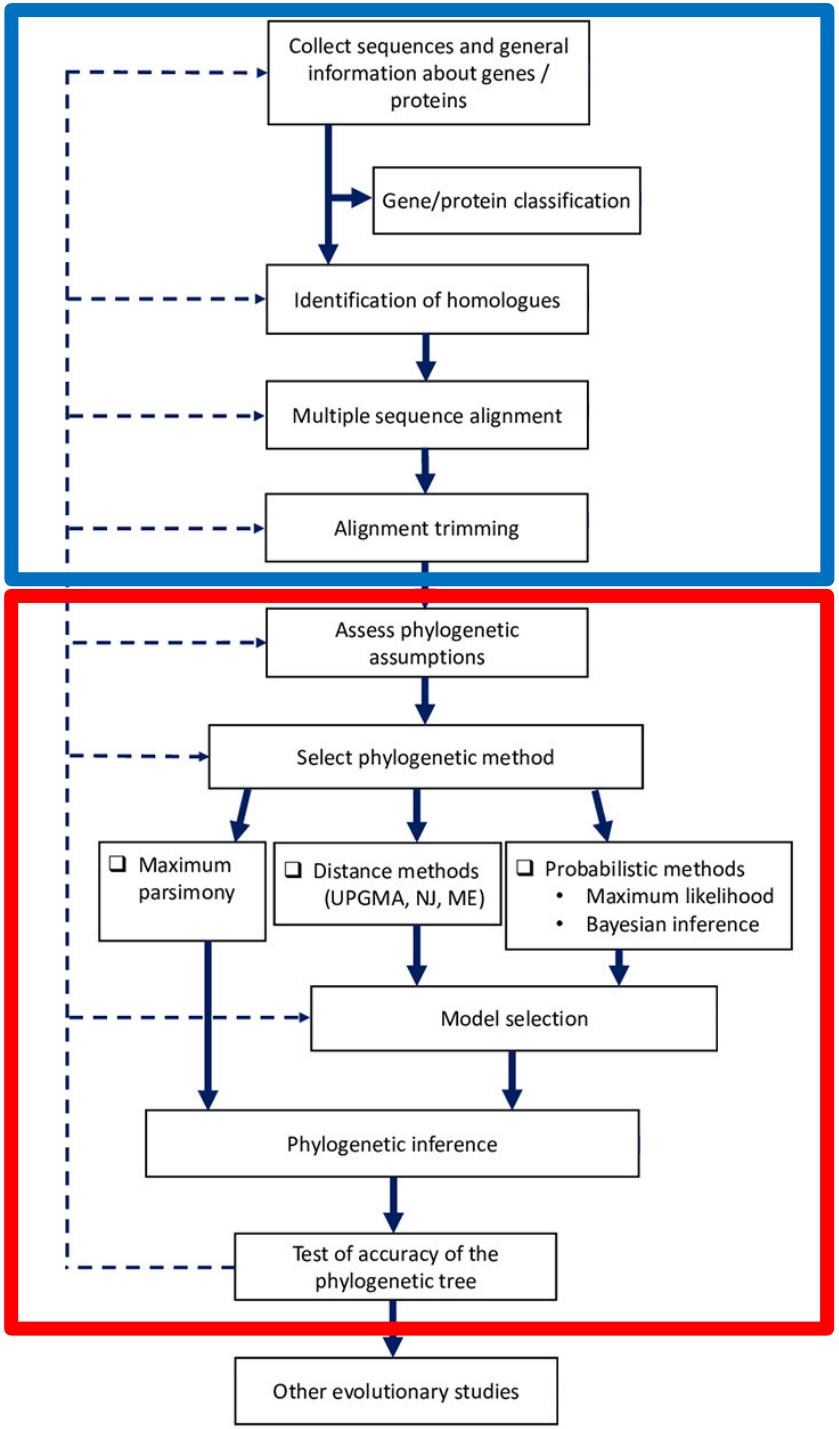
Lecture 1.3

# Phylogenetic Data




## 1. Data preparation

- Taxon and gene sampling
- Sequence alignment (if needed)
- Data filtering

## 2. Phylogenetic inference

- Model selection
- Estimation of tree
- Further analysis and interpretation

Jacques *et al.* (2023) *PLOS ONE*

# Phylogenetic data

- **Select data to optimise signal:noise**
  - Slowly evolving markers for deep evolutionary events
  - Rapidly evolving markers for recent evolutionary events
- **Minimise violations of phylogenetic assumptions**
  - Some data sets can yield biased phylogenetic estimates
- **Take advantage of existing resources**

