Molecular phylogenetics

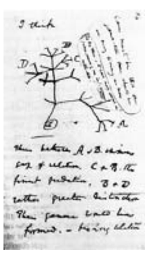
Michael Worobey

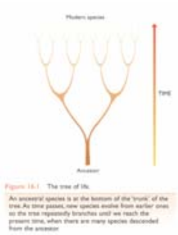
Next two lectures:

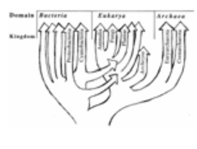
- What is a phylogenetic tree?
- How are trees inferred using molecular data?
- How do you assess confidence in trees and clades on trees?
- Why do trees for different data sets sometimes conflict?
- What can you do with trees beyond simply inferring relatedness?

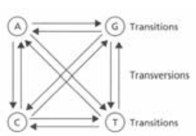
Molecular phylogenetics fundamentals

All of life is related by common ancestry. Recovering this pattern, the "Tree of Life", is one of the primary goals of evolutionary biology. Even at the population level, the phylogenetic tree is indispensable as a tool for estimating parameters of interest. Likewise at the among species level, it is indispensable for examining patterns of diversification over time. First, you need to be familiar with some tree terminology.



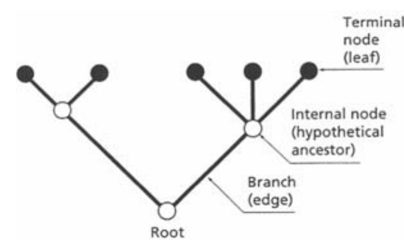






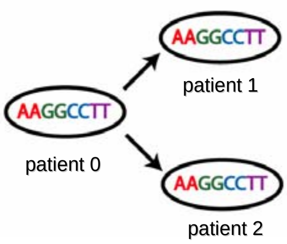
Tree terminology

A **tree** is a mathematical structure which is used to model the actual evolutionary history of a group of sequences or organisms. This actual pattern of historical relationships is the **phylogeny** or **evolutionary tree** which we try and estimate. A tree consists of **nodes** connected by **branches** (also called **edges**). **Terminal nodes** (also called **leaves, OTUs [Operational Taxonomic Units]**, **external nodes** or **terminal taxa**) represent sequences or organisms for which we have data; they may be either extant or extinct.



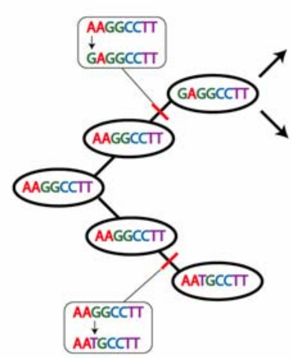
Phylogenetics interlude

All the phylogenetics you need to know, in 5 minutes…



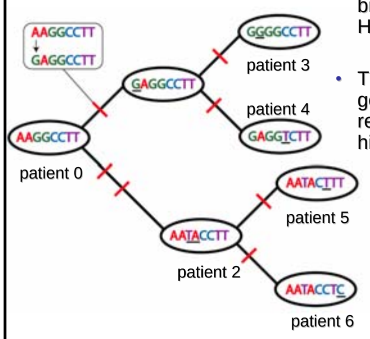
- It's all about ancestor and offspring populations, lineages branching
- The ancestor could be distant great grandmother or a human immunodeficiency virus
- The ancestral form of some gene (a "marker") is inherited in two offspring lineages
- Let's assume that we're looking at virus from a "patient 0" who then infects two others

Phylogenetics interlude



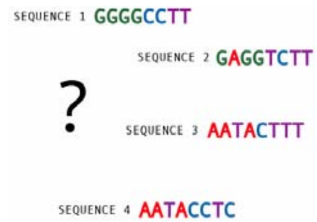
- Mutations happen when genetic material is copied
- Changes accumulate independently along each branch (within each new infectee)
- If one of these patients now infects two new victims, they inherit those changes

Phylogenetics interlude



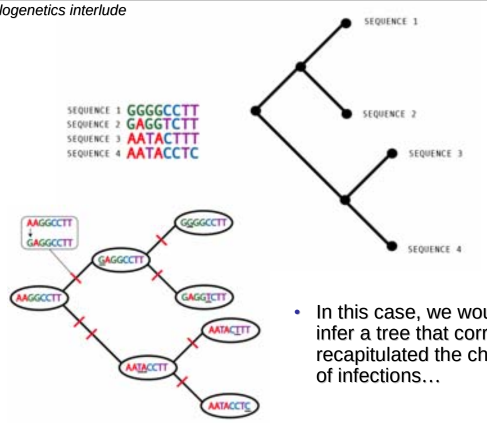
- Eventually, a series of branching events, plus mutations along each branch, lead to 4 current HIV infected patients
- Their viruses display genetic diversity that reflects their evolutionary history

Phylogenetics interlude



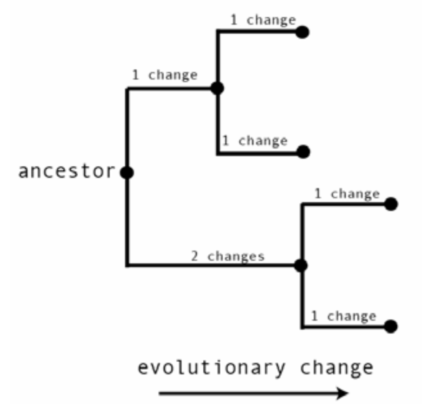
- Unfortunately, we almost never have access to that history
- What we **can** do, is go out into nature and sample genetic markers
- Then we work **backwards** to infer the most likely series of events that gave rise to what we observe

