U.S. FOOD & DRUG ADMINISTRATION

Overview and Introduction to Phylogenetics

Eric L. Stevens, Ph.D.
International Policy Analyst
International Affairs Staff
eric.stevens@fda.hhs.gov
Center for Food Safety and Applied Nutrition

Overview: Phylogenetics

- Understand the principal of bacterial evolution
- Understand the underlying principles in which phylogenetic trees are created
- Have a conceptual understanding of the different ways to create a phylogenetic tree
  - **AND HOW WGS DATA IS USED**
- Understand how to read a phylogenetic tree

Phylogeny

- The evolutionary history of a species or a group of species over time

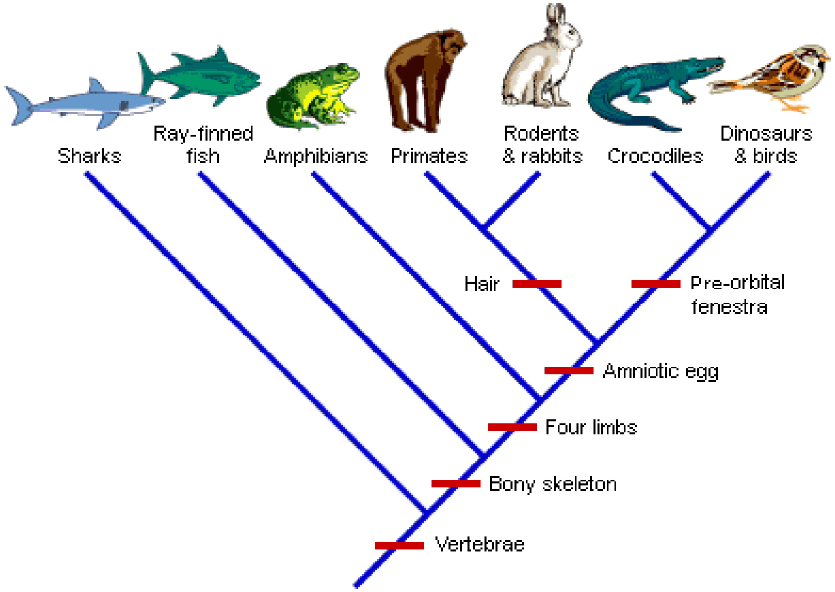
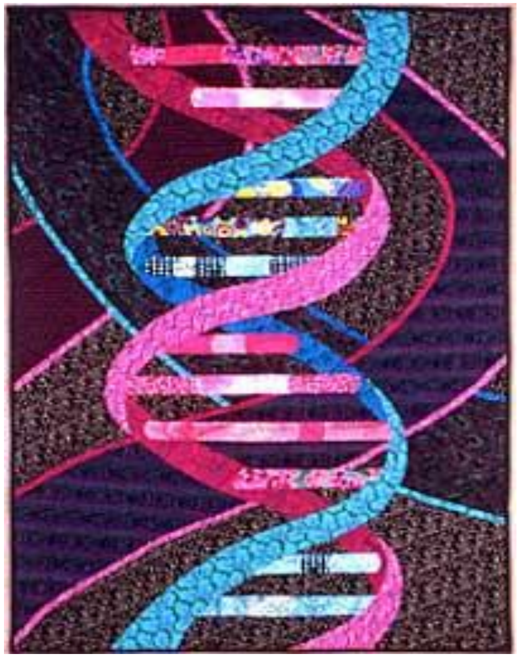


Image courtesy of http://evolution.berkeley.edu/evolibrary/article/0_0_0/evo_03

Office of Regulatory Science

Molecular data **vs.** Morphology / Physiology



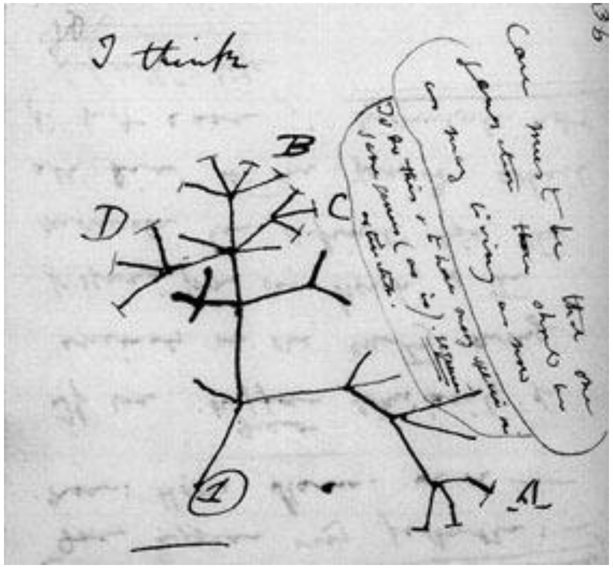
Molecular Data

- Strictly heritable
- Unambiguous data
- Quantitative
- Homology assessment easy
- Relationship of distantly related organisms can be inferred
- Abundant and easily generated

Morphology

- Can be influenced by environmental factors
- Ambiguous modifiers
- Qualitative
- Homology assessment difficult
- Close relationships can be inferred
- Problems when working with reduced visible morphology

WHAT IS PHYLOGENETICS?



Darwin's sketch 1836: the first phylogenetic tree?

- The study of evolutionary history/relationships among organisms or species based on heritable traits (DNA)
  - Homologous sequences
- Includes Taxonomy:
  - The classification and naming of species

Cell Division and Lineages

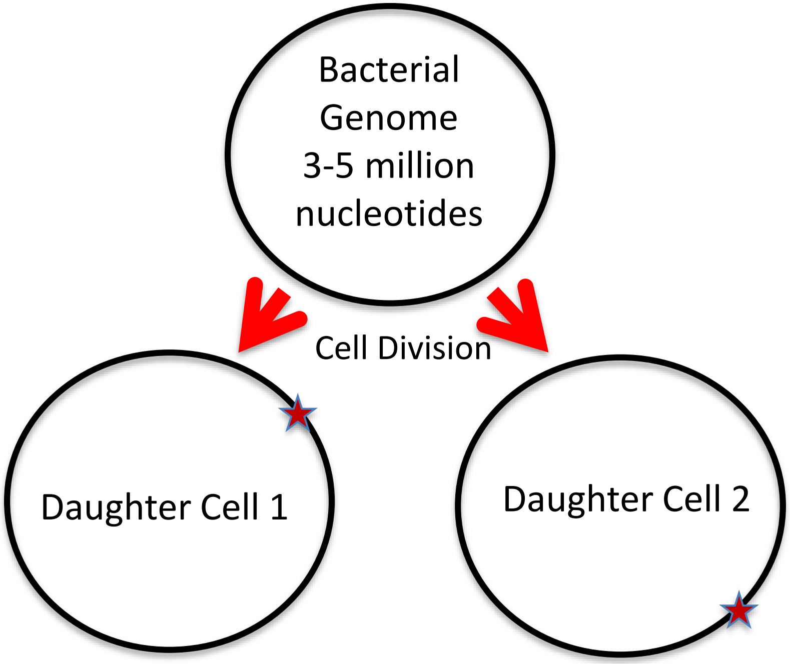
Overview:

- Daughter cells have the same genotype as their parent cell (plus any mutations or plasmid/phage incorporation)

- Overtime the lineages that descended from a single cell will acquire enough mutations (SNPs) to be differentiated from each other

- Estimating the shared ancestor of all lineages is important in determining if the DNA from this bacteria from this clinical patient is related to the DNA from that bacteria that was contaminating that food that the person ate.

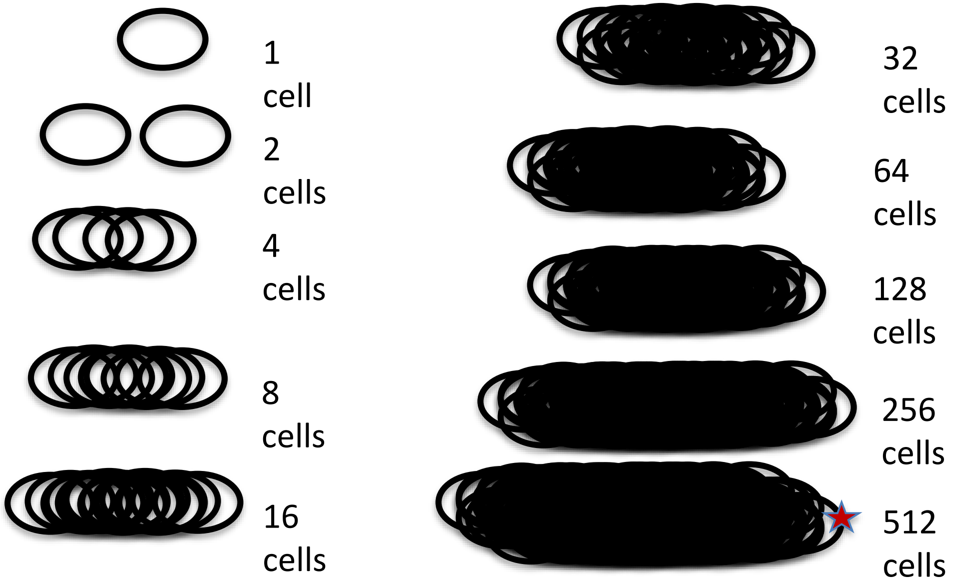
Understanding bacterial evolution and genetic relatedness



- A bacterial cell replicates its genetic material and then divides in half.

- Sometimes during replication a mutation occurs in the DNA and the genome of the daughter cell might be slightly different than the parent

Detecting a mutation based on cell division



Overview:

- Assuming a mutation rate of 0.003 mutations per genome per cell division it would take 9 cell divisions to see a mutation in a single cell (out of 512 cells).

