
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

2

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



3

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}

4

Sequence Data

Sequence data

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



A 454 Illumina sequencing machine, a white and grey laboratory instrument. It features a monitor on top displaying a blue sphere and a bar chart. The machine has a large front-loading compartment and a smaller side compartment. The text '454 Illumina' is visible on the front, and 'Illumina' is on the side.

6

-
- A white GeneSinger Sequencer 2000 with a 454 Illumina logo and a monitor displaying a blue sphere and cube.

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

8

Commonly used DNA sequence loci

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

The diagram illustrates the circular structure of Mitochondrial DNA (mtDNA). The outer circle represents the heavy strand, and the inner circle represents the light strand. The D-loop region is located between the 12S rRNA and 16S rRNA genes. The control region (D-loop) contains the origin of heavy strand replication (O_H) and the origin of light strand replication (O_L). The genes shown include 12S rRNA, 16S rRNA, ND1, ND2, ND3, ND4, ND4L, ND5, ND6, ND7, ND8, ND9, ND10, ND11, ND12, ND13, ND14, ND15, ND16, ND17, ND18, ND19, ND20, ND21, ND22, ND23, ND24, ND25, ND26, ND27, ND28, ND29, ND30, ND31, ND32, ND33, ND34, ND35, ND36, ND37, ND38, ND39, ND40, ND41, ND42, ND43, ND44, ND45, ND46, ND47, ND48, ND49, ND50, ND51, ND52, ND53, ND54, ND55, ND56, ND57, ND58, ND59, ND60, ND61, ND62, ND63, ND64, ND65, ND66, ND67, ND68, ND69, ND70, ND71, ND72, ND73, ND74, ND75, ND76, ND77, ND78, ND79, ND80, ND81, ND82, ND83, ND84, ND85, ND86, ND87, ND88, ND89, ND90, ND91, ND92, ND93, ND94, ND95, ND96, ND97, ND98, ND99, ND100. The genes are color-coded: 12S rRNA (purple), 16S rRNA (blue), ND1 (light blue), ND2 (dark blue), ND3 (red), ND4 (orange), ND4L (yellow), ND5 (green), ND6 (light green), ND7 (dark green), ND8 (brown), ND9 (pink), ND10 (light pink), ND11 (dark pink), ND12 (red), ND13 (orange), ND14 (yellow), ND15 (green), ND16 (light green), ND17 (dark green), ND18 (brown), ND19 (pink), ND20 (light pink), ND21 (dark pink), ND22 (red), ND23 (orange), ND24 (yellow), ND25 (green), ND26 (light green), ND27 (dark green), ND28 (brown), ND29 (pink), ND30 (light pink), ND31 (dark pink), ND32 (red), ND33 (orange), ND34 (yellow), ND35 (green), ND36 (light green), ND37 (dark green), ND38 (brown), ND39 (pink), ND40 (light pink), ND41 (dark pink), ND42 (red), ND43 (orange), ND44 (yellow), ND45 (green), ND46 (light green), ND47 (dark green), ND48 (brown), ND49 (pink), ND50 (light pink), ND51 (dark pink), ND52 (red), ND53 (orange), ND54 (yellow), ND55 (green), ND56 (light green), ND57 (dark green), ND58 (brown), ND59 (pink), ND60 (light pink), ND61 (dark pink), ND62 (red), ND63 (orange), ND64 (yellow), ND65 (green), ND66 (light green), ND67 (dark green), ND68 (brown), ND69 (pink), ND70 (light pink), ND71 (dark pink), ND72 (red), ND73 (orange), ND74 (yellow), ND75 (green), ND76 (light green), ND77 (dark green), ND78 (brown), ND79 (pink), ND80 (light pink), ND81 (dark pink), ND82 (red), ND83 (orange), ND84 (yellow), ND85 (green), ND86 (light green), ND87 (dark green), ND88 (brown), ND89 (pink), ND90 (light pink), ND91 (dark pink), ND92 (red), ND93 (orange), ND94 (yellow), ND95 (green), ND96 (light green), ND97 (dark green), ND98 (brown), ND99 (pink), ND100 (light pink).

-

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

8

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**

9

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						Maximise gene sampling
Taxon 2						
Taxon 3						
Taxon 4						Maximise taxon sampling
Taxon 5						
Taxon 6						

10

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

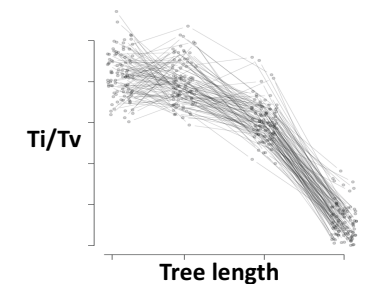
Saturated sites can be removed to improve signal:noise

11

Mutational saturation

- **Plot transitions and transversions**

- Transitions occur more frequently than transversions
- Can plot in various ways

