Bacterial Genomics

The Bacterial Kingdom

First organisms on earth

- 3 billion years ago
- 60% of earth's biomass
- Very diverse metabolic processes

Important to the biological world

- Donated mitochondria to eukaryotes
    - Alpha proteobacteria
- Created the atmosphere we live in today
    - Photosynthetic bacteria
        - Cyanobacteria
    - Evolved O_2 during the light reaction of photosynthesis
- Important ancestors of plants
    - Donated the photosynthetic system to eukaryotes

Found in all environments

- Soil
    - Interact with plants
        - Nitrogen fixation
            - *Rhizobium* species
- High-salt conditions
    - Halophiles
        - Great Salt Lake
- High temperatures
    - Thermal vents
        - *Themus aquaticus*
            - DNA polymerase used in PCR reactions

Mostly known as a pathogen

- Human diseases
    - Pneumonia
    - Blindness
    - Tuberulosis
    - Cholera
    - Plague (Black Death)
        - 25% of European population killed

Food industry

- Dairy products
    - Milk, cheese yogurt

Size varies

- Small
    - Mycoplasmas
        - 5X size of ribosome
- Large
    - *Thiomarginata namibiensis*
        - Size of a fruit-fly eye

Bacterial Clades (Based on Molecular Systematics)

Protobacteria

- Gram negative
- Photoautotrophs
- Chemoautotrophs
- Chemoheterotrophs

Alpha proteobacteria

- Associate with eukaryotic hosts
    - Rhizobium
- Rickettsias (Rocky Mountain Spotted Fever)
    - Endosymbionts

Beta Proteobacteria

- Some soil bacteria
    - Nitrogen recycling
        - Ammonium (NH_4^+) to nitrate (NO_2^-)

Gamma Proteobacteria

- Photoautotrophs and chemoheterotrophs
    - Non-oxygen evolving photobacteria
        - Sulfur bacteria
- Diseases
    - *Legionella* (Legionnaires' disease)
    - Intestinal (enteric) bacteria
        - *E. coli*
    - Cholera (*Vibrio cholerae*)
    - Food poisoning (*Salmonella*)

Delta proteobacteria

- Some are colony forming
- Develop fruiting bodies
    - *Myxobacteria*
    - *Chondromyces*
- Bacterial predators
    - *Bdellovibrio bacteriophorus*

Epsilon proteobacteria

- Close to the delta proteobacteria
- Diseases
    - Stomach ulcers
        - *Helicobacter pylori*

Chlamydias

- Gram negative
- Obligate animals heterotrophs
- Use host as ATP source
- Diseases
    - *Chlamydia trachomatis*
        - Most common sexual transmitted disease in US
        - Blindness

Spirochetes

- Can be:
    - Helical heterotrophs
    - Free-living
- Diseases
    - Syphillis (*Treponema pallidum*)
    - Lyme Disease (*Borrelia burgadorferi*)

Gram-positive Bacteria

- All gram-positive bacteria
  - Plus some gram-negative
- Nearly as diverse as Proteobacteria
- Most are free-living
- Diseases
  - Anthrax *(Bacillus anthracis*)
  - Botulism (*Clostribium botulinum*)
- Decompose organic matter in soil
- Antibiotic source
  - *Streptomyces*
- Some are colony formers
- Diseases
  - Tuberculosis
    - *Mycobacterium tuberculosis*
  - Leprosy
    - *Mycobacterium leprae*
- Mycoplasmas
  - Smallest bacteria
  - 5X size of ribosomes

Cyanobacteria

- Photoautotrophs
- Solitary and colonial
- Abundant in water

What genetic material do you find inside a bacterial cell?

Chromosome

- One chromosome (normally)
    - Multiple chromosome genomes discovered
        - 1989
            - *Rhodobacter sphaeroides*
                - Two chromosomes
- Circular (normally)
    - Linear genomes discovered
        - 1990s
            - *Borelias* and *Streptomyces*

Agrobacterium has a circular and linear chromosome

Chromosome exists as a bacterial nucleoid

- Core
  - Protein
    - HU, IHT, H1
    - Not absolutely required
    - Mutants of these genes viable
  - RNA
    - Function not known

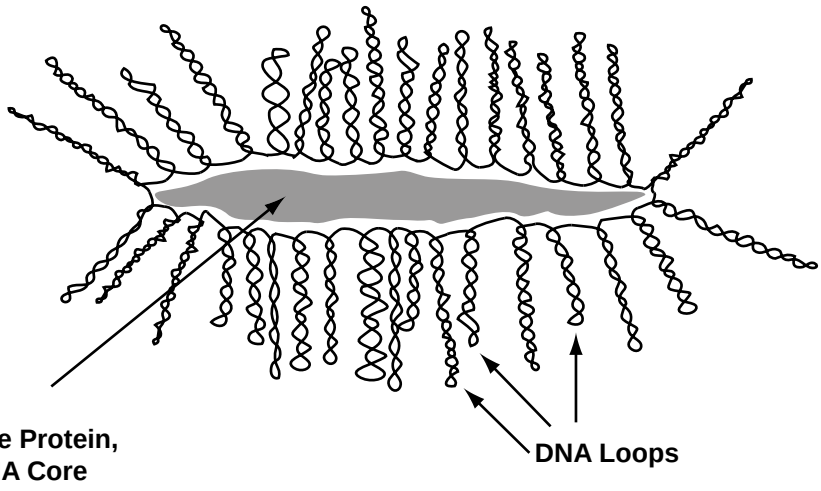
DNA loops

- *E. coli* example
  - Contour length larger than size of cell
    - DNA must be condensed
  - DNA loops major form of condensation
    - 40 kb in length
    - Negative supercoiling
    - Each loop independent

Bacterial Nucleoid

A. The Concept

The bacterial chromosome is generally a single, ***circular molecule***. Unlike the eukaryotic chromosome, it does not have a higher order. Thin-section electron micrographs show a structure in which the DNA is formed into loops. The actual physical features are not known. The DNA appears to be associated with a dense core of unknown nature. Collectively, the structure is called the ***bacterial nucleoid***.



The dense, internal region of the nucleoid consists of both protein and RNA. A number of proteins have been isolated (HU, IHF, H1). Each of these have a role in some DNA functions, but no mutant of any of these proteins is fatal. Therefore, it is unclear what are the critical protein components of the nucleoid. The function of the RNA component of the dense core is unknown

The contour length of a bacterial chromosome, for an organism such as *E. coli*, is about 1100 μm. This is much larger (1-2x) than the length for the entire cell. Therefore, the DNA must be condensed in some manner. ***DNA loops*** are the major condensation feature. The entire chromosome consists of 50-100 loops. The loops are generally 40 kb in length. Each loop cosists of negatively supercoiled DNA, and each loop is structurally independent of the other loops. This means if the supercoiled structure of one loop is released using enzymatic treatment, the neighboring loops will still be supercoiled.

An Old Adage Dispelled. Historically, it was generally reported that bacteria have a single, circular chromosome. This is not a universal truth. Although this is generally true, exceptions do exist. The first exceptions to the "single chromosome" rule was reported in 1989; *Rhodobacter sphaeroides* was discovered to have two large circular chromosomes. The "linear chromosome" rule was conclusively disproven in 1990 with the discovery that the genomes of genera *Borelias* and *Streptomyces* were linear. A dramatic exception is *Agrobacterium tumefaciens*: it contains two circular and two linear chromosomes.

