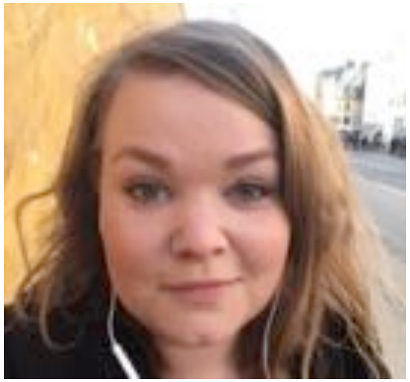
Whole Genome based Phylogeny

Johanne Ahrenfeldt
PhD student
DTU Bioinformatics

Short about me



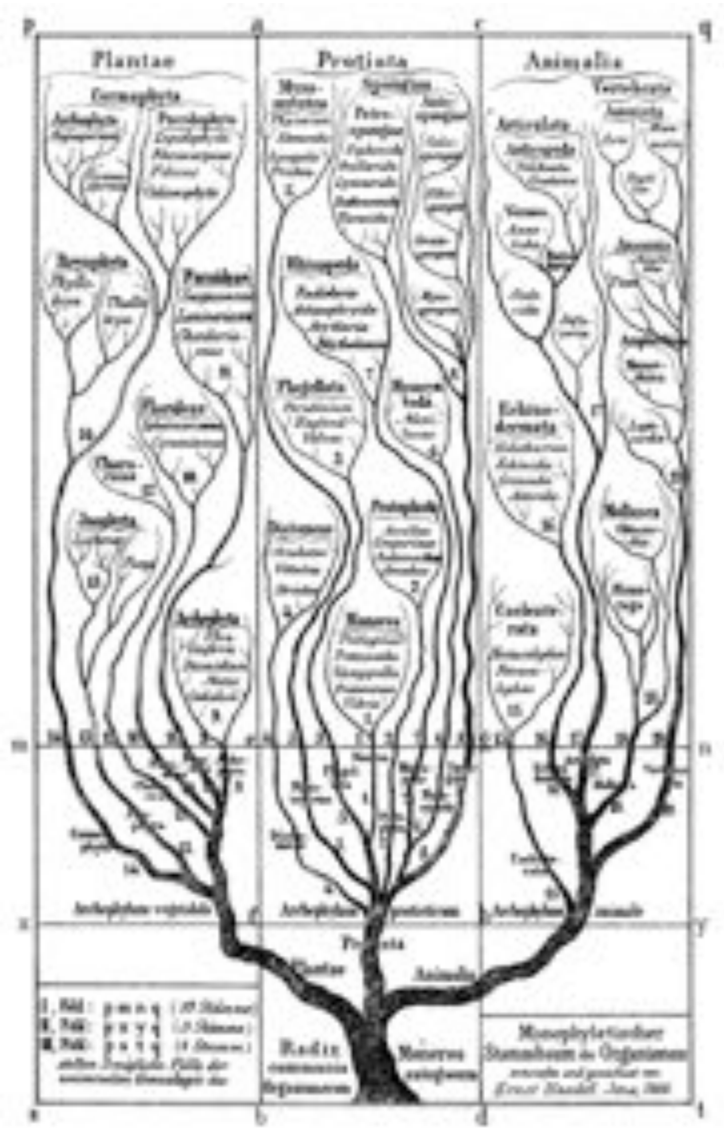
Johanne Ahrenfeldt

ja@bioinformatics.dtu.dk

- PhD student at DTU Bioinformatics
  - Whole Genome based Phylogeny
- Graduate Engineer in Systems Biology and Bioinformatics from Technical University of Denmark
- Working in the CGE project since 2012 – started as a student helper

Overview

- What is Phylogeny
- SNP methods
  - CSI Phylogeny
- Nucleotide Differences
  - NDtree
- Controlled Evolution study
- Good advice



What is phylogeny?

