Supplementary Materials

Testing the Reliability of Genetic Methods of Species Identification via Simulation

Howard A. Ross, Sumathi Murugan and Wai Lok Sibon Li
Bioinformatics Institute, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand

contact: h.ross@auckland.ac.nz

Tables

Table S1. The mean (SD) rate of true or false positive identification of species when sequences were simulated on species trees under different values of rate heterogeneity *a* (0.5, 0.7, and 0.9).

Table S2. The change in the rate of identification outcomes when sequences were simulated under more complex models.

Figures

Figure S1. The rates of successful identification of the query, when the true species was present in the reference sequence dataset, using the four methods for assessing BLAST results. The results are presented as quartile box plots of the observed proportion per replicate.

Figure S2. The rates of successful identification of the query, when the true species was present in the reference sequence dataset, using the two methods for assessing distance-based results. The results are presented as quartile box plots of the observed proportion per replicate.

Figure S3. The rates of successful identification of the query, when the true species was present in the reference sequence dataset, using the tree-based methods for assigning identity: lib = tree (liberal), st = tree (strict). Identical rates of successful identification occurred when a distance threshold was used. The results are presented as quartile box plots of the observed proportion per replicate.

Figure S4. The rates of false positive identification of the query, when the true species was absent in the reference sequence dataset, using the tree-based methods, with and without a distance threshold, for assigning identity: lib = liberal, lib+t = liberal + threshold, st = strict, st+t = strict + threshold. The results are presented as quartile box plots of the observed proportion per replicate.

Figure S5. The variation in the incidence of monophyly among species in relation to the number of sequences sampled (range 2 - 20). In each species, one sequence was subsequently used as the query and the remaining N-1 sequences (range 1 - 19) formed the reference set.

Figure S6. The variation in the outcome when BLAST was used to identify a query in relation to the number of reference sequences (range 1 - 19) for the species.

Figure S7. The variation in the outcome when the distance method was used to identify a query in relation to the number of reference sequences (range 1 - 19) for the species.

Figure S8. The variation in the outcome when the liberal tree method was used to identify a query in relation to the number of reference sequences (range 1 - 19) for the species.

Figure S9. The variation in the outcome when the strict tree method was used to identify a query in relation to the number of reference sequences for the species. This method is only defined when there are two or more reference sequences (range 2 - 19).

Table S1. The mean (SD) rate of true or false positive identification of species when sequences were simulated on species trees under different values of rate heterogeneity *a* (0.5, 0.7, and 0.9).

Method	**Evolutionary Scenario (theta, species birth rate)**	**True Positive ID (species present)**			**False Positive ID (species absent)**		
		0.5	**0.7**	**0.9**	**0.5**	**0.7**	**0.9**
BLAST	0.01, 1	0.998 (0.006)	0.999 (0.003)	0.998 (0.006)	0.300 (0.086)	0.281 (0.086)	0.258 (0.087)
	0.01, 30	0.938 (0.040)	0.951 (0.033)	0.938 (0.036)	1.000 (0.000)	1.000 (0.002)	1.000 (0.000)
	1, 1	0.772 (0.061)	0.754 (0.058)	0.742 (0.062)	0.195 (0.075)	0.171 (0.066)	0.156 (0.068)
Distance	0.01, 1	0.996 (0.009)	0.997 (0.007)	0.996 (0.009)	0.042 (0.047)	0.044 (0.045)	0.048 (0.047)
	0.01, 30	0.938 (0.038)	0.952 (0.034)	0.944 (0.036)	0.714 (0.127)	0.736 (0.120)	0.746 (0.117)
	1, 1	0.847 (0.050)	0.839 (0.054)	0.833 (0.052)	0.365 (0.092)	0.325 (0.094)	0.291 (0.079)
Liberal tree-based	0.01, 1	0.998 (0.006)	0.999 (0.004)	0.998 (0.007)	0.640 (0.098)	0.645 (0.105)	0.637 (0.096)
	0.01, 30	0.914 (0.046)	0.928 (0.045)	0.916 (0.046)	0.665 (0.065)	0.659 (0.065)	0.669 (0.060)
	1, 1	0.833 (0.054)	0.828 (0.055)	0.822 (0.055)	0.843 (0.103)	0.879 (0.103)	0.908 (0.089)
Strict tree-based	0.01, 1	0.943 (0.035)	0.944 (0.029)	0.948 (0.032)	0.011 (0.020)	0.010 (0.018)	0.012 (0.021)
	0.01, 30	0.800 (0.061)	0.811 (0.062)	0.805 (0.075)	0.213 (0.057)	0.220 (0.068)	0.227 (0.069)
	1, 1	0.631 (0.065)	0.626 (0.074)	0.606 (0.075)	0.152 (0.061)	0.137 (0.062)	0.119 (0.057)

Table S2. The change in the rate of identification outcomes when sequences were simulated under more complex models. For each evolutionary scenario, the first value is the change in the outcome when sequences were simulated using (a) variable rather than fixed values of theta, and the second value is the change when simulations used (b) variable theta and a complex model of evolution rather than a fixed theta and simple model. Variable theta values were drawn from a Gaussian distribution with mean equal to the fixed value and standard deviation equal to 20% of the mean. The complex model of sequence evolution used GTR + G + I rather than HKY.

