Class discussion on 04/12/2016

Hosner P.A., Faircloth B.C., Glenn T.C., Braun E.L., Kimball R.T. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes) *Molecular Biology Evolution*, Advanced access published December 29, 2015.

Figures taken directly from the manuscript. Slides and presentation are my own interpretation of the manuscript's content.

Motivation

Given molecular sequence data on a set of m taxa and n genes/loci, we construct an $m \times n$ matrix such that each column $i \in 1, \ldots, n$ is a multiple sequence alignment on the m taxa.

This matrix can be expanded to larger numbers of taxa and/or genes by allowing missing data, i.e., molecular sequences.

Question: Is more data always better?

- Do the benefits of a larger data matrix at the cost of missing data, benefit or harm existing species tree estimation methods?
- How should researchers select/justify a threshold for gene/locus inclusion given the presence of missing data?

Motivation

The $m \times n$ matrix may be expanded by including missing data of

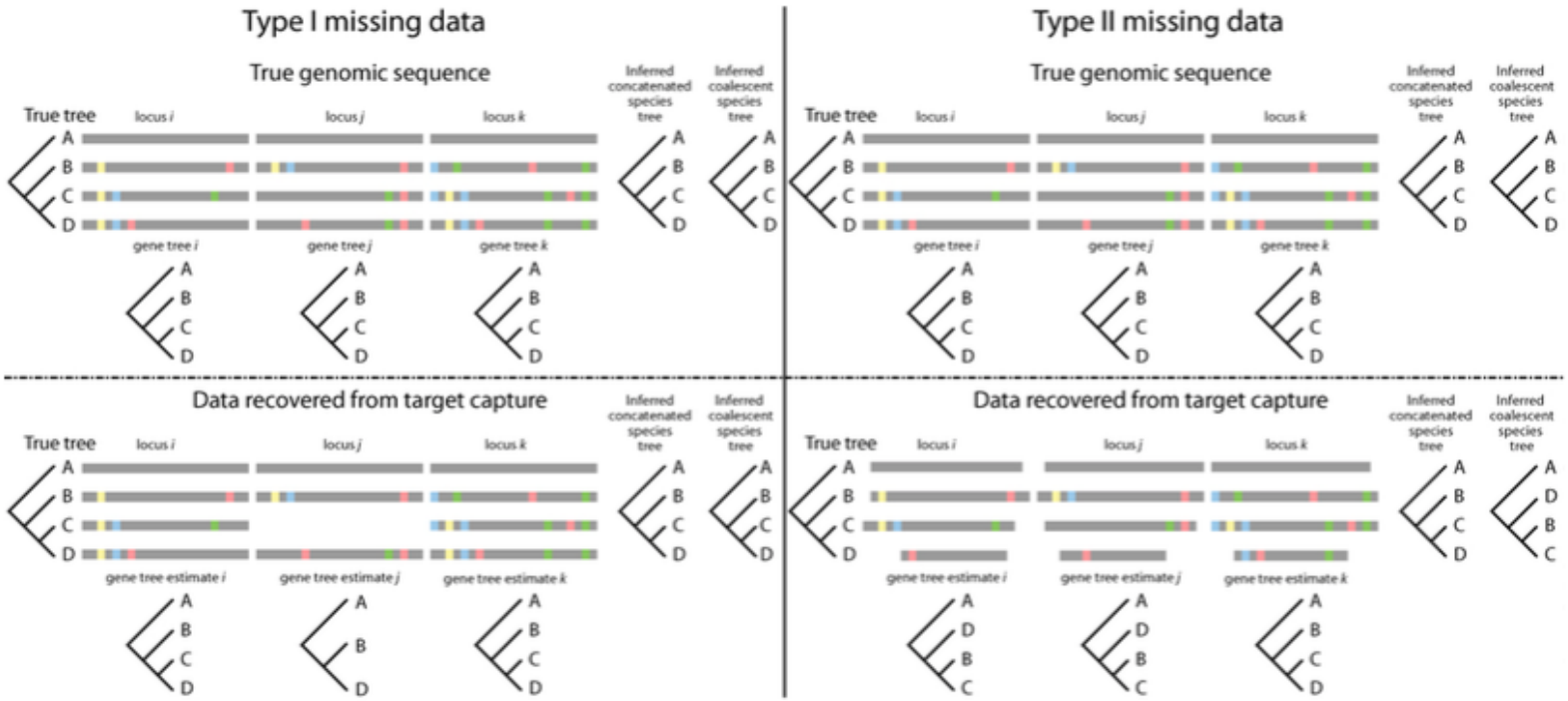
- **type I:** genes/loci for which some taxa have **no** molecular sequence data
- **type II:** genes/loci for which some taxa have **partial** molecular sequence data

which result in

- an incomplete set of bipartitions
- greater gene tree estimation error as taxa with incomplete sequences may become "rogue"

Question: Is a particular type of missing data more insidious when estimating species trees?

