Bacterial comparative genomics: the tips and tricks

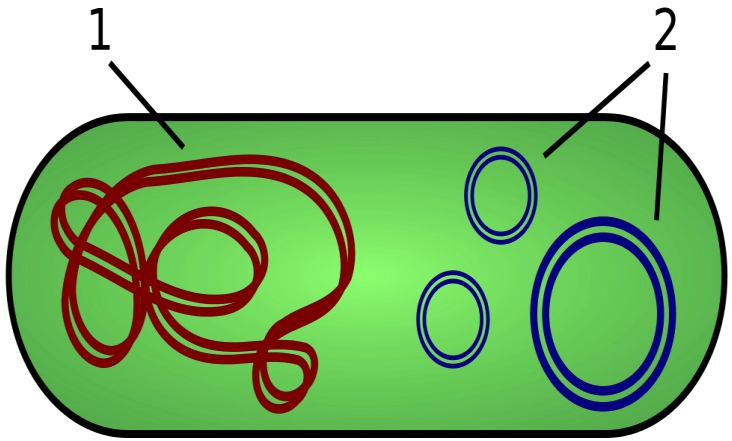
Olga Bochkareva

postdoc, Kondrashov group

Bacterial genome structure

1. Chromosome

- circular, double-stranded DNA molecule
- 1Mb – 6 Mb
- one origin of replication
- carry house-keeping genes and a lot of others



2. Plasmids

- circular, double-stranded DNA molecules
- 15 kb – 200 kb
- replicate independently
- carry genes that benefit the survival of the organism.

3. Secondary replicons
(megaplasmid/chromid)

- circular, double-stranded DNA molecule
- 300 Mb – 3 Mb
- plasmid type origin of replication
- may have small part of core genes
- is subjected to regulation by chromosomally encoded mechanisms

Bacterial genomic data

Bacteria

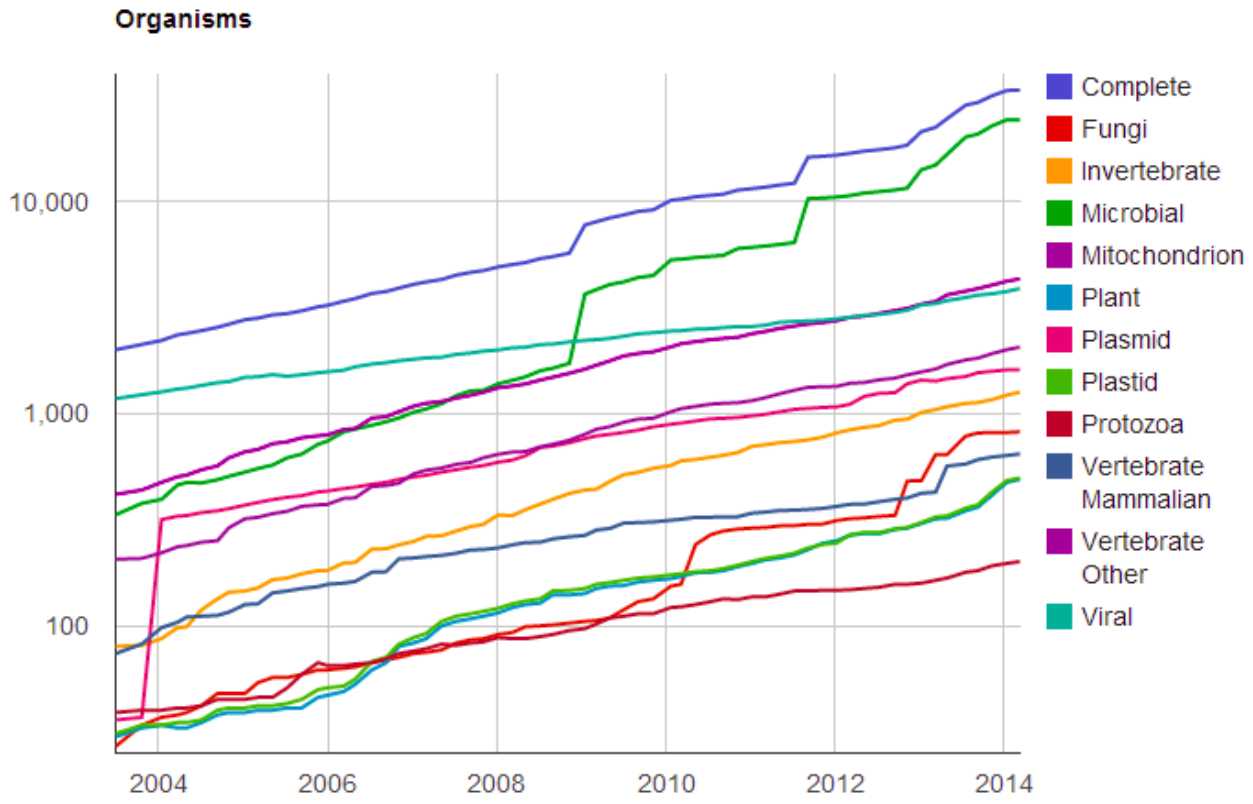
479 320 assemblies

↓

16 527 complete genome assembly level

↓

1 631 representative genomes

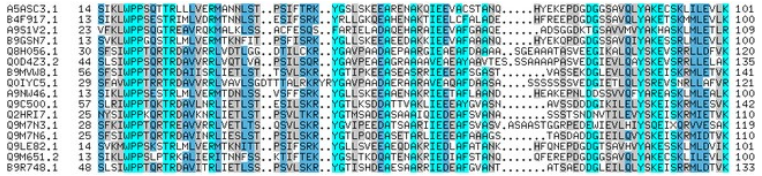


https://www.ncbi.nlm.nih.gov/refseq/statistics/

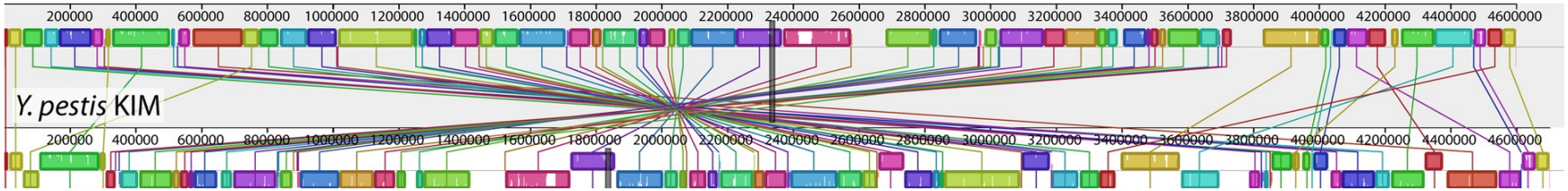
How do prokaryotes evolve?

What can we observe in genomes?

- Mutations – genes sequences



- Gene gains/losses – genes content
- Rearrangements - genes order



What are the mechanisms?

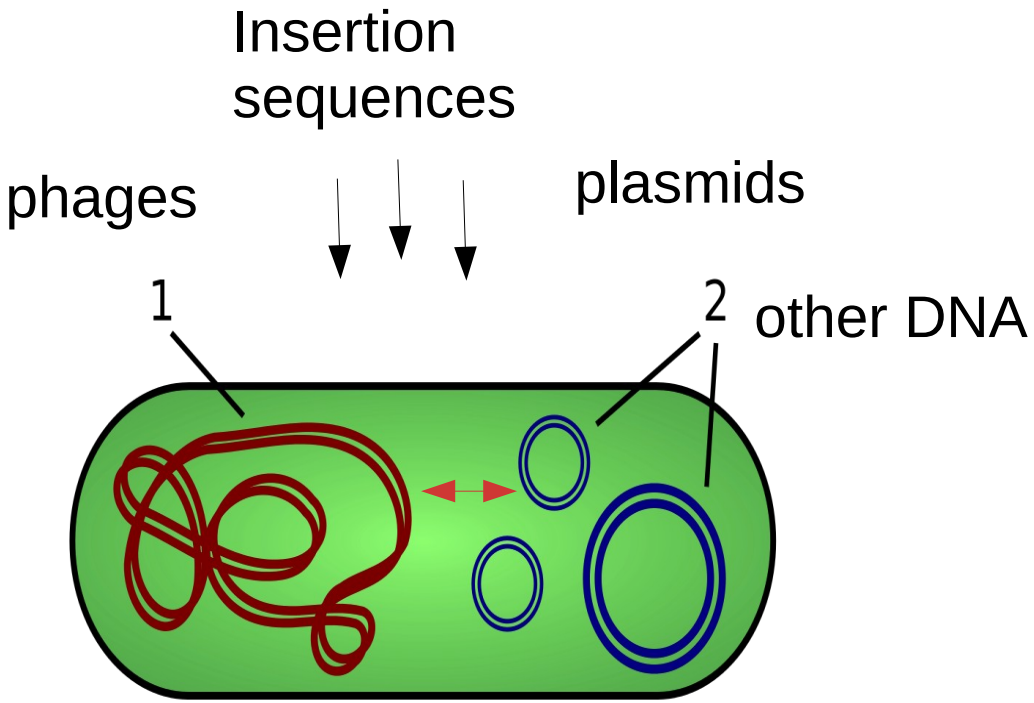
1. Homologous recombination

2. Intra-genomic recombination

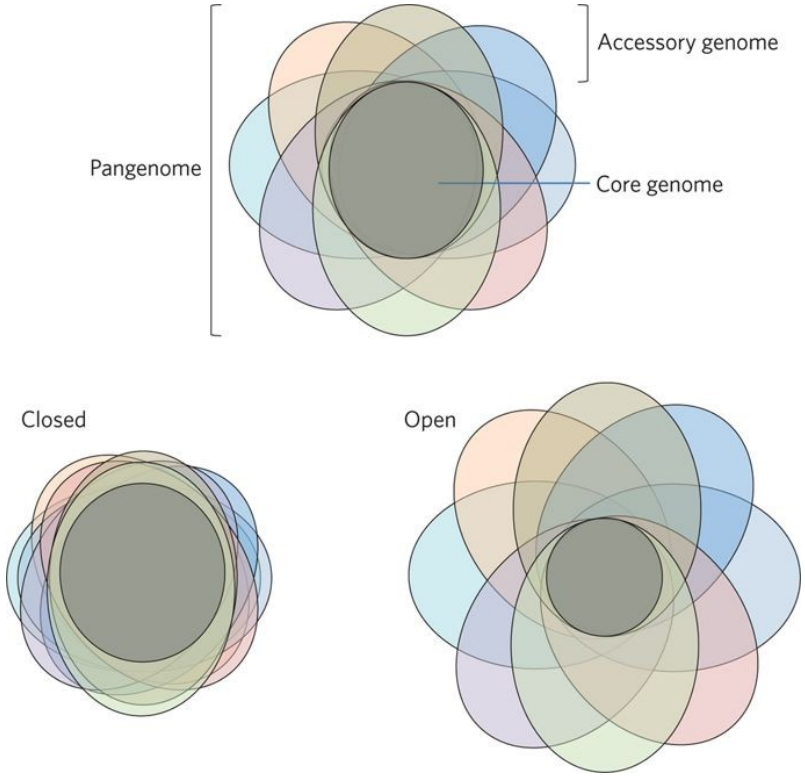
3. Horizontal gene transfer:

- transformation
- transduction
- bacterial conjugation

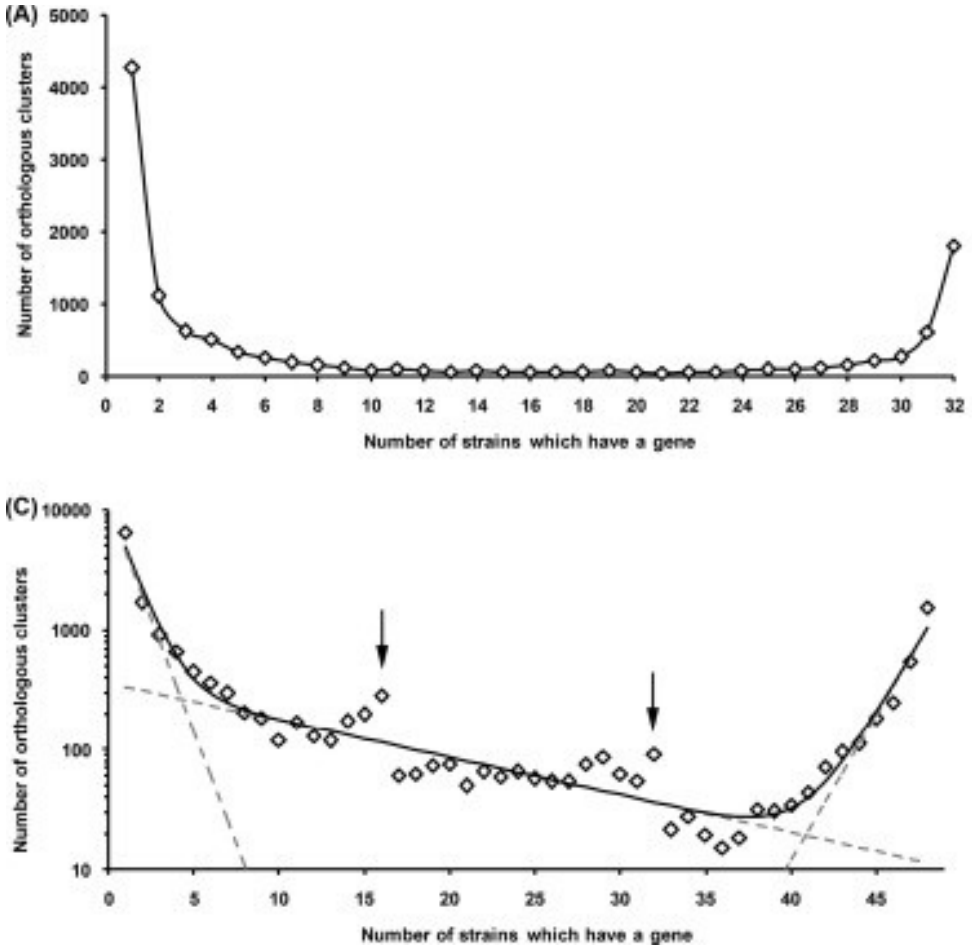
How do prokaryotes evolve?



