

PhyDesign

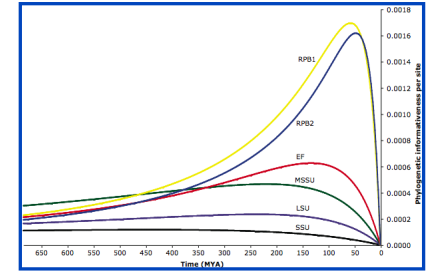
A web application for
profiling phylogenetic
informativeness of genes



Tools for marker selection are evolving

Theory

Phylogenetic Informativeness (PI) profiles.

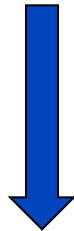


Townsend 2007, *Syst. Biol.*

Tools for marker selection are evolving

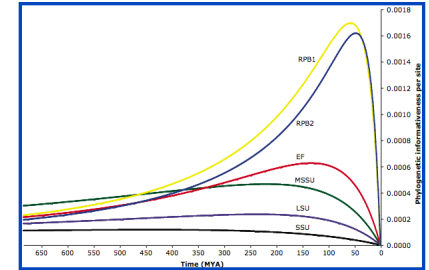
Theory

Phylogenetic Informativeness (PI) profiles.



Tools

PhyDesign web application.



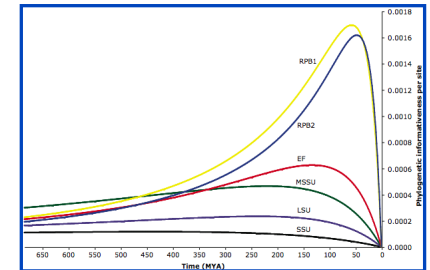
Townsend 2007, *Syst. Biol.*

Lopez-Giraldez & Townsend 2011,
BMC Evol. Biol.

Tools for marker selection are evolving

Theory

Phylogenetic Informativeness (PI) profiles.



Townsend 2007, *Syst. Biol.*

Tools

PhyDesign web application.

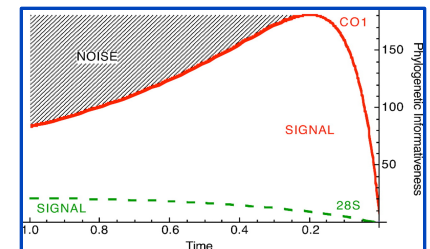
A screenshot of the PhyDesign web application interface. The interface is divided into three main sections: 1. INPUT DATA, 2. RATES RESULTS, and 3. PI PROFILES. The 1. INPUT DATA section is active, showing options for 'Alignment file' (upload or paste), 'Chromosome file' (upload or paste), and 'Choose program' (dropdown menu). The 2. RATES RESULTS section shows a table of results. The 3. PI PROFILES section shows a graph of PI profiles. The interface is titled 'PhyDesign: Profiling phylogenetic informativeness'.

Lopez-Giraldez & Townsend 2011, *BMC Evol. Biol.*

New Theory

Signal and noise

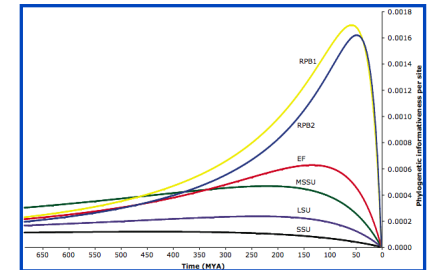
Taxon sampling vs character sampling
Complex models of molecular evolution



Tools for marker selection are evolving

Theory

Phylogenetic Informativeness (PI) profiles.



Townsend 2007, *Syst. Biol.*

Tools

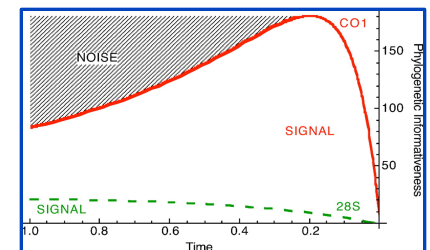
PhyDesign web application.

Lopez-Giraldez & Townsend 2011, *BMC Evol. Biol.*

New Theory

Signal and noise

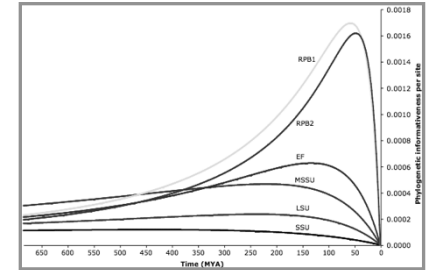
Taxon sampling vs character sampling
Complex models of molecular evolution



Tools for marker selection are evolving

Theory

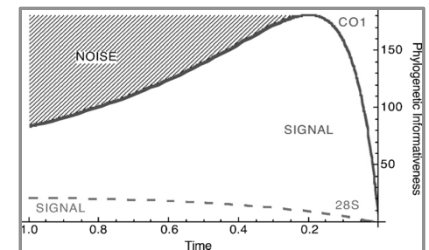
Phylogenetic Informativeness (PI) profiles.



Townsend 2007, *Syst. Biol.*

GOAL: To develop theoretical and practical methods to guide performance of **more cost-effective and accurate phylogenetic inference.**

Lopez-Giraldez & Townsend 2011,
BMC Evol. Biol.



New Theory

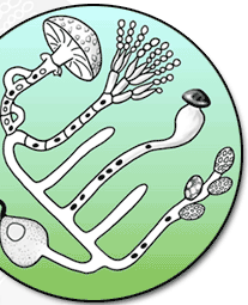
Signal and noise

Taxon sampling vs character sampling

Complex models of molecular evolution

Only a few of many possible genes have been used for phylogenetics

AFToL-1 Assembling the Fungal Tree of Life

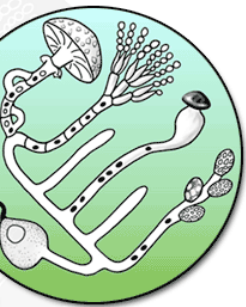


12 initial candidates.