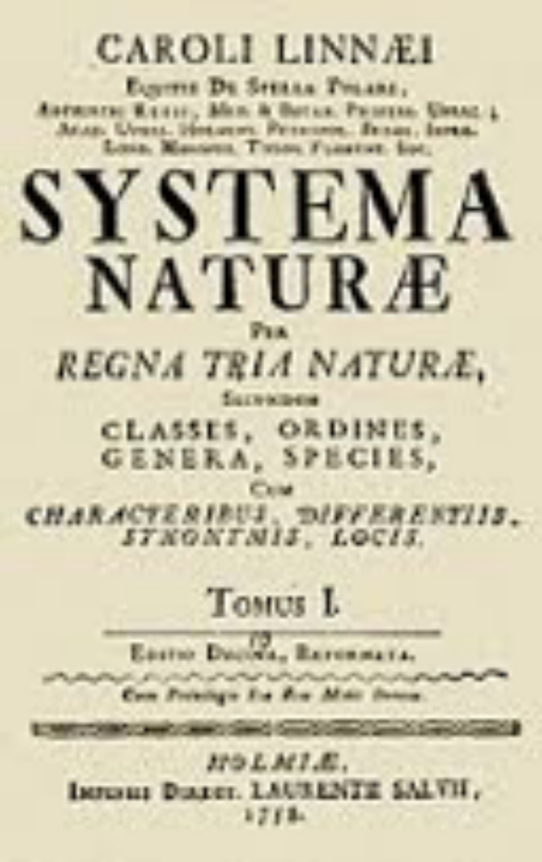
- Early phylogeny
  - Classification
  - Based on phenotypes
- Current phylogeny
  - Based on genotypes
  - DNA mutations as basis for evolution

Classification

Carl Linnaeus 1707-1778

Hierarchical system

- Kingdom
- Phylum
- Class
- Order
- Family
- Genus
- Species



CAROLI LINNÆI

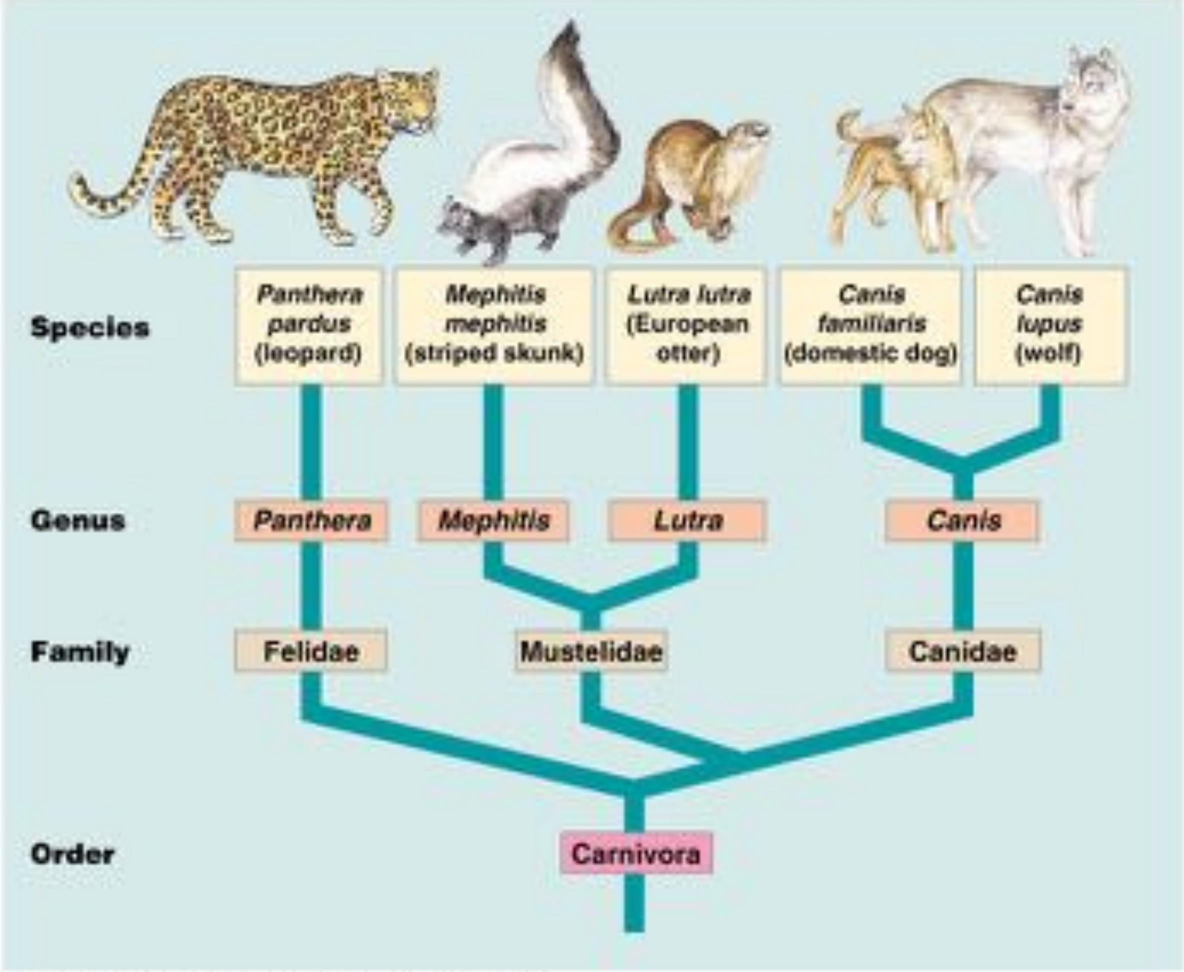
SYSTEMA
NATURÆ
PER
REGNA TRIA NATURÆ,
SECUNDUM
CLASSES, ORDINES,
GENERA, SPECIES,
CUM
CHARACTERIBUS, DIFFERENTIIS,
SYNONYMIS, LOCIS.

