Developing Tools for Rapid and Accurate Post-Sequencing Analysis of Foodborne Pathogens

Mitchell Holland, Noblis

Agenda

- Introduction
- Whole Genome Sequencing
  - Analysis Pipeline
  - Sequence Alignment
  - SNPs and Phylogenetic Trees
  - Current Challenges
- BioVelocity – A high-speed sequence alignment platform
  - Capabilities and Advantages
  - Integration into WGS pipeline
  - Application to Foodborne Disease Outbreaks
- Conclusions

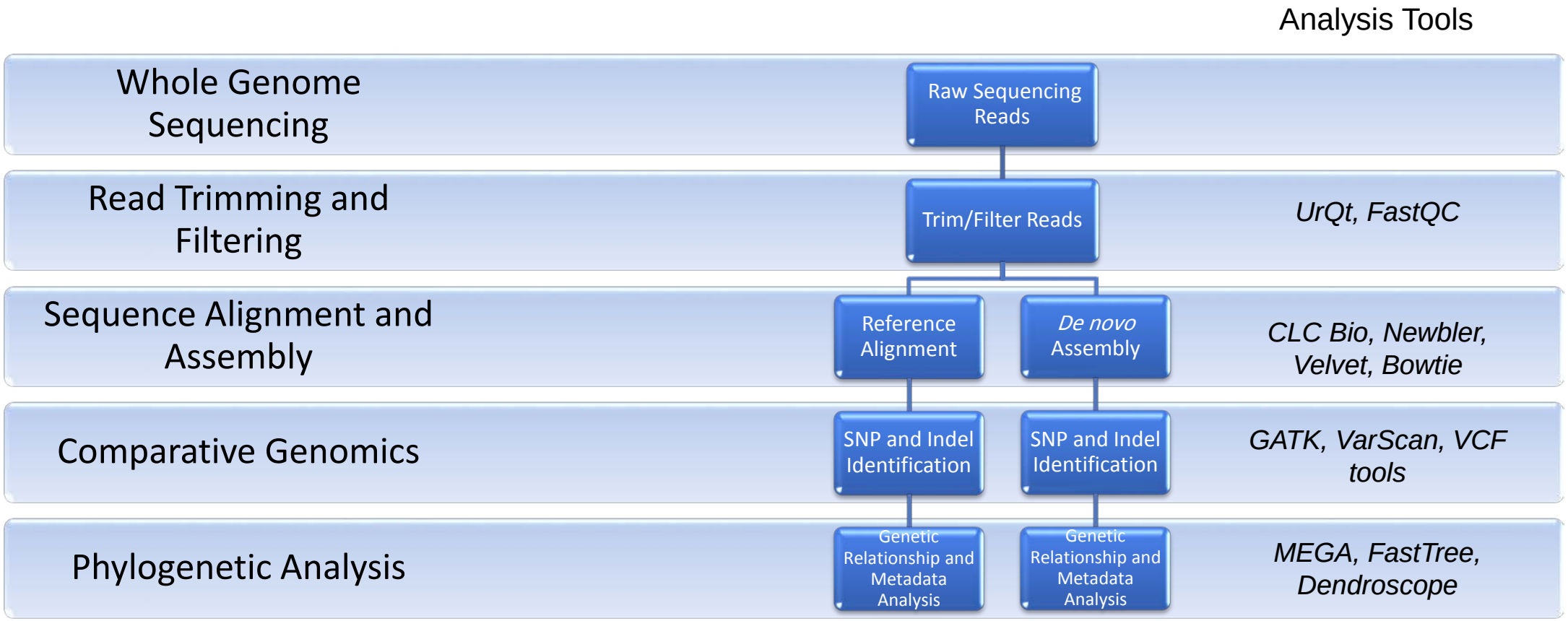
Introduction

- Whole Genome Sequencing (WGS) is capable of generating a wealth of data and is becoming cheaper and more readily available to industries outside of academia

- While many bioinformatics tools have been developed to address the needs of analyzing this data, time to process data remains the rate limiting step

To make progress we have to speed up the process!

