最尤系統樹の信頼性に対する樹形探索労力量と初期樹形の影響

筑波大・院・生命環境科学
田辺晶史

系統樹ってなに？

系統樹ってなに？

系統樹の例：類人猿の系統関係

系統樹ってなに？

- 生物の系統関係を樹形図で表したもの

系統樹の例：類人猿の系統関係

系統樹ってなに？

- 生物の系統関係を樹形図で表したもの
- 生物の進化の歴史を反映

系統樹の例：類人猿の系統関係

系統樹ってなに？

系統樹の例：類人猿の系統関係

- 生物の系統関係を樹形図で表したもの
- 生物の進化の歴史を反映

進化生物学に必須のツール

系統樹の推定方法（最節約法）

系統樹の推定方法（最節約法）

データ配列

アカゲザル	A	C	C	G	T	T	A	C	C	G	A	T	T	A	C	C	G	A	T	T	A	C	T	T	A	C
テナガザル	A	C	C	G	A	T	A	C	C	G	A	A	T	A	C	C	G	A	A	T	A	C	A	T	A	C
オランウータン	T	C	C	C	A	T	T	C	C	C	A	A	T	T	C	C	C	A	A	T	T	C	A	T	T	C
ゴリラ	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
チンパンジー	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
ヒト	T	C	T	A	C	T	T	C	T	A	A	C	T	T	C	T	A	A	C	T	T	T	C	T	T	C

系統樹の推定方法（最節約法）

データ配列

アカゲザル	A	C	C	G	T	T	A	C	C	G	A	T	T	A	C	C	G	A	T	T	A	C	T	T	A	C
テナガザル	A	C	C	G	A	T	A	C	C	G	A	A	T	A	C	C	G	A	A	T	A	C	A	T	A	C
オランウータン	T	C	C	C	A	T	T	C	C	C	A	A	T	T	C	C	C	A	A	T	T	C	A	T	T	C
ゴリラ	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
チンパンジー	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
ヒト	T	C	T	A	C	T	T	C	T	A	A	C	T	T	C	T	A	A	C	T	T	T	C	T	T	C

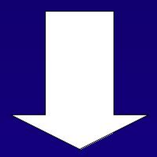
あり得る全ての系統樹で
データが実現するために必要な最小置換回数
を求める

系統樹の推定方法（最節約法）

データ配列

アカゲザル	A	C	C	G	T	T	A	C	C	G	A	T	T	A	C	C	G	A	T	T	A	C	T	T	A	C
テナガザル	A	C	C	G	A	T	A	C	C	G	A	A	T	A	C	C	G	A	A	T	A	C	A	T	A	C
オランウータン	T	C	C	C	A	T	T	C	C	C	A	A	T	T	C	C	C	A	A	T	T	C	A	T	T	C
ゴリラ	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
チンパンジー	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
ヒト	T	C	T	A	C	T	T	C	T	A	A	C	T	T	C	T	A	A	C	T	T	T	C	T	T	C

あり得る全ての系統樹で
データが実現するために必要な最小置換回数
を求める



最小置換回数が最小の系統樹を採用

系統樹の推定方法（最尤法）

データ配列

アカゲザル	A	C	C	G	T	T	A	C	C	G	A	T	T	A	C	C	G	A	T	T	A	C	T	T	A	C
テナガザル	A	C	C	G	A	T	A	C	C	G	A	A	T	A	C	C	G	A	A	T	A	C	A	T	A	C
オランウータン	T	C	C	C	A	T	T	C	C	C	A	A	T	T	C	C	C	A	A	T	T	C	A	T	T	C
ゴリラ	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
チンパンジー	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
ヒト	T	C	T	A	C	T	T	C	T	A	A	C	T	T	C	T	A	A	C	T	T	T	C	T	T	C

系統樹の推定方法（最尤法）

データ配列

アカゲザル	A	C	C	G	T	T	A	C	C	G	A	T	T	A	C	C	G	A	T	T	A	C	T	T	A	C
テナガザル	A	C	C	G	A	T	A	C	C	G	A	A	T	A	C	C	G	A	A	T	A	C	A	T	A	C
オランウータン	T	C	C	C	A	T	T	C	C	C	A	A	T	T	C	C	C	A	A	T	T	C	A	T	T	C
ゴリラ	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
チンパンジー	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
ヒト	T	C	T	A	C	T	T	C	T	A	A	C	T	T	C	T	A	A	C	T	T	T	C	T	T	C

あり得る全ての系統樹で
データが実現する確率＝尤度
を求める

系統樹の推定方法（最尤法）

データ配列

アカゲザル	A	C	C	G	T	T	A	C	C	G	A	T	T	A	C	C	G	A	T	T	A	C	T	T	A	C
テナガザル	A	C	C	G	A	T	A	C	C	G	A	A	T	A	C	C	G	A	A	T	A	C	A	T	A	C
オランウータン	T	C	C	C	A	T	T	C	C	C	A	A	T	T	C	C	C	A	A	T	T	C	A	T	T	C
ゴリラ	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
チンパンジー	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
ヒト	T	C	T	A	C	T	T	C	T	A	A	C	T	T	C	T	A	A	C	T	T	T	C	T	T	C