TOMUS I.

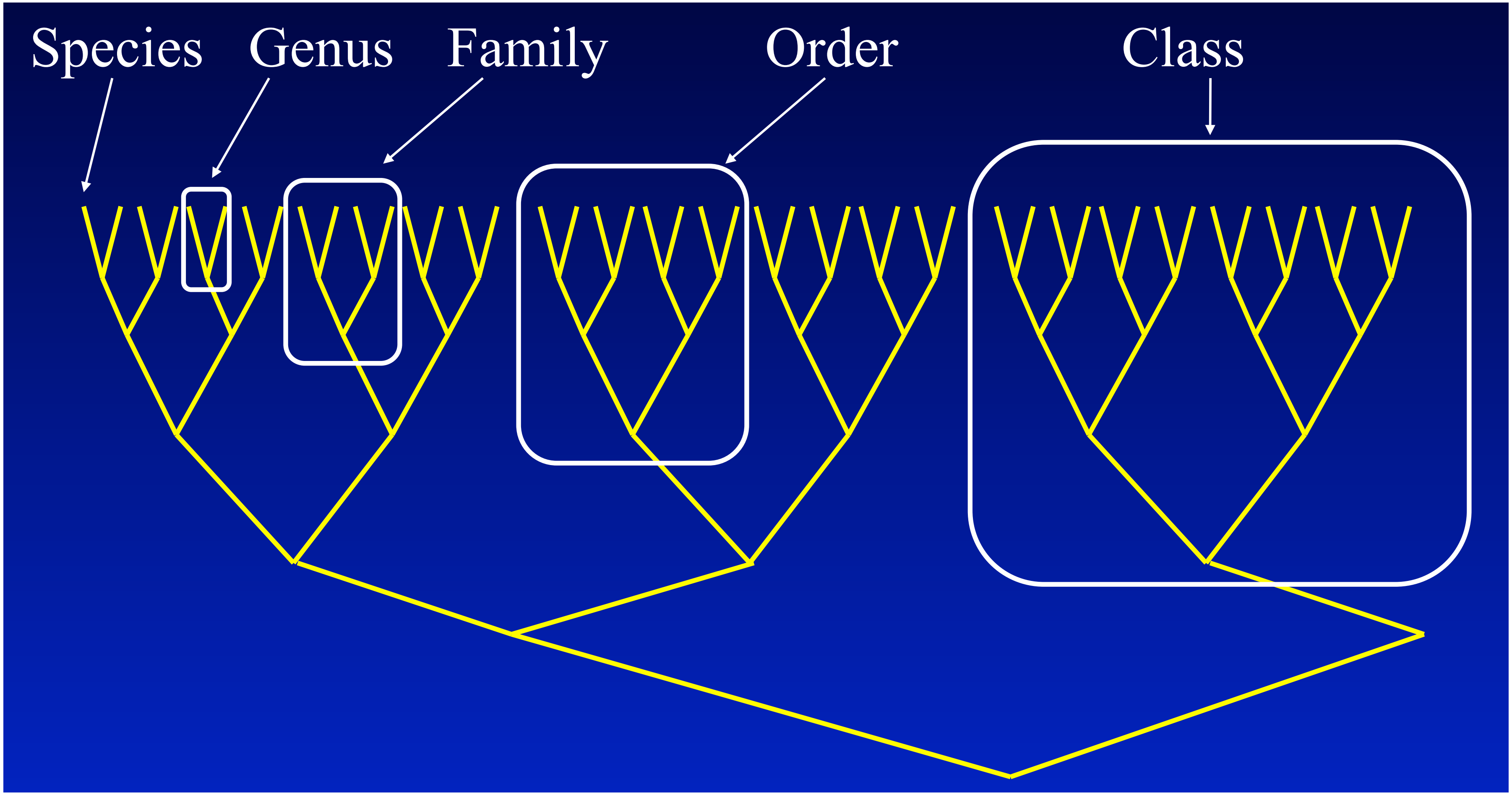
EDITIO DECIMA, REFORMATA.

