Comparative Genomics Team 2

Kara Lee, Kristine Lacek, Courtney Astore, Ujani Hazra, Jianshu Zhao

Presentation outline

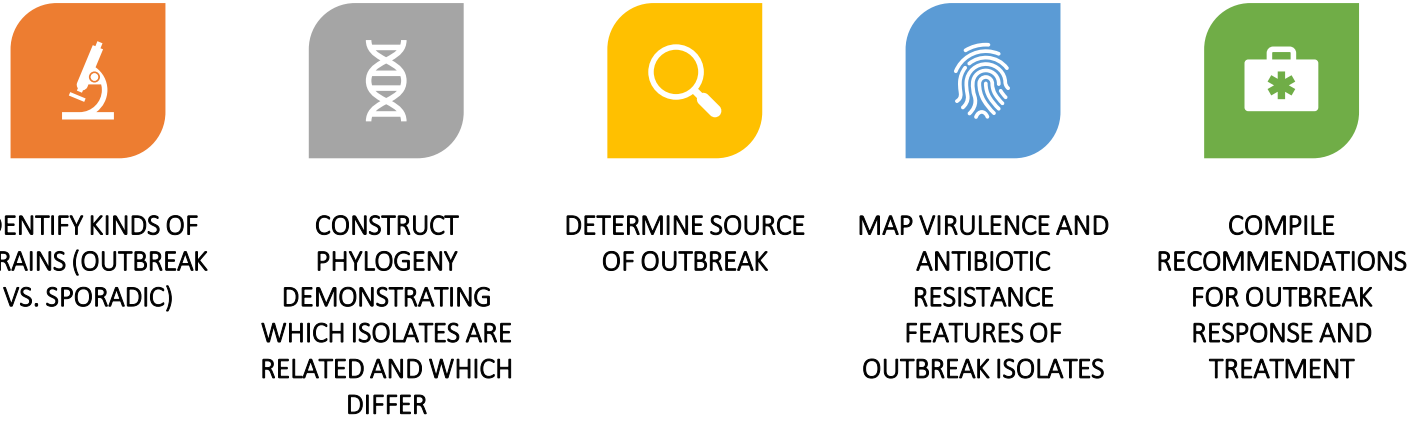
- Pathogenic organism
- Overview of Comparative Genomics & Objective
- Comparative Genomics pipeline & Software Selection
  - ANI
  - MLST
  - SNP Typing
- Future directions & deliverables

Campylobacter jejuni

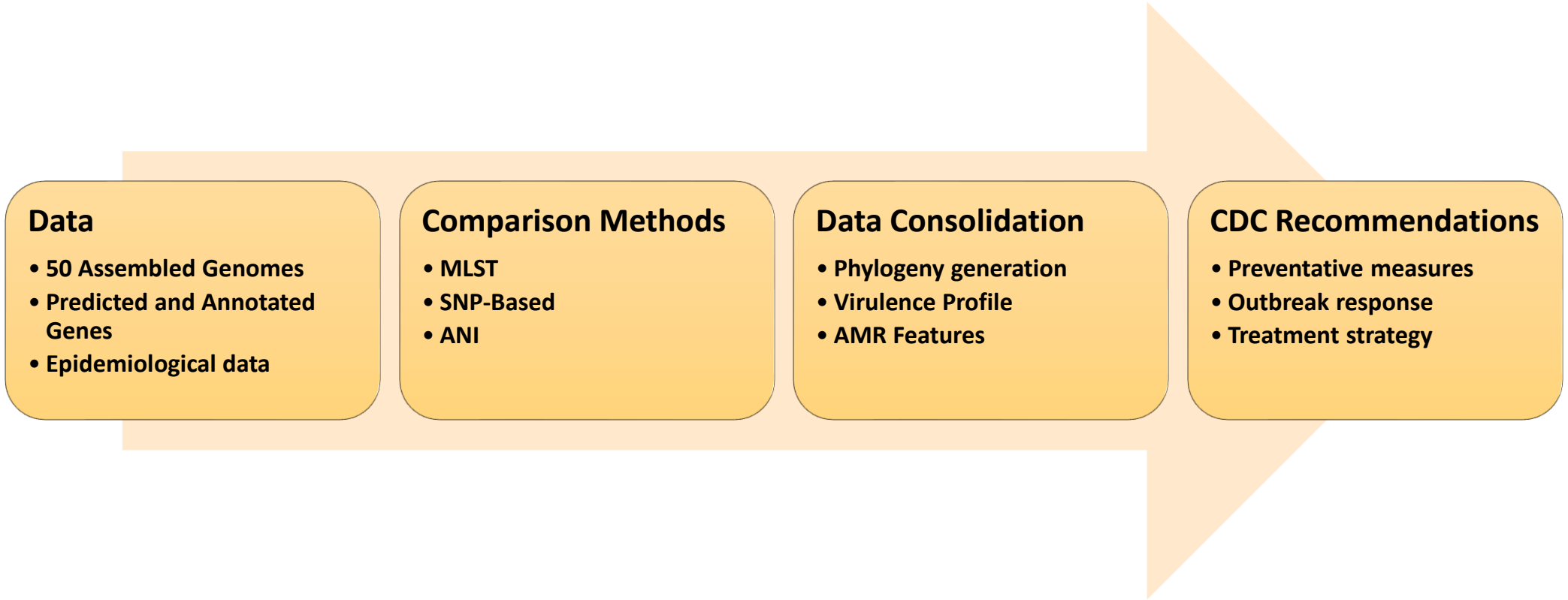
- Shares high sequence homology to *Campylobacter coli*
- It colonizes the intestinal mucosa
- Highly associated with acute gastroenteritis in humans causing global bacterial food poisoning
- We are interested in
  - Virulence factors
  - AMR Profiles

Comparative Genomics Overview

- Comparison of whole genome sequences for determining how closely related organisms are to one another
- Genomes can be compared by the following features:
  - Genomic sequence
  - Strand asymmetry
  - Genes
  - Gene order
  - Genomic structural landmarks (functional annotations)
  - And more...!



Comparative Genomics Pipeline Summary

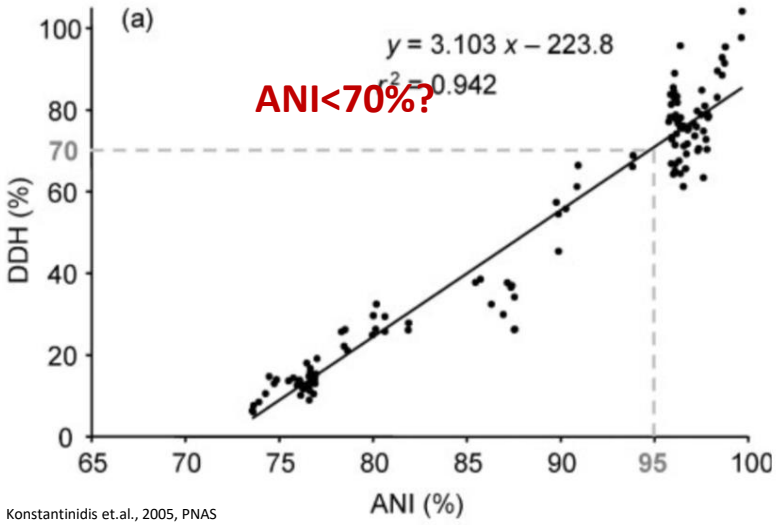


We will benchmark each software for each category prior to finalizing our selection

Types of comparative genomics techniques

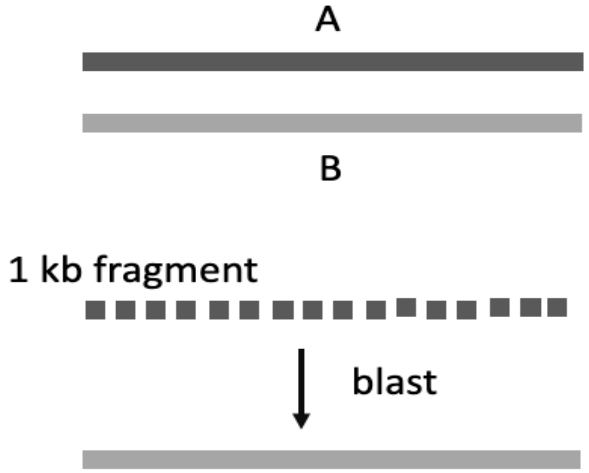
Average Nucleotide Identity (ANI)	Multi Locus Sequencing Typing (MLST)	Single Nucleotide Polymorphisms (SNP) Typing
Classification of bacterial species.	Estimates relationships between bacteria based on *allelic variations*	Compares base-by-base alignments to ascertain similarity

A value of 70 % DDH (DNA-DNA hybridization, 1 kb fragments of genome) was proposed as a recommended standard for delineating species (Wayne et.al.,1987)

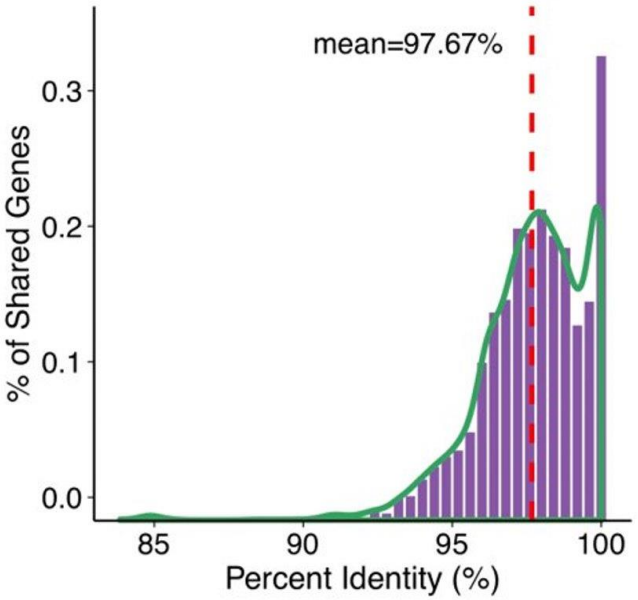


Konstantinidis et.al., 2005, PNAS

ANIb: blast based ANI



Two inportant factors affecting ANI: gene identity threshold, sequence alignment fraction



Average Nucleotide Identity (ANI)

Average Nucleotide Identity (ANI) is a measure of nucleotide-level genomic similarity between the coding regions of two genomes (A,B): define bacterial species?