Phylogenetics interlude



- In this case, we would infer a tree that correctly recapitulated the chain of infections…

Phylogenetics interlude



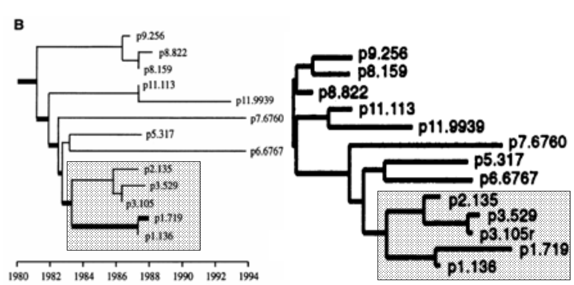
Phylogenetics interlude

Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis

Thomas Leitner*, David Escanilla†, Christer Franzén‡, Mathias Uhlén§, and Jan Albert*¶

TRUE TRANSMISSION HISTORY AND SAMPLING TIMES

INFERRED TREE FROM GENE SEQUENCES



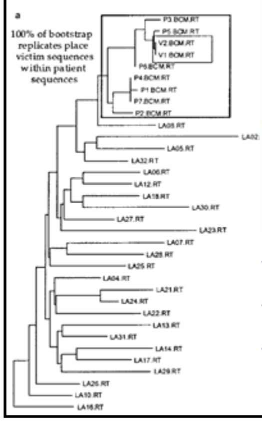
Phylogenetics interlude

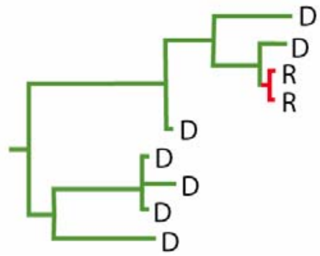
Molecular evidence of HIV-1 transmission in a criminal case

Michael L. Metzker*†, David P. Mindell‡, Xiao-Mei Liu*§, Roger G. Ptak¶‖, Richard A. Gibbs*, and David M. Hillis**

A gastroenterologist was convicted of attempted second-degree murder by injecting his former girlfriend with blood or blood-products obtained from an HIV type 1 (HIV-1)-infected patient under his care. Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in this case, representing the first use of phylogenetic analyses in a criminal court case in the United States.

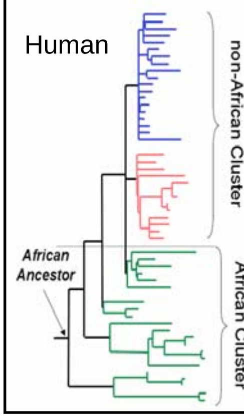
Phylogenetics interlude

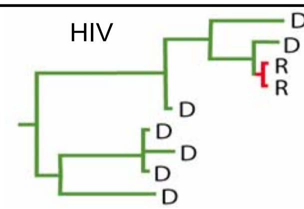




- Sequences recovered from the **victim**
- Sequences recovered from the **patient**
- Sequences also recovered from **other** HIV-positive individuals from the **same city**

Phylogenetics interlude

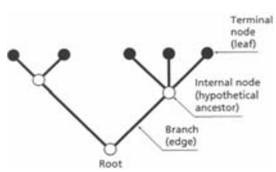




- The evolutionary pattern in this HIV phylogeny is just like the pattern in human mtDNA
- In both, we see a subpopulation that has recently emerged from a more diverse "source" population
- A few years of HIV evolution = 1 million years of human mtDNA evolution

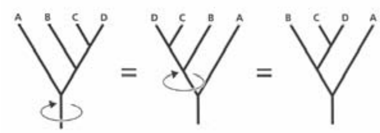
Tree terminology

Internal nodes represent hypothetical ancestors; the ancestor of all the sequences that comprise the tree is the **root** of the tree. Edges can also be classified as internal (leading to an internal node) or external (leading to an external node). Most methods try to estimate the amount of evolution that takes place between each node on the tree, which can be represented as **branch length**. The branching pattern of the tree is its **topology**.

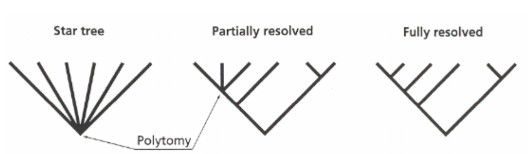


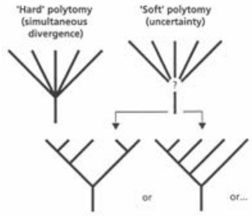
Tree styles

There are many different ways of drawing trees, so it is important to know whether these different ways actually reflect differences in the kind of tree, or whether they are simply stylistic conventions. Think of the tree as a mobile:



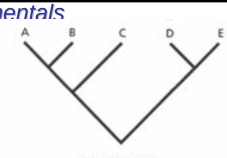
polytomies





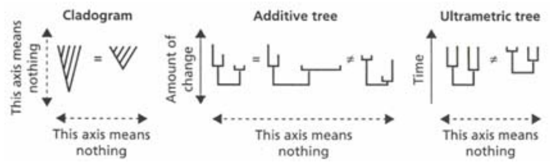
These polytomies can represent two different situations; first they may represent simultaneous divergence- all the descendants evolved at the same time (a '**hard**' polytomy); alternatively, they may indicate uncertainty about phylogenetic relationships (a '**soft**' polytomy).