WGS Analysis Pipeline

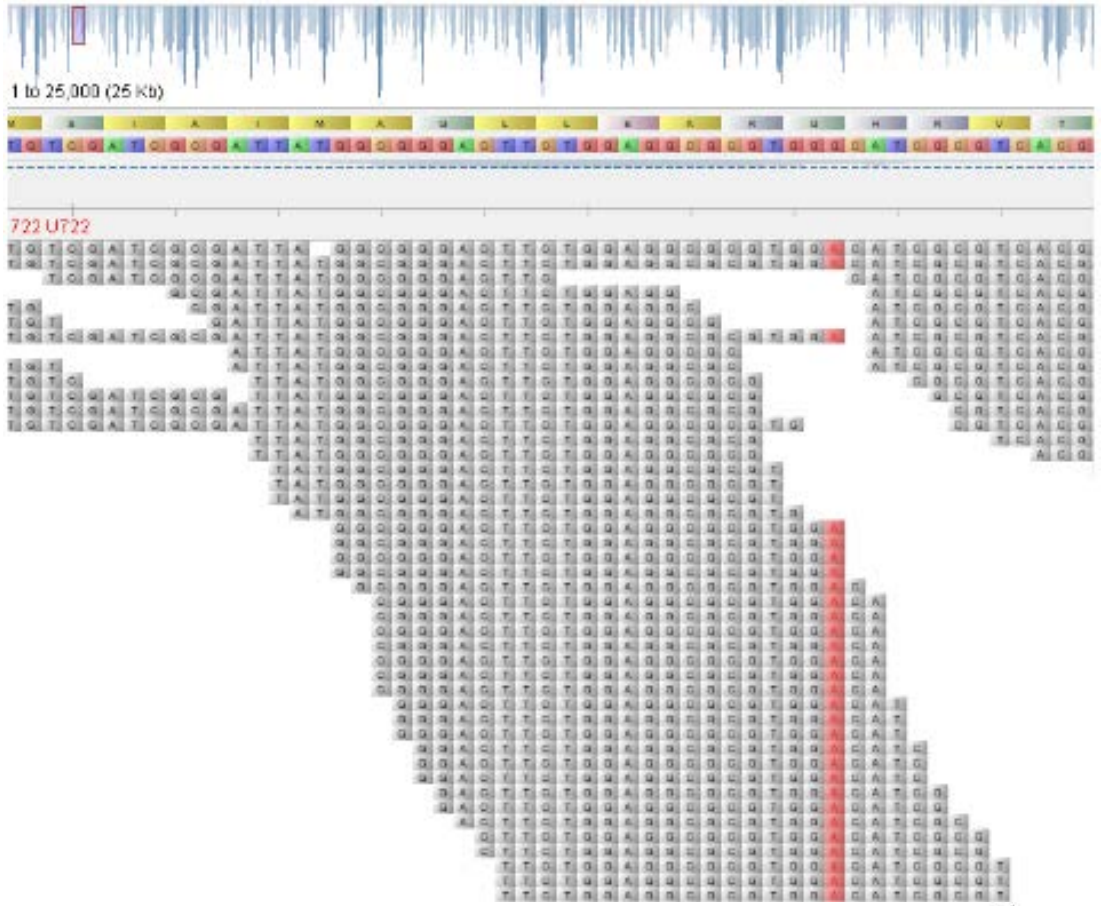


Sequence Alignment / Assembly

- The process of aligning and combining small sequence fragments (reads) to reconstruct the original sequence

- Two types:
  - **Reference-assisted**: Comparing the reads against a known reference genome
  - ***De novo***: Aligning the reads together into contigs without a known reference to use as a guide

