B3 Background: Genotyping, Taxonomic and Quality Assessment

By: Celine Al-Noubani, Kyungbeom Kim, Lydia Keller, Eunsu Hwang, & Krisha Shetty

Georgia Tech

Timeline

- 2/25/25 Initial Meeting
  - Researched tools and assigned tools
  - Set 2/27 as the deadline to finish running small and large files
- 2/27/25 Meeting
  - Finalizing command scripts
  - Set 3/2 midnight deadline to push prelim results to Github
- 3/03/25 Meeting
    - Finalize preliminary presentation
    - Divide up slides and practice presenting
- 3/04/25 To Do
    - Presentation day!
    - Record respective section for background presentation & edit video
    - Deadline @11:59pm to submit background presentation and finalize Github

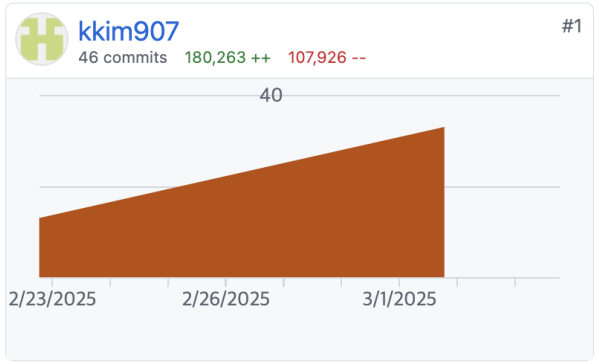
Timeline Moving Forward

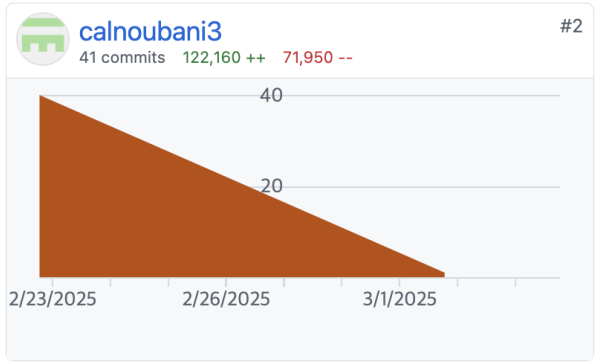
- 3/10/25 Meeting
  - Finalize the pipeline
  - Divide up files
  - Get to work!
- 3/24/25 Meeting
  - Finalize presentation
  - Divide up slides and practice presenting
- 3/25/25
  - Presentation day!
  - Deadline @11:59pm to finalize Github

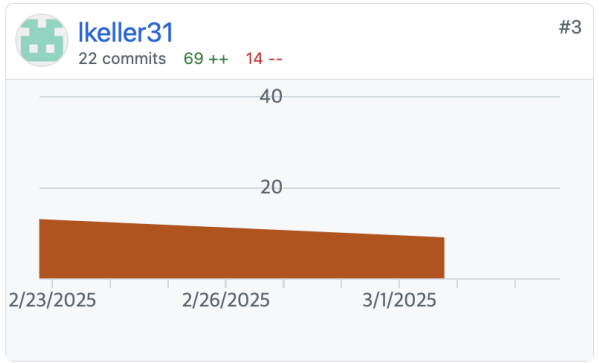
Breakdown and Student Roles

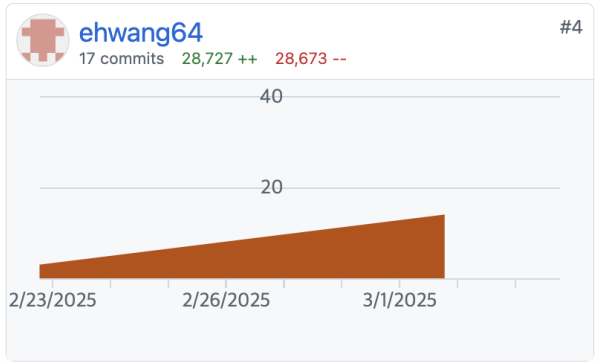
1. Genotype our samples with **MLST** - Lydia
2. Taxonomic classification methods to determine identity of samples
   - Genus Level:
     - **Mash** - Lydia
     - **ANI Calculator** - Kyungbeom
   - Species Level:
     - **FastANI** - Krisha
     - **Skani** - Eunsu
3. Quality Assessment Results
   - Contamination and completeness of assemblies (**CheckM**) - Celine
   - (Fine) Contig-by-contig (**Kraken2**) - Kyungbeom

