



Evolution of Bacterial Genomes

Mikhail Gelfand
Skoltech, IITP

Evolution of Complexity...
Yerevan, 30 VI 2022

Levels of genome evolution

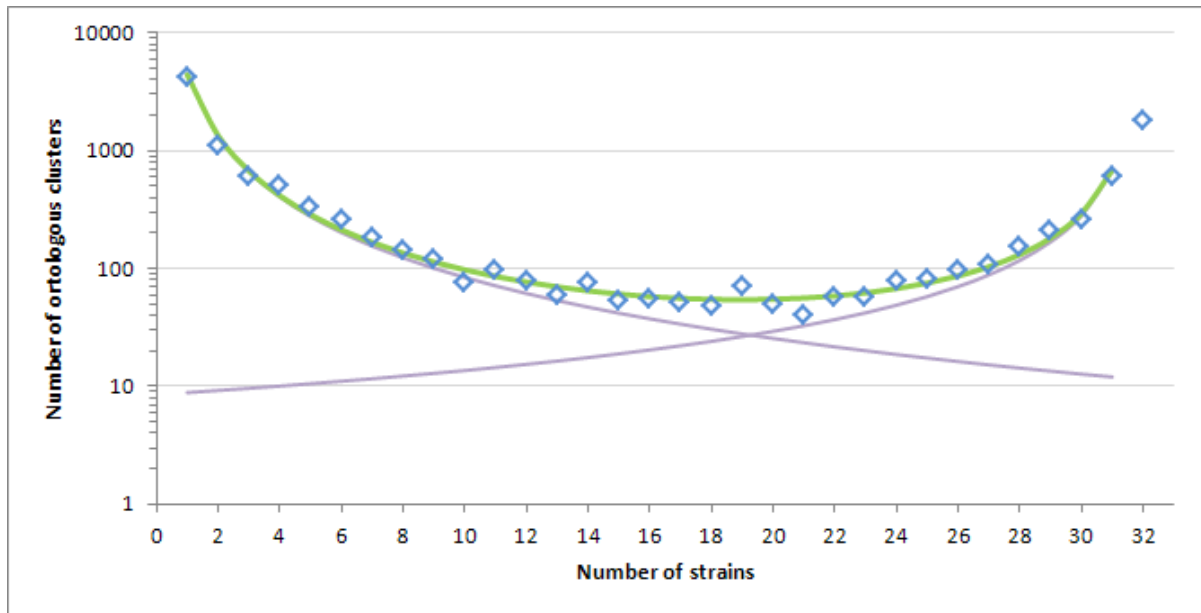
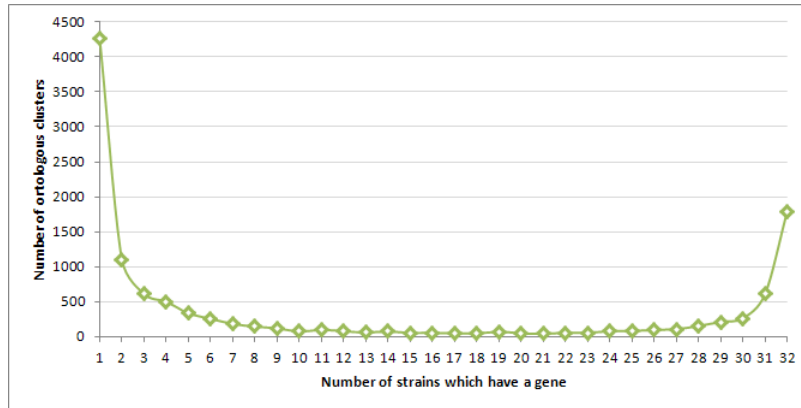
- Genome as a set of genes:
 - Losses, gains, duplications
 - Horizontal transfer
- Genome as a sequence:
 - Macro level: genome rearrangements
 - Macro level: homologous recombination
 - Micro level: mutation signatures
- Genome as an instruction: evolution of regulatory networks

Basic assumptions:

- Conservation implicates function
- Parallel events indicate positive selection

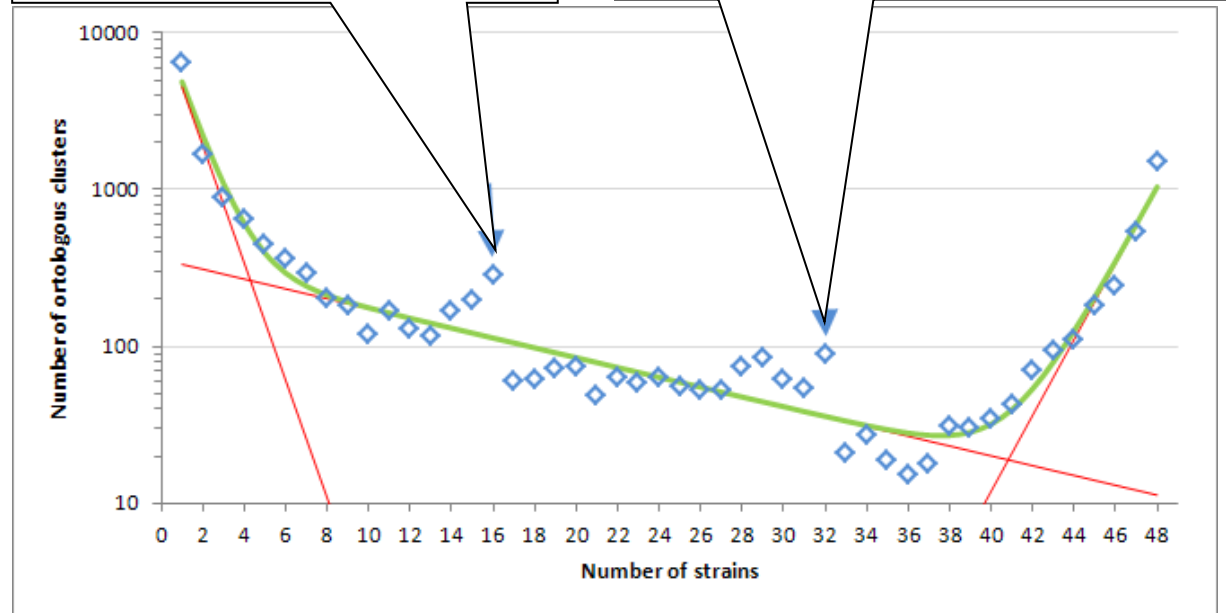
Sprectral function (U-curve)

E. coli +
Shigella

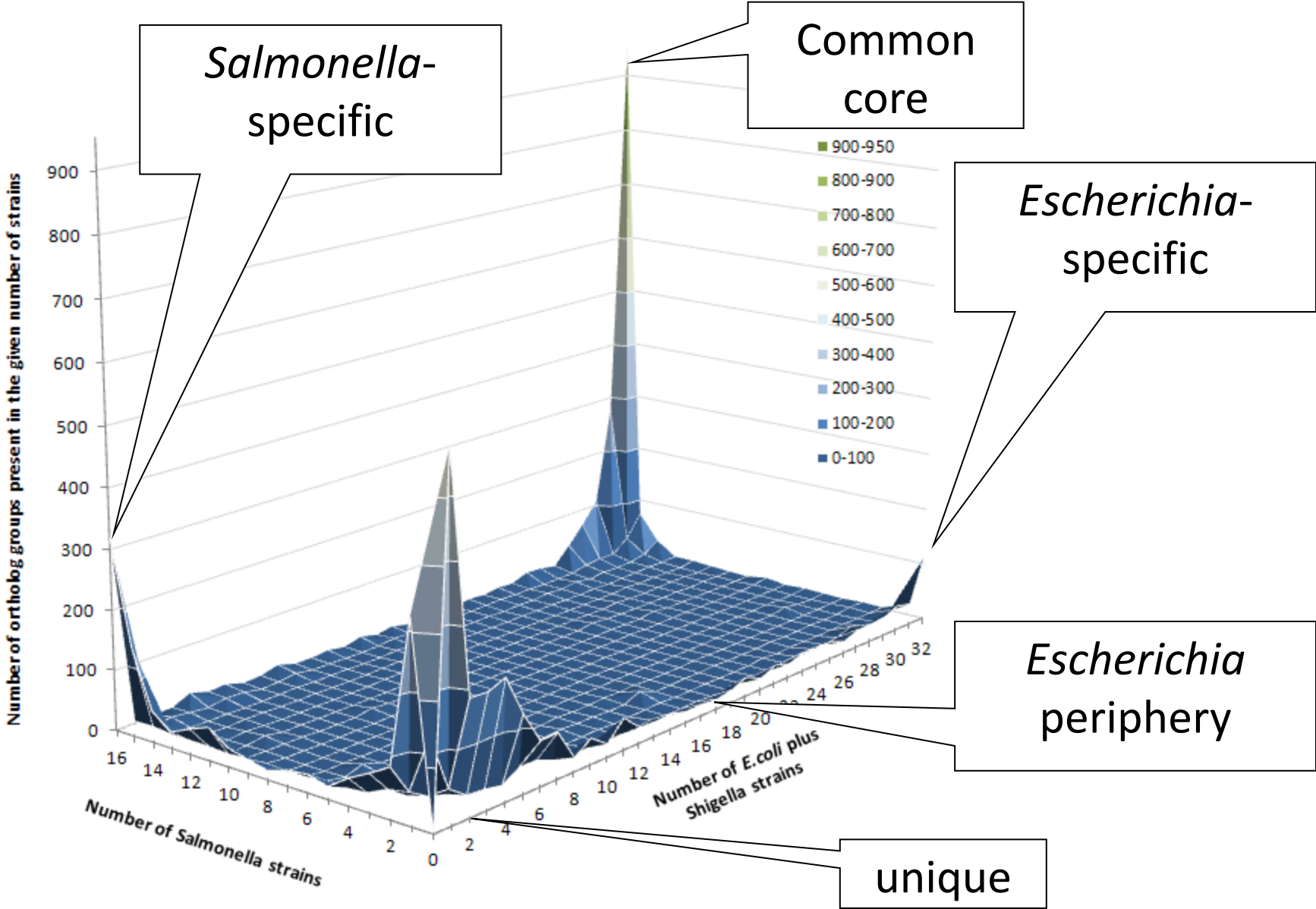


Salmonella-specific

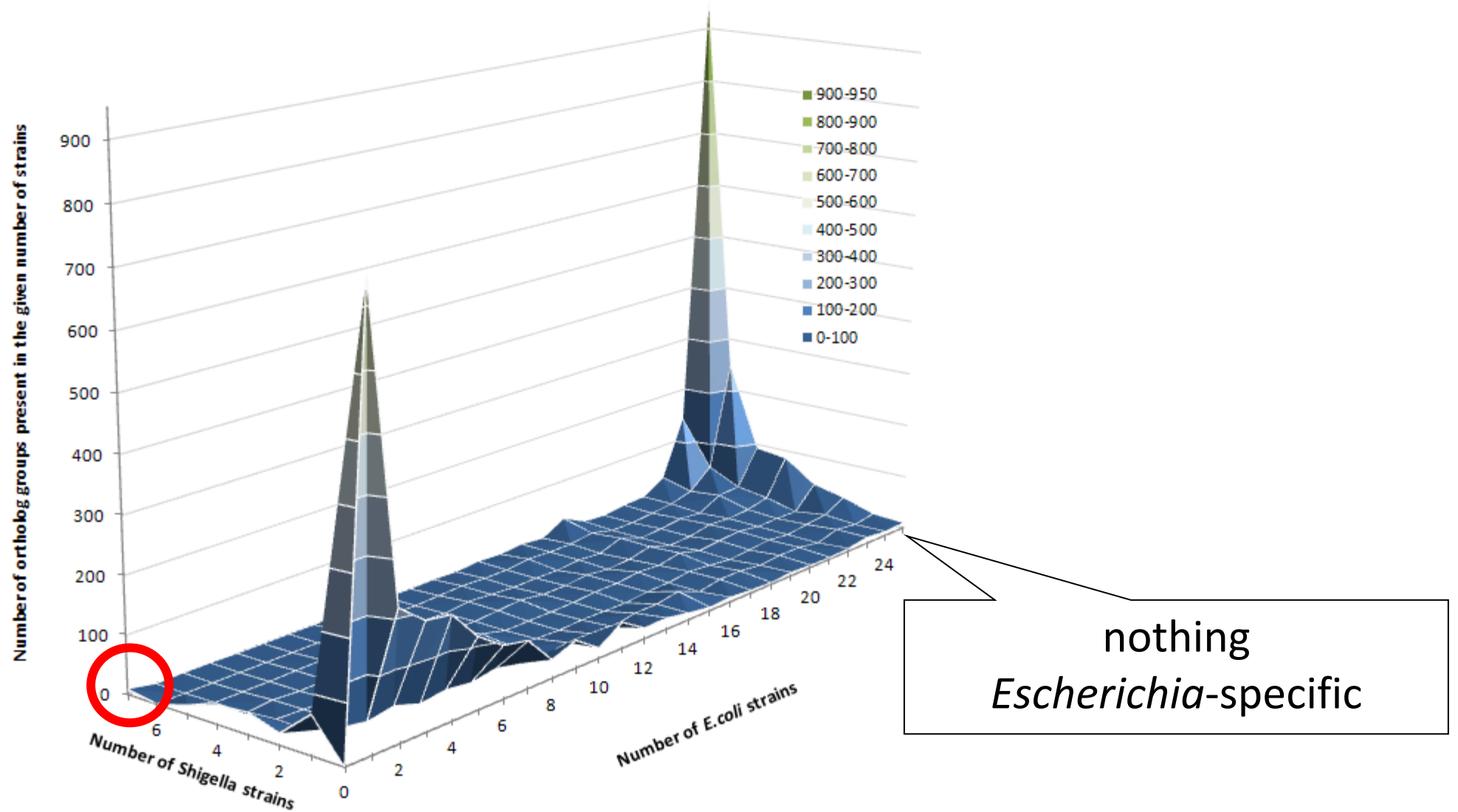
Escherichia-specific



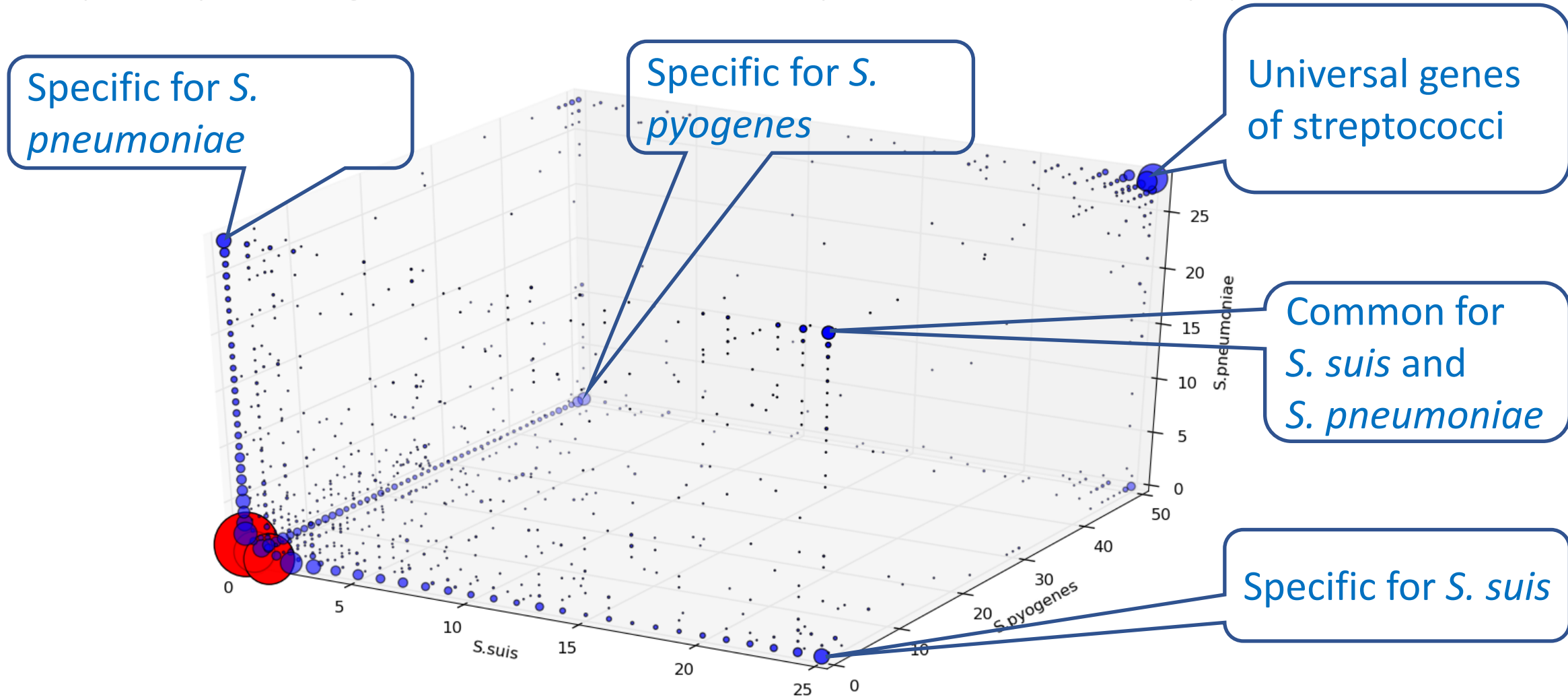
Double pan-genome (*Escherichia* vs *Salmonella*)



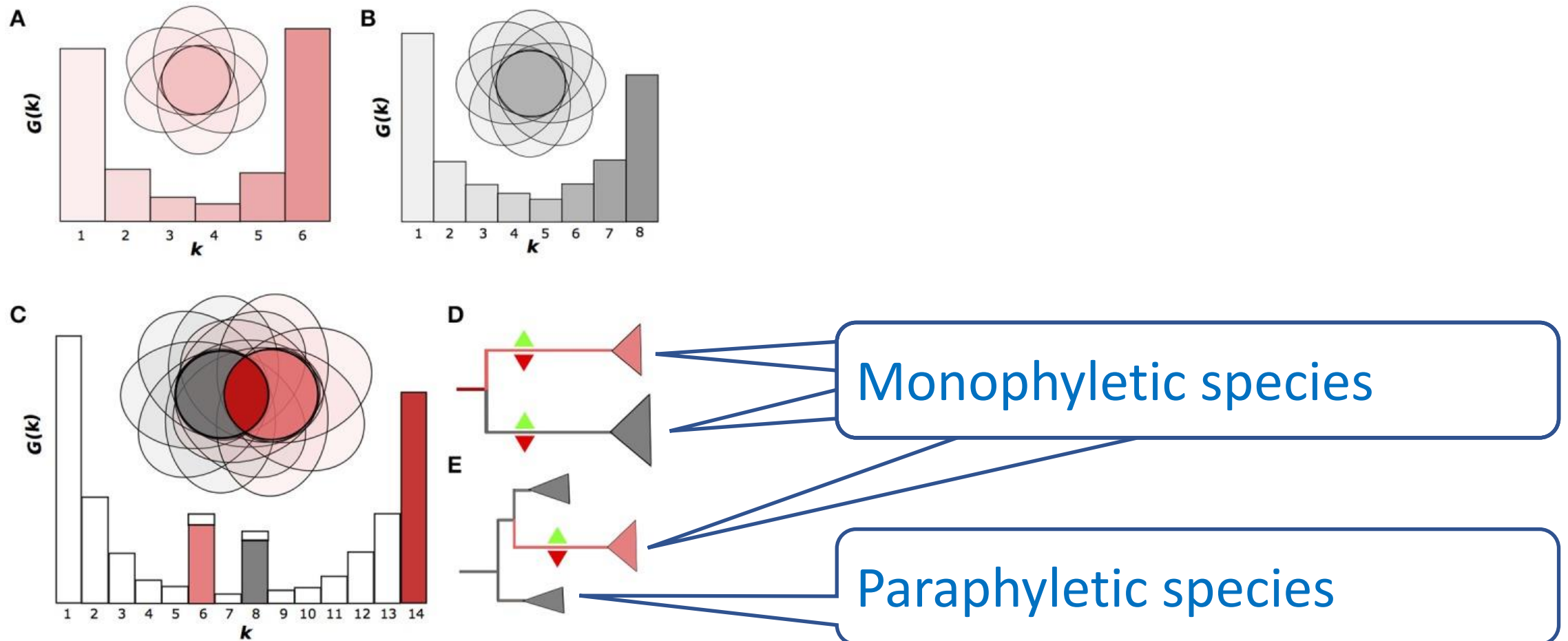
Double pan-genome (*Escherichia* vs *Shigella*)