Review, Nature 2019
https://www.nature.com/articles/nmicrobiol201740

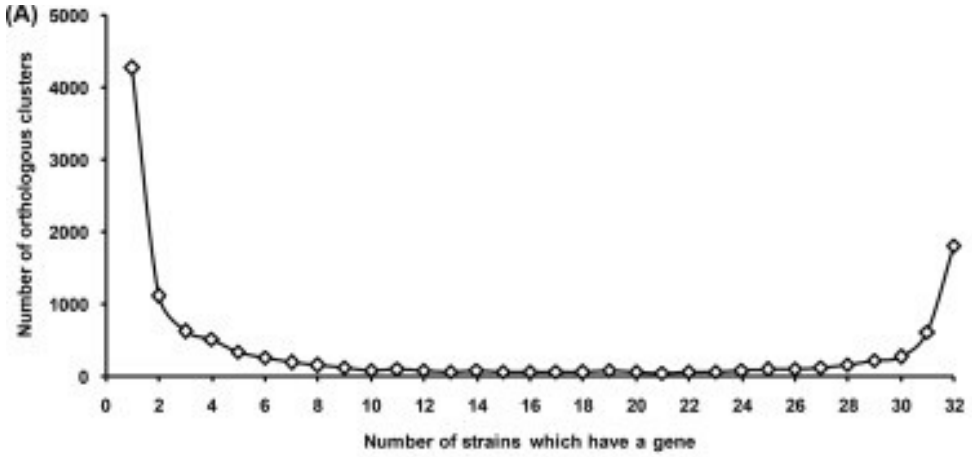


Genome composition (pangenome)

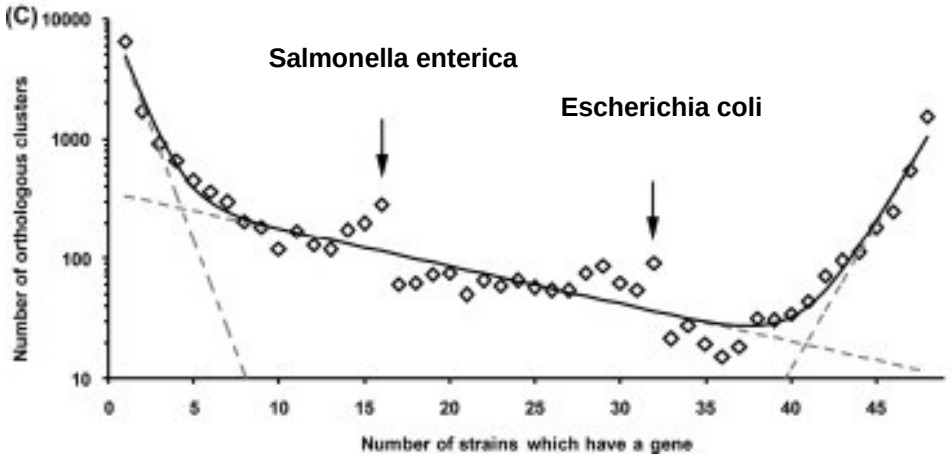


Gordienko EN, Kazanov MD, Gelfand MS. **Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica.** J Bacteriol. 2013;195(12):2786-2792. doi:10.1128/JB.02285-12

Genome composition (pangenome)



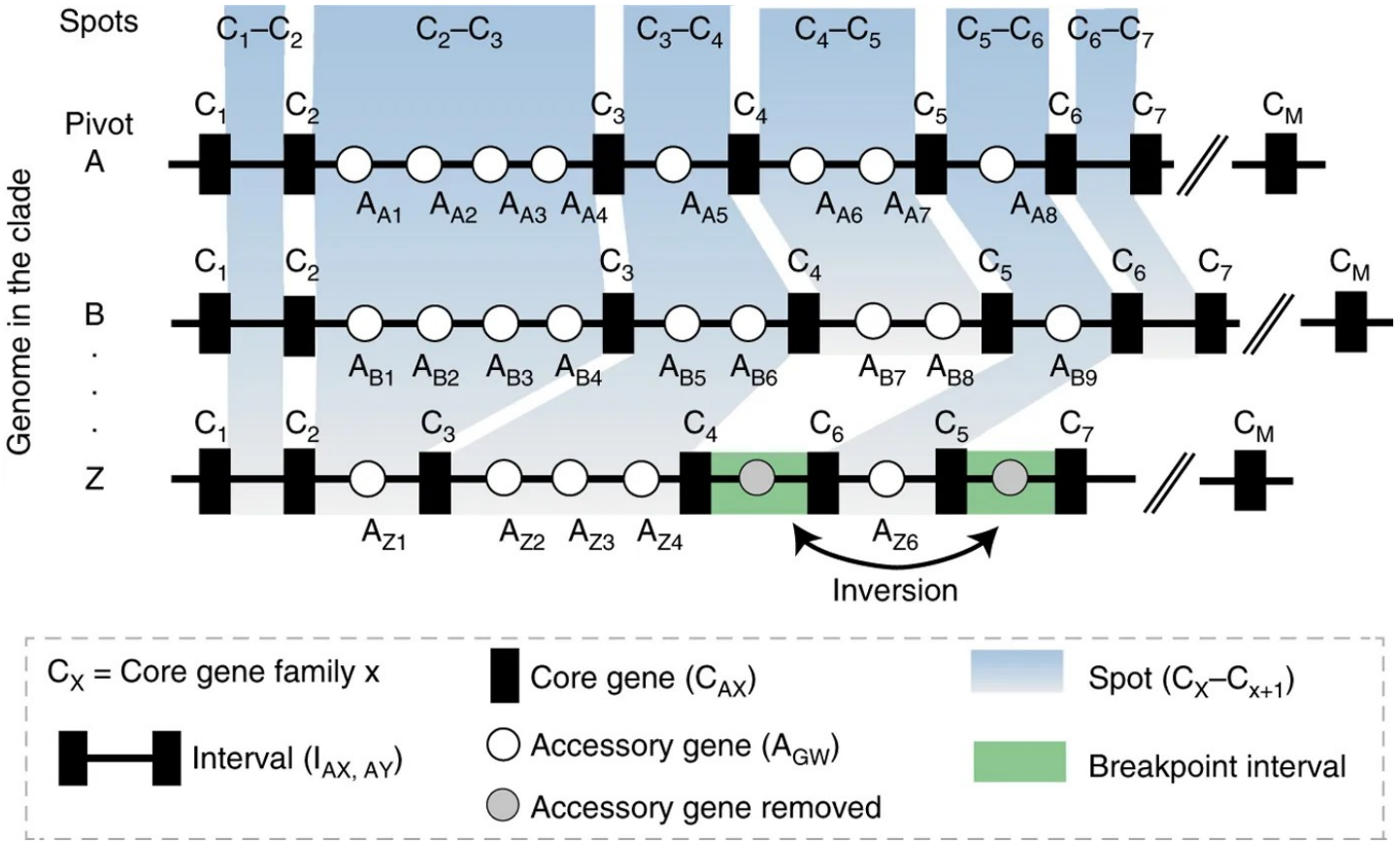
Gordienko EN, Kazanov MD, Gelfand MS. **Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica.** J Bacteriol. 2013;195(12):2786-2792. doi:10.1128/JB.02285-12



Pangenomic Definition of Prokaryotic Species

Moldovan MA, Gelfand MS. **Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp.** Front Microbiol. 2018;9:428

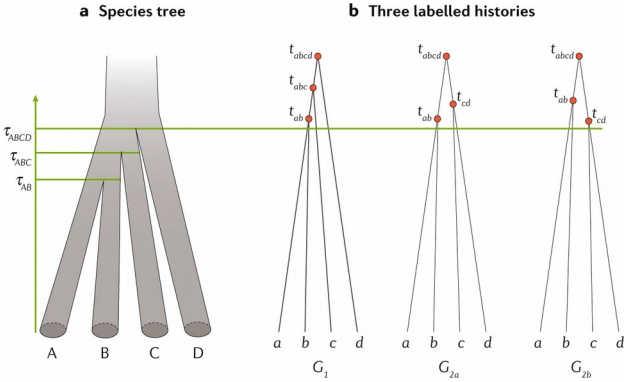
Genome composition (pangenome)

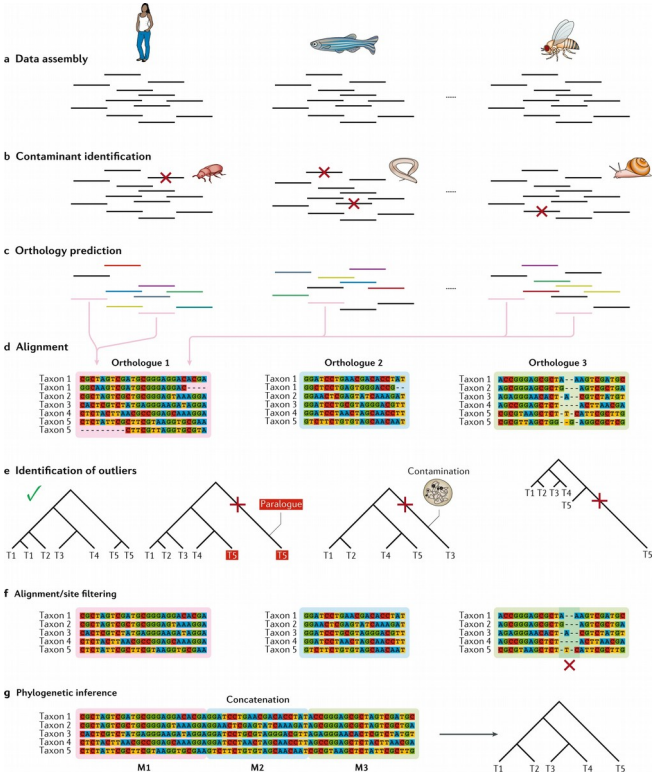


Oliveira, P.H., Touchon, M., Cury, J. et al. The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun. 2017

Genome composition (pipeline)

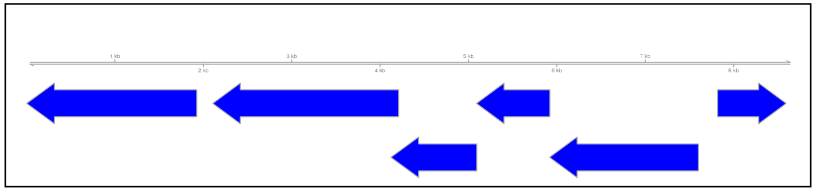
1. gene annotation (gff files)
2. orthologs (ProteinOrtho)
3. pangenome statistics (biopython)
4. core genes alignments (muscle)
5. phylogenetic tree (PhyML)
6. phyletic patterns (itol)





