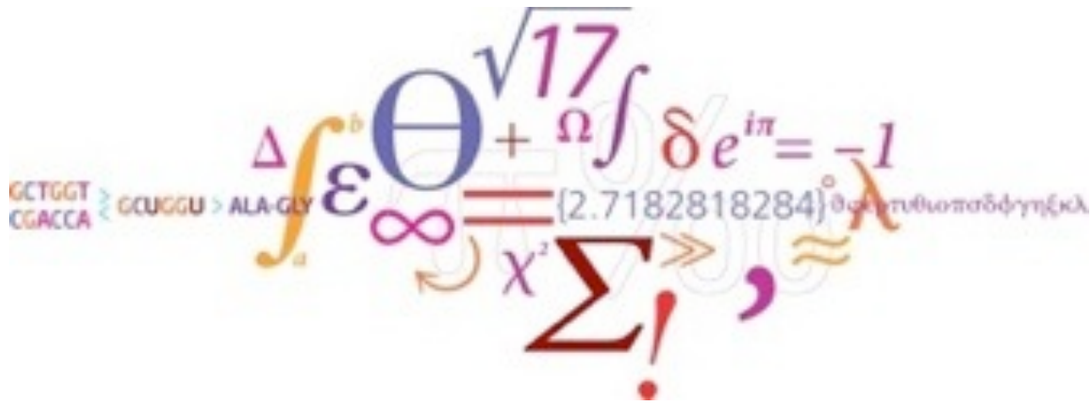
# Comparative Genomics

## with an emphasis on bacterial genomes....




**Dave Ussery**
UiO course #MBV-INF 4410
Bioinformatics for Molecular Biology

Comparative Genomics lecture
Friday, 10 September, 2010

**Center for Biological Sequence Analysis**
Department of Systems Biology


DFDS SEAWAYS

# Center for Genomic Epidemiology

## Three questions -

### from a bacterial genome sequence, can we compute:

### 1. What is it?

(*e.g., E. coli, S. aureus*)

### 2. Have we seen this before?

(*S. aureus* strain M1 from Hvidovre hospital)

### 3. How do we treat it?

(Gylcopeptides, such as vancomycin or teicoplanin;
- NOTE: don't use β-lactames)

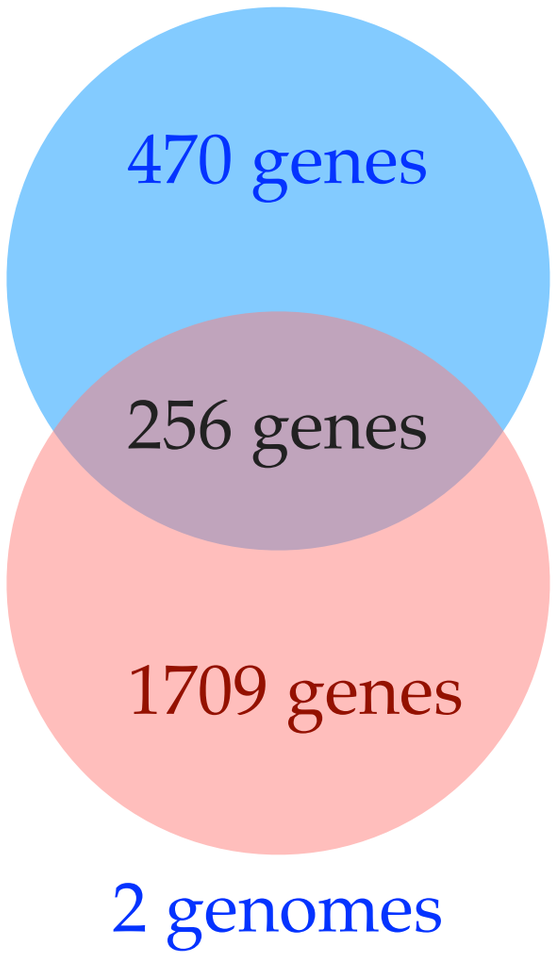
*Proc. Natl. Acad. Sci. USA*
Vol. 93, pp. 10268–10273, September 1996
Evolution

# 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)*




**ABSTRACT** The recently sequenced genome of the parasitic bacterium *Mycoplasma genitalium* contains only 468 identified protein-coding genes that have been dubbed a minimal gene complement [Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* 270, 397–403]. Although the *M. genitalium* gene complement is indeed the smallest among known cellular life forms, there is no evidence that it is the minimal self-sufficient gene set. To derive such a set, we compared the 468 predicted *M. genitalium* protein sequences with the 1703 protein sequences encoded by the other completely sequenced small bacterial genome, that of *Haemophilus influenzae*. *M. genitalium* and *H. influenzae* belong to two ancient bacterial lineages, i.e., Gram-positive and Gram-negative bacteria, respectively. Therefore, the genes that are conserved in these two bacteria are almost certainly essential for cellular function. It is this category of genes that is most likely to approximate the minimal gene set. We found that 240 *M. genitalium* genes have orthologs among the genes of *H. influenzae*. This collection of genes falls short of comprising the minimal set as some enzymes responsible for intermediate steps in essential pathways are missing. The apparent reason for this is the phenomenon that we call nonorthologous gene displacement when the same function is fulfilled by nonorthologous proteins in two organisms. We identified 22 nonorthologous displacements and supplemented the set of orthologs with the respective *M. genitalium* genes. After examining the resulting list of 262 genes for possible functional redundancy and for the presence of apparently parasite-specific genes, 6 genes were removed. We suggest that the remaining 256 genes are close to the minimal gene set that is necessary and sufficient to sustain the existence of a modern-type cell. Most of the proteins encoded by the genes from the minimal set have eukaryotic or archaeal homologs but seven key proteins of DNA replication do not. We speculate that the last common ancestor of the three primary kingdoms had an RNA genome. Possibilities are explored to further reduce the minimal set to model a primitive cell that might have existed at a very early stage of life evolution.

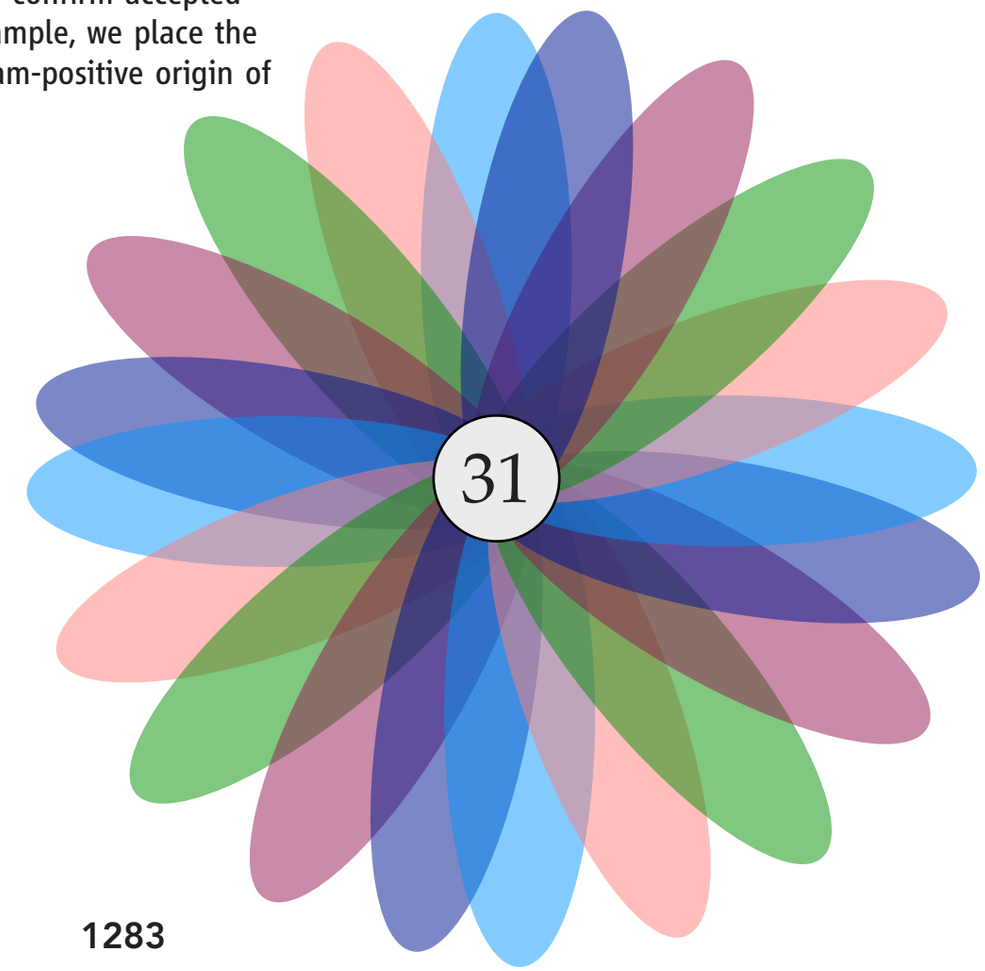
# Toward Automatic Reconstruction of a Highly Resolved Tree of Life

Francesca D. Ciccarelli,[1,2,3]* Tobias Doerks,[1]* Christian von Mering,[1] Christopher J. Creevey,[1] Berend Snel,[4] Peer Bork[1,5]†

We have developed an automatable procedure for reconstructing the tree of life with branch lengths comparable across all three domains. The tree has its basis in a concatenation of 31 orthologs occurring in 191 species with sequenced genomes. It revealed interdomain discrepancies in taxonomic classification. Systematic detection and subsequent exclusion of products of horizontal gene transfer increased phylogenetic resolution, allowing us to confirm accepted relationships and resolve disputed and preliminary classifications. For example, we place the phylum Acidobacteria as a sister group of δ-Proteobacteria, support a Gram-positive origin of Bacteria, and suggest a thermophilic last universal common ancestor.