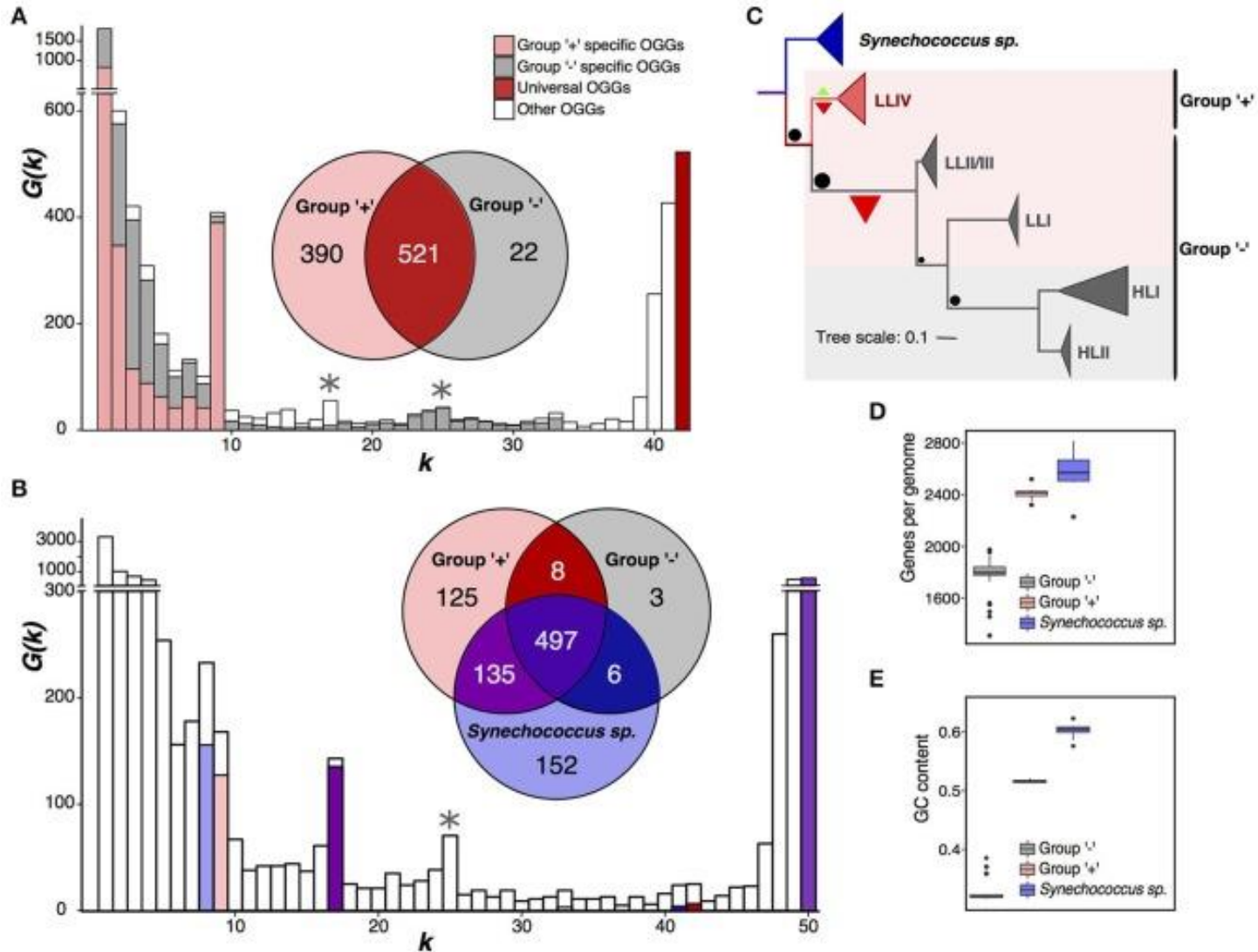
Triple pan-genome of *Streptococcus* spp.

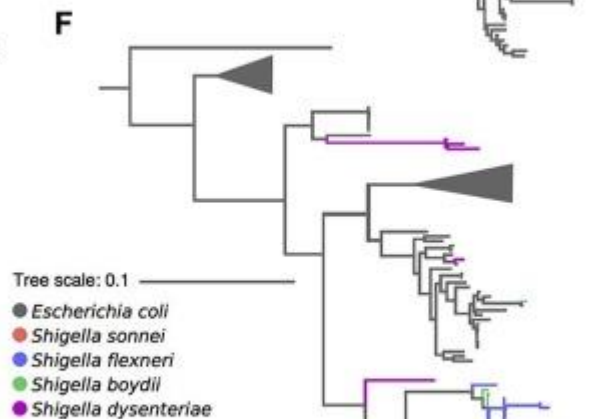
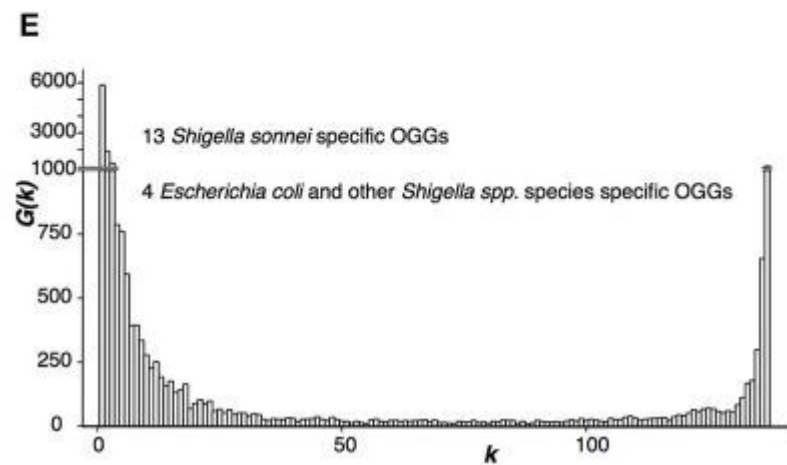
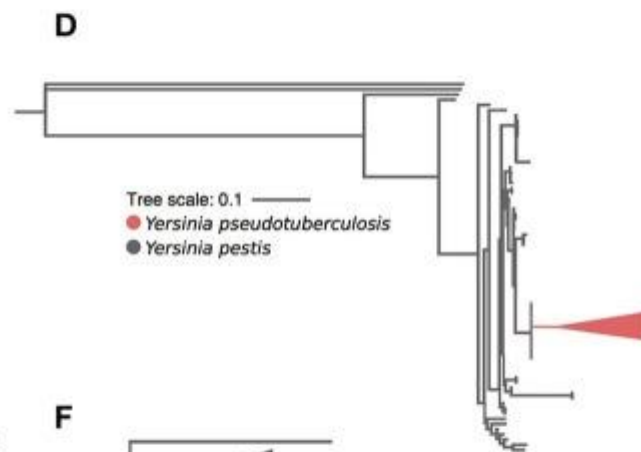
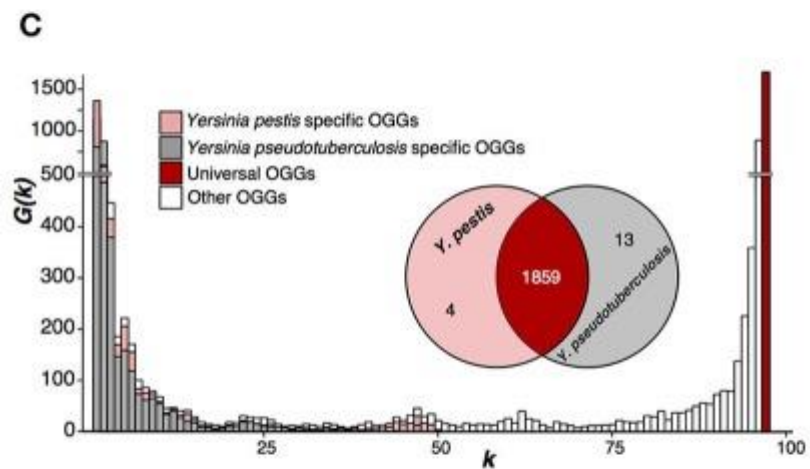
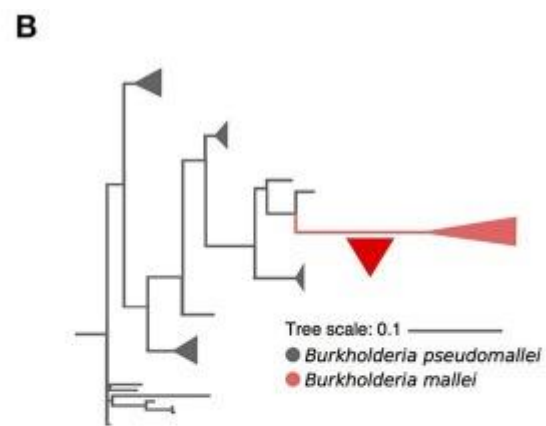
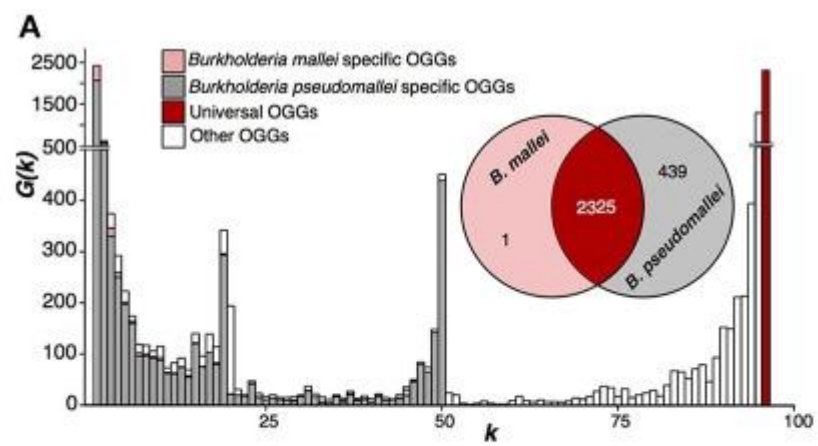


Definition of species via gene composition

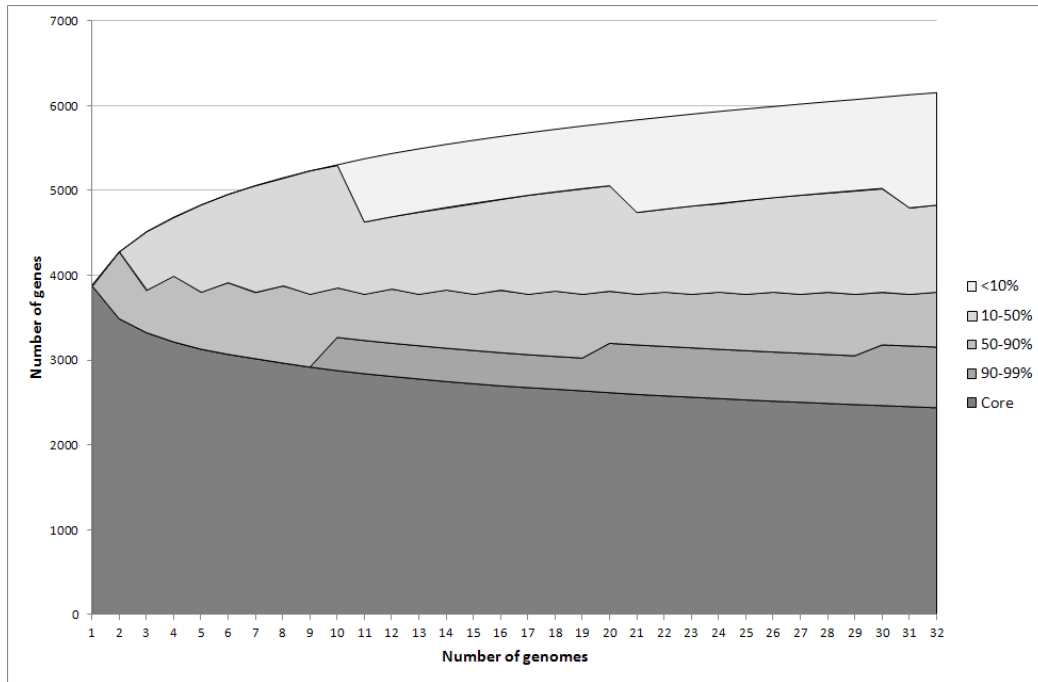


Example: *Prochlorococcus* spp.

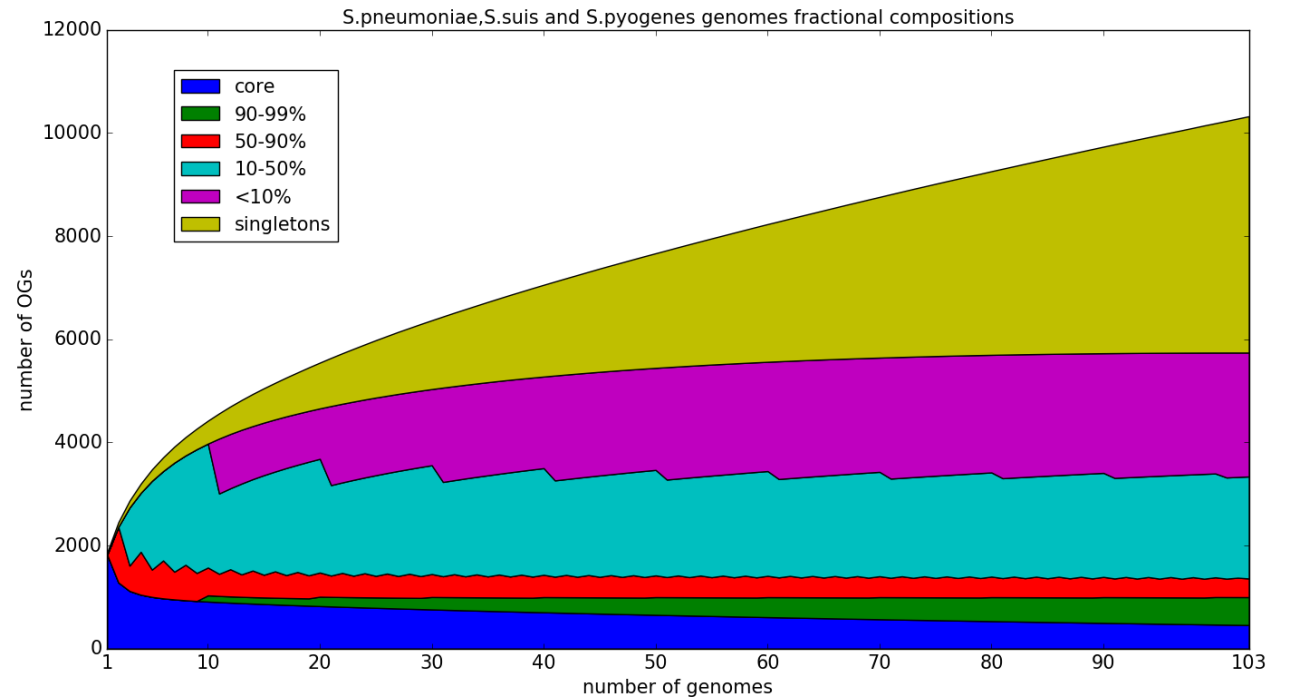




Fractional pan-genomes

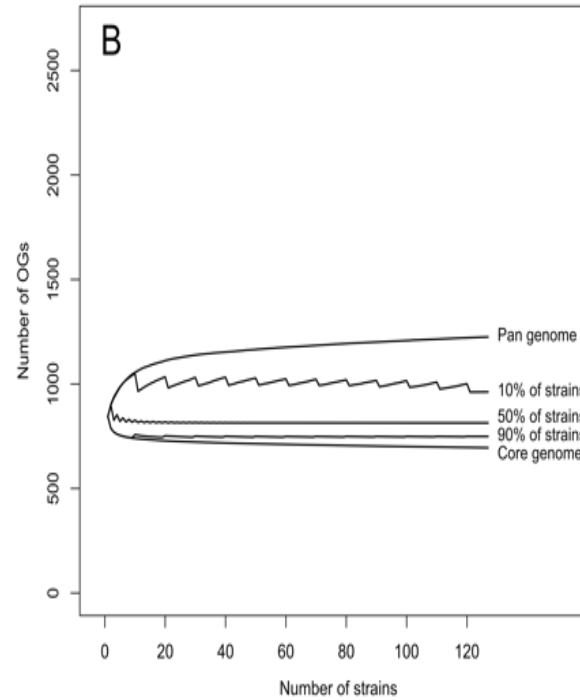
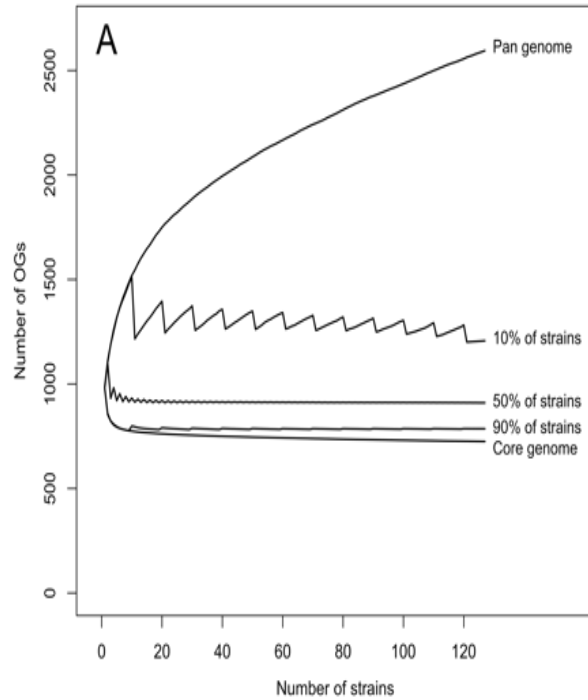


E. coli + *Salmonella*

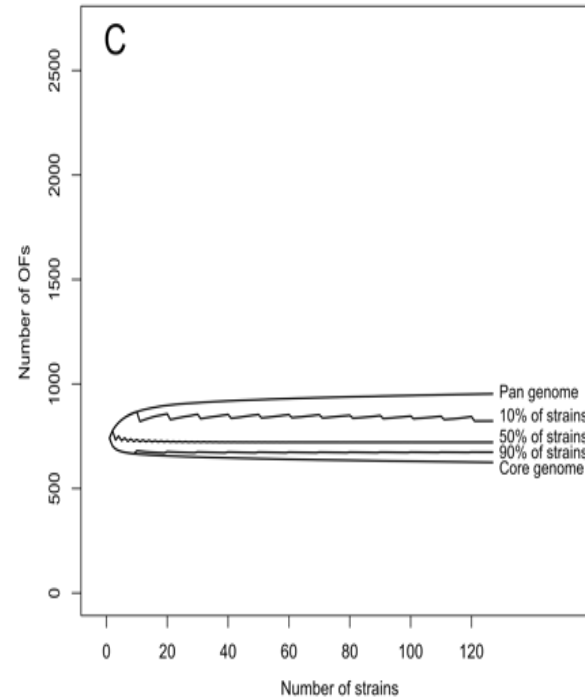


Streptococcus spp.

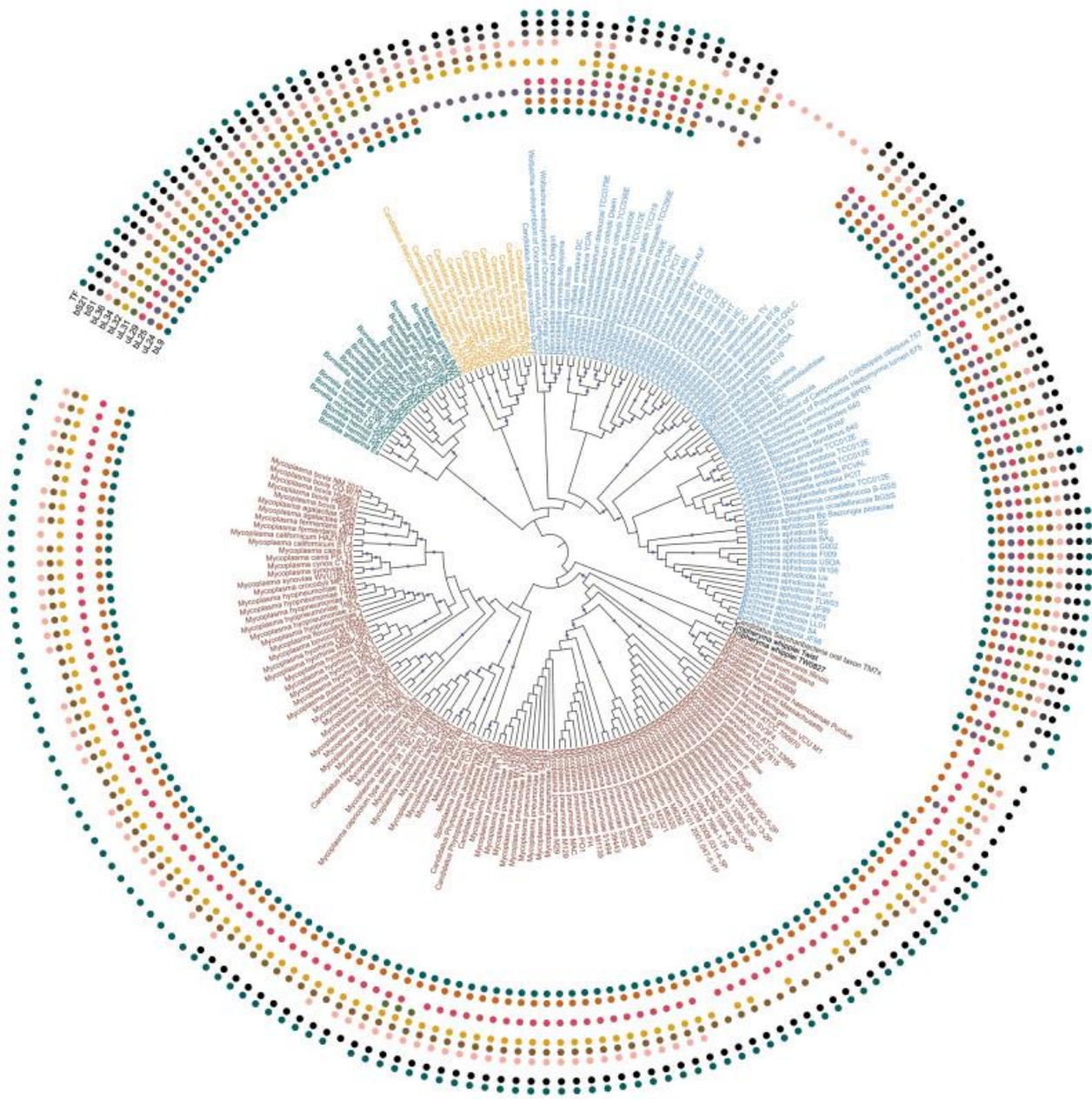
Chlamidia spp.



