A Bayesian Supetree Model for Genome-Wide Species Tree Reconstruction Syst. Biol (2014)

De Oliveria Martins & Posada

Centre national de la recherche scientifique (CNRS)
(French National Center for Scientific Research)
Station d'Ecologie Expérimentale du CNRS à Moulis

Last updated: March 13, 2015

1. Introduction
2. Materials and Methods
3. Simulations
4. Results
5. Discussion

Guenomu

Essentially, the paper is presentation of the software guenomu that is used to predict a set of likely species trees, while estimating uncertainty associated with the input gene trees [Martins et al., 2014]

- computationally reasonable
  (447 gene families $\Longrightarrow \sim$ 6 hours on a single processor)
- Gene trees can be of any input format (i.e. CAT family of models)
  - Posterior distributions of unrooted gene tree topologies
- Allows for multiple leaves on a species branch

- Gene tree evolution is not an exact representation of a species tree:
  DL- duplication/loss, **ILS**- incomplete lineage sorting, **HGT**- horiz. gene transfer
- Classes of species tree inference based on gene trees/alignments:
  - Supermatrix - also a 'supergene' approach where genes are concatenated into a single large alignment
  - Supertree - phylogenetic analysis for each alignment followed by species tree inference
  - Model-based - Probabilistic modeling of the incongruence between species and gene trees (subclass of other classes)
- **Robinson-Foulds (RF) supertree** – finds a species tree that minimizes disagreement with gene tree collections without explicitly taking into account biological phenomena
  [Bansal and Eulenstein, 2013]
- **Gene tree parsimony (GTP)** – Finding the species tree that minimizes the reconciliation cost [Guigó et al., 1996]

Model-based inference of species trees

- Most supertree approaches neglect gene tree branch lengths
- Multispecies coalescent - Another class of methods closely related to supertree methods that try to reconstruct a species tree based on a matrix of distances between species (for review [Liu et al., 2009])
- Gene trees are often assumed to be known – in other words supertree methods generally do not account for uncertainty

guenomu approach

Input: a set of unrooted gene tree distributions (one per family)* and a list of species names
Output: posterior distribution of the rooted species as well as a posterior distribution of gene trees for each gene family

* Posteriors of gene trees from programs like PhyloBayes [Lartillot and Philippe, 2004] can be used as input

- Let gene family i be a set of homologous sequences that compose an alignment D_i, which can comprise paralogs and orthologs
- Each member of the gene family needs to have a map to a species
- Gene tree (G_i) - any phylogenetic tree connecting all sampled members of gene family i

as a reminder

Likelihood:$P(X|\theta)$ is the probability of the evidence given the parameters
Posterior: $P(\theta|X)$ is the probability of the parameters given the evidence
Prior: $P(\theta)$ is the probability describing the uncertainty about the parameters before evidence is taken into account

Bayesian Hierarchical Model

The joint distribution

$$P(D, \theta|G) = P(D|(\theta|G)) \times P(\theta|G) \tag{1}$$

The posterior distribution of a species tree S is given by

$$P(S, \mathbf{\Theta}|\mathbf{D}) \propto P(\mathbf{D}, \boldsymbol{\theta}|\mathbf{G})P(\mathbf{G}|\boldsymbol{\lambda}, S)P(\boldsymbol{\lambda}|\boldsymbol{\lambda}_0)P(\boldsymbol{\lambda}_0)P(S) \tag{2}$$

$$\propto P(\boldsymbol{\lambda}_0)P(S)\prod_{i=1}^{N} P(D_i, \boldsymbol{\theta}_i|G_i)P(G_i|\lambda_i, S)P(\lambda_i|\lambda_0) \tag{3}$$

where $\Theta = (\boldsymbol{\theta}, \mathbf{G}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_0)$. $\boldsymbol{\lambda} = \lambda_{ij}$ is a matrix with penalty parameters j for gene family i. These penalty parameters are distributed according to $\boldsymbol{\lambda_0}$. They assume $P(G_i|S)$ follows an exponential distribution s.t. $d(G_i, S)$ can be written based on this distribution along with penalty parameter λ_i (see text).