Fig 1: Missing data schematic



A model (biological) problem

The challenges in reconstructing the landfowl tree of life

- landfowl seem to have rapidly increased in number and diversity of species
- a small subset ($\sim 10\%$) of landfowl are endangered/critically endangered species

may be mitigated through the utilization of ultraconserved elements (UCEs) which

- can be sampled from thousands of loci, including those of historical specimens in museums
- contain sufficient phylogenetic signal to resolve previously ambiguous nodes

Methods

Data matrix is analyzed allowed for five **thresholds** of missing data:

- 0% type I missing data
- 5% type I missing data
- 25% type I missing data
- 50% type I missing data
- total evidence: all loci regardless of the type/amount of missing data

Here "25% type I missing data" means loci with sequences from at least 75% of taxa are included in species tree estimation.

This threshold approach differs from simulation studies of missing data by Xi et al., where "25% missing data" means 25% of sequences are missing from the complete data matrix.

Table 1: Concatenated 90-taxon dataset summary

	Alignment				
	0%-missing taxa in loci	< 5%-missing taxa in loci	< 25%-missing taxa in loci	< 50-missing taxa in loci	Total-evidence
Loci	140	1,740	3,361	3,919	4,817
Informative loci	140	1,739	3,329	3,868	4,638
Base pair	75,998	838,164	1,416,592	1,573,308	2,208,355
Variable sites	16,623	171,986	270,380	289,840	363,562
Informative sites	10,506	105,886	161,224	170,913	179,676
Avg. bootstrap %	95.0	98.8	98.6	98.8	98.8
Partitions	7	28	27	45	34
% missing sites	10%	12%	15%	18%	39%
% informative sites	14%	13%	12%	11%	8%
% variable sites	22%	21%	19%	19%	17%
Avg. locus length	536	475	417	397	452

Methods

Sequences were

- aligned using MAFFT 7
- ends trimmed when missing 35% of cells (?)

and then used to estimate species trees via

- Concatenated maximum likelihood approach (500 bootstrap replicates)
  - fast (RAxML 8.1.1 GTRCAT)
  - thorough (RAxML 8.1.1 GTRGAMMA)
  - partitioned (PartitionFinder 1.1, RAxML 8.1.1 GTRGAMMA)

Methods, cont.

- "Quartet-based" approach (Quartets are estimated from concatenated data matrix)
  - SMRT-ML
    - Estimate rooted triples, i.e., all quartets contained Oxyura jamaicensis (RAxML 8.1.1 GTRGAMMA)
    - Apply MRP (PAUP* 4.0b10) or QMC (QMC 3.0)
  - SVDquartets (PAUP* 4.1a146 + QMC 3.0)
- "Gene tree reconciliation" approach
  - Estimate gene trees (RAxML 8.1.1, GTRGAMMA, 100 bootstrap replicates)
  - Consider phylogenetic informativeness of gene trees
    - Compute phylogenetic informativeness as number of parsimony informative sites (PhyDesign)
    - Include the 100%, 50%, 25%, 5% of phylogenetically informative trees, i.e., trees with the most
  - Apply ASTRAL 4.4.4
    - ASTRAL-I did not support missing data through adding the missing taxa to their respective gene trees.
  - Apply ASTRID (no version specified on Github)

Biological Results

Most of the results emphasize **relationships between taxa** as determined by the species tree estimation methods on different amounts of missing data, their level bootstrap support, and their support from previous studies.

Authors seemed to prefer results from concatenation followed by quartet-based methods.

Evaluation of the impact of missing data

- Compute majority-rule consensus trees (PAUP* 4.0b10)
- Create a pairwise RF tree distance matrix (R 3.1)
- Visualize the multidimensional scaling (MDS) of the distance matrix (Hillis et al. 2005)

Analysis of the impact of missing data, cont.

Essentially, MDS is used to visualize observed distances between trees by mapping the pairwise distance matrix into a two-dimensional space.

