



Геном прокариот

Н.Н. Колесников

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia



Таблица 1. Состав сложных геномов бактерий [19]

| Бактерия | Геном |
|--------------------------------------|---|
| <i>Agrobacterium tumefaciens</i> C58 | 1 линейная хромосома; 1 кольцевая хромосома; 2 плазмиды |
| <i>Bacillus cereus</i> F0836 76 | 1 кольцевая хромосома; 1 мегаплазида |
| <i>Brucella melitensis</i> | 2 кольцевых хромосомы |
| <i>Leptospira interrogans</i> | 1 кольцевая хромосома; 1 мегаплазида |
| <i>Rhizobium meliloti</i> | 1 кольцевая хромосома; 2 мегаплазмиды |
| <i>Rhodobacter sphaeroides</i> | 2 кольцевых хромосомы |
| <i>Rhodococcus facians</i> | 1 линейная хромосома; линейная плазида |
| <i>Streptomyces ambofaciens</i> | 1 линейная хромосома |
| <i>Streptomyces lividans</i> 66 | 1 линейная хромосома; линейные плазмиды |

Таблица 2. Характеристики геномов, определенные по их полной нуклеотидной последовательности

| Видное название | До- мея | Размер генома (п. н.) | % копир- руемых последо- ватель- ностей | % GC | Колоче- ство: ORF | Характеристика (для тер- мофилов указана оптима- льная температура роста T_{opt}) | Лит. |
|---|------------|---|---|---|--|---|------|
| <i>Mycoplasma genitalium</i> | B | 580070 | 89.7 | 52 | 479 | Вызывает воспаленная мочеполовых путей, грам+ ^h | [19] |
| <i>Mycoplasma pneumoniae</i> | B | 816394 | 88.7 | 40 | 677 | Возбудитель пневмонии, грам+ ^h | [22] |
| <i>Borrelia burgdorferi</i> | B | 910725 ^l 532000 ^{h*} | 93 | 28.6 ^h от 23.6 до 28.1 ^{h*} | 85.5 ^h 430 ^{h*} | Возбудитель клещевого спирохетоза (болезнь Лайма), грам-. Хозяева: грызуны, ископаемые клещи, человек | [23] |
| <i>Chlamidia trachomatis</i> | B | 1042519 + плаз- мида 7493 | | 41.5 | 894 | Вызывает трахому и воспаление мочеполовых путей, грам- ^{h*} | [24] |
| <i>Rickettsia prowazekii</i> | B | 1111523 | 76 | 29.1 | 834 | Возбудитель сыпного тифа, грам- | [25] |
| <i>Treponema pallidum</i> | B | 1138006 | 92.9 | 52.8 | 1041 | Возбудитель сифилиса, грам- | [26] |
| <i>Chlamidia pneumoniae</i> | B | 1230230 | | 40.6 | 1073 | Возбудитель пневмонии, грам- ^{h*} | [27] |
| <i>Aquifex aeolicus</i> | B | 1551335 | 93 | 43.4 | 1512 | Хемолитотроф, микроаэро- фил, термофил, T_{opt} 85°C, грам- | [28] |
| <i>Helicobacter pylori</i> , штамм 26695 | B | 1667867 | 91.0 | 39 | 1552 | Вызывает язвенную болезнь, гастрит, грам- | [29] |
| штамм J99 | | 1643831 | 90.8 | 39 | 1495 | | [30] |
| <i>Methanococcus jannaschii</i> | A | 1660000 плазмиды: 58000 | | 31.4 | 1738 | Анаэроб, метаноген, обитает в глубоководных термальных источниках, T_{opt} 85°C | [31] |
| | | 16000 | | 28.2 | 44 | | |
| | | | | 28.8 | 12 | | |
| <i>Pyrococcus horikoshii</i> | A | 1738505 | 90.7 | 41.9 | 2061 | Анаэроб, гипертермофил, T_{opt} 98°C | [32] |
| <i>Methanobacterium thermoautotrophicum</i> | A | 1751377 | | | 1855 | Термофил, T_{opt} 65°C | [33] |
| <i>Haemophilus influenzae</i> | B | 1830137 | | 38 | 1073 | Вызывает отиты, ОРЗ, возбу- дитель менингитов, грам- | [34] |
| <i>Thermotoga maritima</i> | B | 1860725 | 95 | 46 | 1877 | Один из древнейших видов зубактерий, T_{opt} 65°C | [35] |
| <i>Archaeoglobus fulgidus</i> | A | 2178400 | 92.2 | 48.5 | 2436 | Метаболизует серу, T_{opt} 83°C | [36] |
| <i>Synechocystis</i> sp. PCC6803 | B | 3573470 | 87 | | 3168 | Фотосинтезирующая цианобактерия, грам- | [37] |
| <i>Bacillus subtilis</i> | B | 4214810 | 87 | 43.5 | 4100 | Автотроф, грам- | [38] |
| <i>Mycobacterium tuberculosis</i> H37Rv | B | 4411529 | 91 | 65.6 | 4000 | Возбудитель туберкулеза, грам+ | [39] |
| <i>Escherichia coli</i> K-12 | B | 4639221 | 88.6 | 50.8 | 4288 | Прототрофная аэробная бактерия, грам- | [40] |
| <i>Saccharomyces cerevisiae</i> | E | 12068000 (16 хромосом) | | | 5885 | | [41] |

Примечания. А - архей, В - бактерии, Е - эукариоты. Обозначения грам+ и грам- указывают тип клеточной стенки.

^h данные относятся к линейному геному *B. burgdorferi*.

^{h*} суммарные данные по 11 плазмидам *B. burgdorferi*, нуклеотидная последовательность которых определена. Всего в клетке боррелии может присутствовать до 20 различных плазмид.

^{*} относится к ветви грам+ бактерий, не имеет клеточной стенки.

^{h*} не имеет муреинового слоя, клеточная стенка грам- типа.

