Lecture 1.3

Phylogenetic Data

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

Phylogenetic data

- **Select data to optimise signal:noise**
  - Slowly evolving markers for deep evolutionary events
  - Rapidly evolving markers for recent evolutionary events
- **Homoplasy**
  - Taxa share similarities that do not reflect evolutionary history
- **Take advantage of existing resources**

NCBI

DRYAD

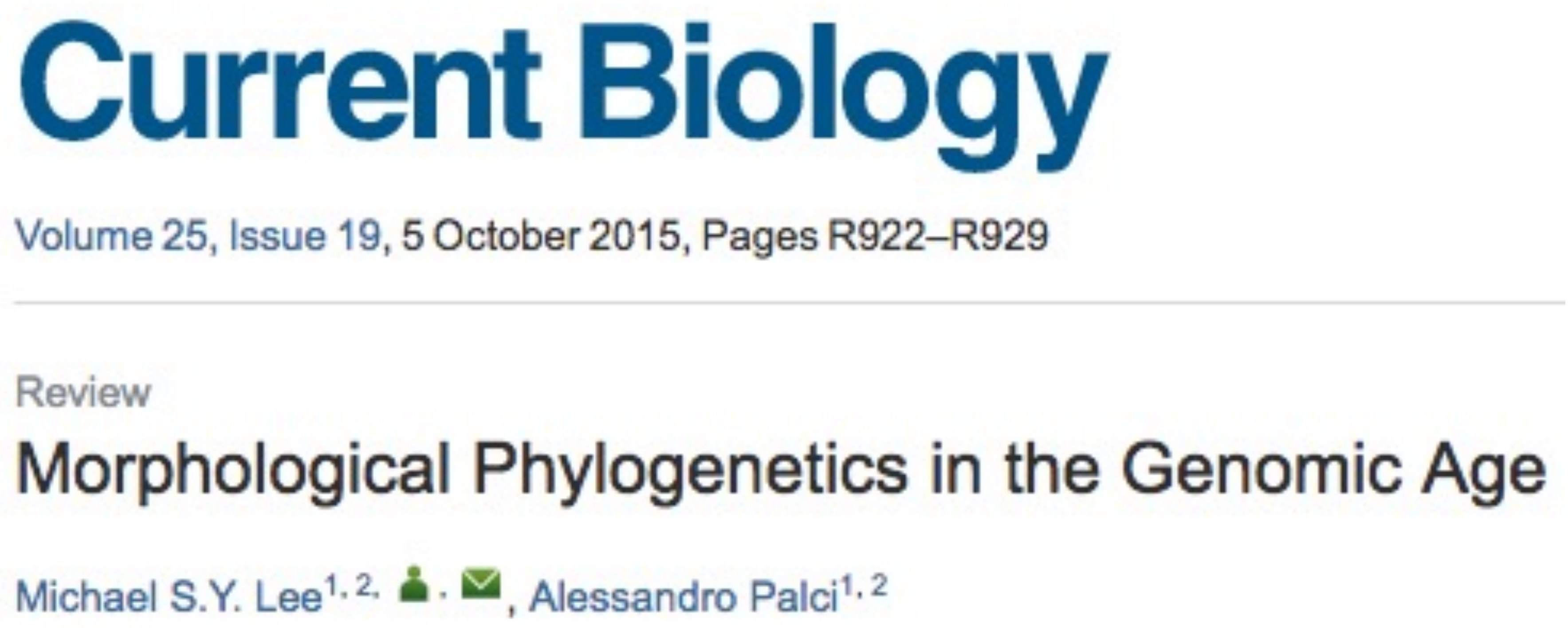
TreeBASE

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



Current Biology

Volume 25, Issue 19, 5 October 2015, Pages R922–R929

Review

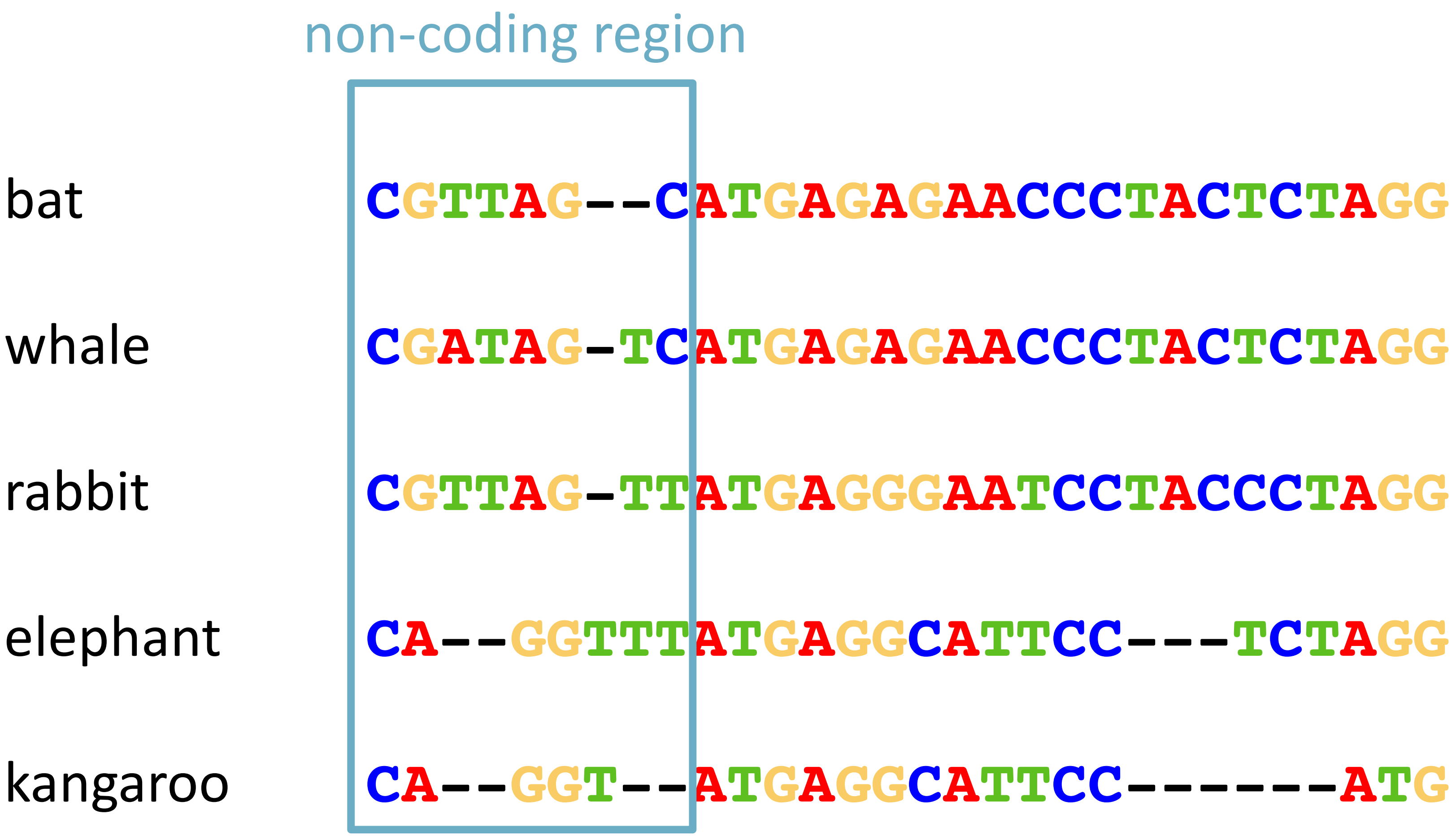
Morphological Phylogenetics in the Genomic Age

Michael S.Y. Lee[1, 2,] 👤 . ✉ , Alessandro Palci[1, 2]