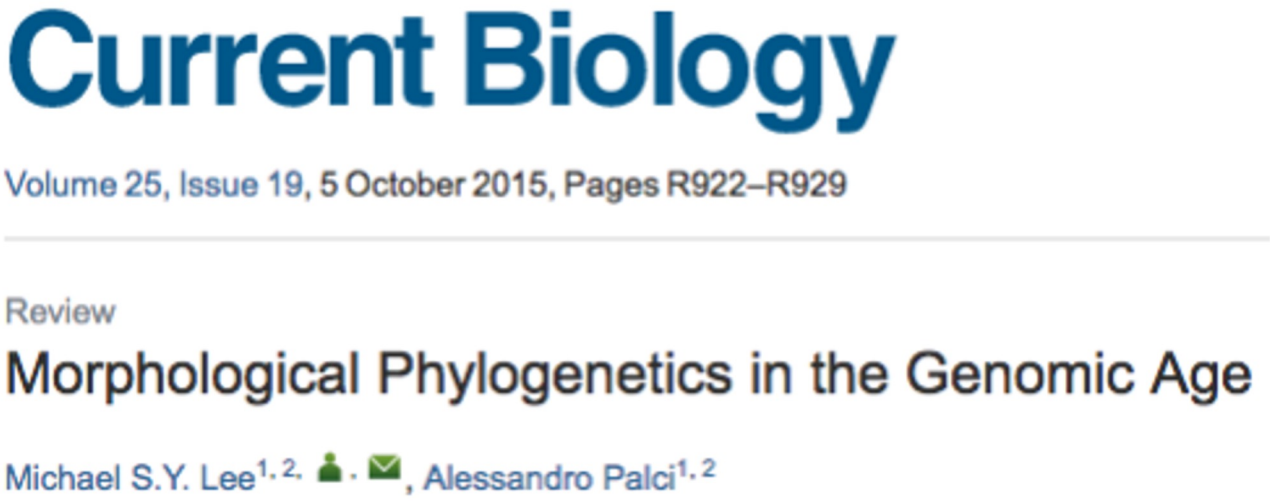


# Data types

- **Sequence data**
  - Nucleotides
  - Amino acids
- **Binary data** (presence/absence of genomic features)
- **Microsatellites** (repeat numbers)
- **Single-nucleotide polymorphisms** (SNPs)
- **Reduced-representation sequences**

