The Basics of Understanding Whole Genome Next Generation Sequence Data

Heather Carleton, MPH, Ph.D.

ASM-CDC Infectious Disease and Public Health Microbiology Postdoctoral Fellow

PulseNet USA Next Generation Subtyping Unit

NCEZID/DFWED/EDLB

June 2nd, 2014

APHL 2014

National Center for Emerging and Zoonotic Infectious Diseases

Division of Foodborne, Waterborne and Environmental Diseases

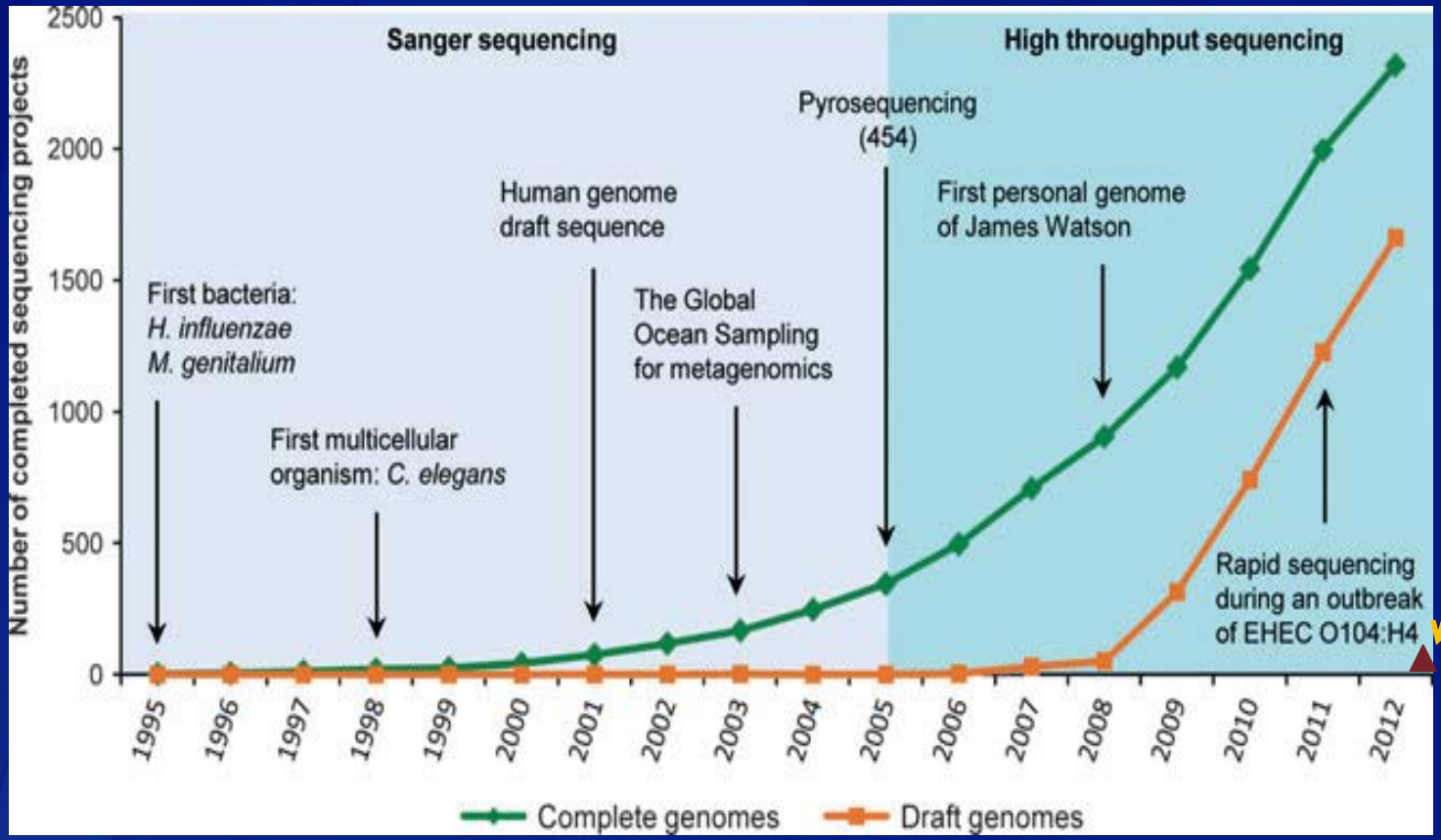
Objectives

- Provide a basic overview of the terminology surrounding whole genome sequence (WGS) data
- Explain ways to analyze WGS data to characterize isolates
- PulseNet vision for implementing whole genome sequencing

The evolution of whole genome sequencing

- First generation – dye terminator (Sanger) sequencing
  - ABI 3730xl
- Second generation – Massive parallel sequencing by synthesis
  - Roche 454, GS Junior - pyrosequencing
  - Illumina GAIIx, Hiseq, Miseq, NextSeq – synthesis with reversible terminators
  - IonTorrent PGM, Proton – semiconductor sequencing
- Third generation – Single cell sequencing by synthesis
  - PacBioRS
  - Nanopore

Evolution of whole genome sequencing (cont'd)



Real-time sequencing for outbreak surveillance

Milestones in whole genome sequencing

Bertelli, C. and Greub, G. (2013) Clin. Microbiol. Infect. Epub Apr 24

Next Generation Sequence Data Generation

Leading NGS Benchtop Sequencers

Sequence output →

Millions of reads
Gigabytes sequencing data per run

Ion Torrent PGM

Illumina MiSeq

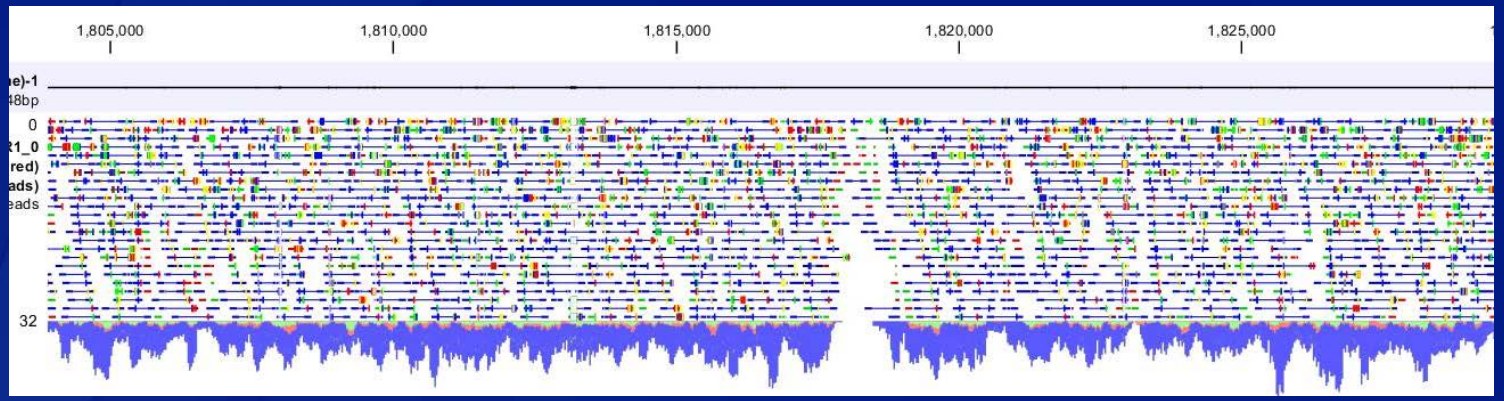
What do you do with it?