Sequence data

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

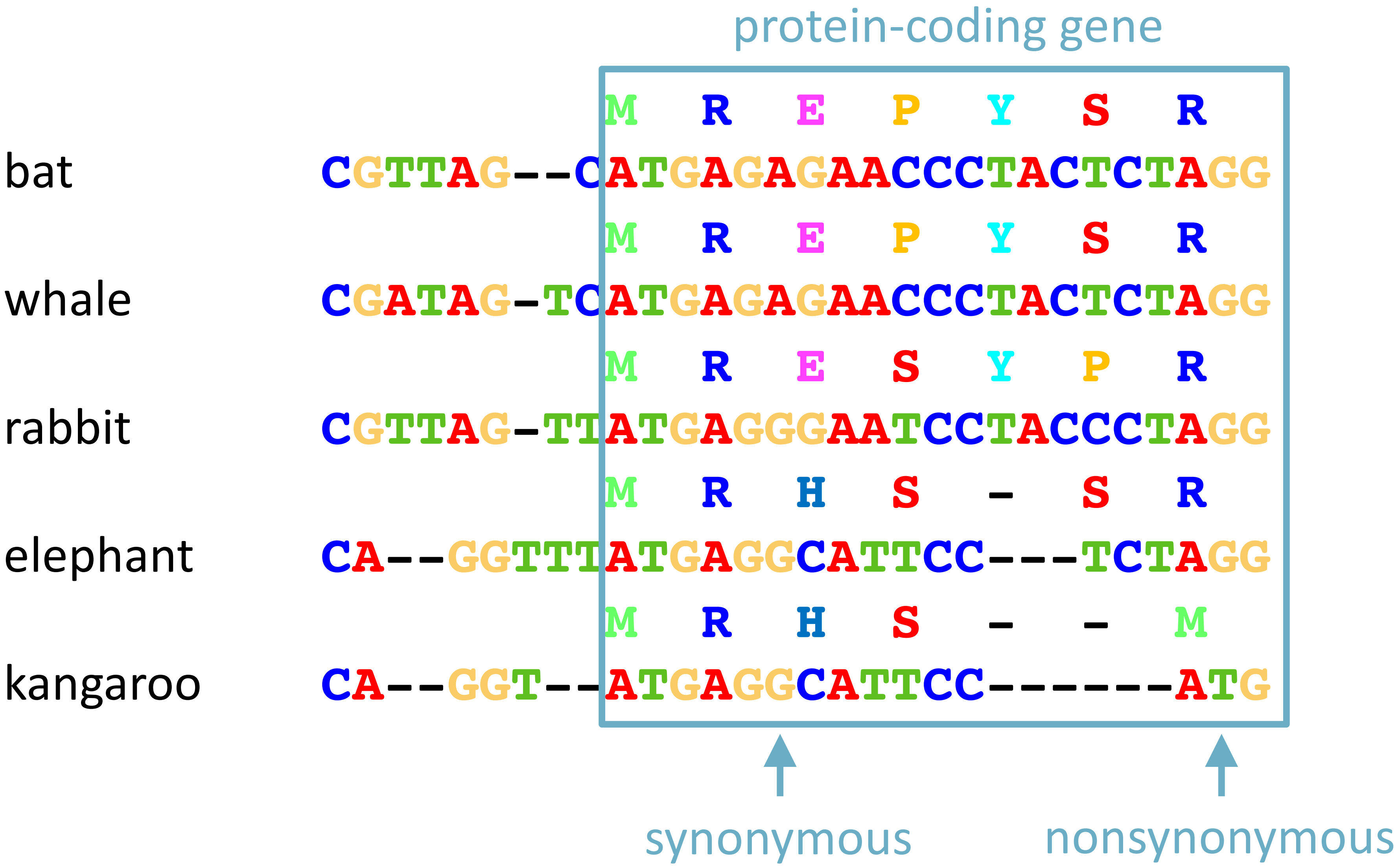
Sequence data



non-coding region

```
bat         CGTTAG--CATGAGAGAACCCTACTCTAGG
whale       CGATAG-TCATGAGAGAACCCTACTCTAGG
rabbit      CGTTAG-TTATGAGGGAATCCTACCCTAGG
elephant    CA--GGTTTATGAGGCATTCC---TCTAGG
kangaroo    CA--GGT--ATGAGGCATTCC------ATG
```

