Lecture 1.3

Phylogenetic Data

Mark de Bruyn

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 characters

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]

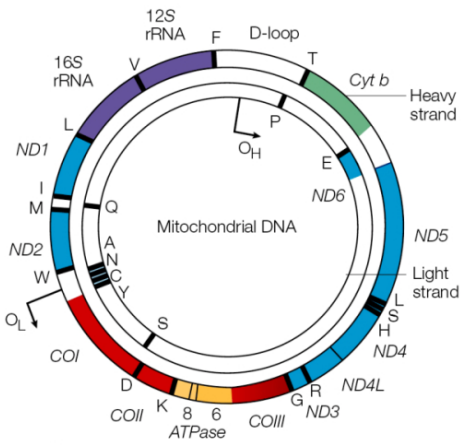
Sequence Data

Sequence data

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

Commonly used DNA sequence loci

- **Mitochondrial genome**
  - Maternally inherited
  - Protein-coding genes
  - rRNA genes
  - Control region



Commonly used DNA sequence loci

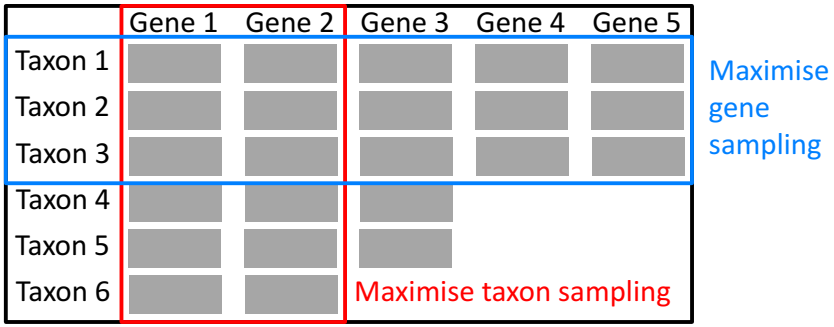
- **Nuclear genome**
  1. **Microsatellites (STRs)**: highly polymorphic, mutation model, size homoplasy
  2. **EPICs**: highly variable, indels
  3. **Anonymous loci**: highly variable, no ascertainment bias, neutral?
  4. **NPCLs**: conserved, lack indels, 'ancient' events
  5. **rRNA 'arrays'**: tandemly duplicated, variable conservation

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



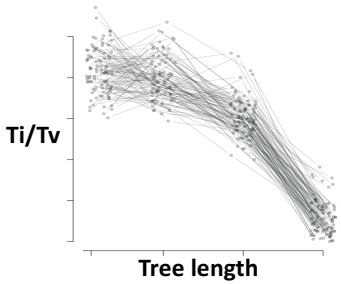
Mutational saturation

- 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

Saturated sites can be removed to improve signal:noise

Mutational saturation

- **Plot transitions and transversions**
  - Transitions occur more frequently than transversions
  - Can plot in various ways



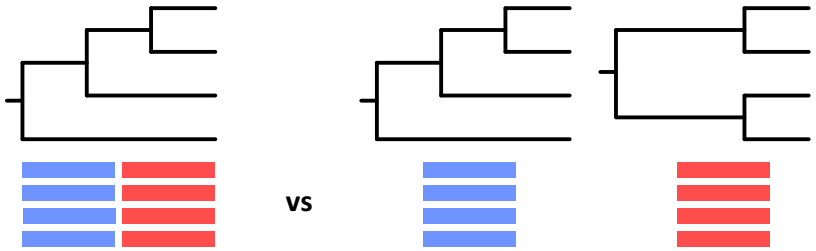
- **Xia's method (I_{SS})**
  - When sequences are fully saturated (I_{SS} = 1), expected base frequencies at each site are equal to the global frequencies
  - Compare I_{SS} with critical value calculated via simulation

Duchêne, Ho, & Holmes (2015) *BMC Evol Biol*

Xia *et al.* (2003) *Mol Phylogenet Evol*

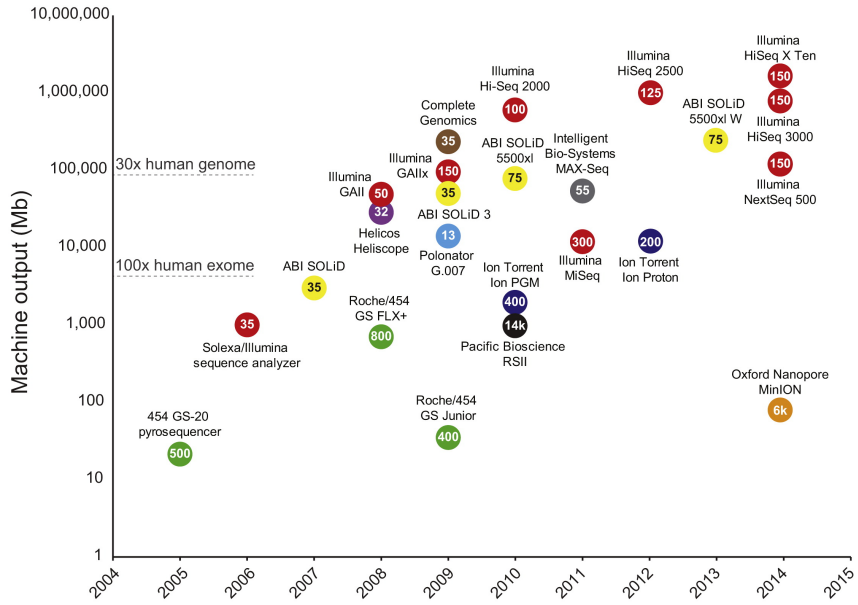
Partition-homogeneity test

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



138 steps 78 steps + 57 steps

High-Throughput Data



Reuter *et al.* (2015) *Molecular Cell*

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

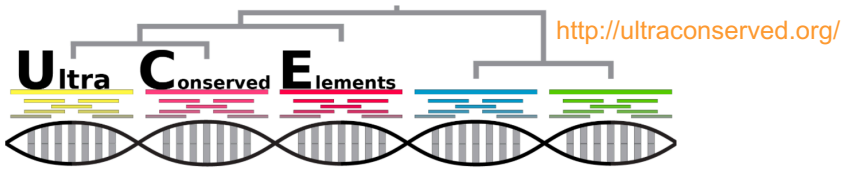
- Markers identified by cutting genome with restriction enzymes
- Process creates binary data and short sequences
- Examples include RADseq and DArTseq

Diversity Arrays Technology
IT'S A BULLSEYE WITH DArT

- 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

Ultra-conserved elements

- Genomic elements perfectly conserved in mammals (>200 bp)
- Gene 'deserts' – long range regulators?
- Flanking regions: phylogenetic markers varying in conservation
- Purifying selection?
  - Increase rate of lineage sorting: gene trees ≈ spp. trees
  - Inference of historical demographic parameters problematic – Ne



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?
  - Sampled loci have independent gene trees?
  - Historical recombination?

Useful references