Pangenome Structure?

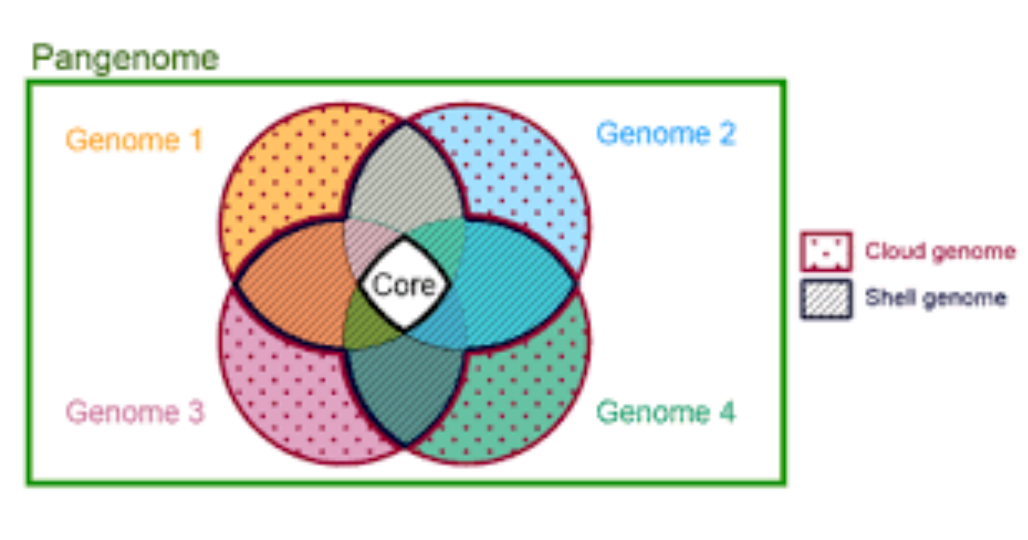
Ryan C. Fink

Let’s have some fun first…

- Divide in groups of three
- Get a piece of paper and draw a line to get three columns
- Title the three columns “Essential”, “Optional”, and “Unique”
- Think about your car, your parent’s car, or your dream car
- ESSENTIAL: ALL cars must have them to be functioning vehicles
- OPTIONAL: not necessary, but you would love to have them in your car
- UNIQUE: they make a car different from any other (good or bad)

Pangenome structure

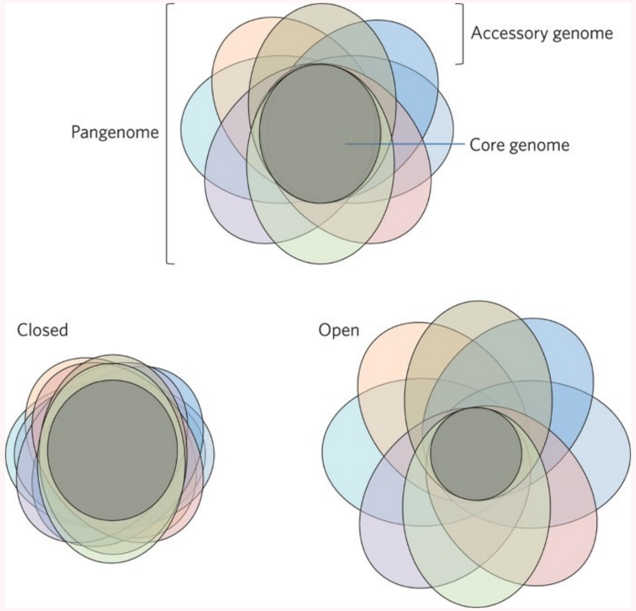
- **PANGENOME** = ESSENTIAL + OPTIONAL + UNIQUE
  - Πας, πασα, παν = all, every, each
- **OPTIONAL** = ACCESSORY (SHELL + CLOUD)
- **UNIQUE** = CLOUD

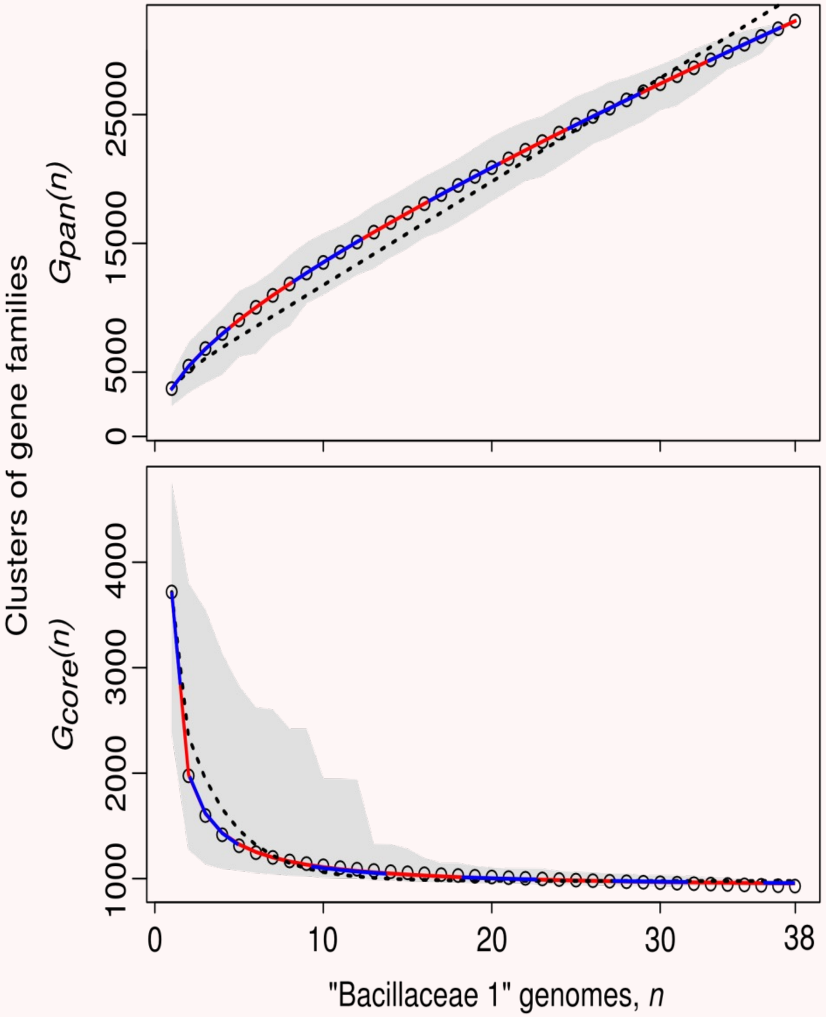


Open vs. Close

CLOSED: The pangenome size tends to a maximum as number of genomes increases

OPEN: The pangenome keeps increasing as you add new genomes



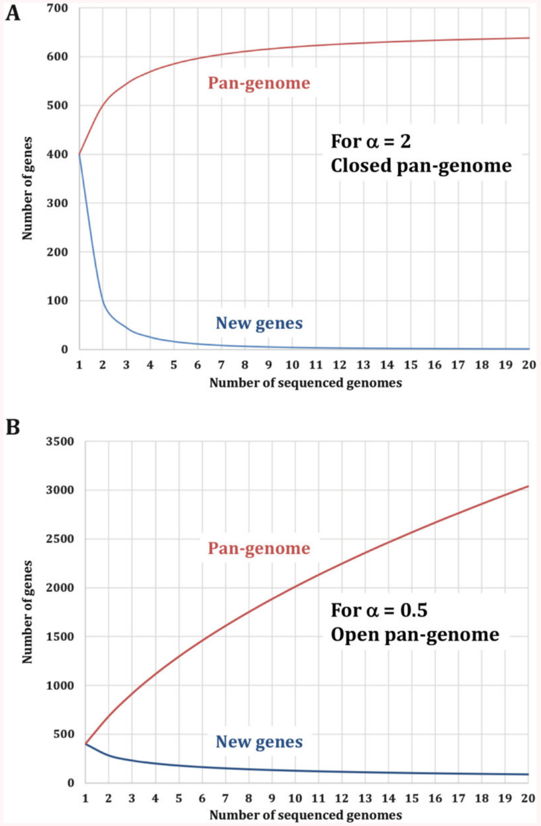


Open or closed?

Pangenome measurements follow Heap's law

"As more and more books are read, the number of different words grows as a power law of the total number of books read"

$$n = \kappa N^{-\alpha}$$



Core, Shell, Cloud… 🤪

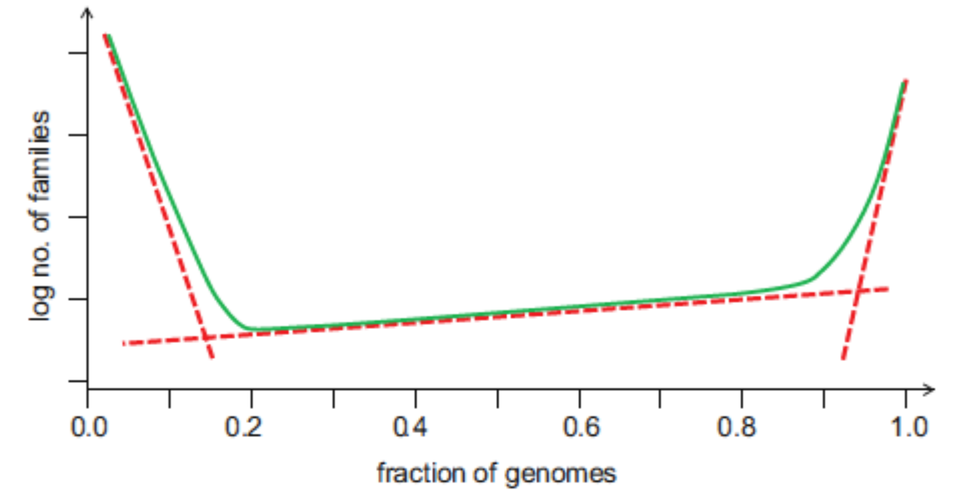


FIGURE 4 | The universal distribution of gene commonality in the microbial genomic universe: a generalized schematic. The three broken lines represent three exponential functions that fit the core (on the right), the shell (in the middle) and the cloud (on the left) of prokaryotic genes (O'Malley and Koonin, 2011).

