What Is Phylogenetic/Phylogenomic Analysis?

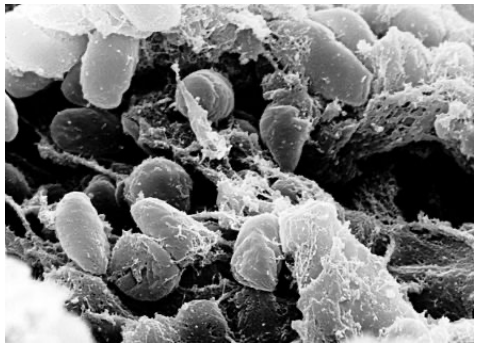
Phylogeny = Evolutionary history

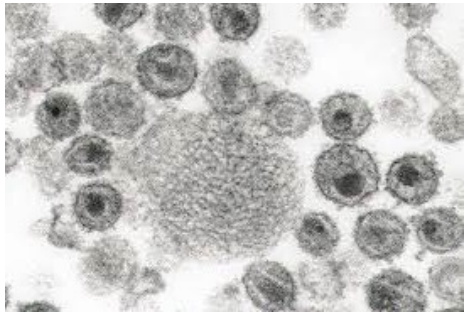
Reconstruction of evolutionary history using shared traits among organisms.

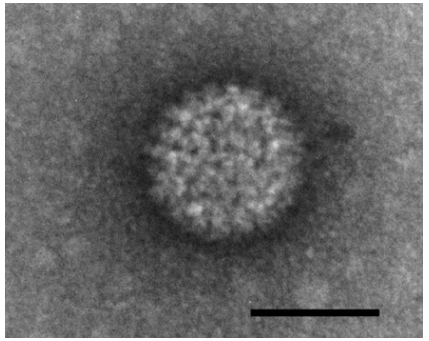
- Tree of life *http://www.tolweb.org/tree/*
- Time Tree *http://www.timetree.org*

DNA Sequencing of Pathogens

- Genome organization, structure and function
- Estimation of pattern and rate of mutation within each pathogen gene
- Estimation of genetic diversity within the pathogen
- Identification of emergent strain
- Geographic and evolutionary origin of emergent strain(s)
- Patterns of global transmission
- Co-evolution and adaptation
- Selection
- Fitness traits-replication rate, transmissibility, immune recognition
- Intra-host diversification and horizontal gene transfer among strains
- Vaccine and drug therapy development



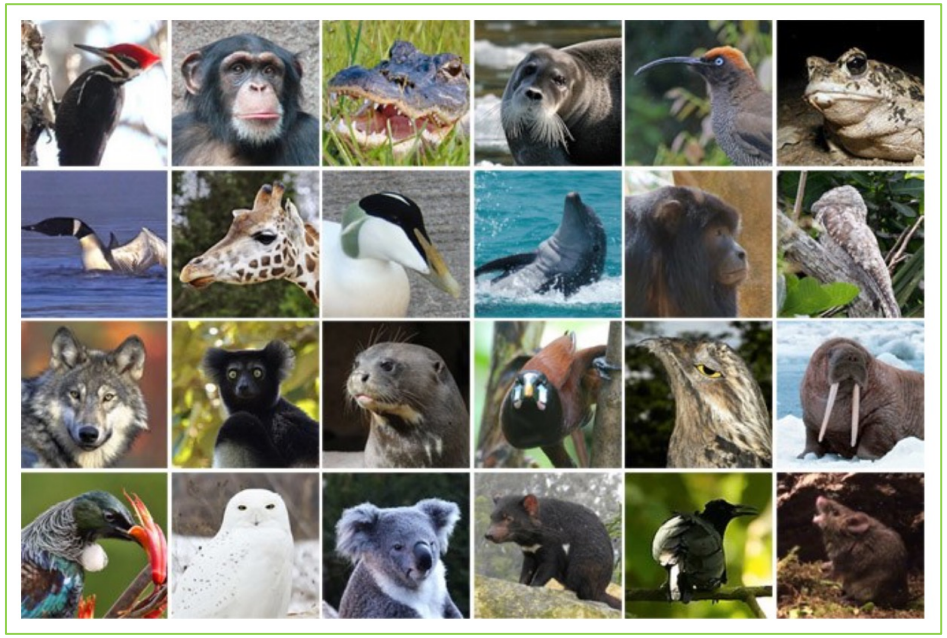




GDW
Genomics of Disease in Wildlife

DNA Sequencing of Host Species

- Genome structure, content, function, evolution
- Evolutionary history of species
- Population structure and phylogeography
- Historic and ongoing patterns of migration
- Genetic diversity
- Inbreeding
- Domestication
- Hybridization
- Endangered or relic species, subspecies and populations
- Identification of genes involved in disease resistance and progression
- Predictive effects of pathogen emergence in naïve host populations



