Comparative Genomics

Background and Strategy
Team1

Huyen T Nguyen, Jinkinson Payne Smith, Junkai Yang, Linglin Zhang, Gabriel Leventhal-Douglas, and Monica Isgut

Introduction

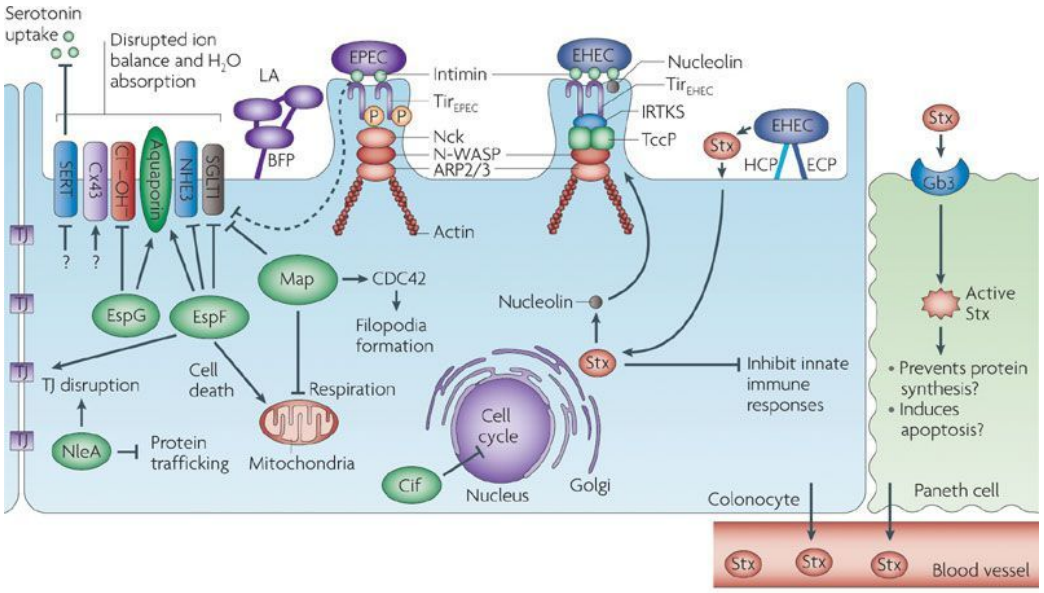
- Microbiologists often work with bacterial isolates taken from an infectious disease outbreak
- They may need to identify the species and strain of bacteria responsible for the outbreak based on the isolate
- Many approaches exist to attempt to accomplish this; often the goal will also be to identify clusters of strains associated with the outbreak
- The methods we will be discussing can be broken down into three categories:

1. Whole Genome Distance
2. Gene by Gene Distance (MLST)
3. SNP-Based Distance

So we have *E. coli*...

- Virulence genes most likely acquired by horizontal gene transfer via plasmids, bacteriophages, pathogenicity islands, and transposons
- Antibiotic resistance is also highly prone to HGT

- Shared virulence strategies between strains
- Widely studied pathotypes may help us differentiate our isolates



Nature Reviews | Microbiology

- Most pathogenic strains belong to diarrheagenic E. coli
- Can cause extra-intestinal infections
- Diagnostic targets may serve as guiding points for strain analysis