Assemble genomes

Whole genome analyses

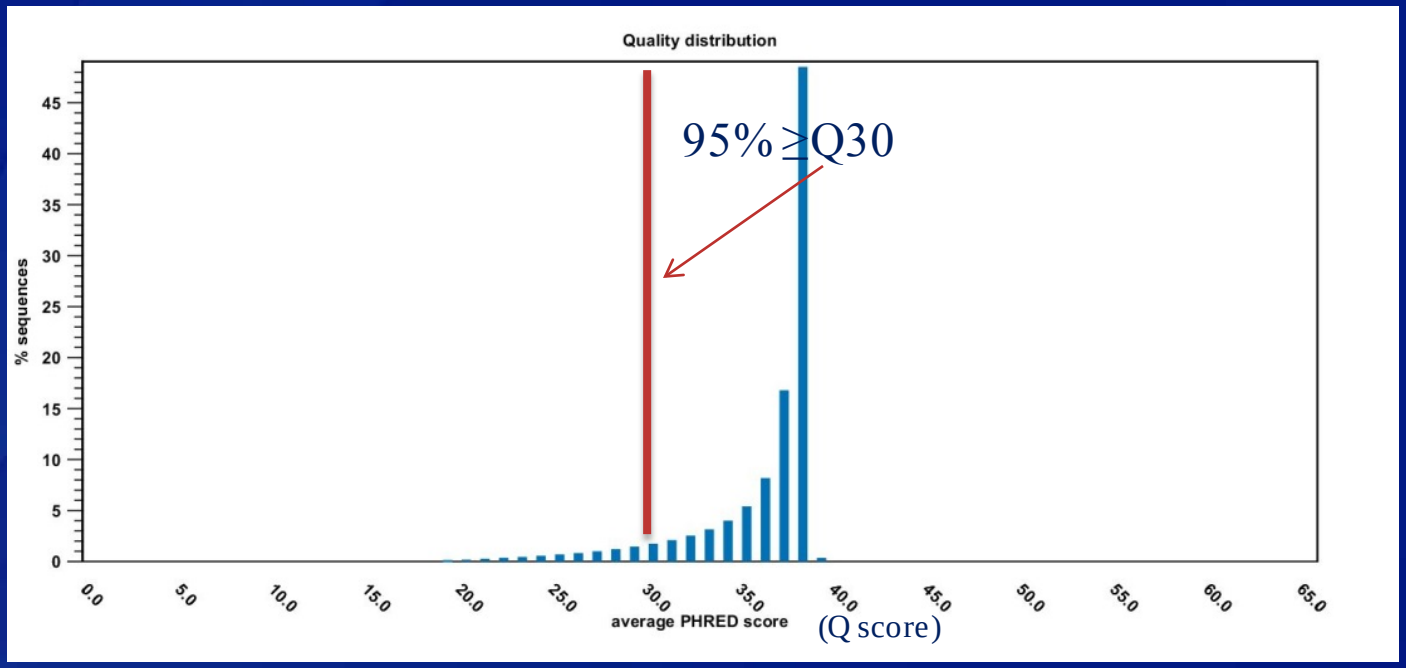
WGS terms: Raw Read

- **Raw Read**
  - Single sequencing output from your NGS machine; length depends on sequencing chemistry
  - Generally 100 thousand – millions of raw reads are generated per isolate sequenced using NGS



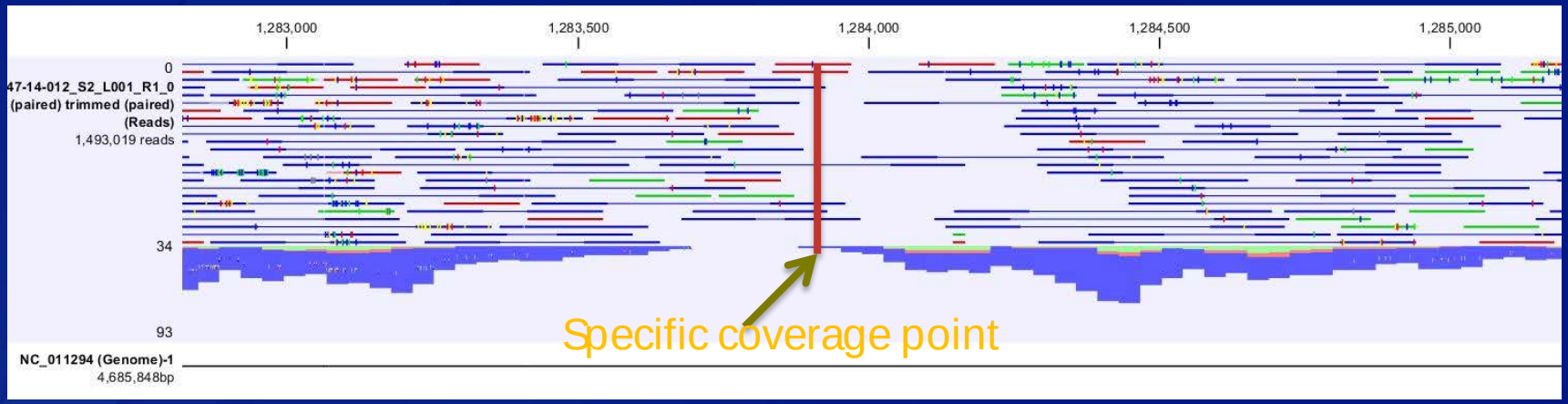
WGS terms: Quality Scores

- **Quality scores**
  - Likelihood the base call is correct
    - Phred – part of fastq file generated from sequencer that scores base call quality
    - Q30 – the percentage of base calls that have a 1 in 1000 chance or less of being incorrect (Q20 – 1 incorrect in 100 base calls)
      - indicates how much usable data you have from a run



