Finding relationships – AKA clustering

Dr Kristy Horan
Austrakka bioinformatician

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

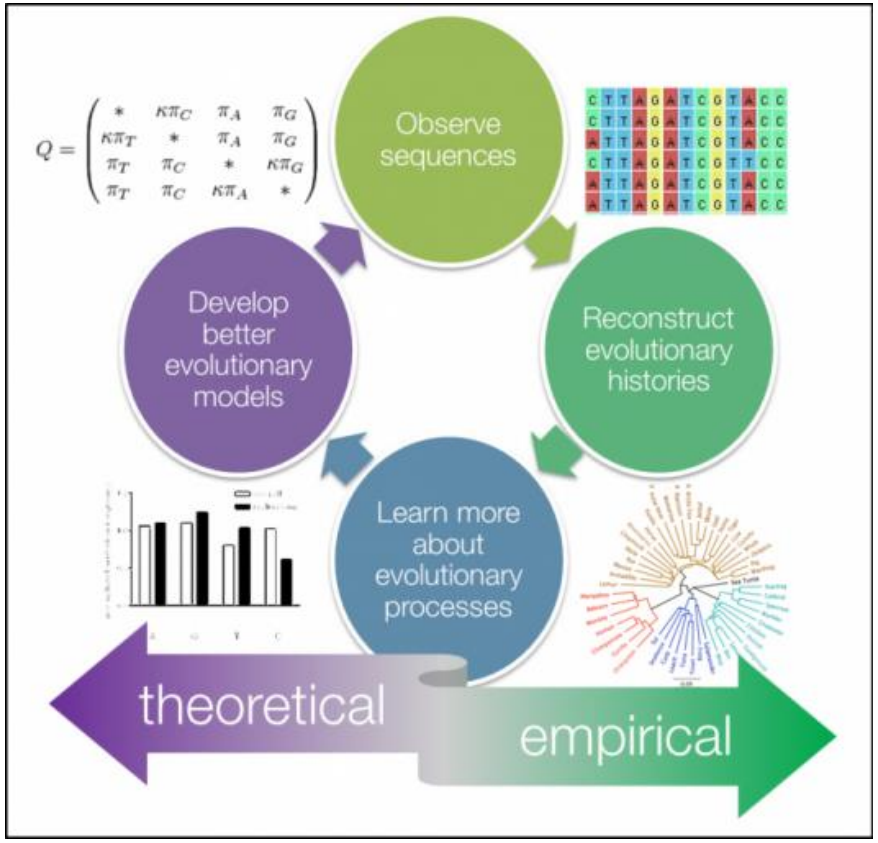
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Recap

Phylogenetics

Slides by Dr Jake Lacey

- Multi-disciplinary field
- Study of evolutionary relationships among biological entities – often species, individuals or genes
- **Molecular phylogenetics** utilises sequence data
- In **pathogen genomics** phylogenetics can play important role in understanding **evolution and outbreaks**



