

91學年度上學期「生物資訊學」課程

Microbial genomics

張傳雄

國立陽明大學 遺傳學研究所

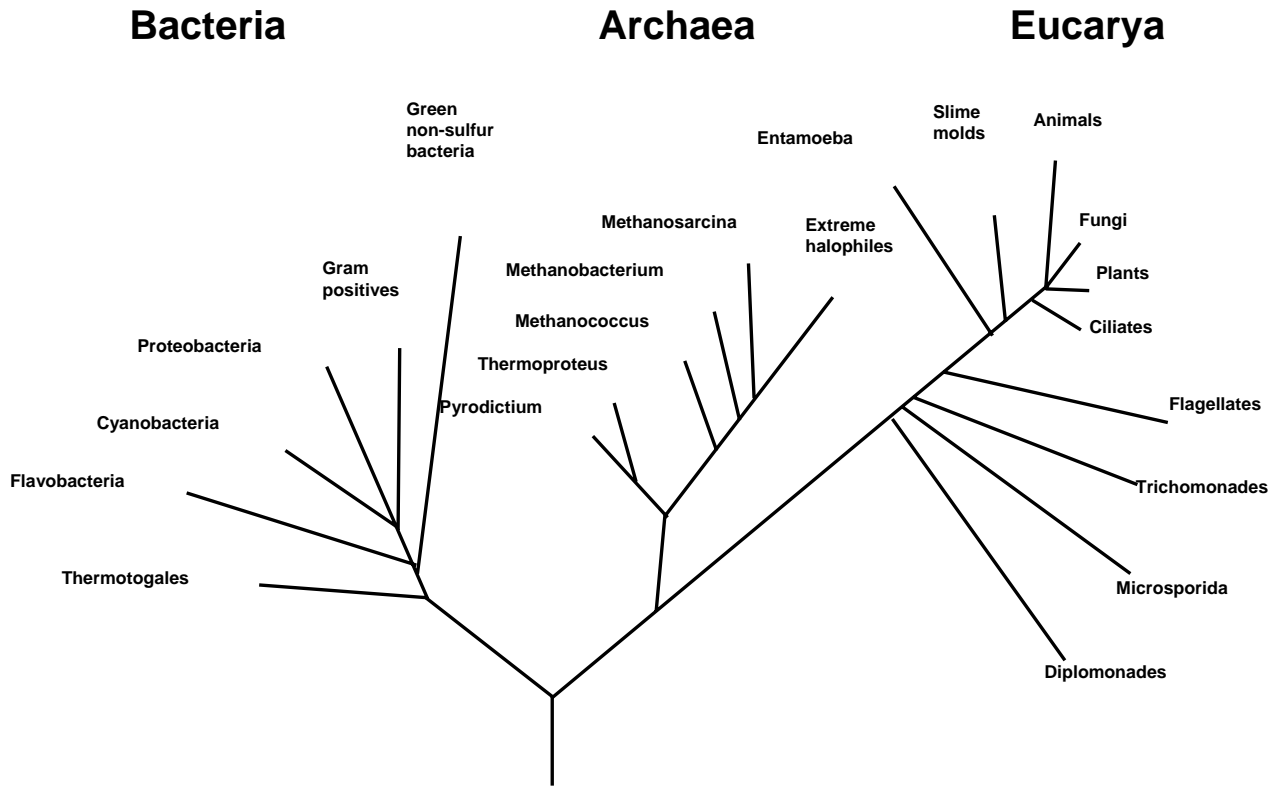
01-06-2003



Why study microbial genomes?

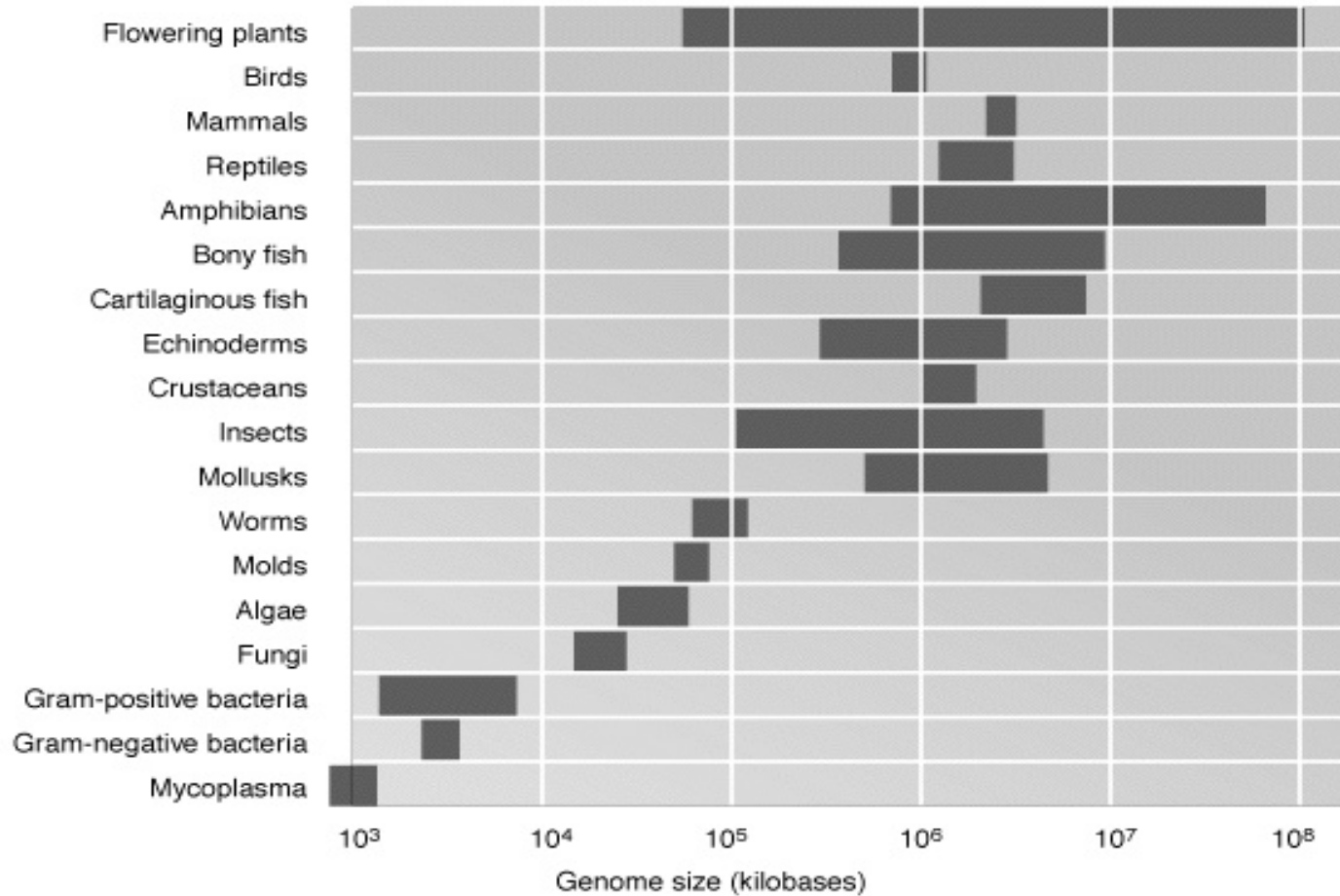
- studies on whole genome sequences give us a complete genomic blueprint for an organism. We can examine how all the parts operate cooperatively to influence the activities and behavior of an entire organism – a complete understanding of the biology of an organism. Microbes provide an excellent starting 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 Carl R. Woese)