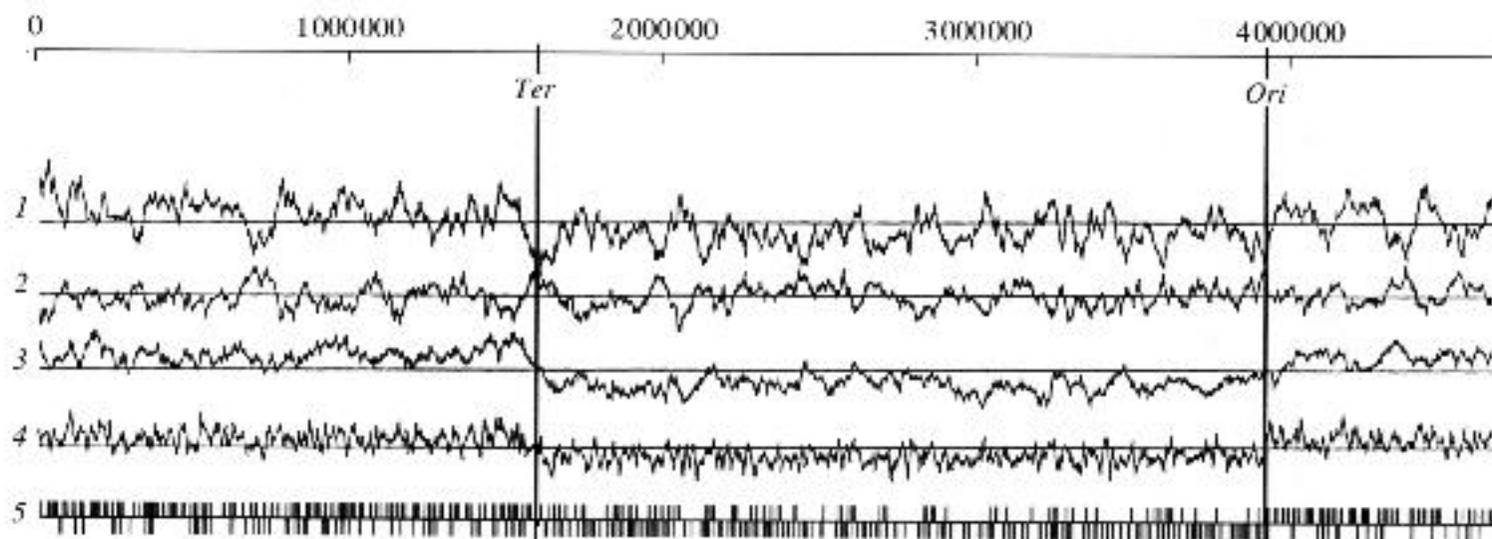


Рис. 1. Некоторые характеристики генома *E. coli*, установленные по полной нуклеотидной последовательности. Вертикальные линии указывают положение участков *ori* и *ter* репликации. Наверху указаны позиции нуклеотидов в соответствии с их нумерацией в геноме. Линии 1–4 – профиль распределения G/C для всех позиций кодона (определен как отношение $(G - C)/(G + C)$ на интервалах 10 т.п.н. по одной нити ДНК):

1 – 1-ая позиция кодона; 2 – 2-ая позиция; 3 – 3-я позиция; 4 – по всем позициям. Содержание G повышено в ведущей нити, причем преимущественное присутствие G наиболее четко выражено в 3-й позиции кодона. Линия 5 – расположение октамеров GCTGGTGG (Chi). Черточки указывают позицию октамера, а направление вверх или вниз от горизонтальной линии указывает нить, содержащую данную последовательность [40].

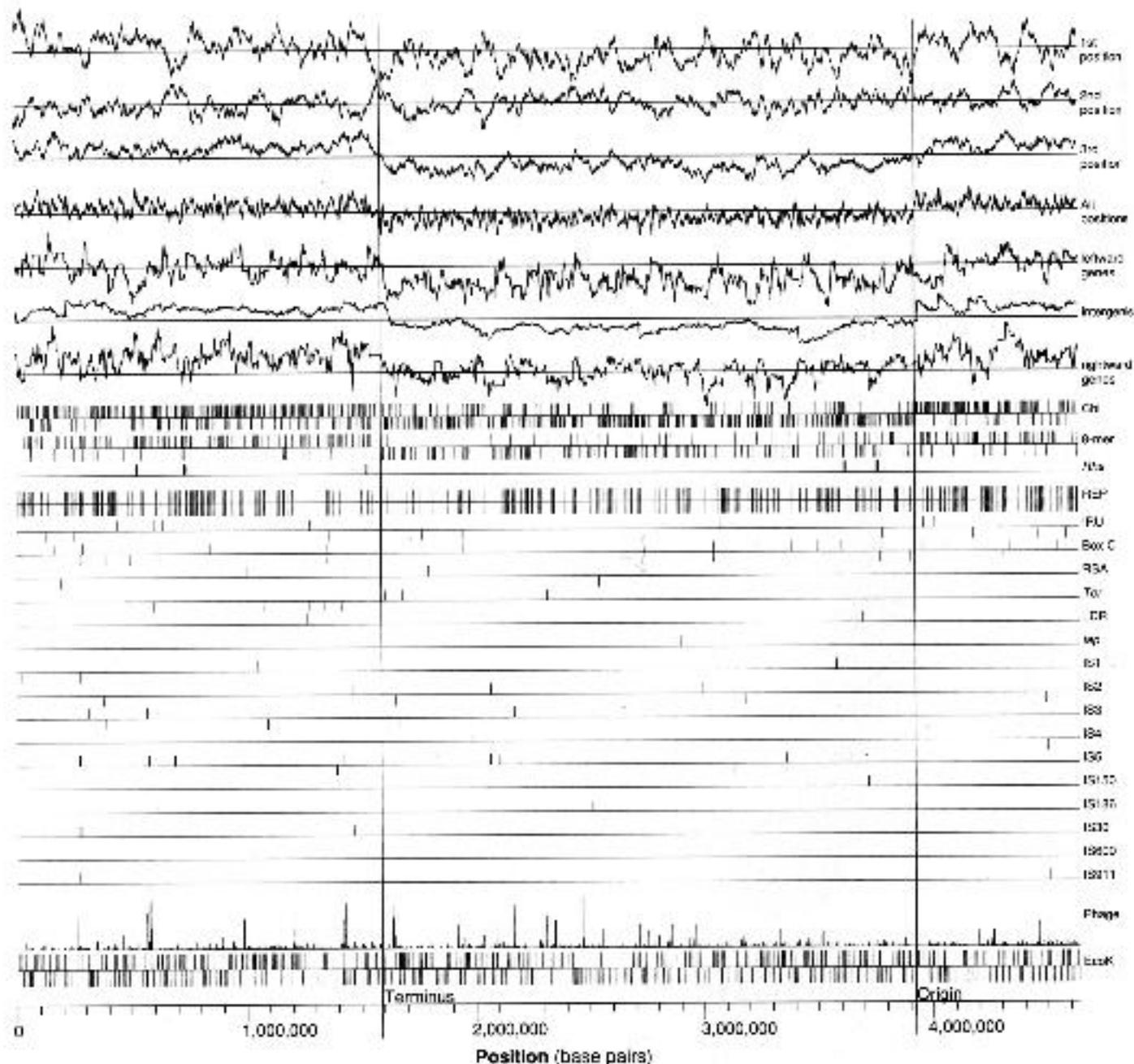


Fig. 2. Base composition is not randomly distributed in the genome. G-C skew $(G - C)/(G + C)$ is plotted as a 10-kb window average for one strand of the entire *E. coli* genome. Skew plots for the three codon positions are presented separately: leftward genes, rightward genes, and non-protein-coding regions are shown in lines 5, 6, and 7. The two horizontal lines below the skew plots show the distribution of two highly skewed octamer sequences, CCTCGTCC (CH1) and GCAGGGCG (3-mer). Tick marks indicate the position of each copy of a sequence in the complete genome; and are vertically offset to indicate the strand containing the sequence. The

next 15 horizontal lines correspond to distinct classes of repetitive elements. The penultimate line contains a histogram showing the similarity (the product of the percent of each protein in the pairwise alignment and the percent amino acid identity) across the signed region of known shape proteins to the proteins encoded by the complete *E. coli* genome. The last line indicates the position and orientation of the EcoK restriction-modification site AACNNNNNGTGC (N, any nucleotide). Two vertical lines through the plots show the location of the origin and terminus of replication.