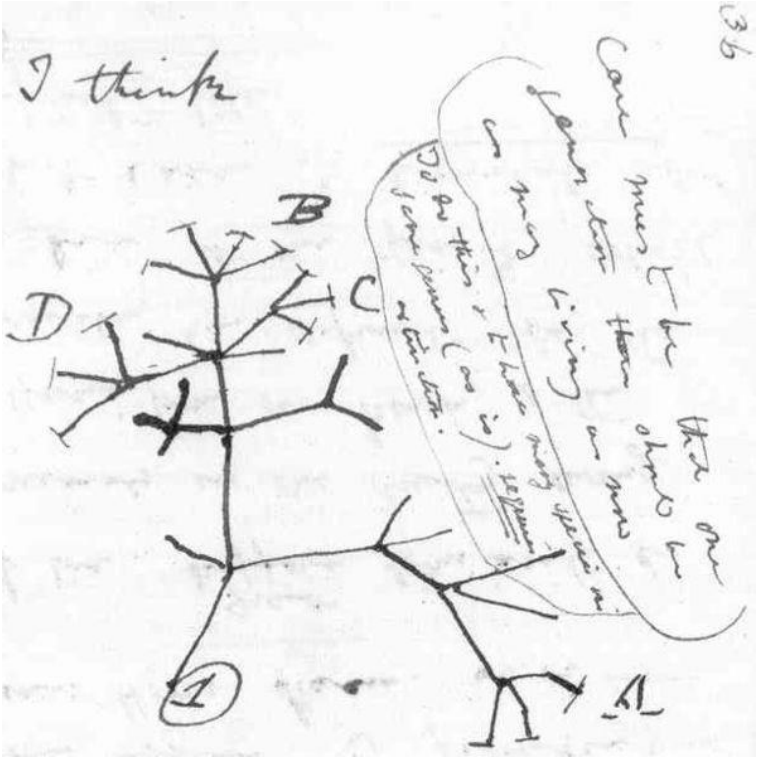
Trees – a visual representation

Slides by Dr Jake Lacey

“Common” graphical representation of phylogenetic analysis are Trees

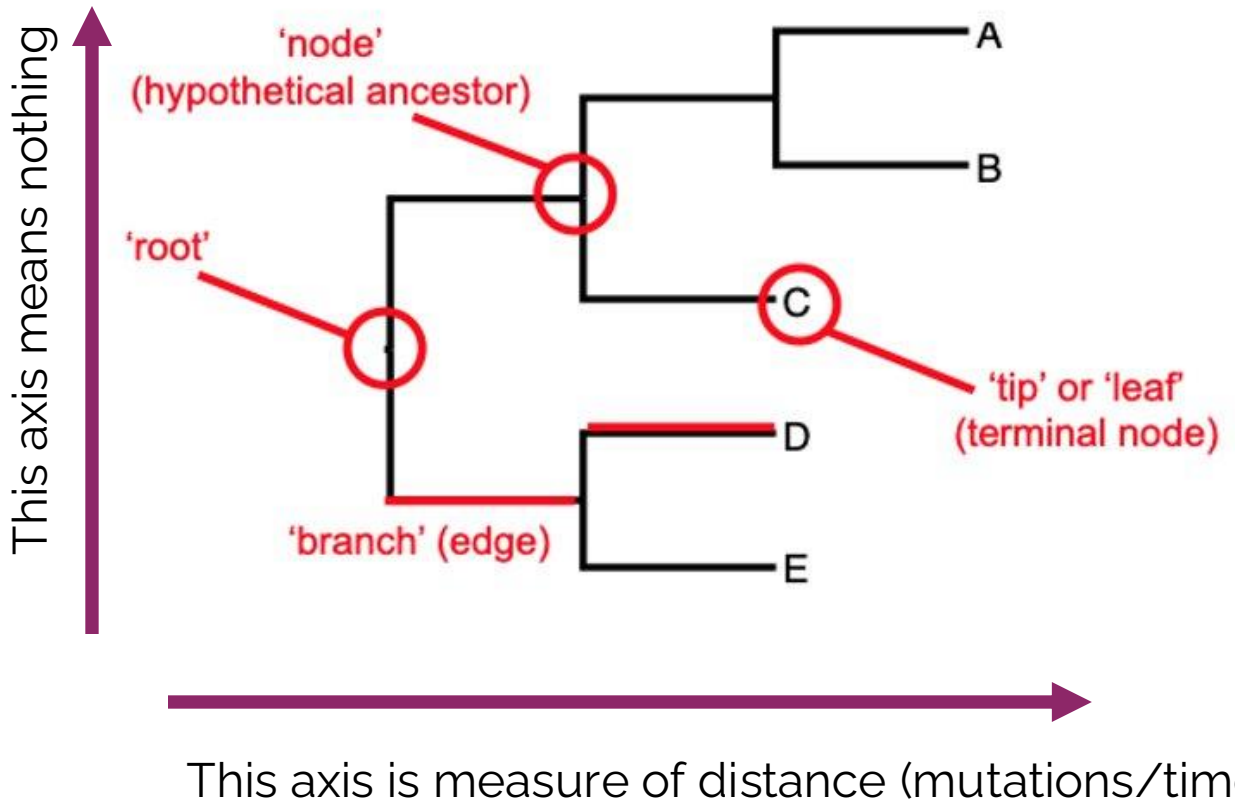
Many kinds of phylogenetic trees

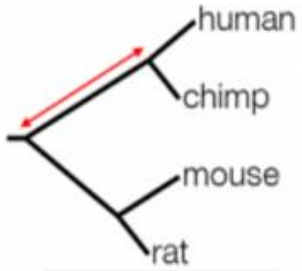
- Cladograms
- Phenograms
- Distance based (NJ, UPGMA)
- Maximum Parsimony
- Maximum likelihood
- Bayesian approach



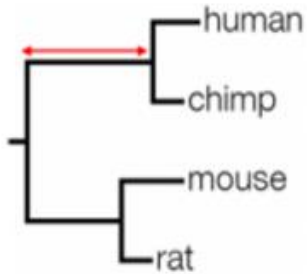
Trees – a visual representation

Slides by Dr Jake Lacey

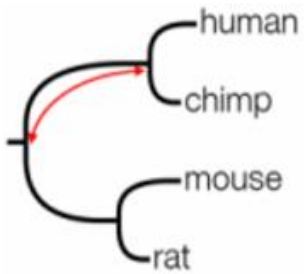




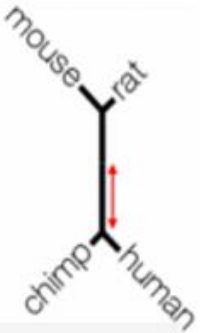
Diagonal



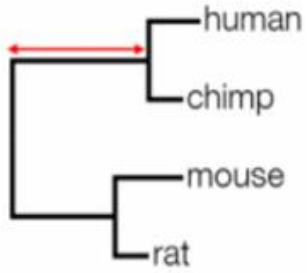
Rectangular



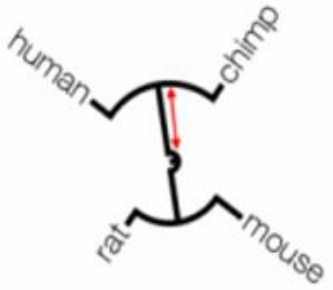
Curved



Radial



Unrooted



Circular

Planting the seeds (inputs)