Genome size variation

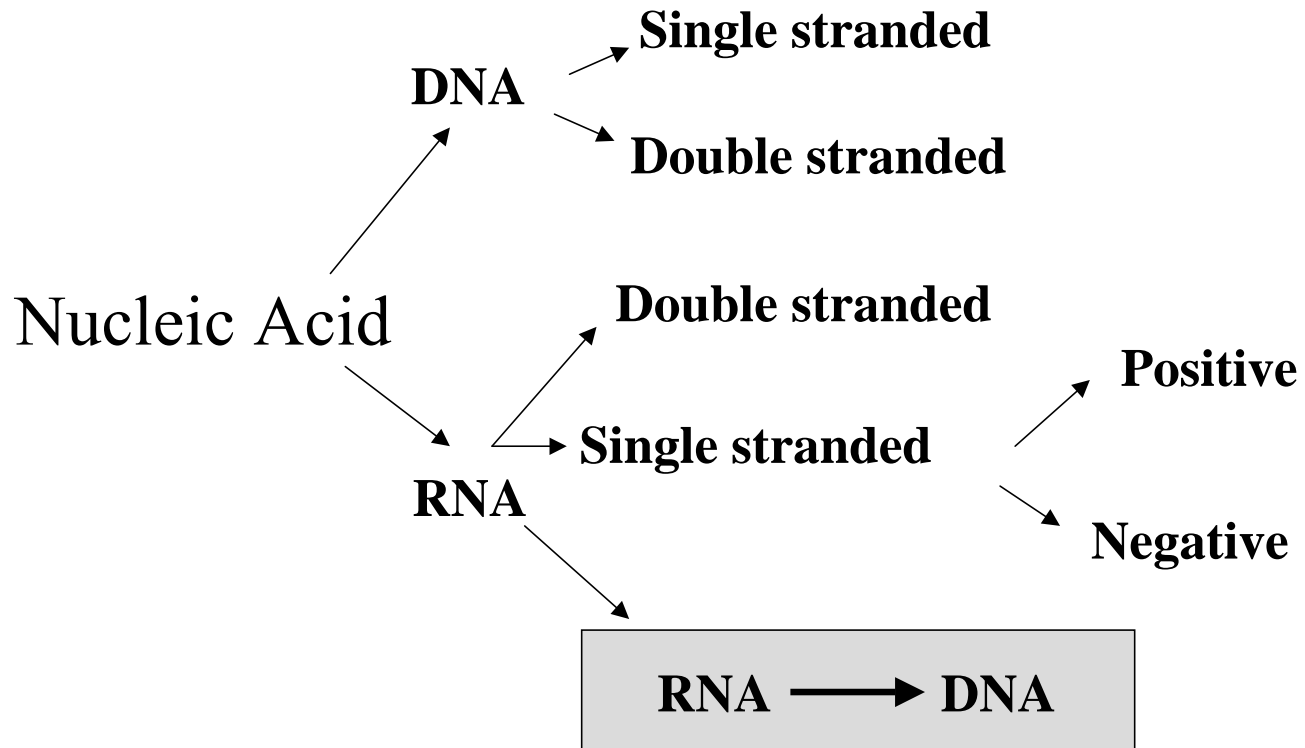


From McAllister, UTA

Early microbial genome sequencing

Organism/organelle	Genome Size	Date	Comment
Bacteriophage MS2	5.6 knt	1976	First organism (ss-RNA)
Bacteriophage ϕ X174	5.4 kbp	1977	Nature 265:687, 1977
SV40	5.8 kbp	1979	First virus
CaMV	8.0 kbp	1980	Cauliflower mosaic virus
TMV	6.3 knt	1982	Tobacco mosaic virus
Bacteriophage λ	48.5 kbp	1982	
chloroplast	121 kbp	1986	Tobacco
Vaccinia virus	192 kbp	1990	
CMV	229 kbp	1991	Cytomegalovirus
<i>Marchantia polymorpha</i>	187/121 kbp	1992	liverwort mitochondria & chloroplast
Variola	186 kbp	1993	Smallpox virus (automated sequencing)

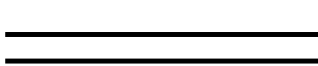
Viral genomes



Genome types

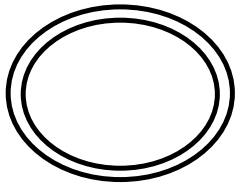
- **RNA Genome**
 - i) ssRNA
 - *e.g.*, Poliovirus, Rabies virus, HIV retrovirus
 - ii) dsRNA
 - *e.g.*, Reovirus
 - This virus' genome consists of 10-12 linear pieces of ssRNA
- **DNA Genome**
 - ssDNA
 - linear → paroviruses
 - circular → M13 phage

ii) dsDNA genome



- linear

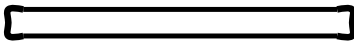
- *T4 phage*



- circular

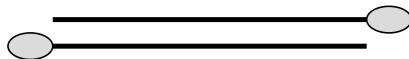
- *Herpes viruses*

- *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

NCBI

Genome

BLAST Pubmed Nucleotide Protein

Search Genes for [] [Go] [Clear]

Limits Index History

Entrez Genomes

Entrez Help | FAQ

Submitting genome sequences

All Organisms

Prominent Organisms

Microbial genomes

BLAST

List of projects

Archaea

Genome

Plasmids

unfinished

Viruses Taxonomy / List **1222**

A-2 plague virus	NC_009988	7374 bp	Jul 15 2000
African swine fever virus	NC_001499	5894 bp	Jan 21 1998
African mosaic virus	NC_001928	2629 bp	Nov 12 1990
African swine fever virus	NC_006290	5234 bp	Feb 1 2002
African swine fever virus	NC_002795	8657 bp	Oct 11 2000
Acute liver necrosis virus	NC_002548	9491 bp	Sep 14 2000
Acryphobion phage virus	NC_003780	10035 bp	Dec 9 1997
Adeno-associated virus 1	NC_002077	4718 bp	Apr 26 1999
Adeno-associated virus 2	NC_001401	4675 bp	Apr 27 1999
Adeno-associated virus 3B	NC_001863	4722 bp	Jan 12 1998
Adeno-associated virus 4	NC_001822	4767 bp	Aug 21 1997
Adeno-associated virus 6	NC_001862	4683 bp	Jan 12 1998
African filovirus demersus	NC_006285	4176 bp	Jul 7 1994
African swine fever virus	NC_001467	2779 bp	Mar 2 2001
African swine fever virus	NC_001659	170101 bp	Apr 26 1996
African swine fever virus	NC_003434	2746 bp	Mar 4 2002
African yellow fever virus	NC_004090	2768 bp	Jul 30 2002
African yellow fever virus	NC_002981	2748 bp	Jul 4 2001
African yellow fever virus-associated DNA Rep	NC_003414	1360 bp	Feb 14 2002
African yellow fever virus-associated DNA beta	NC_003403	1347 bp	Jun 10 2000
Aichi virus	NC_001918	8251 bp	Jul 9 1998
Albanian hemorrhagic fever virus	NC_002531	130008 bp	Aug 21 1997

<http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/vis.html>

<http://wit.integratedgenomics.com/GOLD/>



GOLD™ Genomes OnLine Database



Contact: <u>GOLD</u>	Last Update: January 6, 2003	Sponsored by <u>Integrated Genomics Inc.</u>
	<u>Search GOLD: 706 genome projects</u>	
118 <u>Published Complete Genomes</u> including 2 chromosomes	352 <u>Prokaryotic Ongoing Genomes</u>	236 <u>Eukaryotic Ongoing Genomes</u> including 3 chromosomes

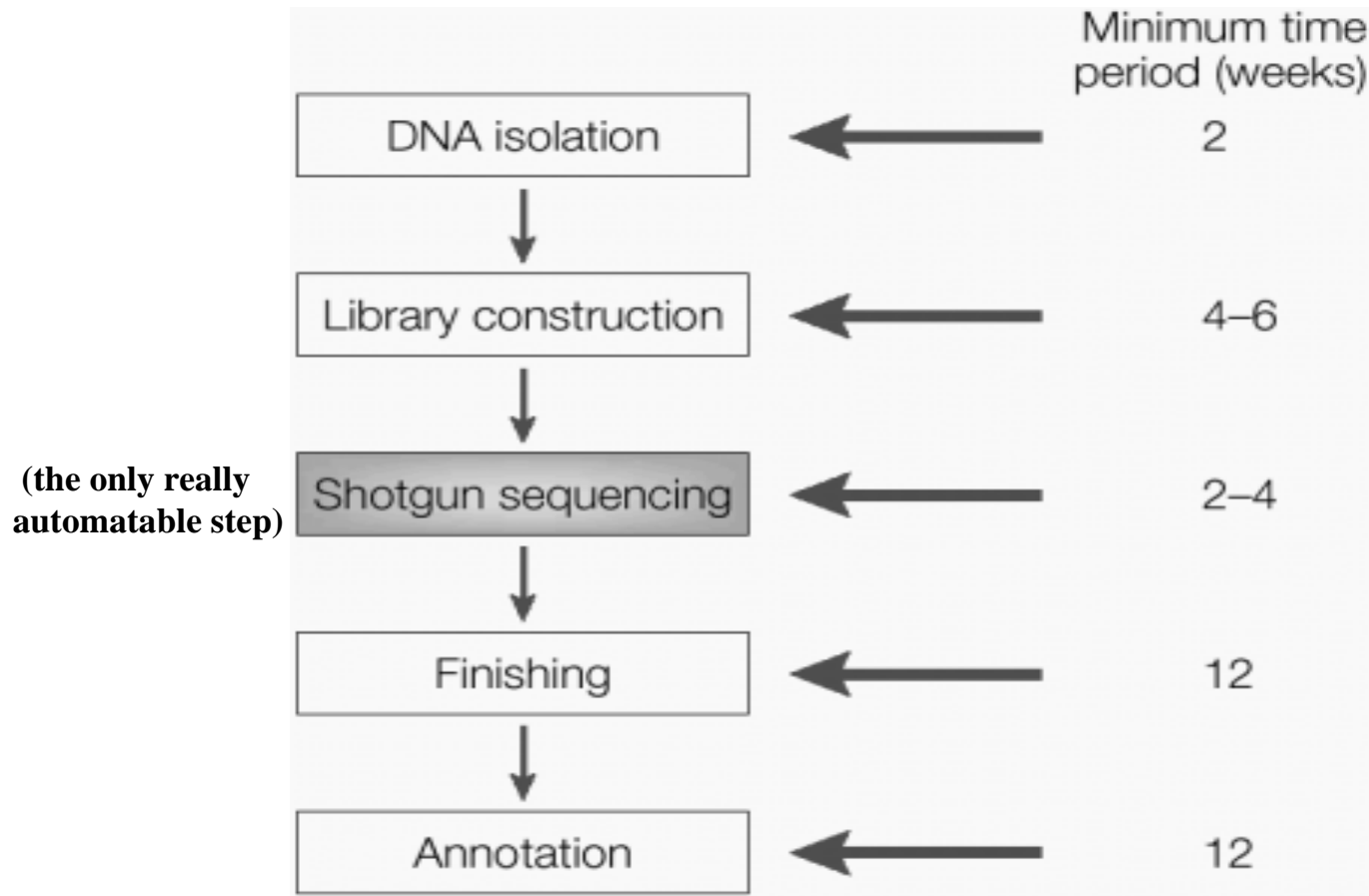
▶ 118 Published Complete Genomes:

- 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



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

..ACGATTACAATAGGTT..

