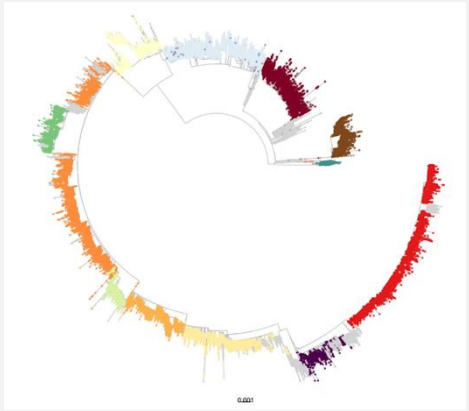
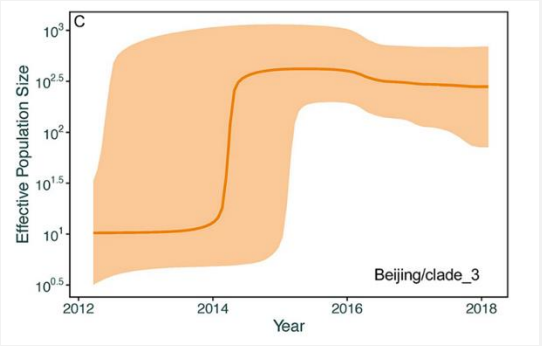
GENOMIC ANALYSIS AND PHYLODYNAMICS

Lecture 4: Advanced Applications of WGS





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

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Monday 5th May (Advanced)	9:00–10:30	**Lecture 4: Advanced Applications of WGS**	• Phylogeography and phylodynamics • Recombination • Average Nucleotide Identity (ANI) • Mixed infection • Fitness and selection
	10:45–12:00	**Practical Session 5: Mixed Infection, Recombination and ANI**	• Identifying mixed infection • Calculating ANI • Testing for recombination
	12:00–13:00	**Lunch Break**	
	13:00–13:30	**Practical Session 5 (cont.): Mixed infection, Recombination and ANI**	• (cont.) Identifying mixed infection • Calculating ANI • Testing for recombination
	13:30–15:00	**Practical Session 6: Phylogeography and Phylodynamics**	• Phylogeography (ancestral state reconstruction) • Phylodynamic analysis with BEAST2 (Skyline analysis)
Tuesday 6th May (Advanced)	9:00–12:00	**Practical Session 7: Fitness and Selection**	• Strain-specific fitness (LBI) • Site-specific selection (homoplasy, dN/dS) • GWAS
	12:00–12:30	**Closing Remarks - Advanced Course**	• Full course summary and feedback collection

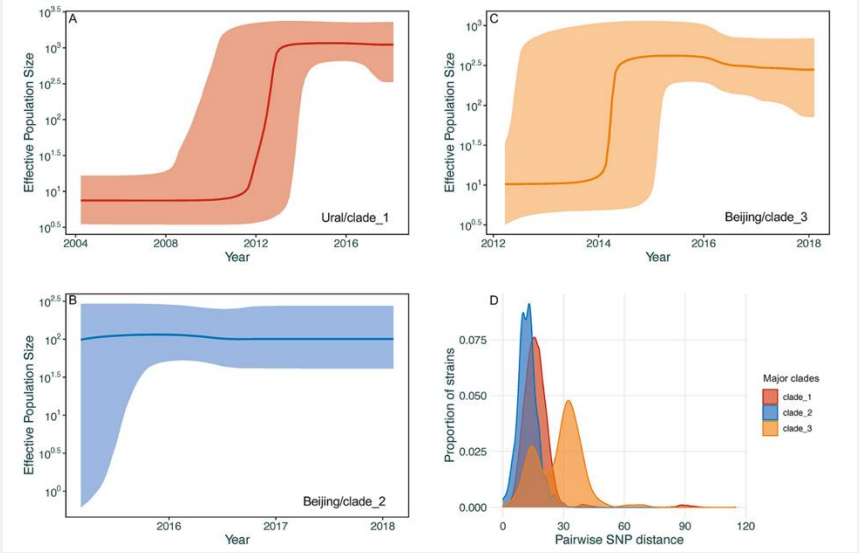
LECTURE 4: ADVANCED APPLICATIONS OF WGS

- Phylodynamics and phylogeography
- Recombination analysis
- Average Nucleotide Identity (ANI)
- Mixed infection
- Fitness and selection

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Phylodynamics

- Phylodynamics is the study of how epidemiological and evolutionary processes act to shape phylogenies
- Phylodynamic approaches typically involve the reconstruction of phylogenetic trees, combined with mathematical models that describe evolutionary processes and population dynamics