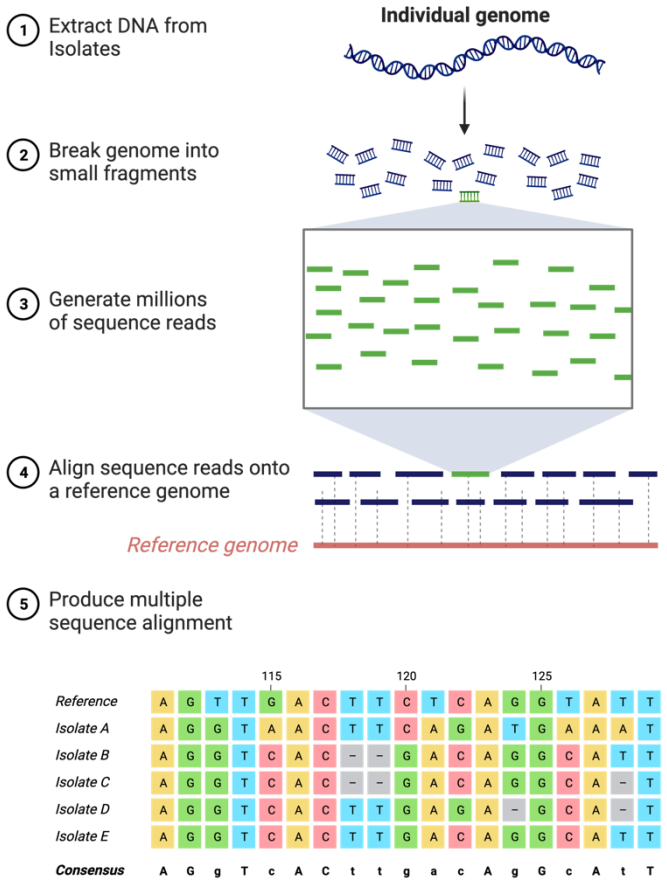
Slides by Dr Jake Lacey

Reference method:

- Find mutations (SNPs) by aligning reads of genome to a reference
- Repeat with other genomes
- Combine SNP calls into multiple sequence alignment (MSA)
- Extract core genome

Reference-free method:

- Directly compare reads between genomes (pairwise)
- Determine distance between each genome (matrix or a list)



Nurturing your tree (parameters)

Slides by Dr Jake Lacey

Multiple different methods to build a tree

- Distance based reference-free approaches = UPGMA, NJ
- Sequence alignment with model and parameters applied = Maximum likelihood

Reliable trees need appropriate parameters

- Model that describes the **rates of change** of one nucleotide to another
- Proportion of **invariant sites** (sites that do not change)
- **Gamma distribution** - gamma distributed rate variation among sites
- **Bootstrapping** (replication of tree structure from subsampling)
- **Molecular clock** estimates (natural SNP variation and accumulation)

Here the be rabbit holes

Tree models – don't get lost

Slides by Dr Jake Lacey

Different models used to build trees have different assumptions around nature of mutations (100s models)

Substitution mutations are grouped hierarchically:

- general base substitution,

transitions and **transversions**

- Model use different rates

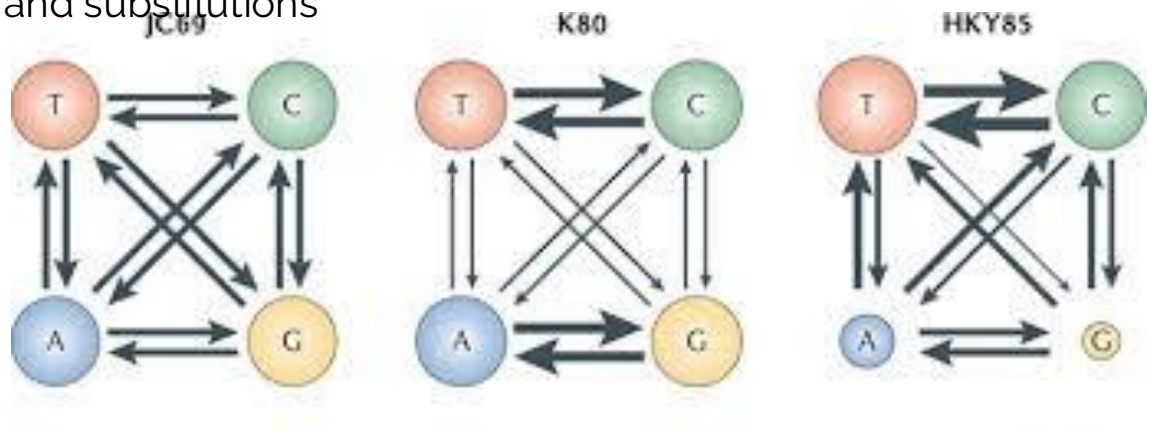
Models do not account for:

- insertions/deletions
- Recombination
- Horizontal gene transfer

Jukes-Cantor (JC69)
All nucleotides **equal** frequency and substitutions

Kimura (K80)
Transitions are more common

Hasegawa-Kishono-Yano (HKY85)
Variable frequencies and transitions more common



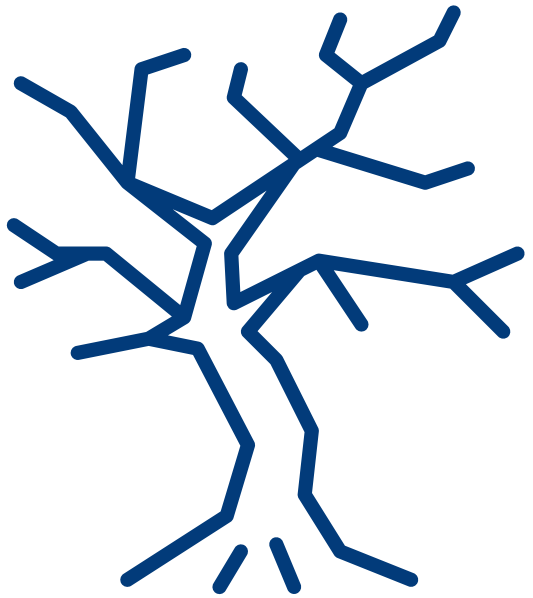
Infinite tree space

The tree you see are not the only possible tree

- Infinite tree space
- You only see the local maxima
- You may get a subtly different tree everytime you run an analysis

Do you need to be worried about this?