Laboratory tools for studying whole genomes

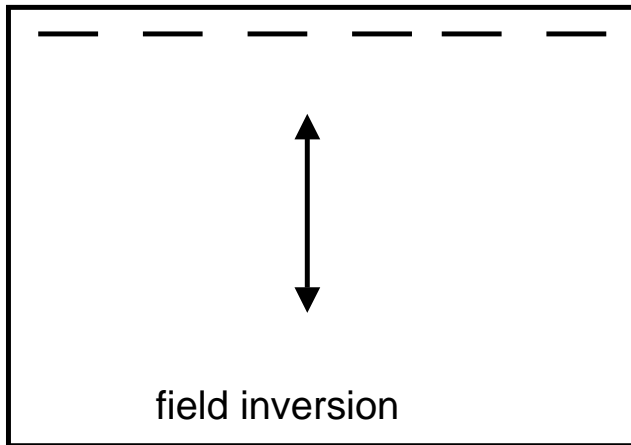
Pulsed Field Gel Electrophoresis (PFGE)

- 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)

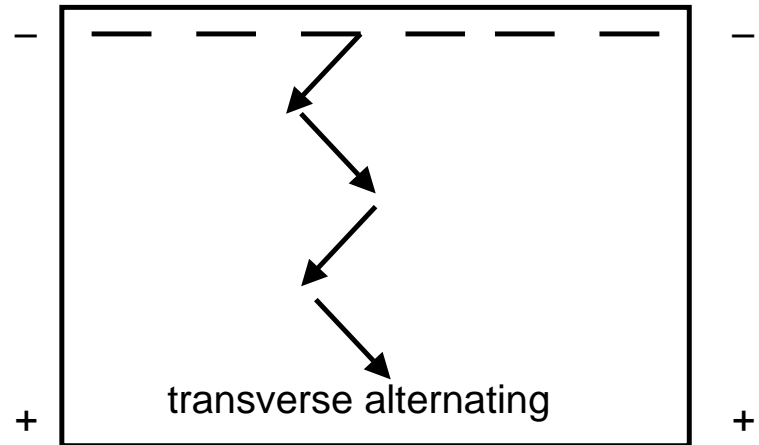
Separating large fragments

Pulsed field gel electrophoresis (PFGE)

Alternating electric fields



FIGE



CHEF

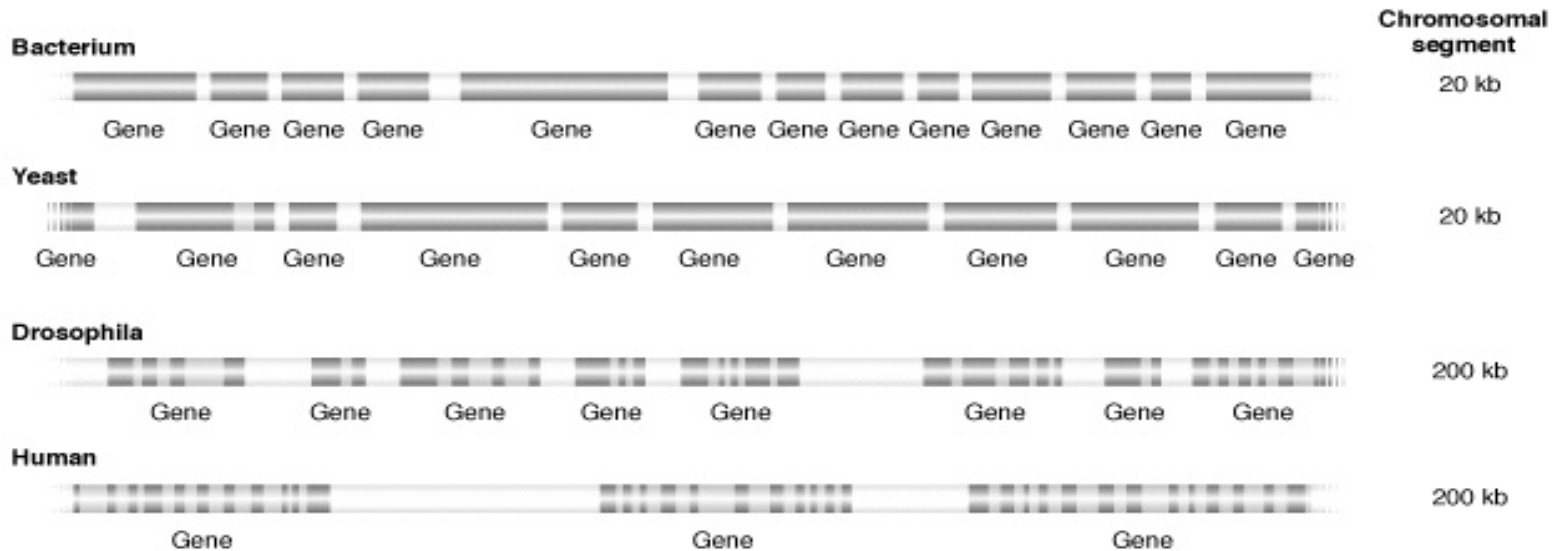
Large insert cloning vectors – BAC & PAC

- 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

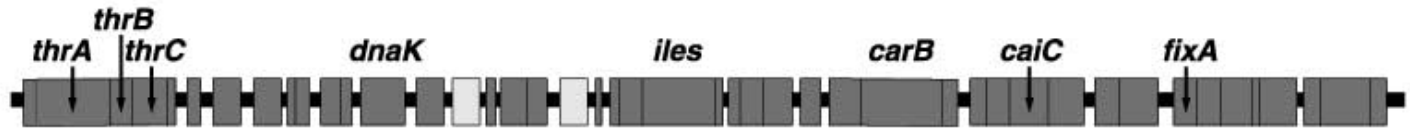
Example - *Haemophilus influenzae*

- 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)



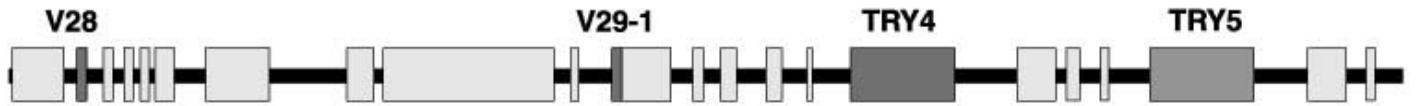
(a) *Escherichia coli*



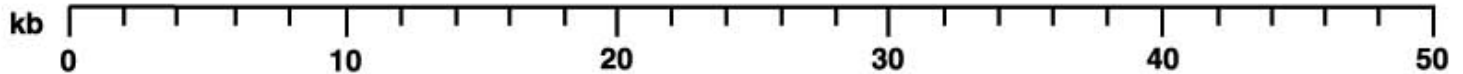
(b) *Saccharomyces cerevisiae*



(c) Human



(d) Maize



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 1Mb)

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)]

(Nature 390:580-586, 1997)

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)

o.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, \rightarrow) 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

Tn3	3.5kb	<i>bla</i>
Tn2513	7 kb	<i>mer</i>
Tn21	19 kb	<i>mer sul str</i>
Tn4	22 kb	<i>bla* sul str (*Tn3)</i>

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

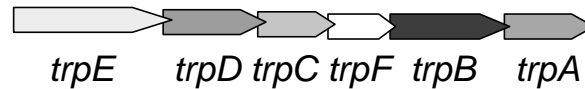
Bacterial chromosome numbers

- Most bacteria contain a single chromosome
(± extrachromosomal elements)
- Some bacteria have been found also to contain 2-3 replicons which can be considered either megaplastids 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)

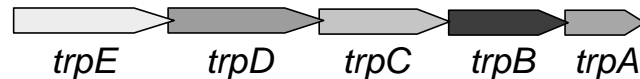
Gene order & orientation

- 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



Escherichia coli



Methanobacterium sp.



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: 6×10^5 ~ 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