# Morphological data

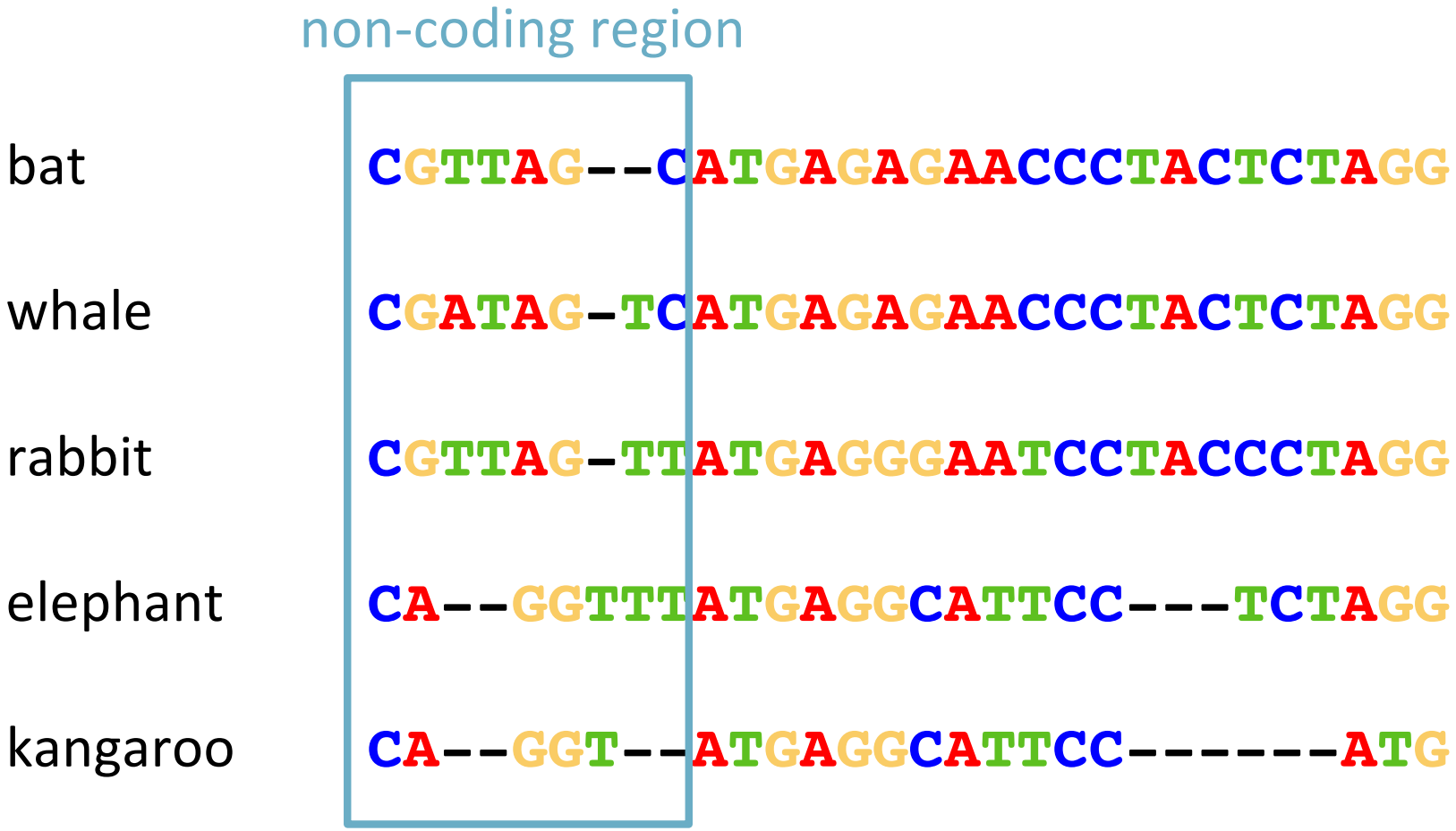
- Morphological characters from extant and extinct taxa



