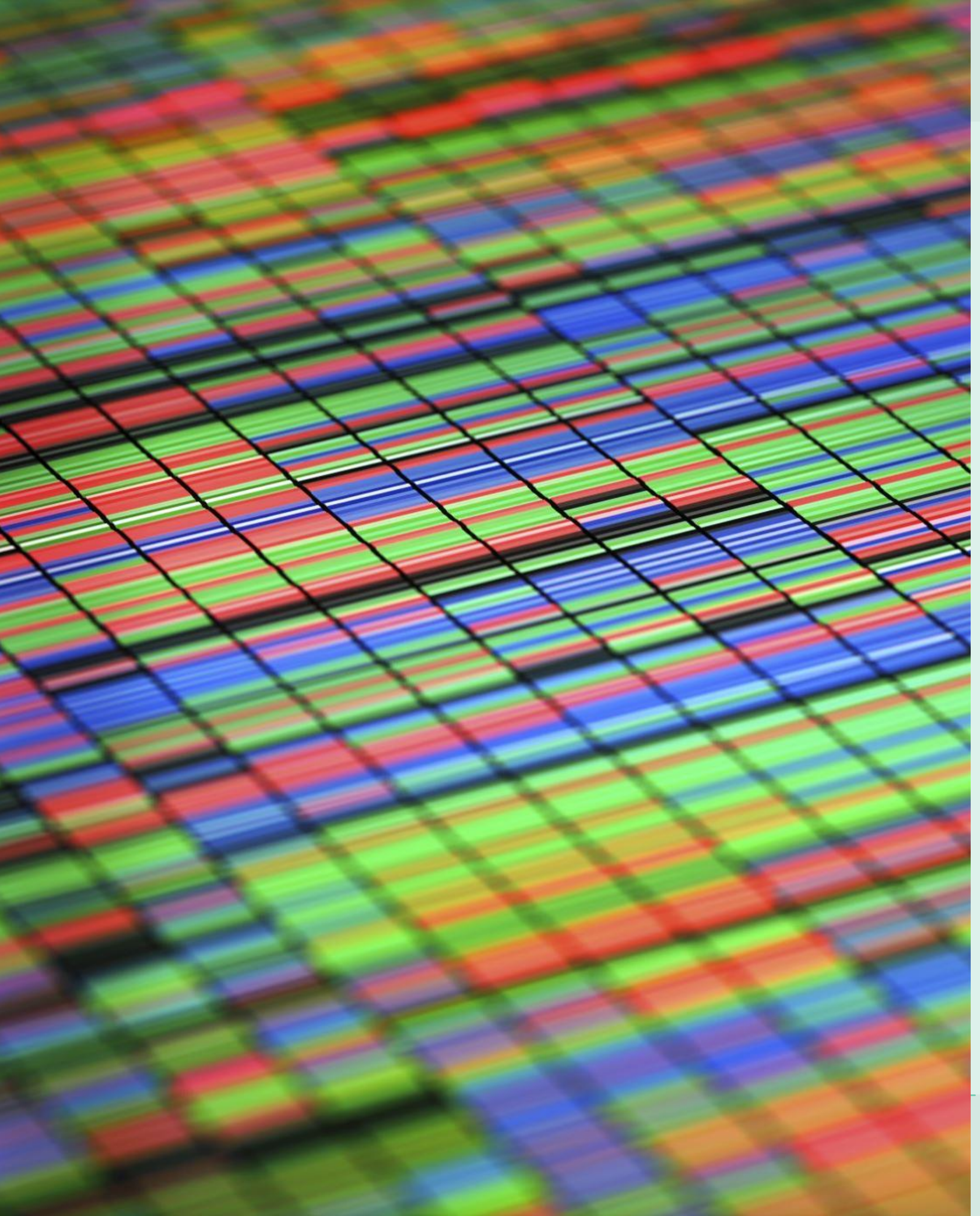


Genotyping & Taxonomy

Background and Strategy

Team F (Group 3)