あり得る全ての系統樹で
データが実現する確率＝尤度
を求める

尤度が最大の系統樹を採用

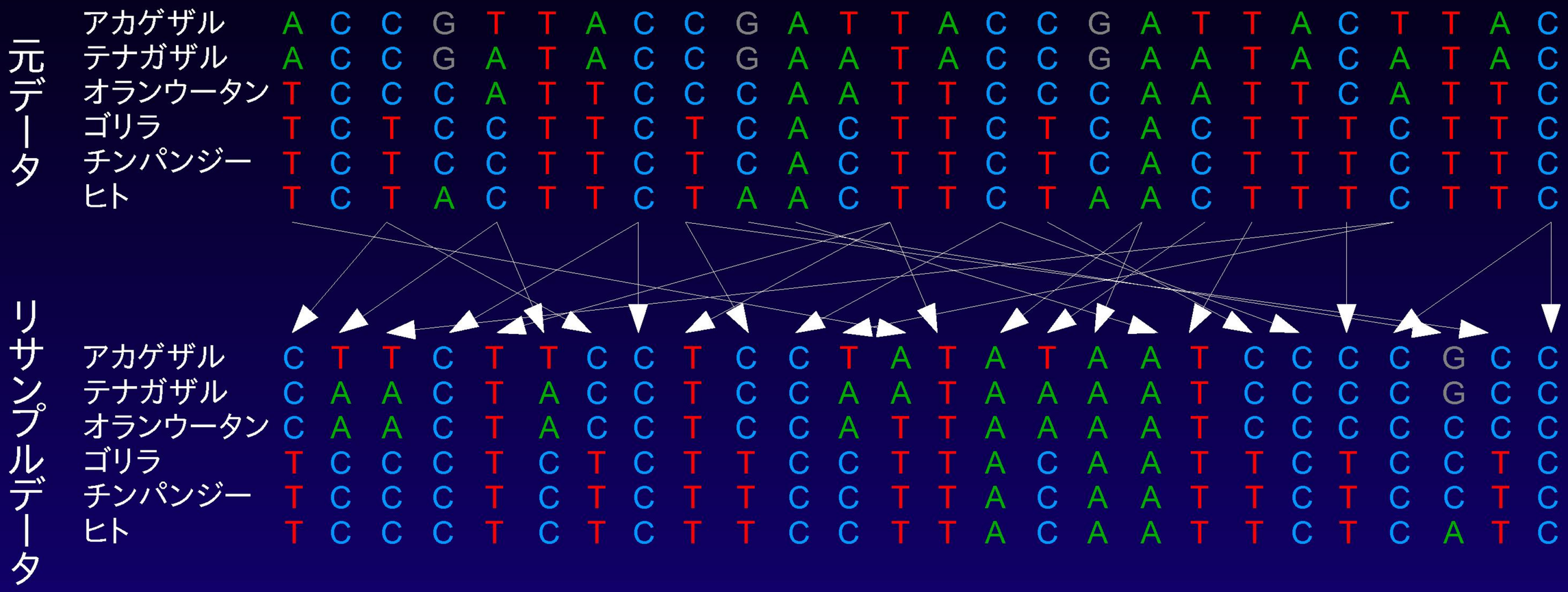
系統樹の信頼性を推定する

系統樹の信頼性を推定する

元データ

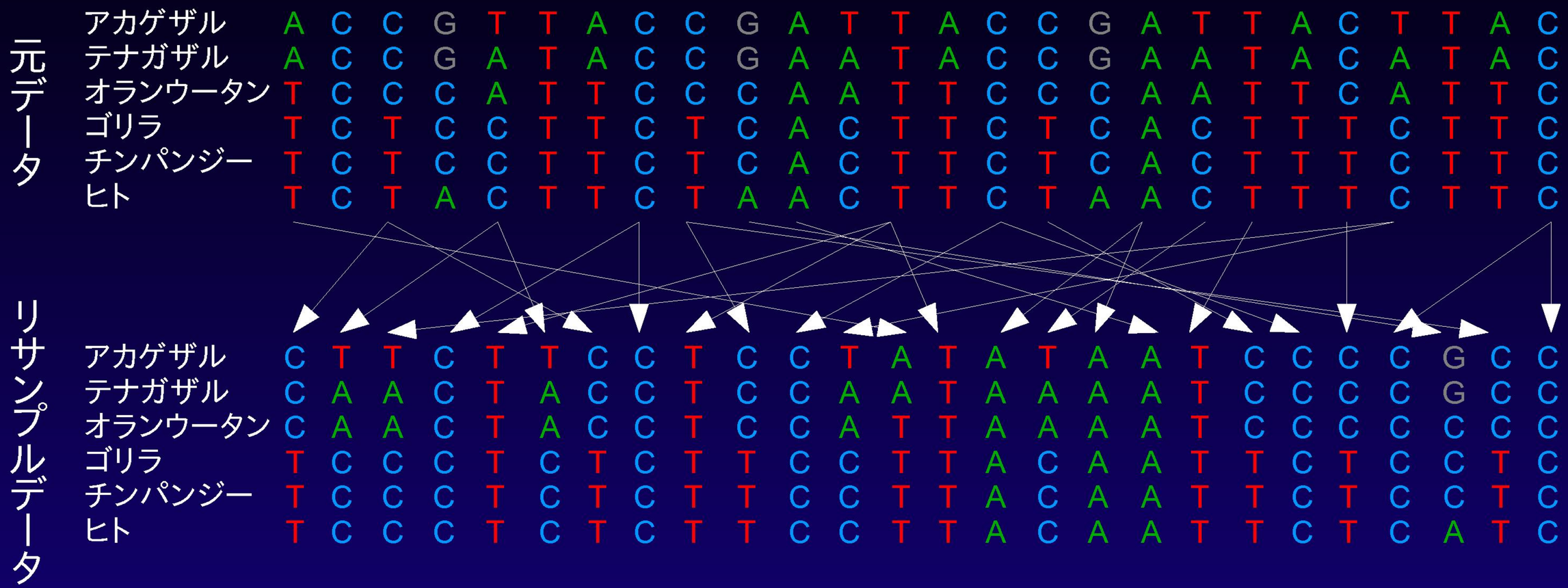
アカゲザル	A	C	C	G	T	T	A	C	C	G	A	T	T	A	C	C	G	A	T	T	A	C	T	T	A	C
テナガザル	A	C	C	G	A	T	A	C	C	G	A	A	T	A	C	C	G	A	A	T	A	C	A	T	A	C
オランウータン	T	C	C	C	A	T	T	C	C	C	A	A	T	T	C	C	C	A	A	T	T	C	A	T	T	C
ゴリラ	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
チンパンジー	T	C	T	C	C	T	T	C	T	C	A	C	T	T	C	T	C	A	C	T	T	T	C	T	T	C
ヒト	T	C	T	A	C	T	T	C	T	A	A	C	T	T	C	T	A	A	C	T	T	T	C	T	T	C