31 genes

191 genomes



# The Vanishing Set of Conserved "Core" Genes

1996 - 256 conserved proteins, based on 2 genomes

2006 - 31 conserved proteins, 191 genomes

2010 - 0 conserved proteins, 1000 genomes

**Genome update: the 1000th genome – a cautionary tale**

There are now more than 1000 sequenced prokaryotic genomes deposited in public databases and available for analysis. Currently, although the sequence databases GenBank, DNA Database of Japan and EMBL are synchronized continually, there are slight differences in content at the genomes level for a variety of logistical reasons, including differences in format and loading errors, such as those caused by file transfer protocol interruptions. This means that the 1000th genome will be different in the various databases. Some of the data on the highly accessed web pages are inaccurate, leading to false conclusions for example about the largest bacterial genome sequenced. Biological diversity is far greater than many have thought. For example, analysis of multiple *Escherichia coli* genomes has led to an estimate of around 45 000 gene families – more genes than are recognized in the human genome. Moreover, of the 1000 genomes available, not a single protein is conserved across all genomes. Excluding the members of the *Archaea*, only a total of four genes are conserved in all bacteria: two protein genes and two RNA genes.

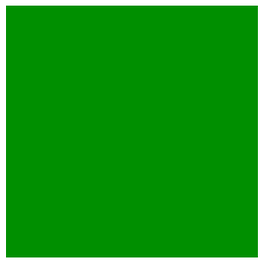
omes listed at NCBI and more than twice as many genomes (2307) listed as 'in progress'; the Genomes Online Database (GOLD) web pages (see below and link in Table 1) boast more than 6400 microbial genome sequencing projects.

## The 1000th genome(s)

In principle, there should be a list where one could go to find the 1000th genome; however, as several genomes are processed and submitted on an almost daily basis to databases, determining the 1000th genome is not as easy as might appear at first. Table 1 lists the set of genomes for the various databases. According to GOLD (which

Microbiology, **156**:603-608, (2010).