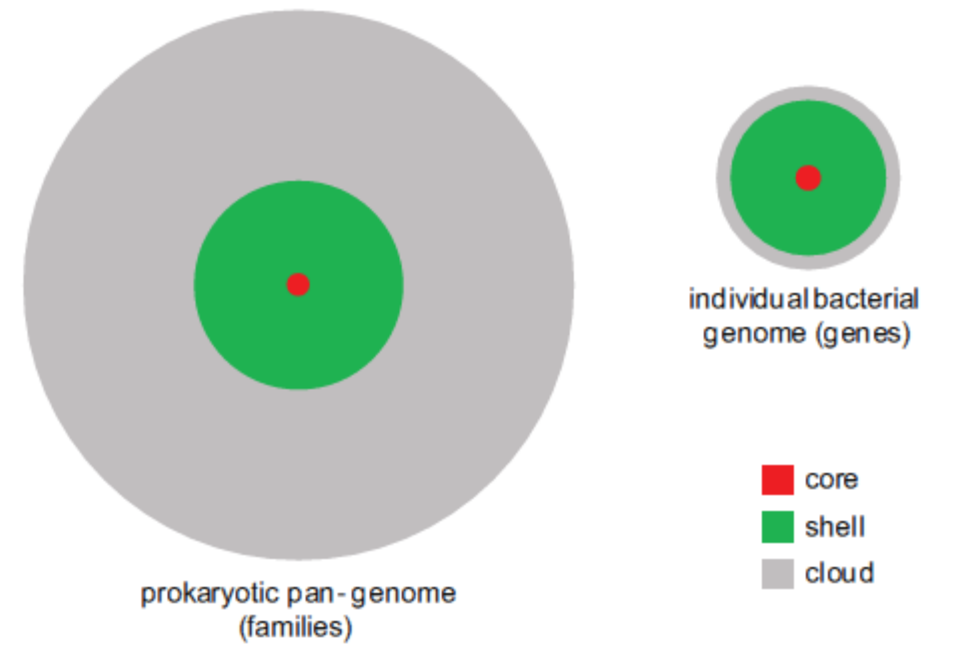
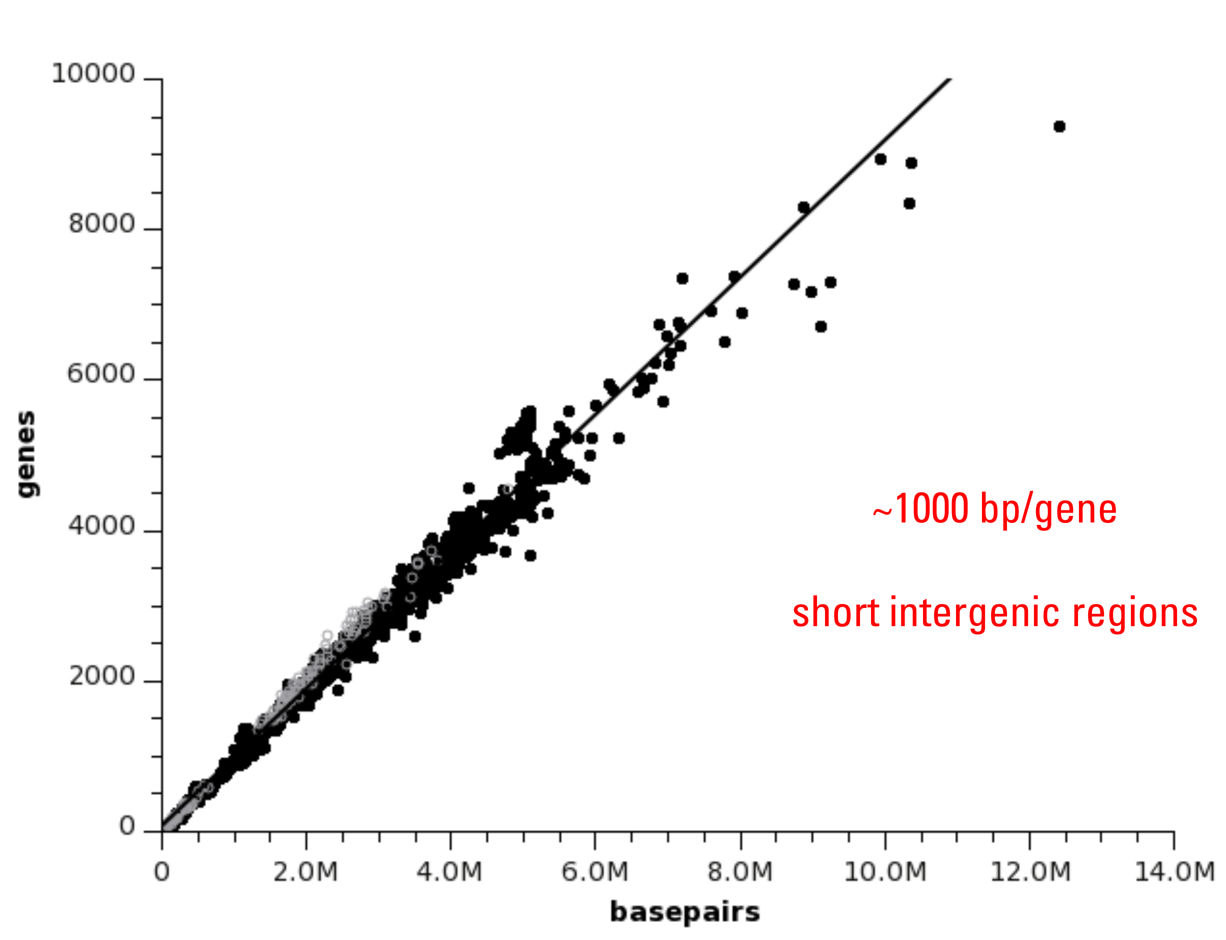

individual bacterial genome (genes)
prokaryotic pan-genome (families)
🟥 core
🟩 shell
⬜ cloud

FIGURE 5 | The core, shell, and cloud of microbial genes. A generalized schematic showing the approximate contributions of the core, shell, and cloud to the pangenomes of prokaryotes and individual genomes.

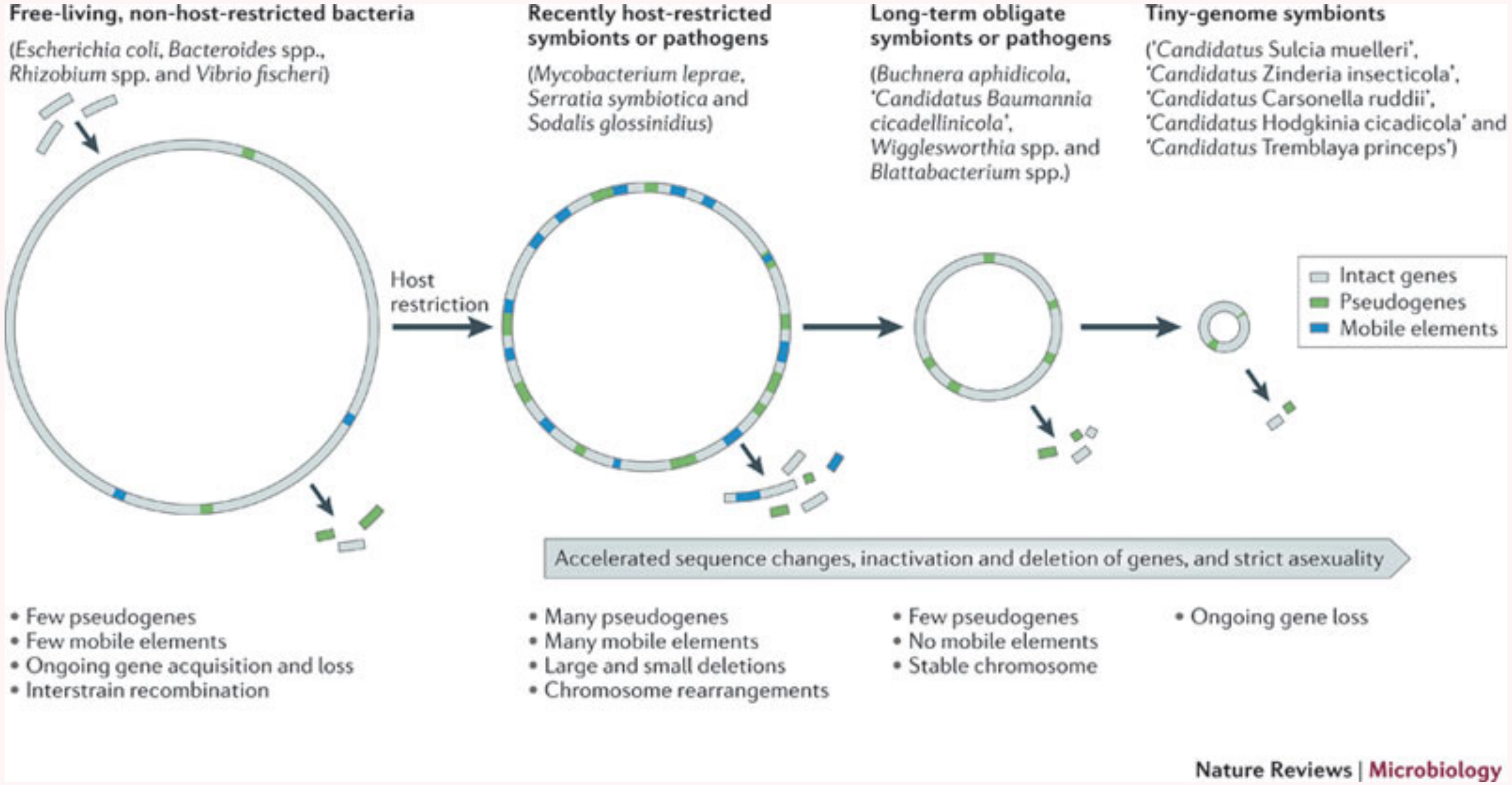
Genomic streamlining in prokaryotes: a needed little tangent



~1000 bp/gene

short intergenic regions

It will make sense by the end of this slide…



Free-living, non-host-restricted bacteria

(*Escherichia coli*, *Bacteroides* spp., *Rhizobium* spp. and *Vibrio fischeri*)

Recently host-restricted symbionts or pathogens

(*Mycobacterium leprae*, *Serratia symbiotica* and *Sodalis glossinidius*)

Long-term obligate symbionts or pathogens

(*Buchnera aphidicola*, '*Candidatus* Baumannia cicadellinicola', *Wigglesworthia* spp. and *Blattabacterium* spp.)