Remove ORFs
shorter than
100 codons

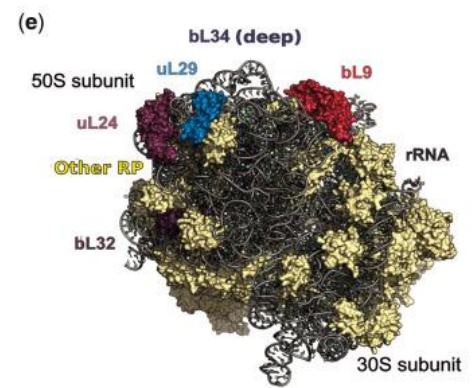
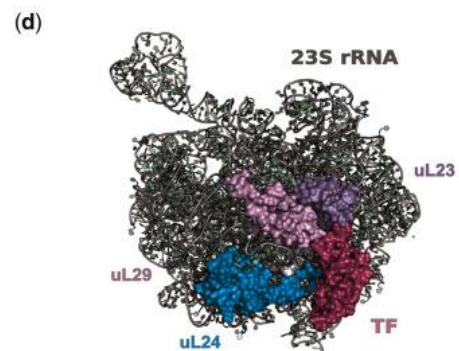
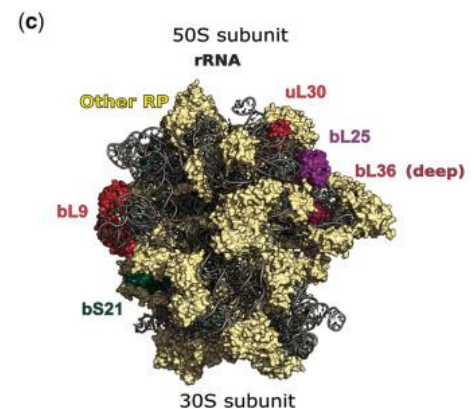
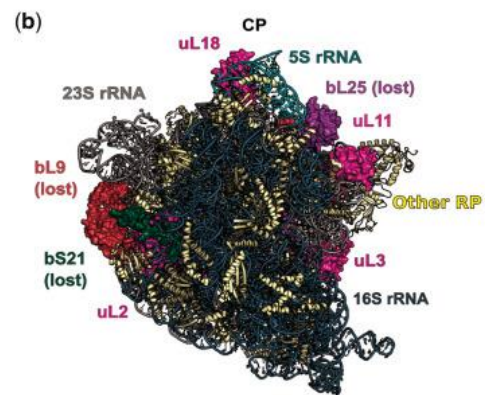


Collect
families

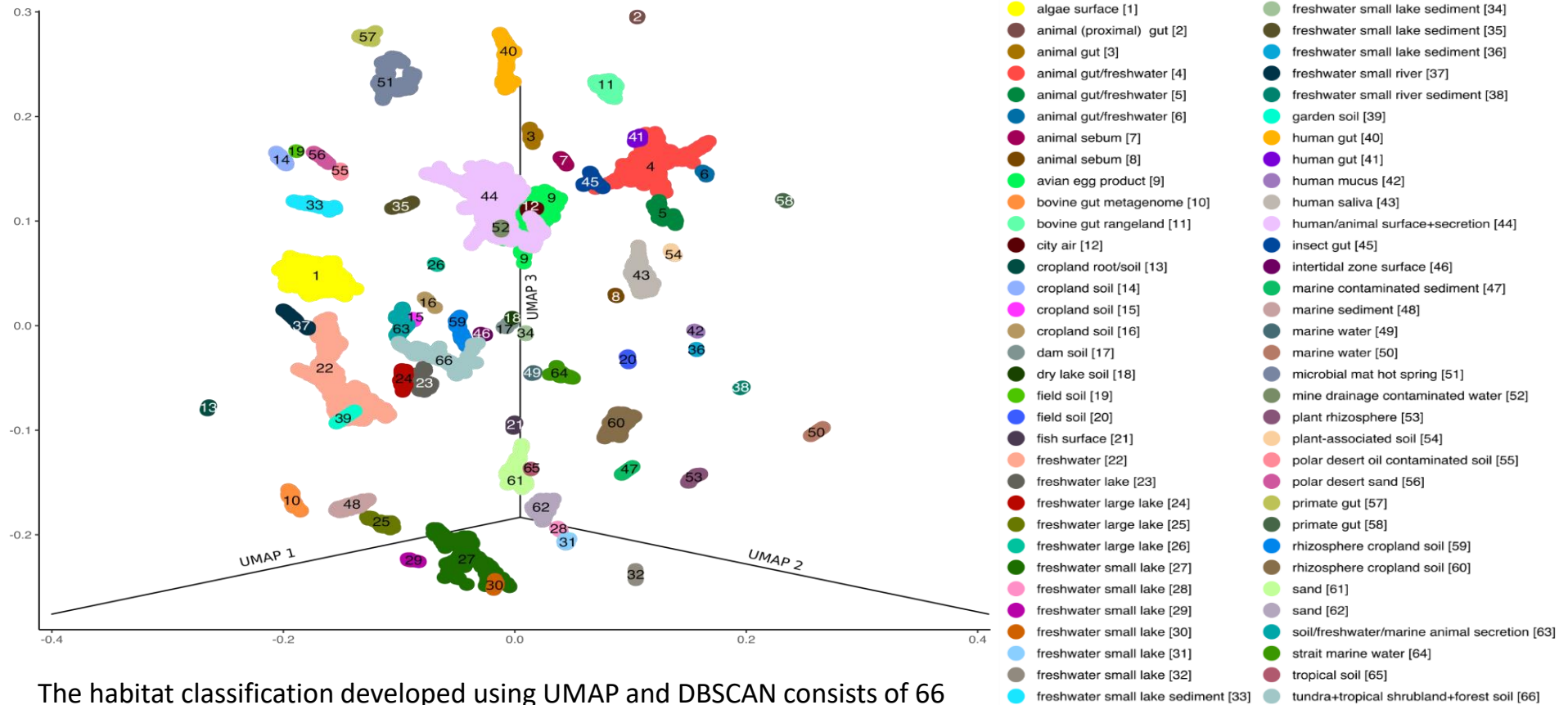


(a)

TF	bL36	bL32	uL29	bL34	uL24	bL9	uL30	bS21	bL25	bS1
bS1	1	1	1	0.97	0.11	0.65	0.1	0.1	0.02	0
bL25	1	1	1	0.77	0.1	0.57	0.04	0.26	0	0.02
bS21	1	1	1	0.88	0.17	0.25	0.22	0	0.26	0.1
uL30	1	1	1	0.73	0.68	0.67	0	0.22	0.04	0.1
bL9	1	1	1	0.84	0.01	0	0.67	0.25	0.57	0.65
uL24	1	1	1	0.5	0	0.01	0.68	0.17	0.1	0.11
bL34	1	1	1	0	0.5	0.84	0.73	0.88	0.77	0.97
uL29	1	1	1	0	1	1	1	1	1	1
bL32	1	1	0	1	1	1	1	1	1	1
bL36	1	0	1	1	1	1	1	1	1	1
TF	0	1	1	1	1	1	1	1	1	1

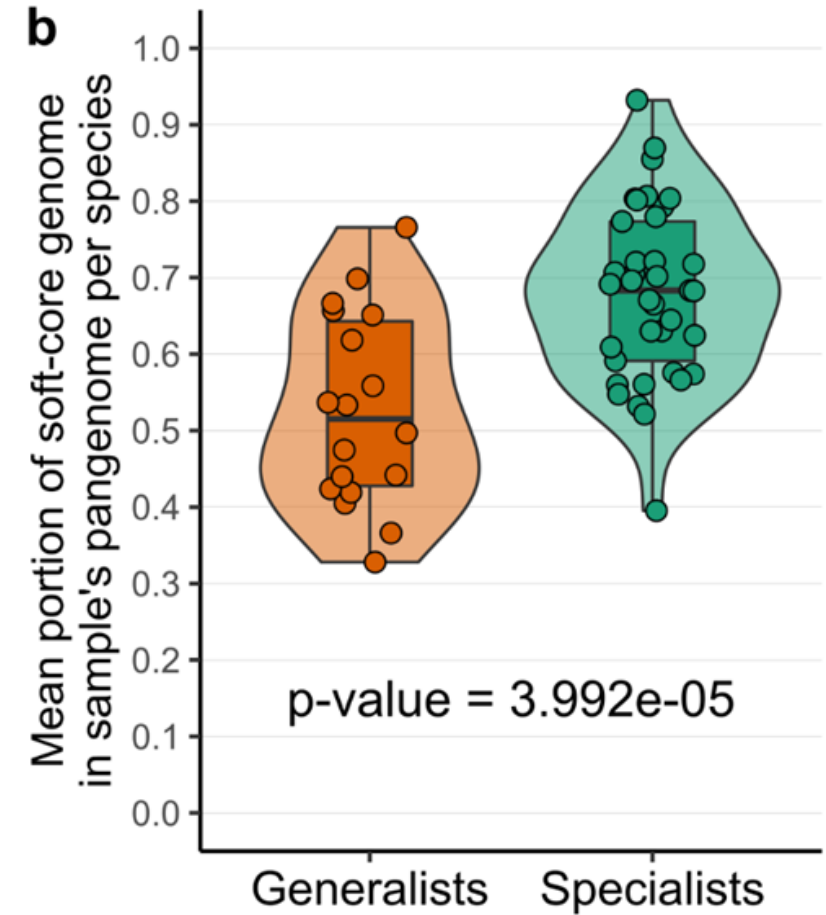
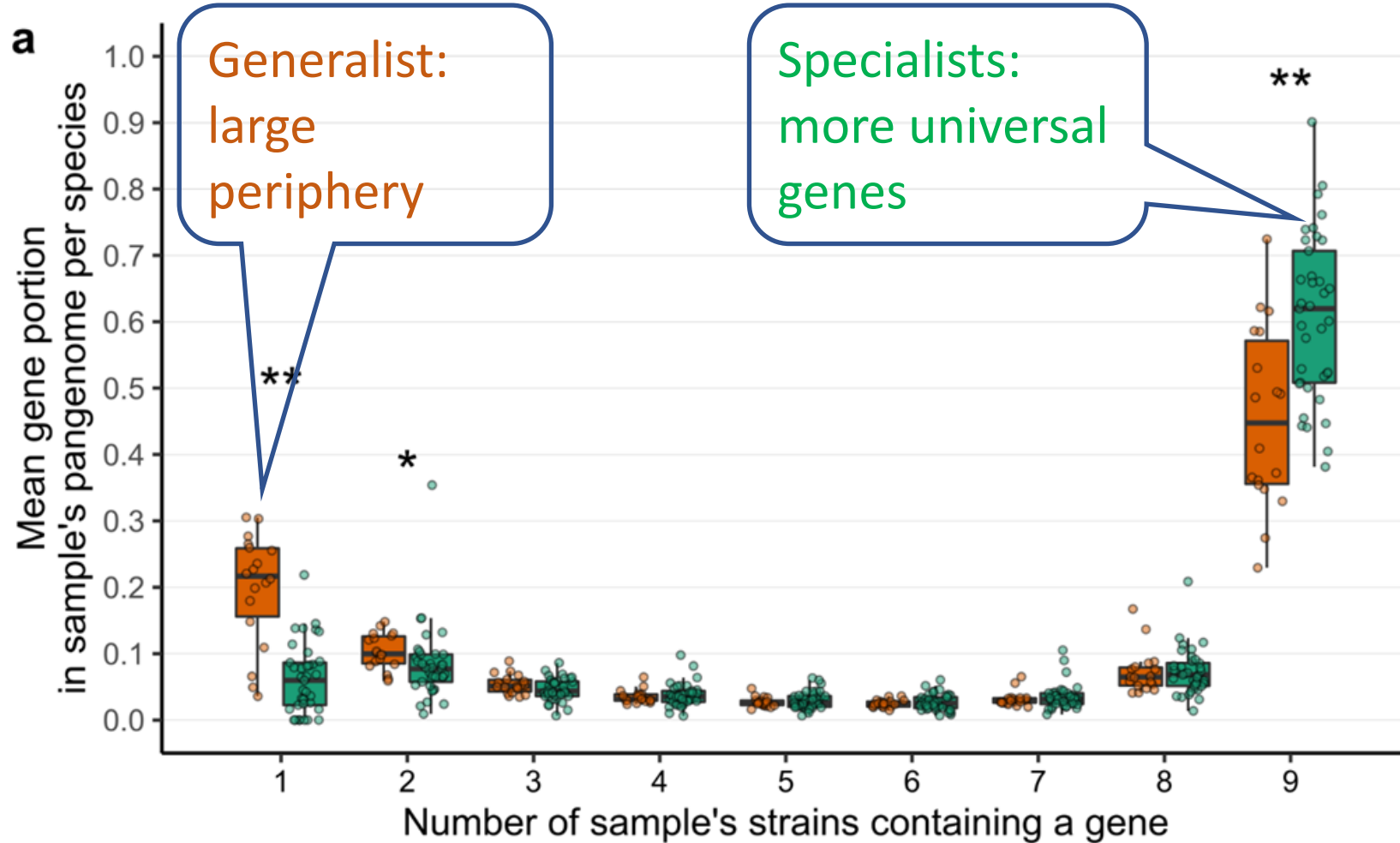


Metagenomics – classification of habitats



The habitat classification developed using UMAP and DBSCAN consists of 66 clusters.

Generalists have larger periphery



Ubiquitin ligases IpaH – functional signature of *Shigella* spp.

Phyletic patterns of ubiquitin ligases IpaH and functionally linked T3SS secretion systems

Color: *Shigella* species

Colored dots: IpaH subfamilies
 red-yellow: chromosomal
 green-blue: plasmid

Blue lines:
 no co-localized T3SS genes

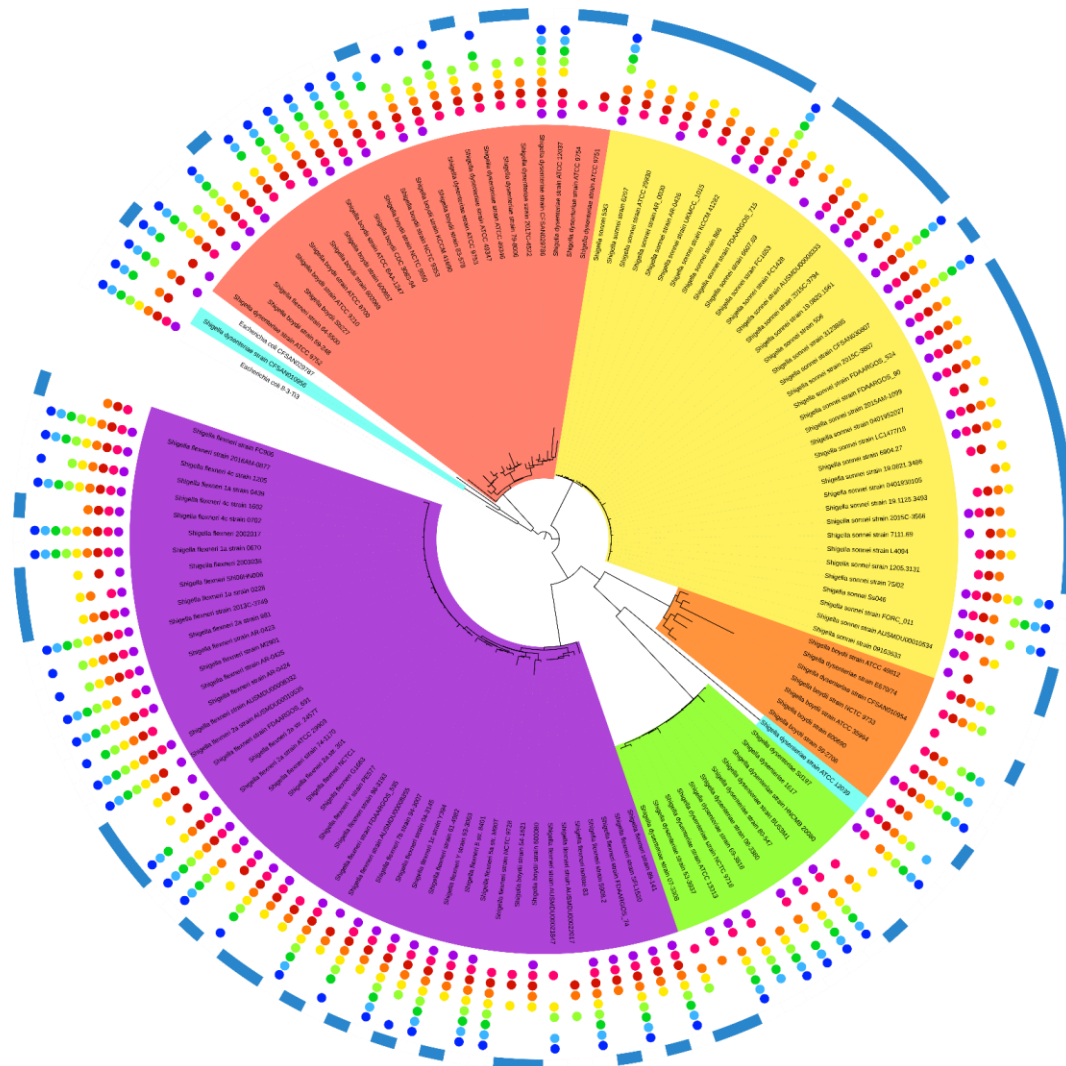
Tree scale: 0.01

Clades

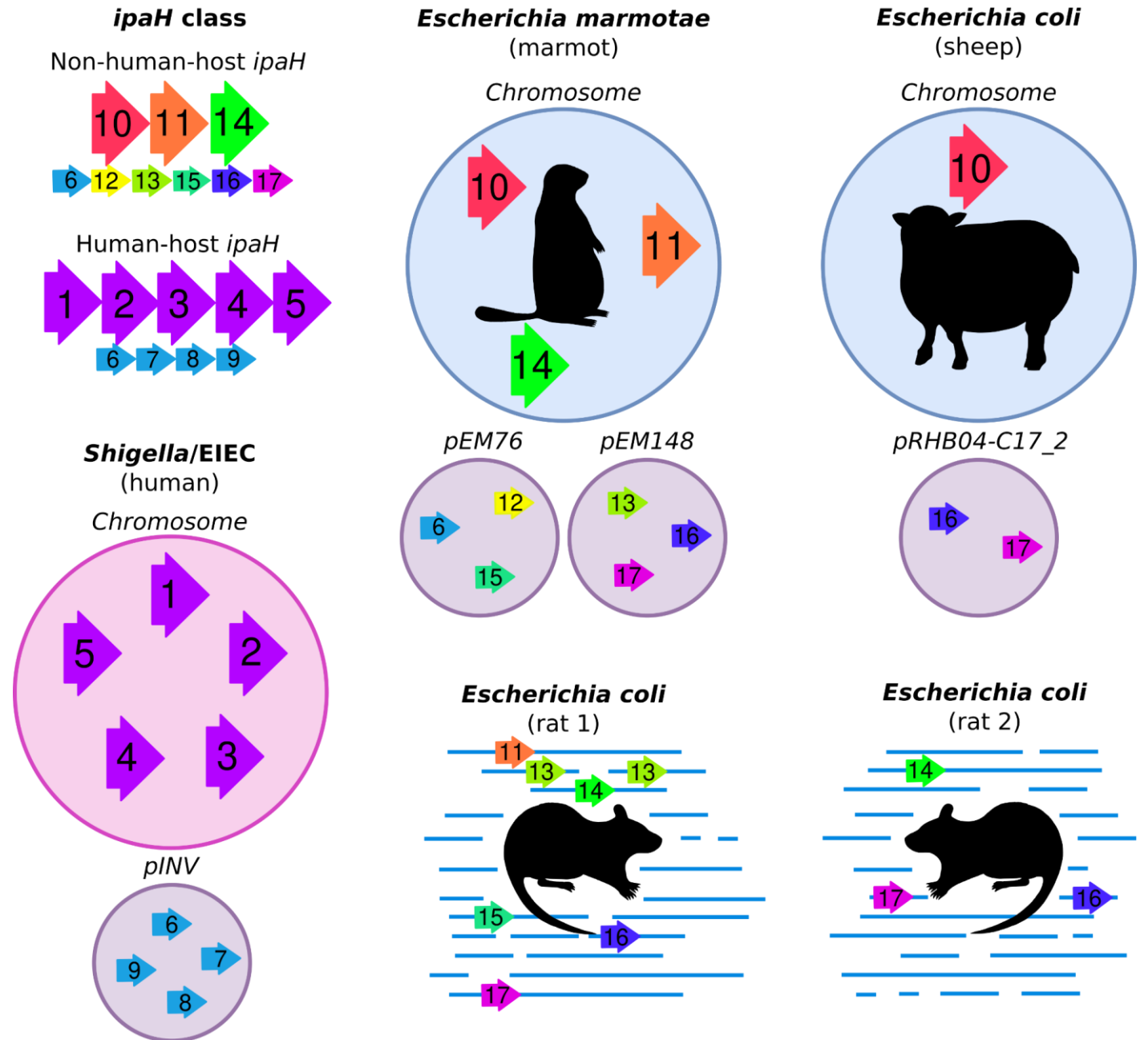
- S. sonnei*
- S. boydii* and *S. dysenteriae*
- S. flexneri*
- S. boydii* and *S. dysenteriae*
- S. dysenteriae*
- S. dysenteriae*
- EIEC

