EVE 161:
Microbial Phylogenomics

Class #12:
Annotation

UC Davis, Winter 2018
Instructor: Jonathan Eisen
Teaching Assistant: Cassie Ettinger

- Presenters

Topics

General Workflow for Annotation

- Prokaryotic vs. Eukaryotic major differences

Lots of jargon / terms

- NCBI
- Complete vs Draft
- Pan genome
- Blast
- ncRNAs
- CRISPR
- Prokaryotic vs. eukaryotic annotation

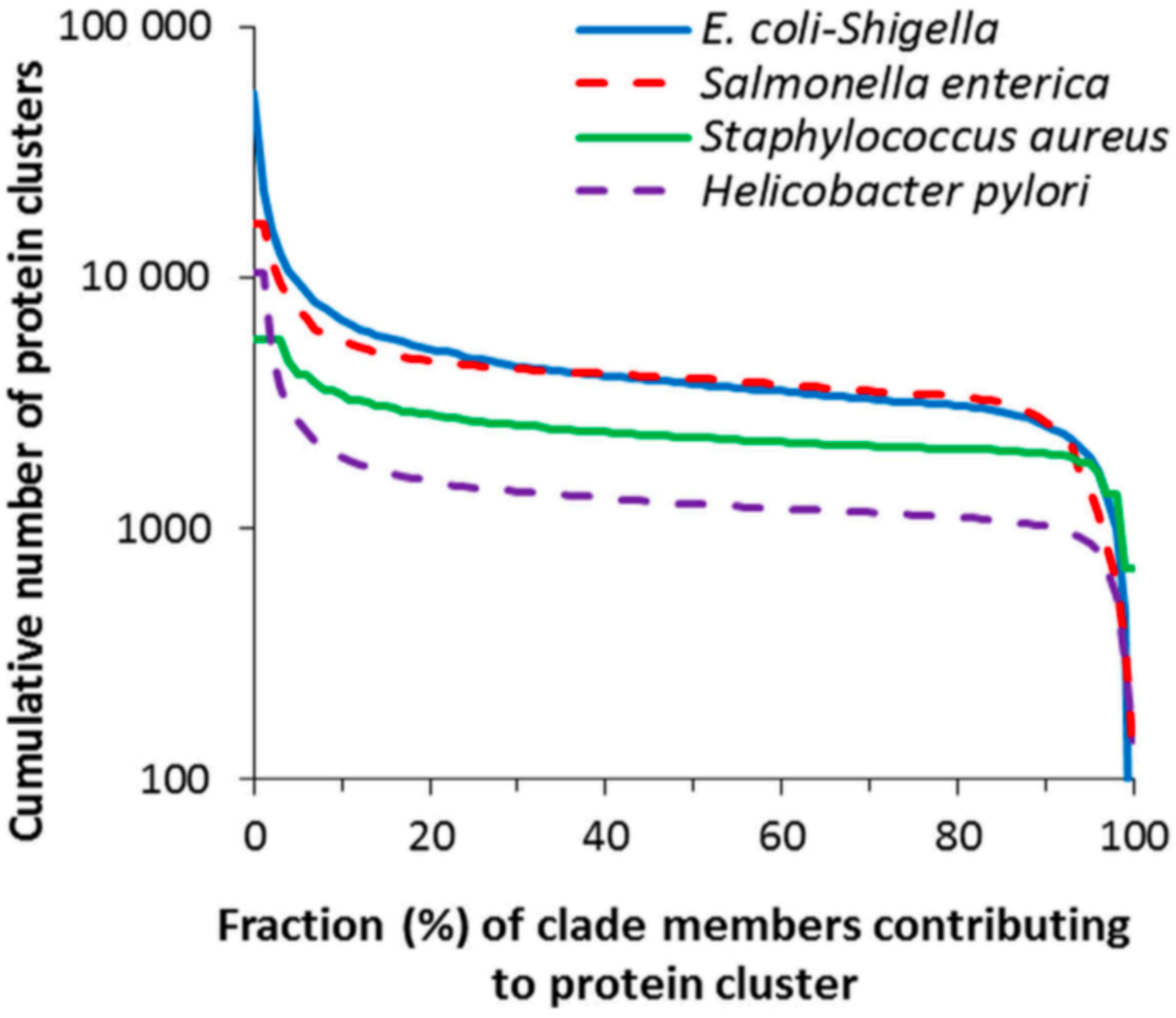
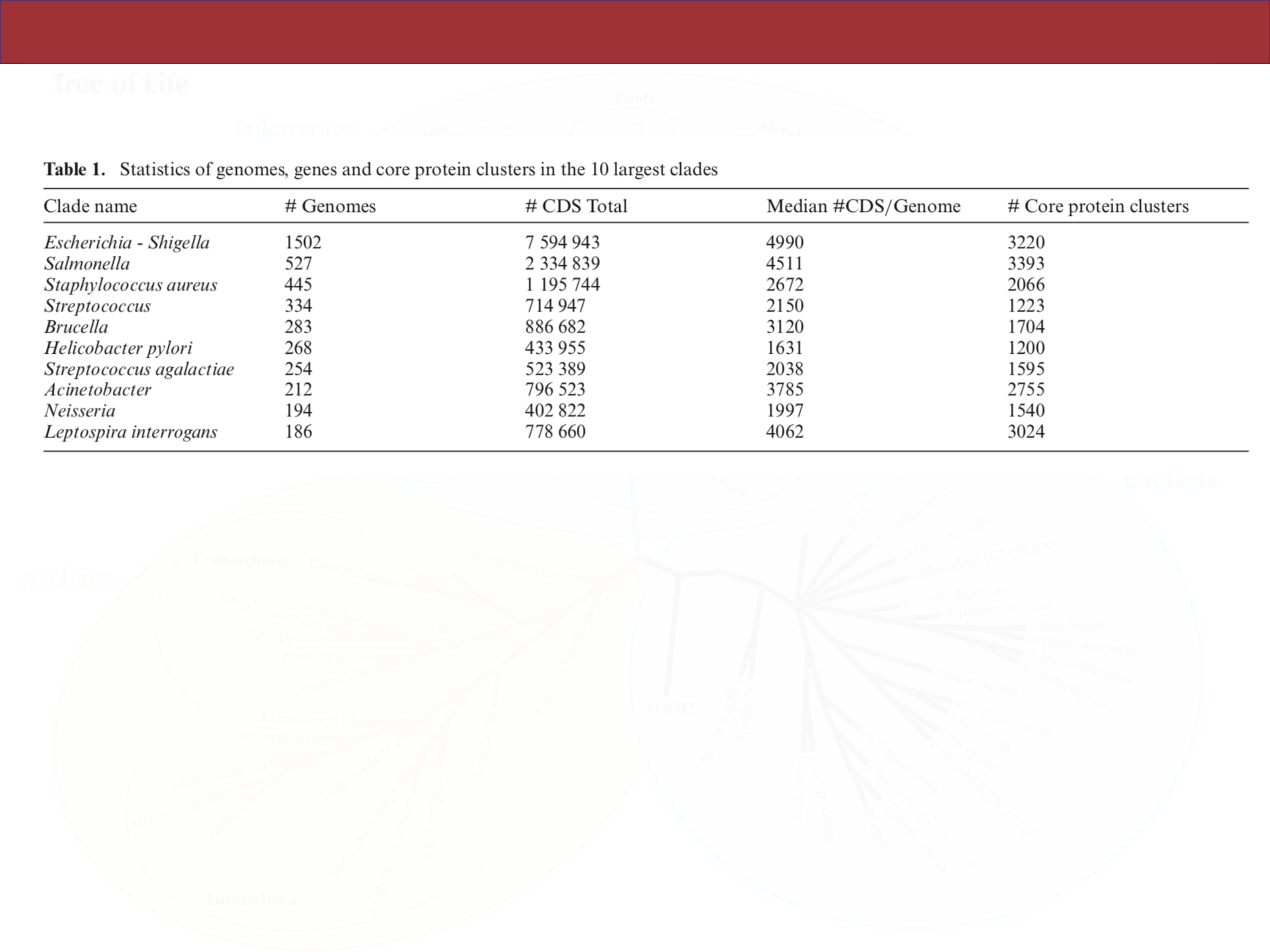


Figure 1. Cumulative number of protein clusters (Y) is defined for a given X (%) as the number of clusters containing proteins from fraction $x \geq X$ of all members of the clade. Data are presented for the four well studied clades.

