Comparative Genomics - Team 1 Background & Strategy

Heather Patrick, Lawrence McKinney, Laura Mora, Manasa Vegesna, Kenji Gerhardt, Hira Anis

April 7, 2020

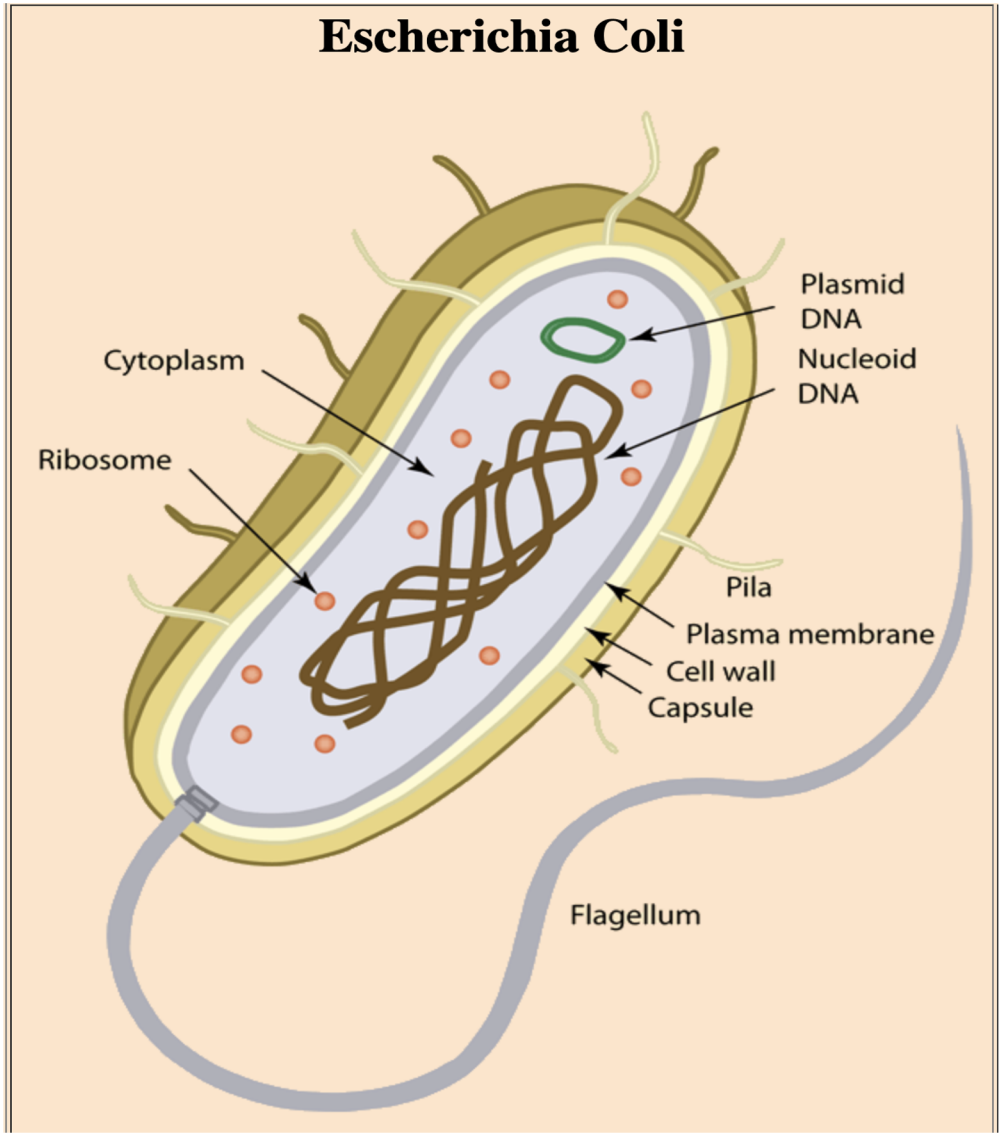
Outline

- Objectives
- Background
- Comparative Genomics Approaches
- Proposed pipeline
- References

Objectives

- Compare and contrast functional & structural features of isolates.
  - Antibiotic Resistance profile
  - Virulence profile
- Differentiate outbreak vs. sporadic strains.
- Characterize the virulence and antibiotic resistance functional features of outbreak isolates.
- Identify the source and spread of the outbreak.
- Recommend outbreak response and treatment.

Background



Enger, Eldon D. and Ross, Frederick C., Concepts in Biology, 10th Ed., McGraw-Hill, 2003.

- *Escherichia coli (E. coli)* is a gram-negative bacterium composed of numerous strains and serotypes.
- *E. coli* contains plasmids (mobile genetic elements) which generate genome diversity by promoting homologous recombination, horizontal gene transfer between bacteria, and can confer antimicrobial resistance and virulence.
- About ~46% of *E. coli* genome is conserved among all strains (core genome)

Bacterial Strain Typing

- Identifying bacteria at the strain level, is particularly important for diagnosis, treatment, and epidemiological surveillance of bacterial infections.
- Bacterial epidemiological typing generates isolate-specific genotypic or phenotypic characters that can be used to elucidate the sources and routes of spread of bacteria.
- Especially important for bacteria exhibiting high levels of antibiotic resistance or virulence.
- Strain typing also has applications in studying bacterial population dynamics.

