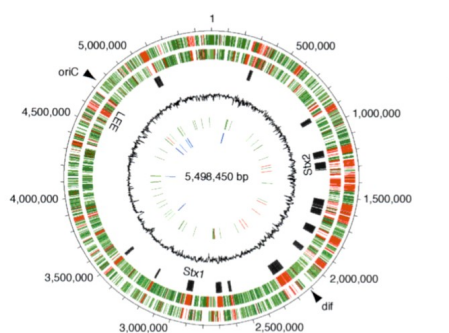


Multiple Genome Alignment for Identifying the Genomic Core Among Moderately Related Microbial Genomes

Ikuo Uchiyama
National Institute for Basic Biology
Okazaki, Japan

Species genome concept Core genome and pan genome

Escherichia coli O157

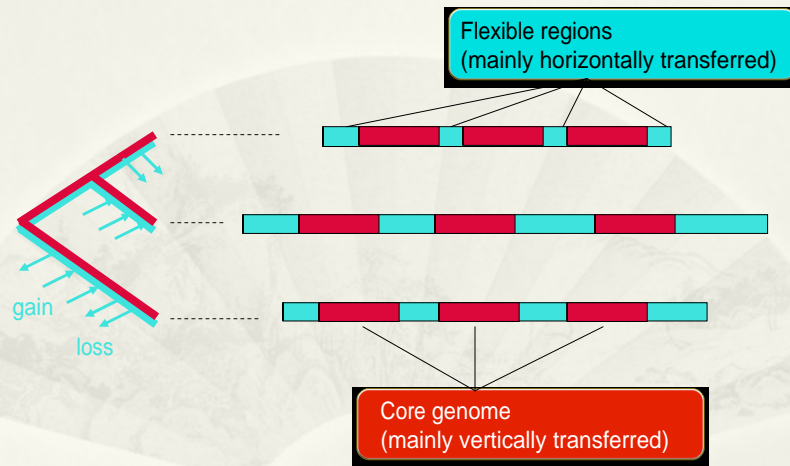


■ Shared with *E. coli* K12 (75%)
■ Unique to *E. coli* O157 (25%)

Escherichia coli

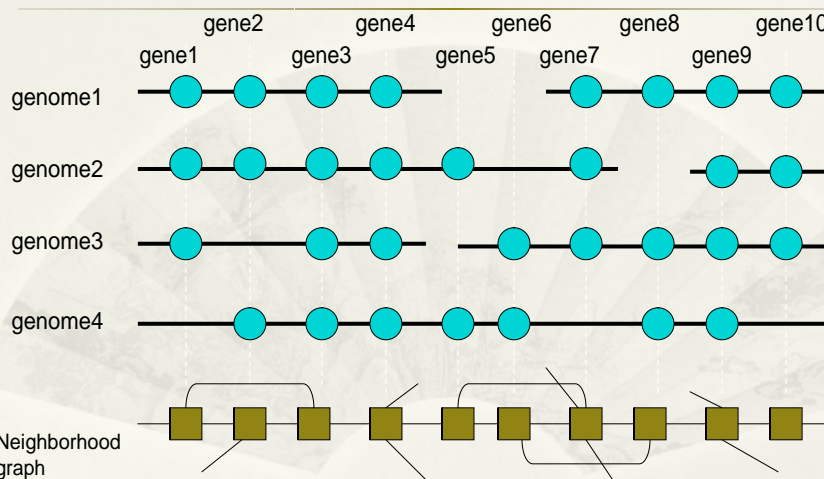


Core genome concept



Purpose: to identify the core structure of moderately related genomes using the information of synteny conservation

