Bacterial Genomes

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

- ***Soi***l
  - Interact with plants
    - Nitrogen fixation
      - *Rhizobium* species
- ***High-salt conditions***
  - Halophiles
    - Great Salt Lake, Utah, USA
- ***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

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 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
- **October 2020**
  - 20,475 species publicly released complete genomes
    - 20,172 bacteria
    - 403 archaea
- **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 and NCBI

Species	Interest	Size (nt)	G+C (%)	Coding density	bp/gene	# Genes (% unique)*
Candidatus Hodgkinia cicadicoola	Smallest genome	105,760	46			378
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	1,374 (36)
Haemophilus influenzae	Flu	1,830,138	39	85	1070	1,709 (14)
Lactococcus lactis	Cheese starter	2,365,589	36	84	1043	2,266 (25)
Clostridium tetani	Tetnus	3,799,251	29	85	1179	2,373
Clostridium perfringens	Gangrene	3,031,430	29	83	1139	2,660
Mycobacterium leprae	Leprosy	3,268,203	58	76	1201	2,720 (27)
Clostridium acetobutylicum	Solvent producer	3,940,880	31	85	1073	3,672 (26)
Vibrio cholerae (Total)	Cholera	4,033,464	48	86	1053	3,828 (20)
Chromosome 1		2,961,149	48	87	1082	2,736
Chromosome 2		1.072,315	47	84	982	1,092
Mycobacterium tuberculosis	Tuberculosis	4,411,529	66	90	1126	3,918 (20)
Escherichia coli	Model organism					
K-12 MG1655		4,639,221	51	87	1081	4,289
K-12 W3110		4,641,433	51	87	1057	4,390
CFT073		5,231,428	51	87	972	5,379
O157 RIMD0509952		5,498,450	51	87	1025	5,361
O157 EDL93		5,528,970	51	87	1033	5,349 (25)
Yersinia pestis	Plague	4.653,728	48	83	1161	4,008 (17)
Xanthomonas campestris	Citrus canker	5,076,188	66	84	na	4,181
Vibrio parahaemolyticus	Gastroenteritis	5,165,770	46	86		4,832
Chromosome 1		3,288,558	46	86	1067	3,080
Chromosome 2		1,877,212	46	86	1071	1,752
Bacillus anthracis	Anthrax	5,227,293	36	84	800	5,508
Agobacterium tumefaciens	Transformation	5,673,563	60	89	1070	5,299
Chromosome 1 (circular)	vector	2.841,581	60	88	1040	,2721 (16)
Chromosome 2 (linear)		2,074,782	60	90	1131	1,833
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	5,471
Anabaena nostic	Photosynthesis	7,211,893	42	82		,6129
Streptomyces avermitilis	Antibiotics	9,025,608	71	86	1191	7,575
Bradyrhizobium japonicum	N2 fixation	9,105,828	65	86	1094	8,317
Minicystis rosea	Largest genome	14,782,125	72			10,452

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

General Features of Bacterial Genomes

- ***140X differences in genome sizes***
  - *Candidatus Hodgkinia cicadicoola vs. Minicystis rosea*
- 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

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
  - 14.0% *Candidatus Carsonella ruddii*
  - 75.6% *Cellulomonas* sp. JZ18
- 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

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:

- ***E. coli Lac Operon***
    - Features
        - Promoter
        - Operator
        - Genes
            - LacZ (beta-galactosidase)
            - LacY (beta-galactoside permease)
            - LacA (beta-galactoside transacetylase)
        - **Activation of Lac operon genes**
            - 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%

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 proteins
        - **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**