СТРУКТУРА ПРОКАРИОТИЧЕСКИХ ГЕНОМОВ

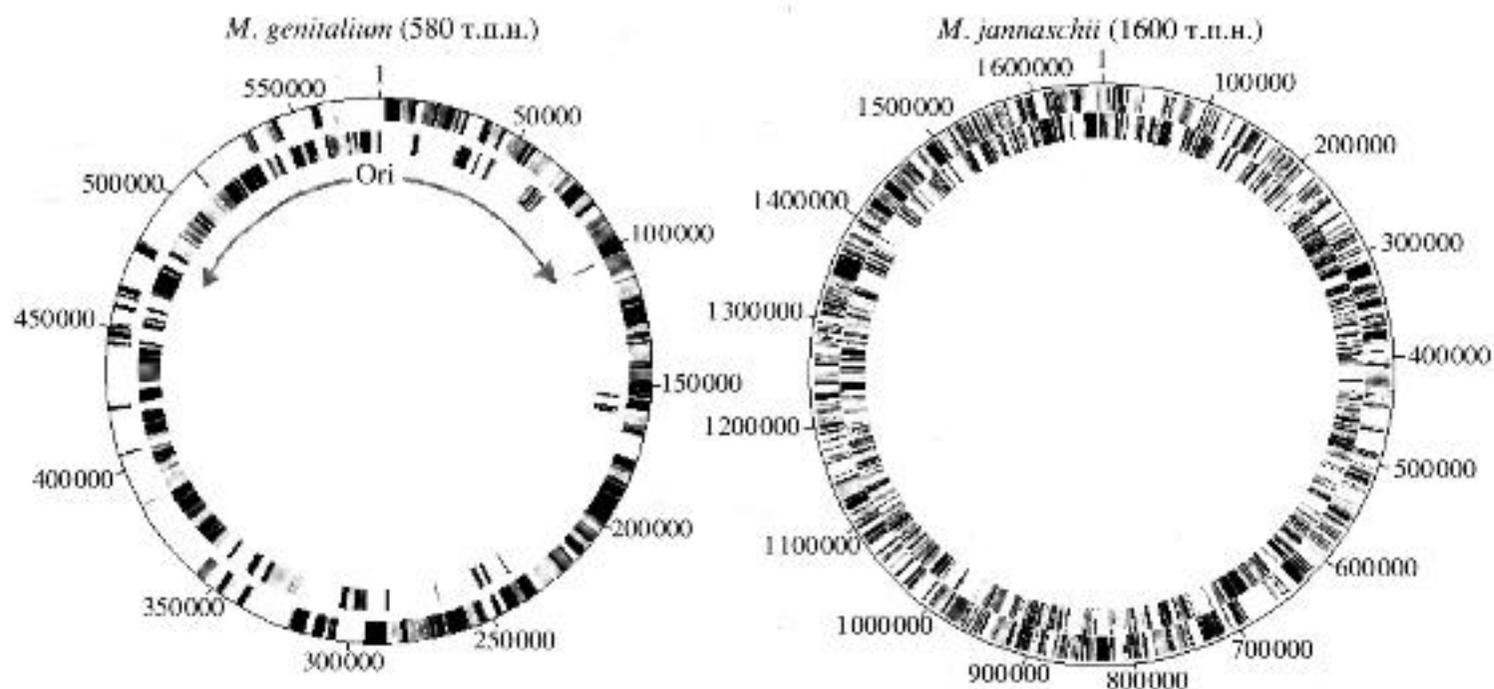


Рис. 2. Распределение кодирующих последовательностей в геномах *M. genitalium* и *M. jannaschii*. Кодирующие районы, указанные на наружном кольце, транскрибируются с "+" нити, а на внутреннем с "-" нити. Для *M. genitalium* стрелки, начинающиеся возле *ori*, указывают направление репликации [19, 31].



Рис. 3. Распределение функций генов *E. coli*, выявленных по полной нуклеотидной последовательности генома [50].



Таблица 3. Количество ORF у зубактерий и архебактерий (лит. источники см. в табл. 2)

| Видовое название | Всего ORF | Функция известна | Функция неизвестна | |
|------------------------------|-----------|------------------|------------------------------|------------------------------|
| | | | есть гомологи в базах данных | нет гомологов в базах данных |
| <i>M. genitalium</i> | 479 | 468 (97%) | | |
| <i>M. pneumoniae</i> | 677 | 603 (89%) | | |
| <i>B. burgdorferii</i> | 853* | 59% | 12% | 29% |
| | 430** | 70 (16%) | 100 (23%) | 250 (58%) |
| <i>Ch. trachomatis</i> | 894 | 604 (68%) | 35 (4%) | 255 (28%) |
| <i>R. prowazekii</i> | 834 | 63.7% | 12% | 24.8% |
| <i>T. pallidum</i> | 1041 | 577 (55%) | 177 (17%) | 287 (28%) |
| <i>Ch. pneumoniae</i> | 1073 | 636 (60%) | 251 (23%) | 186 (17%) |
| <i>A. aeolicus</i> | 1512 | 849 (56%) | 256 | 407 |
| <i>H. pylori</i> | 1590 | 1091 (69%) | | |
| <i>M. jannaschii</i> | 1738 | 38% | | |
| <i>P. korikoshii</i> | 2061 | 406 (19.7%) | 453 (22%) | 1202 (58.3%) |
| <i>H. influenzae</i> | 1073 | 58% | 20% | 335 (22%) |
| <i>T. maritima</i> | 1877 | 1014 (54%) | 407 (22%) | 373 (20%) |
| <i>A. fulgidus</i> | 2436 | 1797 | | 639 |
| <i>B. subtilis</i> | 4100 | | 42% | |
| <i>M. tuberculosis</i> H37Rv | 3924 | 40% | 44% | 16% |
| <i>E. coli</i> K-12 | 4288 | 62% | | |

* – число ORF в хромосоме.