Method	**Evolutionary Scenario (theta, species birth rate)**	**Identification Outcome**			
		Correct ID	**Uncertain ID**	**Incorrect ID (species present)**	**False Positive ID (species absent)**
BLAST	0.01, 10	0.003, 0.002	0.000, 0.000	-0.003, -0.002	0.000, 0.029
	0.03, 15	0.000, -0.018	0.000, 0.000	0.000, 0.018	-0.002, 0.004
	0.1, 30	0.007, 0.004	0.000, 0.000	-0.007, 0.004	0.000, 0.000
Distance	0.01, 10	0.004, 0.000	-0.001, 0.000	-0.003, 0.000	0.005, -0.010
	0.03, 15	-0.004, -0.019	0.001, 0.000	0.003, 0.020	0.018, 0.042
	0.1, 30	0.006, -0.008	0.000, 0.000	-0.006, 0.008	0.001, 0.001
Liberal tree-based	0.01, 10	-0.002, 0.006	0.002, 0.005	0.000, -0.001	-0.006, 0.001
	0.03, 15	-0.003, -0.016	0.003, 0.001	0.000, 0.016	-0.016, -0.016
	0.1, 30	0.005, -0.008	-0.003, -0.003	-0.003, 0.010	0.003, -0.008
Strict tree-based	0.01, 10	-0.006, -0.006	0.005, 0.007	0.000, -0.001	-0.002, -0.012
	0.03, 15	0.012, -0.004	-0.011, 0.006	-0.001, -0.002	-0.003, 0.017
	0.1, 30	-0.005, 0.000	0.004, -0.001	0.001, 0.001	-0.011, -0.013