IpaH

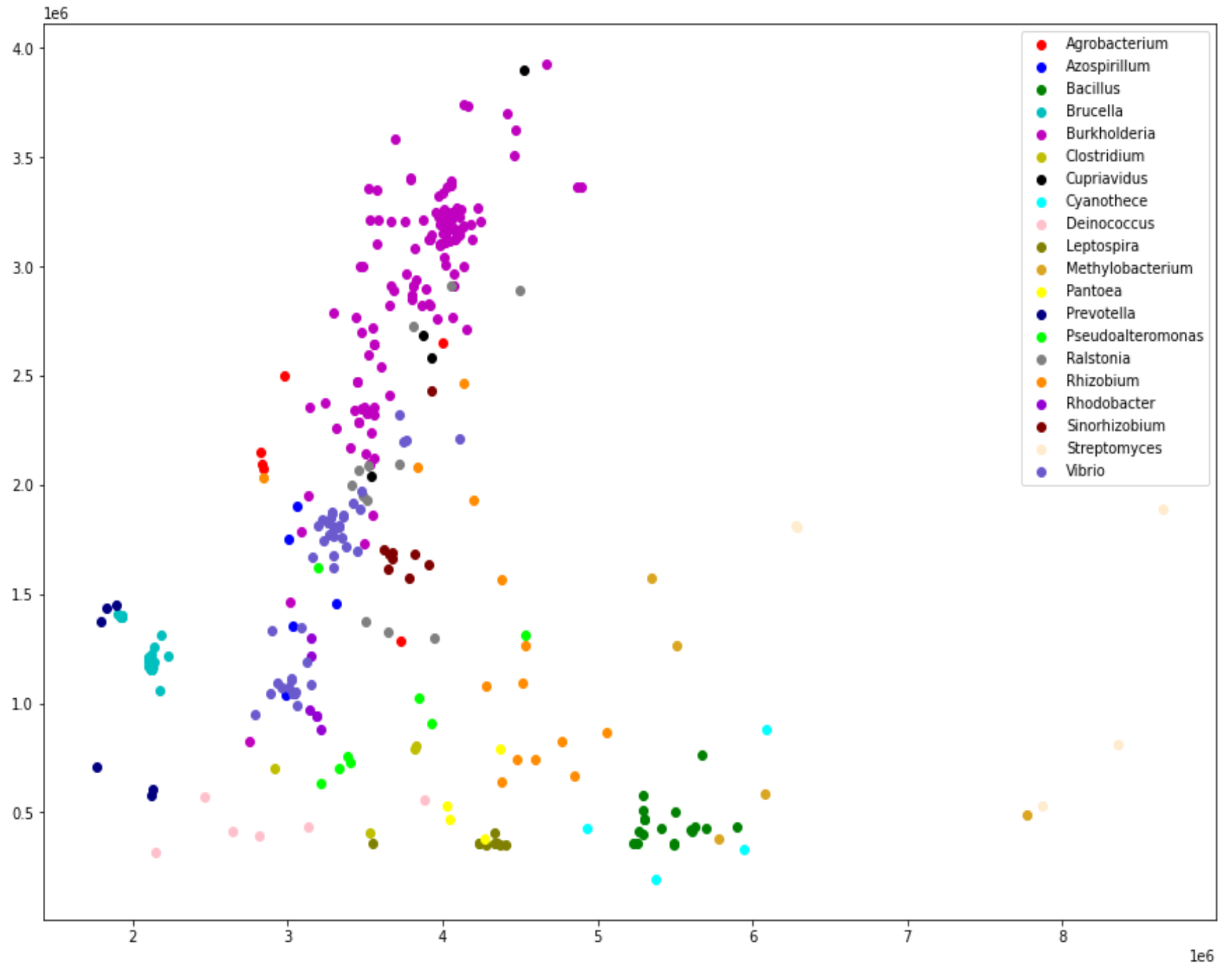
- Class 1, chromosomal
- Class 2, chromosomal
- Class 3, chromosomal
- Class 4, chromosomal
- Class 5, chromosomal
- Class 6, plasmid
- Class 7, plasmid
- Class 8, plasmid
- Class 9, plasmid



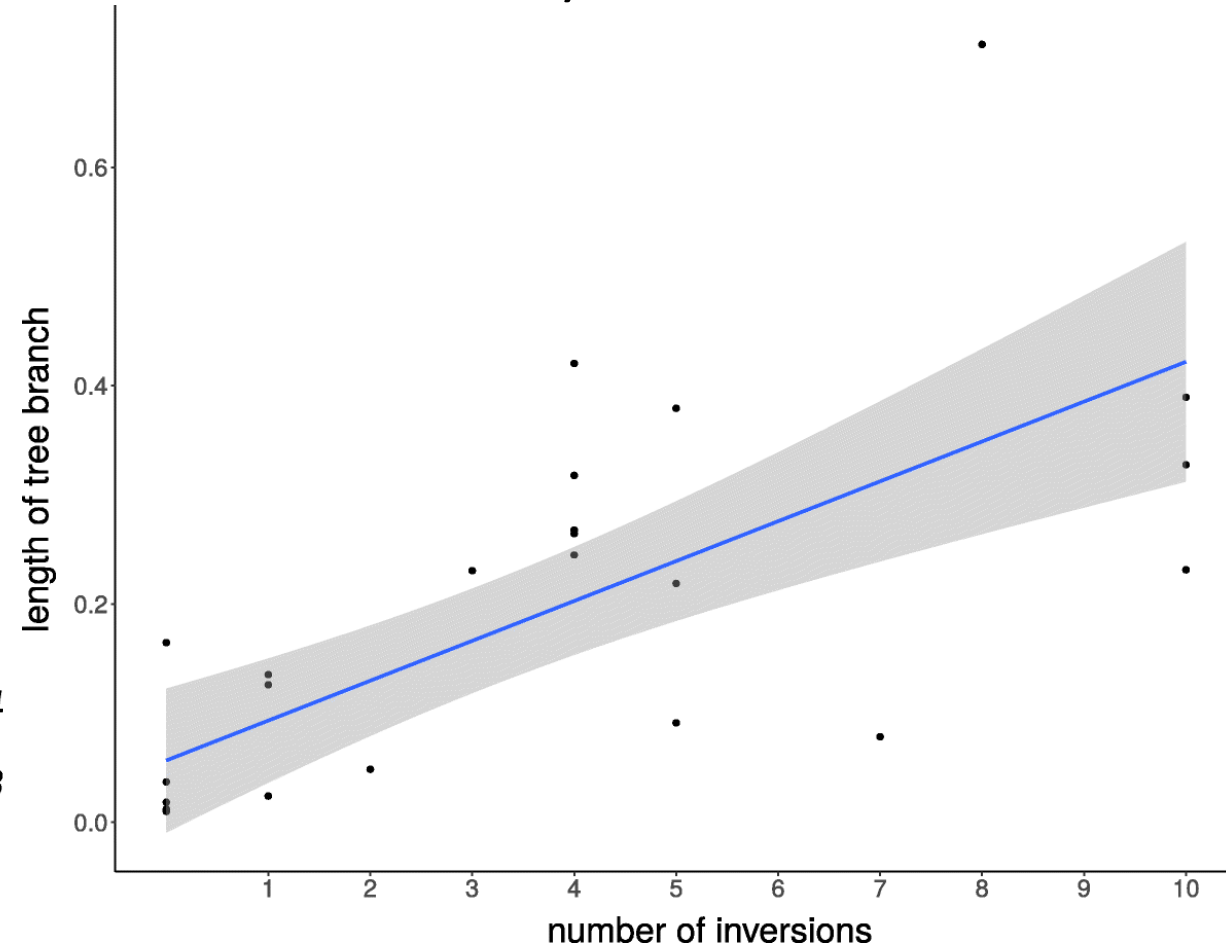
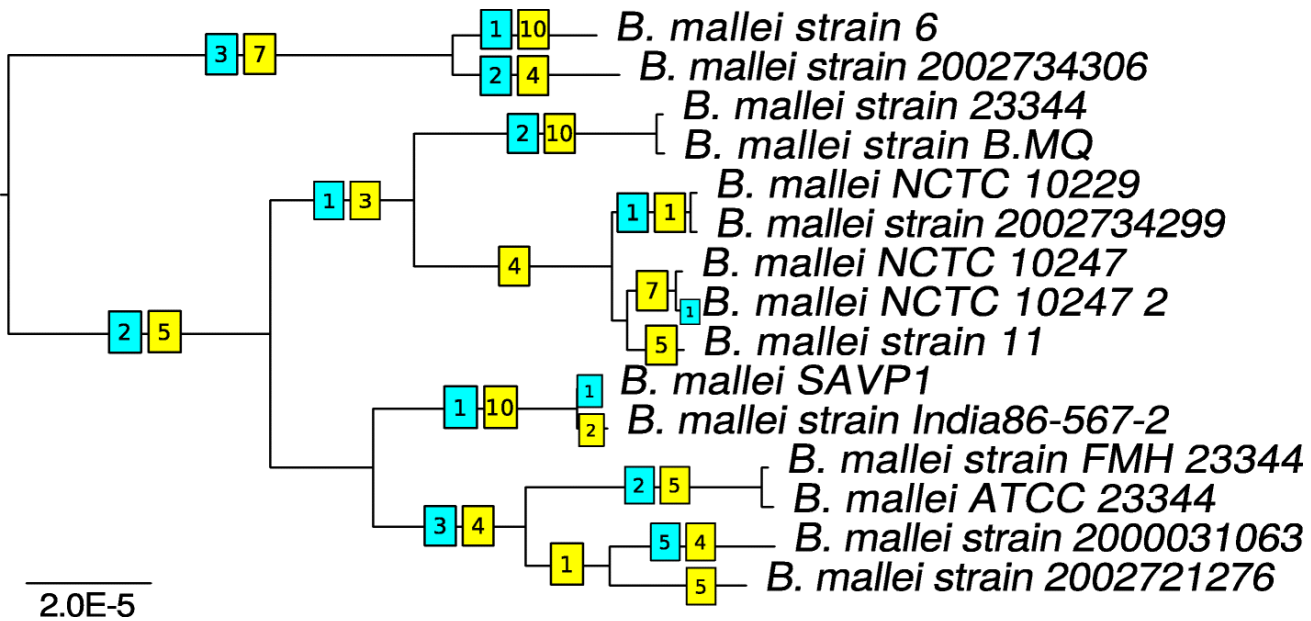
More subfamilies in non-human *E. coli*

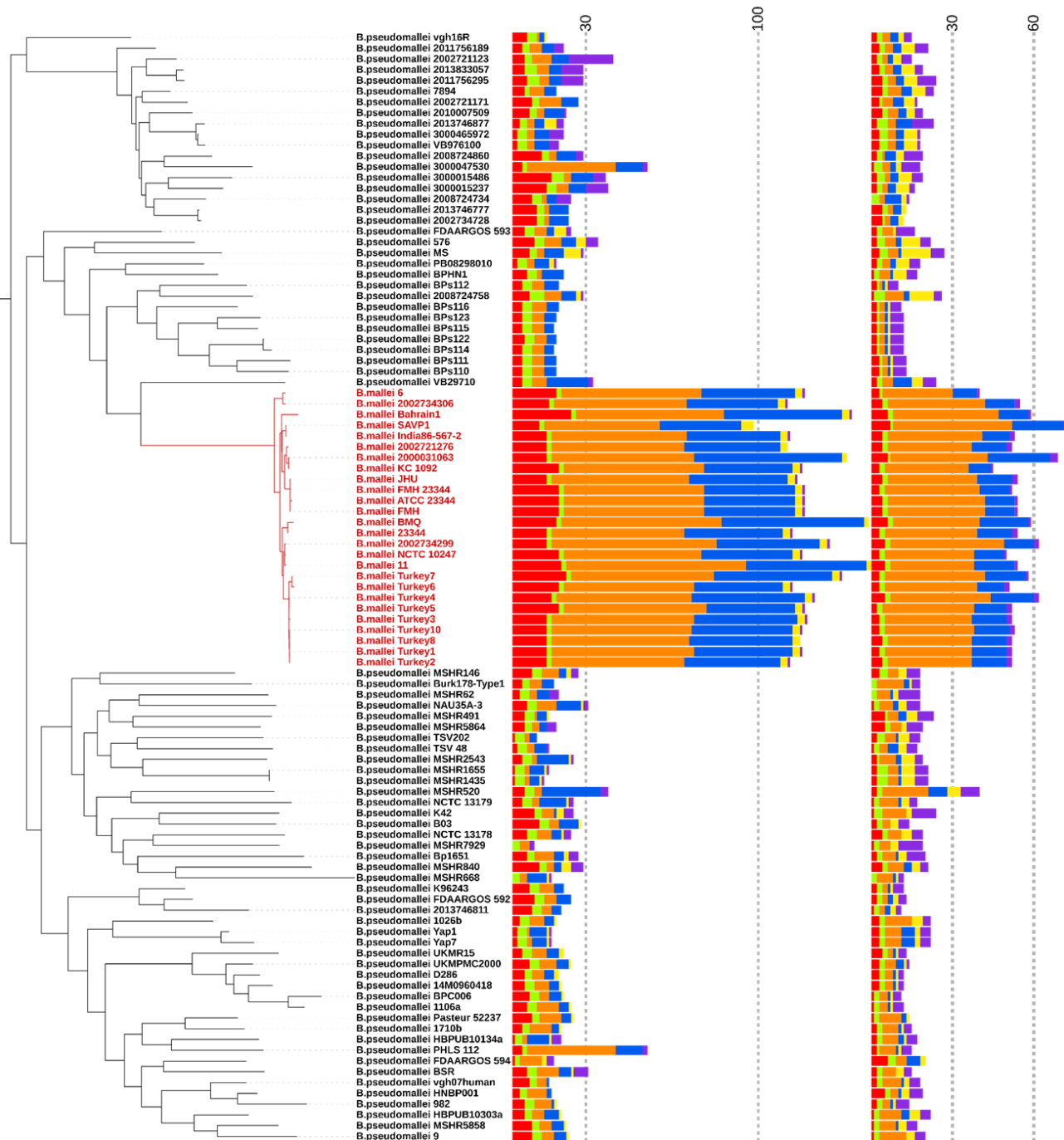








Two- component genomes



Increased inversion rate in a young pathogen *Burkholderia mallei* (a clade of paraphyletic *B. pseudomallei*)





IS type	<i>B. pseudomallei</i> , ch1	<i>B. mallei</i> , ch1	<i>B. pseudomallei</i> , ch2	<i>B. mallei</i> , ch2
 ISL3	5,4+-2,3	16,4+-2,7	2,3+-0,9	3,9+-0,9
 IS5_ssgr_IS427	3,6+-0,6	2	2,1+-0,5	1,9+-0,2
 IS3_ssgr_IS407	5,5+-2,6	57,4+-2,8	4,2+-1,8	34,1+-2,7
 IS1182	5,7+-2,3	41,5+-4,7	2,0+-1,3	12,9+-2,1
 IS3_ssgr_IS150	0,9+-0,9	2,9+-0,3	2,8+-1,5	0,0
 IS3_ssgr_IS3	1,9+-2,1	0,7+-0,4	3,9+-1,5	1,7+-0,5

Burst of inversions in *Shigella* spp.

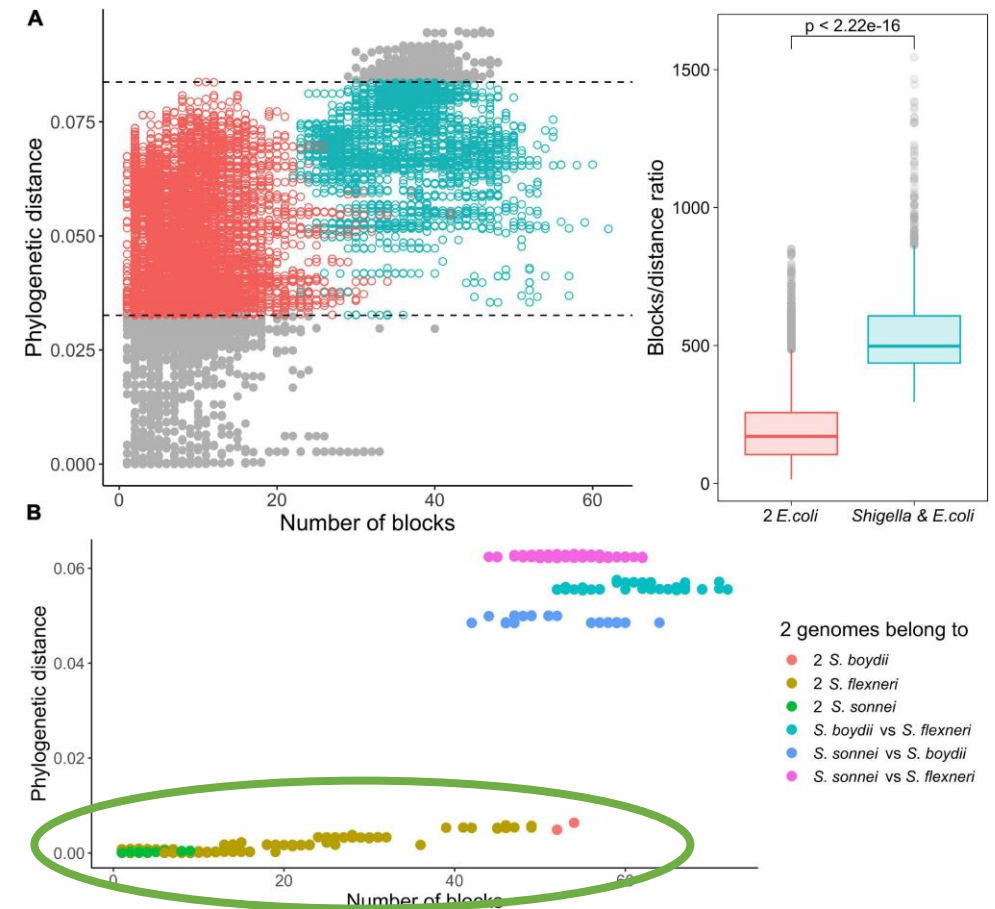
of syntenic groups in comparisons:

Red: *E. coli* vs *E. coli*

Blue: *E. coli* vs *Shigella*

Bottom: *Shigella* vs *Shigella*

(Many blocks = Numerous breakpoints even at short phylogenetic distance)



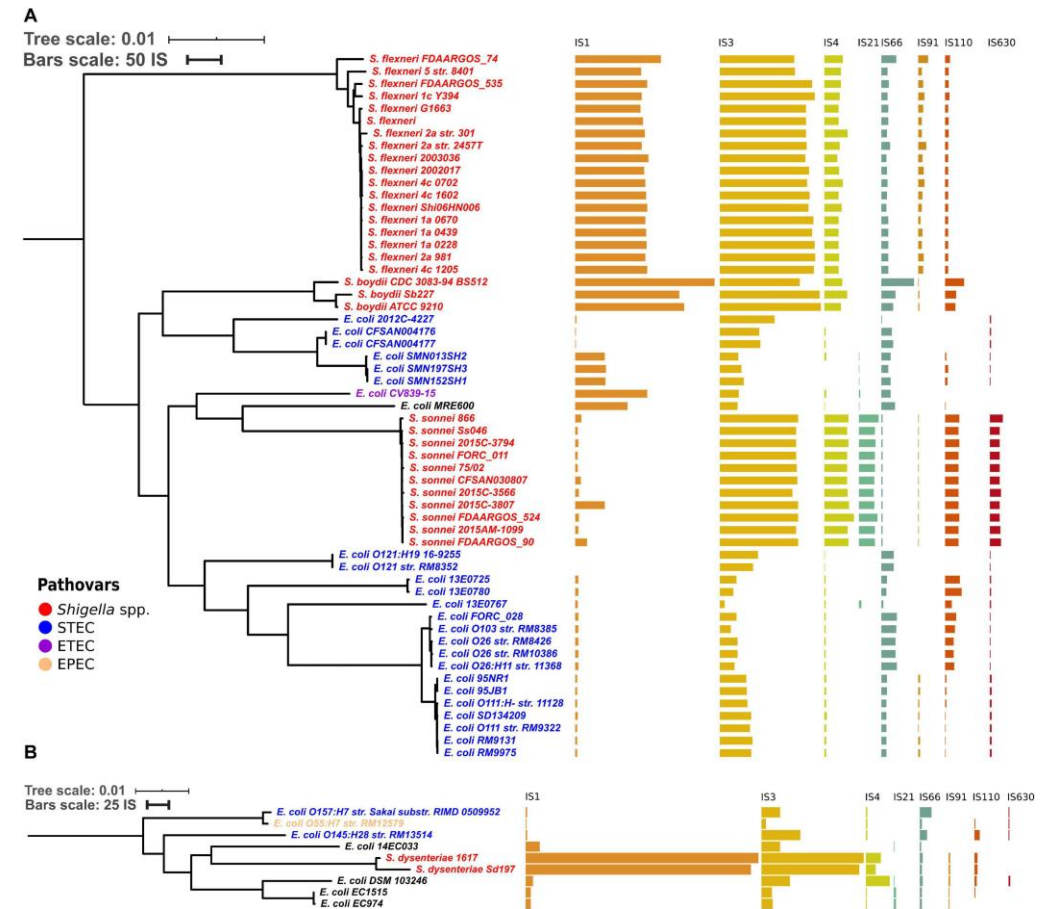
Burst of IS elements in *Shigella* spp.