ANI Tools

Alignment based

ANIb, ANIm (faster than ANIb)

OrthANIb, OrthANIm, OrthANIu

gANI, genome wide ANI (predicted gene based, no rRNA and tRNA, faster than ANIm)

Non-Alignment based

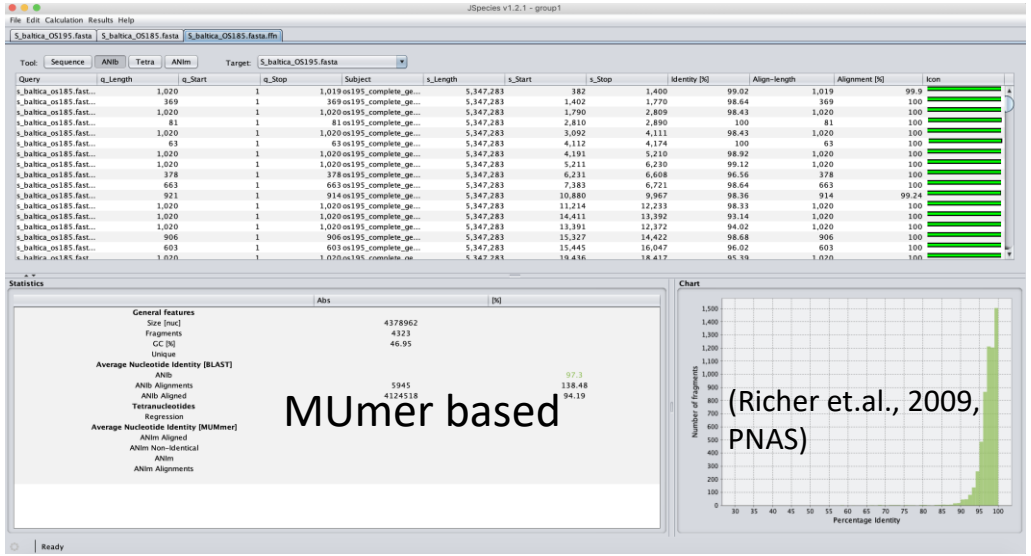
ARTICLE

DOI: 10.1038/s41467-018-07641-9 OPEN

High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries

Chirag Jain[1,2], Luis M. Rodriguez-R[3,4], Adam M. Phillippy[2], Konstantinos T. Konstantinidis[3,4] & Srinivas Aluru[1,5]

Jspecies (Java implementation of ANIb, ANIm)



MUmer based (Richer et.al., 2009, PNAS)

OrthANI (Java based)

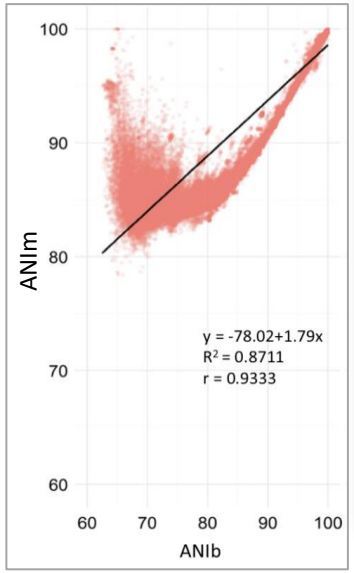
Program	Version	Parameters
USEARCH	8.1.1861_i86linux32	-usearch_local -id 0.5 -strand both -evalue 1.0E − 15 -maxaccepts 1 -xdrop_g 150 -mismatch -1 -match 1 -dbaccelpct 100 -qmask none -dbmask none
BLAST+	ncbi-blast-2.2.30+	blastn -evalue 1.0E−15 -dust no -xdrop_gap 150 -penalty -1 -reward 1
MUMmer	3.23	nucmer –mum -l 20 -b 200 -c 65 -g 90 –optimize -p

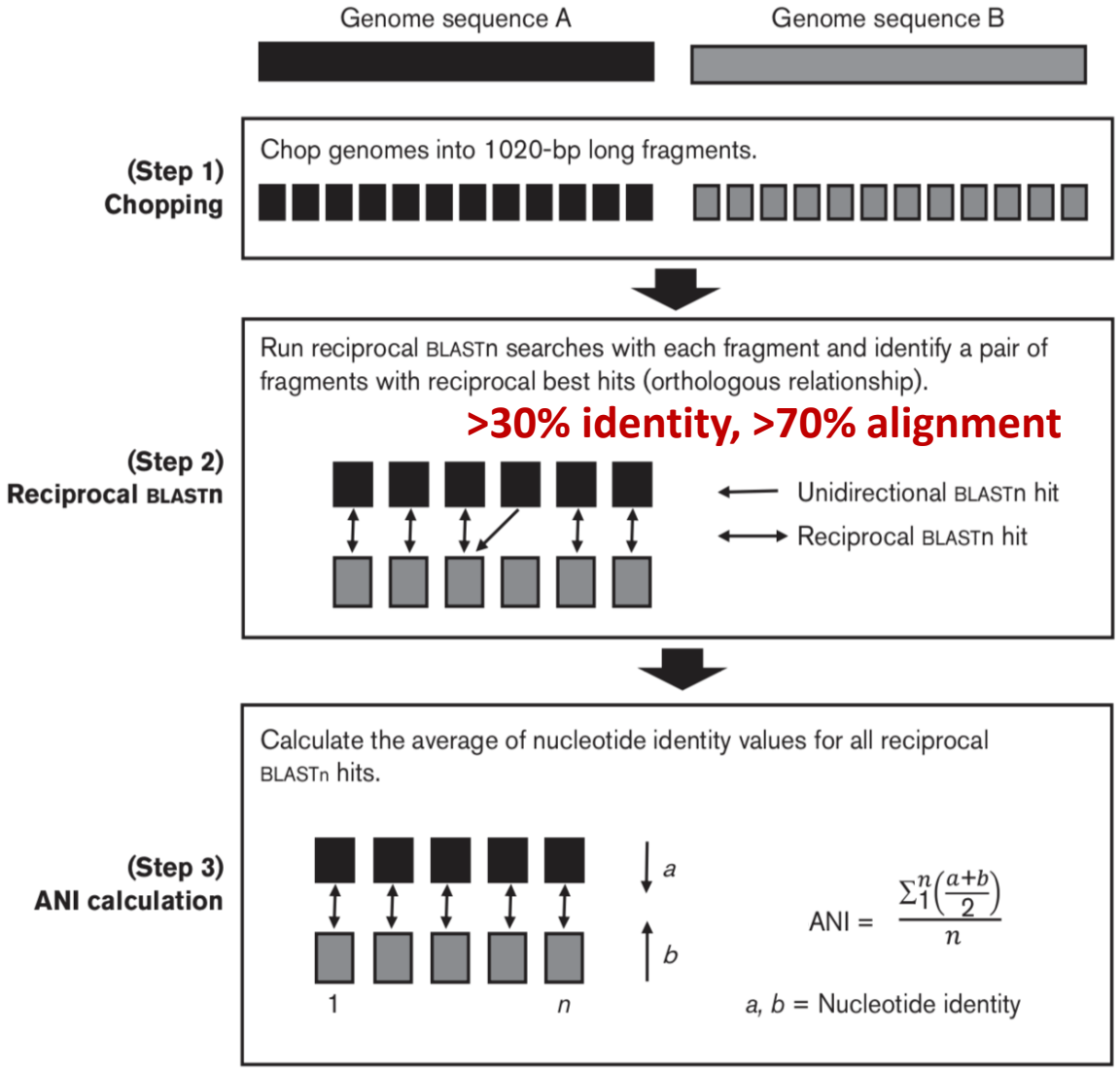
ANI_Calculator (gANI)

pyani (Jan, 2020), python implantation of ANIb, ANIm, OrthANIb, OrthANIm

OrthANIb, OrthANIm, OrthANIu

- **Blast-based**
  - based on a large number of genes
  - better measure of genomic relatedness than single gene, 16S rRNA gene
  - Not affected by varied evolutionary rates or HGT
  - Computationally intensive for large datasets
- **Usearch-based**
  - Usearch, a faster local alignement tool than blast for short sequences
- **MUMer-based**
  - MUMer uses an efficient data structure, suffix trees to calculate alignments.
  - These suffix trees can rapidly align sequences containing millions of nucleotides with precision.



