
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**



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**

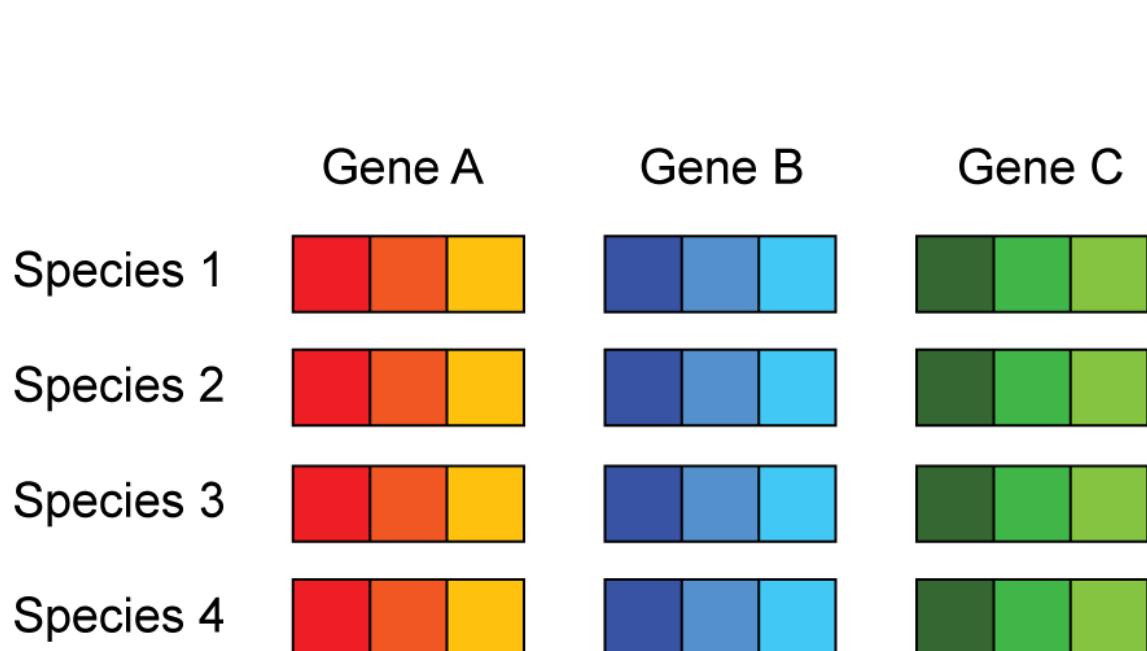


Sequence data

		protein-coding gene							
		M	R	E	P	Y	S	R	
brown bear	CGTTAG--CATGAGGGAACCTACTCTAGG	M	R	E	P	Y	S	R	
cave bear	CGATAG-TCATGAGGGAACCTACTCTAGG	M	R	E	S	Y	P	R	
black bear	CGTTAG-TTATGAGGGAATCCTACCCCTAGG	M	R	H	S	-	S	R	
giant panda	CA--GGTTIATGAGGCATTCC---TCTAGG								

Data partitioning

- 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: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses

Robert Lanfear,^{*,1} Brett Calcott,^{1,2} Simon Y. W. Ho,³ and Stephane Guindon⁴

2012 – *Molecular Biology and Evolution*, 29: 1695–1701.

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
- **Treat gaps as a 5th (nucleotide) or 21st (amino acid) state**
 - Not appropriate when there are long gaps
- **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	grey	grey	grey	grey	grey
Taxon 2	grey	grey	grey	grey	grey
Taxon 3	grey	grey	grey	grey	grey
Taxon 4	grey	grey	grey		
Taxon 5	grey	grey	grey		
Taxon 6	grey	grey			

