Phylogenetic Trees cont.

Stefan Wyder

May 2016
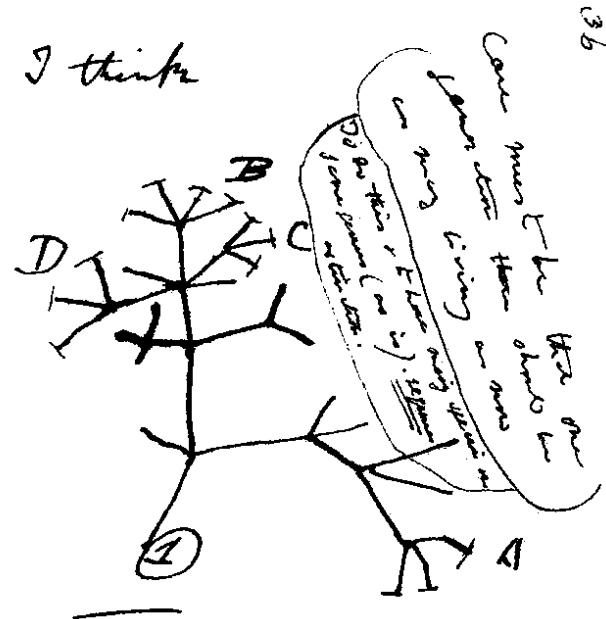
Universität Zürich UZH
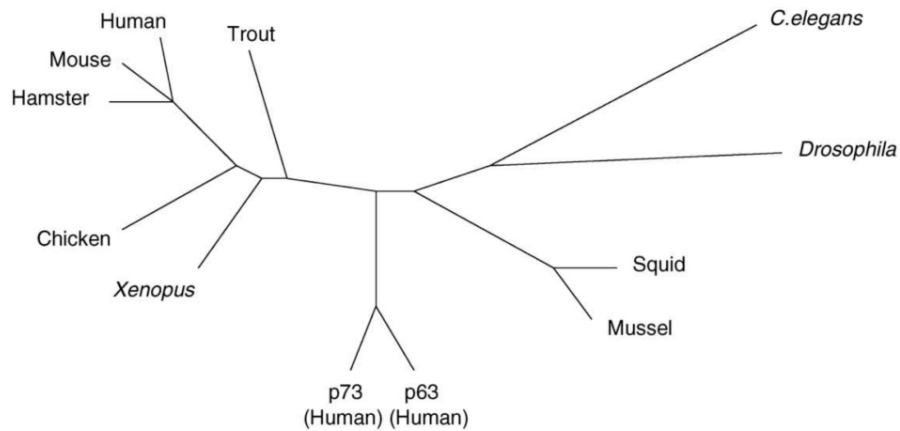
URPP Evolution in Action