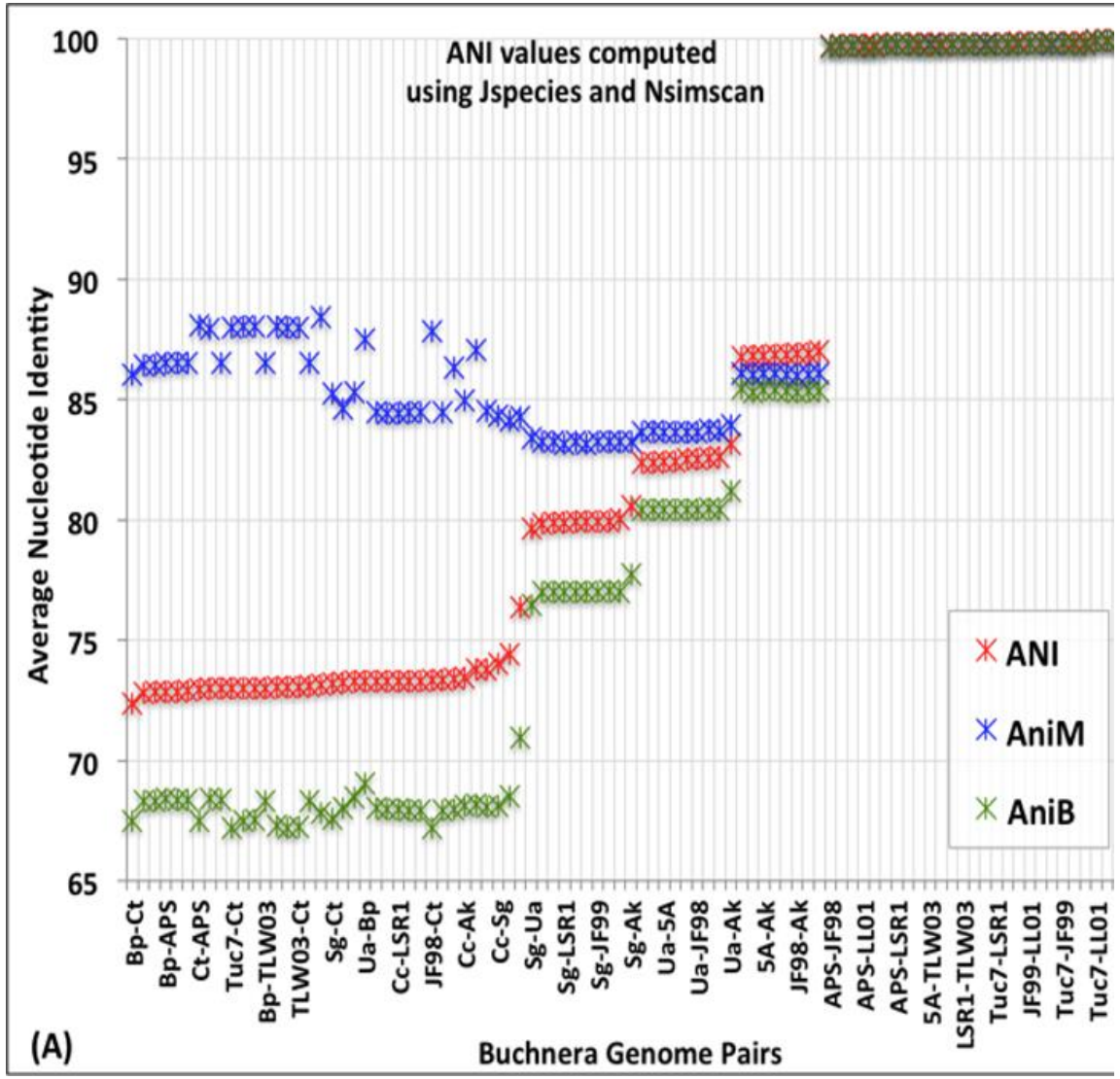


(Goris et.al., 2007; Lee et.al., 2016)

gANI (genome wide ANI)

- high performance similarity search tool NSimScan: protein-coding genes (A, B) were compared at the nucleotide level
- High speed: query aggregation, use of optimized bitwise operations in alignment computing, and by avoidance of dynamic programming
- Can be used for a large number of genome pairs

- **gANI (Varghese et.al., 2015, Nucleic Acid. Res)**



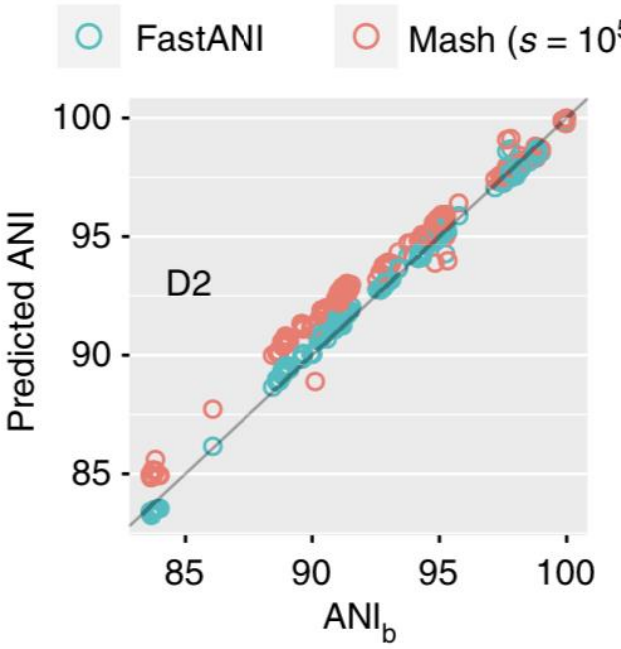
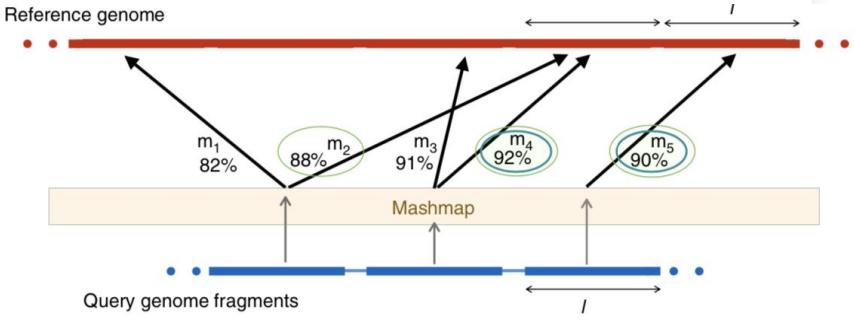


Table 3 Comparison of execution time of FastANI vs. ANI_b

Dataset	FastANI		ANI_b (s)	Speedup
	Indexing (s)	Compute (s)		
D1	468.2	16.76	13,113	782x
D2	195.7	264.8	18,155	69x
D3	1538	1981	99,317	50x
D4	128.8	214.5	11,051	52x
D5	2784	14.88	68,571	4608x

Speedup in the last column is measured as the ratio of ANI_b's runtime and FastANI's compute time



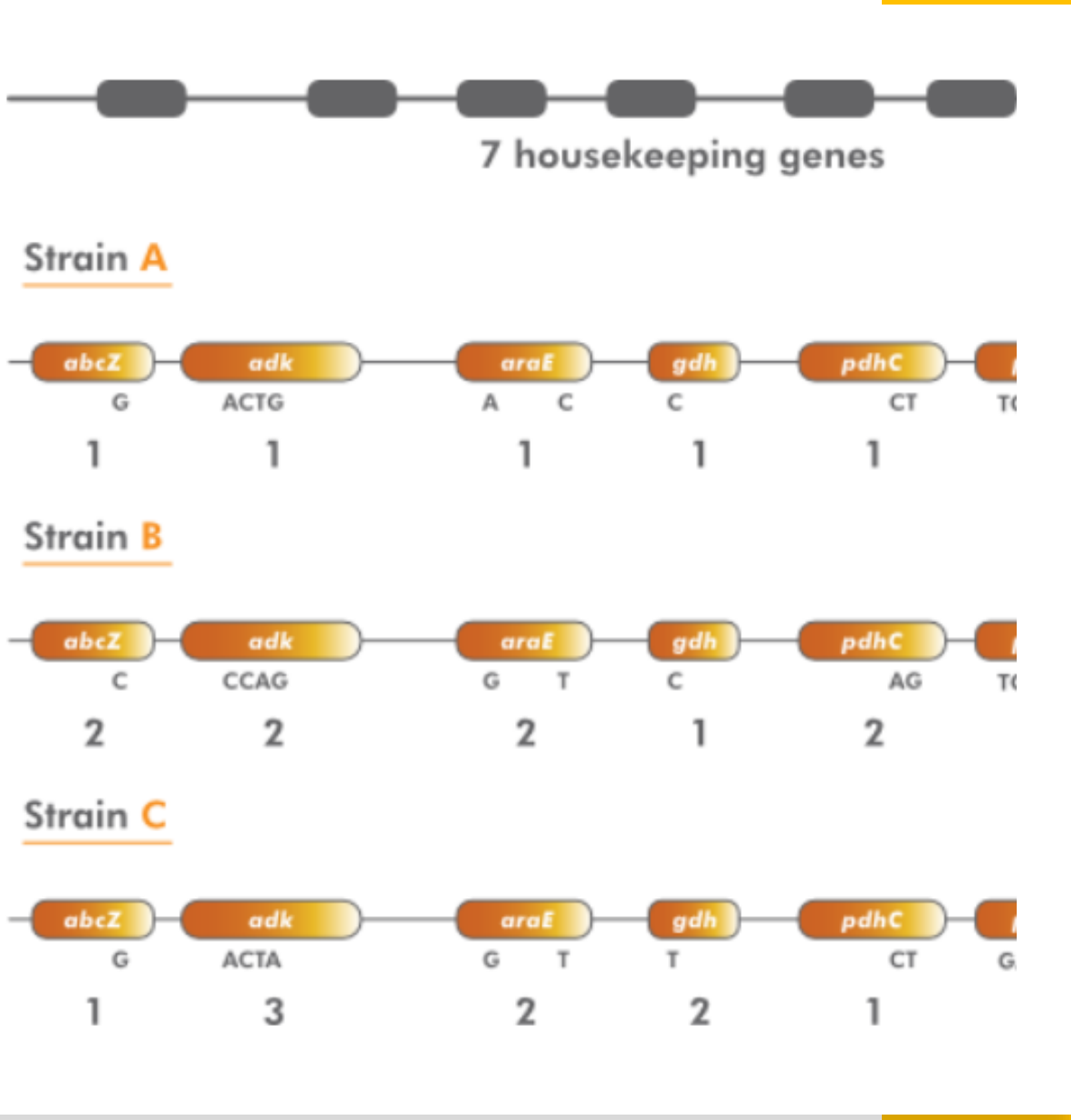
FastANI

(Jain et.al., 2018, NC)

