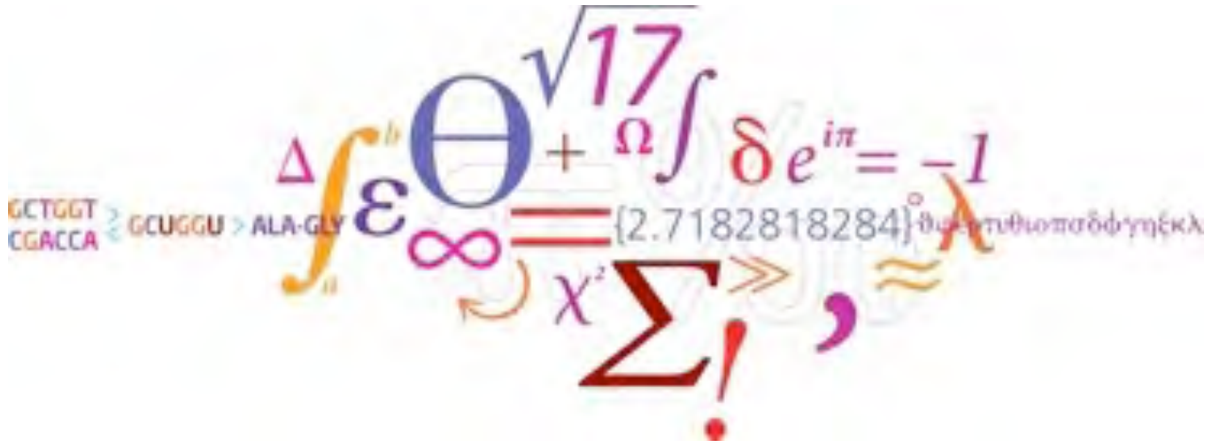
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

DTU

Pan- and core-Genomics



Dave Ussery

Genomics of Prokaryotes Workshop
Universidad Miguel Hernández
Alicante, Spain

Wednesday, 14 December, 2011

Center for Biological Sequence Analysis
Department of Systems Biology

Wednesday, 14 December, 2011

http://egg.umh.es/course.html

Course outline

	Monday, 12 December	Tuesday, 13 December	Wednesday, 14 December	Thursday, 15 December	Friday, 16 December
9:30-10:30	Introduction to Genomics *E. coli* O104 as an example	3 Generations of Genome Sequencing	Pan- and Core-genomics	Introduction to Metagenomics	Tools to compare bacterial genomes
10:30-10:45 (short Break)					
10:45-11:30	Sequences as biological information	BLAST matrices and atlases	*Veilonella*: halfway between Gram positive & Gram neg.	Transcriptomics and metranscriptomics	***Vibrio***
11:30-12:00 (Coffee Break)					
12:00-13:00	**Computer exercises:** Introduction to UNIX	**Computer exercises:** Extract DNA, find rRNAs	**Computer exercises:** Gene-finding	**Computer exercises:** BLAST Matrix	***Bifidobacterium Pseudomonas***
13:00-15:00 (Lunch break)					
15:00-16:30	**Computer exercises:** GenomeAtlas construction	**Computer exercises:** Construction of 16S rRNA trees	**Computer exercises:** Amino acid and Codon Usage	**Computer exercises:** Pan- core-genome plots	***E. coli, Bacillus, Yersinia***
16:30-17:00 (Coffee break)					
17:00-18:30	**Computer exercises:** GenomeAtlas construction	**Computer exercises:** Construction of 16S rRNA trees	**Computer exercises:** Amino acid and Codon Usage	**Computer exercises:** Pan- core-genome plots	certificate presentation, closing

http://www.cbs.dtu.dk/staff/dave/genomics_course/Alicante2011.php

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 10268–10273, September 1996
Evolution

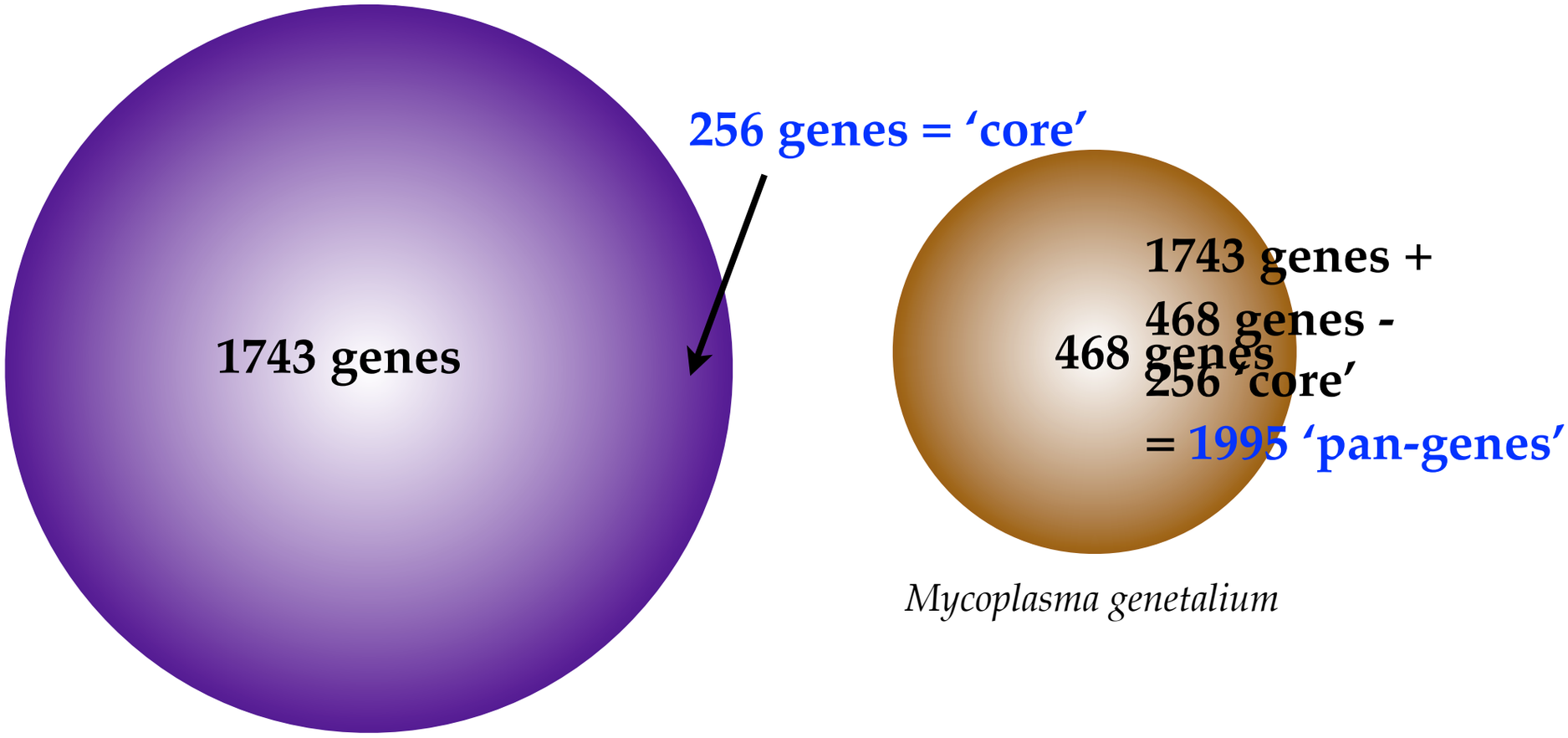
1. What is a core-genome? What is a pan-genome?

A minimal gene set for cellular life derived by comparison of complete bacterial genomes

ARCADY R. MUSHEGIAN AND EUGENE V. KOONIN*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Communicated by Clyde Hutchinson, University of North Carolina, Chapel Hill, NC, May 17, 1996 (received for review March 11, 1996)



Haemophilus influenza

Mycoplasma genetalium

1. What is a core-genome?

JOURNAL OF BACTERIOLOGY, July 2005, p. 4627–4636 Vol. 187
0021-9193/05/$08.00+0 doi:10.1128/JB.187.13.4627–4636.2005

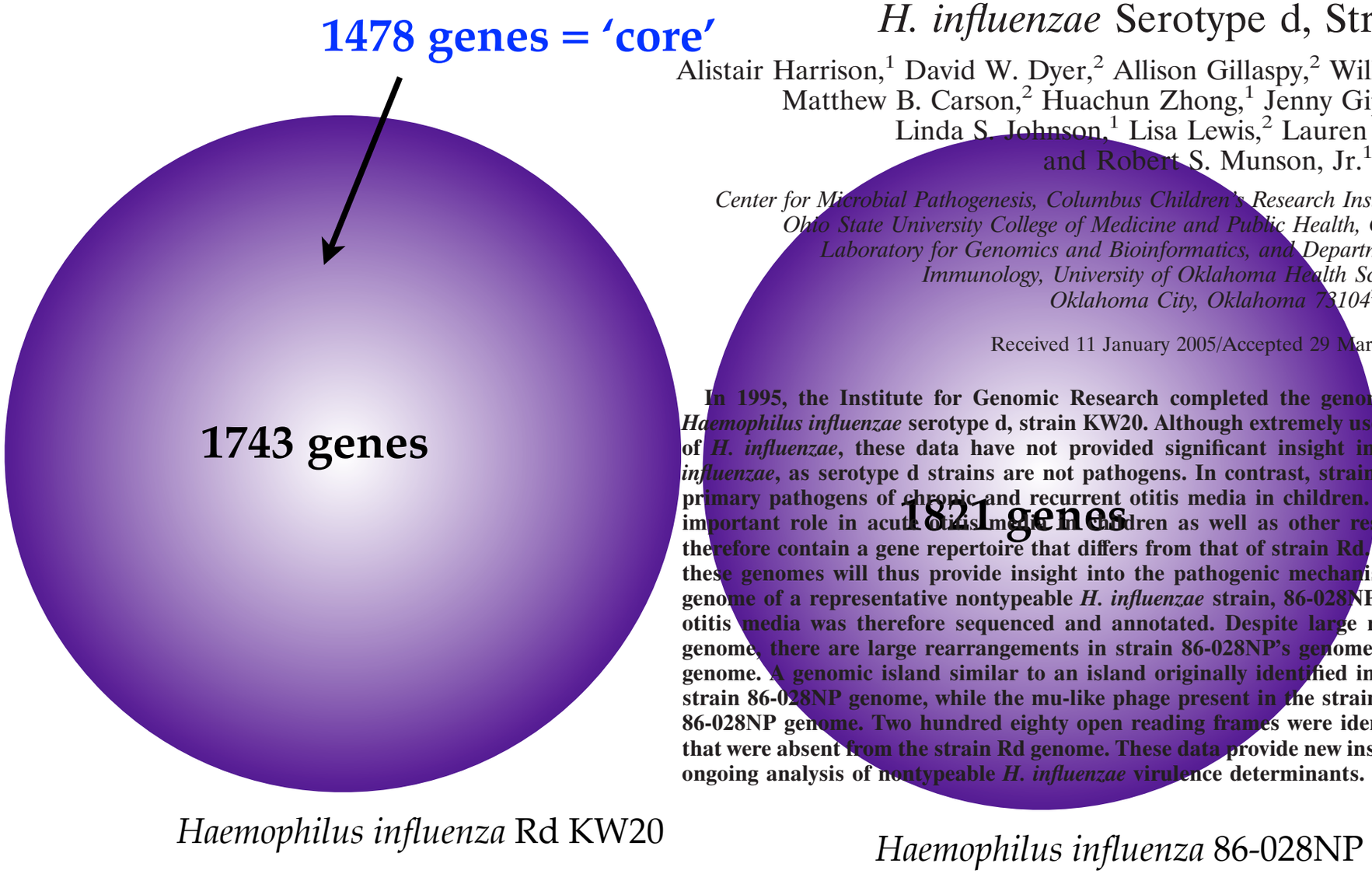
Genomic Sequence of an Otitis Media Isolate of Nontypeable *Haemophilus influenzae*: Comparative Study with *H. influenzae* Serotype d, Strain KW20

Alistair Harrison,[1] David W. Dyer,[2] Allison Gillaspy,[2] William C. Ray,[1] Rachna Mungur,[1] Matthew B. Carson,[2] Huachun Zhong,[1] Jenny Gipson,[2] Mandy Gipson,[2] Linda S. Johnson,[1] Lisa Lewis,[2] Lauren O. Bakaletz,[1] and Robert S. Munson, Jr.[1]*

Center for Microbial Pathogenesis, Columbus Children's Research Institute, and Department of Pediatrics, Ohio State University College of Medicine and Public Health, Columbus, Ohio 43205,[1] and Laboratory for Genomics and Bioinformatics, and Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73104[2]