Branch color – species
 red – *Shigella* spp.
 blue – *E. coli* strains

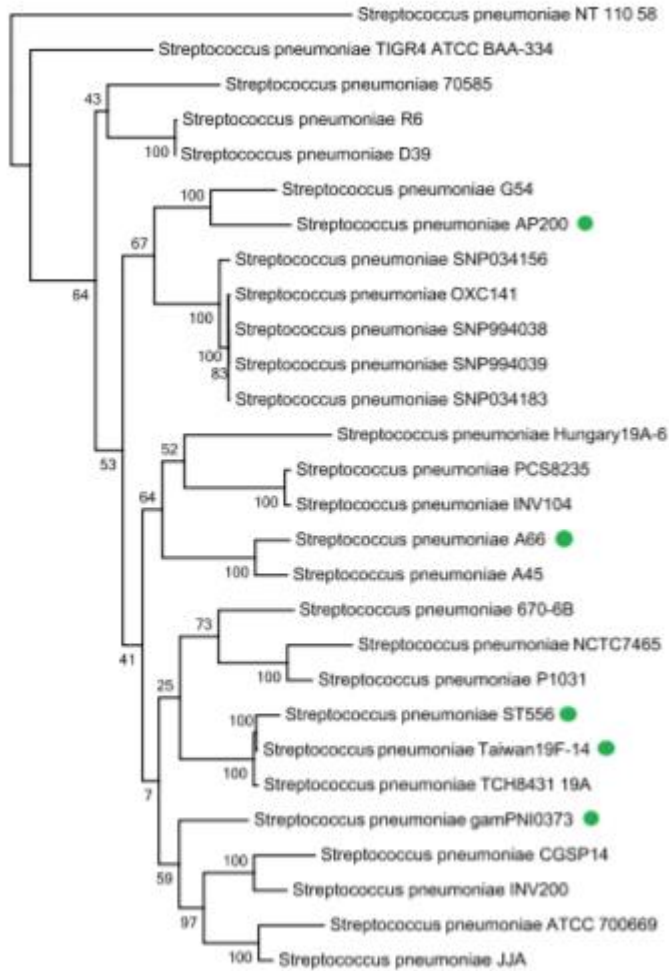
Bars –
 # of IS elements

Top: phylogroup B1

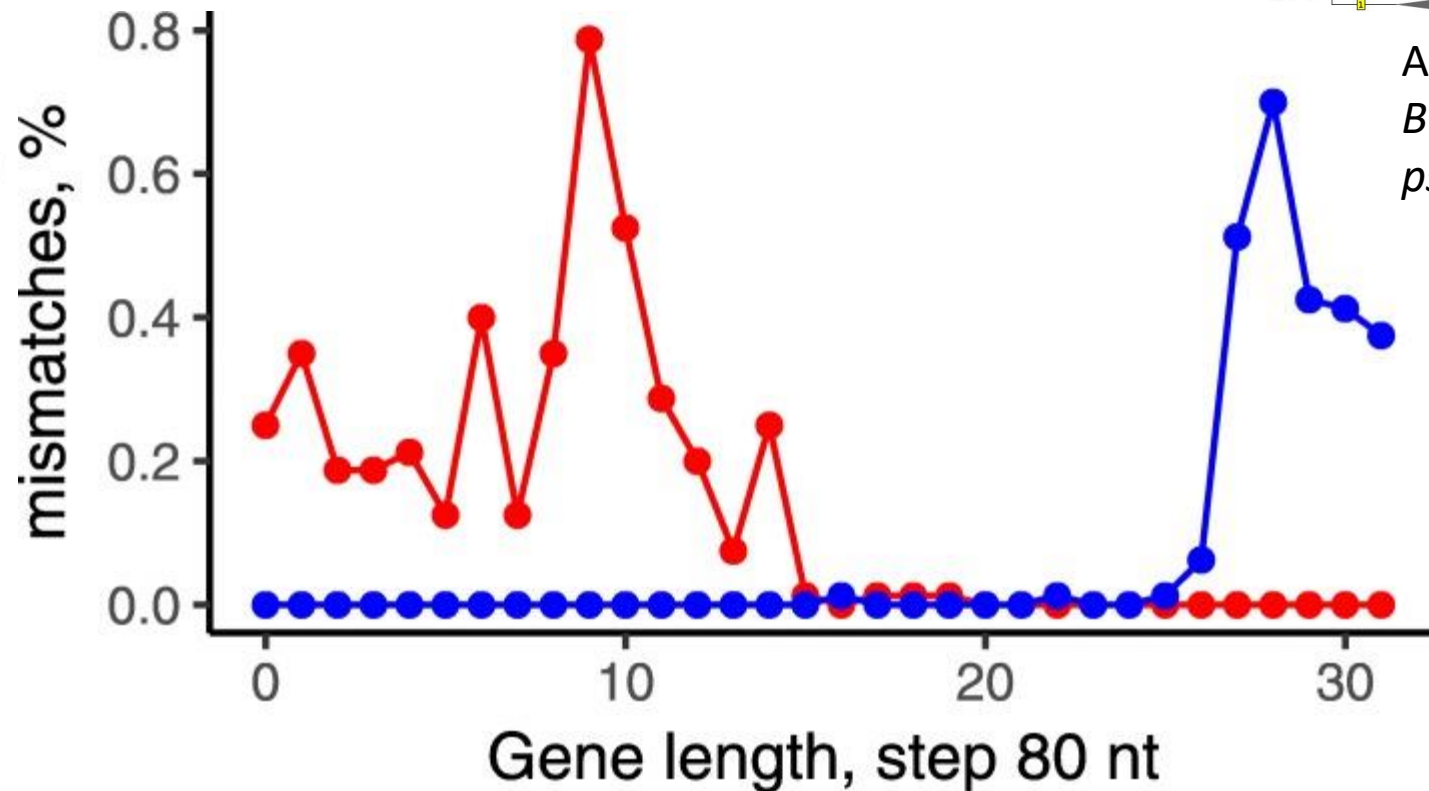
Bottom: phylogroup E



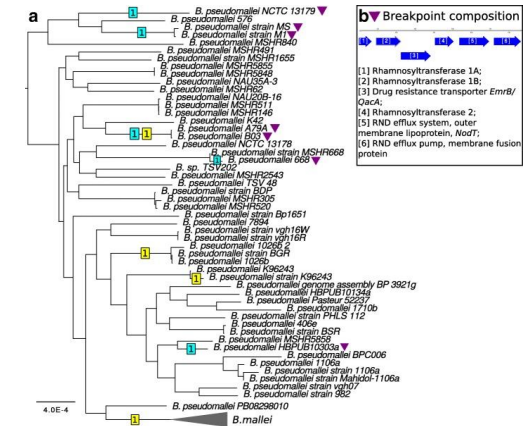
Parallel inversions as a mechanism of phase variations: *Streptococcus pneumoniae*



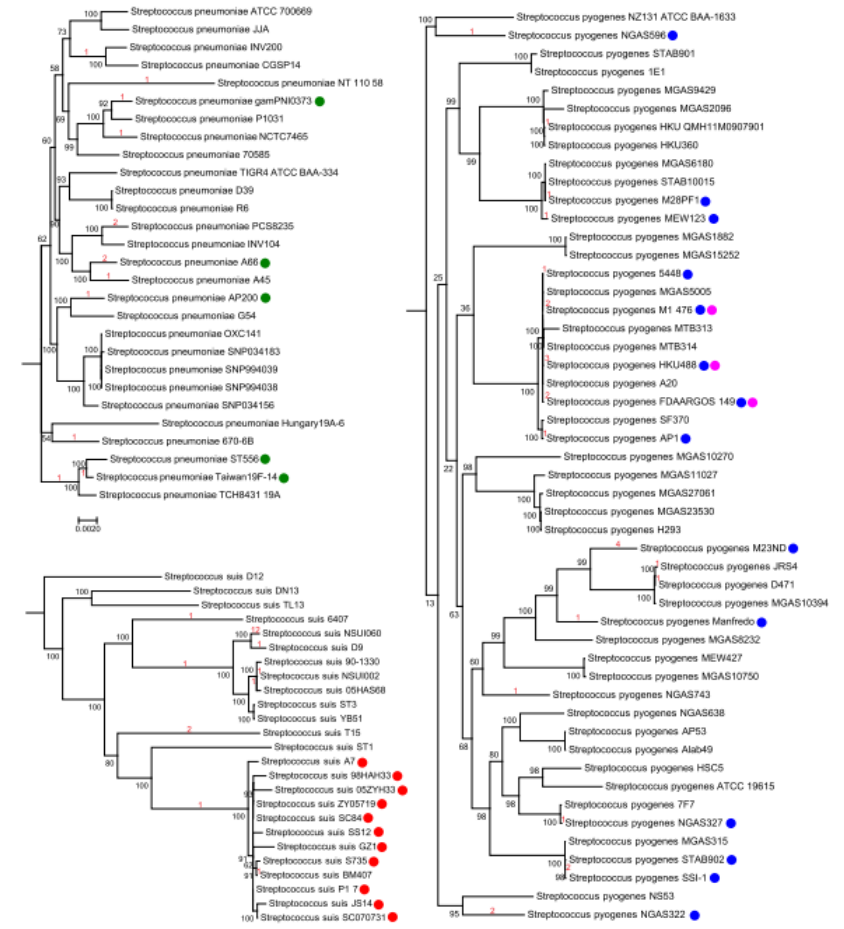
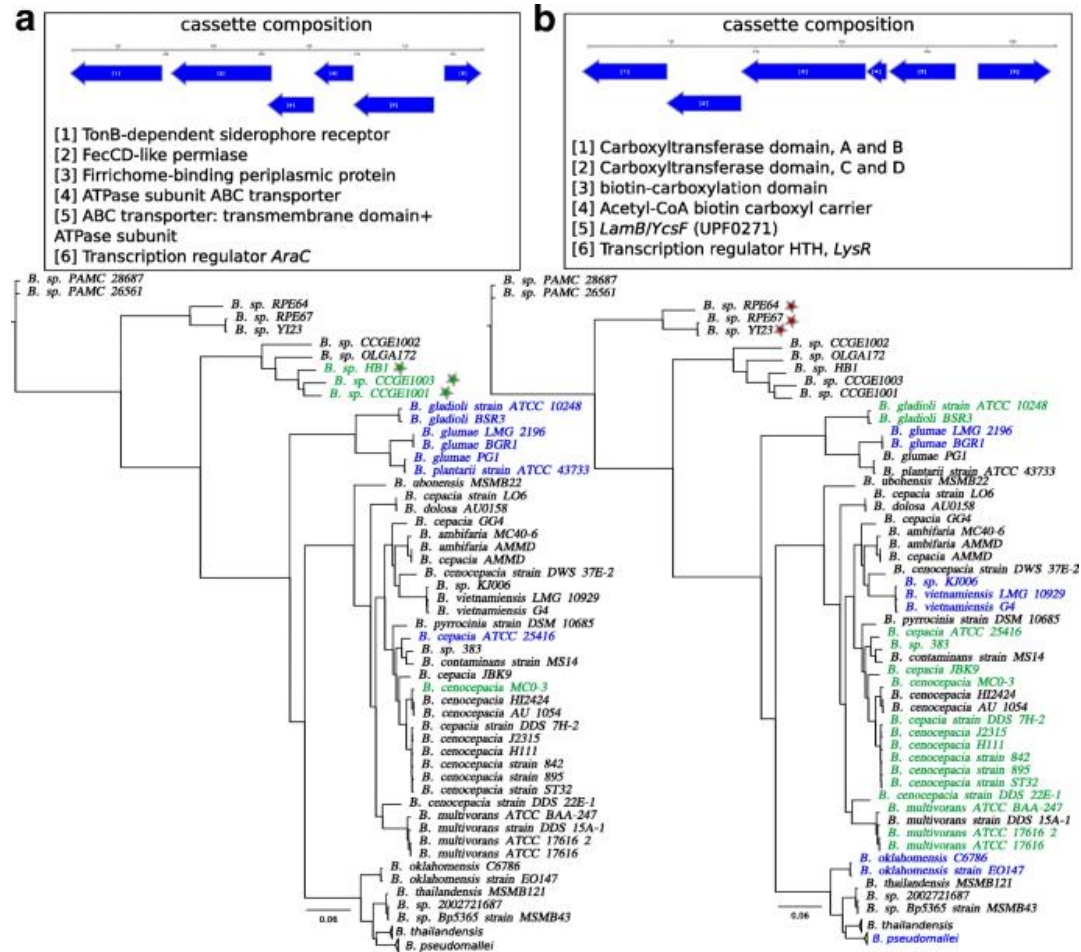
0.0010



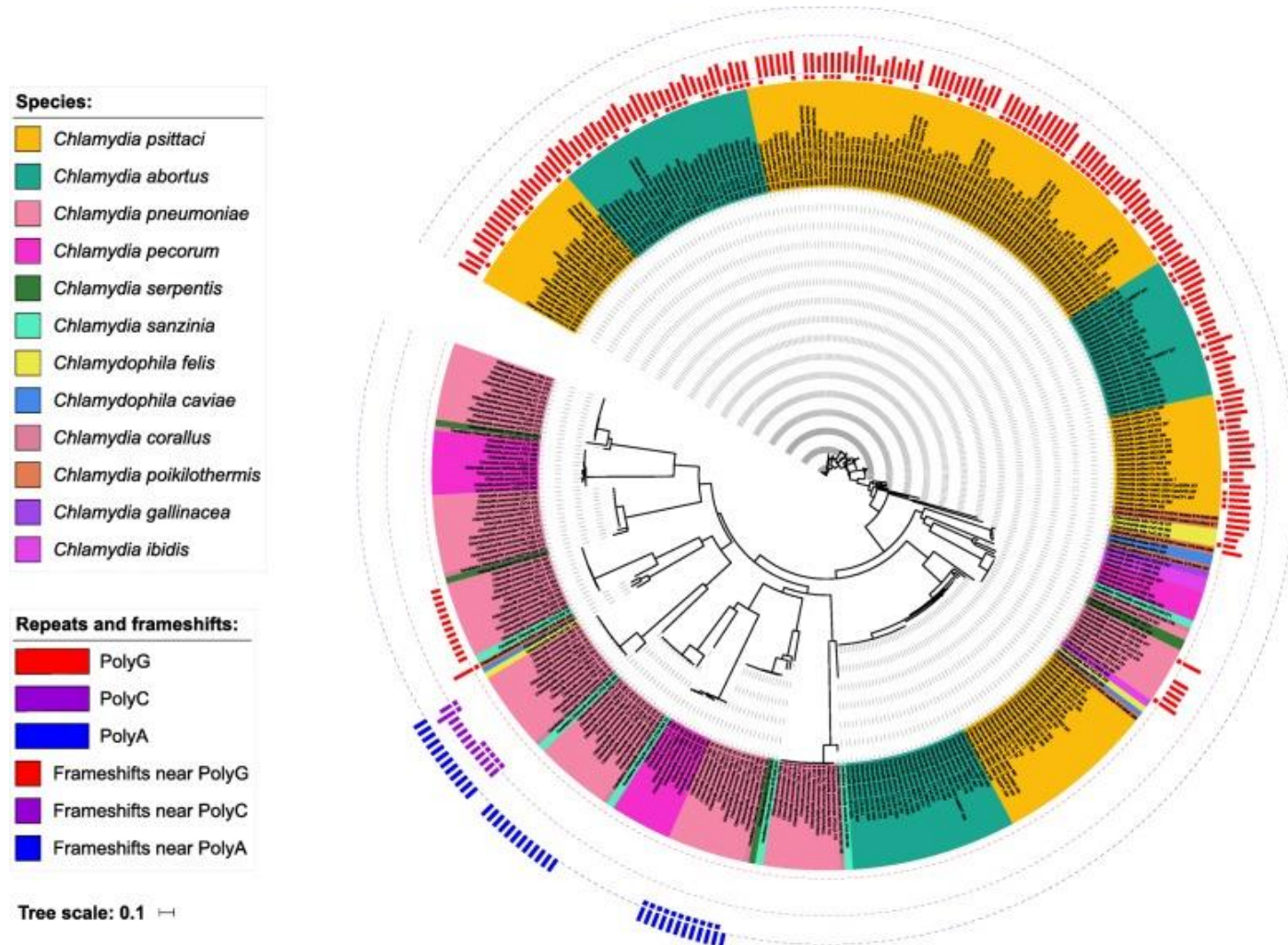
And in
Burkholderia
pseudomallei



... and in many other pathogens (work in progress)



Phase variations in polymorphic membrane proteins of *Chlamydia* spp.



Phylogenetic tree of OG1 (subfamily of PmpG).