- Mashmap: (A) fragments are mapped to the reference genome (B) using Mashmap. Mashmap first indexes the reference genome and subsequently computes mappings as well as alignment identity estimates for each query fragment, one at a time
- Reciprocal way, fastest and parallelized
- Only for identity around 80% or higher

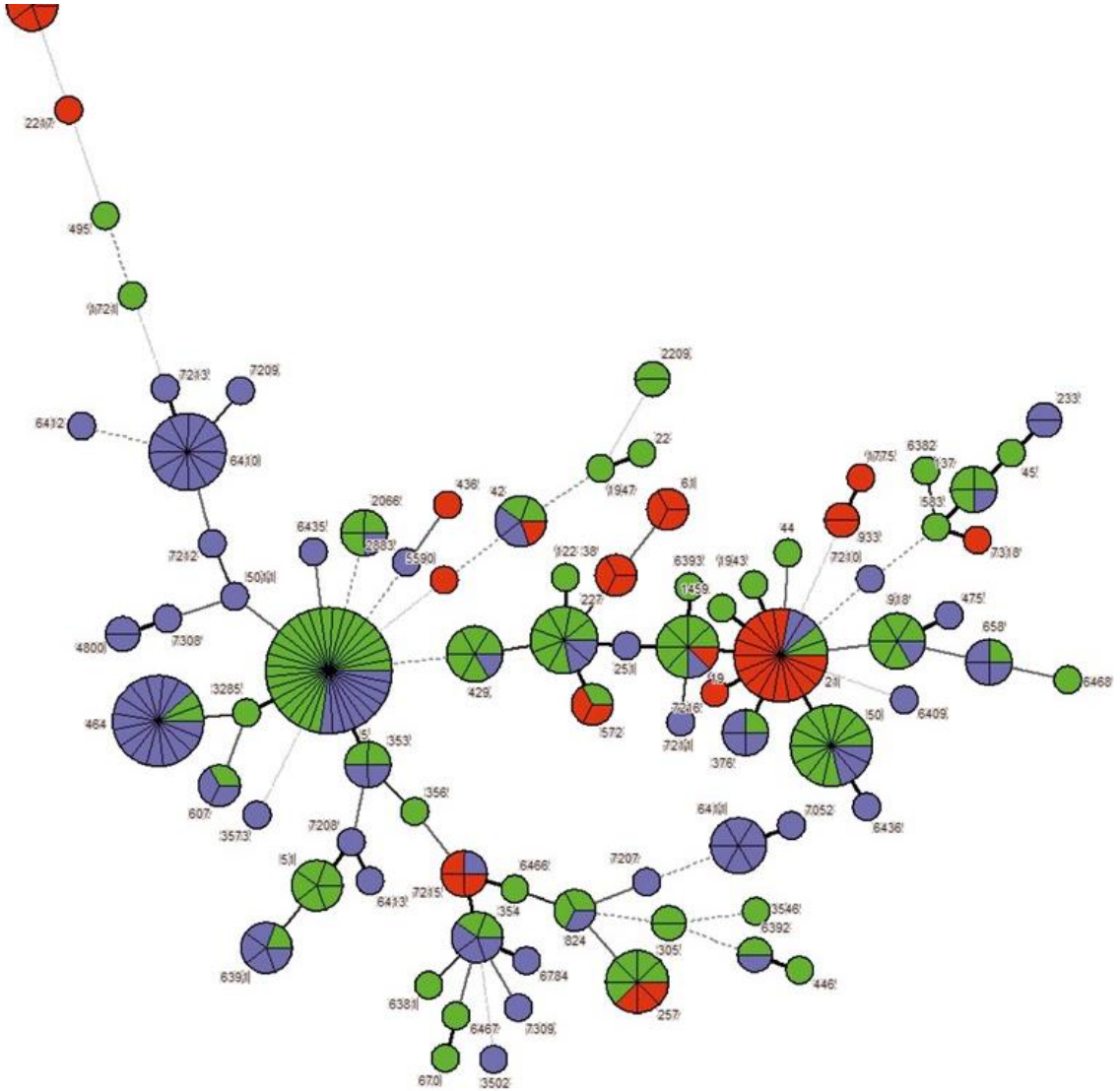
MLST: Multi-locus Sequence Typing

- Identify **a set of loci (genes)** in the genome and compare each locus in a genome against the set of loci
- Estimates relationships between bacteria based on *allelic variations*
- Profile of alleles ("sequence type" or ST) by calling the alleles
- MLST has been used successfully to study population genetics and reconstruct micro-evolution of epidemic bacteria and other micro-organisms.

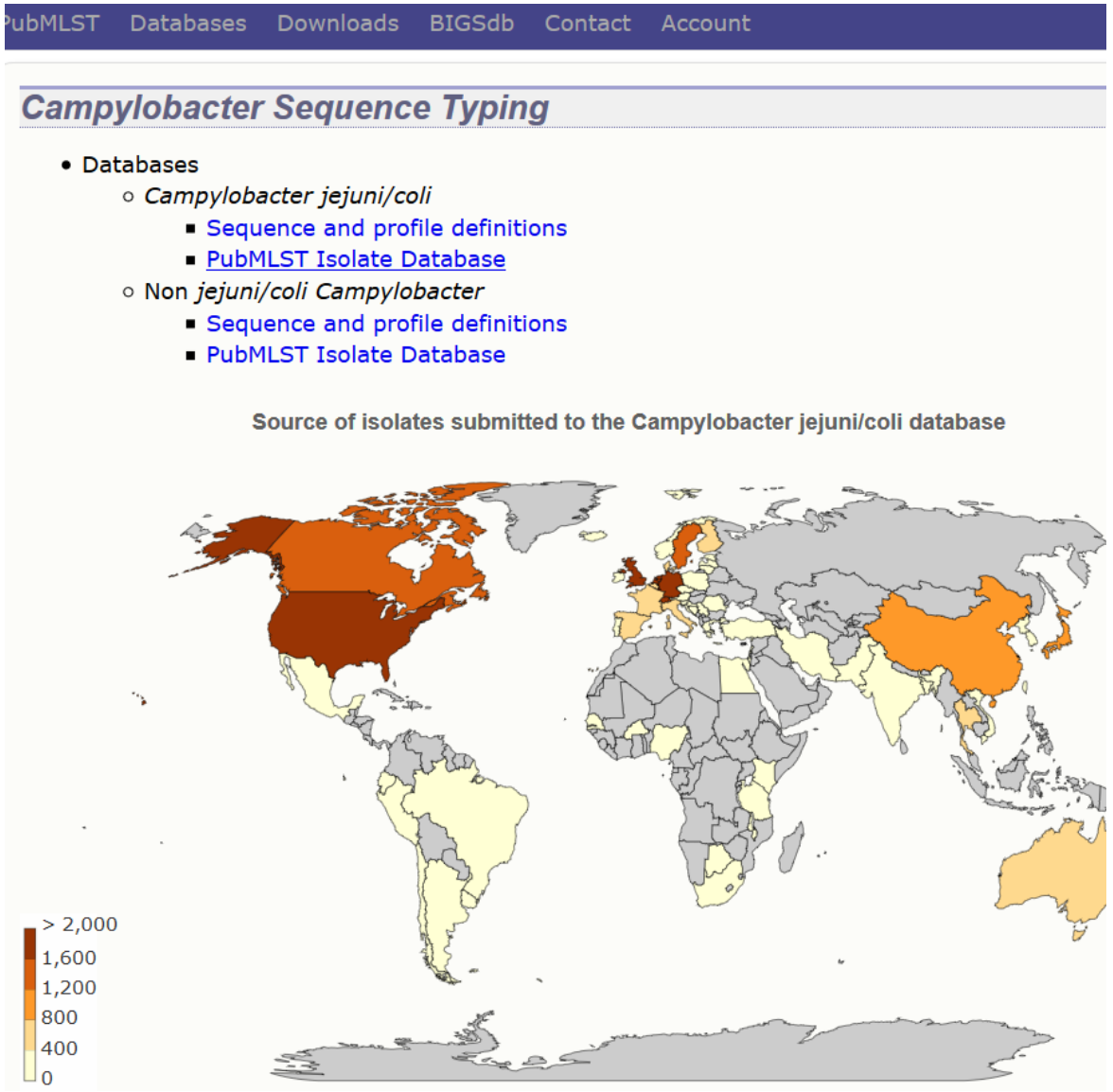


MLST: Multi-locus Sequence Typing

- **Whole-genome MLST (wgMLST)** – all the loci of a given isolate compared to equivalent loci in other isolates
- **Core-genome MLST (cgMLST)** – focused on only the core elements of the genomes of a group of bacteria
- **7-gene MLST** - choose 7 loci in the genome and compare all genomes to these 7 loci
- **Ribosomal MLST (rMLST)** – based on 53 loci that code for ribosomal proteins in most bacteria