human

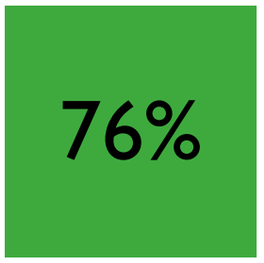


23,621 genes

chimp


19,829 genes
19,568 orthologs

chicken

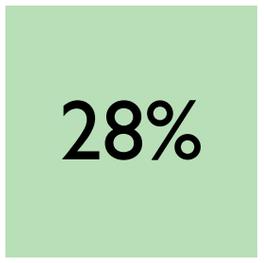



18,529 genes
14,080 orthologs

worm

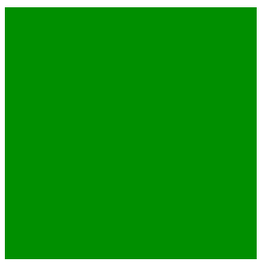

19,404 genes
10,673 orthologs

yeast




5,885 genes
1,647 orthologs

*C. botulanium*
type A, strain ATCC 3502

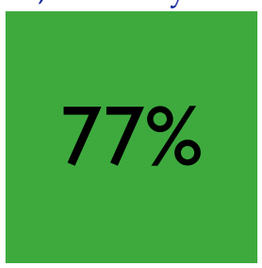


3606 genes

*C. botulanium*
type A, strain ATCC 19397


3553 genes
3202 orthologs

*C. botulanium*
type A, strain Kyoto



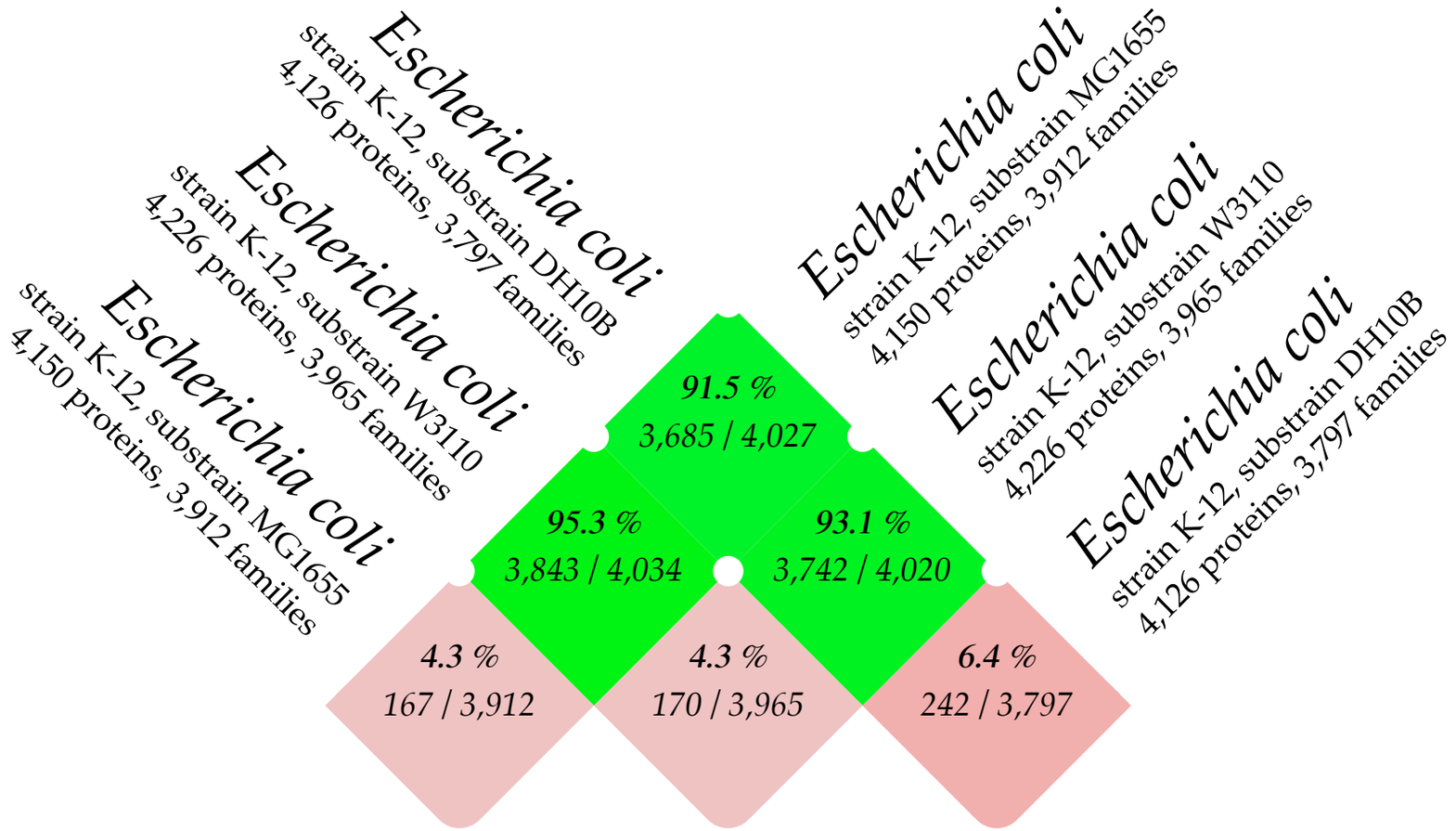

3874 genes
2974 orthologs

*C. botulanium*
type C, strain Ecklund

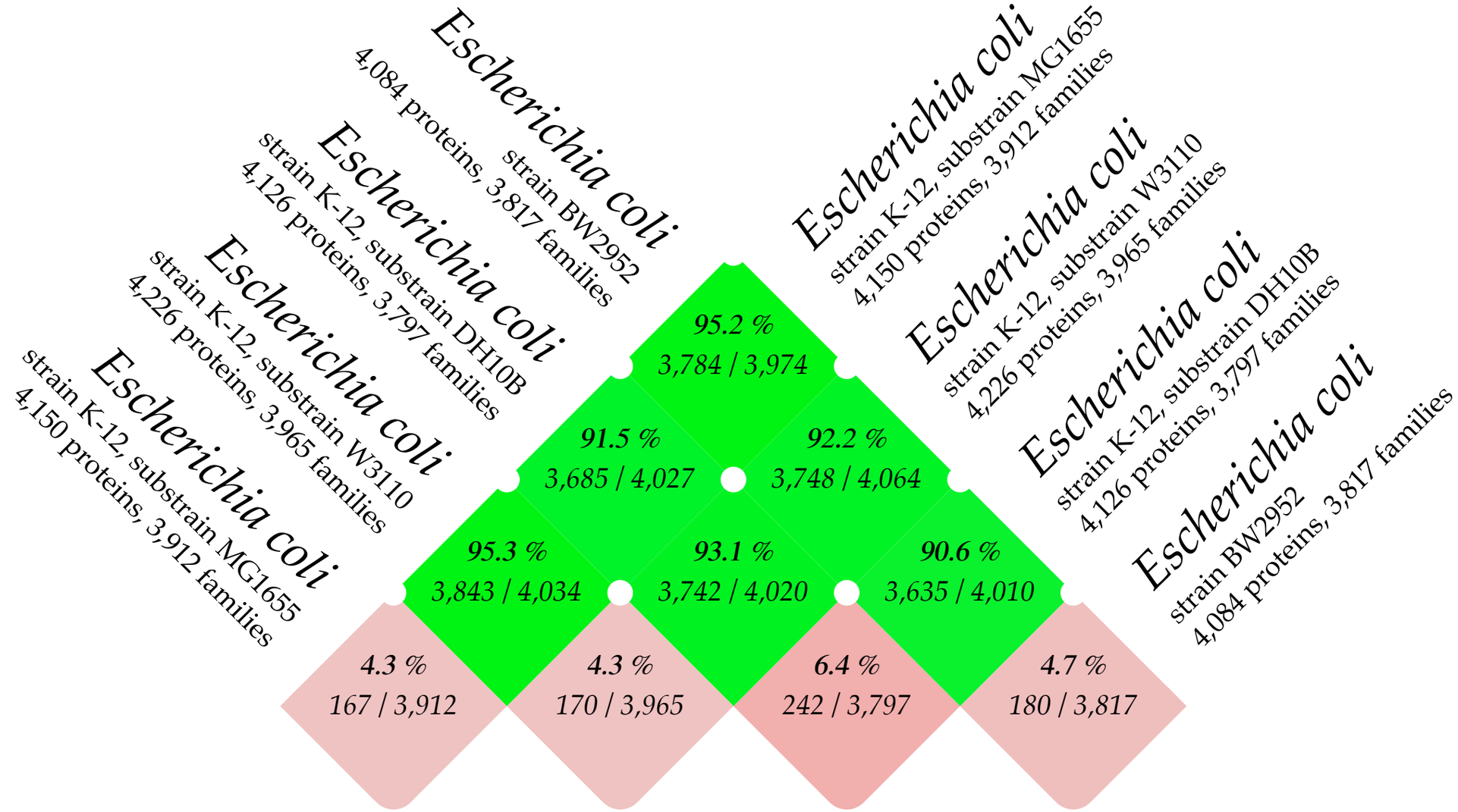

2801 genes
1126 orthologs

*C. botulanium*
type E1, strain BoNT E Beluga

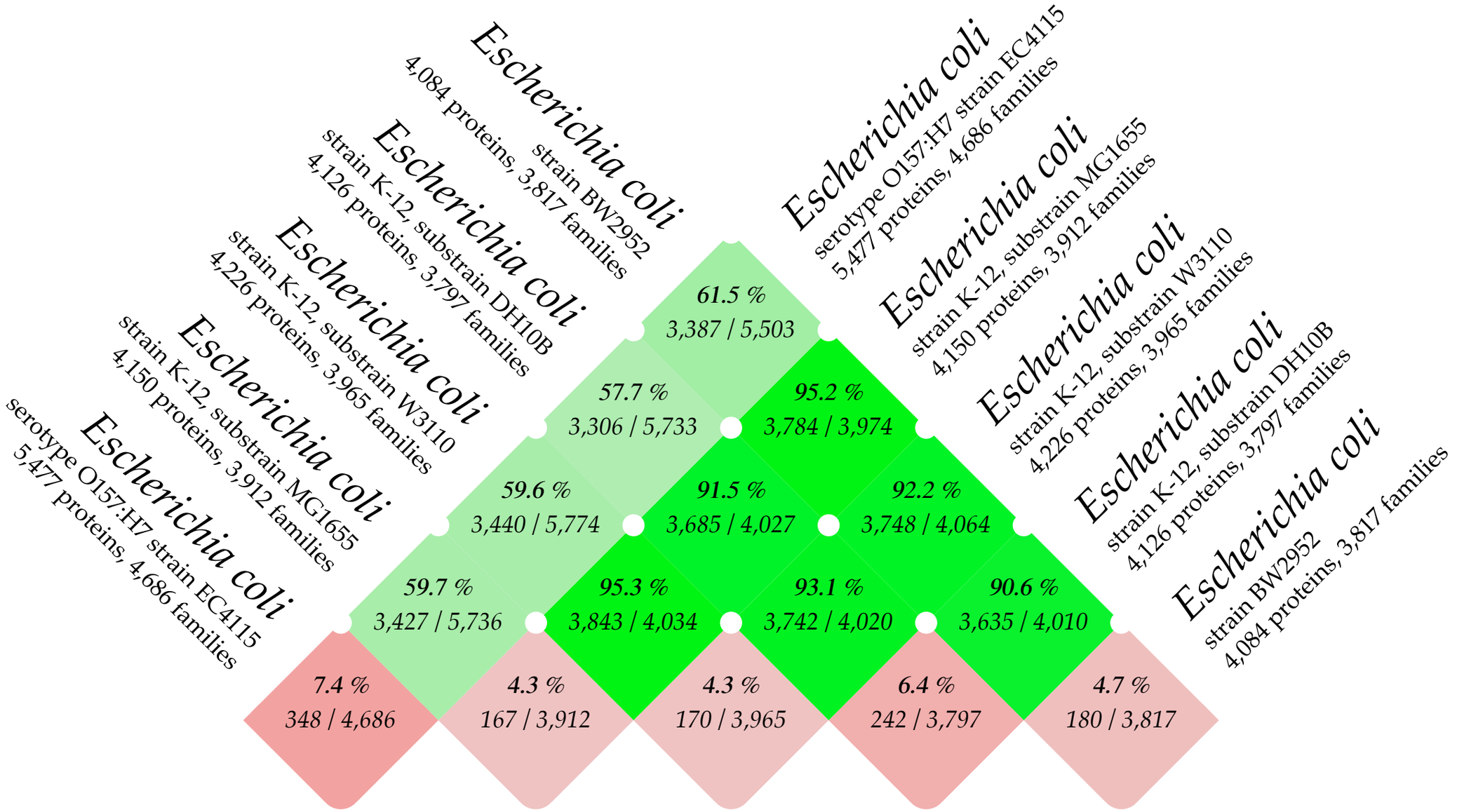

3760 genes
1092 orthologs


Escherichia coli
strain K-12, substrain DH10B
4,126 proteins, 3,797 families
Escherichia coli
strain K-12, substrain W3110
4,226 proteins, 3,965 families
Escherichia coli
strain K-12, substrain MG1655
4,150 proteins, 3,912 families
91.5 %
3,685 / 4,027
95.3 %
3,843 / 4,034
93.1 %
3,742 / 4,020
4.3 %
167 / 3,912
4.3 %
170 / 3,965
6.4 %
242 / 3,797

