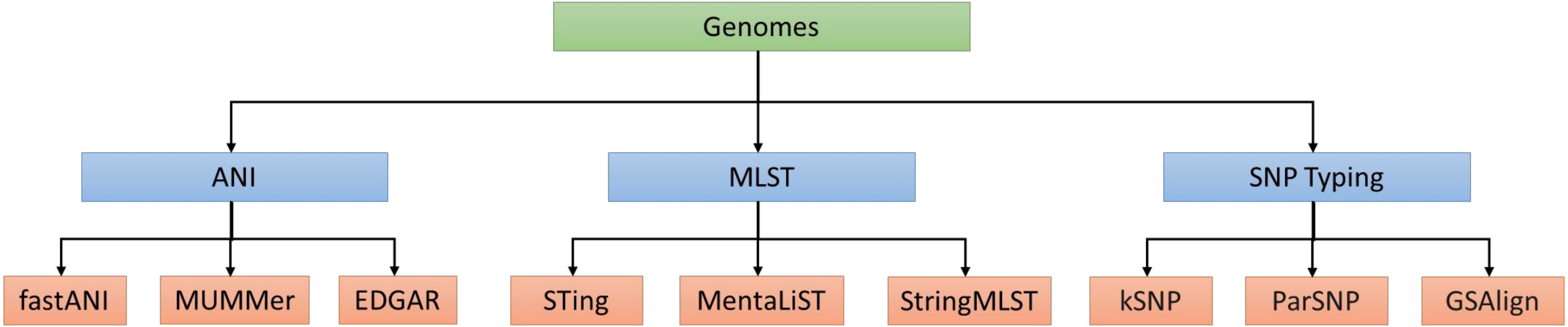
Comparative Genomics: Background and Strategy

Team 2 (Group 4): Ishika Verma, Jyothi Guruprasad, Gautham Krishna Sankar Ramalaxmi, Pushti Dhananjay Mehta, Shreya Rajasekar, Varsha Srinivasan

Pipeline



Genomes

ANI

MLST

SNP Typing

fastANI

MUMMer

EDGAR

STing

MentaLiST

StringMLST

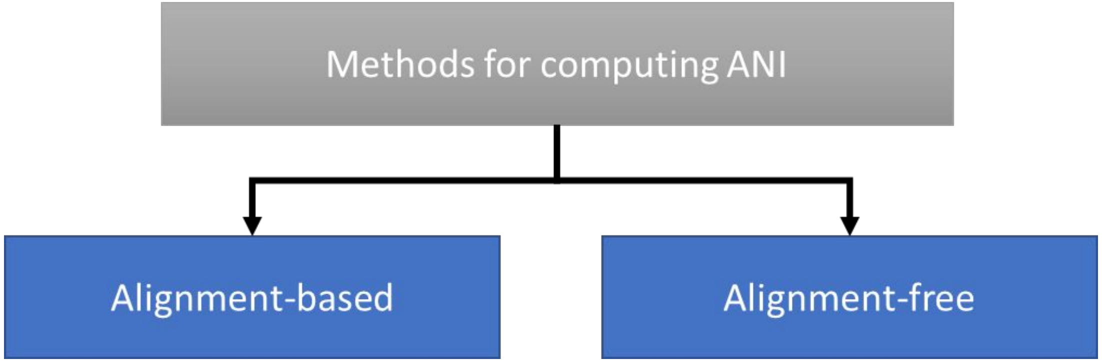
kSNP

ParSNP

GSAlign

ANI-Based Methods

- Measures the average nucleotide identity between homologous genomic regions shared between two genomes



Methods for computing ANI

Alignment-based

Alignment-free

FastANI

- FastANI is a software tool which calculates the average nucleotide identity (ANI) between two genomes
- Sequence alignment-free method based on k-mer sampling and hashing to compute ANI values, which reduces the computational complexity and memory requirements of the algorithm
- Computationally efficient and scalable, hence suitable for analyzing large numbers of bacterial genomes
- It has three modes - one to one, one to many, many to many
- Input: FASTA or multi-FASTA format
- Output: .out file

MUMMer

- Stands for "Maximal Unique Matches Mapper"
- The main algorithm used in MUMmer is based on the concept of maximal exact matches (MEMs)
- One of the main strengths of MUMmer is its speed and efficiency
- It is able to handle very large datasets
- Results produced by MUMmer are highly accurate and reliable
- Two main executables in MUMmer are nucmer and promer
- Input: FASTA format & multi-FASTA format files containing multiple sequences
- Output: .delta, .coords, .map, and .mums.