Background

Genotyping

- Genotyping involves the analysis of sequencing data to identify and characterize genetic variants, including SNPs, indels, and structural variants
- Provides insight into genetic variations within sample populations
- Genotyping is utilized in many fields from molecular epidemiology and tracking the spread of infectious diseases and potential outbreaks to precision medicine and the identification of genetic variants associated with disease susceptibility

Taxonomy

- Taxonomic sequence classifiers are used to classify all input reads by aligning/matching reads information to databases containing comprehensive nucleotide, protein, or whole genome datasets.
- Aggregated per-read classifications can be used to produce taxonomic profiles with relative abundance estimates (often based on read counts).
- Taxonomic classification follows a hierarchical system, organizing organisms into groups based on shared characteristics such as: domains, kingdoms, phyla, classes, orders, etc.
- This classification can be used to infer evolutionary relationships (phylogenetic trees).

Genotyping

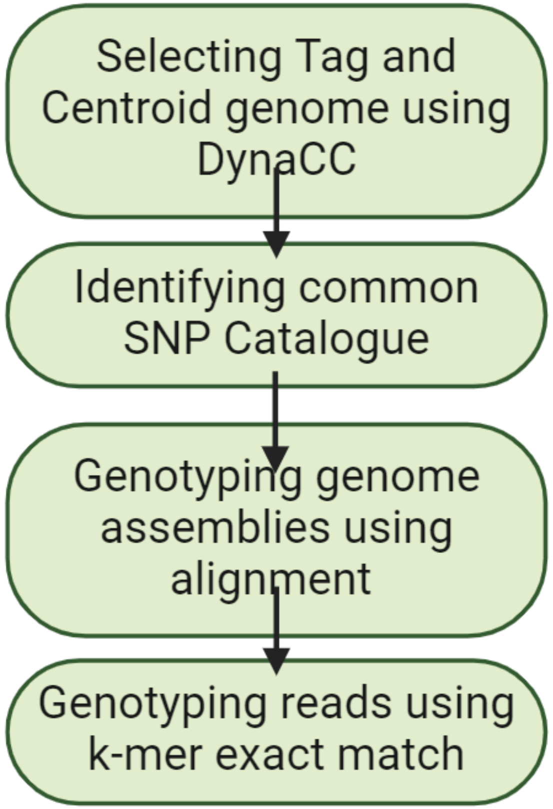
MLST

- Initially developed in the late 1990's, MLST was designed for typing bacterial strains
- Modern MLST utilizes PubMLST, a database of genomic bacterial sequences
- Looks at variation in highly conserved housekeeping genes to look at bacterial sequence types (ST)
- Pros:
    - High discriminatory power for species and ST's identification
    - Serves as a standardized, highly recognized method
- Cons:
    - May lack the resolution to discriminate between closely related strains
    - Dependent on database which cannot always be up to date on latest genetic variations

Genotyping

MAAST: (Microbial agile accurate SNP Typer):

- Open-source bioinformatics pipeline that fully automates SNP calling and genotyping for microbial species.
- Written in Python and C++.
- Maast has two components:
- (i) Constructing reference panel of SNPs using a reduced set of non-redundant genomes .
- (ii) Ultra-fast, *in-silico* genotyping of reference SNPs from large scale genome collections.
- Pros: Higher accuracy and faster performance compared to traditional SNP genotyping methods.
- Cons: May miss rare variants due to focus on common SNPs and dependent on the quality and diversity of the input genomic datasets.