Maximise gene sampling

Maximise taxon sampling

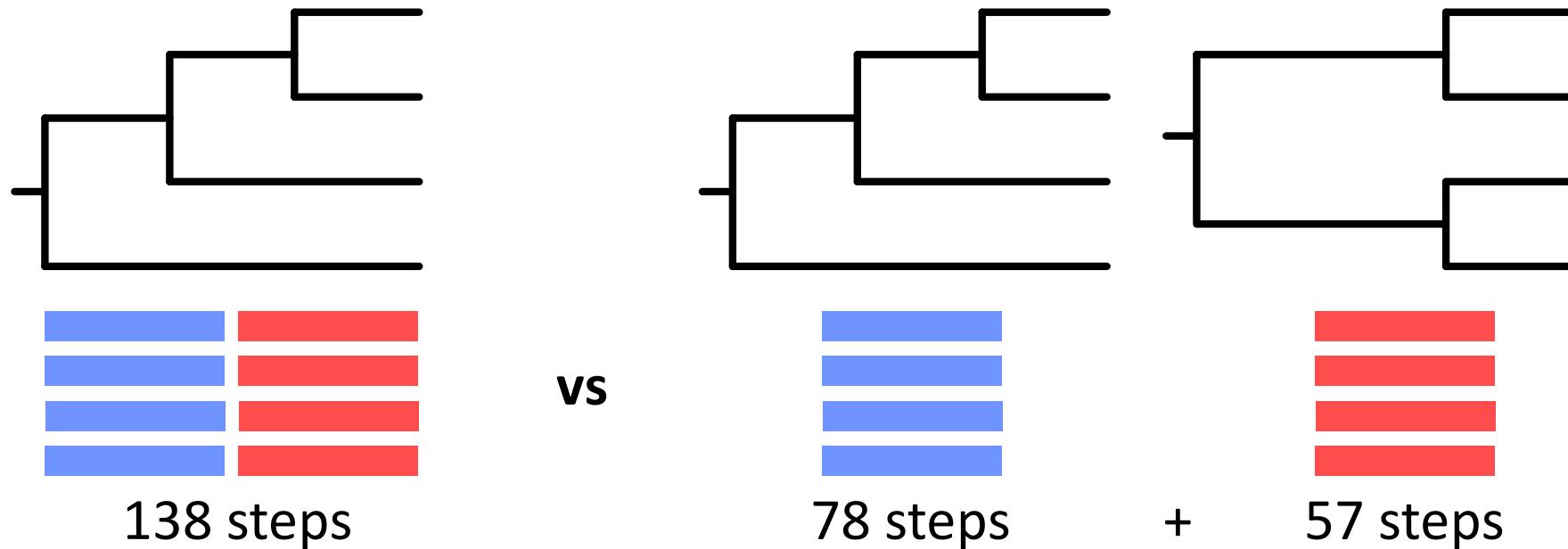
Mutational saturation

- 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)

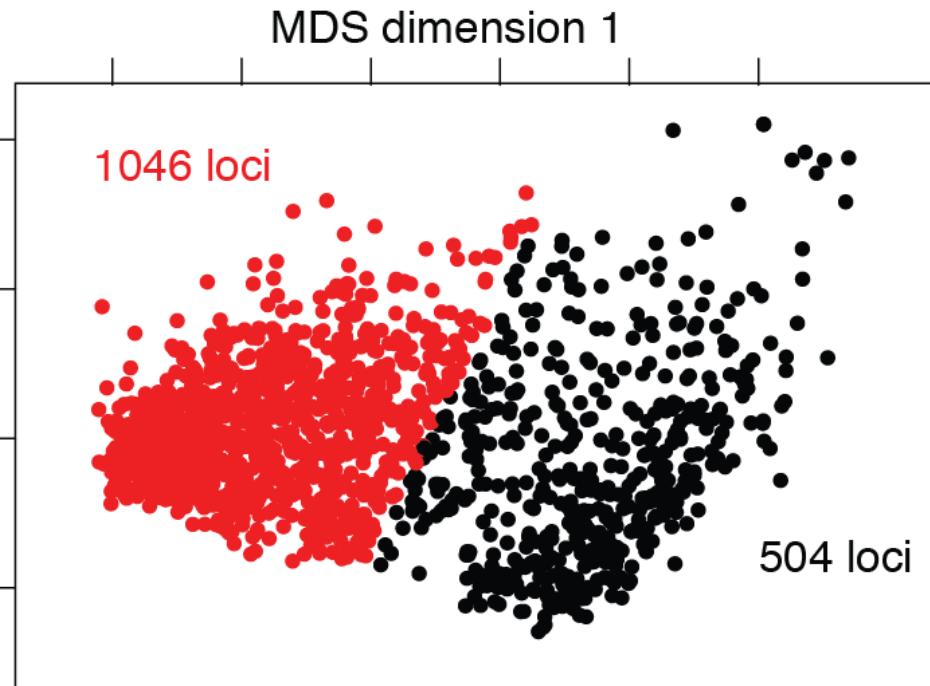
Saturated sites can be removed to improve signal:noise