Bacterial Plasmids

- Autonomous molecules in bacterial genomes
- Structure
    - Circular
    - Linear
- Size variation exists
    - Megaplasmid
        - Hundreds of kilobases in size
        - Contain hundreds of genes
        - Example
            - *Agrobacterium* (plant pathogen/plant transformation vector)
                - pAT
                    - 543 kb, 547 genes
                - pTI
                    - 214 kb, 198 genes
    - "Miniplasmid"
        - A few to tens of kb
        - A few to tens of genes
- Function of plasmids
    - Virulence
        - Attack other organisms
        - *Agrobacterium*
    - Drug resistance
    - Toxin production
    - Conjugation (transfer of DNA) with other bacteria
    - Metabolic degradation of molecules
        - Environmenal importance

Bacterial Genome Examples

Escherichia coli

- Circular chromosome
    - strain specific sizes
        - 4.5-5.5 megabases
        - 4300-5300 genes

Agrobacterium tumefaciens

- Chromosomes
    - Circular chromosome
        - 2.8 megabases
        - 2721 genes
    - Linear chromosome
        - 2.0 megabases
        - 1833 genes
- Megaplasmids
    - plasmid pAT
        - 543 kilobases
        - 547 genes
    - plasmid pTi
        - 214 kilobases
        - 198 genes

Borrelia burgdorferi

- Linear chromosome
  - 911 kilobases
  - 850 genes
- 21 plasmids
  - 12 circular plasmids
    - 9-31 kilobases
    - 11-45genes
  - 9 linear mini plasmids
    - 5-54 kilobases
    - 6-76 genes

Microbial Genome Sequencing

The Beginnings

- *Hemophilius influenzae*
  - 1995
  - First microbe genome sequenced
  - Shotgun sequence approach used
    - Showed utility of shotgun sequencing

Genomes sequenced

- September 2003
  - 136 species publicly released
    - 120 bacteria
    - 16 archaea
  - TIGR (http://www.tigr.org/tdb/mdb/mdbcomplete.html)
    - The Institute for Genome Research
    - 32 genomes sequenced
    - Focuses on microbes
    - Promoted importance of microbe sequence information
      - Useful in bioterrorism defenses
  - Private industry
    - Don’t know how many sequenced
    - Industrial uses are being investigated

Table 1. Features of selected bacterial genomes. Data compiled from http://www.cbs.dtu.dk/services/GenomeAtlas/Bacteria/index.html

Species	Interest	Size (nt)	G+C (%)	Coding density	bp/gene	# Genes (% unique)*
Mycoplasma genitalium		580,074	32	90	1208	480 (20)
Bruchnera aphidicola		609,132	26	81	1222	504 (1)
Rickettsia conorii	Spotted fever	1,268,755	33	80	923	1374 (36)
Haemophilus influenzae	Flu	1,830,138	39	85	1070	1709 (14)
Lactococcus lactis	Cheese starter	2,365,589	36	84	1043	2266 (25)
Clostridium tetani	Tetnus	3,799,251	29	85	1179	2373
Clostridium perfringens	Gangrene	3,031,430	29	83	1139	2660
Mycobacterium leprae	Leprosy	3,268,203	58	76	1201	2720 (27)
Clostridium acetobutylicum	Solvent producer	3,940,880	31	85	1073	3672 (26)
Vibrio cholerae (Total)	Cholera	4,033,464	48	86	1053	3828 (20)
Chromosome 1		2,961,149	48	87	1082	2736
Chromosome 2		1.072,315	47	84	982	1092
Mycobacterium tuberculosis	Tuberculosis	4,411,529	66	90	1126	3918 (20)
Escherichia coli	Model					
K-12 MG1655	organism	4,639,221	51	87	1081	4289
K-12 W3110		4,641,433	51	87	1057	4390
CFT073		5,231,428	51	87	972	5379
O157 RIMD0509952		5,498,450	51	87	1025	5361
O157 EDL93		5,528,970	51	87	1033	5349 (25)
Yersinia pestis	Plague	4.653,728	48	83	1161	4008 (17)
Xanthomonas campestris	Citrus canker	5,076,188	66	84	na	4181
Vibrio parahaemolyticus	Gastroenteritis	5,165,770	46	86		4832
Chromosome 1		3,288,558	46	86	1067	3080
Chromosome 2		1,877,212	46	86	1071	1752
Bacillus anthracis	Anthrax	5,227,293	36	84	800	5,508
Agobacterium tumefaciens	Transformation	5,673,563	60	89	1070	5299
Chromosome 1 (circular)	vector	2.841,581	60	88	1040	2721 (16)
Chromosome 2 (linear)		2,074,782	60	90	1131	1833
plasmid pAT		542,869	58	85	992	547
plasmid pTi		214,331	57	86	1082	198
Pseudomonas syringae	Plant pathogen	6,397,126	59	85	1169	5471
Anabaena nostic	Photosynthesis	7,211,893	42	82		6129
Streptomyces avermitilis	Antibiotics	9,025,608	71	86	1191	7575
Bradyrhizobium japonicum	N2 fixation	9,105,828	65	86	1094	8317

*Unique sequences are those not listed as part of a COG (Cluster of Orthologous Genes)

General Features of Bacterial Genomes

- 16X differences in genome sizes
  - *Mycoplasma genitalium* vs. *Bradyrhizobium japonicum*
- Wide range of G+C content
- Low degree of intergenic DNA (10-24%)
- Gene sizes are similar
- Increased genome size means more genes
  - For genomes in Table 1
    - r = 0.98 between genome size and number of genes

Unique Genes

COG

- Cluster of Orthologus Genes
  - If gene similar to two other genes, it is considered part of a cluster
  - Determines if a gene is unique
- Most genes part of a known class
  - Unique genes may define unknown functions specific to a species
- Novel ORFs appear in each newly sequenced genome
  - Source???
    - Rapidly evolving genes
    - Genes of lineage specific function

Relationship Among Genes and Proteins

Essential terms

- Homolog
    - two genes that are related by descent
    - Important to note
        - genes are homologous or they are not homologus
        - there is not percentage homology
        - Proper way of expressing the relationship
            - "Genes A and B are homologous and share X% amino acid (or nucleotide) identity."
        - For amino acids
            - Proteins can be identical or similar
                - Identity
                    - % identical amino acids
                - Similar
                    - % similar amino acids
                    - similar amino acids share similar biochemical properties

Ortholog

- ***Two genes from different species*** that are identical by descent

Paralog

- ***Two genes from the same species*** that have arisen by gene duplication

Table 2. A classificatioin scheme for functional genomics developed by TIGR (http://www.biochem.ucl.ac.uk/~rison/FuncSchemes/Tables/tigr.html)

1 AMINO ACID BIOSYNTHESIS
- 1.1 Other
- 1.2 Serine family
- 1.3 Pyruvate family
- 1.4 Histidine family
- 1.5 Glutamate family
- 1.6 Aspartate family
- 1.7 Aromatic amino acid family