RPB1
RPB2
TEF
SSU
LSU
MSSU

Only a few of many possible genes have been used for phylogenetics

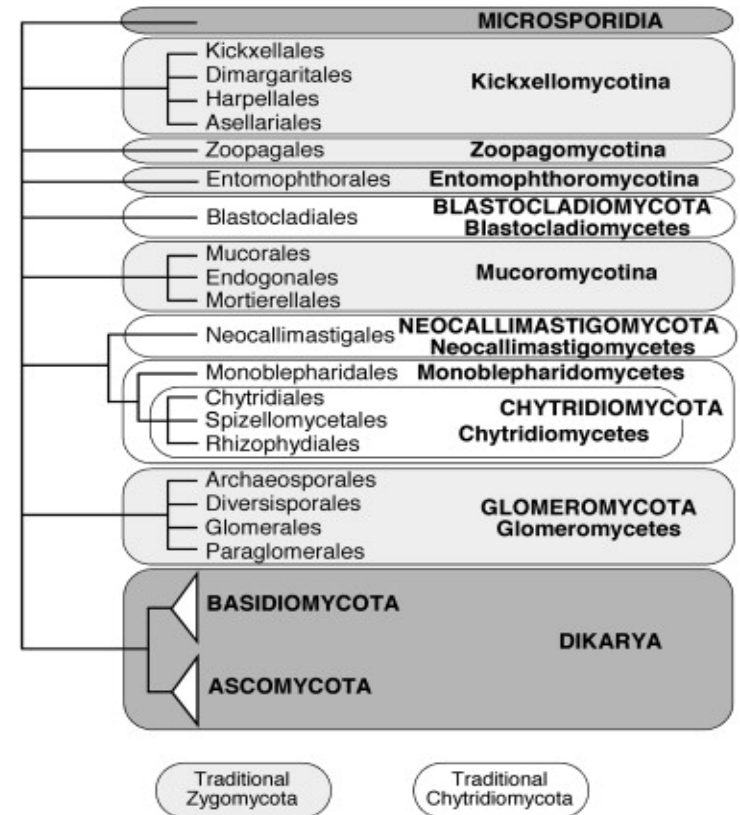


AFToL-1 Assembling the Fungal Tree of Life

12 initial candidates.

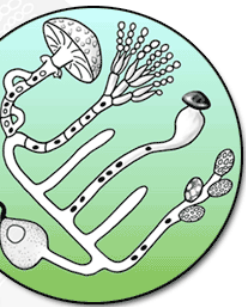


RPB1
RPB2
TEF
SSU
LSU
MSSU



Hibbett et al. 2007, *Mycological Research*

Only a few of many possible genes have been used for phylogenetics

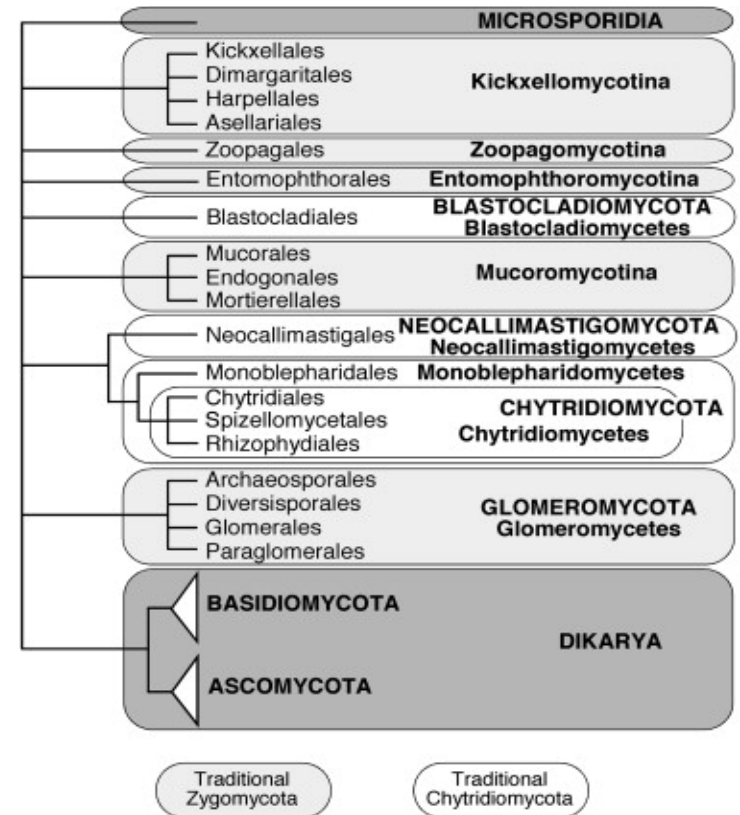


AFToL-1 Assembling the Fungal Tree of Life

12 initial candidates.



RPB1
RPB2
TEF
SSU
LSU
MSSU

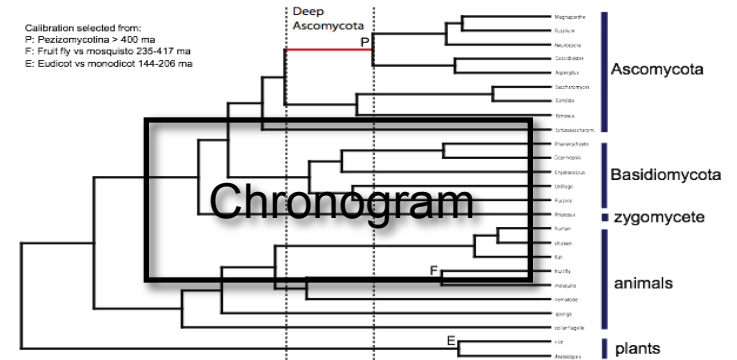
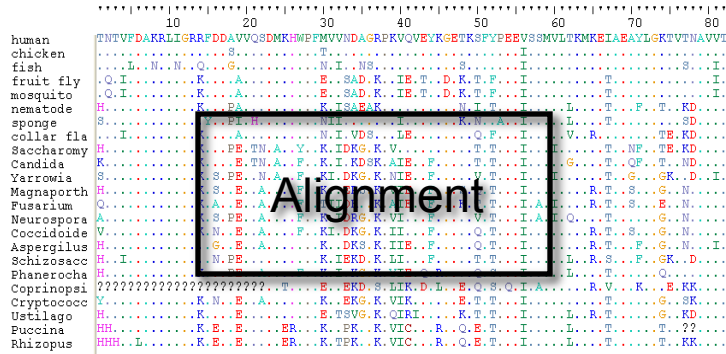