Incongruence among gene trees

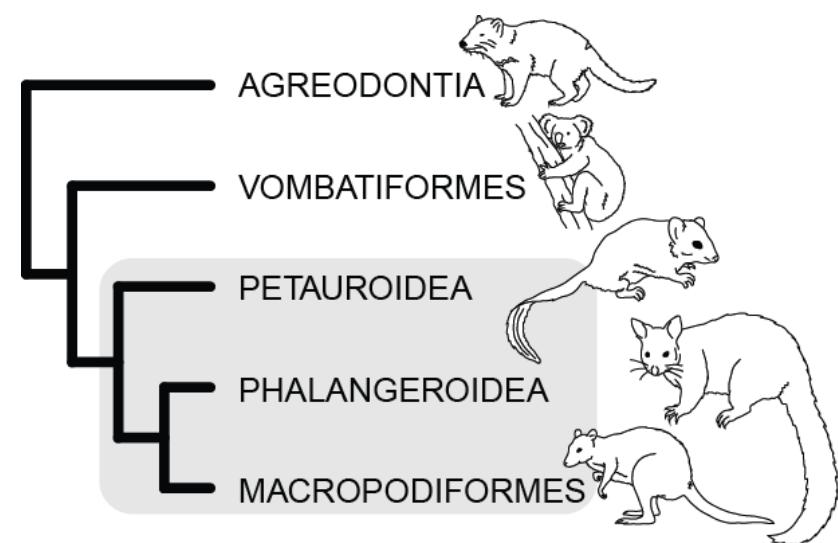
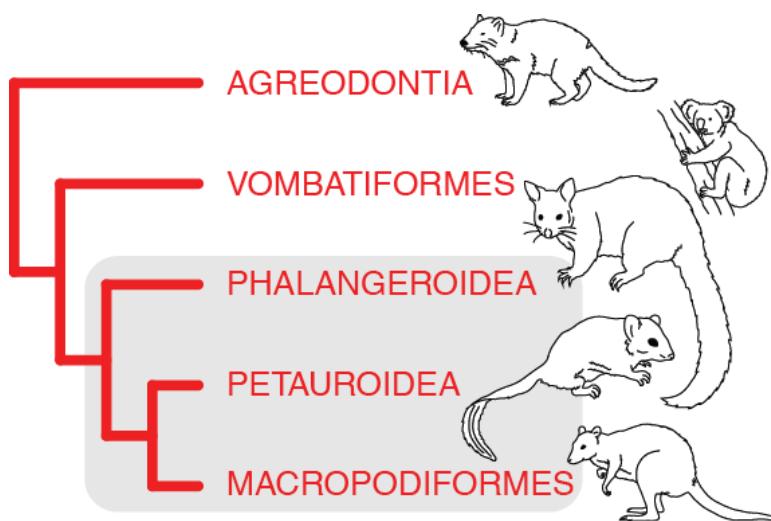
- Unlinked loci can have different gene trees
- Test for phylogenetic congruence across markers
- Partition-homogeneity (incongruence length difference) test