Tree fundamentals



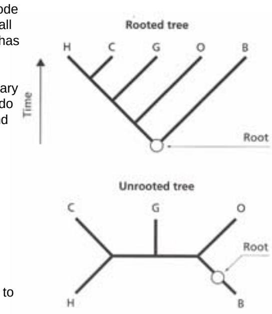
In phylogenetic software, trees are commonly represented in shorthand via parenthetical notation (especially important when loading constraint trees to test hypotheses).

Different kinds of trees can be used to depict different aspects of evolutionary history. The most basic tree is the **cladogram** which simply shows relative recency of common ancestry. **Additive trees** (phylograms) depict the amount of evolutionary change that has occurred along the different branches. **Ultrametric trees** (dendograms) depict the times of divergence.



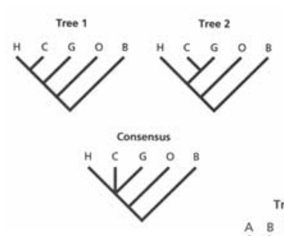
Rooted and unrooted trees

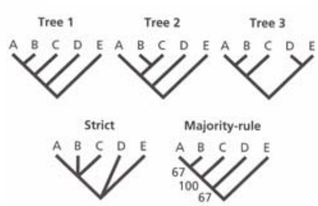
Cladograms and additive trees can either be rooted or unrooted. A **rooted** tree has a node identified as the root from which ultimately all other nodes descend, hence a rooted tree has direction. This direction corresponds to evolutionary time. **Unrooted** trees lack a root, and therefore do not specify evolutionary relationships in quite the same way. They do not allow the determination of ancestors and descendants.



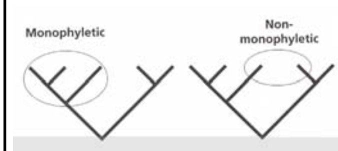
Here we have an unrooted tree for human, chimpanzee, gorilla, orang, and gibbon (B). The rooted tree (above) corresponds to the placement of the root on the branch leading to gibbon.

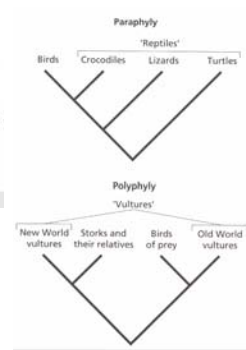
consensus trees





monophyletic clades



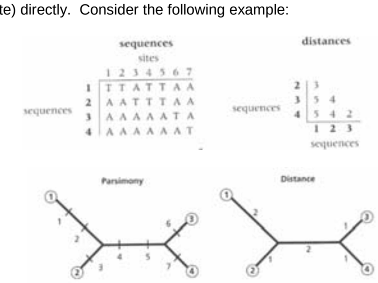


Inferring phylogenies

- All phylogeny reconstruction methods assume you start with a set of aligned sequences.
- The alignment is the statement of **homology**, that is shared ancestry from which historical inferences are made. The alignment, then, becomes critical to reconstructing phylogenies.
- In some cases, the alignment is trivial. In many cases it is not.

Inferring phylogenies

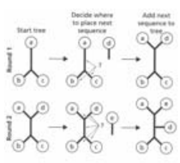
- There are two fundamental ways of treating data; as **distances** or as **discrete characters**.
- Distance methods first convert aligned sequences into a pairwise distance matrix, then input that matrix into a tree building method
- Discrete methods consider each nucleotide site (or some function of each site) directly. Consider the following example:

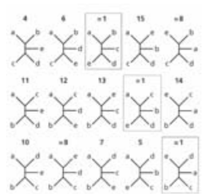


Inferring phylogenies

- There are also two fundamental ways of finding the "best" phylogenetic tree
- **Clustering** methods use some algorithm to cobble together a single tree
- **Optimality** methods survey all possible trees and compare how well they fit the data

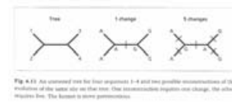
Clustering methods versus optimality methods





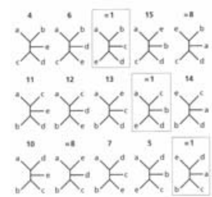
Phylogeny reconstruction: maximum parsimony

The data for maximum parsimony comprise individual nucleotide sites. For each site the goal is to reconstruct the evolution of that site on a tree subject to the constraint of invoking the fewest possible evolutionary changes.



In parsimony we are optimizing the total number of evolutionary changes on the tree or tree **length**. The tree length, then, is the sum of the number of changes at each site. So, if we have *k* sites, each with a length of *l*, then the length *L* of the tree is given by

$$L = \sum_{i=1}^{k} l_i$$



Phylogeny reconstruction: maximum likelihood

The method of maximum likelihood is a contribution of RA Fisher, who first investigated its properties in 1922.

Principle: evaluate all possible trees (topology and branch lengths) and substitution model parameters (TS/TV, base freq, rate heterogeneity etc.). These are the hypotheses. Choose the one that maximizes the likelihood of your data (the alignment)

Likelihood: Given that the coin you're tossing just gave you 15 heads out of 100 tosses, the likelihood that it is fair is very small.

Given the nature of molecular evolutionary data, where evolution has run just once, yielding one data set, maximum likelihood is a powerful framework--evaluate a bunch of different hypotheses to find the one most likely to have generated the observed data!

A non-biological example: coin tossing

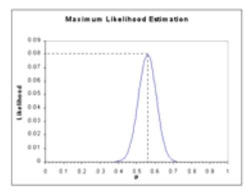
If the probability of an event X dependent on model parameters *p* is written

P (X | p)

then we would talk about the likelihood

L (p | X)

that is, the likelihood of *the parameters given the data*.