Genome alignment based on gene order conservation

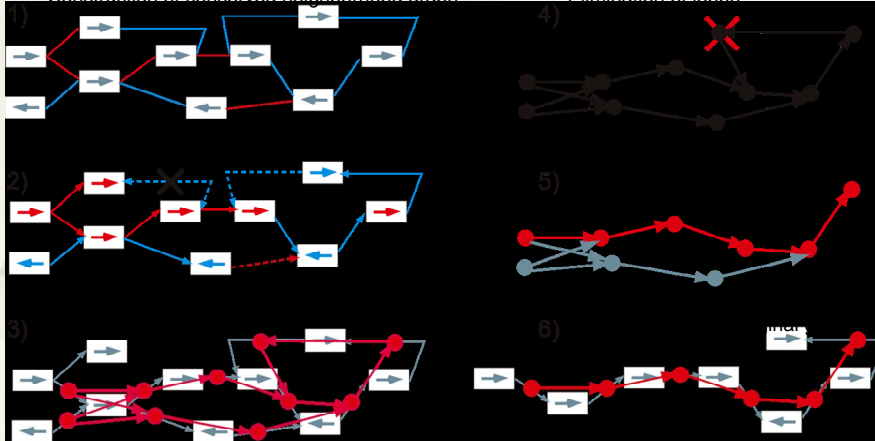


Each edge represents a neighborhood pair of genes that are located within 20 genes in at least 50% of the genomes

CoreAligner algorithm

Construction of conserved neighborhood graph

Elimination of loops



Dataset


* *Bacillaceae*

- * *Bacillus subtilis*¹
- * *Bacillus licheniformis*¹
- * *Bacillus halodurans*²
- * *Bacillus clausii*²
- * *Bacillus anthracis* Ames³
- * *Bacillus cereus* 14579³
- * *Geobacillus kaustophilus*
- * *Oceanobacillus iheyensis*
(outgroup)
- * *Staphylococcus aureus* N315

* *Enterobacteriaceae*

- * *Escherichia coli* K-12 MG1655⁴
- * *Salmonella enterica* CT18⁴
- * *Enterobacter* sp. 638⁴
- * *Erwinia carotovora*
- * *Photobacterium luminescens*
- * *Sodalis glossinidius*
- * *Serratia proteamaculans*
- * *Yersinia pestis* CO92
(outgroup)
- * *Vibrio cholerae*

MBGD (Microbial genome database for comparative analysis)



Microbial Genome Database for Comparative Analysis
National Institute for Basic Biology, National Institutes of Natural Sciences

Introduction
Overview and quick tour

Create/View Ortholog Table
View the default ortholog table or create new grouping on the fly

Enter My MBGD Mode
Add your own genome sequences to MBGD and create your own cluster.

List of Cluster Tables
Cluster tables that you have made

Analysis
[Pairwise comparisons](#)
 Compare closely related genomes using the CGAT interface
[Quick Search](#)

Mode: AND OR

Advanced Search
[Homology Search](#)
[List of Function Categories](#)
[List of Gene Names](#)
[Related Programs](#)

DomClust
The clustering program used for ortholog grouping in MBGD

CGAT

MBGD is a database for comparative analysis of completely sequenced microbial genomes, the number of which is now growing rapidly. The aim of MBGD is to facilitate comparative genomics from various points of view such as ortholog identification, paralog clustering, motif analysis and gene order comparison.

References: *Nucleic Acids Res.* 31:58-62 (2003) / *Viecherle Acids Res.* 35:D343-D346 (2007)