Table 1

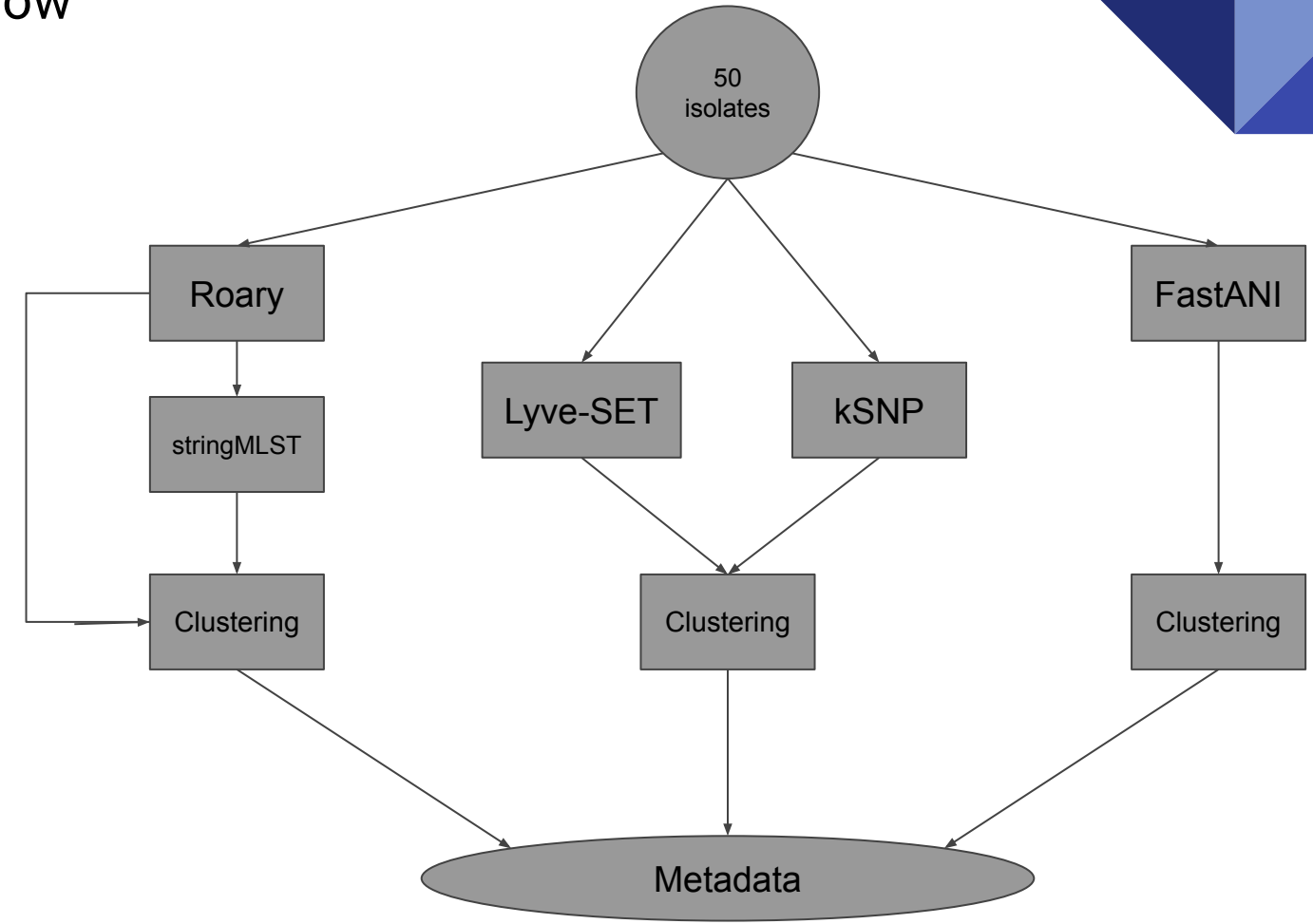
Virulence-associated markers of diarrheagenic *E. coli* from humans.

Pathotype	Defining marker	Essential virulence determinant(s)	Location of essential virulence determinant(s)	Major diagnostic target(s) for PCR	Other diagnostic target(s)
EPEC	LEE PAI	LEE PAI	Pathogenicity island	*eae*	*bfpA*[a]
EIEC/Shigella	pINV	pINV	Plasmid	*ipaH*	Other *ipa* genes
ETEC	ST or LT	ST and/or LT plus colonisation factors	Plasmid; transposon	*elt, est*	
EHEC	Shiga toxin	Shiga toxin 1 and/or 2	Prophages	*stx1, stx2*	*eae*[a], *ehxA*[a]
EAEC	pAA; aggregative adhesion	Not known	Plasmid (probably); possibly others	*aggR, aatA, aaiC*	
DAEC	Afa/Dr adhesins	Not known	Not known	Afa/Dr adhesins[b]	
AIEC	Adherent-invasive phenotype	Not known	Not known	none	none

Objective

- Identify outbreak vs. sporadic strains
- Assess functional & structural similarities and differences between isolates
- Describe the virulence and antibiotic resistance functional features of the outbreak isolates
- Identify the source of the outbreak, and patient(s) zero
- Provide recommendations for outbreak response and treatment

Workflow


50
isolates
Roary
stringMLST
Clustering
Lyve-SET
kSNP
Clustering
FastANI
Clustering
Metadata

Whole Genome Distance

ANI (Average Nucleotide Identity) : average nucleotide identity of all orthologous genes of two genomes

Average Nucleotide Identity (ANI) is computed using the following formula:

$$gANI_{G1 \rightarrow G2} = \frac{\sum_{bbh} (\text{Percent Identity} * \text{Alignment length})}{\text{lengths of BBH genes}}$$