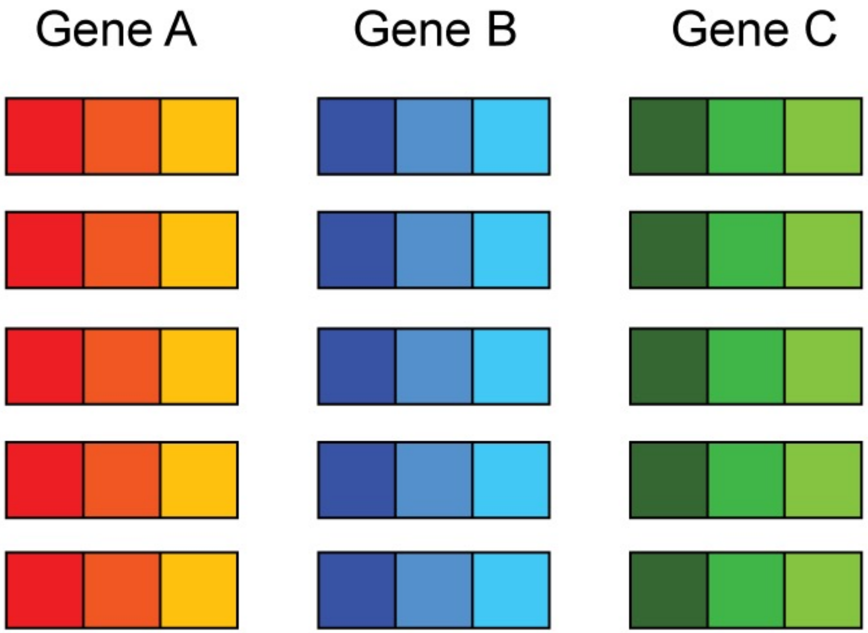
Sequence data



```
                  M  R  E  P  Y  S  R
bat       CGTTAG--CATGAGAGAACCCTACTCTAGG
                  M  R  E  P  Y  S  R
whale     CGATAG-TCATGAGAGAACCCTACTCTAGG
                  M  R  E  S  Y  P  R
rabbit    CGTTAG-TTATGAGGGAATCCTACCCTAGG
                  M  R  H  S  -  S  R
elephant  CA--GGTTTATGAGGCATTCC---TCTAGG
                  M  R  H  S  -  -  M
kangaroo  CA--GGT--ATGAGGCATTCC------ATG
```