系統樹の信頼性を推定する



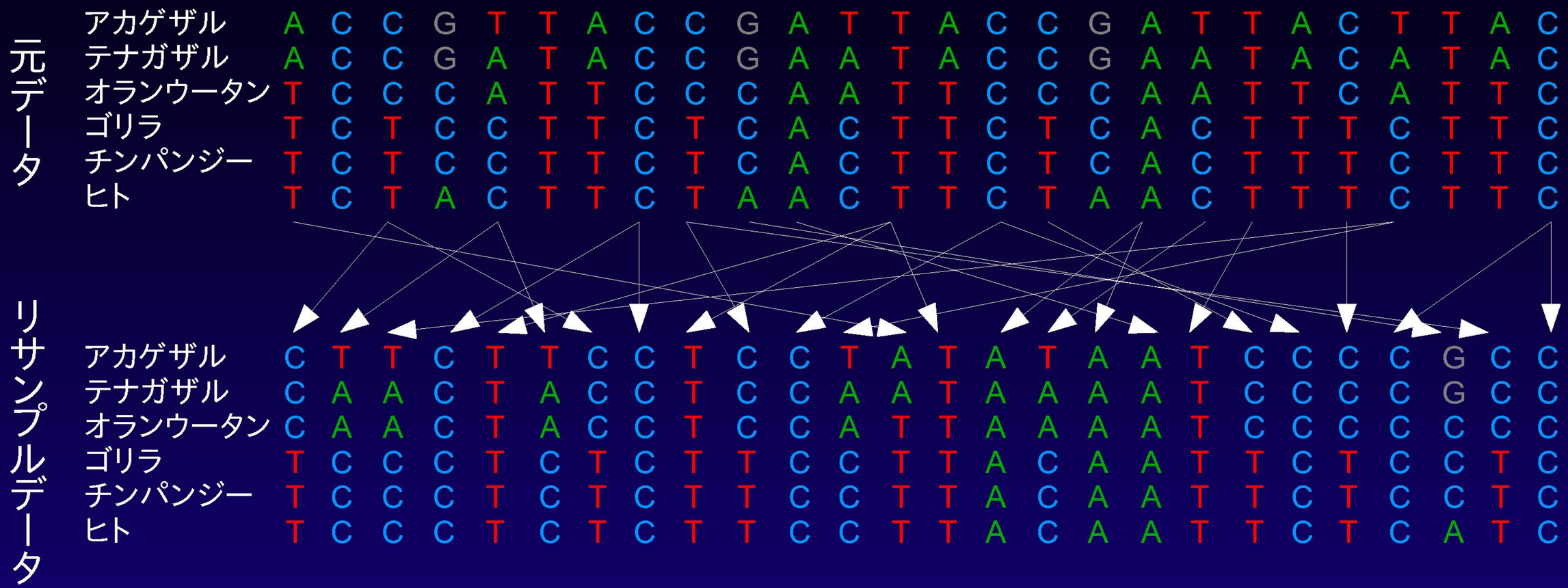
- 元データから重複を許して無作為に形質をリサンプリング

系統樹の信頼性を推定する



- 元データから重複を許して無作為に形質をリサンプリング
- リサンプルデータで系統推定を反復

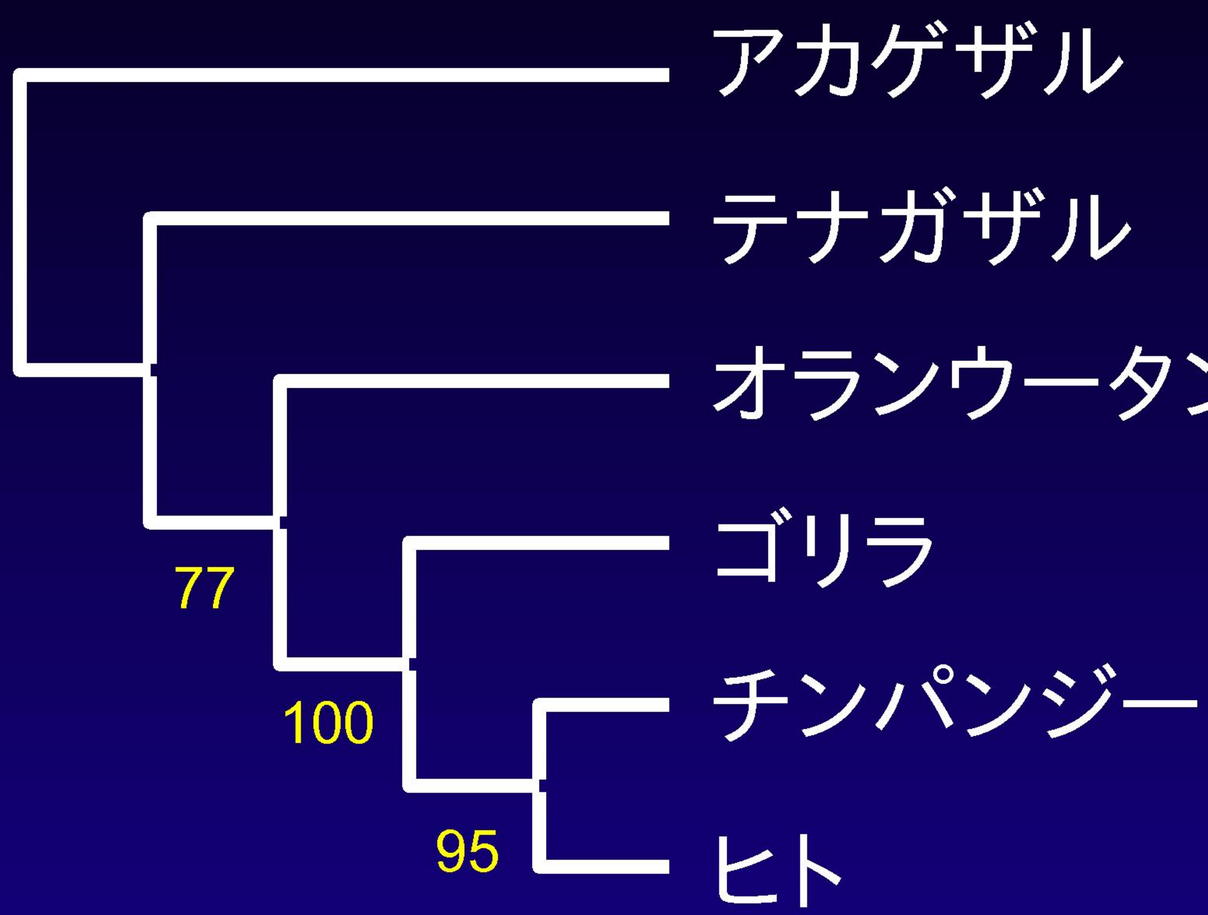
系統樹の信頼性を推定する



- 元データから重複を許して無作為に形質をリサンプリング
- リサンプルデータで系統推定を反復
- 系統関係の再現率を算出して信頼性の指標とする

信頼性の図示

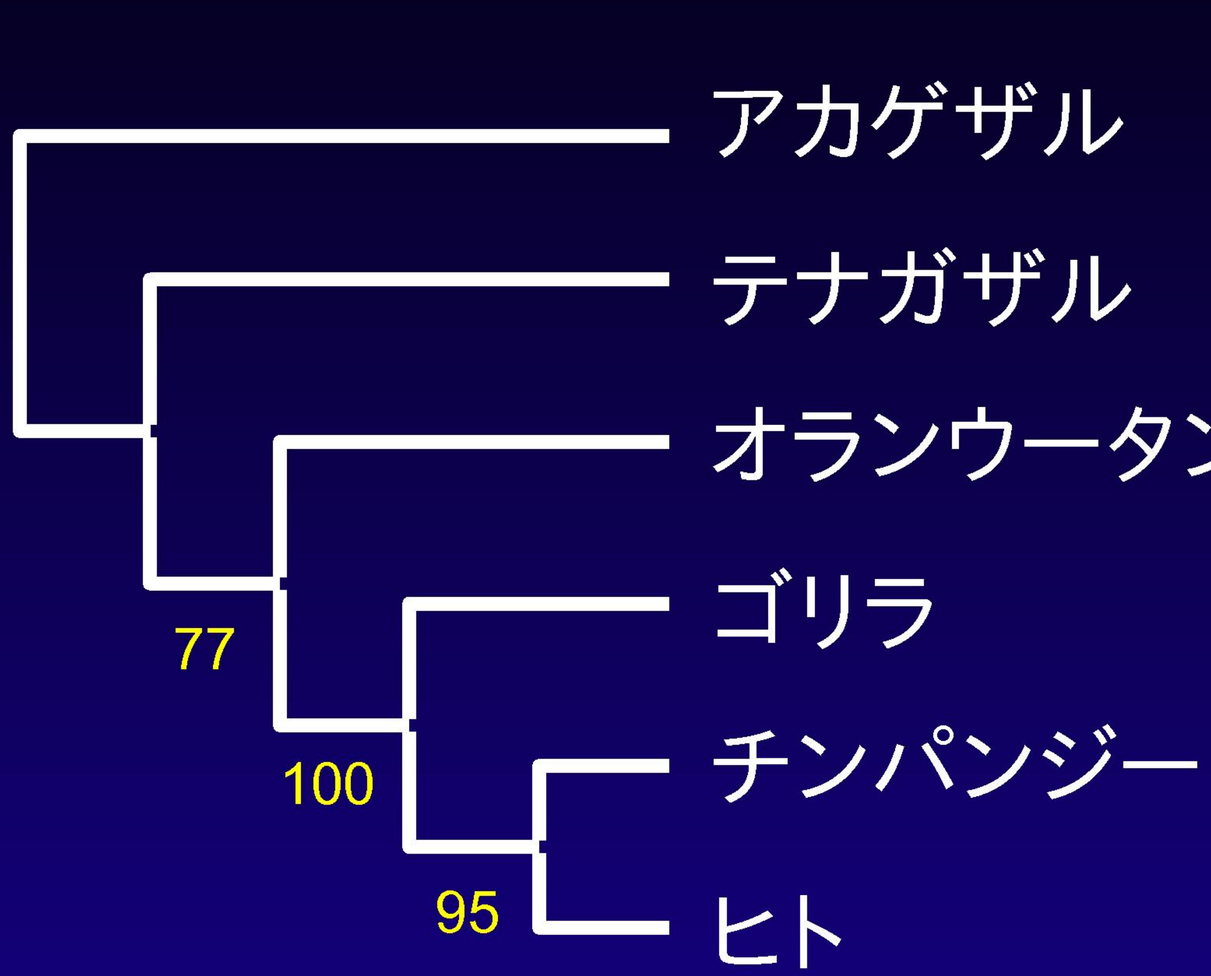
信頼性の図示



信頼性の図示

- 系統樹は OTU (端点) 数 –3 個の仮説を含む

信頼性の図示



- 系統樹は OTU (端点) 数 –3 個の仮説を含む
- それぞれの仮説の信頼性が示される

信頼性の図示

- 系統樹は OTU (端点) 数 –3 個の仮説を含む
- それぞれの仮説の信頼性が示される
- 1/3 の確率で +20 に過大評価されるとしたら, 6OTU の系統樹は +20 の過大評価を平均 1 個含んでいる