Priors for distance penalties

An exponential prior??

$$P(\lambda_{ij}|\boldsymbol{\lambda}_0) = \frac{e^{\lambda_{ij}/\lambda_{0i}}}{\lambda_{0j}} \tag{4}$$

[Steel and Rodrigo, 2008]

Distances

- Reconciliation distances - based on the most parsimonious reconciliations between rooted species tree and gene trees (can be used to find the minimum number of DL or deep coalescences in order to make speces and gene tree match)
- Nonparametric distances - any estimate of disagreement i.e. Robinson-Foulds (RF) – they do not model the outcome only the disagreement.

The *mulRF* distance allows for one of the two trees in the distance calculation to have several leaves with the same label.

MCMC

It is difficult to sample directly from posterior so a variant of MCMC was used...

- Generalized Multiple-try Metropolis (GMTM)
- The denominator of the multivariate exponential is so messy they had to ignore it
- Simulated annealing – instead of sampling directly from $P(S, \mathbf{\Theta}|\mathbf{D})$ they estimate S and Θ

TABLE 1. Parameter values used in the simulations

	Description	Symbol	Distribution
Species tree *(Dendropy)*	Number of species		Uniform(10, 80)
	Number of generations (total tree height)		Uniform(10^2, 10^4)
	Expected number of duplications	E_{dup}	Uniform(10^{-3}, 4)
	Number of gene families	N	Uniform(2, 50)
	Rate heterogeneity multiplier	H_s	Gamma(1, 1)
Locus tree[a] *(SimPhy)*	Gene duplication rate [b]	β	Exponential(E_{dup}/σ)
	Gene loss rate		Uniform(0, $0.75 \times \beta$)
	Rate heterogeneity multiplier	H_l	Gamma(1, 1)
Gene tree[a] *(SimPhy)*	Effective population size	N_e	$2000\times$ LogNormal(0, 0.25)
	Number of individuals per species [c]		Uniform(1, 10)
	Substitution rate (per time unit)		0.001
	Rate heterogeneity multiplier	H_g	Gamma(1, 1)
Gene tree Uncertainty (in-house)	Maximum number of generated trees		160
	Tree dispersion term	L_T	Uniform(2, 5)
	Tree location term	D_T	Uniform(3, 6)
	Frequency of trees with uncertainty	p_T	$1.1\times$ Beta($L_T \times D_T$, D_T)
	Branch location term	L_B	Uniform(1, 5)
	Branch-wise error probability	p_B	$1.5\times$ Beta(L_B, 1)

Notes: Each simulation replicate was parameterized with values sampled from predefined statistical distributions. Given a species tree, simulated with *Dendropy*, the program *SimPhy* was used to simulate gene family trees. An in-house program was then used to mimic gene tree uncertainty, transforming each gene tree in a distribution of gene tree topologies in an attempt to emulate the effect of alignment-based phylogenetic inference in practice.

[a] terminology used by *SimPhy*

[b] the term σ is the total sum of branch lengths (in generations) in the species tree