HOLMIÆ,
LAURENTII SALVII,

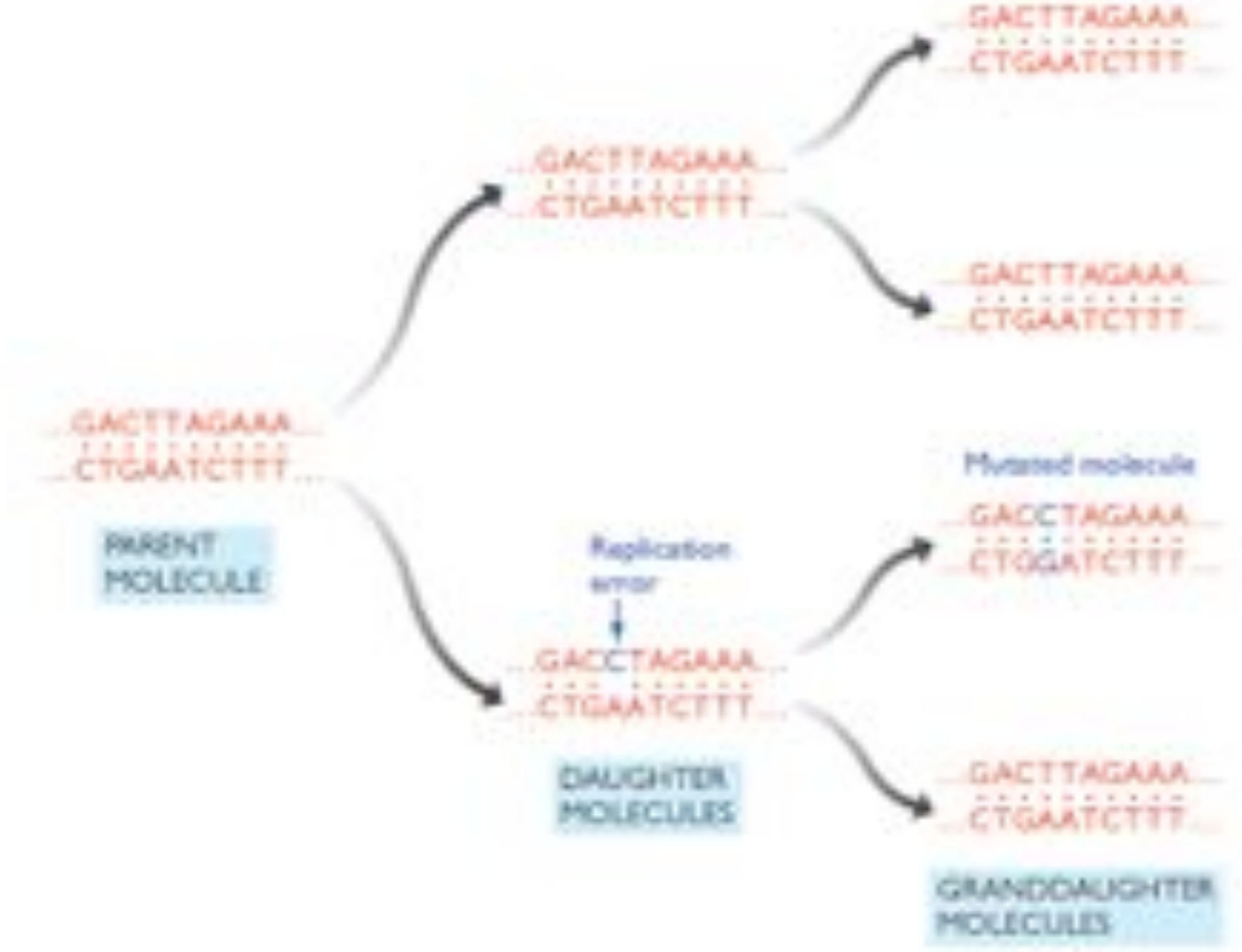
Classification depicted as a tree



Classification depicted as a tree



DNA mutations as basis for evolution



What are phylogenetic trees

- Phylogenetic trees are a visual representation of the genetic relationship between species
- Think of them as family trees
- Phylogeny can also be represented by distance matrices

What are phylogenetic trees

- 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
