Biochem 3BP3

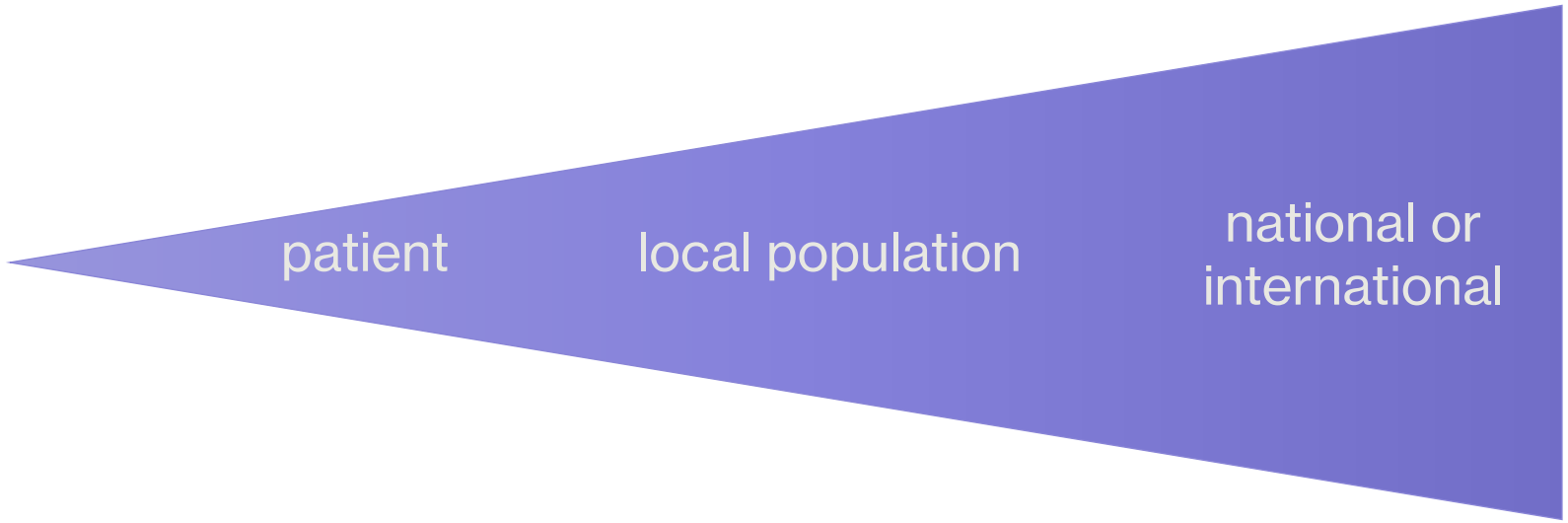
Molecular Epidemiology

Week of Nov 1, 2021

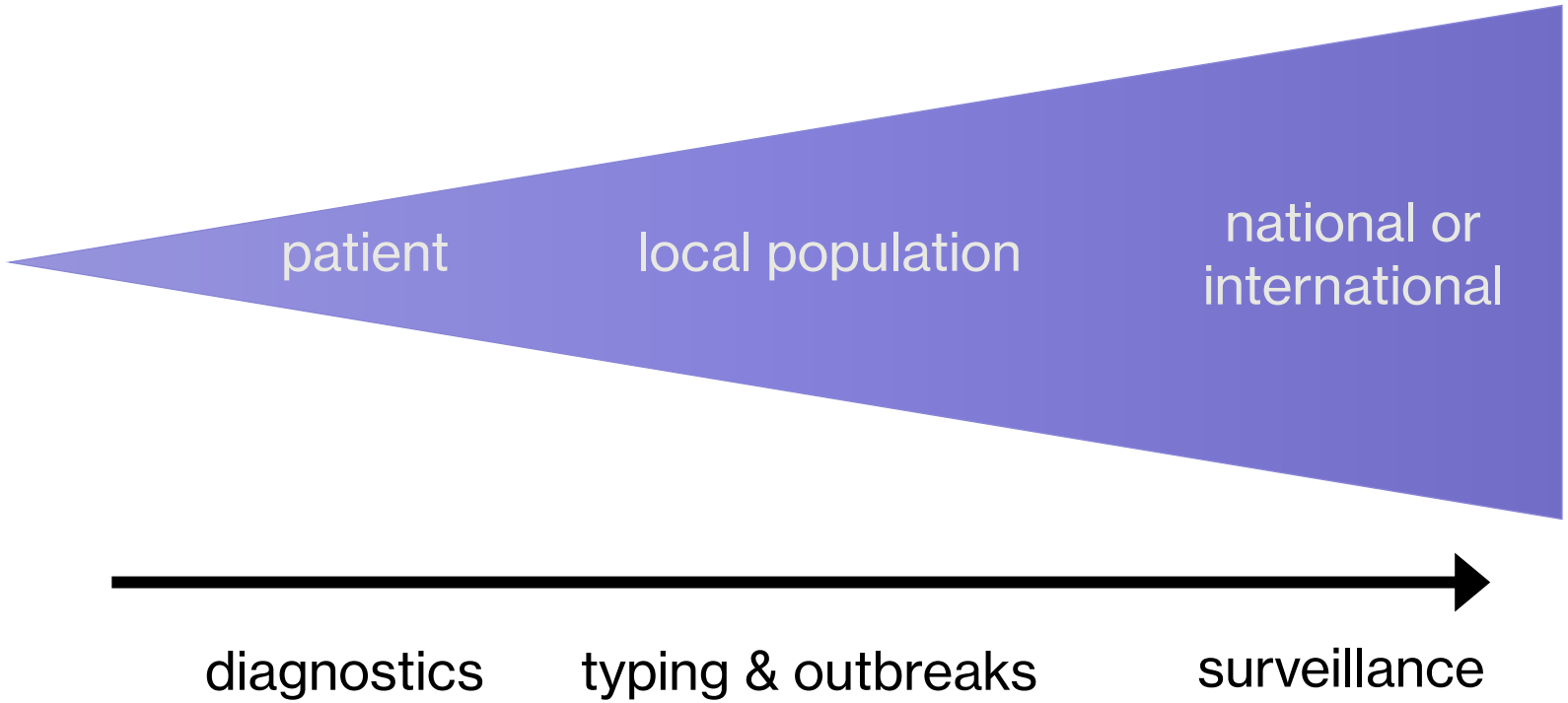
Molecular Epidemiology

- In the broadest sense - examines potential genetic and environmental risk factors, identified at the molecular level, to the etiology, distribution and prevention of disease
- In human disease – cohort studies, longitudinal studies, clinical metadata, exome sequencing, genome-wide association studies (GWAS)
- In animal health – maintain animal productivity & quality of life, determine pathogen & environmental impacts on health, food chain safety, environmental safety
- In infectious disease – our focus this week
  - track the movement, prevalence, and origin of pathogens
  - guide treatment, public health response
  - screen for virulence or drug resistance determinants