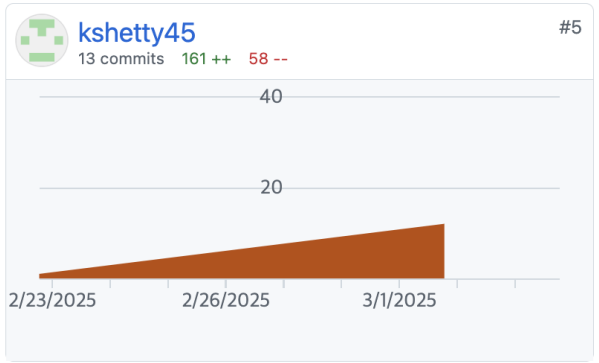
GitHub Activity











Georgia Tech

•Preliminary Data Required

Largest Dataset: B1299860_S01_L001
Smallest Dataset: B1838859_S01_L001
Reference Assembly: Neisseria gonorrhoeae FA 1090

Neisseria gonorrhoeae strain FA1090 was isolated in 1983 from a patient with disseminated gonococcal infection. This whole-genome sequenced bacterial strain has applications in antimicrobial resistance research, infectious disease research, and sexually transmitted disease research.

ATCC

Georgia Tech

Genotyping: MLST

Genotyping: MLST

- Identifies isolates of bacterial species using the sequences of internal fragments of (usually) seven house-keeping genes (https://pubmlst.org/multilocus-sequence-typing)
- Ignores exact sequence differences in favor of giving sequences "allele numbers"
- Seven genes of interest can be identified from PCR products if culturing is not available

B1838859_S01_L001_contigs.fa	neisseria	10314	abcZ(126)	adk(39)	aroE(170)	fumC(111)	gdh(146)	pdhC(153)	pgm(65)

B1299860_S01_L001_contigs.fa	neisseria	10314	abcZ(126)	adk(39)	aroE(170)	fumC(111)	gdh(146)	pdhC(153)	pgm(65)

Each MLST prediction gets a score out of 100. The score for a scheme with N alleles is as follows:

- +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

Georgia Tech

Taxonomic Comparison

MASH: Genus Level

- Mash Distance
  - $D(k,j)=1-1k\ln 2j1+j,\ j=0,\ 0<j\leq 1$
- P-Value
  - $p=1-\sum i=0x-1(si)jir(1-jr)s-i$
- Usage
  - Mash dist <genome1> <genome2>
- Result
  - ANI = 99.6

"Mash distances correlate well with ANI (a common measure of genome similarity), with $D \approx 1 - ANI$ "

**Mash distance ≤0.05 = ANI of ≥95 %

"This threshold roughly corresponds to a 70 % DNA-DNA reassociation value"

Georgia Tech

FastANI: Species Level

- FastANI fragments the query genome into smaller non-overlapping pieces and aligns them with the reference genome
  - Works much faster than traditional alignment-based tools
  - Maintains a high accuracy

- Extra Columns:
  - %Query_Alignment: percent of reference genome fragments that were able to map to query fragments
  - Basepairs_Query_Aligned: the number of reference genome base pairs that the query base pairs can map onto

Query	Reference	%ANI	Num_Fragments_Mapped	Total_Query_Fragments	%Query_Aligned	Basepairs_Query_Aligned
./B1838859/B1838859_problem.fna	./ASM684v1_reference.fna	99.543	649	681	95.301	1947000

Query	Reference	%ANI	Num_Fragments_Mapped	Total_Query_Fragments	%Query_Aligned	Basepairs_Query_Aligned
./B1299860/B1299860_problem.fna	./ASM684v1_reference.fna	99.5571	632	661	95.6127	1896000

Georgia Tech

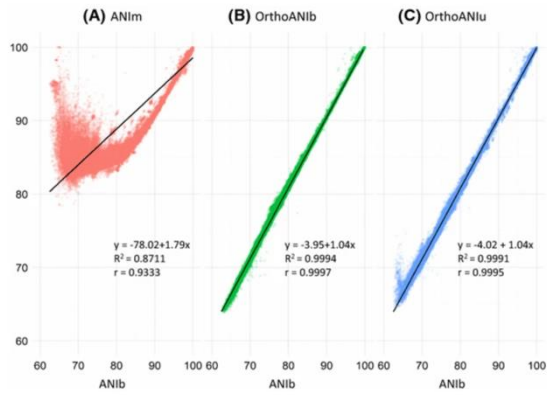
ANI Calculator: Species Level

Why We Chose ANI Calculator?

- Based on the **benchmark study by Yoon et al., 2017**, the **OrthoANIu** algorithm showed **high accuracy and fast processing speed**.
- **OrthoANIu** uses **USEARCH** for pairwise alignment, making it **faster** than BLAST-based methods (e.g., ANI b).
- Web-based tool available through **EzBioCloud**, meaning **no local installation** required.
- Supports **contig-based comparison**, directly matching our fragmented assembly data.