Whole Genome Distance

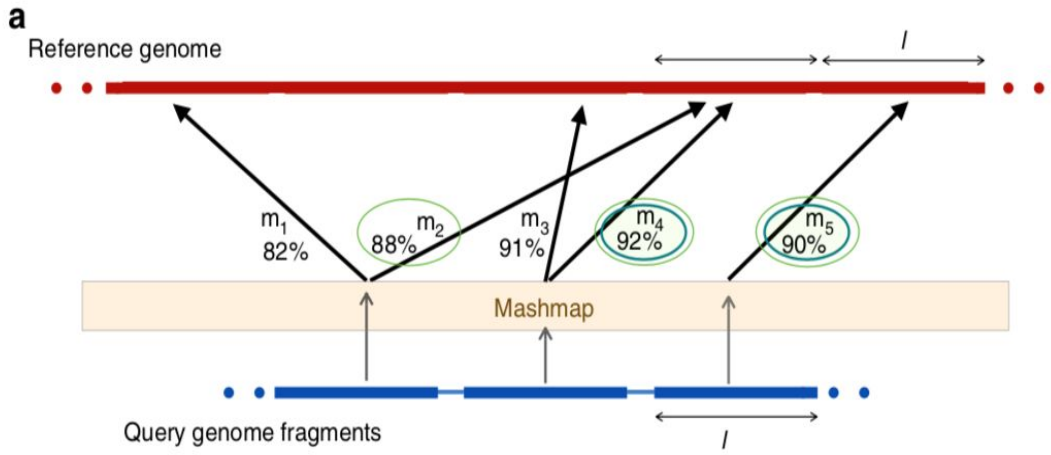
- It is widely accepted that **ANI >95%** indicates the same prokaryotic species
- There seems to be no consensus on ANI threshold to classify strains
- Results from tools consist of distance matrix that can be converted into phylogenetic tree or network graph

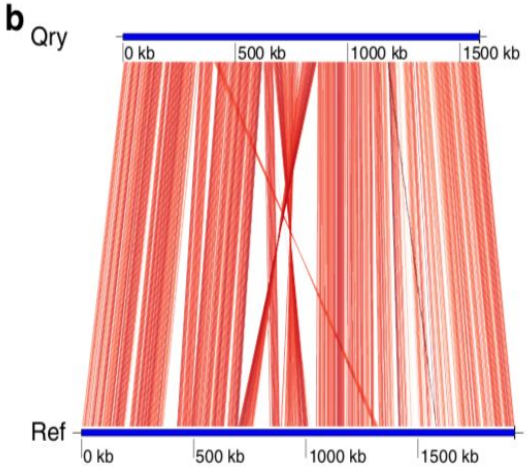
APPROACHES:

- **FastANI** - to get Phylip distance matrix
- **Neighbor Joining** - to get phylogenetic tree
- Visualization of network graph using threshold(s) of choice

FastANI - a MinHash algorithm-based tool for whole genome distance estimation

FastANI: use Mashmap (MinHash based alignment-free sequence mapping algorithm) pairwise comparison for both complete and draft genome assemblies, can estimate ANI in 80-100% identity range





Jain, Chirag, et al. (2018)

FastANI - input and output

Input:

```
./fastANI -q [QUERY_GENOME] -r [REFERENCE_GENOME] -o [OUTPUT_FILE]
./fastANI -q [QUERY_GENOME] --rl [REFERENCE_LIST] -o [OUTPUT_FILE]
./fastANI --ql [QUERY_LIST] --rl [REFERENCE_LIST] -o [OUTPUT_FILE]
```

Output:

Distance matrix in Phylip format

Phylip distance matrix example

- The output of FastANI is a distance matrix in Phylip format.
- Example of a triangular distance matrix:

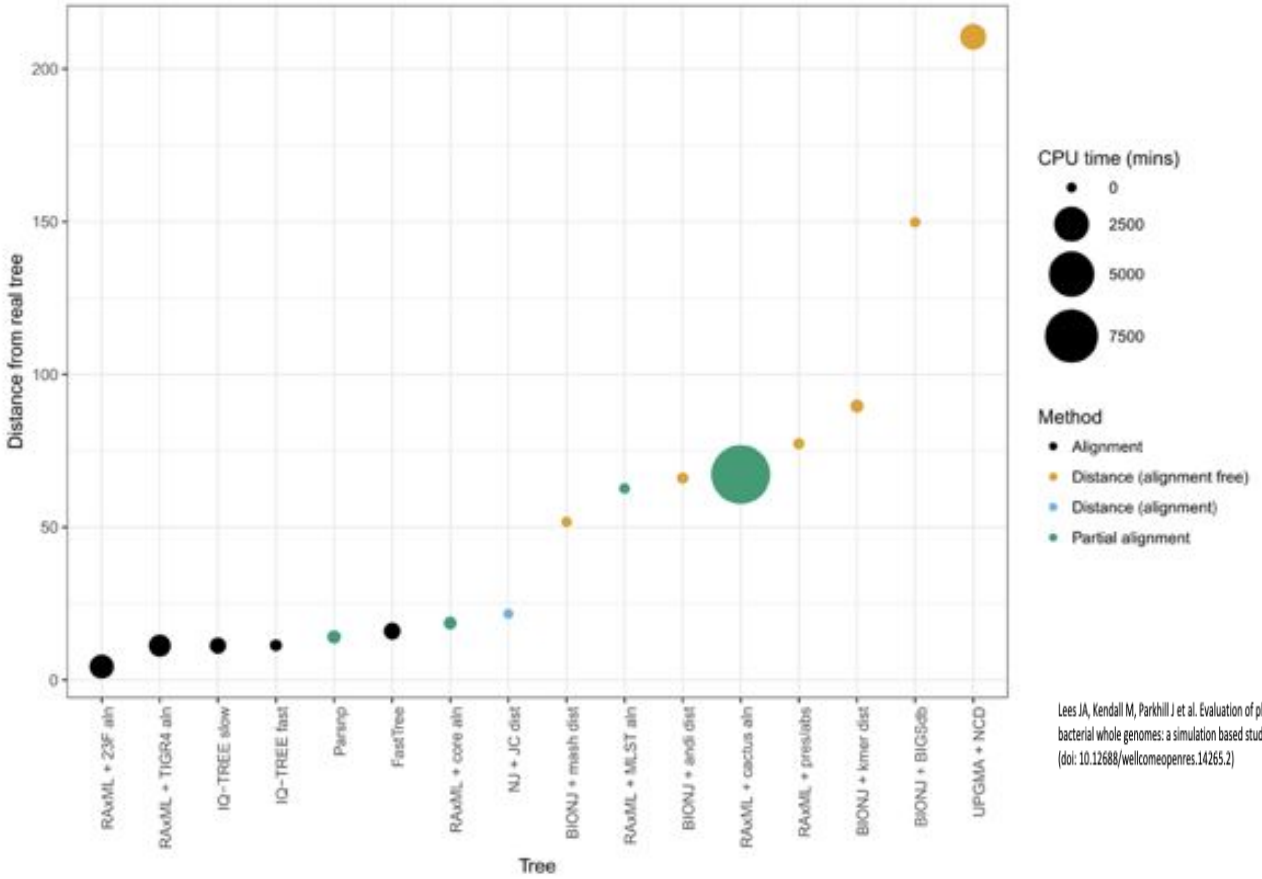
```
98

U68589

U68590      0.3371

U68591      0.3609      0.3782

U68592      0.4155      0.3197      0.4148

U68593      0.2872      0.1690      0.3361      0.2842

U68594      0.2970      0.3293      0.3563      0.3325      0.2768
```

Visualization of genome clusters

Ways to derive insights from a distance matrix:

- **Phylogenetic tree**
- **Network with nodes and edges**

Deciding on tools for phylogenetic tree - 2018 paper



Lees JA, Kendall M, Parkhill J et al. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study [version 2]. Wellcome Open Res 2018, 3:33 (doi: 10.12688/wellcomeopenres.14265.2)

BIONJ - a neighbor joining algorithm-based tool for phylogenetic tree estimation from whole genome distance

- When combined with MASH distance algorithm, performed best out of methods that did not require alignment
- CPU time only 0.75 minutes

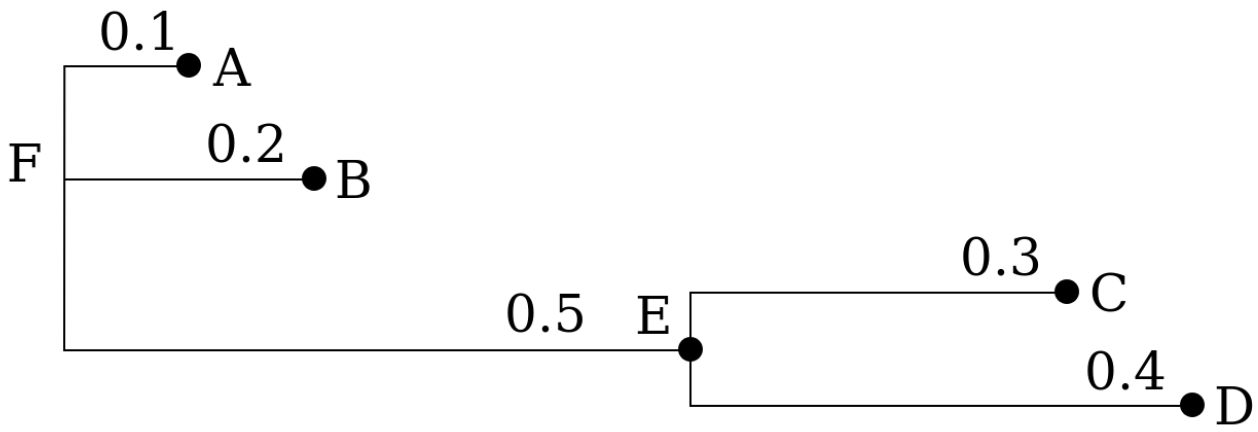
Input:

Distance matrix in Phylip format

Output:

Phylogenetic tree in Newick format

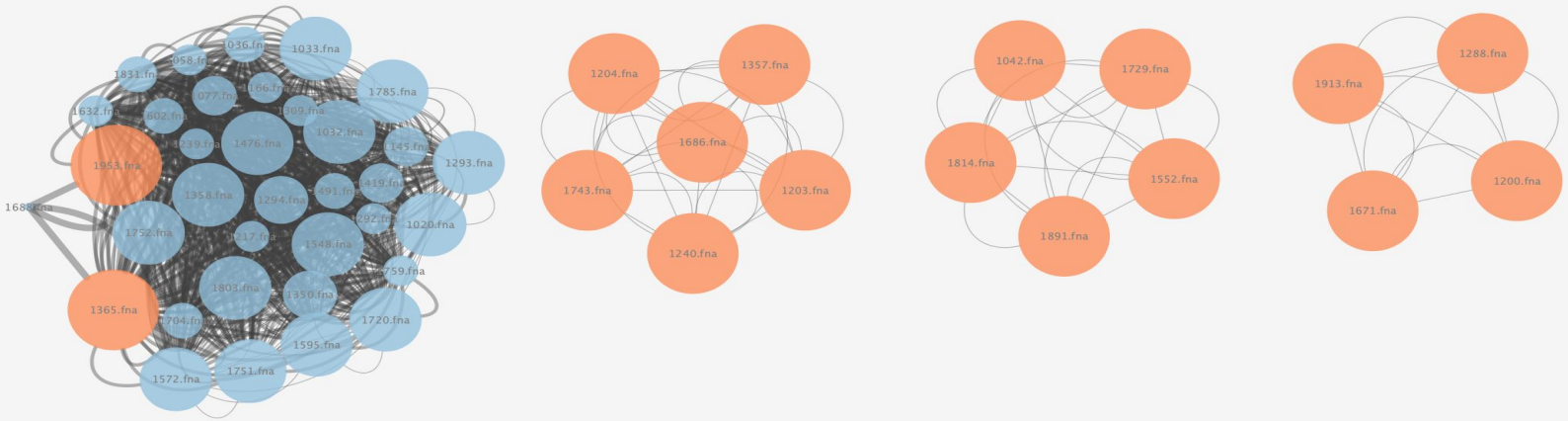
Newick phylogenetic tree example




```
(,,(,));                                no nodes are named

(A,B,(C,D));                            leaf nodes are named

  (A,B,(C,D)E)F;                         all nodes are named
```

https://en.wikipedia.org/wiki/Newick_format

Network with nodes and edges



Chose ANI threshold for strains: >99% by inspection
Orange = low eccentricity (minimum eccentricity = more central node)
Large = higher closeness centrality
Large + Orange = nodes more closely related to the other nodes in cluster

Cluster 1

CGT1020	Human	Stool	12/5/2016	NJ
CGT1032	Human	Stool	12/5/2016	MI
CGT1033	Human	Stool	12/9/2016	GA
CGT1036	Human	Stool	12/14/2016	VA
CGT1058	Human	Stool	12/10/2016	FL
CGT1077	Human	Stool	12/9/2016	SC
CGT1145	Human	Stool	12/24/2016	VA
CGT1166	Human	Stool	12/20/2016	VA
CGT1217	Human	Stool	12/15/2016	VA
CGT1239	Human	Stool	12/15/2016	SC
CGT1292	Human	Stool	12/2/2016	NJ
CGT1293	Human	Stool	12/14/2016	SC
CGT1294	Environmental	Feces	12/9/2016	CA
CGT1350	Human	Stool	12/25/2016	VA
CGT1309	Human	Stool	12/5/2016	MI
CGT1358	Environmental	Burger chain	12/14/2016	SC
CGT1365	Environmental	Water	12/20/2016	SC
CGT1419	Environmental	Water	12/11/2016	CA
CGT1476	Environmental	Leafy Green	12/10/2016	CA
CGT1491	Human	Stool	12/25/2016	VA
CGT1548	Environmental	Prepack Store Sa	12/11/2016	GA
CGT1572	Environmental	Prepack Store Le	12/15/2016	VA
CGT1595	Human	Stool	12/4/2016	MI
CGT1602	Human	Stool	12/9/2016	GA
CGT1632	Human	Stool	12/10/2016	FL
CGT1688	Environmental	Salad Mix	12/3/2016	WV
CGT1704	Human	Stool	12/5/2016	WV
CGT1720	Human	Stool	12/5/2016	NJ
CGT1751	Human	Stool	12/3/2016	NJ
CGT1752	Human	Stool	12/14/2016	GA
CGT1759	Human	Stool	12/11/2016	GA
CGT1785	Human	Stool	12/7/2016	GA
CGT1803	Human	Stool	12/14/2016	GA
CGT1831	Human	Stool	12/13/2016	VA
CGT1953	Human	Stool	11/28/2016	TN