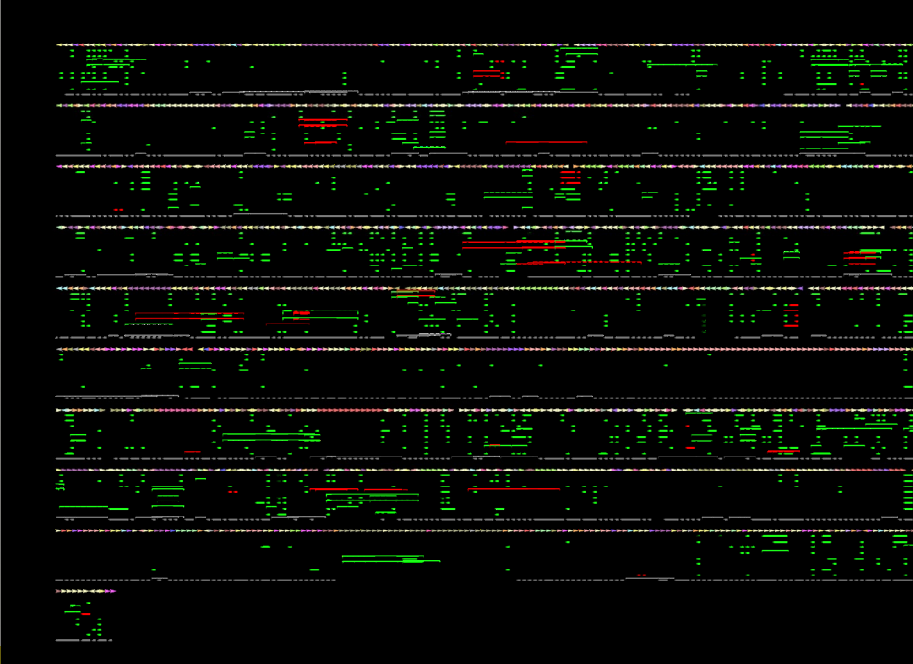
Complete genome sequences [\[Data Sources\]](#)
 [\[Set Default\]](#)

Currently selected organisms (248) are highlighted in red. Please press "Release" button when you return here by "Back" button.

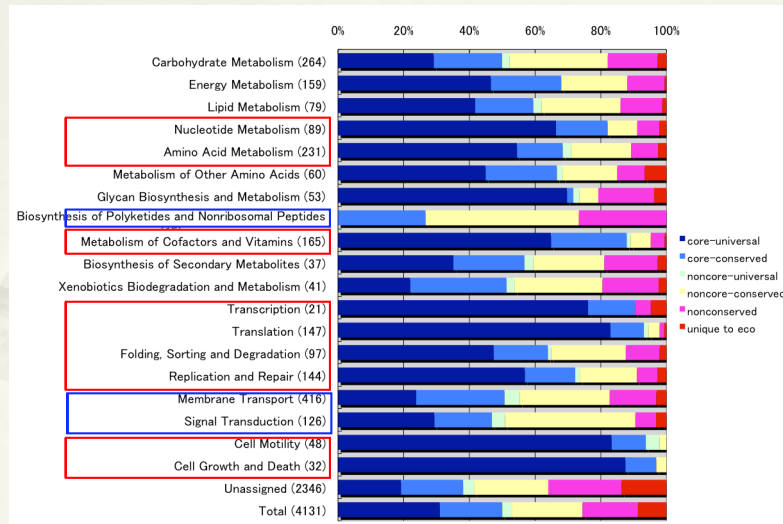
Haemophilus discrevi 35900HP		
Bacteria (548) <i>Gloeobacter</i> <i>Anabaena</i> <i>Nostoc</i> <i>Trichodesmium</i> <i>Prochlorococcus</i> (11) <i>Cyanobacterium</i> <i>Corynebacterium</i> (6) <i>Mycobacterium</i> (14) <i>Nocardia</i> <i>Rhodococcus</i> <i>Acidobacteria</i> <i>Vacterdium</i> <i>Rhodococcus</i> <i>Acidobacterium</i> <i>Frankia</i> (5) <i>Streptomyces</i> <i>Kinococcus</i> <i>Froghyrum</i> (2) <i>Valoniopsis</i> <i>Leifsonia</i> <i>Arthrobacter</i> (2) <i>CGAT</i>	<i>Actinobacteria</i> <i>Thermus</i> (2) <i>Dinococcus</i> (2) <i>Thermus</i> (2) <i>Acidobacteria</i> <i>Acidobacteria</i> <i>Sulfolobus</i> <i>Rhodospirillum</i> <i>Firmicutes</i> <i>Bacillus</i> (6) <i>Rhizobium</i> (2) <i>Sporichthidium</i> (2) <i>Geobacillus</i> (2) <i>Glaucidium</i> (2) <i>Clavibacter</i> <i>Lerifsonia</i> <i>Arthrobacter</i> (2)	<i>Nitrospira</i> (2) <i>Rhodospirillum</i> (5) <i>Bifidobacterium</i> (5) <i>Chlorobacterium</i> <i>Mesorhizobium</i> (2) <i>Halorubrum</i> <i>Parabacterium</i> <i>Agrobacterium</i> (2) <i>Rhizobium</i> (2) <i>Streptococcus</i> (2) <i>Streptococcus</i> (2) <i>Streptococcus</i> (2) <i>Azobacterium</i> (2) <i>Dicellaenomonas</i> <i>Aeromonas</i> (2) <i>Halorubrum</i> <i>Xanthobacter</i> <i>Marinobacter</i> <i>Saccharophagus</i> <i>Calohellia</i> <i>Strombolium</i>