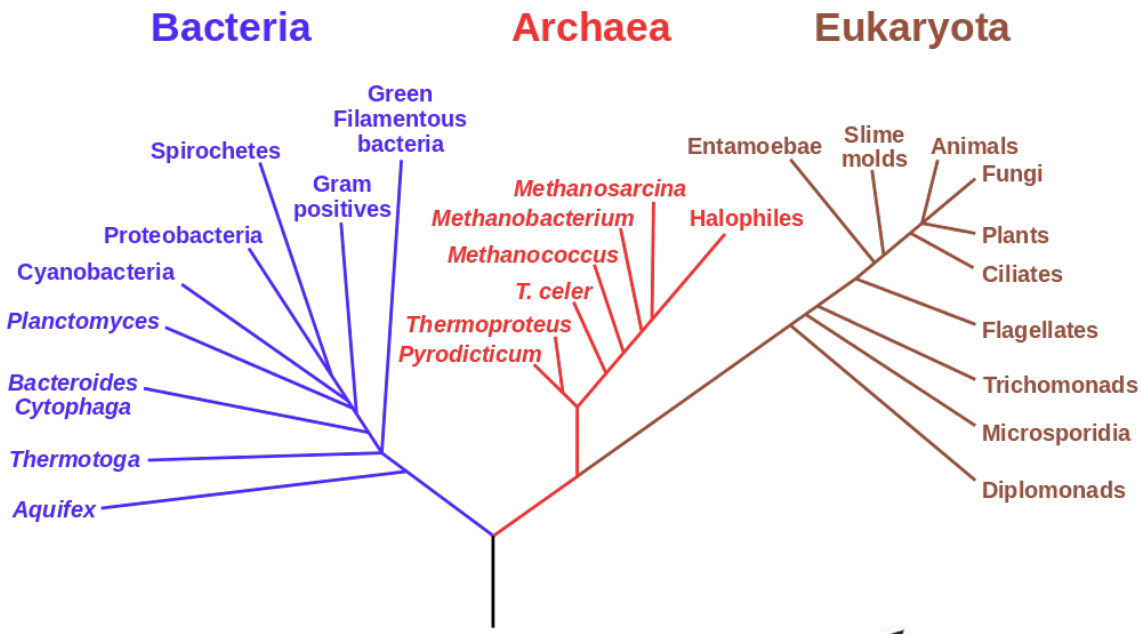
- Challenges:
  - Computationally intensive
  - Difficult to get a good assembly without the correct reference organism



SNP Identification and Phylogenetic Trees

- Single Nucleotide Polymorphism (SNP) Identification
  - Find the genetic differences between your samples and reference organisms
  - SNPs are reported when a single nucleotide position varies by a significant threshold of agreement and with sufficient depth of coverage

- Phylogenetic Tree Construction
  - Use the discovered SNPs to construct a tree showing the inferred evolutionary relationships between your samples
  - This tree will indicate the likely lineage of samples so that an outbreak can be traced back to its source



Challenges in WGS Analysis

- Processing Time
  - The input data files from next-generation sequencing machines contain millions of reads and are gigabytes in size
  - Aligning a read set to many references using traditional tools can take days
- Abundance of Tools, Techniques, and File Formats
  - Hundreds of COTS products and open-source programs to choose from
  - Can be difficult or time consuming to transfer data between the tools
  - May not be compatible and will require retooling
- Accurate Organism Detection
  - Detection is dependent on having the appropriate reference organism in your database
  - Closely related strains can be difficult to distinguish
  - Using the largest possible reference set increases your odds of finding the right match