Comparative Genomics Approaches

MUMmer v.04

- A bioinformatic tool used align and compare entire genomes at varying evolutionary distances.
- It uses "**M**aximal **U**nique **M**atches" as pairwise anchor points to help improve the biological quality of the output alignments.
- Pros:
  - Fast and efficient aligner
  - Optimal for comparing two related bacterial strains
  - Highly cited bioinformatics system in scientific literature (> 900 total citations; + 200 since 2018)
- Cons:
  - Higher false alignment rate (FAR) when compared to similar tools.

Question answered using this tool:

"Which genes do these genomes (reference and query) share and which genes are unique to particular genomes?"

This helps to identify unique regions of the genome that have been documented as being virulent or sporadic in *E. coli.*

MUMmer v.04

Table 5. Performance of Nucmer4, BLASR and BWA MEM on data simulated by pbsim from human and Arabidopsis reference genomes. All numbers are percentages from the total of bases that are in the reads aligned correctly, missed, or aligned incorrectly. The numbers may not add to exactly 100 due to rounding.

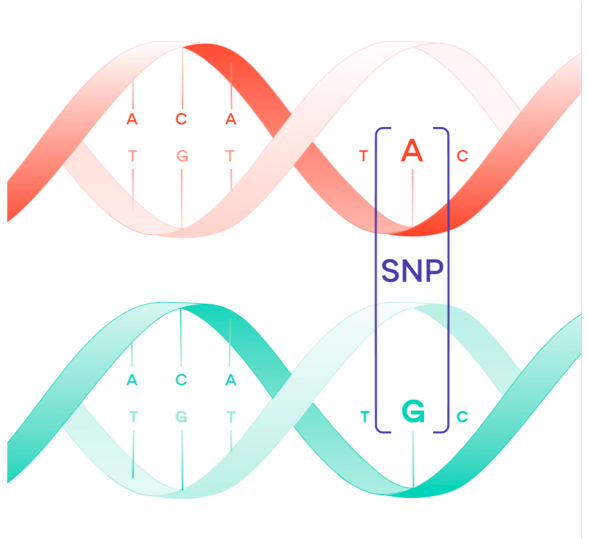
	Arabidopsis			Human		
	Aligned Correctly	Missed	Aligned Incorrectly	Aligned Correctly	Missed	Aligned Incorrectly
nucmer4	94.0	3.5	2.5	84.4	10.9	4.6
blasr	98.2	0.2	1.7	91.8	5.0	3.2
bwa-mem	98.7	0.5	0.8	91.6	5.9	2.5

https://doi.org/10.1371/journal.pcbi.1005944.t005

- MUMmer's sequence aligner feature called "nucmer4" was found to be less sensitive when reads were aligned with BLASTR, nucmer4 and BWA to the corresponding reference genomes.
- Nucmer4 also has marginally higher FAR.
- The sensitivity numbers are consistent with the results on real data.
- MUMmer v4 has a feature (`--maxmatch`) that will account for this error at the expense of run time.

SNP Analysis

- **Single Nucleotide Polymorphisms** are mutations with a single DNA base substitution. When found in exonic regions, they can result in amino acid variants in the protein products or changes in protein length due to their effects on stop codons.



- Identification of SNPs across bacterial genomes is important for **outbreak tracking, phylogenetic analysis and identifying strain differences** that are important to phenotypes such as **virulence and antibiotic resistance**.

- **Main Objective:** Identify SNPs and produce a phylogenetic tree which will help us identify the source and strain of the organism causing the outbreak.

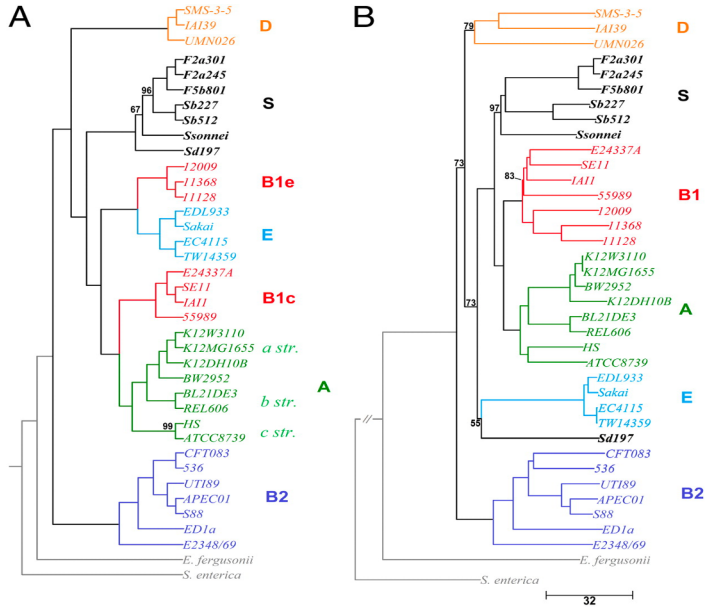


Figure: Whole-genome phylogenies of E. coli/Shigella (Sims *et al.,* 2011)