Data partitioning

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



bat

whale

rabbit

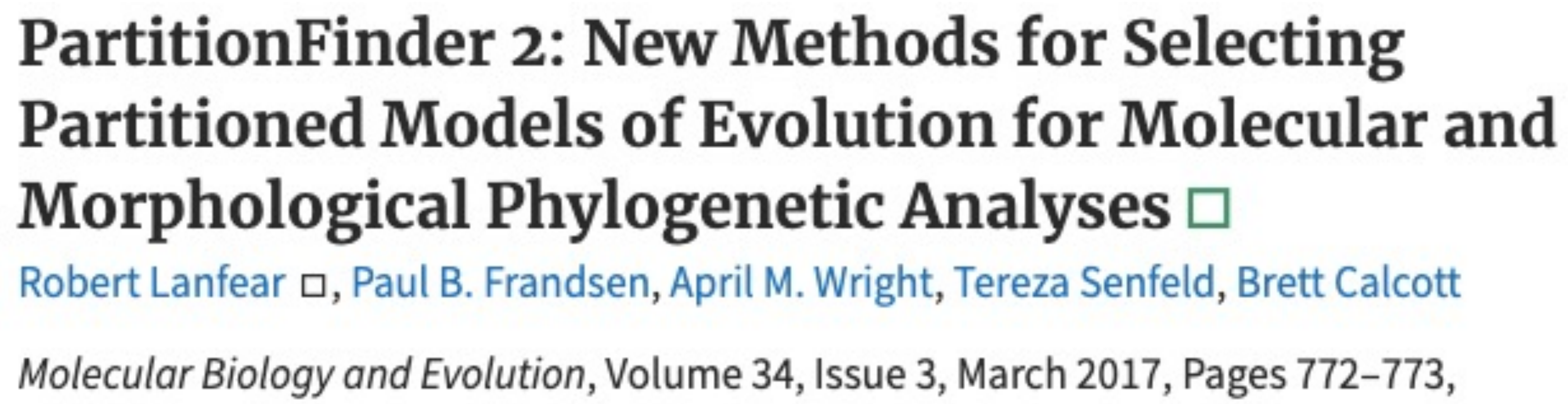
elephant

kangaroo

- **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

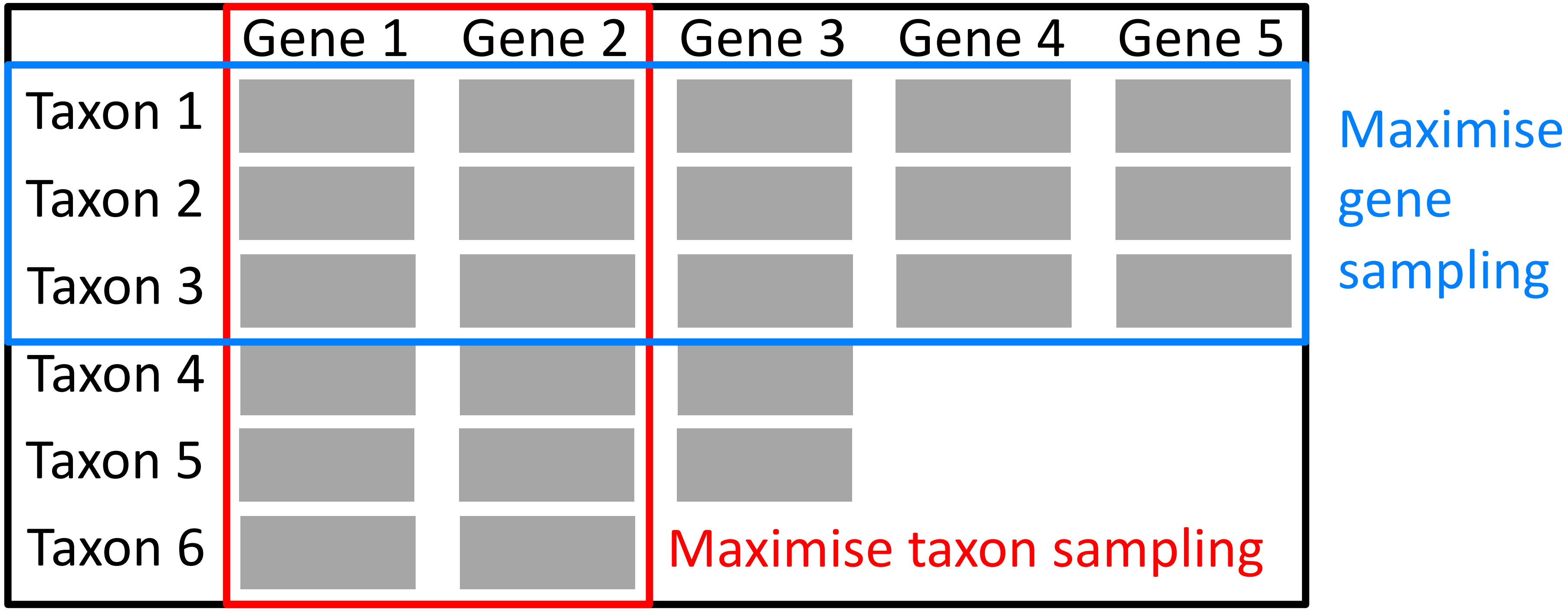
Molecular Biology and Evolution, Volume 34, Issue 3, March 2017, Pages 772–773,

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



	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Taxon 1	▨	▨	▨	▨	▨
Taxon 2	▨	▨	▨	▨	▨
Taxon 3	▨	▨	▨	▨	▨
Taxon 4	▨	▨	▨		
Taxon 5	▨	▨	▨		
Taxon 6	▨	▨			