- Color – chlamydia species
- Lines – polynucleotide tracts (polyG, polyA, polyC).
- Dots – frameshifts / pseudogenes

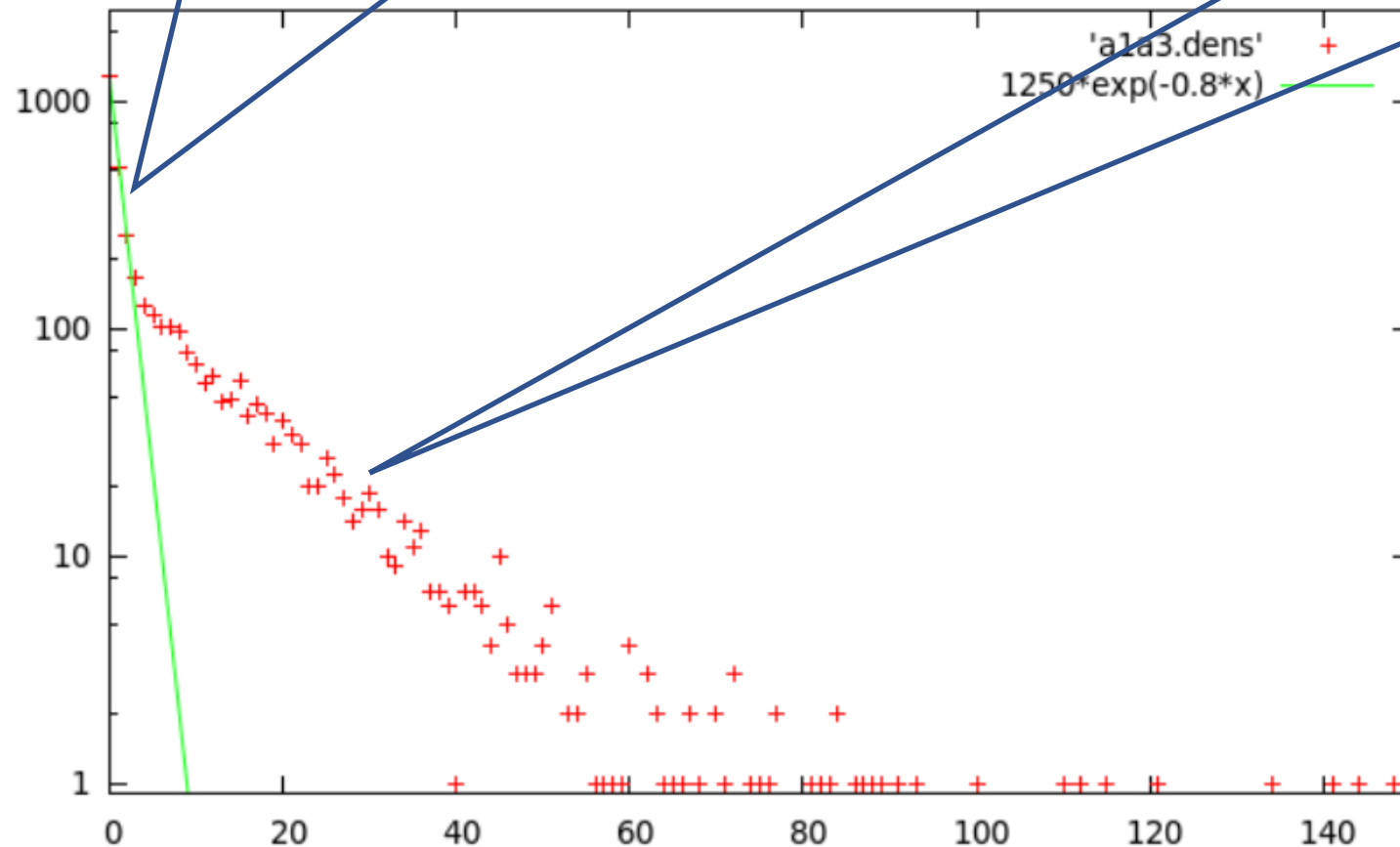
Homologous recombination

Poisson: vertical inheritance

$$\lambda^k e^{-\lambda} / k!$$

Eglang: horizontal gene transfer

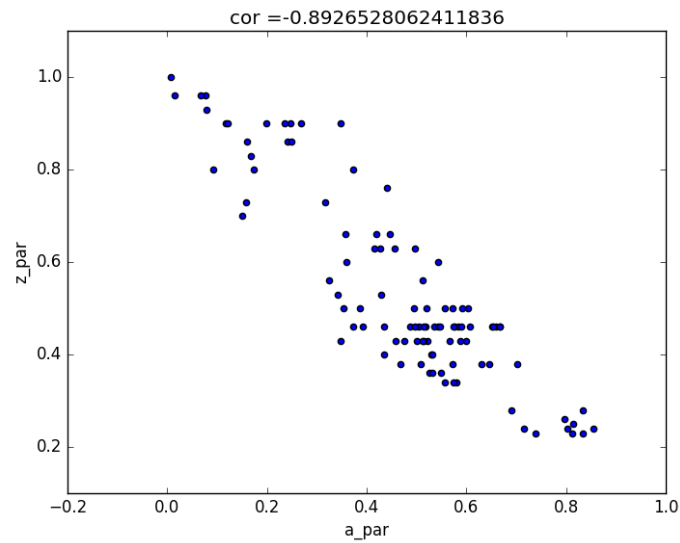
$$\lambda^k (kt)^{k-1} e^{-\lambda kt} / (k-1)!$$



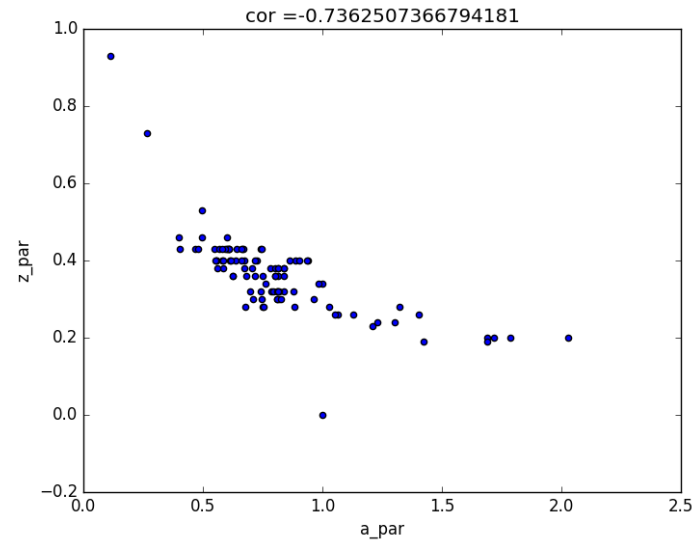
Two strains of *E. coli*:
Histogram of # of
mismatches in 1 kb windows

% of HGT fragments
~ time to the common ancestor
... as it should be

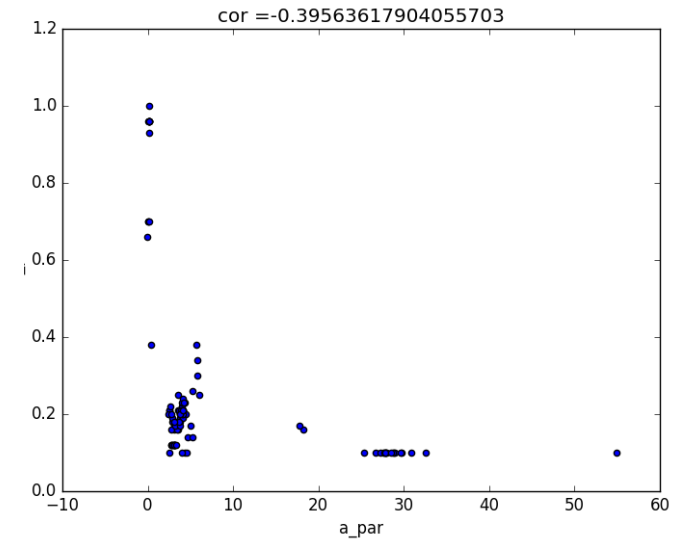
E. coli phylogroups_A



B1

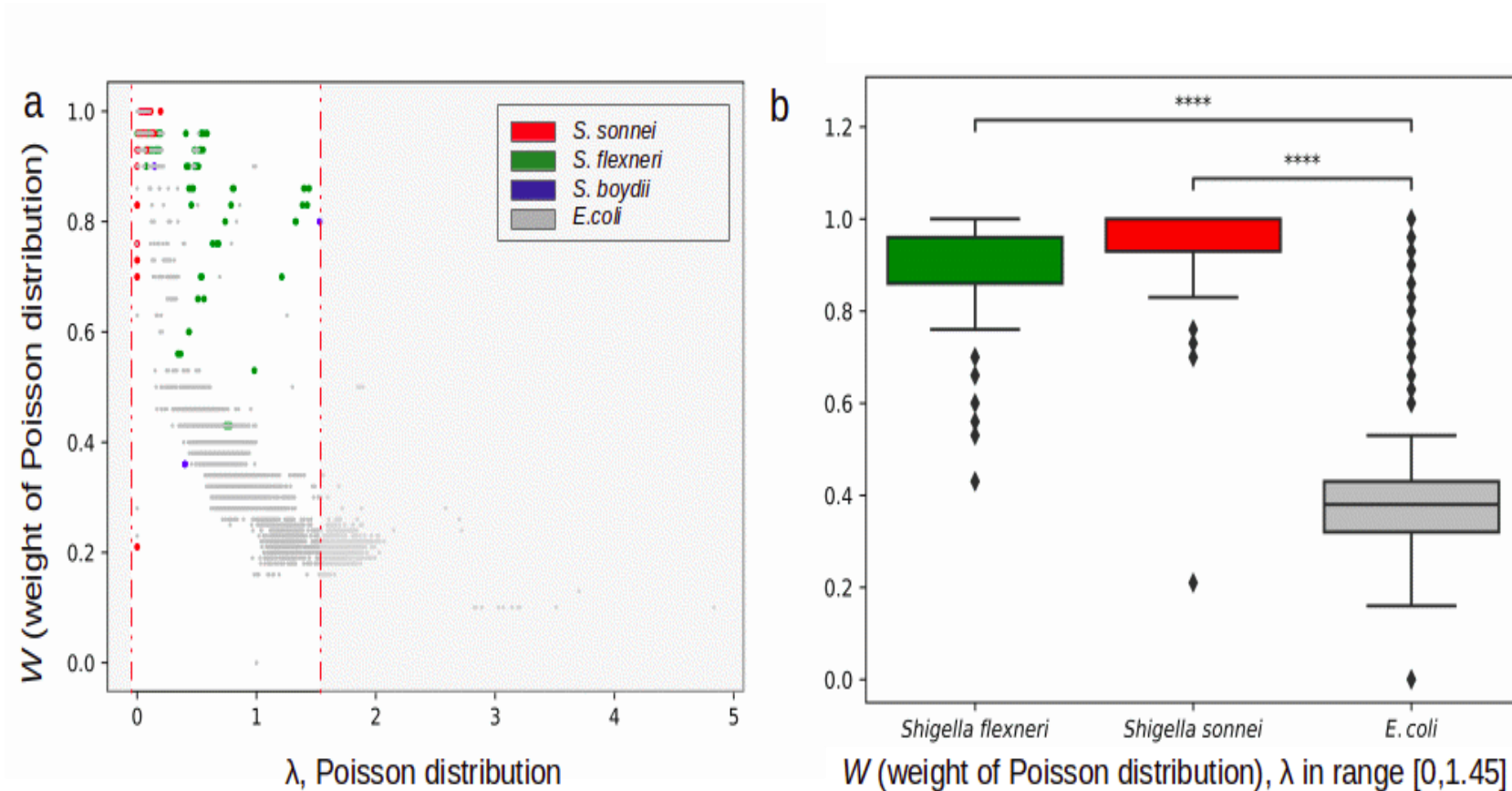


B2

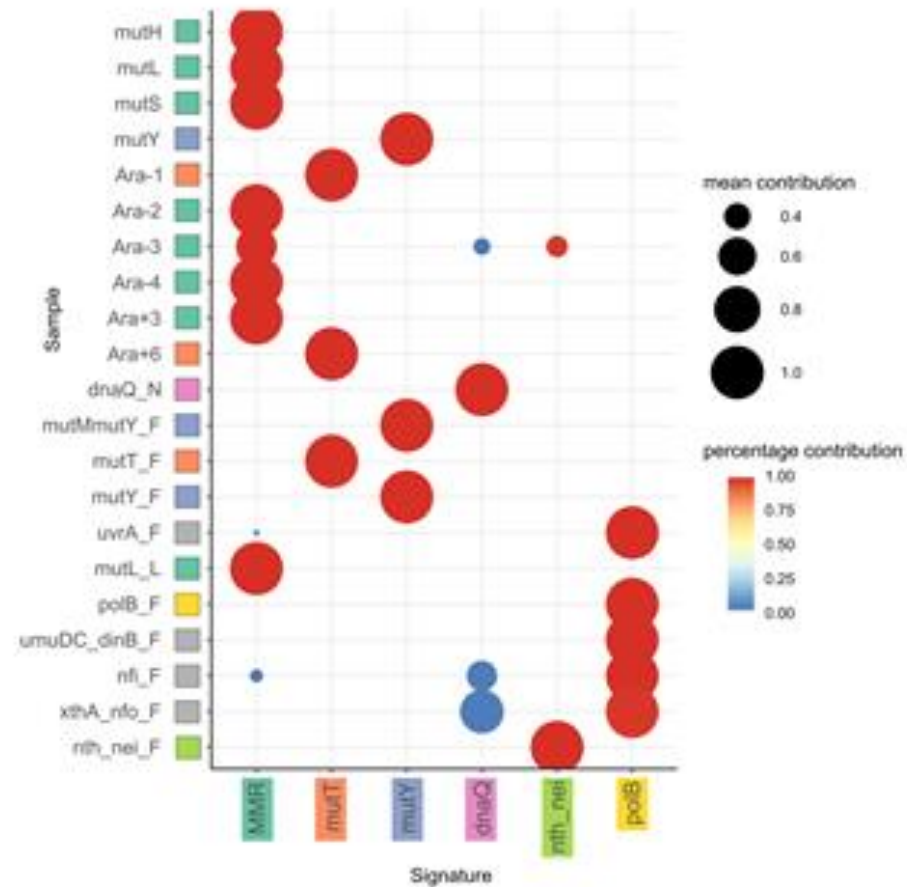


X-axis: Poisson lambda. Y-axis: Poisson weight

Decreased rate of homologous recombination in *Shigella* spp. (due to frequent inversions)

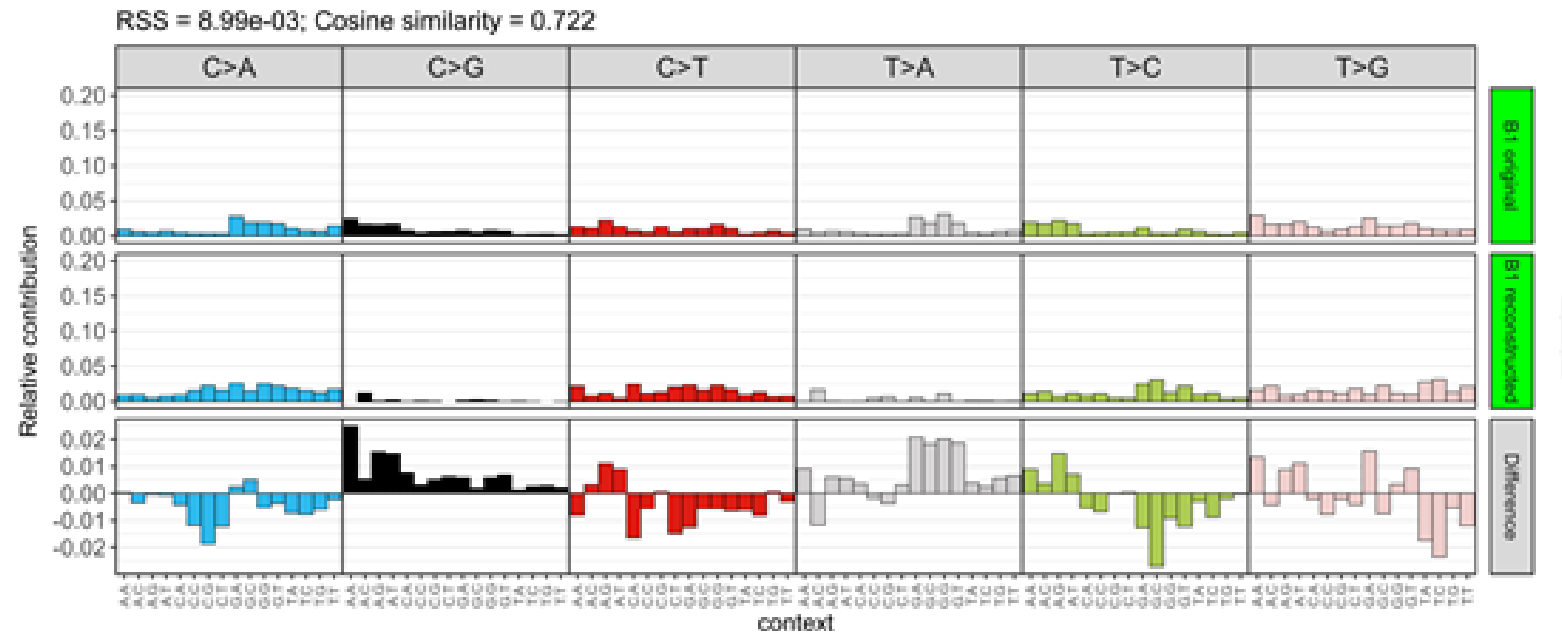
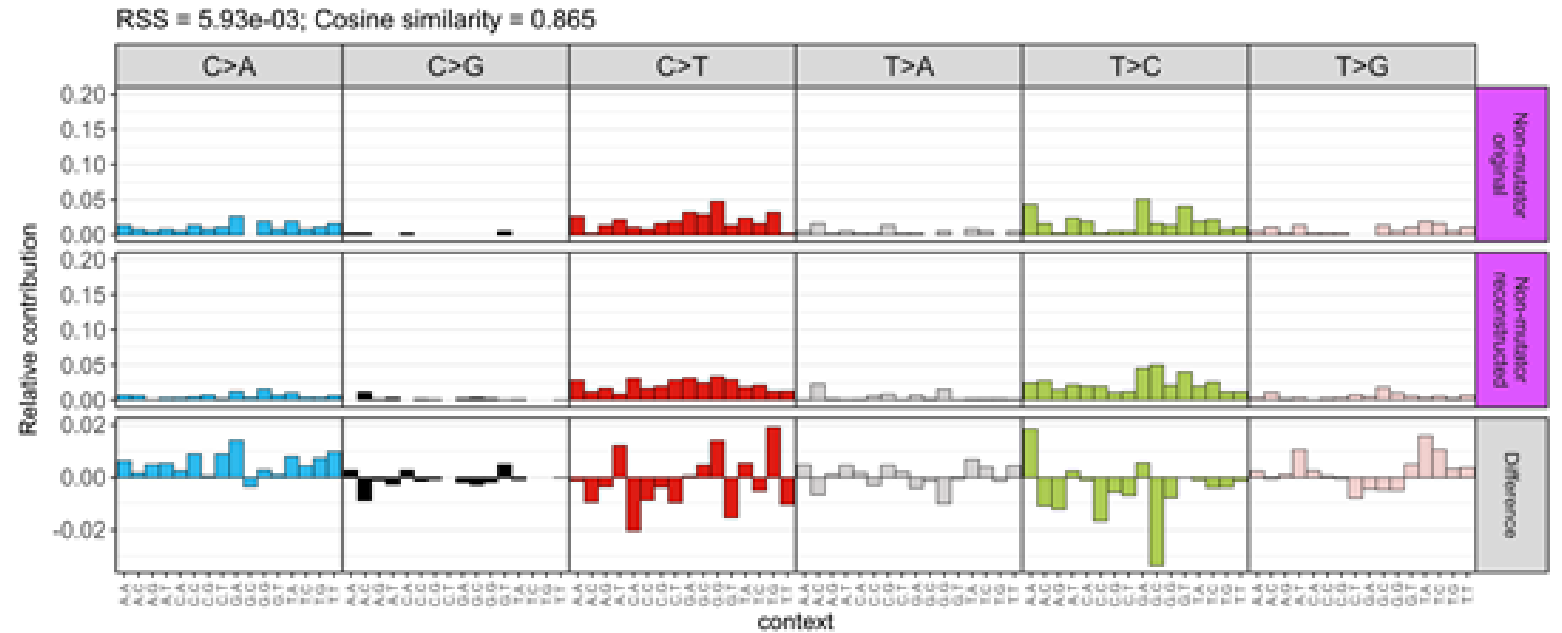


Mutation signatures (*E. coli*)

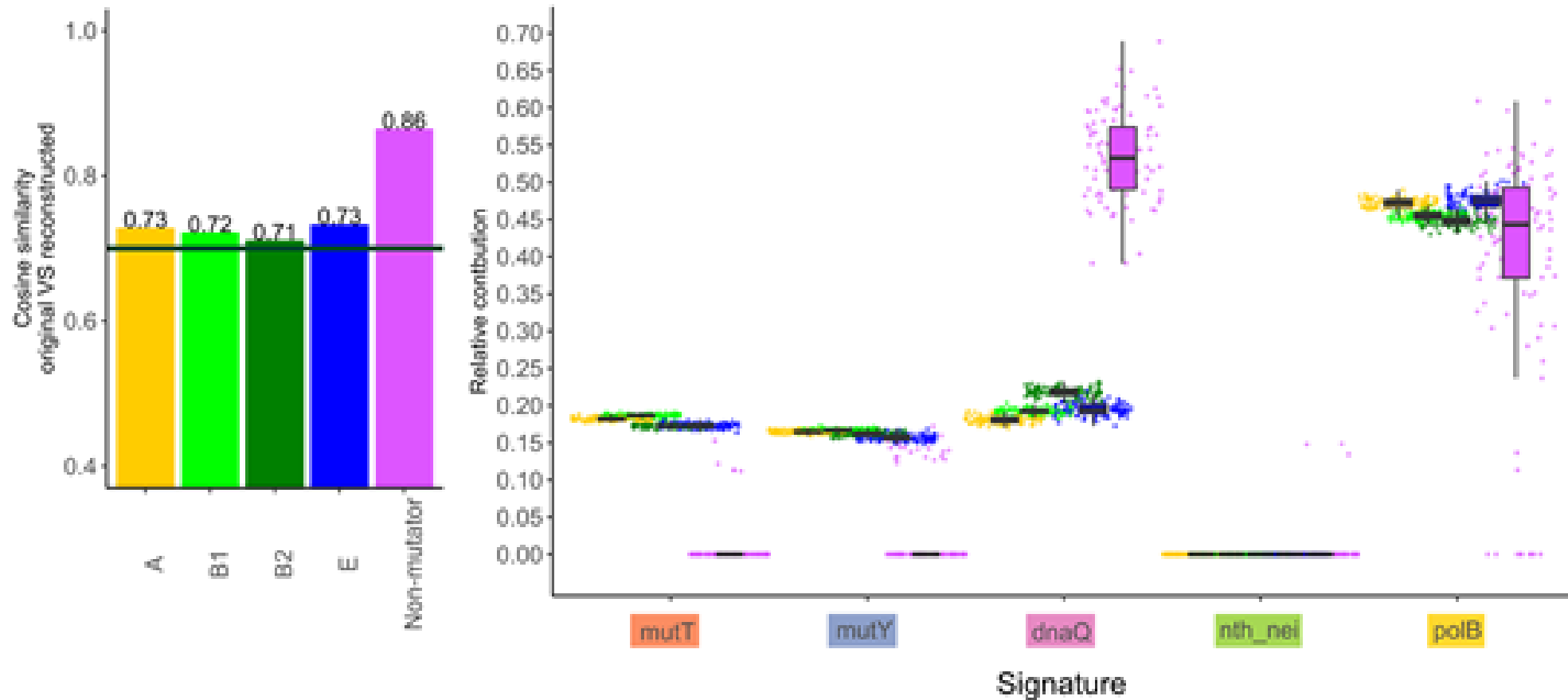


Mutation signatures their fraction in mutator strains

Mutation profiles of wild and lab strains



Contribution of mutation signatures to polymorphism of lab (non-mutator) and wild strains