Received 11 January 2005/Accepted 29 March 2005

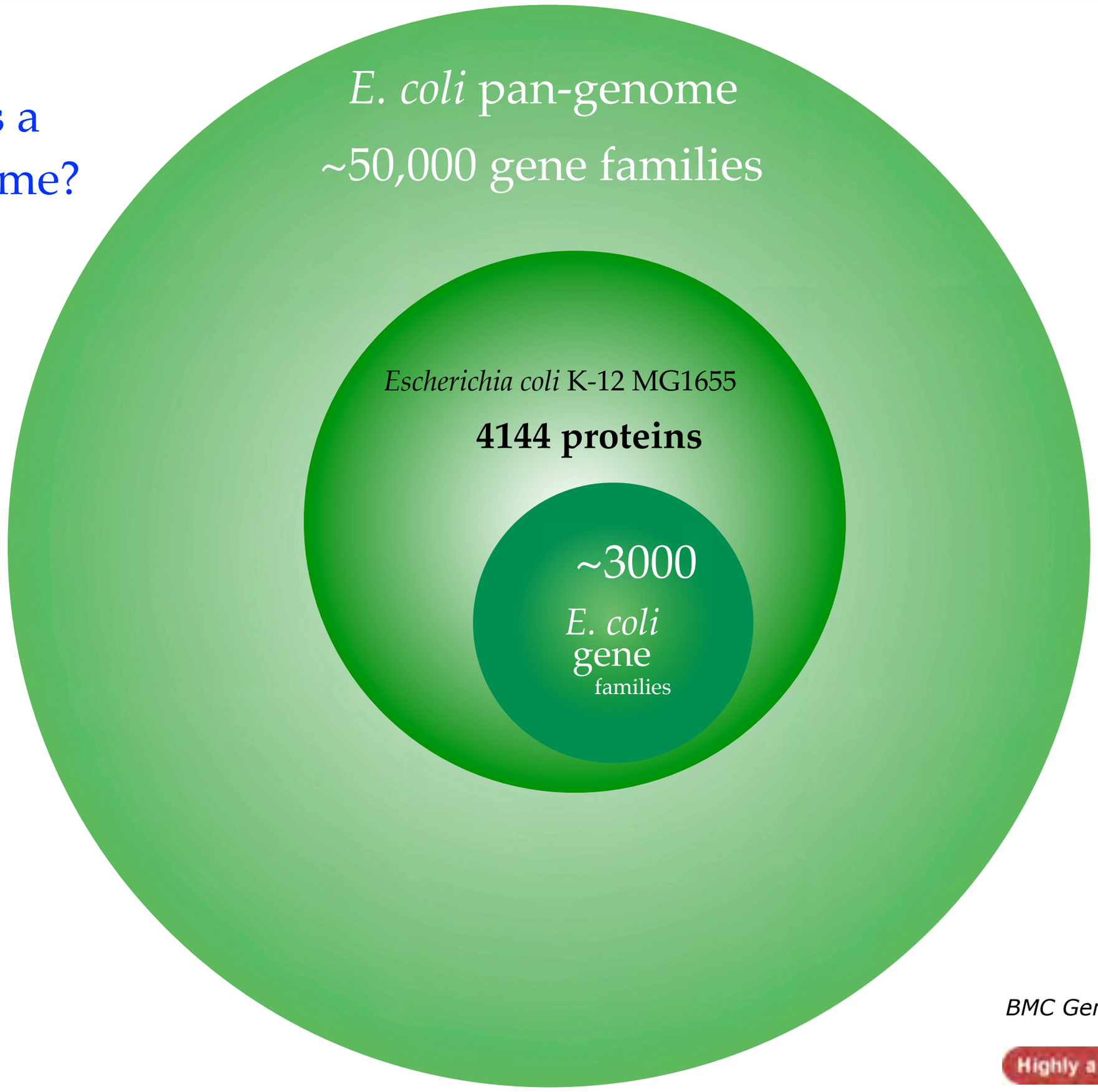
In 1995, the Institute for Genomic Research completed the genome sequence of a rough derivative of *Haemophilus influenzae* serotype d, strain KW20. Although extremely useful in understanding the basic biology of *H. influenzae*, these data have not provided significant insight into disease caused by nontypeable *H. influenzae*, as serotype d strains are not pathogens. In contrast, strains of nontypeable *H. influenzae* are the primary pathogens of chronic and recurrent otitis media in children. In addition, these organisms have an important role in acute otitis media in children as well as other respiratory diseases. Such strains must therefore contain a gene repertoire that differs from that of strain Rd. Elucidation of the differences between these genomes will thus provide insight into the pathogenic mechanisms of nontypeable *H. influenzae*. The genome of a representative nontypeable *H. influenzae* strain, 86-028NP, isolated from a patient with chronic otitis media was therefore sequenced and annotated. Despite large regions of synteny with the strain Rd genome, there are large rearrangements in strain 86-028NP's genome architecture relative to the strain Rd genome. A genomic island similar to an island originally identified in *H. influenzae* type b is present in the strain 86-028NP genome, while the mu-like phage present in the strain Rd genome is absent from the strain 86-028NP genome. Two hundred eighty open reading frames were identified in the strain 86-028NP genome that were absent from the strain Rd genome. These data provide new insight that complements and extends the ongoing analysis of nontypeable *H. influenzae* virulence determinants.



Haemophilus influenza Rd KW20

Haemophilus influenza 86-028NP

1. What is a pan-genome?



BMC Genomics 2009, **10:**385

Highly accessed Open Access

nature International weekly journal of science

Vol 464|4 March 2010|doi:10.1038/nature08821 nature

ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

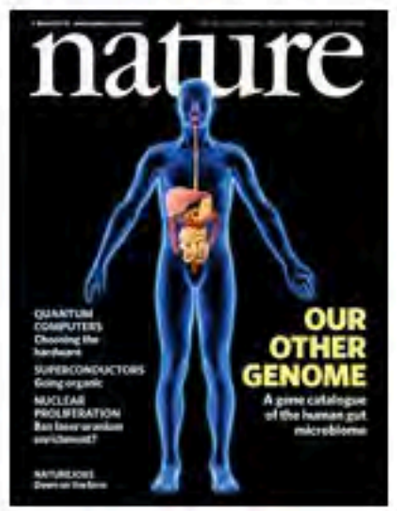
To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, tenfold the number of human cells, and suggested that they encode 100-fold more unique genes than our own genome[1]. The majority of microbes reside in the gut, have a profound influence on human physiology and nutrition, and are crucial for human life[2,3]. Furthermore, the gut microbes contribute to energy harvest from food, and changes of gut microbiome may be associated with bowel diseases or obesity[4–8].

individuals from the United States or Japan[8,16,17]. To get a broader overview of the human gut microbial genes we used the Illumina Genome Analyser (GA) technology to carry out deep sequencing of total DNA from faecal samples of 124 European adults. We generated 576.7 Gb of sequence, almost 200 times more than in all previous studies, assembled it into contigs and predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a

1. What is a meta-genome?

About the cover



The international MetaHIT (Metagenomics of the Human Intestinal Tract) project has published a gene catalogue of the human gut microbiome derived from 124 healthy, overweight and obese human adults, as well as inflammatory disease patients, from Denmark and Spain. The data provide the first insights into this gene set - over 150 times larger than the human gene complement - and permit the definition of both a minimal gut metagenome and a minimal gut bacterial genome. Credit: Roger Harris /Science Photo Library.

2. How is the core- and pan-genome calculated?

Proc Natl Acad Sci USA, **102**:13950-13955 (2005).

Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome"

Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[a], Jonathan Crabtree[a], Amanda L. Jones[g], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a], Marirosa Mora[c], Maria Scarselli[c], Immaculada Margarit y Ros[c], Jeremy D. Peterson[a], Christopher R. Hauser[a], Jaideep P. Sundaram[a], William C. Nelson[a], Ramana Madupu[a], Lauren M. Brinkac[a], Robert J. Dodson[a], Mary J. Rosovitz[a], Steven A. Sullivan[a], Sean C. Daugherty[a], Daniel H. Haft[a], Jeremy Selengut[a], Michelle L. Gwinn[a], Liwei Zhou[a], Nikhat Zafar[a], Hoda Khouri[a], Diana Radune[a], George Dimitrov[a], Kisha Watkins[a], Kevin J. B. O'Connor[h], Shannon Smith[i], Teresa R. Utterback[i], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[e,j], Dennis L. Kasper[e,j], John L. Telford[c], Michael R. Wessels[d,e], Rino Rappuoli[c,k,l], and Claire M. Fraser[a,b,k,m]

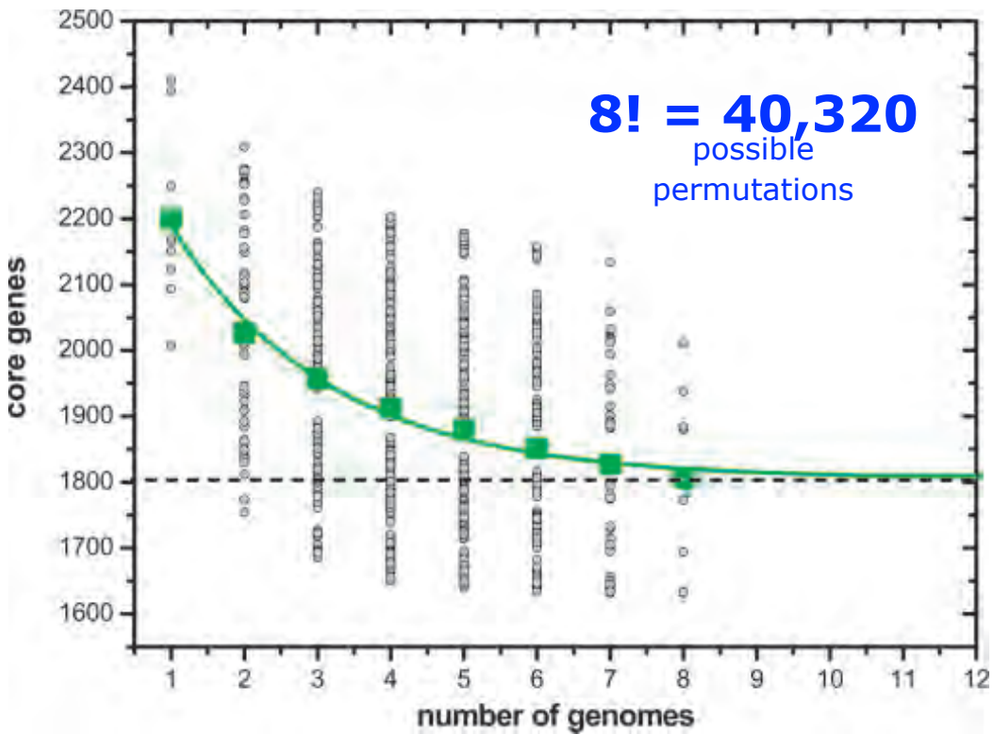


Fig. 2. GBS core genome. The number of shared genes is plotted as a function of the number *n* of strains sequentially added (see *Materials and Methods*). For each *n*, circles are the $8!/[(n-1)!\cdot(8-n)!]$ values obtained for the different strain combinations. Squares are the averages of such values. The continuous curve represents the least-squares fit of the function $F_c = \kappa_c \exp[-n/\tau_c] + \Omega$ (see Eq. **1** in *Supporting Text*) to data. The best fit was obtained with correlation $r^2 = 0.990$ for $\kappa_c = 610 \pm 38$, $\tau_c = 2.16 \pm 0.28$, and $\Omega = 1{,}806 \pm 16$. The extrapolated GBS core genome size Ω is shown as a dashed

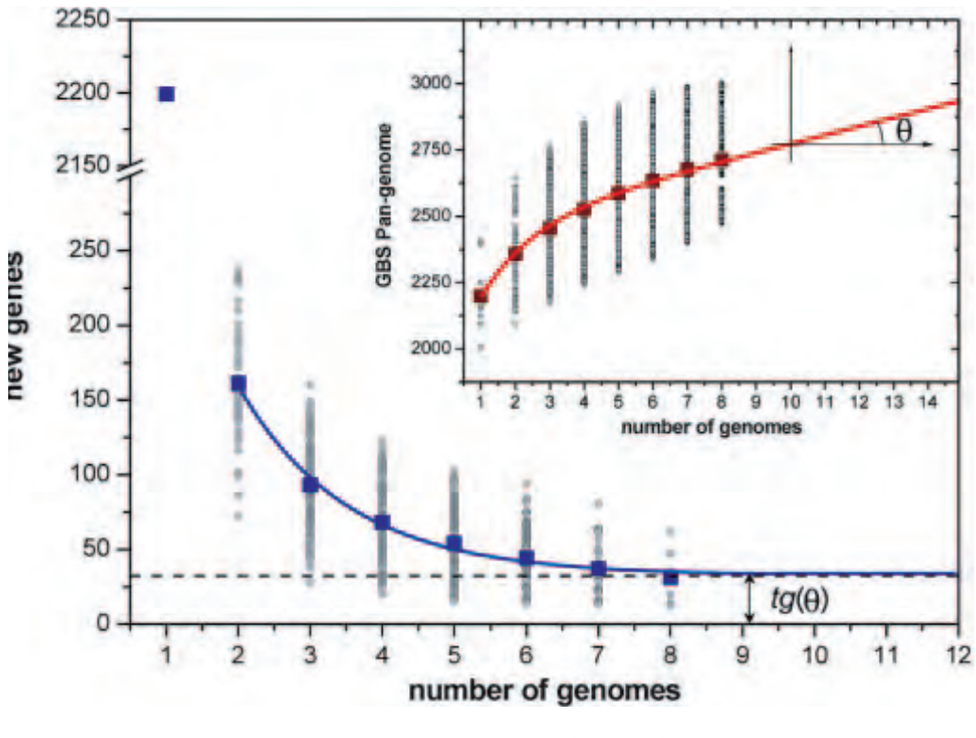
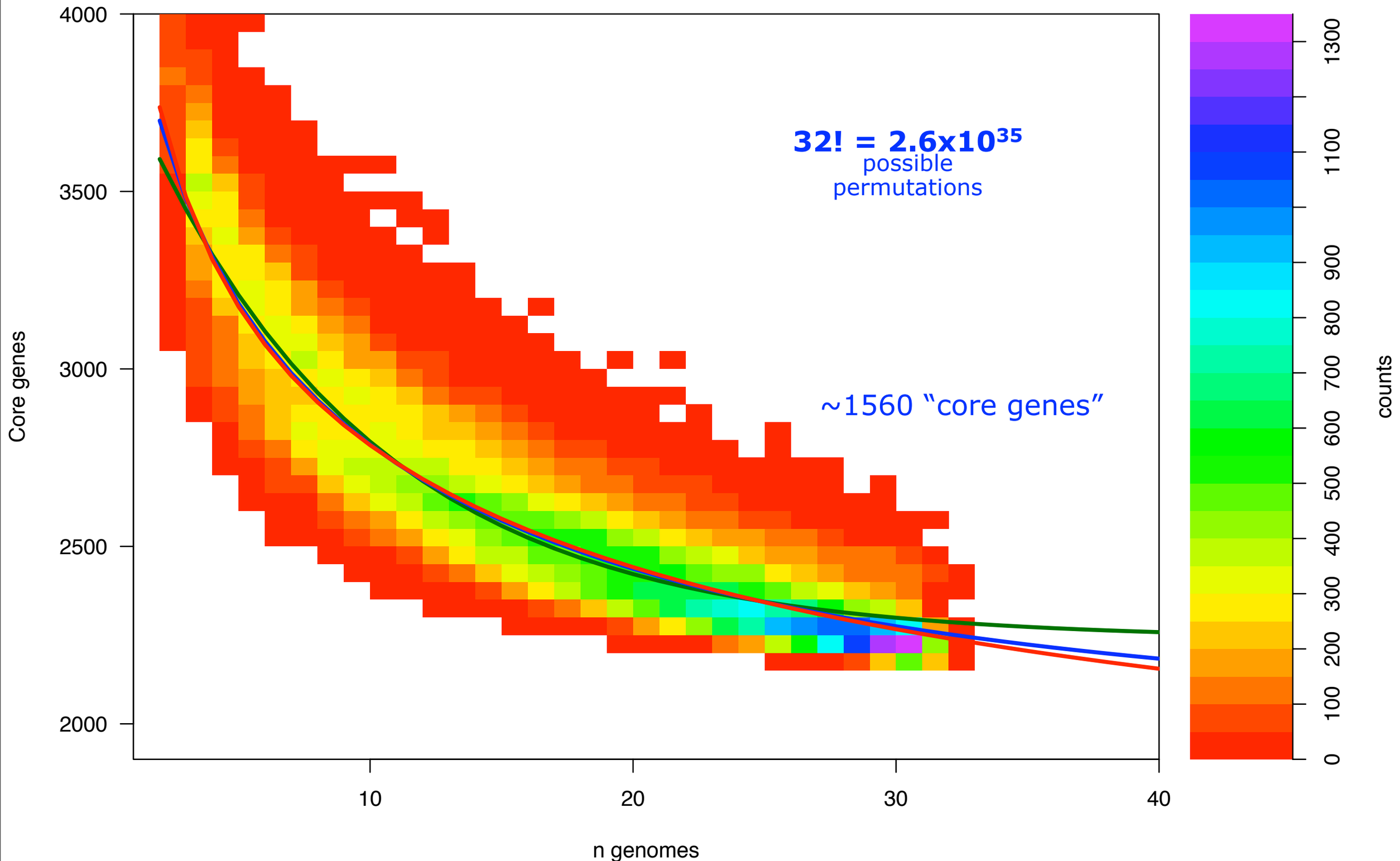