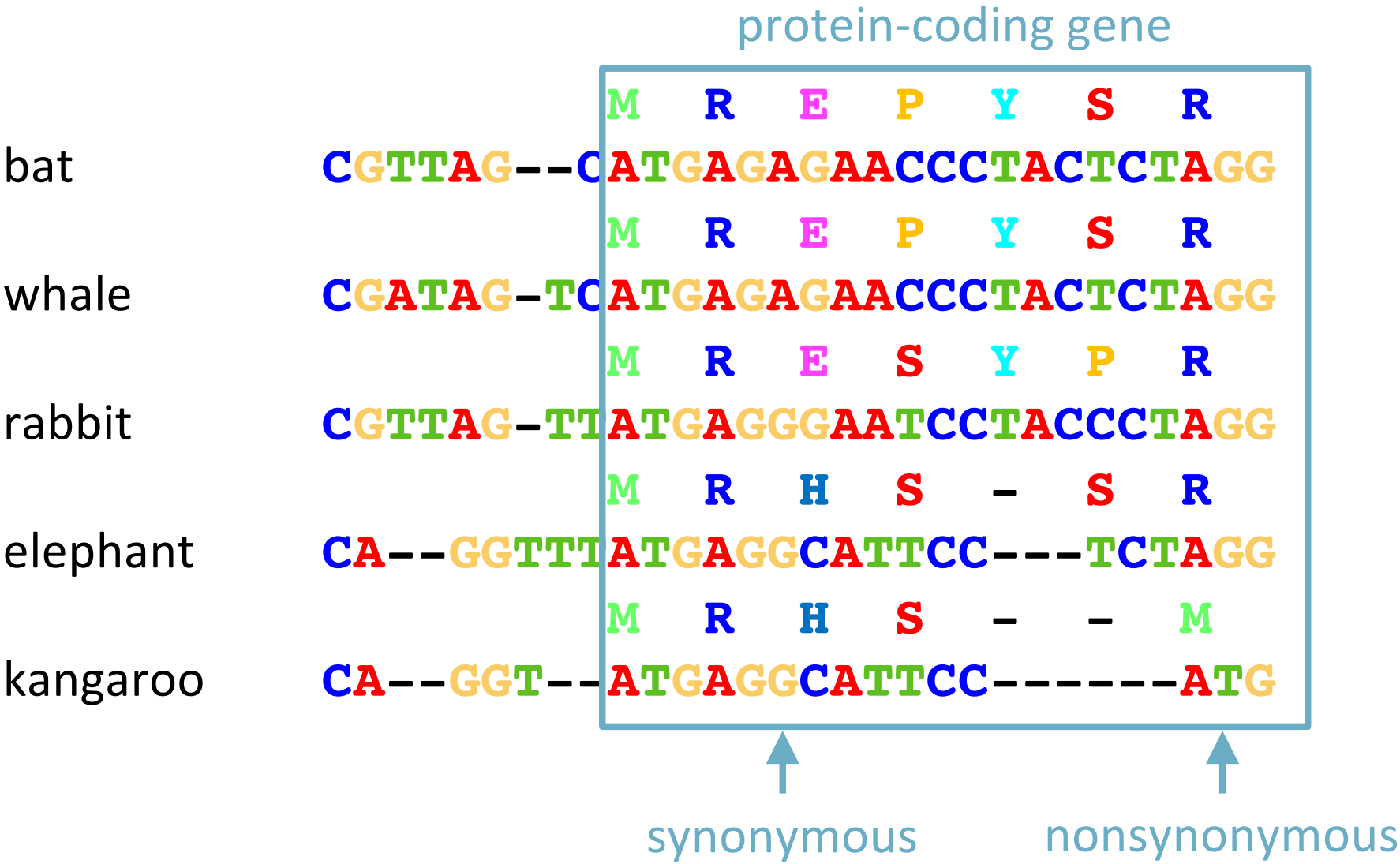
# Sequence data

- **Coding sequences**
  - Ribosomal RNA
  - Protein-coding genes
- **Non-coding sequences**
  - Intergenic sites
  - Introns
- **Amino acid sequences**

# Sequence data

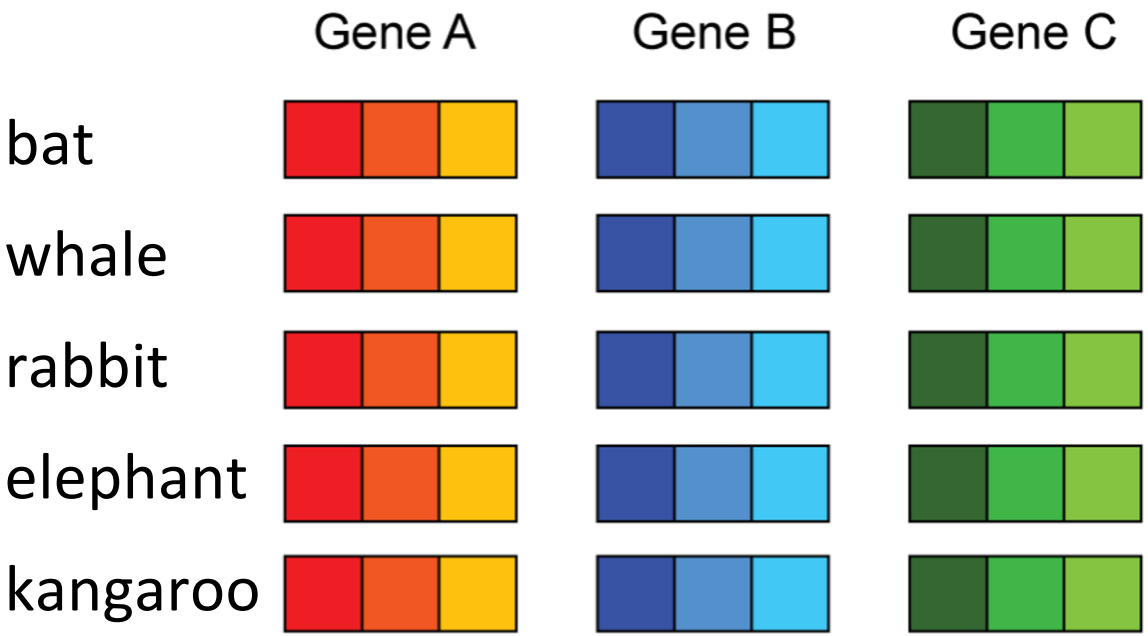


# Sequence data



# Data partitioning

- Sites evolve at different rates
- Separate substitution model for each gene and codon position?