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

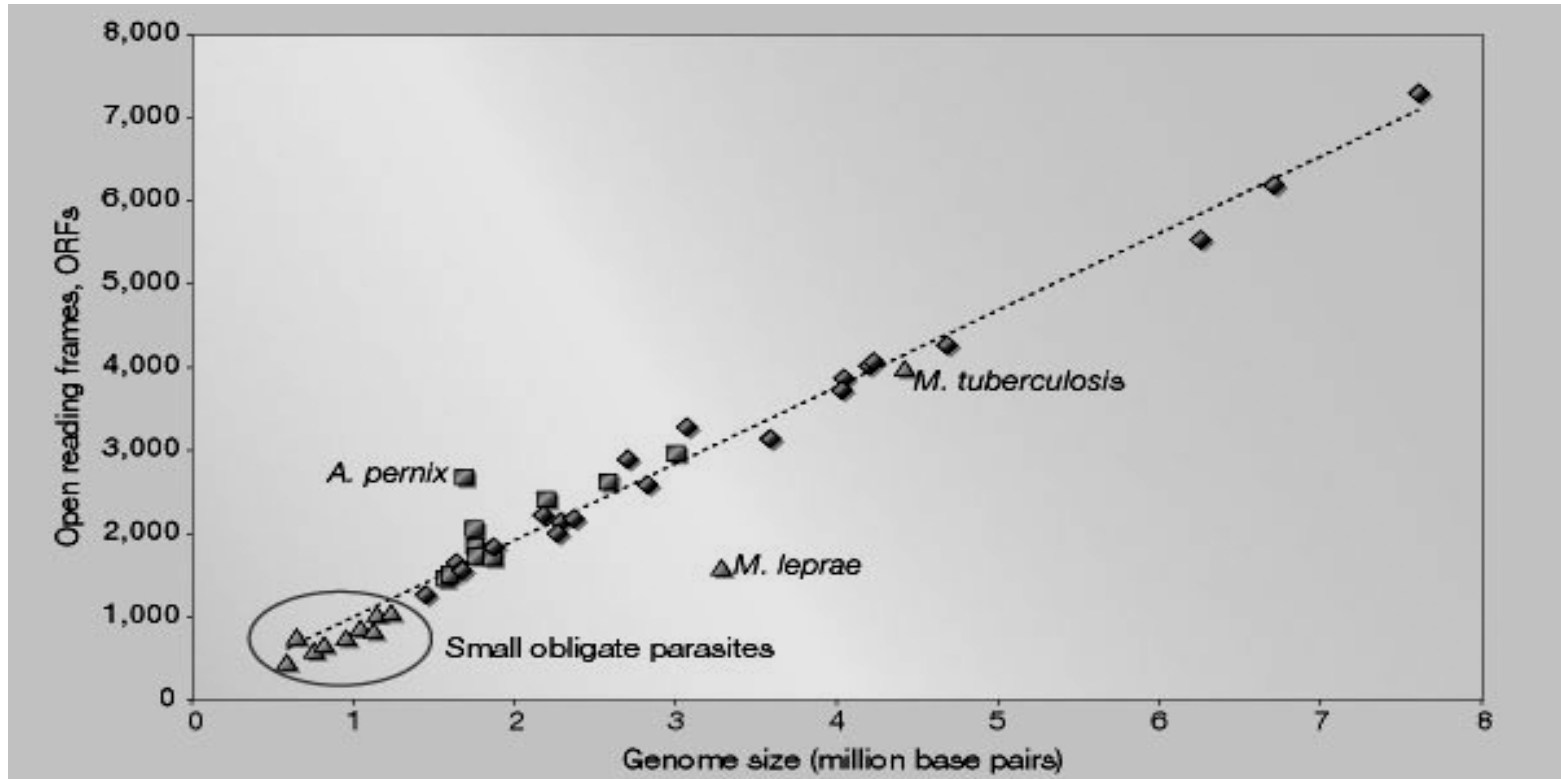
29 bacterial genome sequences finished in 2001

<i>Aerobacterium aerofaciens</i> C58-DuPont	B	ALPHA-PURPLE Taxonomy GEN_report IZME Plant Pathogen	4915 Kb 5402 acfs MAP	AED00608 AED00609 AED00607 AED00606	Univ of Washington DuPont Univ of Campinas	NSF NIH	NCBI NCBI COGs COGs GIB KEGG	Science 294,2317-2323 2001-12-14	Merker E.W
<i>Aerobacterium aerofaciens</i> C58-Cereza	B	ALPHA-PURPLE Taxonomy GEN_report IZME	4915 Kb 5299 acfs MAP	AED07869 AED07870 AED07872 AED07871	Carson Genomics Univ of Richmond Mossanto	Carson Genomics NSF NIH	NCBI NCBI COGs COGs GIB KEGG TIGR	Science 294,2323-2328 2001-12-14	Shuler S
<i>Escherichia coli</i> <i>causalis</i>	E	MICROSPORIDIA Taxonomy GEN_report	2500 Kb 1987 acfs		Genoscope Univ Blaise Pascal	CBRE	NCBI Genoscope	Nature 414, 450-451 2001-11-22	Vivares C
<i>Mycobacterium</i> <i>sp.</i> PCC 7120	B	CYANOBACTERIA Taxonomy GEN_report Cyanobak Info	6413 Kb 5366 acfs MAP	BA000019 NC_003276 NC_003240 NC_003267 NC_003273 NC_003270 NC_003241	Karson DNA Research Institute Michigan State Univ	KDR1 DOE	NCBI Karson GIB EBI-Protosome KEGG	DNA Res. 8,285-313 2001-10-31	Kaneko, I
<i>Listeria monocytogenes</i> EGD-e	B	GRAM+ Taxonomy Literature Biology Food FDA	2944 Kb 2855 acfs MAP	AL591824	EC Consortium	European Union	NCBI Fasterq COGs GIB EBI-Protosome KEGG TIGR	Science 294,849-852 2001-10-26	Conrath P
<i>Listeria monocytogenes</i> Clp11262, ribonuclease negative	B	GRAM+ Taxonomy Control Information	3011 Kb 2981 acfs MAP	AL592022	Institut Pasteur	Institut Pasteur European Union	NCBI Fasterq GIB EBI-Protosome COGs KEGG TIGR	Science 294,849-852 2001-10-26	Conrath P
<i>Salmonella typhimurium</i> , LT 208C1412	B	GAMMA-PURPLE Taxonomy General Disease-1 Microbook Salmonellon	4857 Kb 4597 acfs MAP	NC_003197	Washington Univ	NIH	NCBI Washington Univ COGs GIB EBI-Protosome KEGG TIGR	Nature 413,852-856 2001-10-25	McClelland M
<i>Salmonella typhi</i> CT18	B	GAMMA-PURPLE Taxonomy General Disease-1 Microbook Salmonellon WHO	4809 Kb 4600 acfs MAP	AL513802	Sanger Institute Imperial College	Wellcome Genomics	NCBI Sanger Institute COGs GIB EBI-Protosome KEGG	Nature 413,840-852 2001-10-25	Packhill J
<i>Streptococcus pyogenes</i> R6	B	GRAM+ Taxonomy Antibiotics Disease Microbook Diversity FDA	2030 Kb 2043 acfs MAP	AED07817	Eli Lilly	Eli Lilly	NCBI Eli Lilly GIB COGs KEGG TIGR	J Bacteriol. 183,5709-5717 2001-10-10	Olson R

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

<u>Genome</u>	<u>Size</u> <u>(MM base pairs)</u>	<u>Est. Genes*</u>	<u>Completed</u>	<u>Relevance</u>
Archaea				
Aeropyrum pernix K1	1.67	2,694	1999	Potential source of novel enzymes, etc.
Archaeoglobus fulgidus	2.18	2,407	1997	Potential source of novel enzymes, etc.
Methanobacterium thermoautotrophicum	1.75	1,869	1997	Potential source of novel enzymes, etc.
Pyrococcus abyssi	1.77	1,765	1999	Potential source of novel enzymes, etc.
Pyrococcus horikoshii	1.74	2,064	1998	Potential source of novel enzymes, etc.
Bacteria				
Aquifex aeolicus	1.55	1,522	1997	Potential source of novel enzymes, etc.
Bacillus subtilis	4.21	4,100	1997	Represents sporulating Gram-positive bacteria
Campylobacter jejuni	1.64	1,654	2000	Food-borne pathogen
Chlamydia trachomatis	1.04	894	1998	Human pathogen
Chlamydia pneumoniae	1.23	1,052	1998	Human pathogen
Escherichia coli	4.64	4,289	1998	Key model organism; human pathogen
Haemophilus influenzae	1.83	1,709	1995	Human pathogen; first free-living organism to have genome completely sequenced
Helicobacter pylori	1.67	1,553	1997	Major cause of stomach ulcers
Helicobacter pylori J99	1.64	1,491	1999	Another <i>H. pylori</i> strain
Mycobacterium tuberculosis	4.41	3,918	1998	Causes tuberculosis
Mycoplasma genitalium	0.58	480	1995	Genome is interesting because it is very small
Mycoplasma pneumoniae	0.82	677	1996	Leading cause of "walking pneumonia"
Rickettsia prowazekii	1.11	834	1998	Causes epidemic typhus
Synechocystis PCC6803	3.57	3,169	1996	Should help us understand photosynthesis
Treponema pallidum	1.14	1,031	1998	Causes venereal syphilis
Thermotoga maritima	1.86	1,846	1999	Potential source of novel enzymes, etc.
Ureaplasma urealyticum	0.75	611	2000	Sexually transmitted pathogen
Eukaryota				
Caenorhabditis elegans	~97.0	~19,000	1998	Worm – a key model organism
Saccharomyces cerevisiae	12.07	5,885	1996	Yeast – a key model organism
Human Chromosome 22**	33.46	600+	1999	First human chromosome to be fully sequenced

Source: NCBI; *excludes tRNA and rRNA genes; **euchromatic region

Box 1 | **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
(April 1999–March 2001)^{12–14} <http://www.ludwig.org.br/ORESTES/>
- Complete genome sequence of *Xanthomonas citri*
(September 1999–December 2000) <http://watson.fapesp.br/xantho/main.htm>
- Complete genome sequence of *Xanthomonas campestris*
(September 2000–September 2001)
SUCEST Sugar Cane EST Project
(April 1999–December 2000) <http://sucest.lad.ic.unicamp.br/en/>
- 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

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

FEATURES

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 1. Microbial diversity – known and estimated species³

Group	Estimated total species	Known species ^b	Proportion known of total (%)
Viruses	130 000 ^a	5000	[4] ^c
Archaea	? ^d	<500	?
Bacteria	40 000 ^a	4800	[12]
Fungi	1 500 000	69 000	5
Algae	60 000	40 000	67

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

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

^c[] Indicates that these values are probably gross underestimates.

^d16S rRNA sequence analysis of different biotopes suggests that archaeal species represent a much higher proportion of *in situ* diversity than is indicated by microbial culture studies.