Database: PubMLST for Campylobacter



MLST Tools Overview

Software	Input	Algorithm	Licence	Source	Tests	Installation	Interface
ARIBA	Reads	Assembly	GPL3	GitHub	Yes	Pip, Apt, Docker	Command line
BigsDB [11]	Contigs	BLASTN	GPL3	GitHub	No	Manual	Website
BioNumerics	Reads/ contigs	Proprietary/BLASTN	Bespoke	Proprietary	NA	Manual	GUI
EnteroBase	Reads	UBLAST/USEARCH	NA	NA	NA	NA	Website
MOST [14]	Reads	Mapping	FreeBSD	GitHub	No	Manual	Command line
mlst*	Contigs	BLASTN	GPL2	GitHub	No	Brew	Command line
MLST-CGE [16]	Contigs	BLASTN	Apache 2	Bitbucket	No	Docker	Command line/Website
MLSTcheck [17]	Contigs	BLASTN	GPL3	GitHub	Yes	CPAN, Docker	Command line
SeqSphere+ [18]	Contigs	NA	Bespoke	Proprietary	NA	Manual	GUI
SRST2 (24)	Reads	Mapping	BSD	GitHub	Yes	Apt, pip	Command line
stringMLST [21]	Reads	*k*-mer	Bespoke	GitHub	No	Manual	Command line

MLST tools comparison

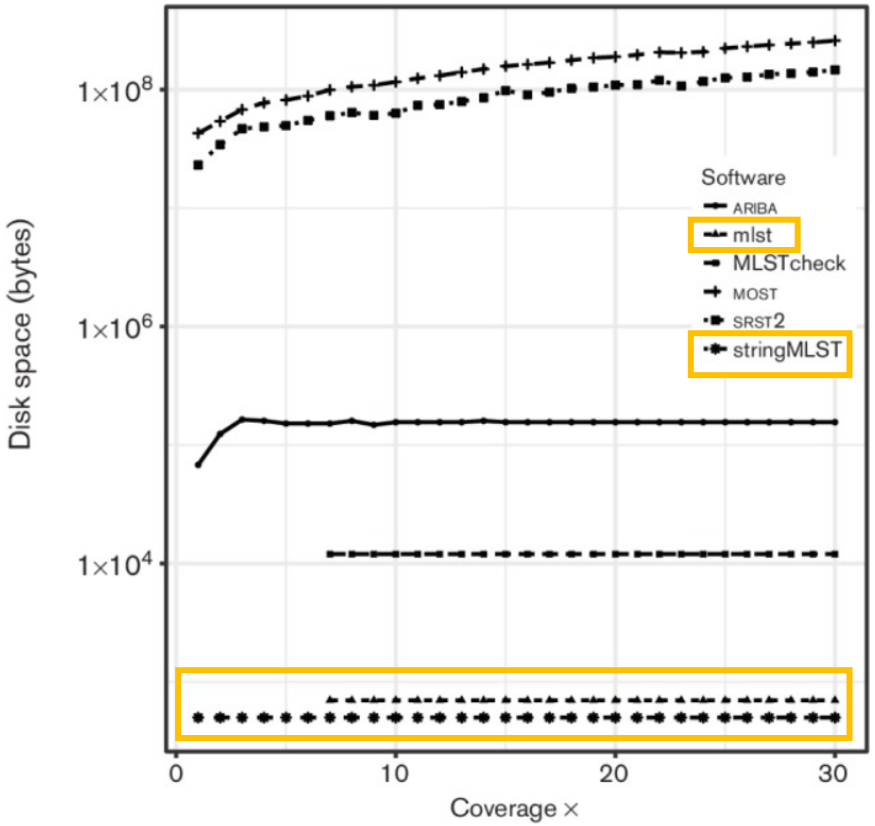


Figure: Disk space requirements in bytes for each software application as the depth of coverage increases. Due to the large difference between applications, a log scale is used.

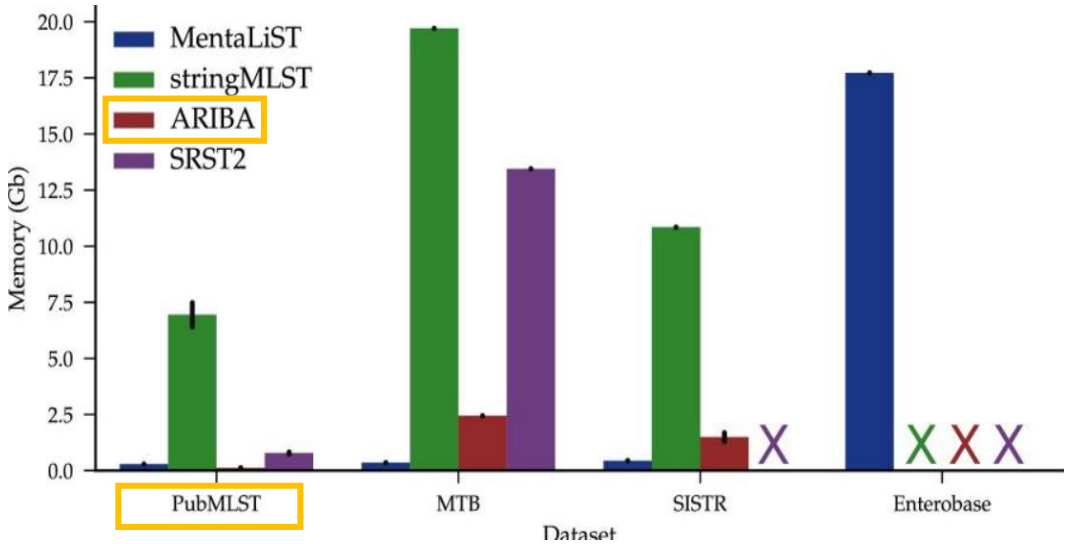


Figure: Peak memory usage for all MLST callers on the different schemes. X indicates that there are no results for the caller on the dataset, either because it failed or took more than 24 h. The bars represent the 95 % confidence interval.

MLST tools comparison

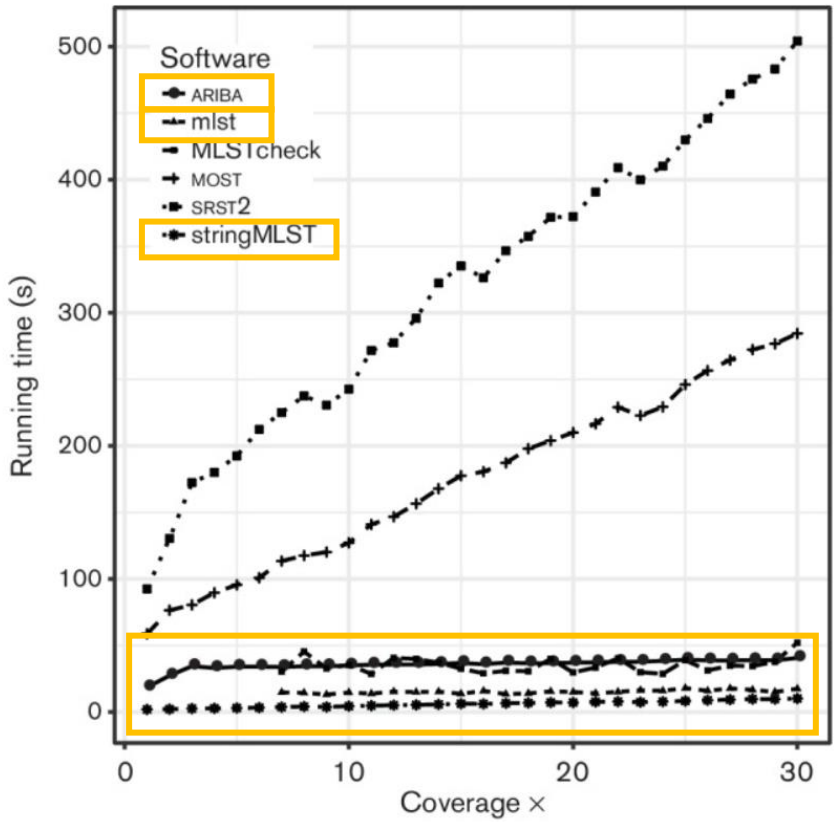


Figure: Running time (s) of each application as the coverage increases to assess the impact of the depth of coverage.