- **Biological**
  - Genome
  - Genes
  - Codon positions
  - RNA stems *vs* loops
  - Hydrophobic *vs* hydrophilic
- **Statistical**

# PartitionFinder

- Too many possible partitioning schemes
  - 15 schemes for 4 genes
  - 52 schemes for 5 genes
  - 203 schemes for 6 genes

**PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses** □

Robert Lanfear □, Paul B. Frandsen, April M. Wright, Tereza Senfeld, Brett Calcott

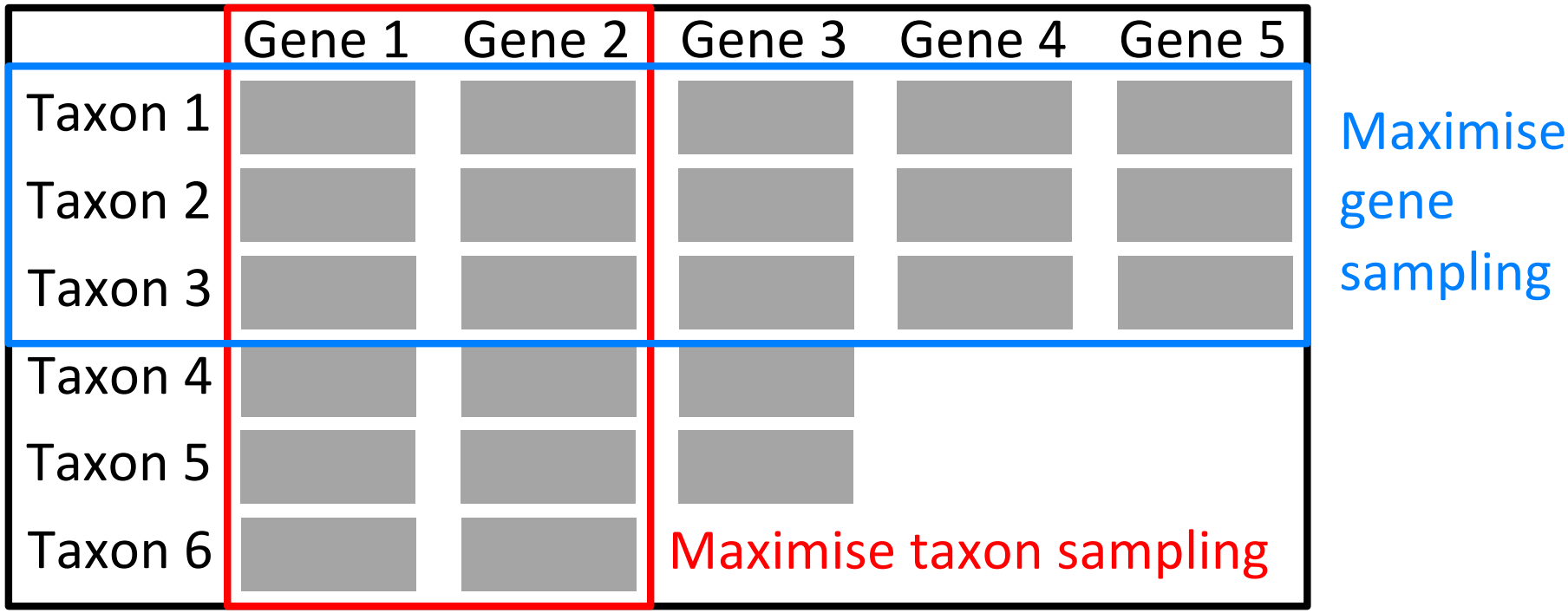
*Molecular Biology and Evolution*, Volume 34, Issue 3, March 2017, Pages 772–773,

# Gaps and missing data

- **Delete sites with any missing data**
  - Potential loss of informative data
  - Problematic in analyses of data supermatrices
- **Treat gaps as unresolved data**
  - Gap is simultaneously A, C, G, and T
  - Most common approach
- **Code gaps as binary characters**

# Gaps and missing data

- Impact of missing data remains poorly understood
- Filter data according to chosen threshold of missing data



# Gaps and missing data

- Impact of missing data remains poorly understood
- Filter data according to chosen threshold of missing data

**Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci**

Nicolás Mongiardino Koch

*Molecular Biology and Evolution*, Volume 38, Issue 9, September 2021, Pages 4025–4038,