SNP Analysis Tool Search

Tool Name	Year	Based On	Advantages	Disadvantages
kSNP v. 3.0	2015	K-mer Analysis	Faster than multiple-alignment and reference-based methods. Has been tested on 68 genomes of E.coli	Cannot identify SNPs which are close to each other
BactSNP	2019	De-novo Assembly and Alignment Information	Can be run without a reference genome and has been benchmarked against other tools/pipelines for bacterial genomes	Doesn't produce phylogenetic trees
ParSNP	2014	Multiple genome alignment	Designed for microbial genomes. Avoids biases from mapping to a single reference	Cannot handle subset data, only works well for core genomes Not as sensitive as the other tools. Should be used in combination with a visualizer
RealPhy	2014	Multiple reference sequence alignment	Avoids biases which come from using one reference genome	Requires a reference genome

kSNP3

- Identifies all pan-genome SNPs in a set of given genome sequences and estimates phylogenetic trees based upon the identified SNPs.
- SNP identification is based on **k-mer analysis**
- kSNP builds **Maximum Likelihood, Neighbor Joining and Parsimony Phylogenetic trees**
- Doesn't require a multiple sequence alignment or the selection of a reference genome
- SNPs are annotated from GenBank files.

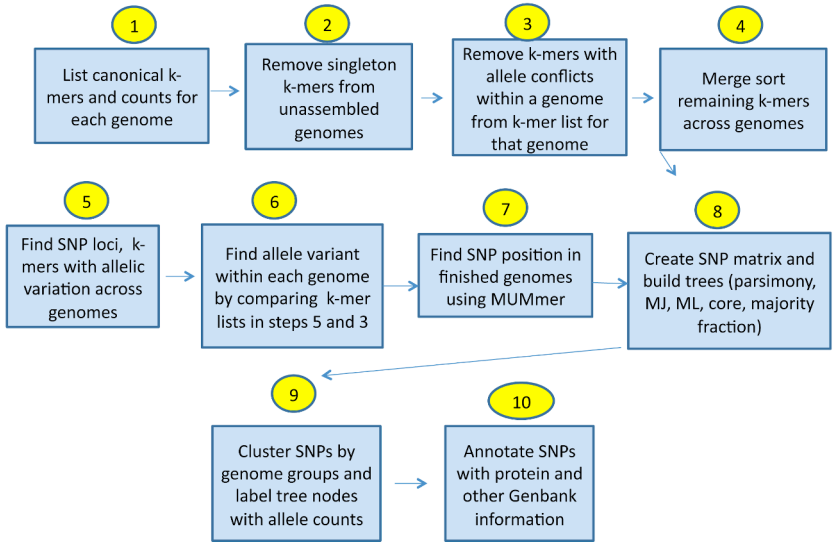
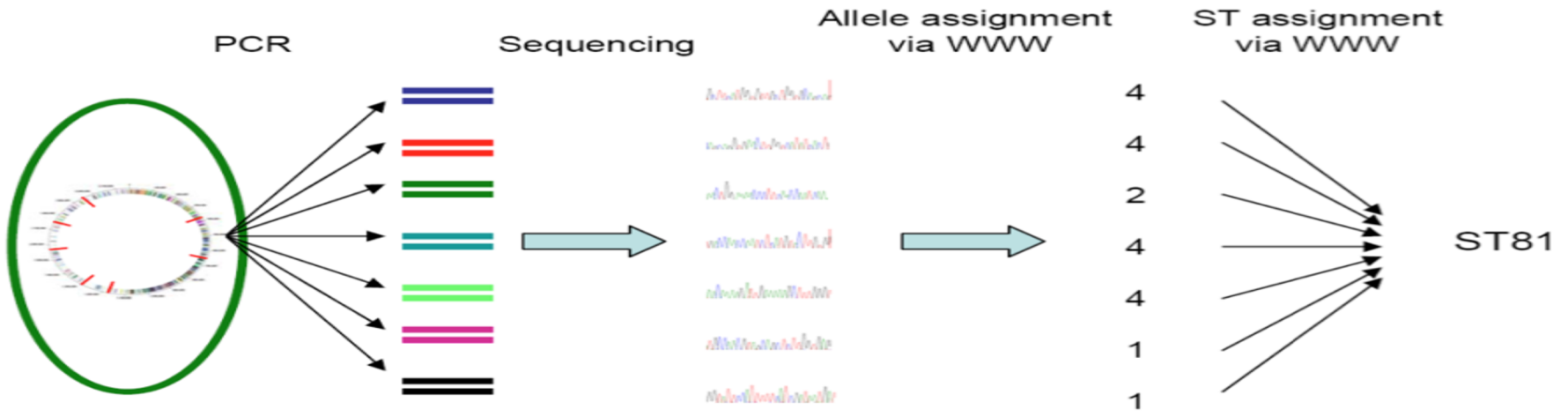


Figure: Diagram of the kSNP process. (Gardner *et al.*, 2013)