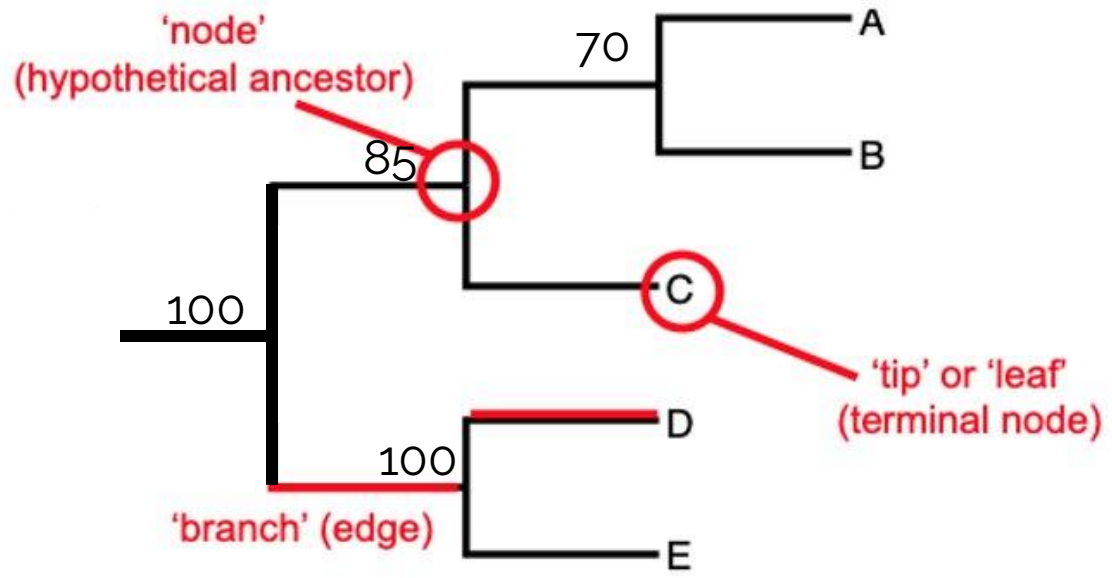
- 99.9999999% no you do not



Bootstrap values

How do you interpret a bootstrap value

- These will look different depending on the method (0.7, 70, 70/100)
- Bootstrap values are related to the branch support
  - The bootstrap values are often displayed on an internal node
  - The value indicates the degree of support for the subtree below that branch
- Broadly reflect the support for the sub-tree below that node
  - In 1000 bootstrap trees that subtree was seen 700 times



In real life

Trees – a visual representation

_

- **Trees can be very useful to demonstrate visually how sequences are related**
- Compresses complex relationships to simple graph
- Relatively simple to understand
- **In practice it is HARD WORK to 'identify relatedness' from a tree.**
- Automation in high throughput setting is very challenging
- Sequence and dataset quality can dramatically impact your interpretations
- How close is close?
- Many pathogens break our assumptions
  - Polyphyletic organisms
  - Recombination
  - Phage

Identification of relatedness in CPHL

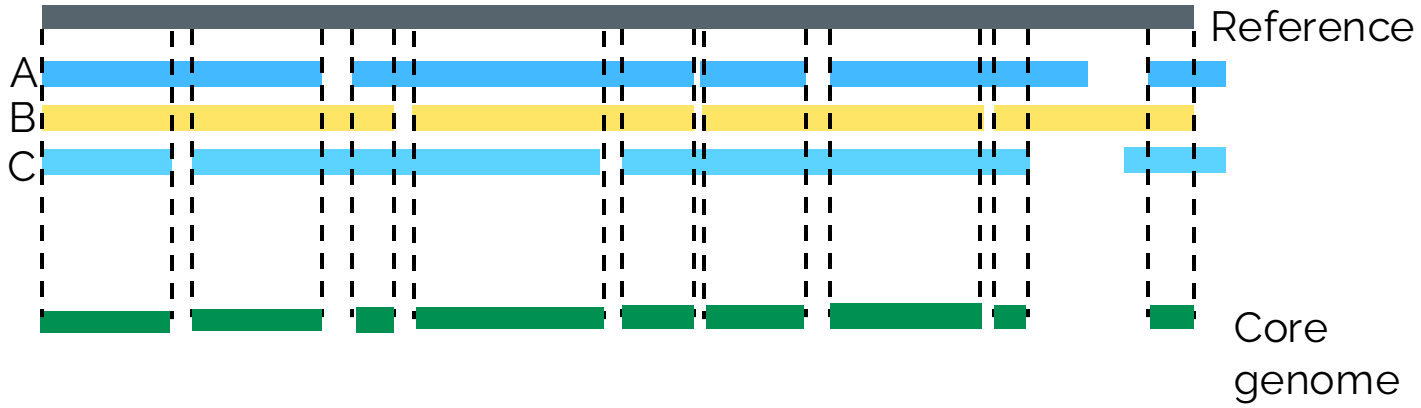
- **Evolution is important**
- Provides context for how to interpret relatedness
- Can provide temporal and geographical insights
- **BUT mostly we are interested in two simple questions**
- Did person A give pathogen X to person B?
- Did this group of people get exposed to pathogen X from the same environmental source?
- **Often over short timeframes which may not be evolutionarily that relevant**
- Weeks to months
- Occasionally years

To cluster or not to cluster?