# Mutational saturation

- Some sites can evolve very rapidly
  - 3rd codon positions
  - Loop regions in RNA
- Multiple hits can erode phylogenetic signal

**Excluding Loci With Substitution Saturation Improves Inferences From Phylogenomic Data**

David A Duchêne ✉, Niklas Mather, Cara Van Der Wal, Simon Y W Ho

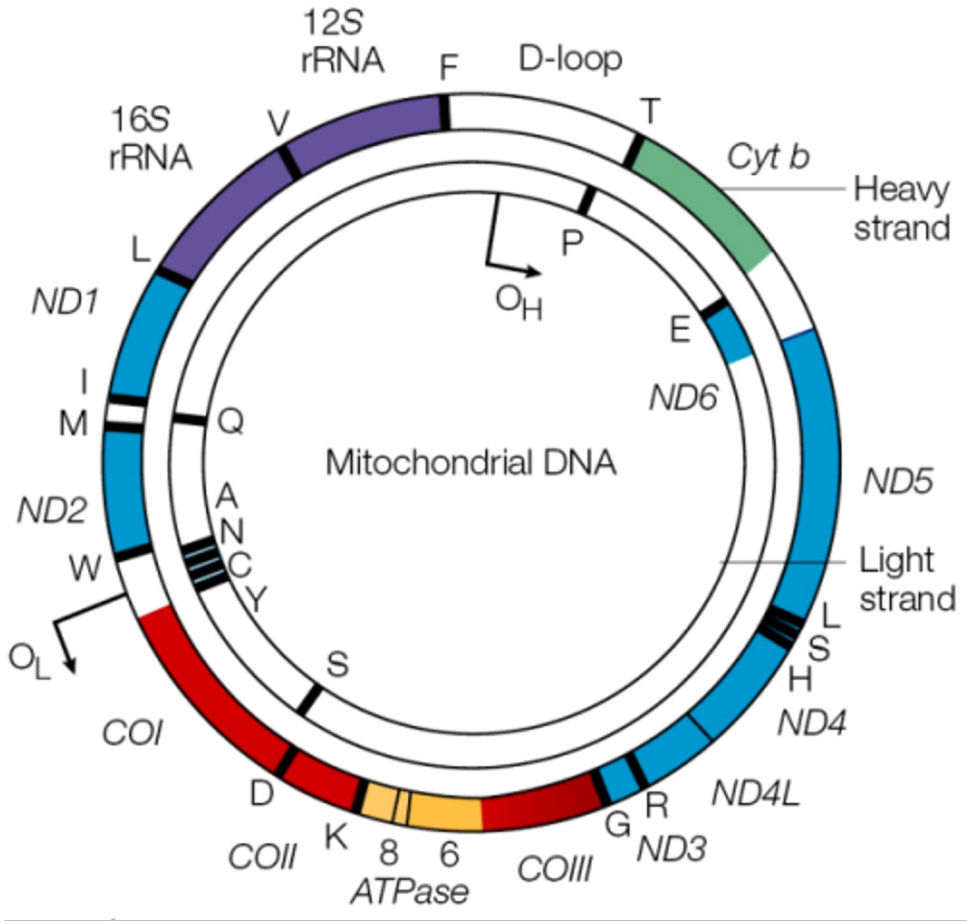
*Systematic Biology*, Volume 71, Issue 3, May 2022, Pages 676–689,

Saturated sites can be removed to improve signal:noise

# High-Throughput Data

# Mitochondrial genomes

- Maternally inherited
- Protein-coding genes (e.g., *COI*)
- RNA genes (e.g., *12S*, *16S*)
- Control region



# Single-nucleotide polymorphisms

- Single sites sampled from throughout the genome
- More common in intraspecific (population) studies

- Issues to consider:
  - **Recombination**
    SNPs are usually unlinked so they are likely to have different (gene) trees
  - **Ascertainment bias**
    SNPs are selected for variability and this can mislead estimates of population sizes, rates, and other parameters

# Reduced-representation sequences

- Markers identified by cutting genome with restriction enzymes
- Process creates binary data and short sequences
- Examples include RADseq and DArTseq


- Issues to consider:
  - **Recombination**
    Markers are usually unlinked so they are likely to have different (gene) trees
  - **Missing data**
    Typically a large proportion of missing data

