Fundamentals of Evolution

Session 7 - 9/24/19

Phylogenomics and the Genome

Quiz recap

- Cambrian explosion occurred ~540 Ma; sudden appearance of most extant animal phyla. Likely associated with environmental changes (e.g., rising oxygen) that facilitated the evolution of hard bodies or shells that fossilize well.

- We know from a few older formations that preserved soft tissues that some of these phyla existed much earlier. Phylogenetic evidence suggests their divergences occur earlier than the Cambrian as well.

Quiz recap

- Sepkoski's Curve demonstrated that the diversity of life has changed dramatically through time, mostly increasing, but with few sharp dips and plateaus.
- Three evolutionary fauna with correlated rise and fall in diversity.
- Recovery from mass extinctions occurs quickly, but not composing the same lineages.
- Many biases exist: older fossils are lost; different times have different size areas of continental shelf (good fossil area); these areas may be tropical in some periods, temperate in others; plate tectonics affect climate, area, and biogeography.

Quiz recap

- Statistical models allow us to better model *homoplasy* (unobserved changes) as well as *rate variation.* This reduces the chance of errors due to long-branch attraction, and provides more accurate estimates of branch lengths.

- We can learn about evolution by estimating model parameters (e.g., branch lengths, rate parameters, ancestral states).

- Likelihood provides a *score* that can be used to compare hypotheses (*i.e., how much better is this tree than that tree?)*

Quiz recap

- A Markov process model describes the probability of discrete character changes (e.g., between DNA states) over a period of time (e.g., branch length) given a matrix of rate parameters.

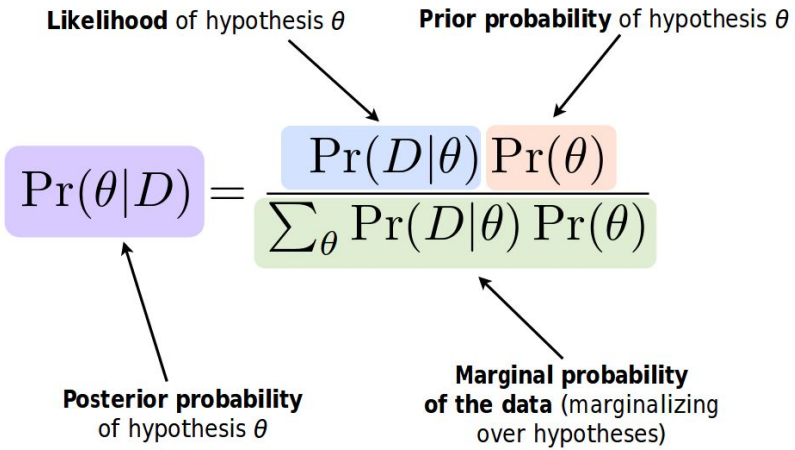
- From this, a likelihood equation is derived that describes the probability that the *observed data (sequence alignment)* could be produced by the *proposed model (tree & rate matrix)* given a set of *parameters (branch lengths & rates).*

- We use heuristic methods to search alternative tree topologies.

- Each site in a sequence alignment is treated independently.

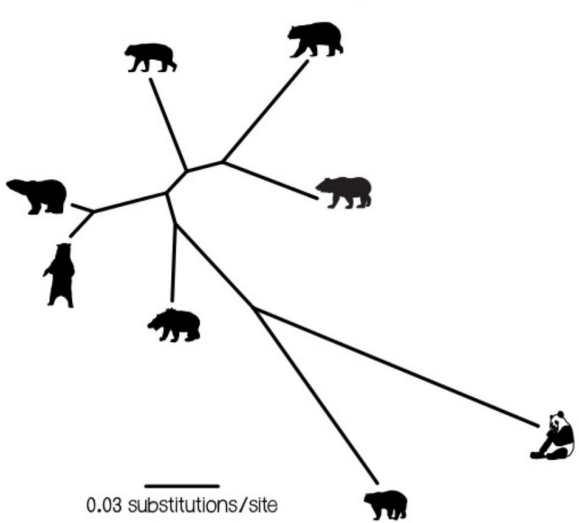
Recap of last session

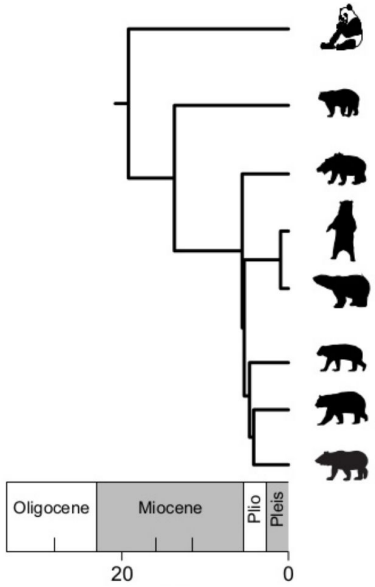
Bayesian statistics asks “what is the probability of my hypothesis given the data?” by incorporating our prior belief as probability.



Why is Bayesian analysis useful for phylogenetics?

Phylogenies with branch lengths in units of time provide more information than unrooted trees with branch lengths in units of rate*time (substitutions/site).

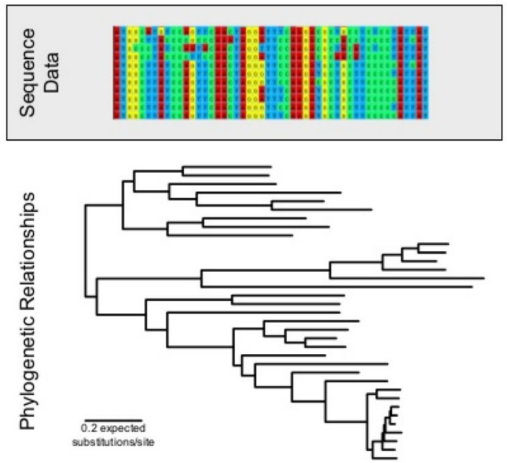




Sequence data provide information about **branch lengths**

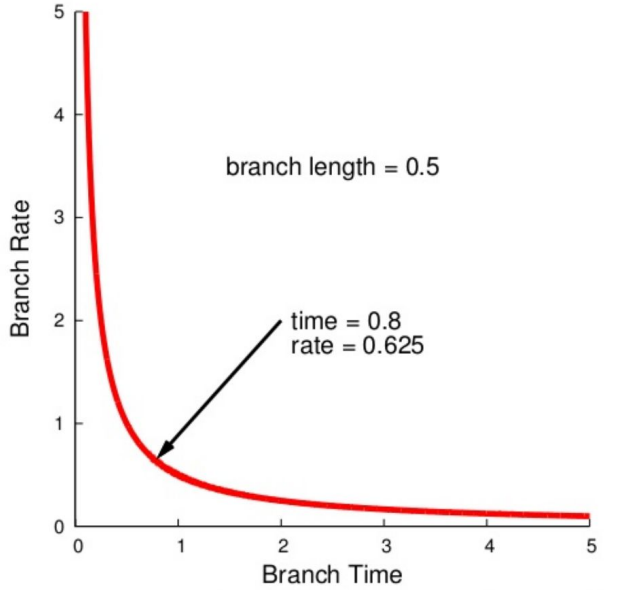
In units of **the expected # of substitutions per site**

branch length = rate × time



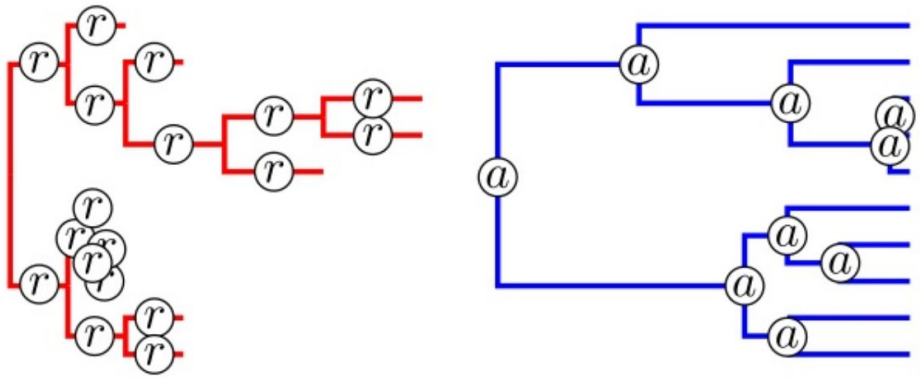
The sequence data provide information about branch length

for any possible rate, there's a time that fits the branch length perfectly



(figure based on Thorne & Kishino, 2005)

Methods for dating species divergences estimate the substitution rate and time separately



length = rate

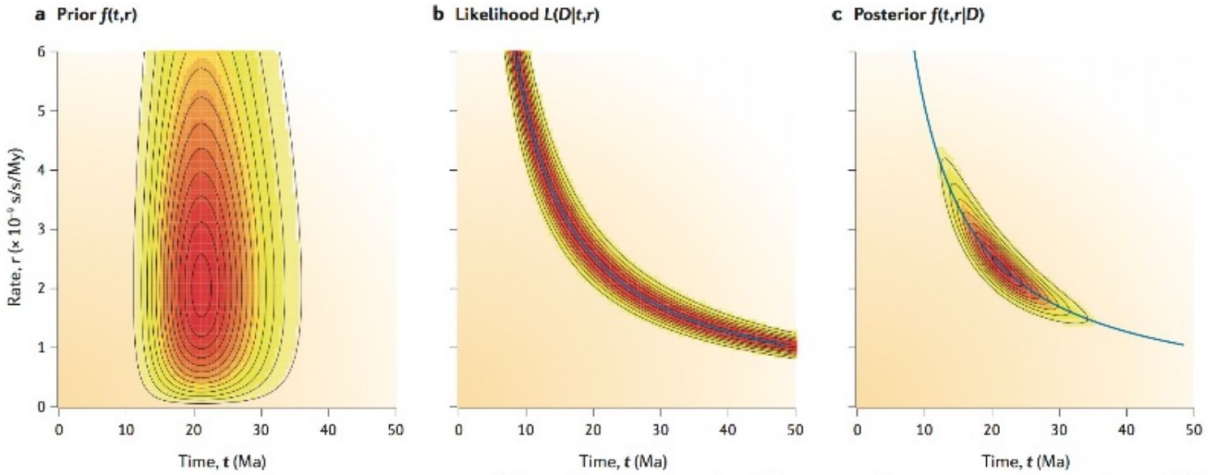
length = time

$$\boldsymbol{\mathcal{R}} = (r_1, r_2, r_3, \ldots, r_{2N-2})$$

$$\boldsymbol{\mathcal{A}} = (a_1, a_2, a_3, \ldots, a_{N-1})$$

$$N = \text{number of tips}$$

Methods for dating species divergences estimate the substitution rate and time separately