Hibbett et al. 2007, *Mycological Research*

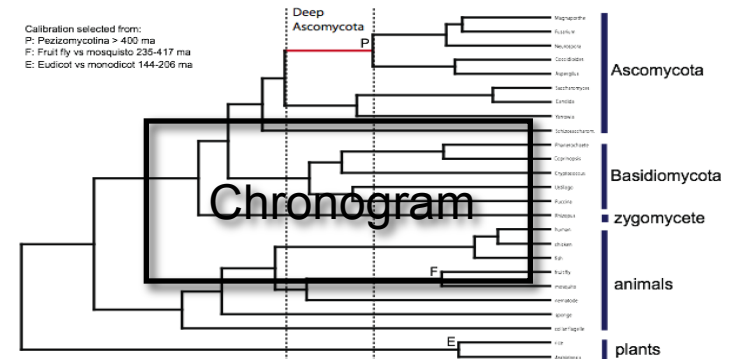
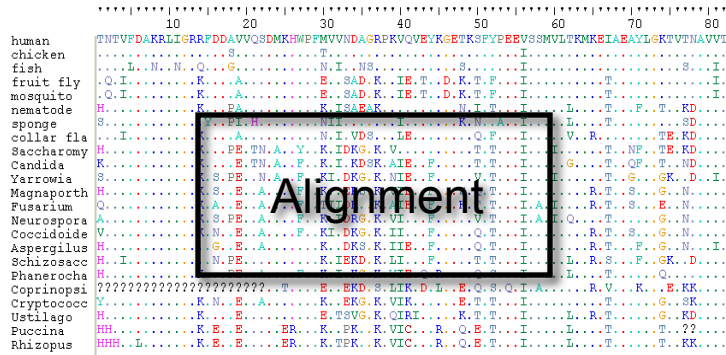
410 Fungal Genome Sequencing Projects

16 completed
168 assembling
226 in progress

PI profiles derives from an alignment and a chronogram

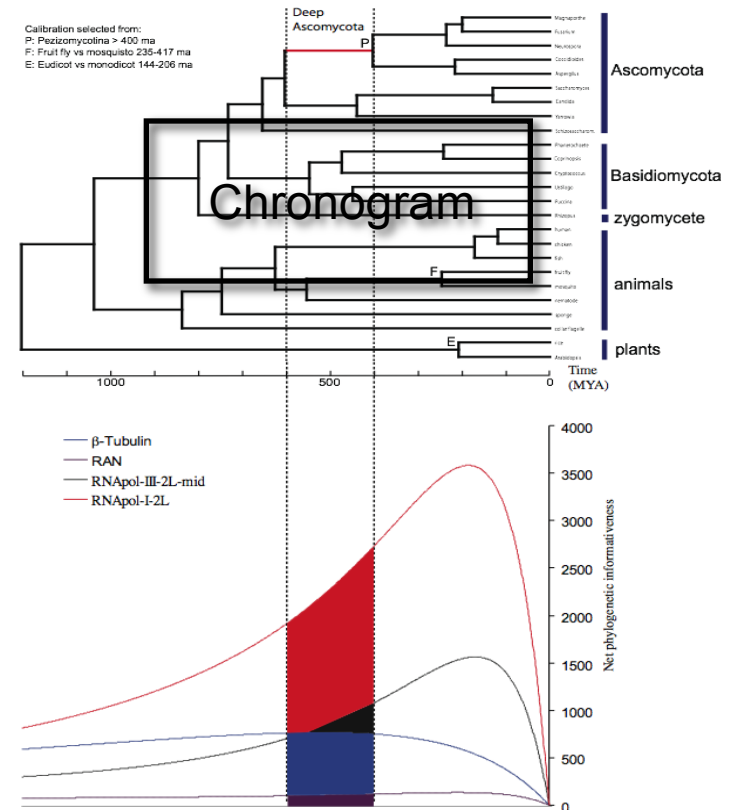
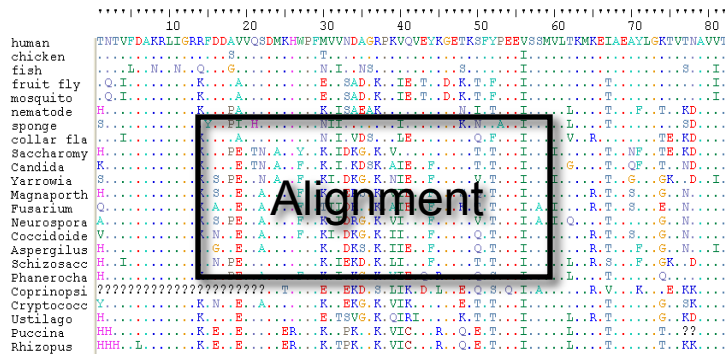


PI profiles derives from an alignment and a chronogram



Substitution rate (λ_i)