Detecting a **mutation** based on cell division

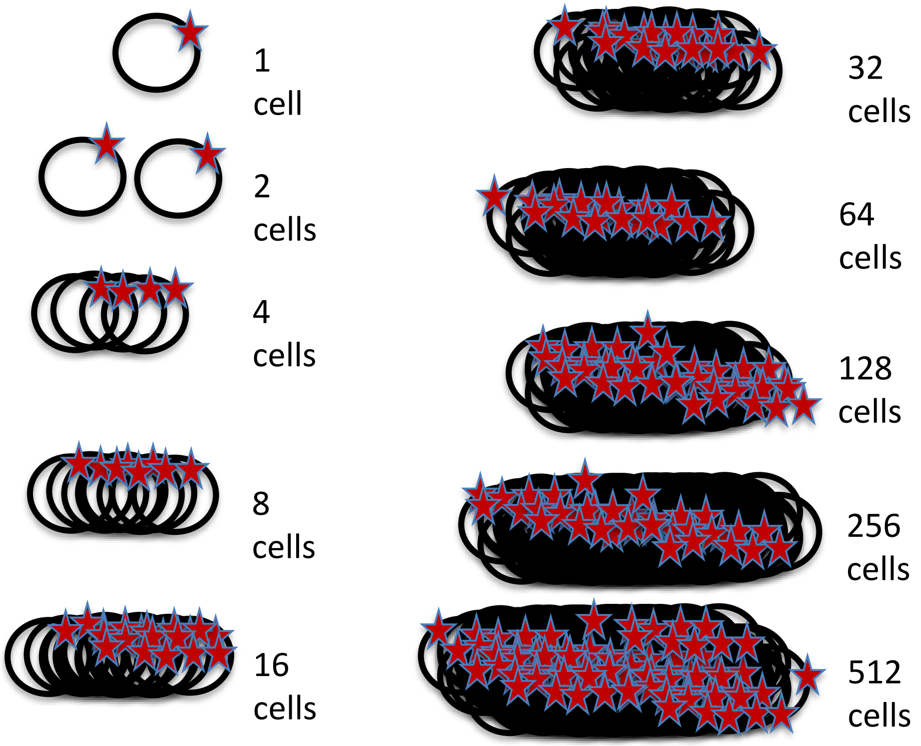
```
AGGATTGTTGGCAG
 GGAATGTTGGCAGT
  GAATGTTGGCAGTC
   AATGTTGGCAGTCG
__________________
AGGAATGTTGGCAGTCG
```

Overview:

- If the group of 512 cells were used for sequencing, then then the fraction of reads with the mutation (or variant) would not be high enough to detect by current sequencing technologies

- Remember, when condensing 4 reads into one genome sequence, if three of the reads show A and the other shows a T, will select the A as what the nucleotide is at that position

Detecting a mutation

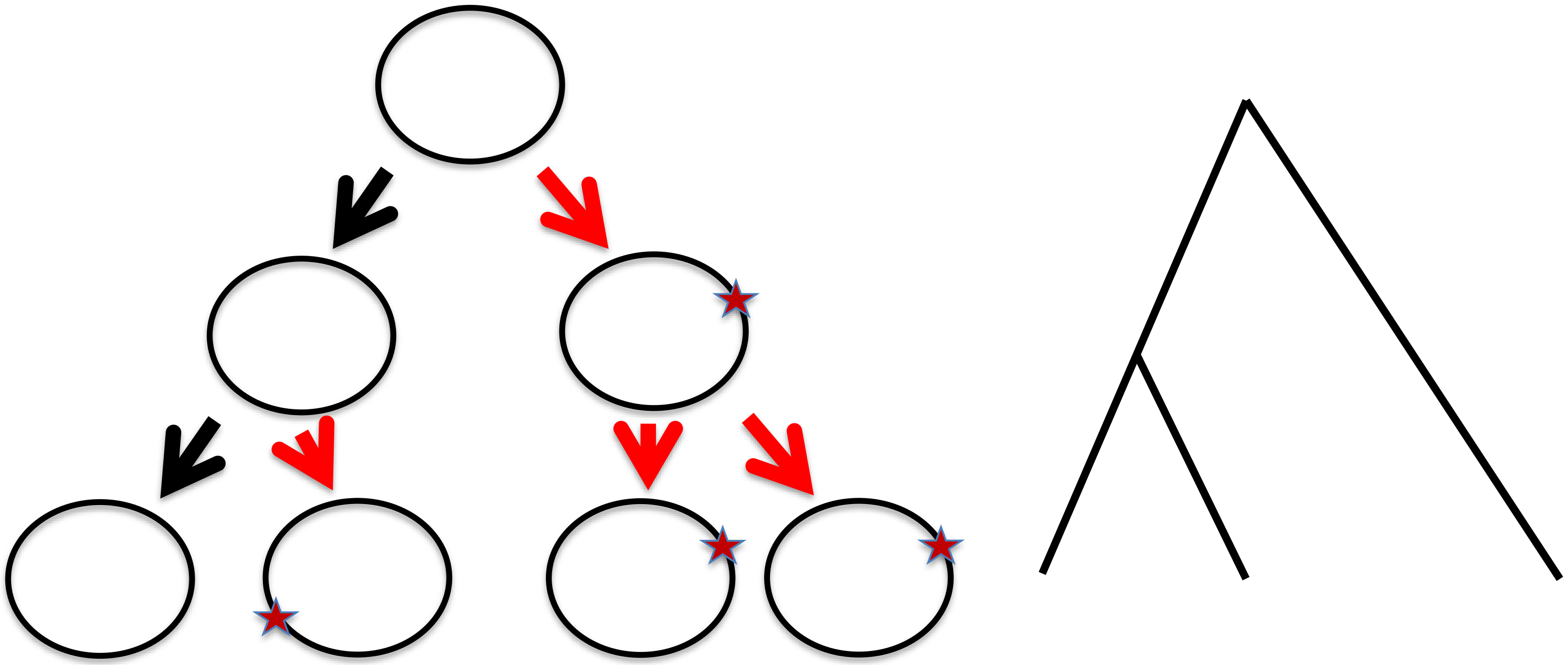


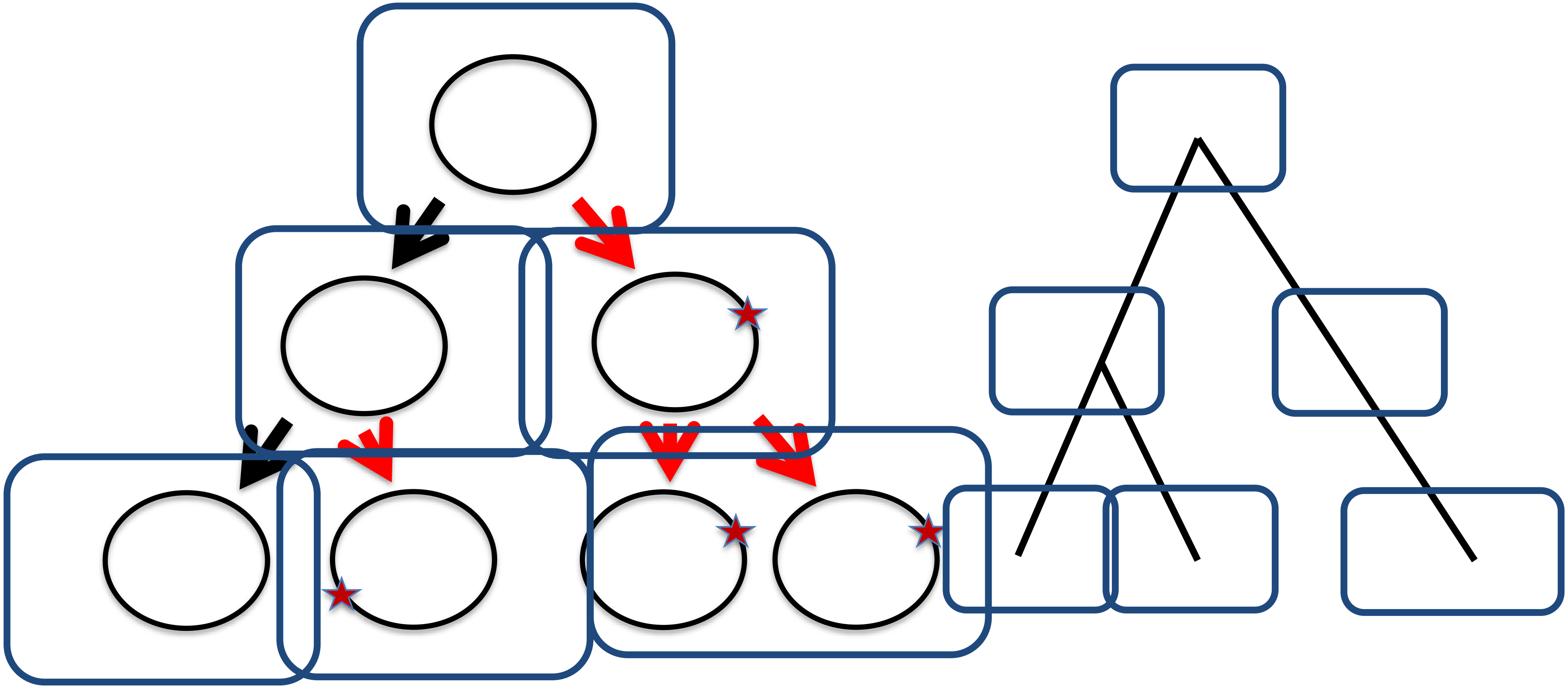
```
AGGATTGTTGGCAG
 GGATTGTTGGCAGT
  GATTGTTGGCAGTC
   ATTGTTGGCAGTCG

AGGATTGTTGGCAGTCG
```