PROS	CONS
● Has been tested on 68 finished E.coli genomes ● Can efficiently analyze distantly-related genomes ● avoids biases stemming from the choice of a reference genome ● finds SNPs which are present in core and non-core regions	● Cannot find SNPs that are too close to each other ● Using a bigger k-mer size will compromise the identification of high density SNPs ● A smaller k-mer size could cause an increase in allele conflicts ● When using raw reads, the tool sometimes cannot distinguish between true SNPs from sequencing errors

MLST: Multi Locus Sequence Typing

- A low-resolution classification to categorize different clonal expressions of pathogens into broad categories.
- The concept is based on allelic variation amongst highly conserved housekeeping genes (the schemes)
- The nomenclature is still widely used by clinicians and microbiologists
- There are bioinformatics tools that use raw sequence reads and others than use de novo assemblies.
- Three schemes available for *Escherichia coli* : **Achtman,Pasteur, Whittam schemes (7:8:15)**
- PubMLST ONLY USES Achtman and Pasteur



MLST : Tool Comparison

MICROBIAL GENOMICS

REVIEW

Page *et al.*, *Microbial Genomics* 2017;3

DOI 10.1099/mgen.0.000124

Comparison of classical multi-locus sequence typing software for next-generation sequencing data

Andrew J. Page,[1,*] Nabil-Fareed Alikhan,[2] Heather A. Carleton,[3] Torsten Seemann,[4] Jacqueline A. Keane[5] and Lee S. Katz[3,6]

Abstract

Multi-locus sequence typing (MLST) is a widely used method for categorizing bacteria. Increasingly, MLST is being performed using next-generation sequencing (NGS) data by reference laboratories and for clinical diagnostics. Many software applications have been developed to calculate sequence types from NGS data; however, there has been no comprehensive review to date on these methods. We have compared eight of these applications against real and simulated data, and present results on: (1) the accuracy of each method against traditional typing methods, (2) the performance on real outbreak datasets, (3) the impact of contamination and varying depth of coverage, and (4) the computational resource requirements.

Tool comparison based on:

- Database availability and updates
- Disk Space
- Time
- Coverage/Quality of Query Sequence
- Software/Dependency Management and Installation
- Efficiency in mixed samples (Doesn't apply in our case since we know we have isolates)