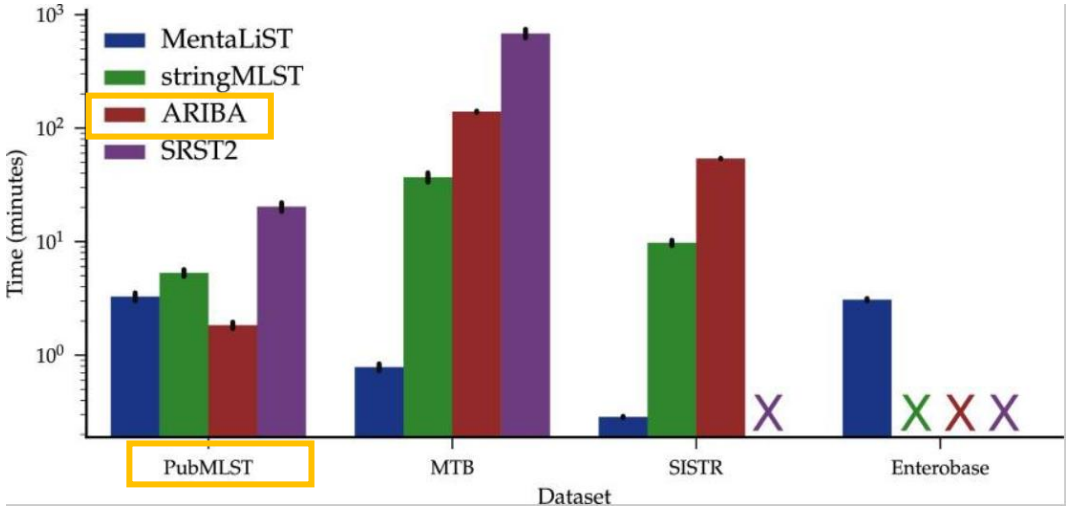


Figure: Running time for all MLST caller programs on the different schemes. X indicates that there are no results for the caller on the dataset, either because it failed or took more than 24 h. The bars represent the 95 % confidence interval.

MLST tools comparison

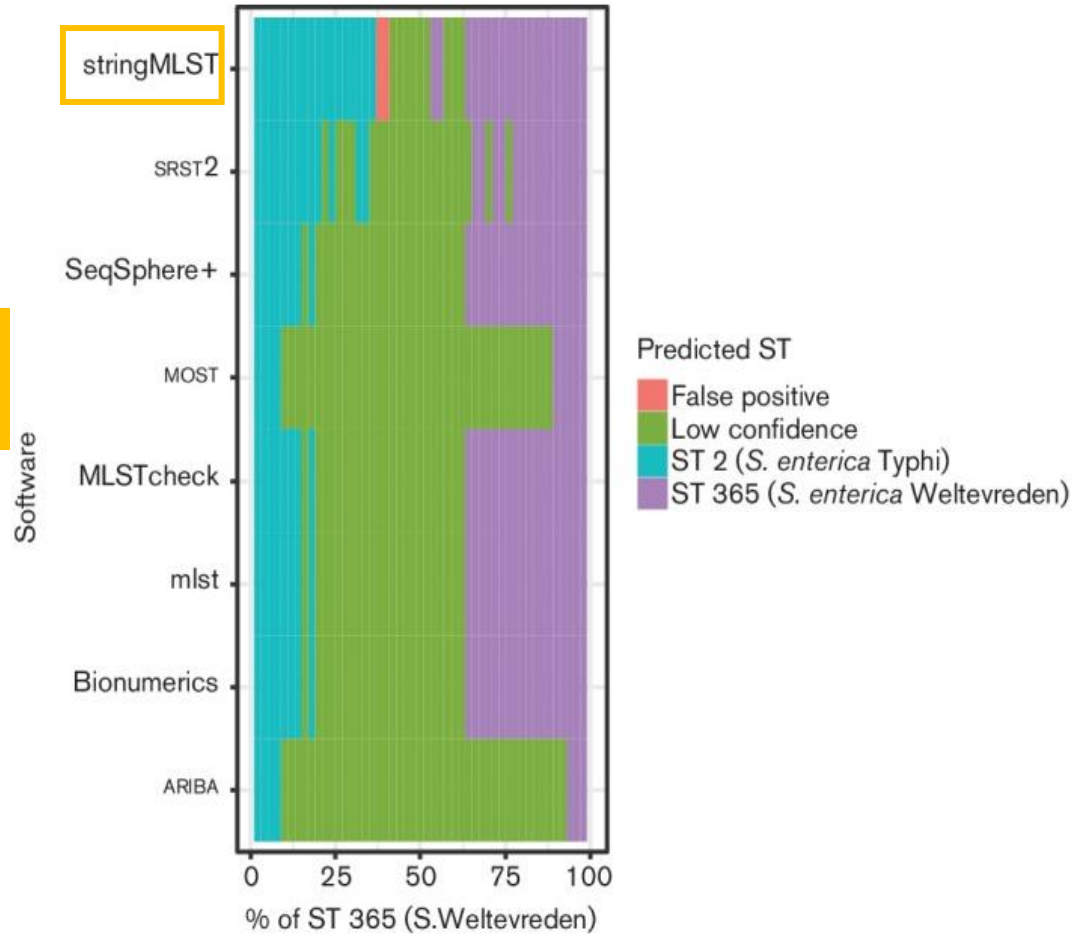


Figure: Tools were tested on simulated dataset consisting of two *Salmonella* samples with different alleles in varying ratios

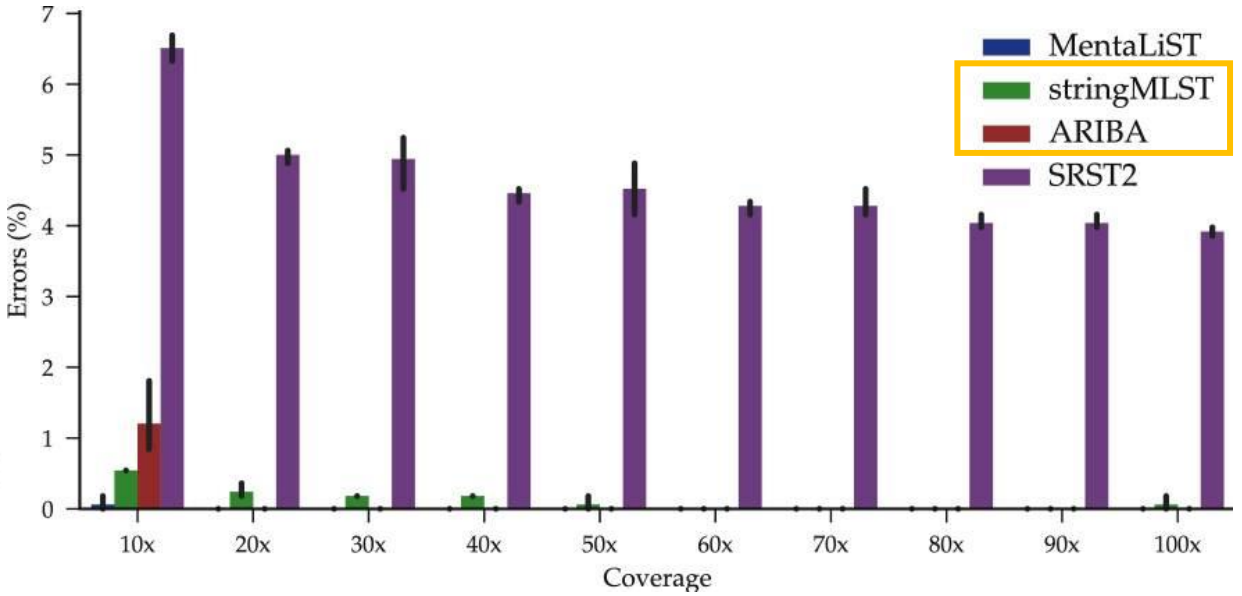


Figure: Average number of calling errors from three *M. tuberculosis* simulated samples, with varying depth of coverage and using the 553 gene ecgMLST scheme. The bars represent the 95 % confidence interval.

MLST tools comparison

Tool	Year of Publication	Citations	Algorithm	Basis
MLST	2012 (version 2.0 in 2018)	912	Assembly based	Stand-alone tool, takes in de novo assemblies, very fast and searches all databases on pubMLST
String MLST	2017	40	k-mer based	Stand-alone tool available, well documented, assembly and alignment free.
ARIBA	2017	154	Assembly based	Stand-alone tool available, well documented.

String MLST

- Tool for detecting the sequence type (ST) of a bacterial isolate directly from the genome sequence reads
- Developed by the Jordan Lab
- Assembly-free & alignment-free
- Faster algorithm compared to traditional MLST tools that maintains high accuracy
- Options to either build a database or use existing online database

Accuracy test (stringMLST; $k = 35$)

#Isolates[c]	#Alleles[d]	#Correctly predicted		Run time[b]	Mem[e]
		STs	Alleles		
1002	7014	1000	7012	40.7	0.67

Larger-scale schemes (stringMLST versus BLAST)

#Isolates[c]	#Alleles[d]	#Correctly predicted		RTR[f]	Sch[g]
		Alleles	%		
20	1060	1009	95.2	516.7	rMLST
20	31 919	28 976	90.8	43.0	cgMLST

a Algorithmic paradigm implemented by the tool.

b Average runtime per sample (in seconds).

c Total number of isolates tested.

d Total number alleles tested.

e Peak memory usage (in GB).

f Run time rate or the rate of processing sequence read files as kb/s.

g Typing scheme.