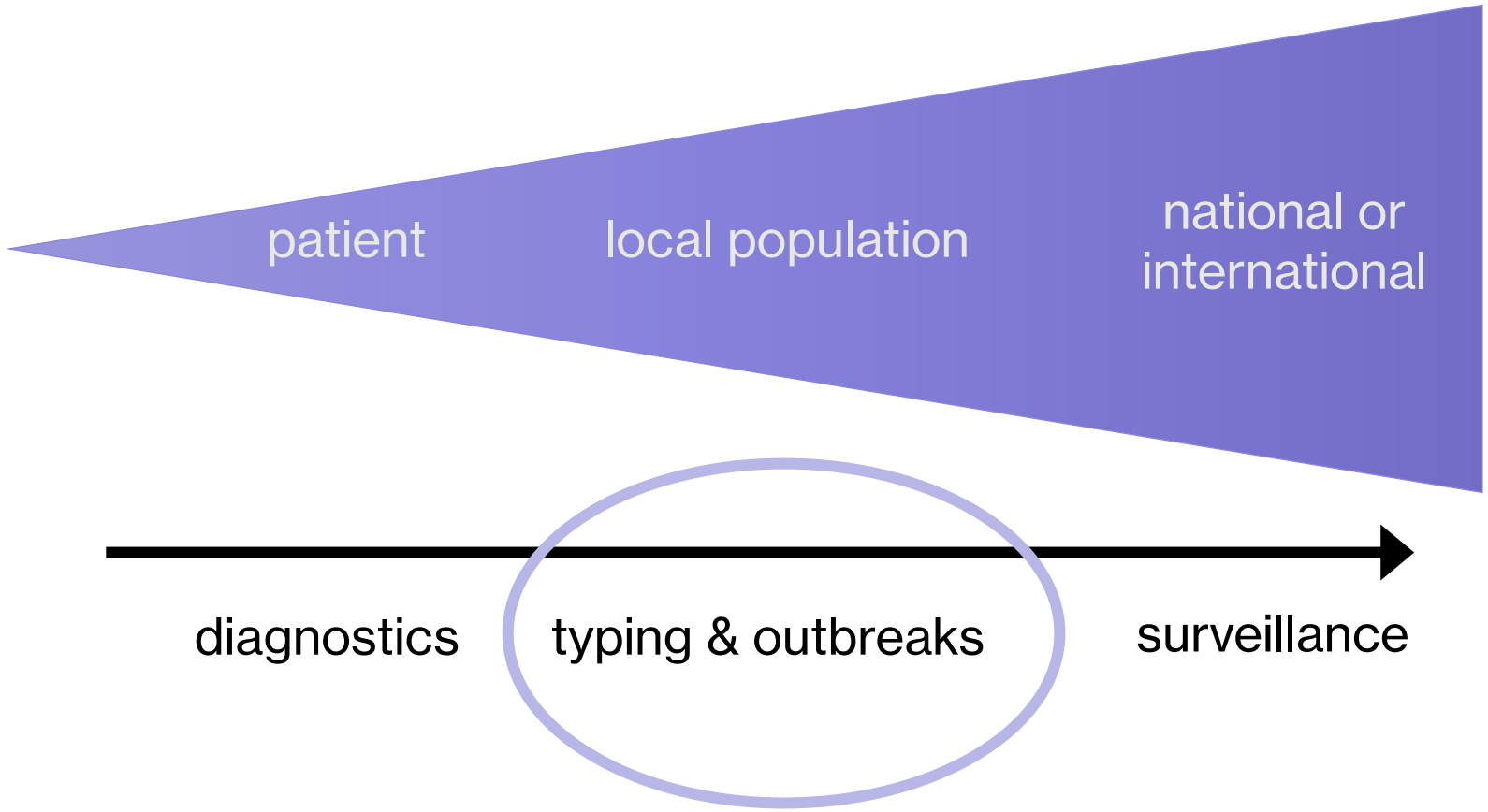
Molecular Epidemiology



Molecular Epidemiology



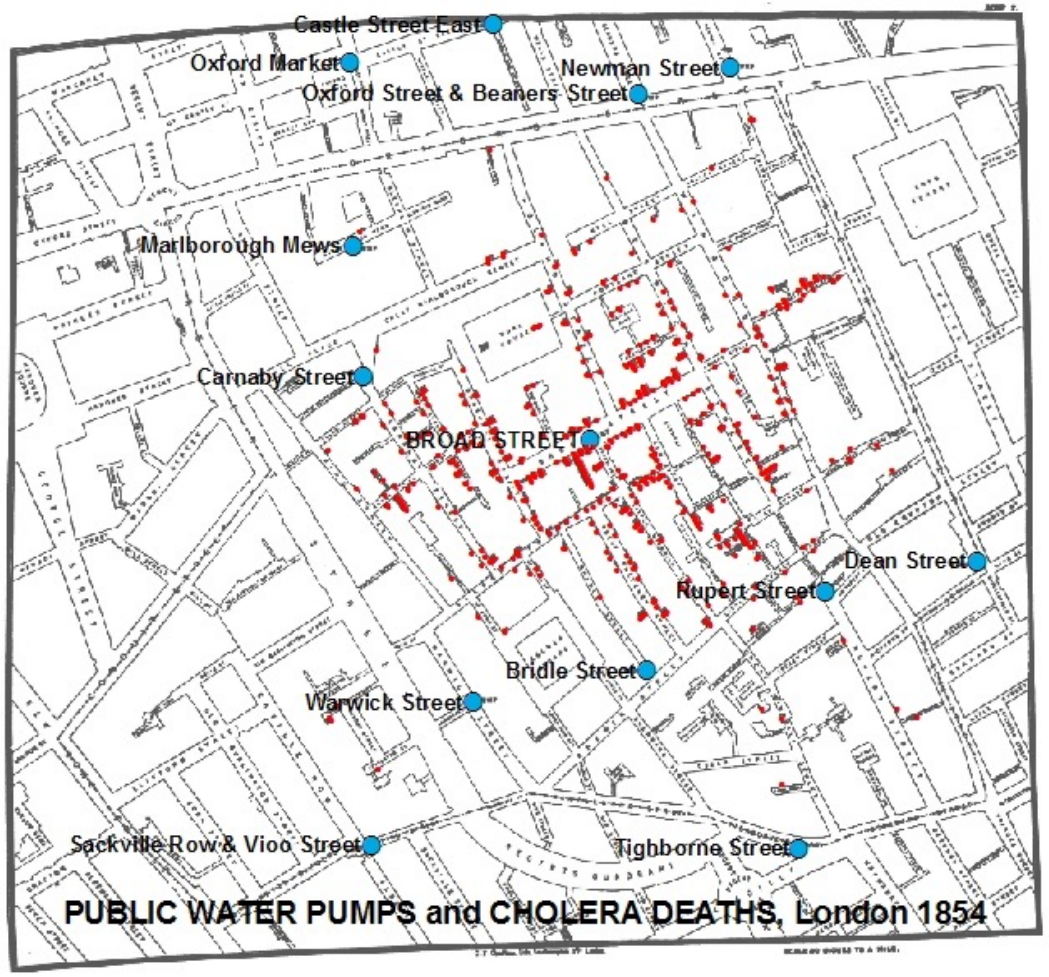
Molecular Epidemiology



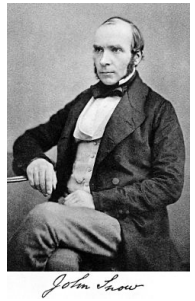
Molecular Epidemiology of Pathogens

- **Who** – which organisms are present?
  - How do we tell them apart?
- **What** – what are they doing?
  - Antibiotic resistance?
  - Virulence?
  - What's happening in their genomes?
- **Where** – where were they sampled?
  - Phylo/biogeography?
  - Route of transmission?
- **When** – when were they sampled?
- **How** – based on the patterns, can we explain how they got there?

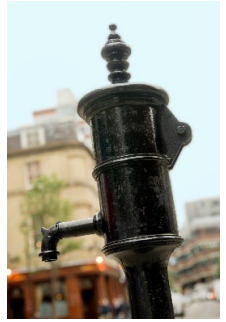
Origin of Epidemiology – Cholera in London



https://ralucanicola.github.io/cholera-map-3D



John Snow



Broad Street Pump

Motivation

Clinical (i.e. patient level)

When an infection is suspected, it can take up to 72 hours to culture from a swab – precious time during which:

- Inappropriate antibiotics are administered
- Condition can worsen
- Unnecessary exposure
- Expensive hospital isolation
- It's now possible (and cheap enough) to directly sequence clinical samples

Public Health (i.e. community level)

- Historical migration patterns, humans and animals
- Identify source of pathogen or contamination of the food chain
- Is there a point source or are cases unrelated?
- **Existing tools lack resolution that DNA sequencing can provide**

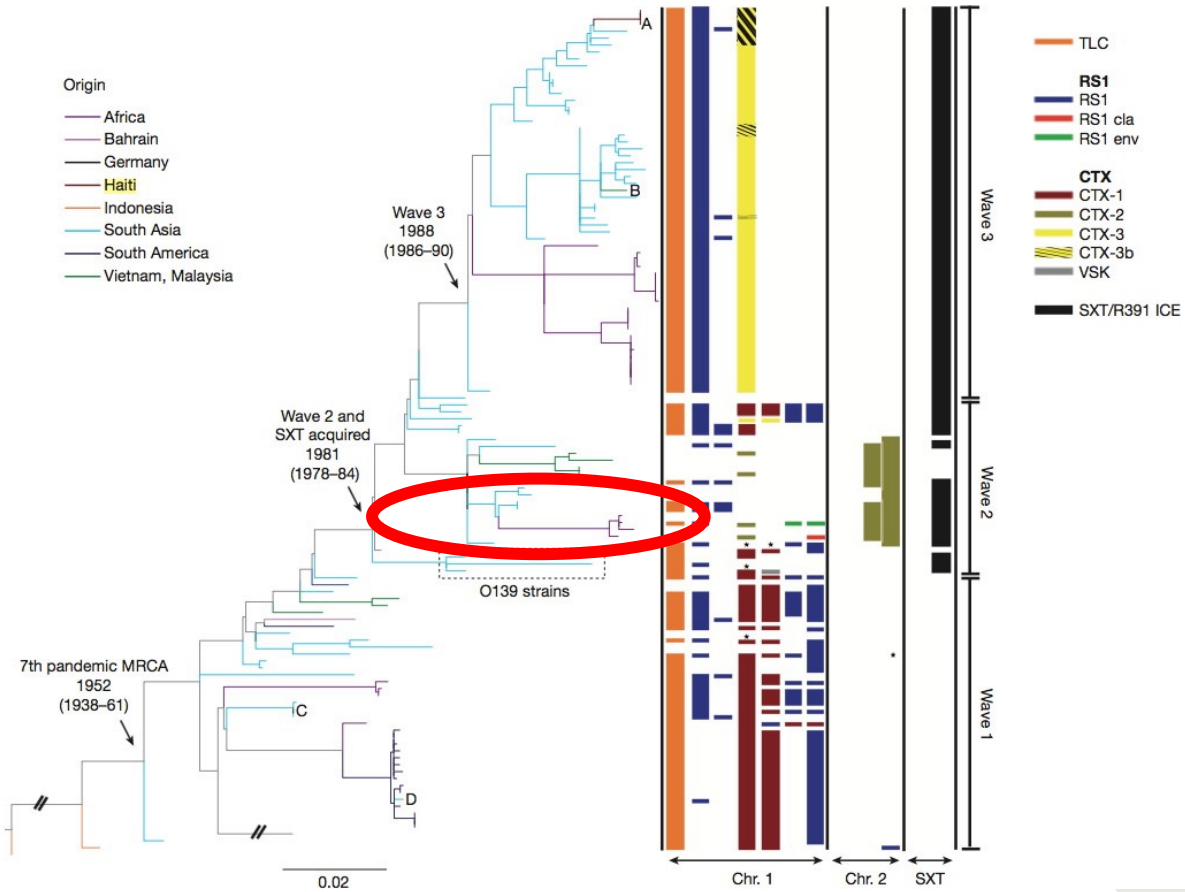
Cholera in Haiti after the 2010 Earthquake

- Cholera is caused by the waterborne bacterium *Vibrio cholerae*
- Prior to 2010, Haiti had not had a cholera outbreak in recorded history
- Post-earthquake, 100,000+ infections & ~7,000 deaths