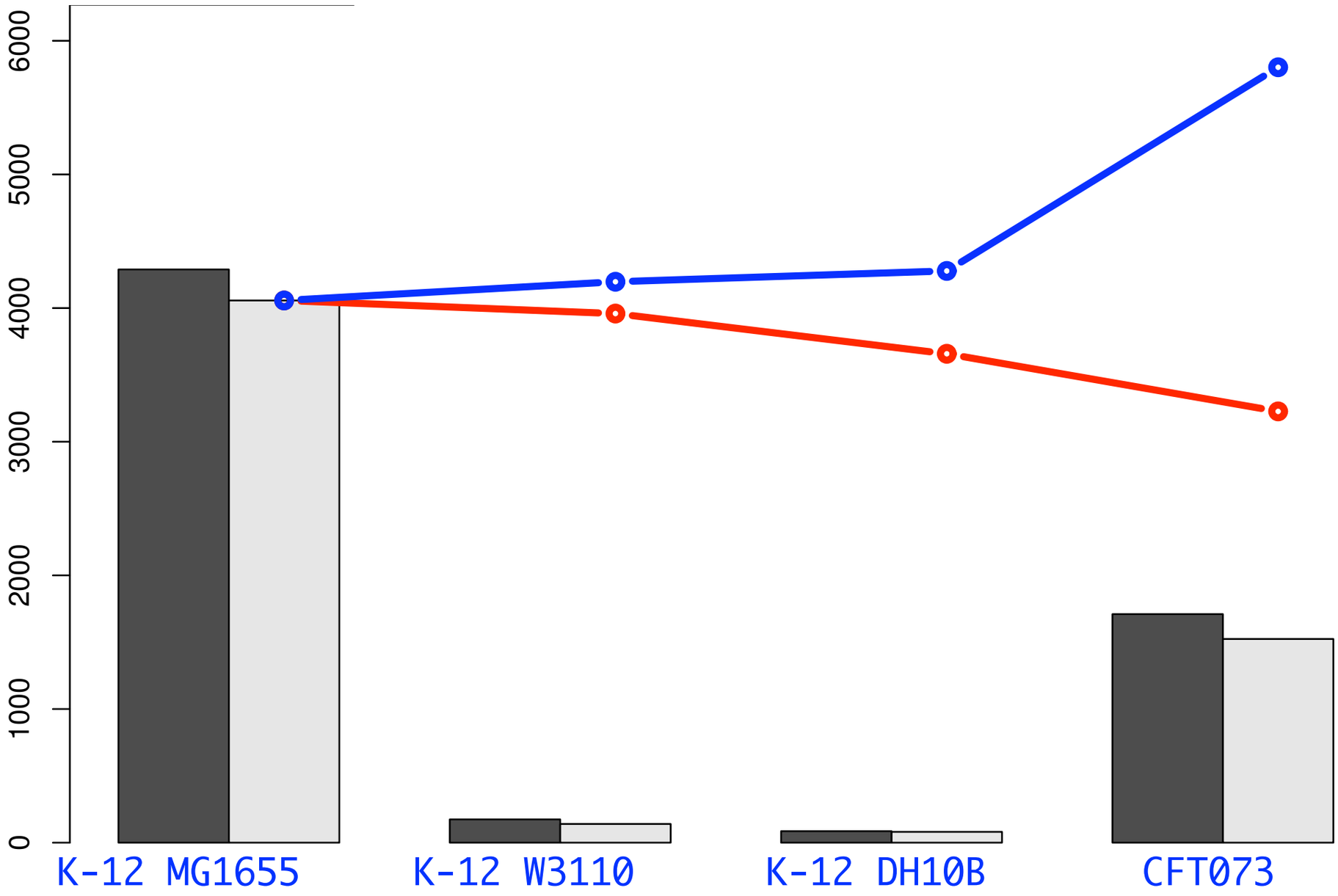
Fig. 3. GBS pan-genome. The number of specific genes is plotted as a function of the number *n* of strains sequentially added (see *Materials and Methods*). For each *n*, circles are the $8!/[(n-1)!\cdot(8-n)!]$ values obtained for the different strain combinations; squares are the averages of such values. The blue curve is the least-squares fit of the function $F_s(n) = \kappa_s \exp[-n/\tau_s] + tg(\theta)$ (see Eq. **2** in *Supporting Text*) to the data. The best fit was obtained with correlation $r^2 = 0.995$ for $\kappa_s = 476 \pm 62$, $\tau_s = 1.51 \pm 0.15$, and $tg(\theta) = 33 \pm 3.5$. The extrapolated average number $tg(\theta)$ of strain-specific genes is shown as a

E. coli "core genes" in 32 genomes



Genome Biology, 2007, 8:R267doi:10.1186/gb-2007-8-12-r267

		#genes	# new genes	# new gene families	pan-	core-
1	E. coli K-12 MG1655	4289	4289	4057	4057	4057
2	E. coli K-12 W3110	4387	174	140	4197	3959
3	E. coli K-12 DH10B	4126	86	81	4278	3658
4	E. coli CFT073	5379	1710	1524	5802	3227



New genes
New gene families
Core genome
Pan genome

number of gene families

15000
10000
5000
0

1 : E. coli K12MG1
2 : E. coli O157RIM
3 : E. coli O157EDI
4 : S. flexneri 2a–3
5 : E. coli CFT073
6 : S. flexneri 2a–2
7 : S. sonnei Ss046
8 : S. boydii Sb227
9 : S. dysenteriae S
10 : E. coli W3110
11 : E. coli UTI89
12 : E. coli 536
13 : S. flexneri 5–8
14 : E. coli APEC–
15 : E. coli 101–1
16 : E. coli 53638
17 : E. coli B171
18 : E. coli B7A
19 : E. coli E22
20 : E. coli E24377
21 : E. coli F11
22 : E. coli HS
23 : E. coli B
24 : E. coli E11001
25 : S. boydii BS51
26 : E. coli 042
27 : E. coli B03
28 : E. coli E2348
29 : E. coli H10407
30 : E. coli RS218

K12

new *E. coli* genomes

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

D.W. Ussery et al., *Computing for Comparative Microbial Genomics*, Computational Biology 8, DOI 10.1007/978-1-84800-255-5_14, © Springer-Verlag London Limited 2009

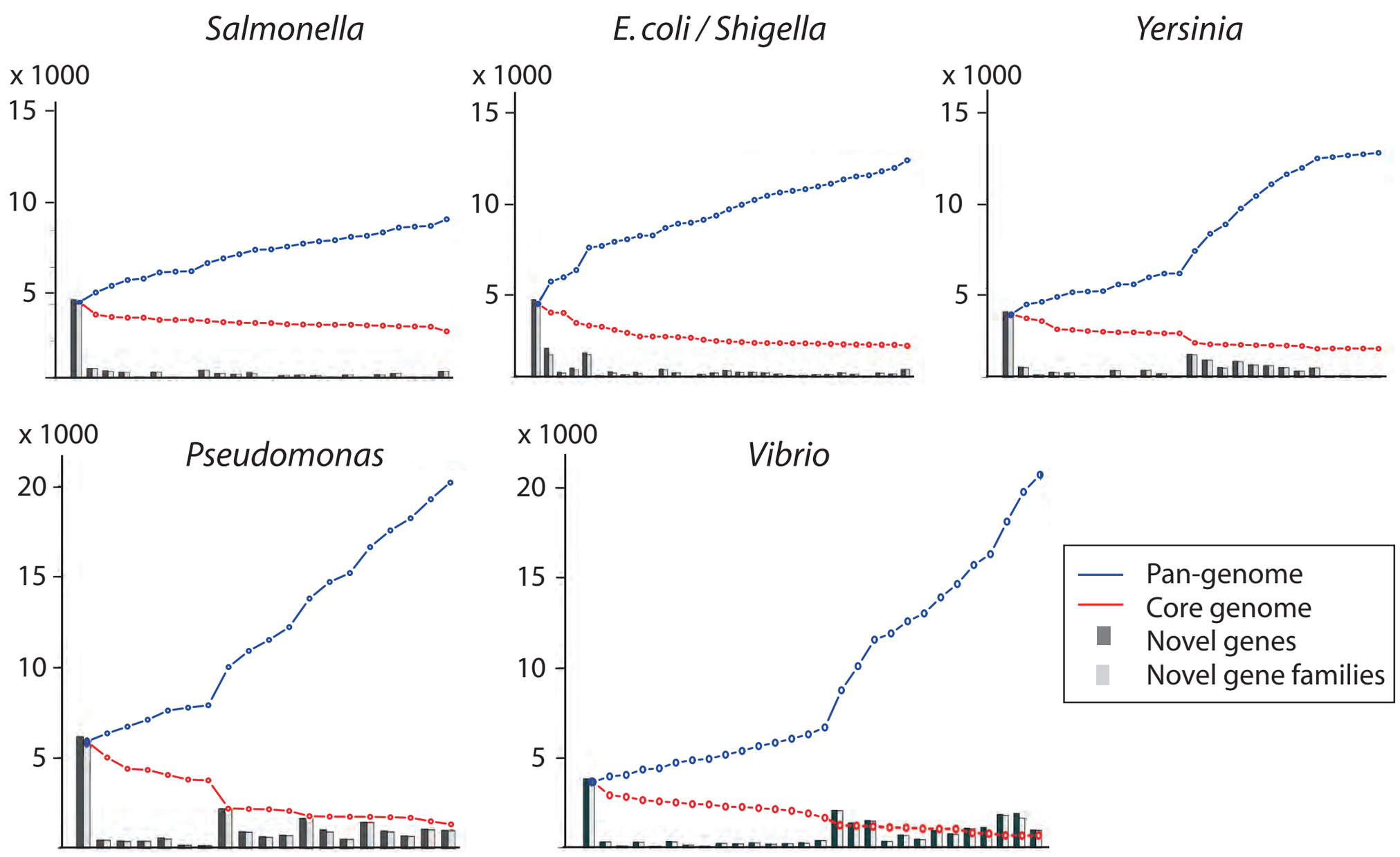
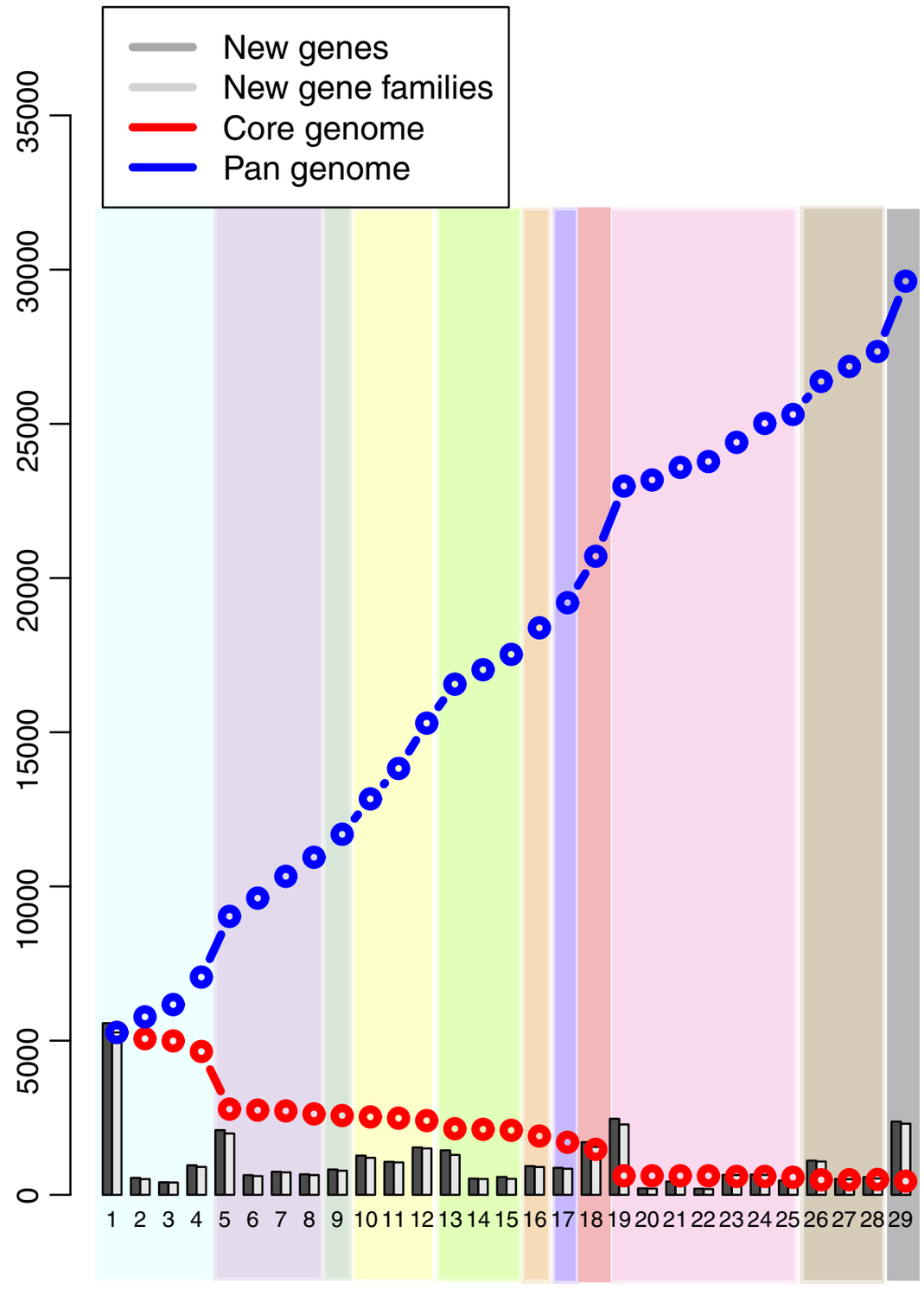