MLST

- MLST tool that scan contig files against traditional PubMLST typing schemes
- Takes *de novo* assemblies as input on the command line and uses BLASTN to align sequences to alleles.
- It is very fast and searches all databases on pubMLST to automatically detect the organism, then calculates the ST.
- Can build DB but also has bundle of all available databases in their software repository, which are regularly updated (every 1-2 months)
- Version 2.x does not just look for exact matches to full length alleles. It attempts to tell you as much as possible about what it found

Symbol	Meaning	Length	Identity
`n`	exact intact allele	100%	100%
`~n`	novel full length allele similar to n	100%	≥ `--minid`
`n?`	partial match to known allele	≥ `--mincov`	≥ `--minid`
`-`	allele missing	< `--mincov`	< `--minid`
`n,m`	multiple alleles		

- +90/N points for an exact allele match *e.g.* `42`
- +63/N points for a novel allele match (50% of an exact allele) *e.g.* `~42`
- +18/N points for a partial allele match (20% of an exact alelle) *e.g.* `42?`
- 0 points for a missing allele *e.g.* `-`
- +10 points if there is a matching ST type for the allele combination

ARIBA

- Assembly based tool
- Primarily developed for identifying Anti-Microbial Resistance - associated genes and single nucleotide polymorphisms directly from short reads
- It provides inbuilt support for and functionality for multi-locus sequence typing (MLST) using data from PubMLST.
- It provides inbuilt support for PlasmidFinder and VFDB (Virulence Factor Databases)
- Can be used in the study of Virulence Profile and AMR features along with the results from the Functional Annotation group

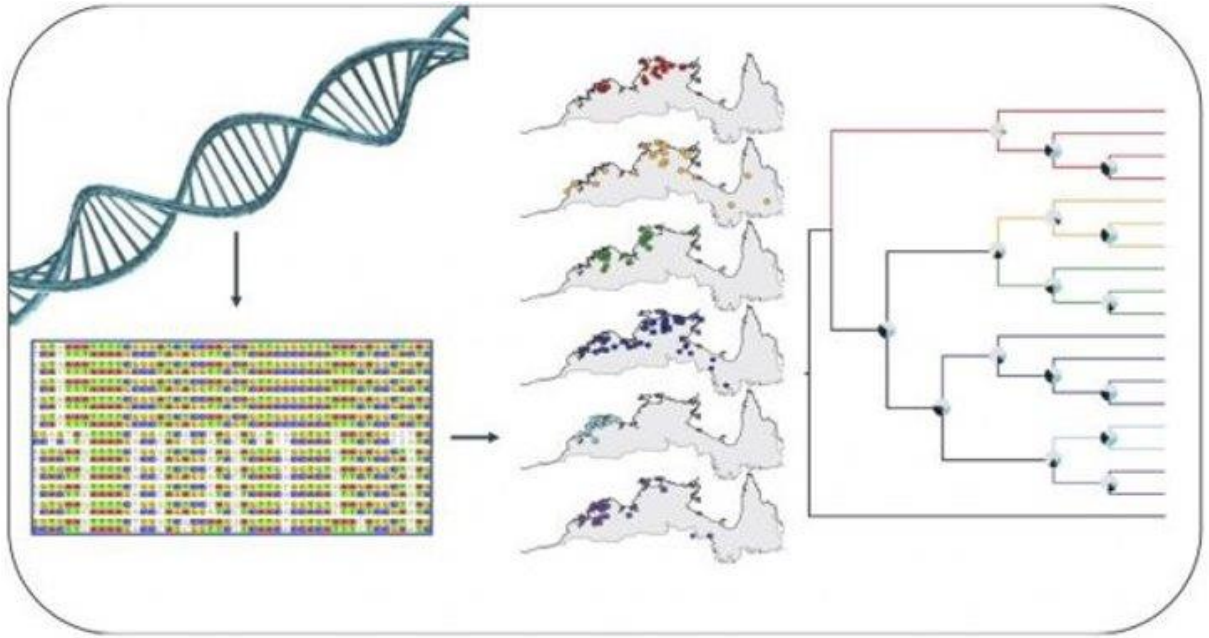
Single Nucleotide Polymporphisms (SNP) Typing

What are SNPs?

- A DNA sequence variation that occurs at a single position in the genome
- Prevalence of each variation > 1%
- Construction of phylogenetic trees based on SNPs for studying genetic and evolutionary factors in various organisms

Algorithm Overview:

- Pre-processing and read cleaning
- Mapping
- SNP calling against reference genome
- Phylogeny based on SNP profiles



SNP tools comparison

Tool	Year	Citations	Algorithm	Basis
kSNP 3.0	2013	214	k-mer based	Stand-alone tool available. Well documented. Multiple software versions created.
Lyve-SET	2017	54	MSA	Stand-alone tool available. Consistent performance. Higher specificity then kSNP.
SNPhylo	2014	186	MSA	Stand-alone tool available. Reduces SNP redundancy.
ParSNP	2014	570	MSA	Stand-alone tool available. Fast.
REALPHY	2014	222	Reference Sequence Alignment	Stand-alone tool available. Poor documentation.
SNVPhyl	2017	48	SNV Alignment	Stand-alone tool available. Can determine outbreak from non-outbreak.

SNP-based tools: kSNP3.1

- kSNP is optimal for situations where whole genome alignments don't work
- MSA-based approaches are computationally expensive and slow
- k-mer-based approaches are alignment-free and have a faster runtime
- Multiple kSNP versions have been created and thoroughly tested

Program	Conditions	Time (h)
kSNP v2	Default (no annotation)	1.04
kSNP3.0	Default (no annotation)	0.89
kSNP v2	Annotation	11.04
kSNP3.0	Standard annotation	2.92
kSNP3.0	Full annotation	11.14

SNP-based tools: Lyve-SET

- Linear regression model (*y =mx+b*) where *m = number of hqSNPs per Lyve-SET hqSNP* and *b = number of hqSNPs when there are no Lyve-SET hqSNPs*
- This represents all pairwise distances comparing Lyve-SET with other pipelines

L. monocytogenes

Pipeline	y=mx+b	R^2
kSNP	y=0.26x+24	0.69
RealPhy	y=1.14x+31	0.96
SNP-Pipeline	y=1.8x-13	0.97
SNVPhyl	y=0.27x+19	0.58

S. enterica

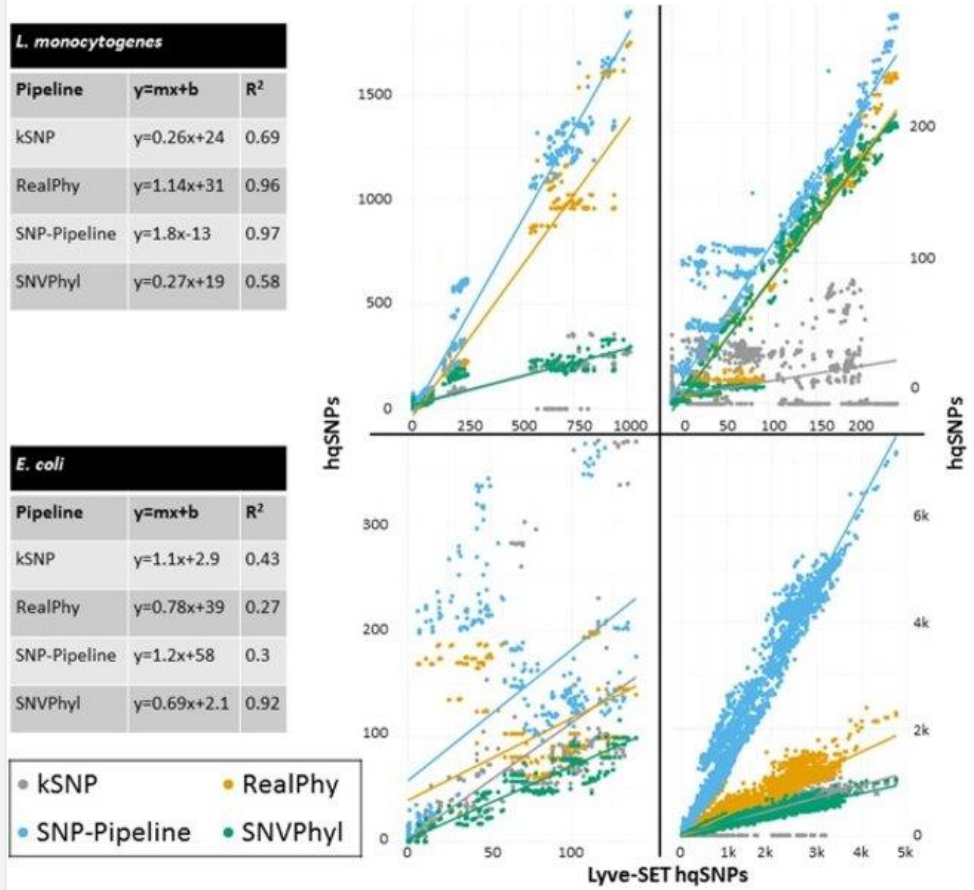
Pipeline	y=mx+b	R^2
kSNP	y=0.11x+4.7	0.23
RealPhy	y=0.92x-5	0.95
SNP-Pipeline	y=1.0x+5.4	0.96
SNVPhyl	y=0.91x-5.1	0.94

E. coli

Pipeline	y=mx+b	R^2
kSNP	y=1.1x+2.9	0.43
RealPhy	y=0.78x+39	0.27
SNP-Pipeline	y=1.2x+58	0.3
SNVPhyl	y=0.69x+2.1	0.92

C. jejuni

Pipeline	y=mx+b	R^2
kSNP	y=0.23x+4	0.89
RealPhy	y=0.4x-15	0.88
SNP-Pipeline	y=1.6x-17	0.97
SNVPhyl	y=0.18x+49	0.92



SNP-based tools: Lyve-SET

- MSA based approaches are computationally expensive!
  - Computationally complex
    - O(Length^{Nseqs})
  - Most use heuristic approaches rather than global optimization

Summary of 12 pipeline comparisons.