Maximise gene sampling

Maximise taxon sampling

Mutational saturation

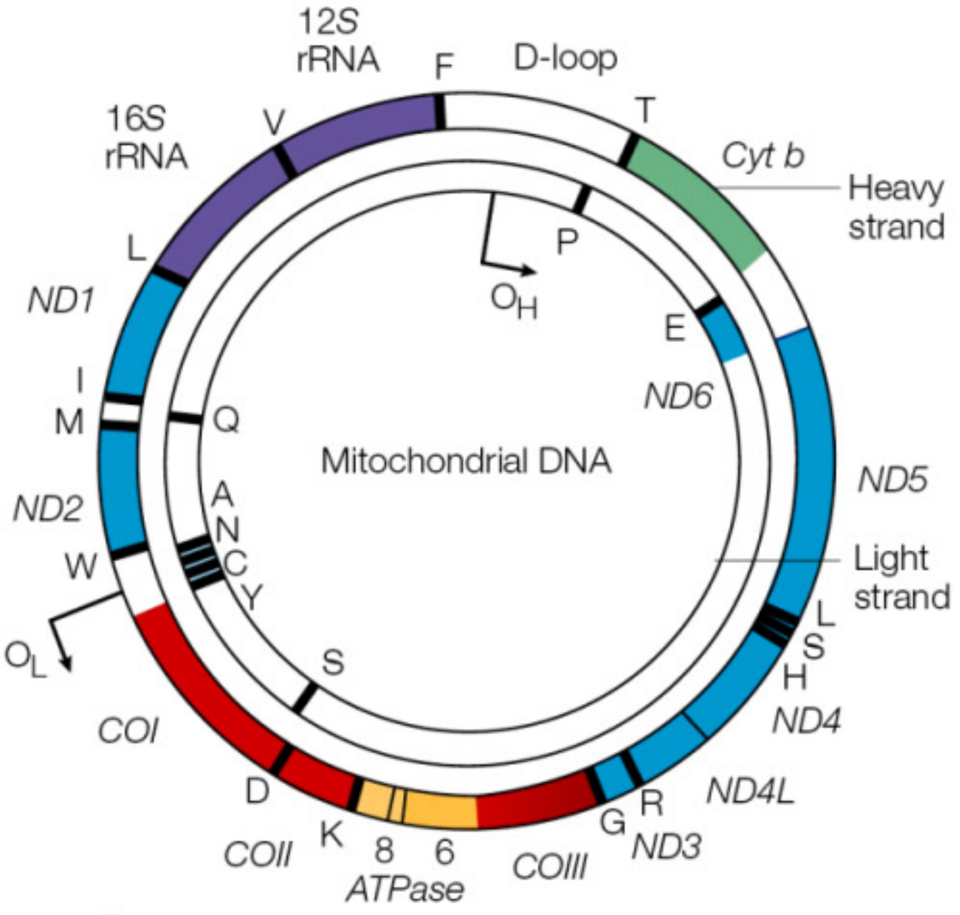
- Some sites can evolve very rapidly
  - 3rd codon positions
  - Loop regions in RNA
- Multiple hits can erode phylogenetic signal
- Various ways of testing for saturation
  (e.g., Xia’s test in DAMBE, PhyloMAd)