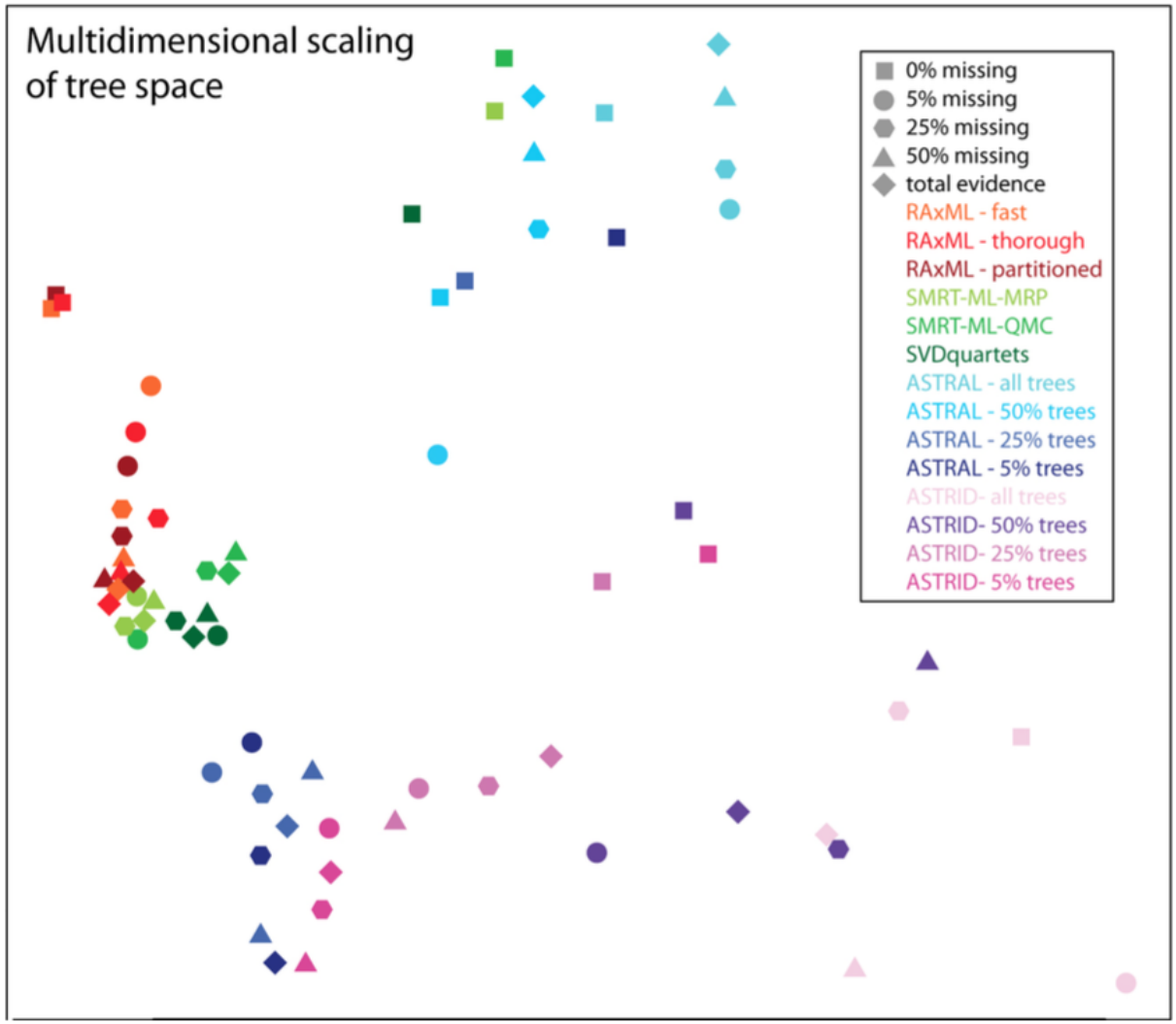
Given an RF distance matrix, D, and point locations in refined space, $\{x\}$, the Kruskal-1 function,

$$S_D(x_1, \ldots, x_n) = \sqrt{\sum_{i=1, i \neq j}^{n} \left(D_{ij} - |x_i - x_j|\right)^2}$$

is minimized to find the optimal solution $\{x\}$ to the MDS problem.

from Hillis D.M., Heath T.A., St. John K. (2005). Analysis and Visualization of Tree Space *Systematic Biology* 54(3): 471–82.

Fig 2: MDS visualization of tree space



Results

Concatenation and quartet-based (SMRT-ML, SVDQuartets) approaches

- less variable with respect to the amount of missing data
- produce similar trees

Gene tree reconciliation approaches (ASTRAL, ASTRID)

- most variable with respect to the amount of missing data
- also highly variable with respect to the collection of trees given as input

Discussion

Hypothesis: Type II missing data

- decreases phylogenetic informativeness of gene trees $\rightarrow$ fewer [informative] sites
- increases gene tree estimation error $\rightarrow$ rogue taxa

which would explain why ASTRAL and ASTRID produce variable results.

New question: "How should researchers decide which gene trees to include?"

Remarks on manuscript

Questions on MSD visualization:

- What is the input to the majority-rule consensus tree?
- Why not use the final landfowl tree of life or a single tree, e.g., thorough concatenation on 0% missing data instead?

Remarks on manuscript, cont.

It is difficult to draw conclusions on the impact of missing data on each method based on the MSD visualization alone. It may be useful to consider additional visualizations, for example,

- Plot RF distance (between species tree on 0% missing data threshold and the other thresholds) versus the threshold of missing data for each method.
- Repeat using the final landfowl tree of life

which are not included in the Supplementary Data.

The effects of type I versus type II data cannot be isolated as type II data only appears in the "total evidence" dataset, which also has type I data.