Comparative analysis of regulatory sites

A.

```

EC  TGCTTTTTACTTTGAGCTACATCAAAAAAAGCTCAAACATCCTTGATGCAAAGCACTATATATAGACTTTAAAATGCGTCCCAACCCAATATGTTGTATT
ST  TGATTTTTACCTTGTTCTACATCAATAAAATTGCAAACATCCTTGATGCAAATCACTTACATATAGACTTTAAAATGCACGCCGACCCAATATGTTGTATT
KP  ACCTTTTTACCTTGTTCTGGGTCAATAAAATCGCAAACATCTTTGATGCAAATCACTTACATATAGAACTTTAAAATGCGCCTCGGCCCAACATATTGTATT
      *****  ***  **  *****  *****  *****  *****  *****  *****  *****  *  *****  **  *****

EC  AATCGACTATAATTGCTACTACAGCTCCCCACG--AAAAGGTGCGGCGTTTGTGGATAAGC-GGATGGCGATTGCGGA-AAGCACCGGAAACGAAACGA
ST  AATTGACTACAATTGCTACAACACCTGTTCACT--CGACACAAGGTGAATTGTGGATAACCTGGGTCAGGATTGCGGG-AAGTCATTGGAAAAGAGATGA
KP  AATCGTCTATTAT-GTACCATATCTTGTGCGATGTCTGGCGGTGATGAGATGTGGATAAAACGGGCGGATCCGAAGGTAAACAGCACGAGCCGTAGCGT
      ***  *  ***  **  *  **  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

EC  AAAAAACCGGAAACGCCTTTCCCAATTTCTTGTGGATAACCTGTTCTTAAAAATATGGAGCGATCATGACACCGCATGTGATGAAACGAGA
ST  ATAAACCTGTTA-TGGCTTCCCCCGGCTCTTGTGGATAACCTGTTCTTACAAATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
KP  GCAGCGCCTTCG-GGATAACCTCCGCTCTTGTGGATAACCTGTTCT--ATATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
      *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
    
```

B.

```

YP  AACAGGAATAACCC-TAACGCC--AATTTCCTTGTTCTAGGTCAACAATATTGGCTATCAGTTGACTGTCACTCATCCAGATACCCATATATAGTGTCT
YE  AACAGGAATAAACC-TAAAGCT--GATTTCCTTGTTCTAGGTCAATTAT-----GTTGACTGTCACTTCGCCATTACCCATATATAGTGTCT
Eca  AAGTTCGATTTATCTACTAGGGAGGAATTTCTTTGCTCTACATCAATTTTGCAGCGATAAAAGTGCAAACACCCCTACGCCAATTTTCAATATATAGTGCCT
Ech  AAGACTGATTTCTCTACGATGCCGGAATTTCCTTGAGCCAGGTCAATTCTAACGCAATAAAACCGGGTCCCCCTCCAGGCGAATTCAATATATAGTGTCT
      **  **  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

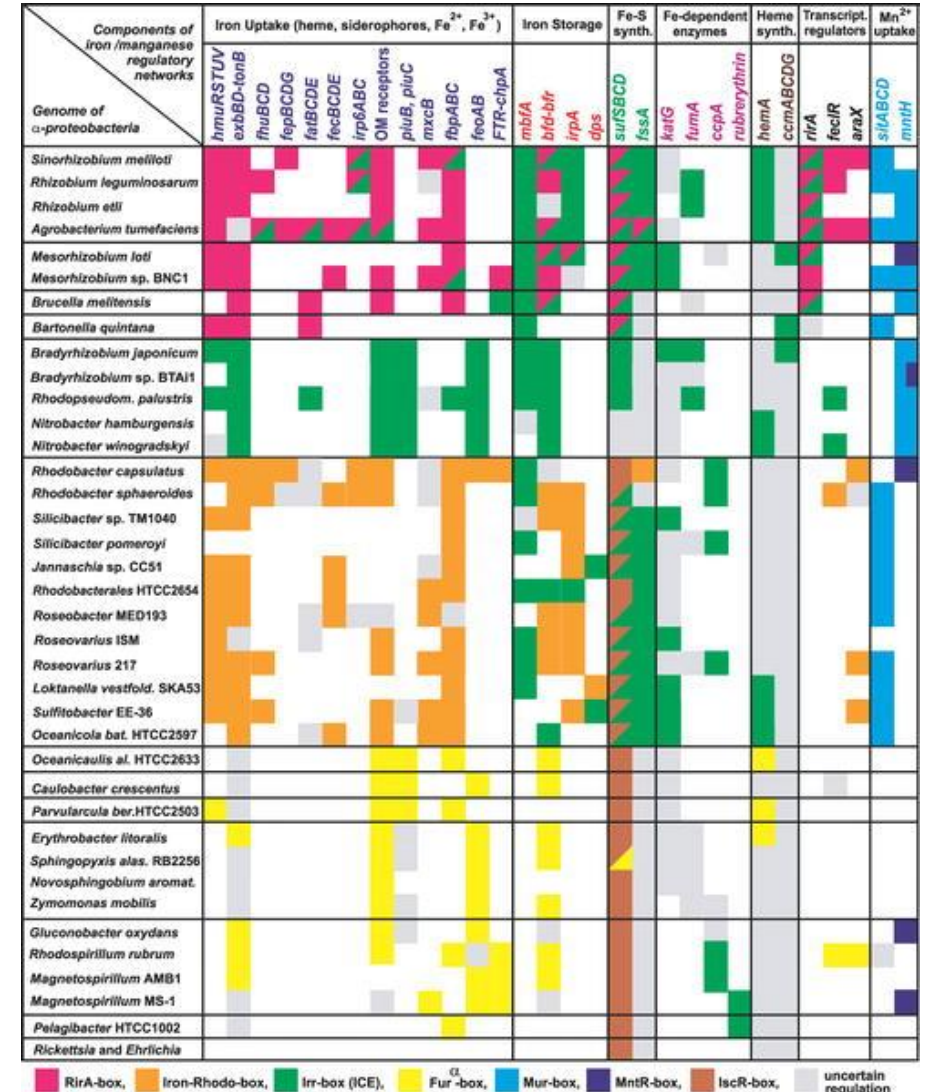
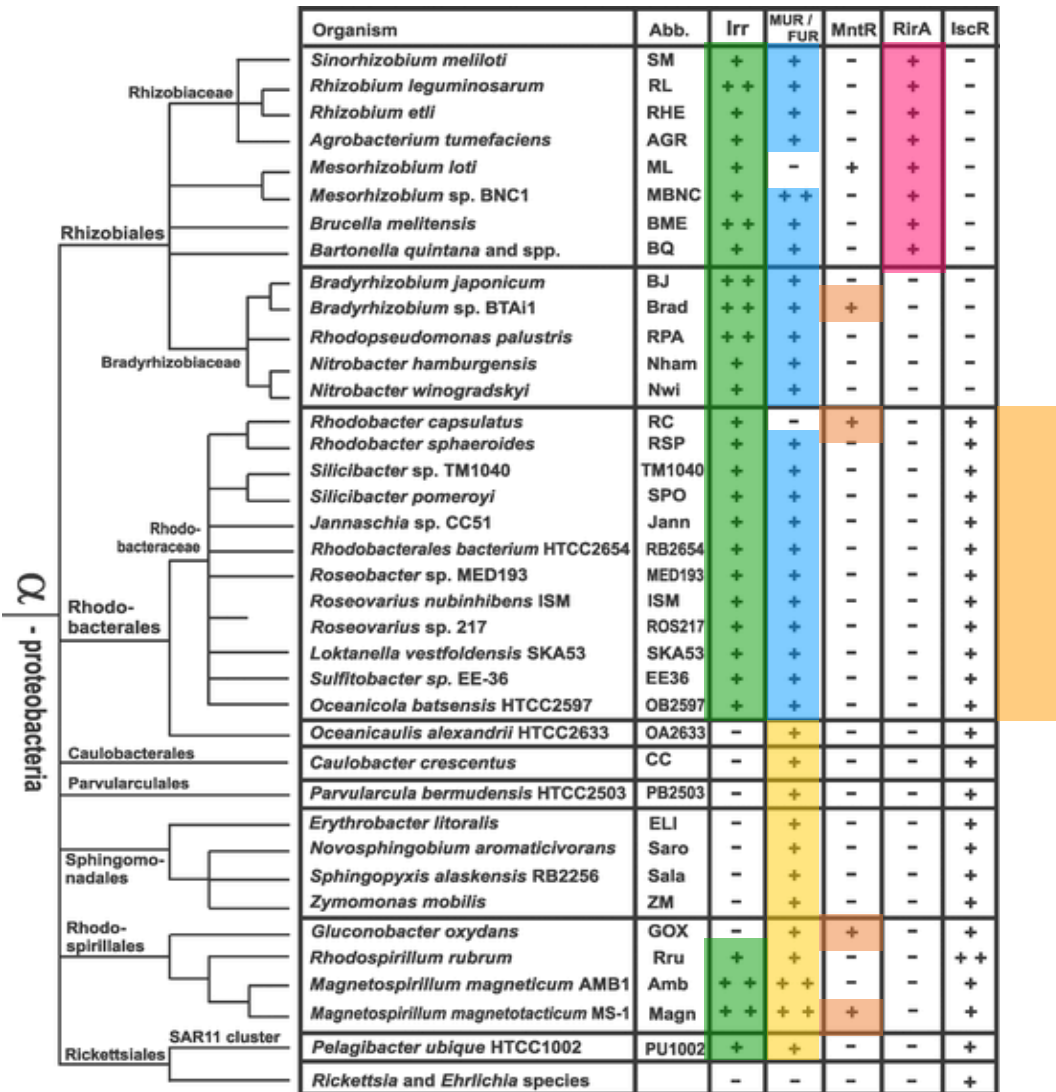
YP  ATAATATTTTAAGCATCTATATGTAGTAGTTATCCACAAAAAGCATCCACATCCCCCTCGCAGCCCTGATGTGCTGCGGGTTGC-CTTGTGGATAA----
YE  ATAGTAATTACGATTACCTATATGTAGTAGTTATCCACAAAACCATCCACA-CCCCCTCGCAGCCCTGATGTGCTGCGGGTTGC-CTTGTGGATAAGATGG
Eca  ATCCGTAAACATTACCTACATATAGTGTTTATCCACAAAAGCATCCACA-GCCCTCTGTAAACCTTGGCAGTTACGGTTCTCGCCTGTGGATAAC----
Ech  ATCTTTTACAAGAACCTACATATTTGTGTTTATCCACAAGACATCCACA-GCC-TCCGCACCCGTTGTGACCCGCGGCTTCCGTCTGTGGATAAC----
      **  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

YP  -----CCCTATGCGGCGGTATACAGGAGTGACATTGTGAAAACAGTAGTGATTAAACGGGACGGCTGCCAGGT
YE  TTTTGGGGCTAATCCTACGCGGCAGGATACAGGAGCGACATTGTGAAAACAGTAGTGATTAAACGGGACGGTTGTCAGGT
Eca  -----CTTTCCAG-----AGGAAGAA-AACTGAAAACAGTAGTGATTAAACGGGACGGTTGCCAGGT
Ech  -----ATCAACAAAGGAAGAACACCGAGGAACAAC---ATGAAAACAGTAGTGATTAAACGGGACGGATGTCAGGT
      *****  *  *****  *****  *****  *****  *****  *****  *****  *****  *****  *****  *****
    
```

Close genomes (~ genus) – regulatory sites as conserved islands.

Distant genomes (~ family, order) – the presence of a site is conserved

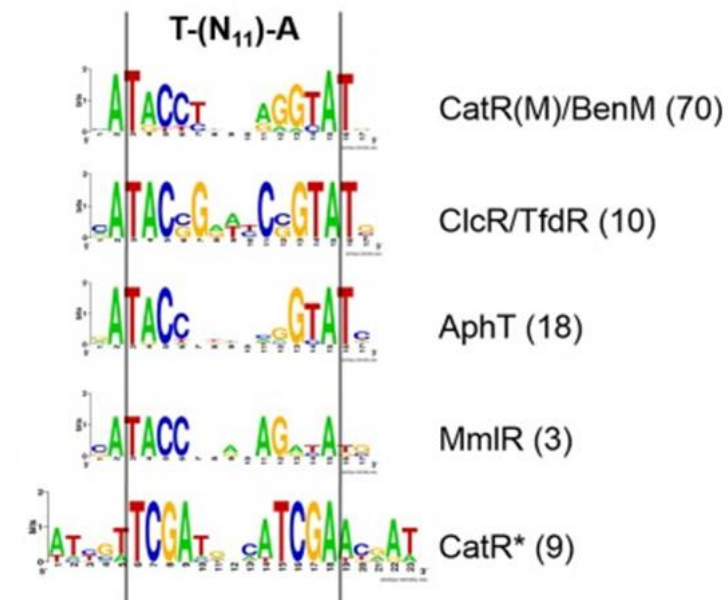
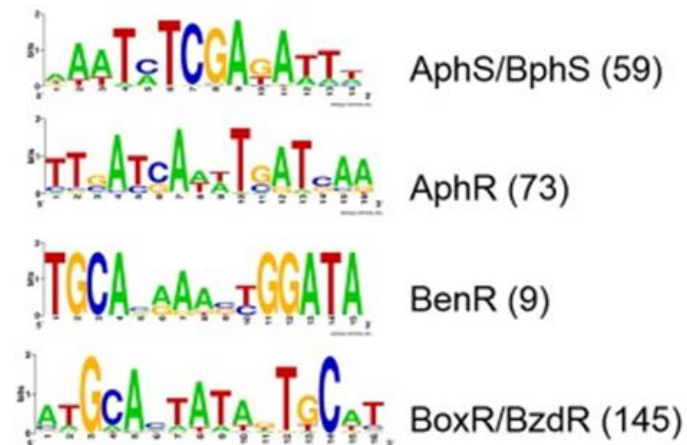
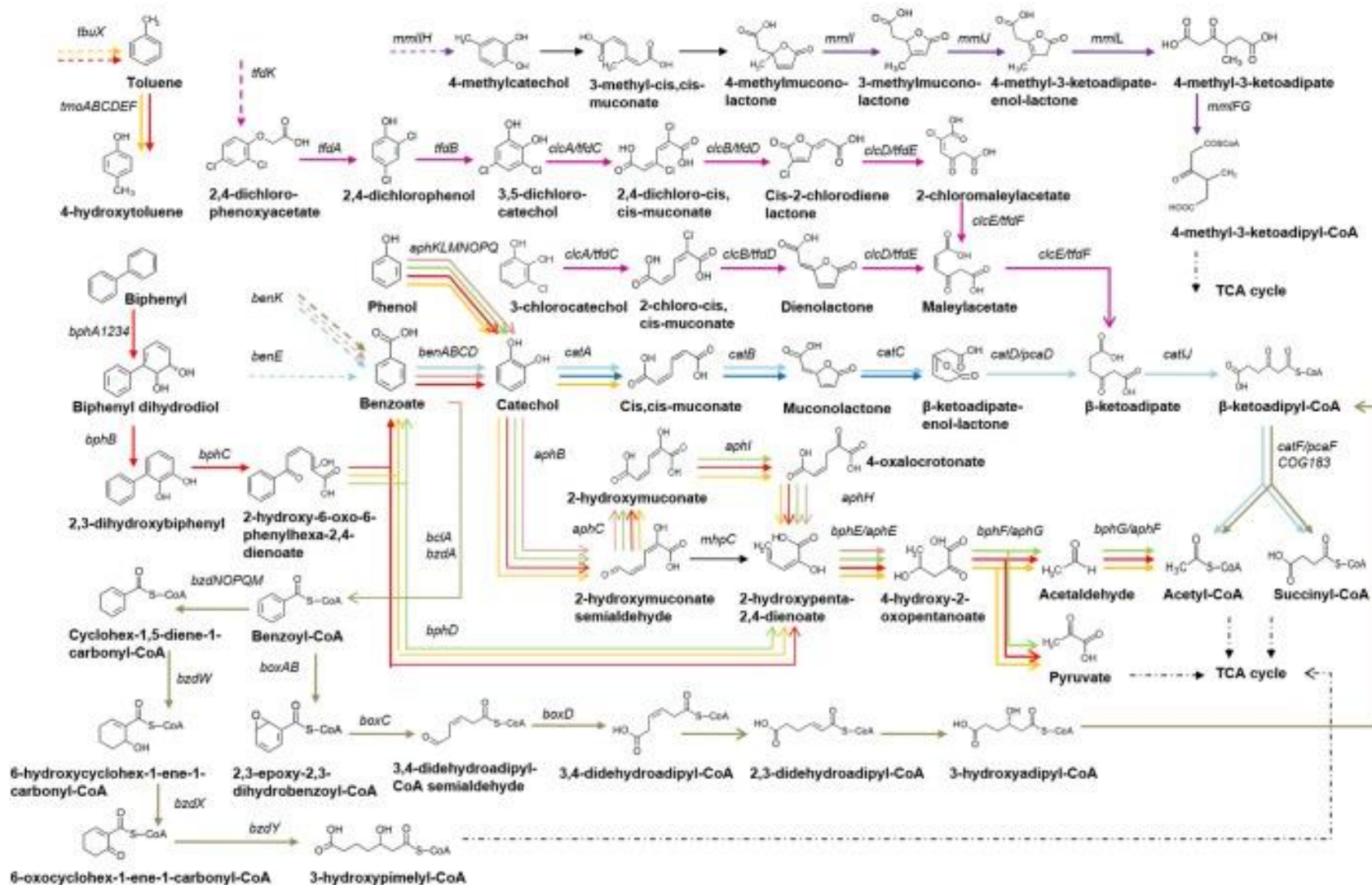
Plasticity of regulatory interactions (iron metabolism in alpha-proteobacteria)