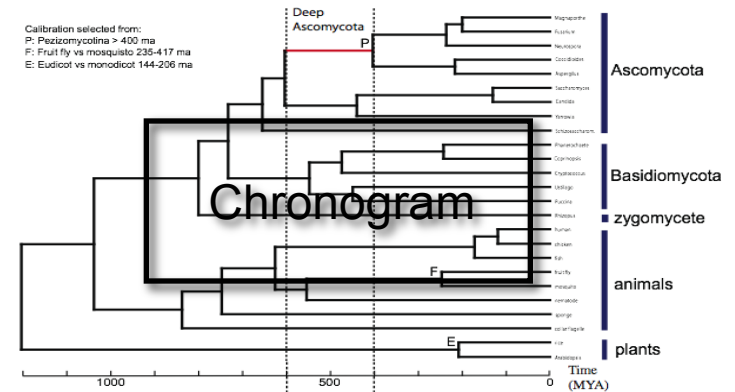
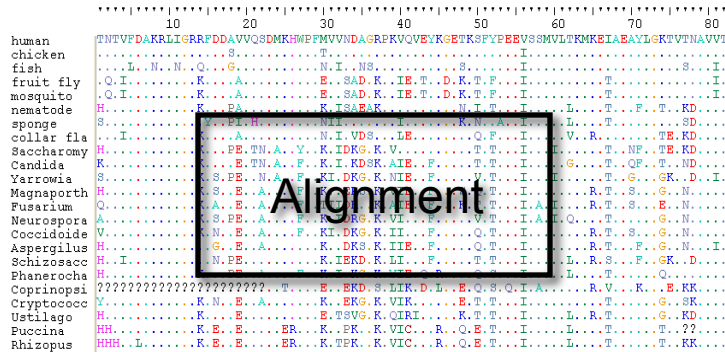
PI profiles derives from an alignment and a chronogram



Substitution rate (λ_i)

$$\rho(T; \lambda_1, \dots, \lambda_n) = \sum_{i=1}^n 16 \lambda_i^2 T e^{-4 \lambda_i T}$$

PI profiles derives from an alignment and a chronogram

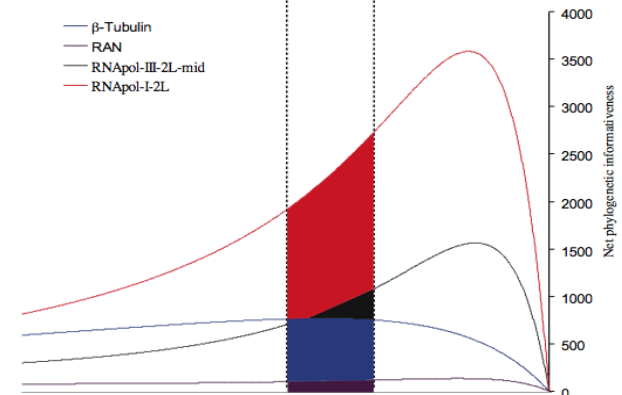


Substitution rate (λ_i)

$$\rho(T; \lambda_1, \dots, \lambda_n) = \sum_{i=1}^n 16 \lambda_i^2 T e^{-4 \lambda_i T}$$

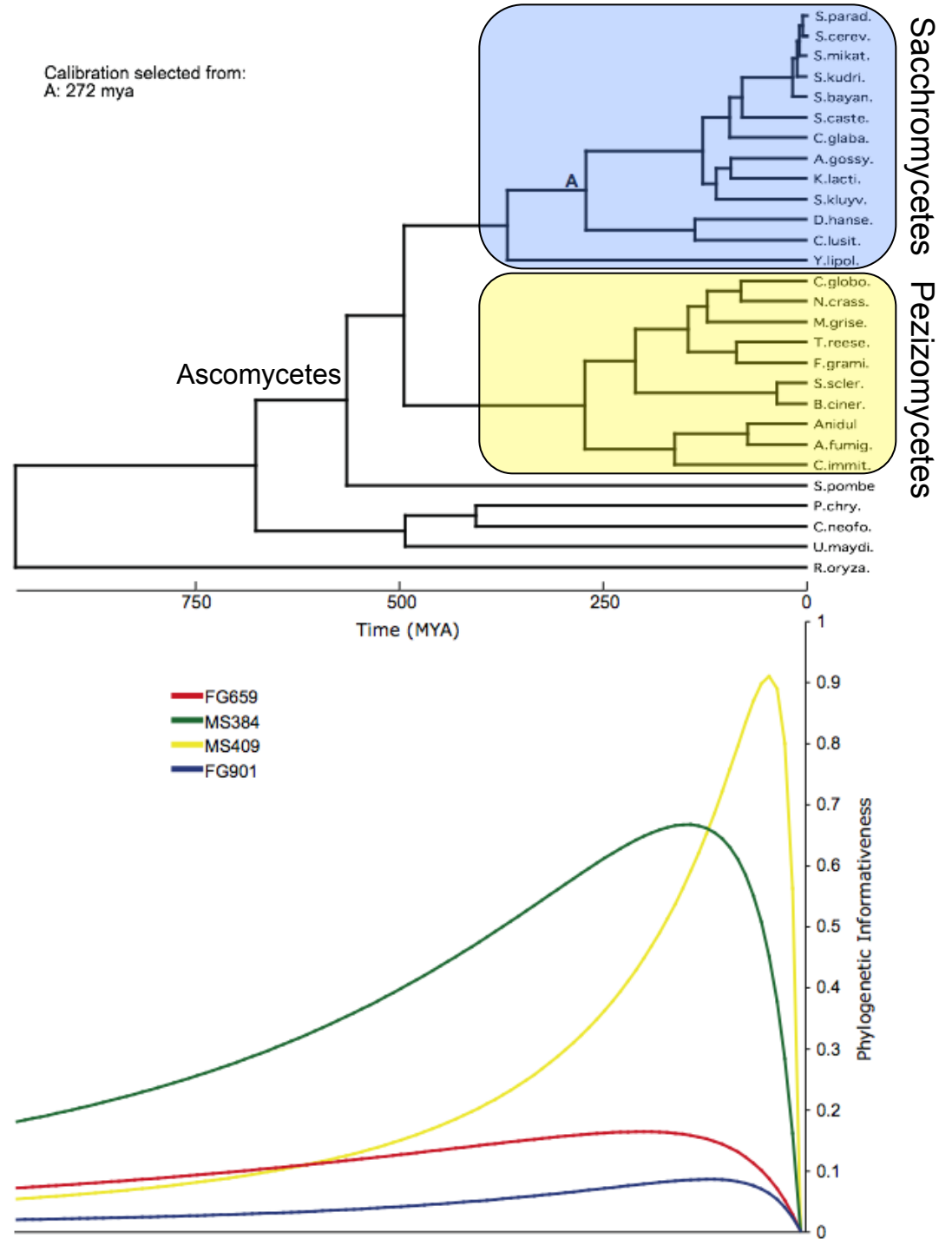
$$\int_{h_1}^{h_2} \rho(T; \lambda) dT$$