GDW
Genomics of Disease in Wildlife

Phylogenetic Tree: Visualizations

- Phylogram
  - Tree branch length proportional to genetic divergence
- Cladogram
  - Topology only, branching order important
  - Branch lengths ignored
  - Character data
- Rooted
  - Outgroup
  - Midpoint rooting
- Unrooted

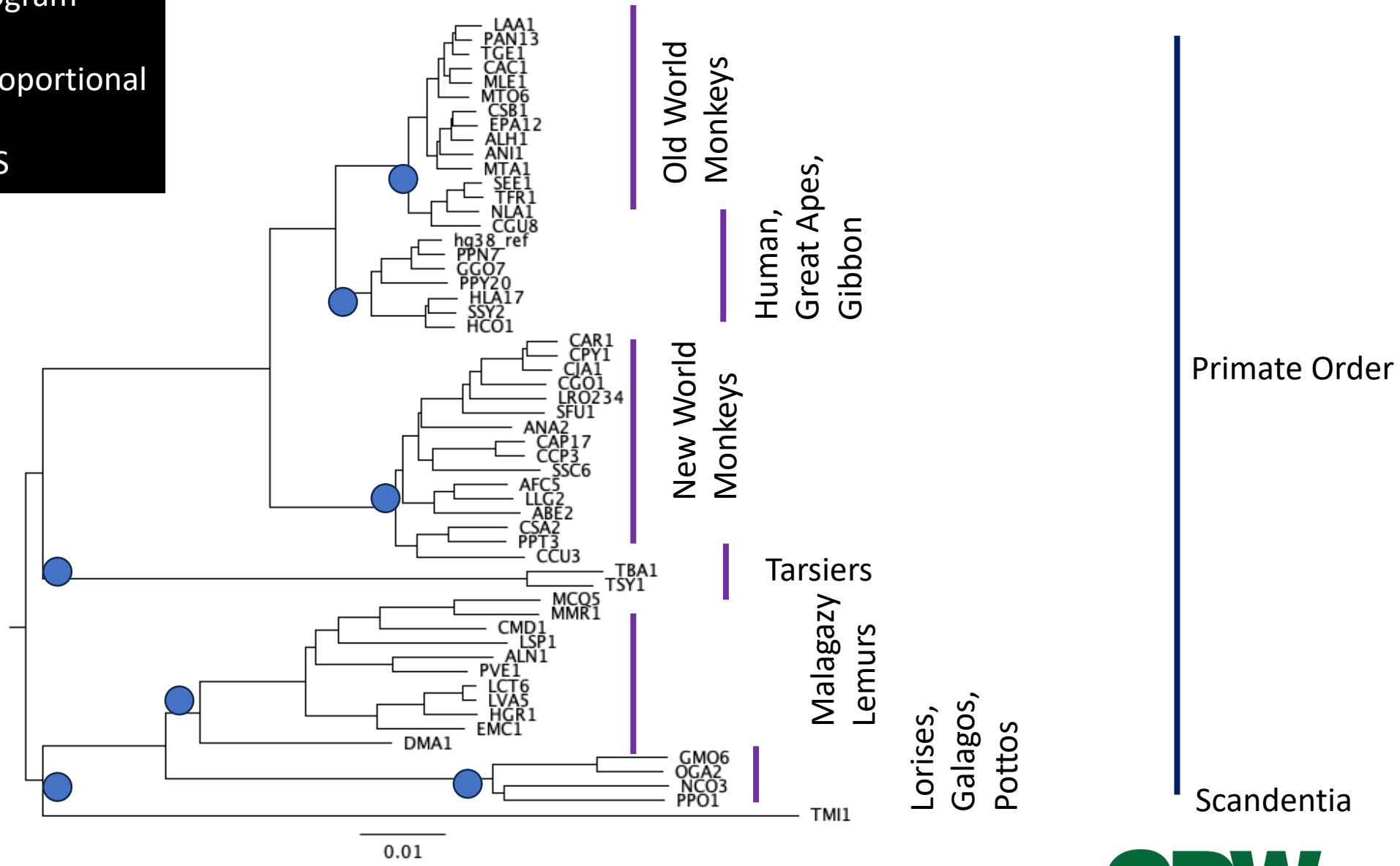
GDW
Genomics of Disease in Wildlife

Rectangular Phylogram

Branch lengths Proportional

All nodes 100% BS

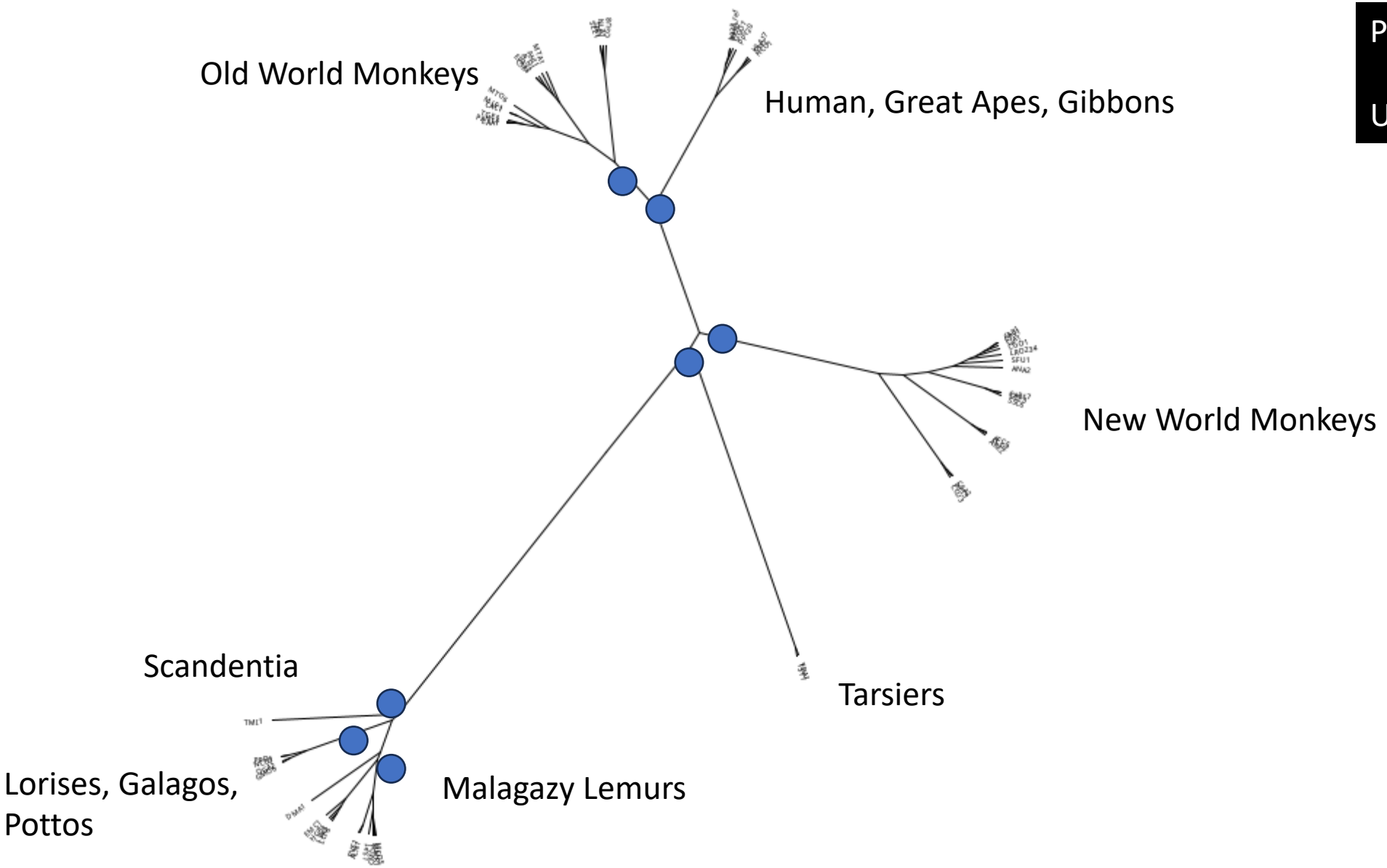
Primate Genera Tree Based on 10,000 genes