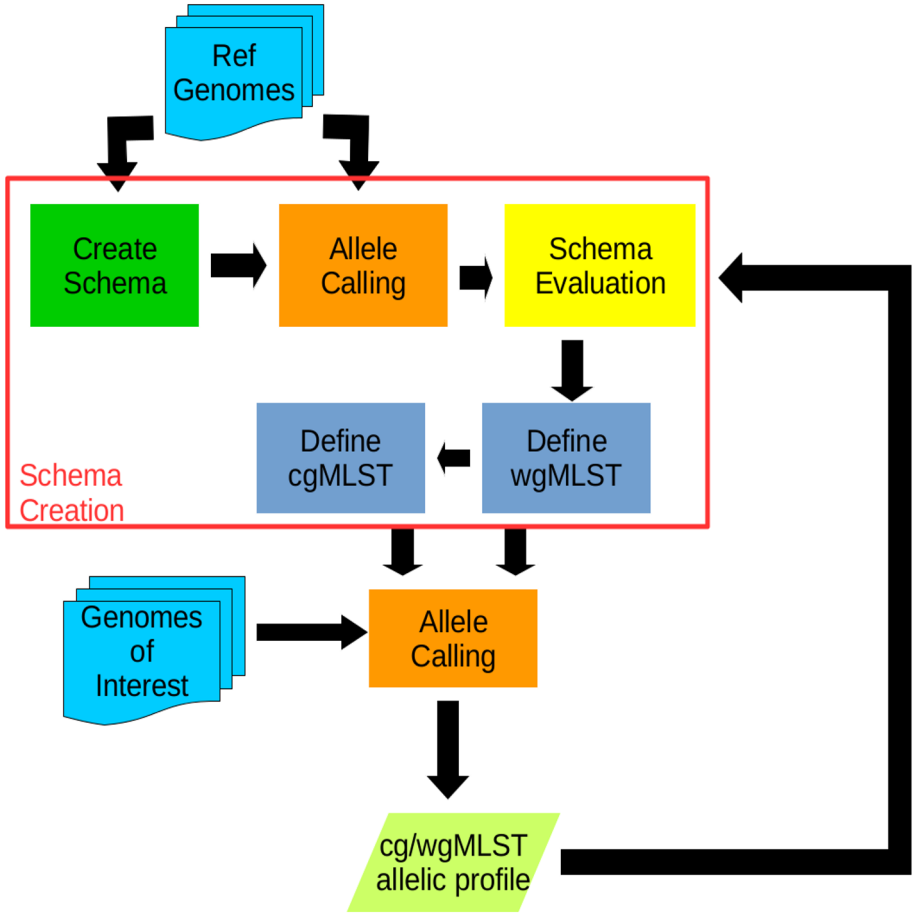
TOOLS TO COMPARE

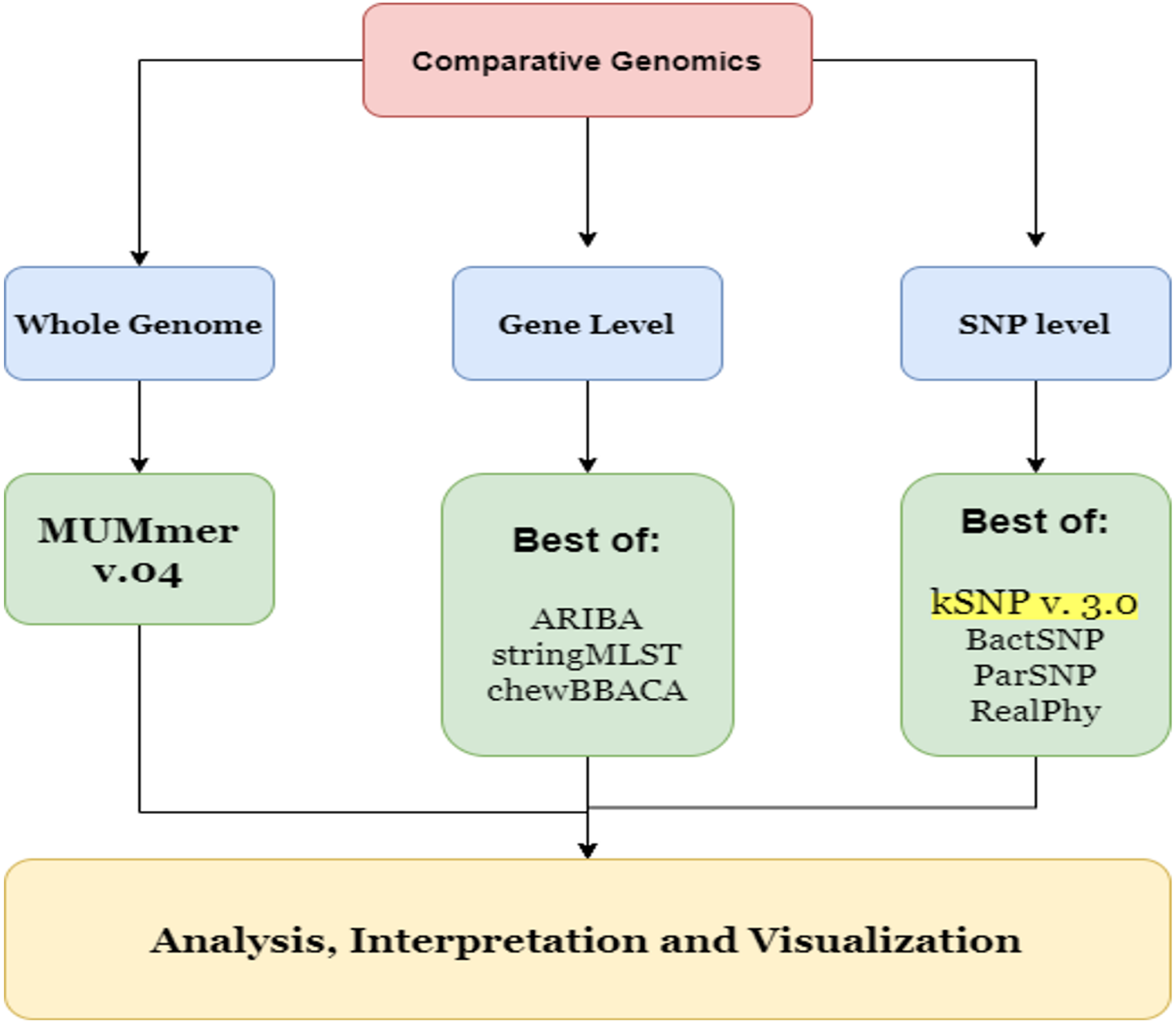
1. ARIBA
2. BigsDB
3. BioNumerics
4. EnteroBase
5. MOST
6. mlst
7. MLST-CGE
8. MLST-check
9. SeqSphere
10. SRST2
11. stringMLST
12. MentaLiST
13. chewBBACA

chewBBACA

A comprehensive pipeline for the creation and validation of whole genome and core genome MLST schemas

- Schema creation and allele calls are done on complete or draft genomes resulting from de novo assemblers
  - The allele calling algorithm is based on BLAST Score Ratio that can be run in multiprocessor settings
- Performs allele calling in a matter of seconds per strain
- Visualizes and evaluates allele variation in the loci