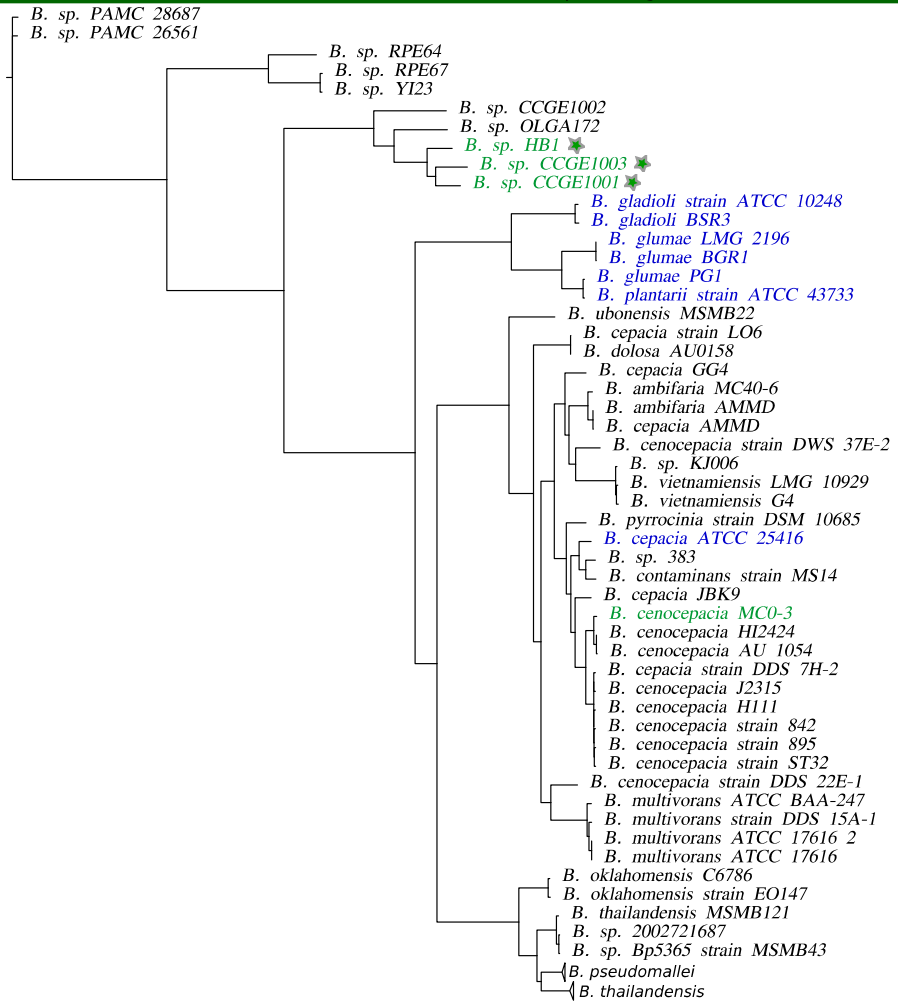
Kapli, P., Yang, Z. & Telford, M.J. **Phylogenetic tree building in the genomic age.** Nat Rev Genet 21, 428–444 (2020). https://doi.org/10.1038/s41576-020-0233-0

Detection of genomic islands

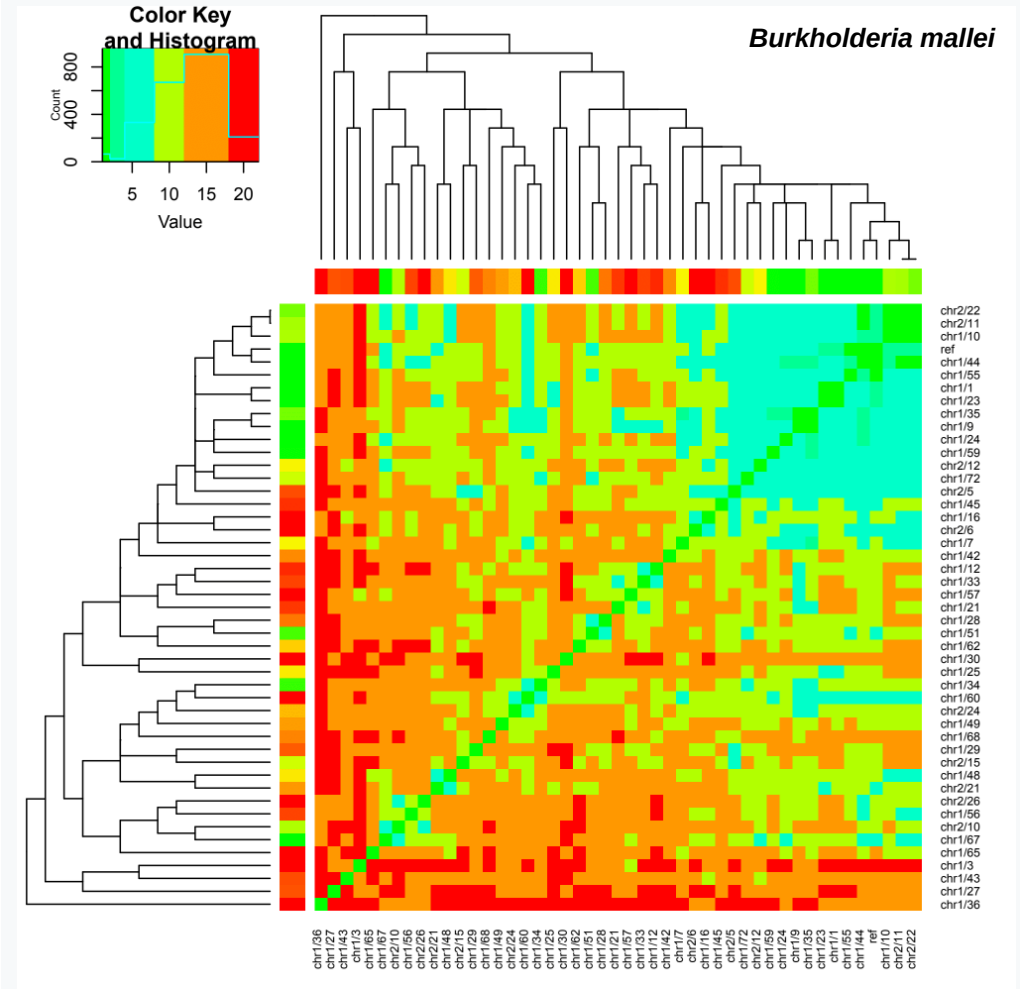


[1] TonB-dependent siderophore receptor
[2] FecCD-like permease
[3]Ferrichome-binding periplasmic protein
[4] ATPase subunit ABC transporter
[5] ABC transporter: transmembrane domain + ATPase subunit
[6] Transcription regulator *AraC*

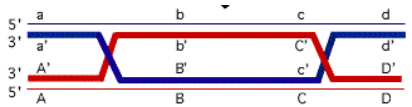
genomic island



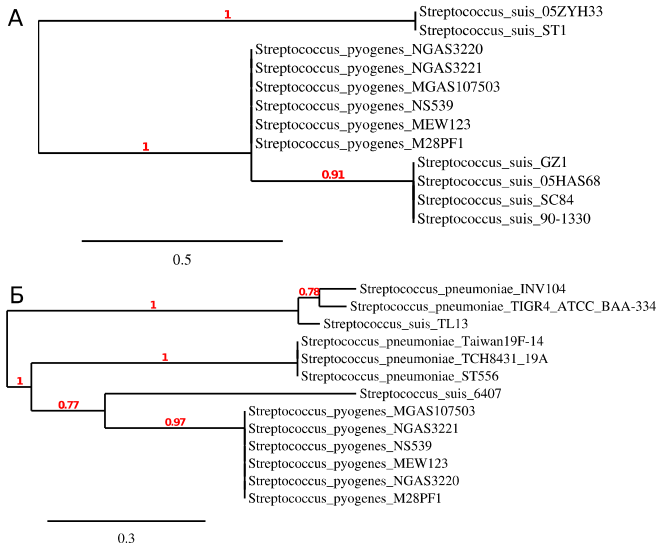
‘To tree or not to tree’



Homologous recombination

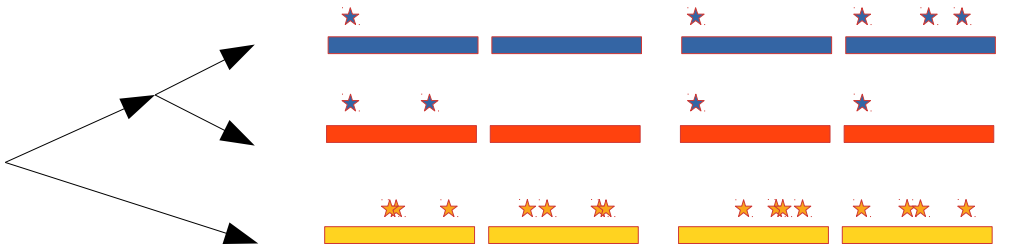


Gene trees (examples)

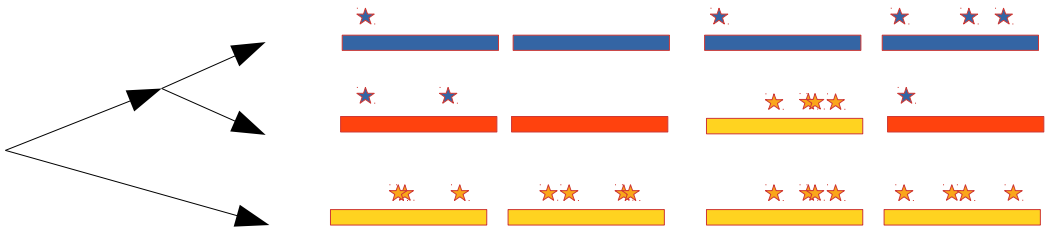


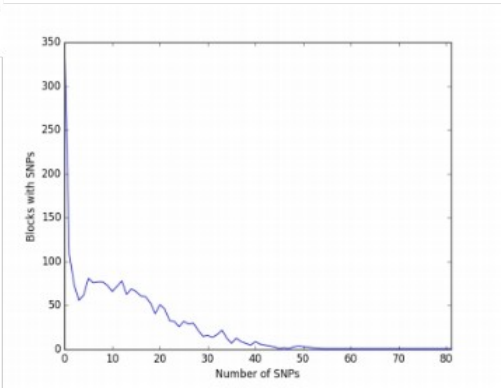
How to detect recombination

vertical inheritance of mutations



after recombination





All pairwise distributions are fitted by the function

$F_{\lambda,k,\mu,W(x)} = W \times P_{\lambda(x)} + (1-W) \times E_{k,\mu(x)}$

Practice 1

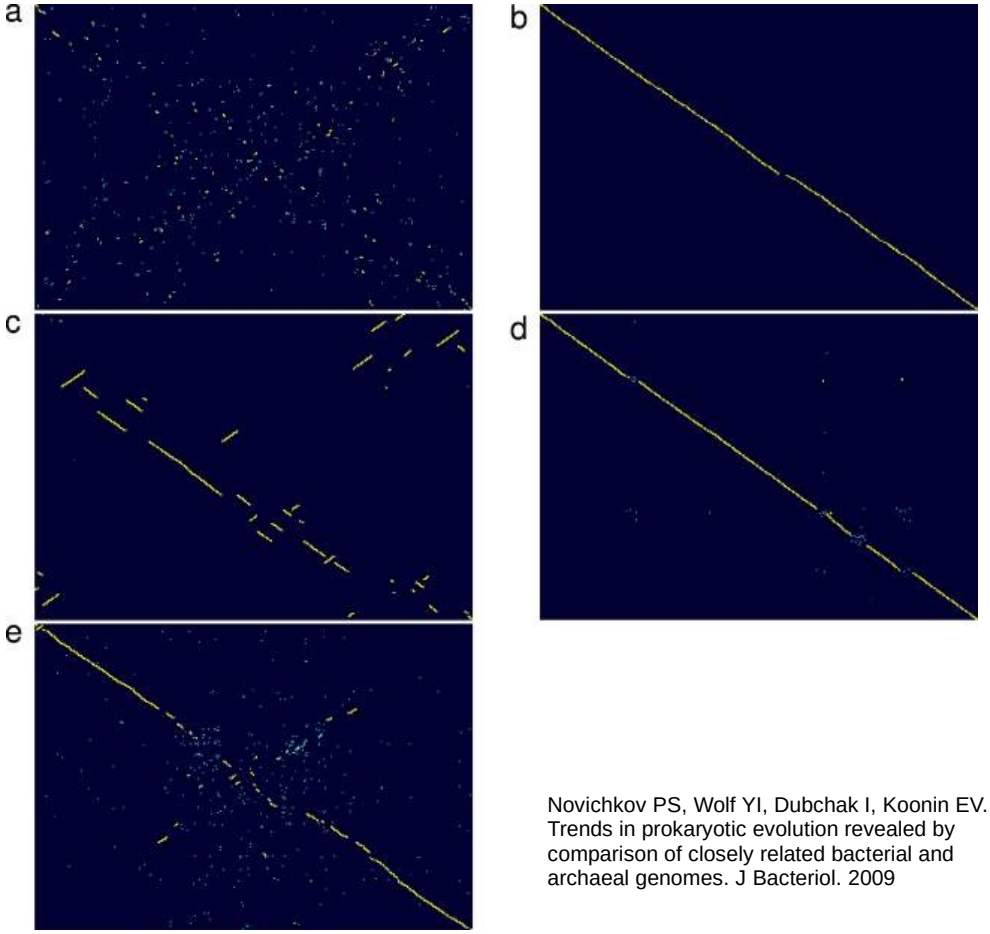
Dot-plot (bioinformatics)

“Dot plots compare two sequences by organizing one sequence on the x-axis, and another on the y-axis, of a plot. When the residues of both sequences match at the same location on the plot, a dot is drawn at the corresponding position. Note, that the sequences can be written backwards or forwards, however the sequences on both axes must be written in the same direction.” https://en.wikipedia.org/wiki/Dot_plot_(bioinformatics)

Software

https://bioinfo.lifl.fr/yass/index.php

How do prokaryotes evolve?