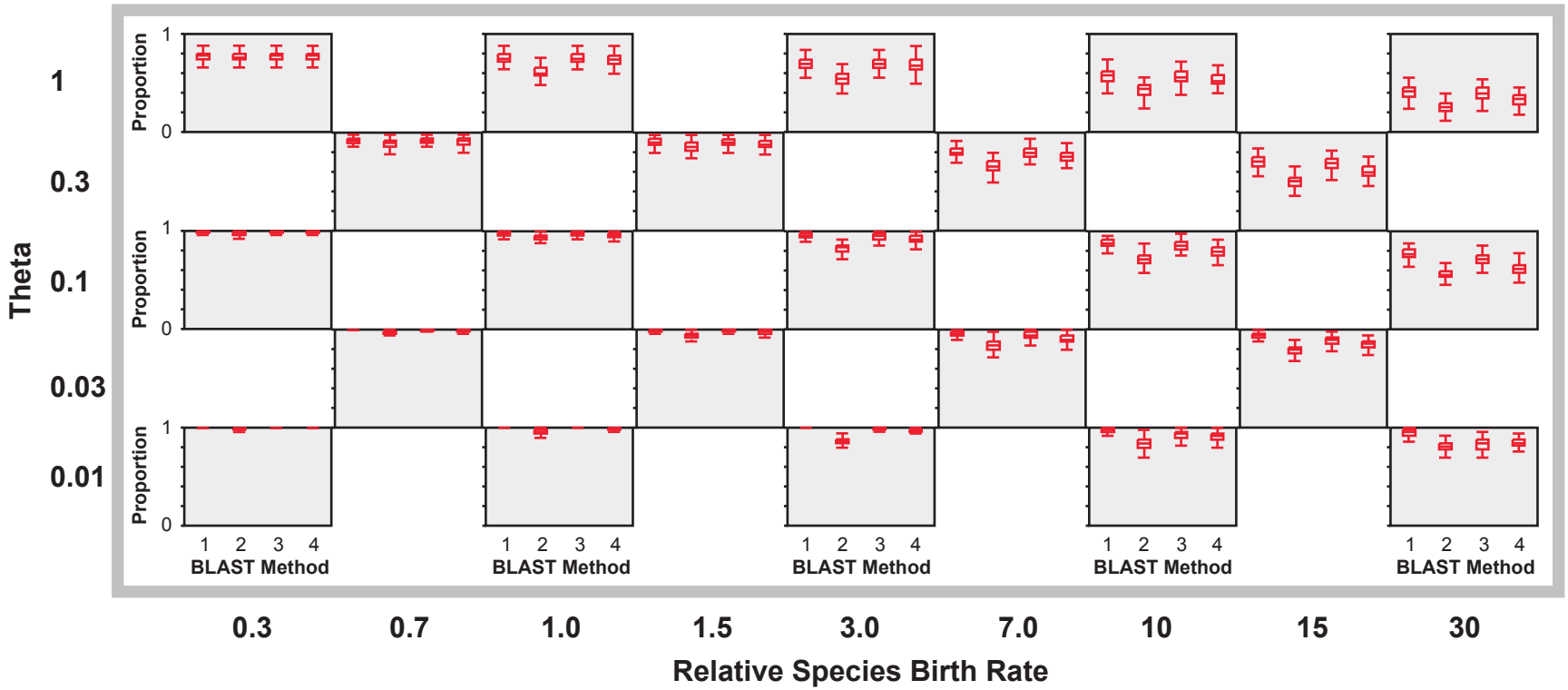

Figure S1. The rates of successful identification of the query, when the true species was present in the reference sequence dataset, using the four methods for assessing BLAST results. The results are presented as quartile box plots of the observed proportion per replicate.

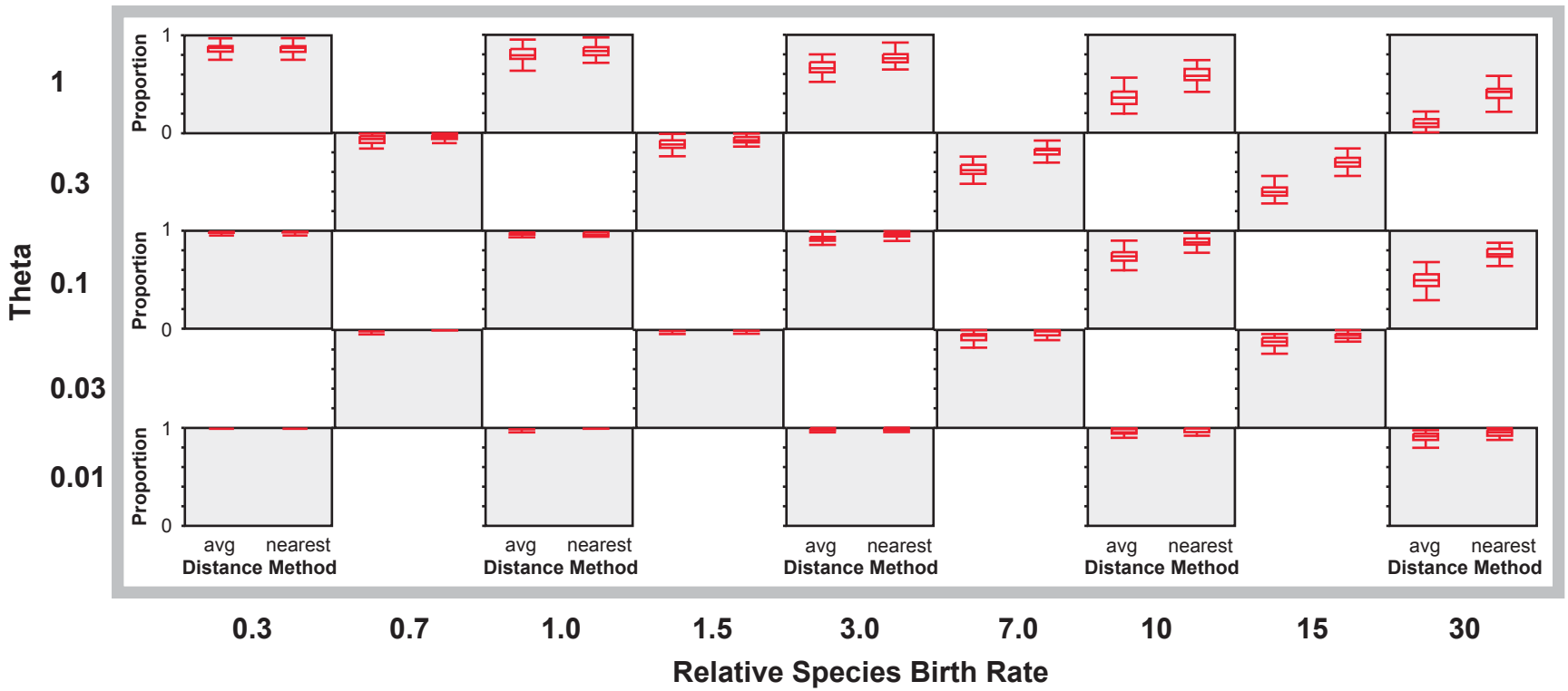

Figure S2. The rates of successful identification of the query, when the true species was present in the reference sequence dataset, using the two methods for assessing distance-based results. The results are presented as quartile box plots of the observed proportion per replicate.