The first tree

Charles Darwin 1837

The phylogenetic tree

unrooted tree

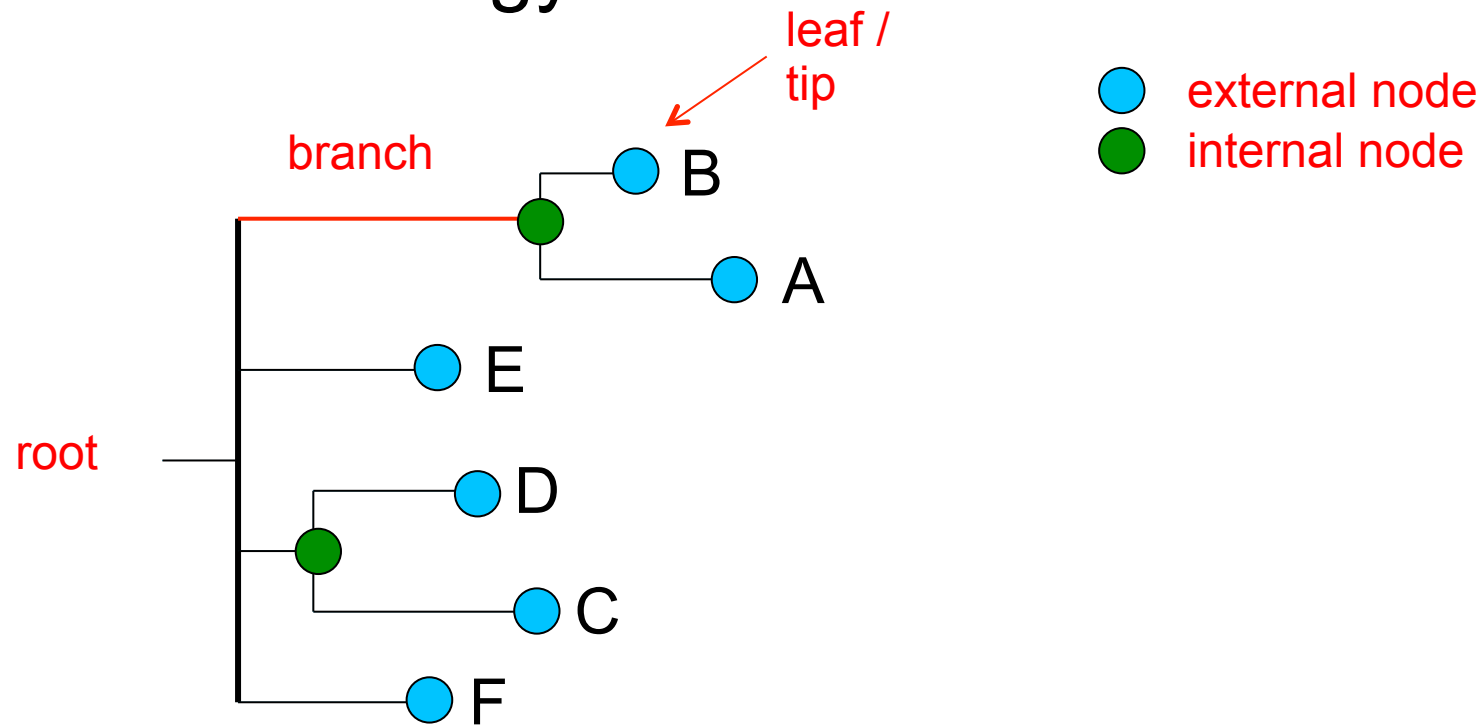evolutionary distance between lineages

rooted tree
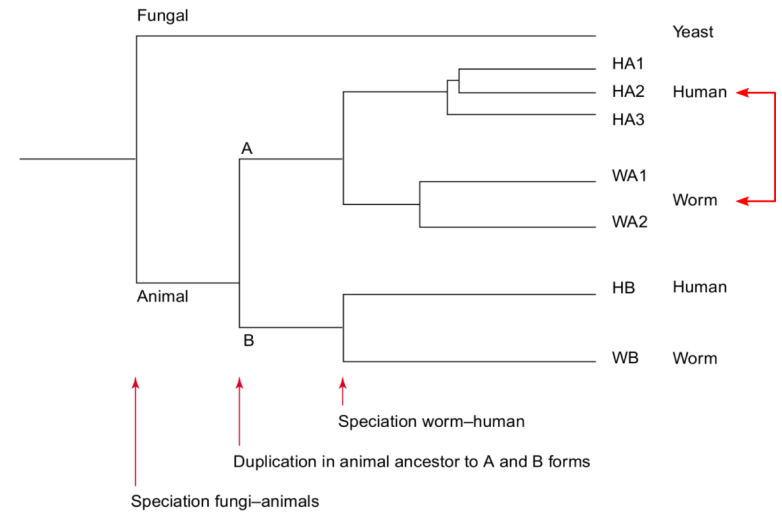
evolutionary change (e.g. substitutions per site) →
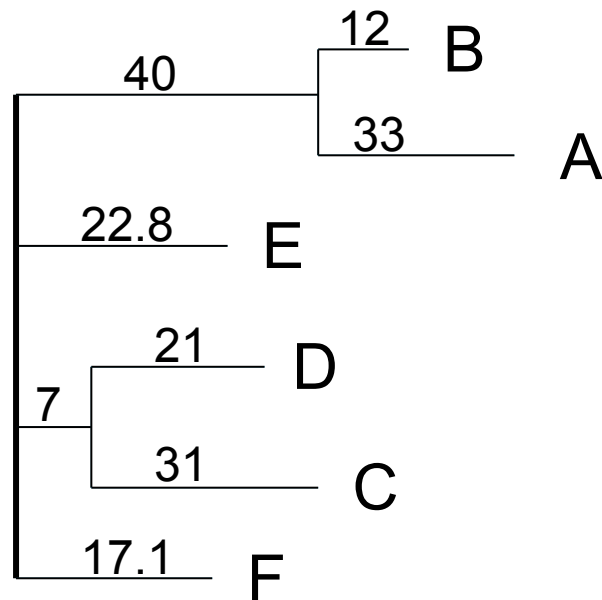
order of evolutionary events

vertical dimension has no meaning

Terminology

The branching pattern is called topology

Newick format

Labels
any string except blanks : ; ( ) [ ]
or quoted
_ are printed as blanks: strainB_2

same number of "(" and ")"
ends with a semicolon
can be multifurcating

(F, (C, D), E, (A, B));

(F:17.1, (C:31, D:21):7, E:22.8, (A:33, B:12):40);

Bootstrap support values (%, sometimes [0,1])
(F:17.1, (C:31, D:21)99:7, E:22.8, (A:33, B:12)94:40);

NEXUS format

```
#NEXUS
BEGIN TAXA;
  TAXLABELS A B C;
END;

BEGIN TREES;
  TREE tree1 = ((A,B),C);
END;
```
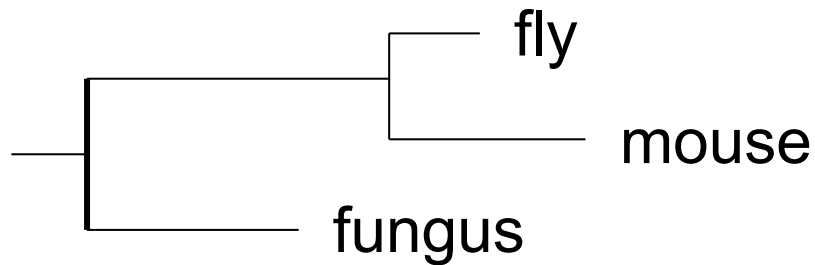
Newick Tree

Blocks
Each block starts with "BEGIN block_name;" and finishes with "END;"

[Comments enclodes in square brackets]

file ending: .nex or .nxs
NEXUS is also an alignment format (BEGIN DATA; ... END;)

Tree rooting

Most methods that reconstruct phylogenies from molecular sequenced do not calculate the root of the tree - the tree is generated with an arbitrary root

Tree rooting needs external evidence
2 methods to find the root:

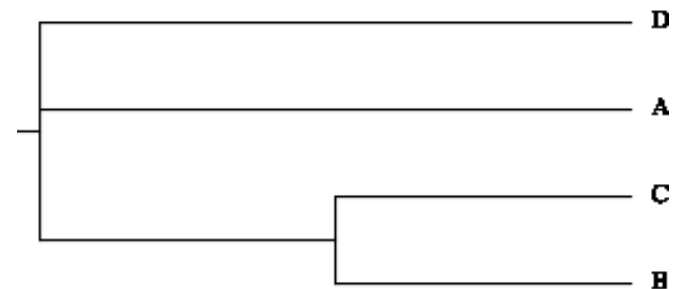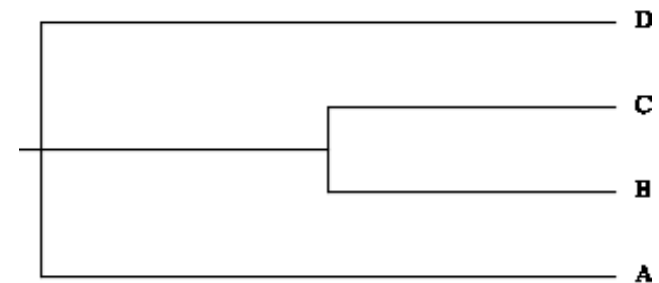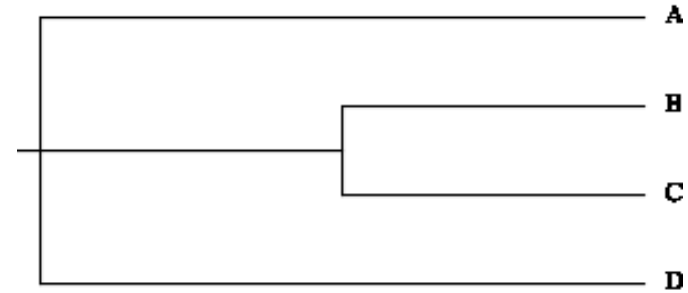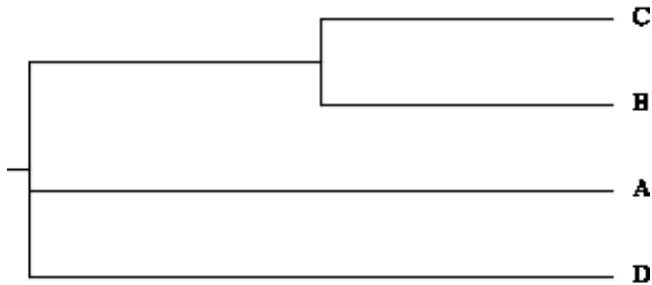1) outgroup: we know where last common ancestor of all samples was
2) molecular clock model

For displaying (arbitrary) midpoint rooting is often used (root in the midpoint of the longest branch)

Not a unique representation of a tree

are all the same tree

(A, (B, C), D);
(A, (C, B), D);
(D, (C, B), A);
(D, A, (C, B));
((C, B),A, D);

left-right order of descendants of a node
affects the representation
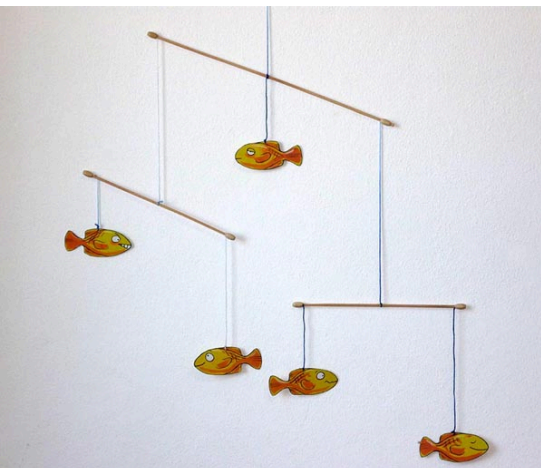
Tree visualization/manipulation

Manipulation: tree (re)rooting, renaming, reordering, pruning/subsetting, collapsing, comparing trees

- **Interactive tree viewers:**
  most popular: FigTree, Dendroscope

- **Qualitative summary of sets of tree:**
  DensiTree shows uncertainty in topology

- **Command Line / Scripting (pipelines):**
  Newick Utilities, biopython, bioperl, R with specialized packages (e.g. ape), other specialized python/perl packages

Potential Pitfalls

- Violating assumptions can lead to systematic errors:
  alignment of orthologous positions
  characters in the alignment are treated as independant

- Long branch attraction (LBA)

- Bootstrap values: a measure of reliability
  tells us what would be expected to happen if we repeated our experiment

  Not a measure of accuracy: does not tell us the probability of our experiment being true

  If the method of reconstruction falls victim to a bias or an artifact, we make it 1000x

Long branch attraction (LBA)

True tree

Reconstructed tree

P

Q

S

R

(A)