[c] common for all gene families, emulating perfect sampling of individuals

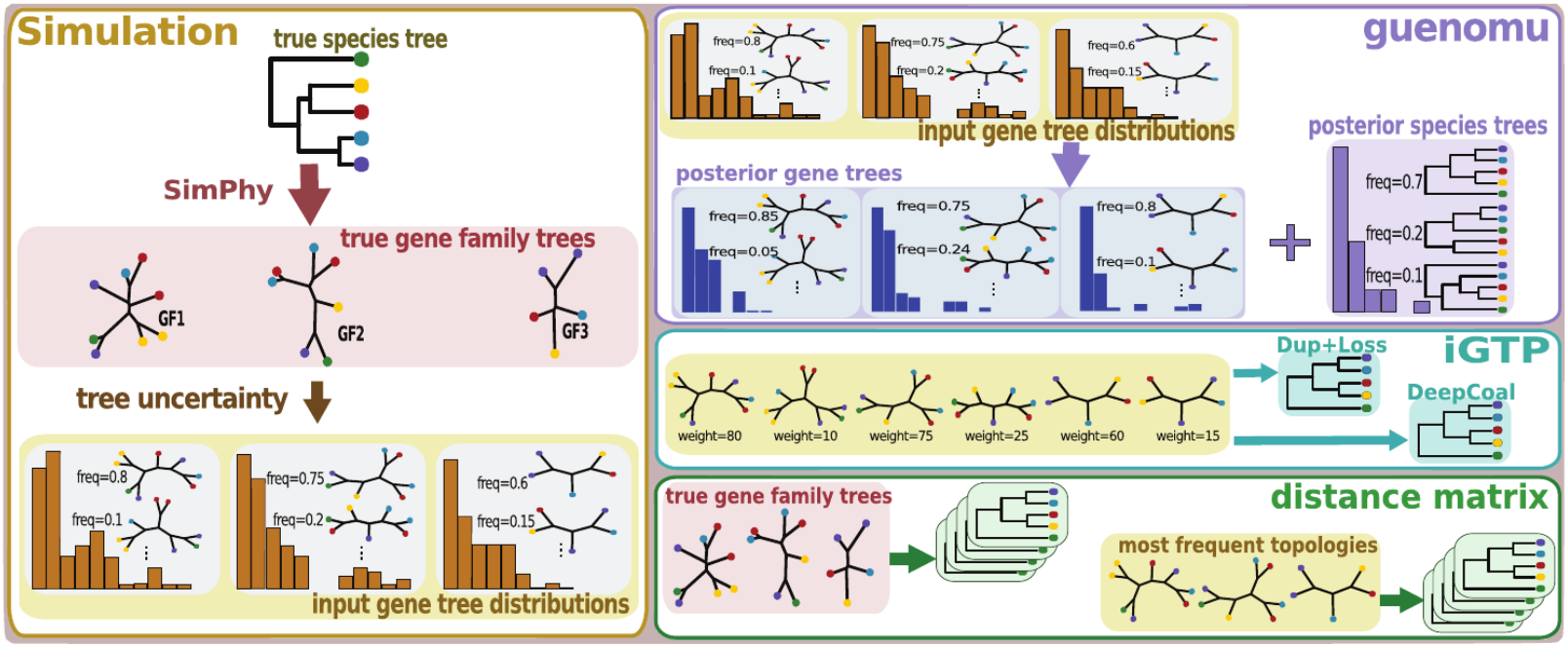

FIGURE 1. Simulation workflow. On the left a single data set is produced using *Dendropy* and *SimPhy*. The true species tree (rooted, with branch lengths) simulated by *Dendropy* is used as input by *SimPhy* to generate several (rooted) phylogenies, one per gene family. Then, uncertainty is added such that we have a distribution of topologies (unrooted, no branch lengths) per gene family. These collections will be used as input to the inference programs in the right panel. Our software *guenomu* estimates the posterior distribution of species trees (rooted, without branch lengths) and the posterior distributions of gene family topologies, based on all input gene tree distributions. The software *iGTP* also uses the input gene trees after transforming the frequencies into integer values representing weights, and estimates two rooted species trees: one under the Duplication and Loss cost, and another under the Deep Coalescence cost. For the distance-based species tree inference algorithms only one tree per gene family is used, and two alternative choices were attempted: one was to use the true gene families, with branch lengths; and the other was to use the most frequent gene trees (topologies only) after introducing phylogenetic uncertainty. In the later case it was assumed that all branches had the same length.