Haemophilus (5)
Histophilus
Mannheimia
Pasteurella
Actinobacterium (2)
Psychrobacter (3)
Psychrobacter (3)
Psychrobacter (3)
Psychrobacter (3)
Acropyrum
Ignavococcus
Staphylothermus
Hyperthermus
Micillophaga
Sulfolobus (3)
Thermoplasma
Thermoplasma
Thermoplasma
Thermoplasma
Pyrobaculum (4)
Euryarchaeota
Archaeoglobus

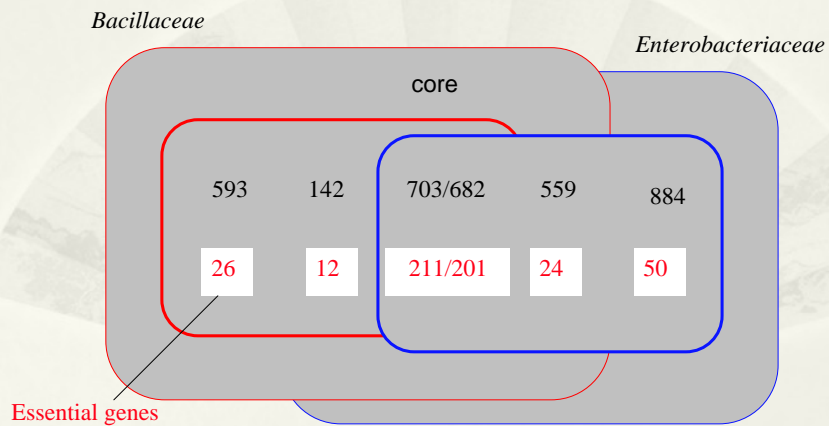
Bacillaceae core alignment



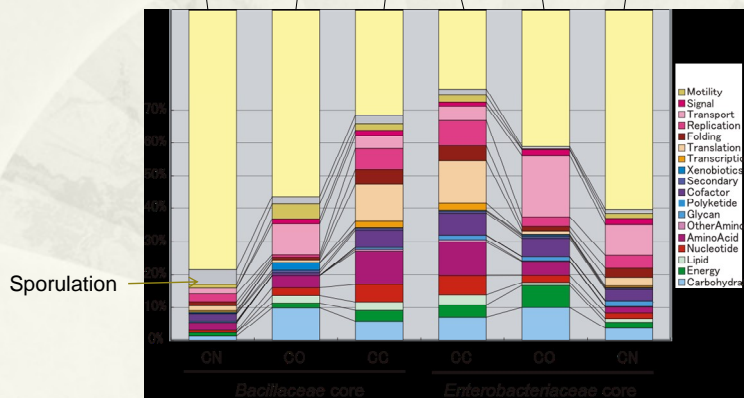
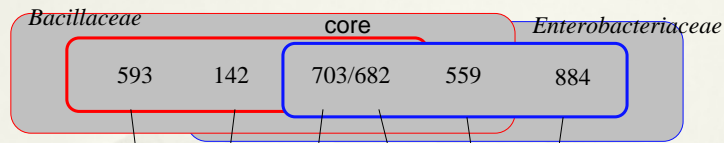
Ratio of core genes in each functional category (*E. coli*)