	Lyve-SET	kSNP	RealPhy	Snp-Pipeline	SNVPhyl	wgMLST
Tree sensitivity (Sn)[a]	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Tree specificity (Sp)[a]	100.0%	90.2%	100.0%	100.0%	100.0%	100.0%
Average of Sn and Sp	100.0%	95.1%	100.0%	100.0%	100.0%	100.0%
Kendall-Colijn ($\lambda = 0$)[b]	–	1.26E-02	7.51E-03	9.28E-03	9.15E-02	1.00E-04
Robinson-Foulds[b]	–	3.16E-69	6.79E-40	5.39E-74	9.61E-49	1.55E-147
Mantel	–	0.60	0.77	0.77	0.79	0.74
SNP ratio[c,d]	–	0.53, 0.78	0.97, 0.84	1.61, 1.75	0.67, 0.84	0.69, 0.72
Goodness-of-fit (R^2)[d]	–	0.46, 0.42	0.7, 0.75	0.77, 0.3	0.83, 0.68	0.75, 0.72
Genome analyzed[e]	25.9%	0.1%	84.8%	0.3%	82.1%	88.2%

SNP-based tools: ParSNP

- MSA based approach that is NOT computationally expensive
- Utilizes Maximal Unique Matches to cluster sample against reference
- Low FDR
- Output includes variant (SNP) calls, core genome phylogeny and multi-alignments
- Uses information provided by multi-alignments flanking SNP sites for QC

Table 1

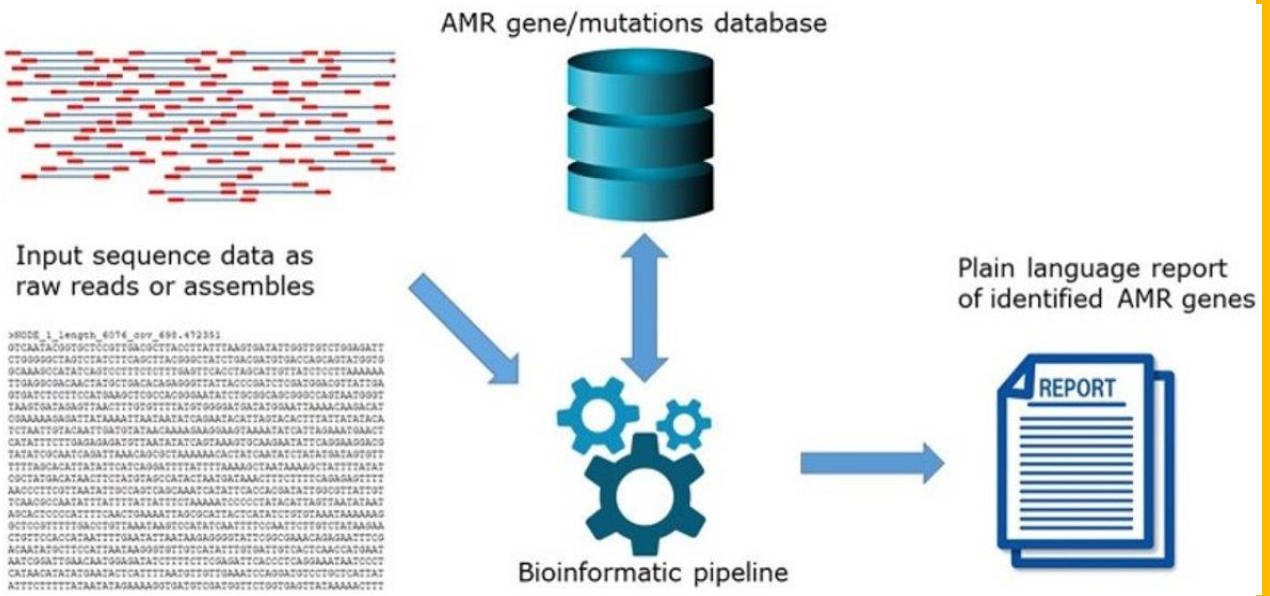
Core-genome SNP accuracy for simulated *E. coli* datasets

Method	Description [a]	FP Low	FN Low	FP Med	FN Med	FP High	FN High	TPR	FDR
Mauve	*WGA*	148	318	198	2,877	100	30,378	0.974	0.0004
Mauve (c)	*WGA*	0	0	2	38	6	649	0.999	0
Mugsy	*WGA*	1,261[b]	395	1,928	3,371	1,335	34,923	0.970	0.0036
Mugsy (c)	*WGA*	2	0	2	0	1	81	0.999	0
Parsnp	*CGA*	23	423	45	3,494	7	35,466	0.970	0.0001
Parsnp (c)	*CGA*	0	24	0	603	0	10,989	0.992	0
kSNP	*KMER*	259	600	908	19,730	1,968	916,127	0.280	0.0086
Smalt	*MAP*	33	110	0	1,307	55	22,957	0.981	0.0001
BWA	*MAP*	0	168	16	1,947	27	27,091	0.9775	0.0000

Data shown indicates performance metrics of the evaluated methods on the three simulated *E. coli* datasets (low, medium, and high). Method: Tool used.

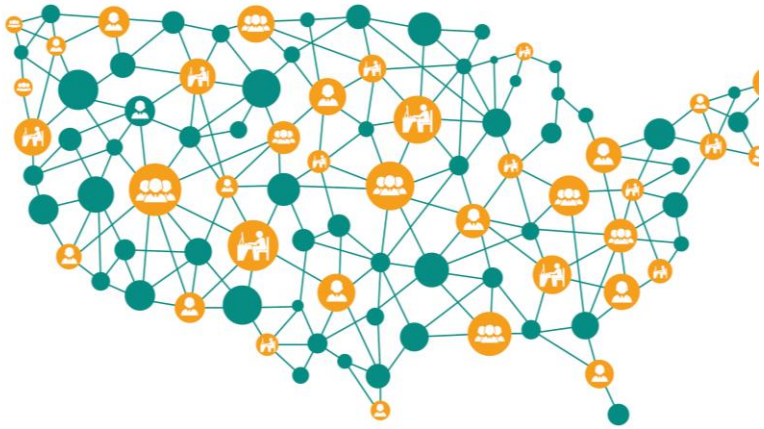
Virulence Profile & AMR Features

- Virulence Factors: Secreted by pathogen to colonize host at cellular level
- Antimicrobial Resistance (AMR) contributes to tens of thousands of deaths each year
- Can be derived from tools utilizing AMR Genes database including ARG-ANNOT, CARD, SRST2, MEGARes, Genefinder, ARIBA, KmerResistance, AMRFinder, and ResFinder
- Results from annotation group most helpful here



Deliverables: CDC Recommendations

- Preventative measures
  - Identify food source of outbreak strains to recommend recalls
  - Determine potential water source shutdown
  - Create PSAs to alert public of risks and hygienic prevention
- Outbreak response
  - Analyze date distribution / geographic outbreak plots
  - Refer related cases to physicians for treatment
  - Alert state labs of heightened related cases
  - Investigate supply chain correlations for specific product
- Treatment strategy
  - Recommend which antibiotics will be most effective and ineffective from AMR profile



Thank you!

Questions?

References

1. Touchman, J. (2010). "Comparative Genomics". *Nature Education Knowledge*. **3** (10): 13.
2. Xia, X. (2013). *Comparative Genomics*. SpringerBriefs in Genetics. Heidelberg: Springer. doi:10.1007/978-3-642-37146-2. ISBN 978-3-642-37145-5.
3. Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole- genome sequence similarities. International Journal of Systematic and Evolutionary Mi- crobiology, 57, 81–91
4. Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences of the United States of America, 102, 2567–2572.
5. Arahal, D.R. (2014). Whole-genome analyses: average nucleotide identity. In: *Methods in microbiology*. Elsevier, pp. 103-122.
6. Richter, M., & Rossello ́-Mo ́ra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences of the United States of America, 106, 19126–19131.
7. Varghese, N.J., Mukherjee, S., Konstantinidis, K.T. & Mavrommatis, K. (2015) Microbial species delineation using whole genome sequences. Nucleic Acid Research, 43, 6761–6771.
8. Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E. & other authors (1987). International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37, 463–464.
9. Jain, C., Dilthey, A., Koren, S., Aluru, S. & Phillippy, A. M. A fast approximate algorithm for mapping long reads to large reference databases. In International Conference on Research in Computational Molecular Biology (Springer, Hong Kong, 2017).
10. https://www.applied-maths.com/applications/mlst
11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5472909/
12. https://pubmlst.org/campylobacter/
13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5610716/
14. https://academic.oup.com/bioinformatics/article/33/1/119/2525695
15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5857373/
16. Lee, T., Guo, H., Wang, X. *et al.* SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 15, 162 (2014). https://doi.org/10.1186/1471-2164-15-162
17. Katz, Lee S et al. "A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens." *Frontiers in microbiology* vol. 8 375. 13 Mar. 2017, doi:10.3389/fmicb.2017.00375
18. Shea N Gardner, Tom Slezak, Barry G. Hall, kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome, *Bioinformatics*, Volume 31, Issue 17, 1 September 2015, Pages 2877–2878, https://doi.org/10.1093/bioinformatics/btv271
19. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6737581/#!po=11.3636
20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5243249/
21. https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/index.html
22. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4262987/