Notation

- DLI - a parameterization of the model that only considers reconciliation distances (min num. DL and ILS)
- DLIR - same except it also considers *mulRF*
- iGTP - a competing software

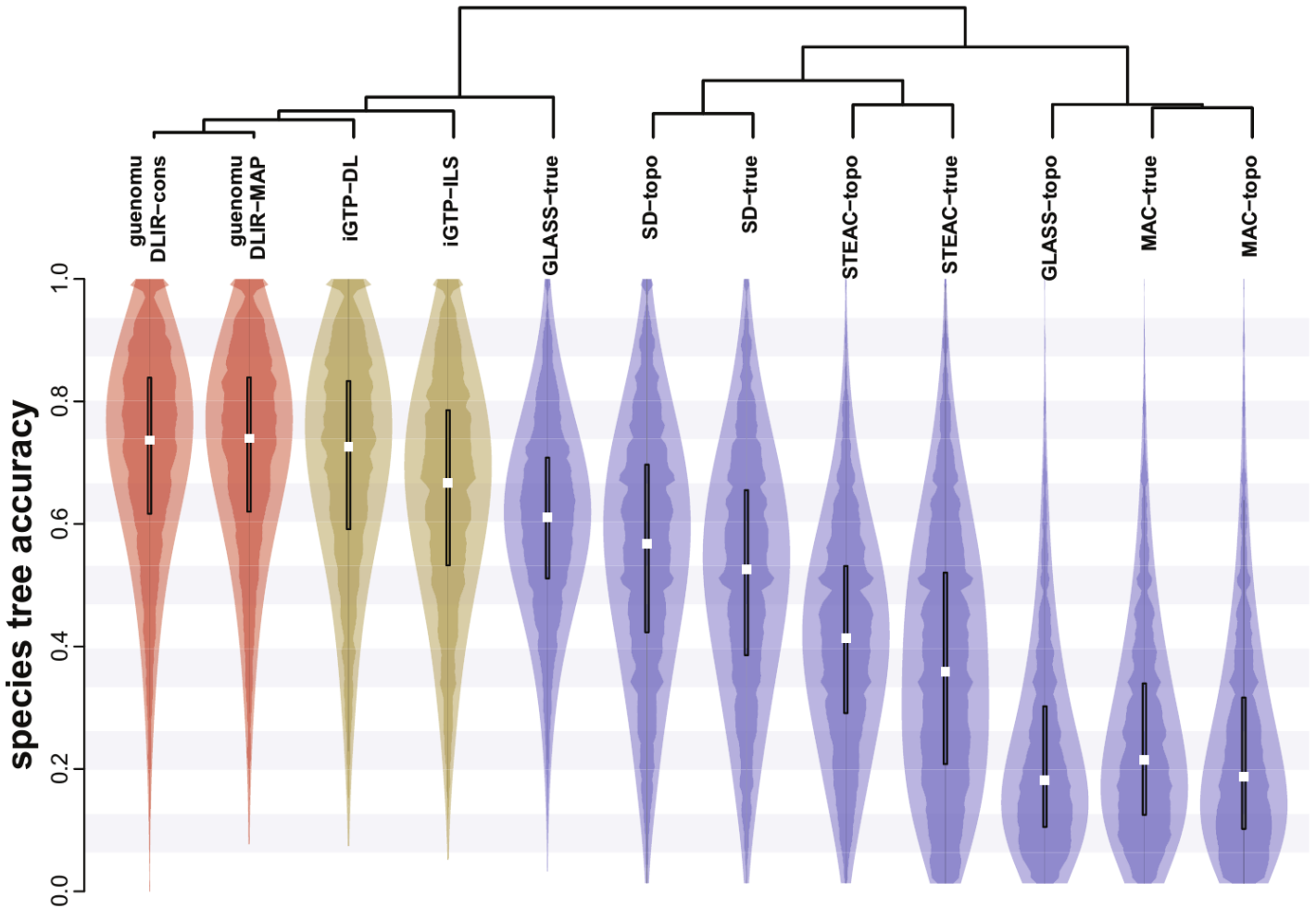

FIGURE 2. Species tree accuracy of several inference methods. Each violin plot (kernel density curves plus a boxplot) represents the distribution of tree accuracy values for the different methods evaluated. From each posterior distribution estimated by *guenomu* under the DLIR parameterization we obtained two point estimates of the species tree: the MAP tree and the consensus tree (labeled 'MAP' and "cons"). We also obtained two estimates of species trees using *iGTP*, running the program under the DL cost and under the ILS cost ("*iGTP*-DL" and "*iGTP*-ILS"). We also ran four distance matrix approaches (GLASS, SD, STEAC, and MAC) using two types of data sets, just the topologies with uncertainty ('topo') or the true simulated gene family trees with branch lengths ("true"). At the top, we show the hierarchical clustering of the different methods based on their tree accuracy values for all replicates.

