Whole genome based phylogeny
Learning objective:

After this lecture you should be able to...

... account for the use of DNA mutations as the basis of distance matrices and hence phylogenetic trees

... account for the generation of WGS based phylogenetic trees by the CGE tools CSIPhylogeny and NDtree
What is phylogeny?

**Early phylogeny**
- Classification
- Based on phenotypes

**Current phylogeny**
- Based on genotypes
- DNA mutations as the basis for evolution
Classification

Carl Linnaeus
(1707-1778)

Hierarchical system
Kingdom
Phylum
Class
Order
Family
Genus
Species
Classification depicted as a tree
Tree terminology

- The tips represent the taxa in the study.
- The “goal” of phylogenetic analysis is to recover “bifurcating” trees, in which each taxon is linked to one other taxon through a node.
- Polychromous trees (multiple branches from one node) are less informative because they indicate that multiple taxa are related to each other, but not how.
Classification depicted as a tree
DNA mutations as the basis for evolution
What are phylogenetic trees

• Phylogenetic trees are a visual representation of the genetic relationship between species
• Phylogeny can also be represented by distance matrices
• Trees were traditionally made using aligned sequences of single genes or proteins
• Whole genome data can be used to create trees based on
  ✦ SNP calling
  ✦ Kmer overlap
  ✦ Alignment of genomes
What is a SNP

A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a Single Nucleotide — A, T, C or G — in the genome differs between members of a biological species or paired chromosomes.
How does it work

Strain A  ATTCAGTAGT
Strain B  ATGCAATTTGA
Strain C  ATGCAATTGT
Strain D  ATCCATTAGC
Construct distance matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain A</td>
<td>AT</td>
<td>TC</td>
<td>AG</td>
<td>TA</td>
</tr>
<tr>
<td>Strain B</td>
<td>AT</td>
<td>GC</td>
<td>AT</td>
<td>TG</td>
</tr>
<tr>
<td>Strain C</td>
<td>AT</td>
<td>GC</td>
<td>A</td>
<td>T</td>
</tr>
</tbody>
</table>
| Strain D | AT | CC | AT | TA | GC
Create the tree

Strain A  \text{ATTCAGTATG}

Strain B  \text{ATGCAGTGAT}

Strain C  \text{ATGCATAATGT}

Strain D  \text{ATCCATTTAGC}

\begin{tabular}{|c|cccc|}
\hline
 & A & B & C & D \\
\hline
A & 0 & 3 & 3 & 3 \\
B & 3 & 0 & 2 & 4 \\
C & 3 & 2 & 0 & 4 \\
D & 3 & 4 & 4 & 0 \\
\hline
\end{tabular}
How to read phylogenetic trees

\[ A = (B + D) = (B + C + E + G) = (B + C + F + H) \]

\[ D = (C + E + G) = (C + F + H) \]
How to read phylogenetic trees

(a)  
(b)
What is phylogeny used for

- Classify taxonomy – The classic use
- Outbreak detection – Increasing with WGS data
What is phylogeny used for

- *Vibrio cholerae* outbreak in Haiti followed the 2010 earthquake
- Rumours said that the outbreak may have come from Nepal, travelling along with UN soldiers from Nepal
- No proof had been given of this until the Hendriksen et al. paper in 2011
Case story

Data

• 24 recent *V. cholerae* strains from Nepal

• 10 previously sequenced *V. cholerae* isolates, including 3 from the Haitian outbreak

Analysis

• Antimicrobial susceptibility testing

• PFGE (pulsed-field gel electrophoresis) to analyse for genetic relatedness

• Whole genome sequencing, SNP identification and phylogenetic analysis
## Case story - Results

<table>
<thead>
<tr>
<th>Resistance profile</th>
<th>Susceptible</th>
<th>Decreased susceptibility</th>
<th>Resistant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nepalese strains</td>
<td>Tetracycline</td>
<td>Ciprofloxacin</td>
<td>Trimethoprim, Sulfamethoxazole Nalidixic</td>
</tr>
<tr>
<td>Hendriksen <em>et al.</em> 2011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haitian outbreak strains</td>
<td>Tetracycline</td>
<td>Ciprofloxacin</td>
<td>Trimethoprim, Sulfamethoxazole Nalidixic</td>
</tr>
<tr>
<td>Centers for Disease Control and Prevention, 2010</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Case story - Results

Pulsed-field gel electrophoresis (PFGE)

