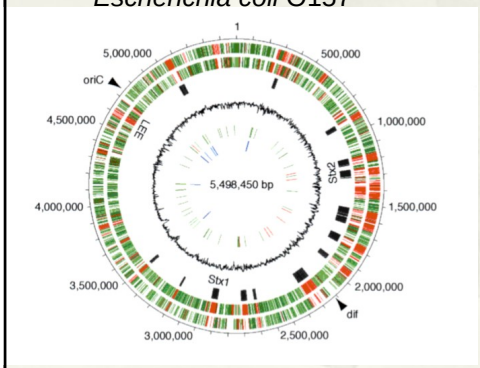
Multiple Genome Alignment for Identifying the Genomic Core Among Moderately Related Microbial Genomes

Ikuo Uchiyama
National Institute for Basic Biology
Okazaki, Japan

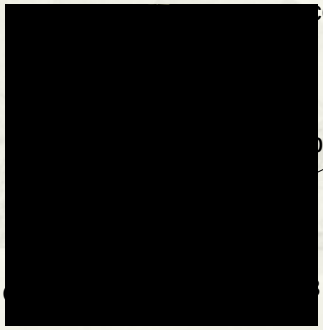
Species genome concept
Core genome and pan genome

Escherichia coli O157



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

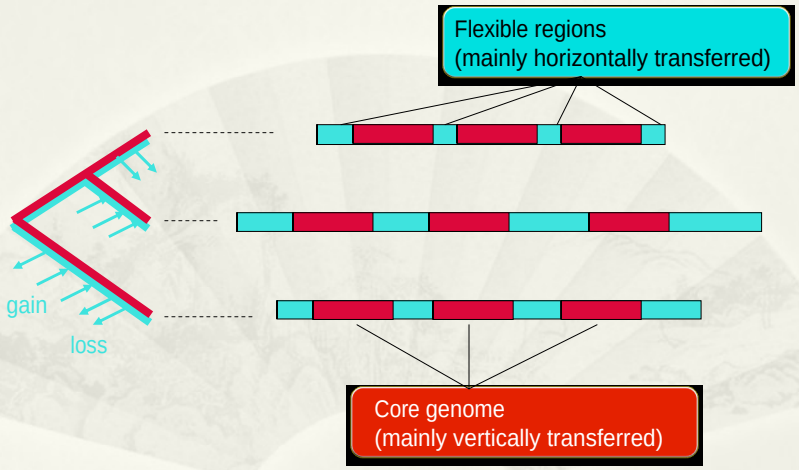
Escherichia coli



core-genome

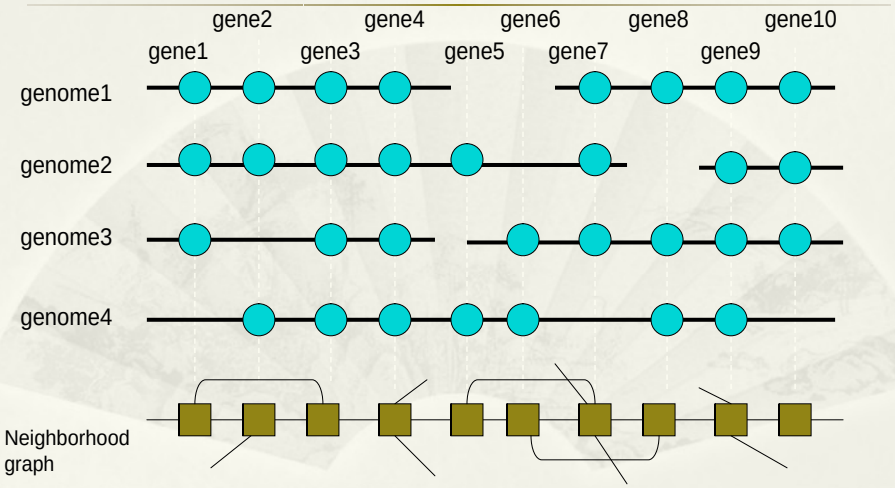
pan-genome

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

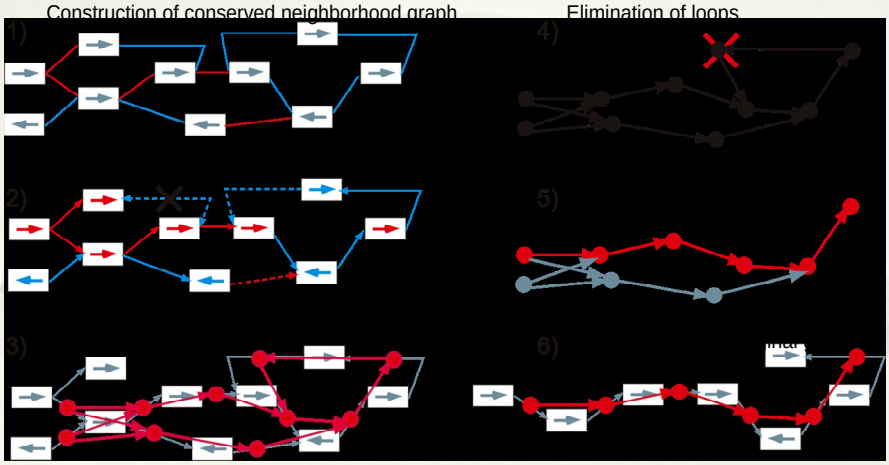


Each edge represents a neighborhood pair of genes that are located within 20 genes in at least 50% of the genomes

