The Basics of Understanding Whole Genome Next Generation Sequence Data

Heather Carleton-Romer, MPH, Ph.D.

ASM-CDC Infectious Disease and Public Health Microbiology Postdoctoral Fellow

PulseNet USA Next Generation Subtyping Unit

NCEZID/DFWED/EDLB

2013 InFORM

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

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

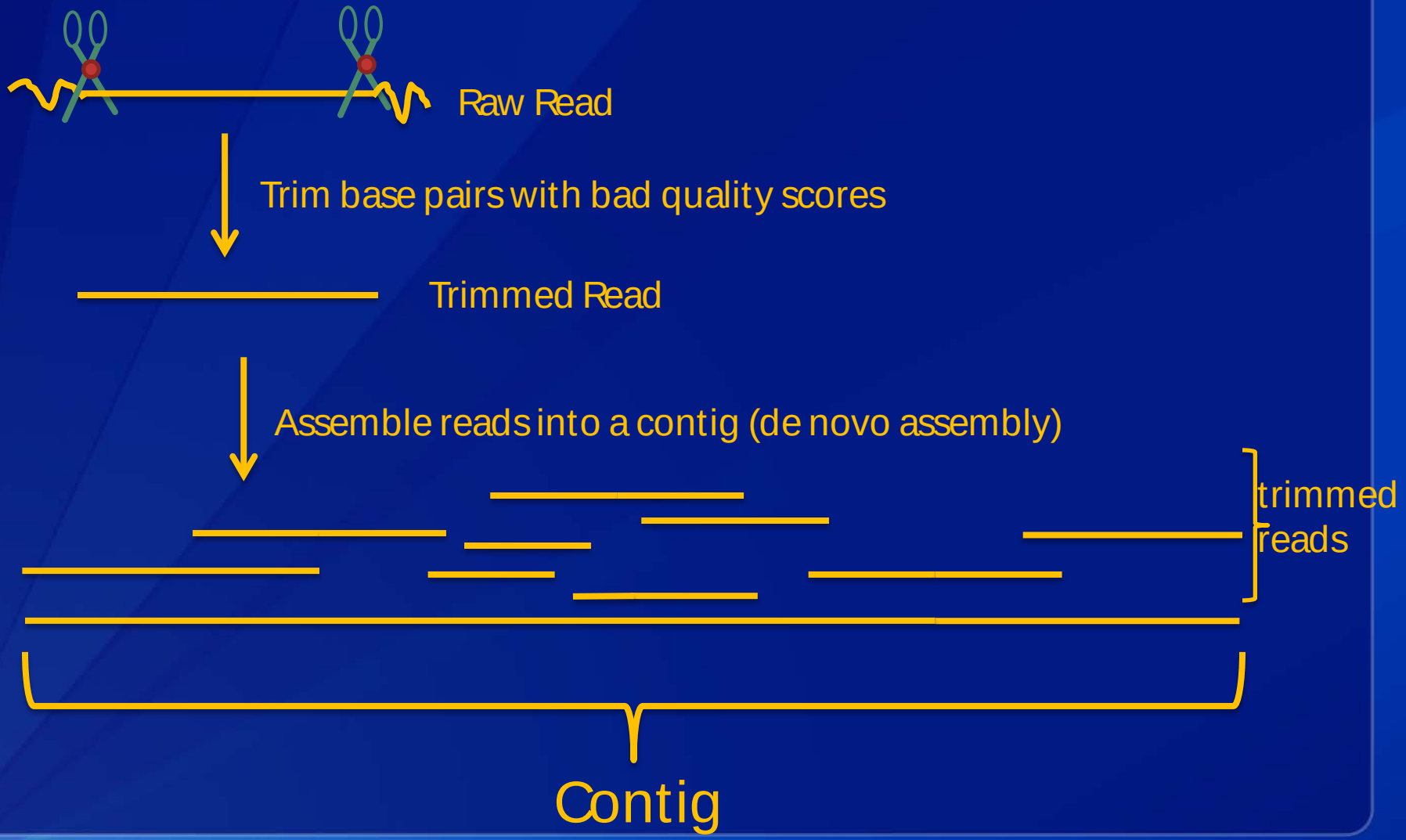
WGS terms for Raw Read Quality

- **Raw Read**
  - Single sequencing output from your NGS machine; length depends on sequencing chemistry
- **Quality scores**
  - Likelihood the base call is correct – each base read from the machine will have a quality score
  - Used for trimming reads
- **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

WGS terms for Assembly

- **Contig**
  - Assembly of overlapping reads into a single longer piece of DNA
- **Reference genome**
  - Well characterized genome that was assembled into 1 contig per chromosome and is fully circularized
- **Assembly – de novo versus reference guided**