Table 2. Estimates of the proportion of 'unculturable' microorganisms in various terrestrial and aquatic biotopes⁴

Biotope	Proportion of culturability (%)
Seawater	0.001–0.100
Freshwater	0.25
Mesotrophic lake	0.1–1.0
Unpolluted estuarine waters	0.1–3.0
Activated sludge	1–15
Sediments	0.25
Soil	0.3

Advantages of using microbial genomes

- Prokaryotic genomes are much smaller than eukaryotic 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

- Itaya, M., FEBS Letters 362(3):257-260, 1995:

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: $4.2 \times 10^6 \text{bp} \times 7.5\% = 3.2 \times 10^5 \text{bp}$

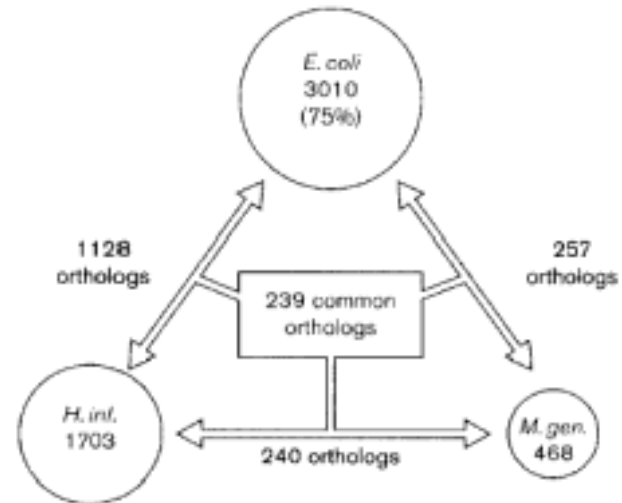
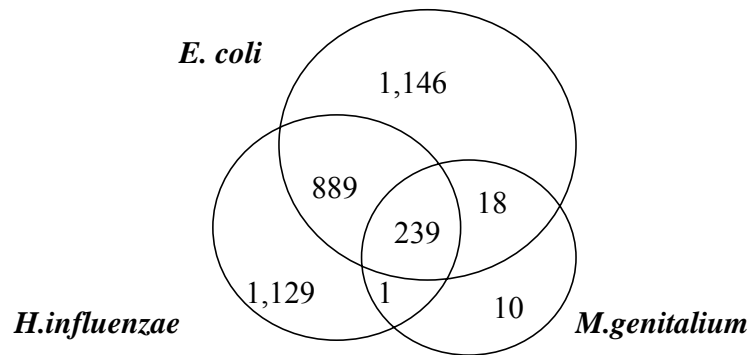
Average gene size is 1.25Kb, so the minimal genome size ≈ 254
genes.

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



[Overlapping orthologous genes (239)] + [non-orthologous gene displacement] – [genes specific to parasitic bacteria or of functional redundancy] = 256 genes



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

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 Venter^{1‡§}

Mycoplasma genitalium with 517 genes has the smallest gene complement of any independently replicating cell so far identified. Global transposon mutagenesis 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 transposon 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

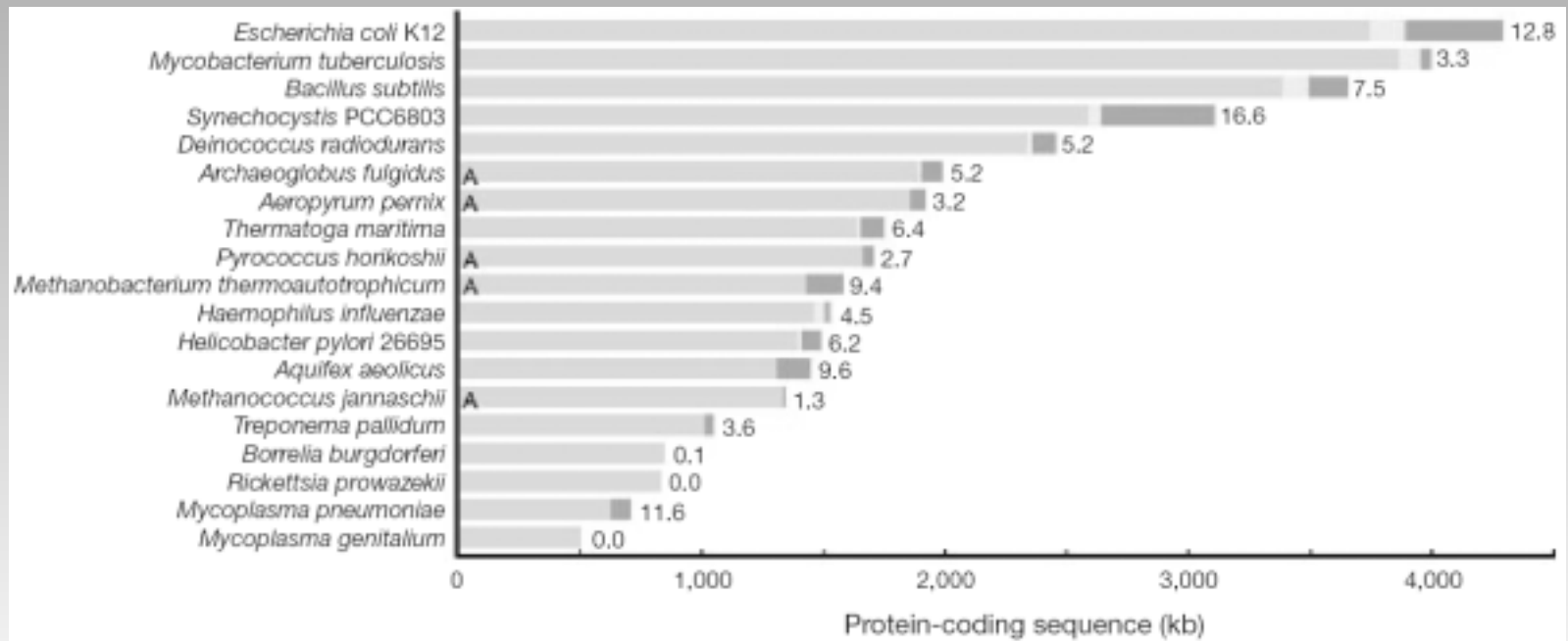
- *M. genitalium* with 480 genes has the simplest known genome
- 33% of its genome is expressed during optimal growth conditions
- Remaining proteins are likely expressed during other conditions, in undetectable amounts, or not isolated

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 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* (*pneumoniae* 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.

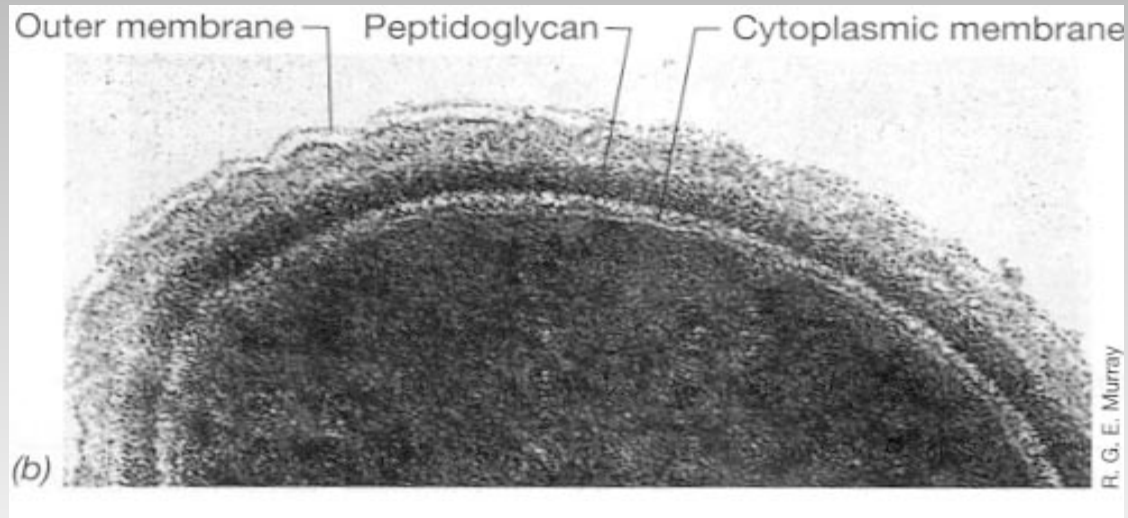
Genome Sequence of the Radioresistant Bacterium *Deinococcus radiodurans* R1

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 *Deinococcus 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.

Molecule	Length	Average ORF length (bp)	Protein coding regions	GC content	Repeat-content
Chromosome I	2,648,638	913	90.8%	67.0%	1.8%
Chromosome II	412,348	1,044	93.5%	66.7%	1.4%
Megaplasmid	177,466	1,100	90.4%	63.2%	9.2%
Plasmid	45,704	928	80.9%	56.1%	13.0%
All	3,284,156	937	90.9%	66.6%	3.8%



Unusual Characteristics of the Cell Wall

Table 6. DNA repair genes and pathways encoded by *D. radiodurans*.