Cholera in Haiti after the 2010 Earthquake

- Cholera is caused by the waterborne bacterium *Vibrio cholerae*
- Prior to 2010, Haiti had not had a cholera outbreak in recorded history
- Post-earthquake, 100,000+ infections & ~7,000 deaths

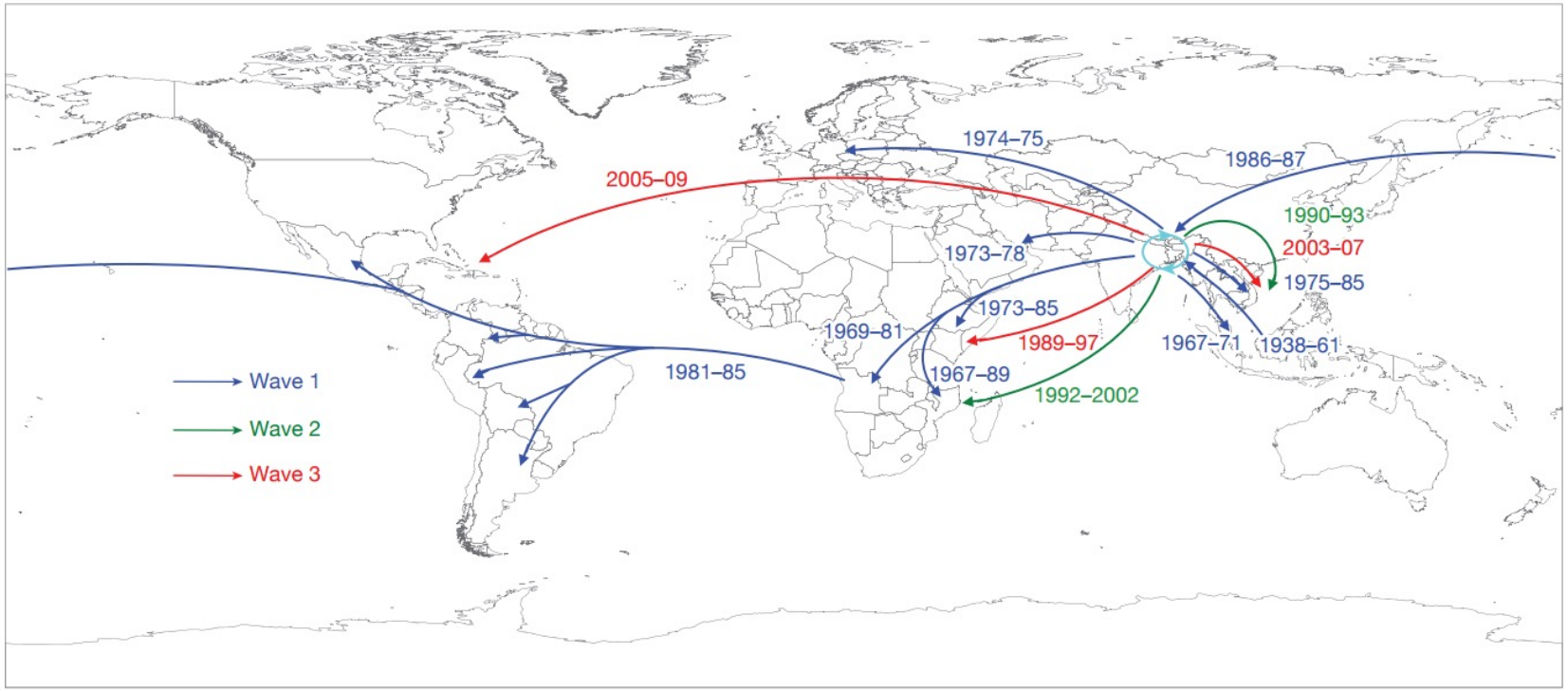
- Sequencing of genomes to determine relationships among *V. cholerae* strains
- Nucleotide difference among strains used in a RAxML analysis
- Haiti strains had their origins in South Asia, likely via Bangladeshi relief workers



Mutreja *et al*. 2011. *Nature* 477: 462-5

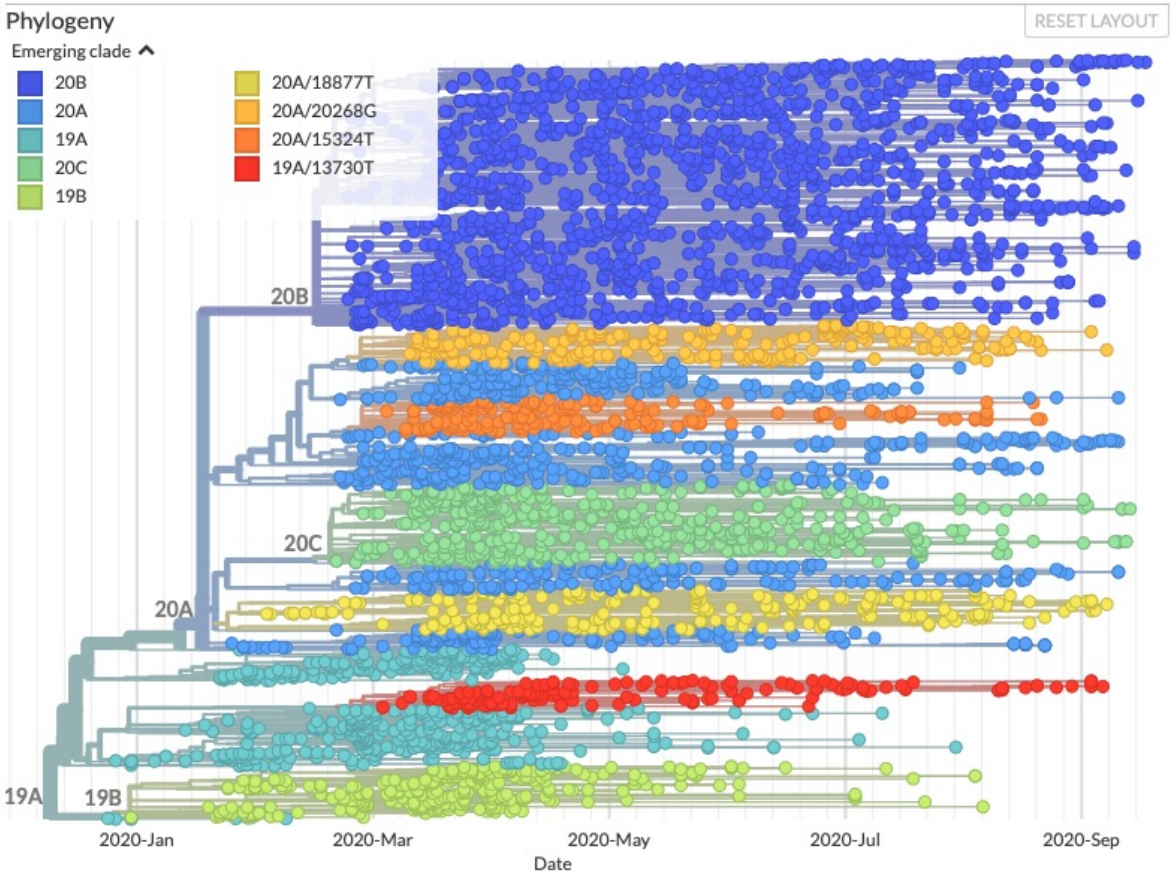
Cholera in Haiti after the 2010 Earthquake

- Cholera is caused by the waterborne bacterium *Vibrio cholerae*
- Prior to 2010, Haiti had not had a cholera outbreak in recorded history
- Post-earthquake, 100,000+ infections & ~7,000 deaths
- DNA sequencing can provide a global perspective on disease

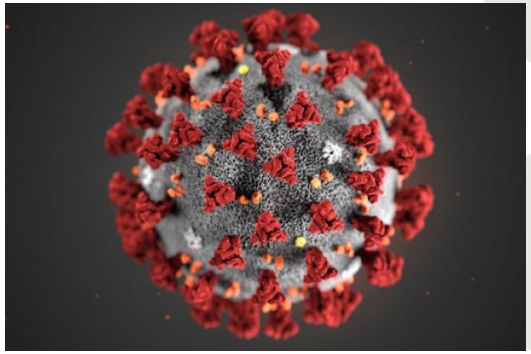


Mutreja *et al*. 2011. *Nature* 477: 462-5

COVID-19



https://nextstrain.org

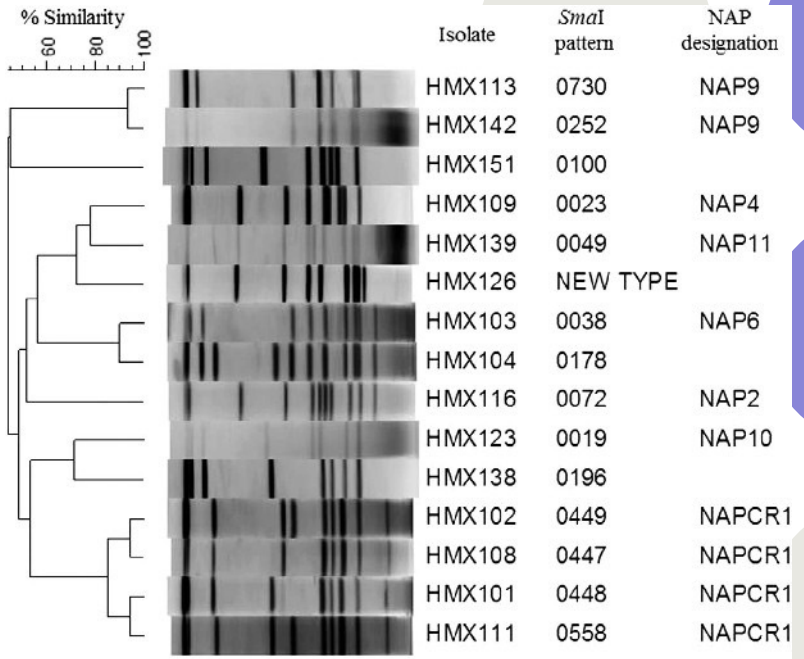


Improving Methods, Improving Resolution

Are these samples the same strain?