EDGAR

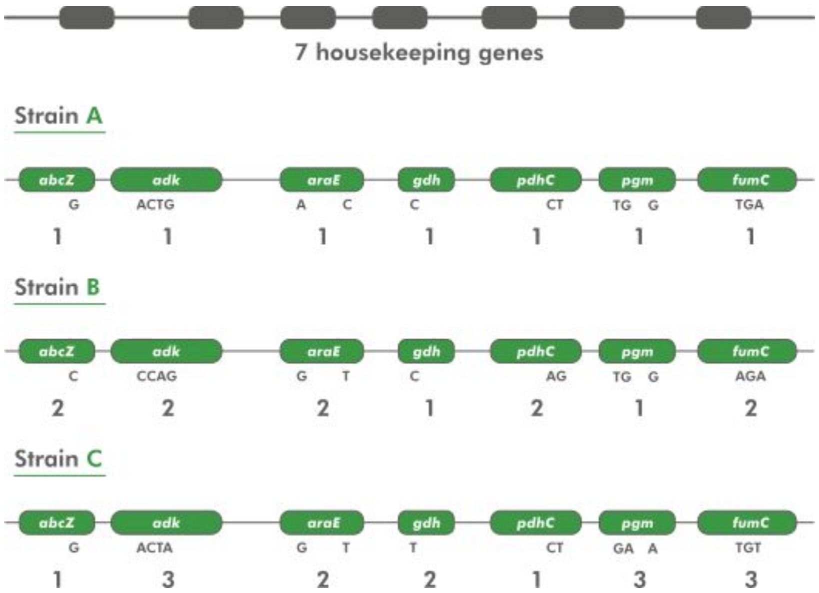
- EDGAR stands for Efficient Database framework for comparative Genome Analyses using BLAST score Ratios
- Developed by the Max Planck Institute in Germany
- It is a tool that enables efficient analysis of large-scale genomic datasets.
- EDGAR is based on a novel approach uses BLAST score ratios (BSRs) to identify homologous regions between genomes
- User-friendly interface for visualization and exploration of data
- Inputs: fasta format
- Output: sql,log files, text files, tsv files

Comparison

	FastANI	MUMMer	EDGAR
Methodology	Alignment-free tool	Alignment-based tool	Alignment-based tool
Algorithm Approach	k-mer frequencies	suffix tree data structures to identify exact matches between sequences	BLAST score ratios to identify homologous regions
Limitations	Genomic rearrangements and other structural variations	Computationally intensive	Requires significant computational resources

MLST-Based Methods

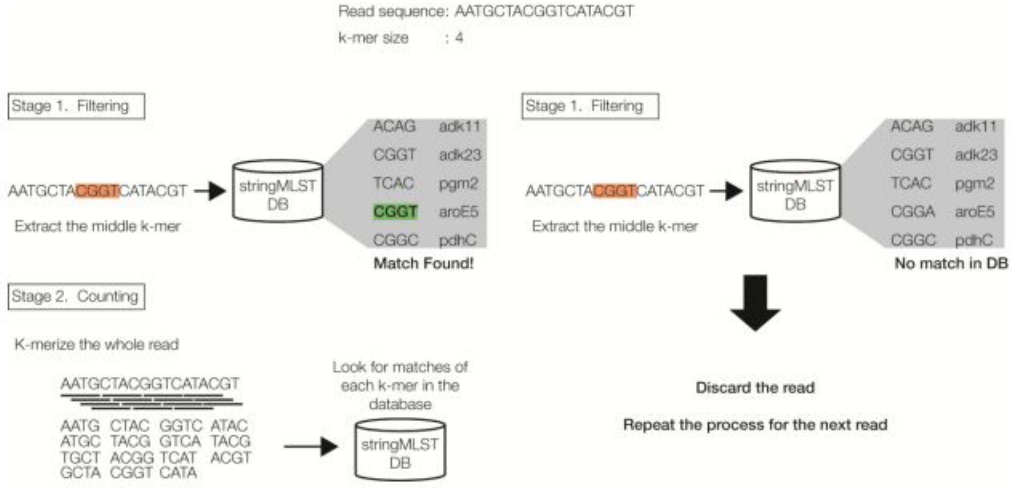
- MLST stands for Multi Locus Sequence Typing
- Molecular typing method used to identify and differentiate bacterial strains based on their DNA sequences
- Based on identifying sequence types from a small number of housekeeping genes



	abcZ	adk	araE	gdh	pdhC	pgm	fumC
Strain A	G	ACTG	A C	C	CT	TG G	TGA
	1	1	1	1	1	1	1
Strain B	C	CCAG	G T	C	AG	TG G	AGA
	2	2	2	1	2	1	2
Strain C	G	ACTA	G T	T	CT	GA A	TGT
	1	3	2	2	1	3	3