OTU 数とあり得る樹形の数

OTU 数とあり得る樹形の数

OTU 数	樹形数
3	1
4	3
8	10,395
16	2.134580e+14
32	2.921561e+40
64	1.037791e+103
128	1.013292e+248

OTU 数とあり得る樹形の数

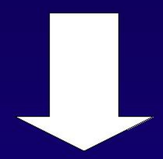
OTU 数	樹形数
3	1
4	3
8	10,395
16	2.134580e+14
32	2.921561e+40
64	1.037791e+103
128	1.013292e+248

OTU の増加に伴って
樹形数が大爆発

OTU 数とあり得る樹形の数

OTU 数	樹形数
3	1
4	3
8	10,395
16	2.134580e+14
32	2.921561e+40
64	1.037791e+103
128	1.013292e+248

OTU の増加に伴って
樹形数が大爆発



網羅的探索は不可能

系統樹の発見的探索

系統樹の発見的探索

樹形空間

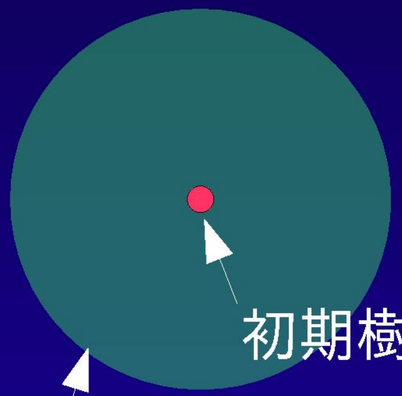
系統樹の発見的探索

樹形空間

初期樹形

系統樹の発見的探索

樹形空間

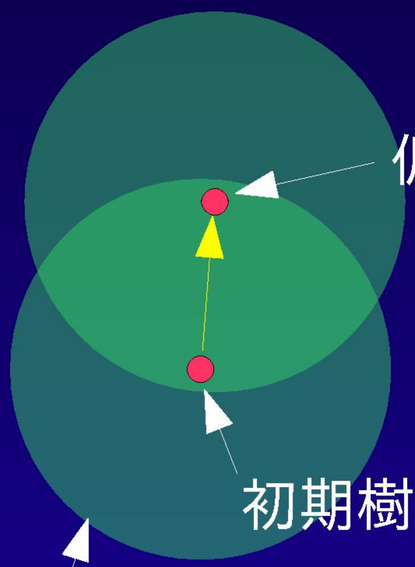


初期樹形

樹形改変によって探索される範囲

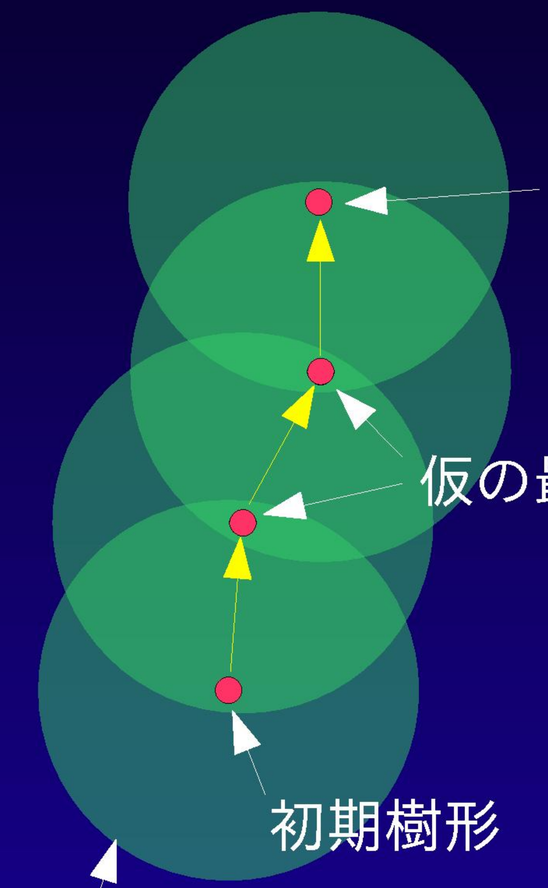
系統樹の発見的探索

樹形空間



系統樹の発見的探索

樹形空間



より尤度の高い系統樹が
見つからなかったら，
その時点で最も尤度が高いものを
最尤系統樹として採用

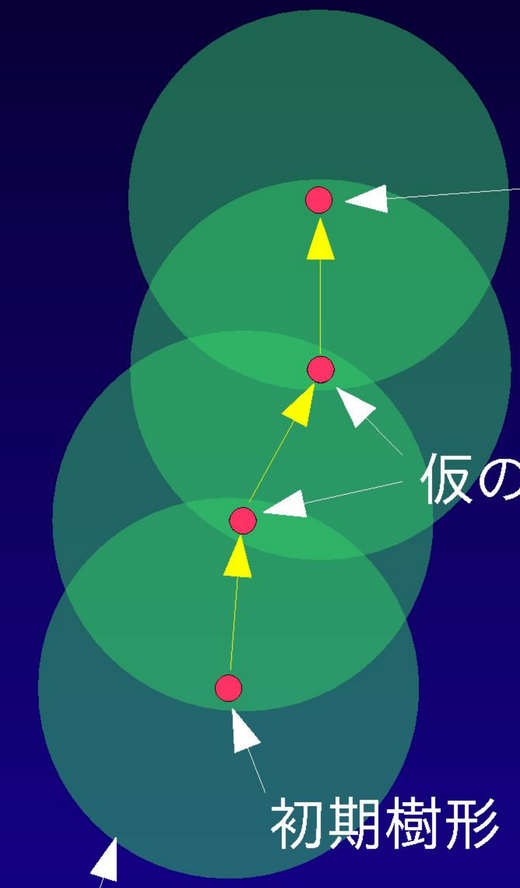
仮の最尤系統樹を更新していく

初期樹形

樹形改変によって探索される範囲

系統樹の発見的探索

樹形空間



より尤度の高い系統樹が
見つからなかったら,
その時点で最も尤度が高いものを
最尤系統樹として採用

仮の最尤系統樹を更新していく

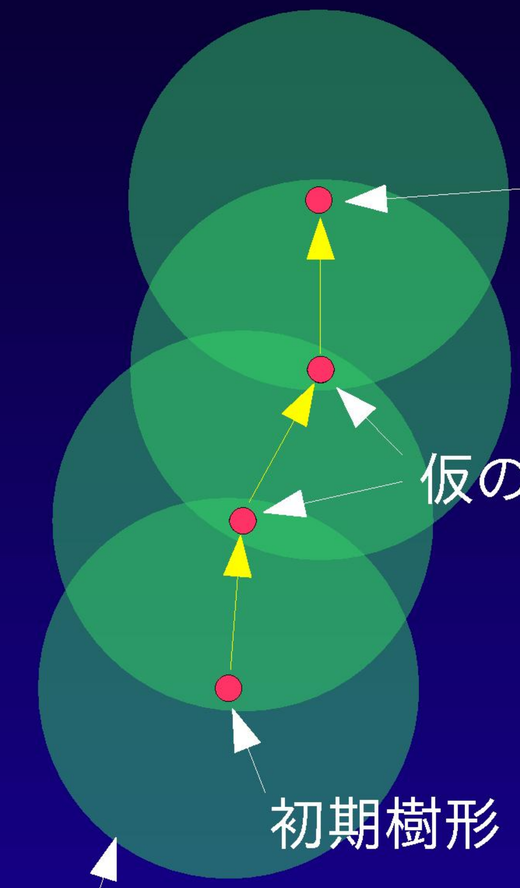
初期樹形

樹形改変によって探索される範囲

初期樹形が最尤系統樹
の樹形に影響する

系統樹の発見的探索

樹形空間



より尤度の高い系統樹が
見つからなかったら,
その時点で最も尤度が高いものを
最尤系統樹として採用

仮の最尤系統樹を更新していく

初期樹形

樹形改変によって探索される範囲

初期樹形が最尤系統樹
の樹形に影響する

探索範囲の広さが
最尤系統樹の樹形に影響する

仮説

仮説

- 初期樹形が元データの最尤系統樹なら信頼性は過大評価される

仮説

- 初期樹形が元データの最尤系統樹なら信頼性は過大評価される
- 初期樹形をリサンプルデータから生成するなら信頼性は過小評価される

仮説

- 初期樹形が元データの最尤系統樹なら信頼性は過大評価される
- 初期樹形をリサンプルデータから生成するなら信頼性は過小評価される
- 樹形探索労力量（探索範囲）を増加させることでバイアスは弱まる