From Yang et. al. 2022

LECTURE 4: ADVANCED APPLICATIONS OF WGS

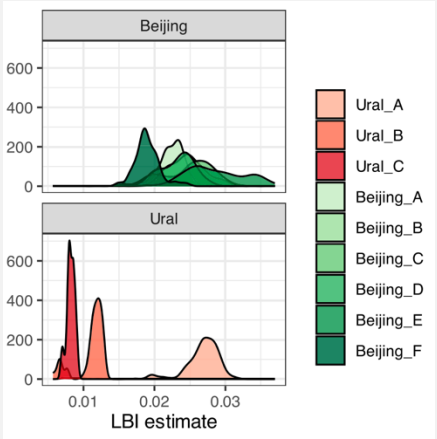
Phylodynamics

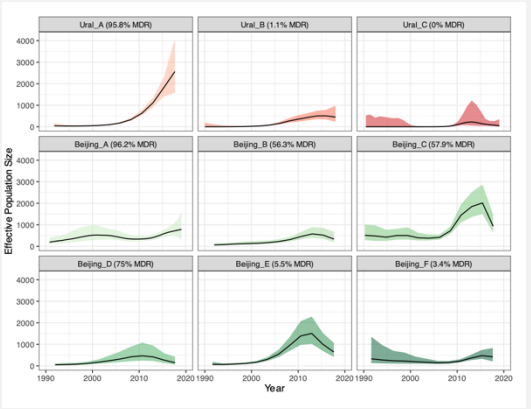
- These analyses can help us to estimate key epidemiological parameters:
  - **Effective population size (Ne)** – Genetic diversity over time
  - **Reproductive number (R or Re)** – Average number of secondary infections caused by an infected individual
  - **Transmission rate (β)** – How quickly the pathogen spreads between hosts
  - **Becoming non-infectious rate (δ)** – Rate at which individuals stop being infectious (recover/death)
  - **Sampling proportion (ρ)** – Fraction of infected individuals whose sequences are sampled
  - **Infection duration (D)** – Can be inferred as the inverse of δ.
  - **Transmission bottlenecks** – Size of founding population during transmission events.
  - **Migration rates** – In structured models, between-host population or region migration.

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Phylodynamics

- Reconstructed past population dynamics (Ne) of 9 clades of M.tb in Moldova
- High local branching index in lineage 4 clade – strong fitness of strains
- Found evidence of recent, rapid expansion of one MDR clade of lineage 4
- Leads to more questions, what's driving this expansion? – mutation and epidemiological analysis



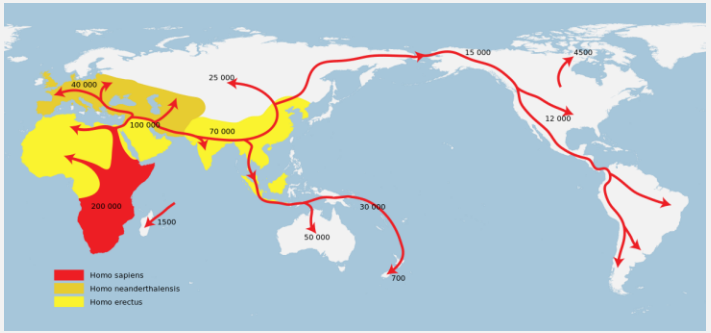


From Chitwood et. al. 2024, Nature Comms.

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Phylogeography

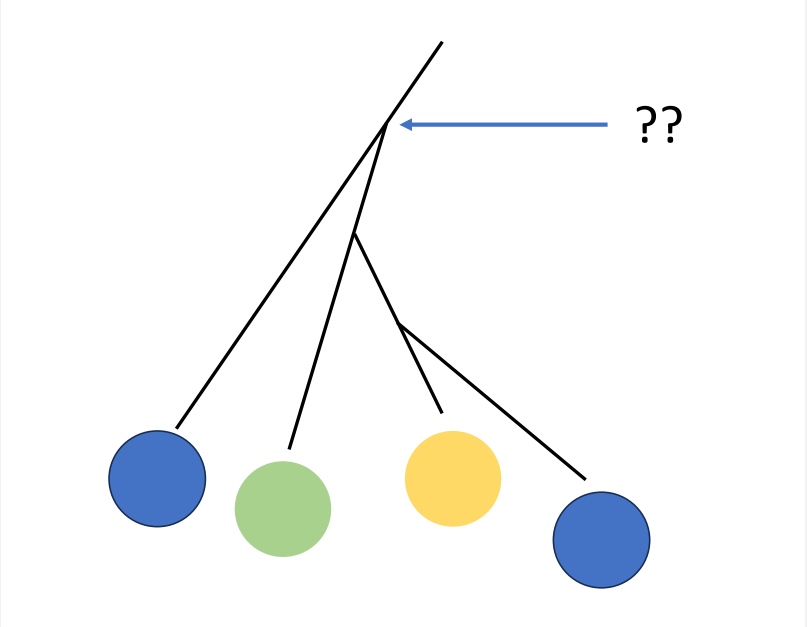
- Phylogeography allows us to reconstruct when and where lineages or clades were present
- Estimate the location of the emergence of new strains or lineages
- Track the migration of different strains or the flow of particular genes or traits