Tiny-genome symbionts

('*Candidatus* Sulcia muelleri', '*Candidatus* Zinderia insecticola', '*Candidatus* Carsonella ruddii', '*Candidatus* Hodgkinia cicadicola' and '*Candidatus* Tremblaya princeps')

Host restriction
□ Intact genes
🟩 Pseudogenes
🟦 Mobile elements
Accelerated sequence changes, inactivation and deletion of genes, and strict asexuality

- Few pseudogenes
- Few mobile elements
- Ongoing gene acquisition and loss
- Interstrain recombination

- Many pseudogenes
- Many mobile elements
- Large and small deletions
- Chromosome rearrangements

- Few pseudogenes
- No mobile elements
- Stable chromosome

- Ongoing gene loss

Nature Reviews | Microbiology

Horizontal Gene Transfer and Genome Stability

Horizontal gene transfer: transfer of genetic information between organisms, as opposed to vertical inheritance from parental organism(s)

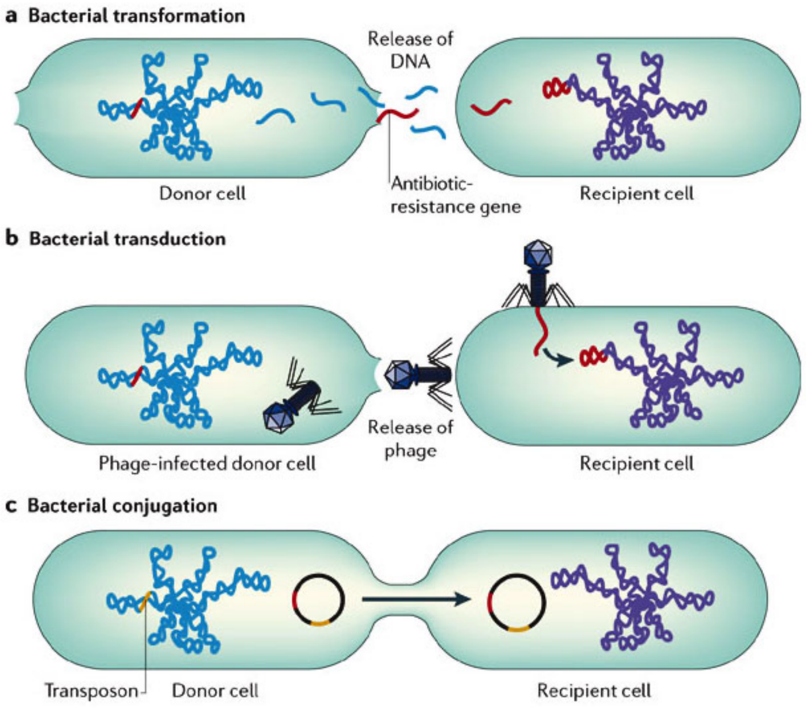
- May be extensive in nature
- May cross phylogenetic domain boundaries

Detecting horizontal gene flow

- Presence of genes typically found only in distantly related species
- Presence of a DNA with GC content or codon bias that differs significantly from remainder of genome

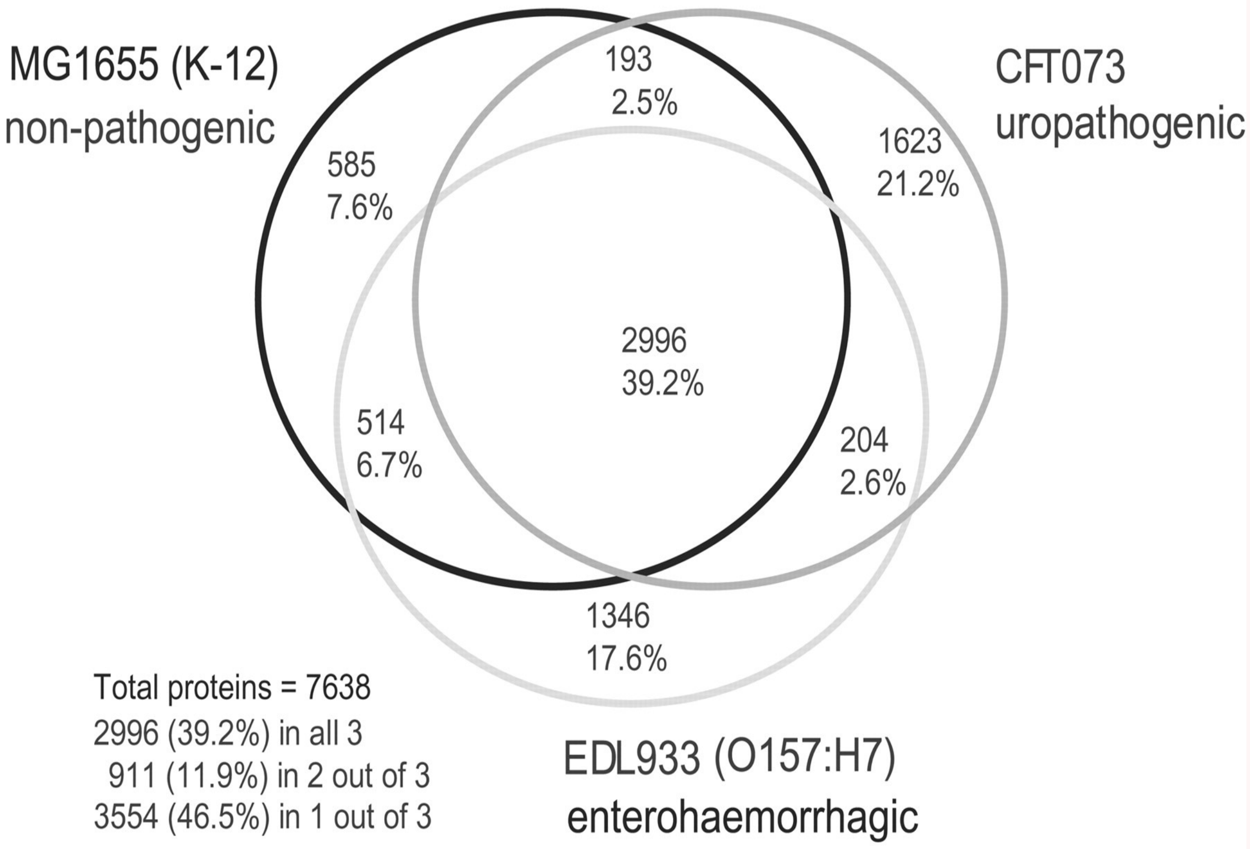
Horizontally transferred genes typically do not encode core metabolic functions

Where are the genes coming from?