2 AUTOTROPHIC METABOLISM
- 2,1 Chemoautotrophy
- 2.2 Photoautotrophy

3 BIOSYNTHESIS OF COFACTORS, PROSTHETIC GROUPS, AND CARRIERS
- 3.1 Pantothenate
- 3.2 Pyridine nucleotides
- 3.3 Pyridoxine
- 3.4 Riboflavin
- 3.5 Thiamine
- 3.6 Other
- 3.7 Molybdopterin
- 3.8 Biotin
- 3.9 Folic acid
- 3.10 Glutathione
- 3.11 Heme and porphyrin
- 3.12 Lipoate
- 3.13 Menaquinone and ubiquinone

4 CELL ENVELOPE
- 4.1 Other
- 4.2 Surface structures
- 4.3 Lipoproteins
- 4.4 Degradation of polysaccharides
- 4.5 Biosynthesis of surface polysaccharides and lipopolysaccharides
- 4.6 Biosynthesis of murein sacculus and peptidoglycan

5 CELLULAR PROCESSES
- 5.1 Other
- 5.2 Transformation
- 5.3 Toxin production and resistance
- 5.4 Protein and peptide secretion
- 5.5 Detoxification
- 5.6 Chaperones
- 5.7 Cell division

6 CENTRAL INTERMEDIARY METABOLISM
- 6.1 Other
- 6.2 Sulfur metabolism
- 6.3 Polyamine biosynthesis
- 6.4 Phosphorus compounds
- 6.5 Nitrogen metabolism
- 6.6 Nitrogen fixation
- 6.7 Amino sugars

7 DNA METABOLISM

- 7.1 Restriction/modification
- 7.2 Degradation of DNA
- 7.3 DNA replication, recombination, and repair
- 7.4 Chromosome-associated proteins

8 ENERGY METABOLISM

- 8.1 Methanogenesis
- 8.2 Pentose phosphate pathway
- 8.3 Polysaccharides
- 8.4 Pyruvate dehydrogenase
- 8.5 Sugars
- 8.6 TCA cycle
- 8.7 Other
- 8.8 Glycolysis/gluconeogenesis
- 8.9 ATP-proton motive force interconversion
- 8.10 Aerobic
- 8.11 Amino acids and amines
- 8.12 Anaerobic
- 8.13 Electron transport
- 8.14 Entner-Doudoroff
- 8.15 Fermentation

9 FATTY ACID AND PHOSPHOLIPID METABOLISM

- 9.1 Biosynthesis
- 9.2 Degradation
- 9.3 Other

10 HYPOTHETICAL

- 10.1 General

11 PURINES, PYRIMIDINES, NUCLEOSIDES, AND NUCLEOTIDES

- 11.1 Other
- 11.2 Sugar-nucleotide biosynthesis and conversions
- 11.3 Salvage of nucleosides and nucleotides
- 11.4 Pyrimidine ribonucleotide biosynthesis
- 11.5 Purine ribonucleotide biosynthesis
- 11.6 Nucleotide and nucleoside interconversions 11.7 2'-Deoxyribonucleotide metabolism

12 REGULATORY FUNCTIONS

- 12.1 General

13 TRANSCRIPTION

- 13.1 Other
- 13.2 Transcription factors
- 13.3 RNA processing
- 13.4 Degradation of RNA
- 13.5 DNA-dependent RNA polymerase

14 TRANSLATION

- 14.1 Other
- 14.2 tRNA modification
- 14.3 Translation factors
- 14.4 Ribosomal proteins: synthesis and modification
- 14.5 Amino acyl tRNA synthetases
- 14.6 Degradation of proteins, peptides, and glycopeptides
- 14.7 Nucleoproteins
- 14.8 Protein modification

15 TRANSPORT AND BINDING PROTEINS

- 15.1 Other
- 15.2 Unknown substrate
- 15.3 Porins
- 15.4 Nucleosides, purines and pyrimidines
- 15.5 Amino acids, peptides and amines
- 15.6 Anions
- 15.7 Carbohydrates, organic alcohols, and acids
- 15.8 Cations

16 OTHER CATEGORIES

- 16.1 Other
- 16.2 Transposon-related functions
- 16.3 Phage-related functions and prophages
- 16.4 Adaptations and atypical conditions

Table 3. Cluster of Orthologous Genes (COG) functional groups.

Code	Function	# Pathways/ functional systems
Information storage and processing		
J	Translation, ribosomal structure and biogenesis	4
K	Transcription	3
L	DNA replication, recombination and repair	2
Cellular processes		
D	Cell division and chromosome partitioning	-
O	Posttranslational modification, protein turnover, chaperones	-
M	Cell envelope biogenesis, outer membrane	1
N	Cell motility and secretion	2
P	Inorganic ion transport and metabolism	1
T	Signal transduction mechanisms	-
Metabolism		
C	Energy production and conversion	7
G	Carbohydrate transport and metabolism	4
E	Amino acid transport and metabolism	10
F	Nucleotide transport and metabolism	5
H	Coenzyme metabolism	11
I	Lipid metabolism	2
Q	Secondary metabolites biosynthesis, transport and catabolism	-
Poorly characterized		
R	General function prediction only	-
S	Secondary metabolites biosynthesis, transport and catabolism	-

Table 4. Translation factors and enzymes involved in translation

Gene	Category	COG	Function
TIF6	[J]	COG1976	Eukaryotic translation initiation factor 6 (EIF6)
TufB	[JE]	COG0050	GTPases - translation elongation factors
eRF1	[J]	COG1503	Peptide chain release factor eRF1
Def	[J]	COG0242	N-formylmethionyl-tRNA deformylase
Map	[J]	COG0024	Methionine aminopeptidase
Fmt	[J]	COG0223	Methionyl-tRNA formyltransferase
Pth	[J]	COG0193	Peptidyl-tRNA hydrolase
PrfA	[J]	COG0216	Protein chain release factor A
PrfB	[J]	COG1186	Protein chain release factor B
Frr	[J]	COG0233	Ribosome recycling factor
FusA	[J]	COG0480	Translation elongation and release factors (GTPases)
EFB1	[J]	COG2092	Translation elongation factor EF-1beta
Efp	[J]	COG0231	Translation elongation factor P/translation initiation factor eIF-5A
Tsf	[J]	COG0264	Translation elongation factor Ts
SUI1	[J]	COG0023	Translation initiation factor (SUI1)
nfB	[J]	COG0532	Translation initiation factor 2 (GTPase)
InfA	[J]	COG0361	Translation initiation factor IF-1
InfC	[J]	COG0290	Translation initiation factor IF3
GCD7	[J]	COG1601	Translation initiation factor eIF-2, beta subunit/eIF-5 N-terminal domain
GCN3	[J]	COG0182	Translation initiation factor eIF-2B alpha subunit
GCD2	[J]	COG1184	Translation initiation factor eIF-2B delta subunit
SUI2	[J]	COG1093	Translation initiation factor eIF2alpha

COGs

- Based on bacterial and yeast genomes
- September 10, 2003 information
- Total number
  - 3307 genes
    - Two species COGs
      - 115 genes
    - Three species COGs
      - 493 genes
    - 26 species
      - 84 genes
  - A **few genes** are widely common
  - **Most genes** are shared with only a few other species

see: http://www.ncbi.nlm.nih.gov/COG/
http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?num=all

Examples of COGs

	# prot	Species	Symbol	Categ		Gene
-	22	--m-k---vd-lb-efgh---j----	RsmC	[J]	COG2813	16S RNA G1207 methylase RsmC
-	64	-------qvdrlbcefghsnujxit-	RsuA	[J]	COG1187	16S rRNA uridine-516 pseudouridylate synthase and related pseudouridylate synthases
-	21	a-mpkz-qvd--b-ef-----j----	LigT	[J]	COG1514	2'-5' RNA ligase
3	59	aompkz-qvdr-bcefghsnujxit-	MiaB	[J]	COG0621	2-methylthioadenine synthetase

Table 5. Distribution of genes by functional classes for selected bacterial genomes.