4 *E. coli* genomes

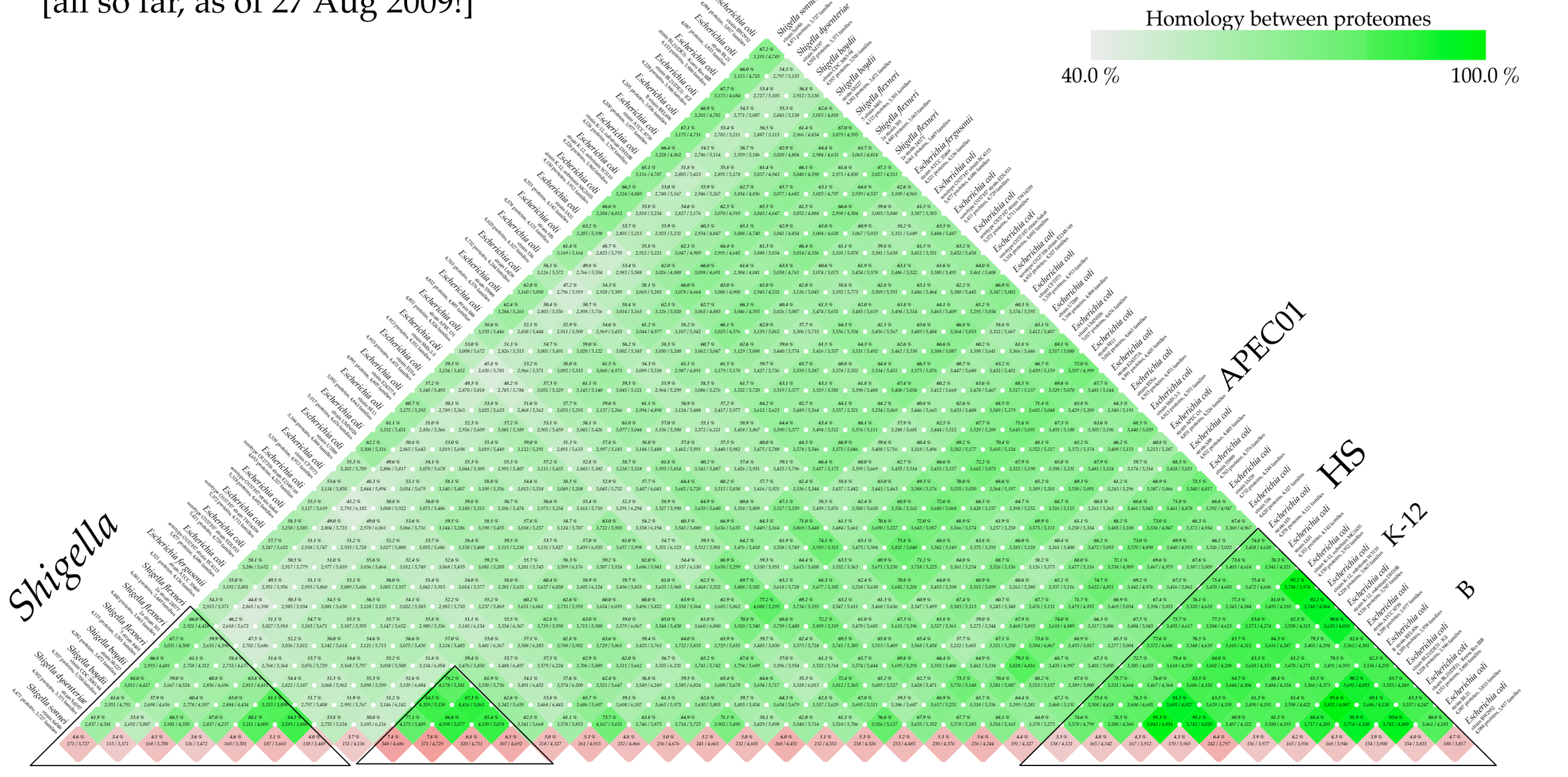


5 *E. coli* genomes



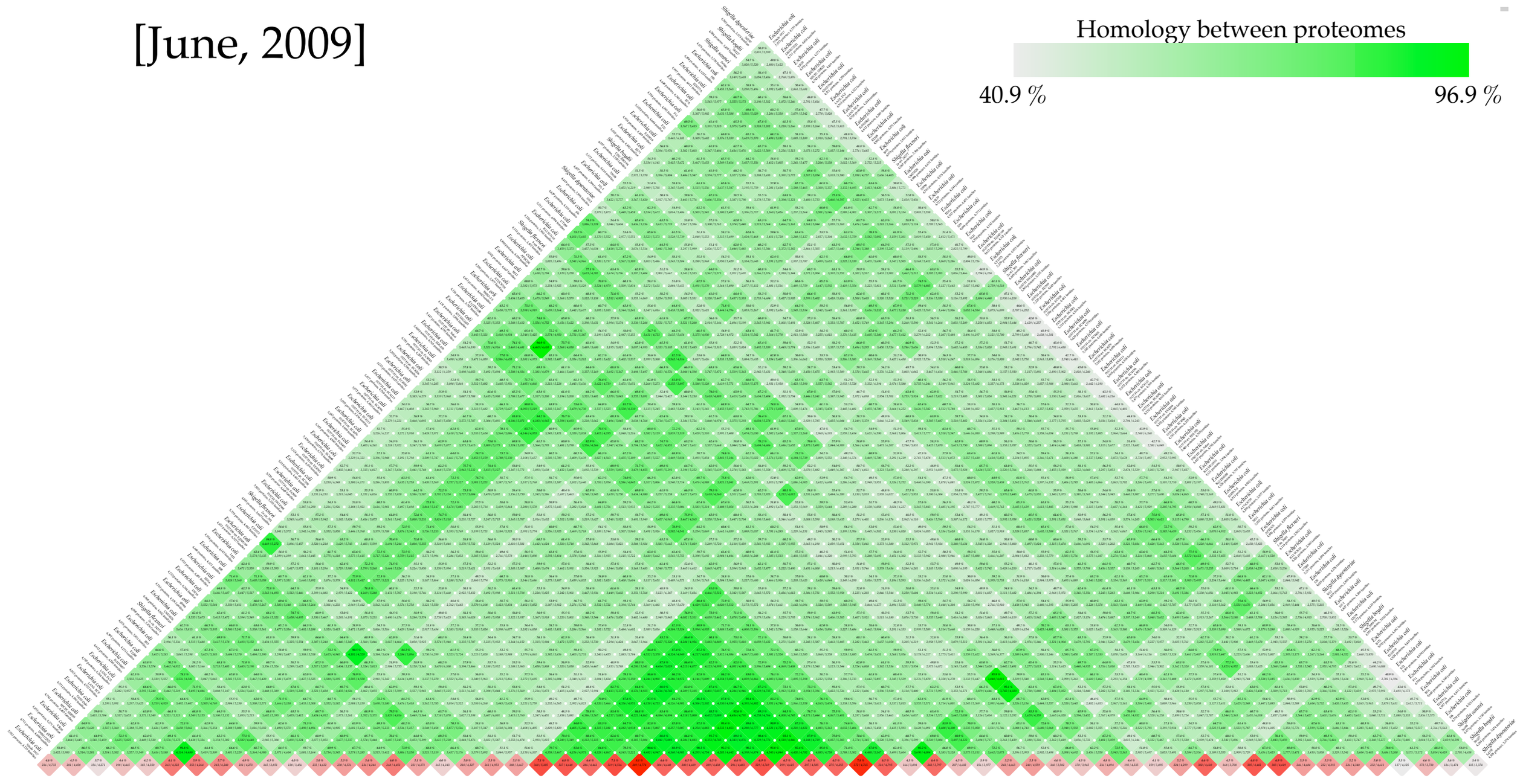
28 *E. coli* genomes

[all so far, as of 27 Aug 2009!]



60 *E. coli* genomes

# [June, 2009]



# *E. coli* pan-genome

~45,000 gene families

*E. coli* K-12
4144 proteins

960

# Microbial comparative pan-genomics using binomial mixture models

Lars Snipen*[1], Trygve Almøy[1]and David W. Ussery[2]

[1]Biostatistics, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Ås, Norway
[2]Centre for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark

Email: Lars Snipen*- lars.snipen@umb.no; Trygve Almøy - trygve.almoy@umb.no; David W. Ussery - dave@cbs.dtu.dk;