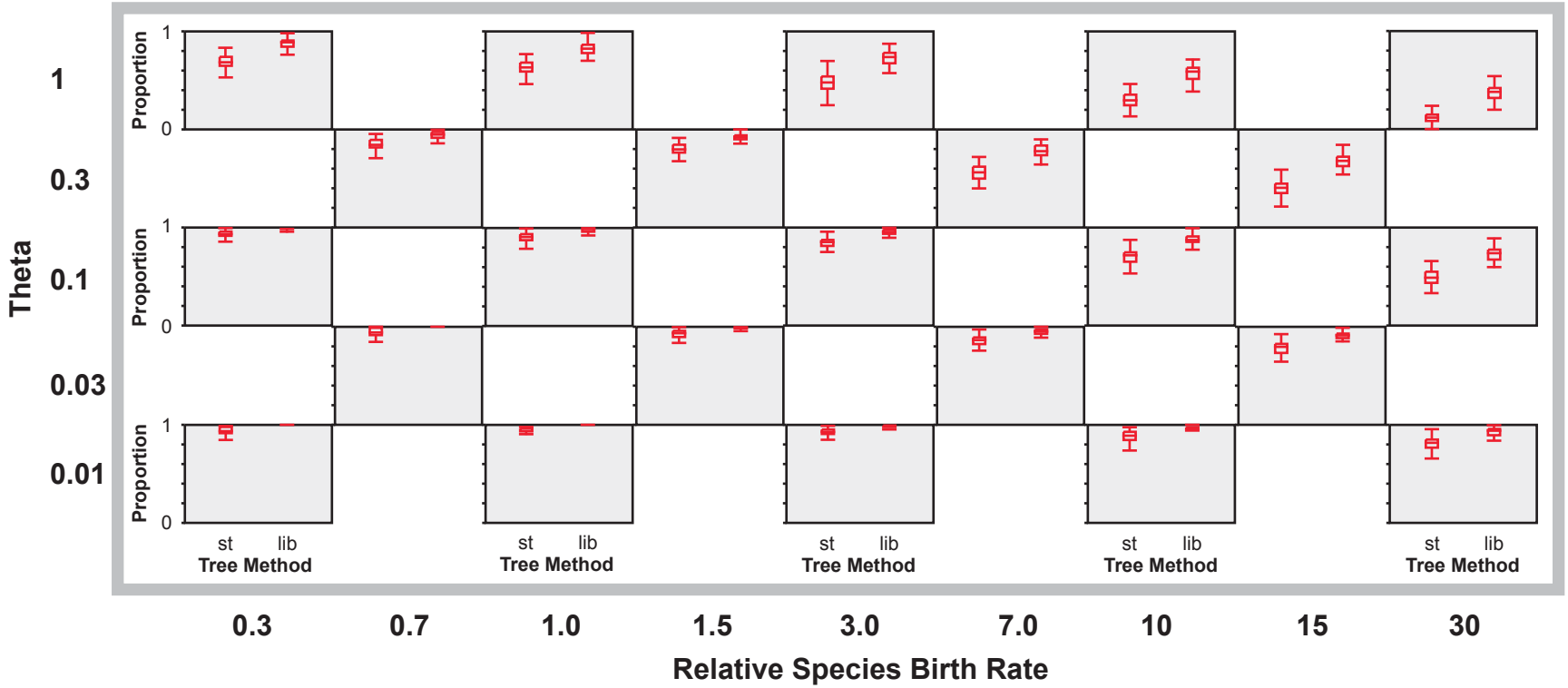

Figure S3. The rates of successful identification of the query, when the true species was present in the reference sequence dataset, using the tree-based methods for assigning identity: lib = tree (liberal), st = tree (strict). Identical rates of successful identification occurred when a distance threshold was used. The results are presented as quartile box plots of the observed proportion per replicate.