	Mycoplasma genitalium	*Rickettsia conorii*	*Haemophilus influenzae*	*Escherichia coli K-12*	*Bradyrhizobium japonicum*
Total Proteins	484	1374	1709	4279	8317
Proteins in COG	385	876	1591	3587	6197
Translation	101	126	149	171	205
RNA procesisng and modification	0	0	1	2	0
Transcription	40	25	73	280	499
Replication, recombination, repair	14	71	111	220	369
Chromatin structure and dynamics	0	0	0	0	2
Cell cycle control, mitosis, meiosis	5	18	24	34	41
Nuclear structure	0	0	0	0	0
Defense mechanisms	8	23	19	48	101
Signal transduction mechanisms	3	22	37	134	328
Cell wall/membrane biogenesis	12	82	122	235	314
Cell motility	0	3	6	107	138
Cytoskeleton	0	0	0	0	0
Extracellular structures	0	0	0	0	0
Intracellular trafficking and secretion	6	32	27	37	74
Posttranslational modification, protein turnover, chaperones	20	56	87	128	233
Energy production and conversion	20	77	95	275	445
Carbohydrate transport and metabolism	26	35	104	368	440
Amino acid transport and metabolism	15	33	154	350	723
Nucleotide transport and metabolism	21	21	57	87	95
Coenzyme transport and metabolism	14	31	72	123	188
Lipid transport and metabolism	9	36	40	83	402
Inorganic ion transport and metabolism	17	22	91	191	298
Secondary metabolites biosynthesis, transport and catabolism	0	13	18	68	179
General function prediction only	40	91	146	338	627
Function unknown	14	59	158	308	496
not in COGs	99	498	118	692	2120

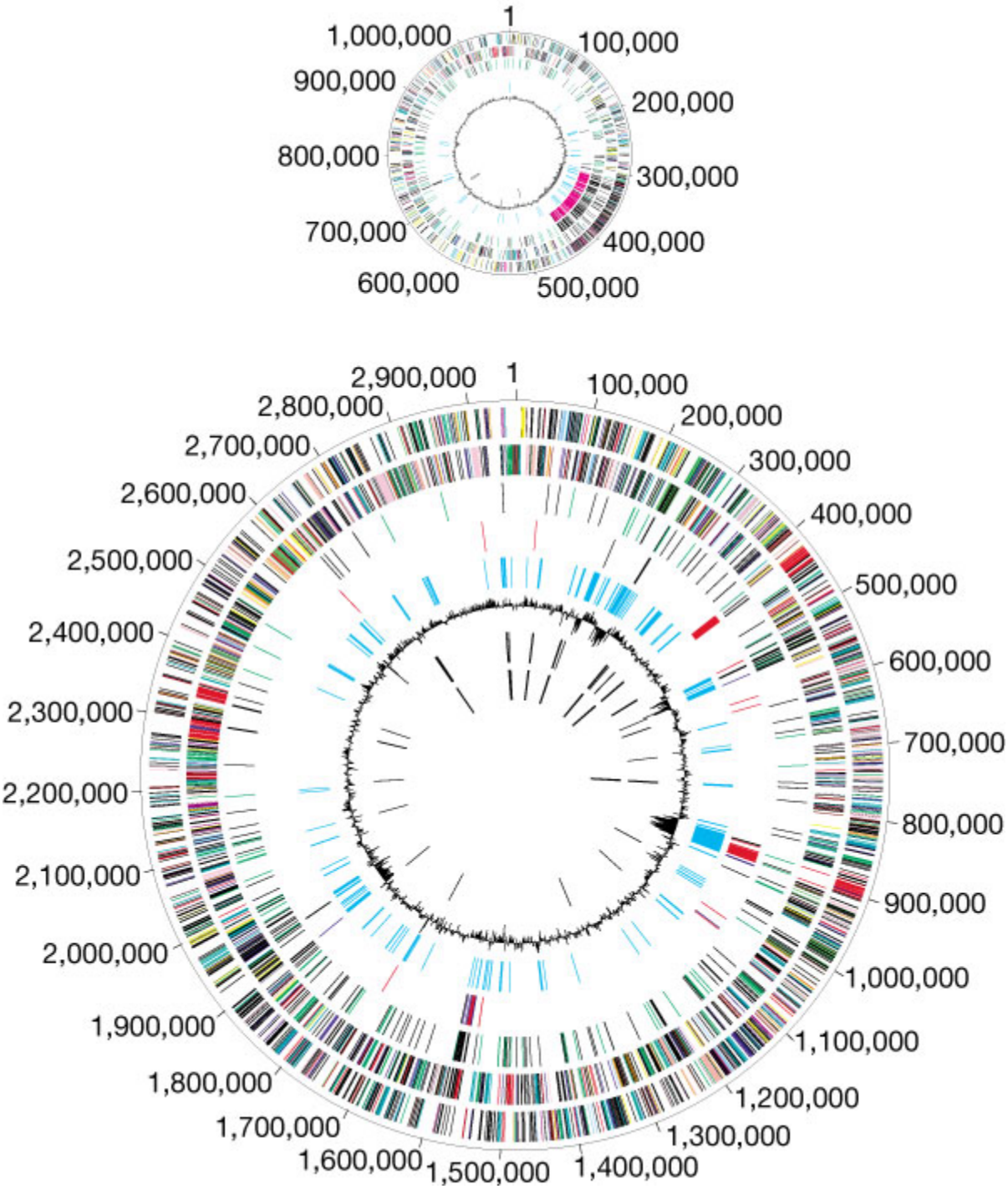


Figure 2 Circular representation of the *V. cholerae* genome. The two chromosomes, large and small, are depicted. From the outside inward: the first and second circles show predicted protein-coding regions on the plus and minus strand, by role, according to the colour code in Fig. 1 (unknown and hypothetical proteins are in black). The third circle shows recently duplicated genes on the same chromosome (black) and on different chromosomes (green). The fourth circle shows transposon-related (black), phage-related (blue), VCRs (pink) and pathogenesis genes (red). The fifth circle shows regions with significant χ^2 values for trinucleotide composition in a 2,000-bp window. The sixth circle shows percentage G+C in relation to mean G+C for the chromosome.The seventh and eighth circles are tRNAs and rRNAs, respectively.

Graphical Representation of Genomes

A. The Concept

Genomic sequencing generates a tremendous amount of sequence data. It is always a challenge to represent that in a manner that is digestable to the scientific public. Chromosome information is typically represented in a linear form. This also is true for genomes, such as for many bacterial speices that have circular chromosomes.