Life of a NGS Read - Assembly



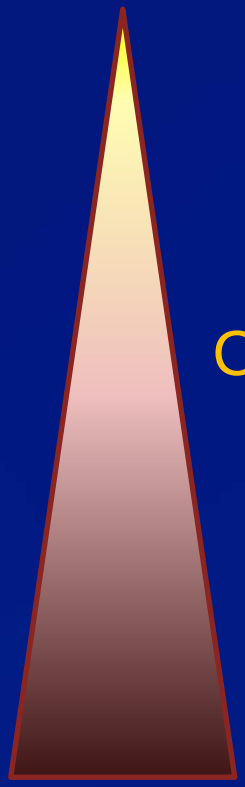
Assessing Assembly Quality

- **Assembly metrics can indicate sequence quality**
  - Number of contigs raw reads assembles into
    - Good: *E coli* <200, *Salmonella* < 100, *Listeria* < 30
  - N50 statistic– is similar to median contig length
    - Good: >200,000 bp

- **Can extract genes and other genetic elements from an assembly for a whole genome multilocus sequence type (wgMLST)**
  - Some genes or genome regions may be missing because that part of the genome did not sequence well and no reads covered it

Ways to Analyze WGS data

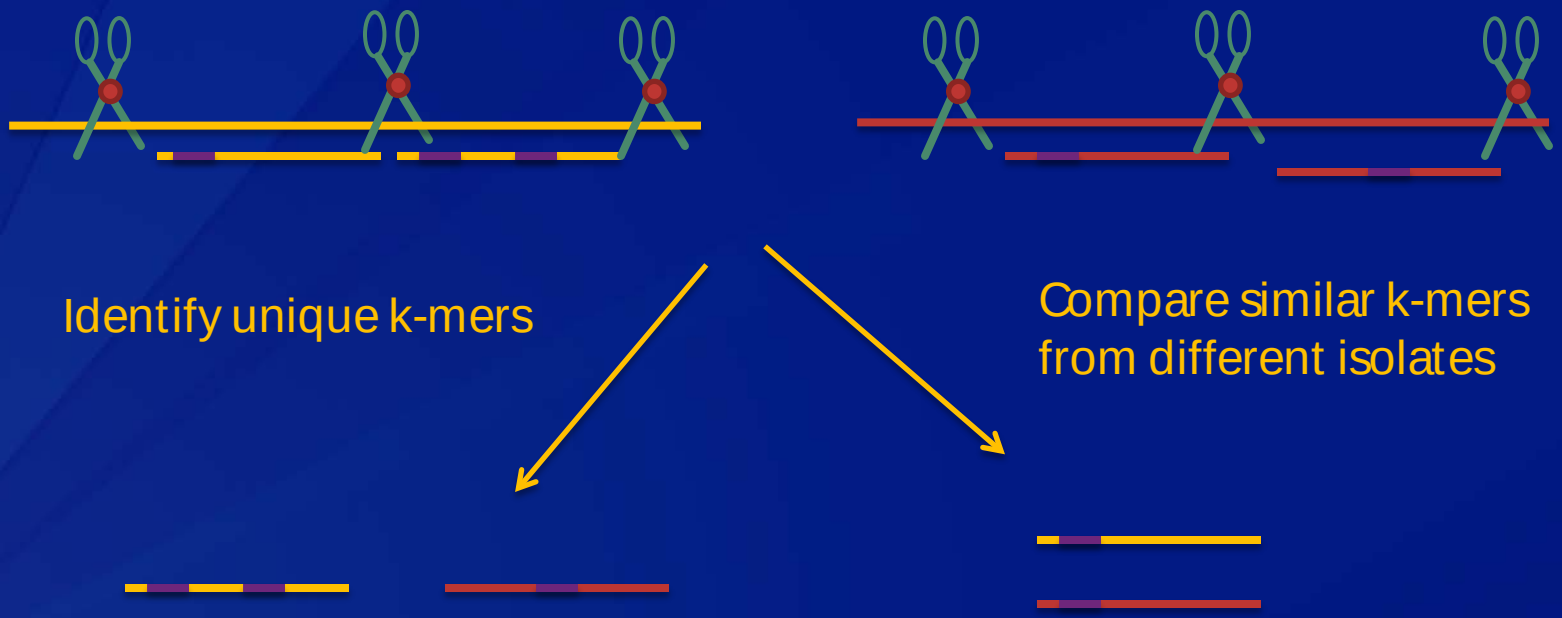
- Kmer analysis
- SNP analyses
  - K-mer SNP
  - hqSNP
- Whole genome multilocus sequence typing (wgMLST)



Computational demands

K-mer analysis

- **K-mer :**
  - Raw reads are chopped up into k-mers of length "k" using computer algorithms
  - k is determined by which length gives you the best specificity and most adequate resolution
  - Identify unique and similar k-mers to determine how related isolates are to each other