Pathway Genes in <i>D. radiodurans</i>	Predicted biochemical activities and comments
Nucleotide excision repair UvrABCD	Corresponds to UV endonuclease α ; <i>uvrA</i> = <i>mtcAB</i> , <i>uvrD</i> = <i>irvB</i> (37)
Transcription repair coupling MFD	Experiments suggest that this process may not be present (34)
UV excision repair UVDE	Corresponds to UV endonuclease β (<i>uvrCDE</i>)
Base excision repair AlkA	3-methyl-guanine glycosylase
MPG-3MG	3-methyl-guanine glycosylase
Ung	Uracil DNA glycosylase
Mug	G:U mismatch glycosylase
Ung2	Uracil DNA glycosylase? (35)
MutM-Fpg	FAPY and 8-oxo-guanine DNA glycosylase
MutY-Nth-1	Likely a G:A glycosylase because most similar to MutYs
MutY-Nth-2	Thymine glycol glycosylase from (36)?
MutY-Nth-3	Second FAPY glycosylase from (36)?
MutY-Nth-4	Unknown
AP endonuclease Xth	May also be an exonuclease
Mismatch excision repair MutLS	Absence of MuthI suggests different strand recognition system than <i>E. coli</i>
Recombinational repair Initiation RecF/NRQ	Nearly complete RecF pathway (RecO missing)
RecD	Absence of RecB and RecC orthologs suggests that this gene functions differently than in <i>E. coli</i>
SbcCD	Homology to Rad50/MRE11 suggests a role in DSB repair (37)
Recombination RecA	Recombinase; may also regulate transcription of other genes
Resolution RuvABC	Likely redundant to the RecG pathway (38)
RecG	Likely redundant to RuvABC pathway (38)
DNA polymerases PolA	Repair replication polymerase (39)
PolC	Chromosomal replication polymerase
PolX	DNA polymerase of unknown function (PolX family)
Ligation Dnlj	Ligation activity is required for all excision and recombinational repair pathways
dNTP pools, cleanup MutT and Nudix family	dNTP cleanup; more copies than any other prokaryote
NrdEFI	Ribonucleotide reductase
NrdX	Ribonucleotide reductase
Induction LexA	Transcription repressor, possibly for SOS response
Other RadA/SMS	DNA damage response?
HepA	Likely role in transcription or DNA repair (or both); member of SNF2 family (24)
MutS2	Possible role in recognizing mismatches but not likely involved in mismatch repair (24)
XseA	Exonuclease VII subunit (but XseB is absent)
UvrA2	Export of damaged DNA?
Extracellular nucleases	Degradation of exported DNA?
SSB	Single-strand DNA binding protein
CinA	May recruit RecA to cell membrane
XerD	Site-specific recombinase

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

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

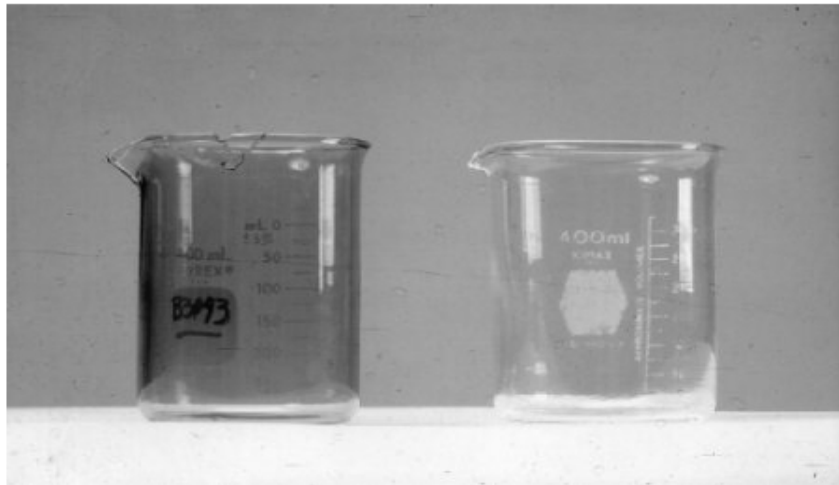
The publication of the fully assembled and annotated sequence of the *Deinococcus radiodurans* R1 genome is expected during 1999 and, if the anecdotal information released to the press¹ thus far is accurate, analysis of the sequence has not revealed much that can be used to explain this organism's extraordinary 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

When exponential-phase cultures of *Deinococcus radiodurans* are exposed to a 5000-Gray dose of γ radiation, individual cells suffer massive DNA damage. Despite this insult to their genetic integrity, these cells survive without loss of viability or evidence of mutation, repairing the damage by as-yet-poorly-understood mechanisms.

J.R. Battista*, A.M. Earl and M-J. Park are in the Dept of Biological Sciences, 508 Life Sciences Bldg, Louisiana State University and A & M College, Baton Rouge, LA 70803, USA.
*tel: +1 225 388 2810,
fax: +1 225 388 2597,
e-mail: jrbattis@unix1.sncc.lsu.edu

to document all forms of ionizing-radiation-induced DNA damage occurring in *D. radiodurans*, it is assumed that all of these types of damage occur. Certainly, measurements of single-strand breaks⁴⁻⁶ and of thymine-glycol production^{7,8} appear to be at levels commensurate with the dose of ionizing radiation administered.

D. radiodurans does not passively protect its genome from the incident radiation. Rather, all the available evidence argues that this organism rapidly and accurately repairs 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 γ -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 γ -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.

http://www.pnl.gov/er_news/08_98/beaker.htm

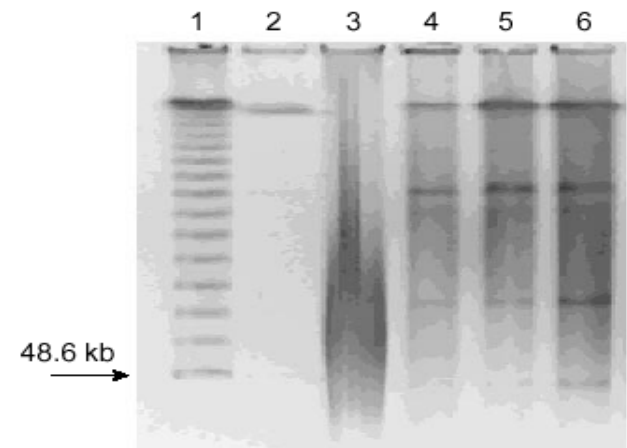


Fig. 1. The ability of *Deinococcus radiodurans* R1 to survive the accumulation of DNA double-strand breaks following exposure to a 3000-Gray (Gy) dose of γ radiation. Lane one contains a lambda size standard; lane two contains chromosomal DNA prepared from an untreated culture; lane three contains chromosomal 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 Mb²², 3000 Gy generates 120 dsbs per genome or, on average, one dsb for every 27 kb²¹.

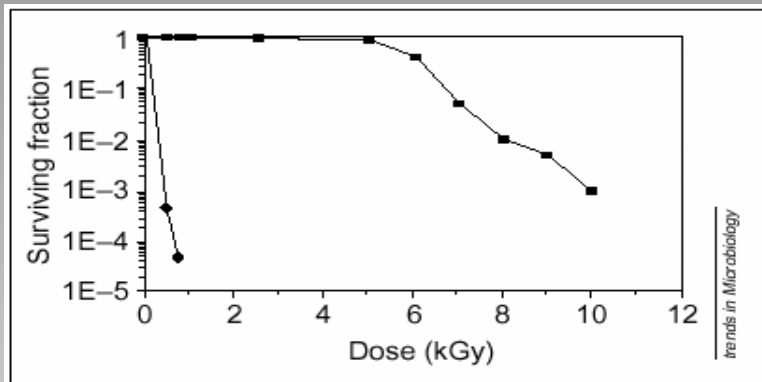


Fig. 1. Representative survival curves for *Deinococcus radiodurans* R1 (squares) and *Escherichia coli* B/r (diamonds) following exposure to γ radiation. $1E-1$ is 1×10^{-1} , or 0.1; each designation on the y -axis therefore represents a reduction in viability by a factor of 10. The D_{37} dose (i.e. the average dose of ionizing radiation that is required to inactivate a single colony-forming unit) for the *E. coli* culture is 30 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 D_{37} dose of between 6000 Gy–7000 Gy for cultures in exponential phase^{1,m}.

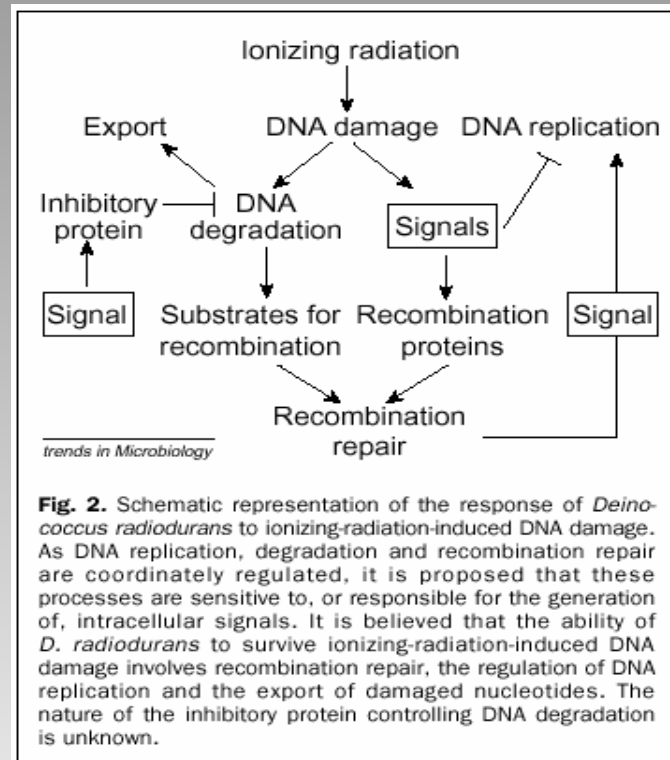


Fig. 2. Schematic representation of the response of *Deinococcus radiodurans* to ionizing-radiation-induced DNA damage. As DNA replication, degradation and recombination repair are coordinately regulated, it is proposed that these processes are sensitive to, or responsible for the generation of, intracellular 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. Wackett², and Michael J. Daly^{1*}

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, Gortner Laboratory, University of Minnesota, St. Paul, MN 55108. ³Pacific Northwest National Laboratory, Richland, WA 99352. *Corresponding author (mdaly@usuhs.mil).