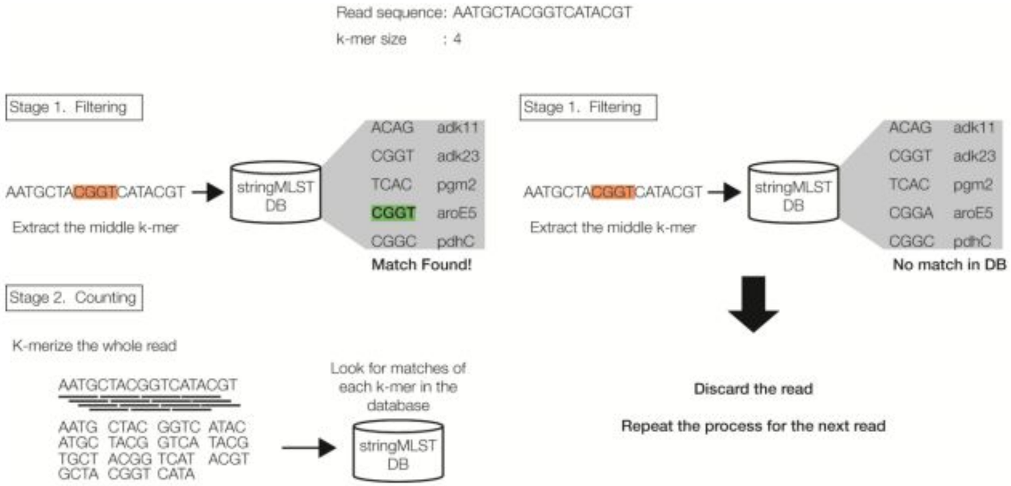
StringMLST

- Tool for detecting MLST of an isolate directly from the genome sequencing reads
- Works on the basis of exact string matching
- Workflow mainly consists of 2 parts - database building & ST discovery
- Input - raw FASTQ files (can be obtained from PubMLST)
- Output - phylogenetic tree



STing

- Uses exact k-mer matching and frequency counting paradigm
- Uses various programming languages:
  - Implementation - C++
  - Downloading database - Python
  - Visualization - R
- Workflow mainly consists of 2 parts - database indexing & sequence variant detection
- Input - raw FASTQ files (can be obtained from PubMLST), WGS sample
- Output - phylogenetic tree

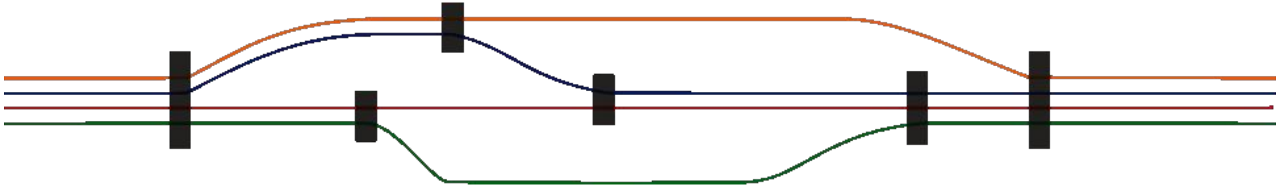


MentaLiST

- An MLST caller based on a k-mer voting algorithm
- Written in Julia, specifically designed and implemented to handle large typing schemes
- General principle: Find all k-mers present on the MLST scheme alleles, for each locus, and store this information as a k-mer hash map in an index file. Then, for each k-mer in the reads of a given sample, all alleles that contain this k-mer will receive one vote. The allele called for each locus is then the one with the most votes.

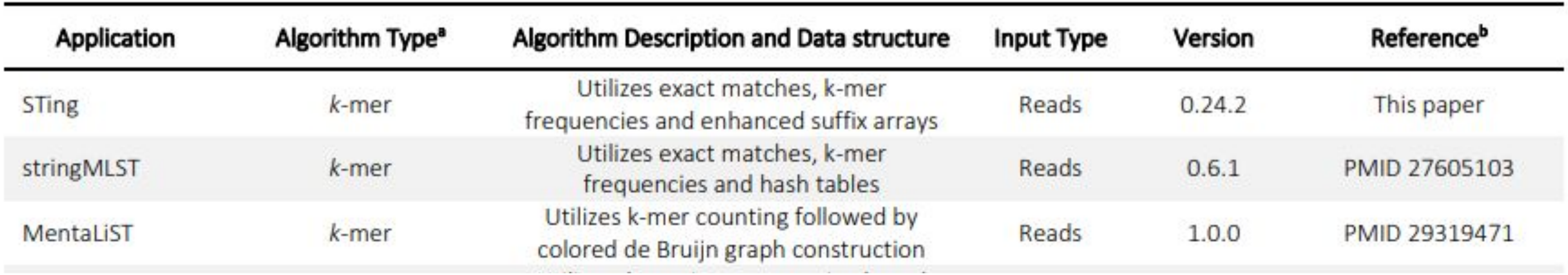
MentaLiST

- Input - Sample file (.fasta) and k-mer database (.db)
- Output - Log file with the number of votes for each allele in each locus and a list of tied alleles