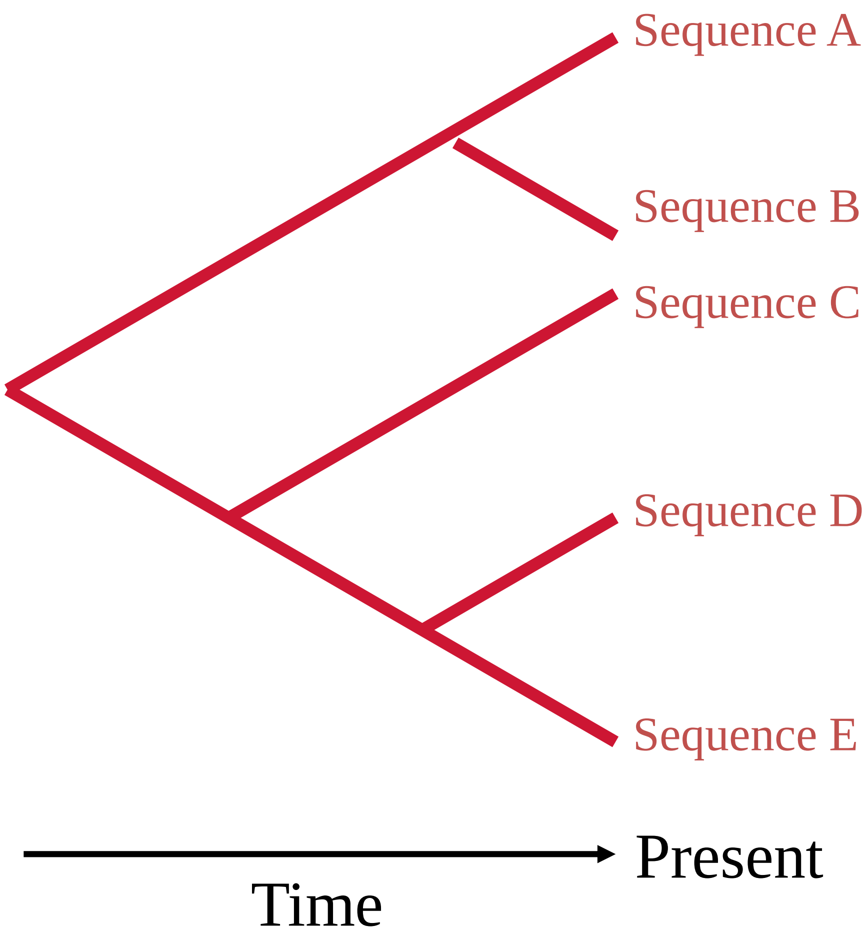
Cell Division and Lineages

Overview: This is where **Phylogenetics** is helpful!





Phylogenetic concepts: Interpreting a Phylogeny

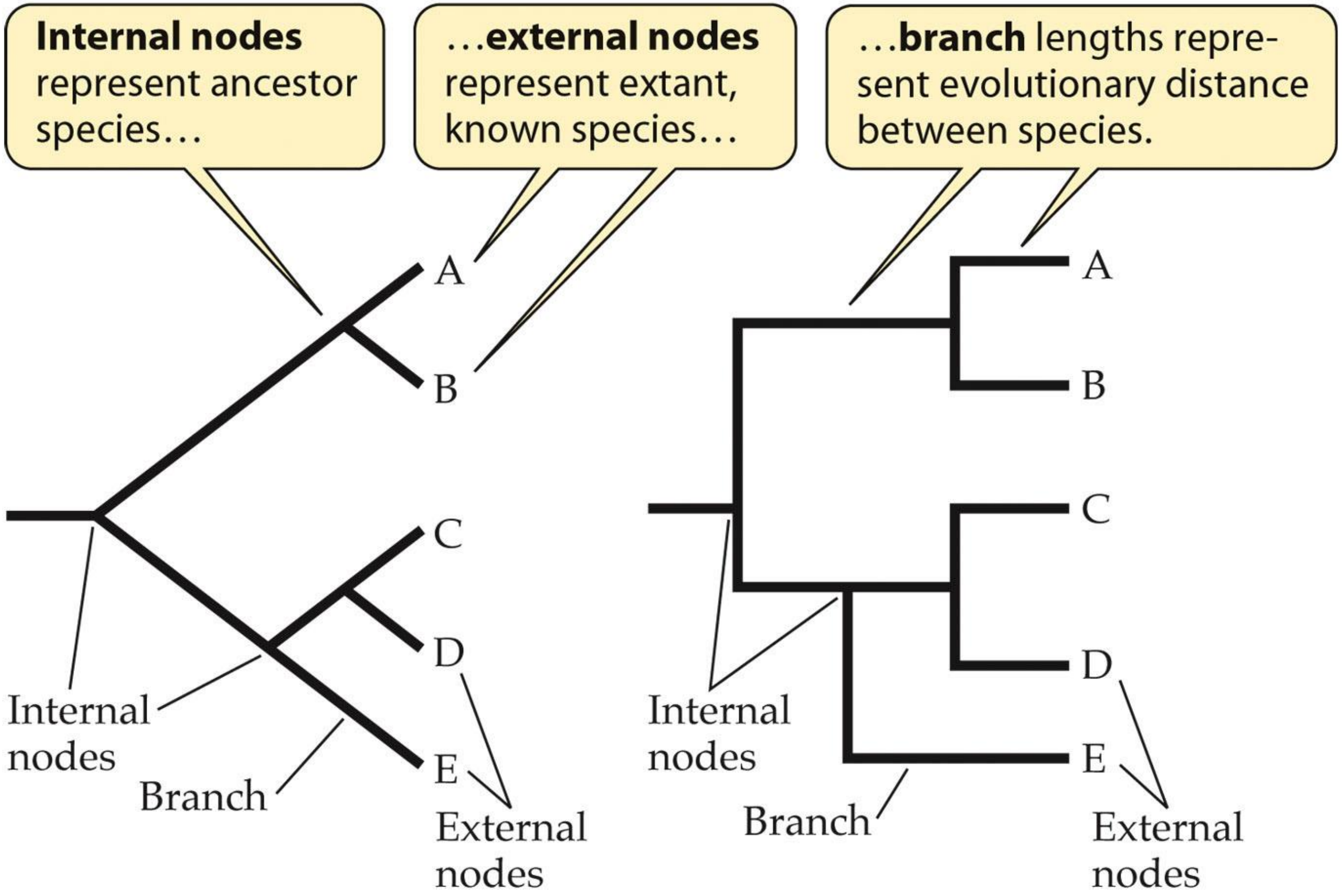


Constructing Phylogenetic trees

- What is the goal of your work?

- Aligning homologous sequences
  - DNA/RNA/protein
  - Are the groups closely related or distantly?
  - What sequences are you choosing?

Phylogenetic trees can be represented in several ways



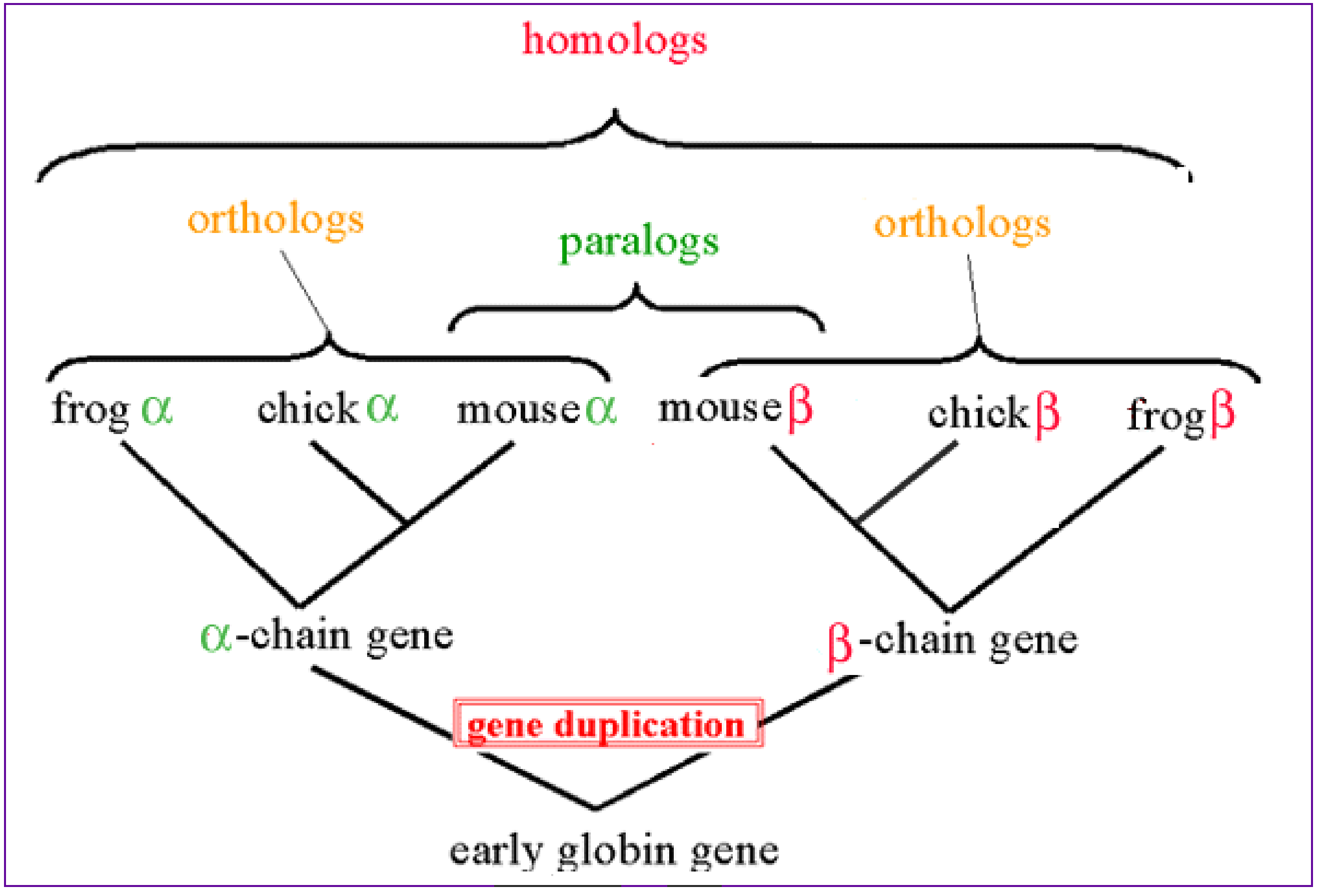
Internal nodes represent ancestor species…

…**external nodes** represent extant, known species…

…**branch** lengths represent evolutionary distance between species.