Fig. 12.5 The pan-genome and core genome for five different Proteobacterial genera. The *Salmonella* graph represents one species (*Salmonella enterica*), whereas the *E. coli/Shigella* figure contains both *E. coli* and four different *Shigella* species. The other graphs represent multiple species per genus. All graphs are drawn on the same scale



1 : Pseudomonas aeruginosa PAO1
2 : Pseudomonas aeruginosa LESB58
3 : Pseudomonas aeruginosa UCBPP–PA14
4 : Pseudomonas aeruginosa PA7
5 : Pseudomonas putida KT2440
6 : Pseudomonas putida F1
7 : Pseudomonas putida GB–1
8 : Pseudomonas putida W619
9 : Pseudomonas entomophila L48
10 : Pseudomonas fluorescens Pf0–1
11 : Pseudomonas fluorescens Pf–5
12 : Pseudomonas fluorescens SBW25
13 : Pseudomonas syringae pv. tomato str. DC3000
14 : Pseudomonas syringae pv. syringae B728a
15 : Pseudomonas syringae pv. phaseolicola 1448A
16 : Pseudomonas mendocina ymp
17 : Pseudomonas stutzeri A1501
18 : Azotobacter vinelandii DJ
19 : Acinetobacter baumannii AB0057
20 : Acinetobacter baumannii AB307–0294
21 : Acinetobacter baumannii ACICU
22 : Acinetobacter baumannii AYE
23 : Acinetobacter baumannii ATCC 17978
24 : Acinetobacter sp. ADP1
25 : Acinetobacter baumannii SDF
26 : Psychrobacter arcticus 273–4
27 : Psychrobacter cryohalolentis K5
28 : Psychrobacter sp. PRwf–1
29 : Cellvibrio japonicus Ueda107

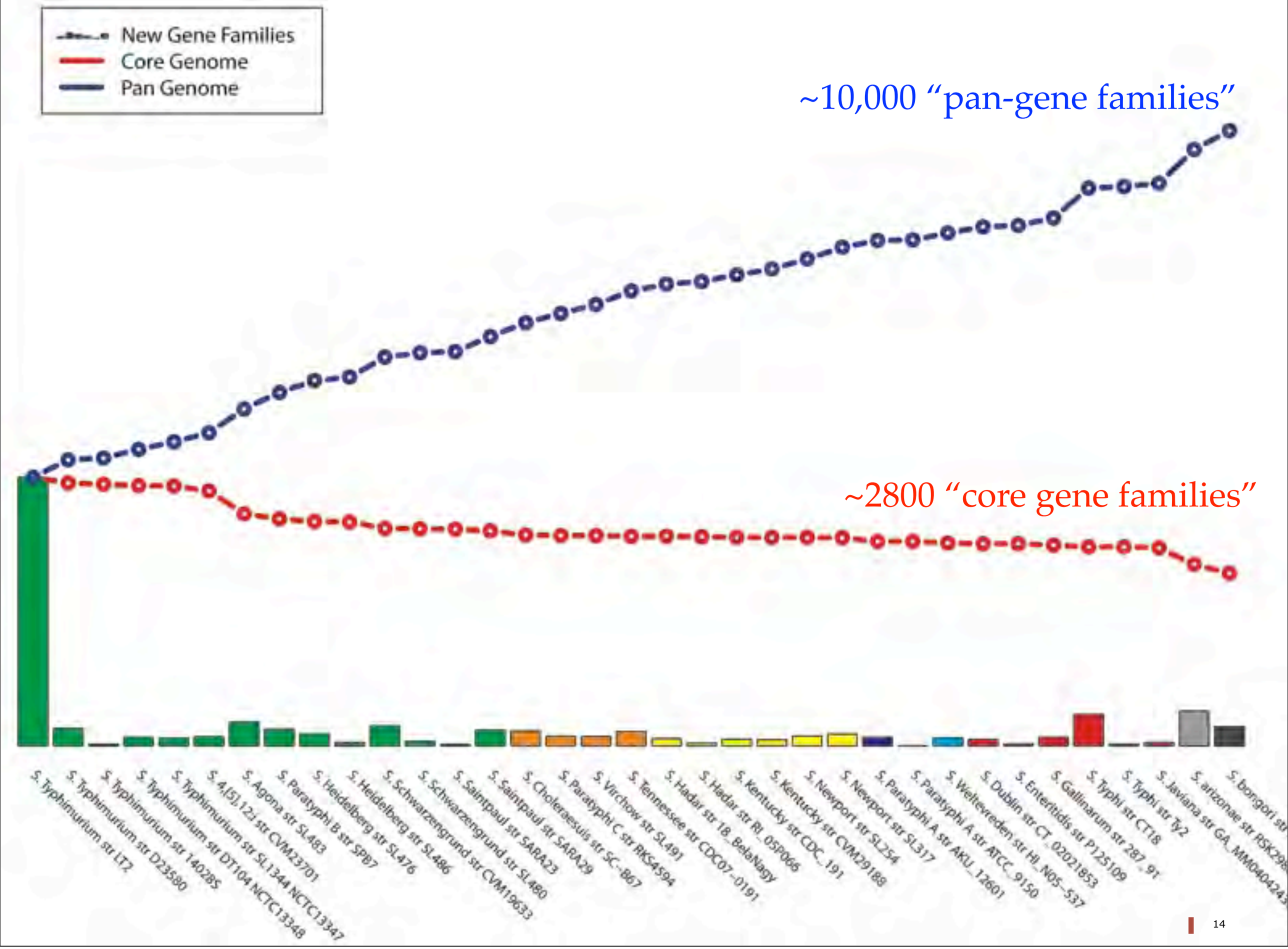
Microb Ecol
DOI 10.1007/s00248-011-9914-8

MINIREVIEWS

Defining the *Pseudomonas* Genus: Where Do We Draw the Line with *Azotobacter*?

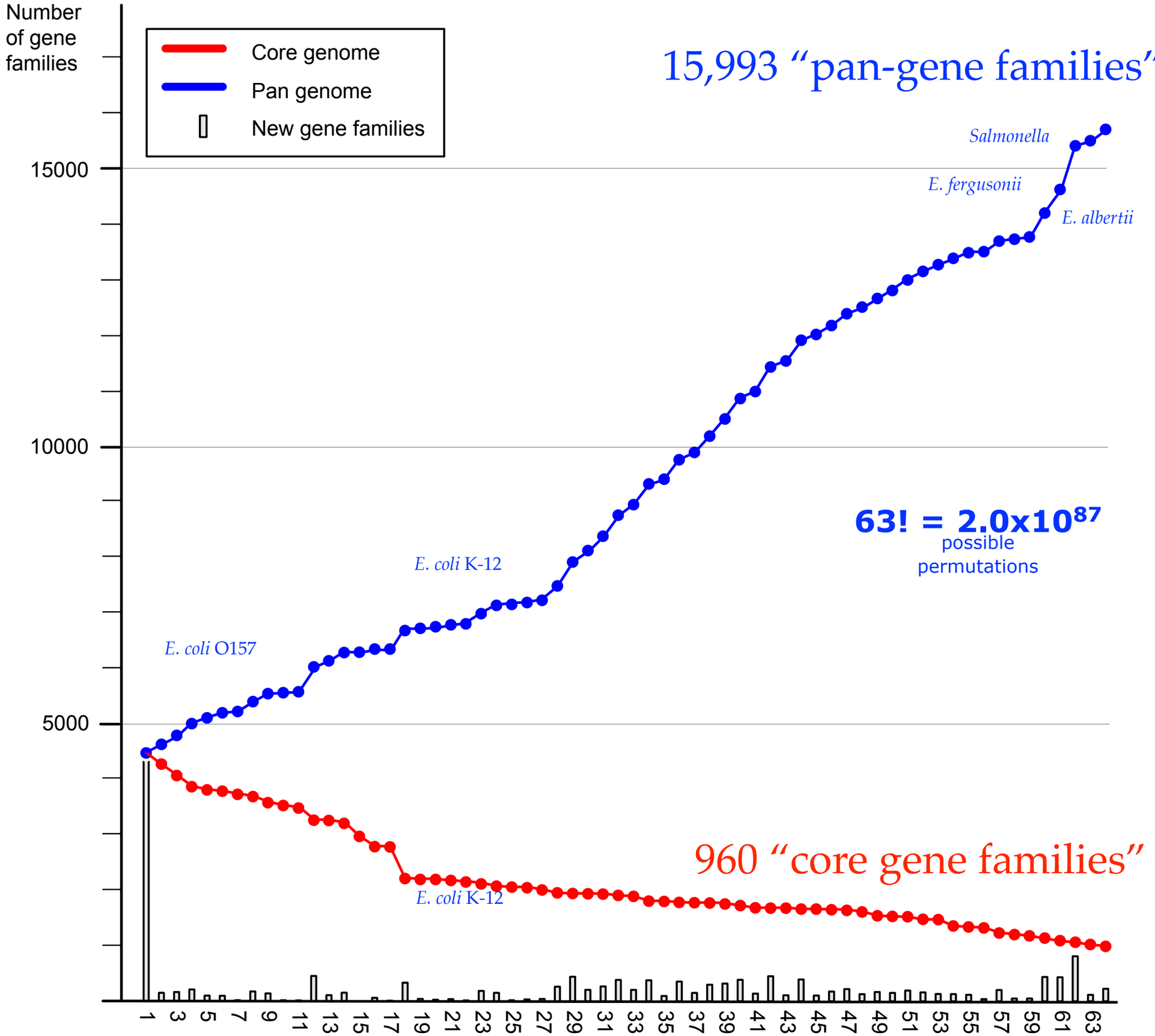
Asli I. Özen · David W. Ussery

Figure 3 Core and pan genome plot of the *Pseudomonas* genomes. *Azotobacter* is *highlighted*. The number of pan genome for all strains used is 29,696 (*dashed line on top*) and the core genome is reduced to 443 (*dashed line below*). The *light purple color* indicates where *P. putida* is added, and the *light red column* indicates where *A. vinelandii* is added


New Gene Families
Core Genome
Pan Genome
~10,000 “pan-gene families”
~2800 “core gene families”
S. Typhimurium str LT2
S. Typhimurium str D23580
S. Typhimurium str 14028S
S. Typhimurium str DT104 NCTC13348
S. 4,[5],12:i- str SL1344 NCTC13347
S. Agona str SL483
S. Paratyphi B str SPB7
S. Heidelberg str SL476
S. Schwarzengrund str SL486
S. Schwarzengrund str CVM19633
S. Saintpaul str SARA23
S. Choleraesuis str SARA29
S. Paratyphi C str RKS4594
S. Virchow str SL491
S. Tennessee str CDC07-0191
S. Hadar str 18_BeloNagy
S. Hadar str RI_05P066
S. Kentucky str CDC_191
S. Newport str CVM29188
S. Newport str SL254
S. Paratyphi A str SL317
S. Paratyphi A str AKU_12601
S. Weltevreden str HI_N05-537
S. Dublin str CT_02021853
S. Enteritidis str P125109
S. Gallinarum str 287_91
S. Typhi str CT18
S. Typhi str Ty2
S. Javiana str GA_MM0404243
S. arizonae str RSK2980
S. bongori str