P

S

R

Q

(B)
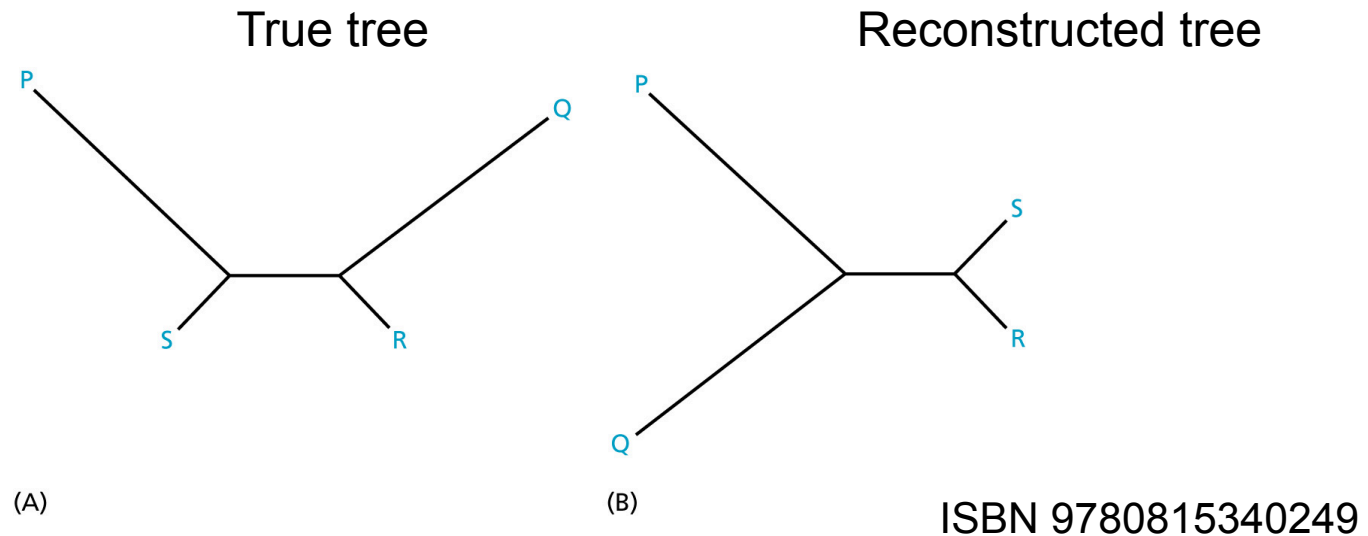
- many tree inference algorithms will incorrectly group the long divergent branches together regardless oft their true relationship
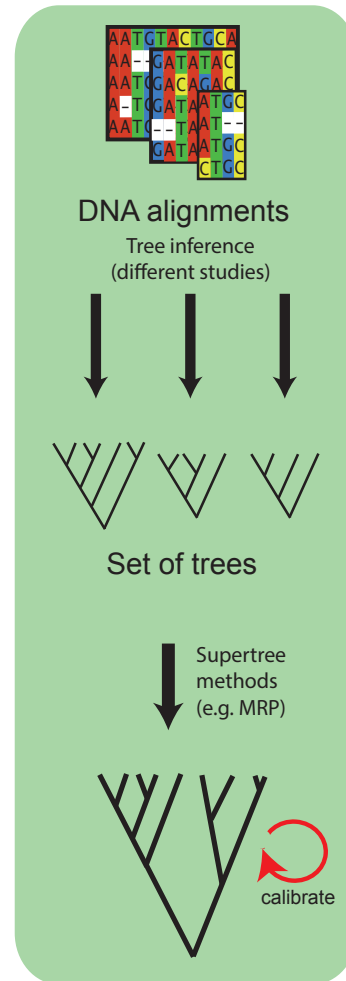- The frequency of LBA is unclear and debated
- all tree construction algorithms are susceptible (parsimony is particularly susceptible)
- Siddal and Whiting method: rerun analysis without species P, then rerun analysis without species Q. If either of the taxa appear at different branch points in the absence of the other, there is evidence of LBA
- remove fast evolving sites from the alignment
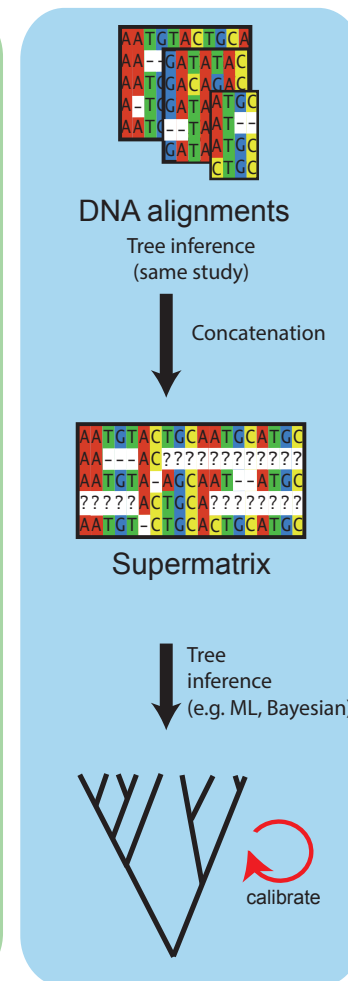- add taxa related to those with the long branches

Supertree / Supermatrix

- can include heterogenous data (e.g. trees from the literature, morphological traits)

- suited for large systematic groups

- lacks any statistical model of evolutionary change

- Methods
  - consensus tree
  - %-majority rule tree
 - MRP/MRL: matrix representation with parsimony or likelihood

Supertree

AATGTACTGCA
AA--GATATAC
AATCGACAGAC
A-TCGATAATGC
--GATAATGC
CTGC

DNA alignments
Tree inference
(different studies)

Set of trees

Supertree
methods
(e.g. MRP)

calibrate

Supermatrix

AATGTACTGCA
AA--GATATAC
AATCGACAGAC
A-TCGATAATGC
--GATAATGC
CTGC

DNA alignments
Tree inference
(same study)

Concatenation

AATGTACTGCAATGCATGC
AA---AC????????????
AATGTA-AGCAAT--ATGC
?????ACTGCA????????
AATGT-CTGCACTGCATGC

Supermatrix

Tree
inference
(e.g. ML, Bayesian)

calibrate

- more popular

- PartitionFinder to find best partitioning scheme and substitution model

- averaging over multiple loci assumes that all loci share a common evolutionary history (e.g. no horizontal gene transfer or incomplete lineage sorting)

