An Annotated Bibliography - Incongruence and Incomplete Lineage Sorting

PhyloSeminar2024

Julia Hobbie

Maddison, Wayne P. 1997. “Gene Trees in Species Trees.” *Systematic Biology* 46 (3): 523–36. <https://doi.org/10.1093/sysbio/46.3.523>.

This is an early review paper that discusses incongruence between gene trees and species trees. It discusses possible causes of these patterns, including horizontal gene transfer, lineage sorting (deep coalescence), and gene duplication, and how to analyze these processes under a parsimony framework. For example, for ILS, the most parsimonious species tree is the one with the fewest “extra” gene tree branches. It addresses how to implement a coalescent model under a maximum likelihood framework. Finally, it discusses reframing phylogenies as “clouds” of gene trees.

Nichols, Richard. 2001. “Gene Trees and Species Trees Are Not the Same.” *Trends in Ecology & Evolution* 16 (7): 358–64. https://doi.org/10.1016/S0169-5347(01)02203-0.

This mostly theoretical paper discusses some modeling/math behind different coalescent times between different gene/species trees. It suggests that these differences can be used to estimate the patterns of population size in deep branches. I uses a number of real examples in the literature (birds, primates) to illustrate that this is a common phenomenon.

Degnan, James H., and Noah A. Rosenberg. 2006. “Discordance of Species Trees with Their Most Likely Gene Trees.” *PLOS Genetics* 2 (5): e68. https://doi.org/10.1371/journal.pgen.0020068.

The authors describe gene trees that don’t match the species trees as “anomalous gene trees”. They show (using models/math) that a species tree cannot always be constructed by simply taking the most common gene tree structures. Adding more data (genes) in these cases does not produce a more “true” tree.

Maddison, Wayne P, and L Lacey Knowles. 2006. “Inferring Phylogeny Despite Incomplete Lineage Sorting.” *Systematic Biology* 55 (1): 21–30. https://doi.org/10.1080/10635150500354928.

This is a simulation paper from before likelihood and Bayesian methods were accessible. The authors simulate data with ILS, then investigate the advantages and disadvantages of sampling deeply (more genes) or broadly (more tips). Generally they found that if there were enough tips, then a majority of genes reflected the “true” tree.

Heled, Joseph, and Alexei J. Drummond. 2010. “Bayesian Inference of Species Trees from Multilocus Data.” *Molecular Biology and Evolution* 27 (3): 570–80.

This paper introduces \*BEAST, which allows phylogenetic inference in a coalescent framework. This simultaneously models individual gene trees and the species tree using Bayesian MCMC.

Steenwyk, Jacob L., Yuanning Li, Xiaofan Zhou, Xing-Xing Shen, and Antonis Rokas. 2023. “Incongruence in the Phylogenomics Era.” *Nature Reviews Genetics* 24 (12): 834–50. <https://doi.org/10.1038/s41576-023-00620-x>.

This review paper gives a broad overview of the causes of incongruence. These include biological factors such as ILS, introgression, and duplication; and analytical factors such as taxon or locus sampling, orthologue assignment, LBA, and (mis-)alignment. It also addresses how to detect incongruence and shares a list of software options for investigating incongruence.

Edwards, Scott V., Zhenxiang Xi, Axel Janke, Brant C. Faircloth, John E. McCormack, Travis C. Glenn, Bojian Zhong, et al. 2016. “Implementing and Testing the Multispecies Coalescent Model: A Valuable Paradigm for Phylogenomics.” *Molecular Phylogenetics and Evolution* 94 (January): 447–62. https://doi.org/10.1016/j.ympev.2015.10.027.

This is a rebuttal to a paper criticizing the use of the multispecies coalescent. This paper shows that most of the error pointed out by the critique do not have an effect on conclusions and offers a defense of the MSC. They emphasize that concatenation can be considered a special case of MSC, where there is zero recombination between genes.

Rivas-González, Iker, Marjolaine Rousselle, Fang Li, Long Zhou, Julien Y. Dutheil, Kasper Munch, Yong Shao, Dongdong Wu, Mikkel H. Schierup, and Guojie Zhang. 2023. “Pervasive Incomplete Lineage Sorting Illuminates Speciation and Selection in Primates.” *Science* 380 (6648): eabn4409. https://doi.org/10.1126/science.abn4409.

This study analyzed ILS in primates. They found high levels of ILS and found that regions of the genome with low recombination had lower levels of ILS. Regions of exons had low ILS, which they represented by showing the effective population size was 30% lower.

Meleshko, Olena, Michael D. Martin, Thorfinn Sand Korneliussen, Christian Schröck, Paul Lamkowski, Jeremy Schmutz, Adam Healey, et al. 2021. “Extensive Genome-Wide Phylogenetic Discordance Is Due to Incomplete Lineage Sorting and Not Ongoing Introgression in a Rapidly Radiated Bryophyte Genus.” *Molecular Biology and Evolution* 38 (7): 2750–66. <https://doi.org/10.1093/molbev/msab063>.

This study was conducted with 12 species of rapidly radiated Sphagnum peat mosses and looked to identify introgression and ILS in this clade. The authors used shotgun sequencing to produce nuclear, chloroplast, and mitochondrial data. In a ML analysis all species groups were recovered to be monophyletic. However, though the chloroplast and mitochondria showed a nearly identical tree topology, they conflicted with the nuclear phylogeny. A coalescent recovered the same incongruence. Analysis of the D-statistic (ABBA-BABA) found evidence of gene flow, which the authors attribute to earlier in the clade’s history because of a lack of evidence for recent gene flow. An analysis with QuIBL (which looks at internal branch lengths for triplets of species) found that only 22% of triplets had evidence of introgression on top of ILS.

Meyer, Britta S., Michael Matschiner, and Walter Salzburger. 2017. “Disentangling Incomplete Lineage Sorting and Introgression to Refine Species-Tree Estimates for Lake Tanganyika Cichlid Fishes.” *Systematic Biology* 66 (4): 531–50. <https://doi.org/10.1093/sysbio/syw069>.

This study identifies ILS and introgression in the radiation of cichlid fishes in Lake Tanganyika. It uses a large dataset with many specimens per species and 40 nuclear markers. First they identify introgression through differences in MRCA age estimates in three-taxon comparison and the f4-statistic. They remove three groups with high signal of introgression, then infer the species tree under the MSC with \*BEAST. Then they used that species tree to estimate gene trees for further inference of introgression.