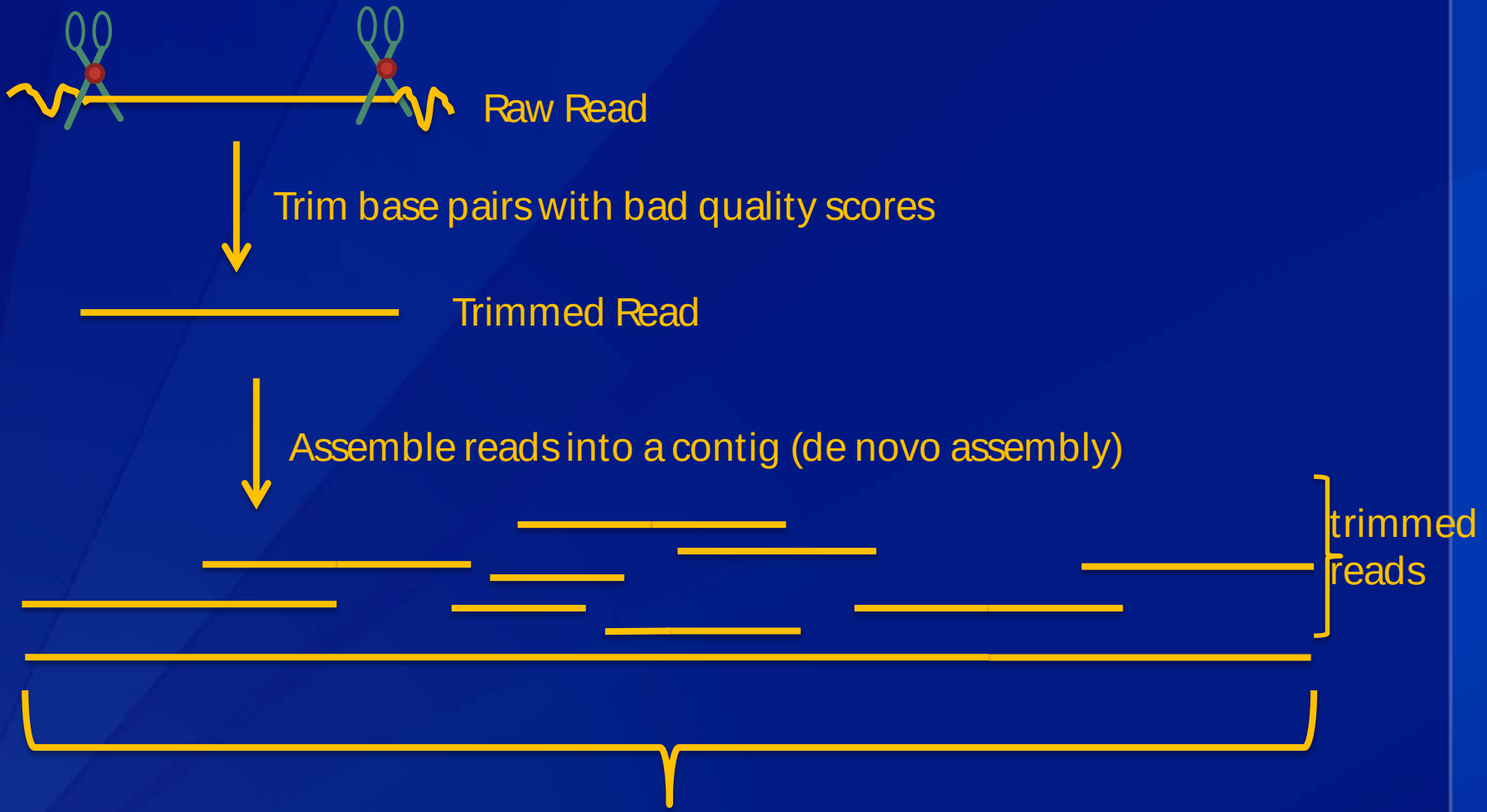
WGS terms: Coverage

- Coverage
  - Average – divide the total # of bases by the genome size (i.e. 156,000,000 (total bases from sequencer)/ 3,000,000 (size of genome = 52x coverage))
  - Specific – how many reads span the 1 base you are looking at



-Average genome coverage of genome example is 41x – coverage at specific coverage point is 7x

WGS terms: de Novo Assembly

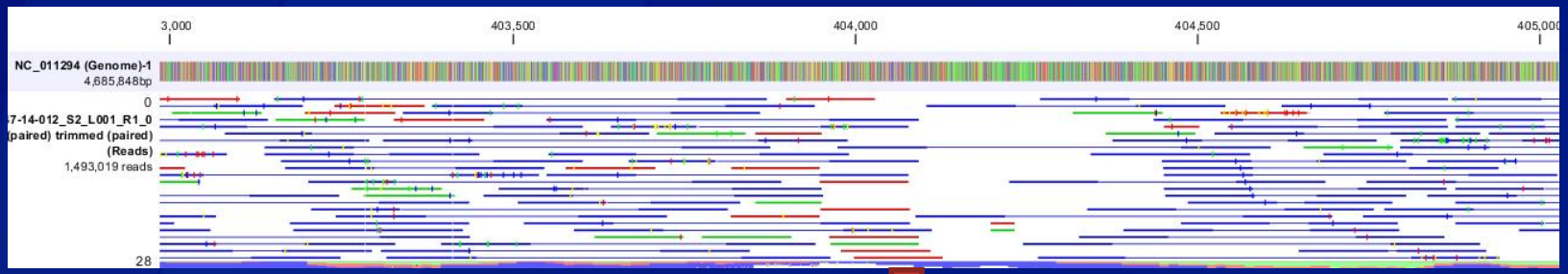


Contig:

Assembly of overlapping reads into a single longer piece of DNA

WGS Terms: Reference-guided Assembly

- Map raw reads to a closely related reference genome



Contigs extracted from read mapping of raw reads

(can set quality and coverage thresholds)