- **Genomic clusters can be established using any feature you can measure in a pairwise fashion**
- Alleles
- SNP
- Branch lengths
- **Most commonly we use hierarchical clustering algorithms**
- Single linkage
- Complete linkage
- Average
- **Thresholds**
- There are NO FIXED THRESHOLDS
- Should be determined by extensive validation or assessment of individual analysis.
  - SNP calling performance – what is the limit of detection?
  - Where is the signal to noise line?

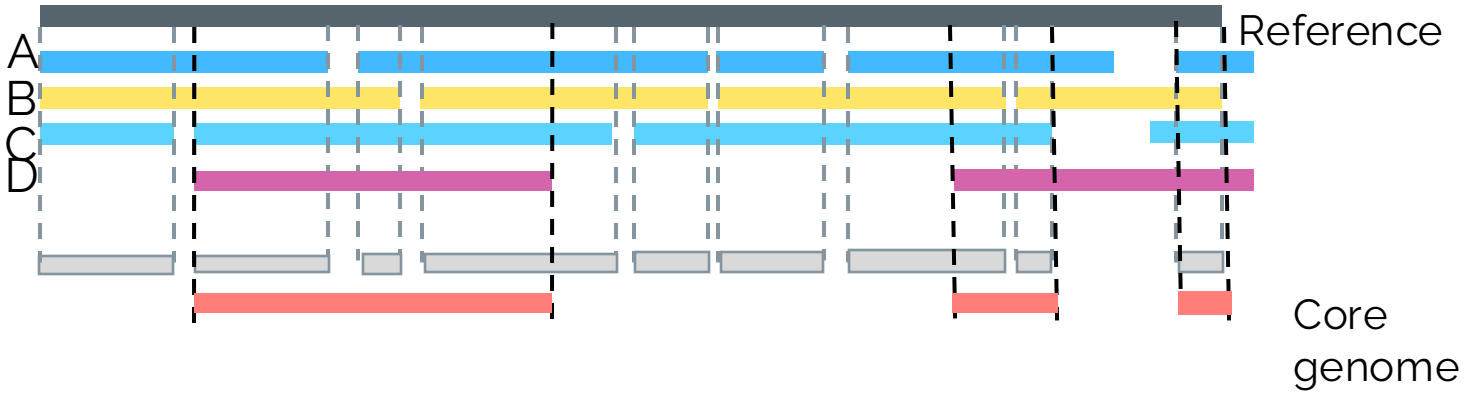
Not all distances are created equally

- **“Distance” often thought of as a ”fixed” feature.**
- Can only compare/calculate distance of what is PRESENT
- **In genomics this is NOT always true.**
- **Core genome is specific to the sequences that are included.**



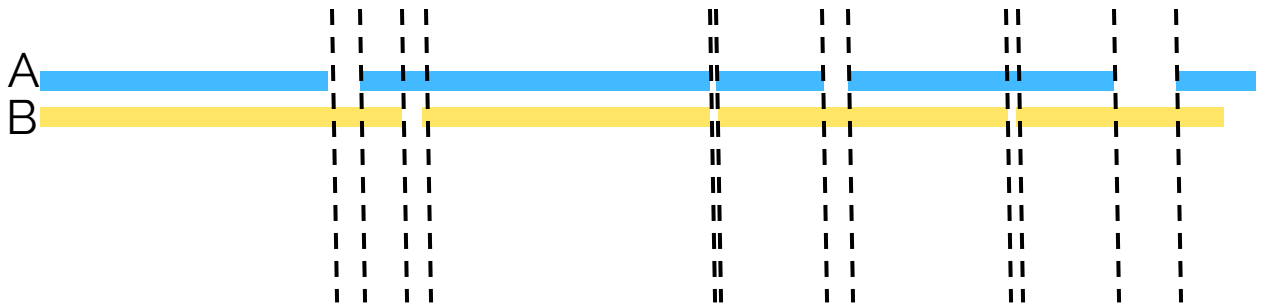
Not all distances are created equally

- **“Distance” often thought of as a ”fixed” feature.**
- Can only compare/calculate distance of what is PRESENT
- **In genomics this is NOT always true.**
- **Core genome is specific to the sequences that are included.**



Not all distances are created equally

- **“Distance” often thought of as a ”fixed” feature.**
- Can only compare/calculate distance of what is PRESENT
- **In genomics this is NOT always true.**
- **Reference free approaches**
- Each pairwise comparisons may be made on DIFFERENT “sequences”



Not all distances are created equally

- **“Distance” often thought of as a ”fixed” feature.**
- Can only compare/calculate distance of what is PRESENT
- **In genomics this is NOT always true.**
- **Reference free approaches**
- Each pairwise comparisons may be made on DIFFERENT “sequences”