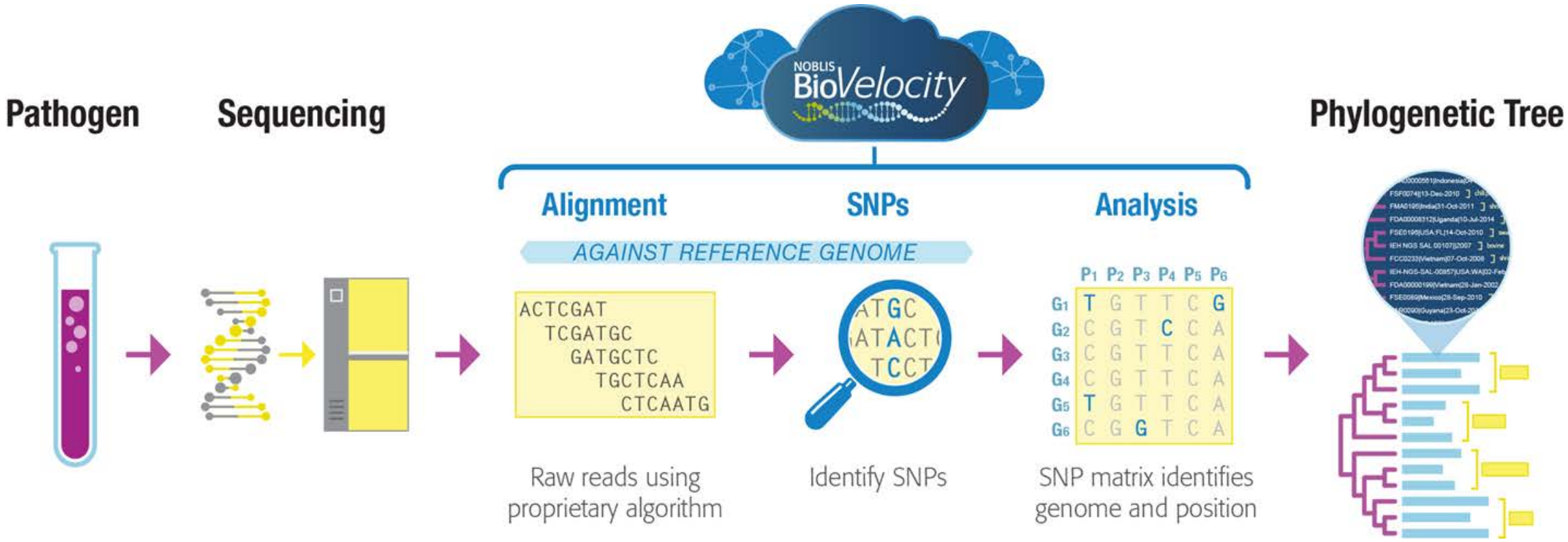
BioVelocity: A Post Sequencing Processing and Analysis Platform

- BioVelocity runs natively on a CRAY-XMT2 supercomputer and uses a unique hashing algorithm for fast sample identification and SNP detection
  - BioVelocity uses a brute force index, built out of all possible base pair sequences of various k-mer lengths

- Capabilities:
  - Alignment of WGS samples to a large public library (e.g., NCBI) of reference genomes (thousands) for strain identification
  - SNP detection: Identify potentially significant evolutionary changes as a matrix of SNPs for comparison
  - Metagenomics analysis: Detecting multiple organisms in a single sample

- Advantages:
  - No need for sequence assembly
  - Simultaneous reference alignment means the job only needs to be run once
  - SNP profiles are output for every genome in your reference
  - BioVelocity, while built initially for the CRAY-XMT2 can run on a variety of architectures, including:
    - TORQUE Cluster
    - IBM POWER8