- Nepalese isolates divided in 4 groups
- Most common Haitian type in same group as four Nepalese strains
Case story
- Results
CGE tools for phylogeny

snpTree - a web-server to identify and construct SNP trees from whole genome sequence data

Pimlapas Leekitcharoenphon$^{1,2*}$, Rolf S Kaas$^{1,2}$, Martin Christen Frølund Thomsen$^2$, Carsten Friis$^1$, Simon Rasmussen$^2$, Frank M Aarestrup$^1$
SnpTree flow

**A**
- Raw reads
  - Pre-processing
    - Reads mapping (using BWA)
    - Identify SNPs (using SAMtools)
      - SNPs filtering (using SAMtools)
        - SNPs tree construction (using Fasttree)

**B**
- Assembled genomes
  - Reference genome alignment (using Nucmer)
    - Identify SNPs (using show-snps from MUMmer)
      - SNPs filtering (using show-snps from MUMmer)
        - SNPs tree construction (using Fasttree)
CSIPhylogeny

https://cge.cbs.dtu.dk/services/CSIPhylogeny/

SNP identification same as snpTree

Strict sorting of SNPs

• Depth
• Relative depth
• Distance between SNPs
• SNP quality
• Read mapping quality

PMID: 25110940
CSIPhylogeny

Requires all SNPs to be significant:

Z-score higher than 1.96 for all SNPs

\[ Z = \frac{X - Y}{\sqrt{X+Y}} \]

X is the number of reads with the most common nucleotide at that position, and Y the number of reads with any other nucleotide.
Output

Tree build by FastTree algorithm, in Newick format

- Branch lengths are substitutions per site at the variable sites

Matrix of SNP pair counts in text (.txt) format
Download the filtered SNP calls in Variant Calling Format (VCF):
Note: VCF files are compressed with gzip.
VCF files

Download matrix of SNP pair counts:
Download matrix as: TXT EPS

Download SNP alignment: FASTA

Percentage of reference genome covered by all isolates: 95.6684818250054
4440598 positions was found in all analyzed genomes.
Size of reference genome: 4641652

Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

<table>
<thead>
<tr>
<th>File</th>
<th>Valid positions</th>
<th>Pct. of reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>_1_1_2_2_1_1_2_1_R1.ignored_snps_4448690</td>
<td></td>
<td>95.8428163076422</td>
</tr>
<tr>
<td>_1_2_1_1_2_1_2_2_R1.ignored_snps_4450004</td>
<td></td>
<td>95.8711251942196</td>
</tr>
</tbody>
</table>
NDtree

https://cge.cbs.dtu.dk/services/NDtree/

**Nucleotide calling**

- A different approach where the main distinction is not between if a SNP should be called or not, but between whether or not there is solid evidence for the nucleotide at the given position
NDtree

Simple mapping approach

- Cuts all reads into K-mers
- Maps all K-mers to reference genome
- Makes ungapped consensus sequences of equal lengths
NDtree - mapping

K-mers

Reference genome

Consensus sequence

Reference genome

Genome 1
Genome 2
Genome 3
Genome 4
Genome 5
Genome 6
NDtree

Nucleotide calling for the individual genomes

For each individual genome, all reads are mapped to the consensus sequence and the significance of the base call at each position is evaluated by calculating the number of reads $X$ having the most common nucleotide at that position, and the number of reads $Y$ supporting other nucleotides.

A $Z$-score threshold is calculated

$$Z = \frac{X - Y}{\sqrt{X+Y}} > 1.96$$

> 90% of reads must support the same base
NDtree

Count nucleotide differences

Each pair of sequences is compared and the number of nucleotide differences in positions called in all sequences is counted.
NDtree

Uses two different algorithms to make two different trees
• UPGMA
• Neighbor Joining

Both algorithms are part of the PHYLIP Neighbor program package and make trees from distance matrices
UPGMA vs. Neighbour Joining

• UPGMA works when samples have been taken the same time

• Neighbor Joining is better when samples have been taken at different times
NDtree

Output

- distance.txt: Distance matrix - tab separated
- dist.mat: Distance matrix - PHYLIP format
- tree.nj.newick: Neighbor Joining tree - Newick format
  - Branch lengths is number of Nucleotide Differences
- tree.upgma.newick: UPGMA tree – Newick format
  - Branch lengths is number of Nucleotide Differences
So... What should I use when?