Hierarchical clustering

- **These algorithms build a hierarchy of clusters.**
- Agglomerative
  - Bottom-up – each sequence starts as is its own cluster
  - Each ‘cluster’ is merged with the closest next cluster until all sequences are in a single ’cluster’
- Linkage method counts
- **Often used in pathogen genomics to identify groups or clusters that closer to each other than to others.**

Hierarchical clustering

	A	B	C	D	E
A	0	5	8	100	8
B	5	0	12	102	9
C	8	12	0	150	8
D	100	102	150	0	10
E	8	9	8	10	0

A B C D E

Hierarchical clustering

	A	B	C	D	E
A	0	**5**	8	100	8
B	**5**	0	12	102	9
C	8	12	0	150	8
D	100	102	150	0	10
E	8	9	8	10	0

Hierarchical clustering

Single-linkage

(A,B)

	A	B	C	D	E
A	0	5	**8**	**100**	**8**
B	5	0	12	102	9
C	**8**	12	0	150	8
D	**100**	102	150	0	100
E	**8**	9	8	100	0

	(A,B)	C	D	E
(A,B)	0	8	100	8
C	8	0	150	8
D	100	150	0	100
E	8	8	100	0

Hierarchical clustering

Single-linkage **(A,B)**

	(A,B)	C	D	E
(A,B)	0	**8**	100	**8**
C	**8**	0	150	**8**
D	100	150	0	100
E	**8**	8	**100**	**0**

Hierarchical clustering

Single-linkage

((A,B),C,E)

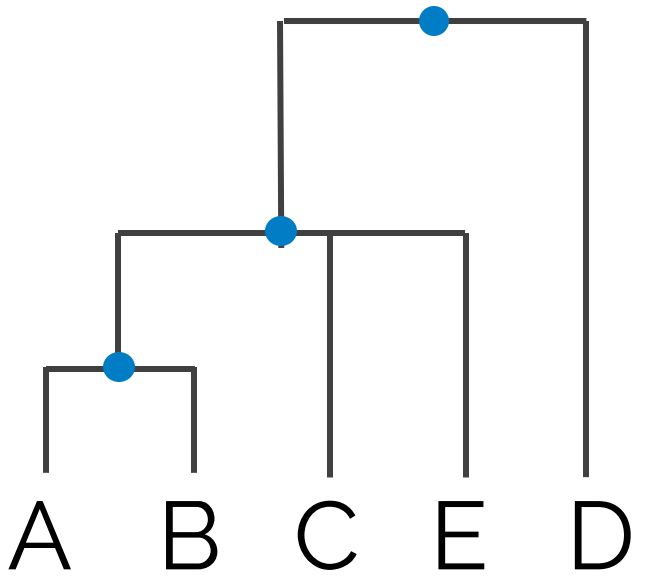
	(A,B)	C	D	E
(A,B)	0	8	**100**	8
C	8	0	150	8
D	**100**	150	0	**100**
E	8	8	**100**	0

	((A,B),C,E)	D
(A,B),C,E)	0	100
D	100	0

Hierarchical clustering

Single-linkage **(((A,B),C,E),D)**

	((A,B),C,E)	D
(A,B),C,E)	0	100
D	100	0



Hierarchical clustering

	A	B	C	D	E
A	0	**5**	8	100	8
B	**5**	0	12	102	9
C	8	12	0	150	8
D	100	102	150	0	10
E	8	9	8	10	0

Hierarchical clustering

Complete-linkage

(A,B)

	A	B	C	D	E
A	0	5	8	100	8
B	5	0	**12**	**102**	**9**
C	8	**12**	0	150	8
D	100	**102**	150	0	100
E	8	**9**	8	100	0

	(A,B)	C	D	E
(A,B)	0	**12**	**102**	**9**
C	**12**	0	150	8
D	**102**	150	0	100
E	**9**	8	100	0

Hierarchical clustering

Complete-linkage **(A,B)**

	(A,B)	C	D	E
(A,B)	0	12	102	9
C	12	0	150	**8**
D	102	150	0	10
E	9	**8**	10	0

Hierarchical clustering

Complete-linkage

(A,B)
(C,E)

	(A,B)	C	D	E
(A,B)	0	**12**	**102**	9
C	**12**	0	**150**	8
D	**102**	**150**	0	100
E	9	8	100	0

	(A,B)	(C,E)	D
(A,B)	0	12	102
(C,E)	12	0	150
D	102	150	0

Hierarchical clustering

Complete-linkage **(A,B),(C,E)**

	(A,B)	(C,E)	D
(A,B)	0	**12**	102
(C,E)	**12**	0	150
D	102	150	0

Hierarchical clustering

Complete-linkage **(A,B),(C,E)**

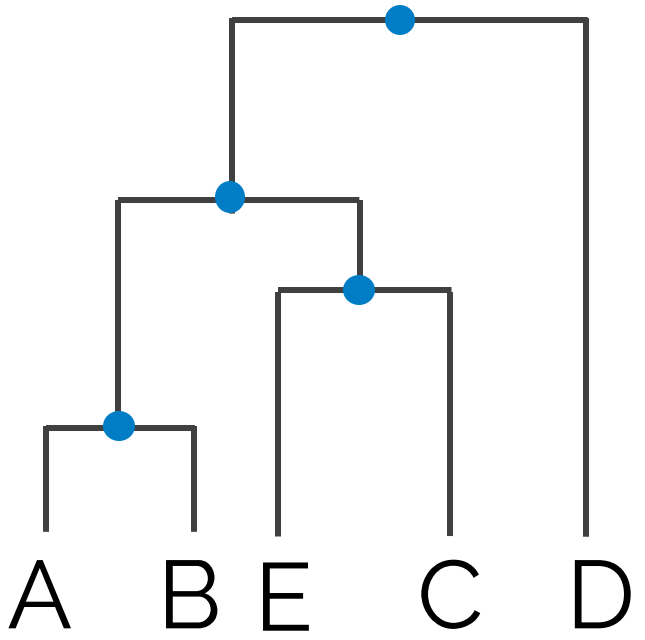
	(A,B)	(C,E)	D
(A,B)	0	12	102
(C,E)	12	0	**150**
D	102	**150**	0