- The first widely used molecular tool was Pulse Field Typing
  - DNA isolated and digested with a rare cutting restriction enzyme
  - Fragments separated in an acrylamide gel
  - Similar pattern assumed to equal related strains

C. difficile in Costa Rica



López-Urēna *et al*. 2016. *Emerg. Microbes Infect*. 5:e42

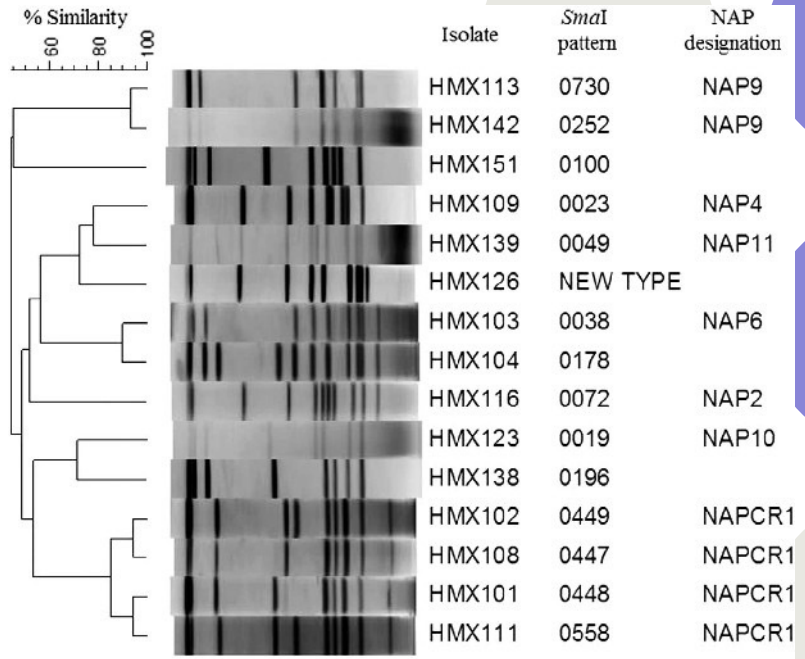
Improving Methods, Improving Resolution

Are these samples the same strain?

- The first widely used molecular tool was Pulse Field Typing
  - DNA isolated and digested with a rare cutting restriction enzyme
  - Fragments separated in an acrylamide gel
  - Similar pattern assumed to equal related strains

- Very low resolution – an entire genome of information reduced to a few restriction fragments
- All isolates in an outbreak look the same – no resolution at all
- Fast & cheap & still a useful indicator, e.g. NAP1 in *C. difficile* is a dangerous hyper-virulent

C. difficile in Costa Rica



López-Urēna *et al*. 2016. *Emerg. Microbes Infect*. 5:e42

Improving Methods, Improving Resolution

Are these samples the same strain?

- A more recent advance is Multiple Locus Sequencing Typing (MLST)
- PCR & sequencing of 7 housekeeping genes:
  - The seven sequences form a typing ‘fingerprint’
  - PubMLST is a repository of MLST fingerprints
  - PCR is being replaced with genome sequencing to determine MLST

adk
atpA
dxr
glyA
recA
sodA
tpi

C. difficile in Ontario (McArthur lab, St. Joseph’s Healthcare, Public Health Ontario)

MLST	Hamilton	Peterborough
Type 1	79 isolates	38 isolates
Type 2	15 isolates	9 isolates
Type 54	5 isolates	12 isolates
Type 8	15 isolates	1 isolate
Type 10	8 isolates	6 isolates
Type 58	9 isolates	
Type 42	5 isolates	2 isolates
other	37 isolates / 22 MLSTs	20 isolates / 10 MLSTs

Improving Methods, Improving Resolution

Are these samples the same strain?

- A more recent advance is Multiple Locus Sequencing Typing (MLST)
- PCR & sequencing of 7 housekeeping genes:
  - The seven sequences form a typing ‘fingerprint’
  - PubMLST is a repository of MLST fingerprints
  - PCR is being replaced with genome sequencing to determine MLST
- MLST provides a higher epidemiological resolution than North American pulsed-field (NAP) typing
- MLST sequences can be used for phylogenetic analysis but not a lot of information in the data
- Can be scaled up to all core genes, e.g cgMLST of *Salmonella* uses 330 genes

C. difficile in Ontario (McArthur lab, St. Joseph’s Healthcare, Public Health Ontario)

MLST	Hamilton		Peterborough
Type 1	79 isolates	NAP1	38 isolates
Type 2	15 isolates	NAP4	9 isolates
Type 54	5 isolates		12 isolates
Type 8	15 isolates		1 isolate
Type 10	8 isolates		6 isolates
Type 58	9 isolates	NAP4, NAP7, NAP 10	
Type 42	5 isolates		2 isolates
other	37 isolates / 22 MLSTs		20 isolates / 10 MLSTs

Improving Methods, Improving Resolution

Are these samples the same strain?

- Whole Genome Single Nucleotide Variant Analysis
- Use DNA sequence variation throughout the entire draft genome sequence of all isolates to determine high resolution strain relationships

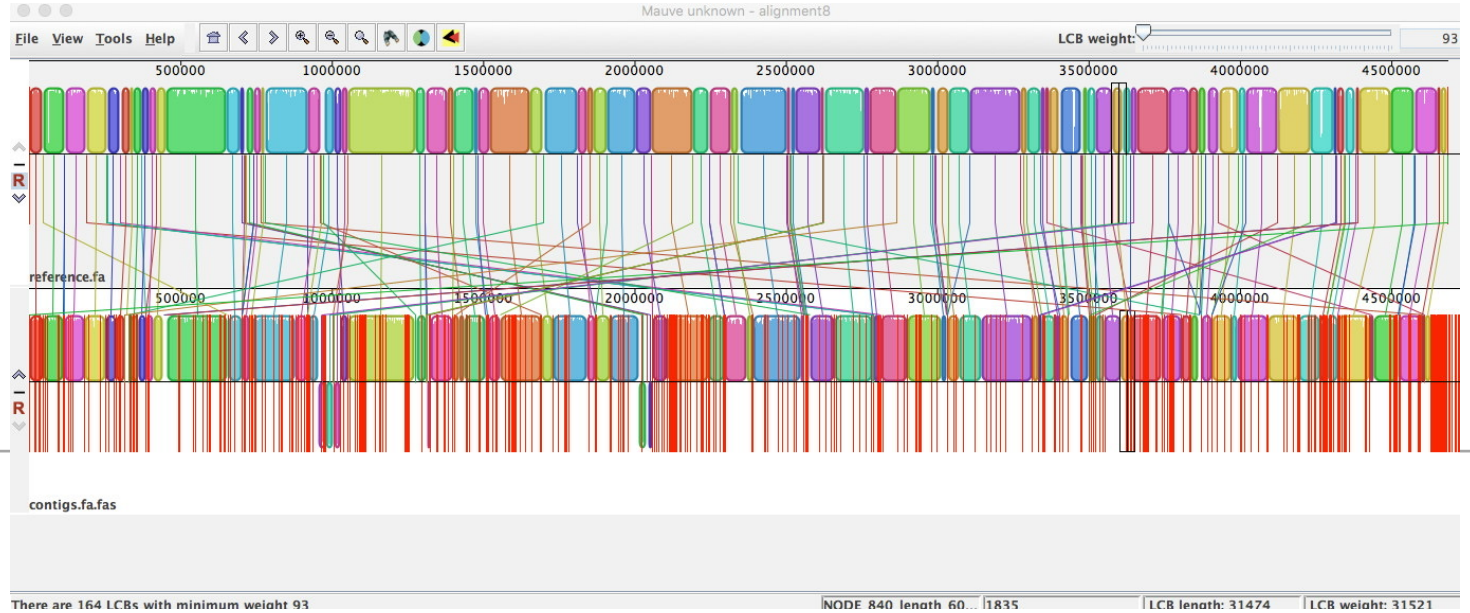
- Use the PARSNP algorithm to:
  - Align draft genomes to a reference genome sequence
  - Extract genome locations with sequence variation among isolates
  - Detect and exclude genome regions involved in horizontal gene transfer or recombination using the PhiPack algorithm

- Reconstruct the phylogenetic history of the isolates using the RAxML algorithm

Improving Methods, Improving Resolution

Whole Genome Single Nucleotide Variant Analysis

```
mlst1_SJHH_2015_210          TATTGATAGTGGTATAACTCGCACTTGGACCAAATTTATCGCTAACAGAATAAAATATTGTAGTTGTCGTATTTTCGGTAATTACCCTAGACGTAGAA
mlst1_SJHH_2015_218          TATTGATAGTGGTATAACTTGCTCTTGGACCAACTTTACCGCTAACAGAATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_noloc_noyear_4         TATTGATAGTGGTATAACTCGCACTTGGACCAAATTTATCGTTAACAAAATAAAATATTGTATACGTCGTATTTTCGGTAATTATCTTGGACGTAGCA
mlst1_noloc_noyear_38        TATTGATAGTGGTATAACTCGCTCTTGGATCAACTTTATCGCTAACAGGATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_SJHH_2010_255          TATTGATAGTGGTATAACTCGCTCTTGGATCAACTTTATCGCTAACAGAATAAAATATTGTATTCACTACTTTTTCGGTAATTATCCTAGACGTAGAA
mlst1_noloc_noyear_24        TATTGATAGTGGTATAACTCGCTCTTGGATCAACTTTATCGCTAACAGAATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_SJHH_2010_267          TATTGATAGTGGTATAACTCGCTCTTGGATCAACTTTATCGCTAACAGAATAAAATATTGTATTCACTACTTTTTCGGTAATTATCCTAGACGTAGAA
mlst1_SJHH_2013_200          TATTGATAGTGGTATAACTCGCTCTTGGACCAACTTTATCGCTAACAGAATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_noloc_noyear_6         TATTGATAGCCGCATAACTCGCACTTGGACCAAATTTATCGCTAACAGAGTAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGAAGAA
mlst1_SJHH_2011_269          TATTGATAGTGGTATAACTCGCTCTTGGATCAACTTTATCGCTAACAGAATAAAATATTGTATTCACTACTTTTTCGGTAATTATCCTAGACGTAGAA
```