Antonelli, 2014
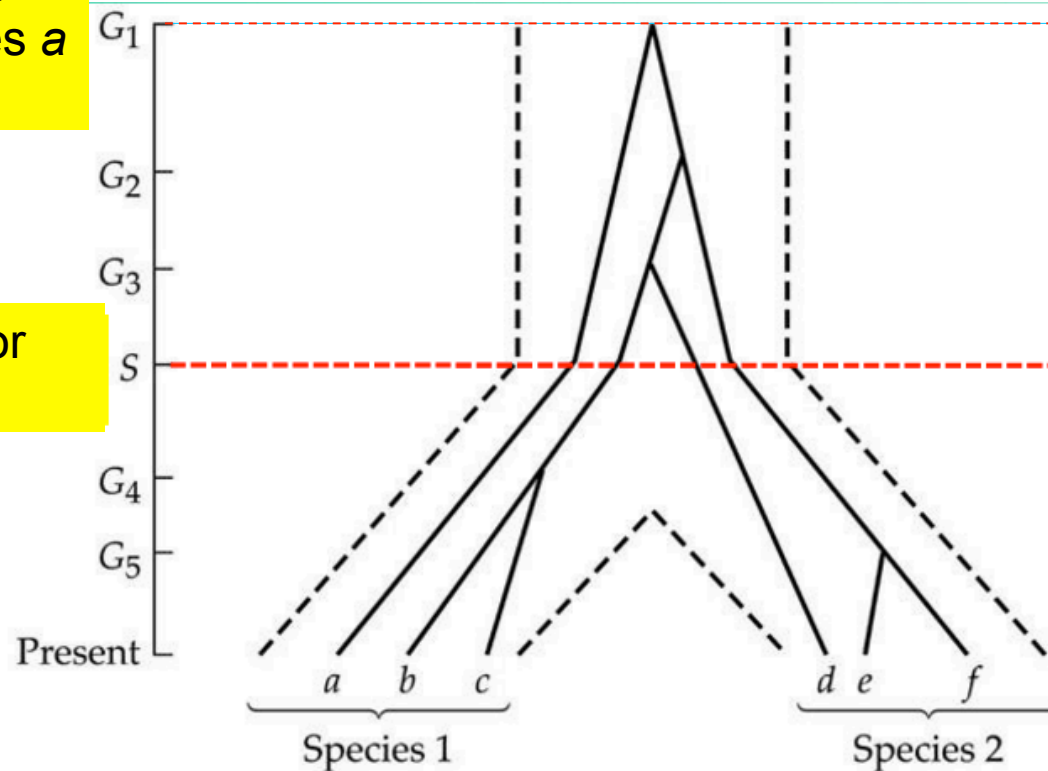
Gene Tree ≠ Species Tree 1
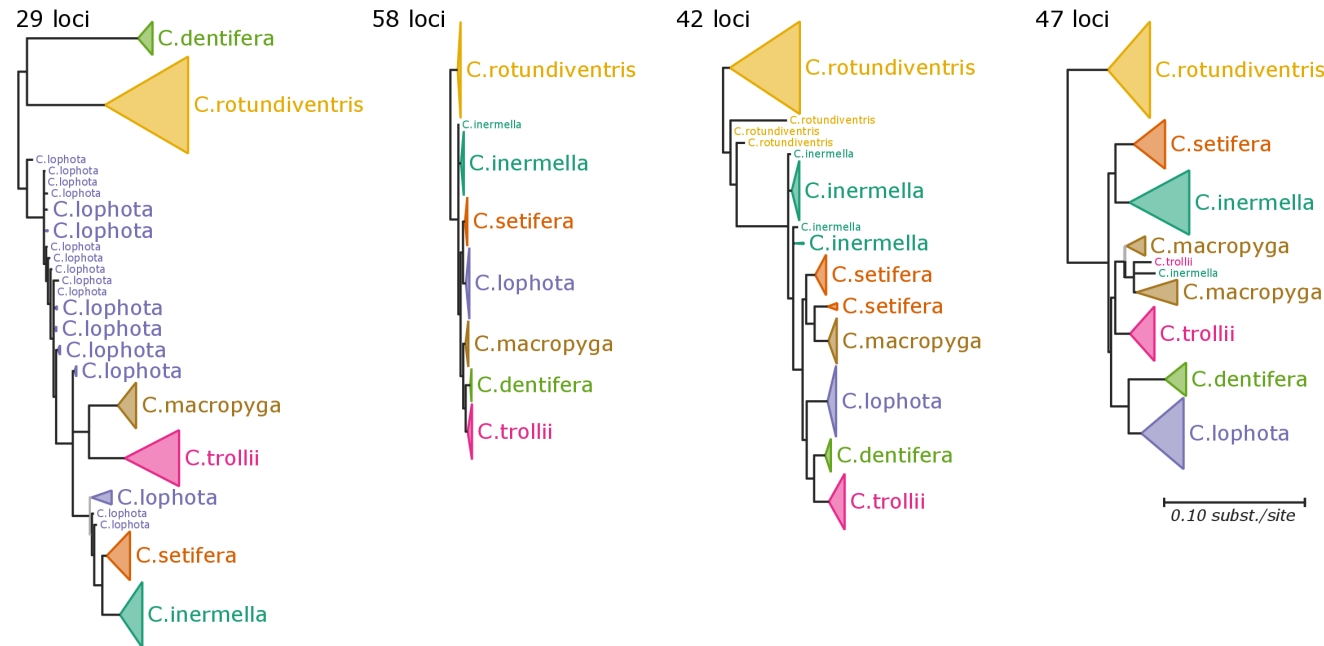
PMID: 26893301