Genomics of Disease in Wildlife

Primate Order

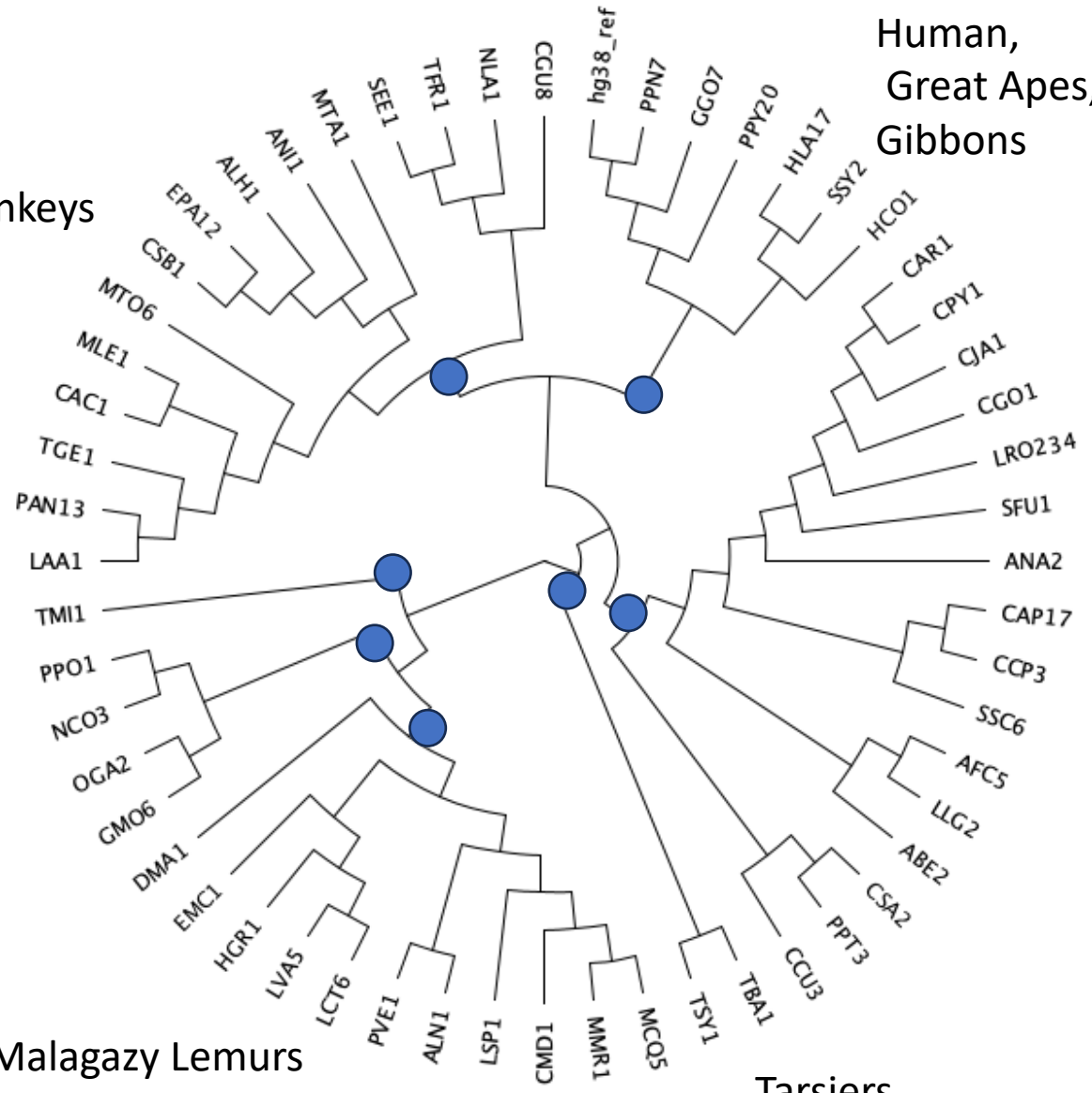
Unrooted Phylogram



Genomics of Disease in Wildlife

Circular Cladogram:

Branching order only
Branch lengths ignored



Human,
Great Apes,
Gibbons

Old World Monkeys

Scandentia

Lorises, Galagos, Pottos

New World Monkeys

Malagazy Lemurs

Tarsiers

Genomics of Disease in Wildlife

Elements for Setting Up the Analyses

Step 1: Determination of informative marker(s) for Hypothesis

Step 2: Determination of sample size and design

- Taxa or OTUs
- Outgroup-root

Step 3: Alignment (sequence data)

Step 4: Model of substitution

- Nucleotide data
  - Among site rate heterogeneity
  - Bias frequency nucleotides
  - Transition: Transversion
  - Codon analyses
- Protein Data
  - Probability matrix [DAYHOFF, DCMUT, JTT, MTREV, WAG, RTREV, CPREV, VT, BLOSUM62, MTMAM, LG] or empirical frequencies of amino acid within data set.