SNP Analysis Terms

- Single Nucleotide Polymorphism (SNP)

```
ATGTTCCTC sequence
ATGTTGCTC reference
```

- Insertion or Deletion (Indel)

```
ATGTTCCCTC sequence
ATGTTC-CTC reference
```

Ways to perform SNP Analysis

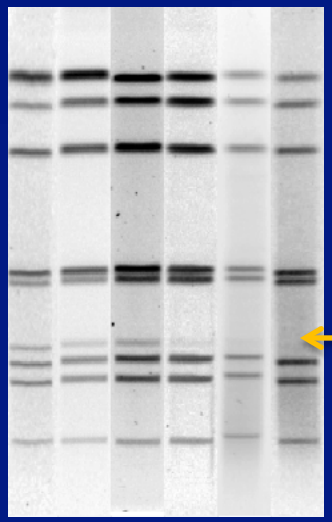
- **Reference-based SNP calling**
  - High quality SNP (hqSNP)
  - Raw reads are mapped to a highly related reference
  - Chosen based on coverage and read frequency at SNP location
  - Shows the phylogenetic relationship

ATGTTACTC
ATGTTCCTC
ATGTTCCTC
ATGTTCCTC
ATGTTCCTC
ATGTTCCTC
ATGTTTCTC
ATGTTCCTC
ATGTTCCTC
ATGTTCCTC

Raw Reads

ATGTTGCTC reference

Is it a SNP?



Is it a band?

Caveats to hqSNP Analysis

Advantages:

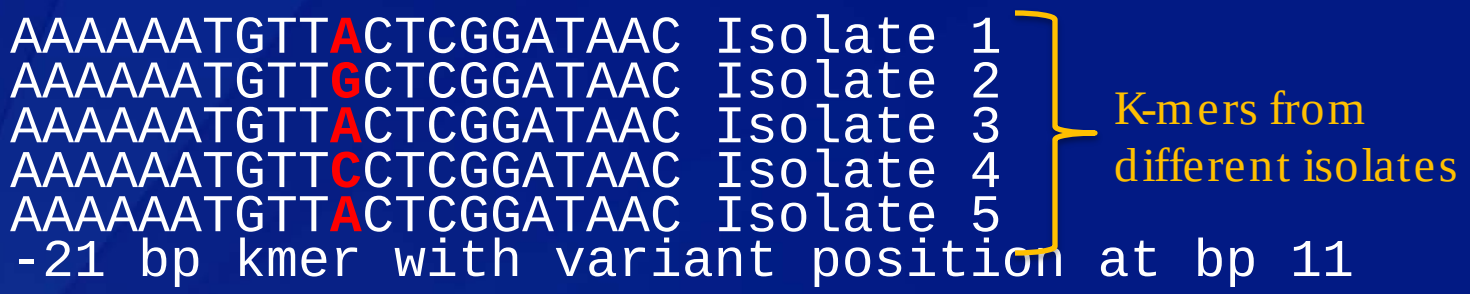
- Phylogenetically informative
- SNP position can be identified on genome to determine what gene or intragenic region contains the SNP

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

Ways to perform SNP Analysis

- **K-mer based SNP calling (kSNP)**
  - Raw reads are chopped up into k-mers of length "k" (k can vary from 13-35bp (or longer), must be an odd number)
  - K-mers screened for variation at the center position



Caveats to k-mer SNP analysis

Advantages:

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

Disadvantages:

- K-mer SNPs from raw reads do not identify the exact location of the SNP on a genome (from raw reads)
- Does not consider sequence quality
- Cannot detect SNPs located close to each other

SNP based analyses – how do you solve a puzzle

Puzzle pieces = raw reads

Different ways to solve?

Start with the edges then fit pieces = aligning to a reference to find hqSNPs

Start in the middle with matching color pieces (kmer) = Aligning kmers to find kmer SNPS

Whole Genome MLST

- Compare gene content between different isolates (can compare over 4000 genes)
- Differences between isolates based on SNPs or Indels within genes both genomes contain
- Can look at virulence genes, serotyping genes, and antibiotic resistance determinants between isolates
- Software like BIGSdb and BioNumerics 7.5 can run these analyses

Concluding remarks WGS

- **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
- **Challenges**
  - Large amounts of data presents storage and analysis issues
  - No standardization for quality metrics for analysis
  - Backwards comparability of WGS data with PFGE difficult to establish
  - Interpretation of data – how to define clusters?

Questions?

For more information please contact Centers for Disease Control and Prevention
PulseNet/CDC
1600 Clifton Road NE, Atlanta, GA 30333
Telephone: 1-404-718-4269
E-mail: hcarleton@cdc.gov Web: http://www.cdc.gov/pulsenet