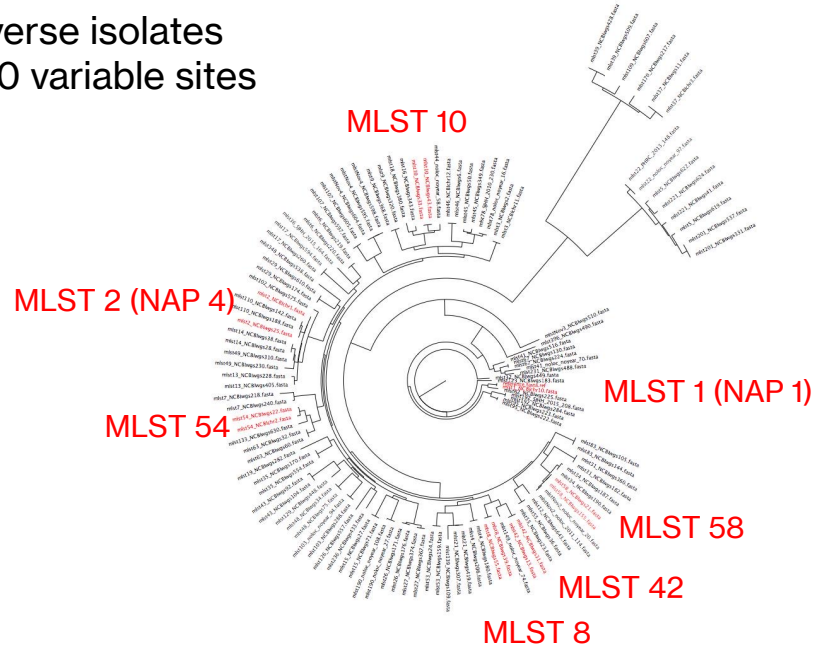


Improving Methods, Improving Resolution

Whole Genome Single Nucleotide Variant Analysis

```
mlst1_SJHH_2015_210          TATTGATAGTGGTATAACTCGCACTTGGACCAAATTTATCGCTAACAGAATAAAATATTGTAGTTGTCGTATTTTCGGTAATTACCCTAGACGTAGAA
mlst1_SJHH_2015_218          TATTGATAGTGGTATAACTTGCTCTTGGACCAACTTTACCGCTAACAGAATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_noloc_noyear_4         TATTGATAGTGGTATAACTCGCACTTGGACCAAATTTATCGTTAACAAAATAAAATATTGTATACGTCGTATTTTCGGTAATTATCTTGGACGTAGCA
mlst1_noloc_noyear_38        TATTGATAGTGGTATAACTCGCTCTTGGATCAACTTTATCGCTAACAGGATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_SJHH_2010_255          TATTGATAGTGGTATAACTCGCTCTTGGATCAACTTTATCGCTAACAGAATAAAATATTGTATTCACTACTTTTTCGGTAATTATCCTAGACGTAGAA
mlst1_noloc_nc                                                                                AATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_SJHH_201                                                                                AATAAAATATTGTATTCACTACTTTTTCGGTAATTATCCTAGACGTAGAA
mlst1_SJHH_201                                                                                AATAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGTAGAA
mlst1_noloc_nc                                                                                AGTAAAATATTGTATTCGTCGTATTTTCGGTAATTATCCTAGACGAAGAA
mlst1_SJHH_201                                                                                AATAAAATATTGTATTCACTACTTTTTCGGTAATTATCCTAGACGTAGAA
```

McArthur *C. difficile* study
121 diverse isolates
90,040 variable sites



- Whole genome SNP analysis provides very high resolution among strains – lots of data!
- Superior to Pulse Field, MLST, or cgMLST
- Relies on well established phylogenetic methods you have been learning

Horizontal Gene Transfer

Detect and exclude genome regions involved in horizontal gene transfer or recombination using the PhiPack algorithm

Horizontal Gene Transfer

Detect and exclude genome regions involved in horizontal gene transfer or recombination using the PhiPack algorithm

HOMOLOGOUS

A homologous trait is shared between two species because they inherited it from a common ancestor.

Information from homologous traits can be used to infer evolutionary relationships.

Non-homologous traits do not reflect evolutionary history but instead convergence. They can mislead inference of evolutionary relationships. Example: octopus eye versus human eye.

Horizontal Gene Transfer

Detect and exclude genome regions involved in horizontal gene transfer or recombination using the PhiPack algorithm

HOMOLOGOUS

A homologous trait is shared between two species because they inherited it from a common ancestor.

Information from homologous traits can be used to infer evolutionary relationships.

Non-homologous traits do not reflect evolutionary history but instead convergence. They can mislead inference of evolutionary relationships. Example: octopus eye versus human eye.

- Like any other phylogenetic analysis, we need our sampled nucleotide variation to be homologous, otherwise we will get the wrong results
- Transposable elements, plasmids, genomic islands all represent horizontal gene transfer
- Two general ways to exclude these from analyses:
  - Only use identified core genes that are essential and uninvolved in horizontal gene transfer, e.g. core genes identified using the chewBACCA algorithms
  - Exclude regions with SNP density, which is a signature of HGT, using the PhiPack or related algorithms

C. difficile Example...

Fecal samples

↓ *culture*

Isolates

↓ *DNA extraction & Illumina sequencing*

raw DNA sequencing results

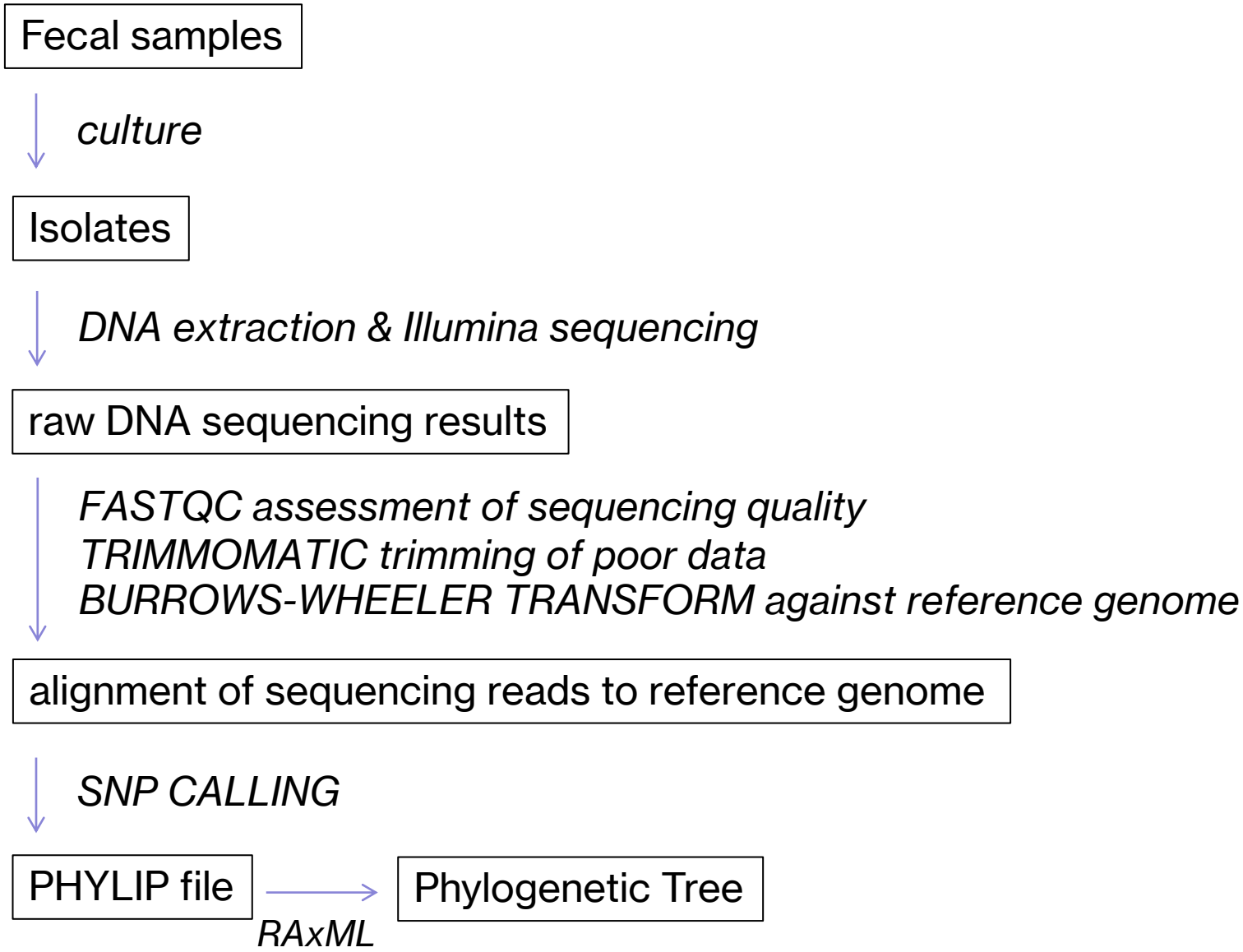
↓ *FASTQC assessment of sequencing quality*
TRIMMOMATIC trimming of poor data
UNICYCLER genome assembly to create a draft genome sequence

whole genome shotgun assemblies

↓ *KRAKEN taxonomic assessment of draft genome sequences**

workable draft genome sequences → *PARSNP* → PHYLIP file → *RAxML* → Phylogenetic Tree

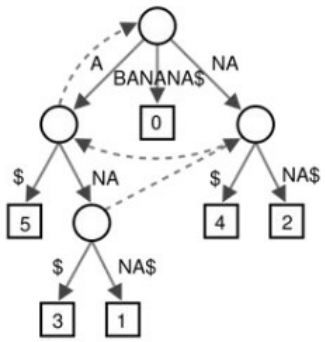
Skipping Genome Assembly...