*Corresponding author

*BMC Genomics* 2009, **10:** *in the press [August 2009]*

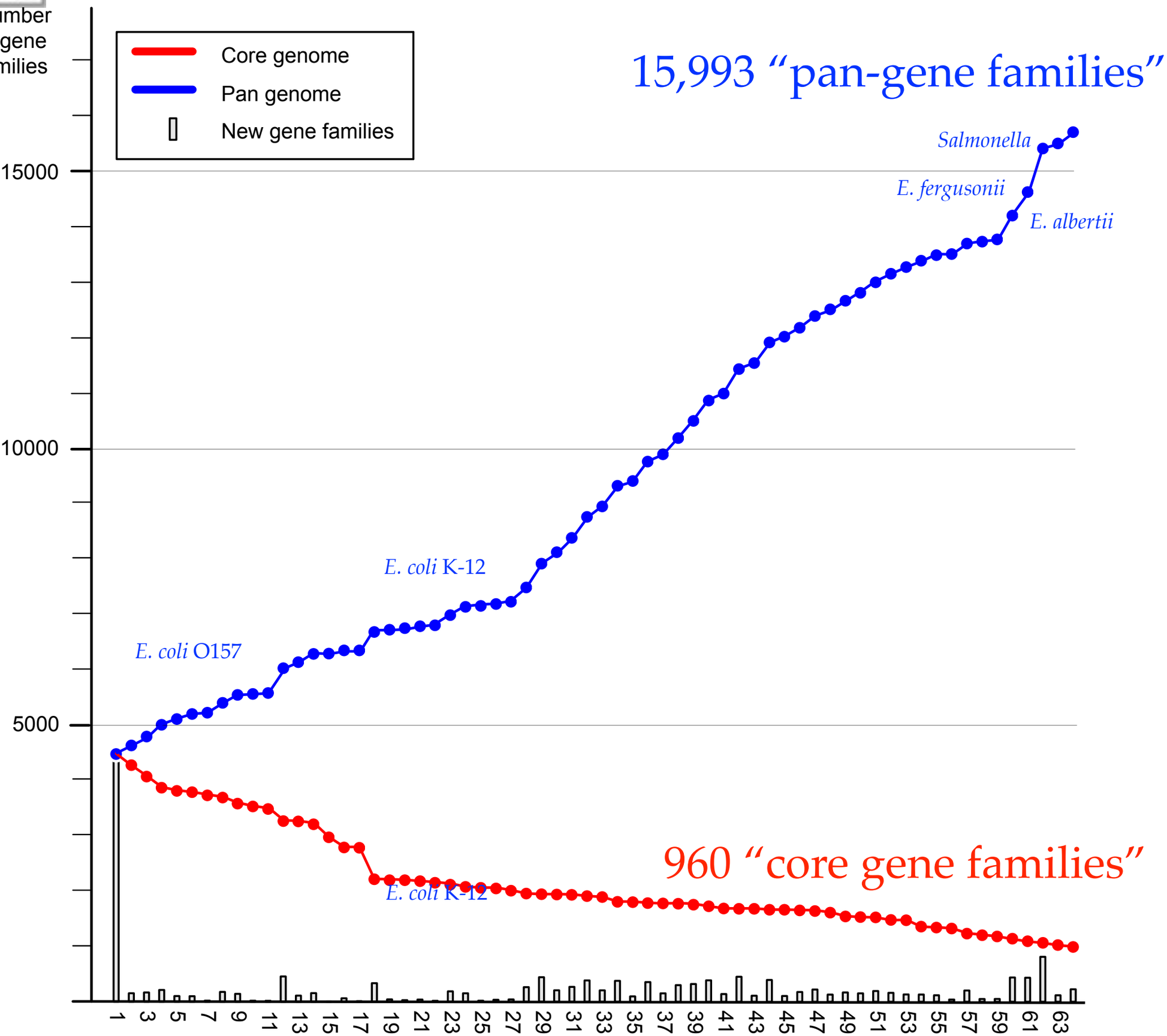
BMC Genomics

## Abstract

**Background:** The size of the core- and pan-genome of bacterial species is a topic of increasing interest due to the growing number of sequenced prokaryote genomes, many from the same species. Attempts to estimate these quantities have been made, using regression methods or mixture models. We extend the latter approach by using statistical ideas developed for capture-recapture problems in ecology and epidemiology.

**Results:** We estimate core- and pan-genome sizes for 16 different bacterial species. The results reveal a complex dependency structure for most species, manifested as heterogeneous detection probabilities. Estimated pan-genome sizes range from small (around 2600 gene families) in *Buchnera aphidicola* to large (around 43000 gene families) in *Escherichia coli*. Results for *Echerichia coli* show that as more data become available, a larger diversity is estimated, indicating an extensive pool of rarely occurring genes in the population.

**Conclusions:** Analyzing pan-genomics data with binomial mixture models is a way to handle dependencies between genomes, which we find is always present. A bottleneck in the estimation procedure is the annotation of rarely occurring genes.




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

Core genome
Pan genome

# 117 enteric genomes ("*E. coli* & friends")

46,058 "pan-gene families"

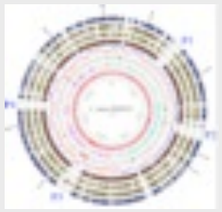
# genes in human genome

38 "core gene families"

0
10000
20000
30000
40000
50000

1 5 9 14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 89 94 99 105 112

1 : Buchnera aphidicola str Cc
2 : Buchnera aphidicola str APS
3 : Buchnera aphidicola str Bp
4 : Buchnera aphidicola str Sg
5 : Buchnera aphidicola str Tuc7
6 : Buchnera aphidicola str 5A
7 : Escherichia coli BL21(DE3 AU)
8 : Escherichia coli BL21(DE3 DOE)
9 : Escherichia coli BL21(DE3 Korea)
10 : Escherichia coli B str REL606
11 : Escherichia coli K–12 MG1655
12 : Escherichia coli K–12 W3110
13 : Escherichia coli K–12 DH10B
14 : Escherichia coli DH1
15 : Escherichia coli BW2952
16 : Escherichia coli ATCC 8739
17 : Escherichia_coli HS
18 : Escherichia coli E24377A
19 : Escherichia coli 536
20 : Escherichia coli UTI89
21 : Escherichia coli APEC O1
22 : Escherichia coli SE11
23 : Escherichia coli SE15
24 : Escherichia coli SMS–3–5
25 : Escherichia coli 042
26 : Escherichia coli CFT073
27 : Escherichia coli IAI1
28 : Escherichia coli S88
29 : Escherichia coli ED1a
30 : Escherichia coli IAI39
31 : Escherichia coli 55989
32 : Escherichia coli UMN026
33 : Escherichia coli O26:H11 str. 11368
34 : Escherichia coli O103:H2 str. 12009
35 : Escherichia coli O111:H– str. 11128
36 : Escherichia coli O127:H6 str. E2348/69
37 : Escherichia coli O157:H7 str. Sakai
38 : Escherichia coli O157:H7 str. EDL933
39 : Escherichia coli O157:H7 str EC4115
40 : Escherichia coli O157:H7 str. TW14359
41 : Escherichia fergusonii ATCC 35469
42 : Yersinia enterocolitica subsp enterocolitica 8081
43 : Yersinia pestis Angola
44 : Yersinia pestis Antiqua
45 : Yersinia pestis CO92
46 : Yersinia pestis D106004
47 : Yersinia pestis D182038
48 : Yersinia pestis KIM 10
49 : Yersinia pestis Nepal516
50 : Yersinia pestis Pestoides F
51 : Yersinia pestis biovar Microtus str 91001
52 : Yersinia pseudotuberculosis IP 31758
53 : Yersinia pseudotuberculosis IP 32953
54 : Yersinia pseudotuberculosis PB1/+
55 : Yersinia pseudotuberculosis YPIII
56 : Shigella boydii CDC 3083–94
57 : Shigella boydii Sb227
58 : Shigella dysenteriae Sd197
59 : Shigella flexneri 2002017
60 : Shigella flexneri 2a str 2457T
61 : Shigella flexneri 2a str 301
62 : Shigella flexneri 5 str. 8401
63 : Shigella sonnei Ss046
64 : Erwinia amylovora ATCC 49946
65 : Erwinia pyrifoliae DSM 12163
66 : Erwinia pyrifoliae Ep1/96
67 : Erwinia tasmaniensis Et1/99
68 : Salmonella enterica subsp. arizonae serovar 62:z4 z23:––
69 : Salmonella enterica serovar Paratyphi A str ATCC 9150
70 : Salmonella enterica serovar Newport str. SL254
71 : Salmonella enterica serovar Schwarzengrund str CVM19633
72 : Salmonella enterica serovar Dublin str CT_02021853
73 : Salmonella enterica serovar Heidelberg str SL476
74 : Salmonella enterica serovar Agona str SL483
75 : Salmonella enterica serovar Paratyphi C strain RKS4594
76 : Salmonella enterica serovar Typhi str CT18
77 : Salmonella enterica serovar Typhimurium str LT2
78 : Salmonella enterica serovar Paratyphi B str SPB7
79 : Salmonella enterica serovar Enteritidis str P125109
80 : Salmonella enterica serovar Gallinarum str 287/91
81 : Salmonella enterica serovar Paratyphi A str AKU_12601
82 : Salmonella enterica serovar Typhimurium str 14028S
83 : Salmonella enterica serovar Typhi str Ty2
84 : Salmonella enterica serovar Typhimurium
85 : Salmonella enterica serovar Choleraesuis str SC–B67
86 : Klebsiella pneumoniae subsp pneumoniae MGH 78578
87 : Klebsiella pneumoniae NTUH–K2044
88 : Klebsiella pneumoniae 342
89 : Klebsiella variicola At–22
90 : Proteus mirabilis HI4320
91 : Proteus mirabilis ATCC 29906
92 : Proteus penneri ATCC 35198
93 : Citrobacter koseri ATCC BAA–895
94 : Citrobacter rodentium ICC168
95 : Citrobacter youngae ATCC 29220
96 : Cronobacter turicensis
97 : Cronobacter sakazakii ATCC BAA–894
98 : Candidatus Blochmannia floridanus
99 : Candidatus Blochmannia pennsylvanicus str BPEN
100 : Candidatus Hamiltonella defensa 5AT
101 : Candidatus Pelagibacter ubique HTCC1002
102 : Dickeya zeae Ech1591
103 : Dickeya dadantii Ech703
104 : Dickeya dadantii Ech586
105 : Edwardsiella tarda EIB202
106 : Edwardsiella ictaluri 93–146
107 : Enterobacter cancerogenus ATCC 35316
108 : Enterobacter sp 638
109 : Pectobacterium carotovorum subsp carotovorum PC1
110 : Pectobacterium wasabiae WPP163
111 : Pectobacterium atrosepticum SCRI1043
112 : Photorhabdus asymbiotica
113 : Photorhabdus luminescens subsp laumondii TTO1
114 : Sodalis glossinidius str morsitans
115 : Serratia proteamaculans 568
116 : Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis
117 : Xenorhabdus bovienii SS–2004



