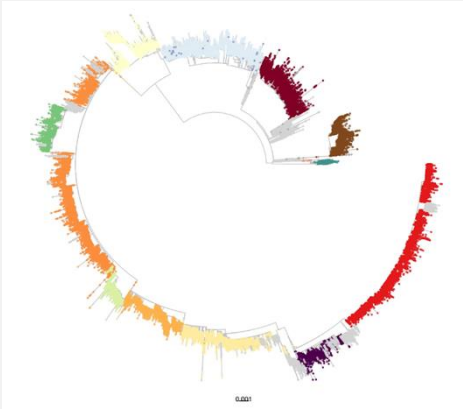
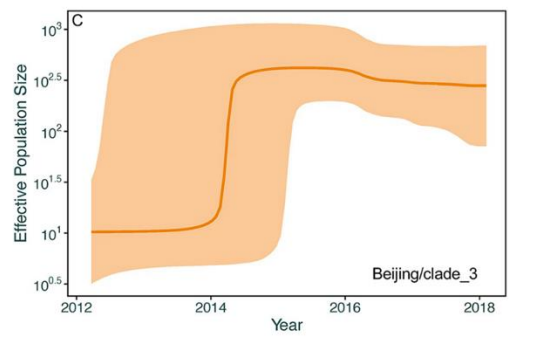
GENOMIC ANALYSIS AND PHYLODYNAMICS

Lecture 3: Practical Applications of WGS and Phylogenetics





Instructor: Dr. Ben Sobkowiak
MRC Senior Research Fellow, University College London

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

- Species, sequence type and lineage identification
- Predicting antimicrobial resistance
- Plasmid profiling
- Inferring transmission and virulence

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Clinical and biological applications of WGS

- After variant calling and/or de novo assembly, we can get consensus sequences + site level information
  - Read depth, coverage, allele frequencies
- Comparison of sequence data and genomic assemblies at different levels can inform different analyses
- Sample-specific variation compared to previously characterised strains and variant databases
  - Species, sequence type and lineage identification
  - Antimicrobial resistance and plasmid profiling
- Genomic variation within sampled isolates
  - Transmission dynamics
  - Novel mutations associated with phenotypes

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

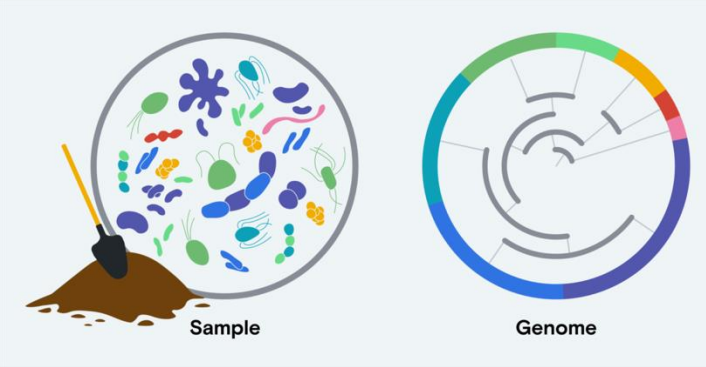
Species identification

- **DNA Barcoding**: Uses short, standardized gene regions (e.g., COI) to match to reference databases (e.g., NCBI BLAST)
- **Alignment-based**: Aligns sequence data (reads or assembled contigs) against curated reference genomes to identify the closest known species (MetaPhlAn, Kraken2, Bracken)
- **Phylogenetic Placement**: Identification based on clade placement even without exact matches (e.g., PhyloPhlAn, UShER, NextStrain)
- **K-mer or Machine Learning Approaches**: Uses patterns of short sequence motifs (k-mers) or trained classification models to assign taxonomic labels (e.g., Kraken2, Bracken, MIDAS)

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Metagenomics

- Identifies species directly from mixed DNA without culturing
- Perform untargeted sequencing (shotgun) or amplify target genes (amplicon)
- Can use reference databases to classify reads or contigs
  - Assembly and binning tools (e.g. MetaBAT2, CONCOCT) reconstruct draft genomes
  - Classified via ANI comparisons or phylogeny
- Enables discovery of uncultured or novel organisms



LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

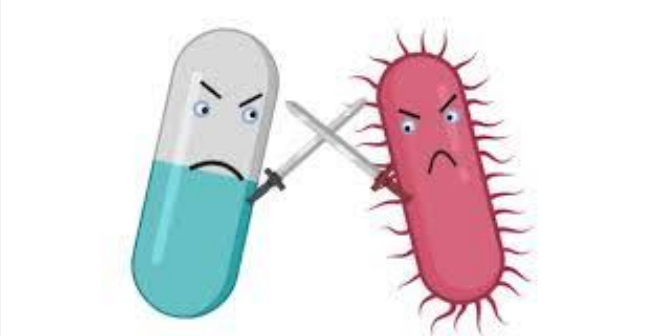
Sequence types and lineage identification

- **MLST and cgMLST**: Assigns sequence types (STs) using alleles from core loci (e.g., PubMLST)
- **Lineage Calling Tools**: Uses curated SNP markers or phylogeny to assign lineages and variant types (e.g., TB-Profiler *M. tuberculosis*, Pangolin SARS-CoV-2)
- **k-mer Based Typing**: Lineage and ST assignment from assemblies (e.g. Kleborate *Klebsiella spp.*)
- **Phylogenetic Reconstruction**: Infers lineages through phylogenetic trees or pan-genome analysis (e.g., UShER, PhyloPhlAn)

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Antimicrobial resistance

- Resistance can arise through chromosomal mutations or horizontal gene transfer, often facilitated by mobile elements like plasmids
- Undermines the effectiveness of antibiotics, making infections harder to treat and increasing the risk of spread
- Major public health threat, with rising resistance in key pathogens like E. coli, Klebsiella, and M. tuberculosis



LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Predicting antimicrobial resistance

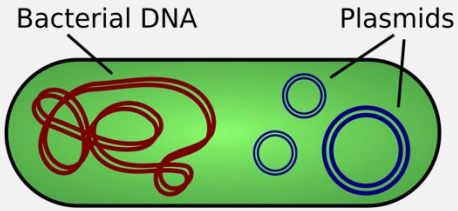
- **Gene Detection:** Identifies known resistance genes from assemblies using databases like CARD, ResFinder, or ARG-ANNOT (e.g., Abricate, ARIBA)
- **Point Mutation Calling:** Detects resistance-associated SNPs in chromosomal genes (e.g. , TB-Profiler for *M. tuberculosis*, Mykrobe).
- **Machine Learning Models**: Predicts resistance phenotypes from genomic features using trained classifiers (e.g., MSDeepAMR, DeepARG).
- **Integrated Prediction Tools**: Combine gene and SNP detection with lineage (e.g. Kleborate, Mykrobe, TB-Profiler)

*Also, investigating genotype/phenotype associations (Advanced course)

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Plasmid detection

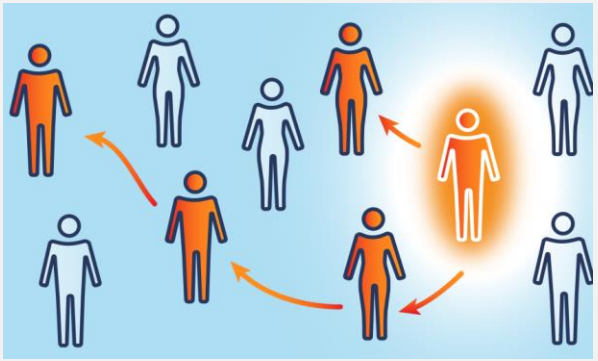
- Extra-chromosomal, typically circular DNA molecules found in bacteria
- Often carry genes for antimicrobial resistance, virulence, or metabolic traits, and can be transferred between bacteria
- Can be detected by identifying replicon sequences (e.g., PlasmidFinder) or clustering plasmid-associated genes (e.g., MOB-suite)
- Tools (e.g., plasmidSPAdes and Recycler) can reconstruct plasmid contigs from short-read data and long-read assemblers (e.g., Unicycler) can generate complete circular plasmid sequences with high confidence



LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

What are transmission networks?