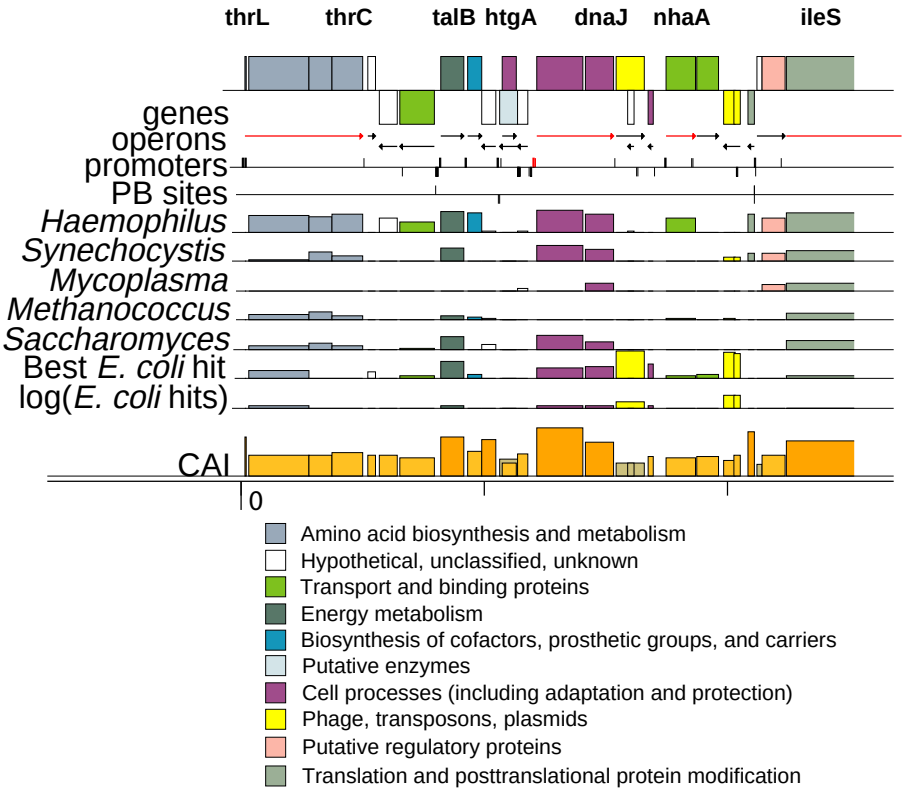


Fig. 3. (PDF). Map of the complete E. coli sequence, its features and similarities to proteins from five other complete genome sequences, proceeding from left to right in 42 tiers. The top line shows each gene or hypothetical gene, color-coded to represent its known or predicted function as assigned on the basis of biochemical and genetic data. Genes are vertically offset to indicate their direction of transcription. Space permitting, names of previously described E. coli genes are indicated above the line. The second line contains arrows indicating documented (red) and predicted (black) operons. Documented operons encoding stable RNAs are blue. Line 3, below the operons, contains tick marks showing the position of documented (red), predicted (black), and stable RNA (blue) promoter sequences. Line 4 consists of tick marks showing the position of documented (red) and predicted (black) protein binding sites. Lines 5 to 9 are histograms showing the results of alignments between E. coli proteins and the products encoded by five other complete genomes. The height of each bar is a simple index of similarity: the product of the percent of each protein in the pairwise alignment and the percent amino acid identity across the aligned region. Line 10 indicates similarity among proteins in E. coli in the same fashion. Line 11 histograms show the logarithm of the number of proteins in the E. coli genome that match a particular protein. Line 12 in each tier is a histogram that indicates the CAI of each ORF. Genes with intermediate CAI values are shown in orange, genes with high CAI values (>90th percentile) are a darker shade of orange, genes with low CAI values (<10th percentile) are light brown, and clusters of four or more genes with low CAI values (<0.25) are yellow. The final line in each tier is a scale showing position (in base pairs).

Figure 2. A graphical representation of the first seven genes of the *Eschericia coli* genome. (The legend is from the original manuscript: Science (1997) 277:1453.