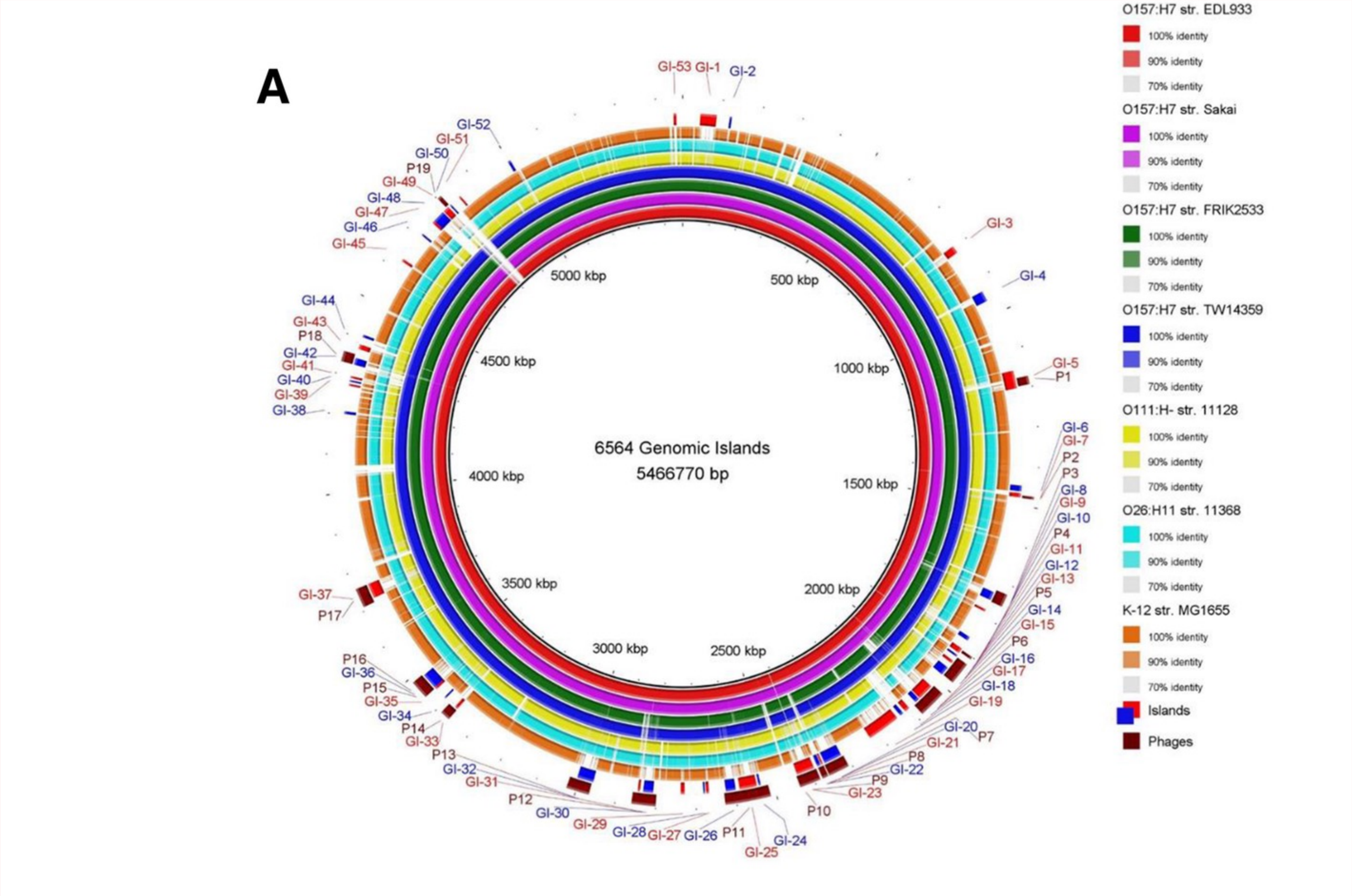


Copyright © 2006 Nature Publishing Group
Nature Reviews | Microbiology

Gene Content Variation among *E. coli* genomes. Evidence for horizontal transfer –



Welch et al (2002).


A
6564 Genomic Islands
5466770 bp
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp
4500 kbp
5000 kbp
GI-53
GI-1
GI-2
GI-3
GI-4
GI-5
P1
GI-6
GI-7
P2
P3
GI-8
GI-9
GI-10
P4
GI-11
GI-12
GI-13
P5
GI-14
GI-15
P6
GI-16
GI-17
GI-18
GI-19
GI-20
P7
GI-21
P8
GI-22
P9
GI-23
P10
GI-24
GI-25
P11
GI-26
GI-27
GI-28
GI-29
GI-30
P12
GI-31
GI-32
P13
GI-33
P14
GI-34
GI-35
P15
GI-36
P16
P17
GI-37
GI-38
GI-39
GI-40
GI-41
GI-42
P18
GI-43
GI-44
GI-45
GI-46
GI-47
GI-48
GI-49
P19
GI-50
GI-51
GI-52
O157:H7 str. EDL933
100% identity
90% identity
70% identity
O157:H7 str. Sakai
100% identity
90% identity
70% identity
O157:H7 str. FRIK2533
100% identity
90% identity
70% identity
O157:H7 str. TW14359
100% identity
90% identity
70% identity
O111:H- str. 11128
100% identity
90% identity
70% identity
O26:H11 str. 11368
100% identity
90% identity
70% identity
K-12 str. MG1655
100% identity
90% identity
70% identity
Islands
Phages

Accessory Genome: a treasure trove of information

- Host specificity
- Lifestyle
- Potential vectors
- Superpowers!

How can we find the accessory, juicy genes?

Different softwares

- Roary
  - 98% ID, 98% match length
  - No easy way to change them
- Panaroo
  - 98% ID, 98% match length
  - Easy way to change them
- PPanGGOLiN
  - It's… different

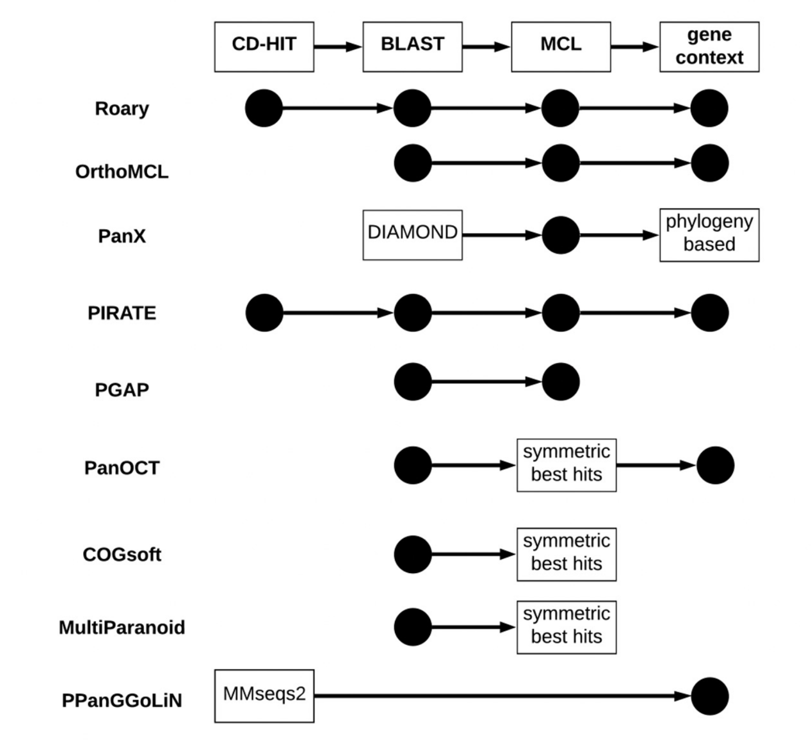
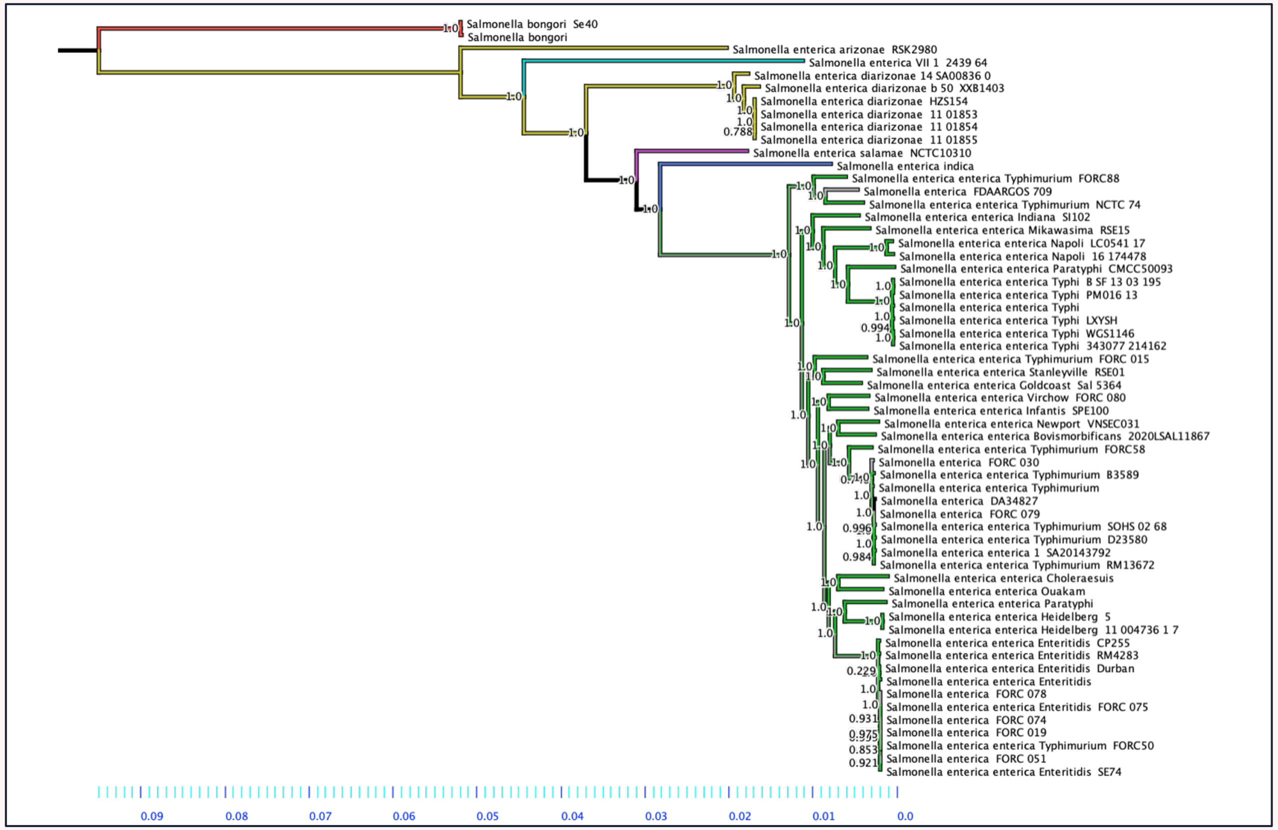