64 *E. coli* genomes

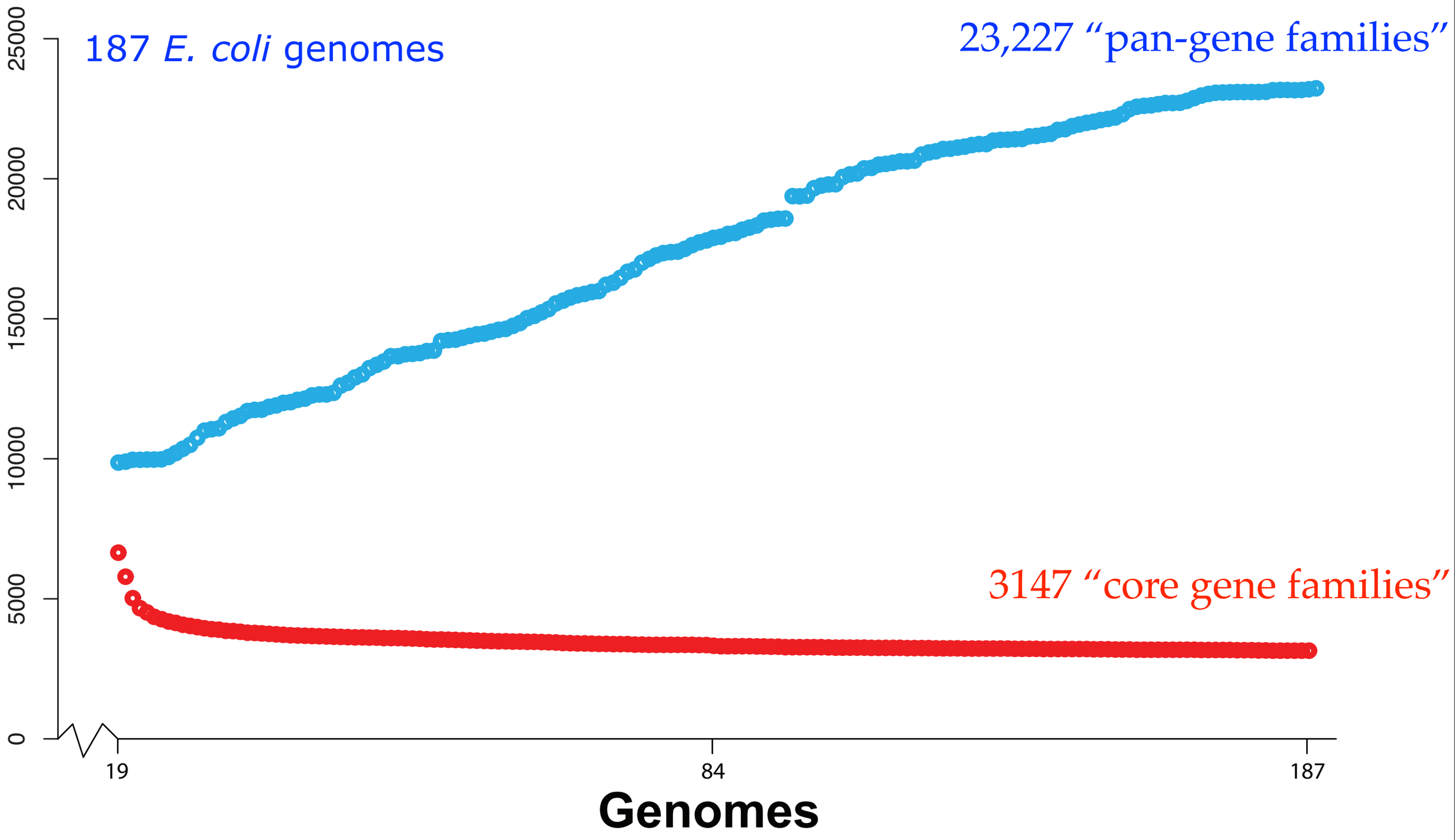
Microbial Ecology, **60**:708-720, (2010).



1 : Escherichia coli 0157:H7 str. EC4196
2 : Escherichia coli 0157:H7 str. EC4113
3 : Escherichia coli 0157:H7 str. EC508
4 : Escherichia coli 0157:H7 str. EC4501
5 : Escherichia coli 0157:H7 str. EC4076
6 : Escherichia coli 0157:H7 str. EC4115
7 : Escherichia coli 0157:H7 str. EC4042
8 : Escherichia coli 0157:H7 str. EC4486
9 : Escherichia coli 0157:H7 str. EC869
10 : Escherichia coli 0157:H7 str. EC4206
11 : Escherichia coli 0157:H7 str. EC4401
12 : Escherichia coli 0157:H7 str. EDL933
13 : Escherichia coli 0157:H7 str. TW14588
14 : Escherichia coli 0157:H7 str. Sakai
15 : Escherichia coli 0157:H7 EC4045
16 : Escherichia coli 0157:H7 str. LANL ECF
17 : Escherichia coli 0157:H7 str. LANL ECA
18 : Escherichia coli K12 str. DH10B
19 : Escherichia coli K12 str. MG1655
20 : Escherichia coli K12 str. W3110
21 : Escherichia coli K12 str. DH1
22 : Escherichia coli BW2952
23 : Escherichia coli ATCC8739
24 : Escherichia coli B REL606
25 : Escherichia coli BL21 (DE3 Korea)
26 : Escherichia coli BL21 (DE3 AU)
27 : Escherichia coli BL21 (DE3 DOE)
28 : Escherichia coli HS
29 : Escherichia coli SE11
30 : Escherichia coli IAI1
31 : Escherichia coli 55989
32 : Escherichia coli E24377A
33 : Escherichia coli O26:H11 str. 11368
34 : Escherichia coli 0127:H6 str. E2348/69
35 : Escherichia coli O103:H2 str. 12009
36 : Escherichia coli O111:H- str. 11128
37 : Escherichia coli 0103 Oslo
38 : Escherichia coli SMS-3-5
39 : Escherichia coli UMN026
40 : Escherichia coli 53638
41 : Escherichia coli IAI39
42 : Escherichia coli UTI89
43 : Escherichia coli S88
44 : Escherichia coli CFT073
45 : Escherichia coli SE15
46 : Escherichia coli 536
47 : Escherichia coli ED1a
48 : Escherichia coli F11
49 : Escherichia coli APECO1
50 : Escherichia coli E110019
51 : Escherichia coli E22
52 : Escherichia coli B7A
53 : Escherichia coli 101-1
54 : Shigella flexneri 2a 2457T
55 : Shigella flexneri 2a 301
56 : Shigella flexneri 5 8401
57 : Shigella boydii CDC 3083-94
58 : Shigella boydii Sb227
59 : Shigella sonnei Ss046
60 : Escherichia fergusonii ATCC 35469
61 : Escherichia albertii TW07627
62 : Salmonella enterica Typhimurium LT2
63 : Shigella dysenteriae Sd197
64 : Shigella dysenteriae 1012

slack of 19 – a core gene family is present in ~90 %
of the genomes to be considered a core gene

Homolog Gene Clusters



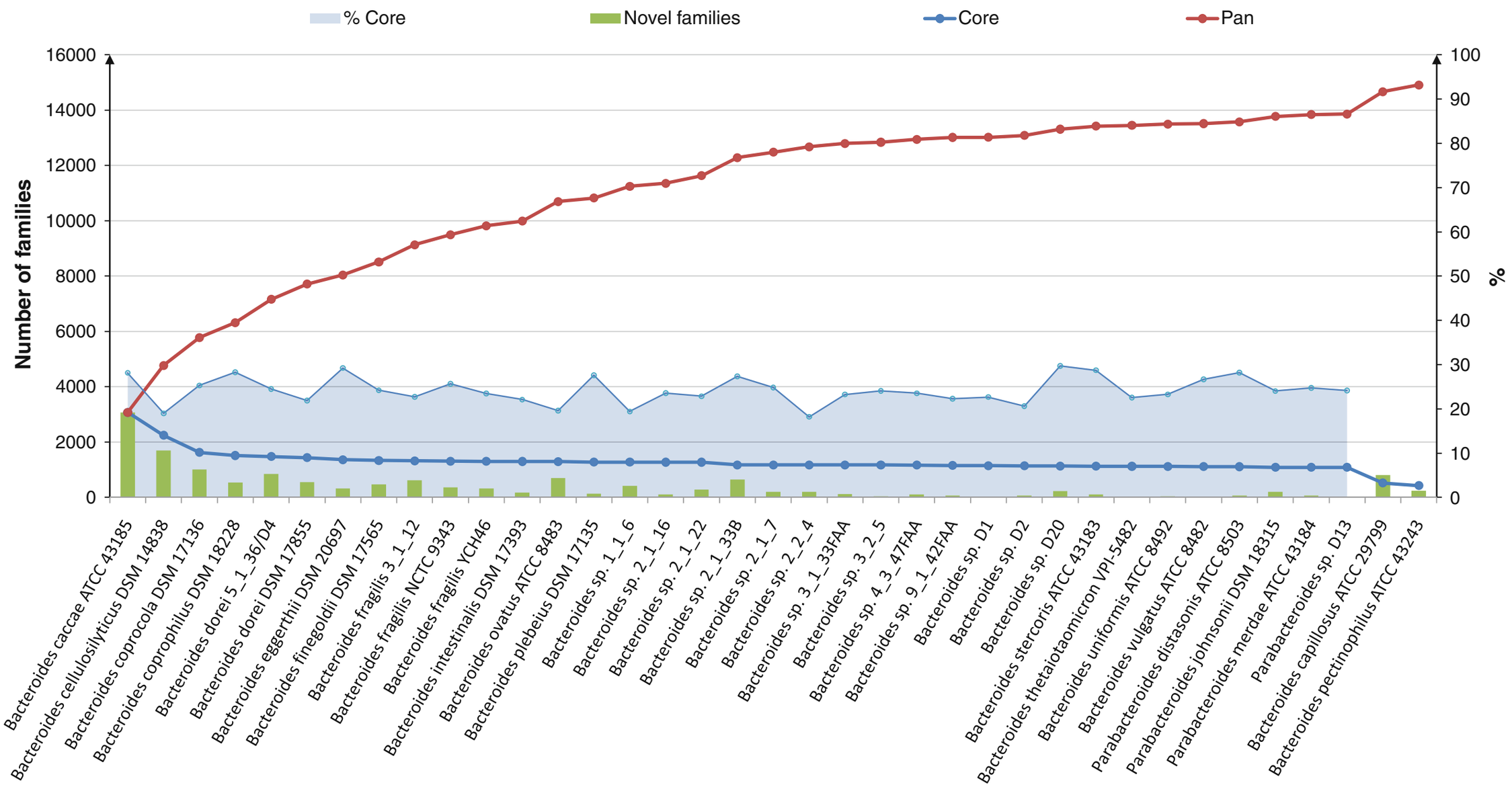
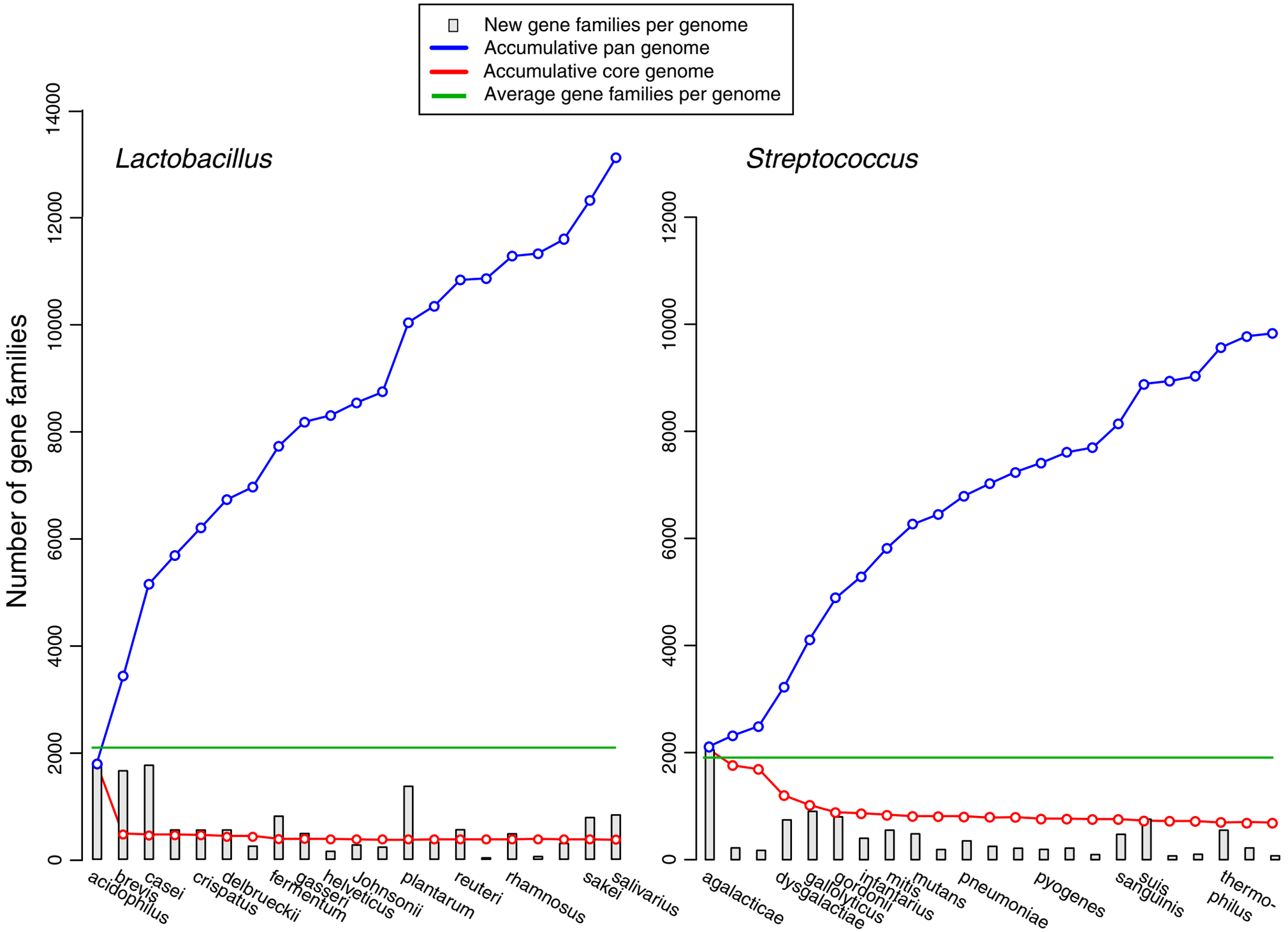