	(A,B),(C,E)	D
(A,B),(C,E)	0	150
D	150	0

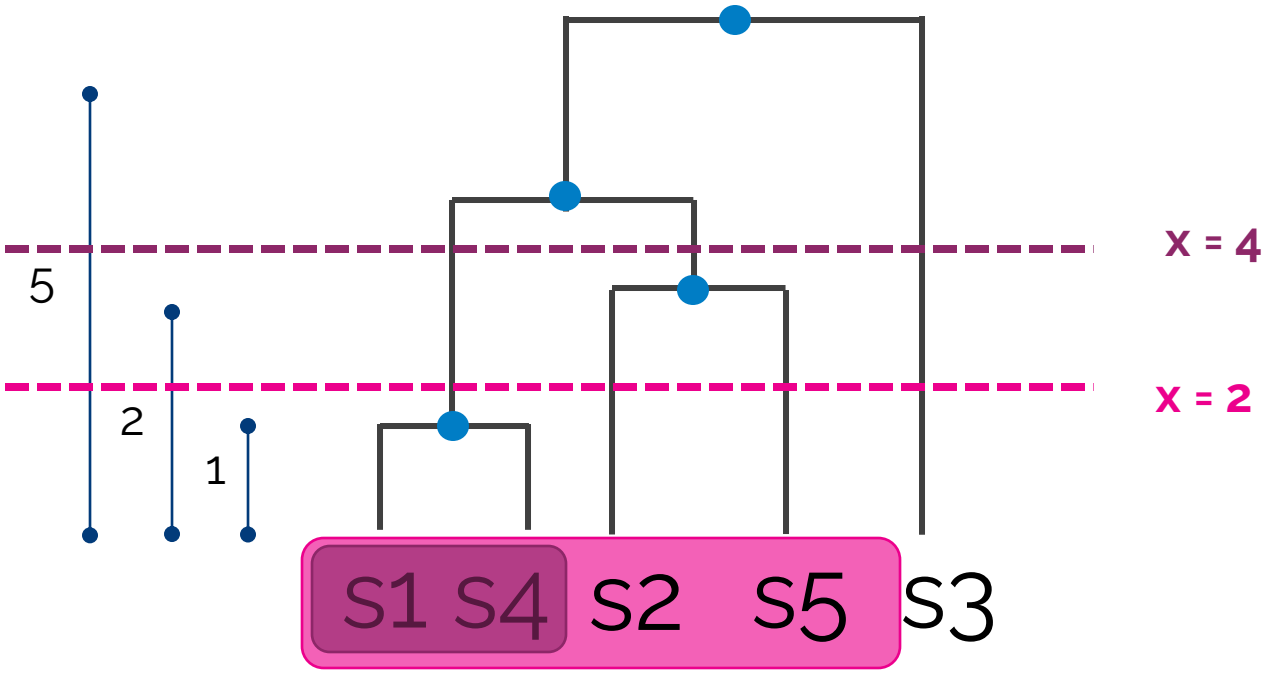
Hierarchical clustering

Complete-linkage **(((A,B),(C,E),D)**

	(A,B),(C,E)	D
(A,B),(C,E)	0	150
D	150	0



Threshold and clusters



What does it mean?

- **Depending on the algorithm…**
- **Single-linkage clustering**
- In order to be part of a single-linkage cluster a sequence must be less than threshold x to at LEAST one other sequence in the cluster.
- Long chains of sequences where the maximum distance between any two sequences can exceed the threshold used.
- **Complete-linkage clustering**
- In order to be part of a complete-linkage cluster a sequences but be less than threshold x to EVERYTHING else in the cluster.
- The maximum distance between any two sequences will not be more than the threshold.
- More discrete clusters