| | strains | unique genes |
|---|---|---|
| *Buchnera* | 6 | 14 |
| *Escherichia + Shigella* | 43 | 8 |
| *Yersinia* | 14 | 68 |
| *Erwinia* | 4 | 201 |
| *Salmonella* | 18 | 59 |
| *Klebsiella* | 4 | 222 |
| *Proteus* | 3 | 292 |
| *Citrobacter* | 3 | 9 |
| *Cronobacter* | 2 | 244 |
| *Dickeya* | 3 | 150 |
| *Edwardsiella* | 2 | 314 |
| *Enterobacter* | 2 | 26 |
| *Pectobacterium* | 3 | 155 |
| *Protorhabdus* | 2 | 402 |
| *Sodalis* | 1 | 390 |
| *Serratia* | 1 | 763 |
| *Wiggiesworthia* | 1 | 212 |
| *Xenorhabdus* | 1 | 1095 |

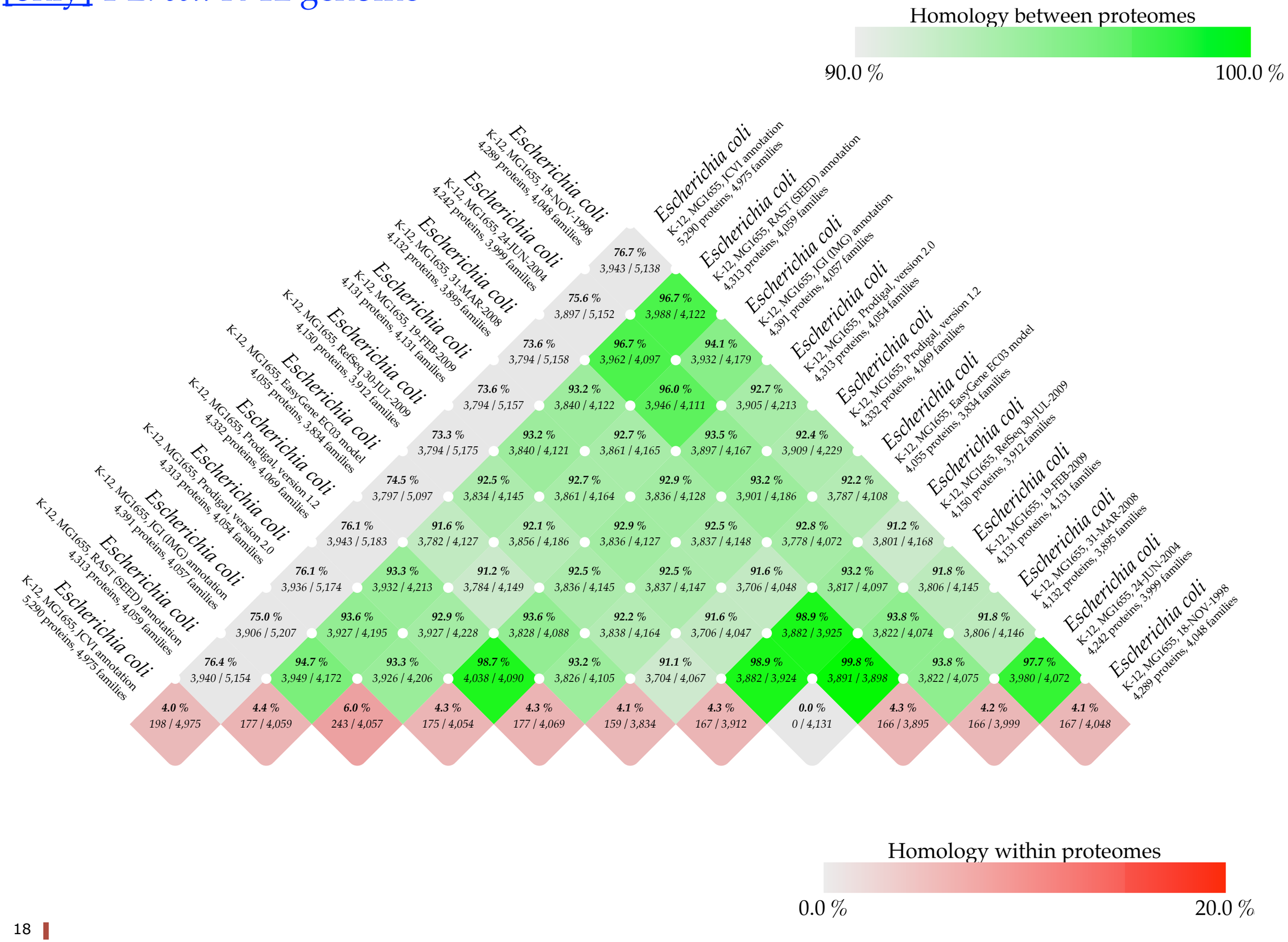
CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