- Simulation and 2 real data sets (Gori et al. MBE 2016)
- RAD-seq: 306 samples from 7 globeflower fly species
- 176 loci (with >100 indiv.)
- Bayesian analysis
- Branches with support values <.9 were collapsed into multifurcations
- -> 7 distinct species (different colors) whose branching order varies substantially across loci suggests incomplete lineage sorting
- 6/7 species thought to have radiated more or less synchronously

Incomplete Lineage Sorting

Inferred divergence time by using alleles *a* and *f*
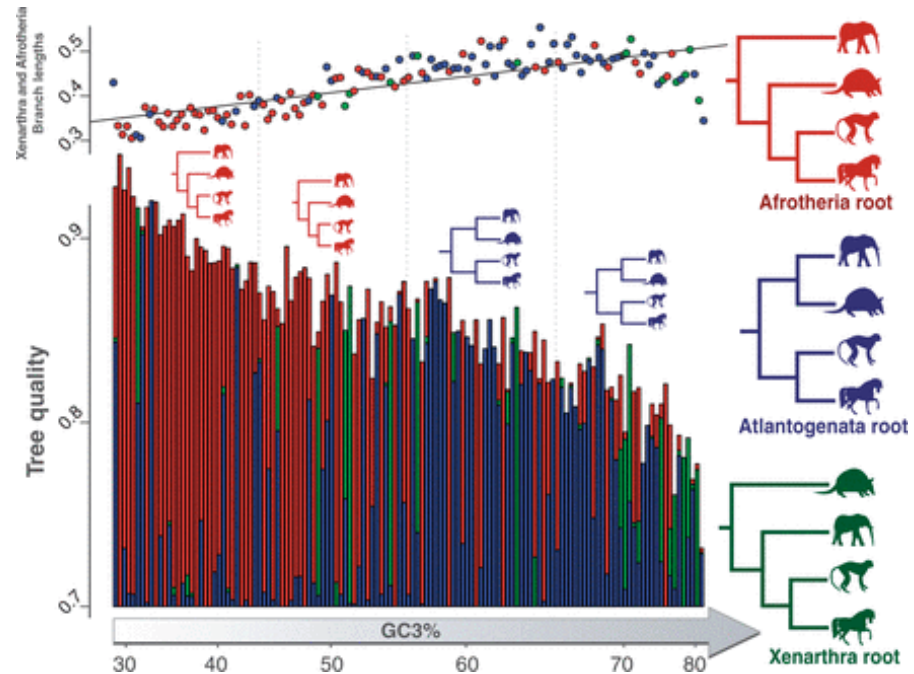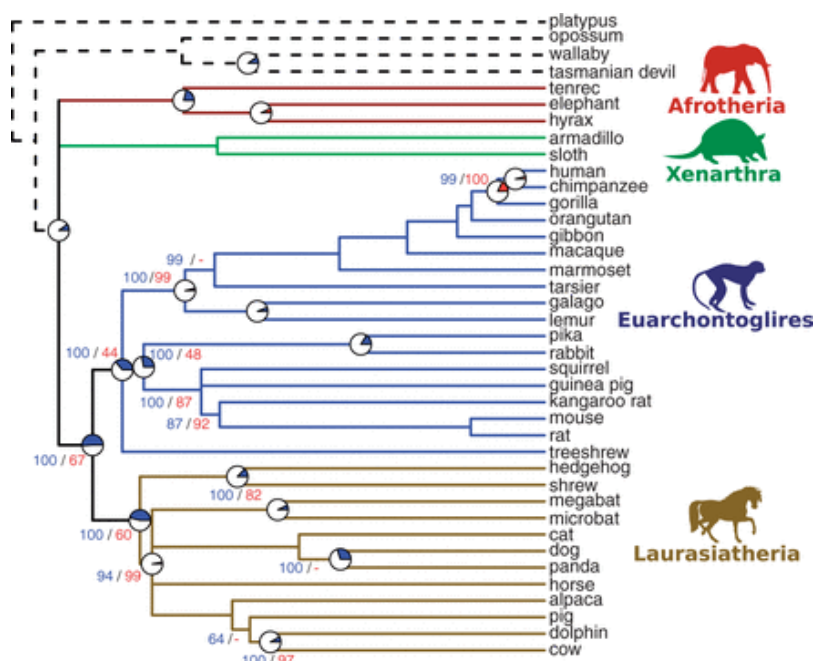
Divergence time for species 1 and 2

Alleles *d* and *b* are closer to each other than alleles *d* and *f*

http://slideplayer.com/slide/3525255/

Incomplete lineage sorting due to polymorphisms at speciation time

Gene Tree ≠ Species Tree 2

PMID:23813978

- Mammalian phylogenomics: rooting of the placental mammal tree is still controversial
- RAxML, Supermatrix and supertree approach (13,111 coding sequence alignments)
- GC-rich genes induced a higher amount of conflict among gene trees
- GC-rich genes performed worse than AT-rich genes in retrieving well-supported, consensual nodes on the placental tree
- GC3-content reflects genome-wide variation in recombination rate?

Summary Gene Tree ≠ Species Tree

- we often observe incongruent trees with different topology

- noise vs non-common evolutionary history

- It can be misleading to infer a single tree
  - Incomplete lineage sorting
  - horizontal gene transfer
  - hybridisation
  - recombination
  - gene duplication
  - migration
  ...