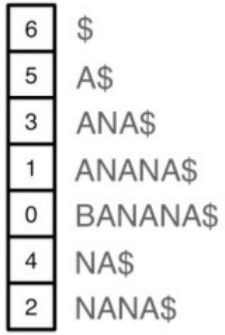
Short Read Alignment / Mapping

- Short reads (supposedly) have low error rates, especially after trimming and filtering
- The goal is to accurately align the short sequencing reads to the reference genomes
- Earlier techniques used a tweaked version of seed + extend (like BLAST) but they did not scale to NGS – too slow and too much memory!
- The two heavyweight short read aligners are BWA and Bowtie2 but both use a technique called Burrows-Wheeler Transform (BWT) to generate data structures that are:
  - Space/memory efficient
  - Suffix sorted, indexed, fast
  - Capable of gapped, inexact search mapping (can be slow – a tradeoff)

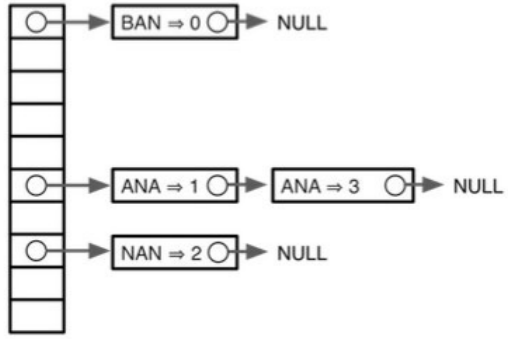
Short Read Alignment / Mapping



Suffix tree



Suffix array

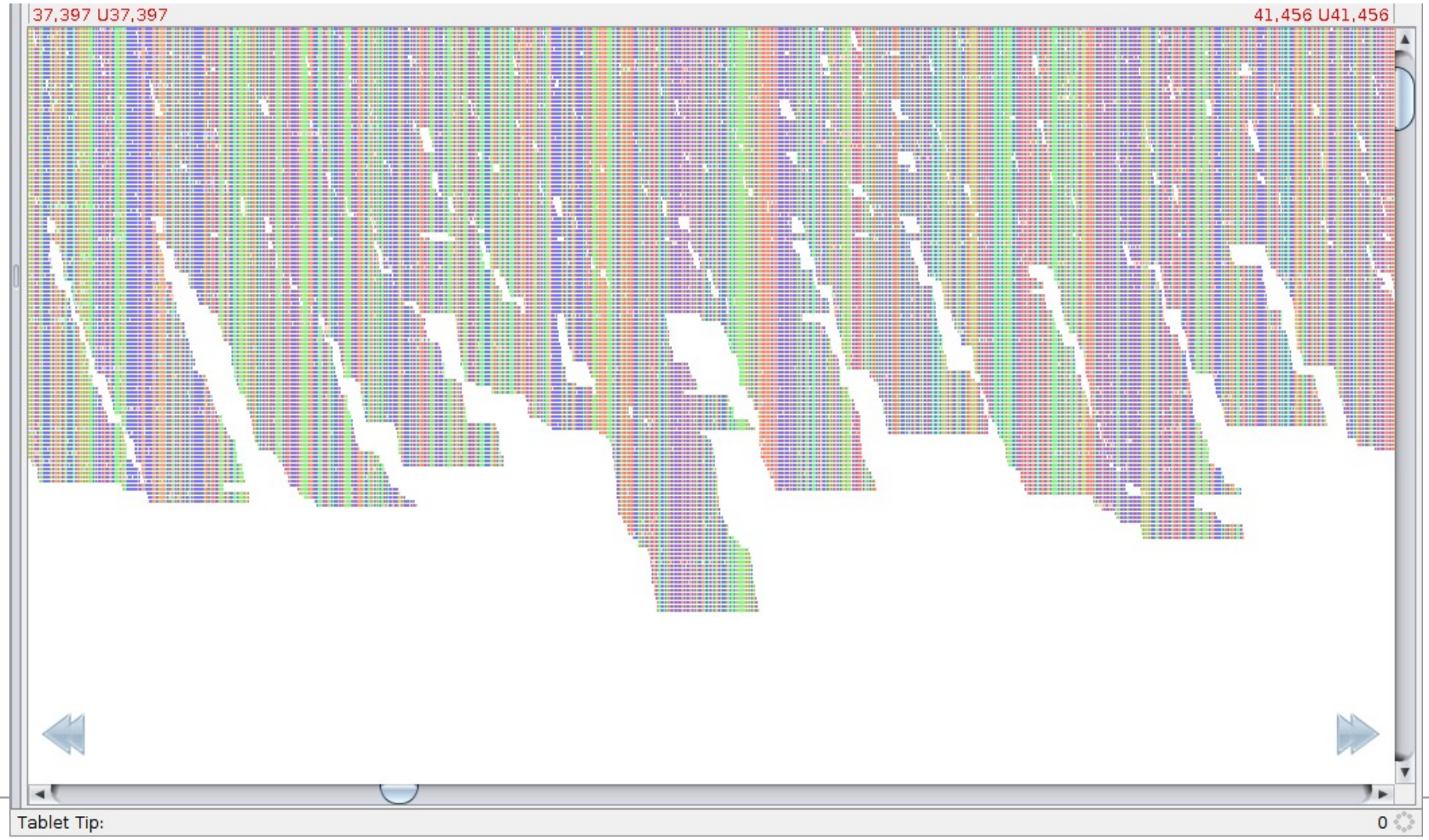


Seed hash tables

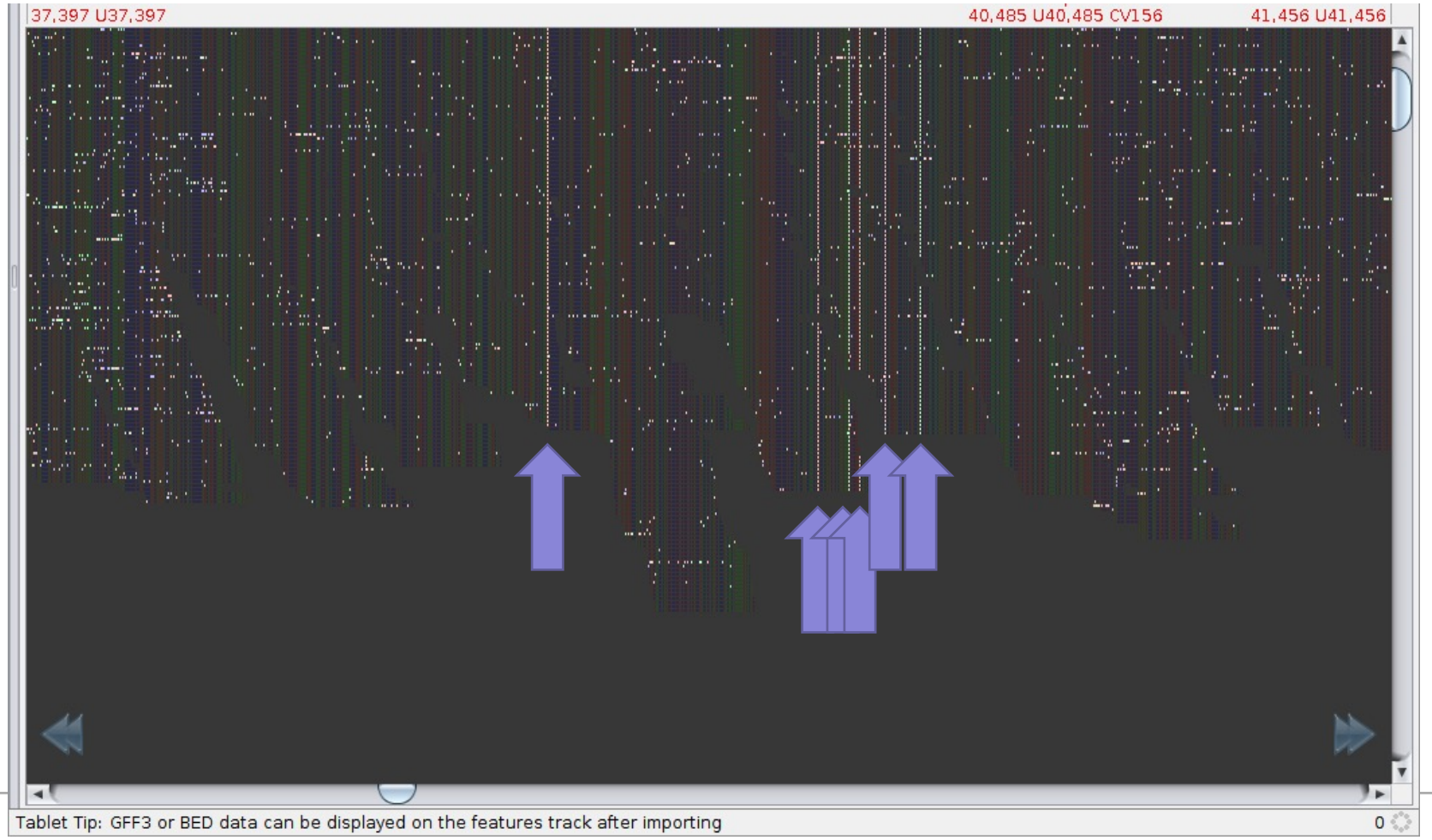
Many variants, incl. spaced seeds

- Due to the very large amount sequencing reads that need to be aligned to the reference genome, BWT uses some advanced computer science:
  - Data compression & indexing
  - “Suffixes” – like BLAST words or K-mers but with an additional emphasis upon the last letter of the sequencing read
  - “Suffix Trees” and reversible, cyclic permutation
- Take home message: BWT is not like BLAST – it emphasizes mapping due to high similarity only! The result is high resolution mapping of reads to reference (default is only tolerant up to ~4% nucleotide divergence)