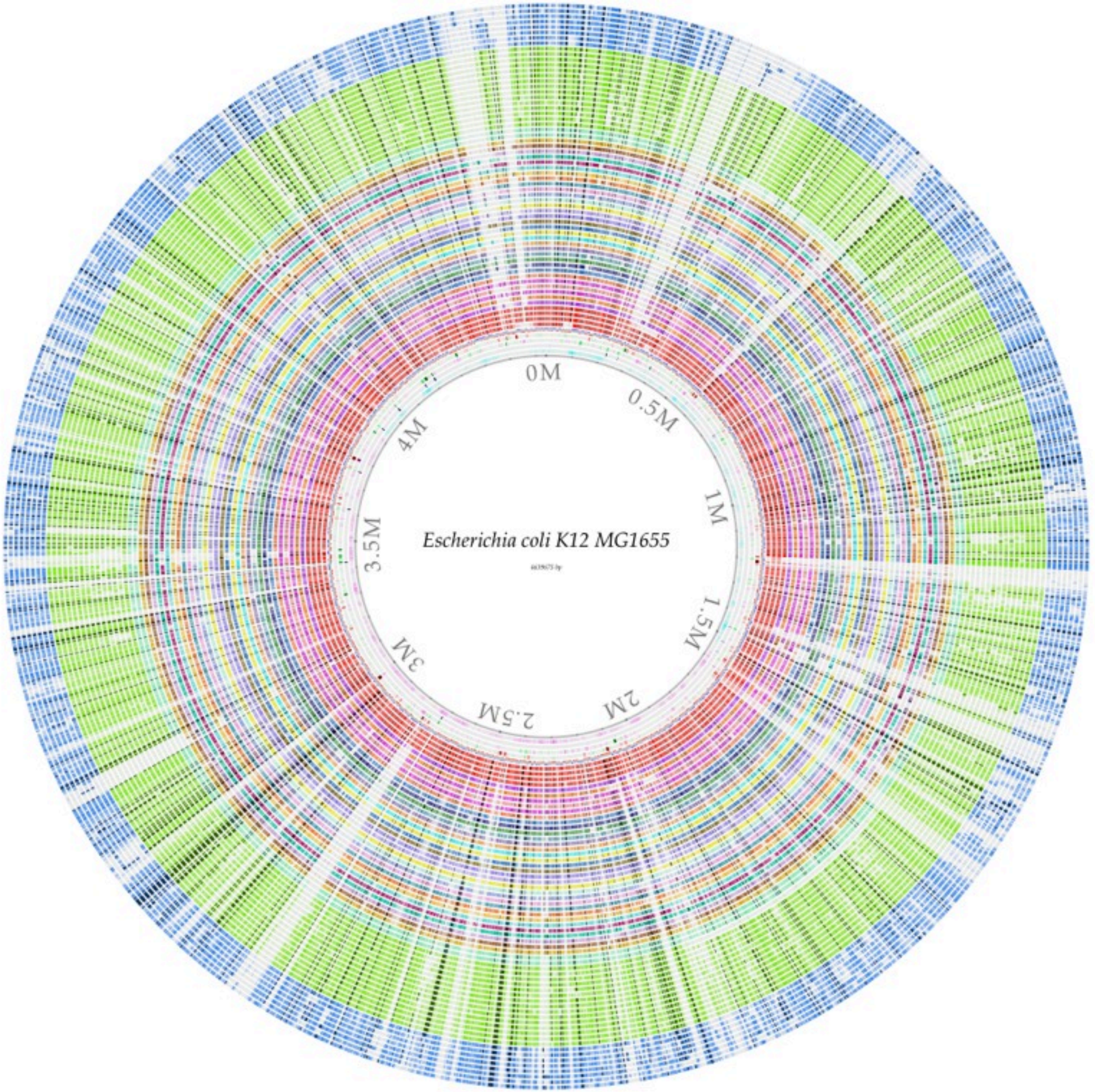
Center for Biological Sequence analysis

Department of Systems Biology, Technical University of Denmark

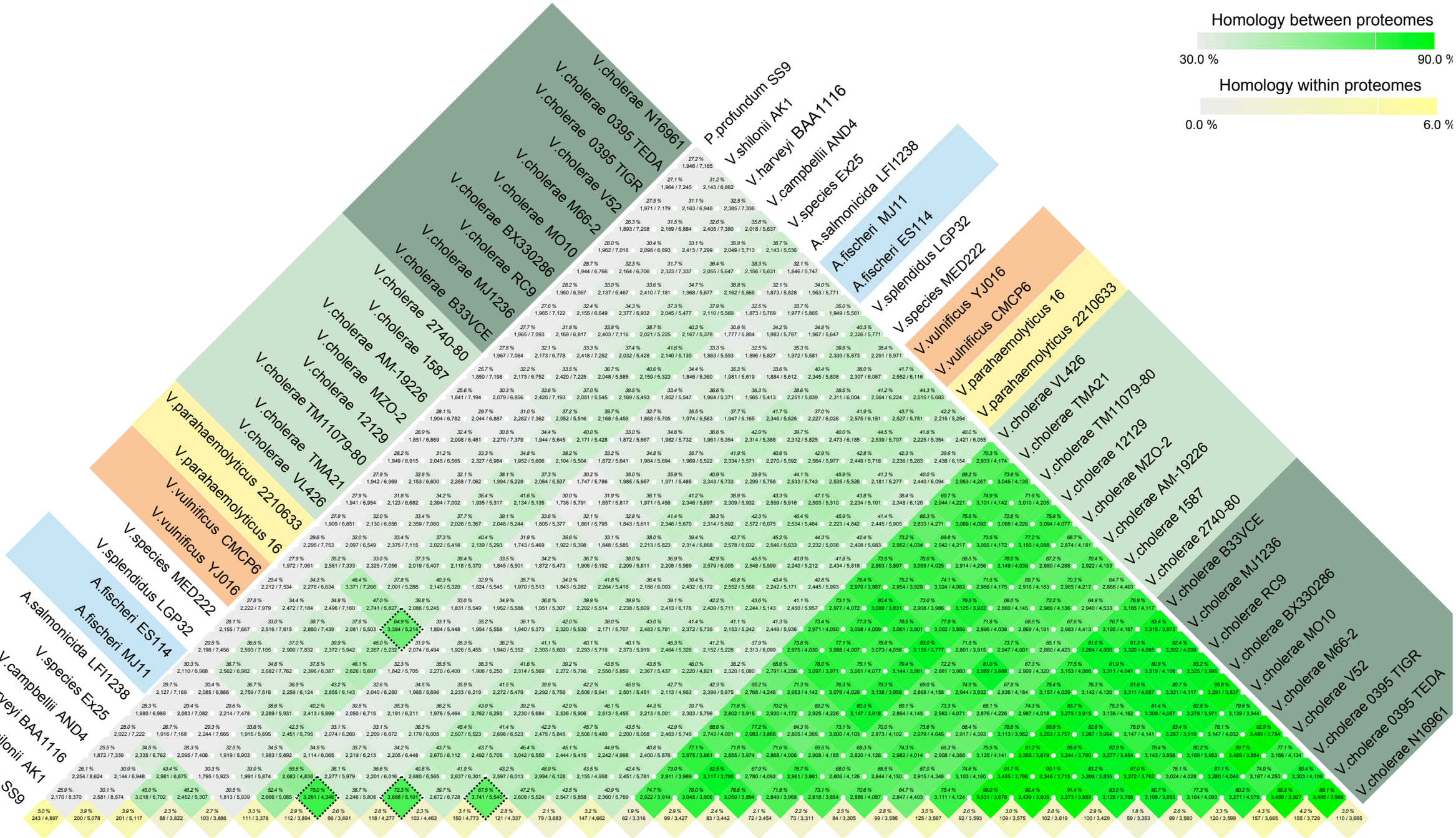
DTU

# [only] 1 *E. coli* K-12 genome

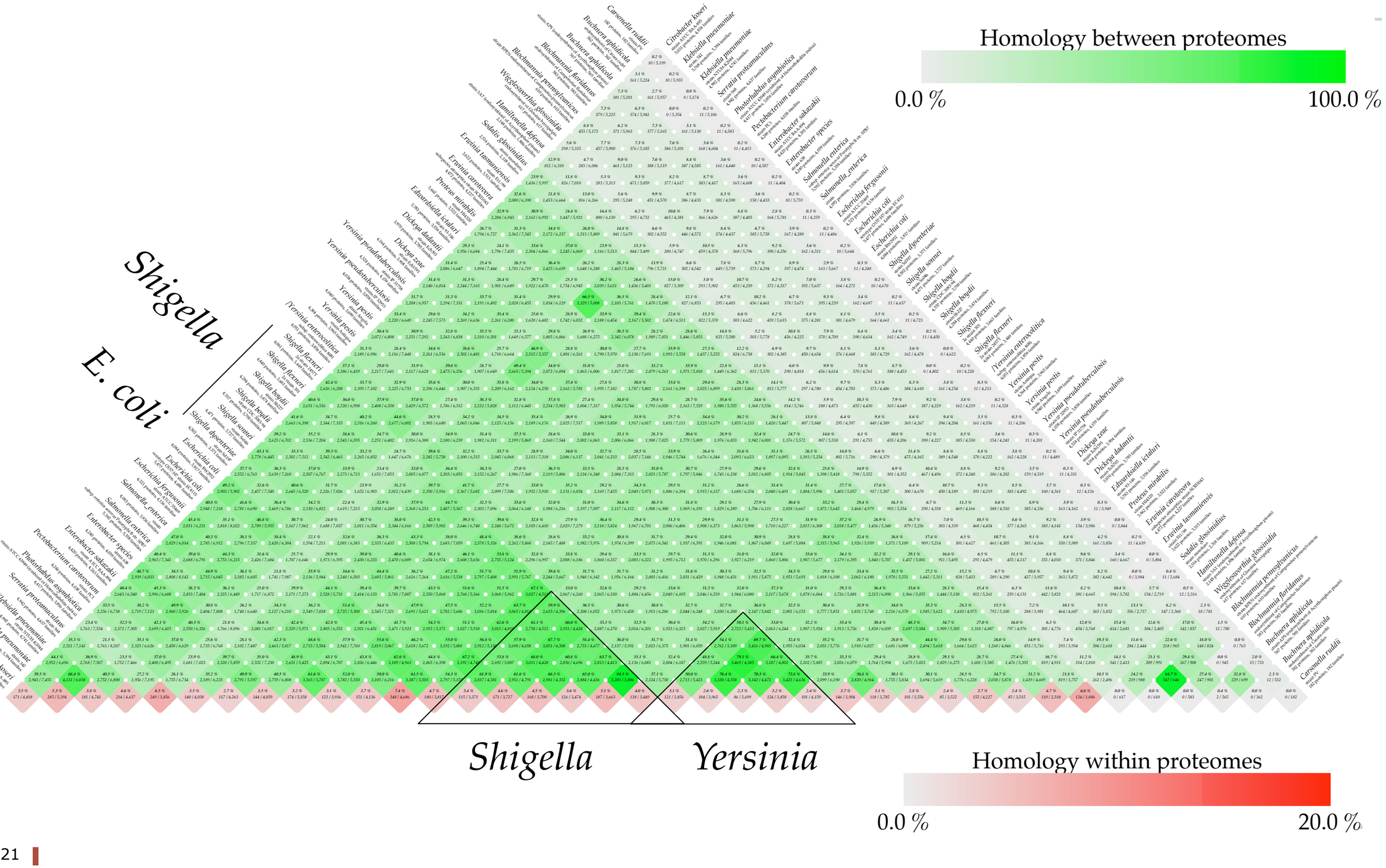



0M
0.5M
1M
1.5M
2M
2.5M
3M
3.5M
4M
Escherichia coli K12 MG1655

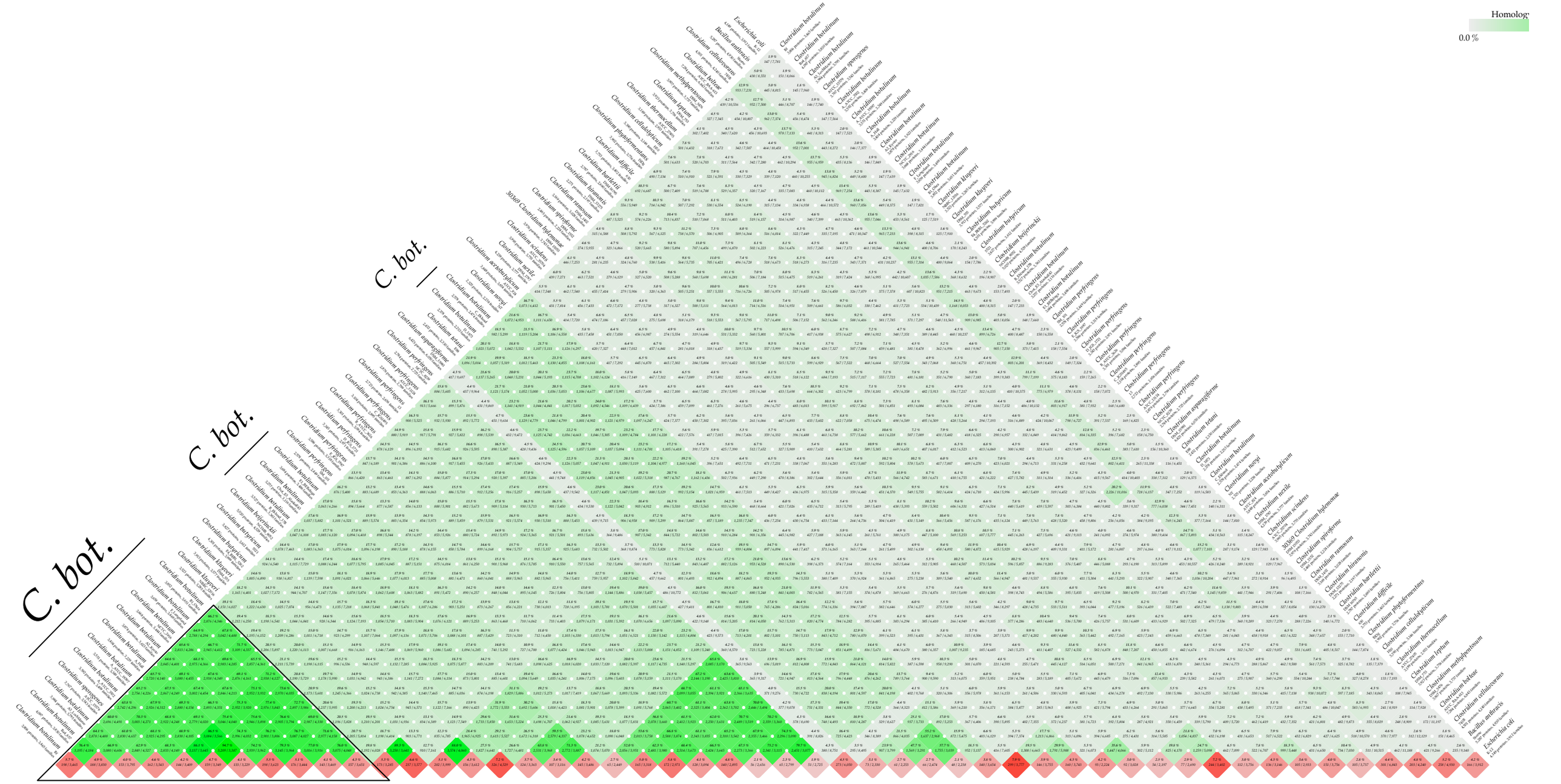
34 *Vibrio* genomes



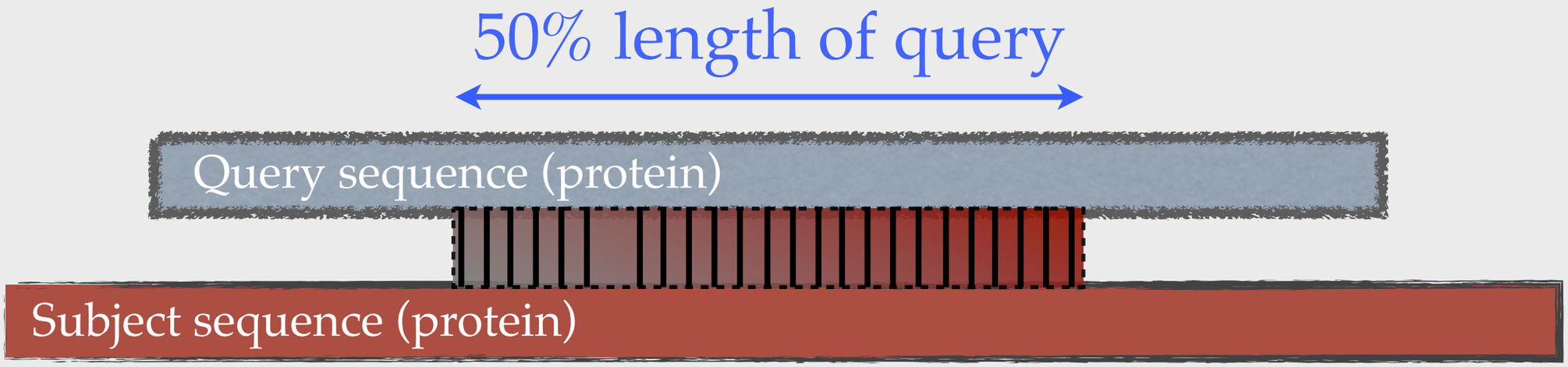
40 *enteric* genomes



60 *Clostridium* genomes



# When are two proteins the same??

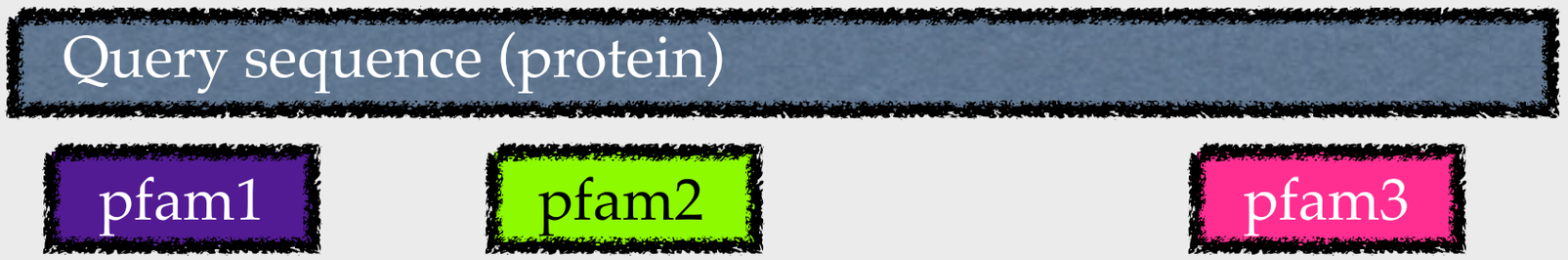



50% identity of match

0 proteins conserved in a thousand genomes

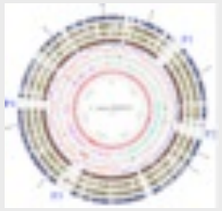
scales poorly

Kristoffer




scales linearly

~150 protein families conserved in a thousand genomes





# Two questions for discussion:

note: I don't know the answer to these!

## 1. When are two proteins the same?

## 2. How to visualize this, on a large scale?

What are some good visualization method would scale to allow comparison of 10,000 or a million bacterial genomes?

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

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.