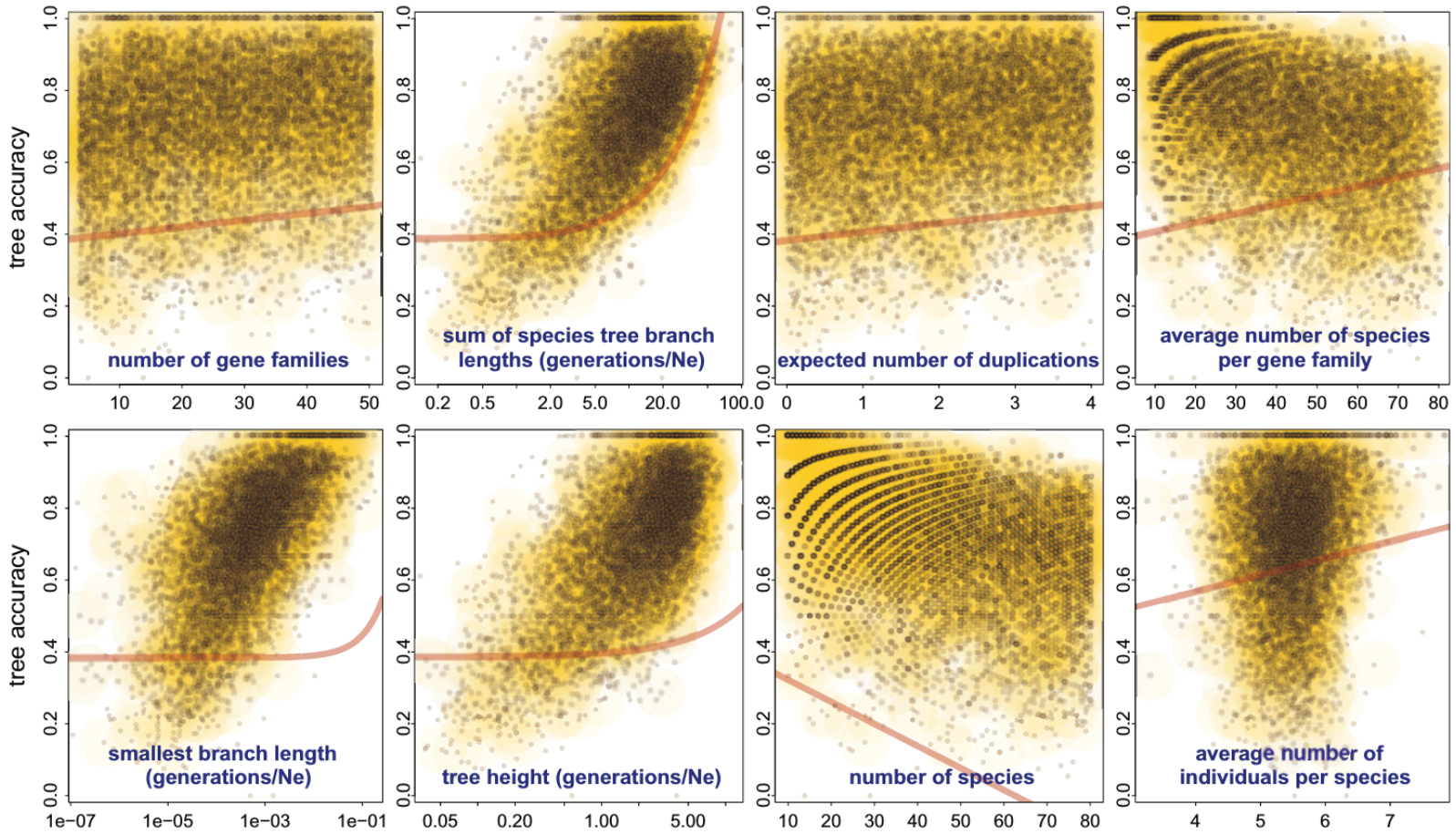

FIGURE 3. *Guenomu*'s tree accuracy with respect to simulation parameters. Here, the species tree was estimated as the consensus tree from the posterior distribution assuming a DLIR parameterization. Based on a multiple linear regression analysis, these are the parameters that most significantly affected tree accuracy: at the top we have, from left to right, the total number of gene families, the total species tree length (in coalescent units), the expected number of duplications per gene family and the average number of species represented by gene family; at the bottom we have the length of the smallest branch in the species tree (in coalescent units), the height of the species tree from root to tips (scaled by the effective population size), the total number of species on the species