Short Read Alignment / Mapping



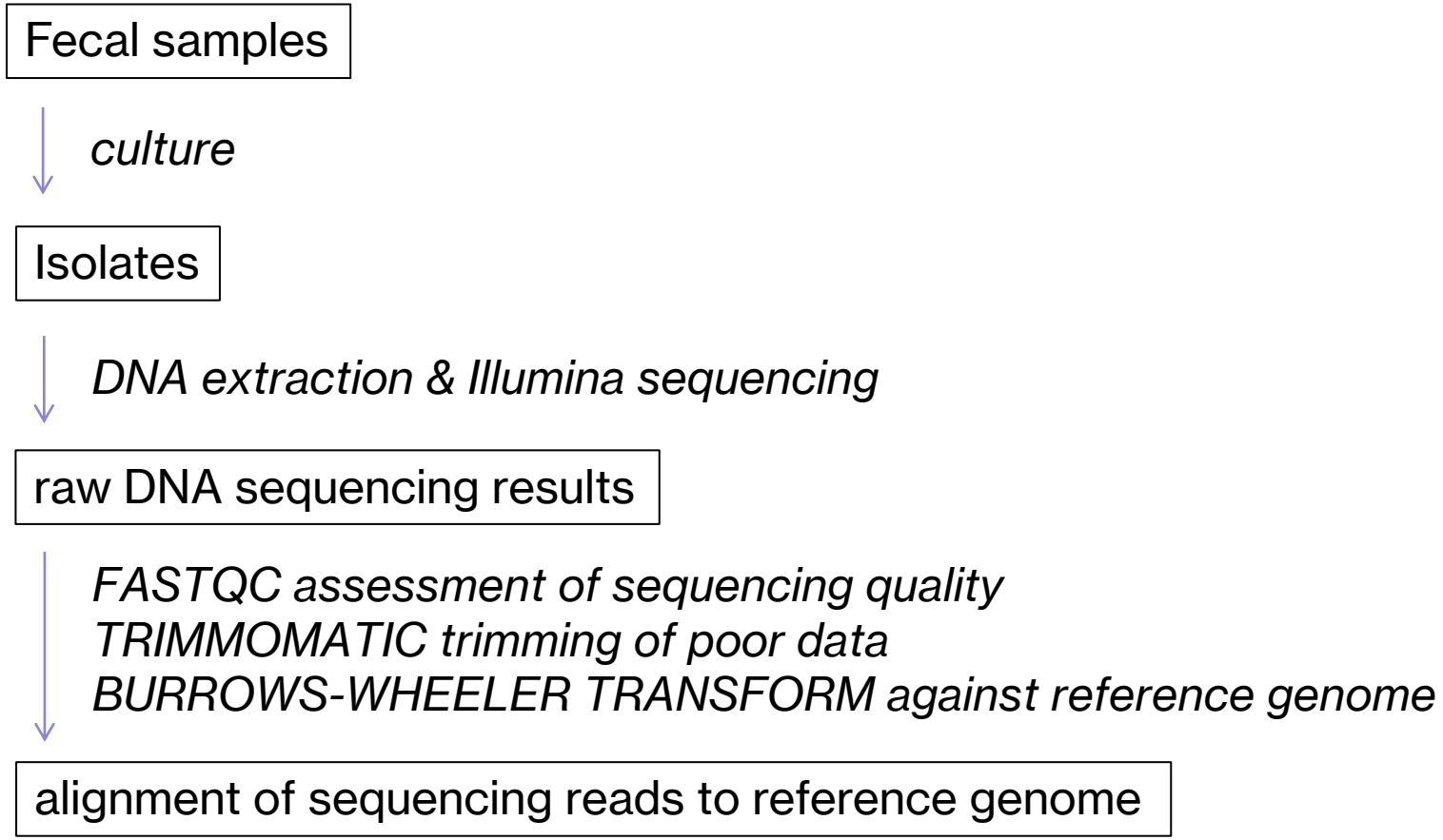
Variants!


37,397 U37,397
40,485 U40,485 CV156
41,456 U41,456
Tablet Tip: GFF3 or BED data can be displayed on the features track after importing
0

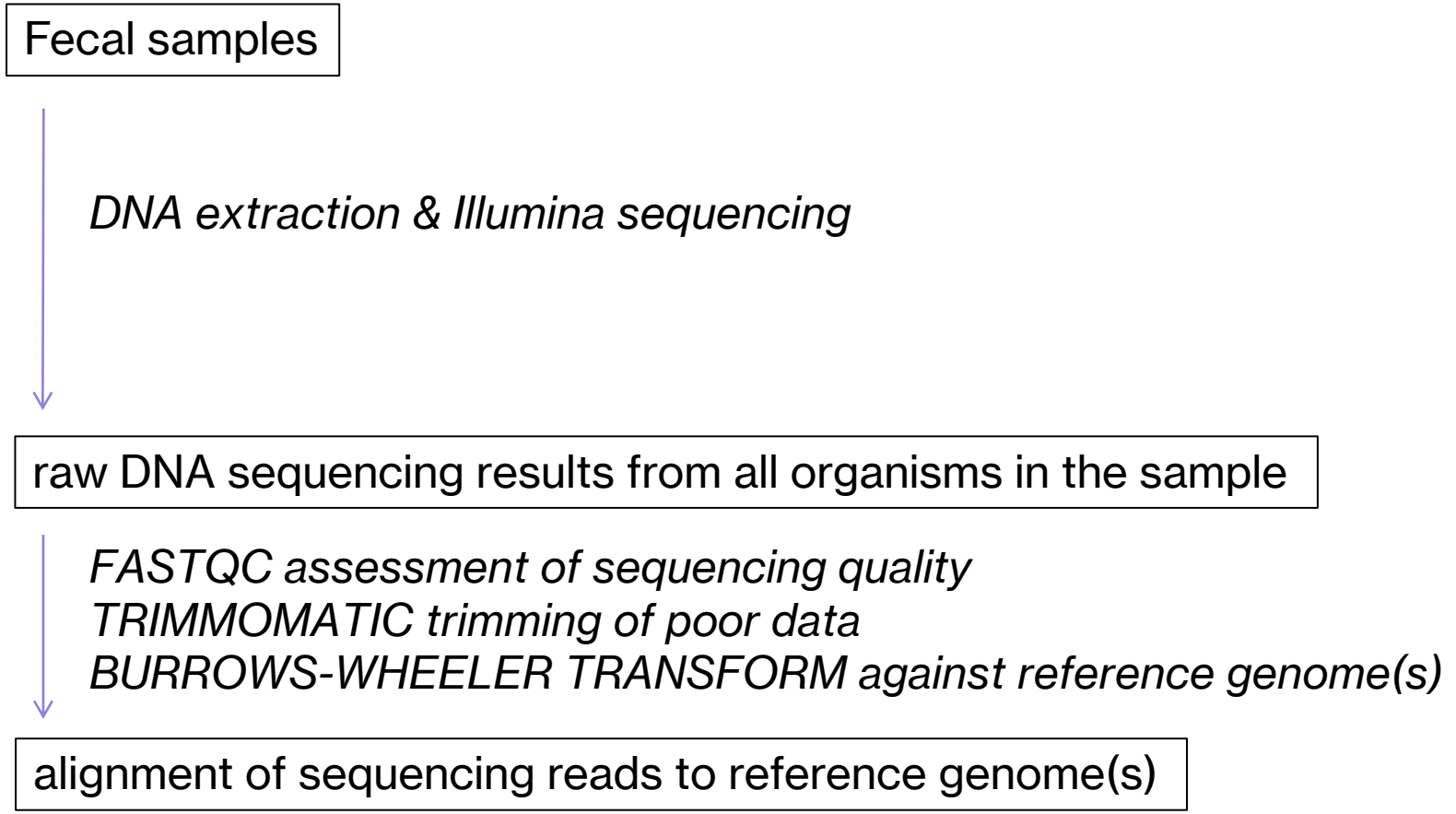
SNP Calling

- Positions where variant genotype does not match reference genotype are potential variants
- Variants are prioritized by:
  - Depth of coverage - more is better!
  - Sequences mapped to both strands – no library bias!
  - High-quality base calls in reads (i.e. PHRED scores)
  - Reads with high mapping quality score (MAPQ) in the Burrows-Wheeler Transform
- A number of software tools and algorithms exist to call SNPs from Burrows-Wheeler Transform mapped reads
  - Allows avoidance of genome assembly step in Molecular Epidemiology studies – saves times, allows lower-coverage sequencing
  - Particularly important for human exome sequencing and genome-wide association studies (GWAS)