The Proposed Preliminary Pipeline

References

1. *Chen X, Zhang Y, Zhang Z, Zhao Y, Sun C, Yang M, Wang J, Liu Q, Zhang B, Chen M, Yu J, Wu J, Jin Z and Xiao J (2018) PGAweb: A Web Server for Bacterial Pan-Genome Analysis. Front. Microbiol. 9:1910. doi: 10.3389/fmicb.2018.01910*
2. Maiden MC, Jansen van Rensburg MJ, Bray JE, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol*. 2013;11(10):728-36.
3. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, et al. (2018) MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology 14(1): e1005944. https://doi.org/10.1371/journal.pcbi.1005944
4. Perez-Losada M, Arenas M, Castro-Nallar E. Microbial sequence typing in the genomic era. Infection, Genetics and Evolution. 2018;63:346-359. http://dx.doi.org/10.1016/j.meegid.2017.09.022
5. Strockbine N, Bopp C, Fields P, Kaper J, Nataro J. 2015. *Escherichia*, *Shigella*, **and** *Salmonella*, p 685-713. *In* Jorgensen J, Pfaller M, Carroll K, Funke G, Landry M, Richter S, Warnock D (ed), *Manual of Clinical Microbiology, Eleventh Edition*. ASM Press, Washington, DC. doi: 10.1128/9781555817381.ch37
6. Sultan, I., Rahman, S., Jan, A. T., Siddiqui, M. T., Mondal, A. H., & Haq, Q. M. R. (2018). Antibiotics, Resistome and Resistance Mechanisms: A Bacterial Perspective. *Frontiers in Microbiology, 9*(2066). doi:10.3389/fmicb.2018.02066
7. Trees E**,** Rota P, Maccannell D, Gerner-smidt P**.**. Molecular Epidemiology, p 131-159. *In* Jorgensen J, Pfaller M, Carroll K, Funke G, Landry M, Richter S, Warnock D (ed), *Manual of Clinical Microbiology, Eleventh Edition*. ASM Press, Washington, DC. 2015. doi: 10.1128/9781555817381.ch10
8. Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics, 31(17), 2877-2878.

Supplementary

Table 1. Overview of MLST software

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

*https://github.com/tseemann/mlst

Table 2. Overview of the MLST databases available with each software application.

Software	Automated download	Bundled DBs	Age of bundled DBs*	DBs ready to use
ARIBA	Yes	0	–	Yes
BioNumerics	Yes	0	–	Yes
mlst	Yes	125	1 month	Yes
MLSTcheck	Yes	0	–	Yes
MOST	No	6	>1 year	Yes
SeqSphere+	Yes	0	–	Yes
SRST2	Yes	0	–	Yes
stringMLST	Yes	128	1 month	Yes

DB, Database.

*The age of the bundled databases was calculated on the 15 March 2017.

Table 3. Summary of performance of each algorithm on real outbreak data for four different species (85 samples)

Software	Total time (min)	Correct ST (%)	No call/low confidence (%)
ARIBA	109.5	98.8	1.2
BioNumerics	NA	**100**	**0**
mlst*	1.9 (+2873)	96.5	3.5
MOST†	1189.7	49.4	50.6
MLSTcheck*	63.8 (+2873)	**100**	**0**
SeqSphere+	NA	96.5	3.5
SRST2	2380.2	95.3	4.7
stringMLST	**80.8**	**100**	**0**

Values in bold indicate the best results in each column.

*The time to assemble with SPAdes before running the applications was 2873 min and is included separately.

†MOST identified the correct ST in 97.6 % of cases, but flagged 48.2 % of these calls as low confidence.

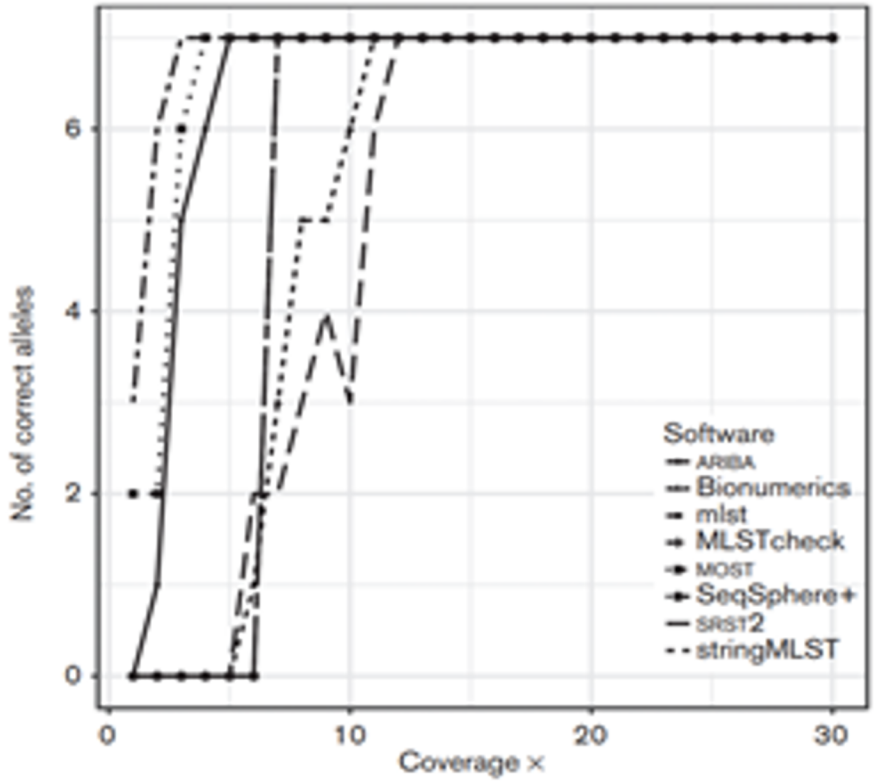


Fig. 1. Number of correct calls of each application as coverage increases. Each ST consists of seven alleles, and all seven must be correctly and confidently called to calculate a ST.

