Transformative potential of model-based analyses in evolutionary biology

L. Lacey Knowles
Dept. of Ecology and Evolutionary Biology
University of Michigan
Applications of model-based analyses:

(i) Species delimitation

(ii) Phylogenetic inference (and beyond the species tree)

(iii) Biogeographic study

(iii) Phylogeography
Genomic data

Model-based analyses

Model Formulation

Composition of competing models
Transformative potential of model-based analyses:

- Codon substitution and analysis of natural selection
- Adaptive molecular evolution
- Divergence time estimation and biogeographic analysis
- Phylogenetic inference
- Species delimitation
- Demographic inference

• All models are flawed..., but...
  models are how we communicate our knowledge to a statistical apparatus
Transformative potential of model-based analyses:

- Decisions/choices we make about model formulation
- Recognizing the subjectivity of model formulation itself when making inferences
- Decisions when applying to empirical data (e.g., all the data, subset of data, what subset of data)
Species delimitation (discovery)

Learning goals:

- Describe applications of the multispecies coalescent (MSC) to species delimitation
- Explain the merit/limitations of the multispecies coalescent (MSC) to delimitation
- Describe (i) how over-estimation of species numbers might occur with applications based on the MSC (ii) what determines the degree of overestimation
- Explain the relevance of the speciation process to delimitation approaches
Isolation is the property that allows species to be recognized genetically.
Coalescence of gene lineages
• Lineages proceed towards monophyly with time

Avise et al. (1987)
• Lineages proceed towards monophyly with time

Avise et al. (1987)
• Lineages proceed towards monophyly with time

Avise et al. (1987)
• Lineages proceed towards monophyly with time

Avise et al. (1987)
Isolation is the property that allows species to be recognized genetically

• Exclusivity criteria (e.g., reciprocal monophyly)

“A group of organisms is exclusive if their loci coalesce more recently within the group than between any member of the group and any organisms outside the group” (Baum & Shaw 1995, p. 296).

• Disconnect between the time of speciation and when taxa reach reciprocal monophyly

FIG. 1. Probabilities of observing reciprocal monophyly with time for populations that are genetically isolated. Curves are shown for a single mitochondrial locus and for samples of different numbers of nuclear loci. Time is measured in units of $N$ generations, where $N$ is the effective population size of each of the two descendant populations. mt, mitochondrial DNA, $n$, number of nuclear loci sampled. We assume that 100 alleles are sampled for each locus.  

Delimiting Species without Monophyletic Gene Trees

L. LACEY KNOWLES AND BRYAN C. CARSTENS

Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Avenue, University of Michigan, Ann Arbor, MI 48109-1079, USA; E-mail: knowles@umich.edu (L.L.K.)

Abstract.—Genetic data are frequently used to delimit species, where species status is determined on the basis of an exclusivity criterium, such as reciprocal monophyly. Not only are there numerous empirical examples of incongruence between the boundaries inferred from such data compared to other sources like morphology—especially with recently derived species, but population genetic theory also clearly shows that an inevitable bias in species status results because genetic thresholds do not explicitly take into account how the timing of speciation influences patterns of genetic differentiation. This study represents a fundamental shift in how genetic data might be used to delimit species. Rather than equating gene trees with a species tree or basing species status on some genetic threshold, the relationship between the gene trees and the species history is modeled probabilistically. Here we show that the same theory that is used to calculate the probability of reciprocal monophyly can also be used to delimit species despite widespread incomplete lineage sorting. The results from a preliminary simulation study suggest that very recently derived species can be accurately identified long before the requisite time for reciprocal monophyly to be achieved following speciation. The study also indicates the importance of sampling, both with regards to loci and individuals. Withstanding a thorough investigation into the conditions under which the coalescent-based approach will be effective, namely how the timing of divergence relative to the effective population size of species affects accurate species delimitation, the results are nevertheless consistent with other recent studies (aimed at inferring species relationships), showing that despite the lack of monophyletic gene trees, a signal of species divergence persists and can be extracted. Using an explicit model-based approach also avoids two primary problems with species delimitation that result when genetic thresholds are applied with genetic data—the inherent biases in species detection arising from when and how speciation occurred, and failure to take into account the high stochastic variance of genetic processes. Both the utility and sensitivities of the coalescent-based approach outlined here are discussed; most notably, a model-based approach is essential for determining whether incompletely sorted gene lineages are (or are not) consistent with separate species lineages, and such inferences require accurate model parameterization (i.e., a range of realistic effective population sizes relative to potential times of divergence for the purported species). It is the goal (and motivation of this study) that genetic data might be used effectively as a source of complementation to other sources of data for diagnosing species, as opposed to the exclusion of other evidence for species delimitation, which will require an explicit consideration of the effects of the temporal dynamic of lineage splitting on genetic data. [Coalescence; genealogical discord; genealogical species concept; gene trees; incomplete lineage sorting.]
Delimiting Species without Monophyletic Gene Trees