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

Diversity
Arrays
Technology
IT'S A BULLSEYE WITH DArT

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

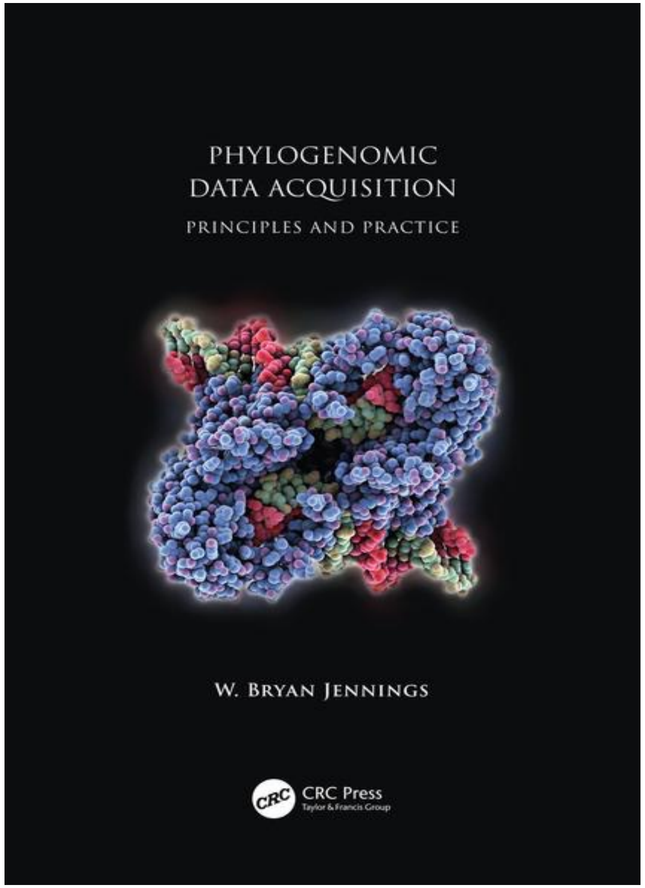
- Issues to consider:
  - **Variability**
    Might not be much variation at the population level
  - **Selection**
    Differences in selection will lead to rate differences across exons

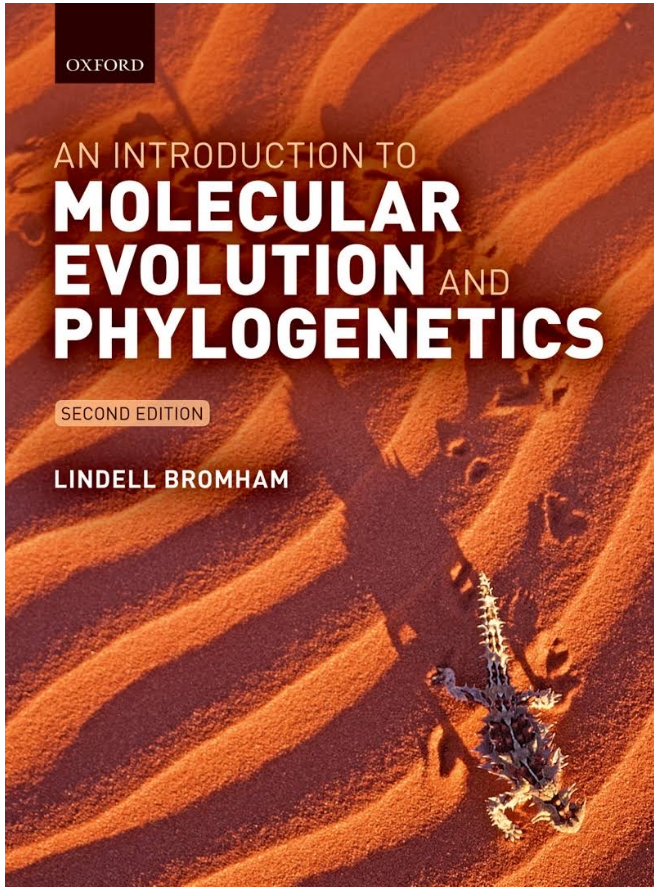
Whole-genome sequencing

- Typically NOT (yet) the entire genome
- Many challenges: Jarvis et al *Science* 2014 >400 years of computing using a single processor
- **Issues to consider**
  - Single-copy genes
  - Selectively neutral
  - Unlinked loci

Does DNA act as a telephone line? p. 1314
Mutations enhancing leukemia development pp. 1290 & 1373
A stable gold support gives sharper resolution p. 1377
Science
510
12 DECEMBER 2014
sciencemag.org
AAAS
SPECIAL ISSUE
Avian genomes
Sequencing across the bird species tree
p. 1308

Useful references


PHYLOGENOMIC
DATA ACQUISITION
PRINCIPLES AND PRACTICE
W. BRYAN JENNINGS
CRC Press
Taylor & Francis Group


OXFORD
AN INTRODUCTION TO
MOLECULAR
EVOLUTION AND
PHYLOGENETICS
SECOND EDITION
LINDELL BROMHAM