** – суммарное число ORF в 11 плазмидях.

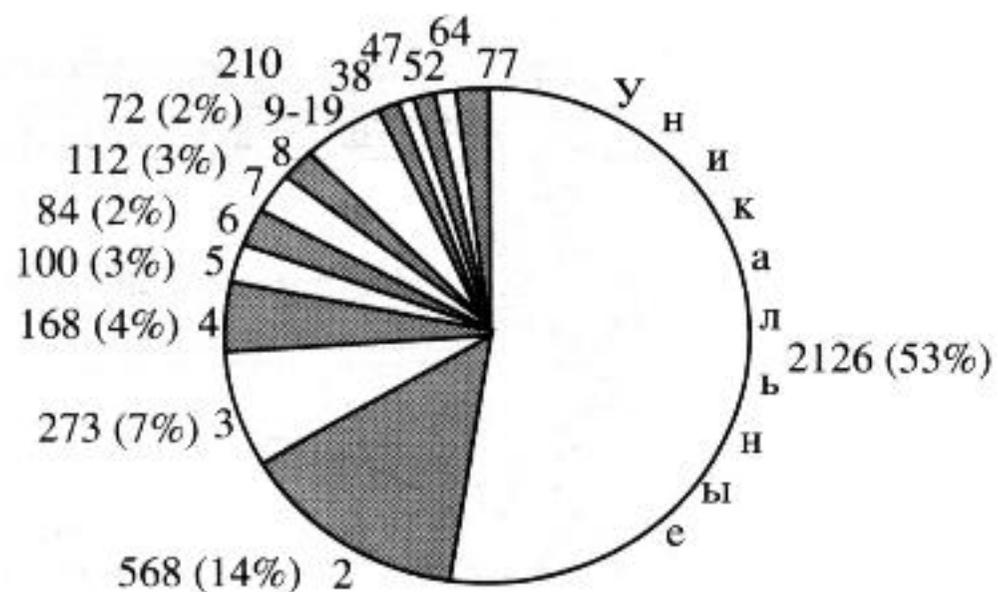


Рис. 4. Распределение числа паралогичных генов в геноме *B. subtilis* [38]. Вдоль окружности указана копияность генов, снаружи – число генов.



БОРИНСКАЯ, ЯНКОВСКИЙ



Рис. 5. Кластеры ортологических генов.

Обозначения: E – *E. coli*, H – *H. influenzae*, G – *M. genitalium*, P – *M. pneumoniae*, C – *Synechocystis* sp. PCC6803, M – *M. jannaschii*, Y – *S. cerevisiae* [66]. Над дробью указано число кластеров, под – общее число генов в каждой функциональной группе.



Таблица 4. Функции белков, соответствующих минимальному набору из 256 генов [72]

| Функция белков | Число белков |
|---|--------------|
| Преобразование энергии | 28 |
| Транспорт и метаболизм аминокислот | 11 |
| Транспорт и метаболизм нуклеотидов | 20 |
| Транспорт и метаболизм углеводов | 5 |
| Метаболизм кофакторов | 8 |
| Метаболизм липидов | 6 |
| Трансляция и биогенез рибосом | 94 |
| Репликация, транскрипция, рекомбинация, репарация | 35 |
| Структурная функция (белки наружной мембраны) | 7 |
| Секреция и адгезия | 5 |
| Шапероны | 13 |
| Транспорт неорганических ионов | 4 |
| Предсказана гипотетическая функция | 15 |
| Функция неизвестна | 4 |



Table 4. Distribution of *E. coli* proteins among 22 functional groups (simplified schema).

| Functional class | Number | Percent of total |
|--|--------|------------------|
| Regulatory function | 45 | 1.05 |
| Putative regulatory proteins | 133 | 3.10 |
| Cell structure | 182 | 4.24 |
| Putative membrane proteins | 13 | 0.30 |
| Putative structural proteins | 42 | 0.98 |
| Phage, transposons, plasmids | 87 | 2.03 |
| Transport and binding proteins | 281 | 6.55 |
| Putative transport proteins | 146 | 3.40 |
| Energy metabolism | 243 | 5.67 |
| DNA replication, recombination, modification, and repair | 115 | 2.68 |
| Transcription, RNA synthesis, metabolism, and modification | 55 | 1.28 |
| Translation, posttranslational protein modification | 182 | 4.24 |
| Cell processes (including adaptation, protection) | 188 | 4.38 |
| Biosynthesis of cofactors, prosthetic groups, and carriers | 103 | 2.40 |
| Putative chaperones | 9 | 0.21 |
| Nucleotide biosynthesis and metabolism | 58 | 1.35 |
| Amino acid biosynthesis and metabolism | 131 | 3.06 |
| Fatty acid and phospholipid metabolism | 48 | 1.12 |
| Carbon compound catabolism | 130 | 3.03 |
| Central intermediary metabolism | 188 | 4.38 |
| Putative enzymes | 251 | 5.85 |
| Other known genes (gene product or phenotype known) | 26 | 0.61 |
| Hypothetical, unclassified, unknown | 1632 | 38.06 |
| Total | 4288 | 100.00* |

*Total of these rounded values is 99.97%.

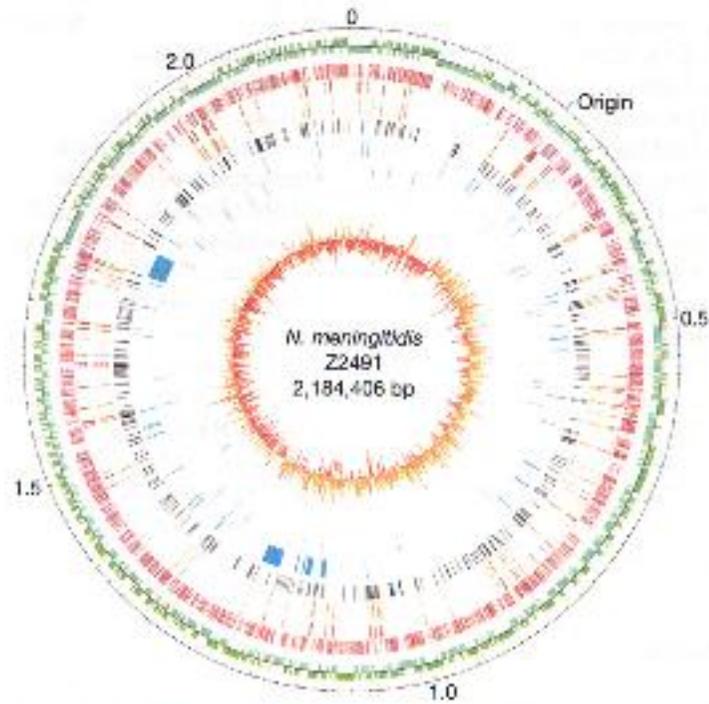


Figure 1 Circular representation of the *N. meningitidis* Z2491 genome. The concentric circles show, reading inwards: the scale in megabases, with the origin of replication indicated; predicted coding sequences clockwise (dark green) and anti-clockwise (light green); neisserial uptake sequences (red); dRSS sequences (dark orange); RS elements (light orange); dispersed repeats (Correla, ATR, REP2-5; black); IS elements and phage (narrow ticks and wide bars respectively; turquoise) and tandem repeats (dark blue). The inner histogram shows plot of (G-C)/(G+C) with values greater than zero in yellow and less than zero in orange. Figure generated with LASERGENE software (DNASTar).