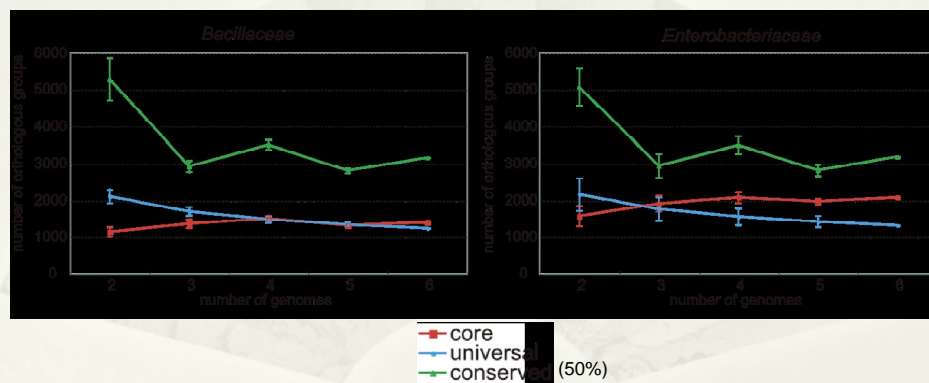
Comparison between the core structures of *Bacillaceae* and *Enterobacteriaceae*



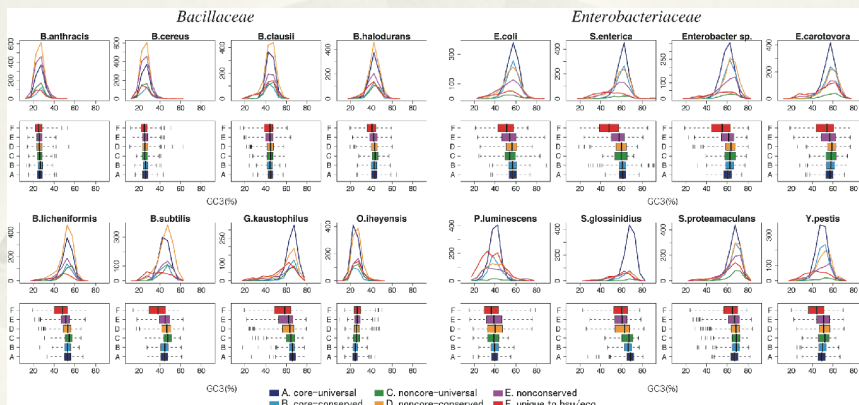
Functional categories of the common core set



Robustness test

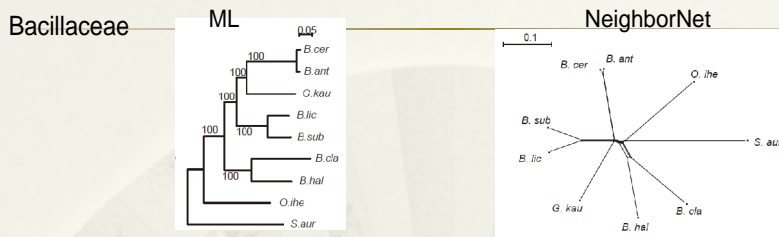


G+C content of the third codon positions



Congruence of phylogenetic trees

Phylogenetic tree of concatenated core genes



Genes in deviated topology (SH-test, 5% significance level)