Selecting Tag and Centroid genome using DynaCC

Identifying common SNP Catalogue

Genotyping genome assemblies using alignment

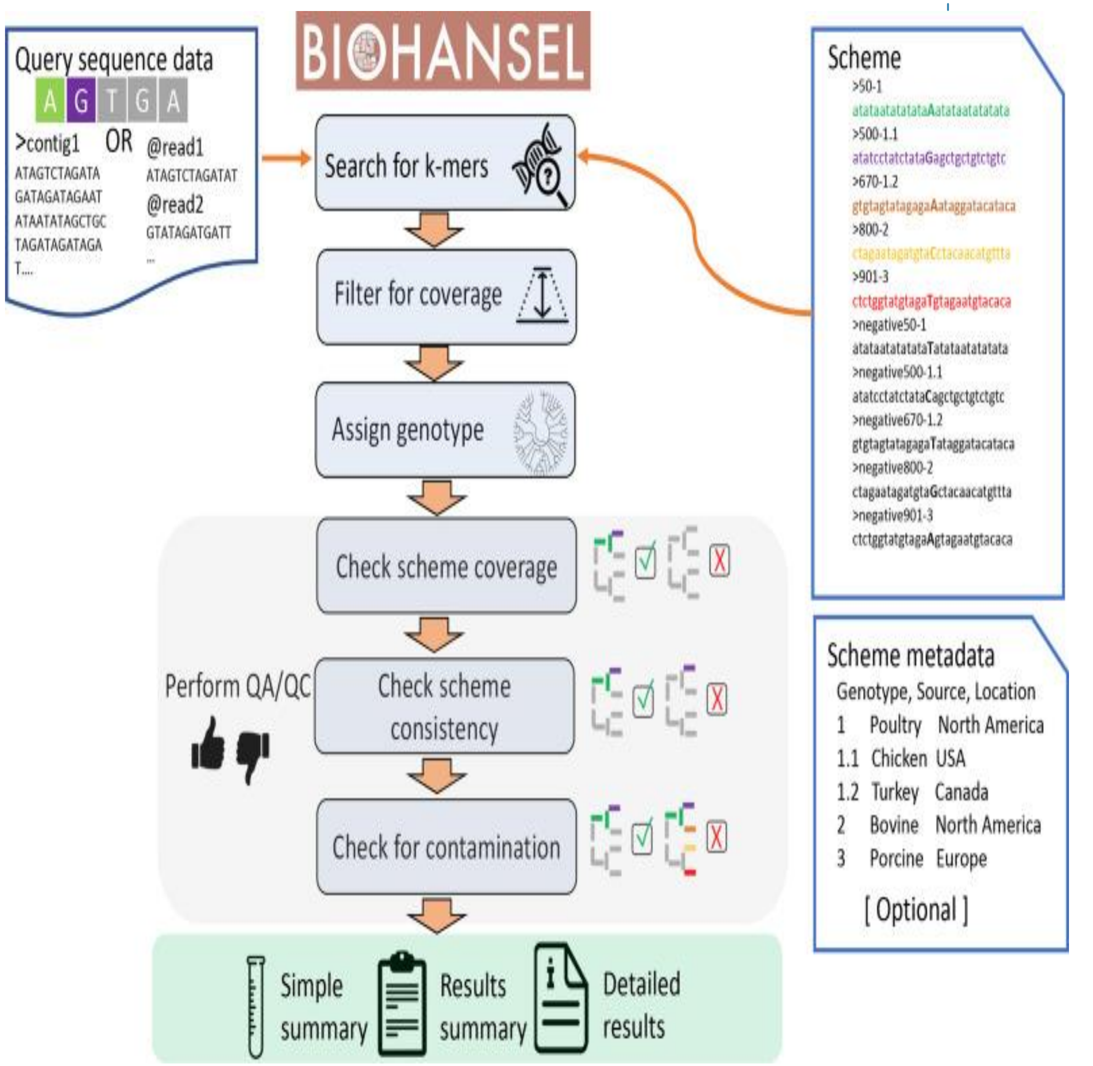
Genotyping reads using k-mer exact match

Maast SNP Genotyping pipeline

Genotyping

BioHansel

- Organism-agnosic, k-mer based genotyping tool for hierarchial SNP schemes.
- Python 3 application available as PyPI, Conda and Galaxy Tool Shed packages.
- Rapidly identifies and types bacterial pathogens in sequencing data using predefined hierarchical SNP schemes for high-resolution genotyping.
- Compatible with both assembled genomes and raw Illumina fastq reads.
- Pros:
  - Rapid genotyping with limited computational resources.
  - Enables detailed contamination detection and quality control.
- Cons:
- Requires well-established, hierarchical SNP schemes for accurate genotyping; might be less suited for highly diverse or recombinant organisms.