LECTURE 4: ADVANCED APPLICATIONS OF WGS

Phylogeography

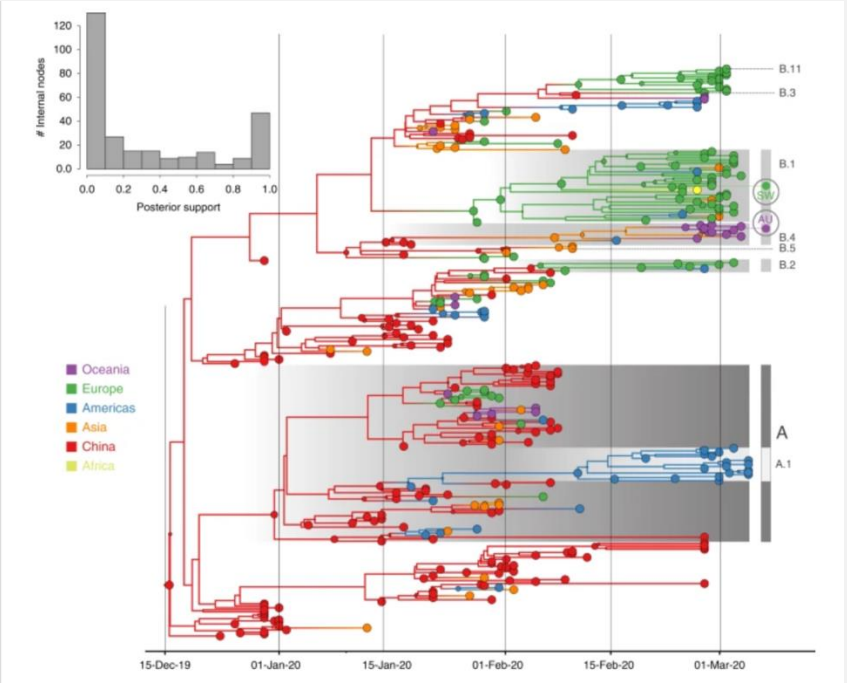
- In which location were past ancestors?
- Inferred using phylogenetic trees and the location data at tips using ancestral state reconstruction
- Employs probabilistic models to infer the most likely states of ancestral nodes, taking into account:
  - the observed character at tips
  - the topology of the phylogenetic tree
  - evolutionary processes governing the character evolution



LECTURE 4: ADVANCED APPLICATIONS OF WGS

Phylogeography

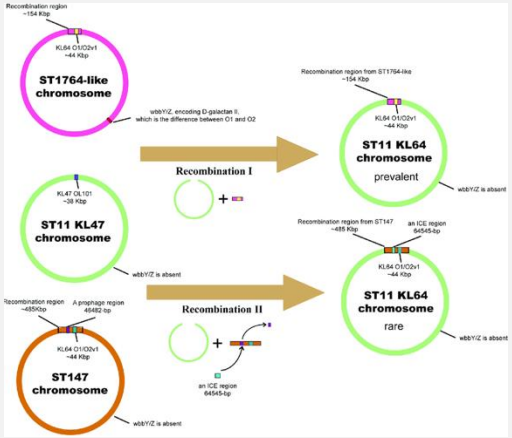
- Lemay et. al. 2020 reconstructed the location of the emergence of the COVID-19 pandemic
- Also inferred the dates in which there were first introductions to other regions



LECTURE 4: ADVANCED APPLICATIONS OF WGS

Recombination

- Exchange of genetic material between different genomes or genomic regions
- Results in new combinations of alleles and contributes to genetic diversity,
- Can obscure phylogenetic signals, as different parts of a genome may have different evolutionary histories
- Can facilitate adaptation (e.g., antimicrobial resistance or immune evasion) by importing beneficial genes or alleles from other strains or species