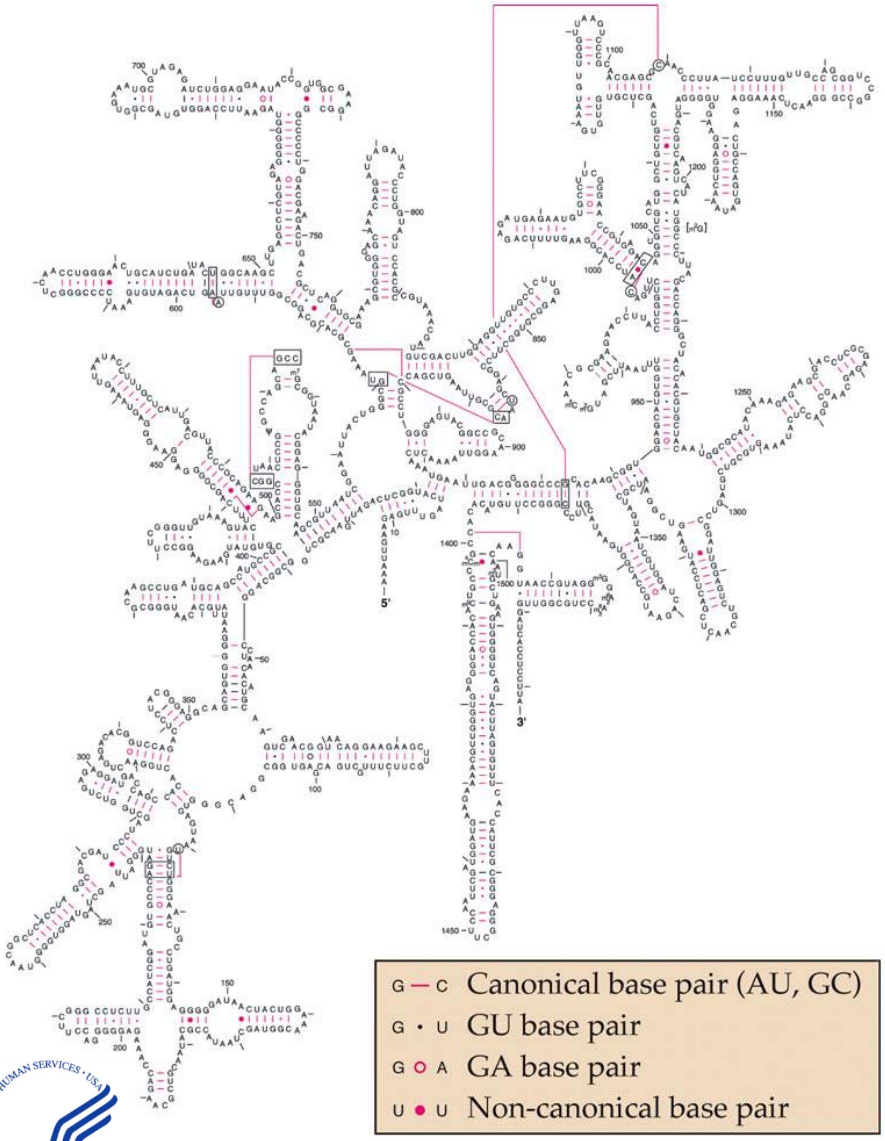
MICROBIAL LIFE, **Figure 17.4** © 2002 Sinauer Associates, Inc.

What data goes into making the tree is important



From Bioinformatics, Baxevanis and Ouellette, 2nd Edition, 2001, p. 327, Wiley Pub.

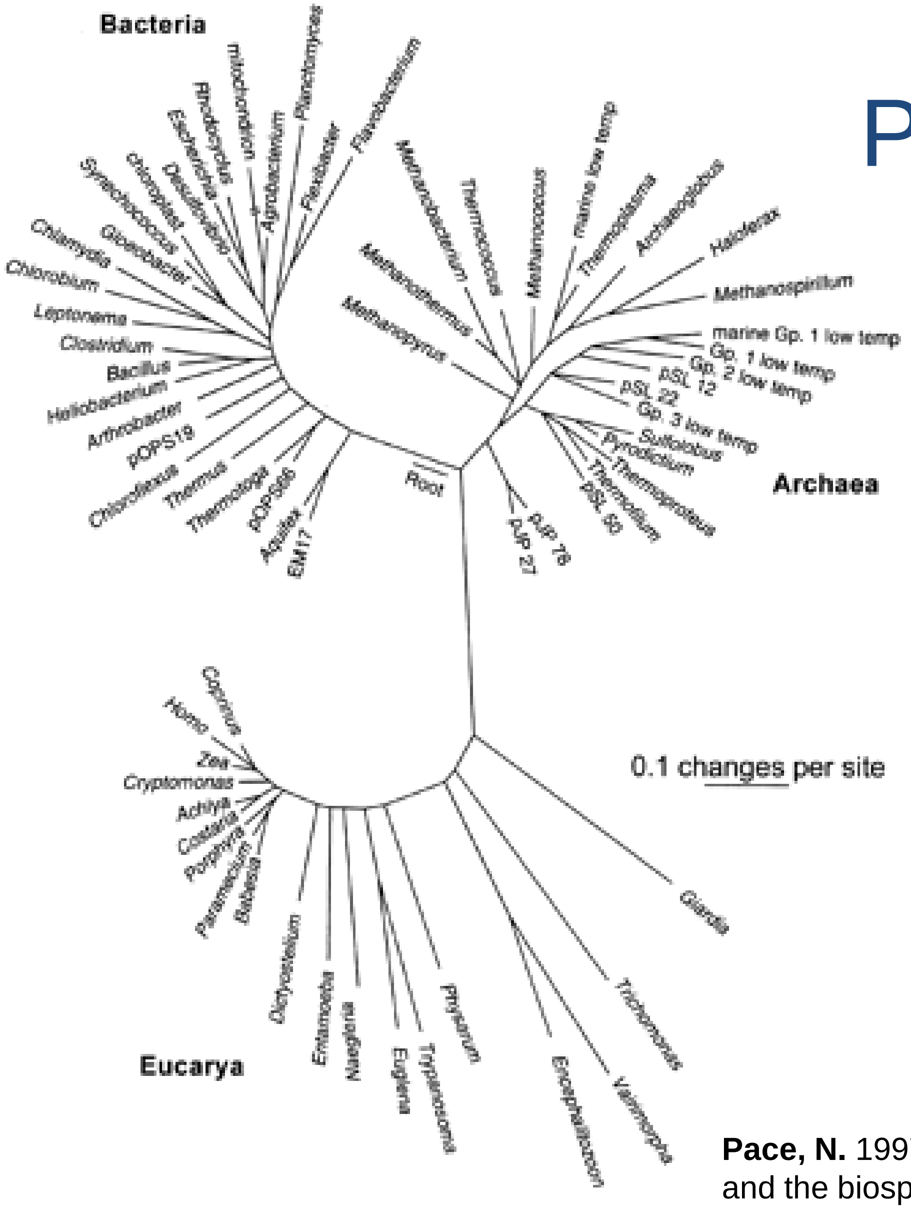
Example: 16S RNA


5'
3'
G — C Canonical base pair (AU, GC)
G • U GU base pair
G ○ A GA base pair
U ● U Non-canonical base pair

MICROBIAL LIFE, Figure 17.1 © 2002 Sinauer Associates, Inc.

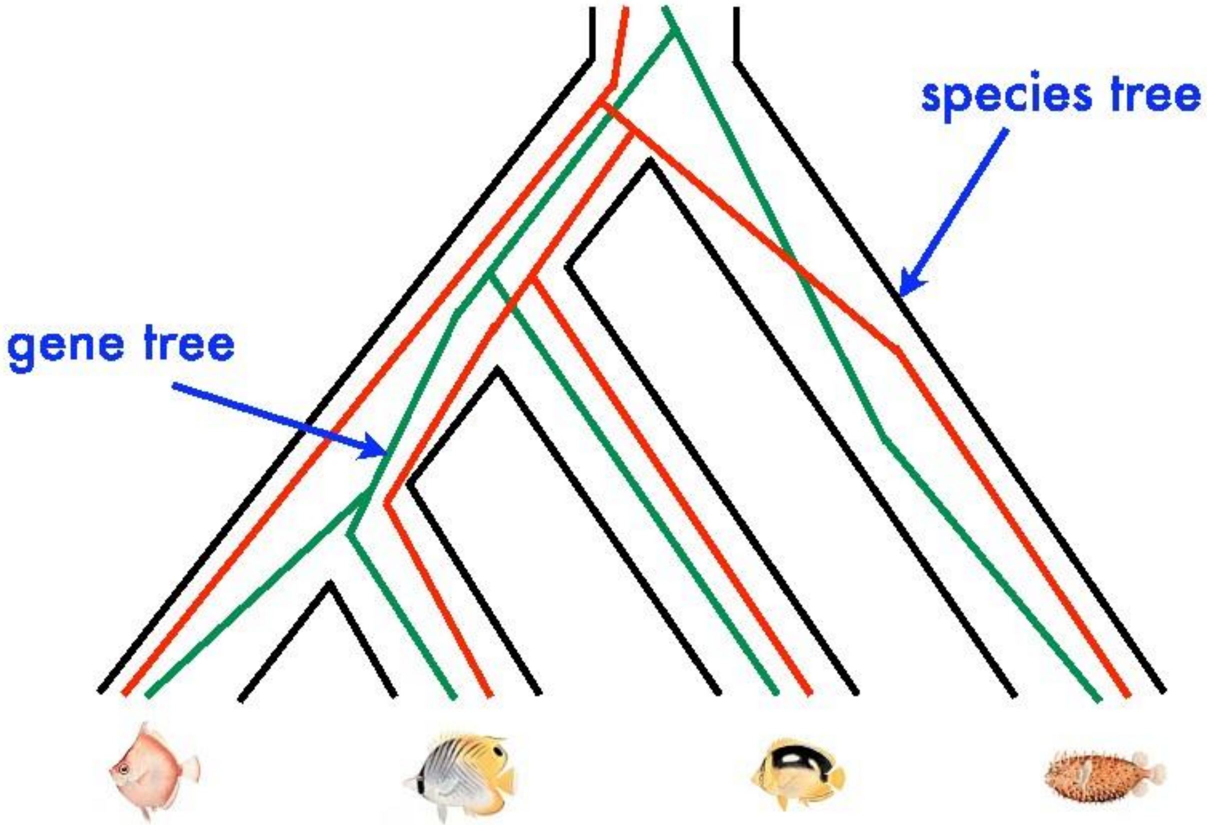
- Secondary structure is shown
- Highly conserved among all species*
  - i.e. homologous sequence (from a common ancestor)