GDW
Genomics of Disease in Wildlife

Building and Assessing the Phylogeny

Step 5. Select optimality criteria and create phylogeny

- Maximum Likelihood (probability scores for observed variation)
- Minimum Evolution (distance-based)
- Maximum Parsimony (character states)
- Bayesian Inference (MCMC)

- Computer Programs
  - IQTree, PAUP, MEGA, RAxML, PhyML, MrBayes

Step 6. *A posteriori* evaluation

- Bootstrap
  - Assessment of tree reliability
  - Re-sampling with replacement
  - Node support
- Jackknife
  - Random sampling without replacement
  - Sites or taxa
- Bayesian Posterior Probability of Model Parameters

GDW
Genomics of Disease in Wildlife

Phylogenetic Data Input

Whole Genome Comparisons

- Comparison of syntenic regions (Conserved Sequence Blocks)

Protein Coding Data (Genes and Exomes)

- Homologs versus Paralogs
  - Gene duplication and Gene Loss
- Amino Acid versus Nucleotide
  - Synonymous substitutions
  - Nonsynonymous substitutions (missense mutations)

MtDNA

- Pattern of Inheritance Bias (Matrillineal)
- Single Evolutionary Unit despite different genes evolving at different rates

Autosomes, X & Y Chromosomes

- Y/X > 1 Male-driven evolution linked with spermatogenesis

Retro-elements

- SINEs, LINEs, Endogenous retroviruses

Microsatellites, SNPs, SNVs

GDW
Genomics of Disease in Wildlife

Data Selection: Hypothesis Driven

Informative: "Garbage In, Garbage Out"

Neutral Evolution: **Does it exist?**

First proposed by Kimura (1968): amino acid changes are selectively neutral driven by mutation rate and genetic drift.

Short answer : unlikely

Genomes and genes under strong selection

Maintain function (purifying selection)

Diversification to adapt or to gain function (positive selection)

Is this really a problem?

Quality of Data:

Robustness: Additional data strengthens overall topology

Accuracy: Data capture 'true' evolutionary and natural history

Precision: Data track similarly throughout phylogeny

Empirical Consensus: Combined Data and/or Multi-Gene Coalescence Methods

GDW
Genomics of Disease in Wildlife

Heuristic Searches To Find the Optimal Solution

Optimum:
Tree with the highest score possible under the different optimality criteria (ML, MP, ME, BAYESIAN)

Problem- Cannot test all possible combinations of taxa.
'Exhaustive search' -not recommended for more than 10 sequences in an alignment
e.g. an alignment of 20 sequences will require testing differences between 2×10^{20} potential trees

Solution: Heuristic searches designed as short cuts through all possible tree space:

Nearest neighbor interchange (NNI)
Subtree-Pruning-regrafting (SBR)
Tree bisection reconnection (TBR)

If suspect getting stuck local optimum:
rearrange input order, star decomposition, iterative step-wise addition.

GDW
Genomics of Disease in Wildlife

Phylogenetic Pitfalls

- Recombination
- Gene conversion
- Gene duplication
- Gene loss
- Selection
- Insufficient elapsed time since last shared common ancestor
- Ancestral polymorphisms
- Incomplete lineage sorting (discordant gene trees)
- Saturation of sites (multiple hits)
- Long branch attraction (inadequate taxon sampling)
- Inadequate models of amino acid or nucleotide substitution

GDW
Genomics of Disease in Wildlife

Lab: Phylogenomics with IQ-Tree: Ver 2.2.0

- **Accuracy**: Proposing novel computational methods that perform better than existing approaches.
- **Speed**: Allowing fast analysis on big data sets and utilizing high performance computing platforms.
- **Flexibility**: Facilitating the inclusion of new (phylogenomic) models and sequence data types.
- **Versatility**: Implementing a broad range of commonly-used maximum likelihood analyses.

IQ-TREE has been developed since 2011 and freely available at http://www.iqtree. org/ as open-source software

Bui Quang Minh, Rob Lanfear, Nhan Ly-Trong Jana Trifinopoulos, Dominik Schrempf, Heiko A. Schmidt March 25, 2022

GDW
Genomics of Disease in Wildlife

Tree-Based Inferences About Gene Function: An Real-World Example

- Hypothesis: What genes are evolving the fastest across the primate order?
  - All the same?
  - All the same among lineages?

- We came up with the top five list for nucleotides and amino acid data
  - Any guesses to what we saw?

Genomics of Disease in Wildlife

The aye-aye

Daubentonia madagascarensis

Genomics of Disease in Wildlife

Nucleotide Substitution Models

- Range from simple to complex
- Models are tested hierarchically, adding more complex elements
- Performs a likelihood test between successive models
- Selects a model based on AIC or BIC criteria

- **Lab:** Download IQRuns from GDW2023 Github to Desktop
- We'll test for a Model
- We'll find a Maximum Likelihood Tree
- We'll do a bootstrap

Genomics of Disease in Wildlife