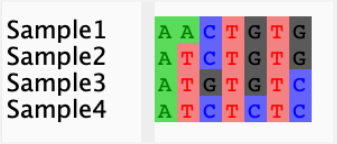
- Reflect the transmission history of infectious diseases
- Classical approaches (e.g., contact tracing) can be labour-intensive and un-reliable
- Genomics and phylogenetics can help to reconstruct transmission networks
  - Who-infected-whom
  - When and where
- Used to identify patterns of transmission (transmission hotspots, vaccination efficacy, individual and group risk factors)



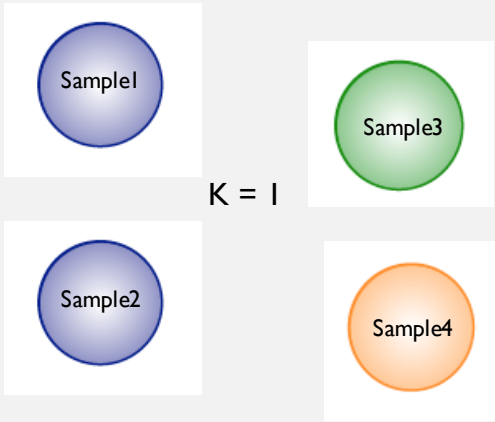
LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

How to reconstruct transmission?

- Compare the pathogen sequences from infected individuals
- Simplest approach – transmission clusters
- Place similar sequences into groups given a threshold to estimate recent transmission between hosts
- Sequences are placed in the same cluster if they differ by *k* SNPs or fewer



	Sample 1	Sample 2	Sample 3	Sample 4
Sample 1				
Sample 2	1			
Sample 3	3	2		
Sample 4	3	2	2	



LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

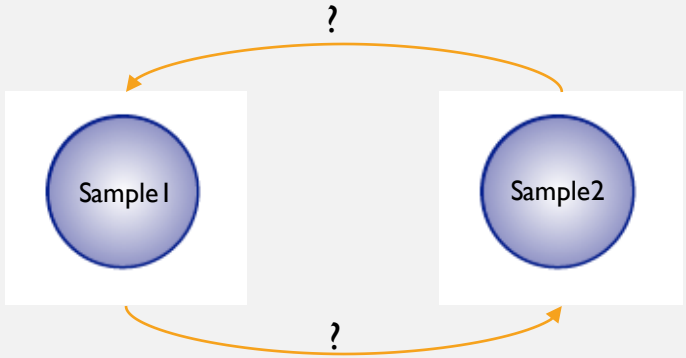
Limitations with transmission clusters

- Should all sites count for the same distance?
- Which threshold to choose?
- Dependent on bioinformatic pipeline

- Can include other information (e.g., dates) – TransCluster (Stimson et. al. 2019)

- Clustering still only tells us who is closely related, not who-infected-whom

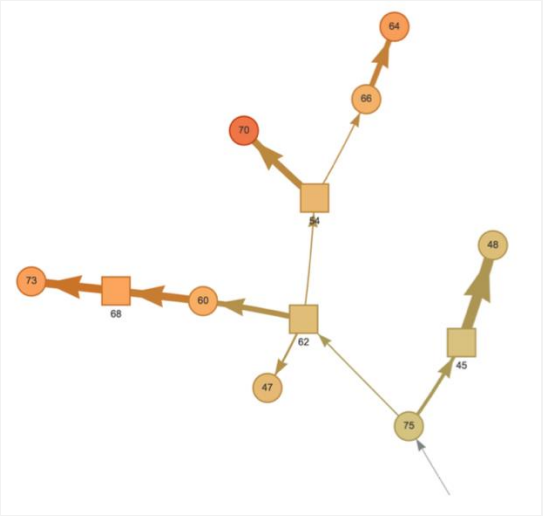
- More complex probabilistic approaches combining genomic data and epidemiological models can estimate transmission networks



LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Reconstructing full transmission networks

- Transmission networks include who-infected-whom and time of transmission events
- Constructed as a graph in which nodes are infected individuals and directed edges correspond to transmission events
- Edges may be associated with times of infection
- Can also be visualized as a transmission tree



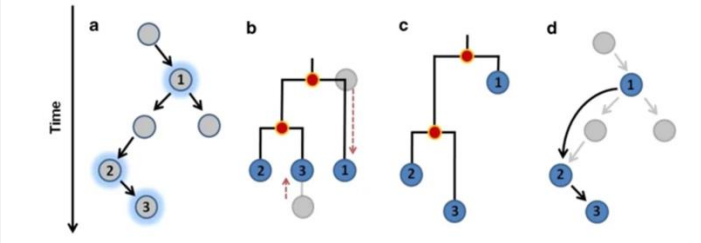
LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Models to reconstruct transmission networks