Figure 3 Pan- and core genome plot of *Bacteroides* and *Parabacteroides* genomes. The *blue line* (core) represents the conserved number of orthoMCL protein families. The *red line* (pan) indicates the cumulative number of orthoMCL protein families in the genomes. *Green bar* indicate the number of novel orthoMCL protein families in the genome. The relative size of the core protein families to the total genome size (% Core) is based on the 1,085 protein families shared by *Bacteroides* and *Parabacteroides* (excluding *B. capillosus* and *B. pectinophilus*), On average, 27% of the proteins is shared in the core protein families

Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9948-y



Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9948-y

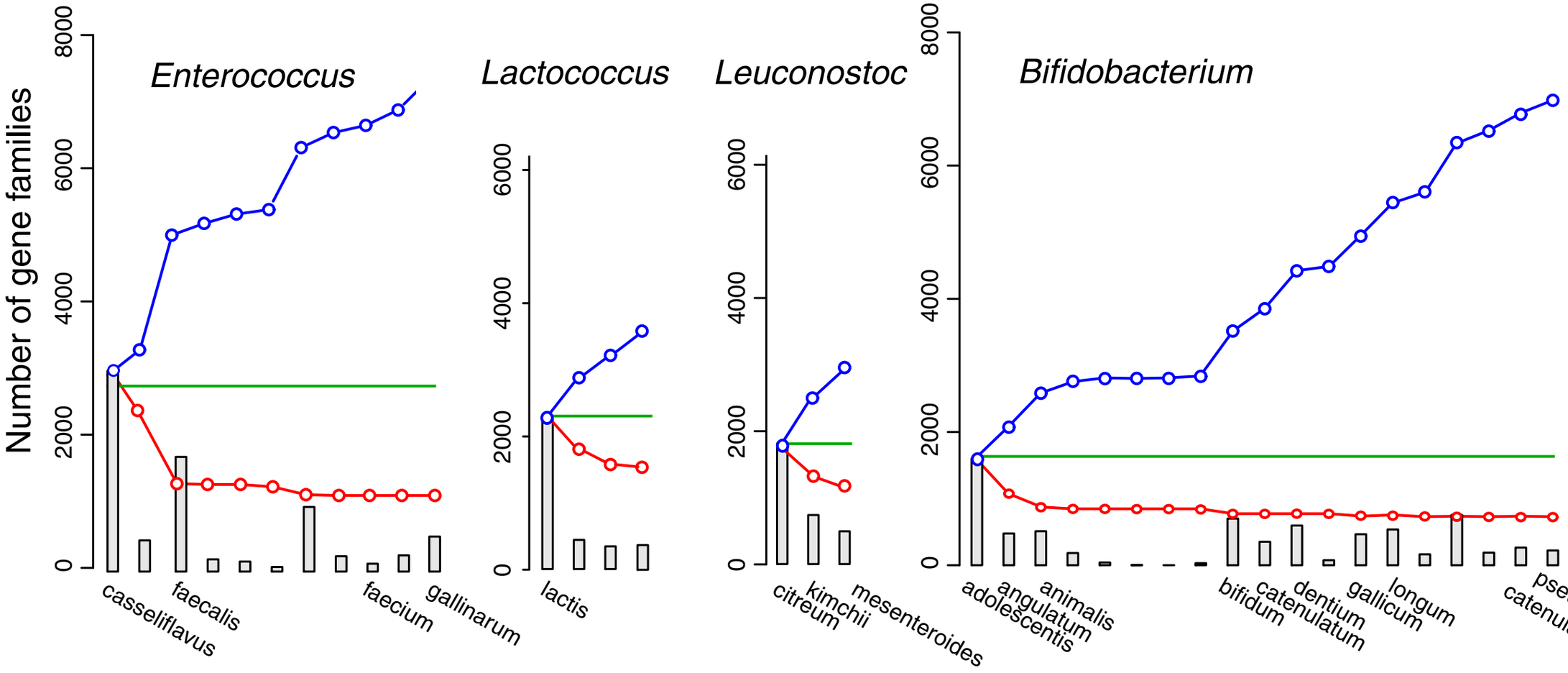


Figure 1 Pan- and core genome plots of the six analyzed genera. The genomes were analyzed in alphabetical order of species names

Bifidobacterium

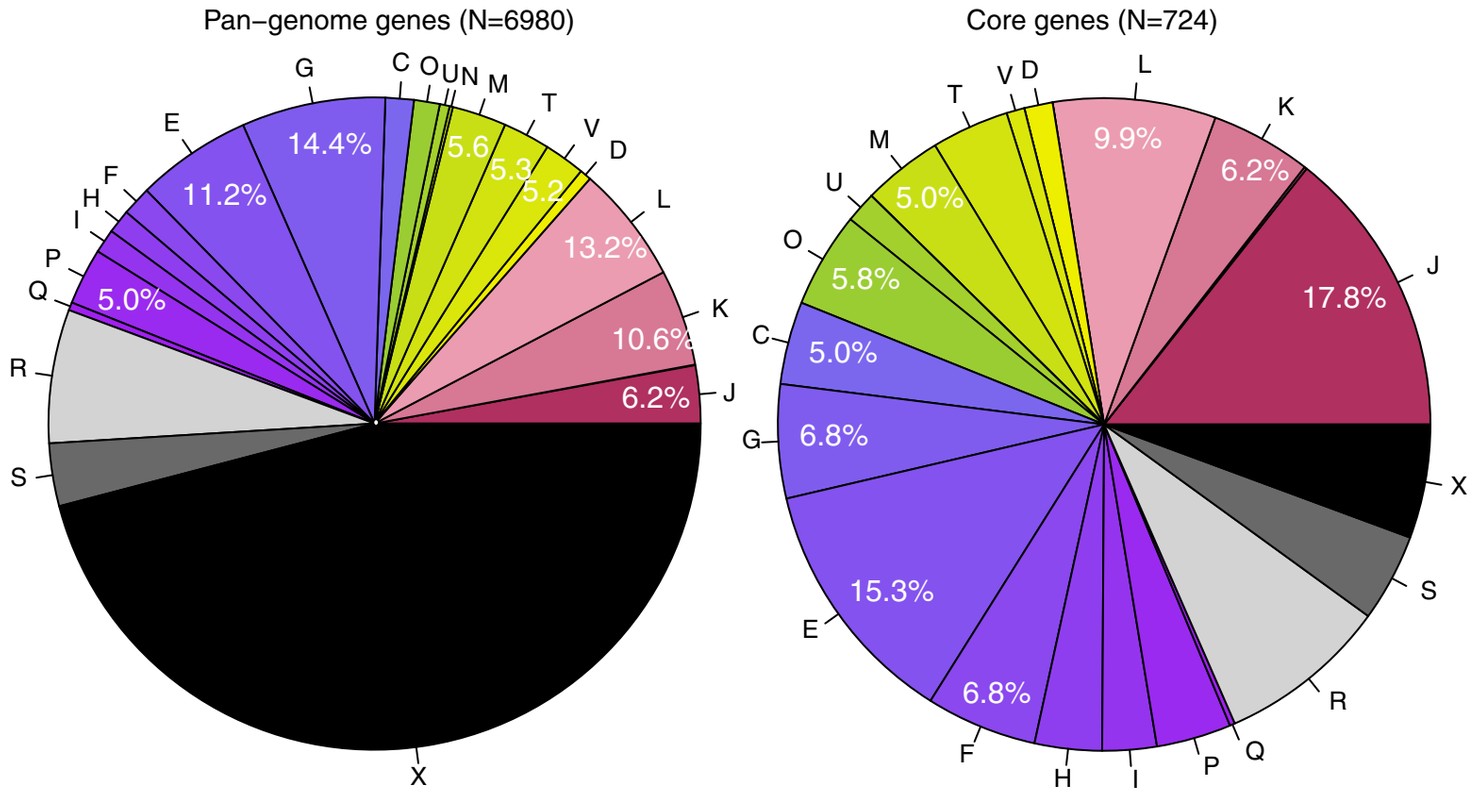
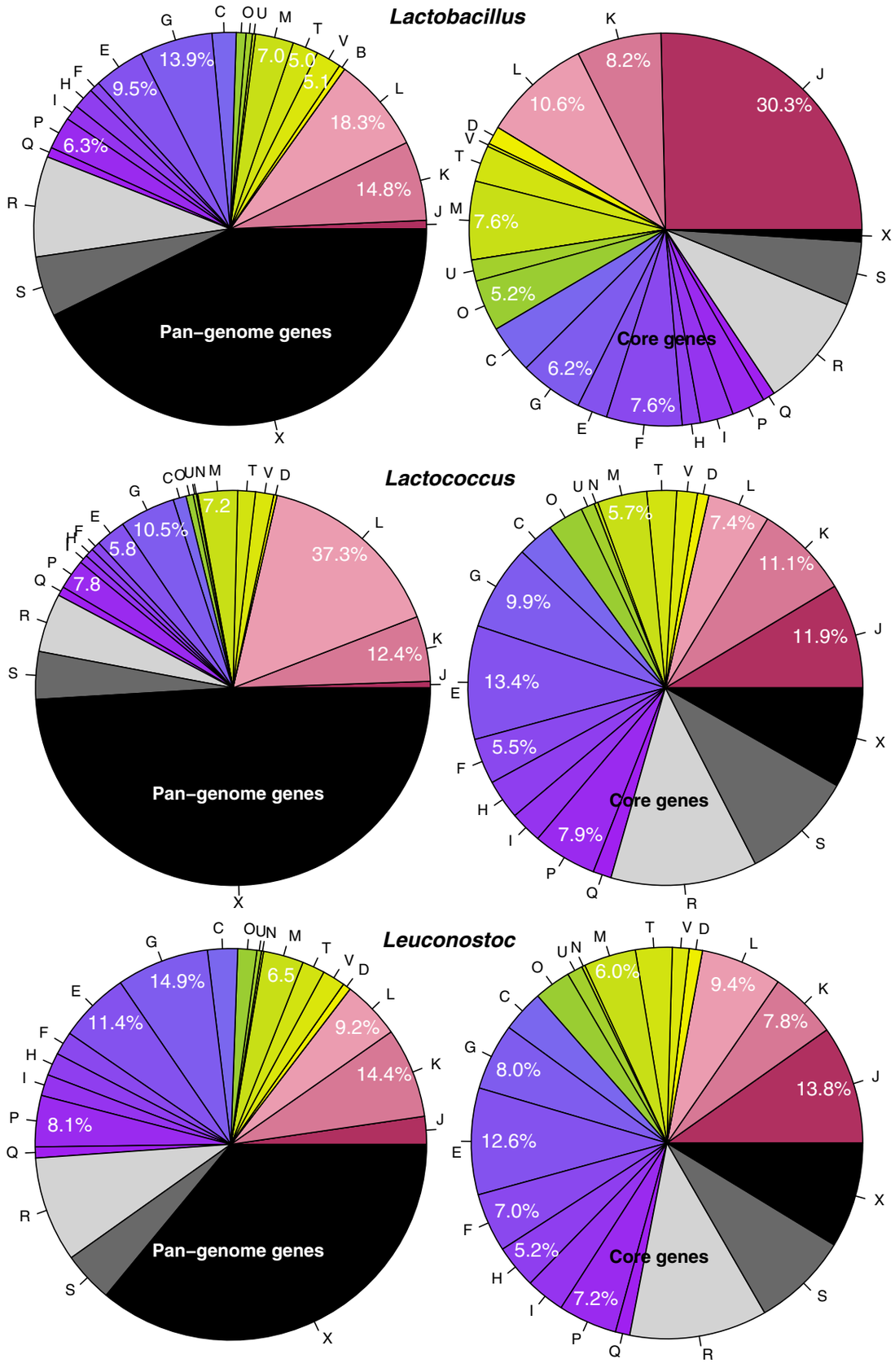


Figure 9 COG statistics for the genes found in the pan-genome (*left*) and core genome (*right*) of *Bifidobacterium* genomes. The key for the COG classes is explained below the pie charts. Percentages given in the pie chart are calculated by exclusion of classes R, S and X. Only values ≥5% are shown

Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9948-y



Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9948-y

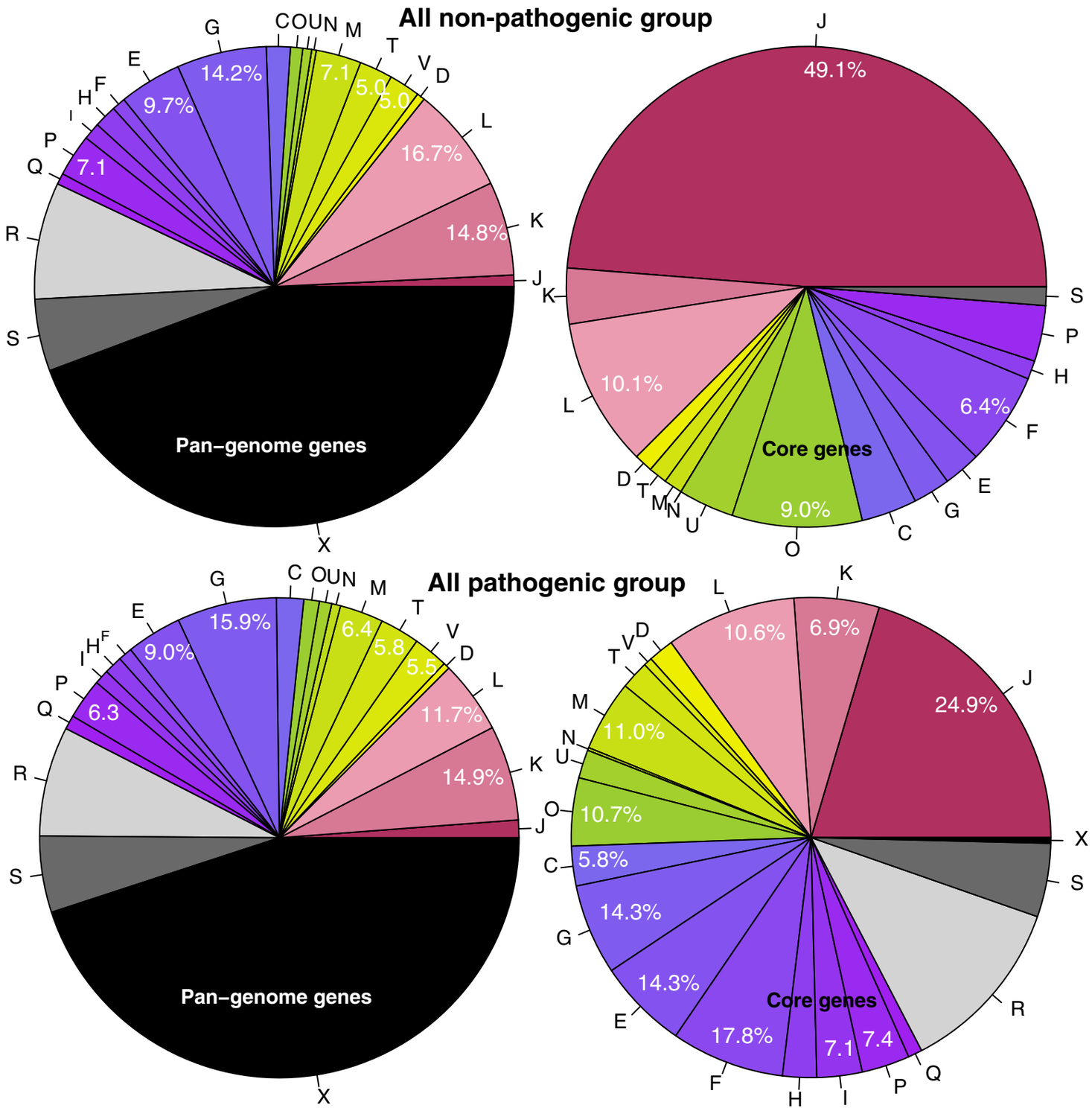


Figure 11 COG statistics for the genes found in the pan-genome (*left*) and core genome (*right*) of the collection of genomes from all included organisms, divided into non-pathogenic isolates (probiotic, fermentative and normal human gut flora) at the *top* and pathogenic isolates at the *bottom*

Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9948-y

3. Who cares? What are pan-genomes good for?

Standard operating procedure for computing pangenome trees

Standards in Genomic Sciences (2010) **2:**135-141

SIGS Standards in Genomic Sciences

An Open Access Journal of the Genomic Standards Consortium

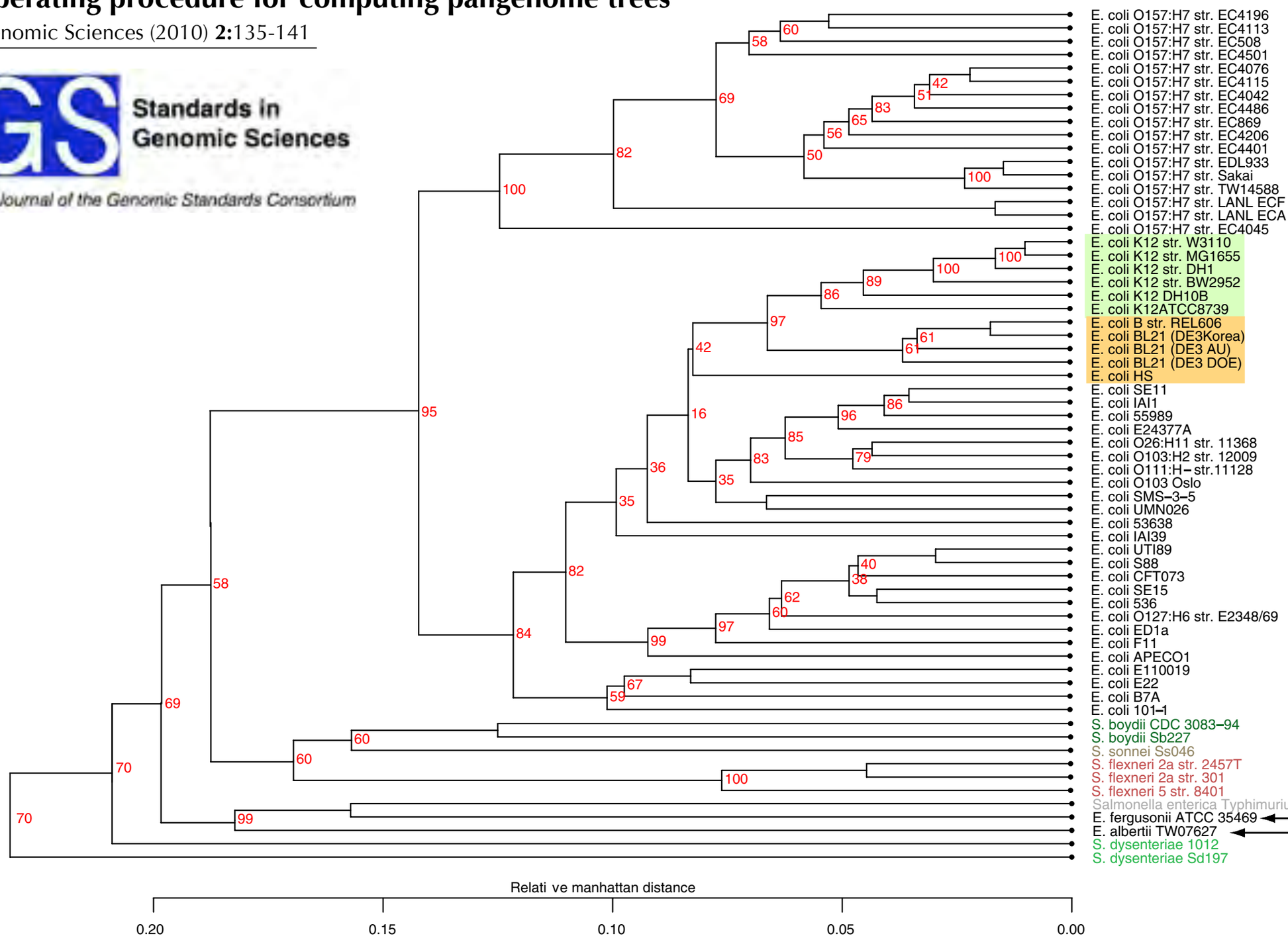


Figure 3 Pan-genome clustering of *E. coli* (*black*) and related species (*colored*), based on the alignment of their variable gene content. The genomes now cluster according to species and a relatedness between *E. coli* K12 derivatives (*green block*) and group B isolates (*orange block*) is visible

Microbial Ecology, **60**:708-720, (2010).

Pangenome trees

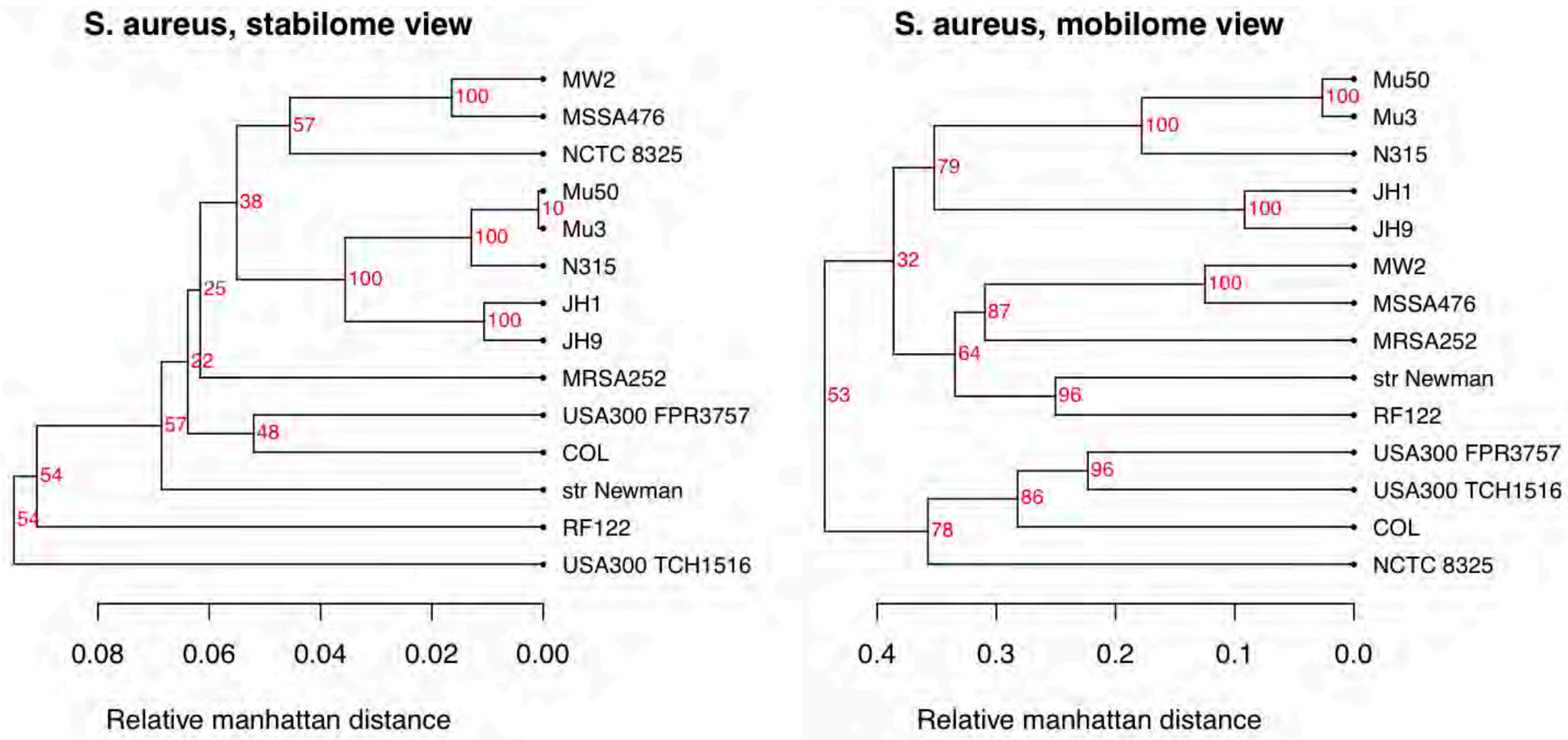


Figure 5. The left panel show the pan-genome family tree for the 14 strains of *Staphylococcus aureus* completed as NCBI. Here the weights have been chosen according to the blue bars in Figure 1, i.e. the "stabilome" genes have been emphasized. In the right panel, the same data have been used, but weights are now chosen to emphasize the "mobilome" genes. In both cases ORFans have been discarded.