Townsend 2007, Syst. Biol.

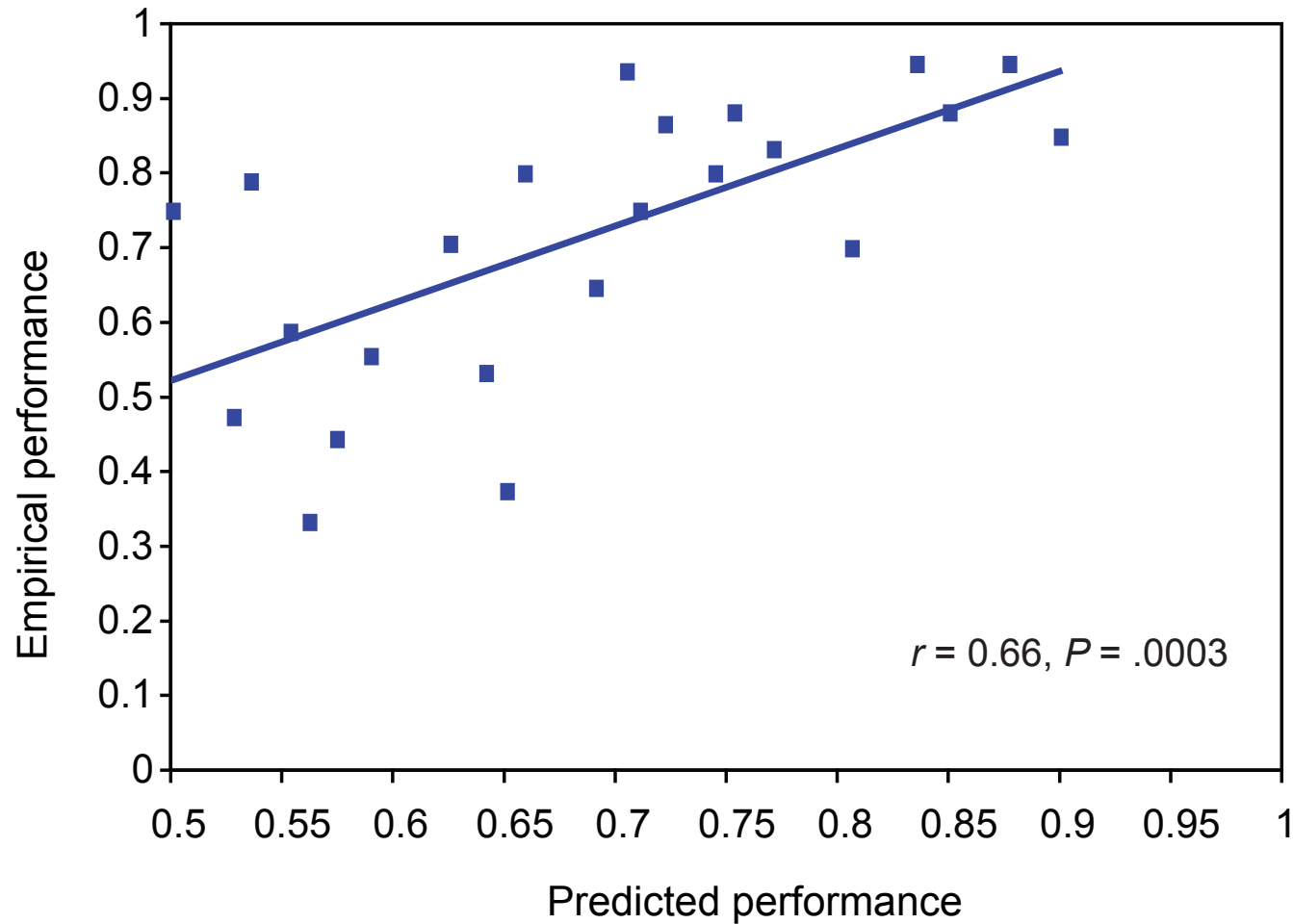


Gene	Length (aa)	Deep Ascomycota (400-600 MYA)
RPA2	484	45.8
RPA1	307	35.9
EF-2	253	23.3
SF3B1	422	26.7
PC	246	19.1
HSP70M	330	18.4
RPC2-mid	230	17.4
ARPC2	129	15.6
β -Tubulin	387	15.3
RAN	114	2.1