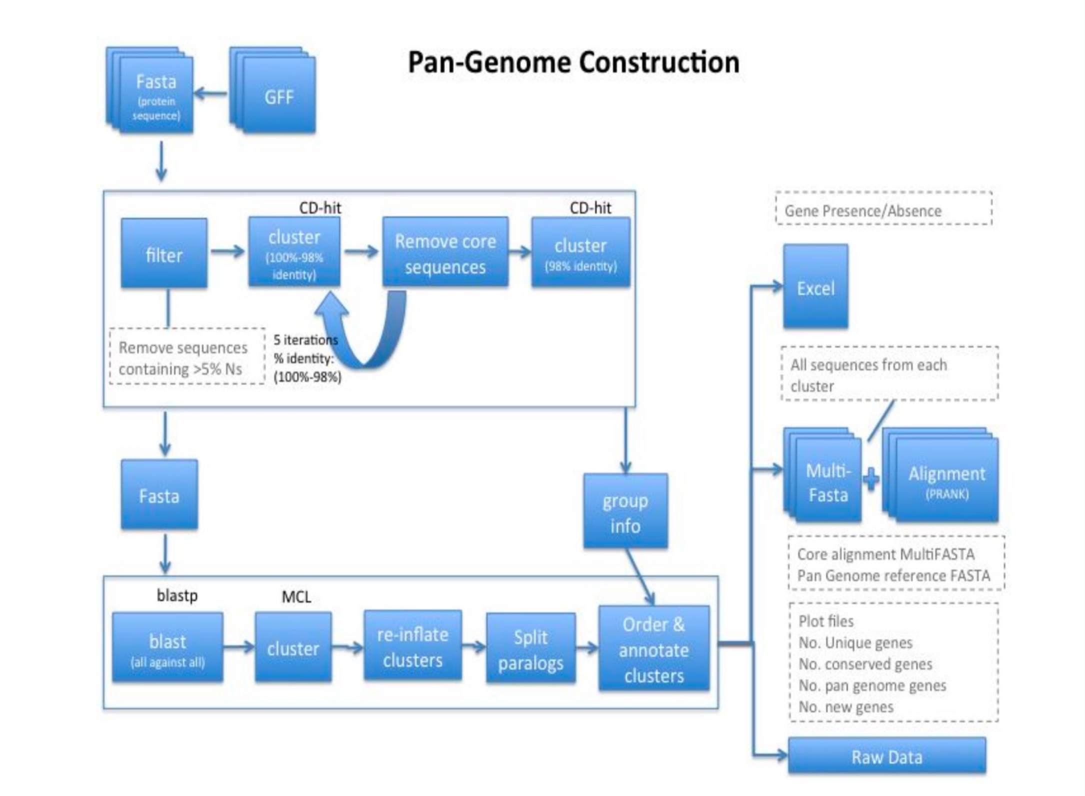
Figure S1: A comparison of the pipelines used by different pangenome analysis tools.

Pet Set of *Salmonella*

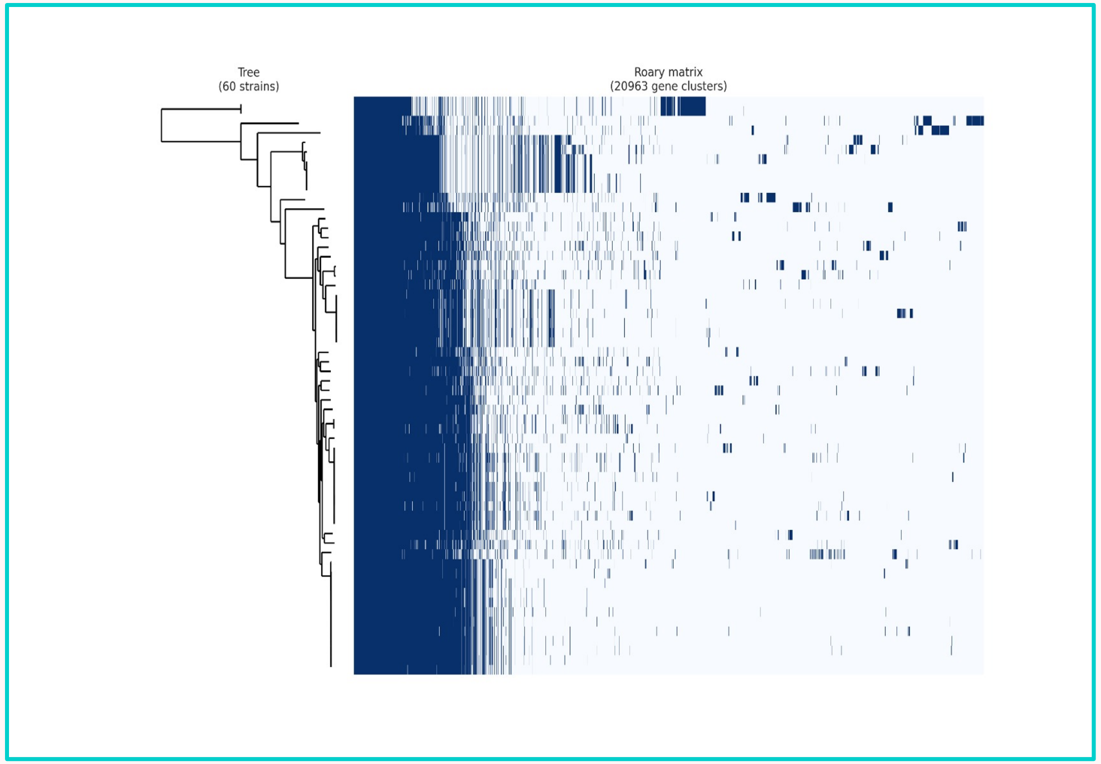

Salmonella bongori Se40
Salmonella bongori
Salmonella enterica arizonae RSK2980
Salmonella enterica VII 1 2439 64
Salmonella enterica diarizonae 14 SA00836 0
Salmonella enterica diarizonae b 50 XXB1403
Salmonella enterica diarizonae HZS154
Salmonella enterica diarizonae 11 01853
Salmonella enterica diarizonae 11 01854
Salmonella enterica diarizonae 11 01855
Salmonella enterica salamae NCTC10310
Salmonella enterica indica
Salmonella enterica enterica Typhimurium FORC88
Salmonella enterica FDAARGOS 709
Salmonella enterica enterica Typhimurium NCTC 74
Salmonella enterica enterica Indiana SI102
Salmonella enterica enterica Mikawasima RSE15
Salmonella enterica enterica Napoli LC0541 17
Salmonella enterica enterica Napoli 16 174478
Salmonella enterica enterica Paratyphi CMCC50093
Salmonella enterica enterica Typhi B SF 13 03 195
Salmonella enterica enterica Typhi PM016 13
Salmonella enterica enterica Typhi
Salmonella enterica enterica Typhi LXYSH
Salmonella enterica enterica Typhi WGS1146
Salmonella enterica enterica Typhi 343077 214162
Salmonella enterica enterica Typhimurium FORC 015
Salmonella enterica enterica Stanleyville RSE01
Salmonella enterica enterica Goldcoast Sal 5364
Salmonella enterica enterica Virchow FORC 080
Salmonella enterica enterica Infantis SPE100
Salmonella enterica enterica Newport VNSEC031
Salmonella enterica enterica Bovismorbificans 2020LSAL11867
Salmonella enterica enterica Typhimurium FORC58
Salmonella enterica FORC 030
Salmonella enterica enterica Typhimurium B3589
Salmonella enterica enterica Typhimurium
Salmonella enterica DA34827
Salmonella enterica FORC 079
Salmonella enterica enterica Typhimurium SOHS 02 68
Salmonella enterica enterica Typhimurium D23580
Salmonella enterica enterica 1 SA20143792
Salmonella enterica enterica Typhimurium RM13672
Salmonella enterica enterica Choleraesuis
Salmonella enterica enterica Ouakam
Salmonella enterica enterica Paratyphi
Salmonella enterica enterica Heidelberg 5
Salmonella enterica enterica Heidelberg 11 004736 1 7
Salmonella enterica enterica Enteritidis CP255
Salmonella enterica enterica Enteritidis RM4283
Salmonella enterica enterica Enteritidis Durban
Salmonella enterica enterica Enteritidis
Salmonella enterica FORC 078
Salmonella enterica enterica Enteritidis FORC 075
Salmonella enterica FORC 074
Salmonella enterica FORC 019
Salmonella enterica enterica Typhimurium FORC50
Salmonella enterica FORC 051
Salmonella enterica enterica Enteritidis SE74
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.0

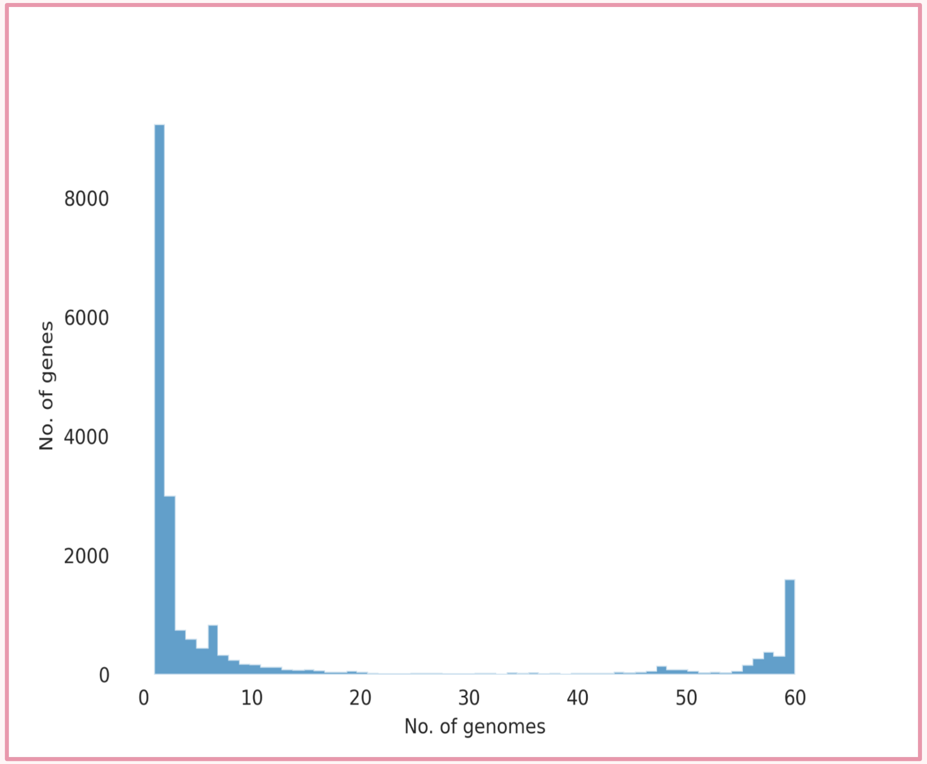
ROARY

- ASSUMPTION:
  - All entries from the same species
- Filter and precluster the proteins
- All against all comparison using BLASTP
- Cluster with MCL