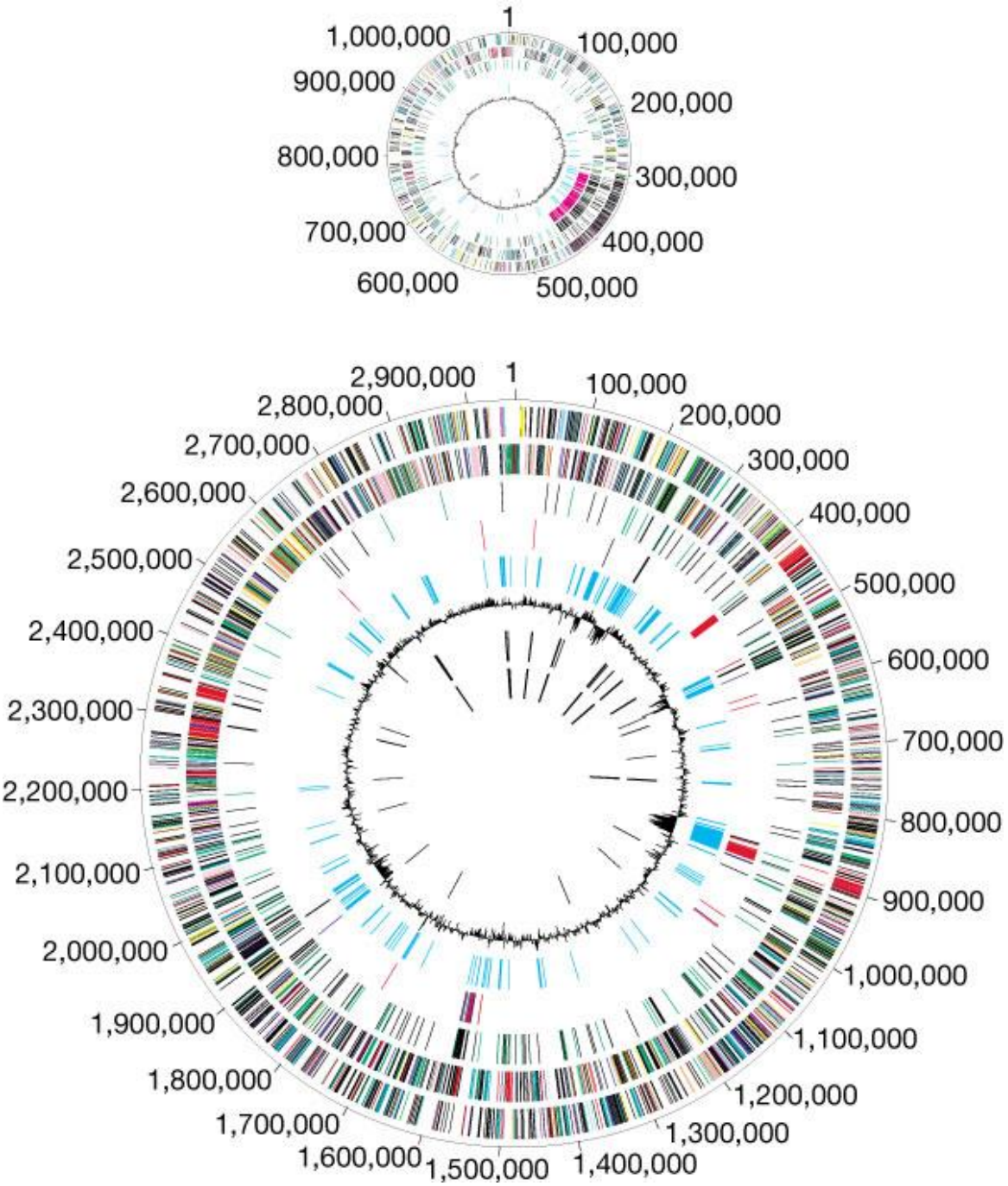

Figure 2 Circular representation of the *V. cholerae* genome. The two chromosomes, large and small, are depicted. From the outside inward: the first and second circles show predicted protein-coding regions on the plus and minus strand, by role, according to the colour code in Fig. 1 (unknown and hypothetical 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.