L. Lacey Knowles and Bryan C. Carstens

Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Avenue, University of Michigan, Ann Arbor, MI 48109-1079, USA; E-mail: knowles@umich.edu (L.L.K.)

Abstract.—Genetic data are frequently used to delimit species, where species status is determined on the basis of an exclusivity criterium, such as reciprocal monophyly. Not only are there numerous empirical examples of incongruence between the boundaries inferred from such data compared to other sources like morphology—especially with recently derived species, but population genetic theory also clearly shows that an inevitable bias in species status results because genetic thresholds do not explicitly take into account how the timing of speciation influences patterns of genetic differentiation. This study represents a fundamental shift in how genetic data might be used to delimit species. Rather than equating gene trees with a species tree or basing species status on some genetic threshold, the relationship between the gene trees and the species history is modeled probabilistically. Here we show that the same theory that is used to calculate the probability of reciprocal monophyly can also be used to delimit species despite widespread incomplete lineage sorting. The results from a preliminary simulation study suggest that very recently derived species can be accurately identified long before the requisite time for reciprocal monophyly to be achieved following speciation. The study also indicates the importance of sampling, both with regards to loci and individuals. Withstanding a thorough investigation into the conditions under which the coalescent-based approach will be effective, namely how the timing of divergence relative to the effective population size of species affects accurate species delimitation, the results are nevertheless consistent with other recent studies (aimed at inferring species relationships), showing that despite the lack of monophyletic gene trees, a signal of species divergence persists and can be extracted. Using an explicit model-based approach also avoids two primary problems with species delimitation that result when genetic thresholds are applied with genetic data—the inherent biases in species detection arising from when and how speciation occurred, and failure to take into account the high stochastic variance of genetic processes. Both the utility and sensitivities of the coalescent-based approach outlined here are discussed; most notably, a model-based approach is essential for determining whether incompletely sorted gene lineages are (or are not) consistent with separate species lineages, and such inferences require accurate model parameterization (i.e., a range of realistic effective population sizes relative to potential times of divergence for the purported species). It is the goal (and motivation of this study) that genetic data might be used effectively as a source of complementation to other sources of data for diagnosing species, as opposed to the exclusion of other evidence for species delimitation, which will require an explicit consideration of the effects of the temporal dynamic of lineage splitting on genetic data. [Coalescence; genealogical discord; genealogical species concept; gene trees; incomplete lineage sorting.]
Different species delimitation models are formulated as competing statistical models and inferred from the genetic data through Bayesian model selection (i.e., through calculation of posterior model probabilities).

Yang and Rannala (2010) *PNAS*
Explosion of applications of the MSC for delimitation

Cryptic species as a window into the paradigm shift of the species concept

Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses

Comparison of Methods for Molecular Species Delimitation Across a Range of Speciation Scenarios

Implementation Bayesian
for species delimitation
species coalescent
Pros of species delimitation under MSC

• Can delimit species before reciprocal monophyly of alleles or fixed differences

• Still detects lineages under low gene flow

• Accuracy of species delimitation to sampling can be evaluated (i.e., will more data change status)

• De facto standardization for objectively delimiting taxa (i.e., data treated equally among all living things and avoid subjectiveness of what characters to measure)
  Fujita et al. (2012) TREE

• Can take into account uncertainty in gene trees
  Yang & Rannala 2010
Coalescent-based species delimitation

Most newly discovered species go undescribed.

• Less than 30% of researchers applying MSC models made taxonomic recommendations!
• Less than 25% of researchers applying MSC models actually use results to describe new species!

Carstens et al. 2013
Measures of evolutionary independence

Pinho and Hey (2010) *Evolution*
Eventually all species concepts agree... so no big deal right?!?
* Not all lineages become species!