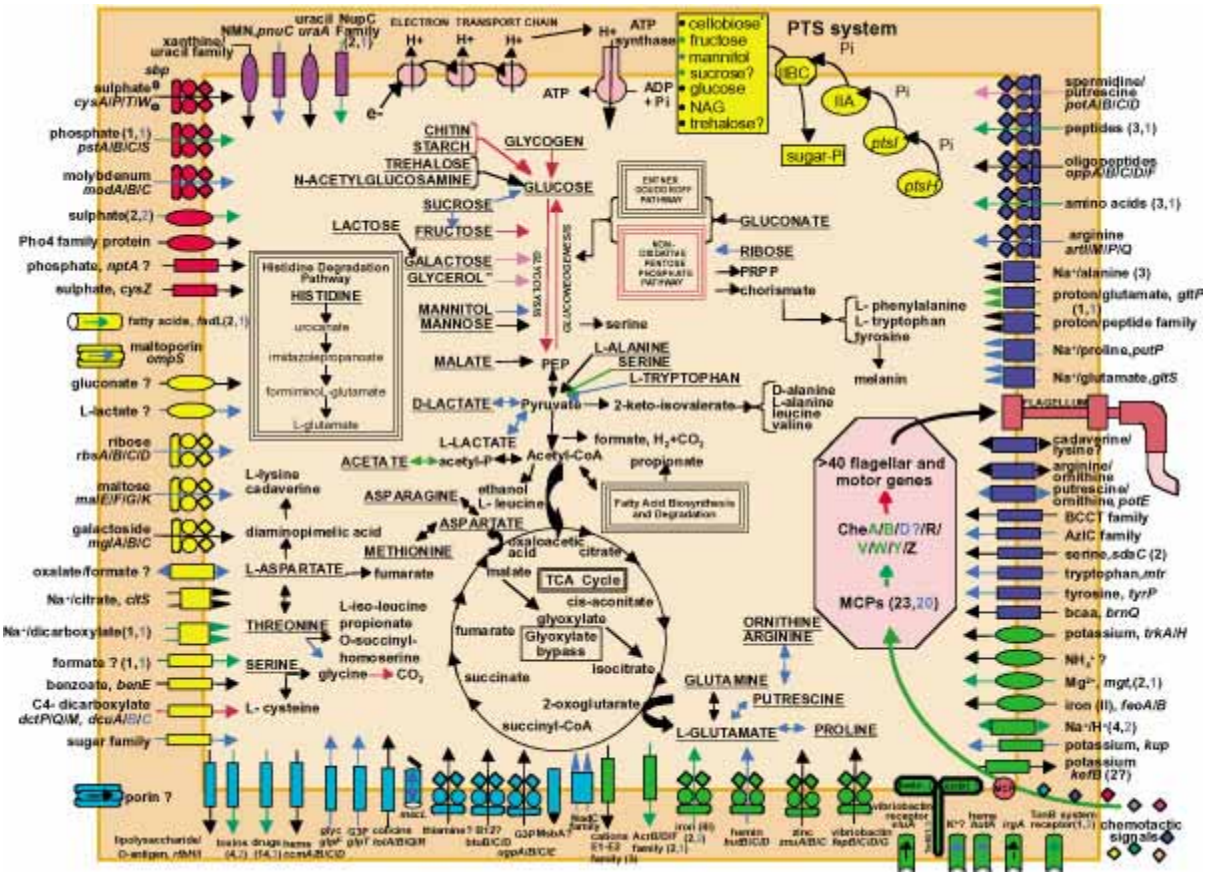


Figure 3 Overview of metabolism and transport in *V. cholerae*. Pathways for energy production and the metabolism of organic compounds, acids and aldehydes are shown. Transporters are grouped by substrate specificity: cations (green), anions (red), carbohydrates (yellow), nucleosides, purines and pyrimidines (purple), amino acids/peptides/amines (dark blue) and other (light blue). Question marks associated with transporters indicate a putative gene, uncertainty in substrate specificity, or direction of transport. Permeases are represented as ovals; ABC transporters are shown as composite figures of ovals, diamonds and circles; porins are represented as three ovals; the large-conductance mechanosensitive channel is shown as a gated cylinder; other cylinders represent outer membrane transporters or receptors; and all other transporters are drawn as rectangles. Export or import of solutes is designated by the direction of the arrow through the transporter. If a precise substrate could not be determined for a transporter, no gene name was assigned and a more general common name reflecting the type of substrate being transported was used. Gene location on the two chromosomes, for both transporters and metabolic steps, is indicated by arrow colour: all genes located on the large chromosome (black); all genes located on the small chromosome (blue); all genes needed for the complete pathway on one chromosome, but a duplicate copy of one or more genes on the other chromosome (purple); required genes on both chromosomes (red); complete pathway on both chromosomes (green). (Complete pathways, except for glycerol, are found on the large chromosome.) Gene numbers on the two chromosomes are in parentheses and follow the colour scheme for gene location. Substrates underlined and capitalized can be used as energy sources. PRPP, phosphoribosyl-pyrophosphate; PEP, phosphoenolpyruvate; PTS, phosphoenolpyruvate-dependant phosphotransferase system; ATP, adenosine triphosphate; ADP, adenosine diphosphate; MCP, methyl-accepting chaemotaxis protein; NAG, *N*-acetylglucosamine; G3P, glycerol-3-phosphate; glyc, glycerol; NMN, nicotinamide mononucleotide. Asterisk, because *V. cholerae* does not use cellobiose, we expect this PTS system to be involved in chitobiose transport.

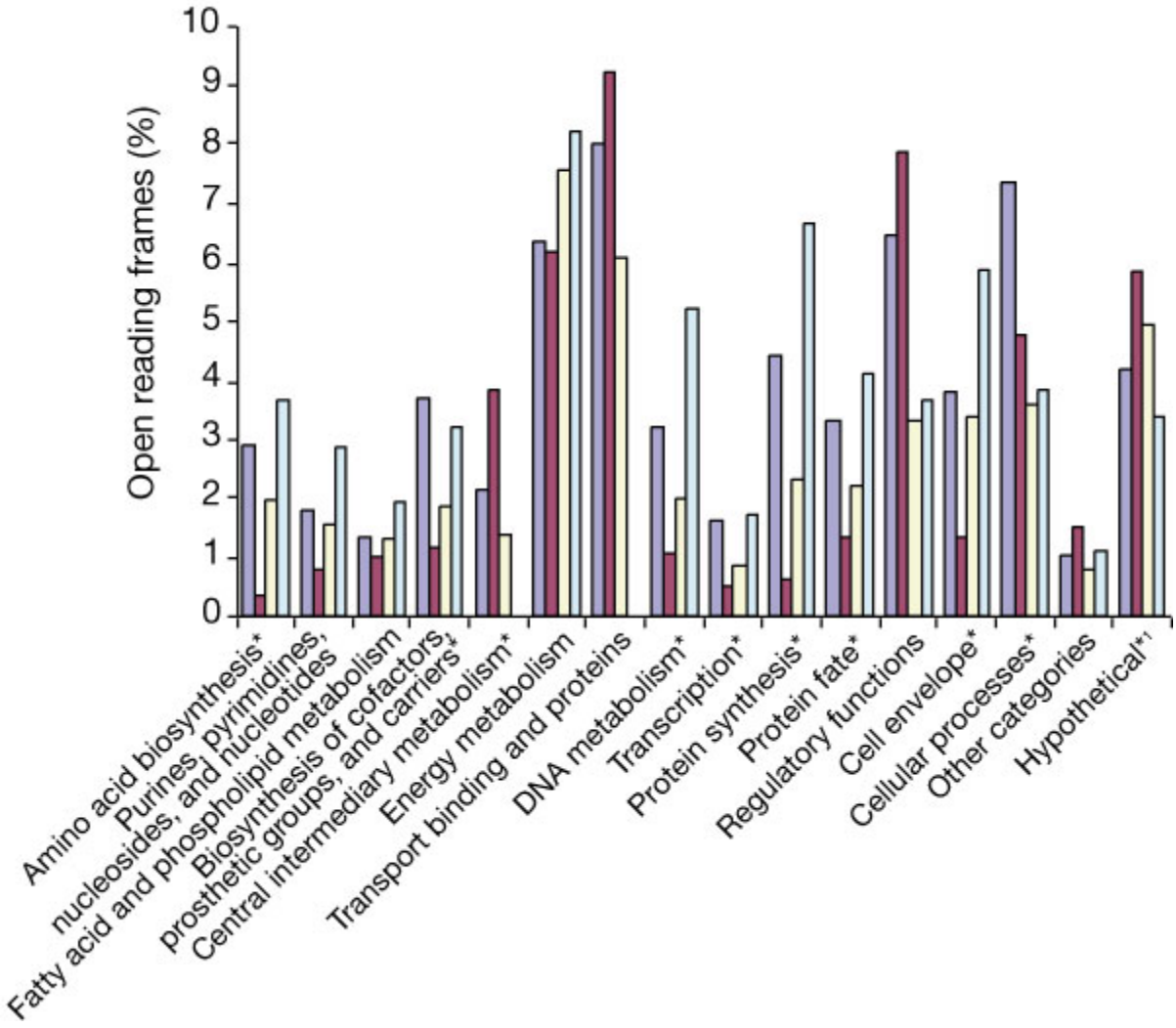


Figure 4 Percentage of total *Vibrio cholerae* open reading frames (ORFs) in biological roles compared with other γ-Proteobacteria. These were *V. cholerae*, chromosome 1 (blue); *V. cholerae*, chromosome 2 (red); *Escherichia coli* (yellow); *Haemophilus influenzae* (pale blue). Significant partitioning ($P < 0.01$) of biological roles between *V. cholerae* chromosomes is indicated with an asterisk, as determined with a χ^2 analysis. 1, Hypothetical contains both conserved hypothetical proteins and hypothetical proteins, and is at 1/10 scale compared with other roles.