CoreAligner algorithm

Construction of conserved neighborhood graph

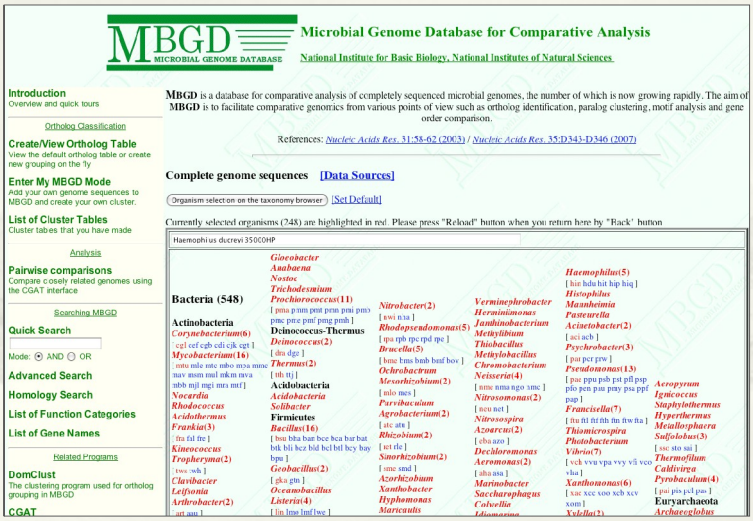
Elimination of loops



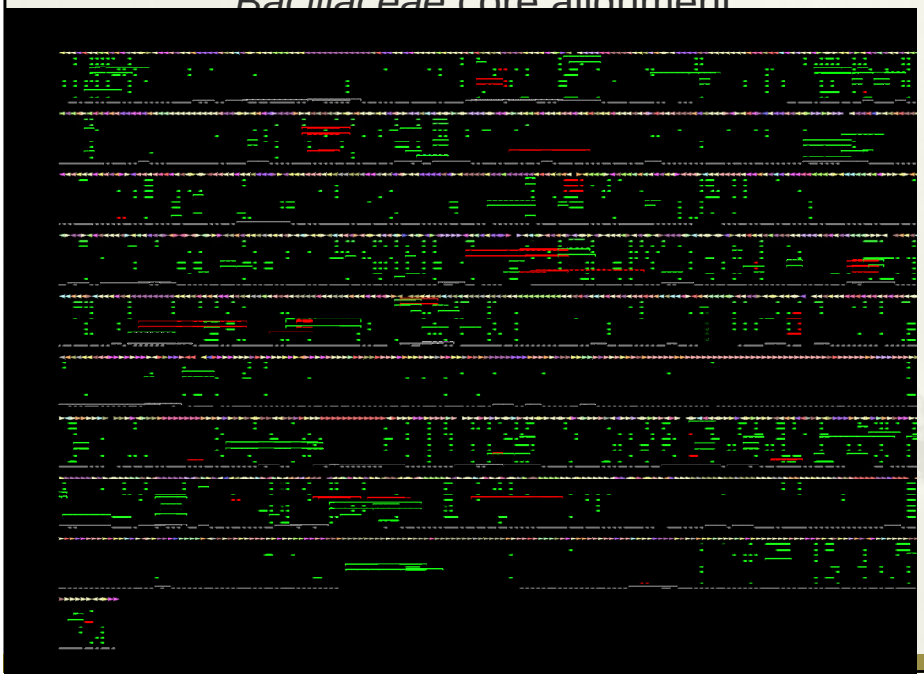
Dataset

- *Bacillaceae*
  - *Bacillus subtilis* [1]
  - *Bacillus licheniformis* [1]
  - *Bacillus halodurans* [2]
  - *Bacillus clausii* [2]
  - *Bacillus anthrasis* Ames [3]
  - *Bacills cereus* 14579 [3]
  - *Geobacillus kaustophilus*
  - *Oceanobacillus iheyensis* (outgroup)
  - *Staphylococcus aureus* N315
- *Enterobacteriaceae*
  - *Escherichia coli* K-12 MG1655 [4]
  - *Salmonella enterica* CT18 [4]
  - *Enterobacter* sp. 638 [4]
  - *Erwinia carotovora*
  - *Photorhabdus luminescens*
  - *Sodalis glossinidius*
  - *Serratia proteamaculans*
  - *Yersinia pestis* CO92 (outgroup)
  - *Viblio cholerae*

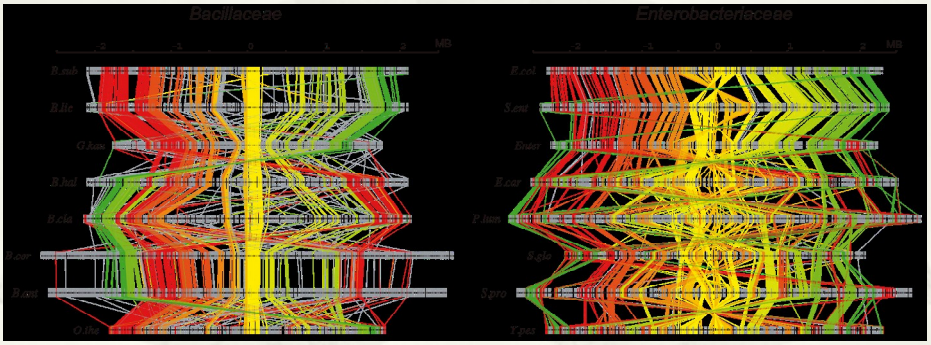
MBGD (Microbial genome database for comparative analysis)