Mosaic of populations and species

* 6 distinguished genetic lineages

* 4 species (represented by different colors)

speciation duration

splitting  extinction  merging
the multispecies COALESCENT

Model-based genetic species delimitation

Sukumaran & Knowles (2017) PNAS
Model-based species delimitation – still long way to go!
Genetic structure: mosaic of population and species lineages (i.e., isolation is a property of not only species, but also populations)

Splitting events such as this are initiation of speciation through, e.g., population isolation.

Color change indicates completion of speciation and development of true species from incipient species (i.e., lineage conversion).

Sukumaran & Knowles (2017) *PNAS*
Simulate data to account for differences in speciation duration (i.e., speciation is not instantaneous)

Splitting events such as this are initiation of speciation through, e.g., population isolation

Color change indicates completion of speciation and development of true species from incipient species (i.e., lineage conversion)

Does the MSC accurate delimit species?

Sukumaran & Knowles (2017) *PNAS*

Most probable delimitation model?

8 vs 3 species
Performance of species delimitation under the MSC for data simulated under different speciation durations

The MSC does not track species, but rather tracks structure of any sort, whether population or true species

Sukumaran & Knowles (2017) *PNAS*
Splitting events such as this are initiation of speciation through, e.g., population isolation.

Color change indicates completion of speciation and development of true species from incipient species (i.e., lineage conversion).

- theoretical demonstration, but not practically relevant since we don’t know how long it takes for speciation

- the PSM model of speciation doesn’t fit my empirical system

- everyone recognizes the MSC doesn’t delimit species per se

- the MSC doesn’t make assumptions about the speciation process
Species discovery

Why are species boundaries NOT accurately delimited under the objective model (the MSC)?
Species discovery: **Subjective** + **Objective**

- Choices during model formulation
Transformative potential of model-based analyses:

- Choices during model formulation
- Recognizing the subjectivity of model formulation itself when making inferences

Need more complex models
Summary

Genetic model-based species delimitation – still long way to go!

• MSC detects structure – not species
  (seeking consensus across MSC-based methods is not a good way to fail)


• “Robustness” to lineage detection with low levels of gene flow is not the same as accurate species delimitation

• Sensitivity to sampling (i.e., more data change status)

• In practice, MSC is not a de facto standardization for objectively delimiting taxa: depending upon speciation process degree of over estimation will vary
In practice, MSC is not a de facto standardization for objectively delimiting taxa: depending upon speciation process degree of over estimation will vary.
Summary

Genetic model-based species delimitation – still long way to go!

- MSC detects structure – not species
  (seeking consensus across MSC-based methods is not a good way to fail)

- “Robustness” to lineage detection with low levels of gene flow is not the same as accurate species delimitation

- Sensitivity to sampling (i.e., more data change status)

- In practice, MSC is not a de facto standardization for objectively delimiting taxa: depending upon speciation process degree of over estimation will vary

- Accurate species delimitation cannot be achieved with genetic data alone (at this time.....stay tuned)
Ad hoc **heuristics** to interpret results from MSC-based models for delimitation

- **Genealogical sorting index**: \(2\tau/\theta\)
  (i.e., population divergence time relative to the population size)

- use population divergence parameters (e.g., distantly related species, lots of migration)

Using diverse sources of data for inferring species boundaries has a long systematic tradition, but not with model-based inference.

Joint analysis of morphology and genetic data!

iBPP: joint analysis of multiple data types

Solis-Lemus et al. 2014

parameter \( \lambda \): models the between-to-within species variance ratio

*Each trait is allowed to have its own \( \lambda \) parameter to model selection acting differently on different traits
A set of simulations with a two-species guide tree to quantify the power from trait data in a simple case

Figure 4. Degree of overlap in trait values between species for three traits simulated with $\lambda = 0.7$. Boxplots along the axes margins summarize the distribution of each trait within each species separately (median, first, and third quartile and range of trait values). Analysis with iBPP gives PP 0.96 for two species.

Solis-Lemus et al. 2014
Quantifying the power from trait data in a simple case

Powerful even without discrete differences among species!

Analyses with one trait:

Analyses with three traits:

**Figure 3.** Posterior probability (PP) for the two-species model given one trait (top) or three traits (bottom), when the true model has one species (left) or two species (right). $p$ is the proportion of replicates for which the PP of the true model was 0.50 or higher.

