Molecular Epidemiology for Epidemiologists

Key Acronyms & Terminology*

**16s sequencing:** sequencing of 16S ribosomal RNA of bacteria. Used for the past few decades as a way of typing bacteria. Typically, 16s typing is good enough to type down to the genus level, and occasionally to the species level.

**Accessory Genome:** see “core genome”

**Alignment:** the alignment of two or more sequences with each other such that the nucleotides (i.e., A, C, G or T) at each position match as closely as possible.

**Allele:** one of two or more variants of a gene; see “MLST”

**ANI:** average nucleotide identity. A system for comparing genomes based on similarity of conserved genes. ANI is often used to identify the species of an organism or to confirm its identity before proceeding to comparisons with other strains of the same species.

**AMD:** Advanced molecular detection

**AR, AMR:** antimicrobial resistance

**BioNumerics:** a proprietary software package by Applied Maths used since the 1990s to create a nationwide database on PFGE data and that is now being used by several programs to create nationwide genomics databases.

**Clade:** generically, a clade is “any monophyletic group”—essentially, any given branch of a phylogenetic tree. In certain domains (e.g., HIV), “clade” has a more specific definition.

**Codon:** a 3-nucleotide sequence in DNA or RNA that codes for a specific amino acid.

**Contig:** a contiguous sequence of DNA or RNA. Sequencing of bacteria, for example, often results not in a single, continuous sequence for the entire genome, but rather in several “contigs” that can’t be linked together into a full genome.

**Core Genome:** The ensemble of genes in a particular species of bacteria that are found in most strains of that species. A species might have, for example, 3000 genes in its “core genome” and another few thousand that are only found in certain strains (the “accessory genome”)

**Graph:** In the context of “graph theory”, a paradigm frequently used in bioinformatics, a “graph” consists of “nodes” (also called “vertices”) connected by “edges” (i.e., circles connected by lines). A phylogenetic tree is a type of “graph”.

**De novo assembly:** the assembly of sequence reads (i.e., the raw output from a sequencing machine) into a genome from scratch, without the assistance of a reference genome. See “reference-based assembly” for the alternative.

**Depth of Coverage (in genome sequencing):** also called “coverage”; the average number of reads representing a given position in the reconstructed sequence. Average coverage can be calculated by (total number of bases generated) / (size of the genome sequenced).

**Flow cell:** the (usually plastic or glass) disposable insert in a next-generation sequencing machine. Each “run” generally requires its own flow cell and reagent kit.

**GenBank:** a global database of nucleotide sequences established in the 1980s at NCBI.

**Genetic distance:** any of several measures of relatedness between two sequences.

**Genetics:** essentially, the study of individual genes in isolation

**Genomics:** the study of the entire genomes

**HTS, High-Throughput Sequencing:** a synonym for NGS, next-generation sequencing

**hqSNP:** high-quality SNPs; one of two systems (see wgMLST for the other) frequently used in public health to compare bacterial genomes

**Indel:** an insertion or deletion in a genome.

**Isolate:** usually refers to bacteria from a single colony on a culture plate (i.e., a colony isolated from the other colonies on the plate)

* The definitions here are tailored to the context of sequencing in public health. Many of the terms defined here have broader or different definitions when used in other contexts.

† Version 2016-11-02
Kmer: a nucleotide sequence of length K nucleotides. Dimers and trimers, for example are 2 or 3 nucleotides in length.

Library prep: The steps in preparation for sequencing, such as nucleic acid extraction, shearing and quality control.

LIMS: Laboratory information management system.

Locus: a generic term meaning the location in a genome. It can refer to a specific gene (as in MLST) or to specific nucleotide position.

Long Read Sequencing: The most common sequencing used today (Illumina [MiSeq]) sequencing typically results in short reads of 150-250 nt in length. Long-read sequencing (e.g., PacBio or nanopore sequencing) can produce long reads in the 10,000nt range.

MALDI-TOF: Matrix-associated laser desorption ionization, time of flight mass spectrometry; most commonly used today to identify bacteria.

Metagenomics: the analysis of the genomic data of an entire population of organisms. Metagenomics in public health is typically used either for diagnostics (e.g., sequencing a sputum sample and pulling out only reads from Mycobacterium tuberculosis) or to study microbial ecology (e.g., looking at the fecal microbiome and how antibiotics disrupt it).

MGE, Mobile Genetic Element: a term most often used in bacteriology to refer to elements such as insertion sequences or bacteriophages that can move around the genome and into and out of the genome; MGEs are often present in multiple copies in a bacterial genome, complicating genome assembly.

MIC, Minimum Inhibitory Concentration: the lowest concentration of antibiotic needed to inhibit growth of a particular bacterial isolate.

MIRU, Multiple Interspersed Repetitive Units. a means of sub-typing by examining the number of short tandem repeats (see STR), usually applied to Mycobacterium tuberculosis. Spoligotyping and MIRU/VNTR are being superseded by whole-genome sequencing.