Bacillaceae core alignment



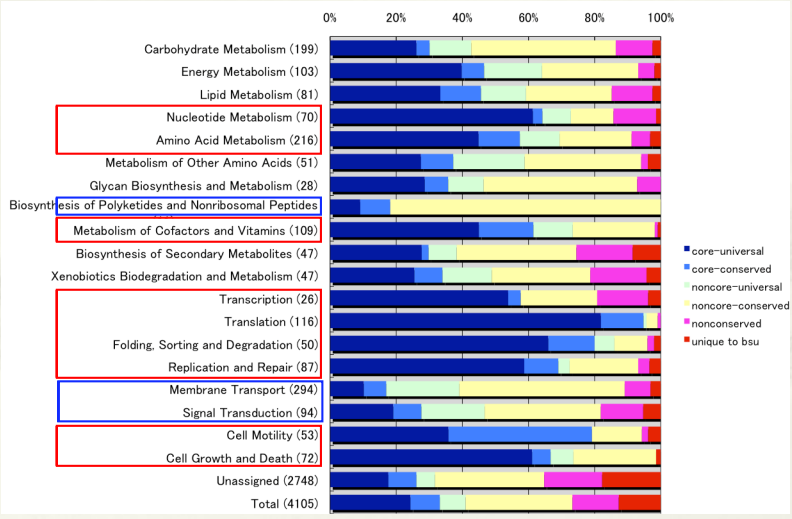
Chromosomal location of core genes



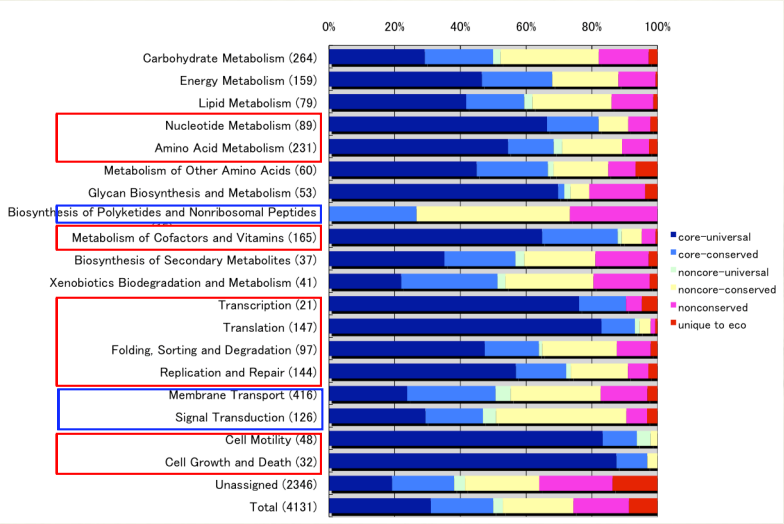
1438 orthologous groups

2125 orthologous groups

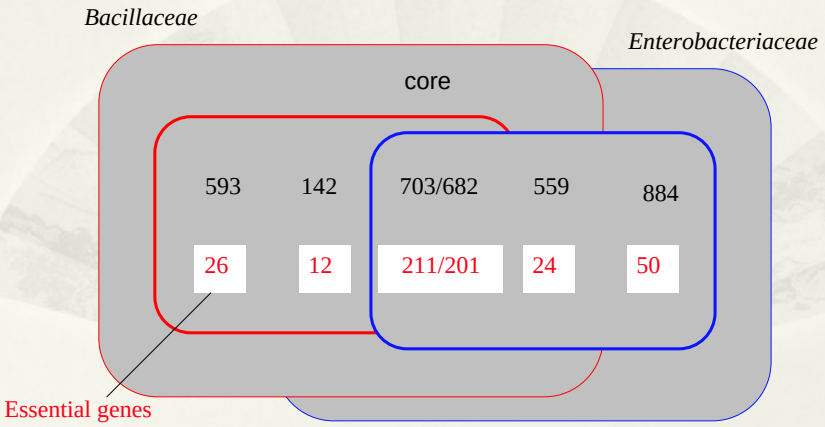