Sequence Analysis Pipeline Using BioVelocity



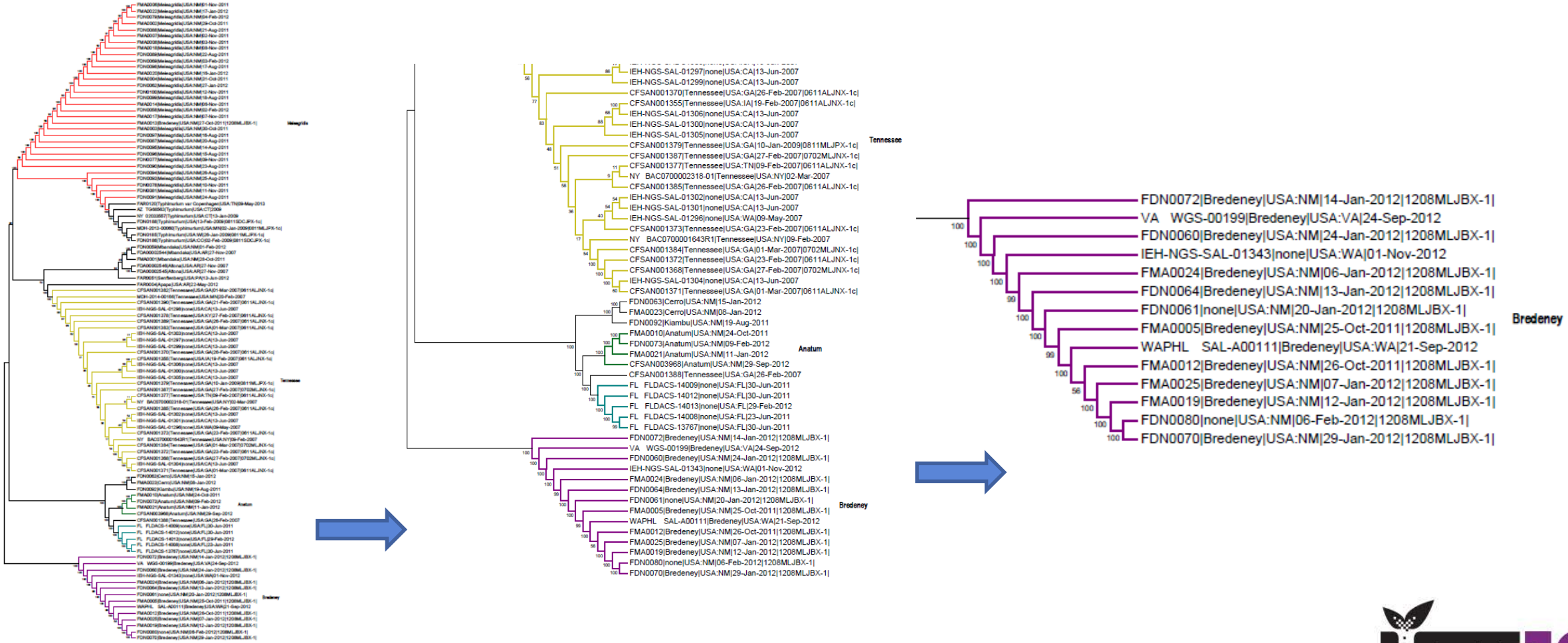
Application to Foodborne Disease Outbreaks

- In 2012, a nationwide outbreak of *Salmonella Bredeney* occurred stemming from Valencia peanut butter products

- Noblis used BioVelocity to analyze 103 samples from NCBI's Genome Trakr SRA database. These were all:
  - *Salmonella enterica* species
  - Isolated from peanut butter
  - Collected since 2007