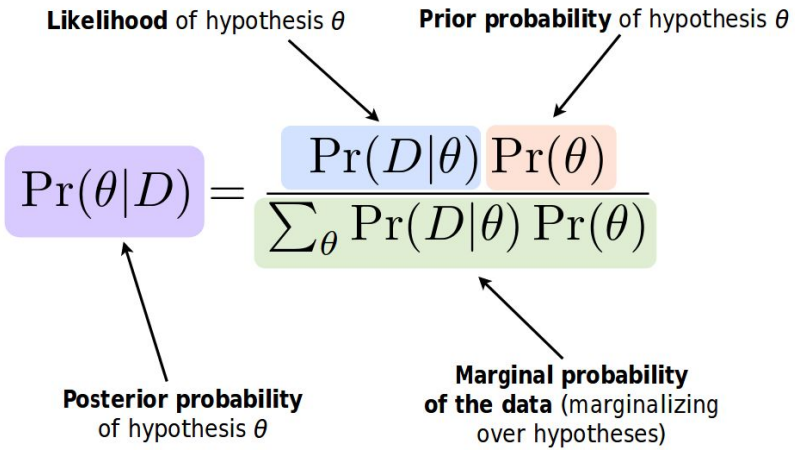


(dos Reis et al. *Nature Reviews Genetics*, 2016)

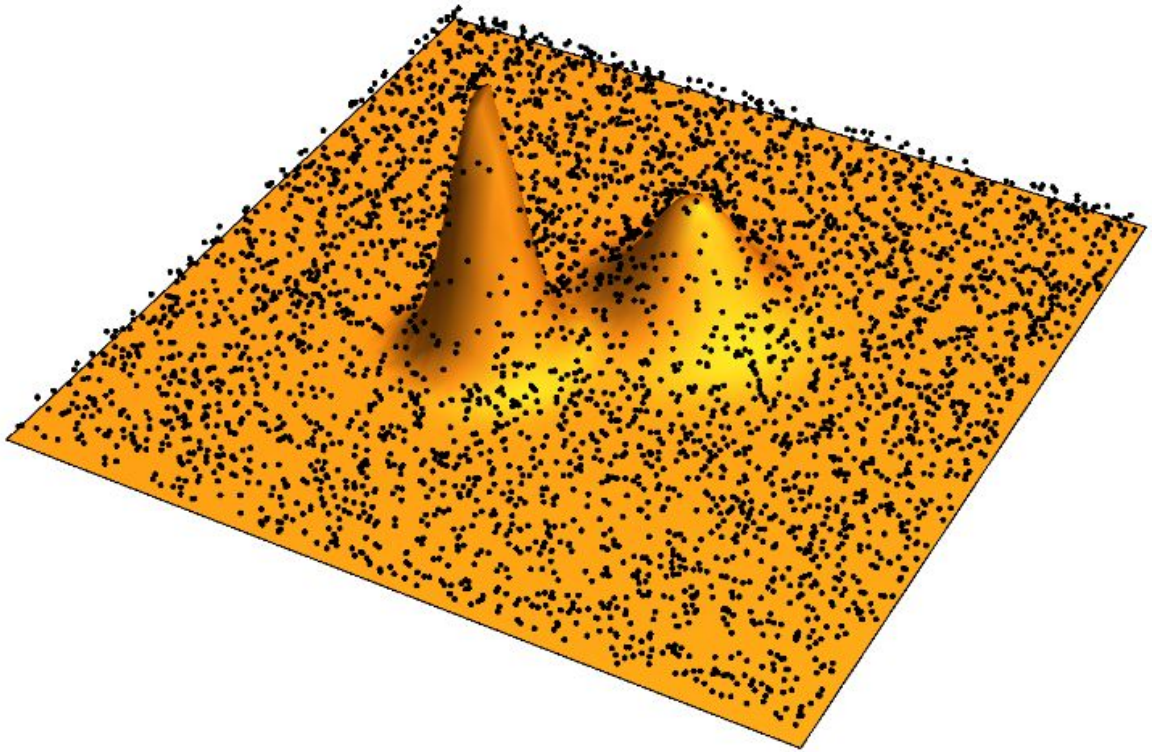
Tree-time priors for molecular phylogenies are only informative on a **relative** time scale

Recap of last session

Bayesian statistics asks “what is the probability of my hypothesis given the data?” by incorporating our prior belief as probability.

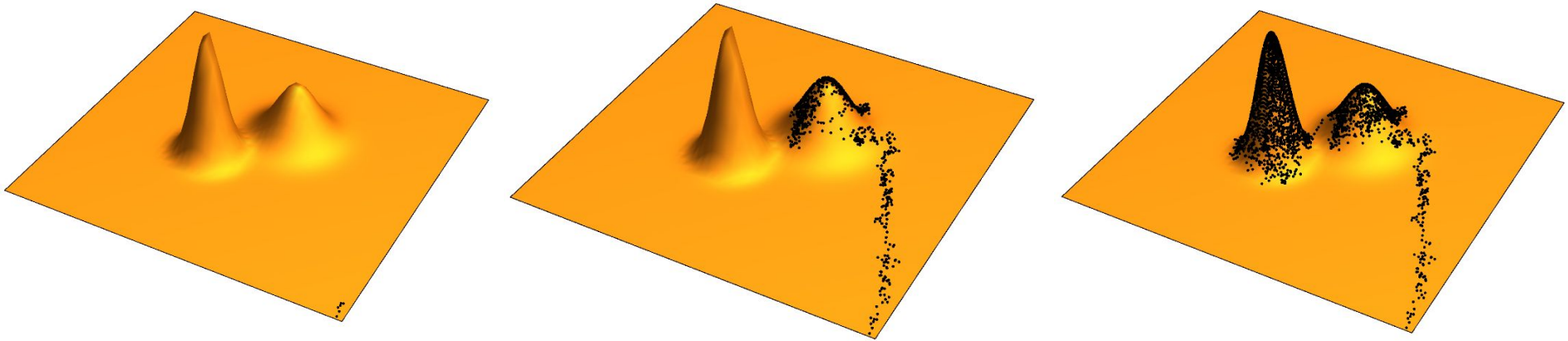


Naive integration approach

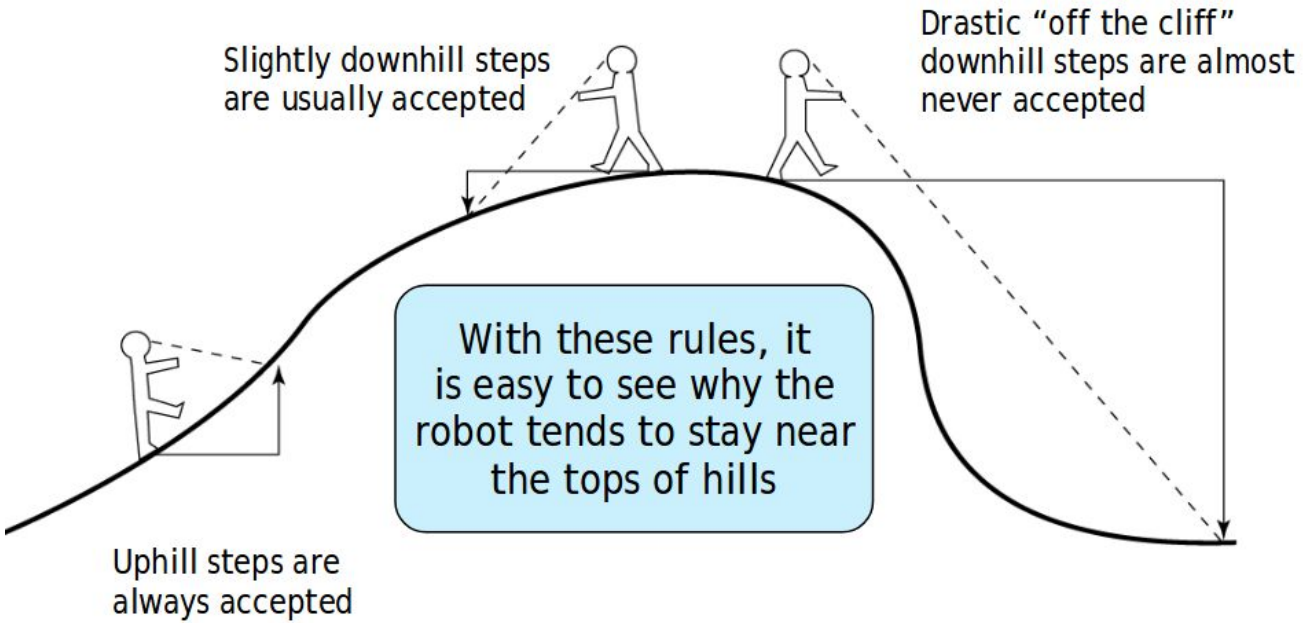


Markov chain Monte-Carlo (MCMC)

Heuristic method of integrating across marginal probabilities. Mechanistic algorithm to search parameter space where the proportion of steps spent in any part of search space reflects the posterior probability support for that parameter. The result is a posterior probability distribution.