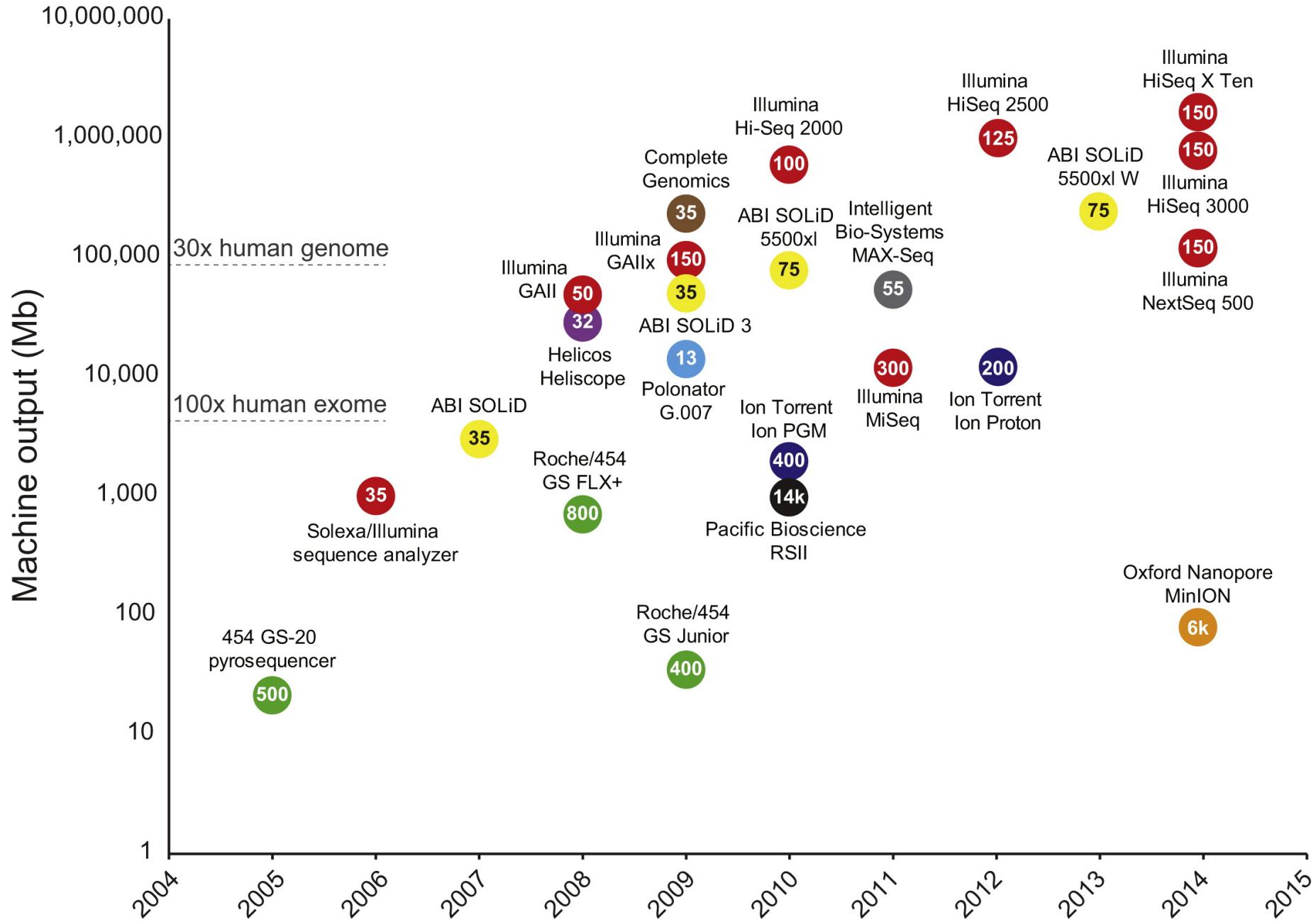
Incongruence among gene trees



Duchene *et al.* (2018)
Syst Biol

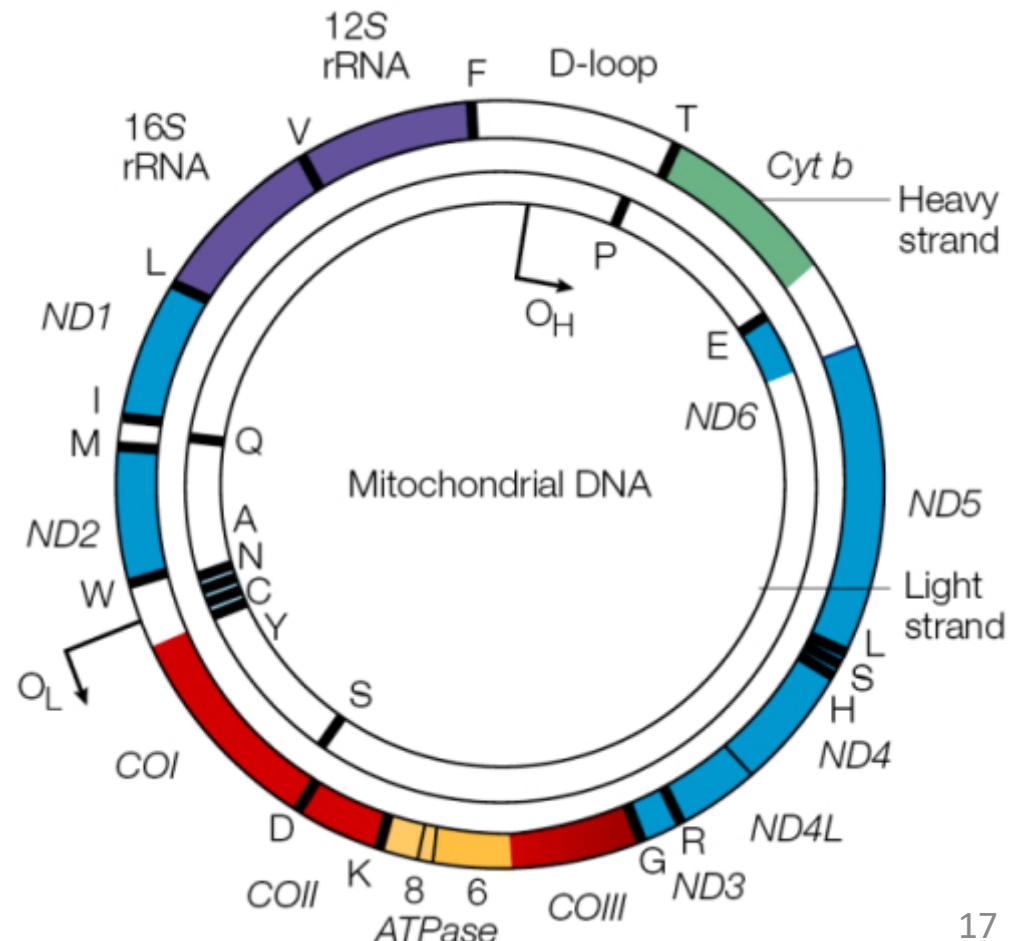


High-Throughput Data



Mitochondrial genomes

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

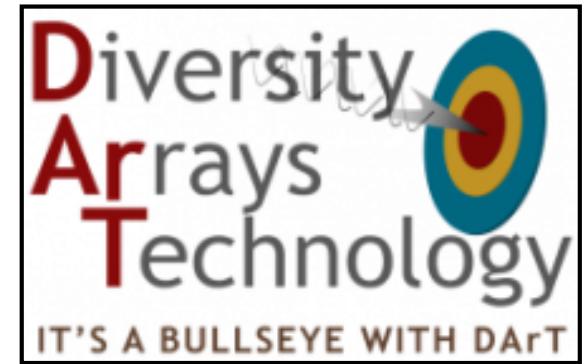


Single-nucleotide polymorphisms

- 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