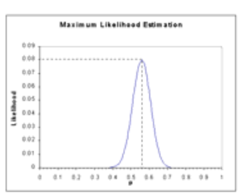


A non-biological example: coin tossing

Say we toss a coin 100 times and observe 56 heads and 44 tails. Instead of *assuming* that *p* is 0.5, we want to find the MLE for *p*. Then we want to ask whether or not this value differs significantly from 0.50.

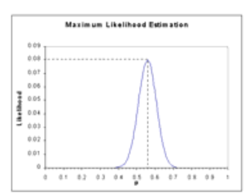
How do we do this? We find the value for *p* that makes the observed data most likely.

p	L
0.48	0.0222
0.50	0.0389
0.52	0.0581
0.54	0.0739
0.56	0.0801
0.58	0.0738
0.60	0.0576
0.62	0.0378

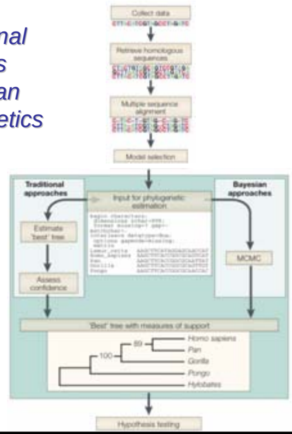


A non-biological example: coin tossing

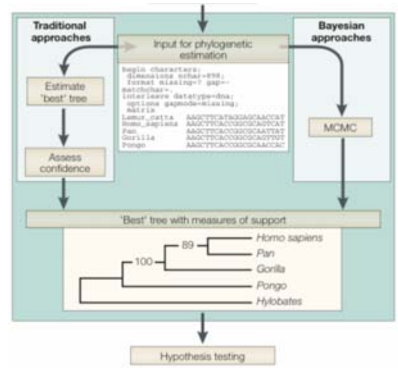
So why did we waste our time with the maximum likelihood method? In such a simple case as this, nobody would use maximum likelihood estimation to evaluate *p*. But not all problems are this simple!



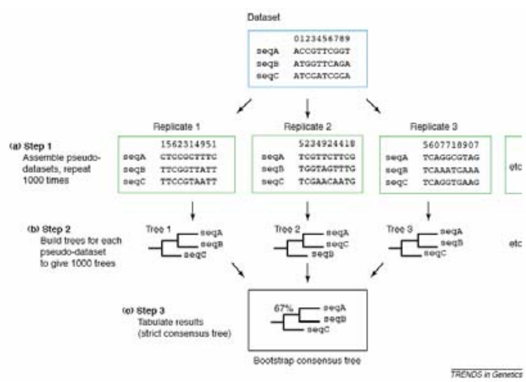
Traditional versus Bayesian phylogenetics

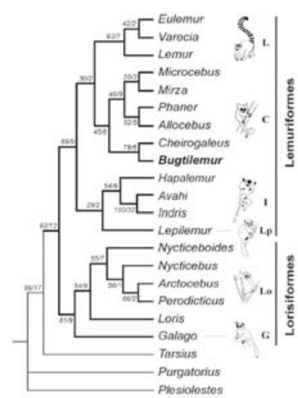


Traditional versus Bayesian phylogenetics



Estimating confidence: Bootstrapping trees





L. Marivaux *et al.* A fossil lemur from the Oligocene of Pakistan. *Science* **294**, 587 (2001).
[The above is only a portion of the figure]

6) Lemuriformes are currently restricted to Madagascar, whereas Lorisiformes are found in Africa and Asia but not Madagascar, and *Tarsius* is Asian. The tree above was generated in order to assess the relationship of a fossil, *Bugtilemur*, found in 30 million year old deposits in Pakistan. Each branch of the tree has been annotated with two numbers, the first of which is the bootstrap percentage, a measure of support. In order to hold that *Bugtilemur* is more closely related to Lorisiformes than to Lemuriformes what is the minimum number of branches, with what bootstrap support, that would need to be incorrect?

a) 1: 92%
b) 2: 78%, 69%
c) 4: 78%, 45%, 30%, 69%
d) 4: 78%, 45%, 30%, 29%

Phylogeny reconstruction: Bayesian methods

But first, Markov Chain Monte Carlo (MCMC)…

A method for integrating complex high-dimensional spaces. In other words, it involves traveling through a set of solutions such that every point is visited at a frequency equal to its likelihood. Basically it's hill climbing, but can head downhill sometimes too--a wandering among states that is biased toward better states.

This allows you to sample from a ridiculously huge hypothesis space. The chain spends most of its time in higher probability regions.

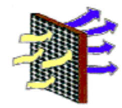
Phylogeny reconstruction: Bayesian methods

The most widely used MCMC method is the Metropolis algorithm:

1. Start at some tree.
2. Pick a neighboring tree in hypothesis space. Call this the proposal.
3. Compute the ratio (R) of the probabilities of the proposed new tree and the old tree.
4. If R >=1, accept the new tree as the current tree.
5. If R < 1, draw a number between 0 and 1. If this number is less than R, accept the new tree as the current tree.
6. Otherwise, reject the new tree and keep the old tree.
7. Return to step 2.

This algorithm never terminates. It is a Markov chain because it is a random process in which the next change depends only on the current state.