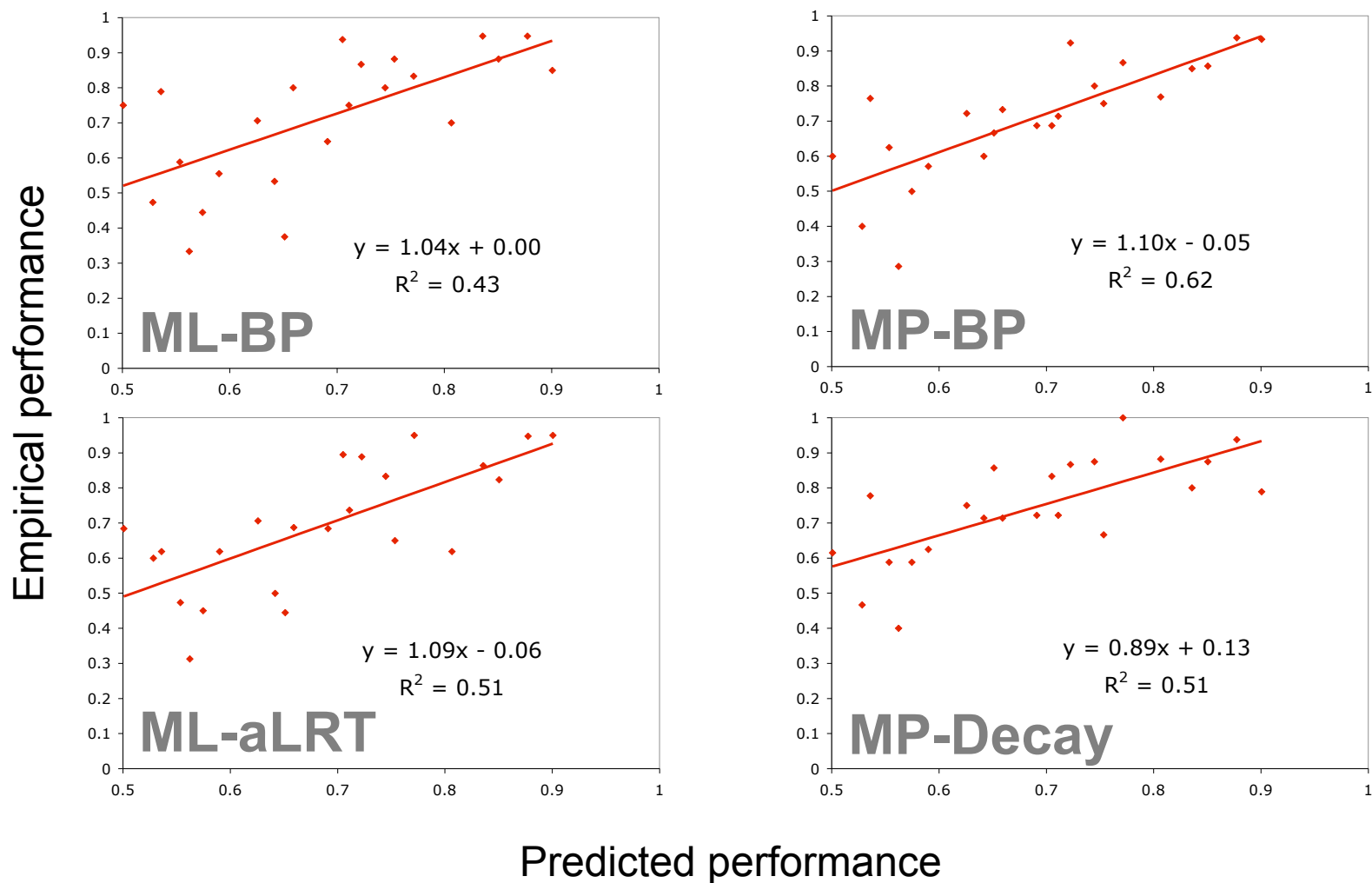
Analysis for 46 genes within Ascomycetes reveals diversity of PI profiles



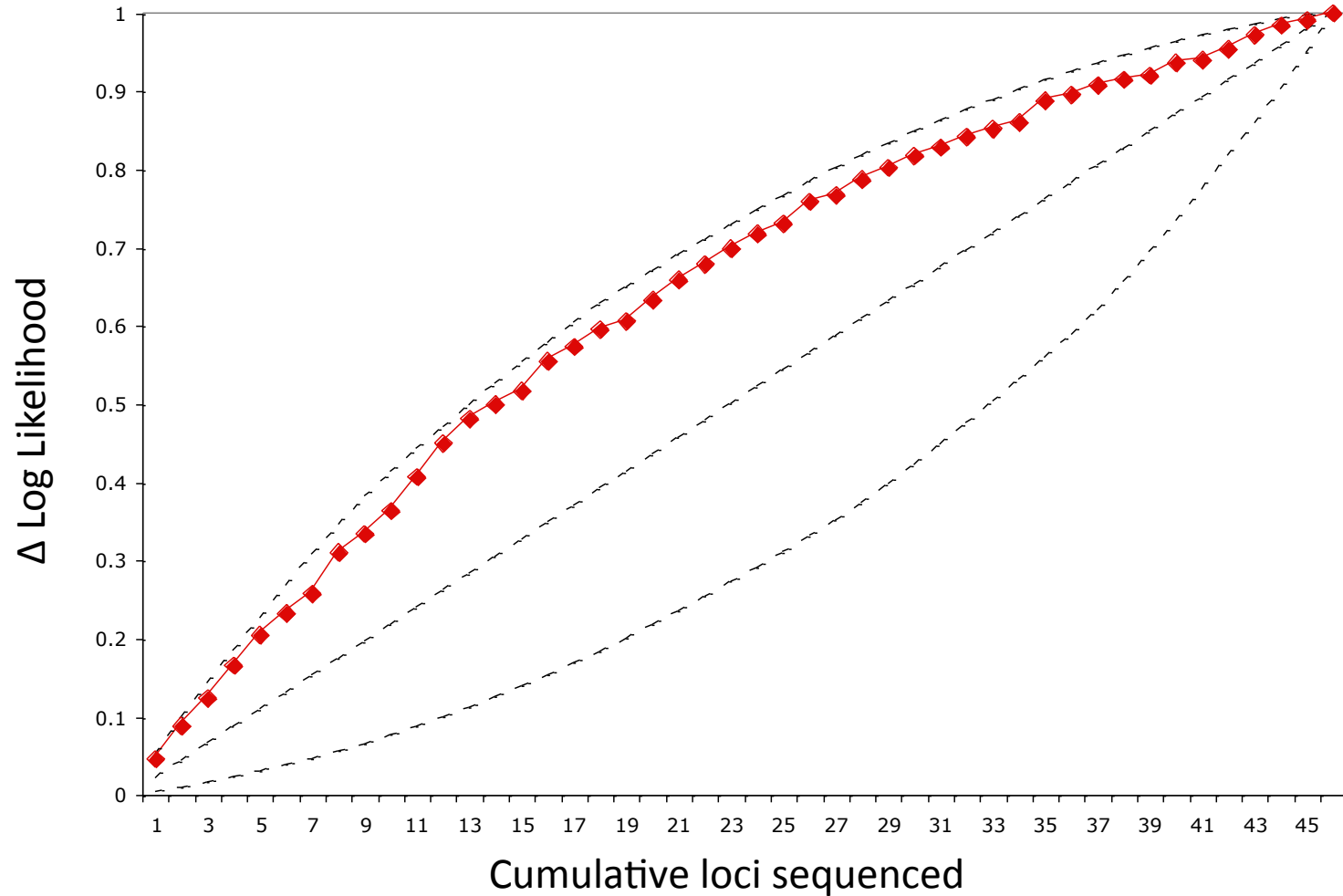
Informativeness of markers predicts support