Leading vs Lagging Strand

- Genes located on both strands
- Most genes on the leading strand
- But highly expressed genes preferentially on leading strand
    - Reason???
      DNA and RNA polymerase functions would collide on lagging strand

G+C Content

- Range
    - 22% *Wigglesworthis glossinidia* (tsetse fly endosymbiont)
    - 67% *Pseudomonas aeruginosa*
- Genome-wide
    - G+C not related to the thermal environment
    - But
        - For species living in elevated temperatures, structural RNAs have higher G+C content in ds regions
        - Aerobic genomes have higher G+C content than anaerobic bacteria
- Within a species
    - G+C content does not vary among genes
    - Genes with unusual G+C content are indicative of lateral transfer of genes

Operons

Definition

- Cluster of gene under the control of a single promoter that are expressed as a single mRNA

Components

- Promoter
- Operator
    - Repressor protein binds to this site
- Gene(s)

Example:

- Lac Operon
- Features
    - Promoter
    - Operator
    - LacZ (beta-galactosidase)
    - LacY (beta-galactoside permease)
    - LacA (beta-galactoside transacetylase)
    - Activation
        - Increased levels of lactose
        - Repressor released

E. coli Operons

Prediction method

- Distance between genes
    - Is there enough distance for a promoter???
        - No: then genes are part of the same operon

Total

- 392 known
- 2192 predicted
- one gene
    - 73% (surprisingly high)
- two genes
    - 16.6%
- three genes
    - 4.6%
- four or more genes
    - 6.0%

E. coli promoters

- 2584 operons
    - 2402 predicted promoters
- one promoter per operon
    - 68%
- two promoters per operon
    - 20%
- three or more promoters
    - 12%

Horizontal (or Lateral) Gene Transfer in Bacterial Genomes

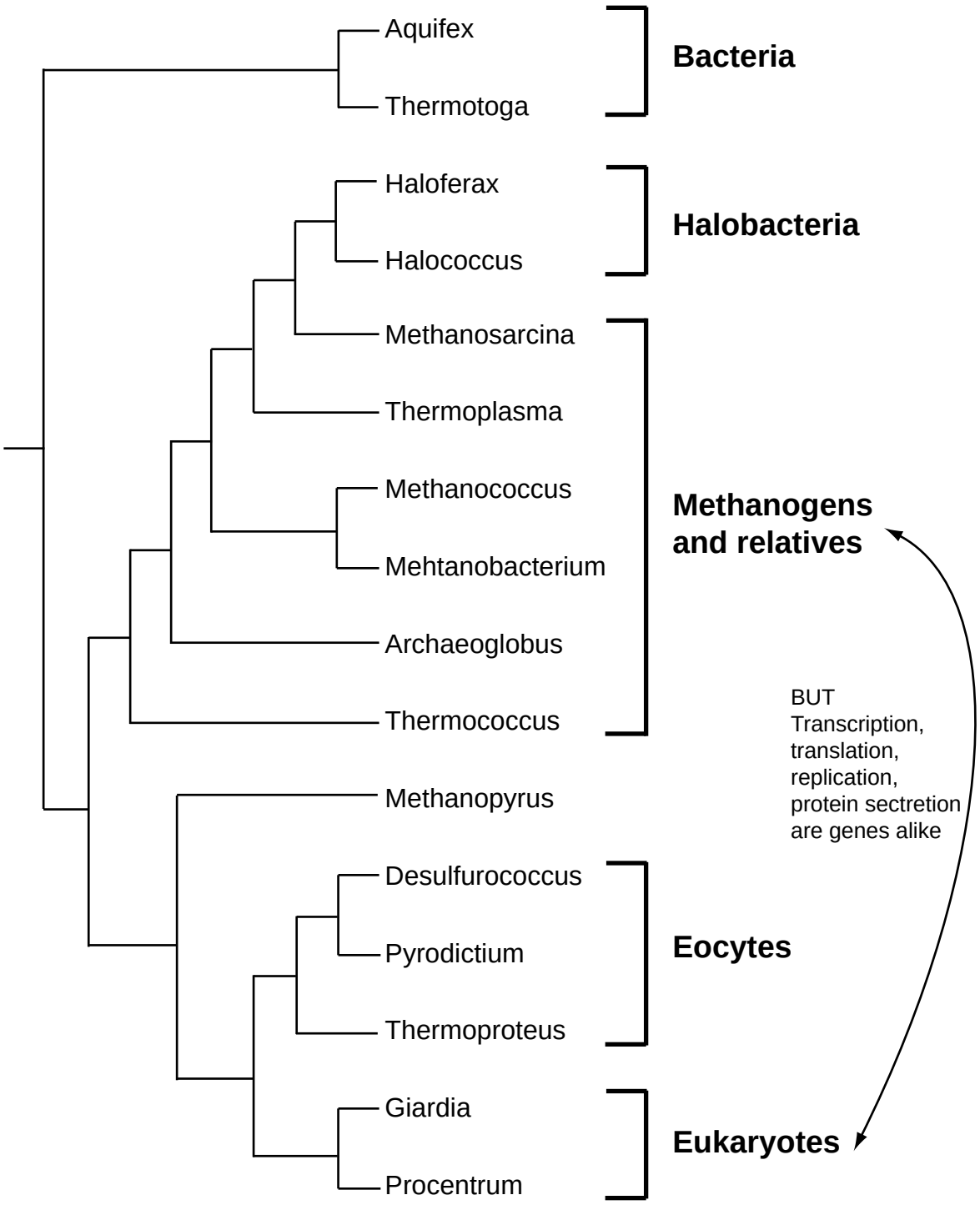
How is DNA transferred between bacterial genomes?

- Mechanisms are known
  - Transformation
    - Free DNA is known to exist in the biological world
      - Bacteria are known to take DNA up from the environment (=transformation)
    - Influenced by high population density
    - Influenced by salt concentration
    - Practical use
      - Introducing foreign DNA into bacteria for cloning
  - Conjugation
    - Well studied biological function of bacteria
      - A pilus is formed between bacteria
      - DNA is transferred between cells via the pilus
      - Can occur between distantly related species
        - *E. coli* and cyanobacteria
        - *E. coli* and yeast
  - Transduction
    - Transferred mediated by viruses
    - Bacterial genes encapsulated in viral genome by mistake
    - Bacterial genes transferred along with the viral gene
    - Detected as “foreign bacteria genes” surrounded by phage sequences

How is HGT detected from genomic sequences

- Proteins
  - Orthology with a distant taxon
  - More similar to eukaryotes (for example) than other bacteria
- Phylogeny
  - Protein groups with eukaryotes (for example) rather than other bacteria (See 16S rDNA tree)
- Phyletic
  - Bacterial COG lineage with archael/eukaryotic species
  - Archael/eukaryotic COG of bacterial origin
    - Example: Bacterial DNA gyrase A subunit found in archael species
- Conserved Gene Order
  - Gene shuffling during evolution is extensive (except for operons)
  - Conservation of order of three genes is unlikely (except for operons)
  - Distantly conserved operons evidence of HGT
    - Example:
      - Nitrate reducatase GHJI
      - Species
        - *Aeropyrum pernix* (archaea)
        - *Pyrococcus abyssi* (archaea)
        - *E. coli* (eubacteria)
        - *Mycobacterium tuberculosis* (eubacteria)
- Unusual localized G+C content
  - Genomes have a specific G+C content
  - Some genes have G+C content drastically different than the average
  - These are considered HGT events

16sRNA Gene Tree of Bacteria



HGT: How often is the occurring

- Range
    - Mycoplasma genitalium
        - 1.6% of genes HGT derived
    - Treponema pallidum
        - 32.6% of genes HGT derived