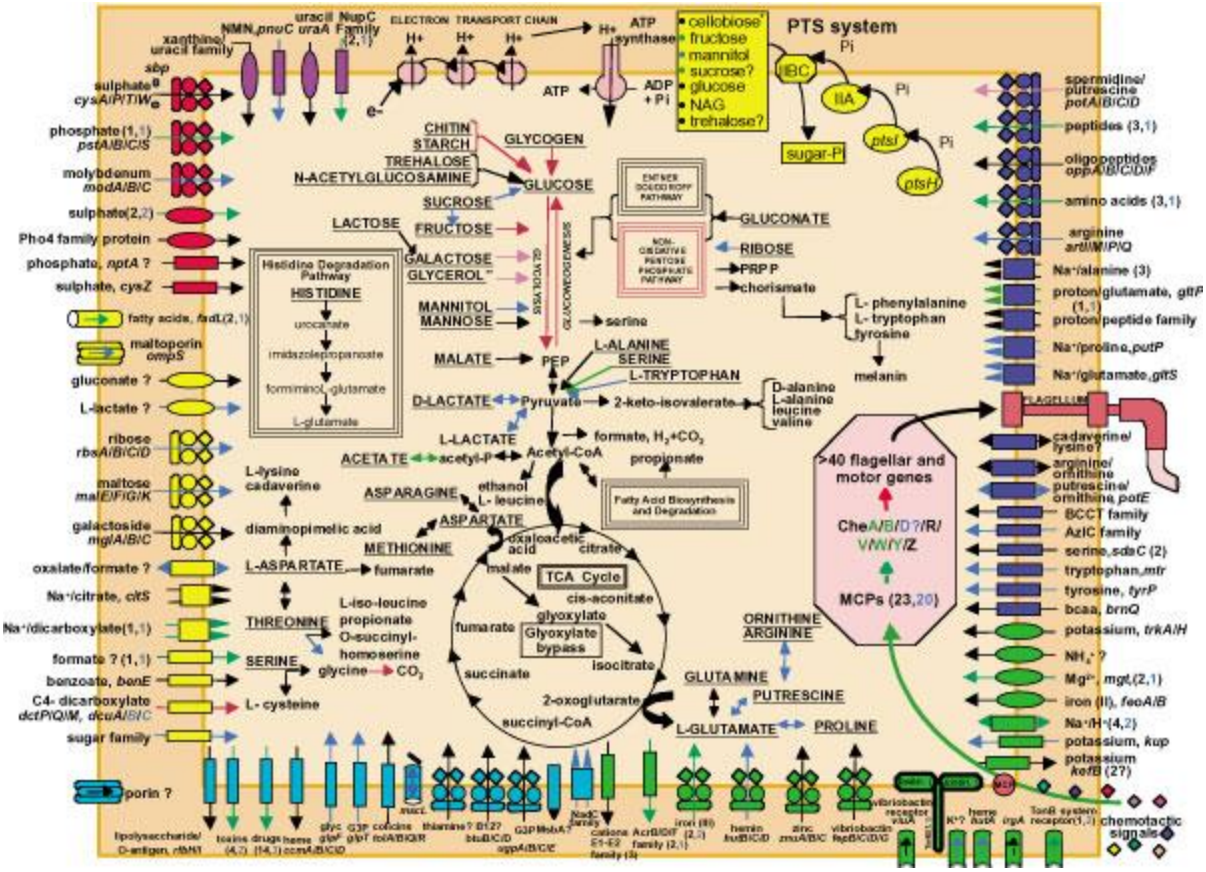

Figure 3 Overview of metabolism and transport in *V. cholerae*. Pathways for energy production and the metabolism of organic compounds, acids and aldehydes are shown. Transporters are grouped by substrate specificity: cations (green), anions (red), carbohydrates (yellow), nucleosides, purines and pyrimidines (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.