Sketch of a coloured de Bruijn graph with four alleles, each represented by a different colour. The branching nodes are marked in grey, and paths between those nodes correspond to contigs. All nodes of the same contig have the same set of colours.

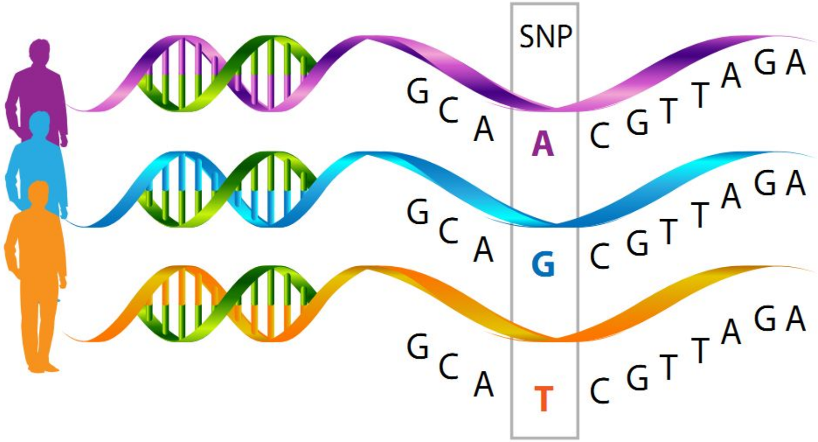
Comparison



Application	Algorithm Type[a]	Algorithm Description and Data structure	Input Type	Version	Reference[b]
STing	*k*-mer	Utilizes exact matches, k-mer frequencies and enhanced suffix arrays	Reads	0.24.2	This paper
stringMLST	*k*-mer	Utilizes exact matches, k-mer frequencies and hash tables	Reads	0.6.1	PMID 27605103
MentaLiST	*k*-mer	Utilizes k-mer counting followed by colored de Bruijn graph construction	Reads	1.0.0	PMID 29319471

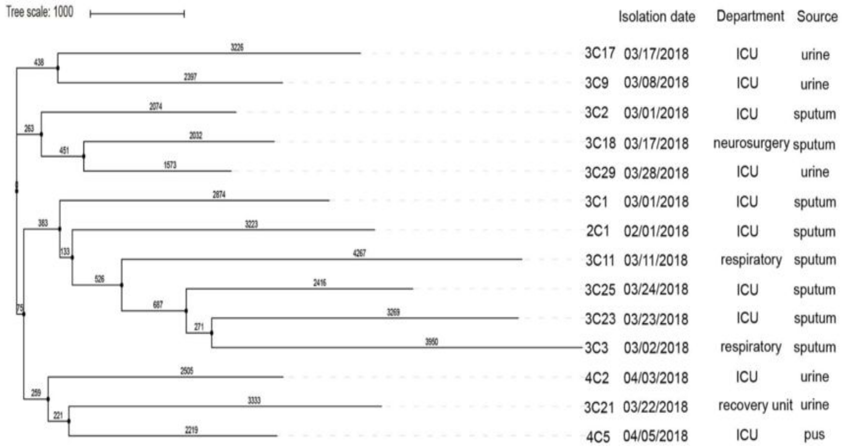
SNP Typing-Based Methods

- SNP typing is a method used in comparative genomics for analyzing genetic variation between different isolates or strains of organisms.
- It is possible to determine the relatedness of the different isolates
- Construct a phylogenetic tree to visualize their evolutionary relationships.