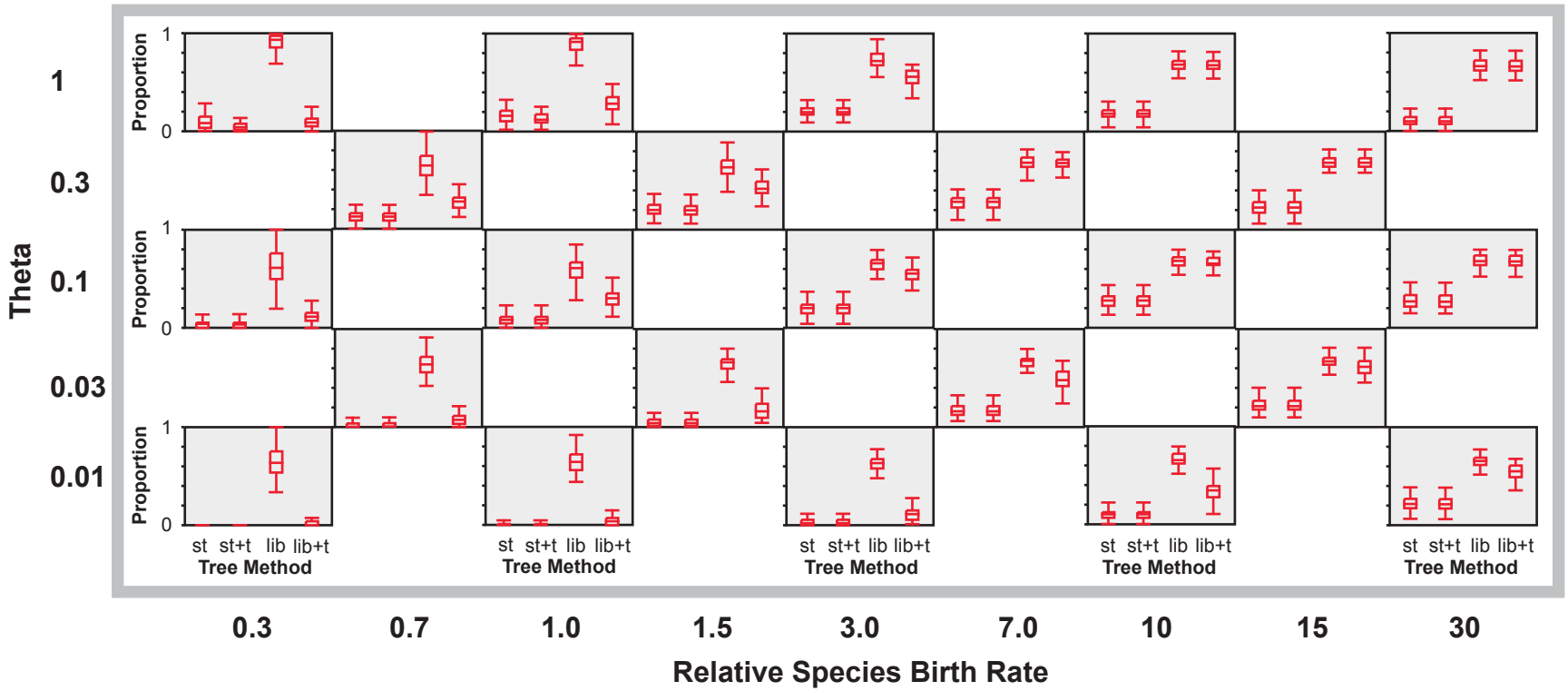

Figure S4. The rates of false positive identification of the query, when the true species was absent in the reference sequence dataset, using the tree-based methods, with and without a distance threshold, for assigning identity: lib = liberal, lib+t = liberal + threshold, st = strict, st+t = strict + threshold. The results are presented as quartile box plots of the observed proportion per replicate.