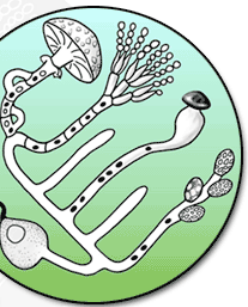
PI predicts support across support measures and methods of inference



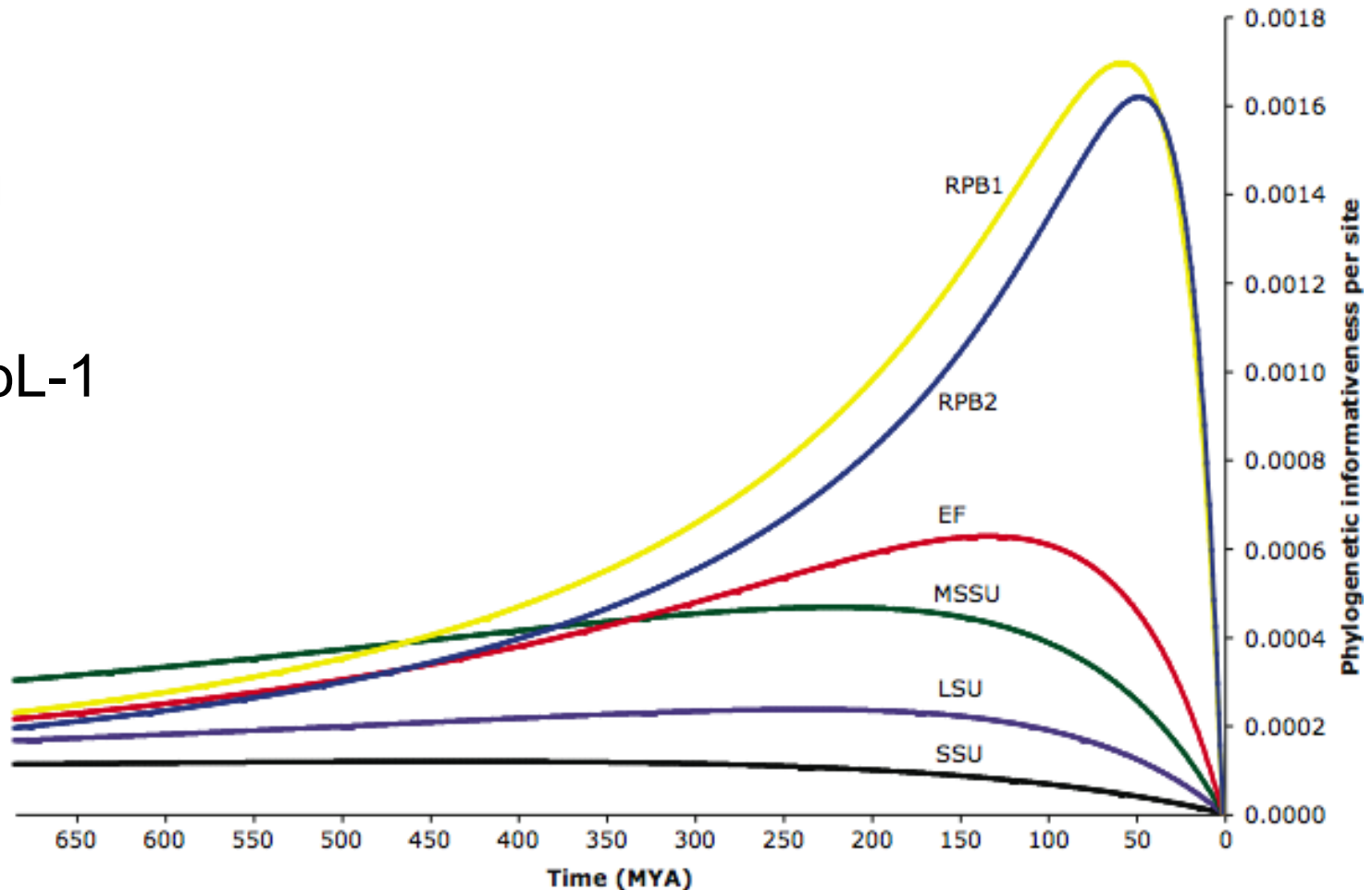
PI predicts a near optimal priority of loci to sequence