16S rRNA Phylogenetic Tree



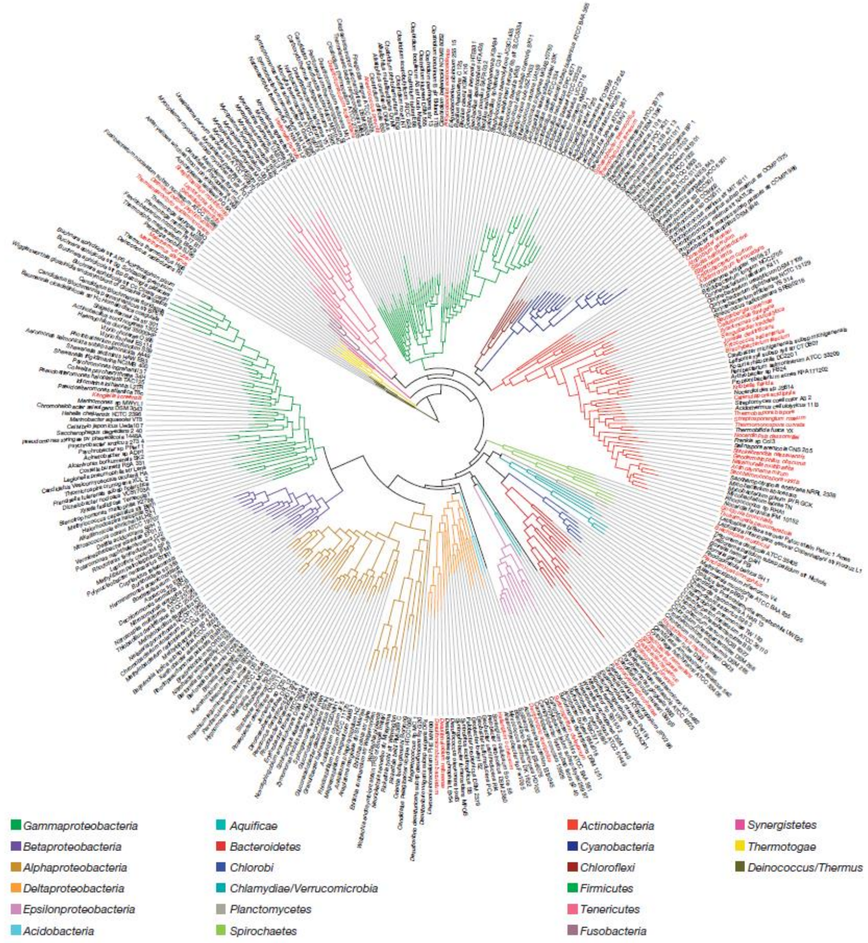
Pace, N. 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734–740.

Gene trees can be different from a Species tree!



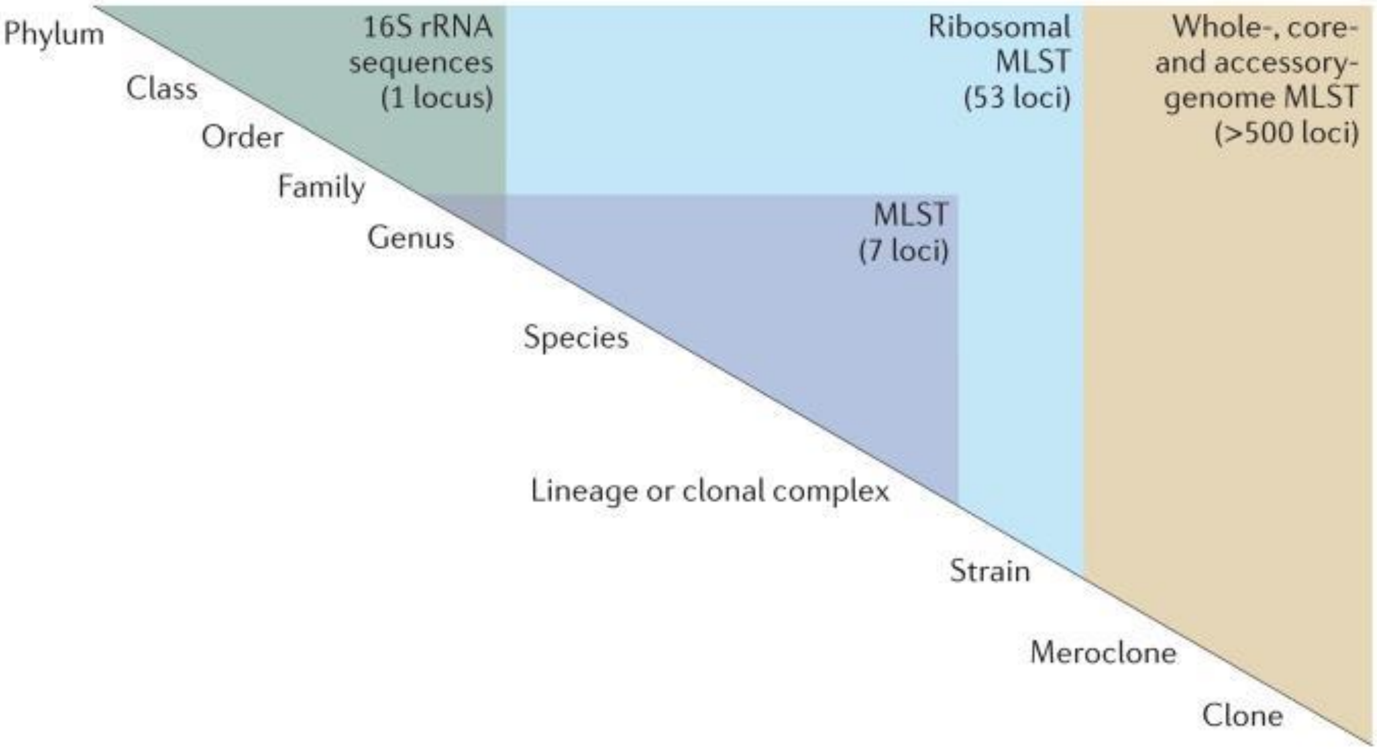
http://www.math.duke.edu/mathbio/proj_stat.html

Bacteria Tree from 31 genes



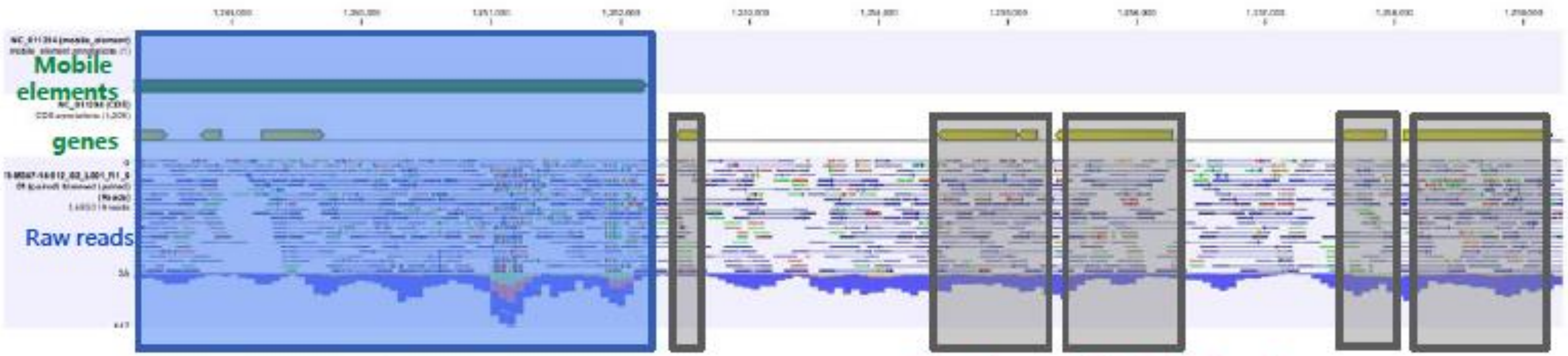
Wu et al. Vol 462|24/31 December 2009| doi:10.1038/nature08656 1056

Discriminating power of increasing sequence data


Phylum
Class
Order
Family
Genus
Species
Lineage or clonal complex
Strain
Meroclone
Clone
16S rRNA sequences (1 locus)
Ribosomal MLST (53 loci)
MLST (7 loci)
Whole-, core- and accessory-genome MLST (>500 loci)

Nat Rev Microbiol. Oct 2013; 11(10): 728–736.

Where to call a SNP?