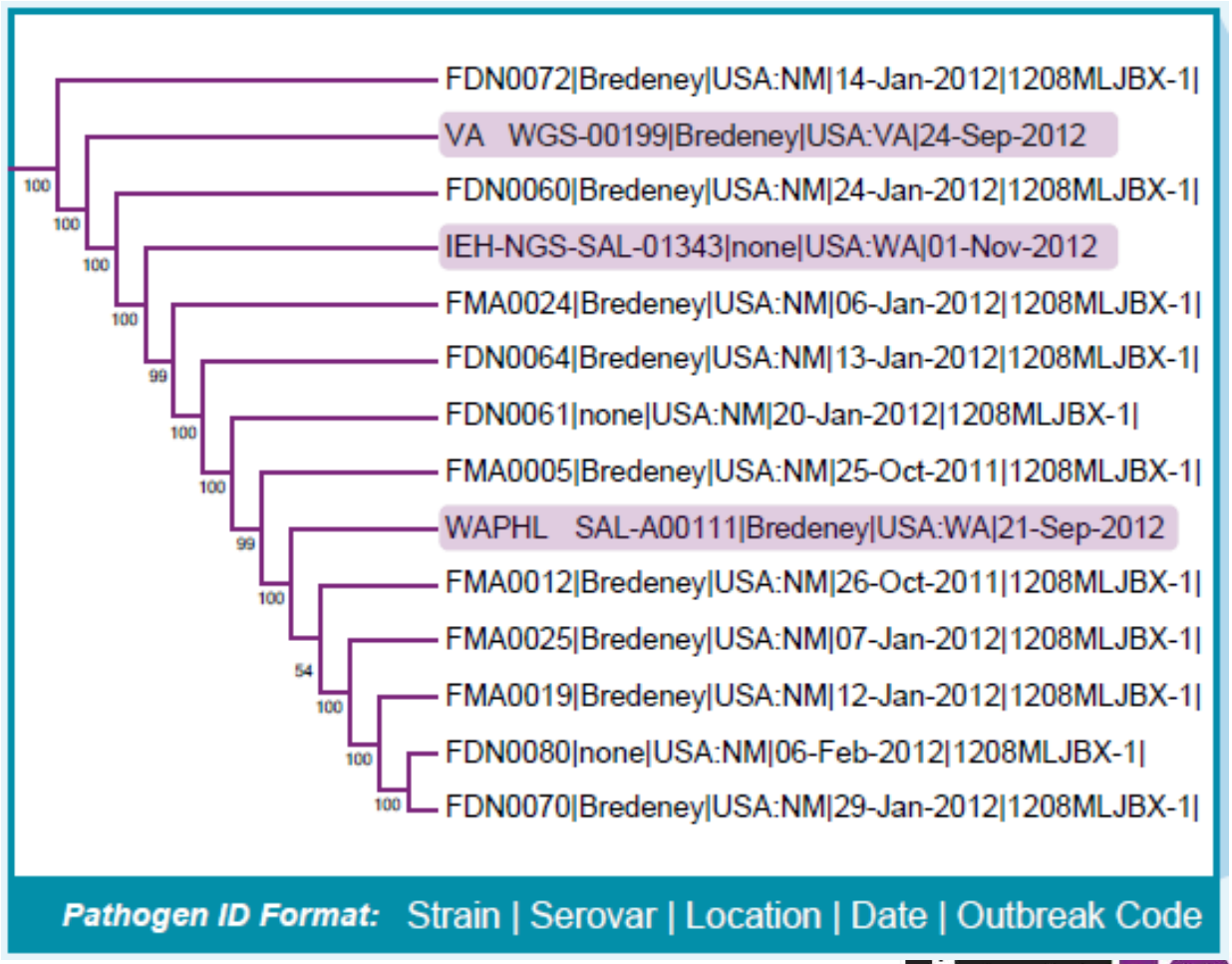
- These samples were all simultaneously aligned to *Salmonella enterica subsp. enterica serovar Typhimurium str. LT2,* a representative *Salmonella* genome
  - A phylogenetic tree was constructed using the resulting SNP matrix

Phylogenetic Tree with 103 *Salmonella* samples



Application to Foodborne Disease Outbreaks

- The resulting phylogenetic tree shows clades for Meleagridis, Tennessee, Anatum, and Bredeney serovars
- The Bredeney serovar was a distinct clade for the 2012 outbreak
- This clade included 3 unassociated samples which can be inferred to be highly related to the outbreak strain based on the location and date of collection included in the metadata



Summary / Key Takeaways

- Current capabilities of WGS for food safety:
  - Better precision: discriminate organisms to the strain level
  - Enhance safety practices: source tracing of contaminations
  - Address gaps: make inferences when the data is incomplete

- Additional applications:
  - Metagenomics
    - Comprehensive assessment of bacterial population
  - Augment traditional identification methods to identify:
    - Difficult/slow growing organisms
    - Phenotypically challenging organisms