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.

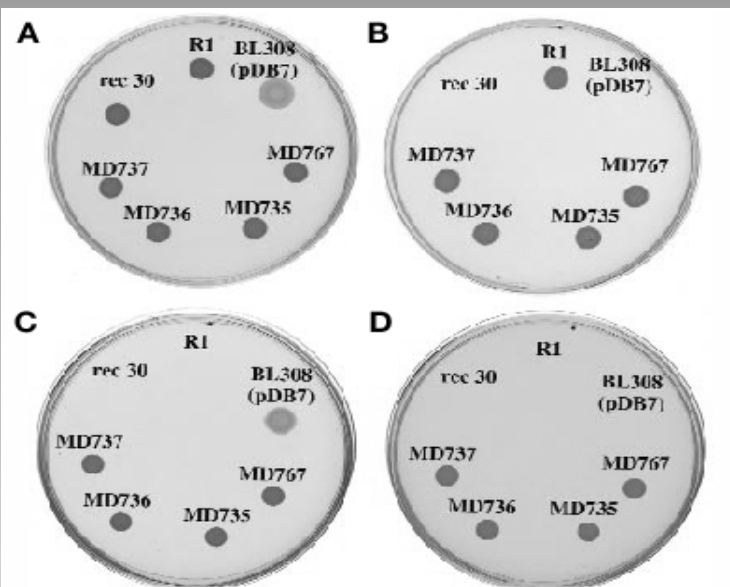


Figure 4. Effect of continuous exposure to γ -radiation and Hg(II) on the growth of strains, containing different copy numbers of the *mer* operon. Two TGY agar plates (A and B), and two TGY agar plates containing 30 μ M Merbromin (C and D) were spotted with 1×10^5 cells of each of the indicated strains. Following plate inoculation, plates B and D were placed into the ^{137}Cs 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.

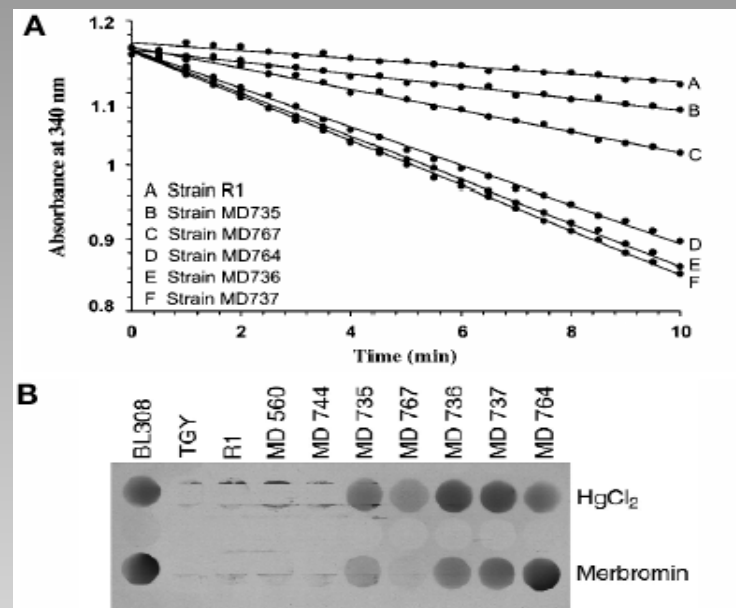


Figure 6. (A) Mercuric reductase assay. Hg(II)-dependent NADPH oxidation catalyzed by cell extracts prepared from the strains R1 (*mer*⁺, *tod*⁺; wild type), MD735 (*mer*⁻), MD767 (*mer*⁻), MD764 (*mer*⁻, *tod*⁻), MD736 (*mer*⁻), and MD737 (*mer*⁻) were monitored spectrophotometrically³⁰. Decreasing absorbance at 340 nm corresponds to a decreasing NADPH concentration. (B) Mercury volatilization by engineered *D. radiodurans*.

Case study - *Neisseria meningitidis*

(Nature 404:502-506, 2000)

- *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*

(Nature 390:580-586, 1997)

- *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.

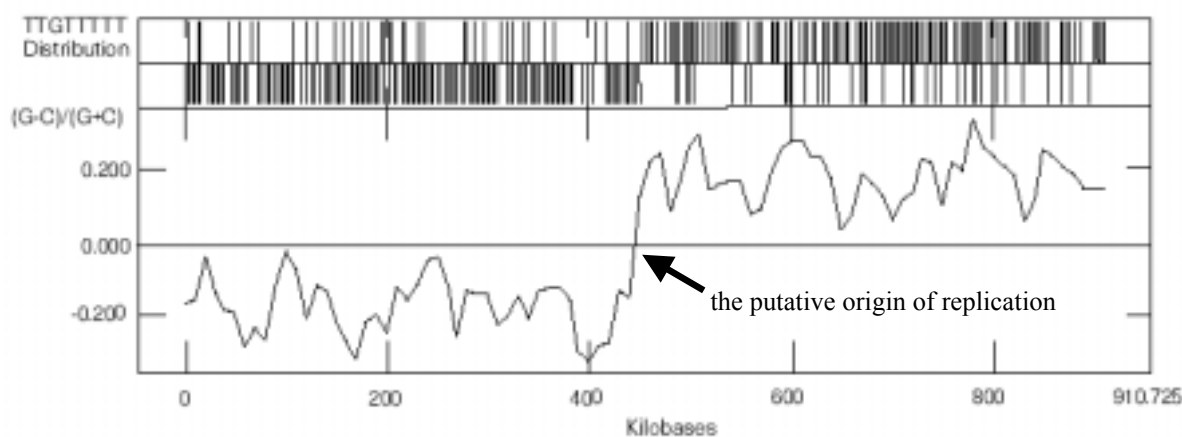


Figure 2 Distribution of TTGTTTTT and GC skew in the *B. burgdorferi* chromosome. Top, distribution of the octamer TTGTTTTT. The lines in the top panel represent the location of this octamer in the plus strand of the sequence, and those in the second panel represent the location of this oligomer in the minus strand of the sequence. Bottom, GC skew.

(Nature 390:580-586, 1997)

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 a 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 chromosomal 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 *Schizopyllum commune* (a basidiomycete fungus), *Bremia lactucae* (oomycete)

Compared to:

- 1% transcribed in humans

The genome sequence of *Schizosaccharomyces pombe*

V. Wood¹, R. Gwilliam¹, M.-A. Rajandream¹, M. Lyne¹, R. Lyne¹, A. Stewart², J. Sgouros³, N. Peat³, J. Hayles³, S. Baker⁴, D. Basham¹, S. Bowman¹, K. Brooks¹, D. Brown¹, S. Brown¹, T. Chillingworth¹, C. Churcher¹, M. Collins¹, R. Connor¹, A. Cronin¹, P. Davis¹, T. Fellwell¹, A. Fraser¹, S. Gentles¹, A. Goble¹, N. Hamlin¹, D. Harris¹, J. Hidalgo¹, G. Hodgson¹, S. Holroyd¹, T. Hornsby¹, S. Howarth¹, E. J. Huckle¹, S. Hunt¹, K. Jagels¹, K. James¹, L. Jones¹, M. Jones¹, S. Leather¹, S. McDonald¹, J. McLean¹, P. Mooney¹, S. Moule¹, K. Mungall¹, L. Murphy¹, D. Niblett¹, C. Odell¹, K. Oliver¹, S. O'Neill¹, D. Pearson¹, M. A. Quail¹, E. Rabbinowitsch¹, K. Rutherford¹, S. Rutter¹, D. Saunders¹, K. Seeger¹, S. Sharp¹, J. Skellon¹, M. Simmonds¹, R. Squares¹, S. Squares¹, K. Stevens¹, K. Taylor¹, R. G. Taylor¹, A. Tivey¹, S. Walsh¹, T. Warren¹, S. Whitehead¹, J. Woodward¹, G. Volckaert⁴, R. Aert⁴, J. Robben⁴, B. Grymonprez⁴, I. Waijers⁴, E. Vanstreels⁴, M. Rieger⁵, M. Schäfer⁵, S. Müller-Auer⁵, C. Gabel⁵, M. Fuchs⁵, C. Fritz⁶, E. Holzer⁶, D. Moestl⁶, H. Hilbert⁶, K. Borzym⁷, I. Langer⁷, A. Beck⁷, H. Lehrach⁷, R. Reinhardt⁷, T. M. Pohl⁸, P. Eger⁸, W. Zimmermann⁸, H. Wedler⁸, R. Wambutt⁸, B. Purnelle⁹, A. Goffeau¹⁰, E. Cadieu¹¹, S. Dréano¹¹, S. Gloux¹¹, V. Lelaure¹¹, S. Mottier¹¹, F. Gallibert¹¹, S. J. Aves¹², Z. Xiang¹², C. Hunt¹², K. Moore¹², S. M. Hurst¹², M. Lucas¹³, M. Rochel¹³, C. Gaillardin¹³, V. A. Tallada^{14,15}, A. Garzon^{14,15}, G. Thode¹⁴, R. R. Daga^{14,15}, L. Cruzado¹⁴, J. Jimenez^{14,15}, M. Sánchez¹⁶, F. del Rey¹⁶, J. Benito¹⁶, A. Dominguez¹⁶, J. L. Revuelta¹⁶, S. Moreno¹⁶, J. Armstrong¹⁷, S. L. Forsburg¹⁸, L. Cerrutti¹, T. Lowe¹⁹, W. R. McCombie²⁰, I. Paulsen²¹, J. Potashkin²², G. V. Shpakovski²³, D. Ussery²⁴, B. G. Barrell¹ & P. Nurse²

We have sequenced and annotated the genome of fission yeast (*Schizosaccharomyces pombe*), which contains the smallest number of protein-coding genes yet recorded for a eukaryote: 4,824. The centromeres are between 35 and 110 kilobases (kb) and contain related repeats including a highly conserved 1.8-kb element. 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. Few similarly conserved genes that are important for multicellular organization were identified, suggesting that the transition from prokaryotes to eukaryotes required more new genes than did the transition from unicellular to multicellular organization.