Ratio of core genes in each functional category (*B. subtilis*)



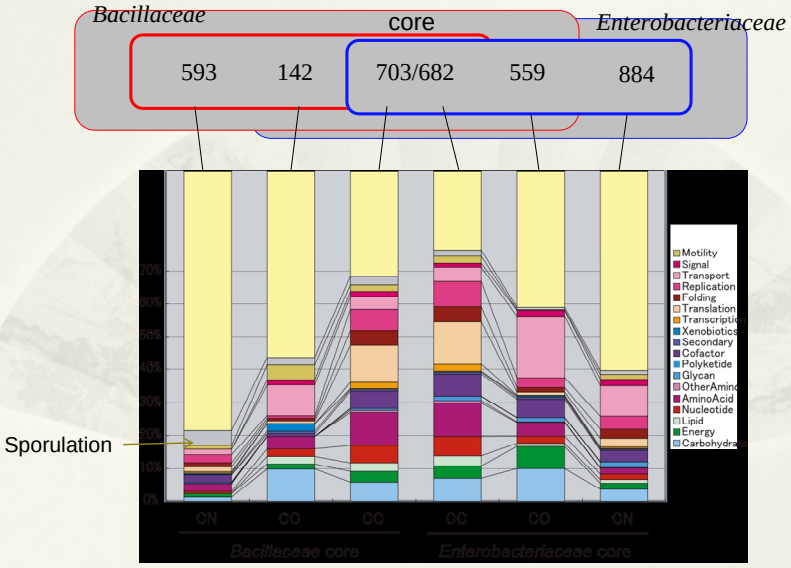
Ratio of core genes in each functional category (*E. coli*)



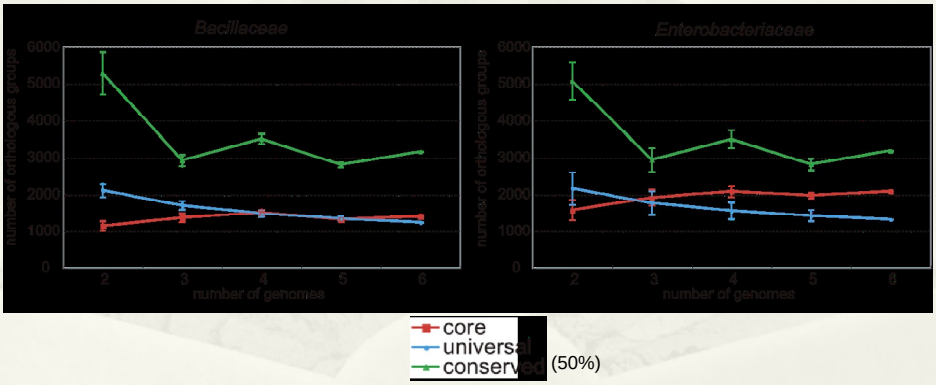
Comparison between the core structures of *Bacillaceae* and *Enterobacteriaceae*