Novichkov PS, Wolf YI, Dubchak I, Koonin EV. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol. 2009

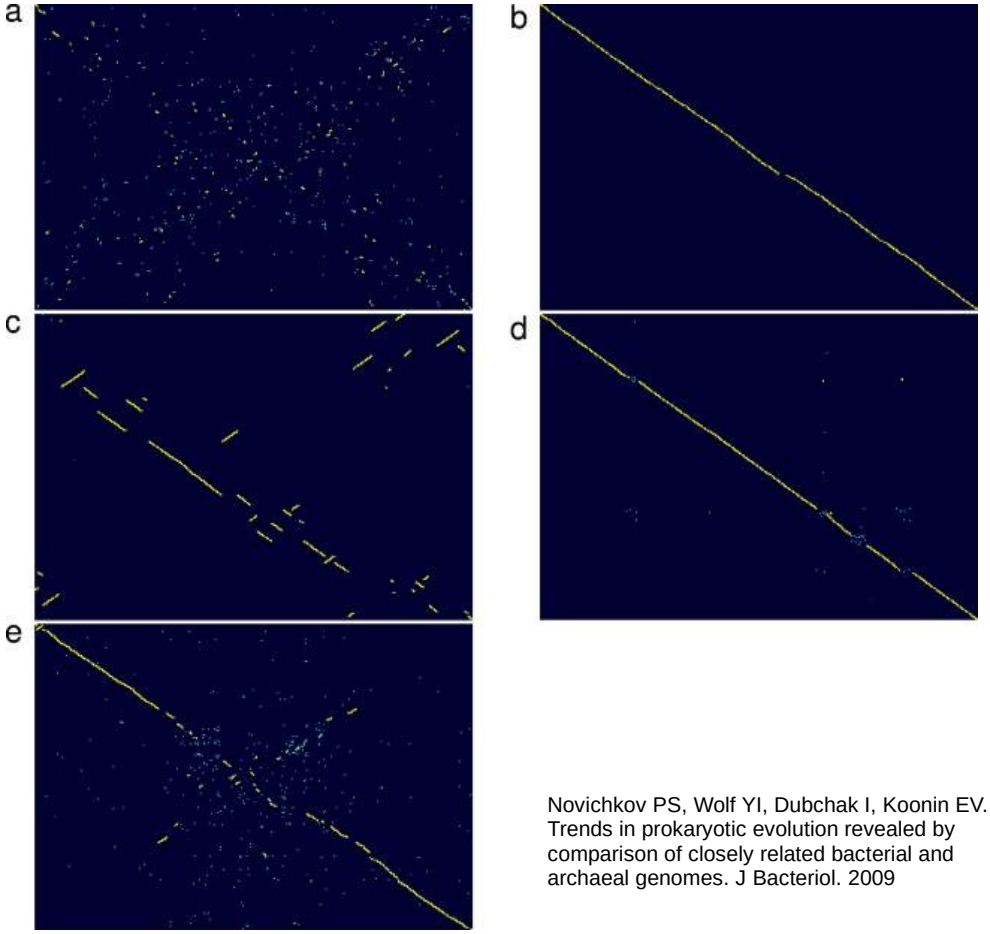
Dot-plot (bioinformatics)

"Dot plots compare two sequences by organizing one sequence on the x-axis, and another on the y-axis, of a plot. When the residues of both sequences match at the same location on the plot, a dot is drawn at the corresponding position. Note, that the sequences can be written backwards or forwards, however the sequences on both axes must be written in the same direction." https://en.wikipedia.org/wiki/Dot_plot_(bioinformatics)

Software

https://bioinfo.lifl.fr/yass/index.php

How do prokaryotes evolve?



Novichkov PS, Wolf YI, Dubchak I, Koonin EV. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol. 2009

(a) Nearly complete decay of synteny; *Streptococcus sanguinis* SK36 and *Streptococcus pneumoniae* R6.

(b) Virtual absence of rearrangements; *Chlamydophila caviae* GPIC and *Chlamydophila abortus* S26/3.

(c) Multiple inversions with limited transposition of individual genes; *Yersinia pestis Antiqua* and *Y. pestis CO92*.

(d) No inversion; hot spots of transposition of individual genes; *P. marinus* AS9601 and *P. marinus* MIT 9215.

(e) Multiple inversions and transposition of individual genes; *Pseudomonas fluorescens* PfO-1 and *P. fluorescens* Pf-5.

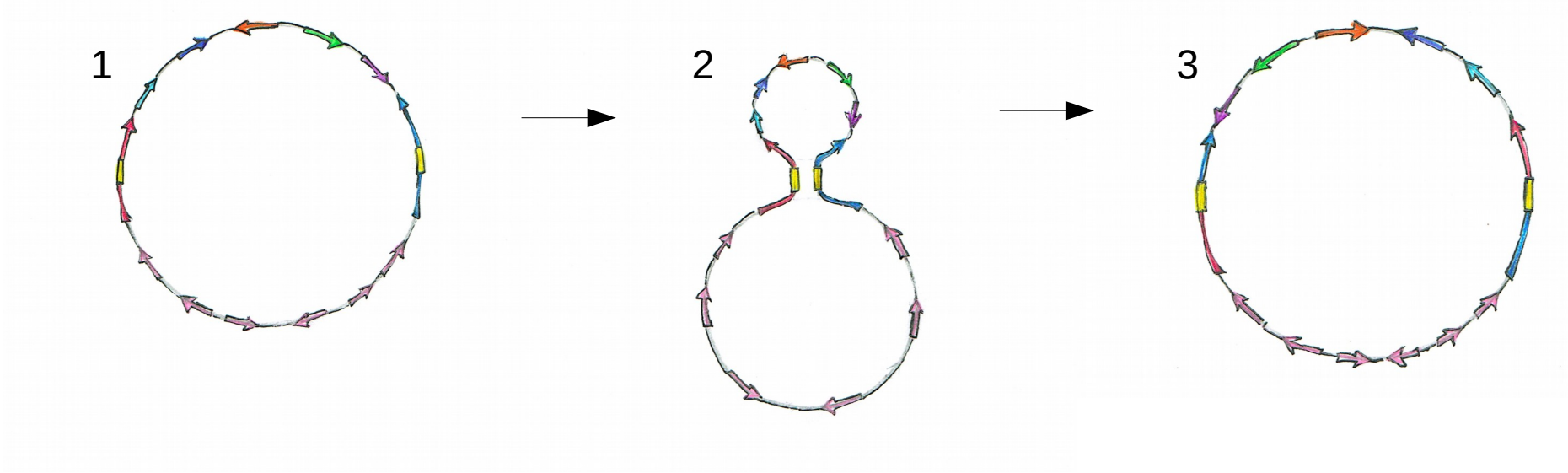
X-shape

Intra-genomic recombination

Circular chromosome with repeats

Intra-chromosome recombination

Inversion of the segment between repeats

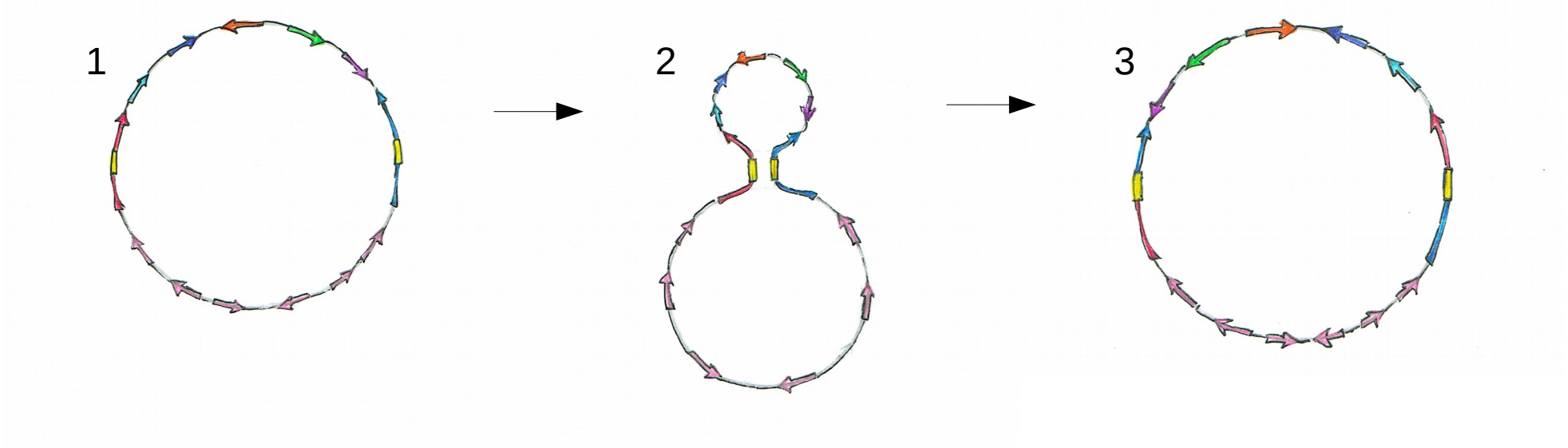


Intra-genomic recombination

Circular chromosome with repeats

Intra-chromosome recombination

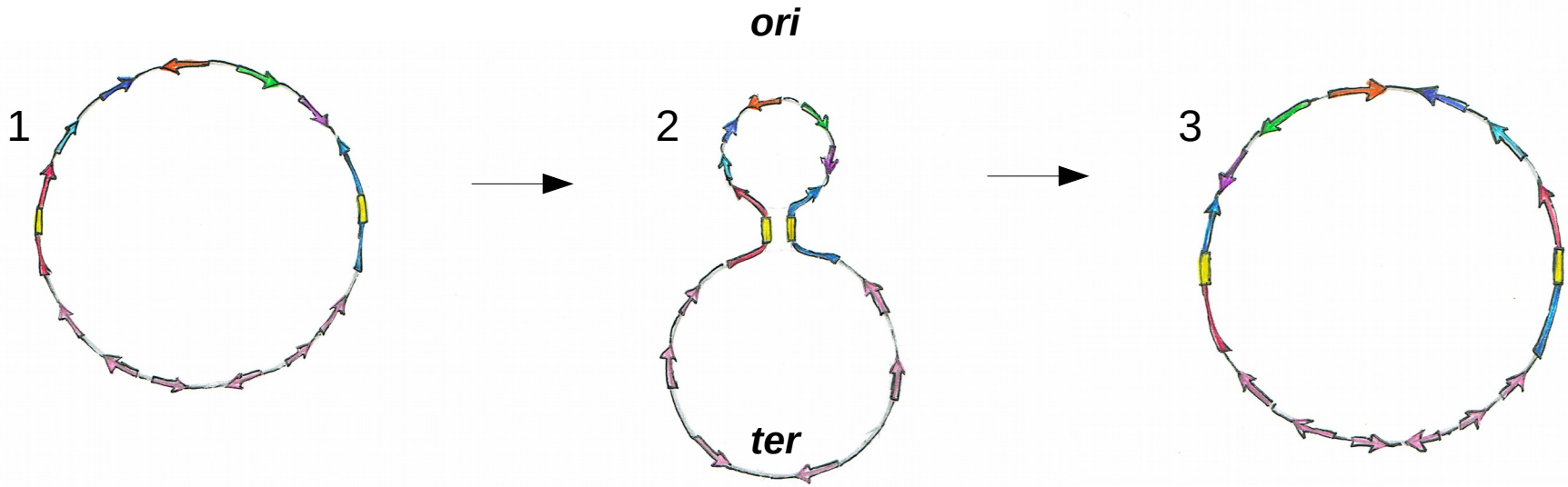
Inversion of the segment between repeats



Genomic repeats:

- mobile elements (transposases) – up to 100 copies per genomes
- rRNA gene operons – up to 15 copies per genomes
- paralogs?