Table 1. Statistics of genomes, genes and core protein clusters in the 10 largest clades

Clade name	# Genomes	# CDS Total	Median #CDS/Genome	# Core protein clusters
Escherichia - Shigella	1502	7 594 943	4990	3220
Salmonella	527	2 334 839	4511	3393
Staphylococcus aureus	445	1 195 744	2672	2066
Streptococcus	334	714 947	2150	1223
Brucella	283	886 682	3120	1704
Helicobacter pylori	268	433 955	1631	1200
Streptococcus agalactiae	254	523 389	2038	1595
Acinetobacter	212	796 523	3785	2755
Neisseria	194	402 822	1997	1540
Leptospira interrogans	186	778 660	4062	3024



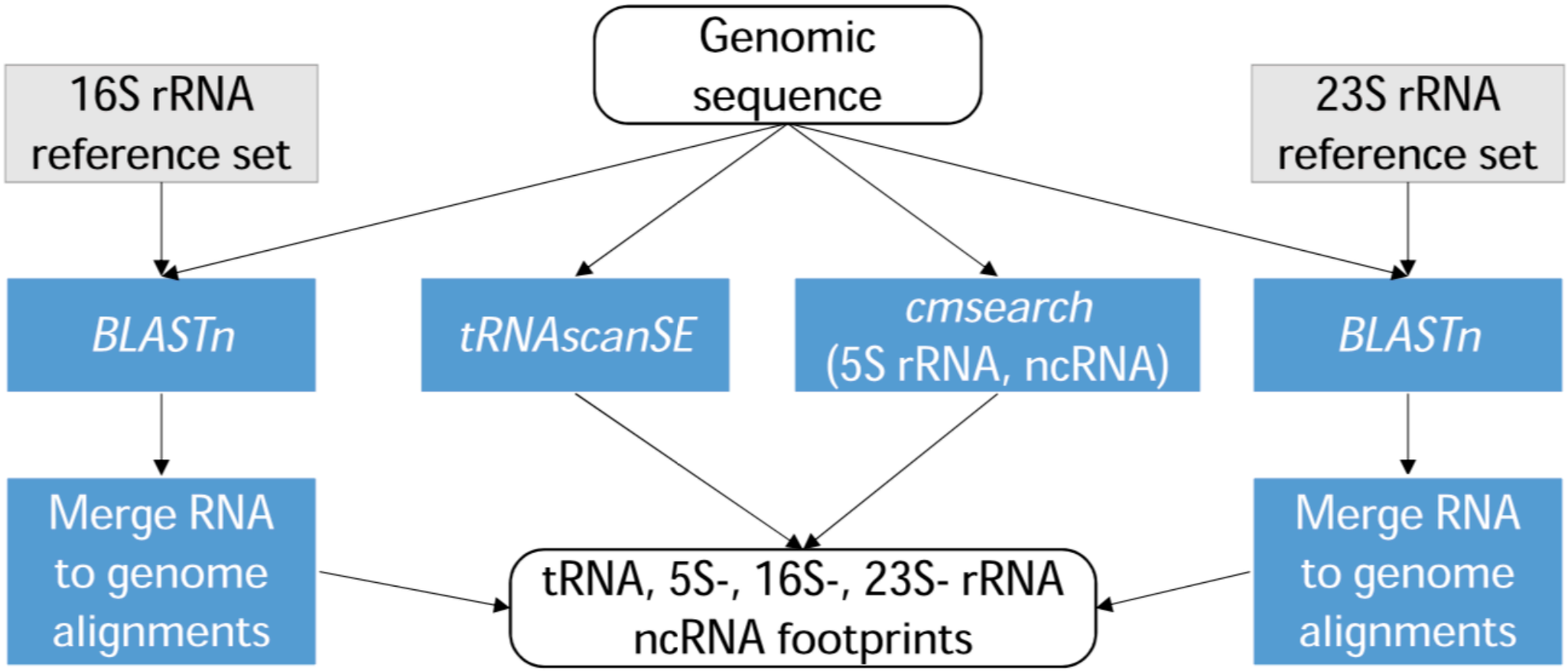


Figure 2. A fragment of the PGAP execution graph: prediction of structural RNA genes (ncRNA, tRNA, 5S-, 16S-, 23S- rRNA).