  - K-mer 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 (or other shared sequence) differs between members of a biological species or paired chromosomes.

How does it work

Strain A	AT**T**CA**G**T**A**G**T**
Strain B	AT**G**CA**G**T**T**G**A**
Strain C	AT**G**CA**A**T**T**G**T**
Strain D	AT**C**CA**T**T**A**G**C**

Construct distance matrix

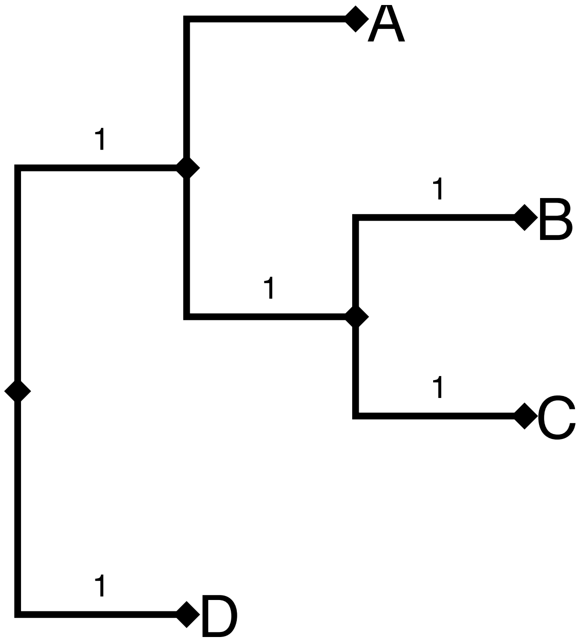
Strain A	AT**T**CA**G**T**A**G**T**
Strain B	AT**G**CA**G**T**T**G**A**
Strain C	AT**G**CA**A**T**T**G**T**
Strain D	AT**C**CA**T**T**A**G**C**