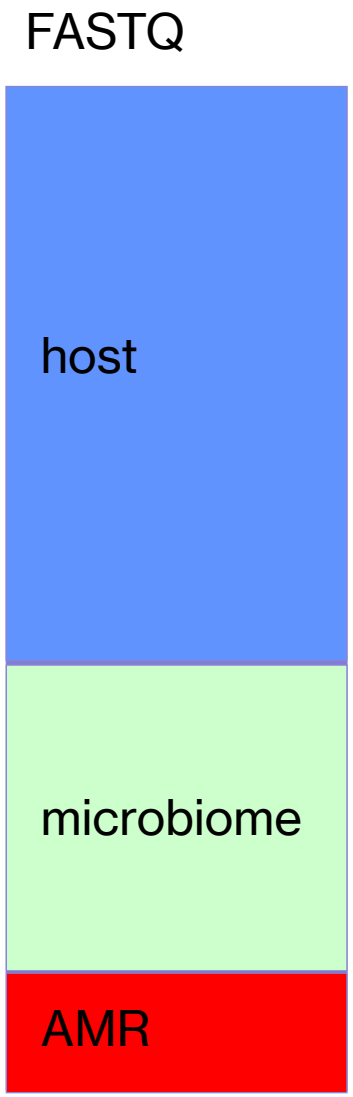
Metagenomics & Burrows-Wheeler Transform



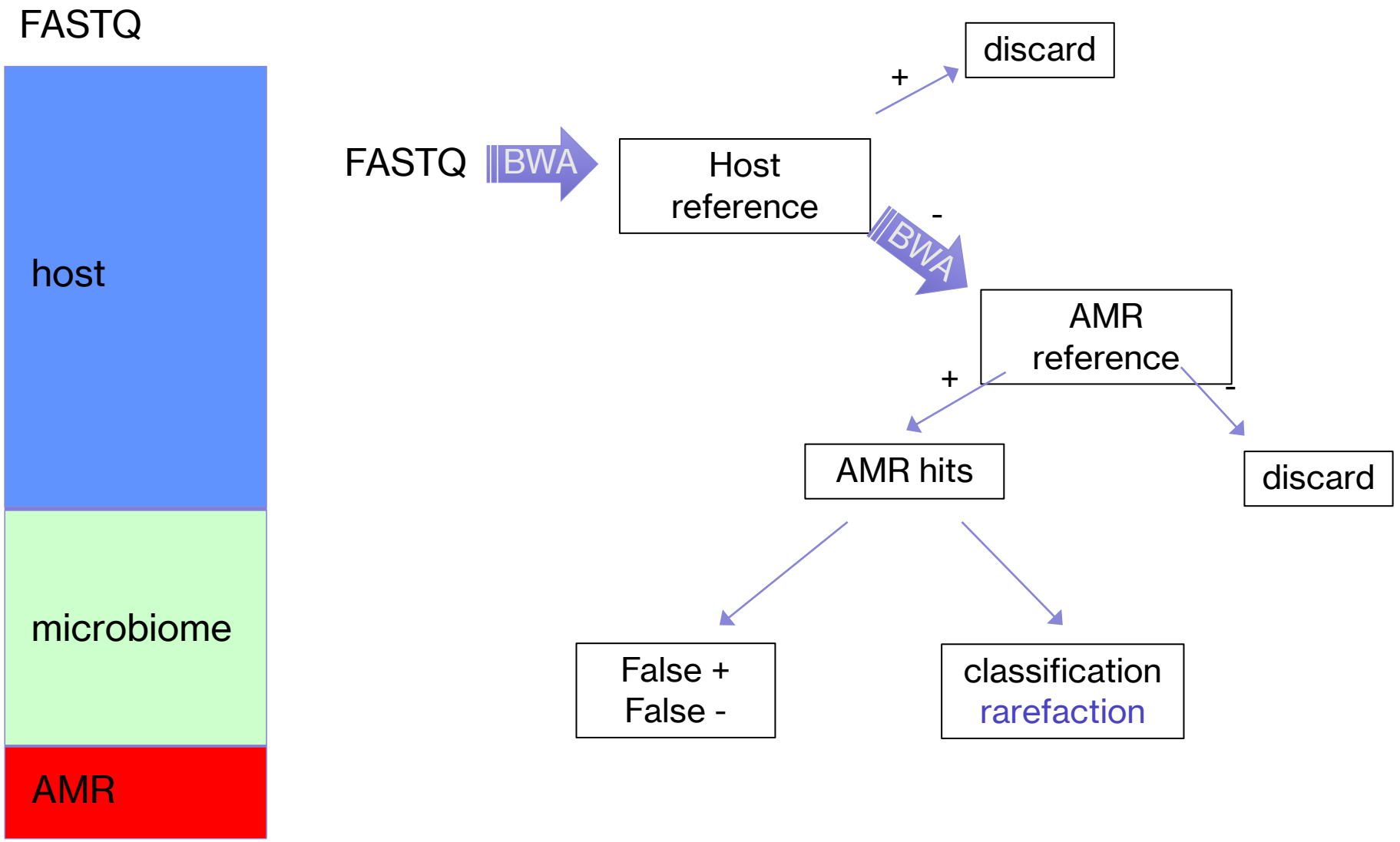
Metagenomics & Burrows-Wheeler Transform



Metagenomics, BWT, & AMR++

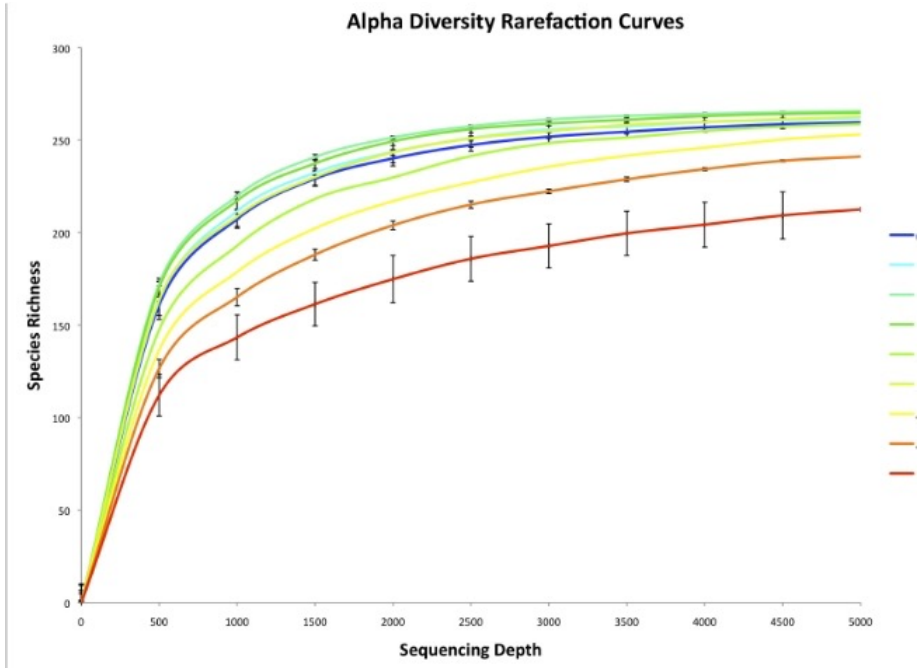


Metagenomics, BWT, & AMR++



Metagenomics – how much do you sequence?

- Microbiome are large and complex so need to be sequenced very heavily to find all the DNA sequences within
- This requires at least an order of magnitude of DNA sequences above normal isolate whole genome assembly – costly & time consuming!
- If you don't know what is in the microbiome, how do you know how much to sequence? – You don't but you can perform rarefaction analysis.



This week...

WEEK 9 (NOVEMBER 1 and 3) - MOLECULAR EPIDEMIOLOGY

LIVE lecture in class Wednesday 12:30pm,

Recorded Content

- Overview of Laboratory #7 - Molecular Epidemiology & Outbreak Analysis
  - Introduction & Task 1, https://web.microsoftstream.com/video/7a0a797b-e56b-43e5-a3de-fc85ab4bfe4b
  - Task 2, https://web.microsoftstream.com/video/3d3a2199-60a9-4183-8637-016bdcaa5a1b
  - Task 3, https://web.microsoftstream.com/video/b1d0c1f8-bf71-4726-88c2-828443a313f2
  - Task 4 & Review of Questions & Problems, https://web.microsoftstream.com/video/aaaf50cd-c358-491f-9760-f83e68f22f20
- Research Focus - Dr. Robyn Lee of the Dalla Lana School of Public Health, https://web.microsoftstream.com/video/8a04da03-951c-42cb-8059-53be0012d1b4

Tutorial

- **LIVE** session with Teaching Assistants and Flash Updates
  - Monday,
  - Wednesday,

Flash Updates

- **SNPs**. Define the term Single Nucleotide Polymorphism (SNP) and explain how these data can be used to determine organism/strain relatedness. Use the spread of MRSA as an example, Science 2010 327: 469-74 [PMID 20093474].
- **Horizontal Gene Transfer**. Define the term Horizontal Gene Transfer (HGT; also known as Lateral Gene Transfer, LGT) and discuss how it could confound determination of organism/strain relatedness using SNP analysis. Use the emergence of MCR-1 as an example, Lancet Infect Dis. 2015 Nov 18. pii: S1473-3099(15)00424-7 [PMID 26603172].
- **Metagenomics**. Introduce metagenomics in the context of molecular epidemiology. See Expert Rev Mol Diagn. 2018 Jul;18(7):605-615. [PMID 29898605].