Cluster 2

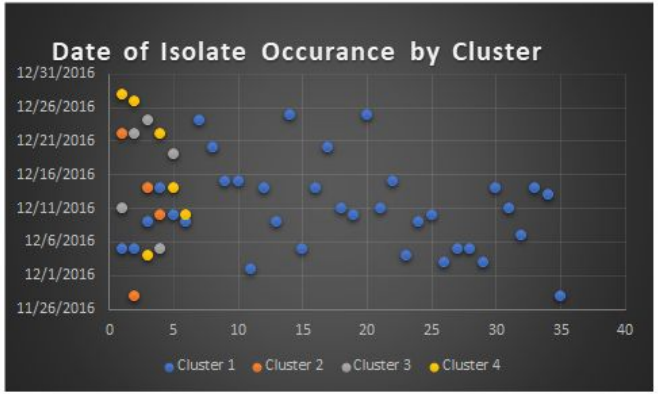
CGT1204	Human	Stool	12/28/2016	SC
CGT1357	Environmental	Water	12/27/2016	GA
CGT1743	Environmental	Salad Mix	12/4/2016	FL
CGT1240	Human	Stool	12/22/2016	TN
CGT1686	Environmental	Salad Mix	12/14/2016	FL
CGT1203	Human	Stool	12/10/2016	FL

Cluster 3

CGT1042	Environmental	Water	12/11/2016	FL
CGT1814	Human	Stool	12/22/2016	VA
CGT1891	Environmental	Water	12/24/2016	SC
CGT1729	Human	Stool	12/5/2016	WV
CGT1552	Human	Unknown	12/19/2016	SC

Cluster 4

CGT1913	Human	Stool	12/22/2016	TN
CGT1288	Human	Stool	11/28/2016	TN
CGT1671	Human	Stool	12/14/2016	TN
CGT1200	Human	Stool	12/10/2016	WV



MLST

- Multilocus sequence typing (MLST) was first proposed in 1998 specifically for inferring genetic relationships between bacteria
- It is often used to analyze isolates from infectious disease outbreaks for surveillance purposes
- It is based on analyzing and comparing multiple alleles from “housekeeping” genetic loci between bacteria
- It estimates relationships between bacteria based on their unique alleles, rather than their nucleotide sequences

High-resolution bacterial genome mapping

- One widely-used method of mapping the genetic relationships between bacteria is to compare their 16S rRNA genes
- This method is widely used and very effective, except when used on closely related bacteria
- This includes comparing different isolates within a species or potentially even different species within a genus
- In such cases, comparing these genes will provide limited resolution, meaning that a new method(s) had to be developed
- MLST is one such method; it is now “the method of choice for typing many organisms” (Maiden et al. 2013)

MLST and pan-genome analysis

- MLST only compares bacterial sequences at a selection of “housekeeping” genes
- By contrast, pan-genome analysis compares “the entire repertoire of genes accessible to the clade studied” (Vernikos et al. 2015)
- Pan-genome analysis and MLST both enable the detailed modeling and prediction of bacterial genomic diversity
- MLST can only be used on closely related bacteria
- Its focus on “housekeeping” genes allows it to account for widespread vertical and horizontal genetic transfer in bacteria

Varieties of MLST

- There are now many different types of MLST that have been developed, including:
- Whole-genome MLST (wgMLST), “in which all the loci of a given isolate are compared to equivalent loci in other isolates” (Maiden et al. 2013)
- Core-genome MLST (cgMLST), focused on only the core elements of the genomes of a group of bacteria
- Ribosomal MLST (rMLST), based only on the 53 loci that code for ribosomal proteins in most bacteria

The “gene-by-gene” approach

- Also known as the “MLST-like” approach, this method involves conducting a *de novo* assembly and annotation
- It can be thought of as applying MLST to whole-genome sequences (WGS)
- It is exceptionally versatile and flexible, and you can increase the level of detail simply by including more genes in the analysis

Roary

- Roary is a tool that builds bacterial pan-genomes based on a large number (potentially thousands) of related isolates
- It takes in annotated *de novo* assemblies, all of which must be from the same species
- “Isolates are clustered based on gene presence in the accessory genome, with the contribution of isolates to the graph weighted by cluster size” (Page et al. 2015)

Samples	Software	Core[a]	Total	RAM (mb)	Wall time (s)
24	PGAP	—	—	—	—
	PanOCT	4522	4991	5313	96 093
	LS-BSR	4451	4843	554	7807
	Roary	4436	4941	444	382