tree and the average number of individuals from same species per gene family. Over each panel we show the regression line overlaid.

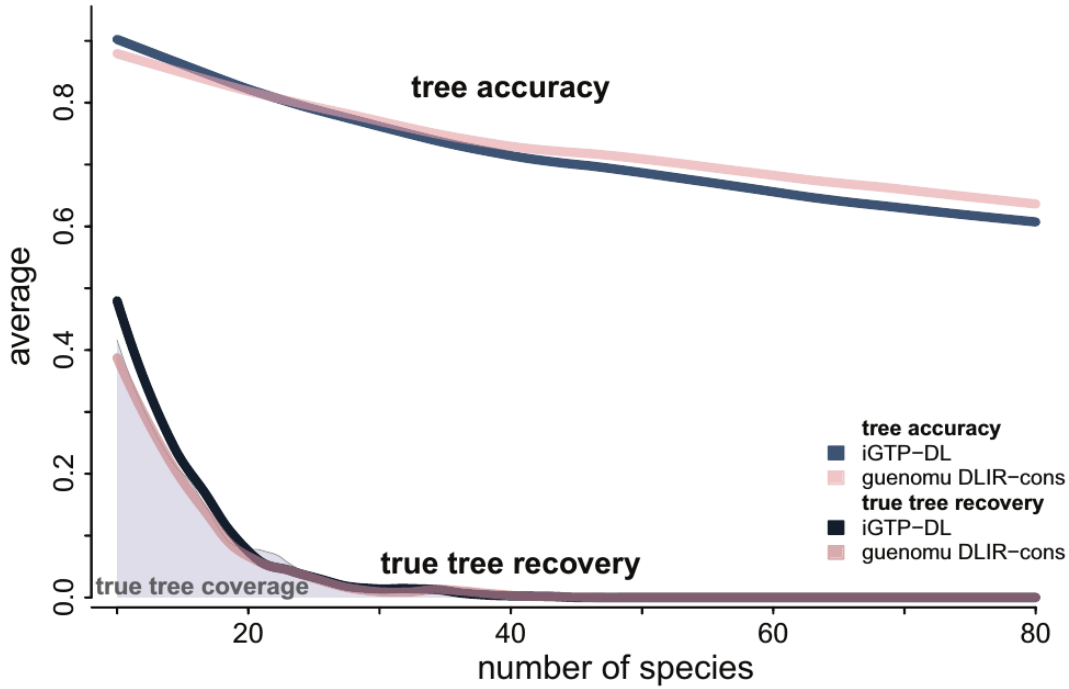

FIGURE 4. Tree accuracy and true tree recovery for *guenomu* and *iGTP* according to the number of species. The lower lines show the true tree recoveries pooled over the number of species, whereas the higher lines show the average tree accuracy under different conditions. The gray area represents the true tree coverage under *guenomu* DLIR model (i.e., the proportion of data sets where the true species tree was within the 95% credible set). Each line is actually a smooth regression line over the individual values. For *iGTP* the species tree was estimated under the DL cost, whereas for *guenomu* we plot the Bayesian consensus under the DLIR parameterization.