Intra-genomic recombination

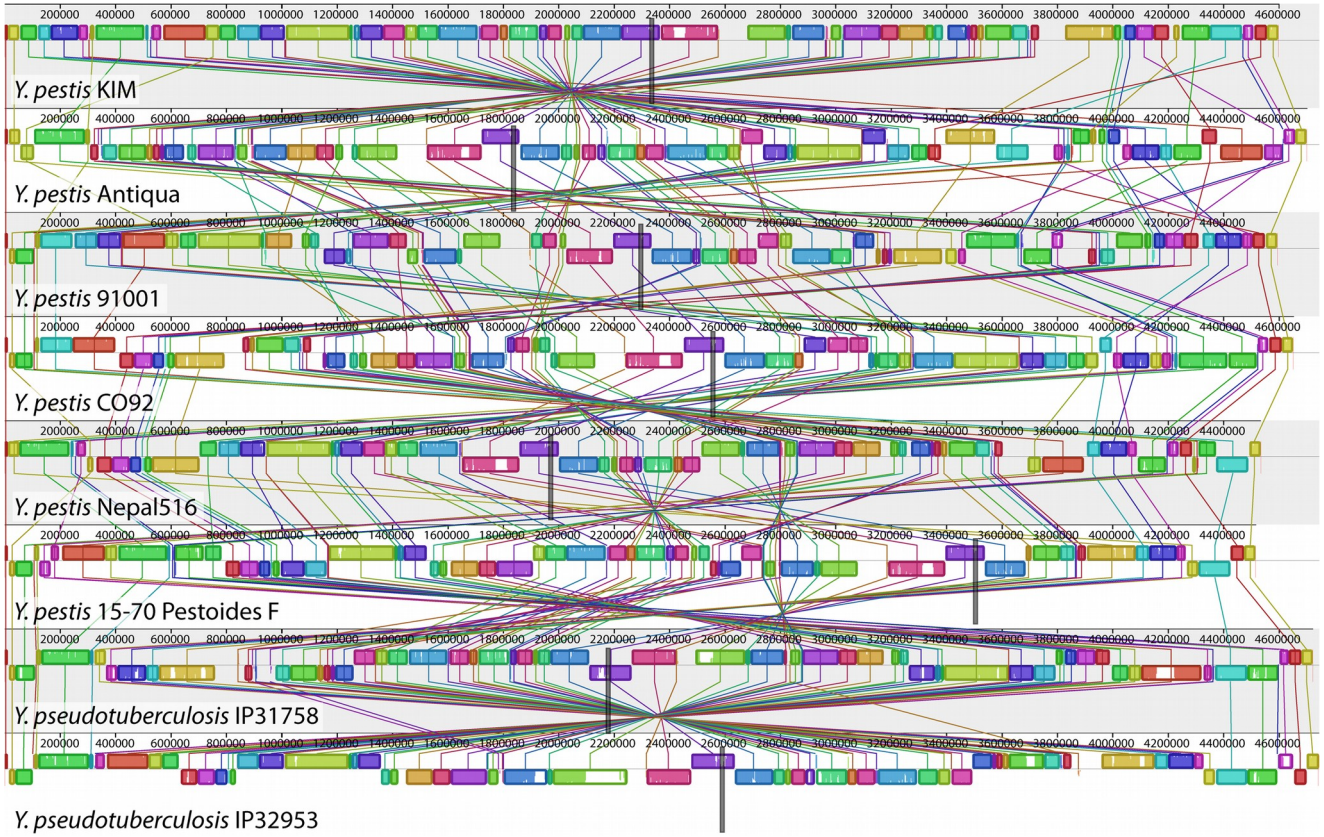


From Biology to Maths

Rearrangement	Math illustration
> Deletion	1 2 -3 4 -5 6 ⇒ 1 2 -3 -5 6
> Insertion	1 2 -3 4 -5 6 ⇒ 1 2 -3 4 7 -5 6
> Translocation	1 2 -3 4 -5 6 ⇒ 1 -3 4 -5 2 6
> Inversion	1 2 -3 4 -5 6 ⇒ 1 2 5 -4 3 6
> Duplication	1 2 -3 4 -5 6 ⇒ 1 2 -3 -3 4 -5 6

Patterns of synteny blocks

***Yersinia pestis* (black death)**

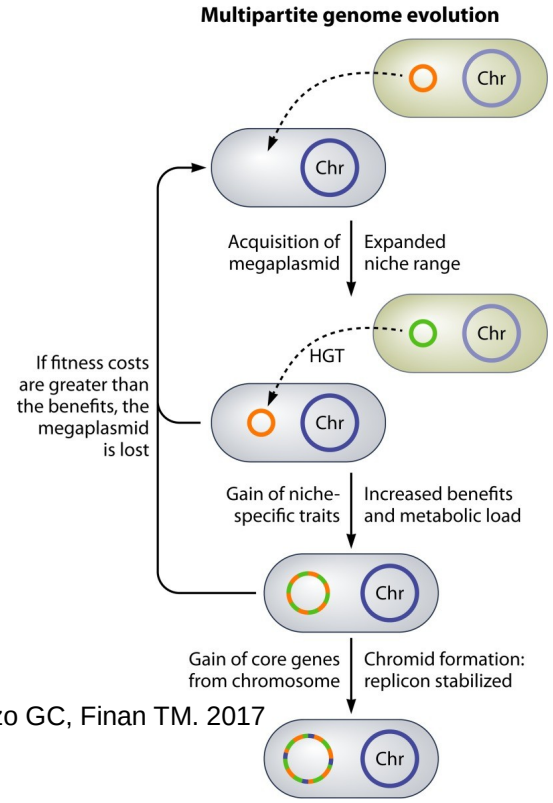


AE, Miklós I, Ragan MA. Dynamics of genome rearrangement in bacterial populations. PLoS Genet. 2008

Story 1. Chromid

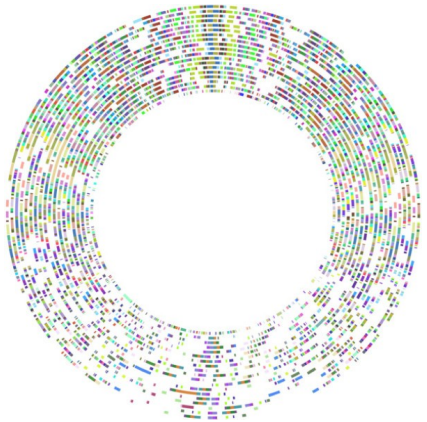
Dikow, Rebecca B, and William Leo Smith. “Genome-level homology and phylogeny of Vibrionaceae (Gammaproteobacteria: Vibrionales) with three new complete genome sequences.” BMC Microbiology 2013

Story 1. Chromid

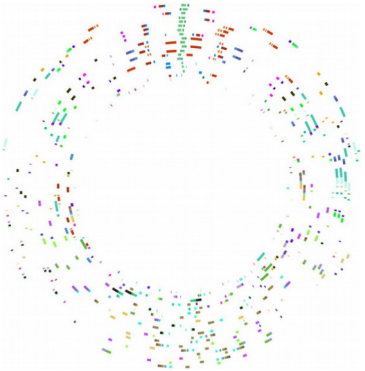


diCenzo GC, Finan TM. 2017

Primary chromosome

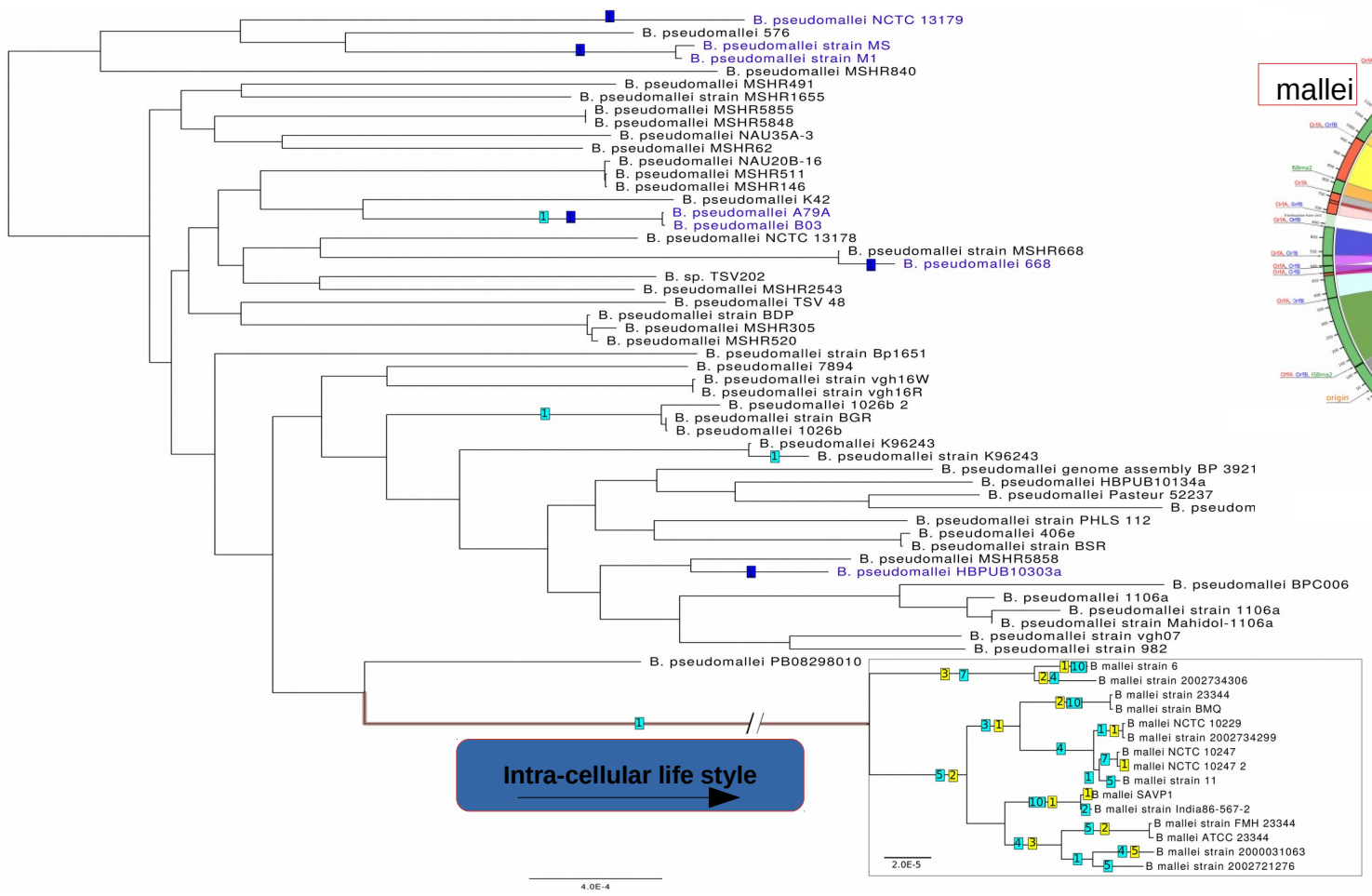


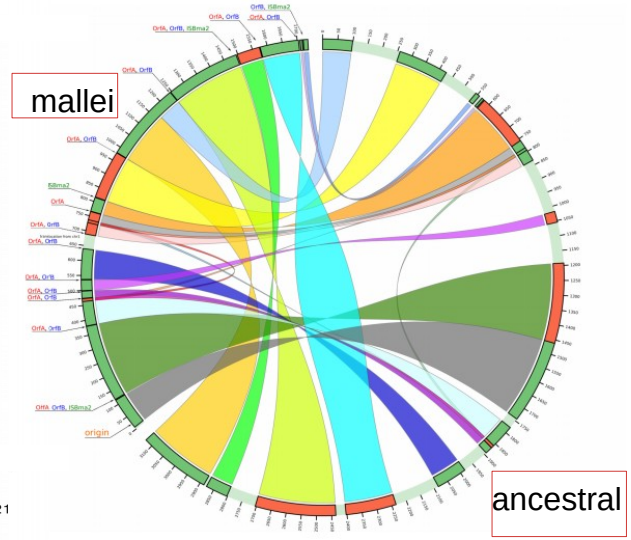
Plasmid → chromosome



Dikow, Rebecca B, and William Leo Smith. “Genome-level homology and phylogeny of Vibrionaceae (Gammaproteobacteria: Vibrionales) with three new complete genome sequences.” BMC Microbiology 2013