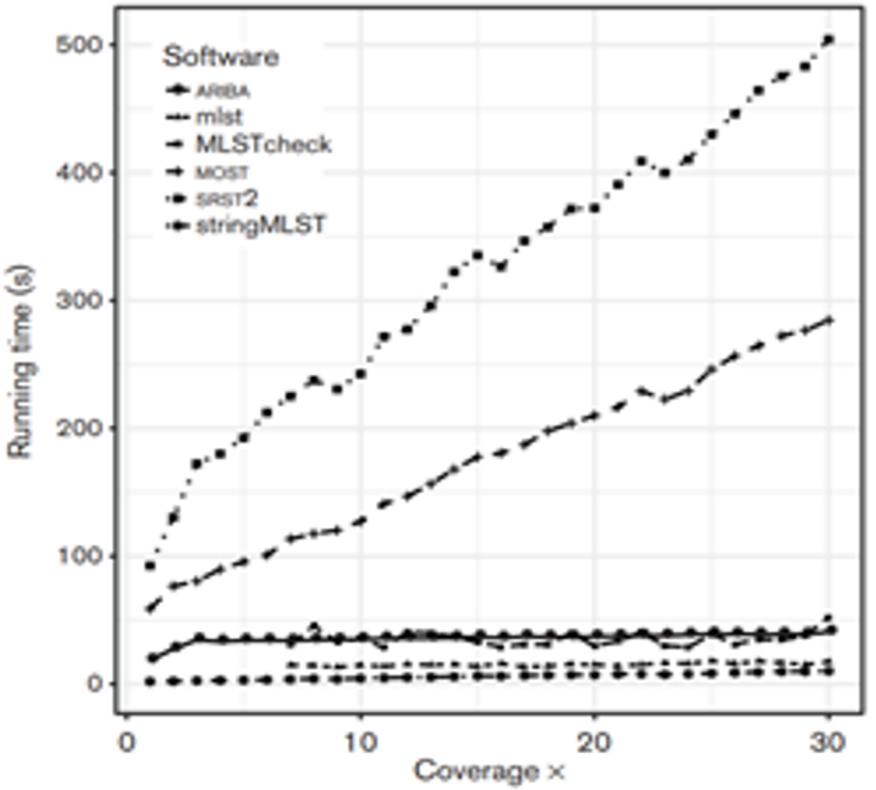


Fig. 2. Running time (s) of each application as the coverage increases to assess the impact of the depth of coverage. No assembled contiguous sequences could be generated where the coverage was less than 7×, as such no data was recorded for the reliant methods (*mlst* and MLSTcheck). No performance results are available for BioNumerics or SeqSphere+.