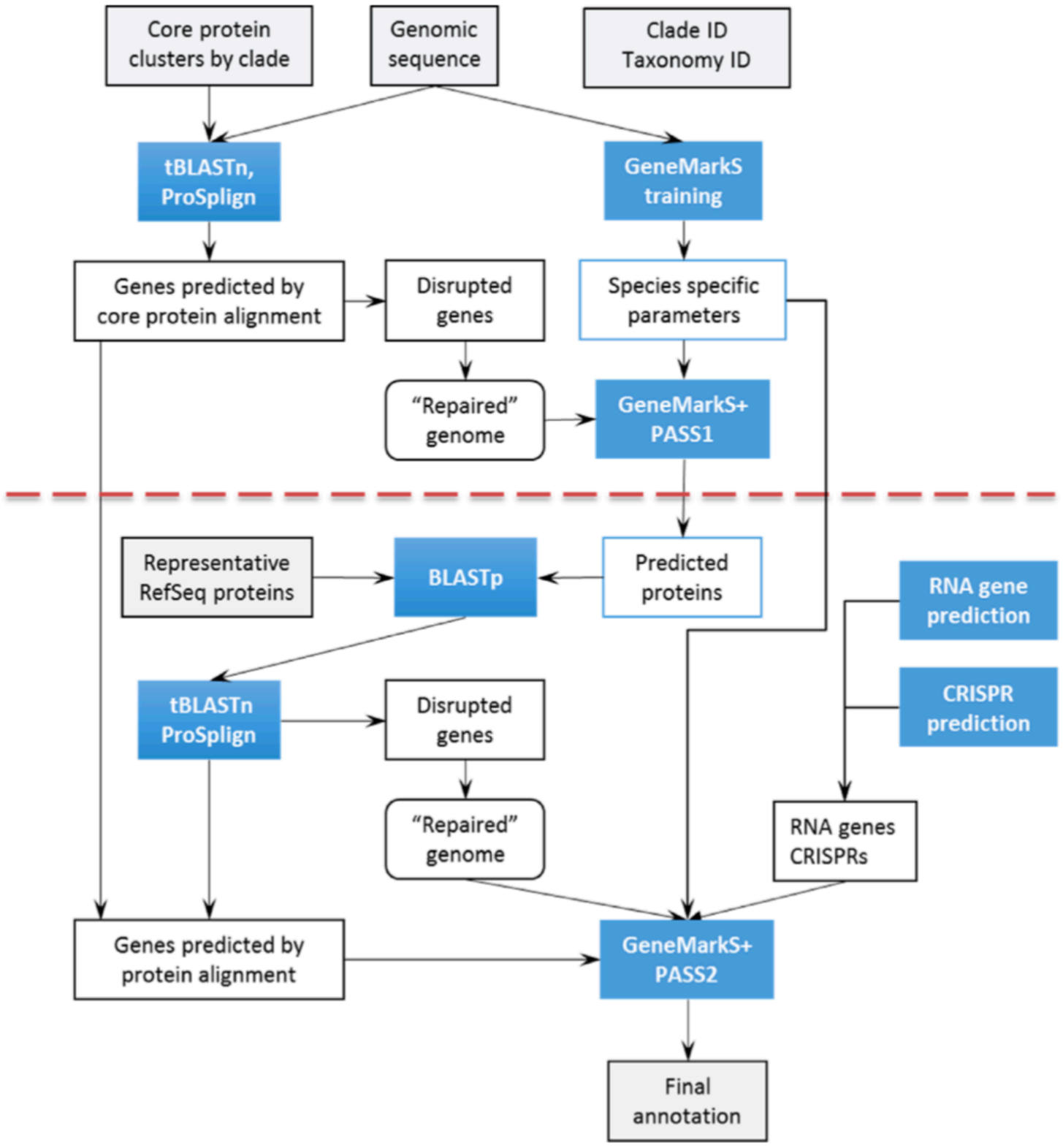


Figure 3. Flowchart of PGAP. The red dotted line indicates separation between pass one and pass two (see text for details).