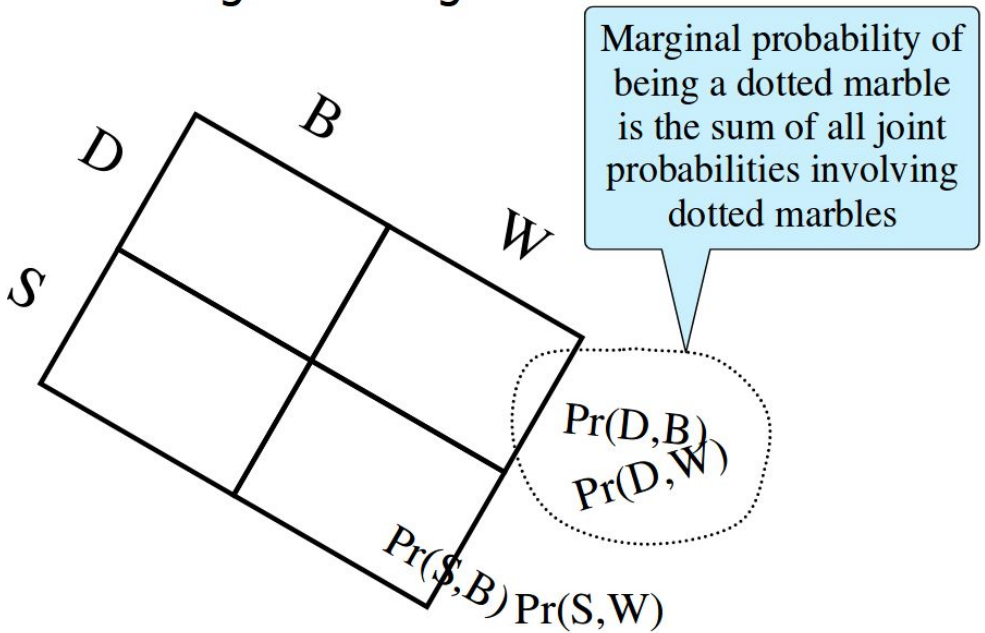
MCMC robot's rules



$$f(\mathcal{R}, \mathcal{A}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s \mid D) = \frac{f(D \mid \mathcal{R}, \mathcal{A}, \theta_s)\, f(\mathcal{R} \mid \theta_{\mathcal{R}})\, f(\mathcal{A} \mid \theta_{\mathcal{A}})\, f(\theta_s)}{f(D)}$$

$f(D \mid \mathcal{R}, \mathcal{A}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s)$	Likelihood
$f(\mathcal{R} \mid \theta_{\mathcal{R}})$	Prior on rates
$f(\mathcal{A} \mid \theta_{\mathcal{A}})$	Prior on node ages
$f(\theta_s)$	Prior on substitution parameters
$f(D)$	Marginal probability of the data

Marginalizing over colors



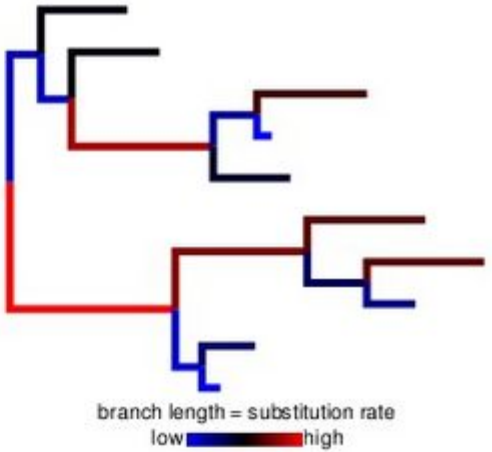

Trees


Trees
rates

- **Global clock** (Zuckerkandl & Pauling, 1962)
- **Local clocks** (Hasegawa, Kishino & Yano 1989; Kishino & Hasegawa 1990; Yoder & Yang 2000; Yang & Yoder 2003, Drummond and Suchard 2010)
- **Punctuated rate change model** (Huelsenbeck, Larget and Swofford 2000)
- **Log-normally distributed autocorrelated rates** (Thorne, Kishino & Painter 1998; Kishino, Thorne & Bruno 2001; Thorne & Kishino 2002)
- **Uncorrelated/independent rates models** (Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)
- **Mixture models on branch rates** (Heath, Holder, Huelsenbeck 2012)

Lineage-specific rates are uncorrelated when the rate assigned to each branch is independently drawn from an underlying distribution

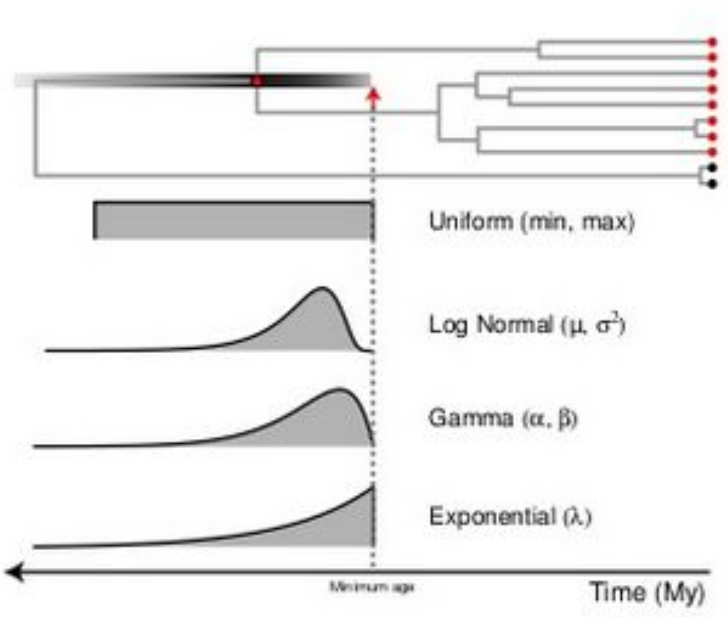
(Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)

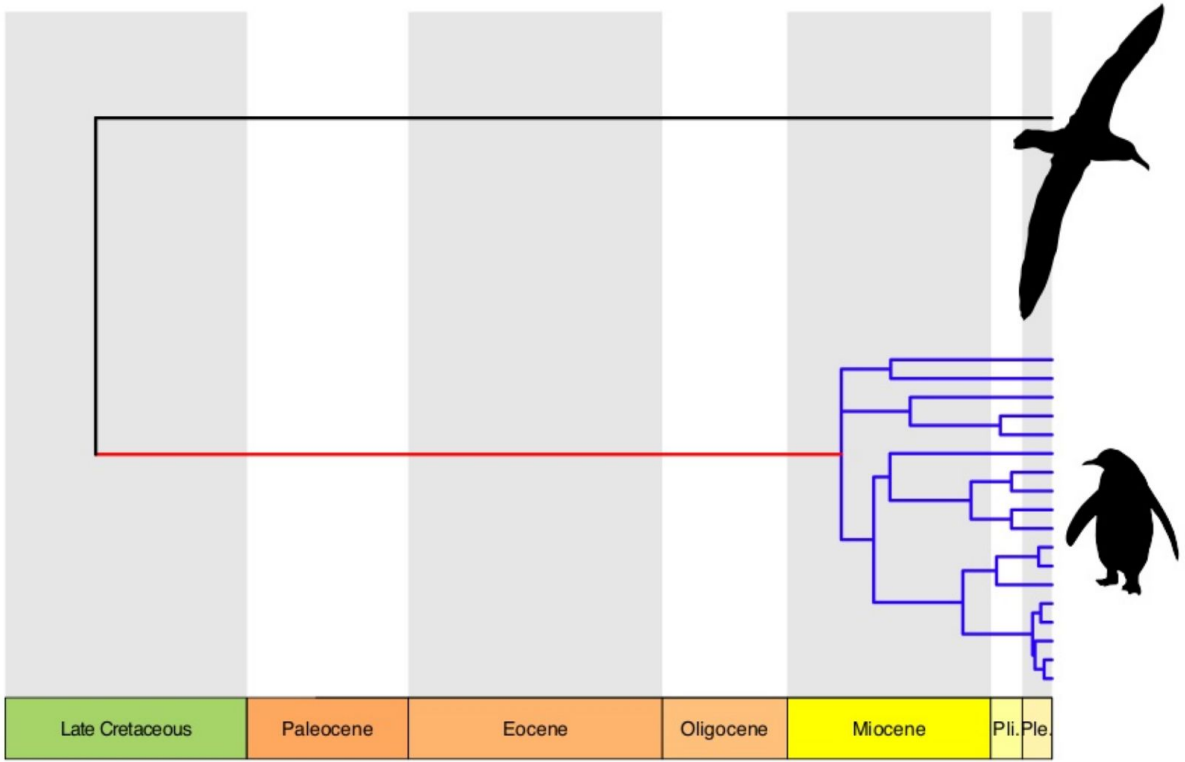


Common practice in Bayesian divergence-time estimation:

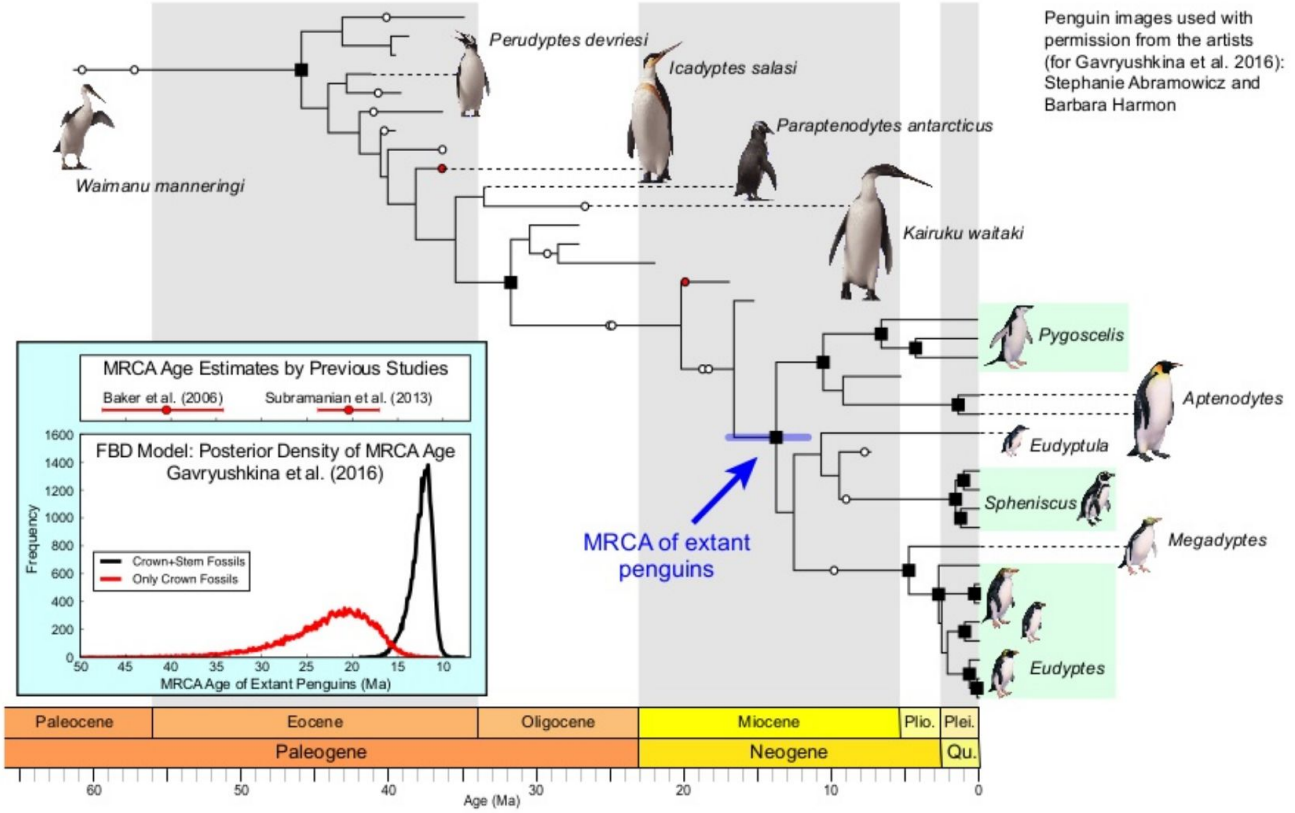
Parametric distributions are typically off-set by the age of the oldest fossil assigned to a clade