- Multiple computational tools developed to combine genomic and epidemiological data to infer transmission networks

- Graph approach – SeqTrack (Jombart et. al., 2011)
  - Optimises branching in a directed graph using genetic distance
  - Very quick but simplistic
  - No uncertainty in the transmission tree or probabilistic parameters
  - Timing not explicitly used to weight graphs
  - No un-sampled individuals in networks

- More complex models developed to better infer transmission



LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

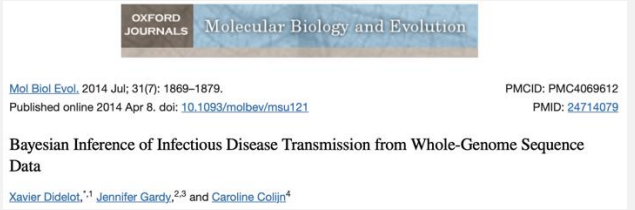
Models to reconstruct transmission networks

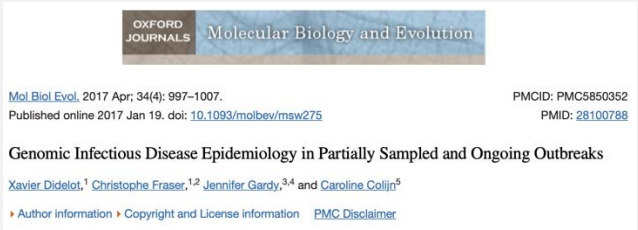
- More complex methods use a Bayesian framework (TransPhylo, Outbreaker, Phybreak, SCOTTi etc.) and runs MCMC to sample many trees
- This allows more estimation and greater flexibility by incorporating prior knowledge, and genetic and epidemiological parameters
- Transmission trees are estimated from the posterior distribution of inferred transmission events and infection times
- Different tools incorporate different underlying models and parameters (generation and sampling time, sampling density, mutation rate, within-host evolution etc.)

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

TransPhylo

- Bayesian approach using a probabilistic model that accounts for within-host diversity in a coalescent model
- Uses a Monte Carlo Markov Chain (MCMC) method to sample from the posterior distribution of transmission trees, given the phylogenetic tree and sampling times
- Captures the uncertainty associated with the inference
- Can handle missing data and ongoing outbreaks

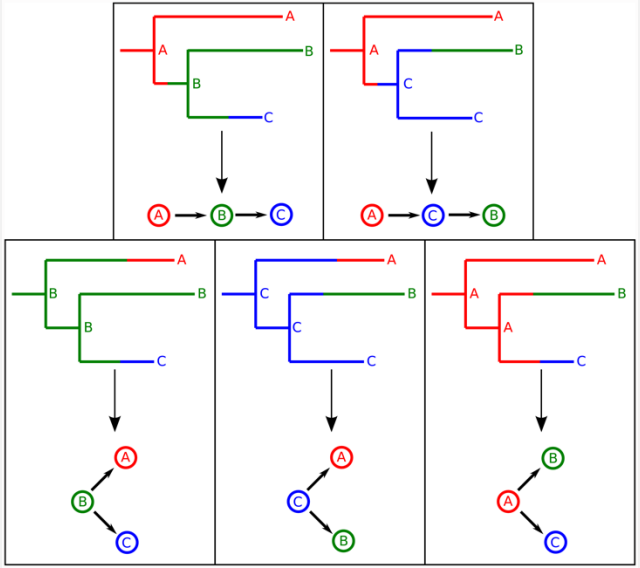




LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

TransPhylo

- Takes a timed phylogeny built from pathogen sequences and dates to infer transmission trees
- Phylogenies consider nucleotide substitution and population demography so accounted for in the model
- Phylogenetic tree ≠ transmission tree in all instances and it might not be clear who-infected-whom
- Phylogeny – tips are sampled individuals, internal nodes common ancestors
- Transmission tree – internal nodes are infected individuals - sampled or not, edges transmission events