Figure 4. A region in the *Deinococcus radiodurans* R1 genome assembly (GCA_000008565.1) contains three overlapping ORFs predicted ab initio as CDSs in the first pass of PGAP. Automatic evaluation of the cross-species protein evidence through the second pass of PGAP reveals proteins bearing homology to all three fragments. Alignment of the proteins to the genome reveals otherwise unpredicted frameshifts. Green bars represent genes, red bars – coding regions; grey bars – alignments with red vertical bars indicating mismatches. (**A**) A region of Chromosome 1 of *D. radiodurans* (AE000513.1) containing the three CDS features is displayed alongside the six-frame translation. (**B**) The same region, updated to include final annotation markup with a frameshifted CDS as well as supporting proteins that demonstrate a consistent pattern and location of two frameshifts (marked by arrows at positions 100 733 and 100 959).

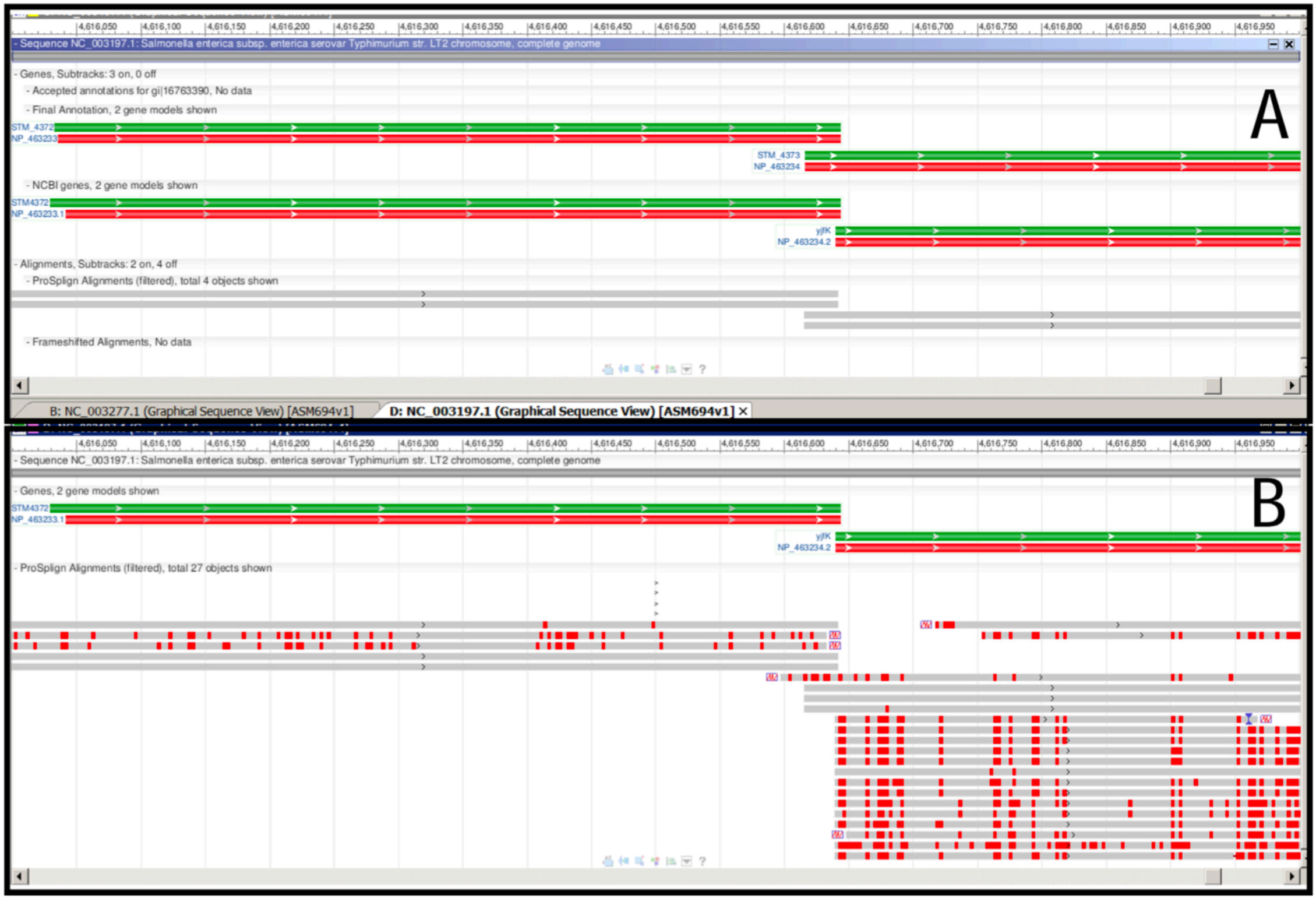


Figure 5. Annotation of genome of *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* str. LT2 (NC_003197). Protein alignment provides support for gene start selection. See legend to Figure 4 for description of the meaning of green, red and gray bars. (**A**) the first round of alignments of protein representatives from the 'core' protein clusters doesn't give enough evidence for gene start selection. (**B**) the second round of alignments clearly supports a shorter gene model which does not overlap with the upstream gene.

```
##Genome-Annotation-Data-START##
Annotation Provider                 :: NCBI
Annotation Date                     :: 02/05/2016 09:49:57
Annotation Pipeline                 :: NCBI Prokaryotic Genome
                                       Annotation Pipeline
Annotation Method                   :: Best-placed reference protein
                                       set; GeneMarkS+
Annotation Software revision        :: 3.1