Relevant to Our Goals

- Provides **species-level taxonomic classification** through genomic comparison.
- Outputs **ANI value, alignment coverage, and other key QC metrics**.
- Suitable for both **largest** and **smallest** datasets in our project.



Paper Summary (Yoon et al., 2017)
Title: A large-scale evaluation of algorithms to calculate average nucleotide identity (ANI)
Key Point:
This paper systematically evaluates **multiple ANI calculation methods** to identify the **most ac curate and efficient algorithm** for species-level classification of prokaryotic genomes.

Comparative Study > Antonie Van Leeuwenhoek. 2017 Oct;110(10):1281-1286.
doi: 10.1007/s10482-017-0844-4. Epub 2017 Feb 15.

A large-scale evaluation of algorithms to calculate average nucleotide identity

Seok-Hwan Yoon [1 2], Sung-Min Ha [1 2], Jeongmin Lim [2], Soonjae Kwon [2], Jongsik Chun [3 4]

Affiliations + expand

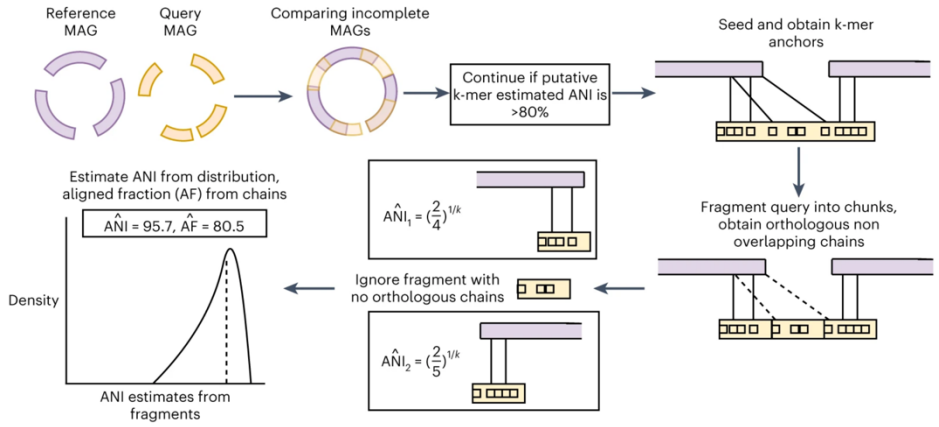
PMID: 28204908 DOI: 10.1007/s10482-017-0844-4

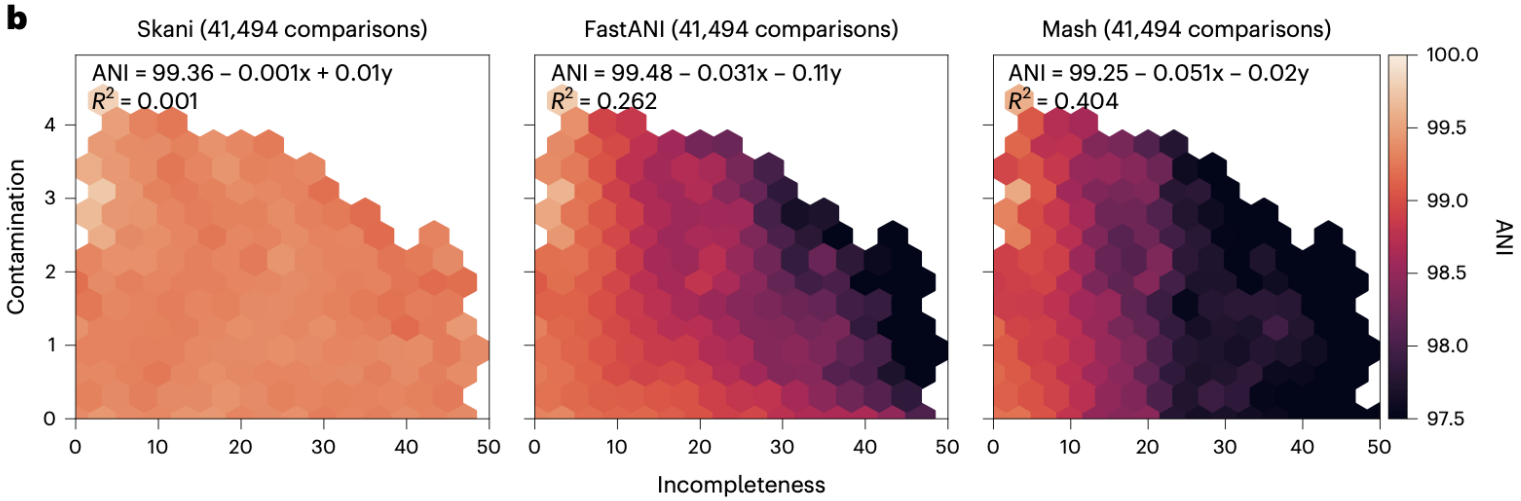
Georgia Tech

Skani: Species Level

- Sequence comparison tool for metagenome-assembled ge nomes (MAGs)
- Fast and robust tool for calculating aligned fraction and ANI in the range>82%
- Skani outperforms fastANI in accuracy and speed
- Useful for extensive, noisy metagenomic datasets
- Find an approximate set of orthologous alignments betwee n two genomes by obtaining a set of minimally overlapping k-mer chains