Contig 1

Contig 2

Choosing de Novo versus Reference-guided Assembly

de Novo

-Computationally costly

-Difficult if there are repeat regions

-Assembles genome and plasmids

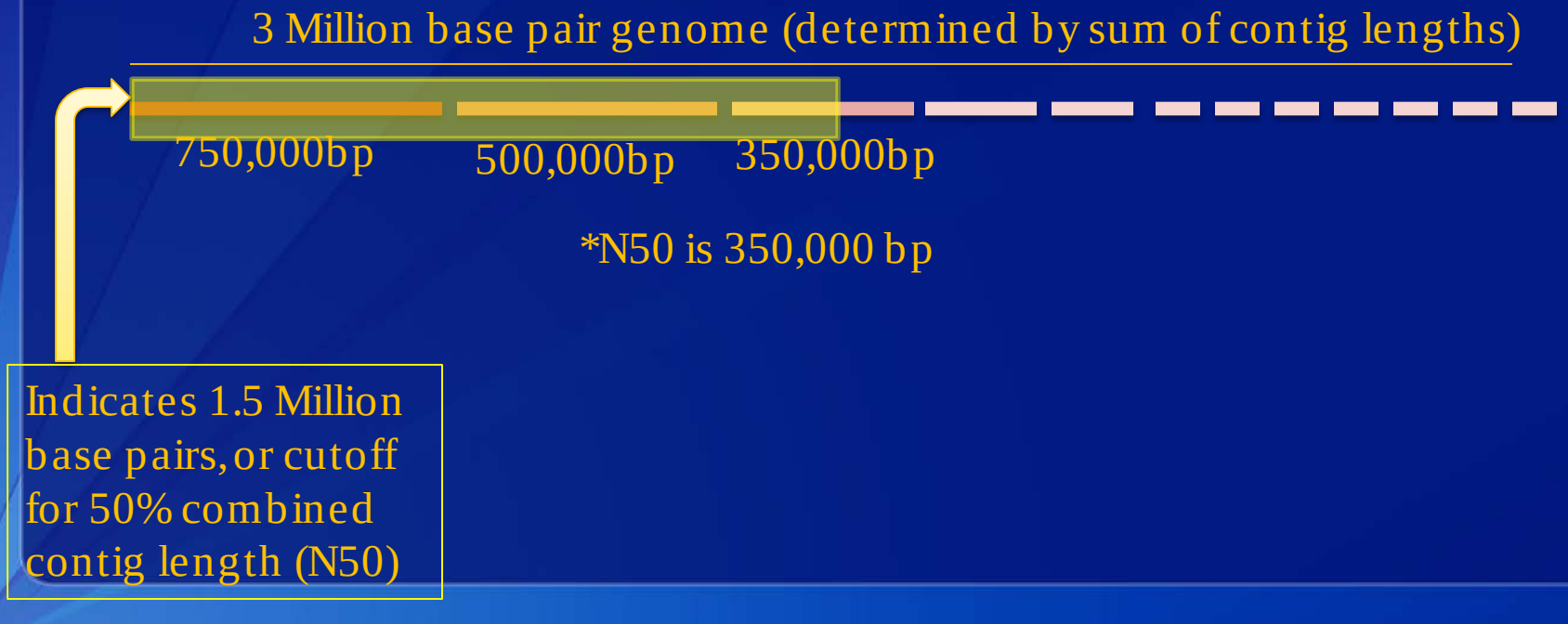
Reference – guided

-Requires closely related good reference genome

-Only assembles reads that match the reference – does not assemble plasmids or insertion elements if there is no reference

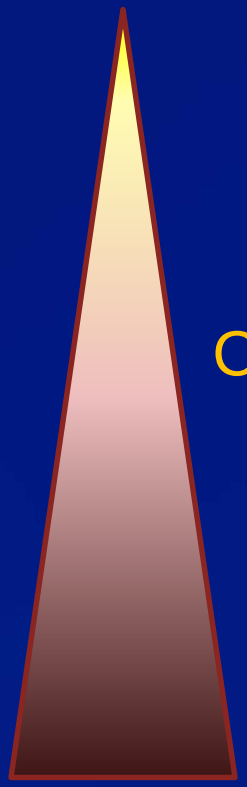
Assessing Assembly Quality

- **Assembly metrics can indicate sequence quality**
  - Number of contigs raw reads assembles into
    - Good: *E coli* <200, *Salmonella* < 100, *Listeria* < 30
  - N50 statistic– Calculated by summarizing the lengths of the biggest contigs until you reach 50% of total combined contig length
    - Good: >200,000 bp



Ways to Analyze WGS data

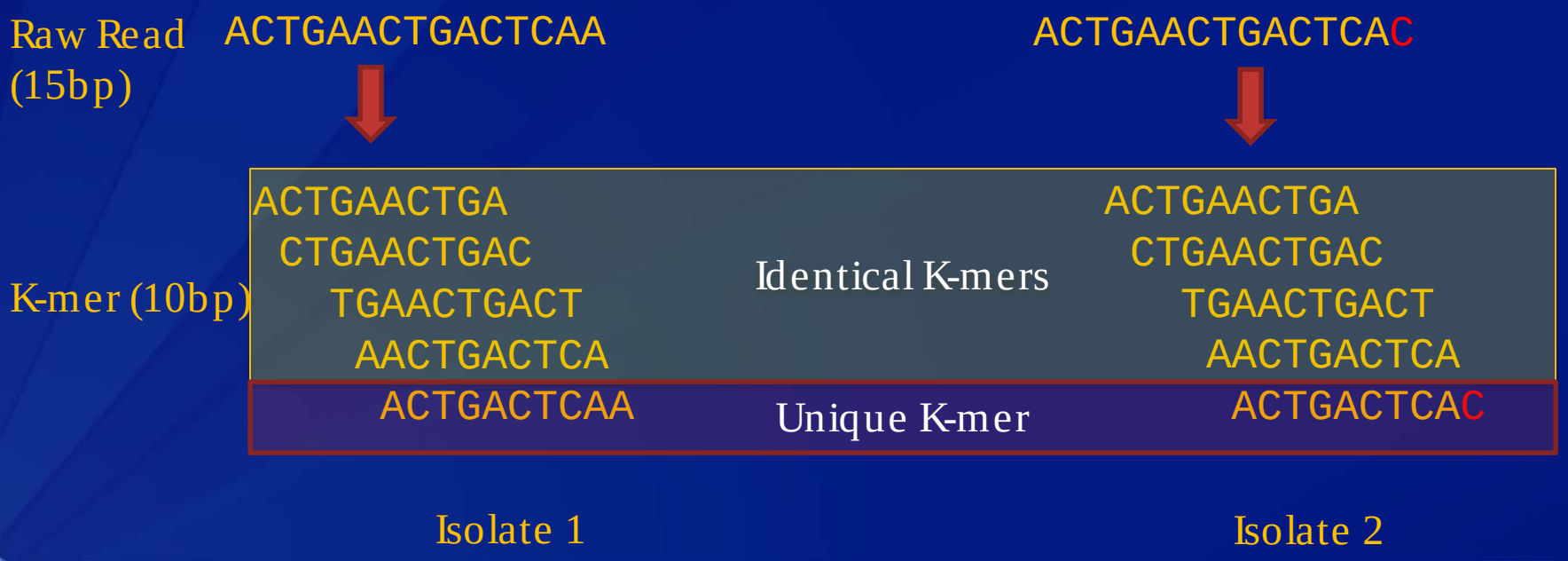
- Kmer analysis
- Whole genome multilocus sequence typing (wgMLST)
- High quality Single Nucleotide Polymorphism (hqSNP) analysis