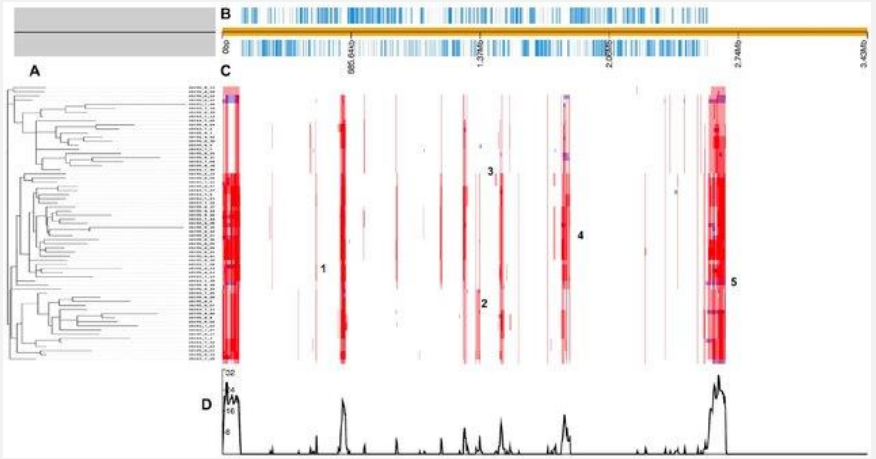


From Chen et. al. 2023

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Recombination

- WGS detects recombination by identifying regions with unusual SNP patterns or conflicting phylogenetic signals
- Bioinformatic tools (e.g., Gubbins, ClonalFrameML) scan genome alignments to find and annotate recombination events
- Recombinant regions can then be masked before phylogenetic or association analyses to reduce bias and false positive associations

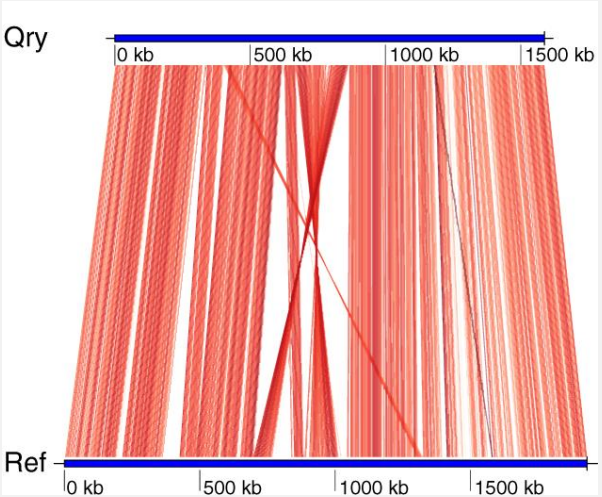


From Hanachi et. al. 2020

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Average Nucleotide Identity (ANI)

- Measures genetic similarity between two genomes
- Commonly used to define bacterial species boundaries
  - (threshold of ~95–96% indicates same species)
- More robust than 16S rRNA for distinguishing closely related microbial taxa
- Reveals overall genomic relatedness, helping to classify species and study divergence

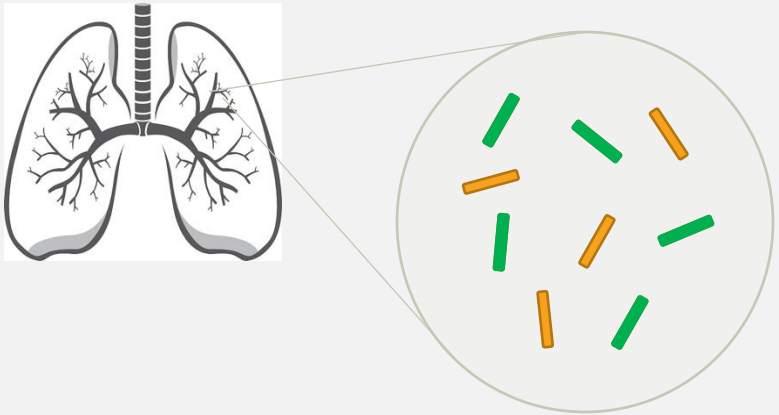


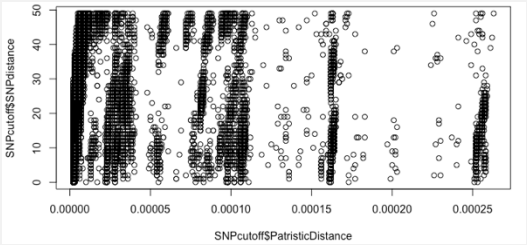
From Jain et. al. 2018

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Mixed infection

- Two or more concurrent, distinct strains or species present in a single individual
- A relatively common occurrence in clinical samples of bacteria, e.g., *M. tuberculosis* up to 21% reported (Micheni et al. 2022)
- Can be clinically important - determine treatment regimen when hetero-resistance present, minor strain transmission
- Important to account for when performing genomic and phylogenetic analysis