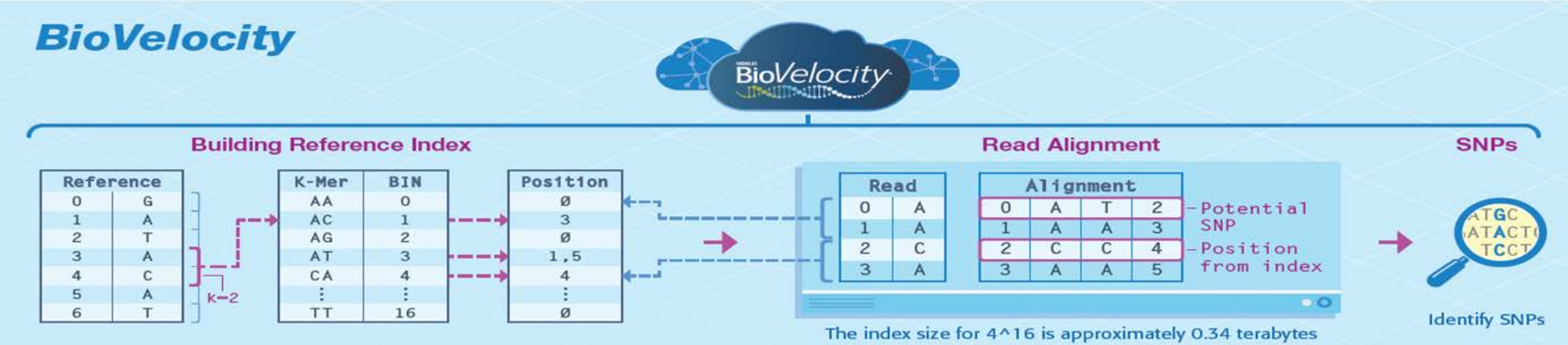
Mitchell Holland
Mitchell.Holland@noblis.org

Jane Tang, Ph.D.
Jane.Tang@noblis.org

Masooda Omari
Masooda.Omari@noblis.org

BACKUP SLIDES

Algorithm Detailed



- BioVelocity is a proprietary algorithm developed by Noblis, optimized for use on a Cray XMT-2 supercomputer, for rapid and accurate processing of complex genome sequences.
- Post-sequence analysis starts once the genome sequences have been downloaded and converted to FASTA file format; assembly or annotation is not required.

- BioVelocity utilizes k-mers for its indexing structure to make alignment rapid and comprehensive
  - Calculates all possible permutations of nucleotides of 16-mers, or 4^{16}, and each is assigned a value.
  - Index structure allows binning of sequence reads to make analysis efficient, including reads that do not align with reference genomes, resulting in a complete analysis.
- First an index array is generated containing a unique bin for each k-mer permutation
- Next the reference is scanned using a sliding window of size k, shifting over each time by 1 position
- Each scan adds the position of the current k-mer to a list at the corresponding index bin

- Then the currnet read is scanned using a sliding window of size k, shifting over each time by k positions
- BioVelocity finds occurrences of the current k-mer using the index
  - If the k-mer exists, then perform a simple alignment at the corresponding position of the genome
    - If the alignment passes, then accept it for subsequent SNP determination and process the next read
    - Otherwise, check a few other positions (if any)
  - Otherwise, check the next scanned k-mer

BioVelocity Configuration

Currently Available Analysis/Jobs

- Standard alignment + FOGSAA/Needleman for gaps
- Metagenomic analysis
- SNP detection
- Conserved sequence detection
- Signature sequence detection
- Read compression

Inputs

- One or more fasta/fastq read files
- Configuration settings (thresholds etc.)
- Type of analysis to be run
- Index containing one or more reference genome(s)

Outputs

- (S)[1] Variant Call Format (VCF)
- (S) Sequence Alignment/Map (SAM)
- (P)[2] Meta-genomic analysis
- (P) Conserved/signature sequences
- (P) Compressed reads

[1] Industry standard format = (S)
[2] Proprietary format = (P)