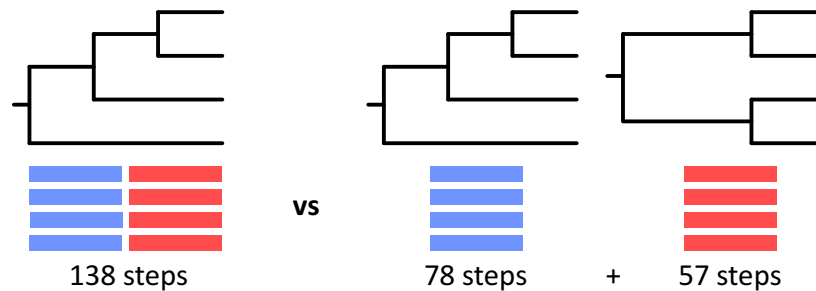


- **Xia's method (I_{SS})**

- 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

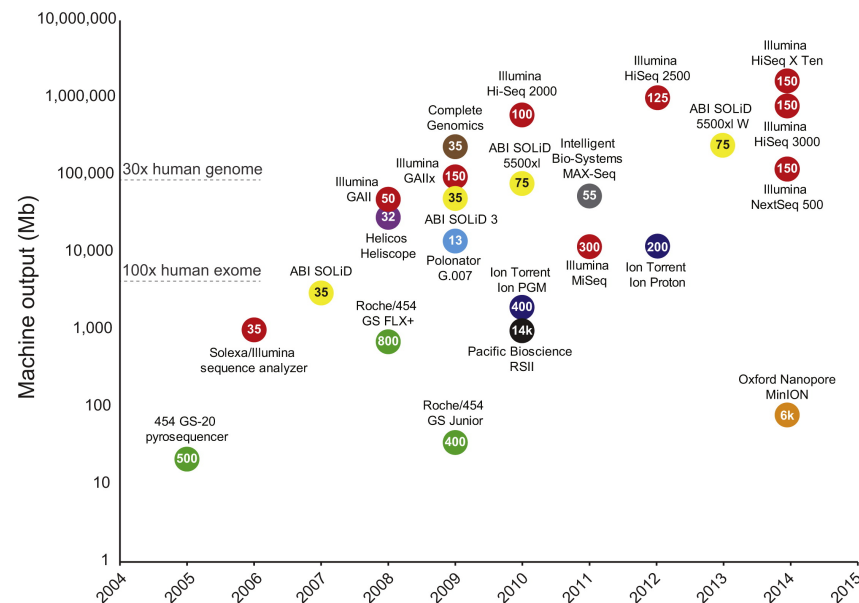
Partition-homogeneity test

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



13

High-Throughput Data



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

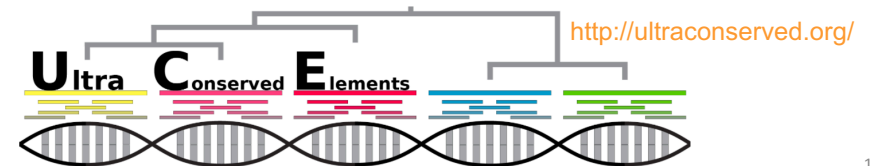
- Markers identified by cutting genome with restriction enzymes
- Process creates binary data and short sequences
- Examples include RADseq and DArTseq
- Issues to consider:
 - **Recombination**
Markers are usually unlinked so they are likely to have different (gene) trees
 - **Missing data**
Typically a large proportion of missing data



17

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 \approx spp. trees
 - Inference of historical demographic parameters problematic – Ne



18

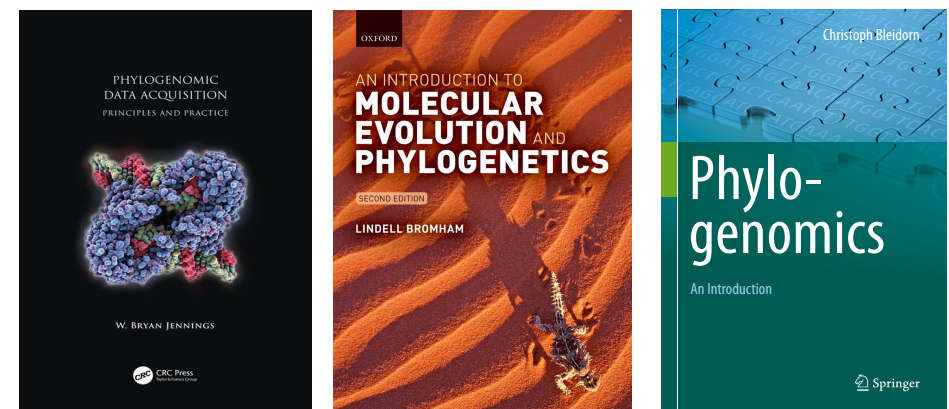
Whole genome sequencing

- 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?



19

Useful references



20