Mask mobile elements
-do no consider SNPs in this location

Only call SNPs in genes

- Not all SNP pipelines are equal – where you call SNPs will affect the total SNP count
- SNPs relevant for phylogenetic analysis are vertically transmitted, not horizontally, so horizontal genetic elements like phages can be masked

Constructing Phylogenetic Trees

- A tree is characterized by how it looks (<u>topology</u>) and its branch lengths
  - Branches represent time or proportional to number of changes

- Three main methods for construction:
  - Parsimony
  - Distance-based
  - Maximum Likelihood

Constructing Phylogenetic Trees

- Trees can be rooted
  - Evolutionary relationship is implied
  - Use an outgroup to “root”
  - Example: *S. bongori* is the outgroup and roots the tree for *S. enterica*

- Trees can be unrooted
  - No evolutionary directionality
  - Want to know which are more alike

Constructing Phylogenetic Trees

Neighbor-joining	Maximum parsimony	Maximum likelihood
Very fast	Slow	***Very*** slow
Easily trapped in local optima	Assumptions fail when evolution is rapid	Highly dependent on assumed evolution model
Good for generating tentative tree, or choosing among multiple trees	Best option when tractable (<30 taxa, strong conservation)	Good for very small data sets and for testing trees built using other methods

Parsimony

- What is the tree that requires the fewest evolutionary changes to explain the data
  - i.e. fewest number of mutations to explain sequence variation

- Directly based on the sequence
  - Does not take into account revertant mutation
  - Does not take into consideration types of mutation (transition vs transversion)

Parsimony Example

		1	2	3	4	5	6	Position
Sequence	1	G	G	G	G	G	G	
	2	G	G	G	A	G	T	
	3	G	G	A	T	A	G	
	4	G	A	T	C	A	T	
		Uninformative	Uninformative	Uninformative	Uninformative	Informative	Informative	

Parsimony Example

		1	2	3	4	5	6	Position
Sequence	1	G	G	G	G	G	G	
	2	G	G	G	A	G	T	
	3	G	G	A	T	A	G	
	4	G	A	T	C	A	T	
		Uninformative	Uninformative	Uninformative	Uninformative	Informative	Informative	

Maximum Likelihood

- Want to find a tree that will maximize the chance that the tree is correct
- Probabilities (likelihoods) are considered for every mutation (nucleotide substitution) for a multiple sequence alignment.
    - Transitions are more likely than transversions (~3:1)
        - If C, T, and G are represented at a site, The sequences that have C and T are probably closer*

Maximum Likelihood

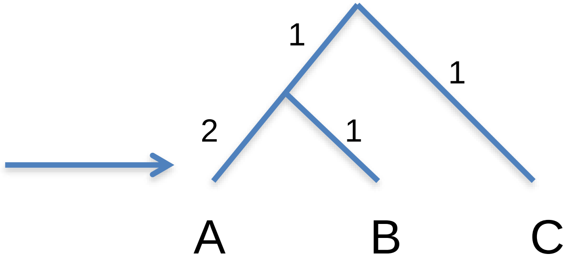
- Each base of every position of every sequence is considered separately (independent) and given a log-likelihood value and the sum is used to estimate branch lengths
  - For every Topology (possible tree)!!!

- Good theoretical background
- General Consistent
- Computationally expensive (a lot of time)

Distance-based

- Construct a distance matrix for each pair of sequences (e.g. how many differences)

A	0		
B	3	0	
C	4	3	0
	A	B	C



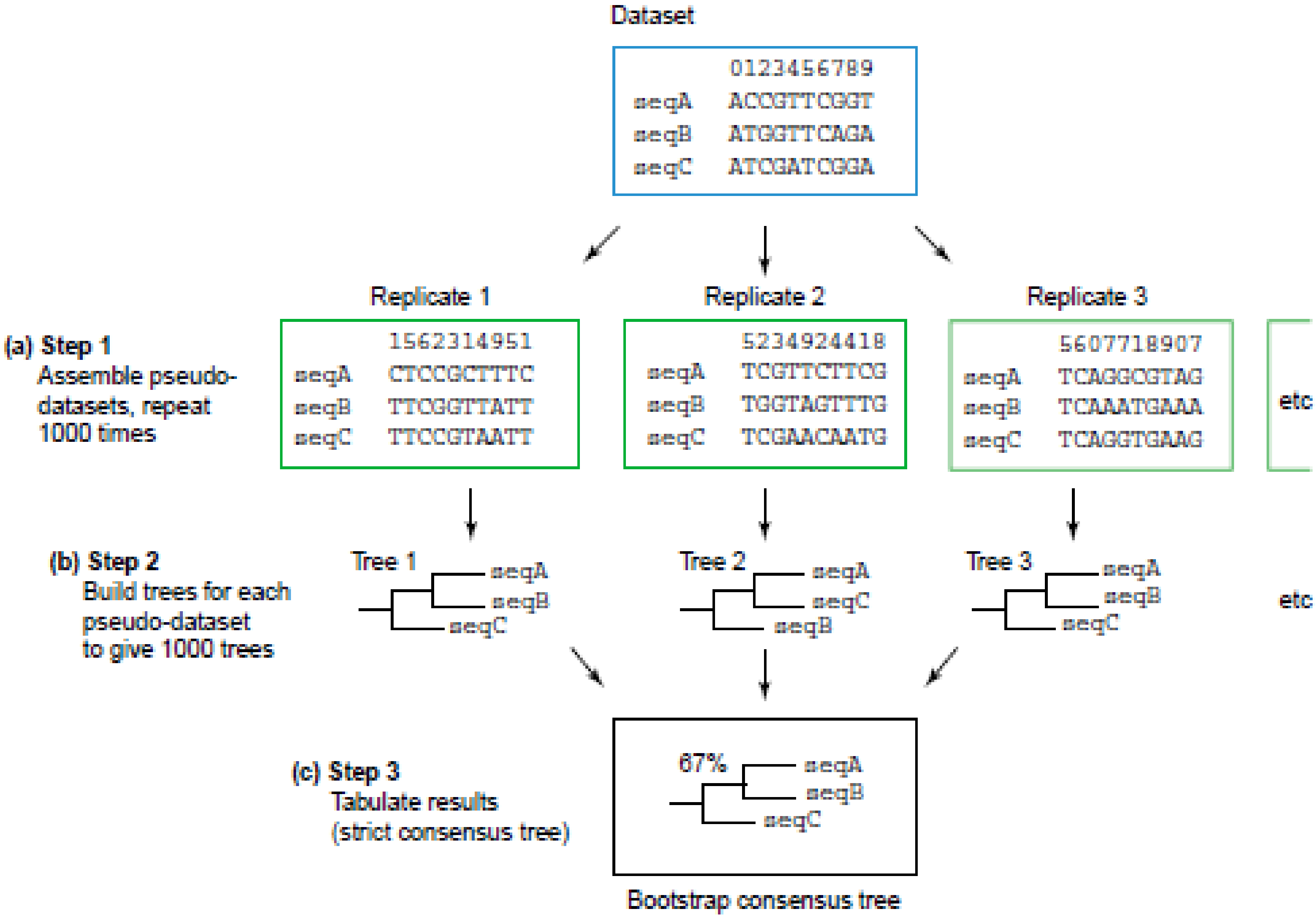
- That distance matrix represents one tree
- Great for very similar sequences
- Very fast
- Loses information

ASSESSING CONFIDENCE IN TREES

- Measure of confidence in the inferred tree.
  - Is the tree likely to change if we got more data, or if we had used slightly different data?
  - Are some parts of the tree more robust than others?

- **Bootstrapping**
  - Create multiple new alignments by resampling the columns of the observed data matrix
  - Construct a tree for the ‘bootstrap’ alignment
  - The bootstrap support for each branch is the % of bootstrap trees that branch appears in.

Assessing Confidence In Trees



Dataset

```
       0123456789
seqA   ACCGTTCGGT
seqB   ATGGTTCAGA
seqC   ATCGATCGGA
```

(a) **Step 1**
Assemble pseudo-datasets, repeat 1000 times

Replicate 1

```
       1562314951
seqA   CTCCGCTTTC
seqB   TTCGGTTATT
seqC   TTCCGTAATT
```

Replicate 2

```
       5234924418
seqA   TCGTTCTTCG
seqB   TGGTAGTTTG
seqC   TCGAACAATG
```

Replicate 3

```
       5607718907
seqA   TCAGGCGTAG
seqB   TCAAATGAAA
seqC   TCAGGTGAAG
```

etc

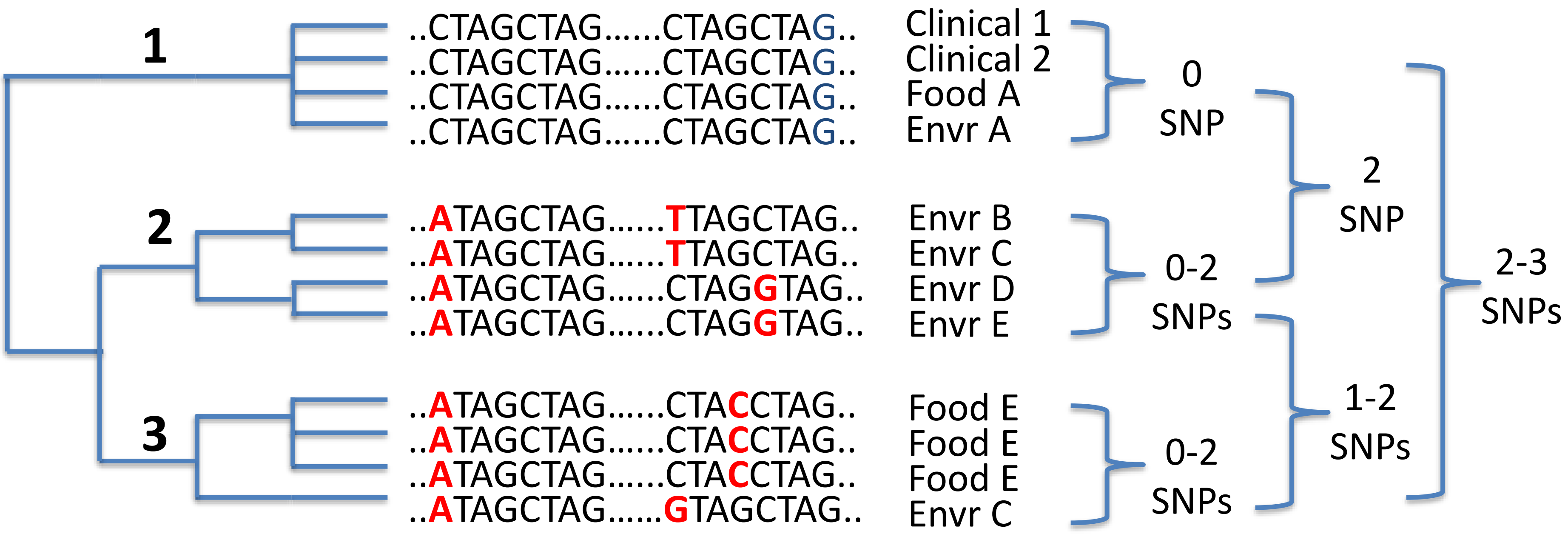
(b) **Step 2**
Build trees for each pseudo-dataset to give 1000 trees