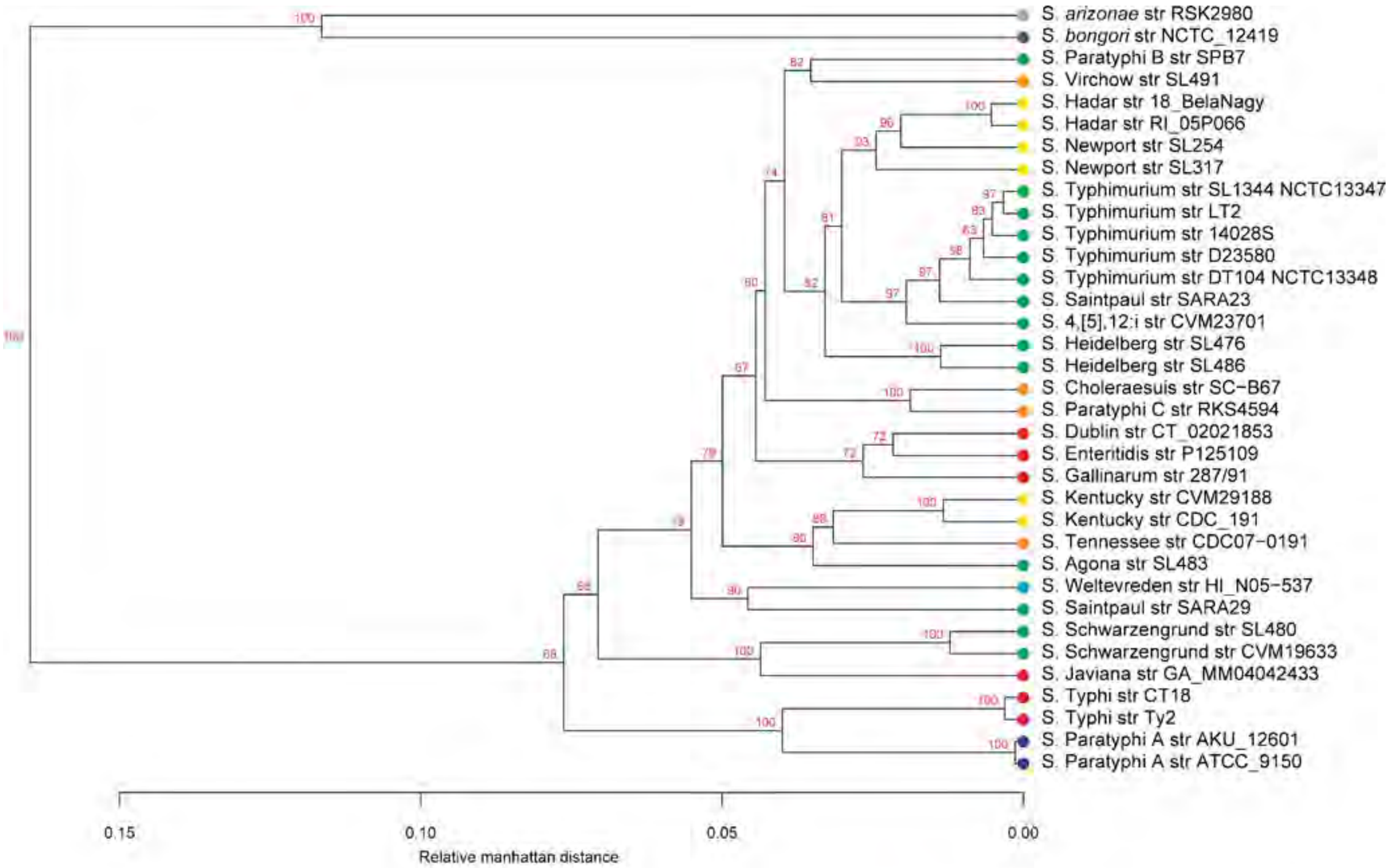
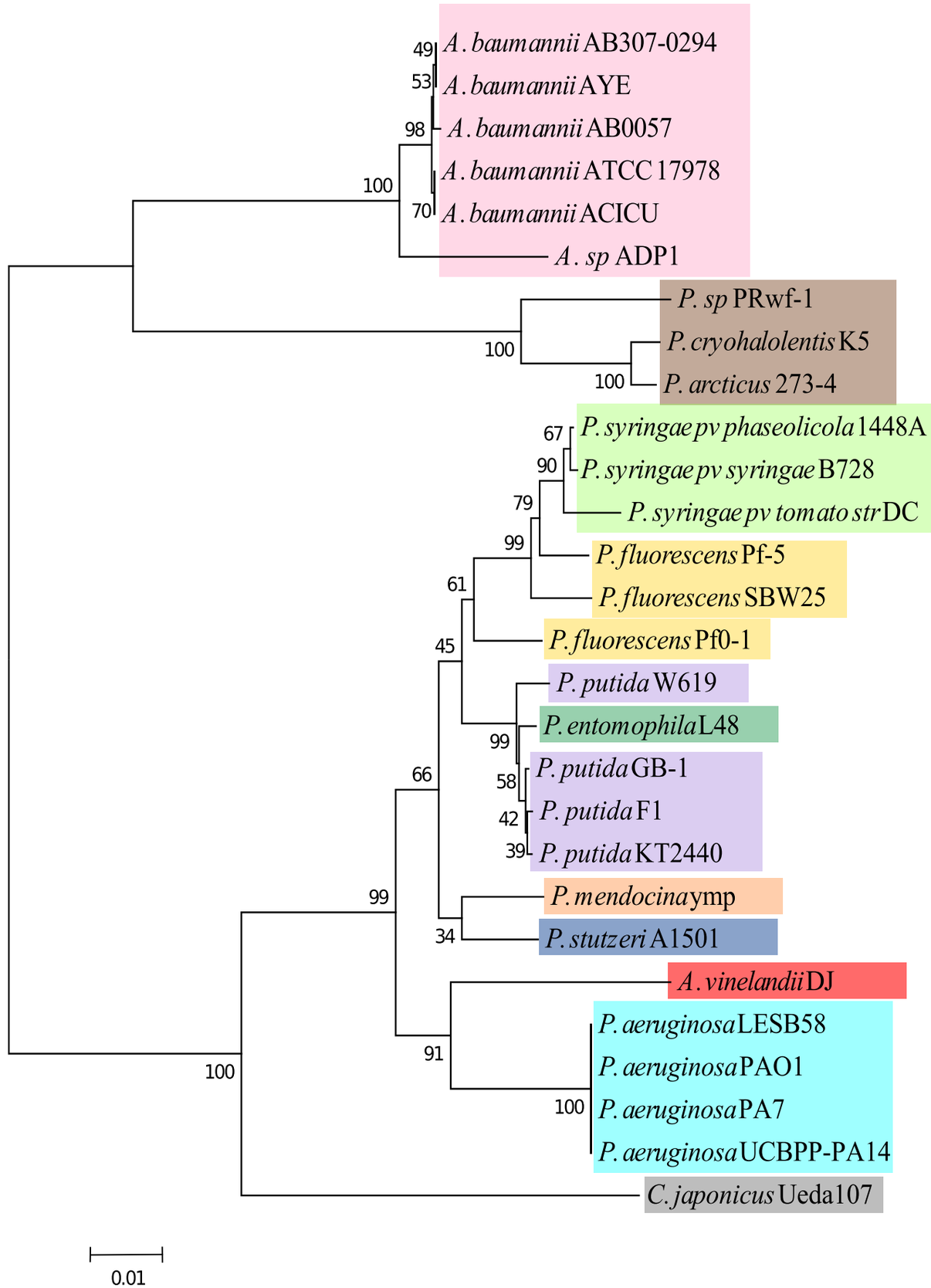

S. arizonae str RSK2980
S. bongori str NCTC_12419
S. Paratyphi B str SPB7
S. Virchow str SL491
S. Hadar str 18_BelaNagy
S. Hadar str RI_05P066
S. Newport str SL254
S. Newport str SL317
S. Typhimurium str SL1344 NCTC13347
S. Typhimurium str LT2
S. Typhimurium str 14028S
S. Typhimurium str D23580
S. Typhimurium str DT104 NCTC13348
S. Saintpaul str SARA23
S. 4,[5],12:i str CVM23701
S. Heidelberg str SL476
S. Heidelberg str SL486
S. Choleraesuis str SC-B67
S. Paratyphi C str RKS4594
S. Dublin str CT_02021853
S. Enteritidis str P125109
S. Gallinarum str 287/91
S. Kentucky str CVM29188
S. Kentucky str CDC_191
S. Tennessee str CDC07-0191
S. Agona str SL483
S. Weltevreden str HI_N05-537
S. Saintpaul str SARA29
S. Schwarzengrund str SL480
S. Schwarzengrund str CVM19633
S. Javiana str GA_MM04042433
S. Typhi str CT18
S. Typhi str Ty2
S. Paratyphi A str AKU_12601
S. Paratyphi A str ATCC_9150
0.15
0.10
0.05
0.00
Relative manhattan distance

Figure 1 16S rRNA tree generated with the neighbor-joining method with 1,000 bootstrap resamplings. The tree shows the evolutionary relationships of *A. vinelandii* with the *Pseudomonas* genus and other Gammaproteobacteria based on their 16S rRNA sequences



Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9914-8

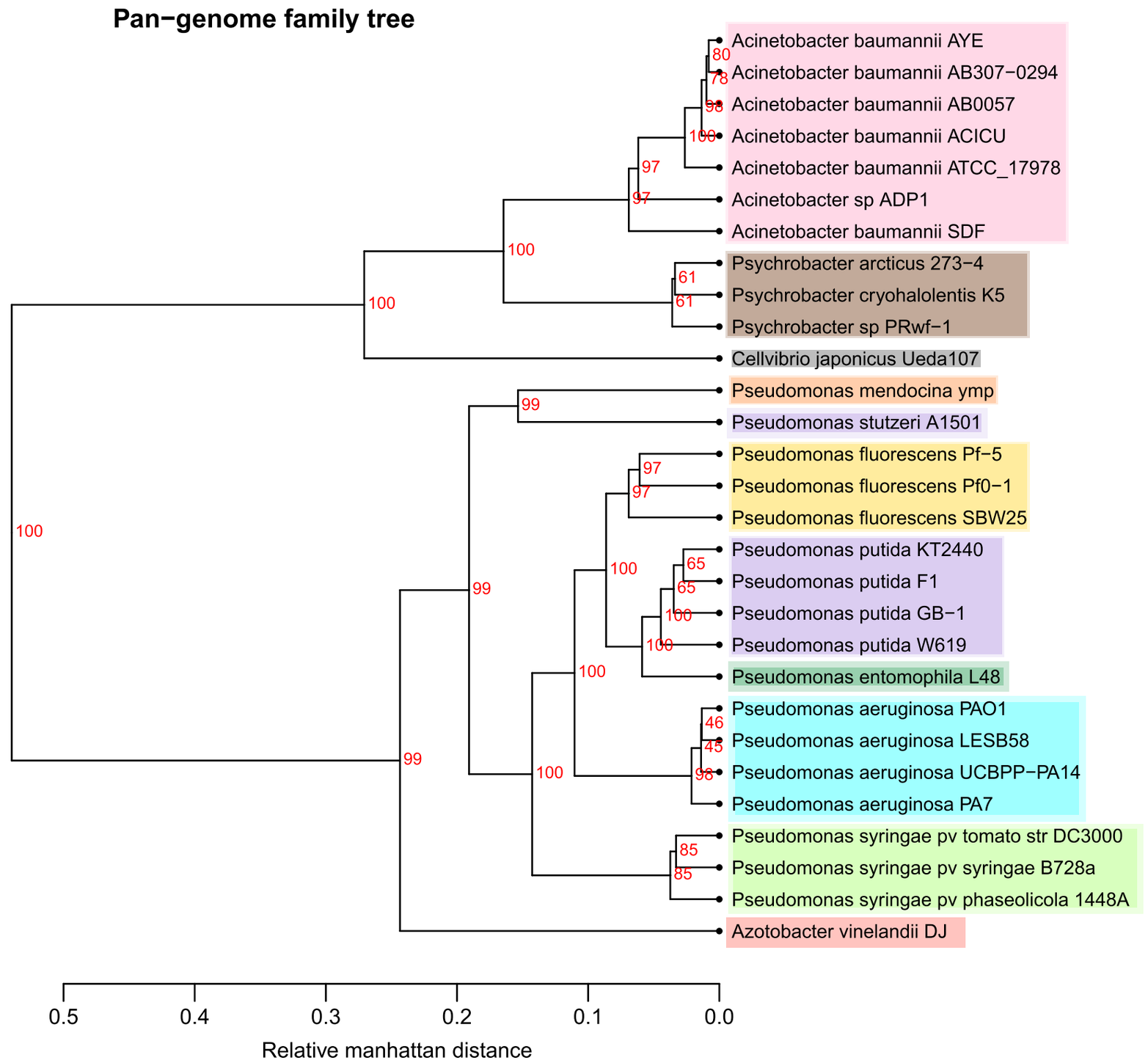


Figure 2 Pan genome family tree. The tree shows the phylogenetic relationships based on the gene families found in the pan genome, excluding the families found in only one genome. Color coding is again based on Table 1

Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9914-8

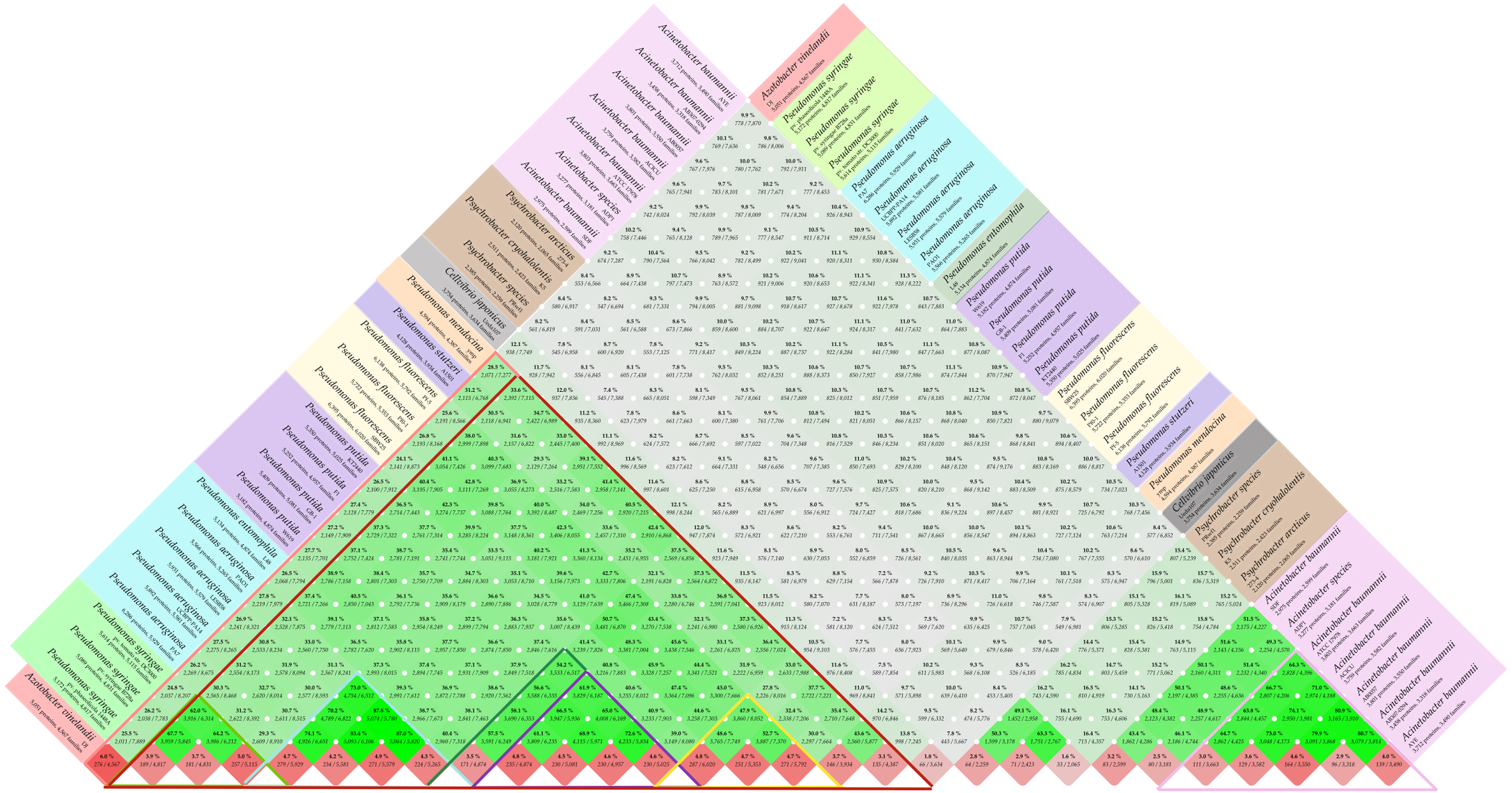


Figure 4 BLASTMatrix shows the homology between the pairs of genomes compared (*green*) and homology within each genome itself (*red*) based on the 50/50 rule

Microbial Ecology, *in the press (2011)*
DOI: 10.1007/s00248-011-9914-8

Wednesday, 14 December, 2011

Questions:

Consider the following taxonomic groups:

Proteobacteria (phyla)
Enterobacteriaceae (family)
Escherichia (genus)
E. coli (species)

- Which core-genome is likely to be the largest?
- Which pan-genome is likely to be the largest?
- How would you expect the addition of a low-quality unfinished genome to affect the size of the core- and pan-genome?