Solis-Lemus et al. 2014
Analyses with traits increase accuracy when divergence occurs with gene flow

Solis-Lemus et al. 2014
Insights beyond the goal of delimiting taxa


5 species

1 species

genus Dynastes
Integrative species delimitation to capture multiple species properties

genus *Dynastes*

Huang & Knowles (2016) *Syst. Biol.*
Capturing divergence process, in addition to capturing multiple properties of species.
Model-based species delimitation

- Subjectivity when choose a model (i.e., cannot be removed from applications of model-based inference)

- the model tells us about support for a hypothesis (i.e., support or don’t support species property as captured by model)

- need more complex models to delimit species; something we haven’t done yet because we ignored subjectivity of interpreting results from MSC

- integrative analyses (i.e., not relying on a single data type) important: (i) capture multiple properties of species, (ii) robustness for evaluating support of hypothesis

Opt for alternatives to model-based inference (e.g., apply heuristics to output from MSC analyses)  
QUESTIONS?

support NSF & the UM
Transformative potential of model-based analyses:

(i) Species delimitation

(ii) Phylogenetic inference (and beyond the species tree)

what I’ll emphasize:

• Decisions/Choices we make about model formulation

• Recognizing the subjectivity of model formulation itself when making inferences

• Decisions when applying to empirical data (e.g., all the data, subset of data, what subset of data)
Model-based phylogenetic estimation

ML tree

\[ \sum_{\text{possible gene trees}} \mathbb{P}(\text{sequences} \mid \text{gene tree}) \]
Model-based phylogenetic estimation

Rate Matrix

\[
\begin{array}{c}
A & \leftrightarrow & G \\
\downarrow & & \updownarrow \\
C & \leftrightarrow & T \\
\end{array}
\]

Base Frequencies

\[
\pi_A + \pi_C + \pi_G + \pi_T = 1
\]

Site Rates

\[+ I + G\]

JC

\[
a = b = c = d = e = f
\]

HKY

\[
a = c = d = f, b = e
\]

GTR

\[
a, b, c, d, e, f
\]

GTR+I+G

\[
a, b, c, d, e, f
\]

partitioned model

ML tree

\[
\sum_{\text{possible gene trees}} [P(\text{sequences} | \text{gene tree})]
\]

Molecular evolution

GTR+H+G

\[
a, b, c, d, e, f
\]

10 free parameters

GTR+H+G

\[
a, b, c, d, e, f'
\]

10 free parameters
Model-based phylogenetic estimation

\[
\sum_{\text{loci}} \sum_{\text{possible gene trees}} [P(\text{sequences} \mid \text{gene tree}) \times P(\text{gene tree} \mid \text{species tree})]
\]

Species tree versus gene trees

- divergence history of a locus and species may differ
Species tree versus gene trees

- the divergence history of individual loci may differ
Relevance of the species-tree gene-tree distinction:

- infer species relationships

“A flock of genomes”

A coalescent-based estimates of the avian species tree of life using a method based on the statistical binning of loci

Mirarab et al. 2014

(from Zhang et al. 2014)
Relevance of the species-tree gene-tree distinction:

- infer species relationships

- study the processes underlying discord
  - diversification history of taxa
  - genome evolution
Relevance of the species-tree gene-tree distinction:

- infer species relationships
- study the processes underlying discord
  - diversification history of taxa
  - genome evolution

EMBRACE THE HETEROGENEITY !
Inherent interest of discordant gene trees

- insight about speciation/diversification process

Maddison 1997

Branch long and narrow: Deep coalescence unlikely

Branch short and wide: Deep coalescence likely
Inherent interest of discordant gene trees

- insight about genome evolution

Xi et al. 2012
Phylogenomics and Next-Generation Inferences: the Future of Phylogenetics in the Era of Big Data

The addition of potential information content for phylogenetic inference comes at the expense of increased data heterogeneity that can result in model misspecification, hindering accurate phylogenetic reconstruction.
Resolved accurate phylogenetic relationships among species?

- NO – recalitrant nodes across the tree of life

Data vs model problem?
Relevance of the species-tree gene-tree distinction:

- infer species relationships
- study the processes underlying discord

Both endeavors face methodological challenges because multiple processes produce discord among gene trees
Relevance of the species-tree gene-tree distinction:

- infer species relationships

- study the processes underlying discord

Both endeavors face methodological challenges because multiple processes produce discord among gene trees

DON’T HAVE ANY APPROACHES FOR PHYLOGENETIC INFERENCE THAT MODEL MULTIPLE PROCESSES THAT PRODUCE DISCORD
There is an inherent increase in data heterogeneity as shift to transcriptomes/genomes and more taxa. Genomic datasets face more than just computational challenges!