Phylogeny reconstruction: Bayesian methods



$$\mathrm{Prob}(H \mid D) = \frac{\mathrm{Prob}(H)\ \mathrm{Prob}(D \mid H)}{\sum_H \ \mathrm{Prob}(H)\ \mathrm{Prob}(D \mid H)}$$

Phylogeny reconstruction: Bayesian methods

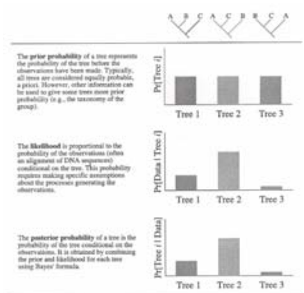
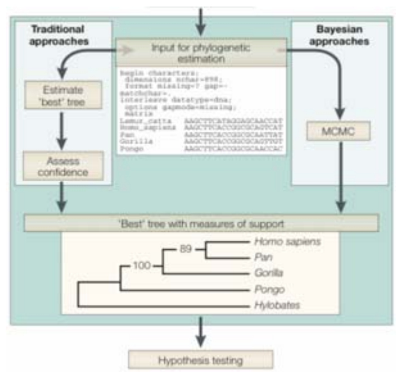


Fig. 1. The main components of a Bayesian analysis.

Traditional versus Bayesian phylogenetics



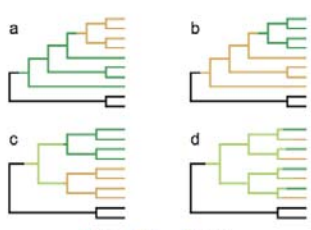
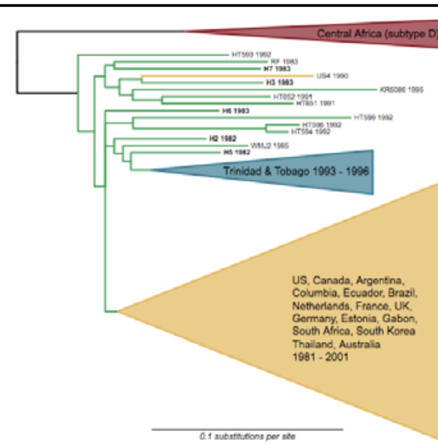
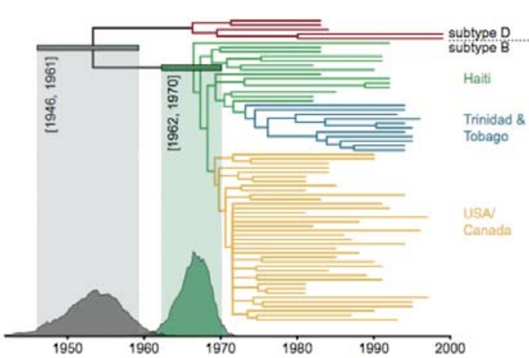


FIG. 1. Schematic diagrams of phylogenetic patterns expected under several hypotheses for the origin and spread of subtype B. (a) If the virus reached Haiti first, then Haitian HIV-1 sequences are expected to branch off from the root part of the subtype B subtree before sequences from elsewhere. Alternatively, (b) the Haitian epidemic could have been imported from the US; (c) both the US and Haitian epidemics could have begun effectively simultaneously then remained distinct; or (d) high levels of migration could have obscured where the virus arrived first.





Why the conflict?

Gene trees vs species trees 1:duplication

Gene trees don't always match species trees. Why is that?

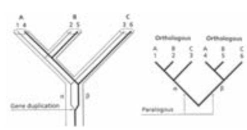
Some genes belong to multigene families which have arisen from duplication events

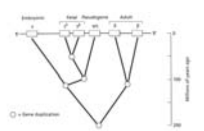
Gene duplication means that some gene pairs are *orthologous* and others *paralogous*

Orthologous genes: MRCA did not undergo duplication

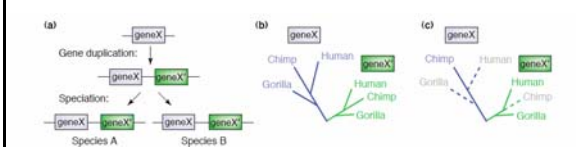
Paralogous genes: MRCA duplicated

1,2,3: ((A,B),C)	4,2,3: (A,(B,C)) *X*
1,2,6: ((A,B),C)	4,2,6: ((A,C),B) *X*
1,5,3: ((A,C),B) *X*	4,5,3: ((A,B),C)
1,5,6: (A,(B,C)) *X*	4,5,6: ((A,B),C)





Gene trees vs species trees 1:duplication



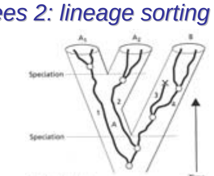
Gene trees vs species trees 2: lineage sorting

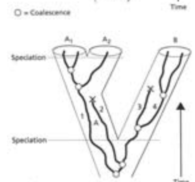
Even restricting analysis to orthologous genes cannot, in principle, guarantee that gene tree = species tree because of ancestral polymorphism and differential survival of alleles (lineage sorting)

At speciation, lineage A was polymorphic, with one allele more closely related to lineage B's allele than to the other lineage A one.

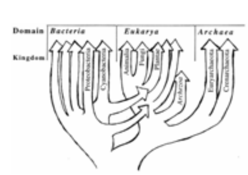
If the polymorphism persists until a subsequent speciation event, gene tree will support $((A_2,B),A)$.

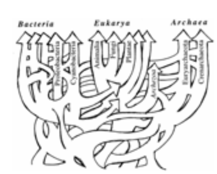
If coalescence times tend to be greater than the intervals between speciation, things will be messy.