Piggy

- Piggy is a modified version of the pan-genome analysis tool Roary
- But instead of assembling large-scale pan-genomes, Piggy only assembles intergenic regions of bacterial genomes
- The advantage of this comes from the fact that most pan-genome tools (including Roary) only focus on protein-coding sequences
- This is despite the fact that non-coding regions are often also phenotypically important, a shortcoming which Piggy addresses (Thorpe et al. 2018)

stringMLST

- stringMLST is a tool for detecting the sequence type (ST) of a bacterial isolate directly from the genome sequence reads.
- Much Faster algorithm compared with tradition MLST tools while still has high accuracy.
- The scale of the analysis is flexible. (manually create database)
- Accept existing database on the internet.

sample	gene1	gene2	gene3	gene4	ST
CGT1	1	1	1	1	1
CGT2	1	2	1	1	2
CGT3	3	1	5	10	12

Comparative test

Tool name	Type[a]	Input	% Correct		Run time[b]
			Alleles	STs	
stringMLST	K-mer	Reads	100.0	100.0	45
CGE/MLST	BLAST	Reads	99.6	97.5	2922
SRST2	Mapping	Reads	98.6	92.5	1887
SRST	BLAST	Assembly	95.0	77.5	2386
Offline CGE	BLAST	Assembly	96.1	80.0	170

stringMLST

Locus	Function
dinB	DNA polymerase
icdA	Isocitrate dehydrogenase
pabB	p-aminobenzoate synthase
polB	Polymerase PolII
putP	Proline permease
trpA	Tryptophan synthase subunit A
trpB	Tryptophan synthase subunit B
uidA	Beta-glucuronidase

- MLST:

  Download Escherichia locus/sequence definitions database from PubMLST.

  But the database only contains housekeeping genes.

- wgMLST

  Build database and definition profile according to the result of pan-genome analysis.

SNP-Based Approach

- Single nucleotide changes can be measured for phylogenetic comparison
- Advantages include detecting host specific intergenic region SNPs
- This is very useful when differentiating closely related strains
- However computationally demanding when providing an exceptionally high subtyping resolution
- Two types:
  - **Reference based** (Parsnp, RealPhy, CSI Phylogeny-1.2, CFSAN, PHEnix, etc.)
    - Higher likelihood of including all information of raw data
    - Computationally expensive, need space to
  - **Non-reference based** (Cortex, Bubbleparse, NIKS, discoSnp, Stacks, etc.)
    - Including lineage-specific regions that may be absent from reference
    - Non-model organisms can be greatly facilitated
    - Smaller groups with fewer resources/ wider phylogenetic groupings
    - Help resolve the data storage and access issues, personal genomics based medicine.
    - **De Bruijn graph** as data-structure for identifying variants

- The tool **kSNP** can be used either with or without reference genome

Reference-free SNP-based approach

De Bruijn graph

```
>read_1
CGATTCTAAGTGTACTGC...
```

1. Break the reads into overlapping bits of length k (k-mers)
2. Make each k-mer a node in the graph
3. Make links between overlapping kmers
4. Follow paths

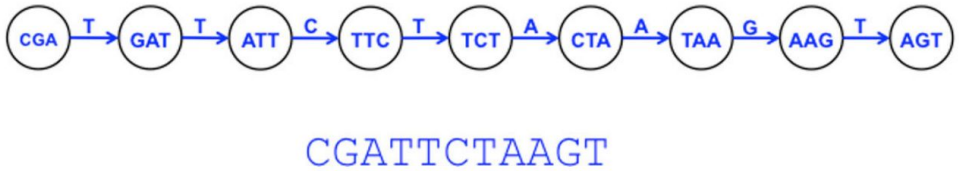
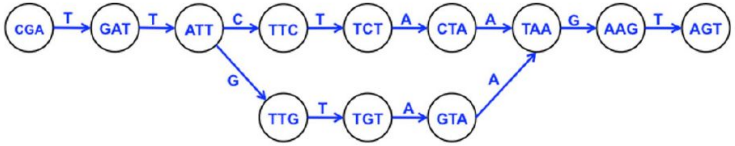


Figure 1 De Bruin graphs constructed from overlapping k-mers. De Bruijn graphs are networks of short overlapping sub-sequences of reads of length *k*. Typically, *k*-mers are set as the nodes in the graph and links are drawn between *k*-mers that have overlap of length *k - 1*, that is they overhang each other by just one nucleotide.