Early example

- *Methanococcus jannaschii* (Archaea)
    - Housekeeping genes
        - Most like *E. coli* and *Syneccoystis* (cyanobacteria)
    - Transcription, translation, replication, protein secretion genes
        - Most like eukaryotes

Eukaryote to Microbial HGT

- Eukaryotes are a source of bacterial genes
- Trend
    - Plant symbionts receive more horizontally transferred genes from plants
    - Animal symbionts receive more horizontally transferred genes from animals
    - Exceptions exists
        - Chlamydia
            - Animal pathogen has more horizontally transferred genes from plants than animals

The Aminoacyl tRNA story

- Genome Research (1999, 9:689)
    - HGT genes replaced the bacterial gene
    - Examples
        - Tyrosine aaRS
            - HGT from gram-positive to E. coli
        - Tryptophan aaRS
            - HGT from eukaryotes to the archea lineage
        - Leucine aaRS
            - no HGT
        - Alanine aaRS
            - HGT from bacteria (via mitochondria) to eukaryotes

Major effects of HGT

- Conservation of the genetic code
    - HGT requires selection for a common mechanism to express genes
    - Genetic code the most basic “common mechanism”
- Difficult to construct a “deep” tree of life
    - Need a “gene-by-gene” approach to phylogeny

Minimal Gene Set Concept

Pioneer

- Eugene V. Koonin

Goal

- Define the minimal set of genes necessary for life
- Koonin definition (Ann Rev Genomics Human Genet. 2000. 1:99)
    - "…the smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favorable conditions imaginable, that is, in the prescence of a full complement of essential nutrients and in the absence of environmental stress."
- How is this derived?
    - Compare the gene sets in small bacterial genomes
- Assumption
    - Small genomes contain the least amount of additional genes beyond the minimal set

Initial Experiment

Compare *Mycoplasma genitalium* and *Haemophilus influenzae*

- Include housekeeping genes because single cell organisms don't take up proteins
- Shared genes among simple genomes are essential
- First result
    - 240 orthologs
- But
    - Some pathways were not complete
- Additional concept
    - NOD: Non-orthologous displacement
        - Same function (a pathway enzyme, for example) is provide by two distinctly different genes
        - Genes may have evolved to the point that they don't appear to be orthologus

The First Minimal Gene Set (PNAS. 1996. 93:10268)

256 genes

- about 5% NOD

Basic Functions/Systems

1. Translation
2. DNA replication
3. Recombination and Repair
4. Transcription
5. Chaperone-like proteins
6. Anaerobic metabolism (glycolysis and phosphroylation)
7. Glutamyl-tRNA to glutaminyl-tRNA conversion
8. Nucleotide salvage pathways (except thymine)
9. Condensation of fatty acids with glycerol
10. Eight cofactor biosynthesis enzymes and eight enzymes requiring complex cofactors
11. Protein export
12. Limited metabolite transport systems involving ATPase and permeases; probably have broad specificity

What is excluded?

1. Amino acid biosynthesis (except glutyamyl-tRNA to glutaminyl-tRNA)
2. *de novo* nucleotide biosynthesis
3. Fatty acid biosynthesis
4. Defense systems

How does this hold up after 25 genomes?

Functional class	**# of genes**
Translation and ribosomal biogenesis	93
Transcription	8
Replication	18
Repair and recombination	11
Chaperone functions	14
Nucleotide metabolism and transport	17
Amino acid metabolism and transport	7
Lipid metabolism	6
Energy production and conversion	35
Coenzymes	8
Cell division, exopolysaccharide metabolism	6
Inorganic ion transport	5
Secretion, protein membrane translocation	6
Miscellaneous	16
Total	250

Ubiguitous (found in all species) COGs of the minimal gene set

Functional class	**# of universal genes**
Translation, and ribosomal biogenesis	53
Transcription	4
Replication, repair and recombination	5
Metabolism	9
Cellular processes: (chaperone functions, secretion, cell division, cell wall biogenesis)	9
Miscellaneous	1
Total	81

Archaea Kingdom: Discovering the Original Gene Set

(abstracted from Genome Biology (2003) 4:115)

Archaea Recognized in 1977

- "Bacterial" species distinct from eubacteria
- Compared species using 16S rDNA sequences
- Abstract of 1977 publication
    - "A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (i) the eubacteria, comprising typical bacteria; (ii) the archaebacteria, containing methanogenic bacteria; and (iii) the ukaryotes, now represented in the cytoplasmic component of eukaryotic cells."
- Equal standing with eubacterial and eukaryote kingdoms

Archaea are divided into three "kingdoms"

- Euryarchaeota
    - Methanogens
    - Extreme halophiles
    - Thermoplasma
- Crenarchaeota
    - Hyperthermophiles
    - Cold dwellers
- Korarchaeota
    - Recently discovered
    - Little known

Closer to Eukaryotes than Bacteria

- Evidence
    - Presence of histones
    - Ribosome structures
    - Sequence conservation of some genes
        - DNA replication
        - DNA repair
        - Transcription
        - Translation
    - But for metabolic genes
        - Arhcaea are closer to eubacteria

Archaea form a unique ancestral biological form

Interestingly

- Some DNA replication factors involved in
    - initiation
    - elongation
    - replication
        - DNA unwinding (helicases)
    - are distinct from eukaryotes and eubacteria
- ***Was double-stranded DNA replication invented twice???***

The Uniqueness of Archaea: Gene Set Perspective

- As of 2003
  - 16 Archaea genomes have been sequenced
    - 12 Euryarchaeota
    - 4 Cenarchaeota
  - Protein set
    - Range
      - 1,482 – 4,540 proteins
      - 59% - 82% found in COGs
    - COGs
      - 313 found in all archaea species
        - Most involved in
          - Translation
          - RNA mdodification
      - 16 COGs only found in archaea
      - 61 exclusive to archaea and eukaryotes
        - Predominantly information processing COGs (except for two)

The Deep Tree of Life: Evolution of Gene Sets

- Phylogeny is typically:
  - Species based
  - Gene based
- Complete genomes alter the approach
  - Genome trees are now possible
  - These trees must account for
    - Gene gain
    - Gene loss
    - Horizontal gene transfer
    - Inherited functionality (shared COGs)
- What does genome tree analysis tell us about evolution of gene sets?
  - Gain and loss of genes occurs at equal probabilities
- Combining two approaches can define ancestral gene sets
  - Species tree phylogeny
  - Investigations of shared COGs

What is observed?

- “Last Universal Common Ancestor”
    - 505 genes
- Eubacteria vs LUCA
    - Gene gain only
    - LUCA was a usable gene set
    - 897 genes
- Archaea/Eukaryote Ancestor vs. LUCA
    - Gene gain only
    - 667 genes
- Eukaryote vs. Archaea/Eukaryote ancestor
    - Significant gain of genes
    - Loss of gene occurred
    - 929 genes
- Archaea vs. Archaea/Eukaryote ancestor
    - Significant gain of genes
    - Minimal gene loss compared to eukaryotes
    - 870 genes

Conclusions

- A shared repertoire of genes with deep roots exists in all species
- Lineage specific gene gain and loss occurs