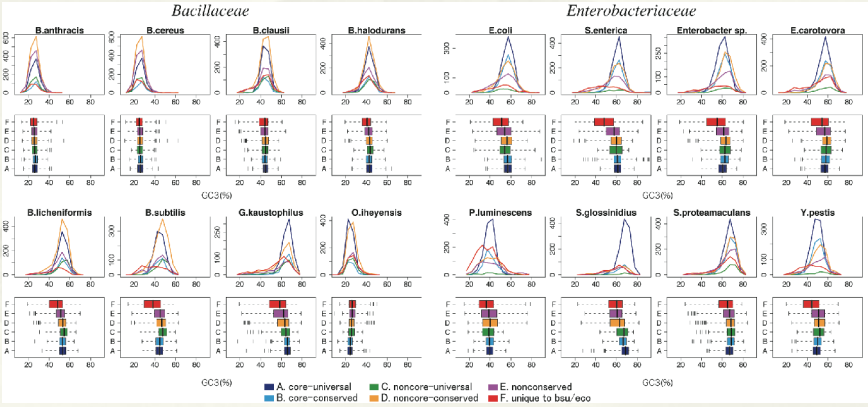
Functional categories of the common core set



Robustness test



G+C content of the third codon positions

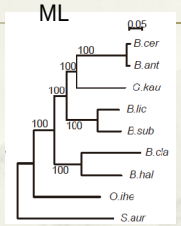


Congruence of phylogenetic trees

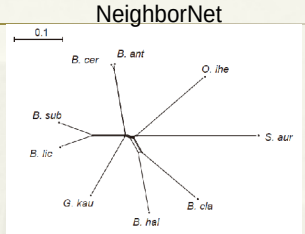
Phylogenetic tree of concatenated core genes

Bacillaceae

ML



NeighborNet



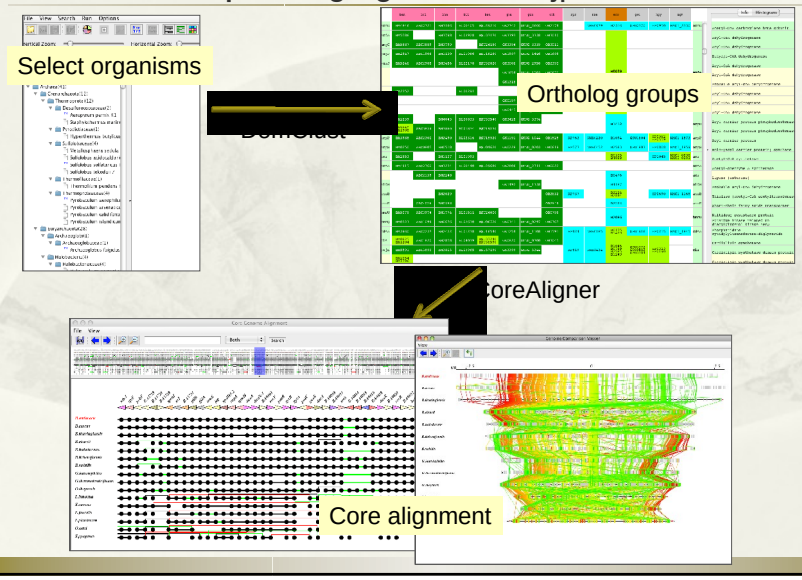
Genes in deviated topology (SH-test, 5% significance level)

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.
- Indigenousness 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
Bacillus subtilis	1362	76	4105
Bacillus licheniformis	1387	51	4152
Bacillus halodurans	1358	80	4066
Bacillus clausii	1296	142	4096
Bacillus anthracis	1300	138	5311
Bacillus cereus	1297	141	5234
Geobacillus kaustophilus	1356	82	3498
Oceanobacillus iheyensis	1265	173	3500
Staphylococcus aureus	883	555	2588

The number of deleted core genes

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