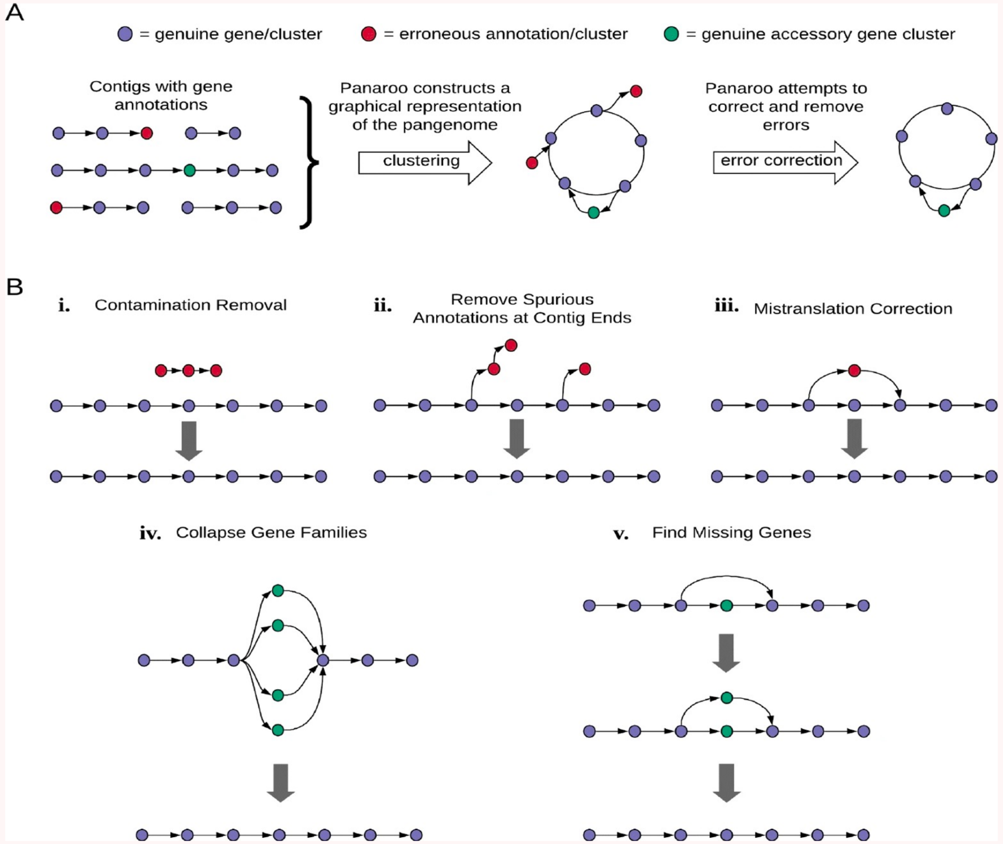
Roary's Output



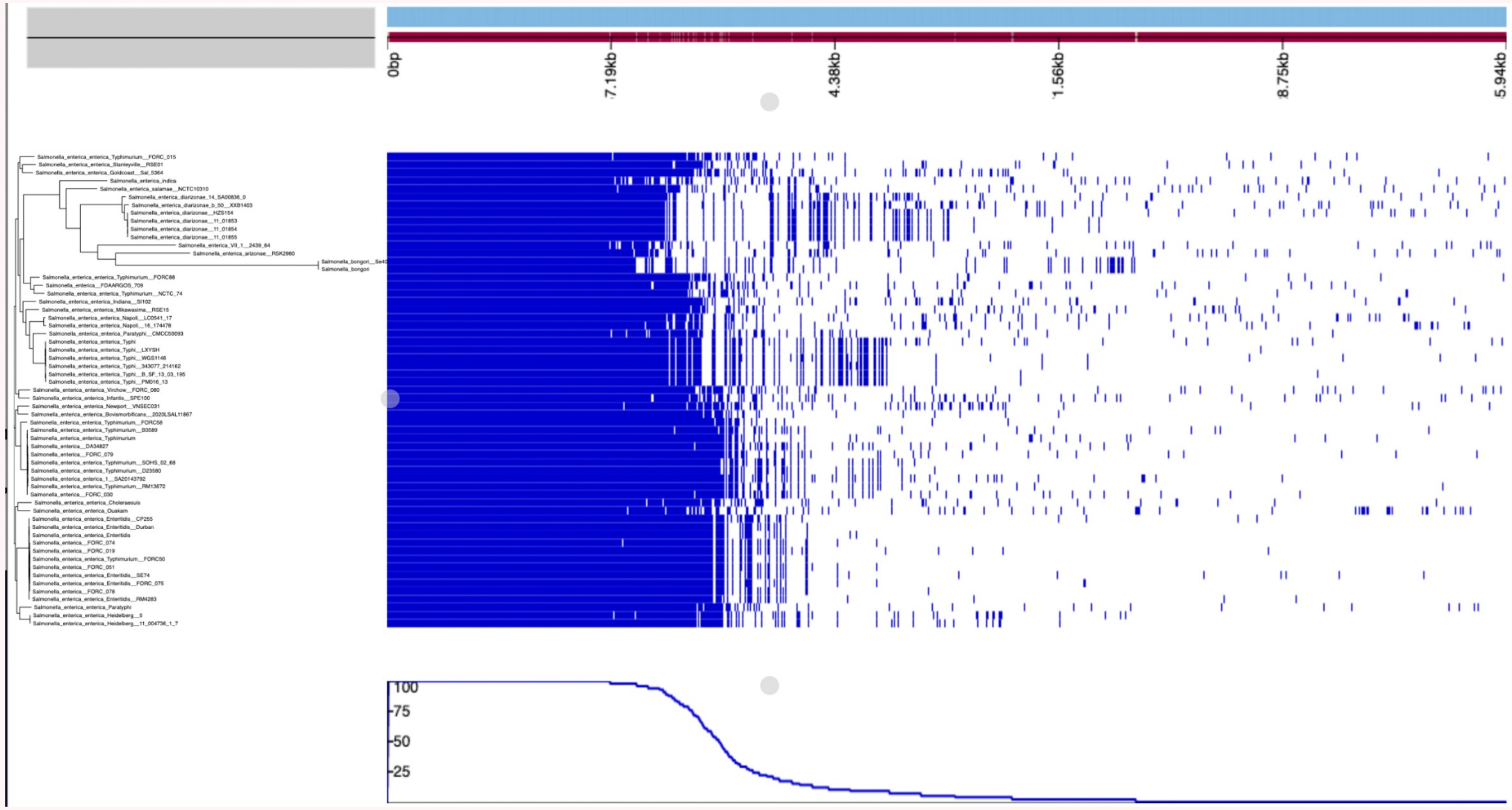


Panaroo

- Panaroo corrects annotation errors
- Contamination appears in the graph as poorly supported components
- Genes are often mis-annotated near contig breaks
- Corrects same DNA sequence translated in multiple reading frames
- Combine diverse gene families into a single gene

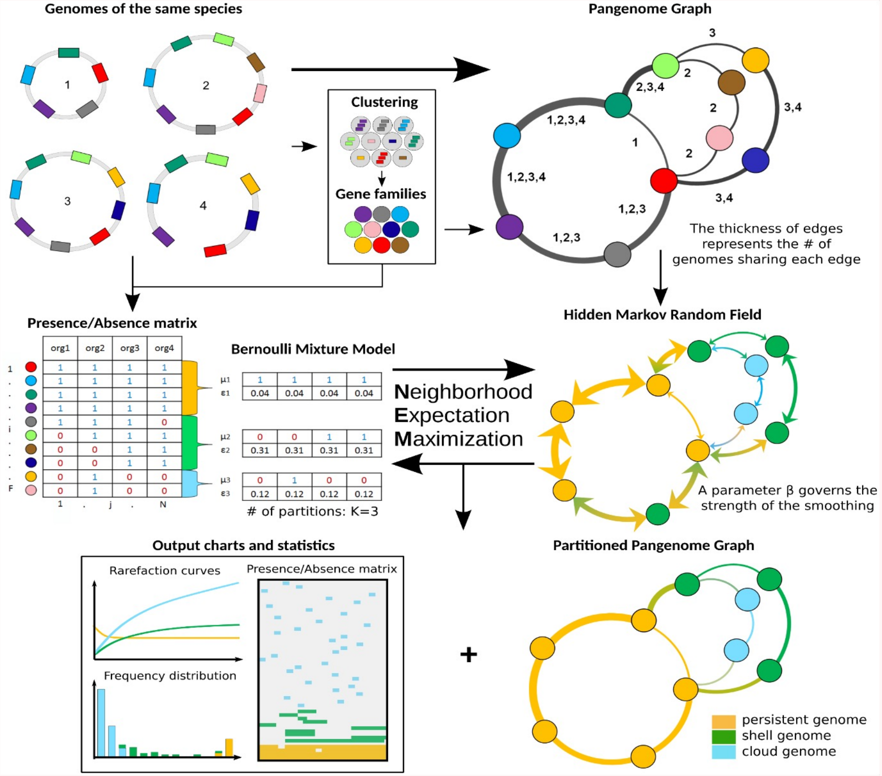

A
● = genuine gene/cluster
● = erroneous annotation/cluster
● = genuine accessory gene cluster
Contigs with gene annotations
Panaroo constructs a graphical representation of the pangenome
clustering
Panaroo attempts to correct and remove errors
error correction
B
i. Contamination Removal
ii. Remove Spurious Annotations at Contig Ends
iii. Mistranslation Correction
iv. Collapse Gene Families
v. Find Missing Genes