MLST, Multi-Locus Sequence Typing: a means of typing bacteria by sequencing a small number of genes (typically 6 to 8) that are believed to be particularly variable between strains. Each gene (“locus”) is assigned to an “allele” based on its sequence.

MRCA: the “most recent common ancestor”, typically on a phylogenetic tree or transmission graph. On a phylogentic tree, the MRCA is found where all the relevant branches of the tree converge. On a transmission graph, on which direction of transmission is often ambiguous, the MRCA is the “node” (i.e., the case or isolate) that appears phylogenetically to be the oldest.

MST, Minimum Spanning Tree: a type of phylogenetic tree in which the sum of the branch lengths is minimized.

NCBI: National Center for Biotechnology Information; the unit within the NIH that houses “GenBank” and related databases.

NGS, Next-Generation Sequencing: high-throughput sequencing technologies that emerged in the 2000s to replace Sanger Sequencing.

Nucleic Acid: DNA or RNA

Nucleotide (nt): One of the 4 building blocks of DNA (A, C, T or G) or RNA (A, C, U, or G)

Optical Mapping: a technology for mapping instances of specific sequences over the length of an entire genome; generally used to create a “finished genome” by joining contigs created by NGS.

Phage, Bacteriophage: viruses that infect bacteria; phage genomes can incorporate into the bacterial genome and are often found when performing bacterial whole-genome sequencing.

Phylogenetics, Phylogenetic Analysis: inference of the evolutionary history of a group of related organisms, generally based on an analysis of their genomic data.

Pipeline: in the bioinformatics context, a succession of bioinformatics software used to process data; for example, bioinformaticians usually create a pipeline to assemble and validate a genome from raw NGS data.

PCR, Polymerase Chain Reaction: a technology developed in the 1980s, usually used to detect the presence of specific organisms by copying their DNA over and over again.

PFGE, Pulse-field Gel Electrophoresis: a system for typing bacteria by cutting their DNA at specific sites (i.e., at specific sequences) and examining the lengths of the pieces that result. PFGE is
largely being replaced now by whole-genome sequencing.

**Proteomics:** the study of an organism’s proteome—the ensemble of all of its proteins. MALDI-TOF is commonly used in proteomics.

**Real Time PCR:** a variation of PCR in which the product of the PCR reaction (the “amplicon”) is detected in real time. Since the early 2000s, real-time PCR has become the predominant variant of PCR used for pathogen detection.

**Reference-Based Assembly:** the assembly of sequence reads (the raw output from a sequencer) by matching them against a known genome of the same species or strain. See “de novo assembly for the alternative.

**Ribotyping:** a means of subtyping bacteria involving slicing with restriction endonucleases, separating on a gel, and then hybridizing with probes targeting ribosomal RNA.

**Sanger Sequencing:** the predominant sequencing technology between its invention in the late 1970s and the advent of next-generation sequencing in 2008.

**Short Read Sequencing:** currently the predominant method of sequencing in public health, which typically results in millions of sequences that are 150-250 nucleotides long

**Beginner’s guide to comparative bacterial genome analysis using next-generation sequence data**
technology used in PacBio instruments

**SNP, Single Nucleotide Polymorphism:** originally synonymous with “point mutation” (i.e., a mutation in a single point in a sequence). In current usage, SNPs typically include more complex mutations such as insertions and deletions of segments of DNA.

**Spoligotyping:** a means of sub-typing, usually applied to *Mycobacterium tuberculosis*. Spoligotyping and MIRU/VNTR are being superseded by whole-genome sequencing.

**ss/dsDNA:** single- or double-stranded DNA; often used to designate the type of genome of a virus.

**ss/dsRNA:** single- or double-stranded RNA; often used to designate the type of genome of a virus.

**STR, Short Tandem Repeats:** short, sequentially repetitive strands of DNA. SRT analysis involves measuring the length of these repeats at specific parts of the genome; see MIRU

**Synthetic Long Read Sequencing:** technologies for using short-read sequencing to

**Typing/Subtyping:** classification of organisms to a finer degree, usually beyond the species level.

**VNTR, Variable Number of Tandem Repeats:** a subtyping method that examines the number of copies of a specific sequence in an organism. VNTR is commonly used to type *Mycobacterium tuberculosis*. Spoligotyping and MIRU/VNTR are being superseded by whole-genome sequencing.

**wgMLST:** whole genome multi-locus sequence typing.

**WGS:** whole genome sequencing

**ZMW, Zero-Mode Waveguide:** part of the technology used in PacBio sequencing.
Publications of interest

**Twenty years of bacterial genome sequencing** (Nicholas Loman and Mark Pallen). http://www.nature.com/nrmicro/journal/v13/n12/full/nrmicro3565.html.


**Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation** (Brendan Jackson et al). http://cid.oxfordjournals.org/cgi/reprint/ciw242?ijkey=oikWr46pgn57udr&keytype=ref