CSI Phylogeny

• Has very good statistics and a good graphical overview
• Advantageous to use when you expect the differences between the isolates to be larger than 5-10 mutations
• Is faster

NDtree

• Is able to find very small differences
• Does not take recombination into consideration
• Works best on raw reads. If given assembled genomes, it simulates reads (including errors!!!)
Choosing a reference genome

For comparison of very closely related isolates, a better level of detail is given by using a closely related reference genome.
What defines an outbreak?

• We can’t tell for certain
• It depends on the species
• A rule of thump is:
  – Within 10 SNPs it is definitely an outbreak
  – Within 30 SNPs it might be an outbreak
  – Above 60 SNPs it is most likely not an outbreak
Table 1
Examples of relatedness criteria for wg/cgMLST and SNP typing schemes of representative clinically relevant bacteria

<table>
<thead>
<tr>
<th>Organism</th>
<th>Relatedness threshold&lt;sup&gt;b&lt;/sup&gt;</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acinetobacter baumannii</td>
<td>≤8</td>
<td>McGladdery et al. [25,26]</td>
</tr>
<tr>
<td>Brucella spp.</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td><a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a></td>
</tr>
<tr>
<td>Campylobacter coli, C. jejuni</td>
<td>≤14</td>
<td>McGladdery et al. [27,28]</td>
</tr>
<tr>
<td>Cronobacter spp.</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td><a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a></td>
</tr>
<tr>
<td>Clostridium difficile</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td>[29], <a href="http://www.cgmst.org/ncs">http://www.cgmst.org/ncs</a>, <a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a></td>
</tr>
<tr>
<td>Enterococcus faecium</td>
<td>≤20</td>
<td>McGladdery et al. [30]</td>
</tr>
<tr>
<td>Enterococcus raffinosus</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td>[31,32], <a href="https://enterobase.warwick.ac.uk/">https://enterobase.warwick.ac.uk/</a></td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>≤10</td>
<td>McGladdery et al. [33,34]</td>
</tr>
<tr>
<td>Franciscella tularensis</td>
<td>≤1</td>
<td>McGladdery et al. [35,36]</td>
</tr>
<tr>
<td>Klebsiella oxytoca</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td><a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a></td>
</tr>
<tr>
<td>Klebsiella pneumonia</td>
<td>≤10</td>
<td>McGladdery et al. [37]</td>
</tr>
<tr>
<td>Legionella pneumophila</td>
<td>≤4</td>
<td>McGladdery et al. [38,39]</td>
</tr>
<tr>
<td>Listeria monocytogenes</td>
<td>≤10</td>
<td>McGladdery et al. [40]</td>
</tr>
<tr>
<td>Mycobacterium abscessus</td>
<td>≤12</td>
<td>McGladdery et al. [41]</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td><a href="http://www.cgmst.org/ncs">http://www.cgmst.org/ncs</a></td>
</tr>
<tr>
<td>Neisseria gonorrhoeae</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td>[42], <a href="http://www.applied-maths.com/applications/wgmlst">http://www.applied-maths.com/applications/wgmlst</a></td>
</tr>
<tr>
<td>Neisseria meningitidis</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td><a href="http://www.cgmst.org/ncs">http://www.cgmst.org/ncs</a></td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>≤14</td>
<td>McGladdery et al. [43]</td>
</tr>
<tr>
<td>Salmonella dublin</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td>[44], <a href="https://enterobase.warwick.ac.uk/">https://enterobase.warwick.ac.uk/</a></td>
</tr>
<tr>
<td>Salmonella typhimurium</td>
<td>Epidemiologic validation in progress&lt;sup&gt;b&lt;/sup&gt;</td>
<td>[46], <a href="https://enterobase.warwick.ac.uk/">https://enterobase.warwick.ac.uk/</a></td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>≤24</td>
<td>McGladdery et al. [47,48]</td>
</tr>
<tr>
<td>Streptococcus suis</td>
<td>≤21</td>
<td>McGladdery et al. [49]</td>
</tr>
<tr>
<td>Vibrio paraheartolyticus</td>
<td>≤10</td>
<td>McGladdery et al. [50]</td>
</tr>
<tr>
<td>Yersinia spp.</td>
<td>0</td>
<td>McGladdery et al. [51]</td>
</tr>
</tbody>
</table>

<sup>a</sup> Data often represent single studies that can be used to begin formulation of species-specific interpretation criteria. Thus, these data should be coupled with newly published similar studies to ensure that resulting values are not atypical and can be generally applied.

<sup>b</sup> Proposed wg/cgMLST schemes are available online (http://www.cgmst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/) but as yet have not been epidemiologically validated.