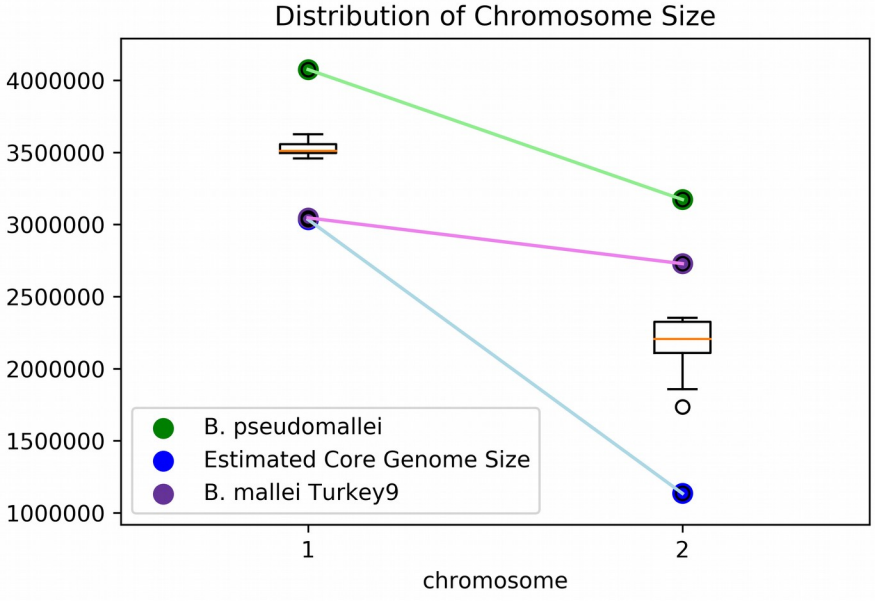
Story 2. Intracellular pathogens

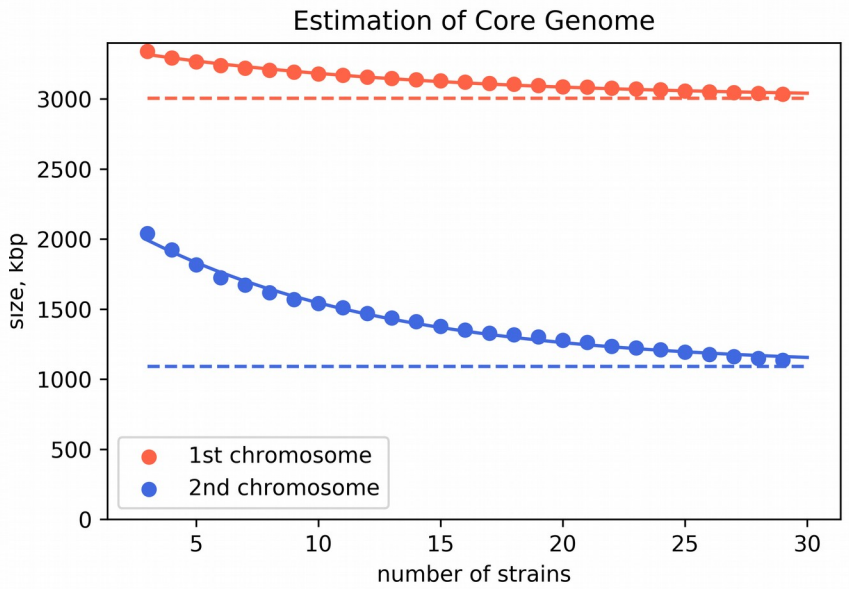




Bochkareva et al.
BMC Genomics
2018

Story 2. Intracellular pathogens



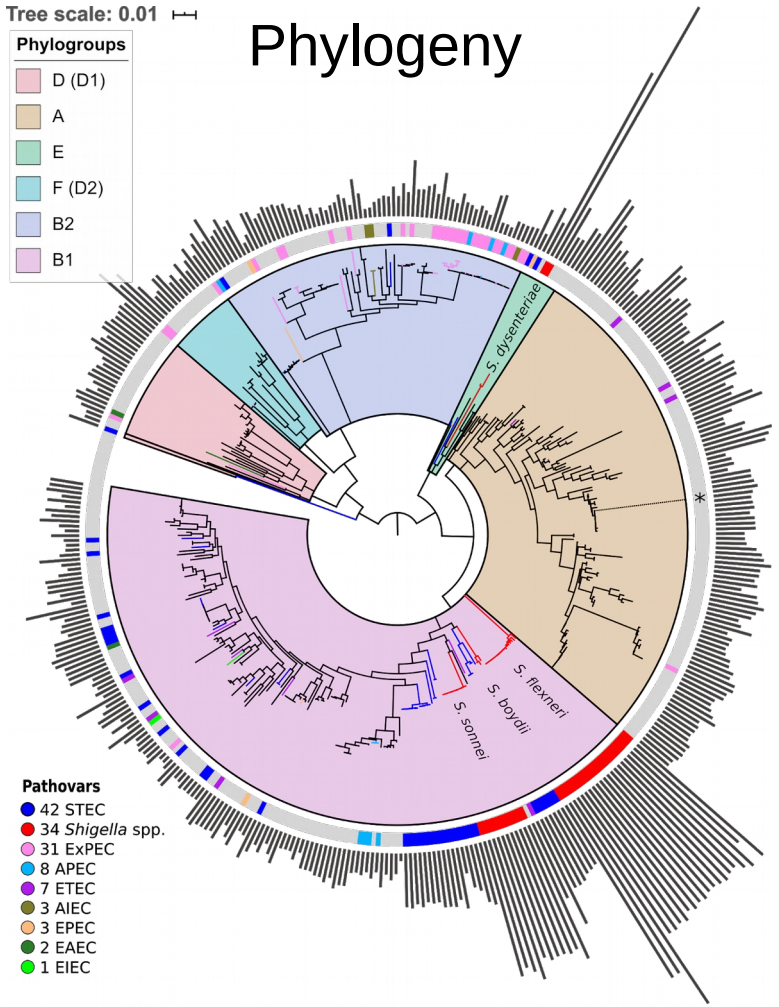


Reduction rates: 25% for the first and 64% for the second chromosome

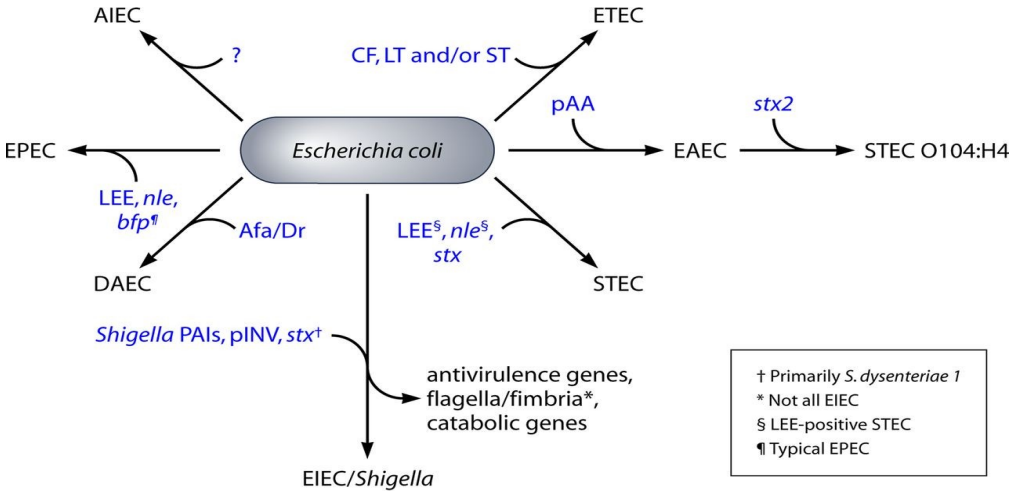
Anna Toidze
SMTB 2019

Story 3. Pathogenic *E.coli*

Phylogeny



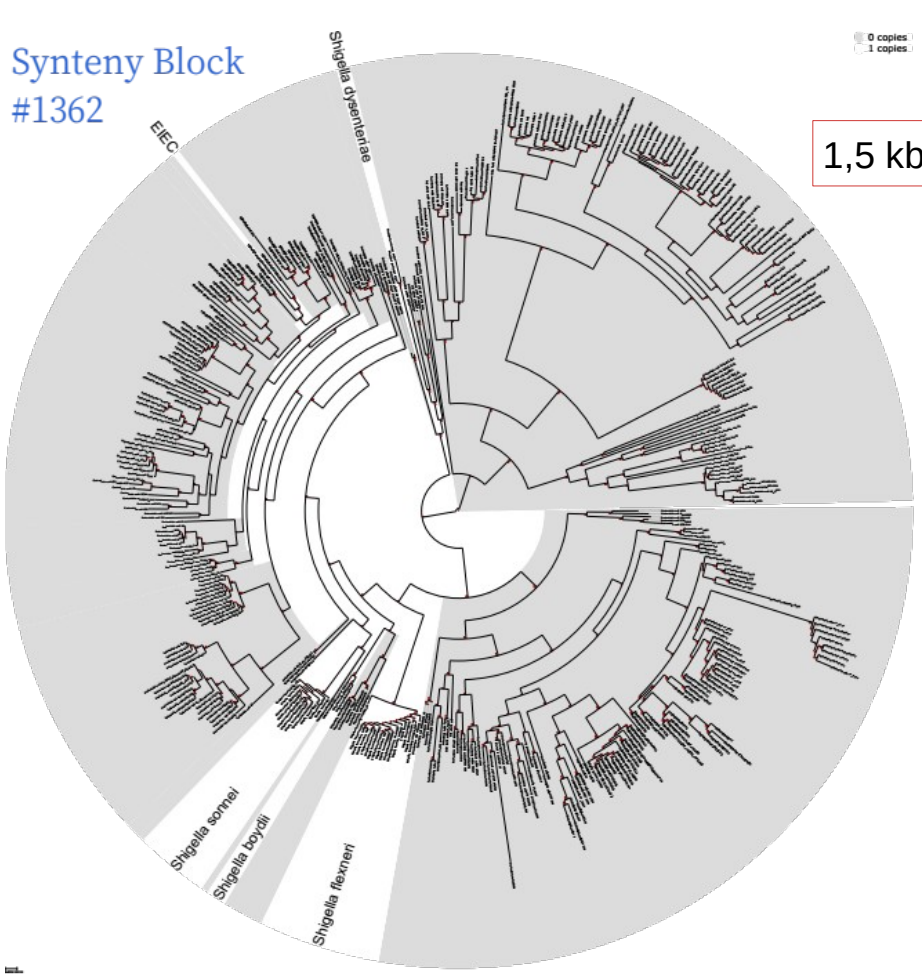
Evolution of pathovars



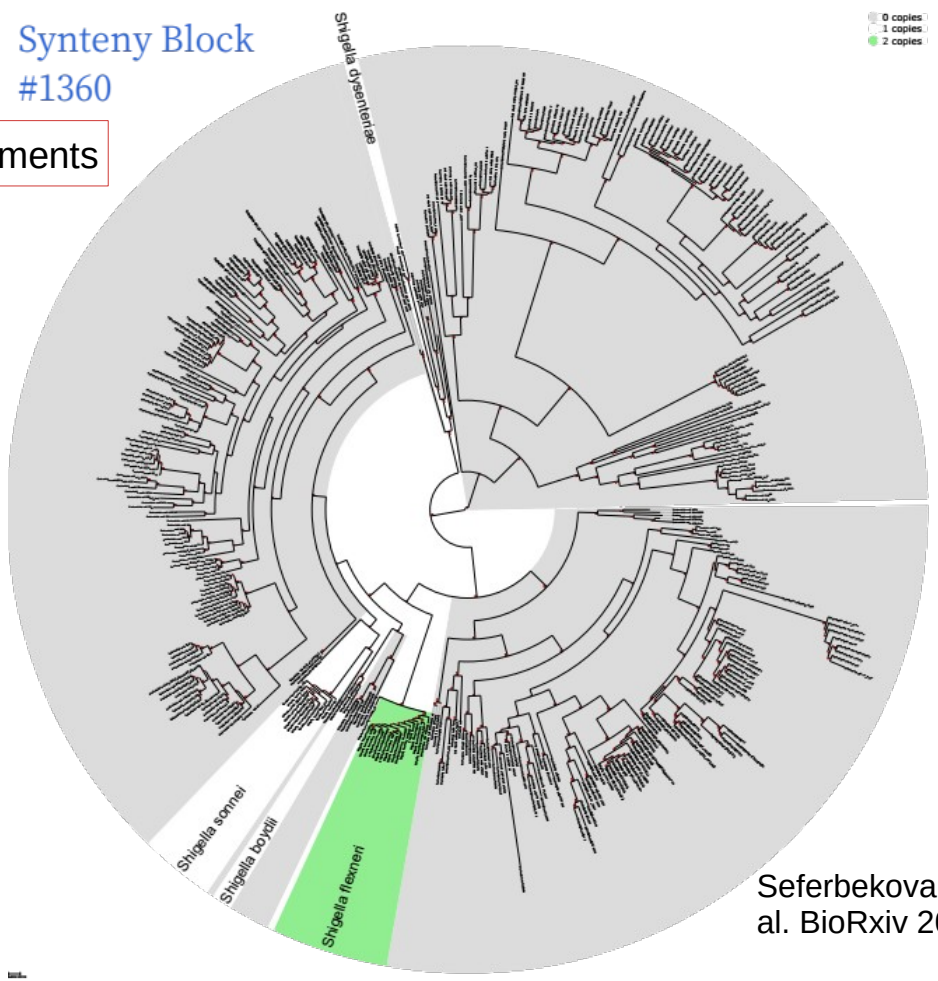
enteropathogenic (EPEC),
enterotoxigenic (ETEC),
enteroinvasive (EIEC),
enteroaggregative (EAEC),
Shiga toxin-producing (STEC),
diffusely adherent (DEAC),
adherent invasive (AIEC).