	A	B	C	D
A				
B				
C				
D				

Make Tree

Strain A	AT**T**CA**G**T**A**G**T**
Strain B	AT**G**CA**G**T**T**G**A**
Strain C	AT**G**CA**A**T**T**G**T**
Strain D	AT**C**CA**T**T**A**G**C**

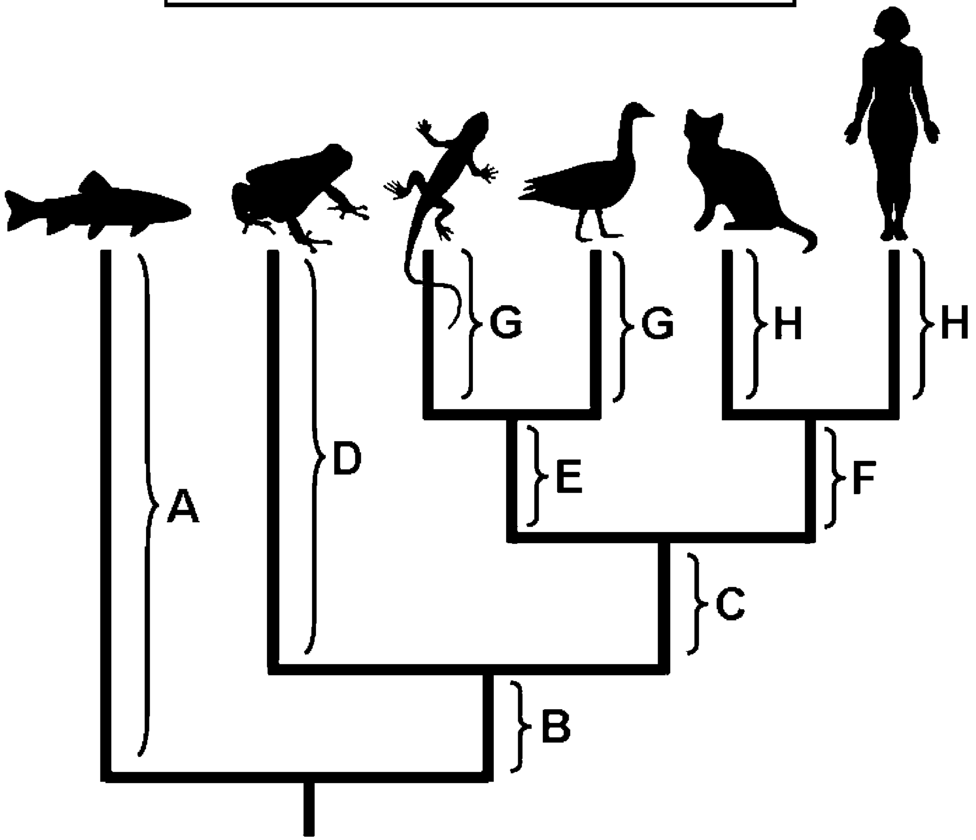
	A	B	C	D
A	0	3	3	3
B	3	0	2	4
C	3	2	0	4
D	3	4	4	0