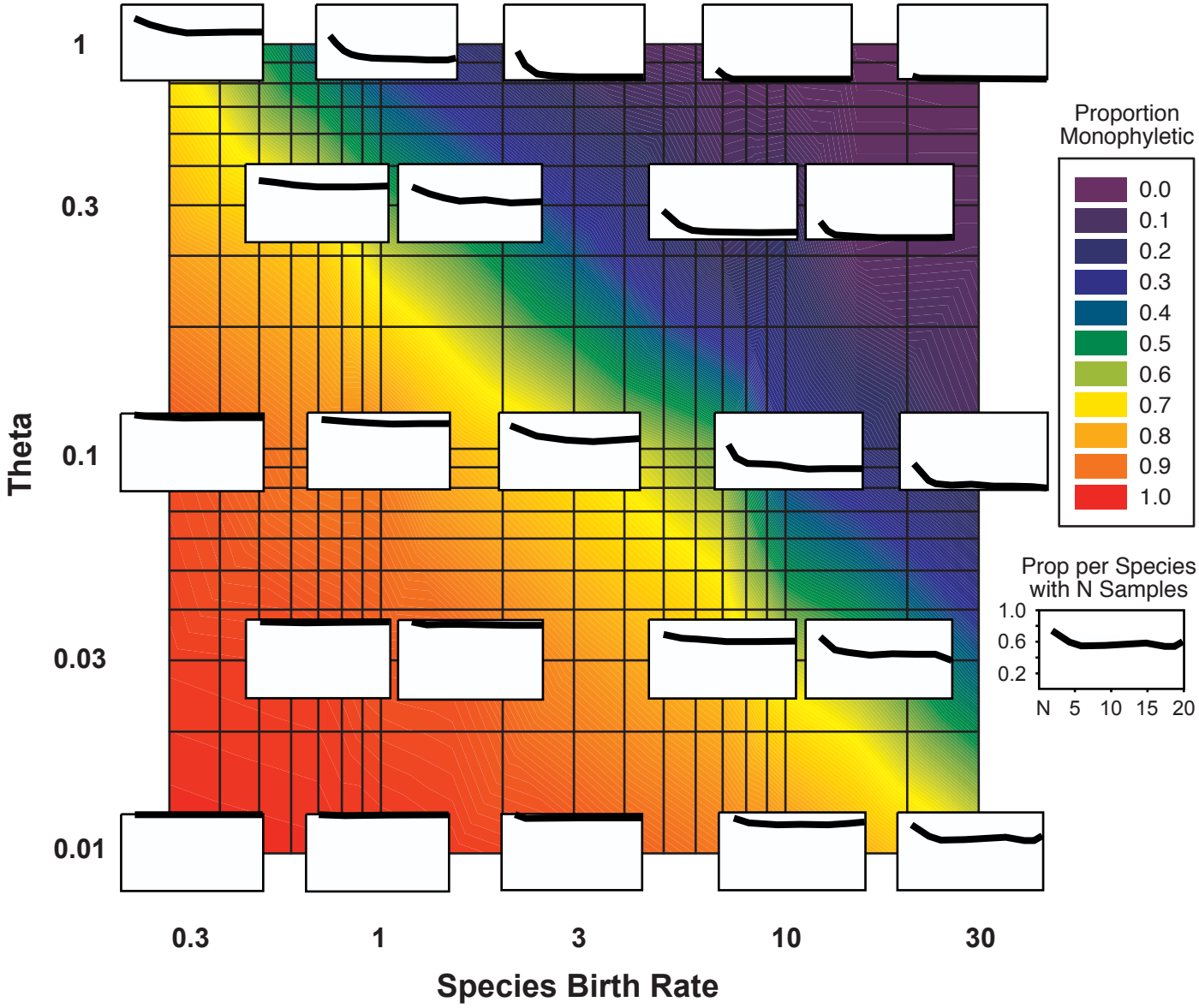

Figure S5. The variation in the incidence of monophyly among species in relation to the number of sequences sampled (range 2 - 20). In each species, one sequence was subsequently used as the query and the remaining N-1 sequences (range 1 - 19) formed the reference set.

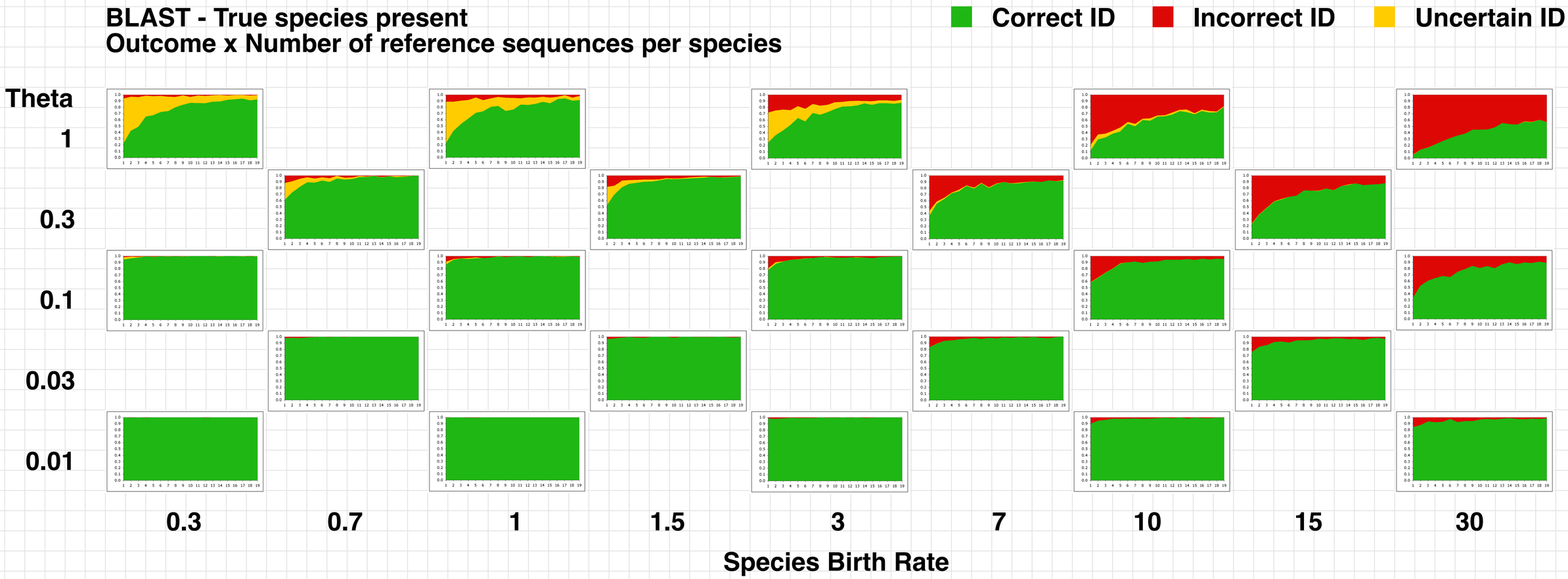

Figure S6. The variation in the outcome when BLAST was used to identify a query in relation to the number of reference sequences (range 1 - 19) for the species.