Seferbekova et al. BioRxiv 2020

Story 3. Pathogenic *E.coli*

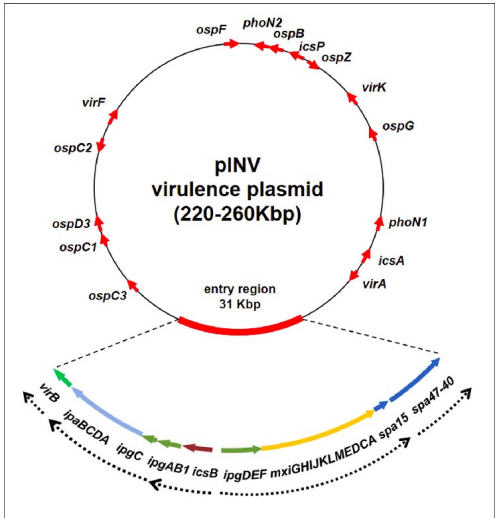


1,5 kb fragments



Seferbekova et al. BioRxiv 2020

Story 3. Pathogenic *E.coli*



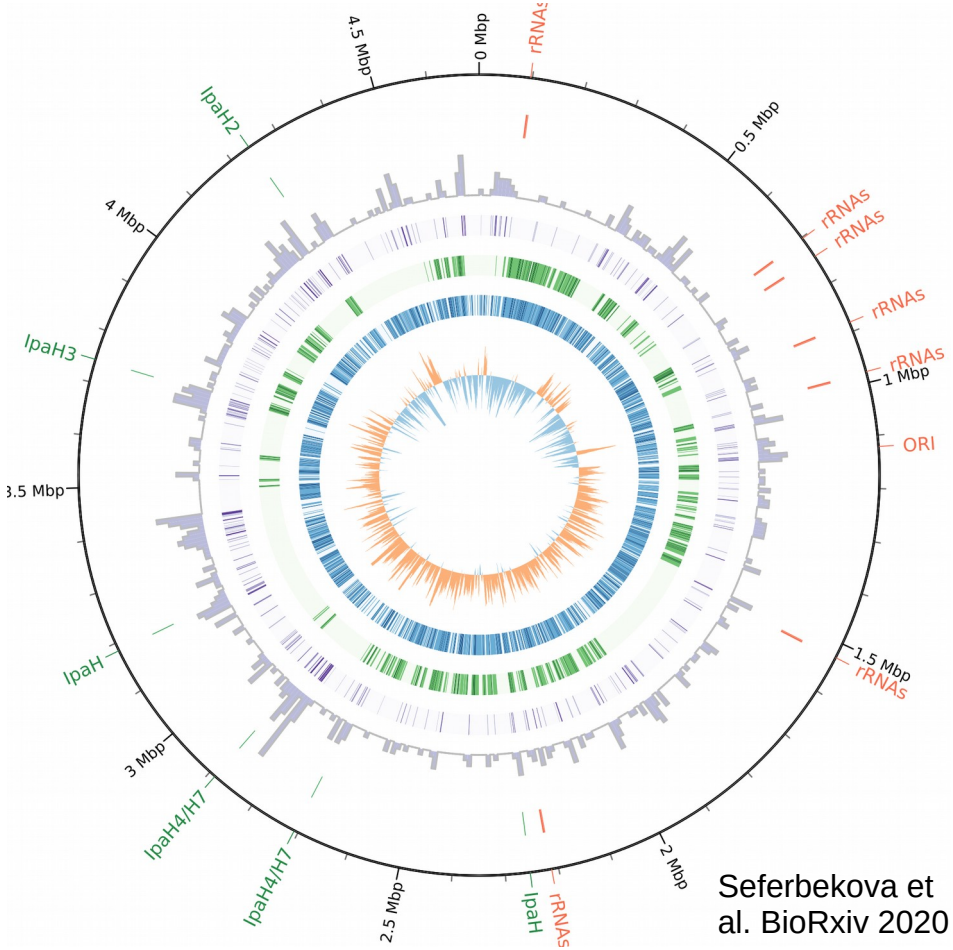
Genetic map of the pINV. The red arrows indicate major virulence determinants. Due to the variability in position and number, the ipaH genes are not shown. *Source: "The Intriguing Evolutionary Journey of Enteroinvasive E. coli (EIEC) toward Pathogenicity"*

- Both Shigella and EIEC spend much of their life cycle within eukaryotic cells and share many invasion-related functional systems.

- The adaptation to an intracellular lifestyle was conferred by the acquisition of the pINV plasmid encoding a Type III secretion system (T3SS)

- The delivery of bacterial virulence proteins into host cells via T3SS plays a crucial role in the infection strategies of Shigella

- Most T3SS effectors are encoded by pINV plasmid genes while the biological role of chromosomally encoded ipaH genes remains obscure.

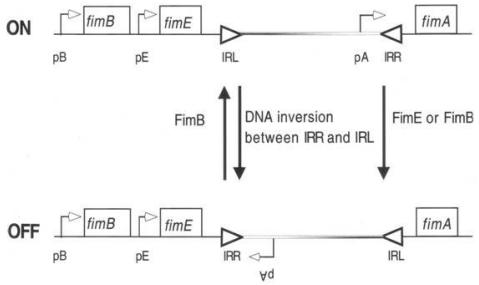


Seferbekova et al. BioRxiv 2020

Practice 2

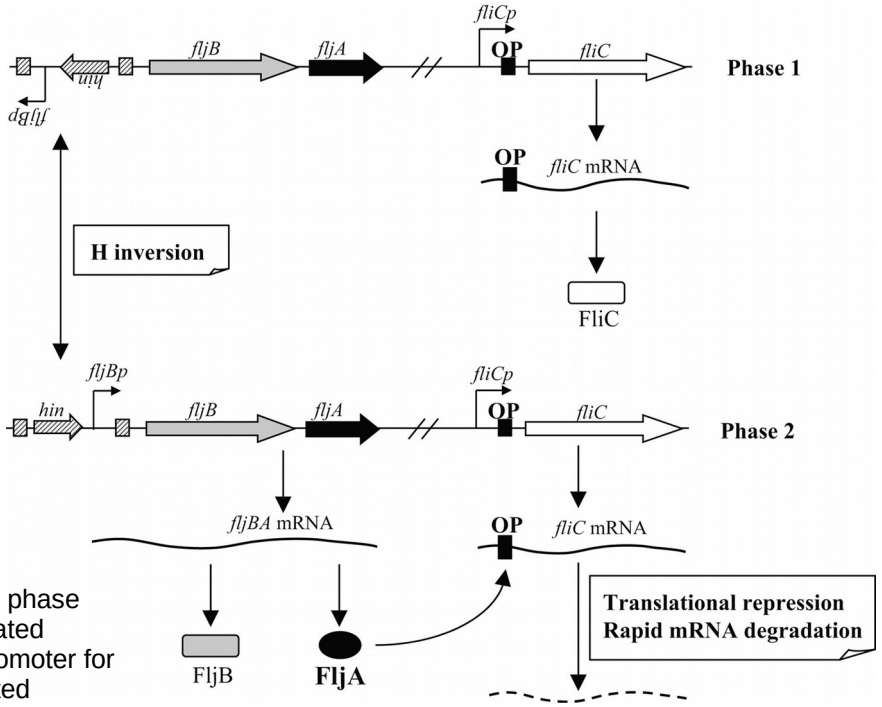
Site-specific micro-inversions

Phase variation – Site specific inversion



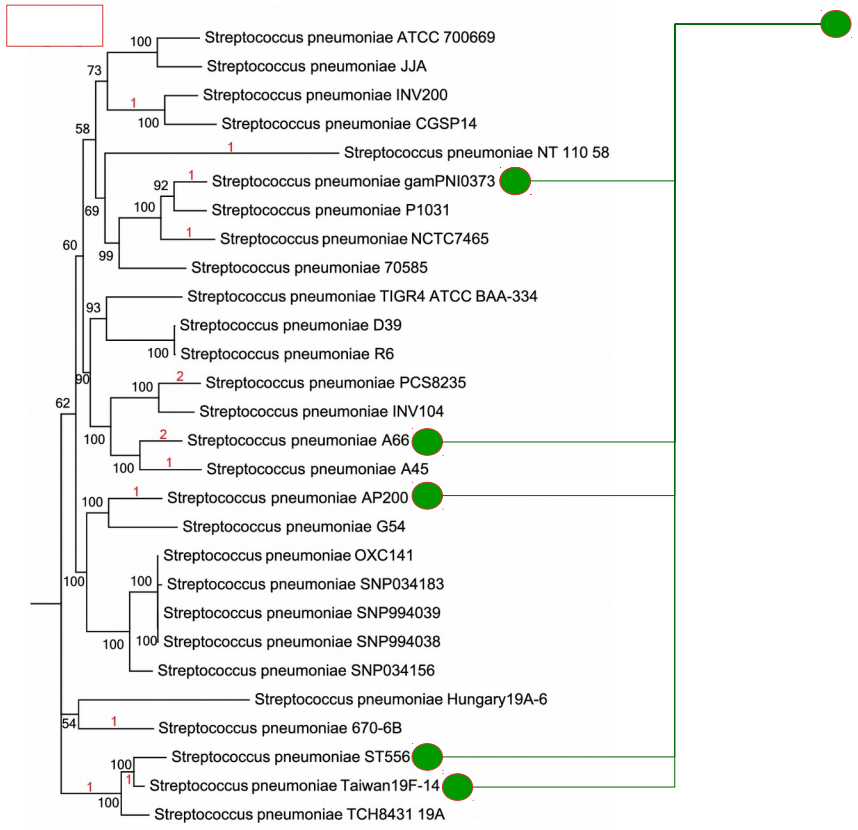
Phase variation – site specific inversion. pB, pE and pA are promoters for the genes *fimB, fimE* and *fimA* respectively. IRL and IRR are inverted repeats. FimB and FimE are recombinases that bind to the open triangles IRL an IRR. The result is an inversion of the DNA sequence (shaded bar) that turns ON or OFF the transcription of *fimA*.

Expression of multiple types of flagellin by *S. typhimurium*



The dual controlling system governs flagellar phase variation; one part of the system is Hin-mediated inversion of the H segment containing the promoter for the fljBA operon, and the other is FljA-mediated inhibition of fliC expression. FljA binds to the 5′-UTR of fliC mRNA, which inhibits its translation and facilitates its degradation.