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

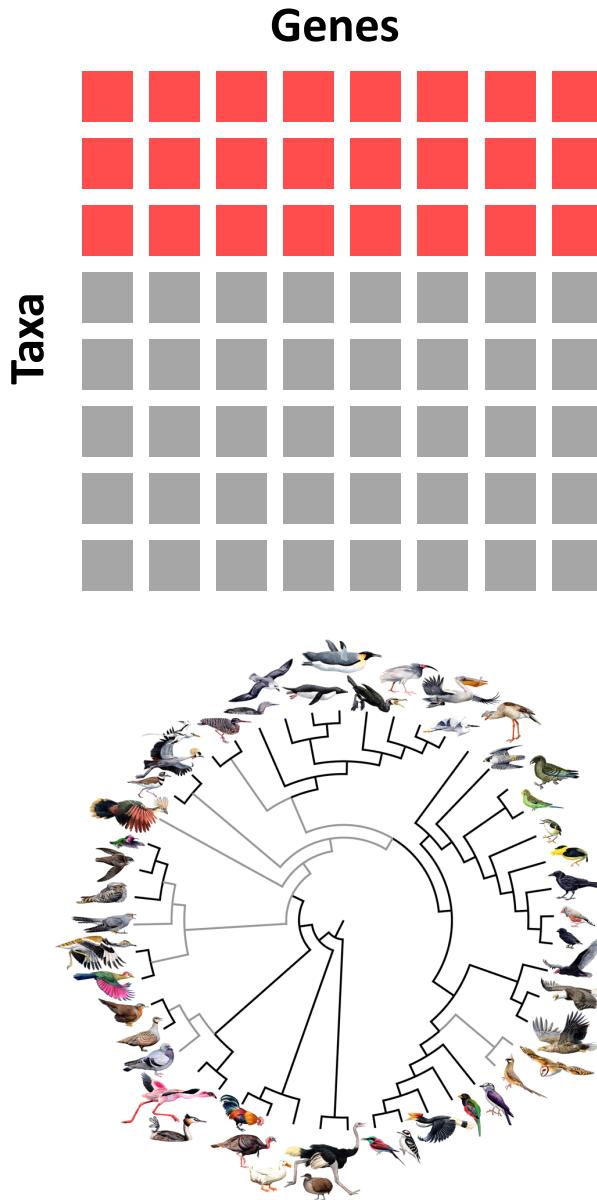
Whole genome sequencing

- Typically NOT (yet) the entire genome
- Many challenges: Jarvis et al *Science* 2014 >400 years of computing using a single processor
- **Issues to consider**
 - Loci are single-copy?
 - Selectively neutral?
 - Loci have discordant gene trees?
 - Historical recombination?