- not reduced by adding more data

- can be modelled by specialized "mechanistic" methods

Complex evolutionary relationships

Phylogenetic Networks

- used to visualize complex evolutionary relationship leading to incompatible phylogenetic signals

- difference from phylogenetic trees: addition of hybrid nodes (nodes with two parents) instead of only tree nodes (nodes with only one parent)

- Software: SplitsTree (uses PhyML or Parsimony), Dendroscope
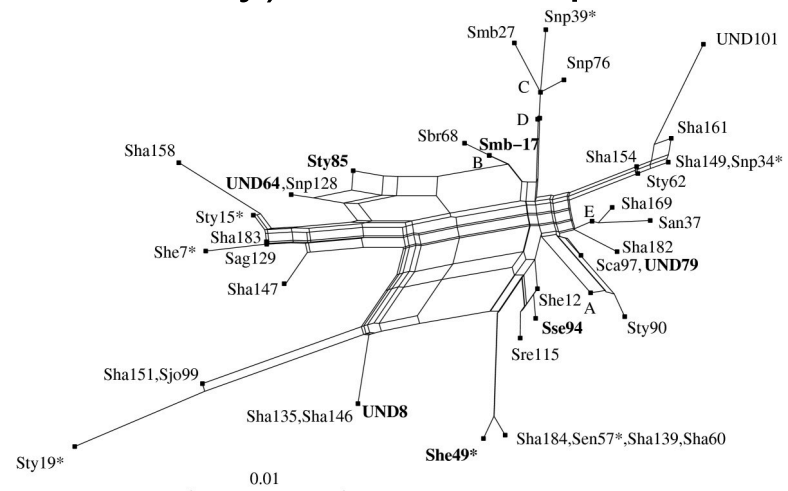
Coalescent-based species tree estimation

- Tools: ASTRAL-II, *BEAST/SNAPP, MP-EST

Further topics

Molecular Phylogenetics

- dating phylogenetic trees

- detecting positive selection on coding genes

- macroevolution (comparative methods)
  i) dating divergence times
  ii) mode and tempo of evolution
  iii) testing key innovations

Comparative methods use the distribution of trait
make interference about traits evolution

Healey, 2014

many methods available
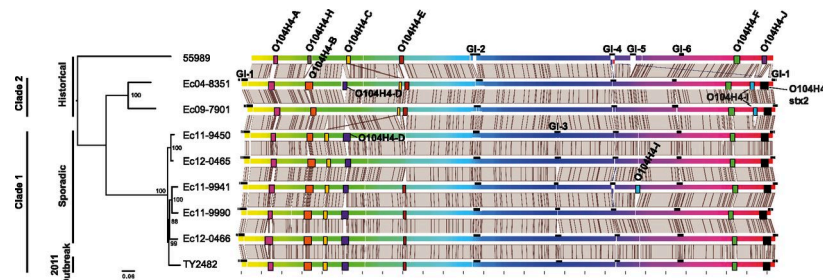

- Species cannot be seen as independent outcomes of evolution
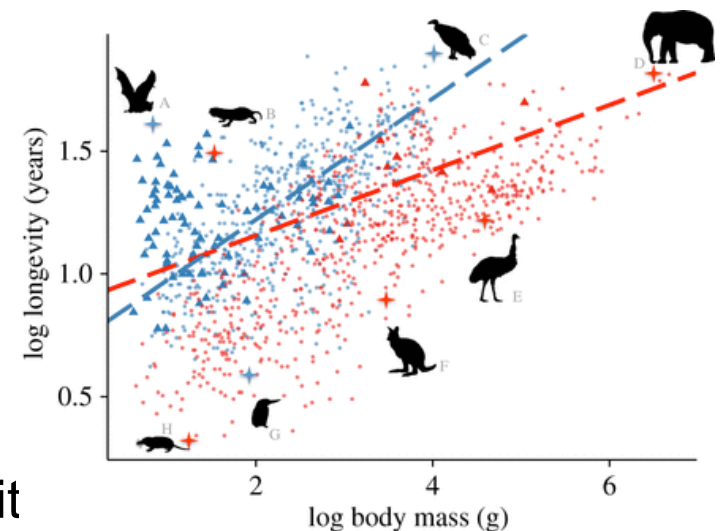
Comparative Genomics

- The study of the genome features* and
  function across different biological species/strains

  * Features are e.g. DNA sequence, genes, gene order, regulatory sequences,
  chromosomal rearrangements, …


Genome rearrangements
of E. coli isolates
[PMID: 23341549]

- Over evolutionary time, non-functional sequences are expected to
  diverge faster than sequences under selective constraint