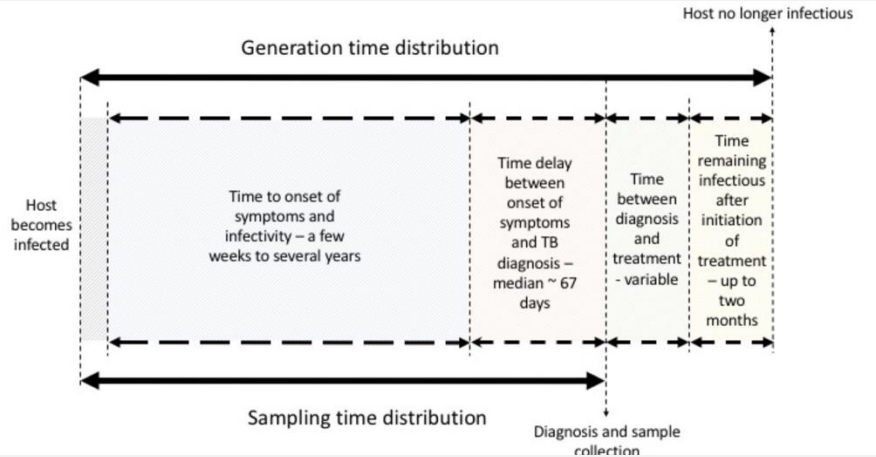


Hall et al, 2015 Plos Comp Bio

LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

TransPhylo

- Can include key epidemiological parameters:
- Generation time distribution
- Sampling time distribution
- Sampling density
- Date of last sample (or ongoing outbreak)



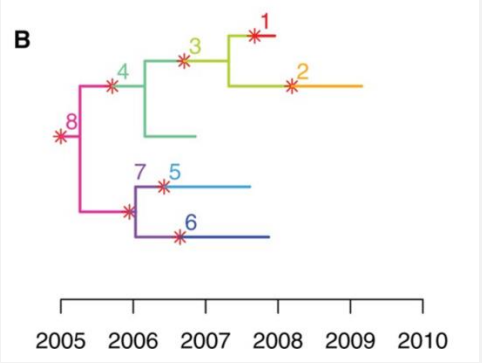
Sobkowiak et. al., 2020 *Microbial Genomics*

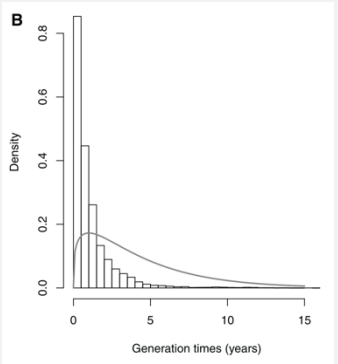
LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

TransPhylo

- Output - posterior collection of transmission events between sampled and un-sampled hosts, along with dates

- From this you can get:
  - Consensus transmission trees
  - Times between infections
  - Time between infection and sampling
  - Number and placement of missing cases
  - Offspring distribution (no. secondary infections)





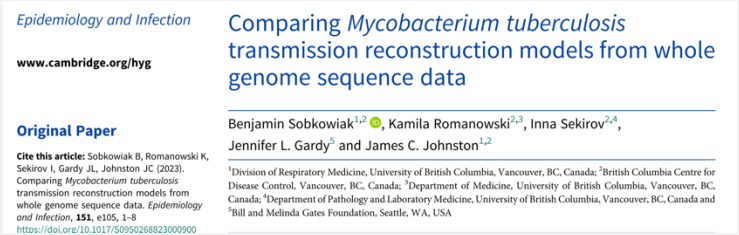
LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Which tool to choose?