Story 4. Phenotype switch



The same inverted fragment.
Breakpoints are formed by genes encoding proteins PhtB and PhtD

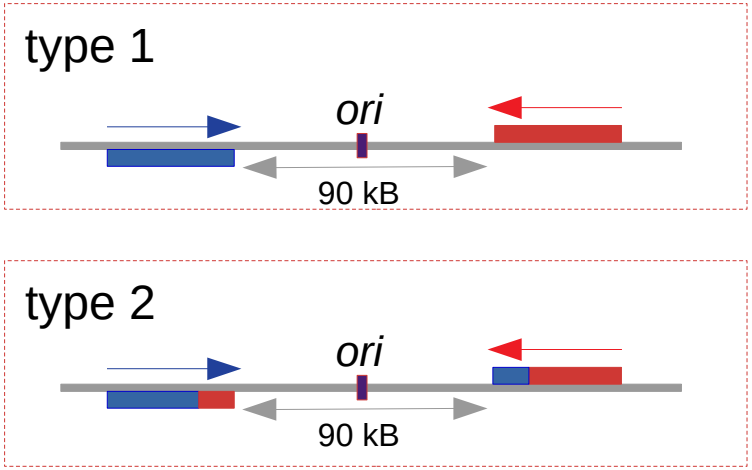
Pneumococcal histidine triad (Pht) proteins

- outer membrane proteins
- are suggested to be involved in Zn^{2+} binding
- antigens

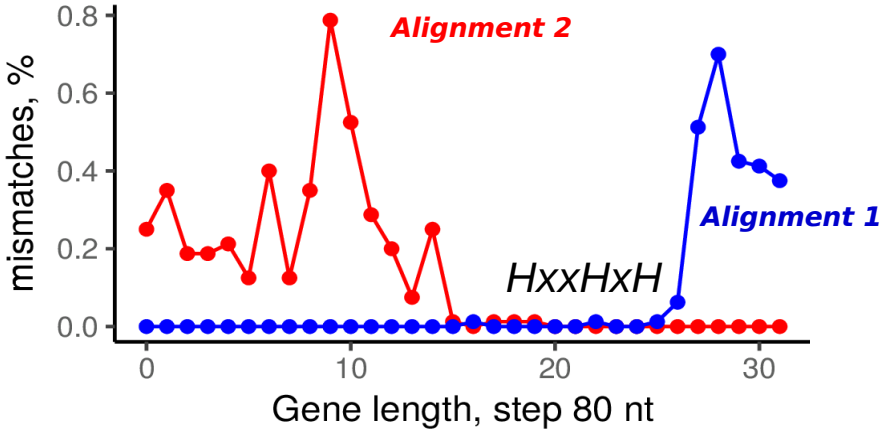
Shelyakin et al.
BMC Evol Biol
2019

Story 4. Phenotype switch

Recombination scheme



Mismatches



Shelyakin et al.
BMC Evol Biol
2019

Story 4. Phenotype switch

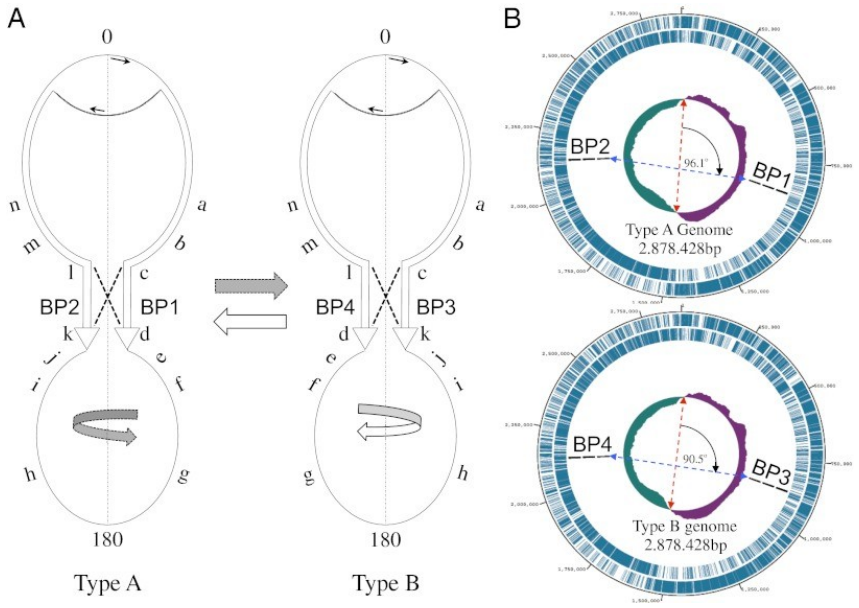
These are two out of four pneumococcal histidine triad (Pht) proteins, which are surface-exposed, interact with human host cells and **are considered to be good vaccine candidates.**

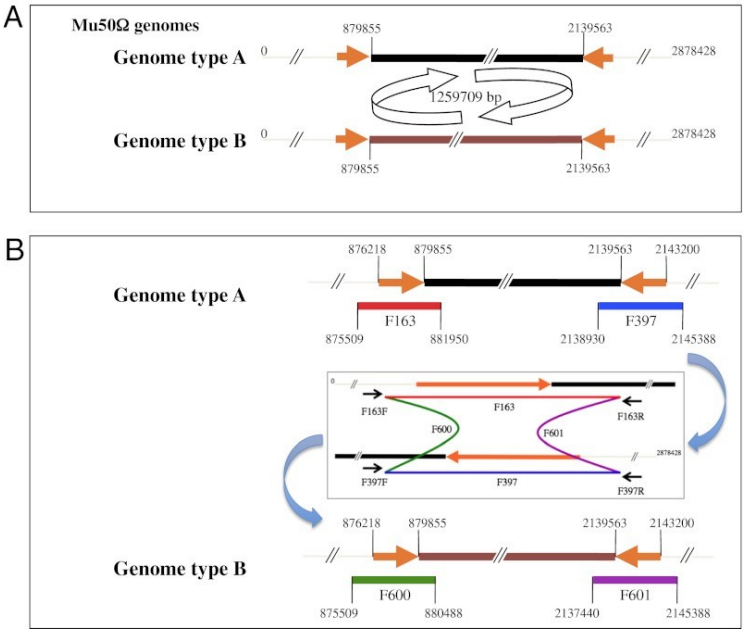
PhtD was already used in several phase I/II clinical trials.

Yun et al.* (PLoS One. 2015) analyzed the diversity of phtD alleles from 172 clinical isolates and concluded that **the sequence variation was minimal.**

Staphylococcus aureus

We report our findings on a bacterium that generates a reversible, large-scale inversion of its chromosome (about half of its total genome) at high frequencies of up to once every four generations. This inversion switches on or off bacterial phenotypes, including colony morphology, antibiotic susceptibility, hemolytic activity, and expression of dozens of genes. Quantitative measurements and mathematical analyses indicate that this reversible switching is stochastic but self-organized so as to maintain two forms of stable cell populations (i.e., small colony variant, normal colony variant) as a bet-hedging strategy.

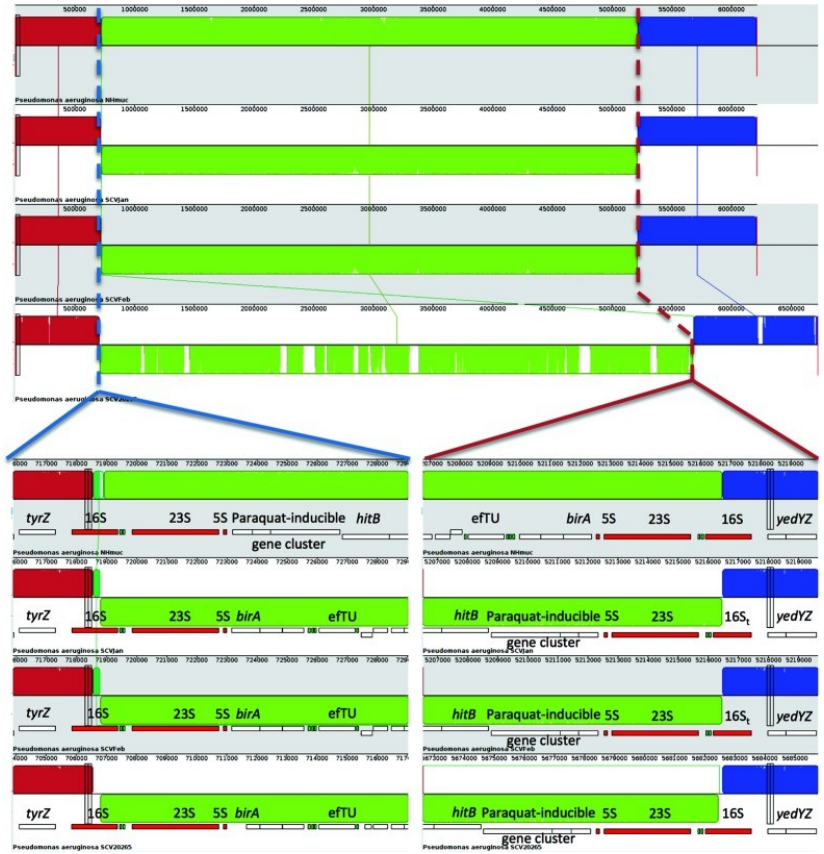




Cui L, Neoh HM, Iwamoto A, Hiramatsu K. Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. Proc Natl Acad Sci U S A. 2012;109(25):E1647–E1656.

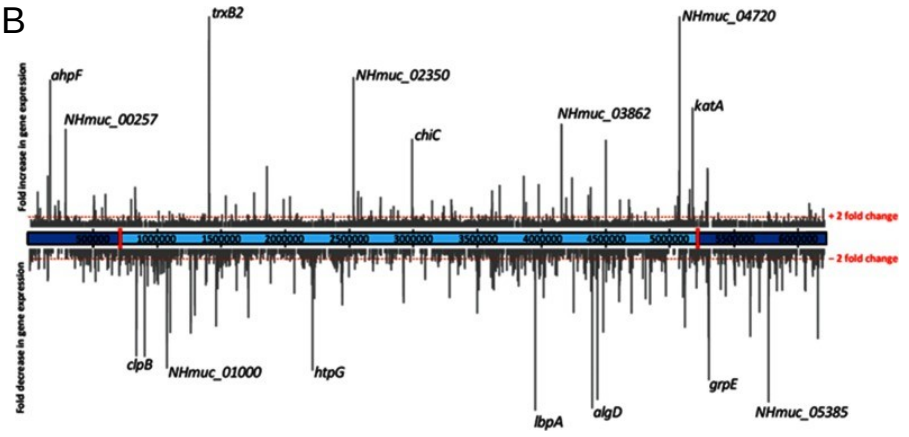
Pseudomonas aeruginosa

A



Using a combination of single-molecule real-time (PacBio) and Illumina sequencing we identify a large genomic inversion in the SCV through recombination between homologous regions of two rRNA operons and an associated truncation of one of the 16S rRNA genes and suggest this may be the genetic switch for conversion to the SCV phenotype. This phenotypic conversion is associated with large-scale transcriptional changes distributed throughout the genome. This global rewiring of the cellular transcriptomic output results in changes to normally differentially regulated genes that modulate resistance to oxidative stress, central metabolism and virulence. These changes are of clinical relevance because the appearance of SCVs during chronic infection is associated with declining lung function.

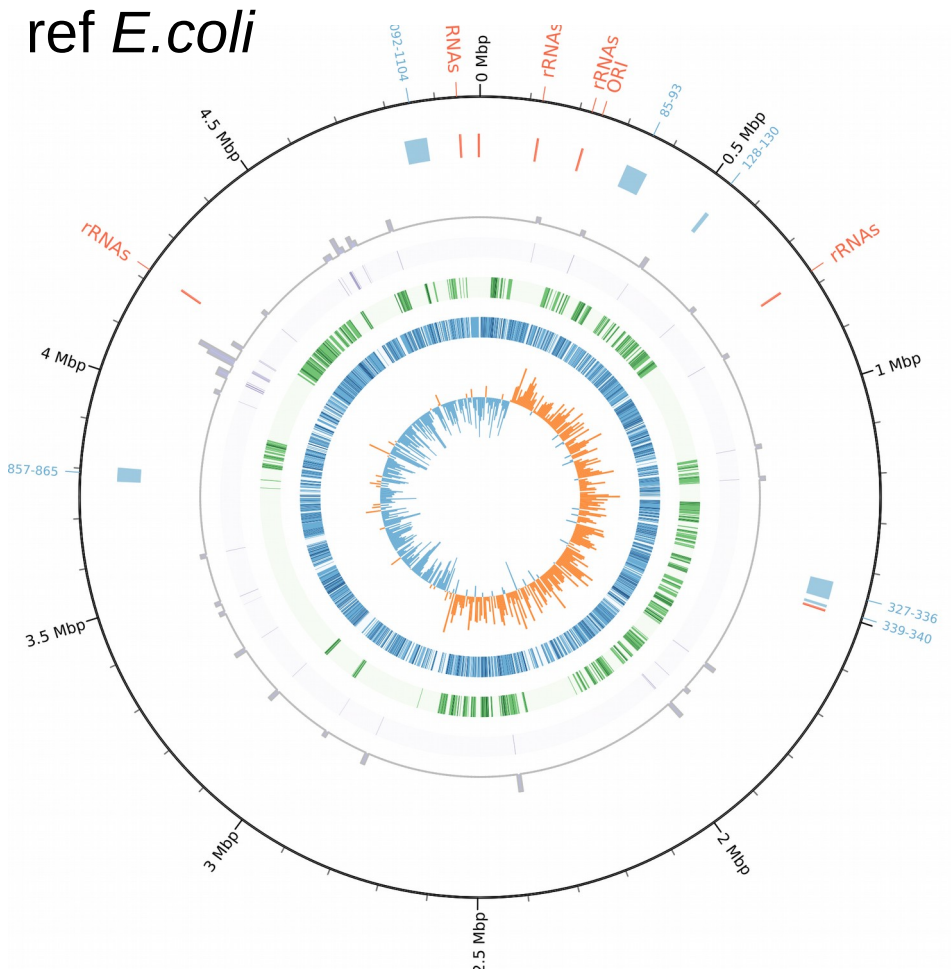
B



Irvine S, Bunk B, Bayes HK, et al. Genomic and transcriptomic characterization of *Pseudomonas aeruginosa* small colony variants derived from a chronic infection model. Microb Genom. 2019;5(4):e000262.

Genomic maps

ref *E.coli*



Shigella