kSNP

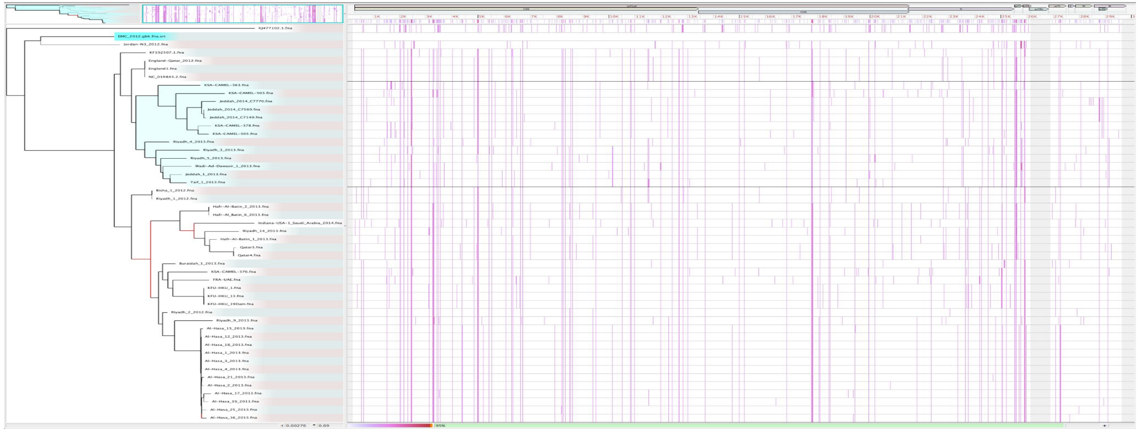
- The tool works by dividing the genomic data into k-mers, which are subsequences of length k, and then identifying SNPs within these k-mers.
- Input - FASTA/FASTQ files
- Output - Phylogenetic tree



	Isolation date	Department	Source
3C17	03/17/2018	ICU	urine
3C9	03/08/2018	ICU	urine
3C2	03/01/2018	ICU	sputum
3C18	03/17/2018	neurosurgery	sputum
3C29	03/28/2018	ICU	urine
3C1	03/01/2018	ICU	sputum
2C1	02/01/2018	ICU	sputum
3C11	03/11/2018	respiratory	sputum
3C25	03/24/2018	ICU	sputum
3C23	03/23/2018	ICU	sputum
3C3	03/02/2018	respiratory	sputum
4C2	04/03/2018	ICU	urine
3C21	03/22/2018	recovery unit	urine
4C5	04/05/2018	ICU	pus

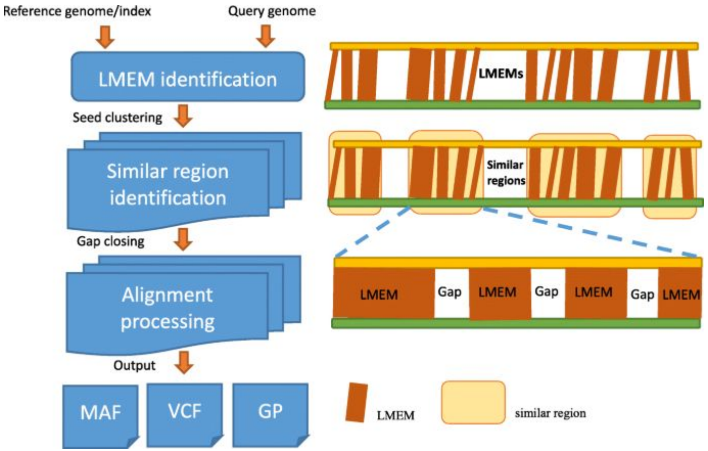
ParSNP

- ParSNP was designed to align the core genome of hundreds to thousands of bacterial genomes within a few minutes to few hours. Input can be both draft assemblies and finished genomes, and output includes variant (SNP) calls, core genome phylogeny and multi-alignments.
- Input - FASTA
- Output -
  - .tree files
  - .vcf files
  - .ggr files
  - .xmfa



GSAlign

- GSAlign is an efficient sequence alignment tool for intra-species genomes. It identifies sequence variations from the sequence alignments. We estimate performance by measuring the correctness of predicted sequence variations.
- Input - FASTA
- Output-
    - maf/aln file
    - vcf file
    - ps file



Comparison

	kSNP	ParSNP	GSAlign
Methodology	K-mer based	K-mer based	Global SA
Scalability	Large dataset	Small dataset	Large dataset
Speed	Relatively fast	Faster than GSAlign	Slow

References

- https://www.applied-maths.com/applications/mlst
- https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000146
- http://jordan.biology.gatech.edu/page/software/stringMLST/
- https://academic.oup.com/nar/article/48/14/7681/5867101
- https://sourceforge.net/projects/ksnp/
- https://github.com/marbl/parsnp
- http://gsalign.sourceforge.net/
- https://mummer.sourceforge.net/manual/

Task Delegation

- ANI-Based Methods - Ishika, Shreya
- MLST-Based Methods - Jyothi, Varsha
- SNP Typing-Based Methods - Gautham Krishna, Pushti