Reference-free SNP-based approach

SNP bubble



CGATTCTAAGT
CGATTGTAAGT

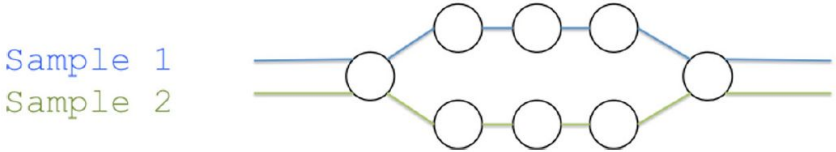


Figure 2 Bubble structures formed in De Bruijn graphs by SNPs. Bubble structures form as the result of a divergence in sequence by one nucleotide, initially at the end of a *k*-mer, that then moves backwards at each progressive node, allowing for a close of the two paths at the end. Colouring the edges in the graph according to sample provenance helps identify inter-sample SNPs.

kSNP 3.0

Why chose this one:

- Annotation of SNPs in all replicons can be provided
- Parsimony tree is consensus, not random
- Input file is a list of paths to genome files, helpful when dealing with raw read genome file sizes >> 500MB
- Automatically detects and incorporate raw-read files
- Option to append a new genome to an existing run, save time compared with repeating the run

Compared to reference genome required methods:Designed to deal with aligning **large numbers of microbial genomes** and **reference genome is not required**, more versatile comparing to Parsnp as one of the substitutes.

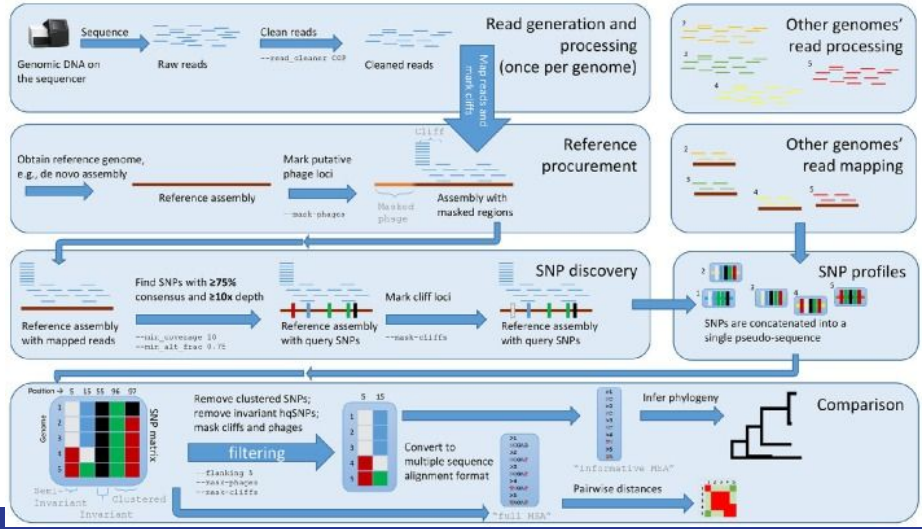
RealPhy **depends on accurate mapping** of raw reads (or contigs) to the reference genomes. **Taxon diverged by > 5–10% the distances to reference** are **underestimated**, leading to **incorrect topologies**.

Limitations:

- Cannot find SNPs that are too close together (closer than one half k).
- Cannot distinguish true SNPs from sequencing errors

Lyve-SET (SNP Extraction Tool)

- High quality SNP pipeline to remove lower-quality SNPs and increase phylogenetic signal
- Map reads to reference using SMALT (hash index)
- Call SNPs from aligned reads with VarScan v2.3.7
  - Empirically determined E. coli settings
- Creation of SNP matrix with bcftools
- MSA FASTA created from SNP matrix
- Phylogeny inferred with RAxML v8
- Independently run tools



References

- Ibarz Pavón AB, Maiden MC. Multilocus sequence typing. *Methods Mol Biol*. 2009;551:129-40.
- 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.
- Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, Julian Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, Bioinformatics, Volume 31, Issue 22, 15 November 2015, Pages 3691–3693, https://doi.org/10.1093/bioinformatics/btv421
- Rong X, Huang Y. Multi-locus Sequence Analysis: Taking Prokaryotic Systematics to the Next Level. In: New Approaches to Prokaryotic Systematics. Vol 41. Methods in Microbiology. Elsevier; 2014:221-251.
- Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. GigaScience. 2018;7(4). doi:10.1093/gigascience/giy015
- Croxen, M. A., & Finlay, B. B. (2009). Molecular mechanisms of Escherichia coli pathogenicity. *Nature Reviews Microbiology,8*(1), 26-38. doi:10.1038/nrmicro2265

References (cont.)

- Gardner SN, Slezak T, Hall BG. kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics. 2015 Apr 25;31(17):2877-8.
- Saltykova A, Wuyts V, Mattheus W, Bertrand S, Roosens NH, Marchal K, De Keersmaecker SC. Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to Salmonella enterica serotype Typhimurium and serotype 1, 4,[5], 12: i:-. PloS one. 2018 Feb 6;13(2):e0192504.
- Leggett RM, MacLean D. Reference-free SNP detection: dealing with the data deluge. Bmc Genomics. 2014 May;15(4):S10.
- Jain, Chirag, et al. "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries." *Nature communications* 9.1 (2018): 5114.
- Katz LS, Griswold T, Williams-Newkirk AJ, et al. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Front Microbiol*. 2017;8:375. Published 2017 Mar 13. doi:10.3389/fmicb.2017.00375

Workflow