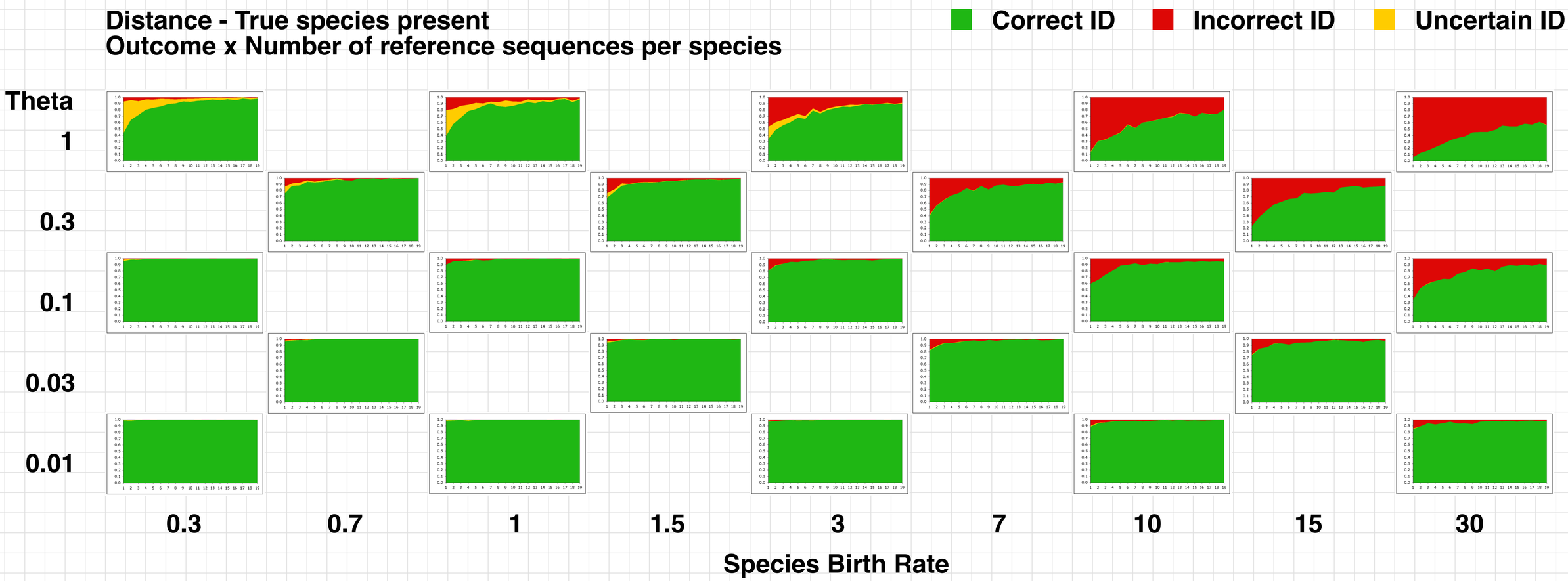

Figure S7. The variation in the outcome when the distance method was used to identify a query in relation to the number of reference sequences (range 1 - 19) for the species.