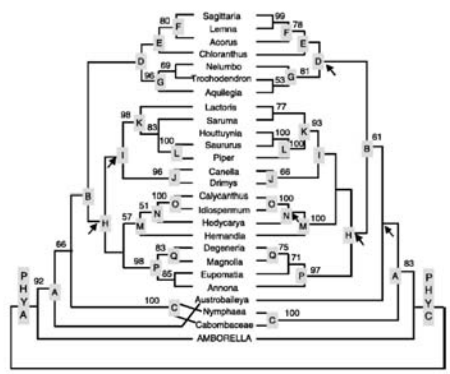




Gene trees vs species trees 3: horizontal transfer







S. Mathews, M. J. Donoghue. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**, 947 (1999).

1) The figure above shows the phylogeny estimated for a sample of flowering plants (angiosperms) from *PHYTOCHROME A* and *PHYTOCHROME C*, a pair of genes that duplicated prior to the origin of the angiosperms. Which of the following sets of taxa constitute a clade (=monophyletic group) on one gene tree but not on the other?

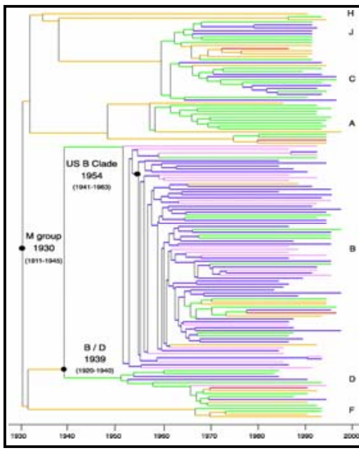
a) *Degeneria-Magnolia-Eupomatia*
b) All angiosperms except *Amborella*
c) *Austrobaileya-Nymphaea*-Cabombaceae
d) *Nelumbo-Trochodendron-Aquilegia*

What can you do with trees beyond simply inferring relatedness? (molecular clocks)

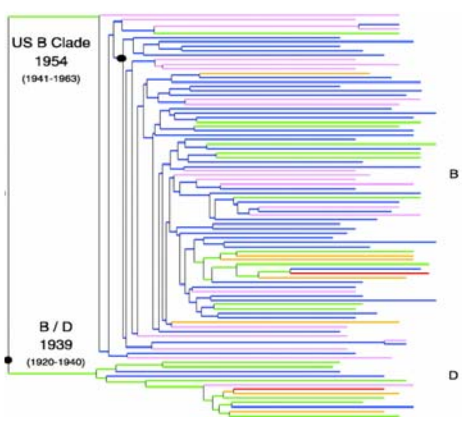
Timing the Ancestor of the HIV-1 Pandemic Strains

B. Korber,[1,2*†] M. Muldoon,[2,3] J. Theiler,[1] F. Gao,[4] R. Gupta,[1] A. Lapedes,[1,2] B. H. Hahn,[4] S. Wolinsky,[5] T. Bhattacharya[1†]

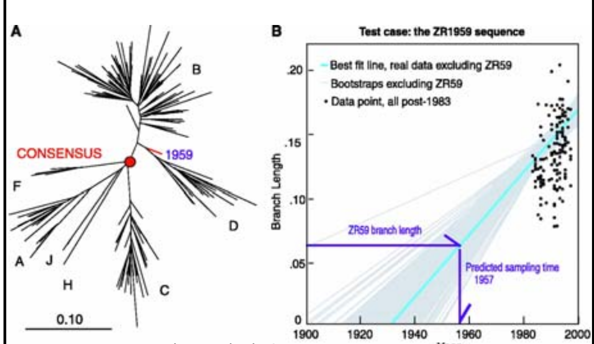
SCIENCE VOL 288 9 JUNE 2000



Korber et al. Fig 4



Korber et al. Fig 4



Korber et al. Fig 2

What can you do with trees beyond simply inferring relatedness? (genome evolution)

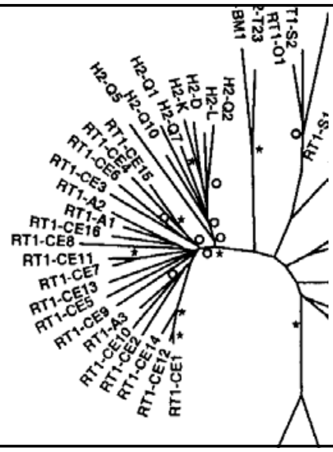
Genome Research

The Genomic Sequence and Comparative Analysis of the Rat Major Histocompatibility Complex

Peter Hurt,[1] Lutz Walter,[2] Ralf Sudbrak,[1] Sven Klages,[1] Ines Müller,[1,3] Takashi Shiina,[4] Hidetoshi Inoko,[4] Hans Lehrach,[1] Eberhard Günther,[2,5] Richard Reinhardt,[1,5] and Heinz Himmelbauer[1,6]

- MHC genes play important roles in immunity
- MHC class I presents antigen from viruses to killer T cells
- These genes are in a brisk arms race with pathogens

Hurt et al. Fig 3



- Phylogenetic trees (in this case a distance/algorithm method was used) can reveal expansion of genes within species
- Here, MHC class I genes show species specific amplification since the split between mouse and rat

Hurt et al. Fig 3

What can you do with trees beyond simply inferring relatedness? (ancestral reconstruction)

Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 1569–1573, February 1994
Evolution

Molecular resurrection of an extinct ancestral promoter for mouse L1

(sequence evolution)

NILS B. ADEY, TRYGVE O. TOLLEFSBOL, ANDREW B. SPARKS, MARSHALL HALL EDGELL, AND CLYDE A. HUTCHISON III*

What can you do with trees beyond simply inferring relatedness?