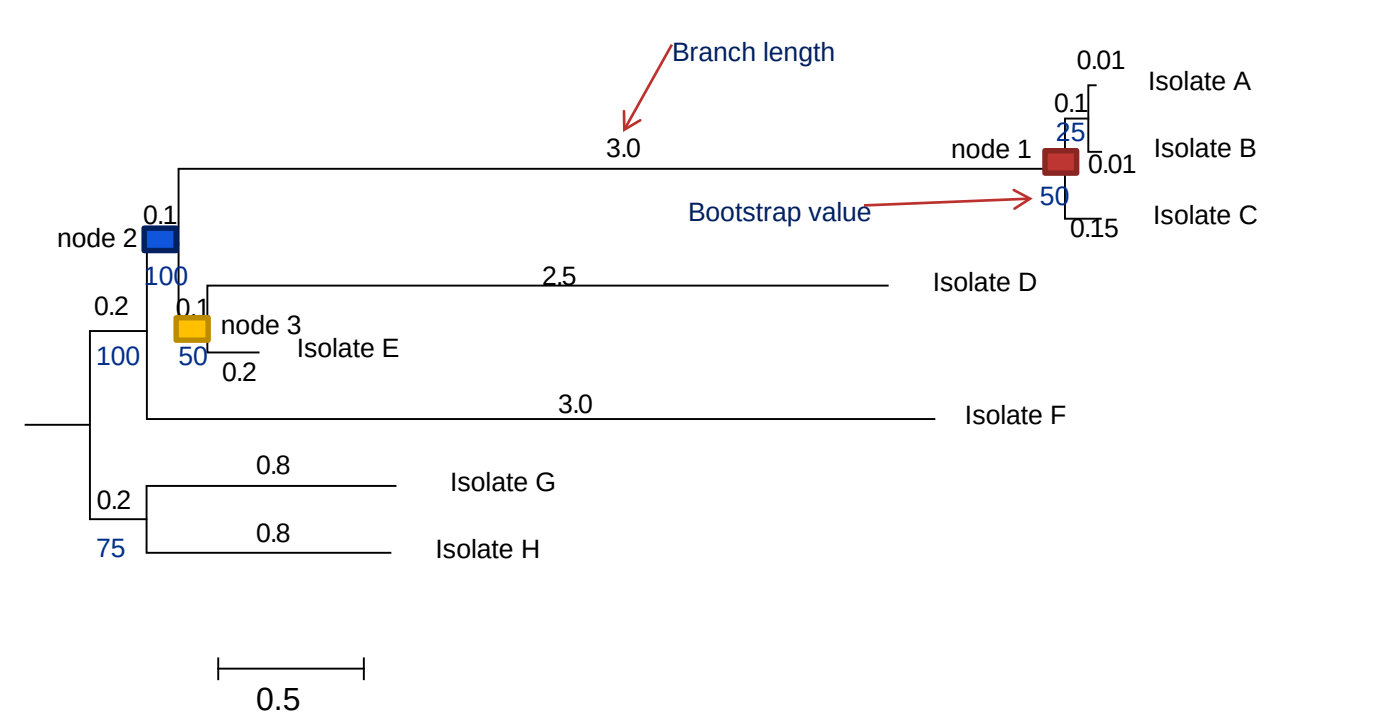
Computational demands

K-mer analysis

- **K-mer :**
  - Computer algorithms use a sliding window to chop up raw reads into shorter lengths (k) of DNA
  - k is determined by which length gives you the best specificity and most adequate resolution
  - Comparing similar and unique kmers gives you a measure of relatedness



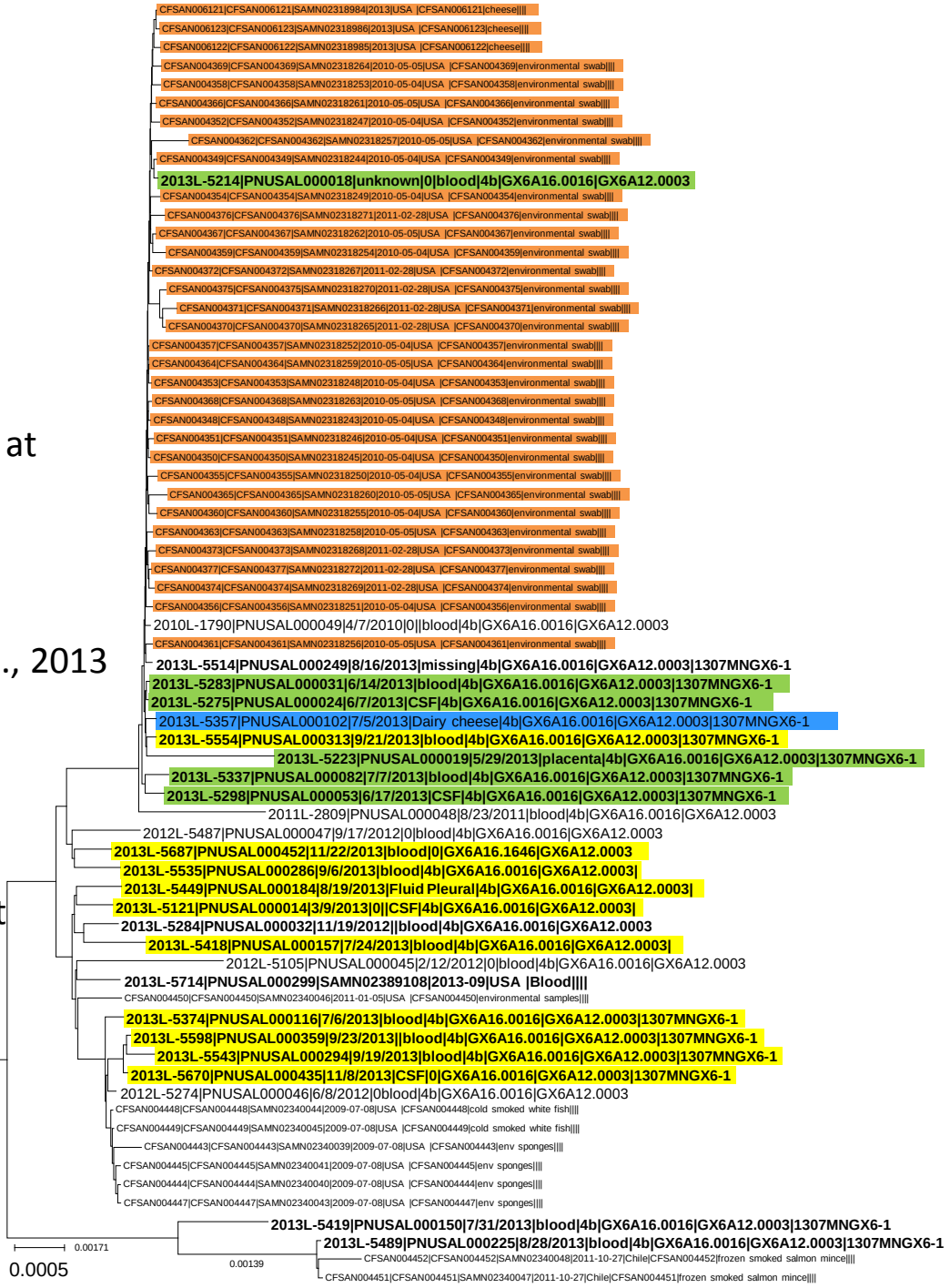
Understanding WGS Data Analysis: Phylogenetic Trees



- Branch length indicates relatedness, shorter horizontal branch length = highly related (isolates in red node 1); longer branch length = less related (yellow node 3)
- Branch length is affected by # of isolates you are comparing as well as relatedness
- Where branches join is referred to as a node, the node indicates a common ancestor (blue node 2), could indicate common transmission source