- Finding often summarized in a tree

- Needs genomes at the appropriate phylogenetic distance for the
  question
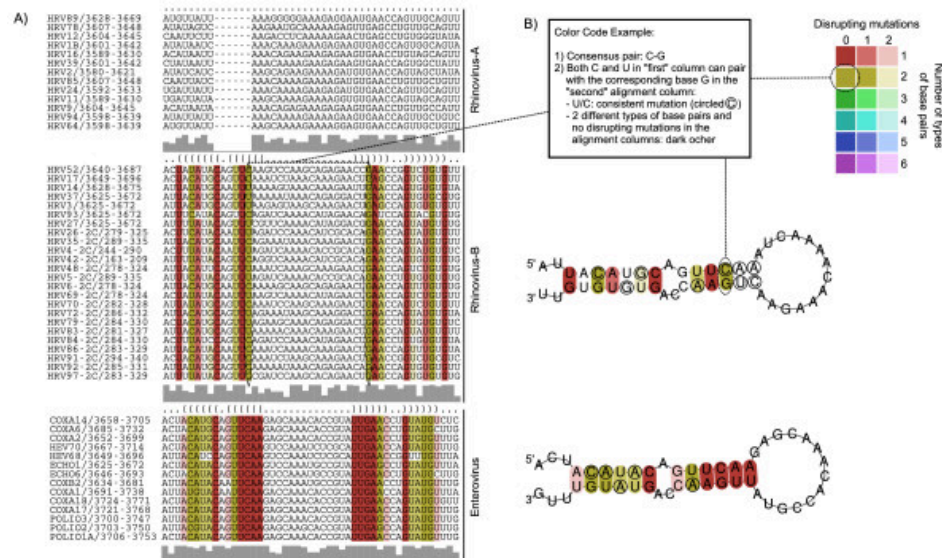
Identification of highly conserved or rapidly evolving sequences:
cis-acting 2C replication elements
in rhinoviruses [17623054]
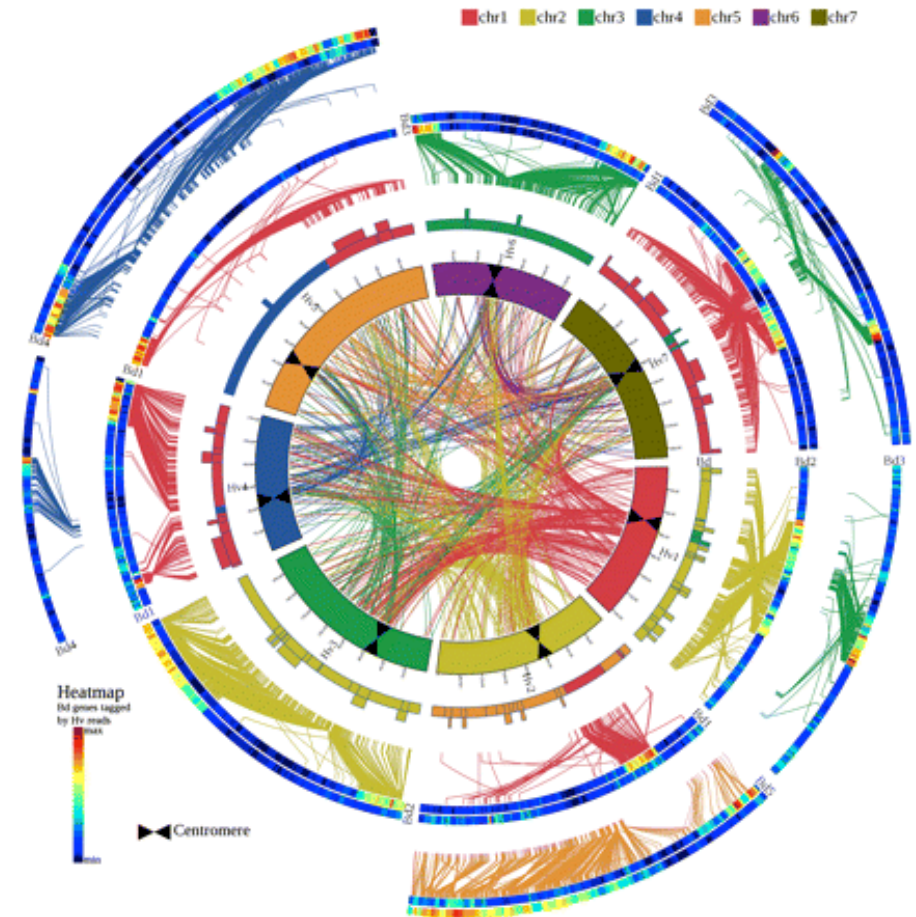
Intron evolution
in eukaryotes
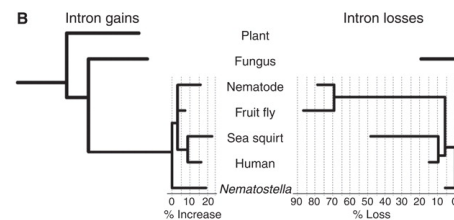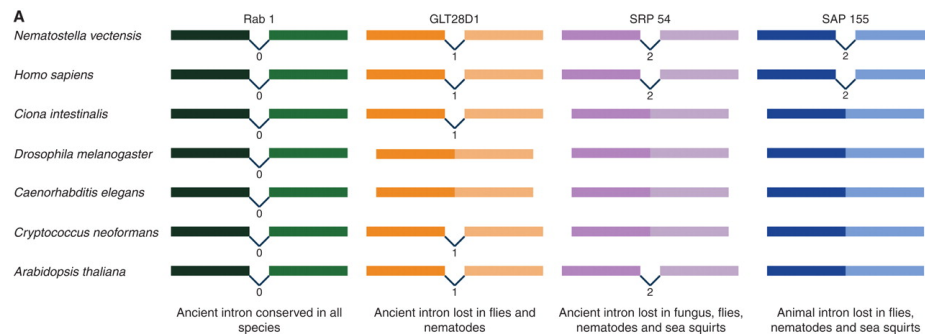
Barley Genome vs Brachypodium
[21467582]

Sources & Links

- Teaching site Prof. N. Salamin (Uni Lausanne)
  http://www2.unil.ch/phylo/course/pmc

- More links see URPP tutorial webpage:
  https://github.com/milchmolch/Phylogenomics