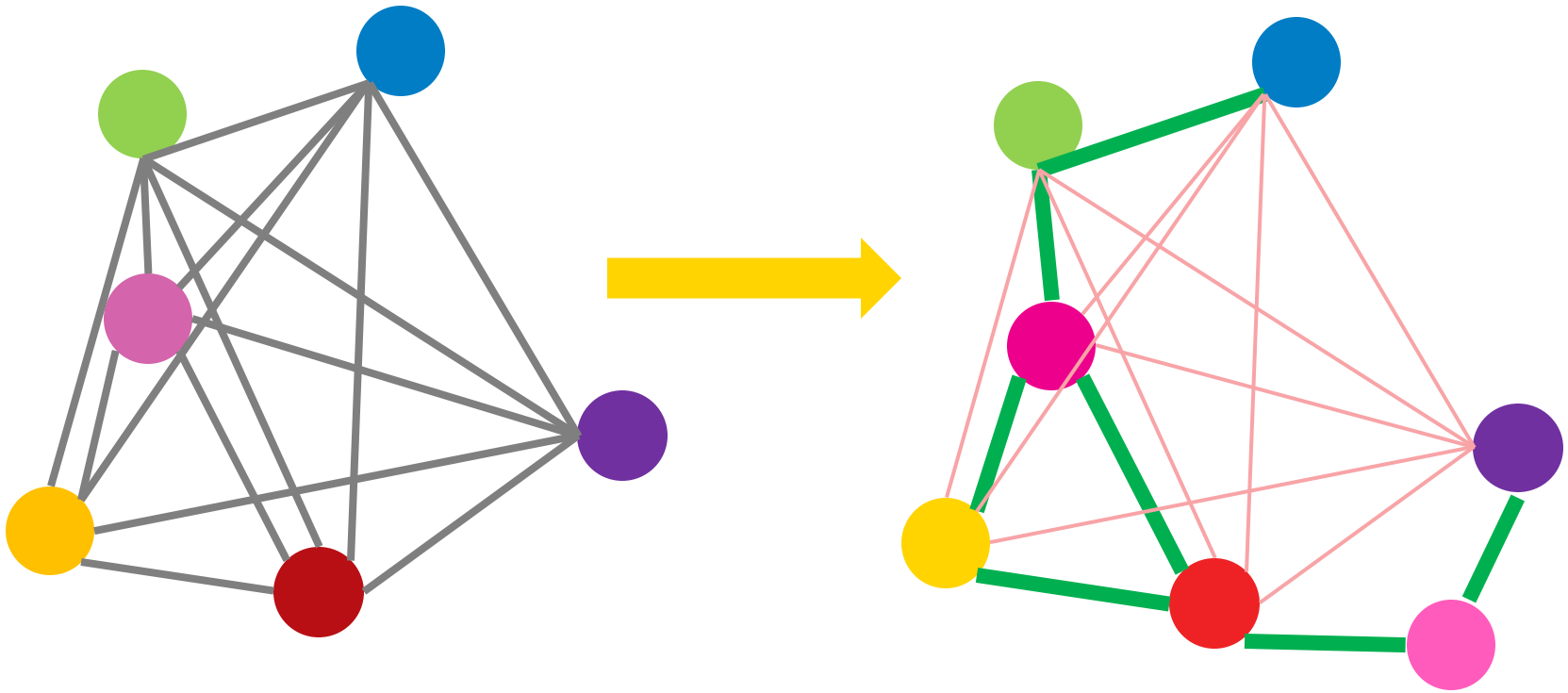
What does it mean?

- **Depending on the algorithm…**
- **Single-linkage clustering**
- In order to be part of a single-linkage cluster a sequence must be less than threshold x to at LEAST one other sequence in the cluster.
- Long chains of sequences where the maximum distance between any two sequences can exceed the threshold used.
- Useful were small changes overtime or transmission events are likely.
- Captures the 'variance' within a population
- **Complete-linkage clustering**
- In order to be part of a complete-linkage cluster a sequences but be less than threshold x to EVERYTHING else in the cluster.
- The maximum distance between any two sequences will not be more than the threshold.
- More discrete clusters.
- More exclusionary can be useful to exclude relationships or for triaging prior to more detailed analysis.

How to choose a threshold?

- **Extensive validation**
- Where an analysis is routine or repeated thresholds may be validated to determine what provides sensible and epidemiologically relevant groups
  - Balance between useful and actionable and missing important relationships.
- These should be revisited regularly to ensure that ongoing utility and appropriateness is maintained.
  - Don't just use 5 cause everyone else does
- **Dataset assessment**
- What method have you used? Reference guided core genome vs reference free?
- What is the diversity of the dataset?
- Quality of the core alignment?
  - Low core genome means that low SNP thresholds are likely not useful.
- Distribution of pair-wise distances
  - If everything is the same – then nothing is closer to anything.. So clustering will not likely reflect any of the actual biology.

When things change



- Each analysis is different.
- Adding or changing an analysis will change the relationships within that analysis.
  - Reference
  - Add sequence
  - Remove sequence.

Sequence quality vs analysis quality

- **Spend A LOT of time assessing quality of sequences. This is vitally important.**
- Actually assessing the quality of a sequence to be included in downstream analysis.
- **This DOES NOT mean a sequence is suitable for inclusion in any subsequent relationship analysis**
- Inappropriate species or serotype or ST
- Too distant/outlier
- Heterogeneous
- Recombination
- Plasmids
- Phage
- Inappropriate/incomplete methodology
- **Many relatedness analysis will be iterative – even the automated ones.**

Reference genome choice

- **Appropriateness of reference WILL determine the quality of an analysis.**
- Poor quality reference (gaps, ambiguity) will lead to poor alignment quality
- Distant reference will impact the core genome and the degree of resolution available
- **Often stated 'gold-standard' == complete reference genome.**
- For local outbreaks – these can be too distant.
- A high quality *de novo* assembly can often provide much better quality results.
  - Low contigs
  - High N50
  - Filter to exclude small rubbish < 500 bp
  - Genome size (number of bases in assembly) should be close to the expected genome size for the species. This ensures that there are few gaps.
  - You may see increases in 'false' SNPs at the ends of contigs were mapping quality is lower – BUT these are most often SNPs that appear in ALL sequences and so are not informative.

That was a lot

Thoughts or questions?

Feel free to reach out

kristy.horan@unimelb.edu.au