- Adey et al. (1994) resurrected an extinct ancestral promotor for a subfamily of retroposons that dispersed in the mouse genome several million years ago
- The retroposons are no longer transcriptionally or transpositionally active
- They hypothesized that the promoter may have accumulated deleterious mutations, used extant sequences to infer the ancestor
- Chemically synthesized it and found it reawakened the retroposon

What can you do with trees beyond simply inferring relatedness?

Molecular Biology and Evolution 19:1483-1489 (2002)

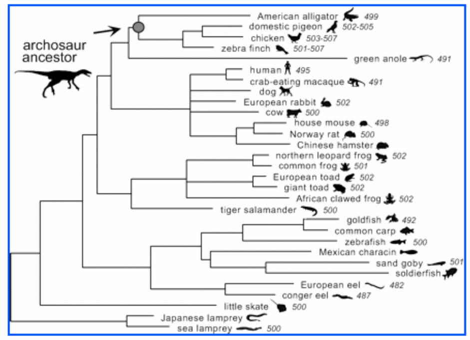
Recreating a Functional Ancestral Archosaur Visual Pigment

Belinda S. W. Chang*, Karolina Jönsson*, Manija A. Kazmi*, Michael J. Donoghue† and Thomas P. Sakmar*

What can you do with trees beyond simply inferring relatedness?

•Chang et al. (2002) used maximum likelihood phylogenetic ancestral reconstruction methods to recreate a putative ancestral archosaur visual pigment (ca. 240 mya)

What can you do with trees beyond simply inferring relatedness?

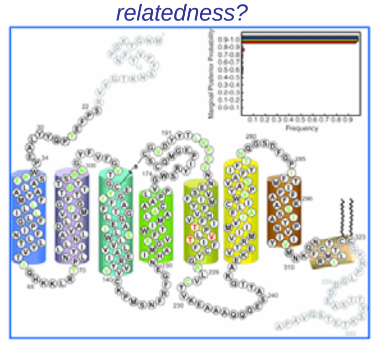


Chang et al. Fig 1

What can you do with trees beyond simply inferring relatedness?

•To determine if these ancestral pigments would be functionally active, the corresponding genes were chemically synthesized and then expressed in tissue culture

What can you do with trees beyond simply inferring relatedness?

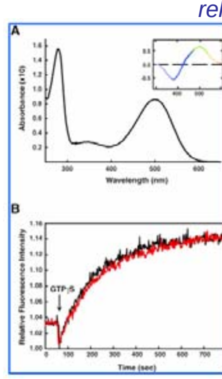


Chang et al. Fig 2

What can you do with trees beyond simply inferring relatedness?

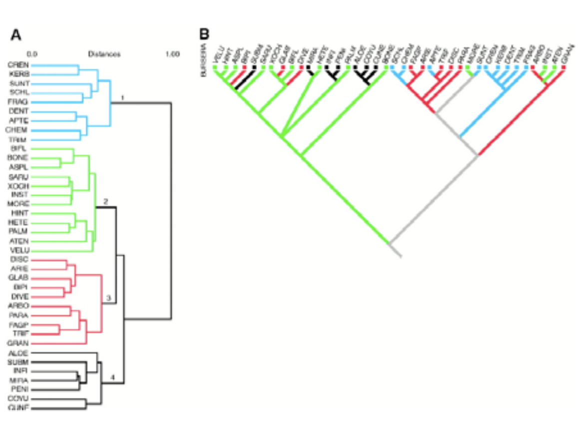
•The expressed artificial genes were all found to yield stable photoactive pigments with max values of about 508 nm, which is slightly redshifted relative to that of extant vertebrate pigments.

What can you do with trees beyond simply inferring relatedness?



Chang et al. Fig 3

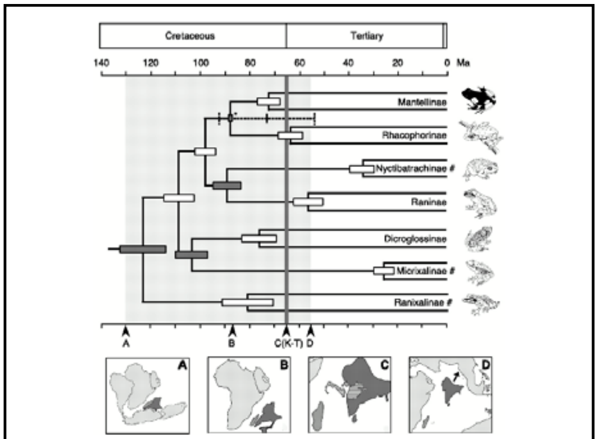
•What might you speculate about the behavior of the ancestral archosaur based on these results?



J. X. Becerra. Insects on plants: macroevolutionary chemical trends in host use. *Science* **276**, 253 (1997).

2) The dendrogram on the left clusters plant species by chemical similarity; each of the four main chemical groups is indicated with a different color. This tree does not depict descent relationships, just degree of chemical similarity. On the right, the evolution of these chemical types is reconstructed on a phylogeny of the plants (this does depict inferred evolutionary relationships). The colors correspond to the chemical groups on the left, and the gray branches indicate uncertainty in character reconstruction. What does a comparison of these two figures tell us about the evolution of plant secondary chemistry?