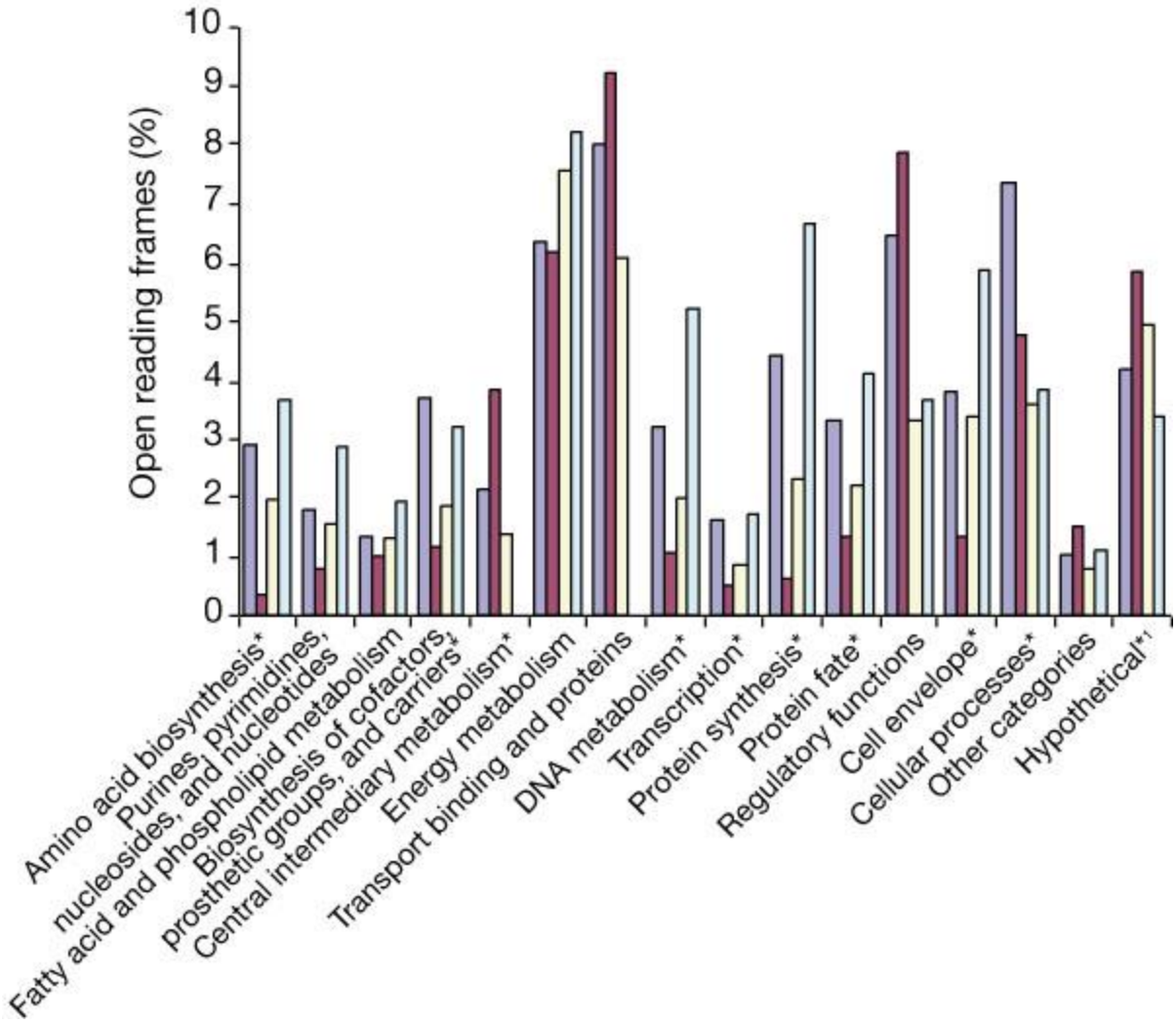

Figure 4 Percentage of total *Vibrio cholerae* open reading frames (ORFs) in biological roles compared with other γ-Proteobacteria. These were *V. cholerae*, chromosome 1 (blue); *V. cholerae*, chromosome 2 (red); *Escherichia coli* (yellow); *Haemophilus influenzae* (pale blue). Significant partitioning ($P < 0.01$) 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.