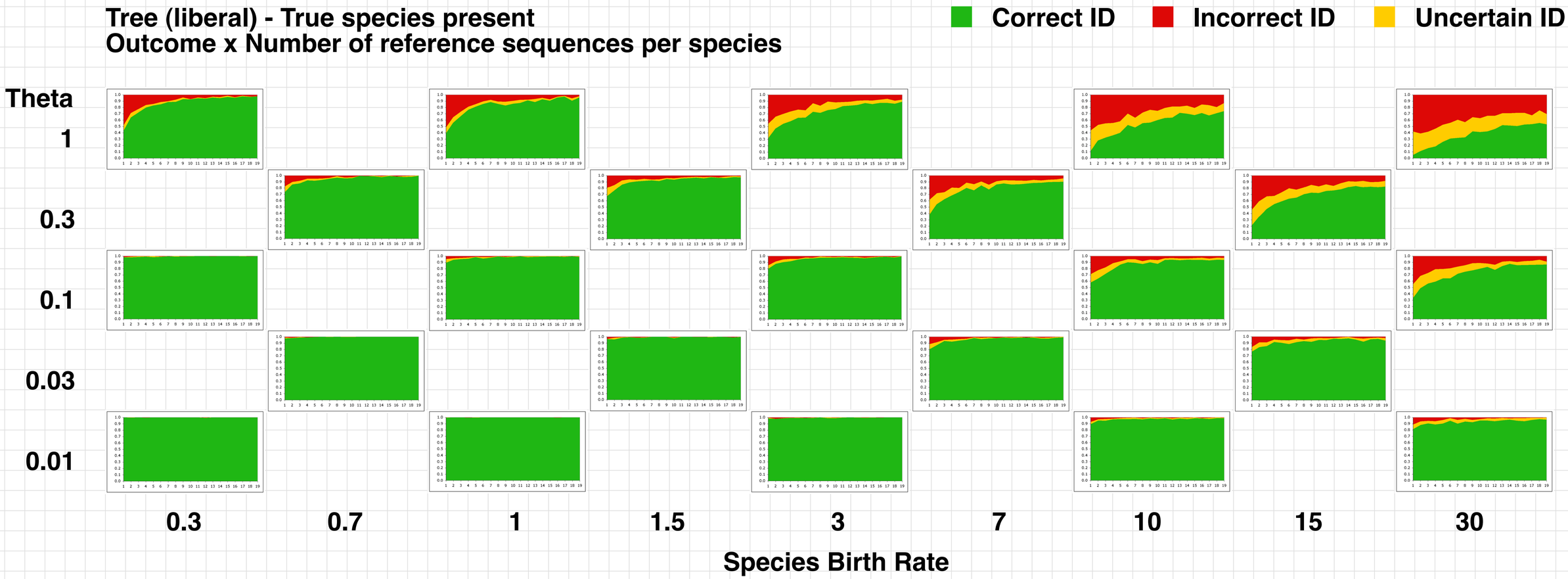

Figure S8. The variation in the outcome when the liberal tree method was used to identify a query in relation to the number of reference sequences (range 1 - 19) for the species.

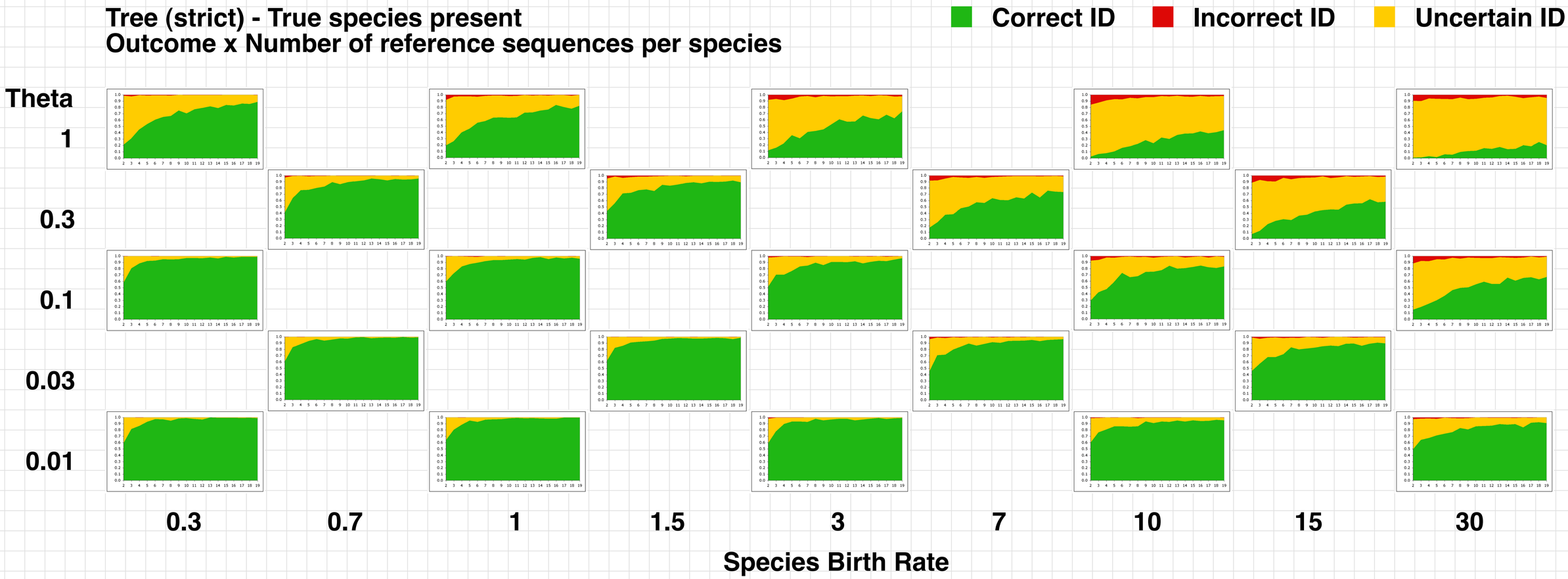

Figure S9. The variation in the outcome when the strict tree method was used to identify a query in relation to the number of reference sequences for the species. This method is only defined when there are two or more reference sequences (range 2 - 19).