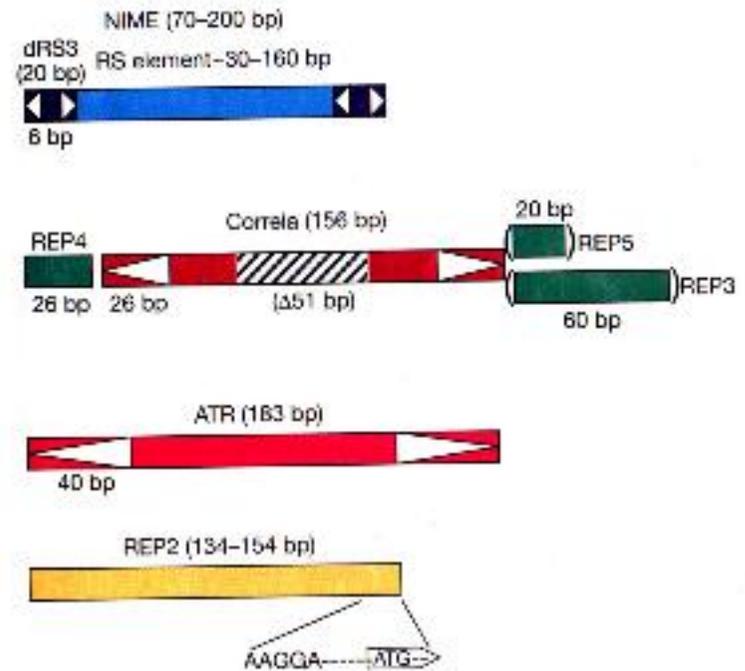


Figure 2 Types of *N. meningitidis* repeat. The name of each repeat type is indicated above the repeat. Inverted repeat sequences are represented by open triangles, and the internal deletion present in some Correla elements by a hatched box. The translational signals within REP2 are indicated below the repeat.

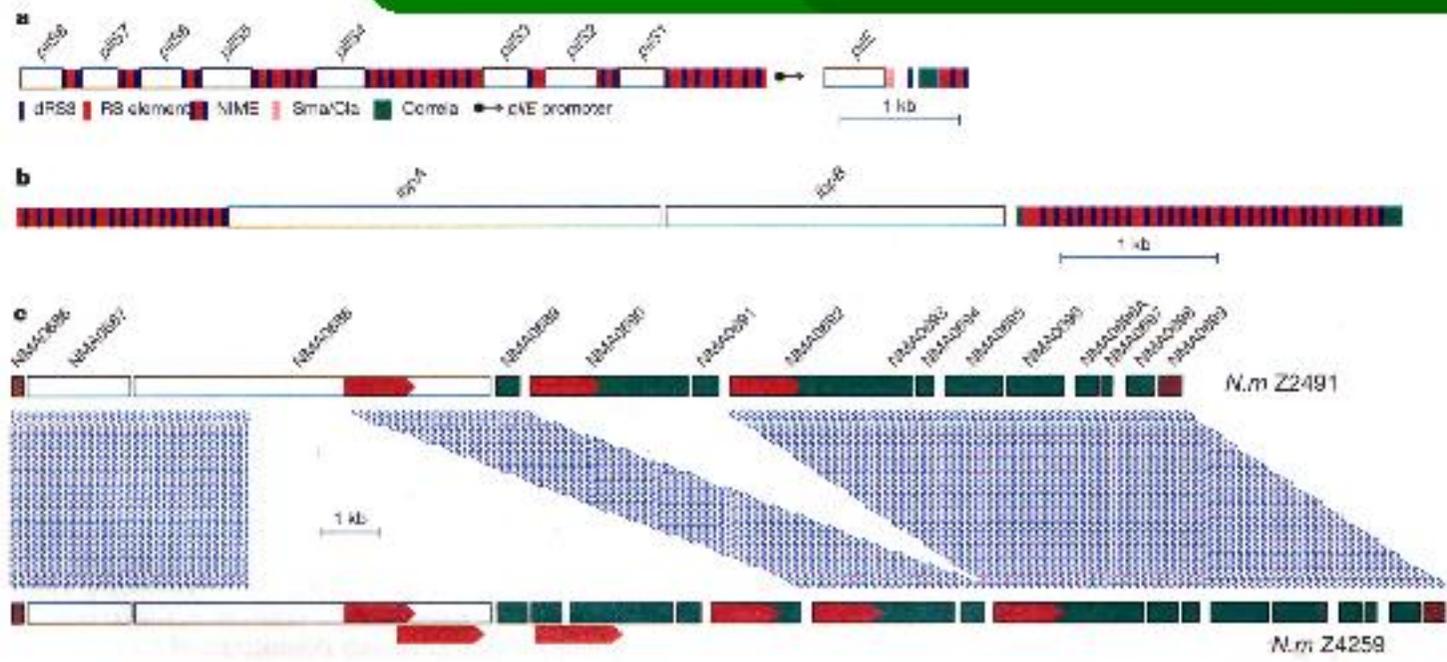


Figure 3 Structure of selected repeat regions. **a**, Repeats around the *pNE5* locus. *pN* genes are indicated by open boxes and repeats by coloured boxes as described in the key. **b**, Large repeat arrays around the *hboA* and *hboB* genes. Repeats are coloured as in **a**. **c**, The filamentous haemagglutinin homologue gene sequences of *N. meningitidis* Z2491 and Z4259. Conserved sequences are indicated by light blue bars, and internal repeat

regions by red arrows. The complex repeat structures have been simplified for clarity. Open boxes represent genes encoding surface-exposed proteins, and green boxes represent genes with no database matches. The brown boxes represent the two halves of an ABC transporter pseudogene.