(c) **Step 3**
Tabulate results (strict consensus tree)

Bootstrap consensus tree

Baldauf 2003. TRENDS in Genetics Vol.19 No.6 June 2003

Goal of Phylogenetic Trees using WGS Data:
Infer evolutionary relationships based on nucleotide differences
And match clinical to food/environmental isolates



..CTAGCTAG……CTAGCTAG..	Clinical 1
..CTAGCTAG……CTAGCTAG..	Clinical 2
..CTAGCTAG……CTAGCTAG..	Food A
..CTAGCTAG……CTAGCTAG..	Envr A
..ATAGCTAG……TTAGCTAG..	Envr B
..ATAGCTAG……TTAGCTAG..	Envr C
..ATAGCTAG……CTAGGTAG..	Envr D
..ATAGCTAG……CTAGGTAG..	Envr E
..ATAGCTAG……CTACCTAG..	Food E
..ATAGCTAG……CTACCTAG..	Food E
..ATAGCTAG……CTACCTAG..	Food E
..ATAGCTAG……GTAGCTAG..	Envr C

0 SNP

0-2 SNPs

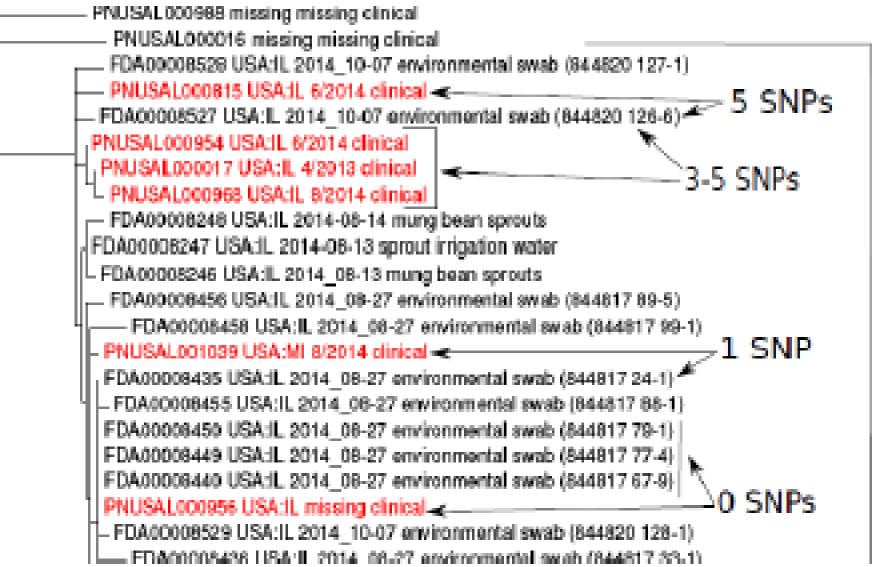
0-2 SNPs

2 SNP

1-2 SNPs

2-3 SNPs

High-Quality Draft


PNUSAL000988 missing missing clinical
PNUSAL000016 missing missing clinical
FDA00008528 USA:IL 2014_10-07 environmental swab (844820 127-1)
PNUSAL000815 USA:IL 6/2014 clinical
FDA00008527 USA:IL 2014_10-07 environmental swab (844820 126-6)
PNUSAL000954 USA:IL 6/2014 clinical
PNUSAL000017 USA:IL 4/2013 clinical
PNUSAL000968 USA:IL 8/2014 clinical
FDA00008248 USA:IL 2014-08-14 mung bean sprouts
FDA00008247 USA:IL 2014-08-13 sprout irrigation water
FDA00008246 USA:IL 2014_08-13 mung bean sprouts
FDA00008456 USA:IL 2014_08-27 environmental swab (844817 89-5)
FDA00008458 USA:IL 2014_08-27 environmental swab (844817 99-1)
PNUSAL001039 USA:MI 8/2014 clinical
FDA00008435 USA:IL 2014_08-27 environmental swab (844817 24-1)
FDA00008455 USA:IL 2014_08-27 environmental swab (844817 88-1)
FDA00008450 USA:IL 2014_08-27 environmental swab (844817 79-1)
FDA00008449 USA:IL 2014_08-27 environmental swab (844817 77-4)
FDA00008440 USA:IL 2014_08-27 environmental swab (844817 67-9)
PNUSAL000956 USA:IL missing clinical
FDA00008529 USA:IL 2014_10-07 environmental swab (844820 128-1)
FDA00008436 USA:IL 2014_08-27 environmental swab (844817 33-1)

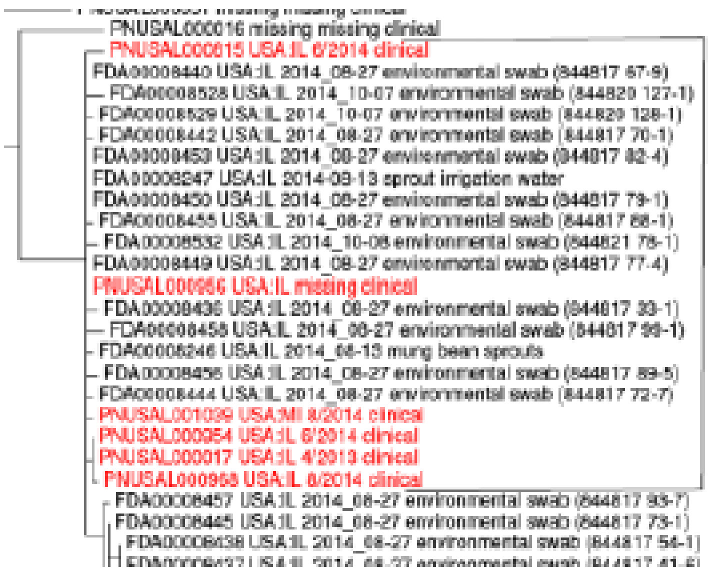
5 SNPs

3-5 SNPs

1 SNP

0 SNPs

Complete (PacBio)


PNUSAL000016 missing missing clinical
PNUSAL000815 USA:IL 6/2014 clinical
FDA00008440 USA:IL 2014_08-27 environmental swab (844817 67-9)
FDA00008528 USA:IL 2014_10-07 environmental swab (844820 127-1)
FDA00008529 USA:IL 2014_10-07 environmental swab (844820 128-1)
FDA00008442 USA:IL 2014_08-27 environmental swab (844817 70-1)
FDA00008453 USA:IL 2014_08-27 environmental swab (844817 82-4)
FDA00008247 USA:IL 2014-08-13 sprout irrigation water
FDA00008450 USA:IL 2014_08-27 environmental swab (844817 79-1)
FDA00008455 USA:IL 2014_08-27 environmental swab (844817 88-1)
FDA00008532 USA:IL 2014_10-08 environmental swab (844821 78-1)
FDA00008449 USA:IL 2014_08-27 environmental swab (844817 77-4)
PNUSAL000956 USA:IL missing clinical
FDA00008436 USA:IL 2014_08-27 environmental swab (844817 33-1)
FDA00008458 USA:IL 2014_08-27 environmental swab (844817 99-1)
FDA00008246 USA:IL 2014_08-13 mung bean sprouts
FDA00008456 USA:IL 2014_08-27 environmental swab (844817 89-5)
FDA00008444 USA:IL 2014_08-27 environmental swab (844817 72-7)
PNUSAL001039 USA:MI 8/2014 clinical
PNUSAL000954 USA:IL 6/2014 clinical
PNUSAL000017 USA:IL 4/2013 clinical
PNUSAL000968 USA:IL 8/2014 clinical
FDA00008457 USA:IL 2014_08-27 environmental swab (844817 93-7)
FDA00008445 USA:IL 2014_08-27 environmental swab (844817 73-1)
FDA00008438 USA:IL 2014_08-27 environmental swab (844817 54-1)