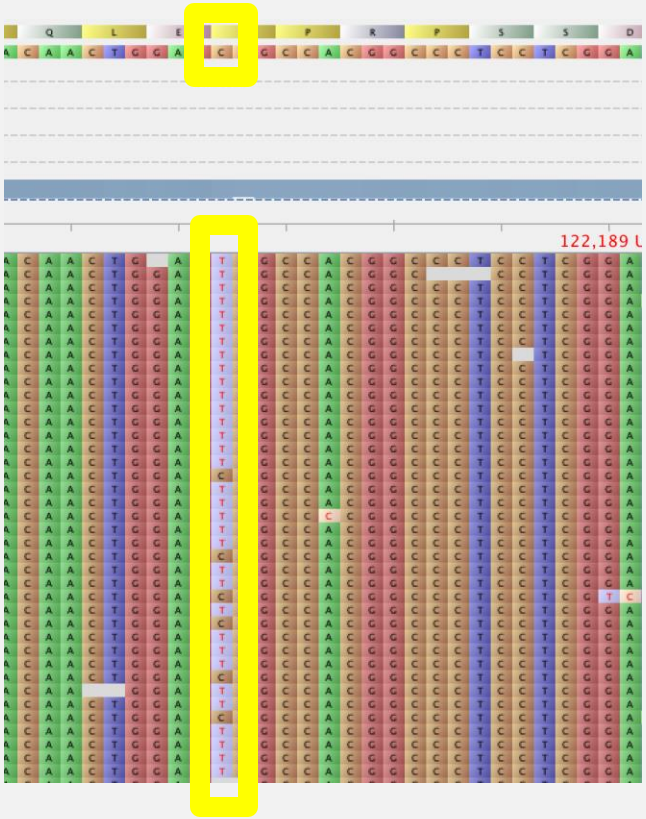




LECTURE 4: ADVANCED APPLICATIONS OF WGS

Mixed infection

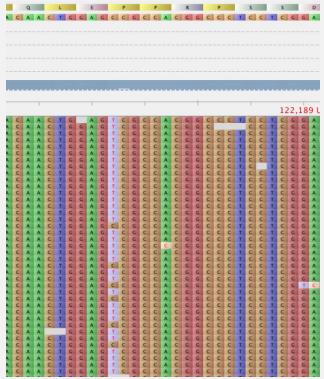
- Whole genome sequencing (WGS) can reveal the signatures of mixed TB infection (Cohen et al. 2012)
- Reference alignment reveals sites with 'heterozygous' calls - more than one allele at a given locus
- Important to distinguish between random noise and true signal of mixes

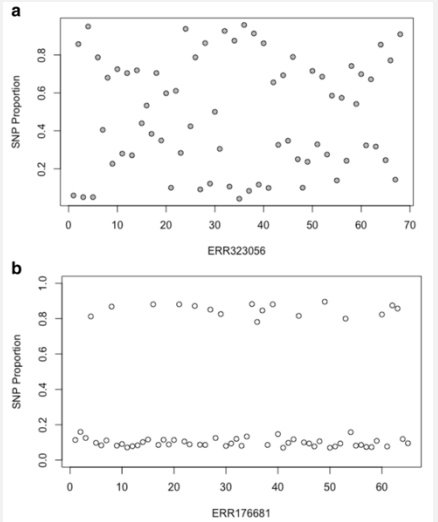


LECTURE 4: ADVANCED APPLICATIONS OF WGS

Mixed infection

- Clustering of read frequencies in mixed sites using Gaussian Mixture models (GMM) can identify mixed infection (Sobkowiak et al. 2018 & 2025, Wang et al. 2023)
- Sub-populations identified in the data consistent with mixed infection, compared with random noise from within-host variation
- Can also predict the constituent strains of mixed infections through assigning alleles in clusters to each strain (though limitations)

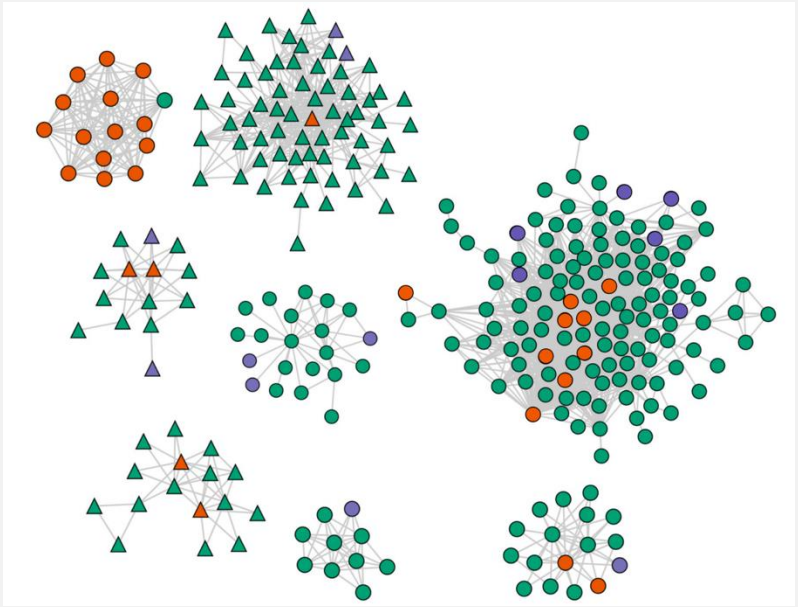




LECTURE 4: ADVANCED APPLICATIONS OF WGS

Mixed infection

- Developed method for detecting mixed infection and reconstructing constituent strains (Sobkowiak et. al. 2025)
- Applied to M. tuberculosis population from Moldova in 2018-2019
- Identified transmission clusters comprised of single strain infections and minor and major proportion constituent strains of mixed infection