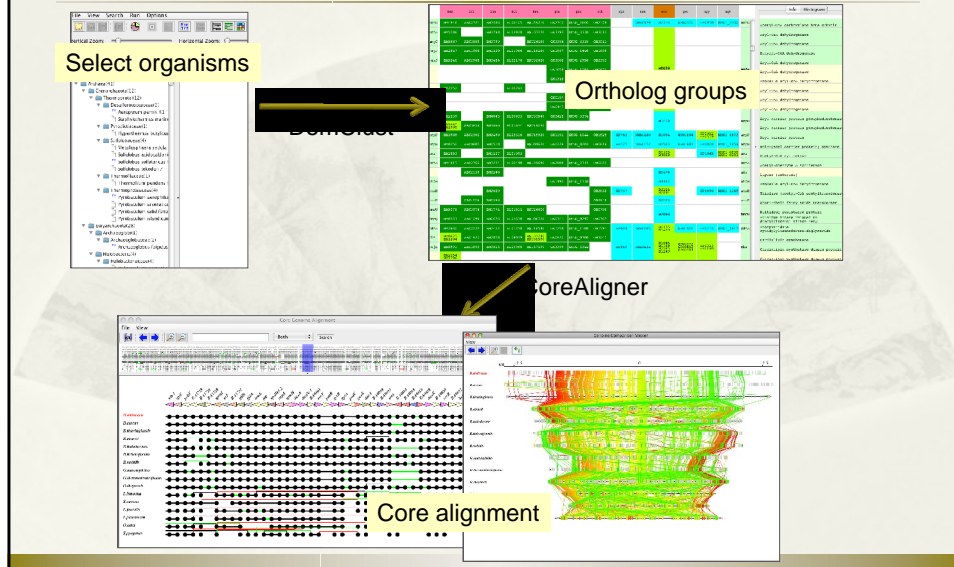
N	Core		Non-core		p-value (core<noncore)
	Rejected	Total	Rejected	Total	
4	3 (3.5)	85	32 (14.0)	228	0.004573
5	11 (6.2)	176	38 (21.6)	176	2.170e-5
6	17 (5.4)	314	34 (29.6)	115	1.899e-10
7	32 (4.3)	739	9 (15.5)	58	0.001705
Total	63 (4.8)	1314	113 (19.6)	577	<2.2e-16

Core genes have more consistent phylogeny than non-core genes

RECOG

(Research Environment for Comparative Genomics)

<http://mbgd.genome.ad.jp/RECOG>



Conclusion

- * We have developed a method (CoreAligner) to extract syntenically conserved regions to construct the core genome alignment.
- * Extracted core set contains most of the functionally important genes including the essential genes.
- * CoreAligner can identify more robust core set than the method based only on gene conservation.
- * Indigenusness of the extrated core is demonstrated by G+C content homogeneity and phylogenetic tree congruence.



The number of deleted core genes

Bacillaceae (1438 core)			
	Conserved core	Deleted core	Total CDS
<i>Bacillus subtilis</i>	1362	76	4105
<i>Bacillus licheniformis</i>	1387	51	4152
<i>Bacillus halodurans</i>	1358	80	4066
<i>Bacillus clausii</i>	1296	142	4096
<i>Bacillus anthracis</i>	1300	138	5311
<i>Bacillus cereus</i>	1297	141	5234
<i>Geobacillus kaustophilus</i>	1356	82	3498
<i>Oceanobacillus iheyensis</i>	1265	173	3500
<i>Staphylococcus aureus</i>	883	555	2588

The number of deleted core genes

<i>Enterobacteriaceae</i> (2125 core)			
	Conserved core	Deleted core	Total CDS
<i>Escherichia coli</i>	2048	77	4131
<i>Salmonella enterica</i>	2002	123	4395
<i>Enterobacter sp. 638</i>	2057	68	4115
<i>Erwinia carotovora</i>	2006	119	4472
<i>Photobacterium luminescens</i>	1769	356	4683
<i>Sodalis glossinidius</i>	1507	618	2432
<i>Serratia proteamaculans</i>	2110	15	4891
<i>Yersinia pestis</i>	2015	110	3885
<i>Vibrio cholerae</i>	1631	494	3835