Features Annotated                  :: Gene; CDS; rRNA; tRNA; ncRNA;
                                       repeat_region
Genes (total)                       :: 3,034
CDS (total)                         :: 2,945
Genes (coding)                      :: 2,921
CDS (coding)                        :: 2,921
Genes (RNA)                         :: 89
rRNAs                               :: 6, 6, 6 (5S, 16S, 23S)
complete rRNAs                      :: 6, 6, 6 (5S, 16S, 23S)
tRNAs                               :: 67
ncRNAs                              :: 4
Pseudo Genes (total)                :: 24
Pseudo Genes (ambiguous residues)   :: 0 of 24
Pseudo Genes (frameshifted)         :: 9 of 24
Pseudo Genes (incomplete)           :: 6 of 24
Pseudo Genes (internal stop)        :: 9 of 24
##Genome-Annotation-Data-END##
```

Figure 6. A summary of PGAP genome annotation process is provided in the COMMENT section of GenBank and RefSeq records. The example is given for *Listeria monocytogenes* strain CFSAN010068, complete genome NZ_CP014250.1.

Table 2. Comparison of the genome annotations generated by PGAP-3.1 with the GenBank annotations of the same genomes (snapshot from February 2016)

Data point	# of genes in GenBank annotation	# of predictions matching annotation in 3′ end	% of annotated genes missed in predictions	# of predictions matching annotation in 5′ and 3′ ends	% of predictions with mismatch in 5′ end	# of hypothetical genes in GenBank annotation
Bacillus subtilis	4185	4044	3.4	3768	6.8	232
Chlamydia trachomatis	892	886	0.7	822	7.2	285
E. coli	4140	4093	1.1	3915	4.3	0
Franciscella tularensis	1602	1589	0.8	1330	16.3	202
Mycobacterium leprae	1605	1599	0.4	1391	13.0	14
Mycobacterium tuberculosis	4018	3954	1.6	3342	15.5	508
Neisseria meningitidis	2063	1958	5.1	1705	12.9	529
Pseudomonas aeruginosa	5571	5531	0.7	5051	8.7	1693
Salmonella enterica	4554	4485	1.5	4031	10.1	10
Yersinia pestis	4083	4031	1.3	3429	14.9	332