仮説

- 初期樹形が元データの最尤系統樹なら信頼性は過大評価される
- 初期樹形をリサンプルデータから生成するなら信頼性は過小評価される
- 樹形探索労力量（探索範囲）を増加させることでバイアスは弱まる

シミュレーションデータの解析により検証

方法

方法

- 完全非対称型系統樹，完全対称型系統樹からモンテカルロ法により 500 塩基対の DNA 塩基配列データを各条件 50 セット生成

方法

- 完全非対称型系統樹，完全対称型系統樹からモンテカルロ法により 500 塩基対の DNA 塩基配列データを各条件 50 セット生成
- ブートストラップ法によりリサンプルデータを 100 セット生成

方法

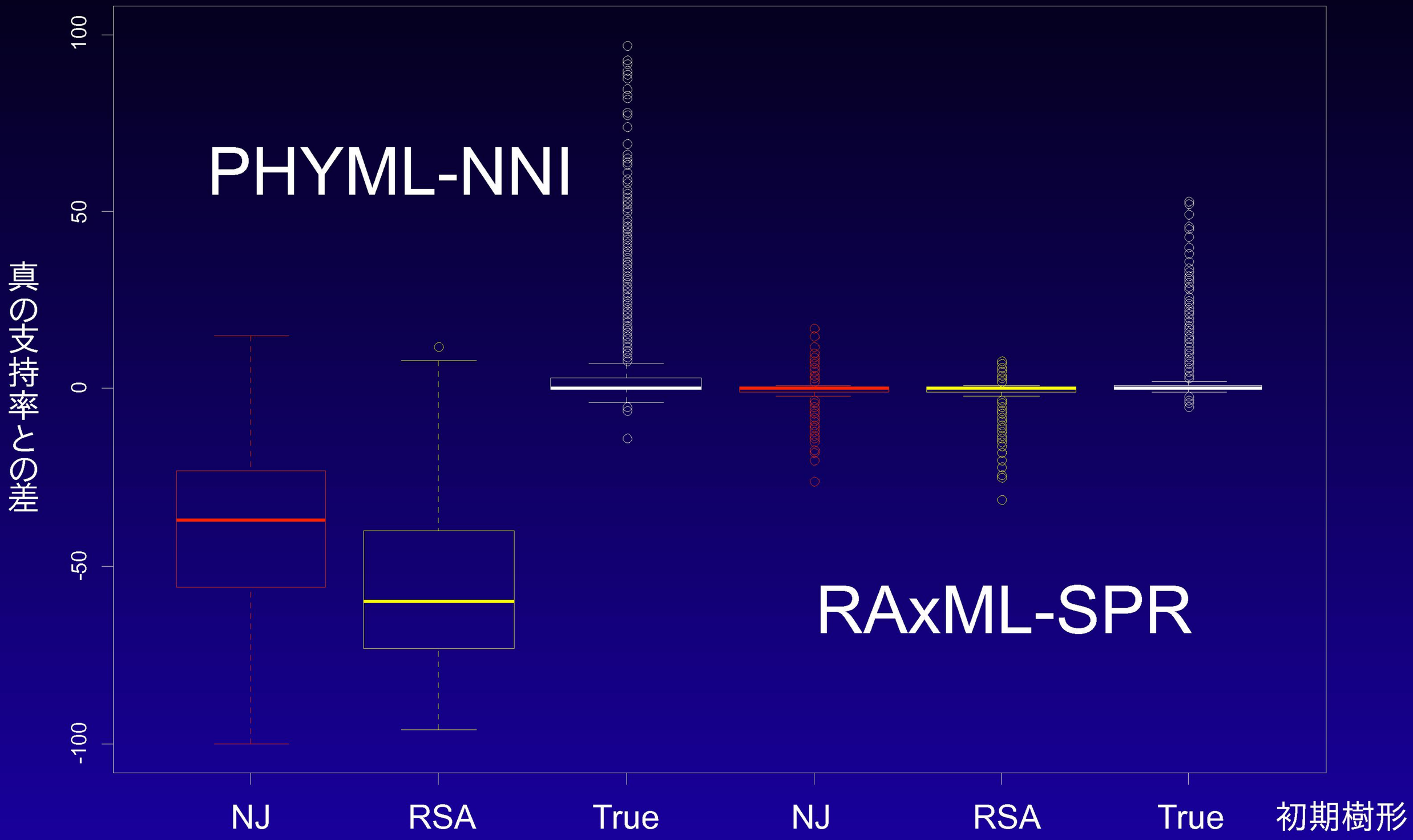
- 完全非対称型系統樹，完全対称型系統樹からモンテカルロ法により 500 塩基対の DNA 塩基配列データを各条件 50 セット生成
- ブートストラップ法によりリサンプルデータを 100 セット生成
- 初期樹形は以下の 3 条件
  - リサンプルデータから近隣結合法で作成 (NJ)
  - リサンプルデータから無作為配列付加法で作成 (RSA)
  - 真の系統樹 (True)