How to read phylogenetic trees

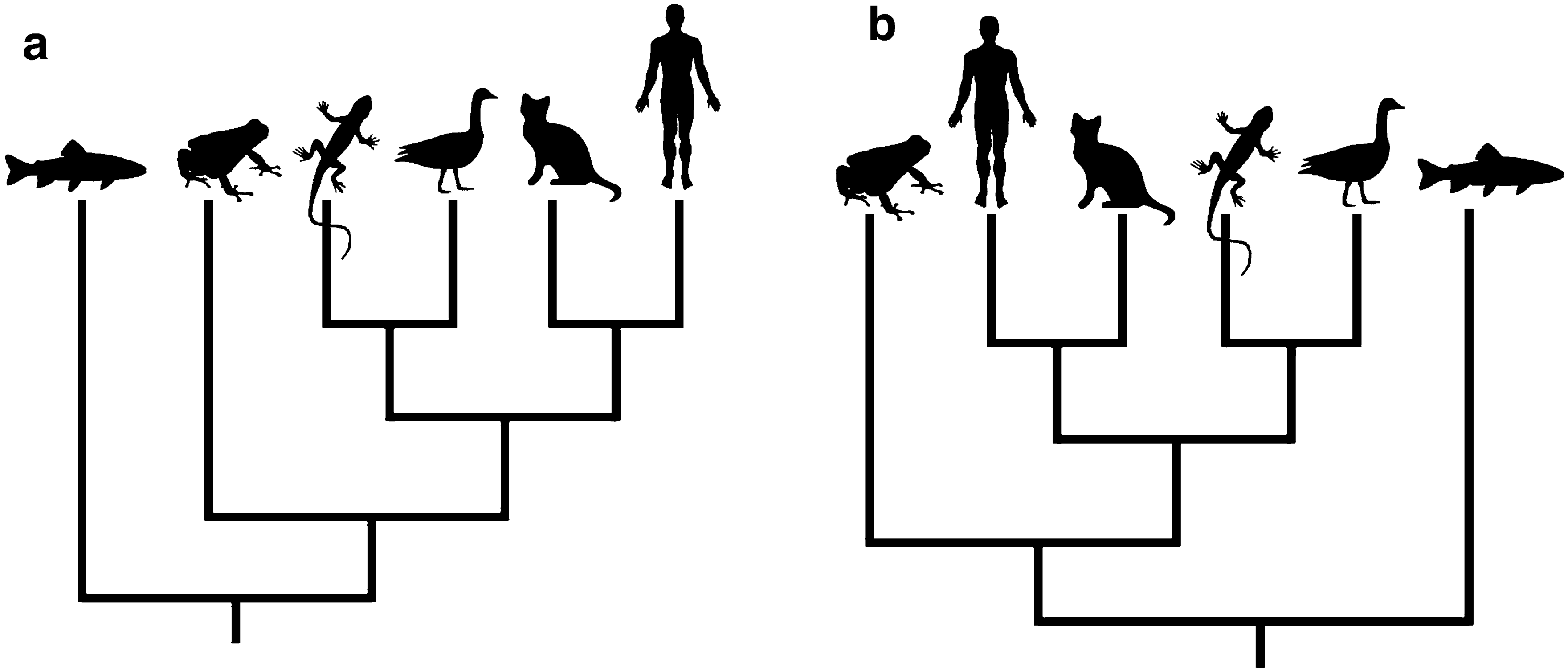
A = (B + D) = (B + C + E + G) = (B + C + F + H)

D = (C + E + G) = (C + F + H)



T. Ryan Gregory. Understanding Evolutionary Trees. Evo Edu Outreach (2008) 1:121–137
DOI 10.1007/s12052-008-0035-x

How to read phylogenetic trees



T. Ryan Gregory. Understanding Evolutionary Trees. Evo Edu Outreach (2008) 1:121–137
DOI 10.1007/s12052-008-0035-x

What is phylogeny used for

- Classify taxonomy – The classic use
- Outbreak detection – Increasing with WGS data

What is phylogeny used for

- Cholera outbreak in Haiti 2010
- Listeria outbreak 2014

  Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014.