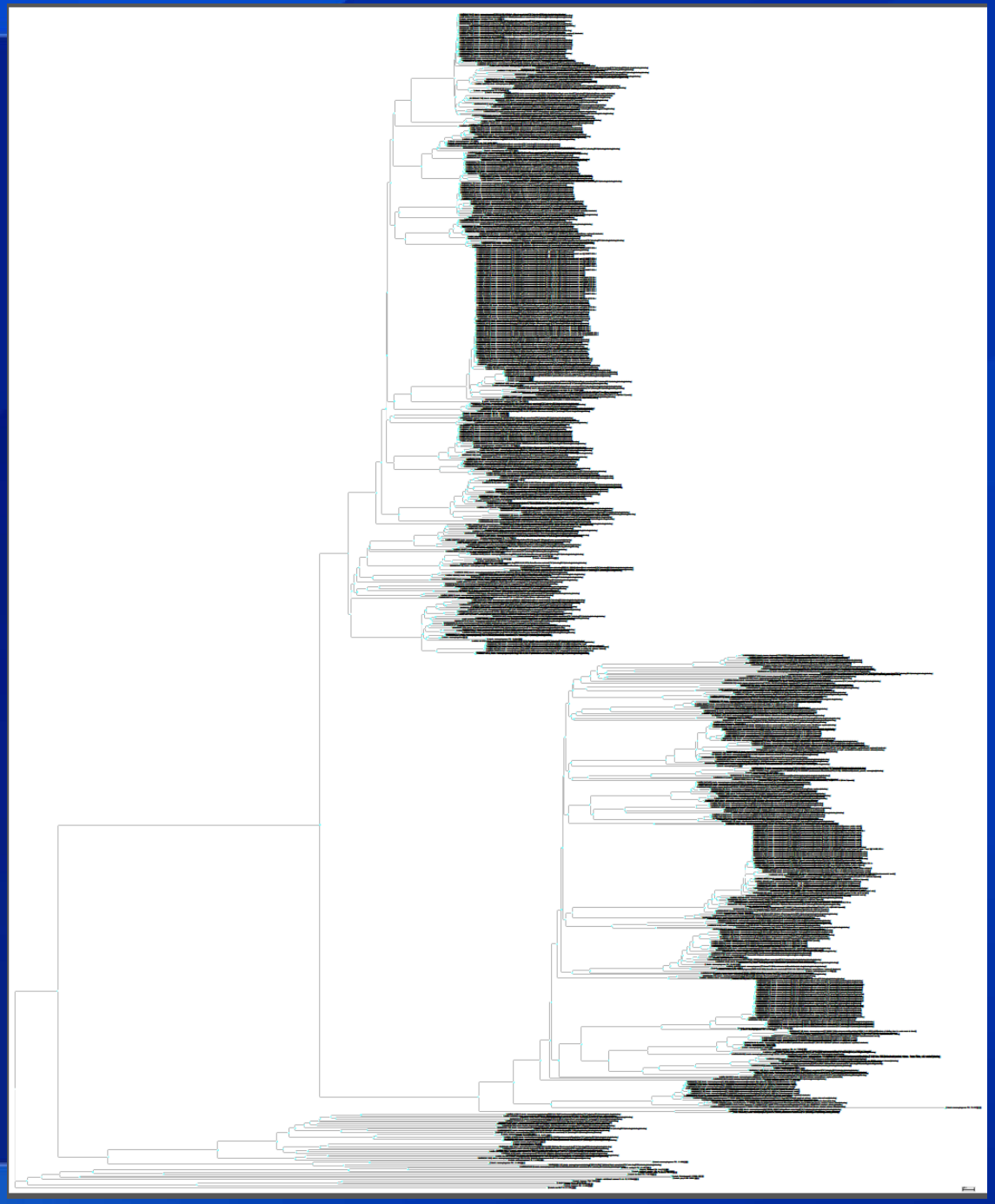
Kmer Tree

- Environmental and food samples that FDA collected at Crave Bros., 2010-2013
- Clinical, Crave Bros. 2013
- Implicated food, Crave Bros., 2013
- New clinical isolates sequenced after closing out the Crave Bros. outbreak



Kmer Tree from NCBI

- As more isolates added to the tree it becomes more difficult to identify clusters



Caveats to K-mer analysis

Advantages:

- Does not require a reference or multiple sequence alignment
- Relatively fast analysis
- Does not require assembly

Disadvantages:

- K-mer analysis does not provide information about where in the genome the differences are
- Does not consider sequence quality*
- Does not provide a true phylogenetic relationship
- Does not lead to strain type nomenclature