方法

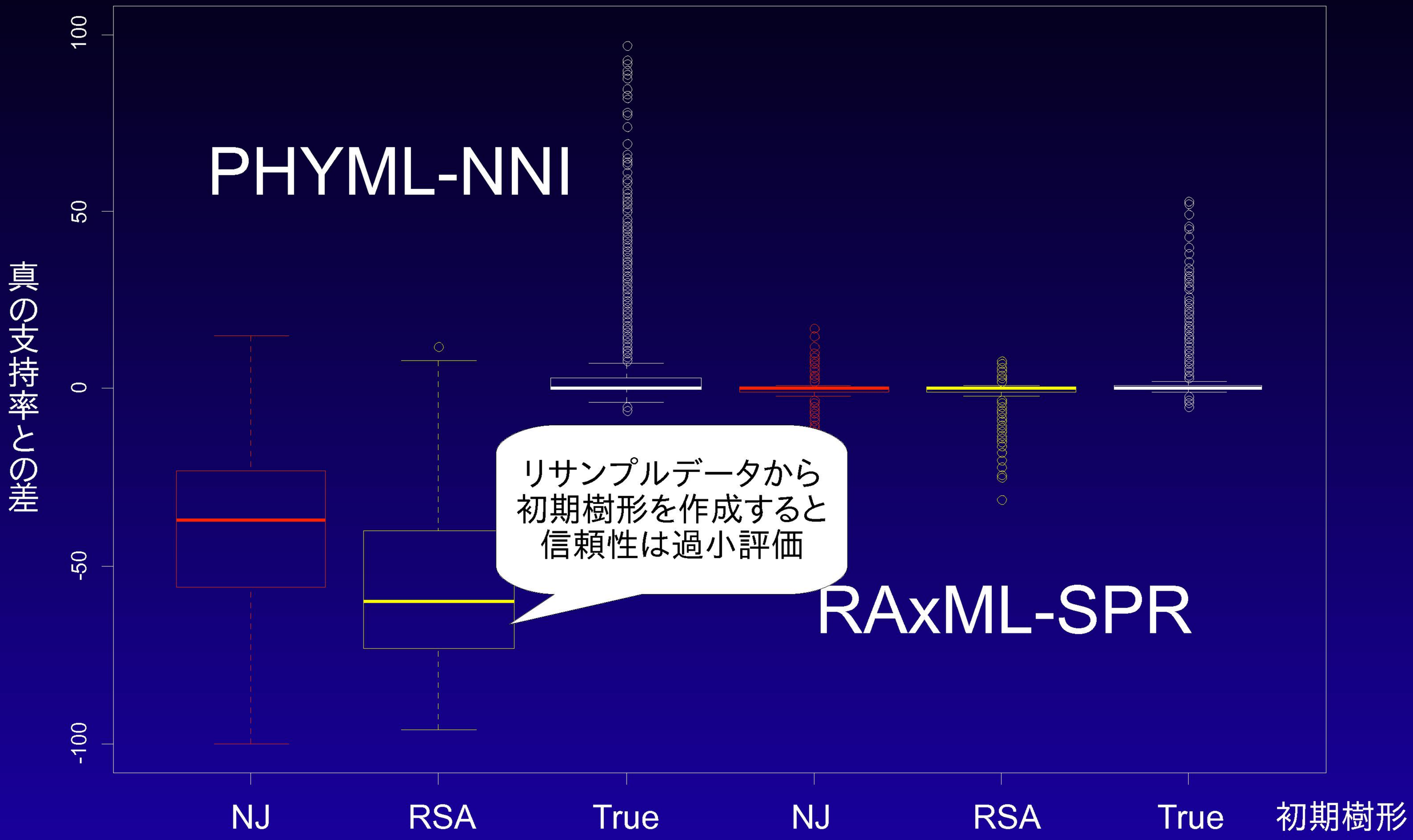
- 完全非対称型系統樹，完全対称型系統樹からモンテカルロ法により 500 塩基対の DNA 塩基配列データを各条件 50 セット生成
- ブートストラップ法によりリサンプルデータを 100 セット生成
- 初期樹形は以下の 3 条件
  - リサンプルデータから近隣結合法で作成 (NJ)
  - リサンプルデータから無作為配列付加法で作成 (RSA)
  - 真の系統樹 (True)
- 樹形探索労力量は以下の 2 条件
  - PHYML の最近隣交換 (PHYML–NNI)
  - RAxML の部分木剪定再接続 (RAxML–SPR)