- There is an inherent increase in data heterogeneity as shift to transcriptomes/genomes and more taxa.

**PROBLEM?**
( Discord not due to just ILS)
Multiple processes contribute to gene tree discord

- highly elevated levels of strongly supported conflict

- Concordant with species tree
- Topologies without strong support
- Support main alternative topology
- Remaining discordant topologies

Caryophyllales

Smith et al. 2015 BMC Evolutionary Biology
Multiple processes contribute to gene tree discord

- highly elevated levels of strongly supported conflict that cannot be explained by ILS alone at some nodes

**Concordant with species tree**
- Topologies without strong support
- Support main alternative topology
- Remaining discordant topologies

Caryophyllales

Smith et al. 2015 BMC Evolutionary Biology
Species tree versus gene trees

... Discord not due to just ILS and may include hidden paralogy, hybridization, lack of phylogenetic signal, recombination, and horizontal gene transfer (HGT)
Species tree inference: theoretical and empirical challenges of today and tomorrow

Laura Kubatko
L. Lacey Knowles
Systematic errors in phylogenetic inference caused by model misspecification

- Quality of phylogenetic inferences might be affected by simplifying assumptions

Hypothesized relationships among major clades of plants

Wickert et al. 2014 PNAS

69 analyses of 92 taxa
Systematic errors in phylogenetic inference caused by model misspecification

- Quality of phylogenetic inferences might be affected by simplifying assumptions

hypothesized relationships among major clades of plants

Wickett et al. 2014 PNAS

69 analyses of 92 taxa

- Leads to empiricists trying to identify the “right” loci in the transcriptomic data for phylogenetic inference
Using all the data versus subsets of data?

“CLASSIPHY” gene trees by cause of discord

• identify data to improve estimates of phylogenetic relationships

  No consensus on how to do this!

• Can we quantify the relative contributions of processes to patterns of discord?
  - Do they differ among clades, across time, among taxa, and is it associated with ecological shifts?

Huang et al. 2018
“CLASSIPHY”: simulation-trained multivariate discriminant analysis classification

Competing models for the cause of gene tree discord

Simulate training data

Huang et al. 2018
Knowles et al. 2018
“CLASSIPHY”: simulation-trained multivariate discriminant analysis classification

Competing models for the cause of gene tree discord

- Simulate training data
- Calculate summary statistics
- Construct discriminant analysis function(s)

Huang et al. 2018
Knowles et al. 2018
“CLASSIPHY”: simulation-trained multivariate discriminant analysis of principal components of summary statistics

- Low discriminatory power with a single summary statistic
“CLASSIPHY”: simulation-trained multivariate discriminant analysis of principal components of summary statistics

- Low discriminatory power with a single summary statistic

- Power of DAPC procedure for simulation trained classification comes from the multidimensional aspect of calculating multiple summary statistics
“CLASSIPHY”: simulation-trained multivariate discriminant analysis of principal components of summary statistics

- Low discriminatory power with a single summary statistic

- Power of DAPC procedure for simulation trained classification comes from the multidimensional aspect of calculating multiple summary statistics

- we calculated 25 summary statistics for each gene tree based on topology
“CLASSIPHY”: simulation-trained multivariate discriminant analysis classification

Competing models for the cause of gene tree discord

- Simulate training data
- Calculate summary statistics
- Construct discriminant analysis function(s)

- the summary statistics are not used directly

Huang et al. 2018
Knowles et al. 2018
"CLASSIPHY": simulation-trained multivariate discriminant analysis classification

Competing models for the cause of gene tree discord

- principal components (PC) extracted from the summary statistics are used in discriminant analysis (hence, the summary statistics can be correlated, and some can even be uninformative for certain divergent histories)

Simulate training data

Calculate summary statistics

Calculate principal components

Construct discriminant analysis function(s) (DAPC; see Jombart et al. 2010)

Huang et al. 2018
Knowles et al. 2018
“CLASSIPHY”: simulation-trained multivariate discriminant analysis classification

Competing models for the cause of gene tree discord

- too many PC will result in overfitting to the training data, whereas too few will result in lack of power

- we select the number of PCs using a heuristic optimization criterion (i.e., we construct an array of discriminant functions using different number of PCs, and re-classify the simulated training dataset using these functions and select the number of PCs that gives the highest proportion of correct re-classification)