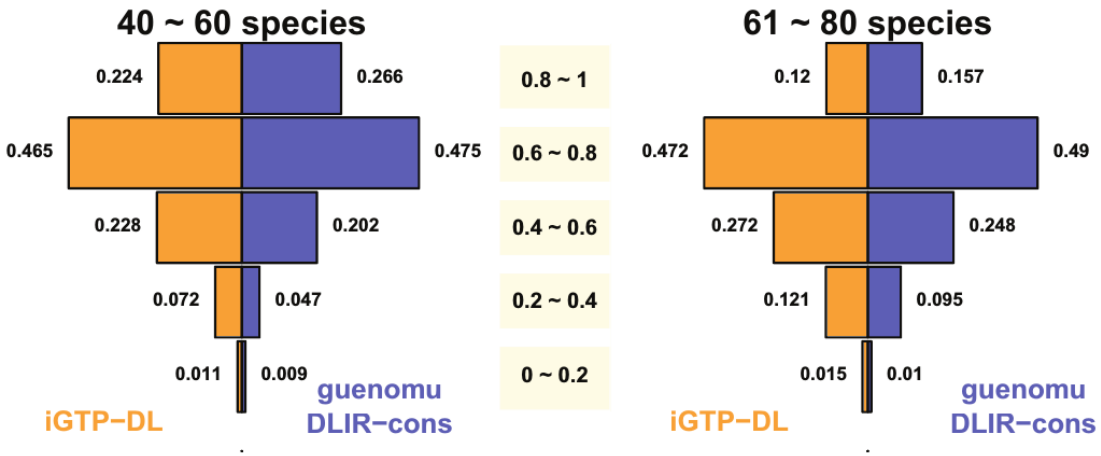

FIGURE 5. Tree accuracies of *iGTP* and *guenomu* for large data sets. On the left we have all data sets where the number of species was between 40 and 60, and on the right we have the pooled replicates with species tree sizes between 61 and 80 leaves. Each bar represents the proportion of data sets with tree accuracies between the ranges described in the middle, such that at the top we have the fraction of simulations where more than 80% of the splits were recovered, while the inferences whereas less than 20% of the true splits were found are the bottom.