結果（64OTU，非対称型系統樹）

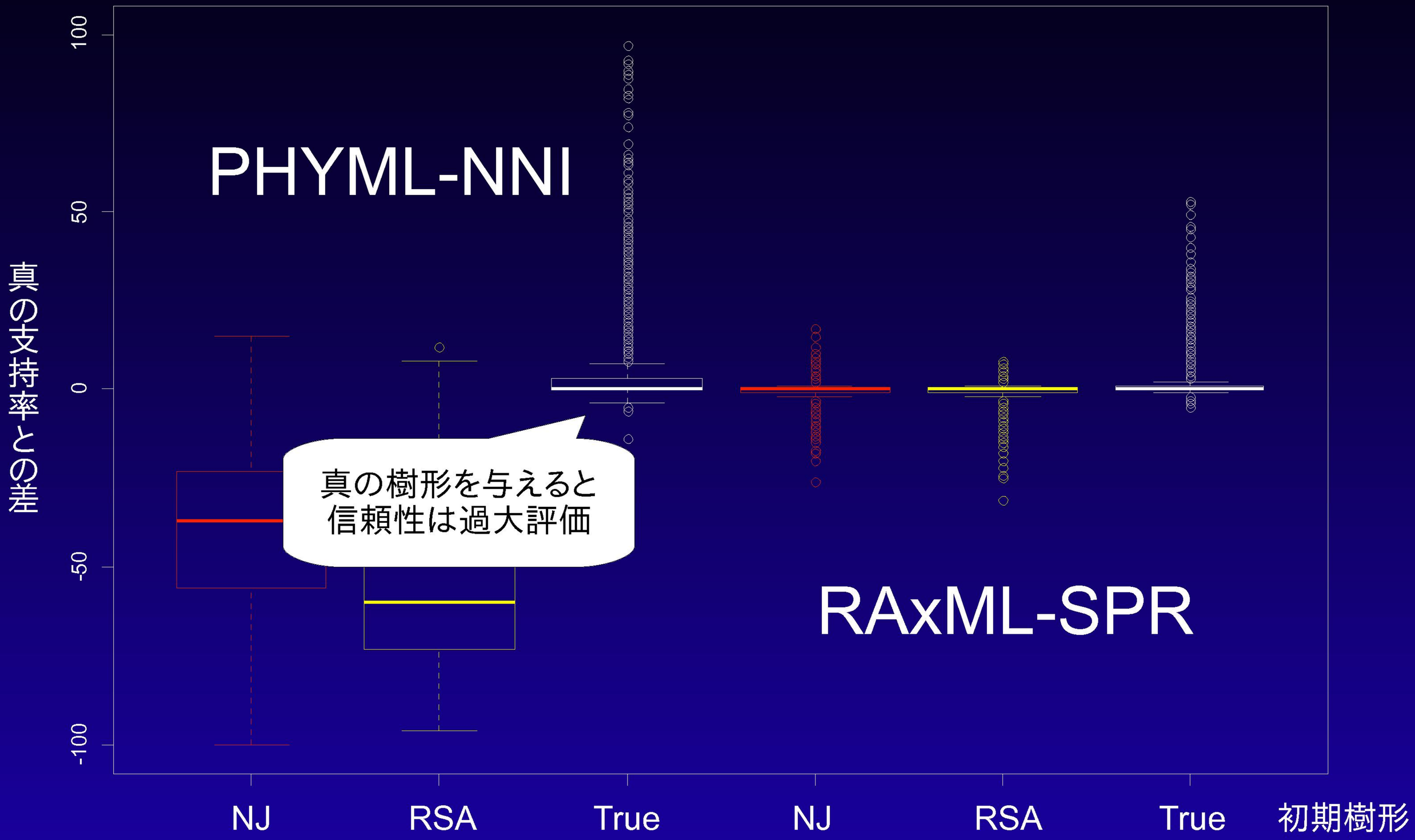
結果（64OTU，非対称型系統樹）



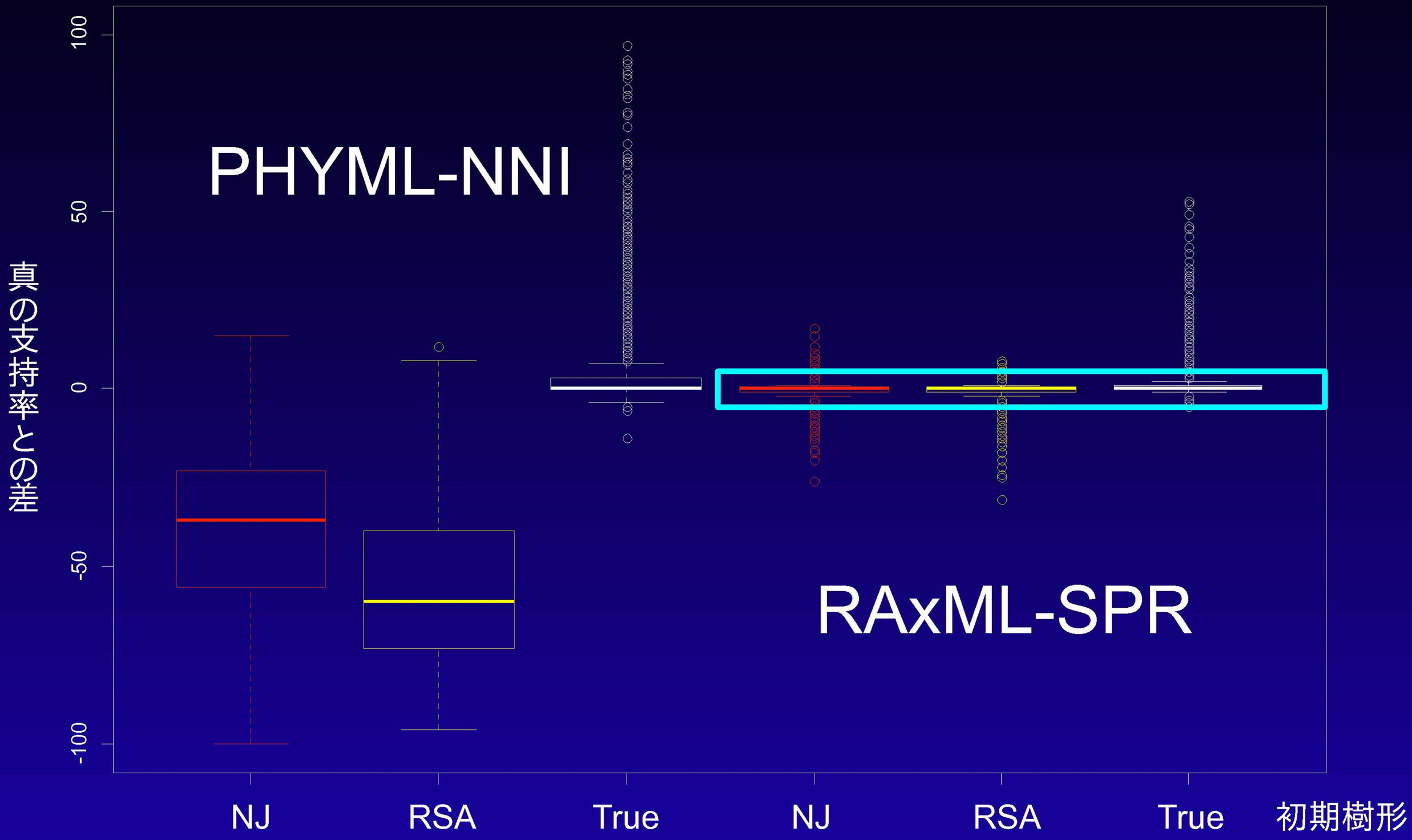
結果（64OTU，非対称型系統樹）



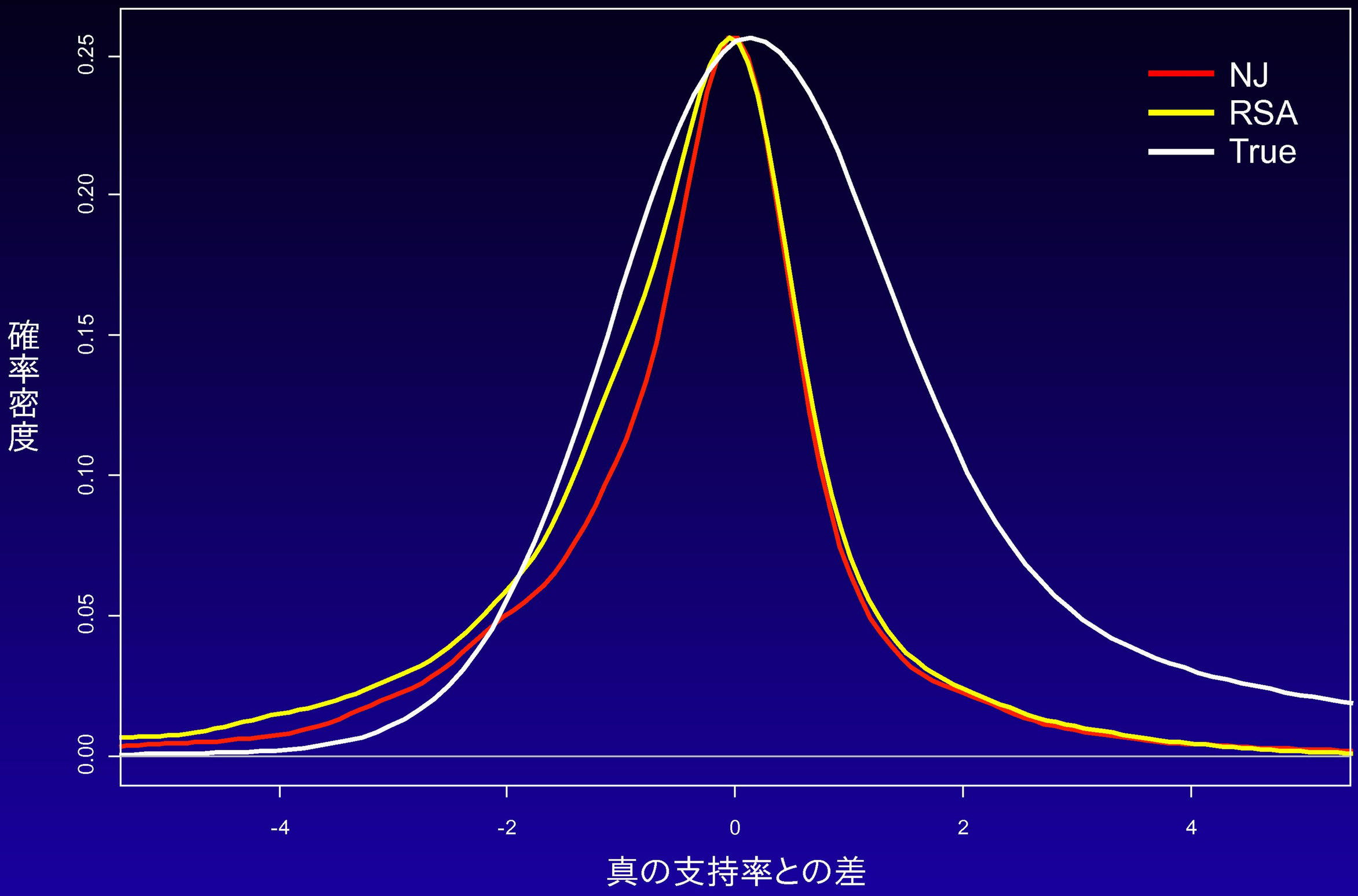
結果（64OTU，非対称型系統樹）



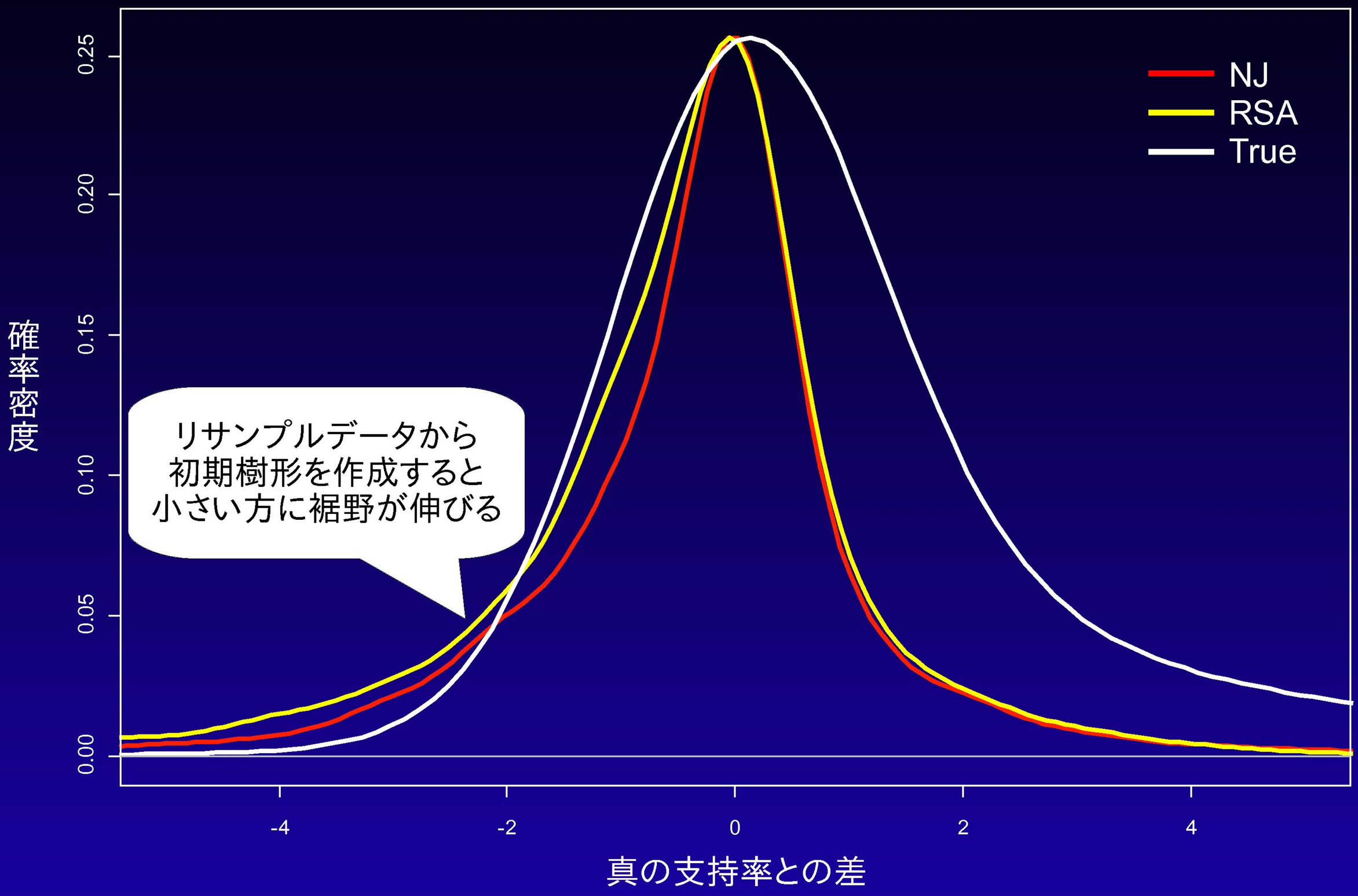
結果（64OTU，非対称型系統樹）



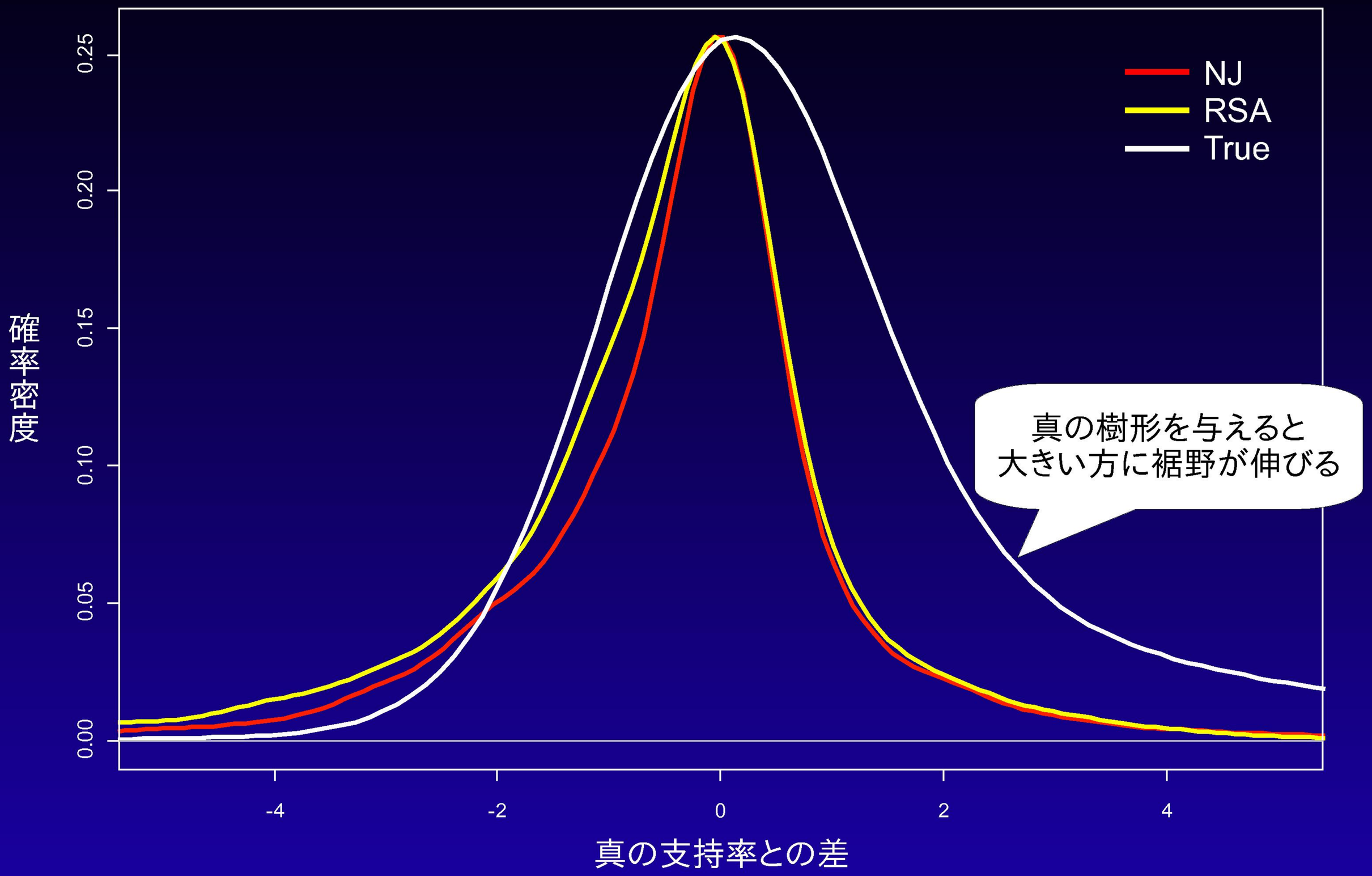
結果（64OTU，非対称型系統樹，RAxML-SPR）



結果（64OTU，非対称型系統樹，RAxML-SPR）



結果（64OTU，非対称型系統樹，RAxML-SPR）



まとめ

まとめ

- 初期樹形によって信頼性が過大評価されたり, 過小評価されることが確認できた

まとめ

- 初期樹形によって信頼性が過大評価されたり, 過小評価されることが確認できた
- バイアスの強さは NNI>>SPR であった

今後の予定

今後の予定

- SPR であっても, より大規模なデータでは無視できないほどのバイアスを受けると予想されるため, これを検証する

今後の予定

- SPR であっても, より大規模なデータでは無視できないほどのバイアスを受けると予想されるため, これを検証する
- 現実のデータでの検証を行う

今後の予定

- SPR であっても, より大規模なデータでは無視できないほどのバイアスを受けると予想されるため, これを検証する
- 現実のデータでの検証を行う
- バイアスを相殺する方法を開発する

終