Panaroo’s Output

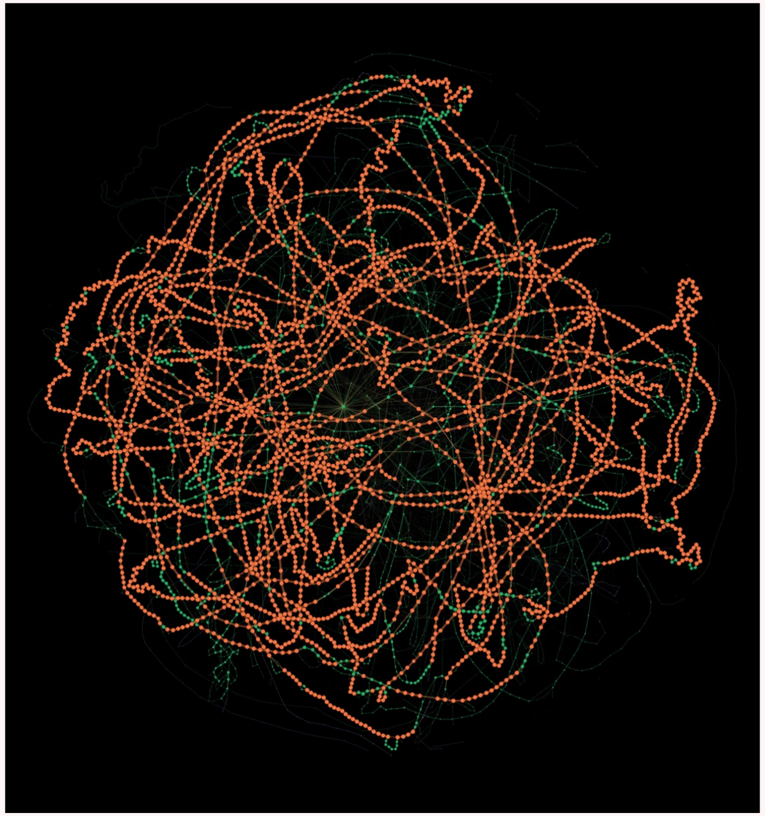

0bp
7.19kb
4.38kb
1.56kb
8.75kb
5.94kb
Salmonella_enterica_enterica_Typhimurium__FORC_015
Salmonella_enterica_enterica_Stanleyville__RSE01
Salmonella_enterica_enterica_Goldcoast__Sal_5364
Salmonella_enterica_indica
Salmonella_enterica_salamae__NCTC10310
Salmonella_enterica_diarizonae_14_SA00836_0
Salmonella_enterica_diarizonae_b_50__XXB1403
Salmonella_enterica_diarizonae__HZS154
Salmonella_enterica_diarizonae__11_01853
Salmonella_enterica_diarizonae__11_01854
Salmonella_enterica_diarizonae__11_01855
Salmonella_enterica_VII_1__2439_64
Salmonella_enterica_arizonae__RSK2980
Salmonella_bongori__Se4
Salmonella_bongori
Salmonella_enterica_enterica_Typhimurium__FORC88
Salmonella_enterica__FDAARGOS_709
Salmonella_enterica_enterica_Typhimurium__NCTC_74
Salmonella_enterica_enterica_Indiana__SI102
Salmonella_enterica_enterica_Mikawasima__RSE15
Salmonella_enterica_enterica_Napoli__LC0541_17
Salmonella_enterica_enterica_Napoli__16_174478
Salmonella_enterica_enterica_Paratyphi__CMCC50093
Salmonella_enterica_enterica_Typhi
Salmonella_enterica_enterica_Typhi__LXYSH
Salmonella_enterica_enterica_Typhi__WGS1146
Salmonella_enterica_enterica_Typhi__343077_214162
Salmonella_enterica_enterica_Typhi__B_SF_13_03_195
Salmonella_enterica_enterica_Typhi__PM016_13
Salmonella_enterica_enterica_Virchow__FORC_080
Salmonella_enterica_enterica_Infantis__SPE100
Salmonella_enterica_enterica_Newport__VNSEC031
Salmonella_enterica_enterica_Bovismorbificans__2020LSAL11867
Salmonella_enterica_enterica_Typhimurium__FORC58
Salmonella_enterica_enterica_Typhimurium__B3589
Salmonella_enterica_enterica_Typhimurium
Salmonella_enterica__DA34827
Salmonella_enterica__FORC_079
Salmonella_enterica_enterica_Typhimurium__SOHS_02_68
Salmonella_enterica_enterica_Typhimurium__D23580
Salmonella_enterica_enterica_1__SA20143792
Salmonella_enterica_enterica_Typhimurium__RM13672
Salmonella_enterica__FORC_030
Salmonella_enterica_enterica_Choleraesuis
Salmonella_enterica_enterica_Ouakam
Salmonella_enterica_enterica_Enteritidis__CP255
Salmonella_enterica_enterica_Enteritidis__Durban
Salmonella_enterica_enterica_Enteritidis
Salmonella_enterica__FORC_074
Salmonella_enterica__FORC_019
Salmonella_enterica_enterica_Typhimurium__FORC50
Salmonella_enterica__FORC_051
Salmonella_enterica_enterica_Enteritidis__SE74
Salmonella_enterica_enterica_Enteritidis__FORC_075
Salmonella_enterica__FORC_078
Salmonella_enterica_enterica_Enteritidis__RM4283
Salmonella_enterica_enterica_Paratyphi
Salmonella_enterica_enterica_Heidelberg__5
Salmonella_enterica_enterica_Heidelberg__11_004736_1_7
100
75
50
25

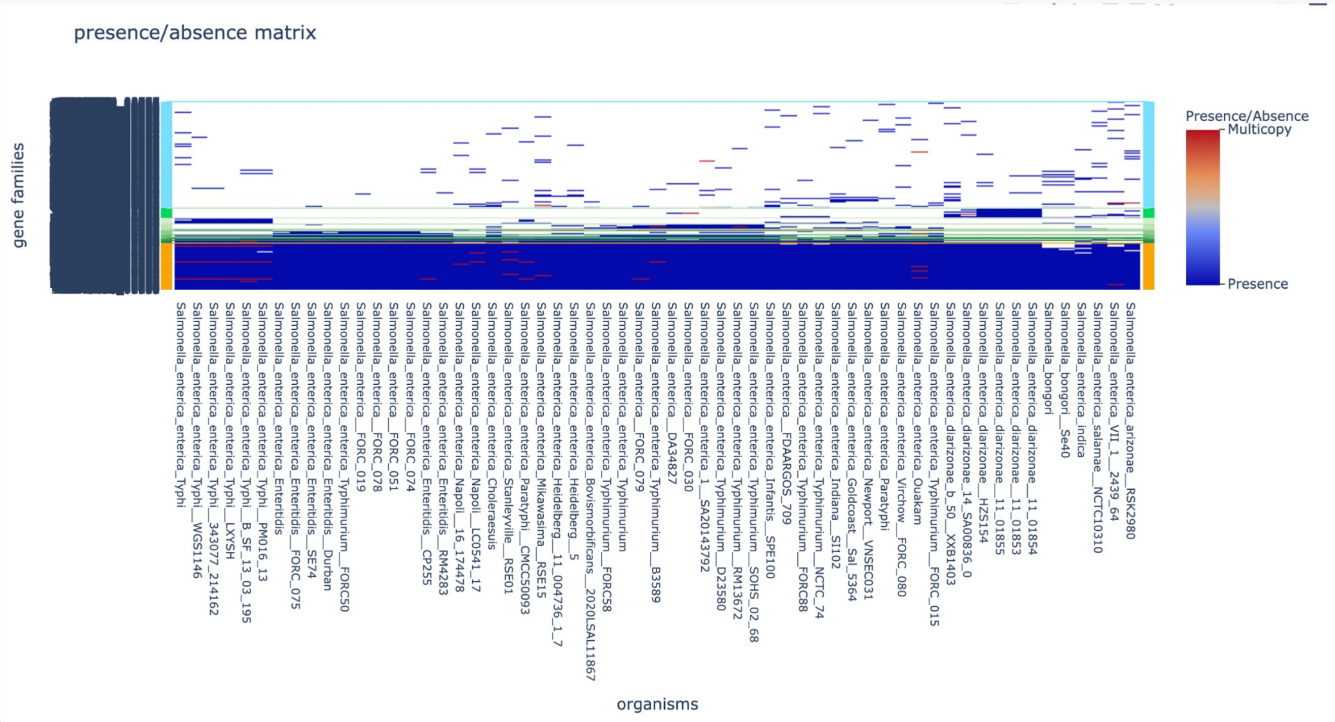
PPanGGOLiN

- Clustering through MMSeqs
- Statistical analysis to assign families to partitions
- No reference tree (but can be produced)
- Can be used for different taxonomic levels
- PPanGGOLiN returns a partitioned pangenome graph
    - persistent, shell and cloud partitions
    - overlaid on the neighborhood graph