Algorithm overview





ANI methods are sensitive to incomplet eness and contamination
However, skani is less affected by the m

https://rdcu.be/eccbP

Quality Assessment

CheckM: Whole Assembly

- **Microbial Focus:**
  CheckM is specifically optimized for bacterial and archaeal genomes, using marker gene sets tailor ed to microbial lineages. BUSCO's universal single-copy orthologs work well across many groups b ut aren't as finely tuned for the diversity within microbes.

- **Lineage-Specific Markers:**
  CheckM employs curated, phylogenetically-informed marker genes that adjust based on the organi sm's taxonomic classification. This allows it to more accurately assess completeness and contamin ation by comparing against expected gene profiles.

- **Contamination Detection:**
  BUSCO mainly focuses on completeness without a dedicated contamination metric.

- **Robust & Scalable:**
  - Leverages thousands of reference genomes to generate robust and reliable marker sets.
  - Easily integrated into high-throughput bioinformatics pipelines.
  - Available through conda, ensuring reproducibility and ease-of-use.

Georgia Tech

CheckM: Whole Assembly

- Parameters

```
checkm \
  analyze \
  --threads 8 -x fa \
  Ng.markers \
  /storage/home/hhive1/calnoubani3/data/checkm/asm/small/ \
  analyze_small_output
```

- Use 8 threads
- Look for bin files with the ".fa" extension (-x fa)
- Used the generated "Ng.markers" marker set
- Input directory is the "small" assemblies folder
- Output results are saved in "analyze_small_output"

```
checkm \
  qa \
  --file checkm.small.tax.qa.out \
  --out_format 1 \
  --threads 8 \
  Ng.markers \
  analyze_small_output
```

Ran CheckM's quality assessment (QA)

Georgia Tech

Kraken2: Contig-by-Contig

What is Kraken2?

Kraken2 is a taxonomic classification tool widely used for metagenomic studies.
It balances **speed and accuracy**, making it ideal for **high-throughput classification**.

Why We Chose Kraken2?

- Efficient classification based on k-mer matching.
- Supports large-scale contig-level classification.
- Lightweight, supports pre-built databases.
- Directly outputs species-level and genus-level classification.

Relevant to Our Goals

- Provides both **genus** and **species** level classification.
- Supports contig-by-contig assessment, which fits our fine-level analysis requirement.

[Kraken2 Standard-8 database]

			* Storage					
Standard-8	Standard with DB capped at 8 GB	9/4/2024	5.5	7.5	.tar.gz	.txt	.tsv	.md5

- Suitable for our machine's **memory limit**.
- Contains all major taxa for comprehensive classification.
- Directly compatible with Kraken2 without further preprocessing.

Georgia Tech

Conclusion

1. MLST identified N. gonorrhoeae ST 10314
2. ANIs > 99%
    - We prefer skani due to consistency
3. CheckM showed low contamination
4. Kraken2 showed contig-by-contig assessment
    - Confident with the final species

Moving Forward

- For all genomes
    - MLST
        - Do other STs come up?
    - CheckM
    - ANI
    - Kraken2

Georgia Tech

Citations

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018 Nov 30;9(1):5114. doi: 10.1038/s41467-018-07641-9. PMID: 30504855; PMCID: PMC6269478.

Jolley, K.A., Maiden, M.C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010). https://doi.org/10.1186/1471-2105-11-595

Ondov, B.D., Treangen, T.J., Melsted, P. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x

Ondov, B., Starrett, G., Sappington, A. *et al.* Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol* **20**, 232 (2019). https://doi.org/10.1186/s13059-019-1841-x

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043-55. doi: 10.1101/gr.186072.114. Epub 2015 May 14. PMID: 25977477; PMCID: PMC4484387.

https://pubmlst.org/multilocus-sequence-typing

Shaw, J., Yu, Y.W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat Methods* 20, 1661–1665 (2023).

https://doi.org/10.1038/s41592-023-02018-3

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek*, *110*(10), 1281–1286. https://doi.org/10.1007/s10482-017-0844-4

Georgia Tech

Thank You!