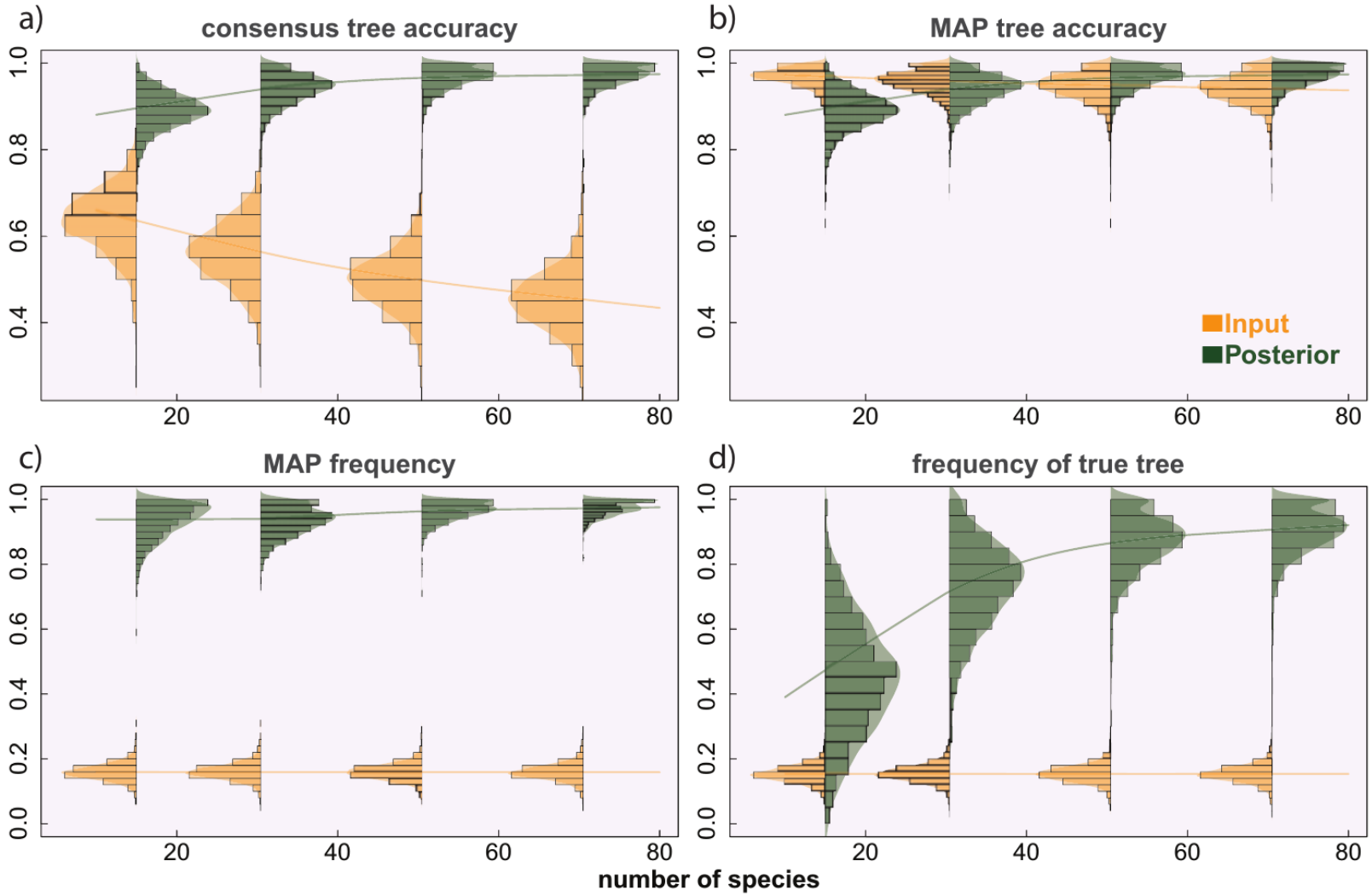

FIGURE 6. Input and posterior gene tree distributions. Each panel shows the distribution of input gene trees (after generation of tree uncertainty using our algorithm) and their posterior counterparts (resampled by *guenomu*) for several ranges of species tree sizes, together with a smooth regression line over all samples. The panels at the top show the accuracies of the consensus (a) and MAP (b) estimates, when compared to the true gene trees simulated by *SimPhy*, whereas the bottom panels display the frequencies of the MAP (c) and true (d) gene trees. All values are averages over all gene families from each replicate.

- Usually different sources of gene tree disagreement are considered separately (e.g. DL,ILS), however unrecognized processes adversely affect gene tree estimation [Rasmussen and Kellis, 2012]
- guenomu does well, GTP methods do well, but coalescence methods perform poorly under the scenarios considered


Bansal, M. S. and Eulenstein, O. (2013).
Algorithms for genome-scale phylogenetics using gene tree parsimony.
IEEE/ACM Trans. Comput. Biology Bioinform., 10(4):939–956.


Guigó, R., Muchnik, I., and Smith, T. F. (1996).
Reconstruction of ancient molecular phylogeny.
Molecular phylogenetics and evolution, 6(2):189–213.


Lartillot, N. and Philippe, H. (2004).
A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.
Molecular biology and evolution, 21(6):1095–109.


Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009).
Coalescent methods for estimating phylogenetic trees.
Molecular phylogenetics and evolution, 53(1):320–8.


Martins, L. D. O., Mallo, D., and Posada, D. (2014).
A bayesian supetree model for genome-wide species tree reconstruction.
Systematic Biology, Advance Access:1–20.


Rasmussen, M. D. and Kellis, M. (2012).
Unified modeling of gene duplication, loss, and coalescence using a locus tree.
Genome research, 22(4):755–65.


Steel, M. and Rodrigo, A. (2008).
Maximum likelihood supertrees.
Systematic biology, 57(2):243–50.