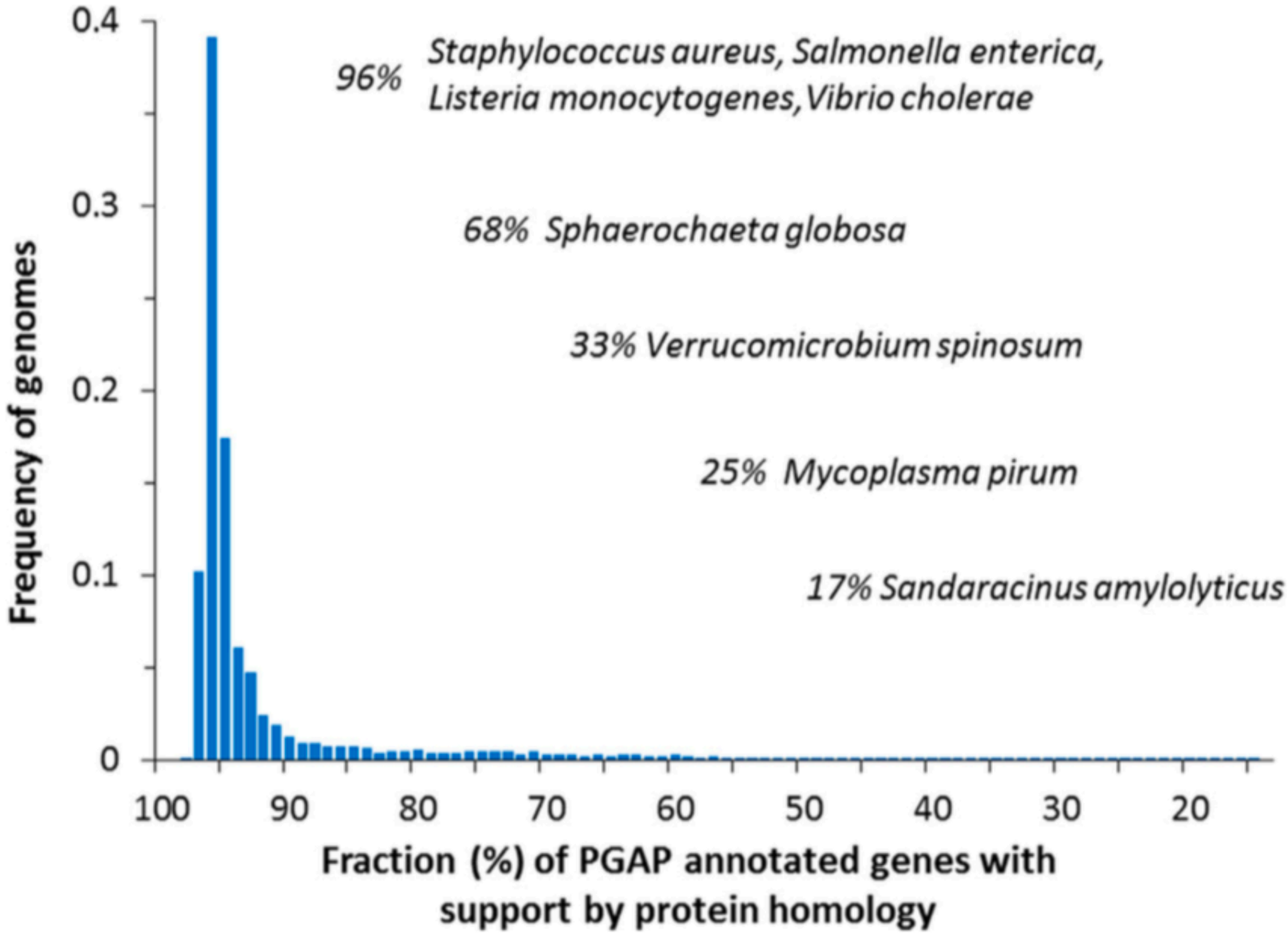


Figure 7. Frequency histogram of genomes with respect to the fraction of the whole complement of genes supported by similarity to proteins in RefSeq. In about 50% of the total set of genomes in consideration, mostly from highly populated clades, more than 95% of protein-coding genes are supported by protein sequence similarity.