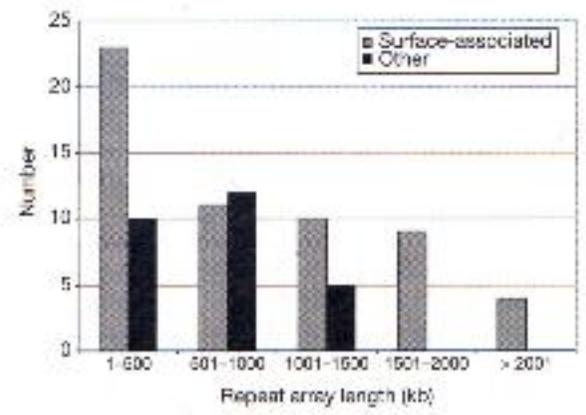


Figure 4 Graph showing the relationship between repeat array length and flanking gene function. For each category of repeat array length the number of flanking genes

associated with surface structures is shown in grey, and the number of genes in all other categories is shown hatched.

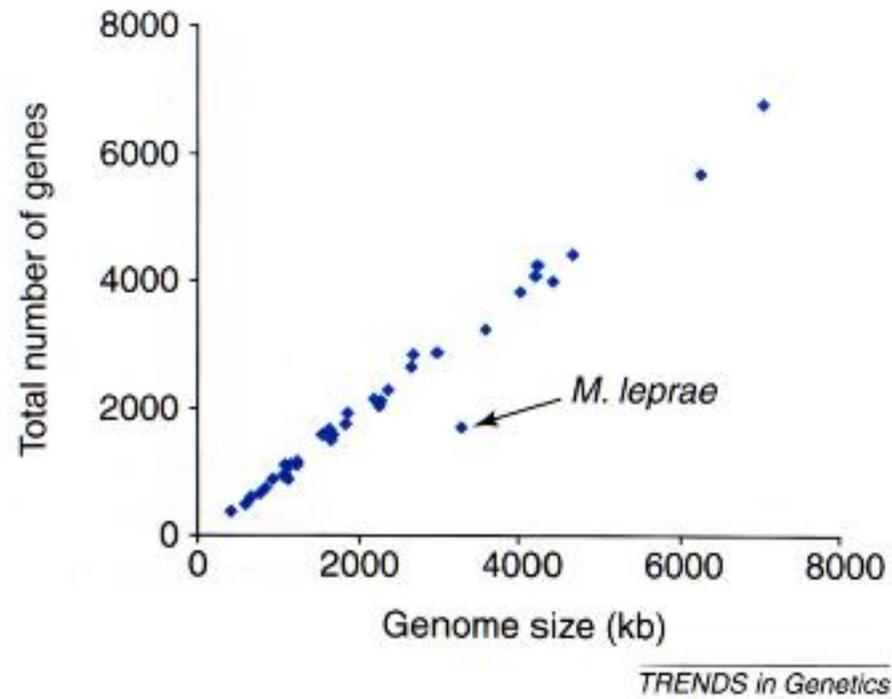


Fig. 1. Association between genome size and gene number in bacteria. Numbers include protein-coding and RNA genes ($R^2 = 0.945$). When the number of annotated pseudogenes is added to the number of functional genes, *Mycobacterium leprae* falls on the regression line. Taxa are listed in Table 1.