  Kvistholm Jensen et al. Clin Infect Dis. (2016) 63 (1): 64-70. doi: 10.1093/cid/ciw192

Case story

- *Vibrio Cholerae* outbreak in Haiti followed the 2010 earthquake
- Rumors 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

Population Genetics of Vibrio cholerae from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. Hendriksen et al. 23 August 2011 mBio vol. 2 no. 4 e00157-11. doi: 10.1128/mBio.00157-11

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 analyze for genetic relatedness
  - Whole genome sequencing, SNP identification and phylogenetic analysis

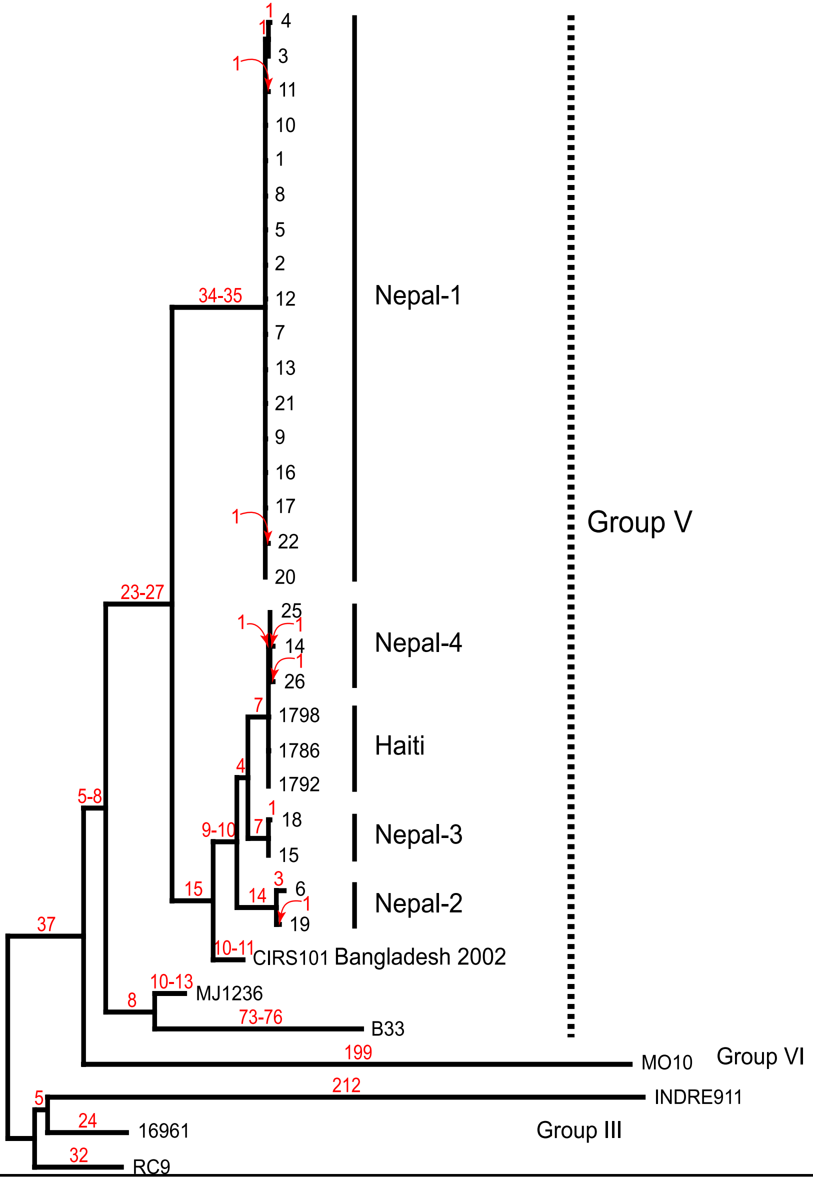
Case story - Results

Resistance profile	Susceptible	Decreased susceptibility	Resistant
Nepalese strains Hendriksen *et al. 2011*	Tetracycline	Ciprofloxacin	Trimethoprim, Sulfamethoxazole Nalidixic
Haitian outbreak strains Centers for Disease Control and Prevention, 2010	Tetracycline	Ciprofloxacin	Trimethoprim, Sulfamethoxazole Nalidixic

Case story - Results

- Pulsed-field gel electrophoresis (PFG)E
  - Nepalese isolates divided in 4 groups
  - Most common Haitian type in same group as four Nepalese strains

Case story - Results



10 minutes break!