Analysing Large Data Sets

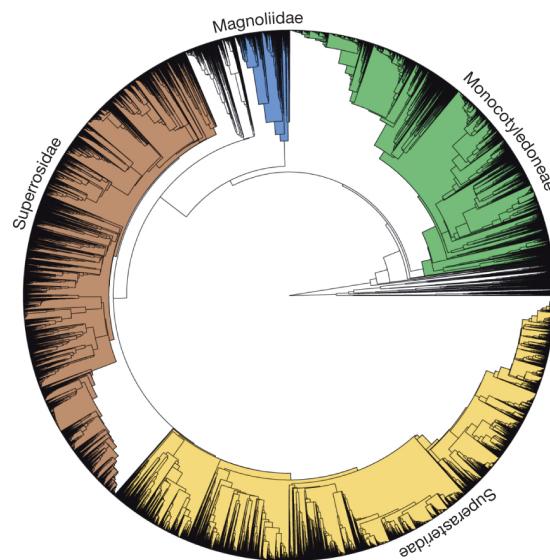
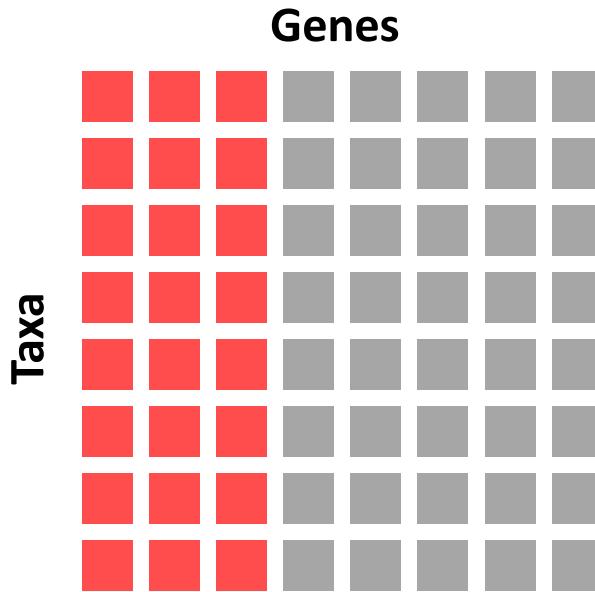
Large data sets



- Calculation of likelihood is expensive
 - Speed up by grouping sites with identical patterns
 - Approximate likelihood calculation
 - Multithreading/parallelisation

48 taxa
8,295 genes
Jarvis *et al.* (2014) *Science*

Large data sets

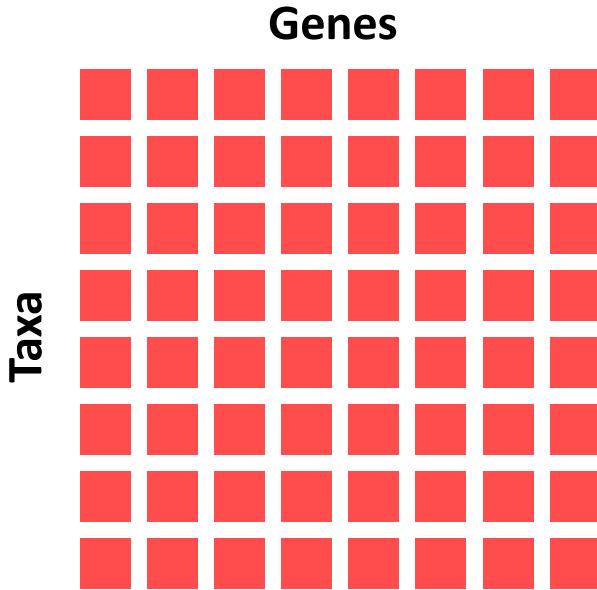


- Tree-space is extremely large
 - Efficient tree-searching heuristics

32,223 taxa
7 genes

Zanne *et al.* (2014) *Nature*

Large data sets



- Analysis is computationally expensive
- Consider filtering the data
 - Phylogenetic signal
 - Mutational saturation
 - Missing data
 - Model fit

Useful references