BioHansel Genotyping Worflow

Taxonomic Classification

- **Kraken**: Uses k-mer-based classification to rapidly classify metagenomic sequencing reads into taxonomic categories by comparing to a pre-built database of k-mers.
  - **Pros**: Fast classification, suitable for large datasets.
  - **Cons**: Requires a large database, may be less accurate for novel or divergent sequences.
- **MEGAN**: Allows taxonomic analysis of metagenomic data by comparing reads to a database of reference sequences.
  - **Pros**: Can provide more detailed taxonomic information.
  - **Cons**: Slower than k-mer-based methods.
- **Centrifuge**: creates significantly smaller databases based on an FM-index and compression of within-species genomes.
  - **Pros**: Faster classification, reduced database size.
  - **Cons**: Requires more computational resources for database creation.

Genome Quality Assessment

- **Whole Assembly Assessment**
  - Is it of the expected or target species?
  - How complete is the genome, and how contaminated is it?
- **Contig-by-Contig Assessment**
  - If there is contamination, which contigs come from other species?
- **Intra-contig Assessment**
  - Is there evidence of mis-assemblies or co-assemblies in the contigs of the assembly and which strains/species are present?

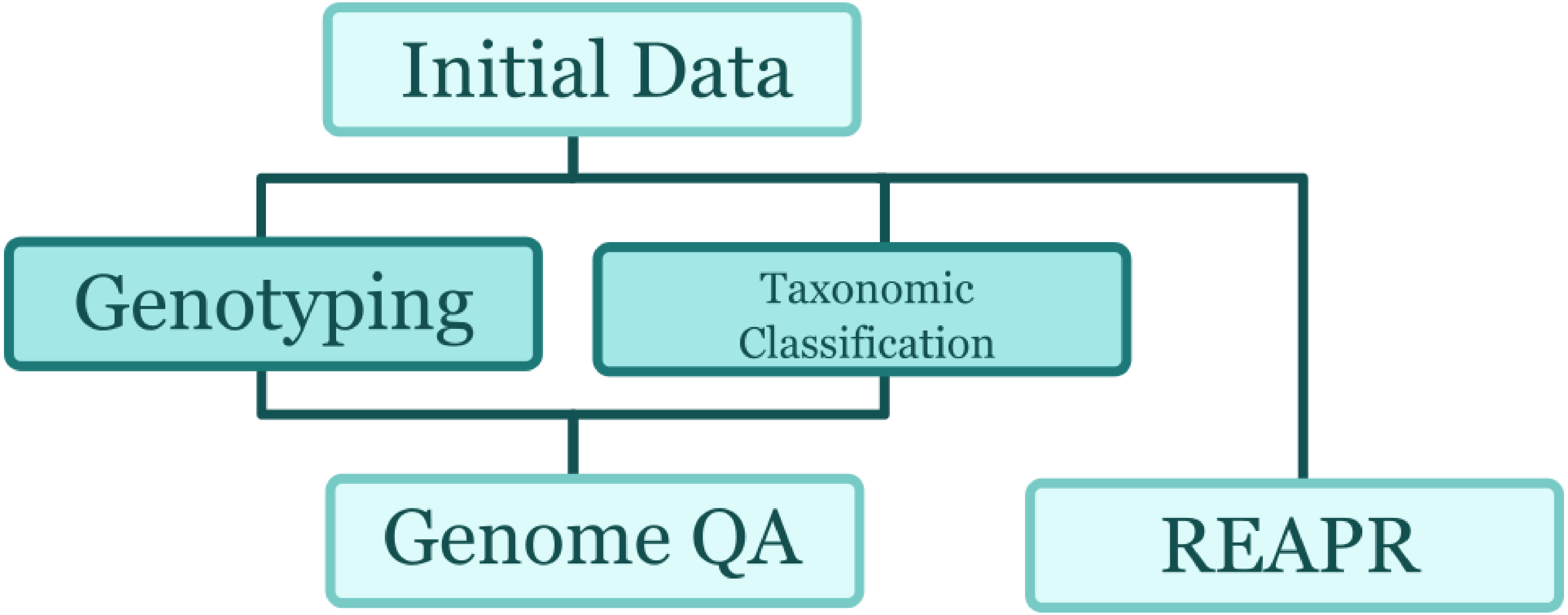
Genome Quality Assessment

- Compare three whole assembly assessment
- **CheckM**
  - Designed in 2015
  - Pros: Use lineage-specific marker genes provide refined estimates of genome completeness and contamination compared to universal or domain-level marker genes
- **REAPR**
  - Designed in 2013 as a tool to assess the quality of genome assemblies
  - Analyzes the alignment of paired-end reads to the assembly and identifies regions with potential errors or mis-assemblies
  - Pro:
    - Identifies errors in genome assemblies **without** the need for a reference sequence
    - Analyzes mis-assemblies which are often overlooked in other Genome QA tools
- **QUAST**
  - Quality Assessment Tool for Genome Assemblies, designed in 2013
  - Pros: Evaluate assembly quality with/without a reference genome
  - Pros: Most time-consuming steps are parallelized → Effectively run on multi-core processors

Genome Quality Assessment

- **Blastn**
  - Useful for contig-by-contig assessment
  - Helps with identifying if specific contigs are classified as different option