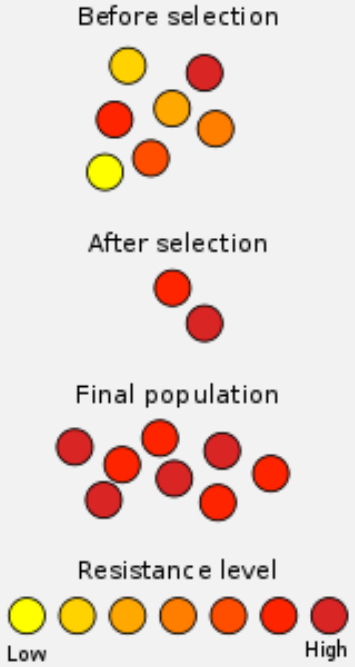


Sobkowiak et. al. 2025

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Testing for sites under selection

- Selection refers to the process by which certain heritable traits become more or less common in a population over time
- This process occurs because individuals with advantageous traits are more likely to survive and reproduce
- Selection can lead to adaptation and the evolution of new traits or the fixation of particular mutations



LECTURE 4: ADVANCED APPLICATIONS OF WGS

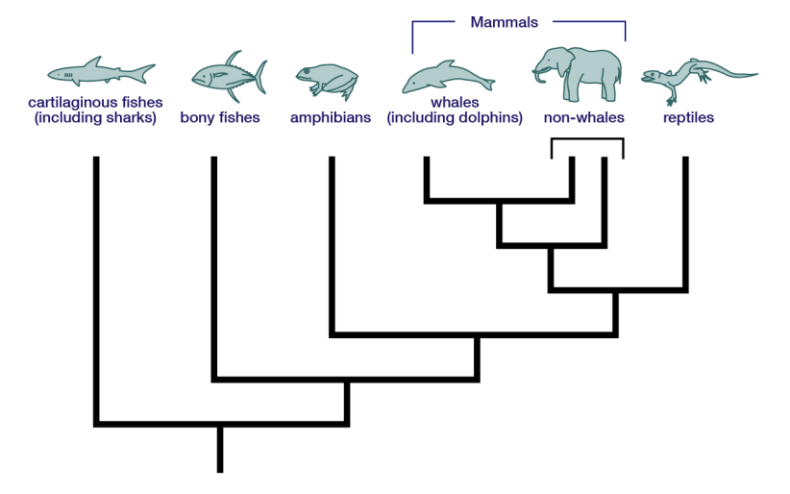
Population genetic tests for sites under selection

- **dN/dS ratio (Ka/Ks)**
  - nonsynonymous to synonymous substitution ratio; dN/dS > 1 suggests positive selection, <1 suggests purifying selection
- **Tajima's D**
  - Compares the number of segregating sites to nucleotide diversity - significantly positive/negative values suggest selection or demographic shifts
- **Fay and Wu's H**
  - Detects an excess of high-frequency derived alleles, a signal of recent positive selection

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Convergent evolution of genomic variants

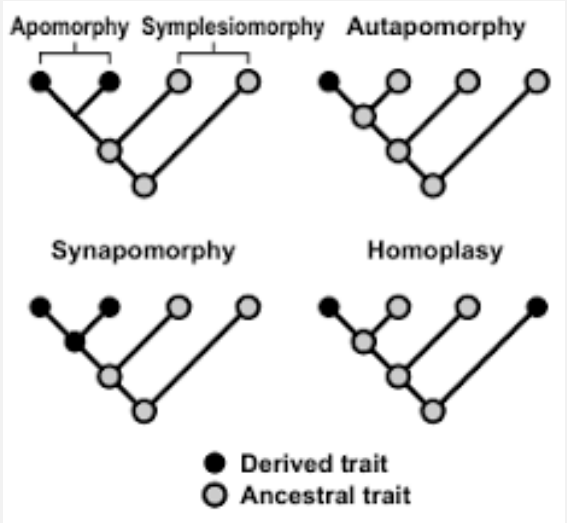
- The independent evolution of mutations or traits in distantly related individuals due to selective pressures or environmental constraints
- In genetics, sites under convergent evolution may represent instances of positive selection
- Mutations that enhance the fitness or adaptability of organisms to similar conditions