These prior densities do not (necessarily) require specification of maximum bounds

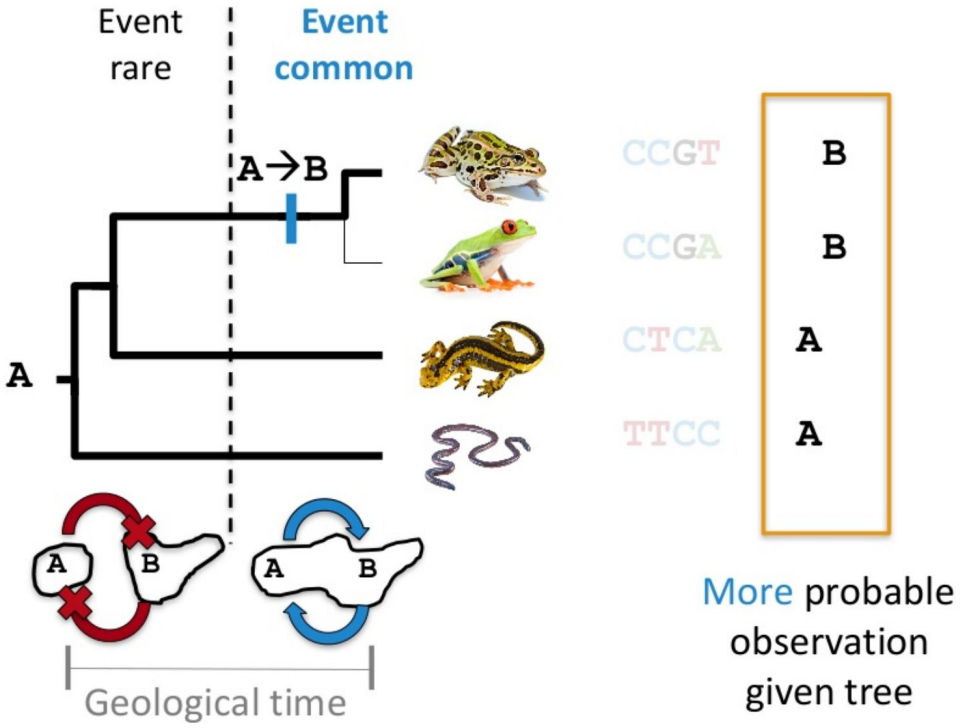



Late Cretaceous
Paleocene
Eocene
Oligocene
Miocene
Pli.
Ple.

Incorporating both fossils and DNA sequences, and informed priors on the fossil placements, Gavryushkina et al. (2016) found the crown age of extant penguins is much younger than previously thought.



Even without fossils, time-informed priors



Phylodynamics

- The study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies.

- Bayesian phylogenetics is highly important because *rate* varies dramatically during viral outbreaks

Summary of Bayesian phylogenetics

- Broadly applicable statistical framework that allows one to combine data from many different sources through defining priors.

- In practice, often used for dated phylogenies because with priors on ages or rates you can better differentiate age from rate (which cannot be done in ML)

- However, it can be rather slow (MCMC search)
- And if you define too strict of priors and your data are not very informative then your results may just return what you put in. Requires careful testing/refining.

Article Discussion: Rates of Molecular Evolution

Discuss the paper in small groups, see if you can answer each others' questions.

Come up with 2-3 questions about (1) things you still do not understand about the paper; or (2) new questions raised by the paper.

Large-scale phylogenetics

- Increasingly, phylogenetic and phylogenomics is a field of informatics, or *data science, and computer science.*

- Data archiving and mining. Researchers focus on specific groups and over time accumulate enough data to span deeper and deeper in time.

- Methods for combining knowledge and minimizing the need to optimization + tree search.

Large-scale phylogenetics

- Super trees:
- Inferring large trees is difficult and time consuming, it is easier to join together smaller trees. Several techniques.
- The largest phylogenies that we have are all supertrees.

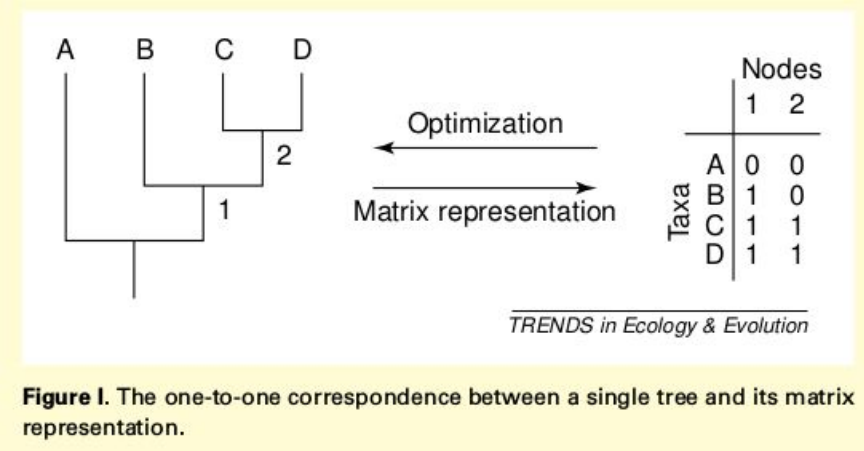


Figure I. The one-to-one correspondence between a single tree and its matrix representation.

Large-scale phylogenetics

- Supermatrices:
- Around the early 2000s common markers were discovered that could be sequenced reliably across many organisms, which made it possible to combine their data into larger analyses. Faster inference methods developed.
- Hundreds of taxa sequenced at one or more of the same genes.

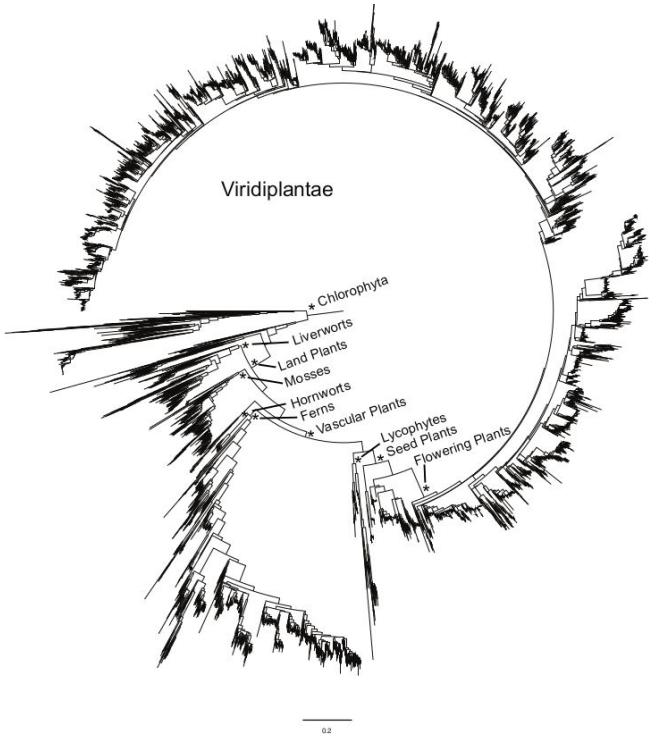
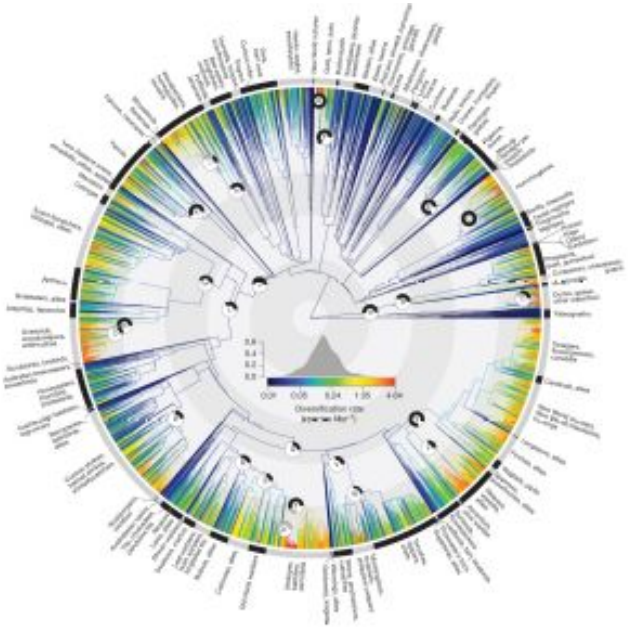


Figure 3
Maximum-likelihood phylogeny for 13,533 species of green plants based on *rbcL* DNA sequences. The data matrix was constructed using the mega-phylogeny method; major clades are labeled and denoted with a star.

Large-scale phylogenetics

- Time-scaled megaphylogenies:
- Bayesian relaxed clock analysis on a reduced set of taxa to infer the backbone.
- Many smaller Bayesian relaxed clock analyses of subclades are added to the estimated backbone.