  - Compares one or more nucleotide query sequences to a subject nucleotide sequence or a database of nucleotide sequences
- **Alvis**
  - Developed in 2021
  - Tools for visualizing alignments of long reads and assemblies from a wide range of different input file types
  - Provide visualization for alignment and genome coverage

Workflow



Initial Data

Genotyping

Taxonomic Classification

Genome QA

REAPR

Internal Scheduling and Deadlines

Genotyping

- Charith - MLST
- Anirudh - Hound
- Anirudh - MAAST

Taxonomy

- Ramya - Kraken
- Yeojin - MEGAN

Genome QA

- Charith - REAPR
- Yeojin - CheckM
- Mehak - QUAST
- Ramya - BLASTn
- Mehak - Alvis

Internal Scheduling and Deadlines

- 3/11 - Meeting: Goal – Assign packages and deadlines
- 3/15 - Meeting: Goal – Early check in and pre-Spring planning / progress update
- 3/22 - Deadline: Taxonomy results
- 3/25 - Meeting: Goal – Post Spring Break progress update
- 3/28 - Deadline: Genotyping and Genome QA
- 3/30 - Meeting: Final Data Deadline

References

- Jolley, K.A., Maiden, M.C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11, 595 (2010).
- Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3), R46.
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome research, 17(3), 377-386.
- Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome research, 26(12), 1721-1729.
- Portik, D.M., Brown, C.T. & Pierce-Ward, N.T. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics* 23, 541 (2022).
- Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 2015 Sep 18;43(16):7762-8. doi: 10.1093/nar/gkv784. Epub 2015 Aug 6. PMID: 26250111; PMCID: PMC4652774.
- Hunt, M., Kikuchi, T., Sanders, M. et al. REAPR: a universal tool for genome assembly evaluation. Genome Biol 14, R47 (2013).
- Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019 Jul 26;7:e7359. doi: 10.7717/peerj.7359. PMID: 31388474; PMCID: PMC6662567.
- Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013) 29 (8): 1072-1075. doi: 10.1093/bioinformatics/btt086 First published online: February 19, 2013
- Labbé, G., Kruczkiewicz, P., Robertson, J., Mabon, P., Schonfeld, J., Kein, D., Rankin, M. A., Gopez, M., Hole, D., Son, D., Knox, N., Laing, C. R., Bessonov, K., Taboada, E. N., Yoshida, C., Ziebell, K., Nichani, A., Johnson, R. P., Van Domselaar, G., & Nash, J. H. E. (2021). Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel. *Microbial genomics*, *7*(9), 000651. https://doi.org/10.1099/mgen.0.000651