- Many computational tools available to reconstruct transmission networks – can give different results

Choice depends on:

- Organism
  - Mutation rate, latency, within-host diversity, recombination, timescale
- Data completeness/availability
  - Sampling density, epi data, sequence data type
- Computational cost
  - Some models require more intensive computation but more complex modelling



Epidemiology and Infection

www.cambridge.org/hyg

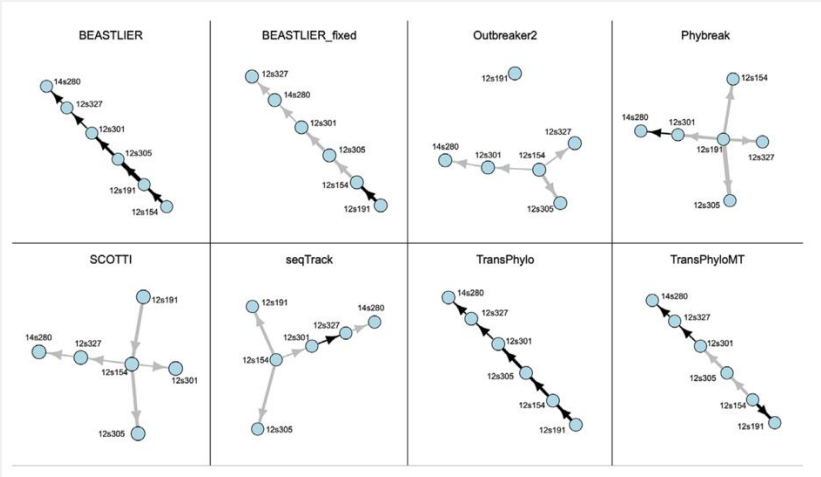
Original Paper

Cite this article: Sobkowiak B, Romanowski K, Sekirov I, Gardy JL, Johnston JC (2023). Comparing *Mycobacterium tuberculosis* transmission reconstruction models from whole genome sequence data. *Epidemiology and Infection*, **151**, e105, 1–8
https://doi.org/10.1017/S0950268823000900

Comparing *Mycobacterium tuberculosis* transmission reconstruction models from whole genome sequence data

Benjamin Sobkowiak[1,2], Kamila Romanowski[2,3], Inna Sekirov[2,4], Jennifer L. Gardy[5] and James C. Johnston[1,2]

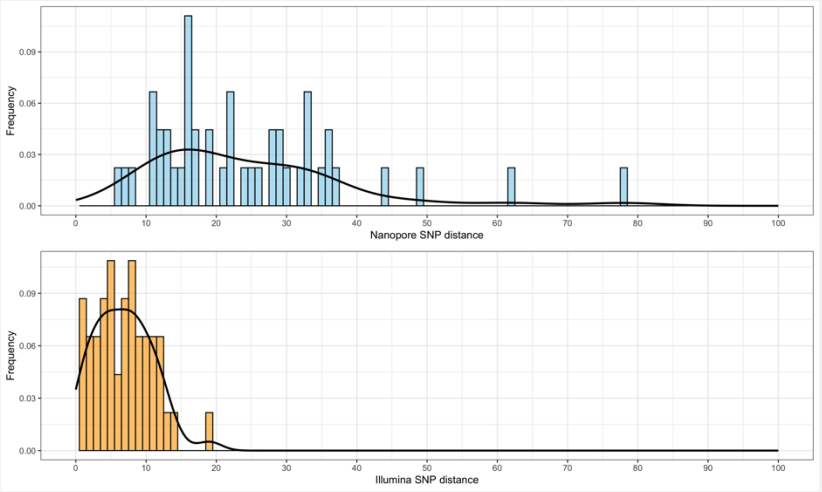
[1]Division of Respiratory Medicine, University of British Columbia, Vancouver, BC, Canada; [2]British Columbia Centre for Disease Control, Vancouver, BC, Canada; [3]Department of Medicine, University of British Columbia, Vancouver, BC, Canada; [4]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada and [5]Bill and Melinda Gates Foundation, Seattle, WA, USA



LECTURE 3: PRACTICAL APPLICATIONS OF WGS AND PHYLOGENETICS

Limitations in transmission reconstruction

- Missing or un-sequenced hosts can lead to incorrect or ambiguous transmission links
- Events may be hard to distinguish genetically, particularly in slowly evolving taxa (e.g., TB); inference is probabilistic rather than definitive
- Within-host diversity, sequencing errors, and variation in mutation rates can impact accuracy



PRACTICAL 3 AND 4: TIMED PHYLOGENETIC TREES; TRANSMISSION AND PROFILING

1. One-step timed phylogenetic tree with BEAST2
2. Two-step timed phylogenies using ML + Bayesian frameworks
3. Identifying species, serotypes, and lineages
4. Inferring transmission networks/clusters