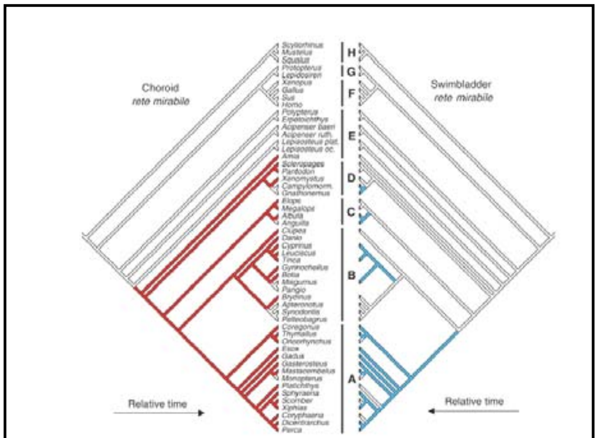
a) The four groups of chemically similar species each constitutes a distinct evolutionary lineage
b) The group colored "black" has the most advanced chemical defenses
c) The red (3) and blue (1) chemical groups are most distantly related
d) The chemical groups have each been gained and/or lost multiple times in evolution



F. Bossuyt, M. C. Milinkovitch. Amphibians as indicators of early tertiary "out-of-India" dispersal of vertebrates. *Science* **292**, 93 (2001).

3) This tree depicts inferred relationships among some major frog groups with branches drawn proportional to absolute time. Error bars on internal nodes depict confidence intervals on the dates of estimated nodes. Assuming this tree and the associated ages are correct which of the following statements is true?

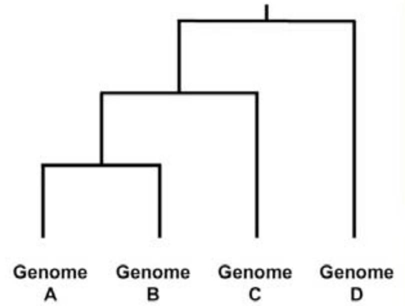
a) No individual living before 70 million years ago is an ancestor of Raninae
b) Raninae and Dicroglossinae shared a common ancestor about 75 million years ago
c) The divergence of Raninae and Nyctibatrachinae occurred more recently than the 85 million year old separation of India from Madagascar
d) The last common ancestor of Micrixalinae and Dicroglossinae lived before India and Madagascar separated (85 million years ago)



M. Berenbrink, P. Koldkjaer, O. Kepp, A. R. Cossins. Evolution of oxygen secretion in fishes and the emergence of a complex physiological system. *Science* **307**, 1752 (2005).

4) *Retia mirabilia* (sing. *rete mirabile*) are vascular bundles that allow fish to secrete O_2. In the above figure, red branches indicate lineages with choroid *retia*, blue branches indicate those with swimbladder *retia*, and white branches indicate absence of *retia*. Assuming the phylogeny and character evolution have been accurately inferred, we can see that:

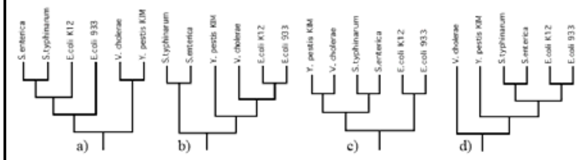
a) Swimbladder *retia* predate choroid *retia*
b) Gains of swimbladder *retia* primarily took place in lineages that already had choroid *retia*.
c) Loss of choroid *retia* causes gain of swimbladder *retia*
d) Choroid *retia* have been gained more often than swimbladder *retia*

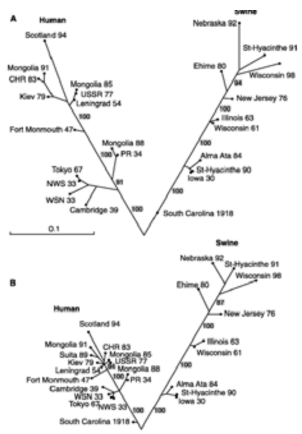


Genome A	Genome B	Genome C	Genome D
E. coli 933	*E. coli K12*	*S. typhimurium*	*Y. pestis KIM*
S. enterica	*S. typhimurium*	*E. coli K12*	*Y. pestis KIM*
E. coli K12	*S. typhimurium*	*Y. pestis KIM*	*V. cholerae*

V. Daubin, N. A. Moran, H. Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science* **301,** 829 (2003).
[The above is only a portion of the figure].

5) Each row in the table above lists a set of four bacterial taxa whose relationship follows the topology shown. Thus each row can be read as a four-taxon tree. Which of the four trees below is compatible with the information in the three rows of the table?





M. J. Gibbs, J. S. Armstrong, A. J. Gibbs. Recombination in the hemaglutinin gene of the 1918 "Spanish flu". *Science* **293**, 1842 (2001).

7) The trees show the phylogeny estimated for the same mammalian H1 influenzas based on portions of the haemaglutin genes. Tree A was estimated from bases 310-870 whereas tree B was estimated from bases 1070-1650. Under the assumption that these trees are <u>correctly rooted</u>, how does this result support the inference that the South Carolina 1918 strain (the only exemplar from the 1918 flu pandemic) arose from recombination between a human and a swine virus?

a) The fact that the South Carolina 1918 strain is near the root of both trees suggests that it is of mixed identity
b) The fact that the South Carolina 1918 strain is more closely related to the swine strains in tree A but to the human strains in tree B
c) The fact that the South Carolina 1918 strain is a direct ancestor of all the swine strains in tree A but was isolated from a human
d) The fact that the South Carolina 1918 strain is more closely related to the Iowa 30 swine strain than to the Scotland 94 human strain in both trees

Worobey *et al.* (2002) *Science.* **296:** 211.