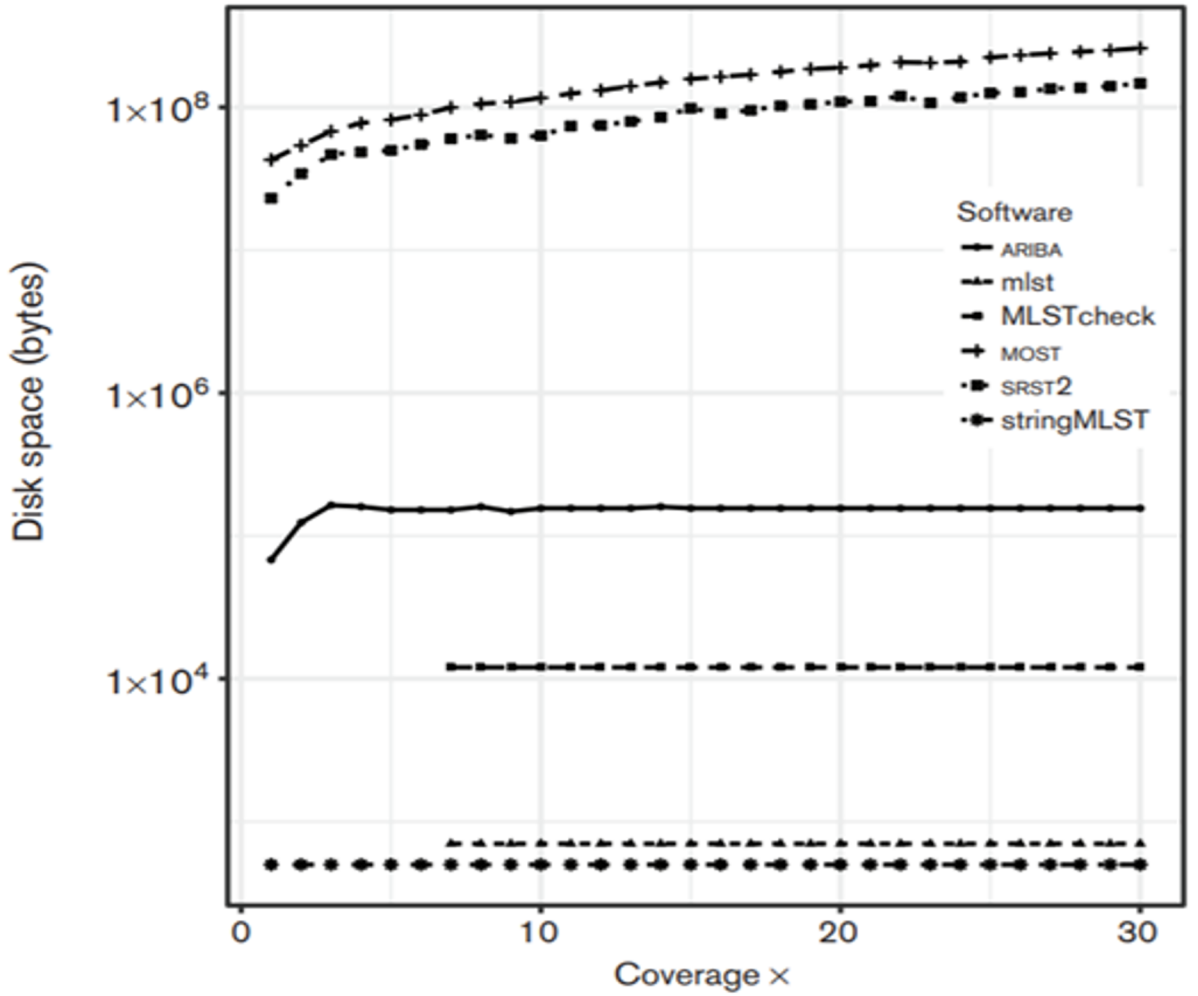


Fig. 3. Disk space requirements in bytes for each software application as the depth of coverage increases. Due to the large difference between applications, a log scale is used.

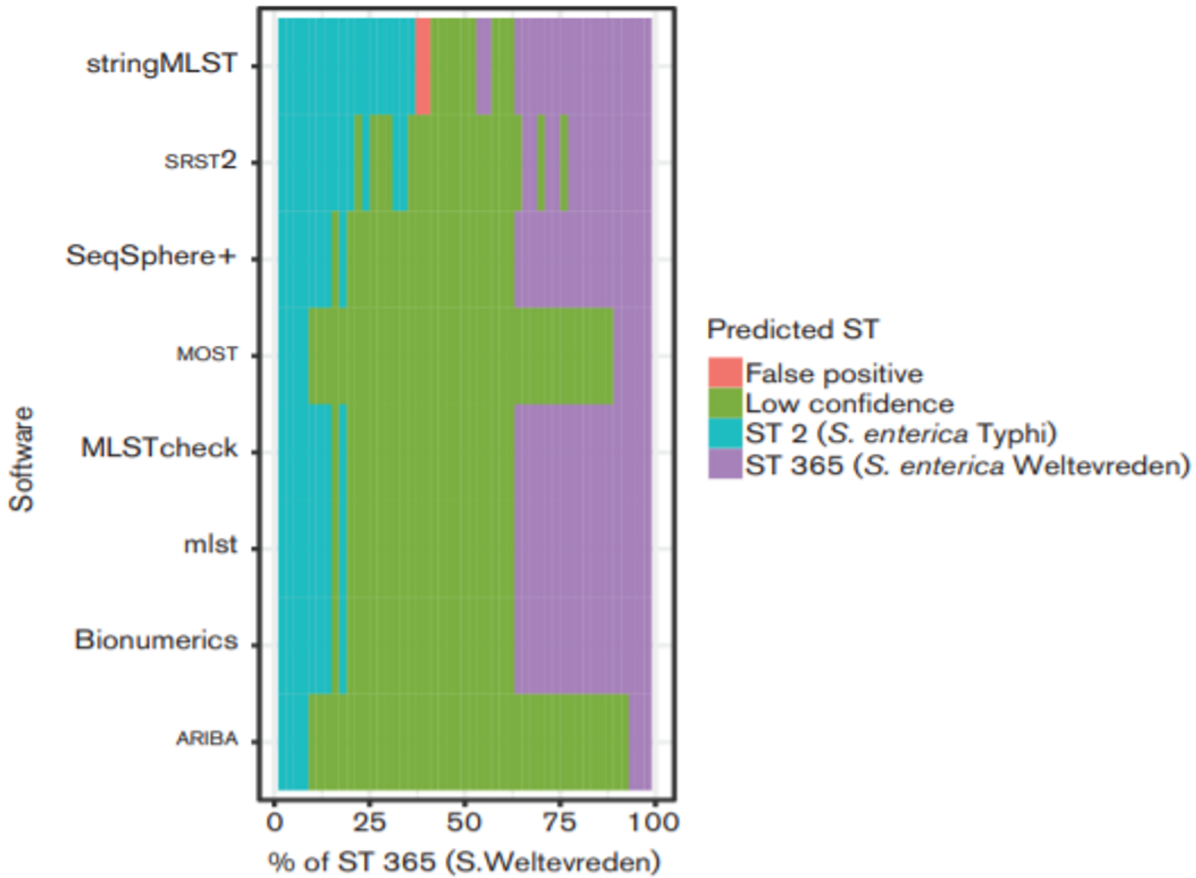


Fig. 4. STs called by each software application when given data containing two different *Salmonella* samples in varying ratios of abundance. Where there is no ST called, or where the ST has any ambiguity at all, it is marked as low confidence. A false positive is where an ST is called with high confidence and is not one of the two samples in the raw data.