Simulate training data

Calculate summary statistics

Calculate principal components

Construct discriminant analysis function(s) (DAPC; see Jombart et al. 2010)

Huang et al. 2018
Knowles et al. 2018
“CLASSIPHY”: simulation-trained multivariate discriminant analysis classification

Competing models for the cause of gene tree discord

Empirical gene trees to be classified
  ↓
Calculate summary statistics
  ↓
Calculate principal components
  ↓
Apply discriminant analysis function(s) to classify empirical sets of gene trees

Simulate training data
  ↓
Calculate summary statistics
  ↓
Calculate principal components
  ↓
Construct discriminant analysis function(s) (DAPC; see Jombart et al. 2010)

• It is the machine learning algorithm (i.e., DAPC in this case) that finds the combination of the summary statistics that can identify HGT-affected loci among loci with ILS-caused discord
“CLASSIPHY” gene trees by cause of discord

- identifying conflicts among gene trees that are inconsistent with ILS
“CLASSIPHY” gene trees by cause of discord

- identifying conflicts among gene trees that are inconsistent with ILS

(this is great - if we actually knew the species tree - which we don’t)

Gene tree distributions under the coalescent

Degnan and Salter 2005
"CLASSIPHY" gene trees by cause of discord

- Integrate over possible species trees for a given number of taxa

Simulating gene trees

- Simulate species trees
  - Species tree
  - Speciess tree
  - Species tree

- Simulate locus trees
  - Locus tree
  - Locus tree

- Simulate gene trees (ILS)
  - Gene tree
  - Gene tree

Sum stats

- Sum stat
- Sum stat
- Sum stat

Optimize the number of PCs in DAPC

Empirical gene trees

- Gene tree
- Gene tree
- Gene tree

Sum stat

- Sum stat
- Sum stat

Testing data matrix

- Testing
data matrix

Training data matrix

- Training
data matrix

Results:

- LGT
- ILS
- ILS

Discriminant function

- Classify loci

Yule model

- 100 taxa
- 50Ne generations

1-5 x 10^9

LGT event

per generation
Test data for classification

- 1000 species trees simulated under BD process with $B = 2 \times D$
  - 100 species
  - total tree depth $50N; \ N_e = 5 \times 10^5$
- 2000 loci per species tree
- HGT rate randomly drawn from a uniform distribution, $1 \times 10^{-9}$ to $5 \times 10^{-9}$

(10%-50% loci had HGT that changed topology)

Each of 2000 gene trees from each species tree was tested (i.e., DAPC analyses were repeated 1000 times to obtain not only the average performance, but also information on the variation across different species trees and HGT rates).
- Training set composed of gene trees that have high to low levels of discord due to HGT and ILS, and differ in the relative contribution of processes to discord.
Performance of “CLASSIPHY”

- the posterior weight is a good predictor for the true regime

ROC curve across all species trees

average AUC (area under curve) is 0.81
Performance of “CLASSIPHY”

- the proportion of gene trees in which the true model was the preferred model

Gene trees were assigned to regimes with posterior probability higher than 0.5 (i.e., 0.5 as the default cutoff)
Performance of “CLASSIPHY”

- the proportion of gene trees in which the true model was the preferred model

(cutoff that maximizes Youden’s index (i.e., sensitivity+specificity) of the ROC curves is 0.36)
Future work for improving/expanding “CLASSIPHY”

- Additional summary statistics to improve performance
  (e.g., summary stats based on covariance across gene trees to exploit correlation predicted for ILS but not HGT)

- Using information available about the species tree
  (e.g., for resolved nodes)

- Map back to gene trees HGT and ILS events
“CLASSIPHY” gene trees by cause of discord

- identify data to improve estimates of phylogenetic relationships

- study the relative contributions of processes to patterns of discord (e.g., what proportion of genome is moving between species boundaries, and does it differ among clades, across time, or if associated with ecological shifts?)


Phylogenomics and Next-Generation Inferences

- Decisions/choices we make about model formulation
- Recognizing the subjectivity of model formulation itself when making inferences
- Decisions when applying to empirical data (e.g., all the data, subset of data, what subset of data)

- The potential information content of big data comes at the expense of increased data heterogeneity that can result in model misspecification, hindering accurate phylogenetic reconstruction.

- Next generation inference - use big data to explore new questions, stimulate new methods for summarizing the data, and generate new insights!