W Jetz *et al. Nature* **000**, 1-5 (2012) doi:10.1038/nature11631

Summary of large-scale phylogenetics

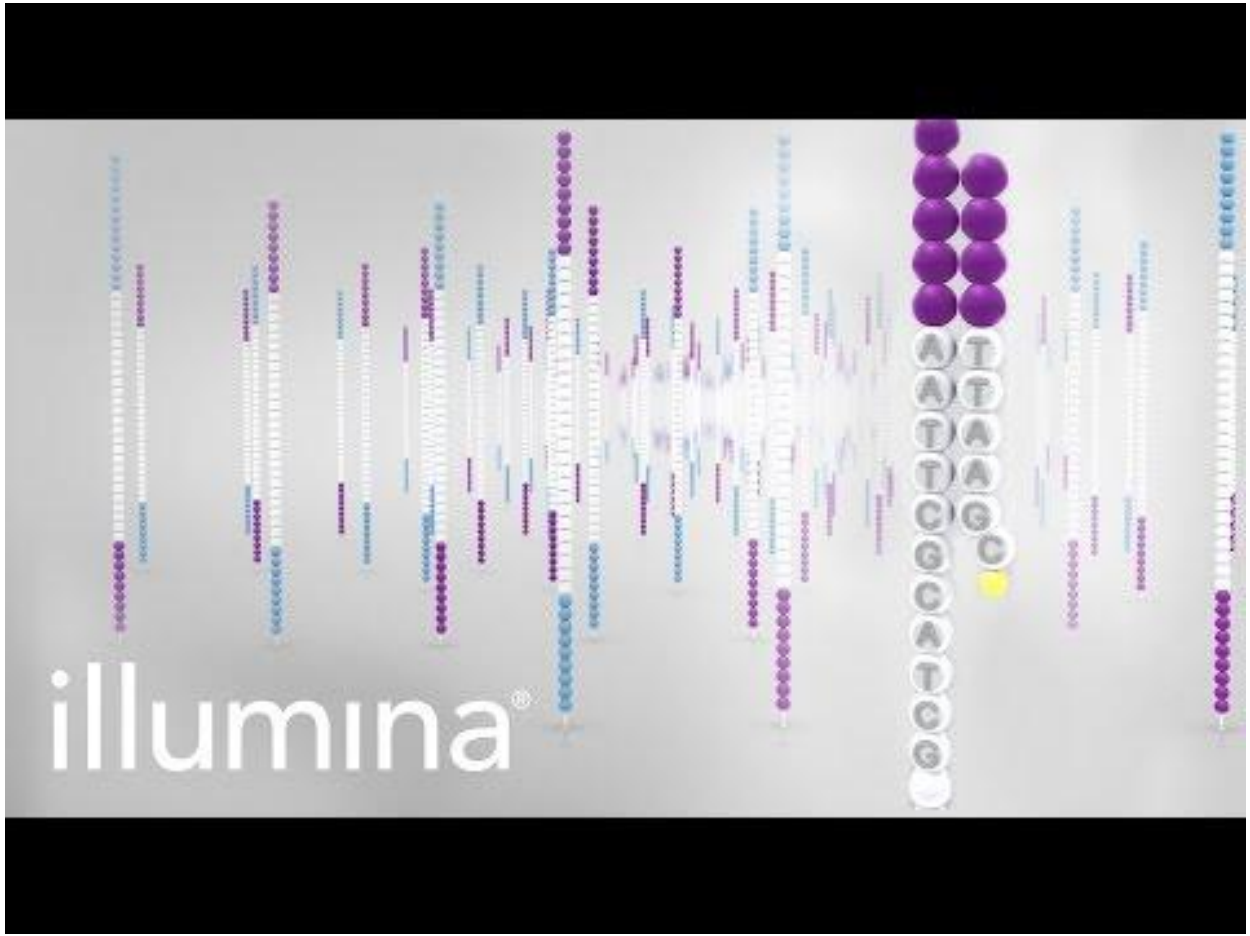
- Supermatrix approaches combine huge numbers of taxa for few genes. Often sparse matrices (missing data). Made possible by algorithmic and computational improvements to likelihood calculations.

- Supertree methods aim to combine information from multiple trees without the need to infer the actual sequence data for all samples at once.

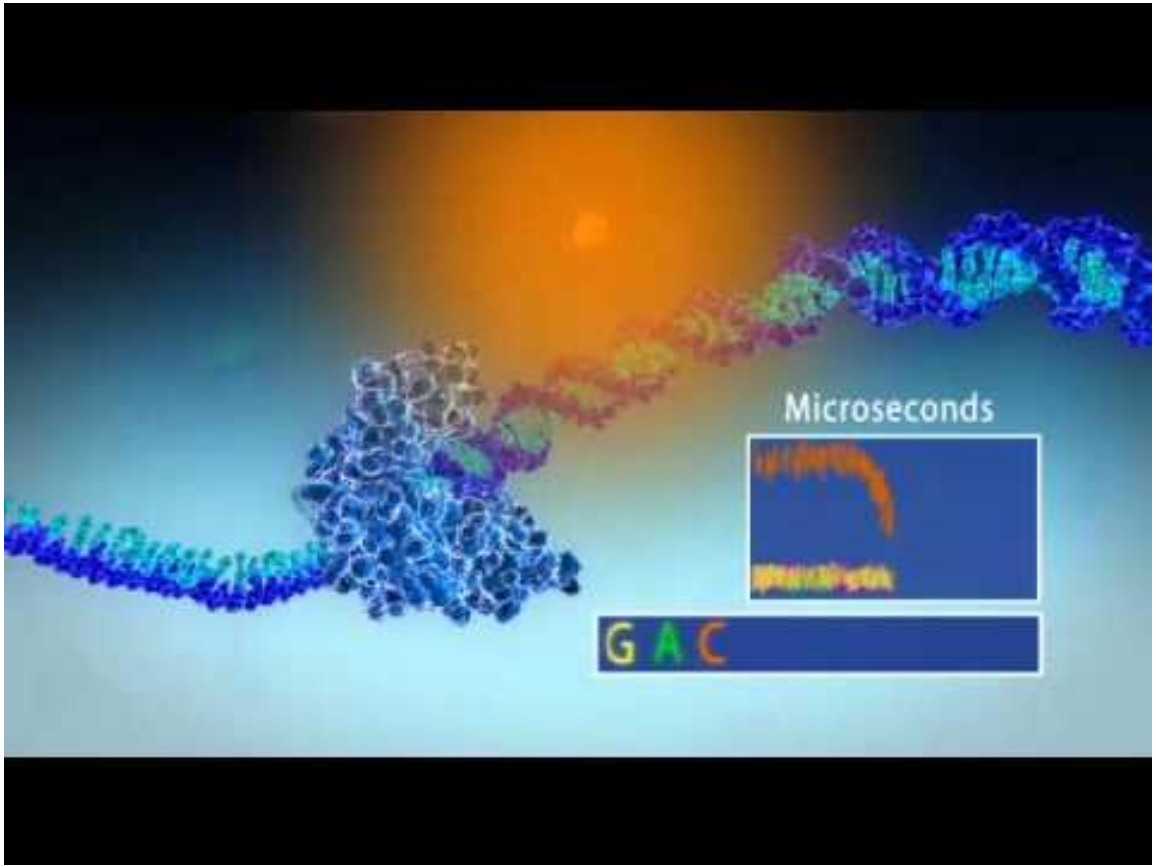
- At the largest scale, both approaches are typically combined to *stitch together* the tree of life with both known (inferred) relationships, and estimated (taxonomy) relationships. *A lot of work remains to be done!*

Phylogenomics

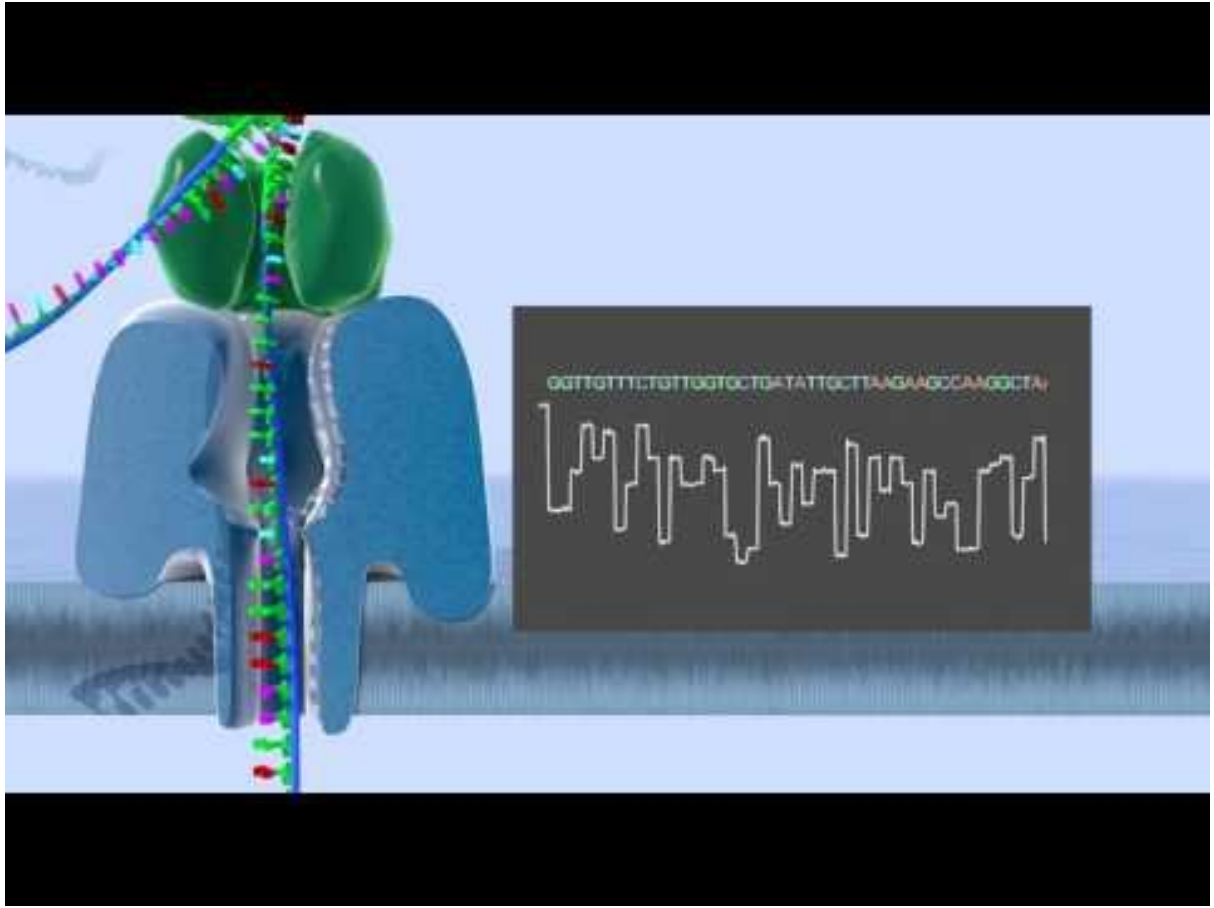
- Of the many ways to sequence genomic data for phylogenetic analyses, how to choose? What methods are available and how do they differ?
    - Whole genome sequencing (WGS)
    - Transcriptome sequencing (RNA-seq)
    - Sequence capture (UCEs; RNA-baits)
    - Amplicon sequencing
    - Restriction-site associated DNA sequencing (RAD-seq)
- It depends on the goal of your study