Leaché *et al.* (2015) *Syst Biol*

# Transcriptomes and exon capture

- Large panels of protein-coding loci
- Sequences are easier to align
- Good for inferring deep relationships

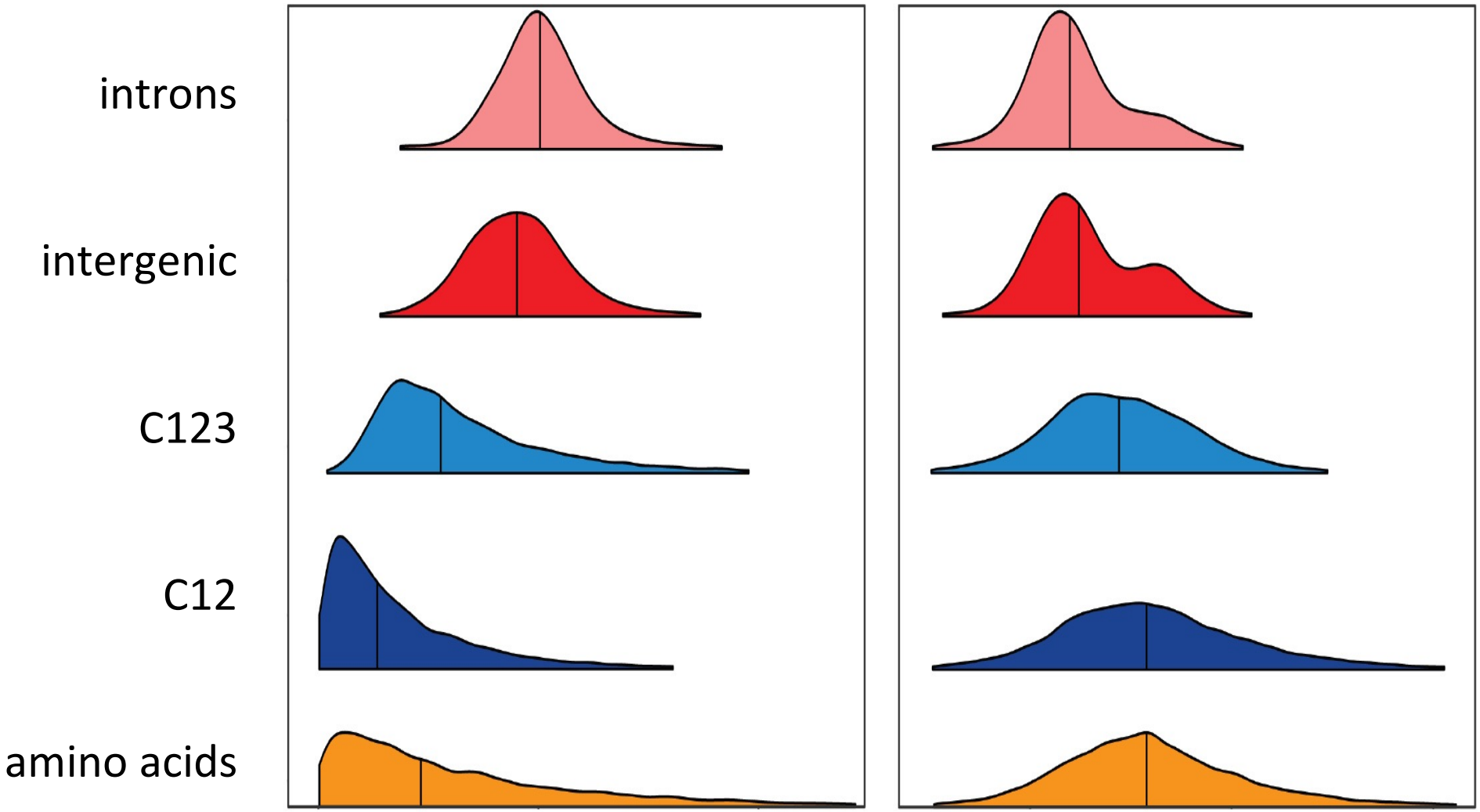
- Issues to consider:
  - **Variability**
    Might not be much variation at the population level
  - **Selection**
    Differences in selection will lead to rate differences across exons

# Whole-genome sequencing

- Choice of data types
- Need to reduce computational burden

- **Issues to consider**
  - Single-copy genes
  - Selectively neutral
  - Unlinked loci

# Bird genomes



# Useful references