Table 1 Genome content for the three chromosomes

	Length (bp)	No. of genes	No. of T2s	No. of pseudo T2s	No. of wtls	No. of lone LTRs	No. of pseudogenes	Mean gene length (bp)*	Gene density†	Coding (%)
Chromosome 1	5,598,923	2,255	8	0	1	77	17	1,446	2,483	58.6
Chromosome 2	4,397,795	1,790	2	1	1	53	9	1,411	2,457	57.5
Chromosome 3	2,465,919	884	1	2	23	50	7	1,407	2,790	54.5
Whole genome	12,462,637	4,929	11	3	25	180	33	1,426	2,528	57.5

* Mean gene length excluding introns.

† Gene density, given as average bp per gene.

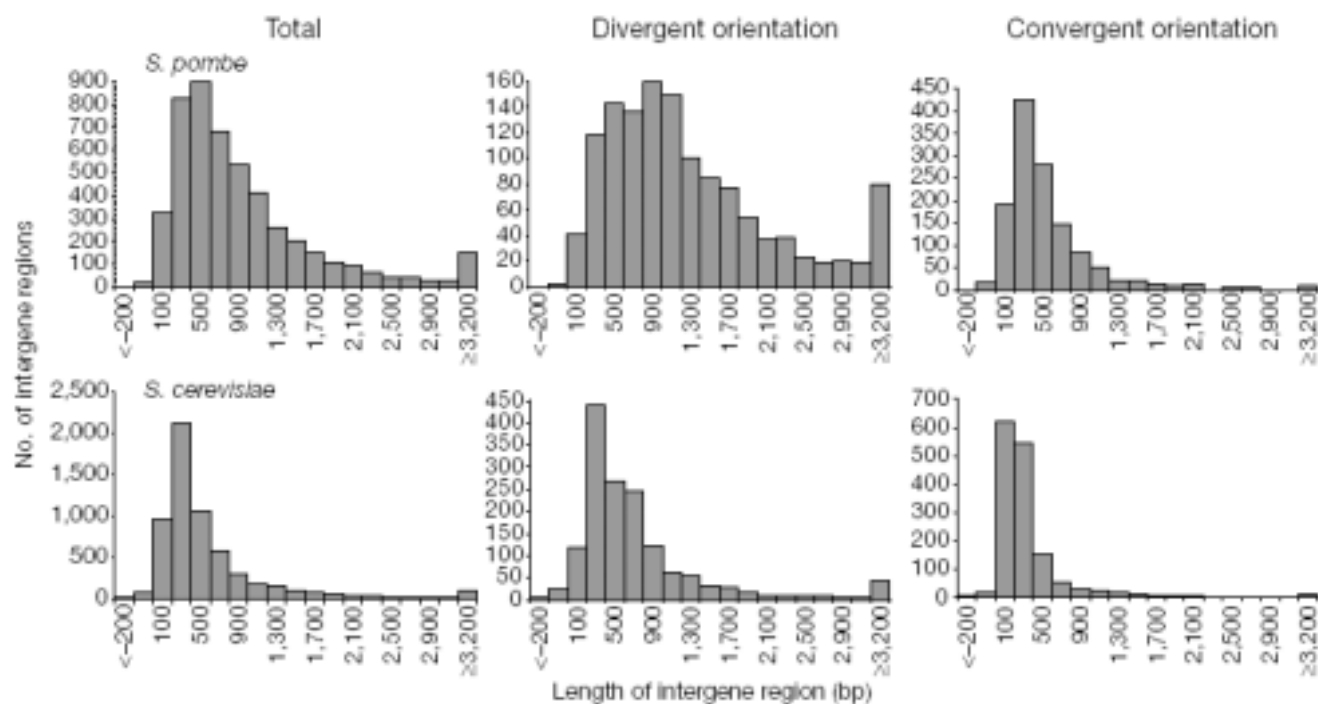


Figure 2 Intergene regions. Distribution of intergene regions given for all genes and for divergent and convergent pairs of genes, for both *S. pombe* and *S. cerevisiae*. A total of 4,890 intergene regions from *S. pombe* were analysed from a database prepared just

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.

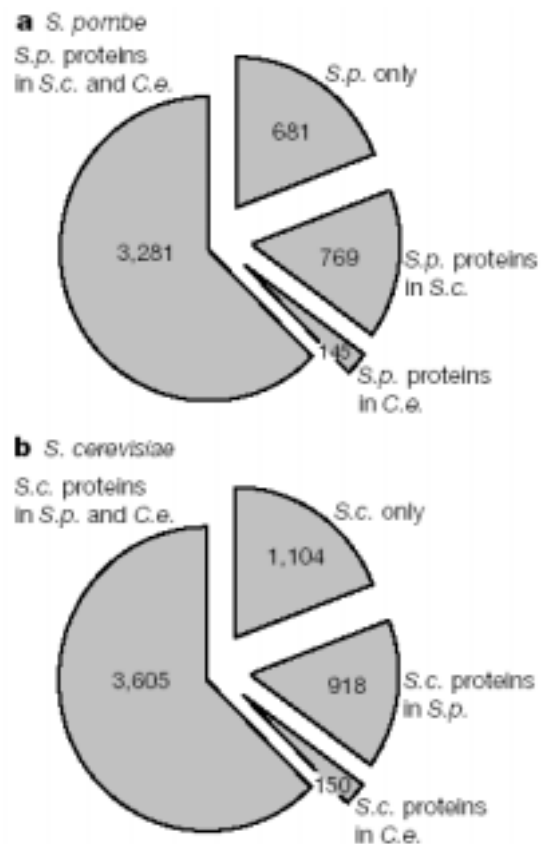


Figure 3 Comparison of proteins in *S. pombe* (*S.p.*), *S. cerevisiae* (*S.c.*) and *C. elegans* (*C.e.*). **a**, Pie chart comparing the homology of proteins of *S. pombe* with those of *S. cerevisiae* and *C. elegans*. **b**, Pie chart comparing the homology of proteins of *S. cerevisiae* with those of *S. pombe* and *C. elegans*. For example, *S.p.* proteins in *S.c.* 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 6 Protein domain analysis and comparison with other eukaryotes

Interpro accession no.	<i>S. pombe</i>		<i>S. cerevisiae</i>		<i>H. sapiens</i>		<i>D. melanogaster</i>		<i>C. elegans</i>		<i>A. thaliana</i>		Interpro name	
	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank		
IPR001687	213	1	267	1	436	5	231	4	191	7	331	5	ATP/GTP-binding site motif A (Ploop)	1
IPR001680	114	2	97	3	277	8	183	5	102	19	210	10	G protein β WD40 repeats	1
IPR000719	111	3	119	2	579	3	377	2	450	2	1,049	1	Eukaryotic protein kinase	1
IPR000604	80	4	61	5	307	7	182	6	97	21	255	8	RNA binding region RNP1	1
IPR001650	67	5	63	4	155	20	101	17	80	27	148	13	Helicase C-terminal domain	2
IPR001841	44	6	33	12	215	15	120	11	126	12	379	4	RING finger	-
IPR001440	38	7	33	12	150	21	92	18	46	43	125	17	TPR repeat	-
IPR001066	36	8	46	8	44	64	45	34	55	37	98	26	Sugar transporter	-
IPR001617	33	9	42	9	75	40	67	28	61	36	103	25	ABC transporter family	-
IPR000822	32	10	51	7	712	2	403	1	154	10	115	20	Zinc finger, C2H2 type	1
IPR001357	14	23	10	30	24	82	17	61	25	60	17	83	BRCT domain	2
IPR000882	8	29	9	31	8	99	9	68	6	79	13	87	Replication factor C conserved domain	2
IPR002064	5	32	5	35	4	102	6	70	3	82	5	95	DNA directed DNA polymerase family β	2
IPR001208	6	31	6	34	12	94	13	64	5	80	8	92	MCM family	2
IPR000002	5	32	3	37	3	103	4	72	2	83	6	94	FIZZY/CDC20 domain	2
IPR001452	21	16	23	18	220	14	82	23	62	35	3	97	Src homology 3 (SH3) domain	3
IPR001849	21	16	26	16	253	11	89	22	75	31	27	73	PH domain	3
IPR000387	9	28	11	29	112	29	47	40	110	16	21	79	Tyrosine-specific protein phosphatase and dual-specificity protein phosphatase family	3
IPR001138	27	13	52	6	0	NA	0	NA	0	NA	0	NA	Fungal transcriptional regulatory protein	-
IPR002293	21	16	32	13	43	65	36	45	32	54	65	42	Permease for amino acids and related compounds	-
IPR000953	7	30	2	38	26	80	20	58	15	70	24	76	Chromodomain	-

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)
Schulte, U., et al. *J. Biotechnology* 94:3-13 (2002)

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)
Schulte, U., et al. *J. Biotechnology* 94:3-13 (2002)