- Illumina short-read technologies
  - https://www.youtube.com/watch?v=fCd6B5HRaZ8



- Pac-Bio SMRT (single molecule real time) sequencing
  - https://www.youtube.com/watch?v=v8p4ph2MAvI



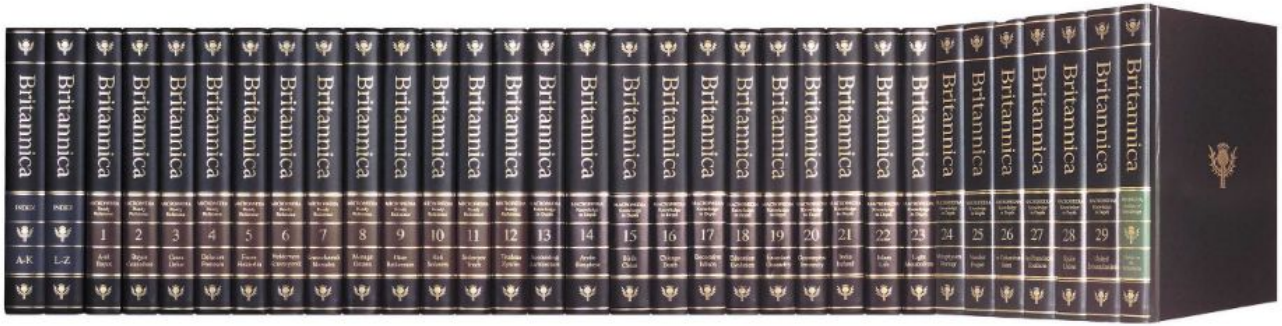
- Nano-pore technology
  - https://www.youtube.com/watch?v=GUb1TZvMWsw

Whole genome sequencing vs. ...

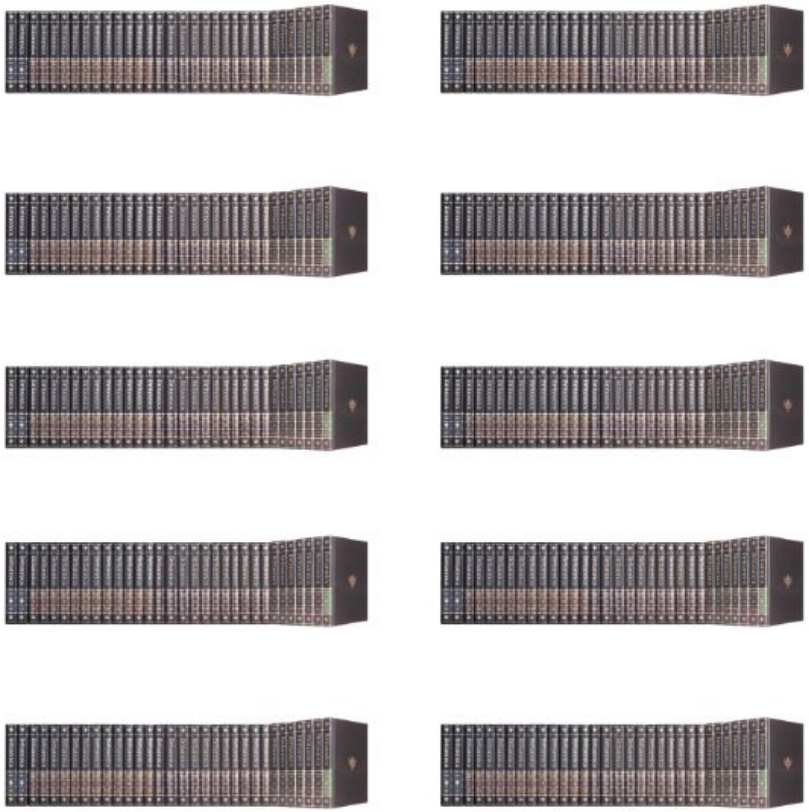
- It is easy to sequence small genomes (e.g., E. coli; 4.6Mb), but very difficult and expensive to sequence large ones (e.g., *Sequoia sempervirens*; 31,000Mb).

- Studies of organisms with small genomes tend to study the whole genome, while large eukaryotic studies tend to subsample the genome, or sequence it to very low depth (~1x), which can introduce many errors.

- Subsampling methods target fewer regions of the genome and typically analyze loci/genes separately (as gene trees).

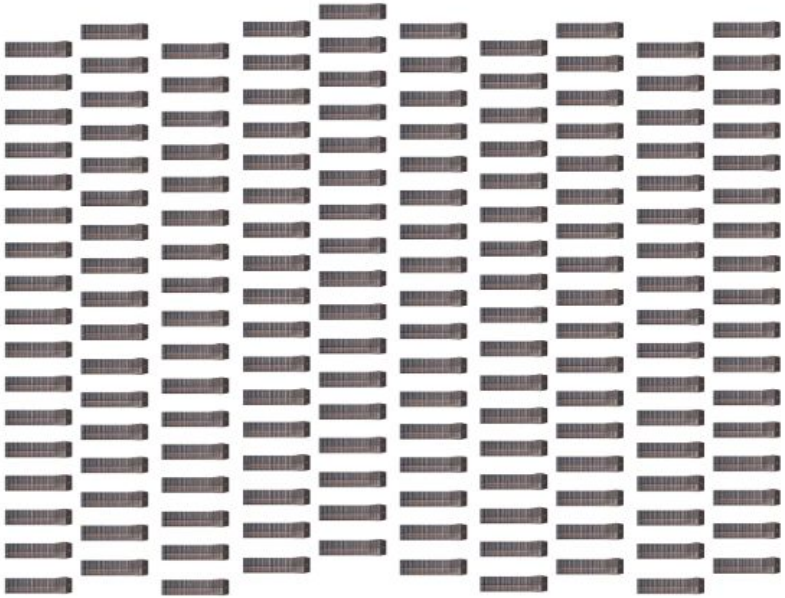


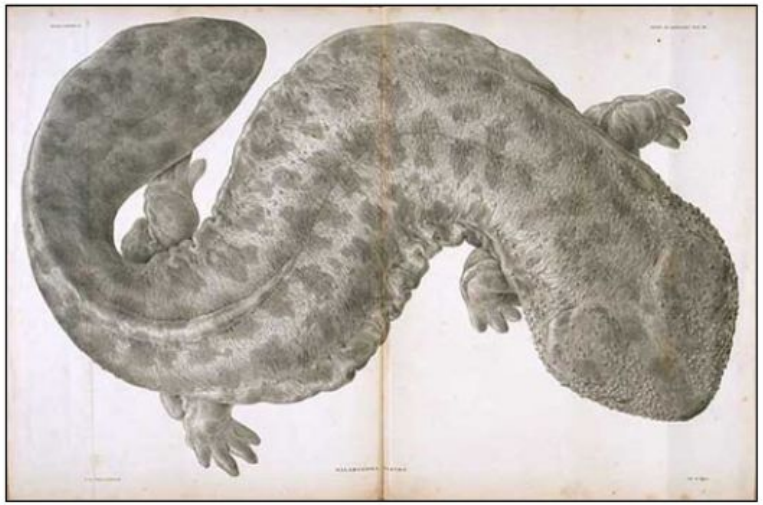
Encyclopedia ≈ *Apis* Genome ≈ 320 million characters



Hominid Genome ≈ 3.2 billion bp

Cryptobranchid Genome ≈ 55 billion bp



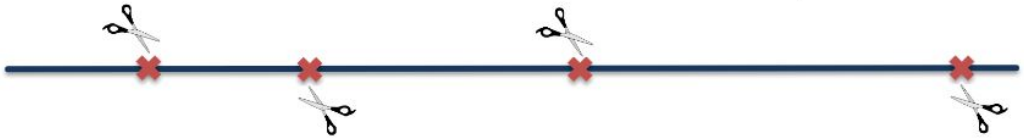


Restriction-site associated DNA

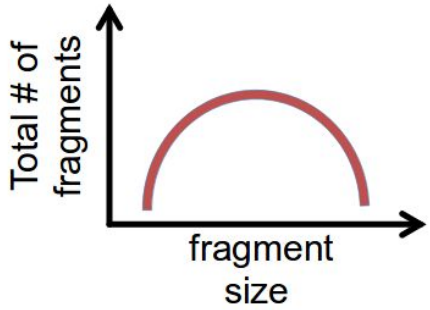
- RAD-sequencing (RAD-seq) and variants (GBS, ddRAD)
- Aims to narrow down the number of sampled regions by targeting a subsample of the genome. In this case, based on the presence of restriction-enzyme recognition sites.
- Subsampling targets fewer regions of the genome and typically treats loci/genes as distinct gene trees.

Restriction enzyme digestion.

= Restriction Enzyme (RE)

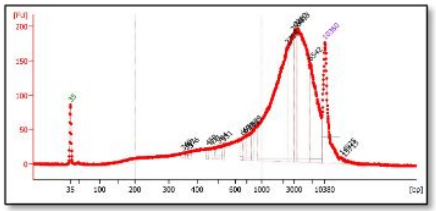


Choice of RE influences the number of fragments produced.