0-6 SNPs

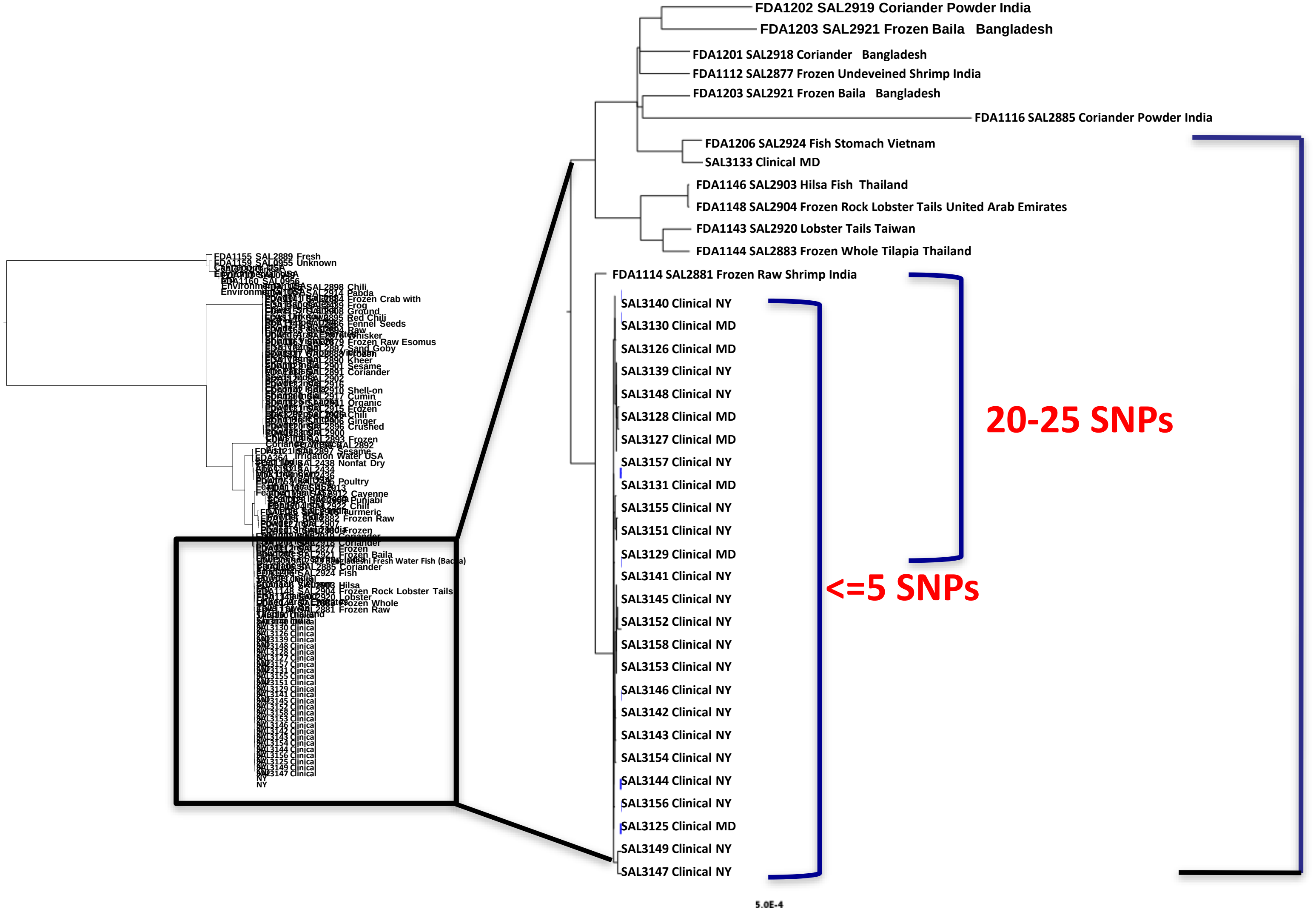
INTERPRETATION OF TREE & SNPS

1. Human mtDNA Forensic Testing Framework

2. Results binned into 3 groups
    1. Include/Match (<=20 SNPs)
    2. Inconclusive (20-100 SNPs)
    3. Exclude/Non-Match (>100 SNPs)

3. Statistical Odds Ratio method in development, databases growing

S. Bareilly Phylogeny



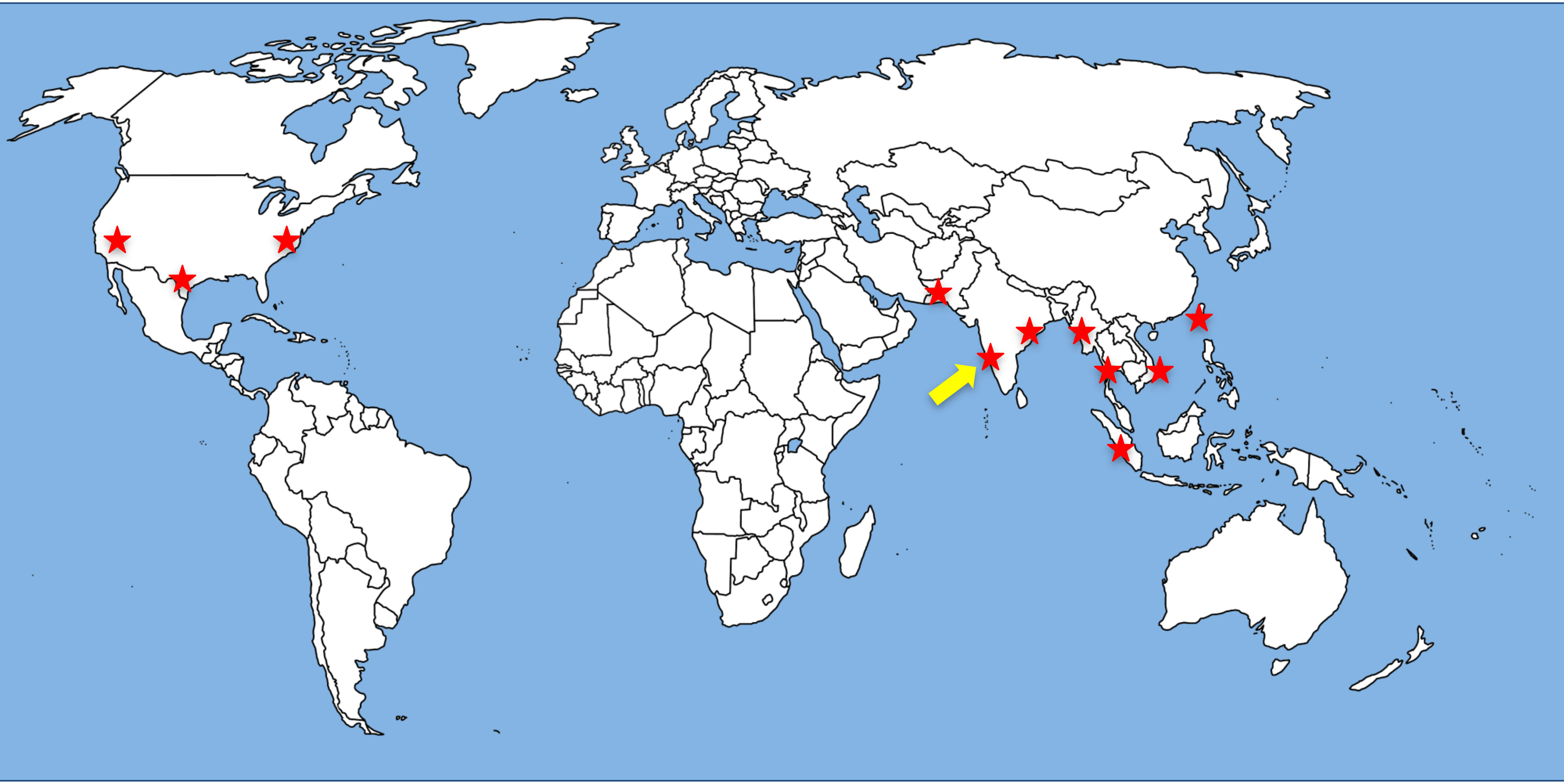
20-25 SNPs

<=5 SNPs

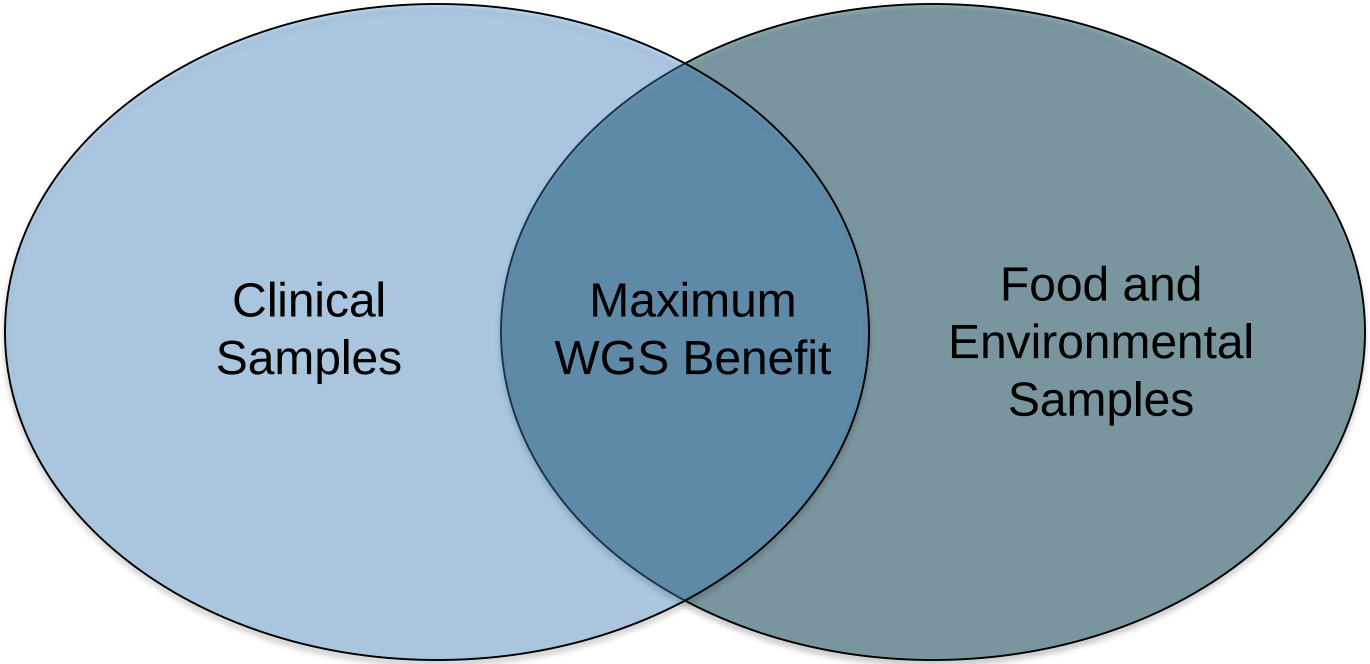
PFGE Match

110-130 SNPs

NGS distinguishes geographical structure among closely related *Salmonella* Bareilly strains



Importance of a Balanced Approach


Clinical Samples

Maximum WGS Benefit

Food and Environmental Samples

Note:

- These slides are for teaching purposes only and have been collected from images that I have made, from the CDC and FDA, and from around the web.

- The findings and conclusions in this report are those of the author and do not necessarily represent the official position of the Food and Drug Administration