Table 1. Characteristics of sequenced prokaryotic genomes

| | Taxonomic group | Genome size (bp) | % occupied by genes ^a | Total gene number | Average gene length (bp) | Pseudogenes number ^b | Mean spacer length ^c (bp) | Median spacer length ^c (bp) |
|---|--|------------------|----------------------------------|-------------------|--------------------------|---------------------------------|--------------------------------------|--|
| Eubacteria | | | | | | | | |
| <i>Mycoplasma genitalium</i> | Low GC Gram - | 580 004 | 80.4 | 451 | 1054 | - | 118.4 | 4 |
| <i>Bacteroides aphidicola</i> | γ-Proteobacteria | 640 001 | 87.9 | 601 | 957 | 8 | 12.75 | 74 |
| <i>Citricoccus unaceticum</i> | Low GC Gram - | 751 719 | 92.5 | 652 | 1269 | - | 86.8 | 37 |
| <i>Mycoplasma pneumoniae</i> | Low GC Gram - | 815 384 | 85.8 | 722 | 812 | - | 120.8 | 21 |
| <i>Bacillus thuringiensis</i> | Sporobacterales | 910 724 | 94.2 | 873 | 1018 | 4 | 60.1 | 19 |
| <i>Chaetomyces fructivorus</i> | Chamysiales | 1 042 519 | 90.9 | 917 | 1013 | - | 101.2 | 53 |
| <i>Chaetomyces murinus</i> | Chamysiales | 1 069 411 | 90.9 | 949 | 1109 | - | 102.6 | 43 |
| <i>Arkhaites proteobactei</i> | α-Proteobacteria | 1 111 523 | 76.0 | 1071 | 971 | 12 | 206.2 | 123 |
| <i>Imposonema pallidum</i> | Sporobacterales | 1 138 201 | 91.1 | 1002 | 985 | - | 72.9 | 36 |
| <i>Citricoccus productionis</i> J-38 | Chamysiales | 1 228 267 | 88.5 | 1110 | 997 | - | 115.5 | 57 |
| <i>Citricoccus productionis</i> AR20 | Chamysiales | 1 228 000 | 88.2 | 1152 | 956 | - | 115.7 | 58 |
| <i>Citricoccus productionis</i> CWGL025 | Chamysiales | 1 230 230 | 88.5 | 1097 | 1000 | - | 128.2 | 76 |
| <i>Aquifex aerophilus</i> | Aquificales | 1 651 336 | 93.7 | 1574 | 930 | - | 62.5 | 10 |
| <i>Dampiera bacterium</i> | α-Proteobacteria | 1 641 451 | 94.3 | 1654 | 938 | 20 | 56.4 | 8 |
| <i>Halobacterium pyrum</i> J98 | ε-Proteobacteria | 1 648 831 | 90.1 | 1491 | 996 | 7 | 108.9 | 23 |
| <i>Halobacterium pyrum</i> Z0696 | ε-Proteobacteria | 1 657 857 | 88.4 | 1553 | 957 | 7 | 24.9 | 25 |
| <i>Halomicrobium intermedium</i> | γ-Proteobacteria | 1 830 138 | 87.0 | 1700 | 814 | - | 136.5 | 68 |
| <i>Thermocoga maritima</i> | Thermotogales | 1 850 725 | 91.0 | 1890 | 927 | - | 90.7 | 6 |
| <i>Moraxella meningitidis</i> strain A | β-Proteobacteria | 2 164 408 | 87.7 | 2121 | 853 | 56 | 178.0 | 85 |
| <i>Moraxella multocida</i> | α-Proteobacteria | 2 251 017 | 89.8 | 2016 | 958 | - | 124.7 | 60 |
| <i>Moraxella meningitidis</i> strain B | β-Proteobacteria | 2 272 301 | 79.6 | 2096 | 853 | 49 | 231.0 | 80 |
| <i>Lactococcus lactis</i> | Low GC Gram + | 2 305 089 | 84.7 | 2258 | 882 | - | 118.7 | 86 |
| <i>Xylella fastidiosa</i> | γ-Proteobacteria | 2 679 306 | 83.8 | 2822 | 707 | - | 104.4 | 75 |
| <i>Deinococcus radiodurans</i> ^d | Thermotoga ^e Deinococcus | 3 060 985 | 85.7 | 2938 | 1019 | - | 122.1 | 44 |
| Averages | | | | | | | | |
| <i>Mycobacterium leprae</i> | Actinobacteria | 3 288 303 | 75.0 | 2770 | 908 | 1081 | 548.4 | 106 |
| <i>Symbioblastus FCC603</i> | Cyanobacteria | 3 573 470 | 107.0 | 3219 | 969 | - | 143.9 | 101 |
| <i>Caulobacter crescentus</i> | α-Proteobacteria | 4 015 947 | 90.5 | 3794 | 981 | - | 201.6 | 62 |
| <i>Wolffella microscopii</i> ^f | γ-Proteobacteria | 4 033 464 | 85.4 | 3949 | 878 | - | 135.6 | 78 |
| <i>Bacillus halodurans</i> | Low GC Gram + | 4 202 253 | 94.9 | 4068 | 870 | - | 165.4 | 87 |
| <i>Bacillus subtilis</i> | Low GC Gram + | 4 214 814 | 97.6 | 4221 | 890 | - | 121.9 | 72 |
| <i>Mycobacterium tuberculosis</i> | Actinobacteria | 4 411 525 | 90.3 | 3570 | 1008 | 9 | 102.2 | 40 |
| <i>Escherichia coli</i> K12 | γ-Proteobacteria | 4 639 221 | 87.9 | 4405 | 1072 | - | 128.8 | 63 |
| <i>Halodomaines aerophilus</i> | γ-Proteobacteria | 6 284 400 | 88.5 | 5600 | 896 | - | 16.8 | 68 |
| <i>Moraxella</i> sp. | α-Proteobacteria | 7 038 071 | 86.4 | 6757 | 906 | - | 142.2 | 73 |
| Averages | | | | | | | | |
| Archaea | | | | | | | | |
| <i>Thermoplasma acidophilum</i> | Thermoplasmatales | 1 564 908 | 88.0 | 1626 | 897 | - | 134.3 | 65 |
| <i>Thermoplasma volcanium</i> | Thermoplasmatales | 1 565 800 | 85.7 | 1648 | 880 | - | 146.1 | 51 |
| <i>Methanococcus jannaschii</i> | Methanococcales | 1 890 900 | 87.8 | 1758 | 854 | - | 115.8 | 49 |
| <i>Pyrococcus horikoshii</i> | Thermococcales | 1 700 905 | 85.4 | 2113 | 814 | - | 130.8 | 13 |
| <i>Methanobacterium thermoautotrophicum</i> | Methanobacteriales | 1 761 377 | 93.8 | 1916 | 852 | - | 113.0 | 37 |
| <i>Pyrococcus abyssi</i> | Thermococcales | 1 785 118 | 101.0 | 1715 | 912 | - | 93.5 | 14 |
| <i>Halobacterium salinarum</i> | Halobacteriales | 2 014 239 | 85.0 | 2110 | 941 | - | 115.0 | 60 |
| <i>Archaeoglobus fulgidus</i> | Archaeoglobales | 2 179 400 | 91.7 | 2420 | 832 | - | 71.5 | 13 |
| Averages | | | | | | | | |
| 87.0 | | | | | | | | |
| 1983 | | | | | | | | |
| 855 | | | | | | | | |
| 112.0 | | | | | | | | |
| 27 | | | | | | | | |

^aIncludes RHA genes.

^bIncludes only non-functional or unusual pseudogenes.

^cSpacers between overlapping genes are scored as gaps.

^dChromosomes 1 and 2 are identical.

^eAlthough no chromosomally encoded pseudogenes were identified, *S. solis* contains many plasmid-borne pseudogenes.

^fSequenced genome contains no annotated pseudogenes, but at least 16 were identified in other studies.

^gSequenced genome contains no annotated pseudogenes, but at least 76 were identified in other studies.