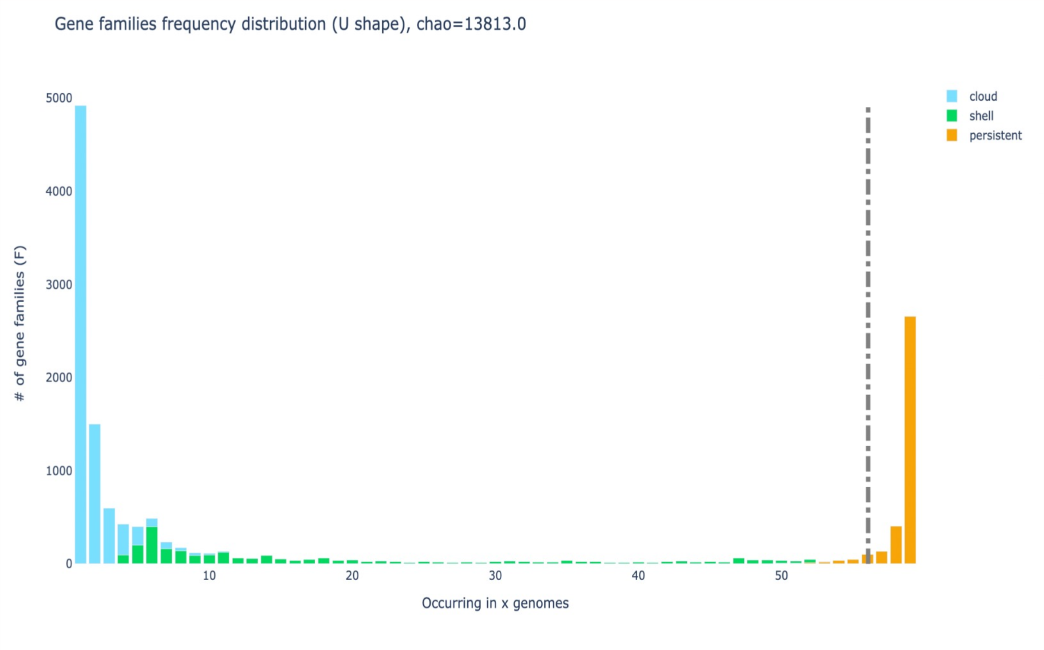


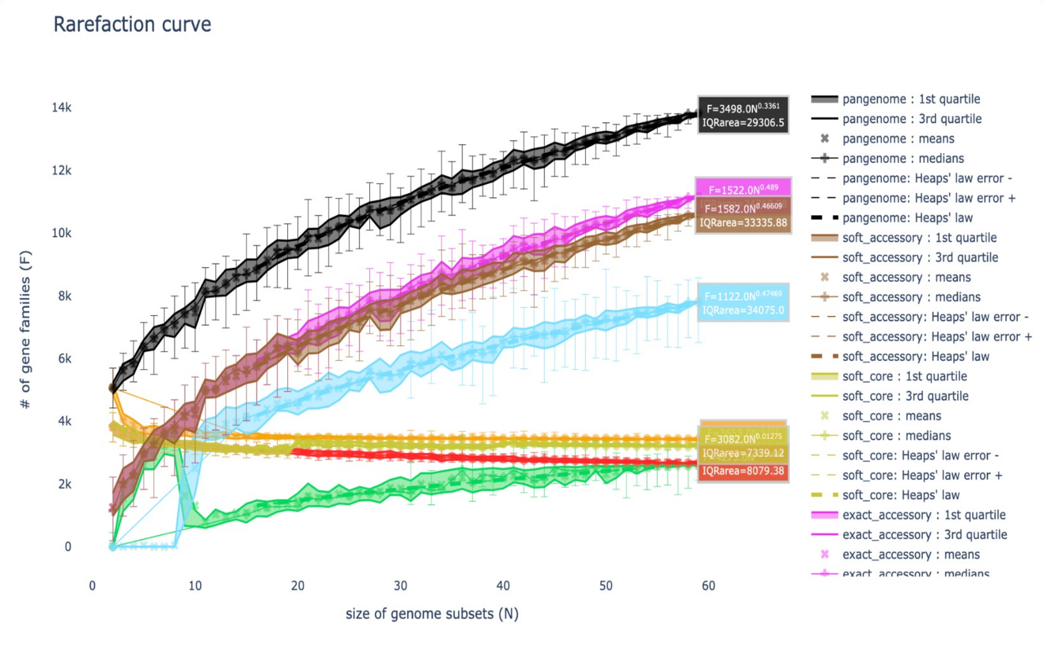
PPanGGOLiN's Output




presence/absence matrix
gene families
Presence/Absence
Multicopy
Presence
Salmonella_enterica_enterica_Typhi
Salmonella_enterica_enterica_Typhi__WGS1146
Salmonella_enterica_enterica_Typhi__343077_214162
Salmonella_enterica_enterica_Typhi__LXYSH
Salmonella_enterica_enterica_Typhi__B_SF_13_03_195
Salmonella_enterica_enterica_Typhi__PM016_13
Salmonella_enterica_enterica_Enteritidis
Salmonella_enterica_enterica_Enteritidis__FORC_075
Salmonella_enterica_enterica_Enteritidis__SE74
Salmonella_enterica_enterica_Enteritidis__Durban
Salmonella_enterica_enterica_Typhimurium__FORC50
Salmonella_enterica__FORC_019
Salmonella_enterica__FORC_078
Salmonella_enterica__FORC_051
Salmonella_enterica__FORC_074
Salmonella_enterica_enterica_Enteritidis__CP255
Salmonella_enterica_enterica_Enteritidis__RM4283
Salmonella_enterica_enterica_Napoli__16_174478
Salmonella_enterica_enterica_Napoli__LC0541_17
Salmonella_enterica_enterica_Choleraesuis
Salmonella_enterica_enterica_Stanleyville__RSE01
Salmonella_enterica_enterica_Paratyphi__CMCC50093
Salmonella_enterica_enterica_Mikawasima__RSE15
Salmonella_enterica_enterica_Heidelberg__11_004736_1_7
Salmonella_enterica_enterica_Heidelberg__5
Salmonella_enterica_enterica_Bovismorbificans__2020LSAL11867
Salmonella_enterica_enterica_Typhimurium__FORC58
Salmonella_enterica_enterica_Typhimurium
Salmonella_enterica__FORC_079
Salmonella_enterica_enterica_Typhimurium__B3589
Salmonella_enterica__DA34827
Salmonella_enterica__FORC_030
Salmonella_enterica_enterica_1__SA20143792
Salmonella_enterica_enterica_Typhimurium__D23580
Salmonella_enterica_enterica_Typhimurium__RM13672
Salmonella_enterica_enterica_Typhimurium__SOHS_02_68
Salmonella_enterica_enterica_Infantis__SPE100
Salmonella_enterica__FDAARGOS_709
Salmonella_enterica_enterica_Typhimurium__FORC88
Salmonella_enterica_enterica_Typhimurium__NCTC_74
Salmonella_enterica_enterica_Indiana__SI102
Salmonella_enterica_enterica_Goldcoast__Sal_5364
Salmonella_enterica_enterica_Newport__VNSEC031
Salmonella_enterica_enterica_Paratyphi
Salmonella_enterica_enterica_Virchow__FORC_080
Salmonella_enterica_enterica_Ouakam
Salmonella_enterica_enterica_Typhimurium__FORC_015
Salmonella_enterica_diarizonae_b_50__XXB1403
Salmonella_enterica_diarizonae_14_SA00836_0
Salmonella_enterica_diarizonae__HZS154
Salmonella_enterica_diarizonae__11_01855
Salmonella_enterica_diarizonae__11_01853
Salmonella_enterica_diarizonae__11_01854
Salmonella_bongori
Salmonella_bongori__Se40
Salmonella_enterica_indica
Salmonella_enterica_salamae__NCTC10310
Salmonella_enterica_VII_1__2439_64
Salmonella_enterica_arizonae__RSK2980
organisms

PPanGGOLiN's Output -2

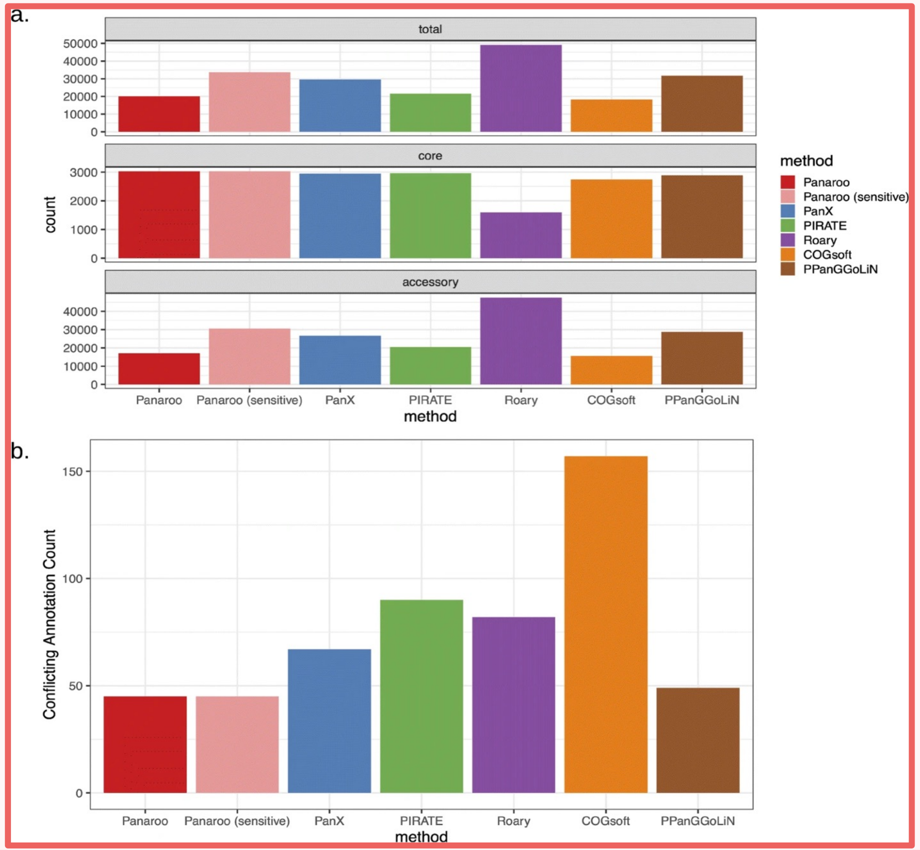

Gene families frequency distribution (U shape), chao=13813.0


Rarefaction curve

How to Choose a Tool?

Considerations:

- Desired visualization(s)
- Data set size and computational resources
- Sensitivity vs specificity


a.
total
core
accessory
count
Panaroo
Panaroo (sensitive)
PanX
PIRATE
Roary
COGsoft
PPanGGoLiN
method
method
Panaroo
Panaroo (sensitive)
PanX
PIRATE
Roary
COGsoft
PPanGGoLiN
b.
Conflicting Annotation Count
Panaroo
Panaroo (sensitive)
PanX
PIRATE
Roary
COGsoft
PPanGGoLiN
method