LECTURE 4: ADVANCED APPLICATIONS OF WGS

Homoplasy

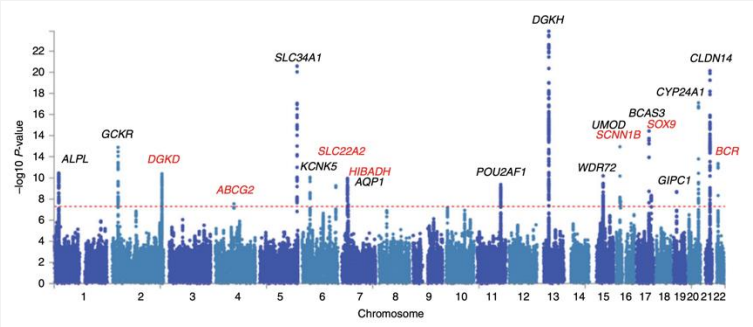
- Convergent evolution can cause homoplasies on a phylogenetic tree
- Shared traits or mutations among taxa that do not accurately reflect their evolutionary relationships
- Most methods use parsimony to infer the most likely ancestral states at internal nodes of the tree
- Identify branches where reversals or independent acquisitions occur, indicating homoplasy



LECTURE 4: ADVANCED APPLICATIONS OF WGS

Genome Wide Association Studies

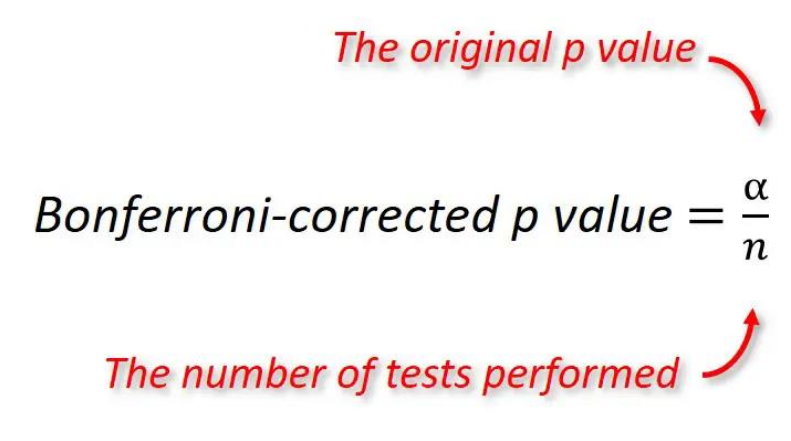
- GWAS is a test to identify genetic variants that are associated with a particular trait
- May be mutations that cause antibiotic resistance, increase virulence or transmissibility, or evolve with host adaptation
- Advancements in sequencing technologies and bioinformatics have significantly enhanced the ability to conduct GWAS in microbial populations



LECTURE 4: ADVANCED APPLICATIONS OF WGS

Genome Wide Association Studies

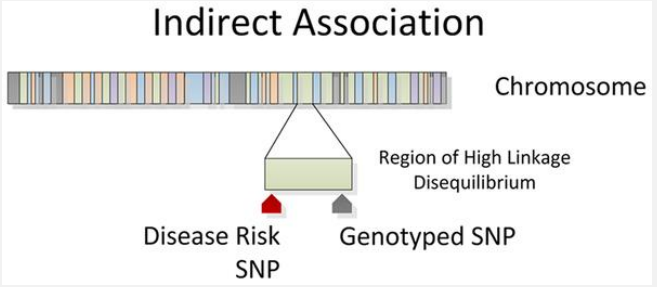
- GWAS statistically analyzes genetic variants to identify any correlations with specific traits or diseases
- Multiple testing correction (such as Bonferroni) is applied to account for potentially thousands of sites being tested
- Also need to account for population structure to remove the confounding effect of genetic substructure in the population



LECTURE 4: ADVANCED APPLICATIONS OF WGS

Genome Wide Association Studies

- GWAS can be complicated in microbial populations for the following reasons:
  - Complex population structures, including clonal lineages, recombination events
  - Linkage disequilibrium, where alleles at different loci are inherited together due to limited recombination leading to spurious associations
  - Causal Inference, traits can be caused by multiple genes and environmental factors

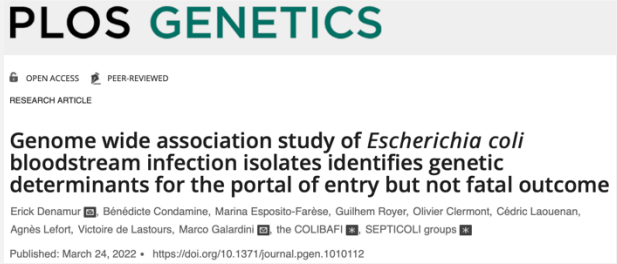


From Bush & Moore, 2012

LECTURE 4: ADVANCED APPLICATIONS OF WGS

Genome Wide Association Studies

- GWAS has led to the discovery of novel variants associated with a variety of traits in a range of pathogens



