intron splice sites, and translation initiation sites
    -- initiation: CAMMATGGCT(ATG)
    -- 5’ intron donor: G^GTAAGTnnYCNYY(GTRNGT)
    -- internal branch point: WRCTRACMnnnnnnYY(CTRAC)
    -- 3’ intron acceptor: WACAG^(YAG)

Stephanie Edelmann, SE & Staben, C. Exp. Mycol. 18:70-81 (1994)