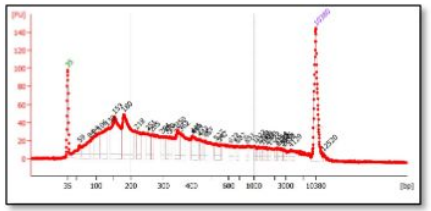
EcoRI

5′...GAATTC...3′
3′...CTTAAG...5′



MluCI

5′...AATT...3′
3′...TTAA...5′



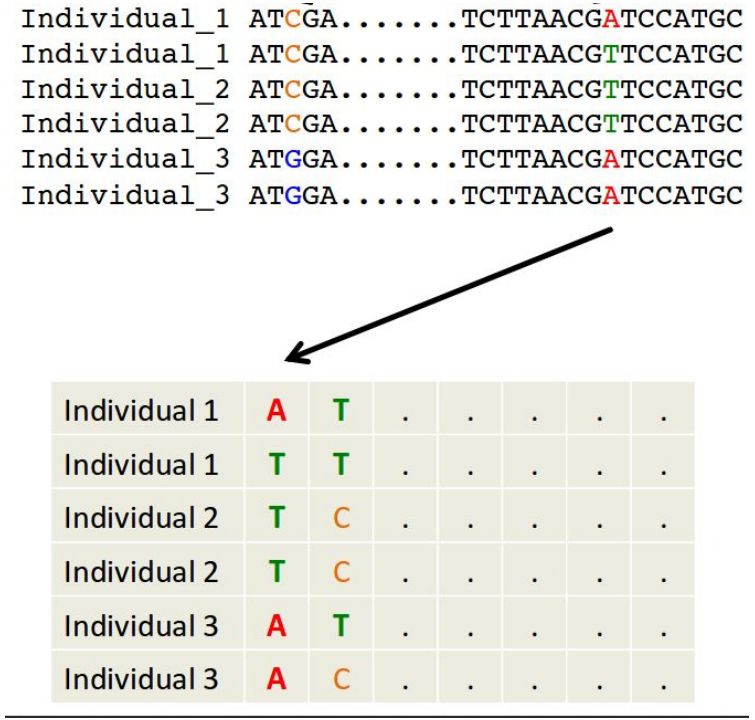
In addition to reducing the genome for sequencing, you also need to sequence each RAD locus enough times to accurately identify genetic variation.

Because of the shotgun nature of NGS, there is no expectation of even coverage across all loci.

Insufficient sequencing can lead to missing one allele at a polymorphic locus, or to completely missing a locus.

Pros and Cons

- RAD-sequencing can target thousands to hundreds of thousands of loci, depending on the enzyme you choose...
- The sequenced loci are short 50-300bp, which may provide little information for inferring gene trees.
- Many new analysis methods make use of only SNPs (single nucleotide polymorphisms), for example, by integrating over all possible gene trees that could produce the SNP (these tend to be slow methods).

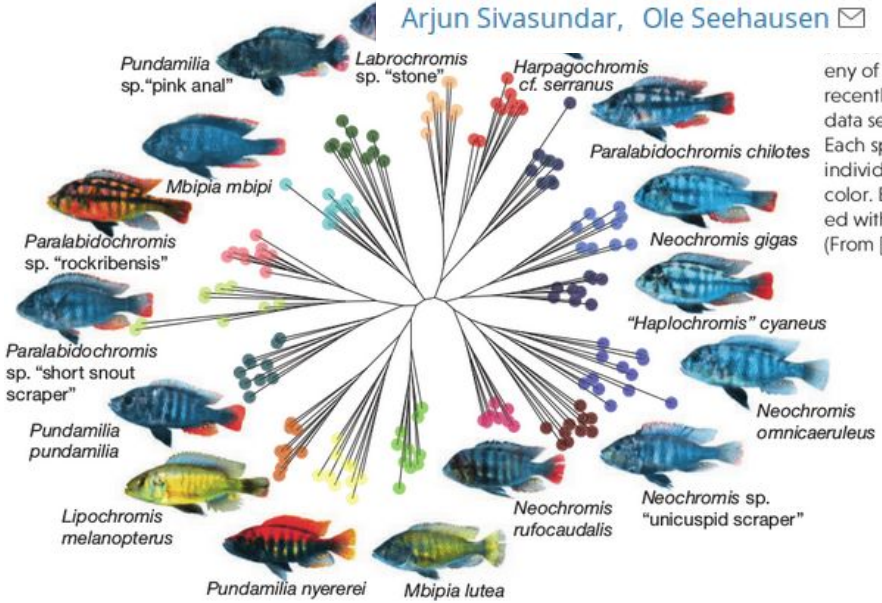


Highly flexible and cheap data type for working across evolutionary scales from very shallow questions to relatively deep-scale questions.

Cichlid radiation includes hundreds of species in just a few thousand years

Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation

Catherine E. Wagner, Irene Keller, Samuel Wittwer, Oliver M. Selz, Salome Mwaiko, Lucie Greuter, Arjun Sivasundar, Ole Seehausen



eny of 16 representative species was recently resolved by using a very large data set of several million DNA bases. Each species is represented by several individuals, shown by dots of the same color. Because this tree was construct- ed without an outgroup, it has no root. (From [36].)

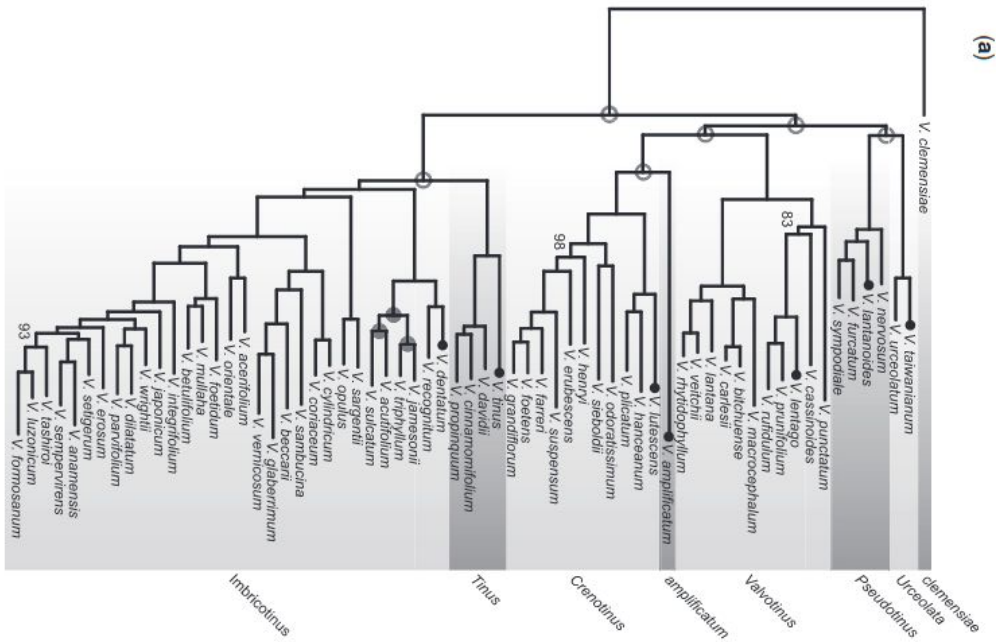
Highly flexible and cheap data type for working across evolutionary scales from very shallow questions to relatively deep-scale questions.

Cichlid radiation includes hundreds of species in just a few thousand years

Viburnum radiation includes a few hundred species over >60 million years.

Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants

DEREN A. R. EATON*, ELIZABETH L. SPRIGGS, BRIAN PARK, AND MICHAEL J. DONOGHUE



Common uses of RAD-seq data

- Phylogenetic inference from hundreds of years to ~100 Ma years.
- QTL mapping, identifying sex-determination loci, loci of large effect.
- Constructing linkage maps (estimating distances between SNPs on chromosomes) based on frequency of recombination throughout the genome.
- Demographic inference: Inferring changes in population sizes and gene flow through time.

Target capture data

- Similarly to RAD-seq this method aims to sample a reduced portion of the genome. Develops “baits” (RNA probes) which capture certain sequences and let non-matching sequences wash away prior to sequencing.

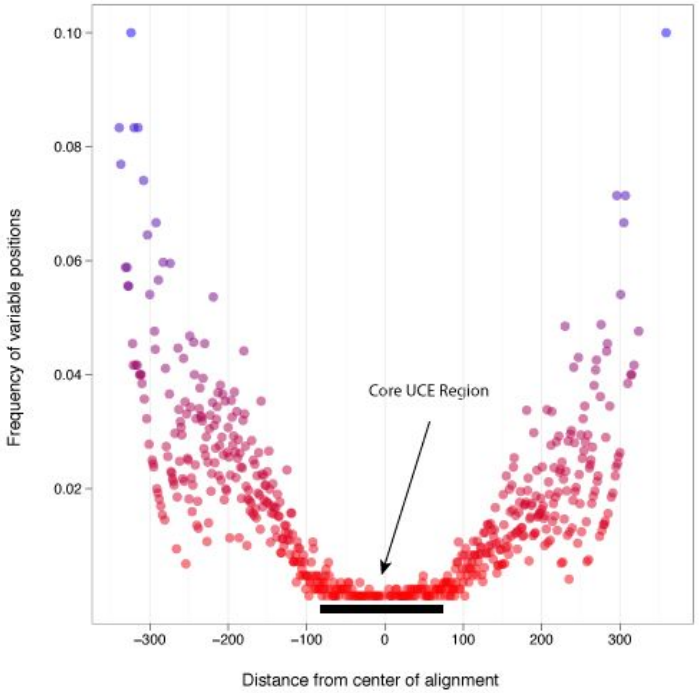
- Requires prior knowledge to develop the baits, typically from a closely related published genome. Or, uses “universal baits” like ultra-conserved elements (UCEs) , which have been found to change very little across huge amounts of time.

- Targets regions downstream from invariant target site. Can target several overlapping regions to build larger *contigs,* spanning up to several thousand base pairs.

- Very repeatable, and creates reusable data across distant taxa.

Pros and Cons

- UCEs target up to a few thousand loci.
- Sequenced loci are longer than RAD loci (200-1000bp), although variation is highly heterogenous. Fewer total SNPs but more informative gene trees on average.
- Hugely useful for deeper phylogenetic analyses because many gene trees can be reliably sampled across very distant taxa (e.g., all birds, all vertebrates).



Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis

John E. McCormack,[1,8] Brant C. Faircloth,[2] Nicholas G. Crawford,[3] Patricia Adair Gowaty,[4,5] Robb T. Brumfield,[1,6] and Travis C. Glenn[7]

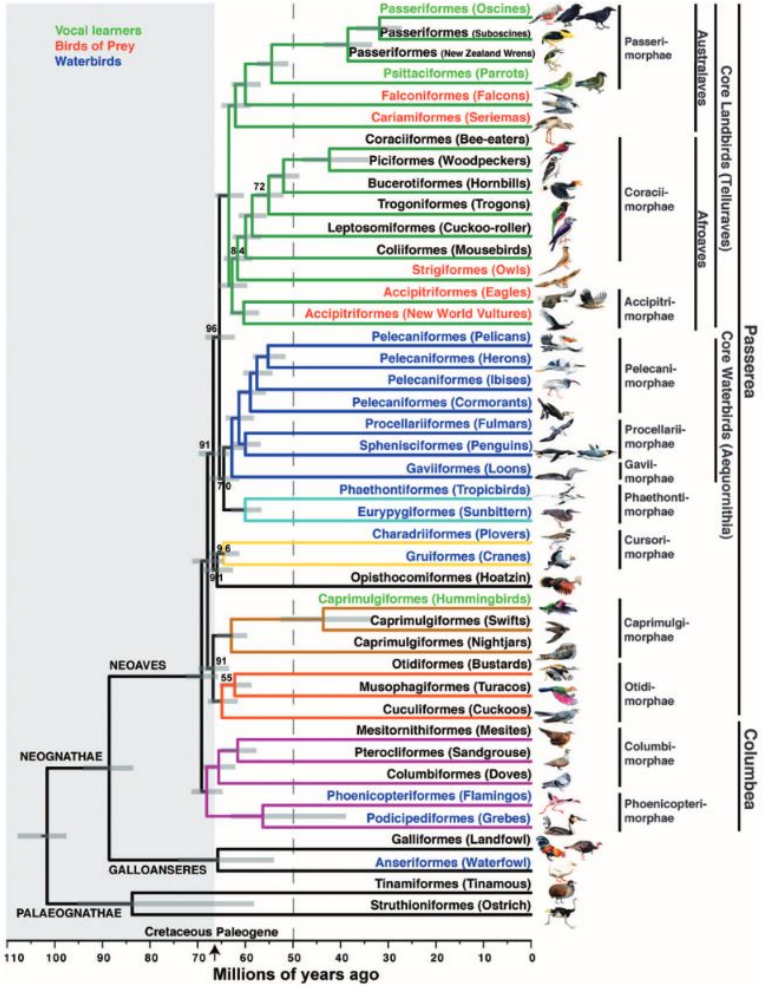
A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing

John E. McCormack ✉, Michael G. Harvey, Brant C. Faircloth, Nicholas G. Crawford, Travis C. Glenn, Robb T. Brumfield

Published: January 29, 2013 • https://doi.org/10.1371/journal.pone.0054848

Whole-genome analyses resolve early branches in the tree of life of modern birds

Erich D. Jarvis,[1*†] Siavash Mirarab,[2*] Andre J. Aberer,[3] Bo Li,[4,5,6] Peter Houde,[7] Cai Li,[4,6] Simon Y. W. Ho,[8] Brant C. Faircloth,[9,10] Benoit Nabholz,[11] Jason T. Howard,[1] Alexander Suh,[12] Claudia C. Weber,[12] Rute R. da Fonseca,[6] Jianwen Li,[4] Fang Zhang,[4] Hui Li,[4] Long Zhou,[4] Nitish Narula,[7,13] Liang Liu,[14] Ganesh Ganapathy,[1] Bastien Boussau,[15] Md. Shamsuzzoha Bayzid,[2] Volodymyr Zavidovych,[1] Sankar Subramanian,[16] Toni Gabaldón,[17,18,19] Salvador Capella-Gutiérrez,[17,18] Jaime Huerta-Cepas,[17,18] Bhanu Rekepalli,[20] Kasper Munch,[21] Mikkel Schierup,[21] Bent Lindow,[6] Wesley C. Warren,[22] David Ray,[23,24,25] Richard E. Green,[26] Michael W. Bruford,[27] Xiangjiang Zhan,[27,28] Andrew Dixon,[29] Shengbin Li,[30] Ning Li,[31] Yinhua Huang,[31] Elizabeth P. Derryberry,[32,33] Mads Frost Bertelsen,[34] Frederick H. Sheldon,[33] Robb T. Brumfield,[33] Claudio V. Mello,[35,36] Peter V. Lovell,[35] Morgan Wirthlin,[35] Maria Paula Cruz Schneider,[36,37] Francisco Prosdocimi,[36,38] José Alfredo Samaniego,[6] Amhed Missael Vargas Velazquez,[6] Alonzo Alfaro-Núñez,[6] Paula F. Campos,[6] Bent Petersen,[39] Thomas Sicheritz-Ponten,[39] An Pas,[40] Tom Bailey,[41] Paul Scofield,[42] Michael Bunce,[43] David M. Lambert,[16] Qi Zhou,[44] Polina Perelman,[45,46] Amy C. Driskell,[47] Beth Shapiro,[26] Zijun Xiong,[4] Yongli Zeng,[4] Shiping Liu,[4] Zhenyu Li,[4] Binghang Liu,[4] Kui Wu,[4] Jin Xiao,[4] Xiong Yinqi,[4] Qiuemei Zheng,[4] Yong Zhang,[4] Huanming Yang,[48] Jian Wang,[48] Linnea Smeds,[12] Frank E. Rheindt,[49] Michael Braun,[50] Jon Fjeldsa,[51] Ludovic Orlando,[6] F. Keith Barker,[52] Knud Andreas Jønsson,[51,53,54] Warren Johnson,[55] Klaus-Peter Koepfli,[56] Stephen O'Brien,[57,58] David Haussler,[59] Oliver A. Ryder,[60] Carsten Rahbek,[51,54] Eske Willerslev,[6] Gary R. Graves,[51,61] Travis C. Glenn,[62] John McCormack,[63] Dave Burt,[64] Hans Ellegren,[12] Per Alström,[65,66] Scott V. Edwards,[67] Alexandros Stamatakis,[3,68] David P. Mindell,[69] Joel Cracraft,[70] Edward L. Braun,[71] Tandy Warnow,[2,72†] Wang Jun,[48,73,74,75,76†] M. Thomas P. Gilbert,[6,43†] Guojie Zhang[4,77†]



LETTER

doi:10.1038/nature15697

A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing

Richard O. Prum[1,2]*, Jacob S. Berv[3]*, Alex Dornburg[1,2,4], Daniel J. Field[2,5], Jeffrey P. Townsend[1,6], Emily Moriarty Lemmon[7] & Alan R. Lemmon[8]

Although reconstruction of the phylogeny of living birds has progressed tremendously in the last decade, the evolutionary history of Neoaves—a clade that encompasses nearly all living bird species—remains the greatest unresolved challenge in dinosaur systematics. Here we investigate avian phylogeny with an unprecedented scale of data: >390,000 bases of genomic sequence data from each of 198 species of living birds, representing all major avian lineages, and two crocodilian outgroups. Sequence data were collected using anchored hybrid enrichment, yielding 259 nuclear loci with an average length of 1,523 bases for a total data set of over 7.8×10^7 bases. Bayesian and maximum likelihood analyses yielded highly supported and nearly identical phylogenetic trees for all major avian lineages. Five major clades form successive sister groups to the rest of Neoaves: (1) a clade including nightjars, other caprimulgiforms, swifts, and hummingbirds; (2) a clade uniting cuckoos, bustards, and turacos with pigeons, mesites, and sandgrouse; (3) cranes and their relatives; (4) a comprehensive waterbird clade, including all diving, wading, and shorebirds; and (5) a comprehensive landbird clade with the enigmatic hoatzin (*Opisthocomus hoazin*) as the sister group to the rest. Neither of the two main, recently proposed Neoavian clades—Columbea and Passerea[1]—were supported as monophyletic. The results of our divergence time analyses are congruent with the palaeontological record, supporting a major radiation of crown birds in the wake of the

It has long been recognized that phylogenetic confidence depends not only on the number of characters analysed and their rate of evolution, but also on the number and relationships of the taxa sampled relative to the nodes of interest[9–11]. Theory predicts that sampling a single taxon that diverges close to a node of interest will have a far greater effect on phylogenetic resolution than will adding more characters[11]. Despite using an alignment of >40 million base pairs, sparse sampling of 48 species in the recent avian genomic analysis may not have been sufficient to confidently resolve the deep divergences among major lineages of Neoaves. Thus, expanded taxon sampling is required to test the monophyly of neoavian clades, and to further resolve the phylogenetic relationships within Neoaves.

Here, we present a phylogenetic analysis of 198 bird species and 2 crocodilians (Supplementary Table 1) based on loci captured using anchored enrichment[12]. Our sample includes species of 122 avian families in all 40 extant avian orders[2], with denser representation of non-oscine birds (108 families) than of oscine songbirds (14 families). Effort was made to include taxa that would break up long phylogenetic branches, and provide the highest likelihood of resolving short internodes at the base of Neoaves[11]. We also sampled multiple species within groups whose monophyly or phylogenetic interrelationships have been controversial—that is, tinamous, nightjars, hummingbirds, turacos, cuckoos, pigeons, sandgrouse, mesites, rails, storm petrels, petrels, storks, herons, hawks, hornbills, mousebirds, trogons, king-

Transcriptome sequencing

- Sequence only the transcribed portion of the genome by extracting RNA, and from that making cDNA.

- Con: Extracting and storing RNA can be very difficult.
- Con: Assembling transcripts can be difficult due to differential splicing.
- Con: Markers are too conserved, too little variation.

- Pro: Phylogenetic markers may also provide insight into the evolution of functional differences, or convergence.
- Pro: Markers are conserved across very deep time-scales.
- Pro: Paralogs (duplicated genes) are typically easier to detect.

Summary of phylogenomic methods

- Whole genome sequencing is difficult for organisms with large genomes where it is preferable to sequence fewer regions to high/reliable depth. Whole genomes must be split into non-recombining loci, or analyzed by sliding window.

- RAD-seq targets regions on the basis of restriction-sites, and can generate many thousands of short markers. Used for younger clade analyses (0-80Ma).

- UCEs target regions on the basis of designed baits chosen to target known conserved regions. Generates a few hundred or thousand markers of longer length and informativeness than RAD. Very useful for deep scale phylogenetic analyses (20-300 Ma).

Reading assignments

- Textbook chapters: 15
- Articles on courseworks:
  - Hoekstra & Coyne 2007
  - Craig 2009