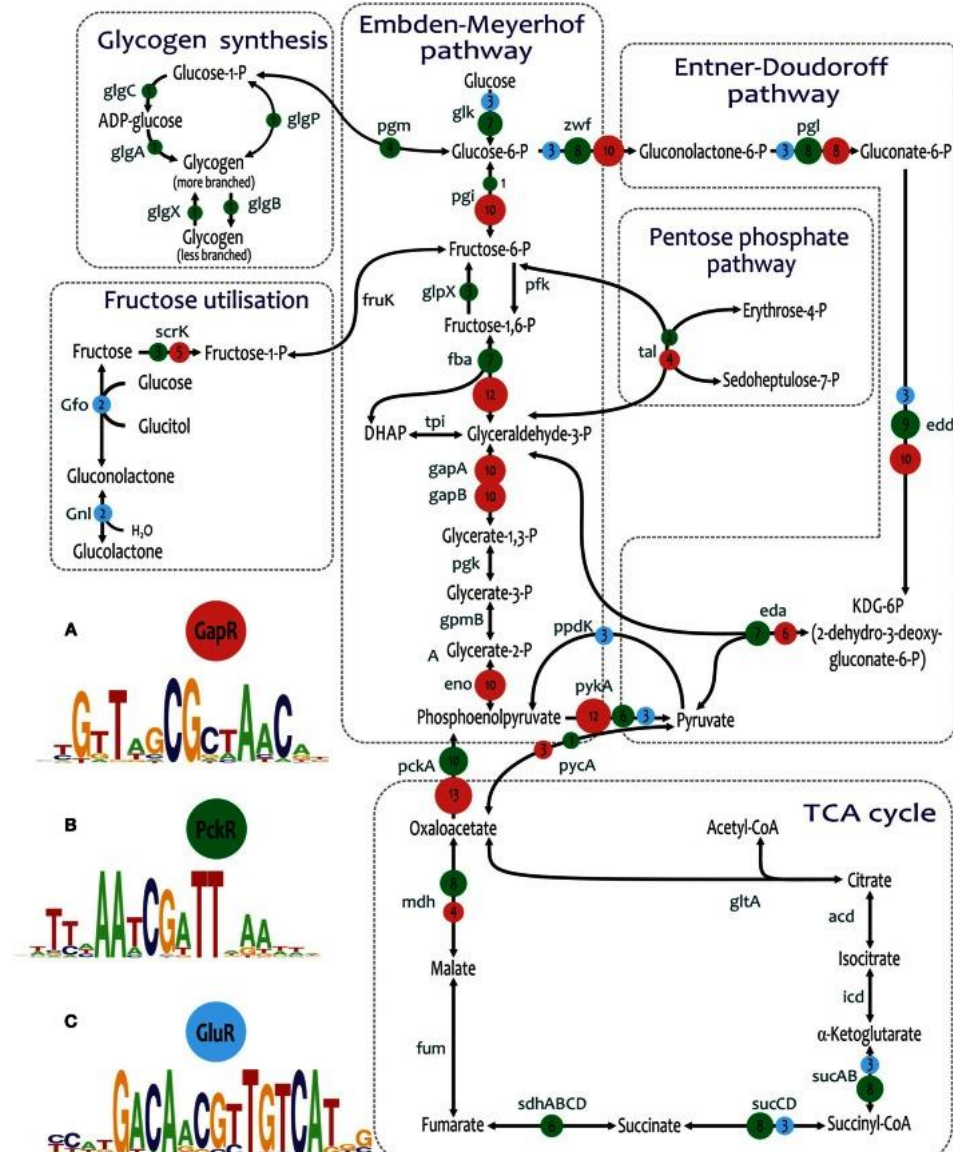
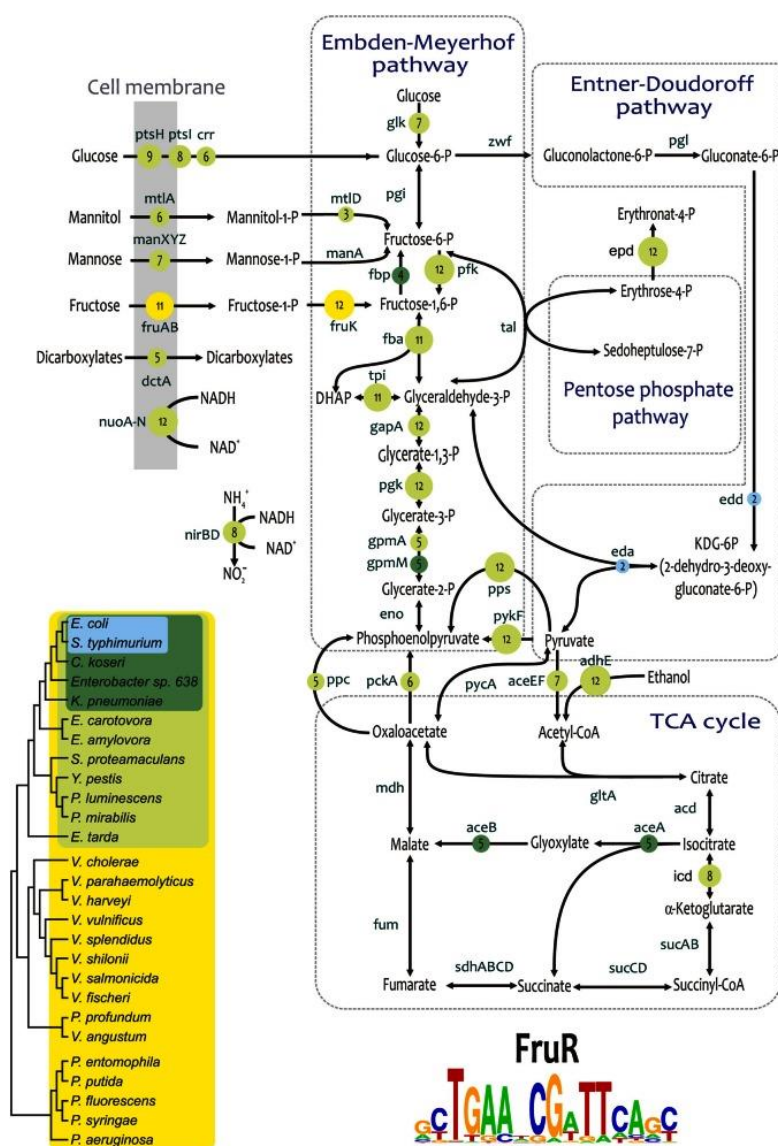
Legend: RirA-box (pink), Iron-Rhodo-box (orange), Irr-box (ICE) (green), Fur-box (yellow), Mur-box (blue), MntR-box (purple), IscR-box (brown), uncertain regulation (grey).

Evolutionary plasticity

(catabolism of aromatic compounds in gamma-proteobacteria)

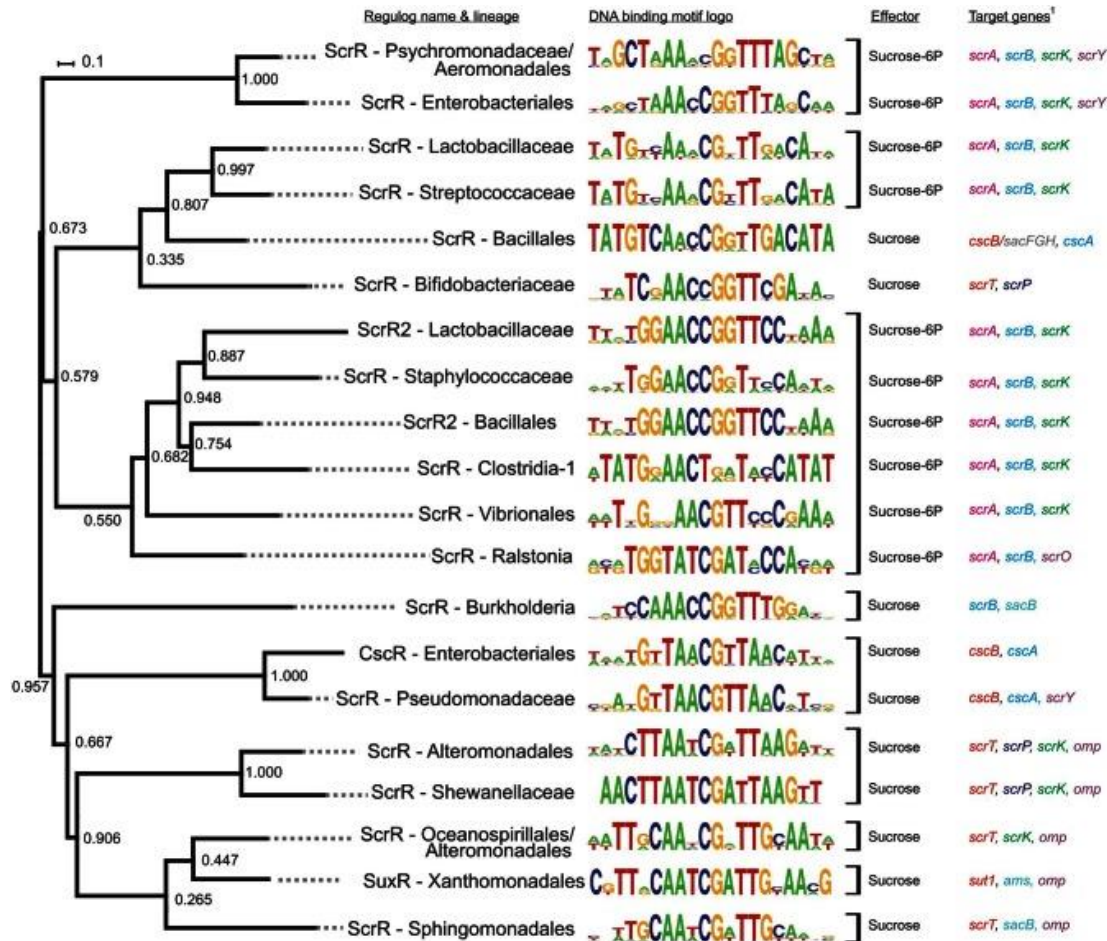


Expansion of regulatory systems (central metabolism of γ - and α -proteobacteria)



Evolution of binding motifs

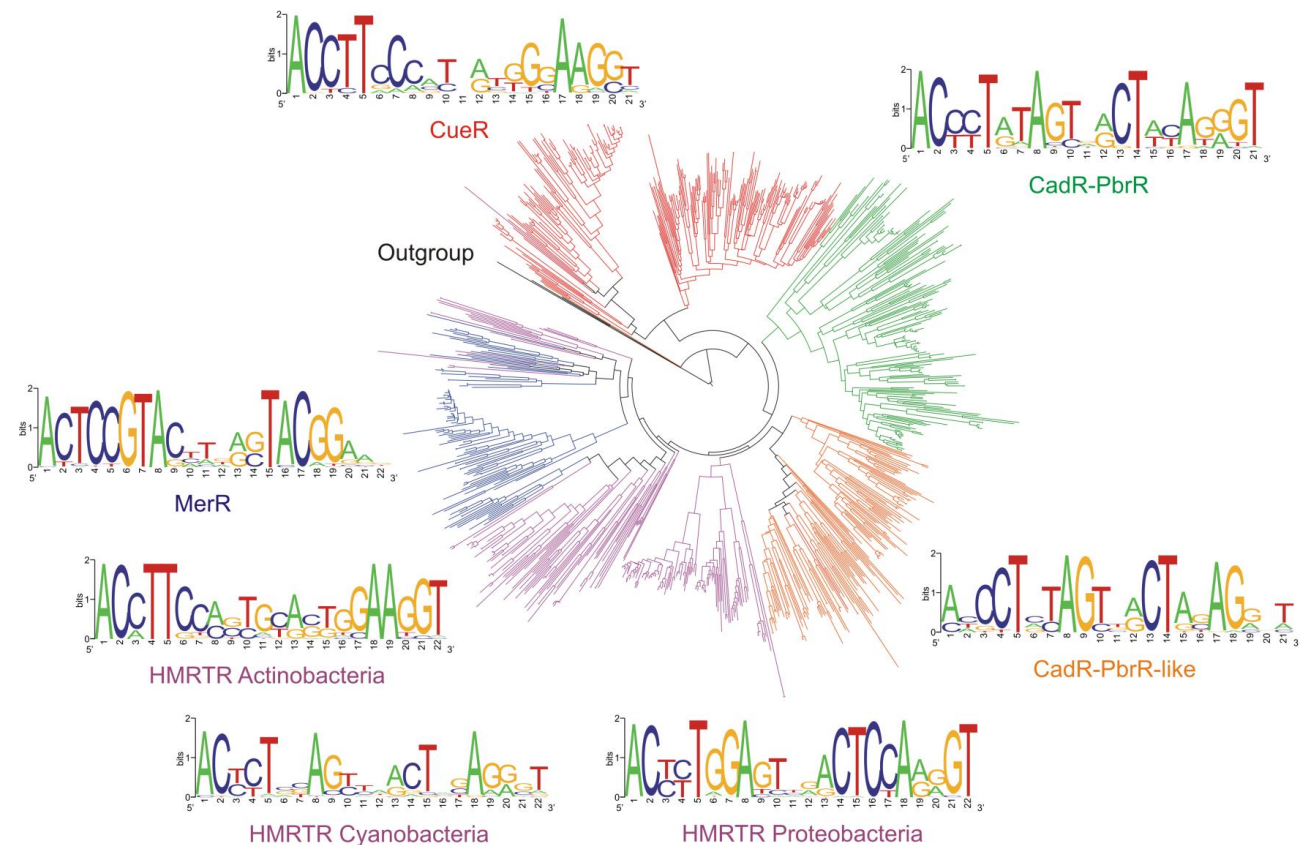
LacI/ScrR



¹Sucrose-specific transporters: permeases, PTS, porins; Sucrose or Sucrose-6P hydrolases, Sucrose phosphorylase, Levansucrase, Fructokinase

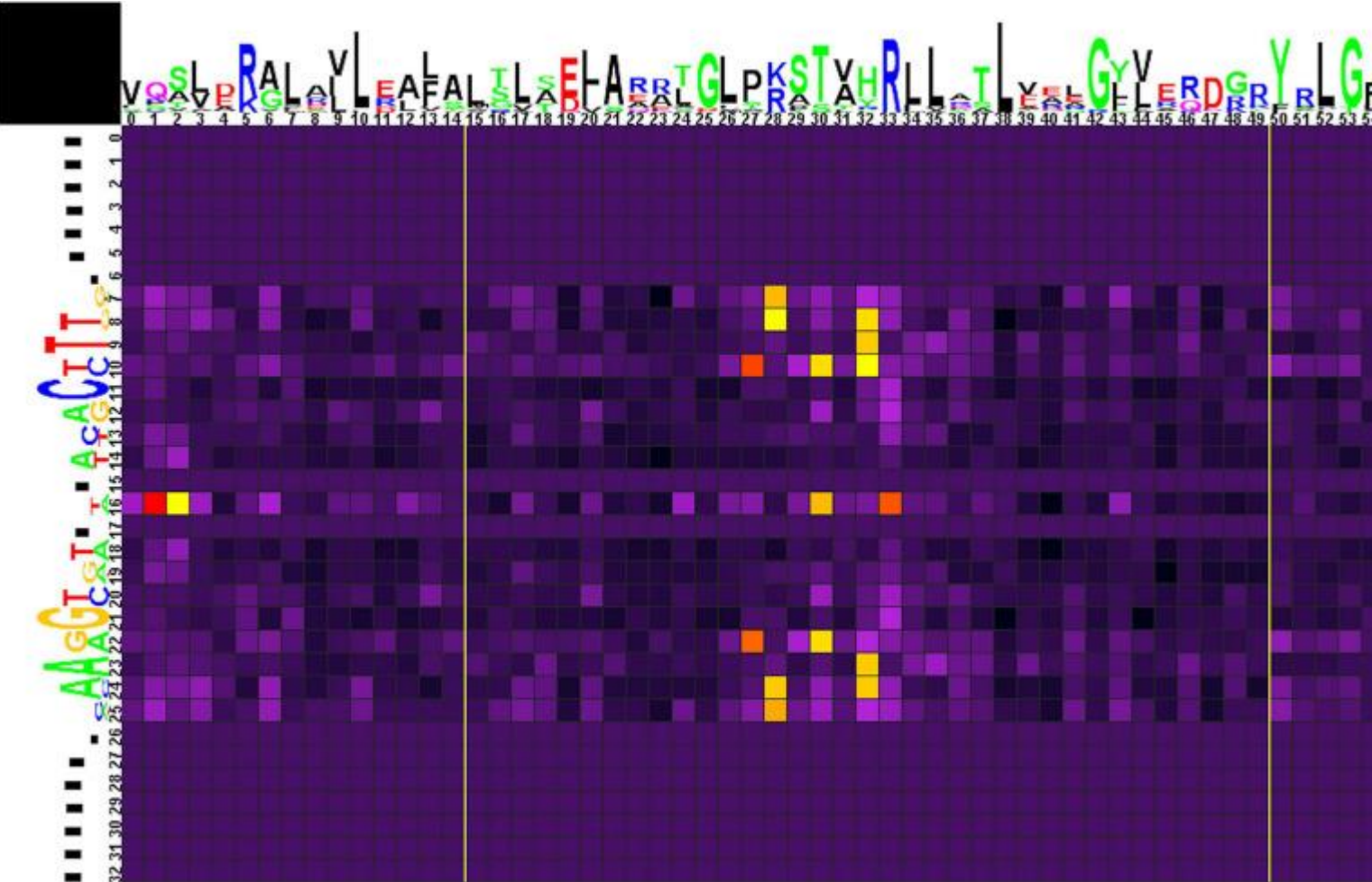
MerR

Phylogenetic tree of HMR transcriptional regulators from MerR family

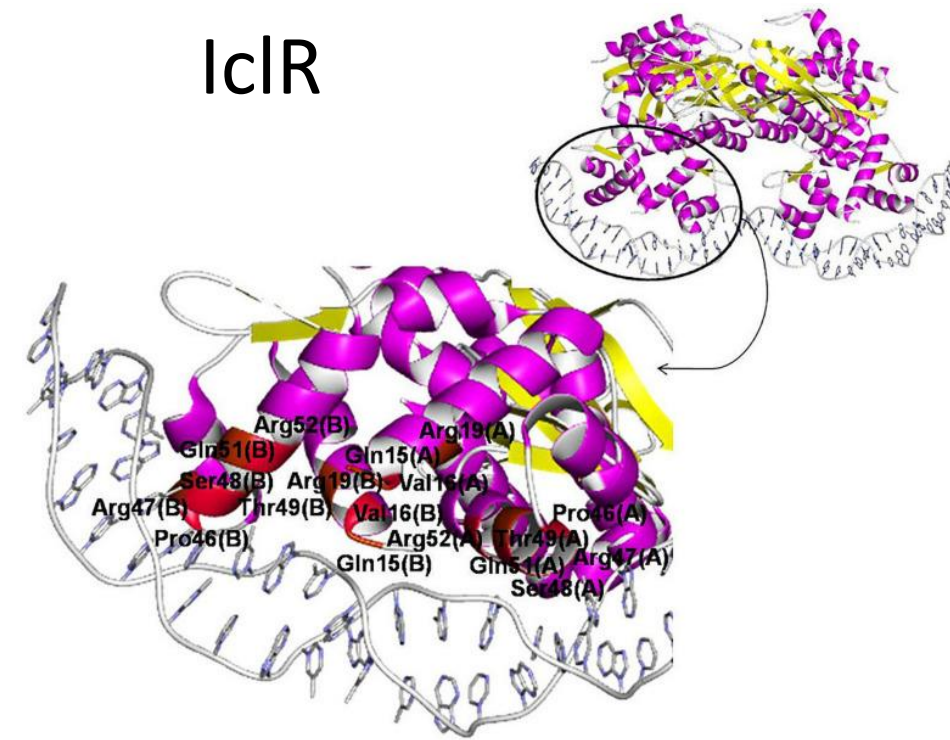


First 3 positions in sequence logos are the 3' end of 10 promoter boxes.

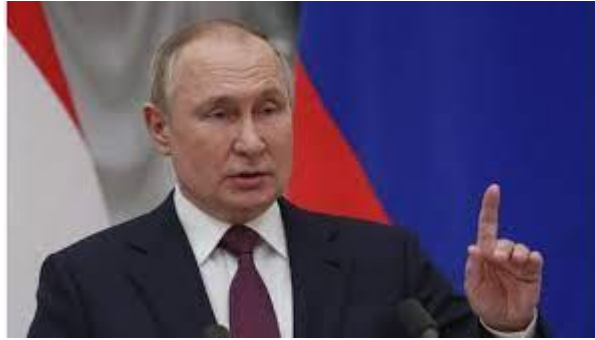
Co-evolution of TFs and motifs



ICLR



- Robert Afasizhev (IITP) – homologous recombination
- **Olga Bochkareva** (IITP => IST Austria) – pan-genomes, rearrangements, phase variations
- Natalya Dranenko (Skoltech => IITP) – pan-genomes, IS-elements, phase variation
- **Sofya Garushyants** (IITP => NBCI USA) – tiny genomes, specialists and generalists, mutation signatures
- Ilya Zharov (IITP) – co-evolution of TFs and motifs
- Yuri Korostelev (IITP) – co-evolution of TFs and motifs
- **Mikhail Moldovan** (MSU => Skoltech) – pan-genomes, definition of species
- Darya Nikolaeva (MSU => Skoltech) – tiny genomes, specialists and generalists
- Dmitry Ravcheev (IITP => Galway U. Ireland) – regulatory sites
- **Dmitry Rodionov** (IITP => Burnham Inst.) – regulatory sites, evolution of regulatory systems
- Zaira Seferbekova (MSU) – rearrangements, homologous recombination
- Ariadna Semenova (MSU) – inversions and IS-elements
- Pavel Shelyakin (IITP) – pan-genomes, intergenic regions
- Olga Sigalova (IITP => EMBL, Germany) – phase variations
- **Inna Suvorova** (IITP) – evolution of regulatory systems, co-evolution of TFs and motifs
- **Maria Tutukina** (IITP, Skoltech) – regulation of transcription



2000: Мы знаем цену миру, знаем, что мир — это, прежде всего, прочность экономики и благополучие людей.

2009: Современная цивилизованная политика должна основываться на общих моральных и единых правовых принципах.

2017: Россия всегда будет на стороне сил мира, с теми, кто выбирает путь равноправного партнерства, кто отрицает войны как противные самой сути жизни и природе человека. Мы обязаны передать потомкам стабильность и мир на планете.

2019: Война принесла столько невыносимых испытаний, горя и слез, что забыть это невозможно. И нет прощения и оправдания тем, кто вновь замышляет агрессивные планы.