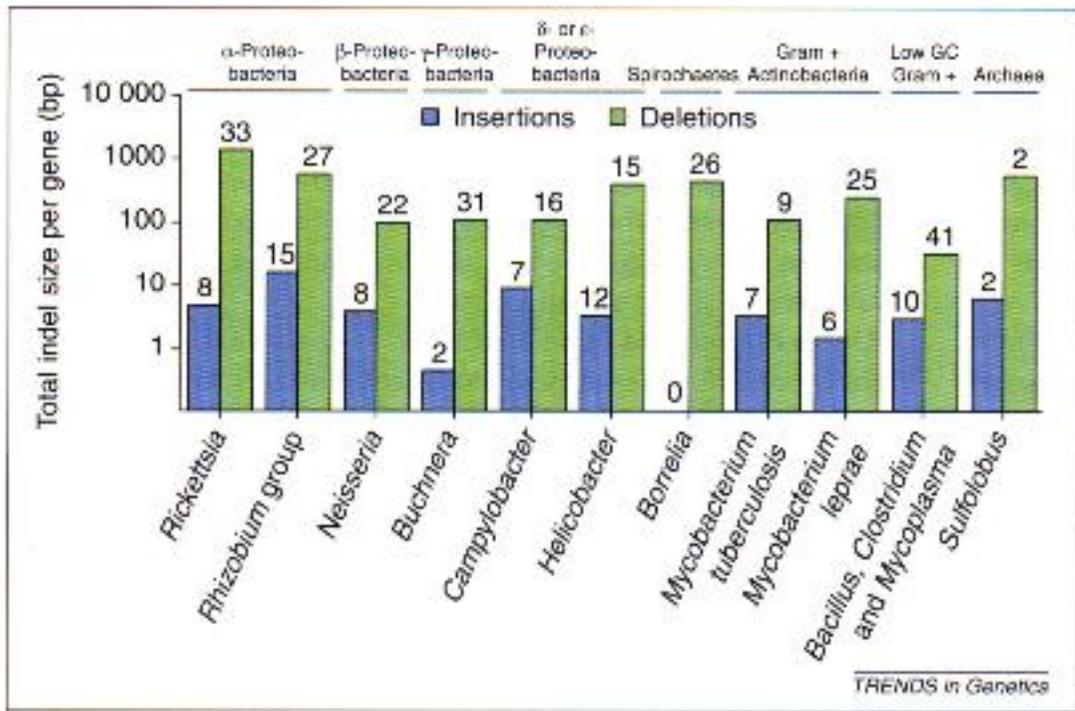


Fig. 2. Frequency of deletions and insertions in bacterial genomes. Frequencies based on comparative analyses of pseudogenes with their functional counterparts from a closely related species, generally from the same genus, and with at least one functional gene in a closely related outside reference species. Bars represent the average total size of deletions and insertions per pseudogene. Numbers at tops of bars represent the numbers of each type of event. Analyzed pseudogenes were: *hmbR*, *opaA*, *opaB*, *yedI*^{*}, *thuD*, *porA*, *pilC2*, *opcB* (*Neisseria meningitidis*, *Neisseria gonorrhoeae*); *kdpA*, *kdpC*, *ast*, *glpT*^{**}, *cj0565*, *cj794*^{**} (*Campylobacter jejuni*); *thuD*, cytochrome P-450, IS1380, *egl*, *groEL* (*Rhizobium* sp.); *vacA*, *olpA*, *iceA*, *rfaJ*, OMP29, HP1589 (*Helicobacter pylori*); *ahcY* (*Sulfolobus solfataricus*); *groES2* (*Rhodobacter sphaeroides*); *msp1b1pg* (*Anaplasma marginale*); ORF3, *aatA* (*Agrobacterium rhizogenes*); *epsD* (*Azospirillum brasilense*); *oxyR*, Rv1503c-Rv1504c, *hypB*, Rv3349c^{**} (*Mycobacterium tuberculosis*); *bfrB*, *csp*, *ackA-pta*, *fadE7*, *cysM* (*Mycobacterium leprae*); *vmp*, *vlp*, BBQ20^{**}, BBQ71^{**}, BBQ55 (*Borrelia* sp.); *hmw2*, *ippA* (*Mycoplasma* sp.); *treP*, *hblB*, transposase, *s14*, *thuC* (*Bacillus* sp.); *cpe*, *p-21* (*Clostridium* sp.); *recombinase*, *sat4* (*Staphylococcus* sp.); *hisG*, *hisC* (*Lactococcus lactis*); *ace* (*Enterococcus faecalis*). Genes marked with an asterisk are those in which a functional equivalent in the same bacterial group could not be found; two asterisks indicate that no suitable homolog could be found in an outside reference species.

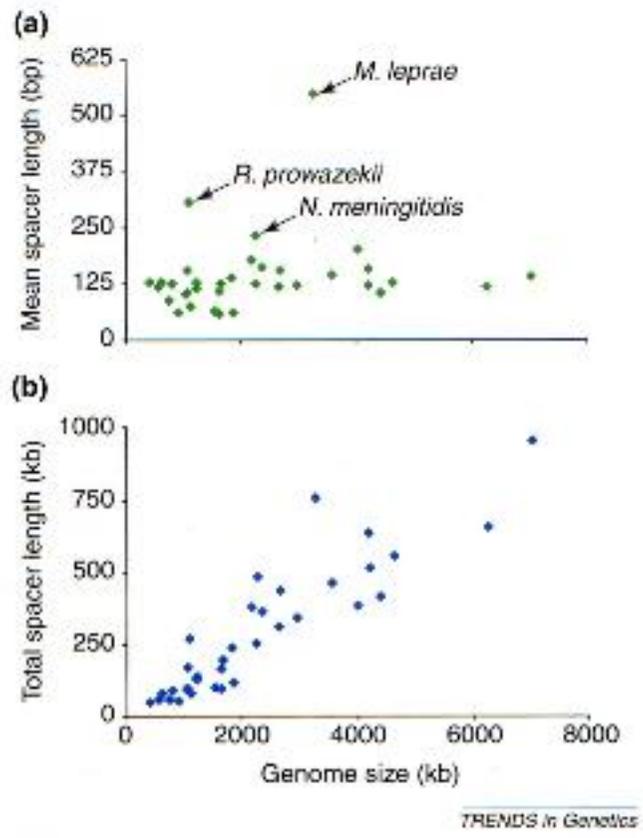


Fig. 3. The relationship between genome size and noncoding DNA in eubacteria. (a) Relationship between genome size and mean spacer length. The three species with exceptionally long spacers are bacteria with high numbers of pseudogenes: *Mycobacterium leprae*, *Rickettsia prowazekii* and *Neisseria meningitidis*. (b) Relationship between genome size and summed spacer length in eubacteria ($R^2 = 0.762$). Taxa are listed in Table 1.

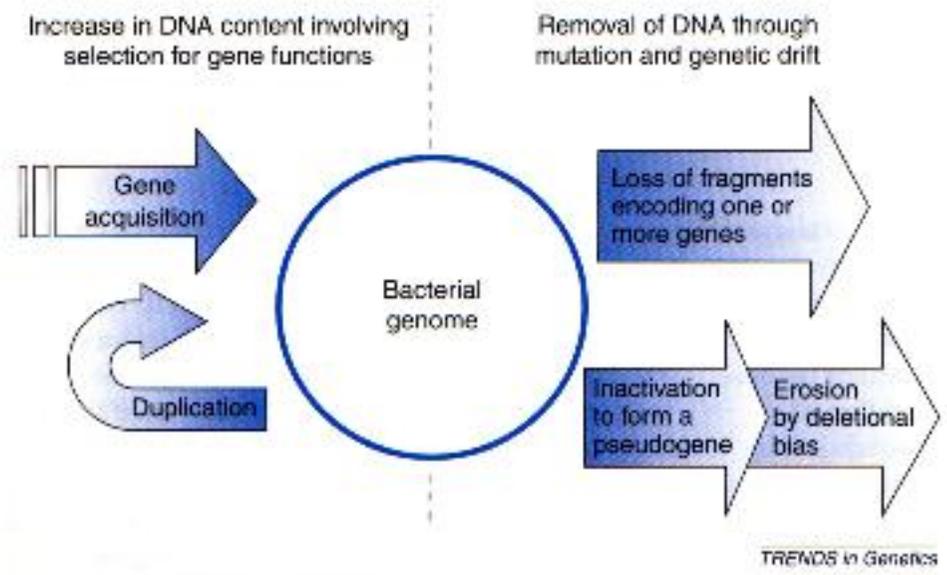


Fig. 4. Processes involved in the evolution of genome size in bacteria. New sequences are acquired by DNA transfer and gene duplication, the former being the predominant mode of DNA increase within most species. DNA loss can be produced by large deletions eliminating one or more genes in a single event, or by loss of function followed by subsequent deletions of the resulting pseudogenes.