PLOS GENETICS

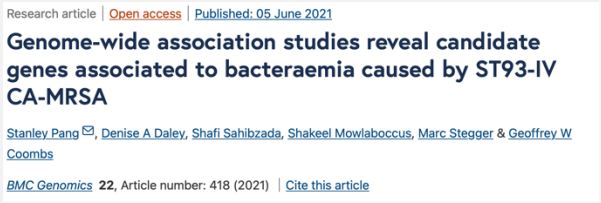
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Genome wide association study of *Escherichia coli* bloodstream infection isolates identifies genetic determinants for the portal of entry but not fatal outcome

Erick Denamur, Bénédicte Condamine, Marina Esposito-Farèse, Guilhem Royer, Olivier Clermont, Cédric Laouenan, Agnès Lefort, Victoire de Lastours, Marco Galardini, the COLIBAFI, SEPTICOLI groups

Published: March 24, 2022 • https://doi.org/10.1371/journal.pgen.1010112

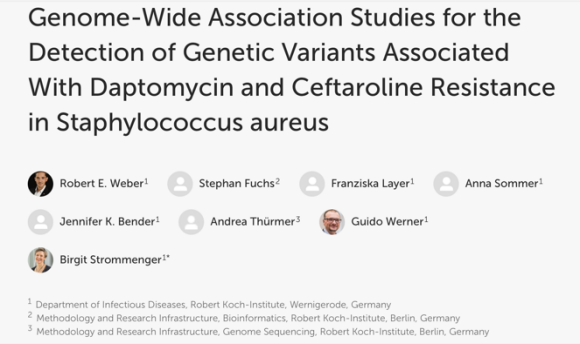


Research article | Open access | Published: 05 June 2021

Genome-wide association studies reveal candidate genes associated to bacteraemia caused by ST93-IV CA-MRSA

Stanley Pang, Denise A Daley, Shafi Sahibzada, Shakeel Mowlaboccus, Marc Stegger & Geoffrey W Coombs

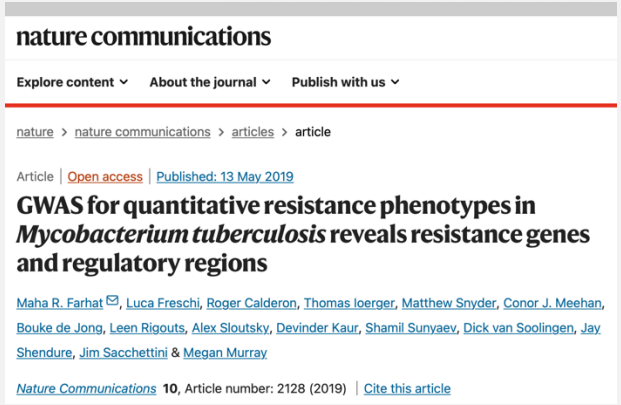
BMC Genomics **22**, Article number: 418 (2021) | Cite this article



Genome-Wide Association Studies for the Detection of Genetic Variants Associated With Daptomycin and Ceftaroline Resistance in Staphylococcus aureus

Robert E. Weber[1], Stephan Fuchs[2], Franziska Layer[1], Anna Sommer[1], Jennifer K. Bender[1], Andrea Thürmer[3], Guido Werner[1], Birgit Strommenger[1*]

[1] Department of Infectious Diseases, Robert Koch-Institute, Wernigerode, Germany
[2] Methodology and Research Infrastructure, Bioinformatics, Robert Koch-Institute, Berlin, Germany
[3] Methodology and Research Infrastructure, Genome Sequencing, Robert Koch-Institute, Berlin, Germany



nature communications

Explore content | About the journal | Publish with us

nature > nature communications > articles > article

Article | Open access | Published: 13 May 2019

GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions

Maha R. Farhat, Luca Freschi, Roger Calderon, Thomas Ioerger, Matthew Snyder, Conor J. Meehan, Bouke de Jong, Leen Rigouts, Alex Sloutsky, Devinder Kaur, Shamil Sunyaev, Dick van Soolingen, Jay Shendure, Jim Sacchettini & Megan Murray

Nature Communications **10**, Article number: 2128 (2019) | Cite this article

PRACTICAL 5, 6 AND 7

Practical session 5: Mixed Infection, Recombination and ANI

Practical session 6: Phylogeography and Phylodynamics

Practical session 7: Fitness and Selection