SNP Analysis Terms

- **Single Nucleotide Polymorphism (SNP)**

  ATGTT**C**CTC sequence

  ATGTT**G**CTC reference

  *phylogentically informative differences

- **Insertion or Deletion (Indel)**

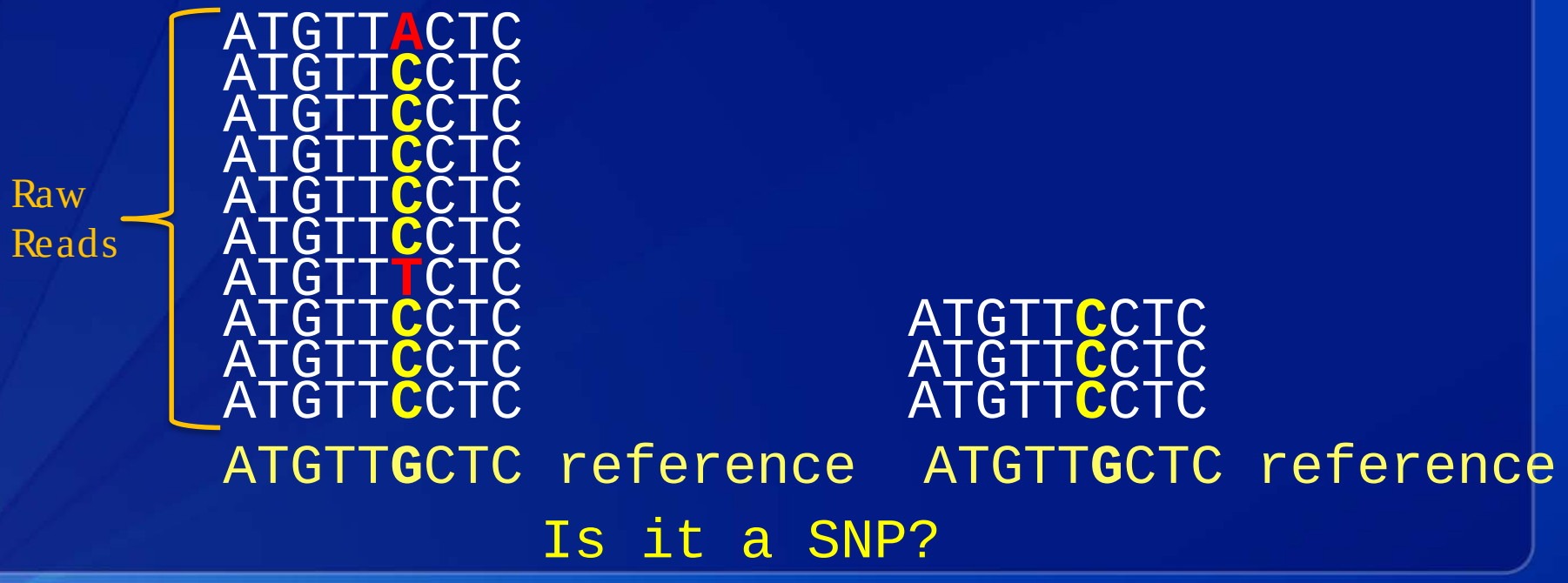
  ATGTT**CC**CTC sequence

  ATGTT**C-**CTC reference

  *differences not used in hqSNP analysis

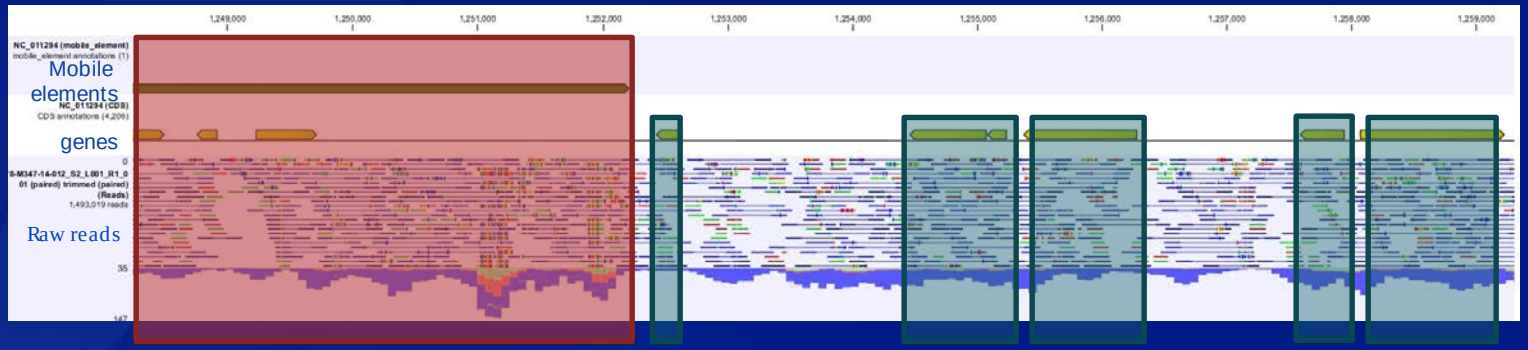
Ways to perform SNP Analysis

- **Reference-based SNP calling**
  - High quality SNP (hqSNP)
  - Raw reads are mapped to a highly related reference
  - Called based on coverage and read frequency at SNP location
  - Shows the phylogenetic relationship



Where to call SNPs

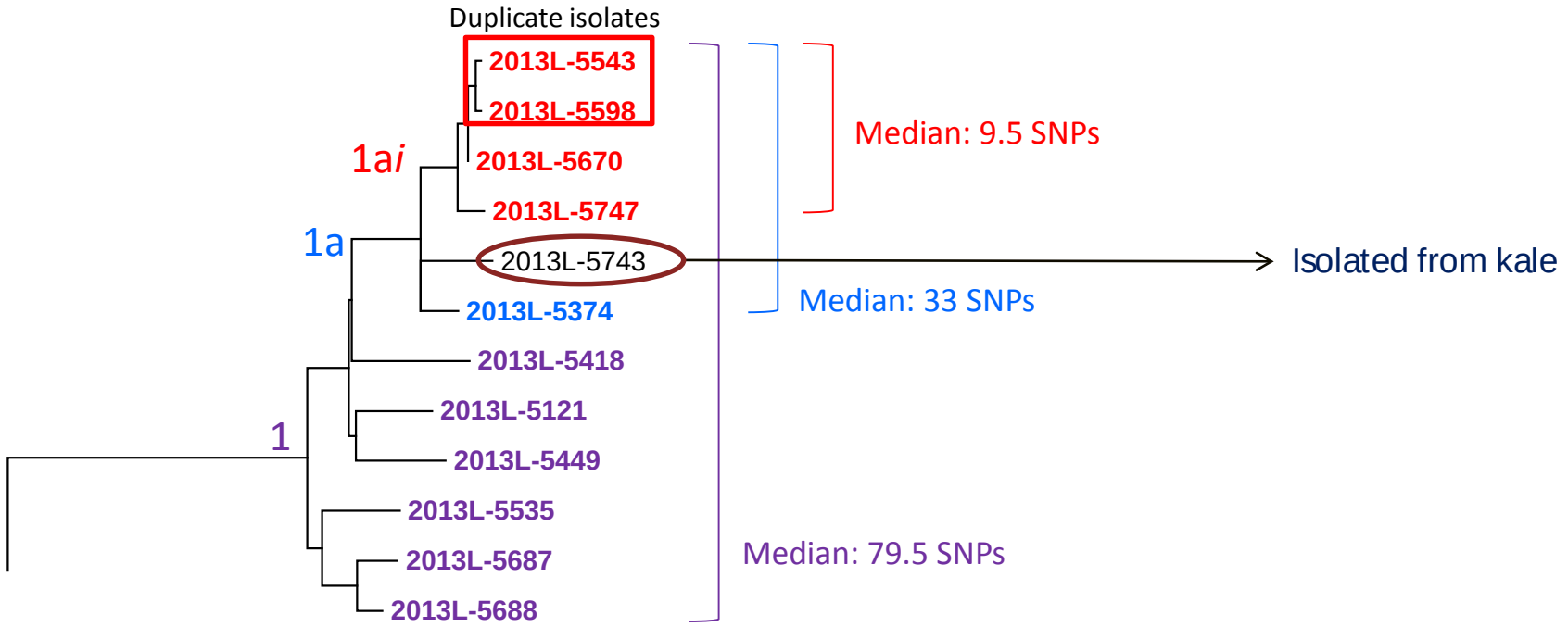
- Focusing on different parts of the genome will give you different SNP counts
  - Can look at SNPs in whole genome, in core genes only, or even mask part of the genome and not consider any SNPs found there.

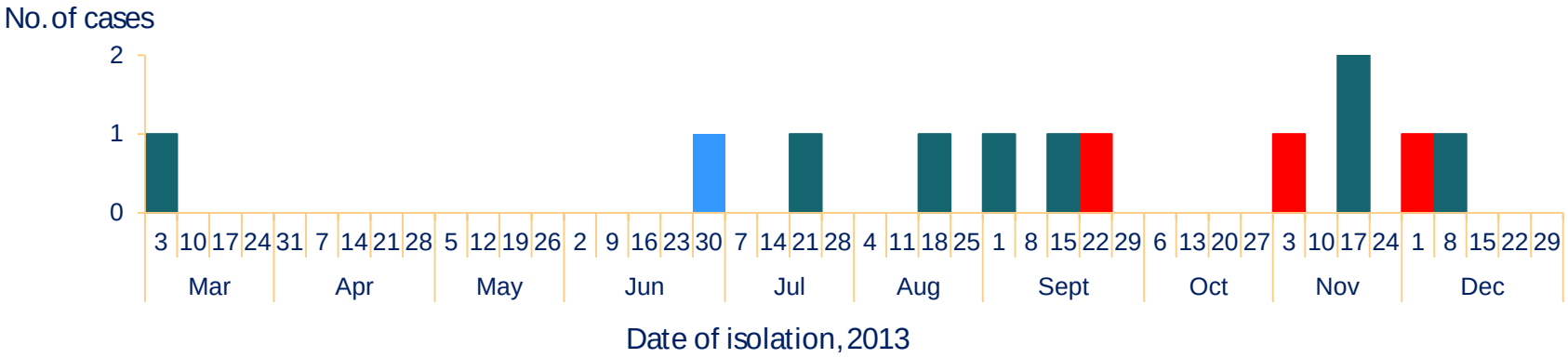


Mask mobile elements
-do no consider SNPs in this location

Only call SNPs in genes

Cluster 1 (1312MLGX6-1): Discriminatory Power





Caveats to hqSNP Analysis

Advantages:

- Phylogenetically informative
- SNP position can be identified on genome to determine what gene or intragenic region contains the SNP

Disadvantages:

- Requires a closed reference or good draft genome
  - Recent closed references from all serotypes are not available
- Computationally costly
  - Requires multiple sequence alignment to a reference
- Does not lead to strain type nomenclature
- Mutational hotspots, due to recombination or mobile elements, can make SNP counts artificially high

Whole Genome MLST (wgMLST)

- Compare gene content between different isolates (can compare over 5000 genes in *Listeria*)
- 1 or more differences (SNP or indel) equal to a new allele name
- Can categorize genes into subgroups: virulence profiles, serotypes, antimicrobial resistance determinants, housekeeping gene MLST, ribosomal MLST, core genome MLST, etc.
- Software like BIGSdb and BioNumerics 7.5 can run these analyses

Locus 1	ACTAGAGGGAAA allele 1	ACTAGAGGCTAA allele 2	ACT-GAGGGTAA allele 3

wgMLST Tree

*wgMLST tree made for Crave Bros and 1312MLGX6-1 cluster highlighting what the dendrogram looks like and the different allele calls

Caveats to wgMLST Analysis

Advantages:

- Phylogenetically informative
- All subtyping genes, virulence genes, and antibiotic resistance genes are pulled out as part of the analysis
- Can create a standardized nomenclature based on allele calls

Disadvantages:

- Computationally costly to initially assign alleles
- Comparing character data, not actual sequence data
  - SNPs and indels treated equally
  - No difference between 1 or more SNP or indel differences in naming an allele

Caveats for WGS Analysis

- **Opportunities**
  - Universal high resolution subtyping method
  - All information currently obtained by traditional methods contained in the sequence data
    - Can use to identify serotype, virulence genes, resistance genes, etc
    - Huge savings opportunity by replacing traditional methods with NGS
- **Challenges**
  - Large amounts of data presents storage and analysis issues
  - Currently no standardization for quality metrics or analysis pipelines
  - Backwards comparability of WGS data with PFGE difficult to establish
  - Interpretation of data – how to define clusters?

Comparison between NGS platforms for *Listeria* project

- Selected 22 isolates of *Listeria* from different serotypes, sporadic and outbreak isolates
- Sequenced the same 22 isolates on the Ion Torrent PGM and Illumina MiSeq
- Determined variability in assembly metrics, hqSNP calls, and allele calls for wgMLST

Findings

Factor	MiSeq	PGM
Coverage	128 (58x-266x)	47x (21x-73x)
Contigs per assembly	22 (assembled using CLC)	28 (assembled using MIRA)
N50	391,927	306,604
hqSNP calls	0-2 differences	
wgMLST loci detected	16 more identified by MiSeq	
wgMLST allele call diferences	0-2 discrepancies	

Platform Comparison Discussion

- Preliminary analysis suggests data generated from the 2 platforms is compatible to use in surveillance and outbreak detection
- Additional comparisons are being done looking at *Salmonella* and *Escherichia coli* data compatibility between the 2 platforms
- Determine if loci with missing allele calls from PGM data are important for outbreak detection
- Use Sanger sequencing to determine which platform made the correct base call where there were discrepancies

Vision for Implementation of WGS into PulseNet and Enteric Reference Activities

- **Advanced Molecular Detection (AMD)**
  - 5 year initiative
  - FY2014 funding: $30 million
    - Most will stay at CDC
    - Limited reagent support for the labs that already have Illumina Miseq
      - Sequence all STEC, selected Campy and *Salmonella*
  - FY2015 projected funding: $30 million
    - Increasing support for PHLs in transitioning to WGS
  - By the end of 2018, every PulseNet lab will sequence all foodborne isolates received replacing all current conventional workflows
    - Will be used for strain identification, serotyping, pathotyping, virulence characterization, AR monitoring, PulseNet subtyping

Vision for Implementation of WGS into PulseNet

- **Sequence data analysis within the AMD initiative**
  - wgMLST using the BioNumerics 7.5 as a primary surveillance tool
    - User friendly workflows
    - No need for specific bioinformatic expertise
    - Raw data storage at NCBI
    - Allele calls and metadata stored in sql-database at a CDC server
  - First pilot testing with selected labs in spring 2015
  - If funds available, the whole PulseNet network upgraded to BN 7.5 at the same time

Role of CDC Laboratories In The World of WGS

- Data management & data analysis
- Surge capacity for WGS
- WGS Troubleshooting
- National organism specific subject matter expertise
- 'Center for Classical Microbiology'
  - When WGS fails or new strains emerge
  - Sentinel surveillance using classical methods
- More integration of laboratory and epidemiology
  - Laboratory expertise is needed to use and interpret the data in epidemiological contexts

Questions?

For more information please contact Centers for Disease Control and Prevention
PulseNet/CDC
1600 Clifton Road NE, Atlanta, GA 30333

E-mail: pfge@cdc.gov Web: http://www.cdc.gov/pulsenet

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

- Resources:

Program	What for?	Where to find it	Cost?	Platform
BioNumerics 7.5	Assembly, wgMLST, SNP analysis	http://www.applied-maths.com/	Yes	Windows
CLC Bio Genomics Workbench	Workflows, read metrics, assemblies, etc, SNP analyses	http://www.clcbio.com/products/clc-genomics-workbench/	Yes	Windows/ Linux
Geneious	Assemblies, trees, SNP analysis	http://geneious.com/	Yes	Windows
MEGA5	Phylogenies	megasoftware.net/	No	Windows
Lasergene	Assemblies, read metrics, analysis	http://www.dnastar.com/	Yes	Windows
Genome Workbench	Viewing trees, analysis	http://www.ncbi.nlm.nih.gov/tools/gbench/	No	Windows/ Linux
CG-Pipeline	Assembly, read metrics, assembly metrics, read cleaning, etc	sourceforge.net/projects/cg-pipeline	No	Linux
Snp Extraction Tool	Creating Phylogenies	github.com/lskatz/lyve-SET	No	Linux

* List of some analysis tools for WGS data