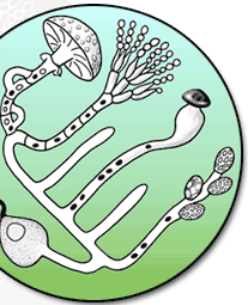
The most frequent genes used in phylogenetics exhibit a diversity of informativeness



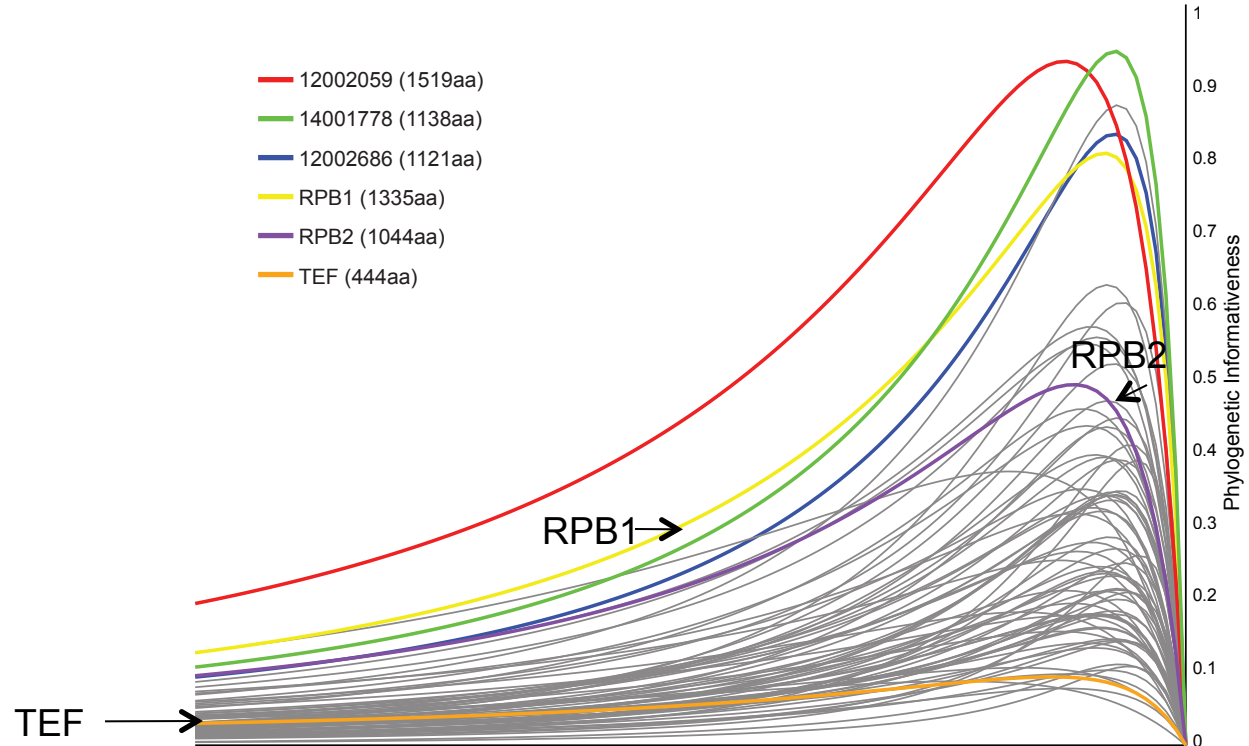
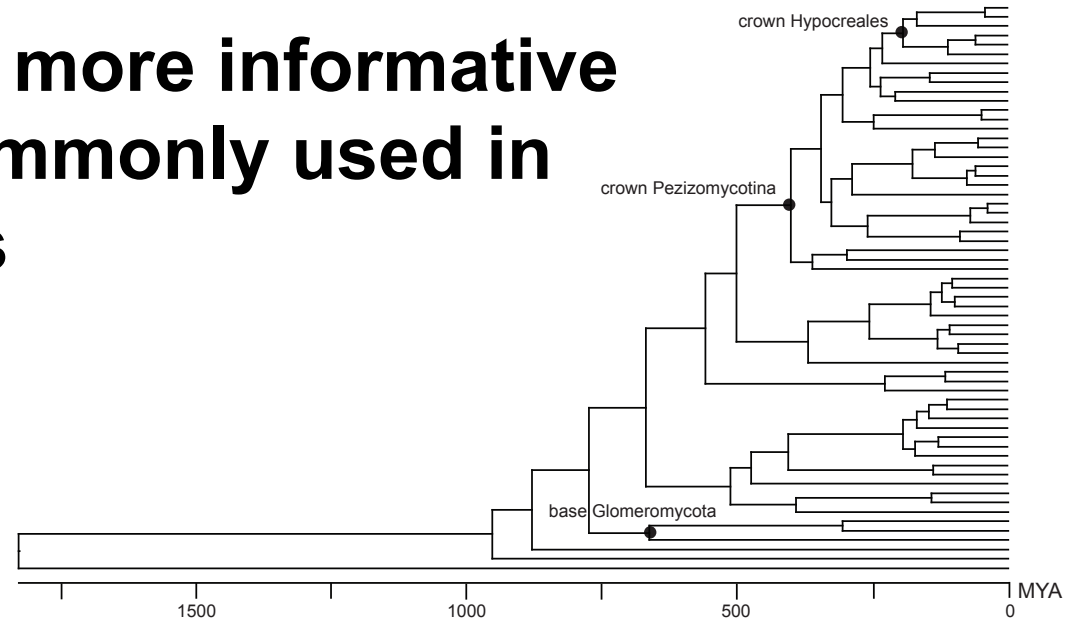
AFToL-1



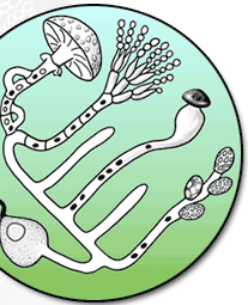
There are potentially more informative genes than those commonly used in fungal phylogenetics



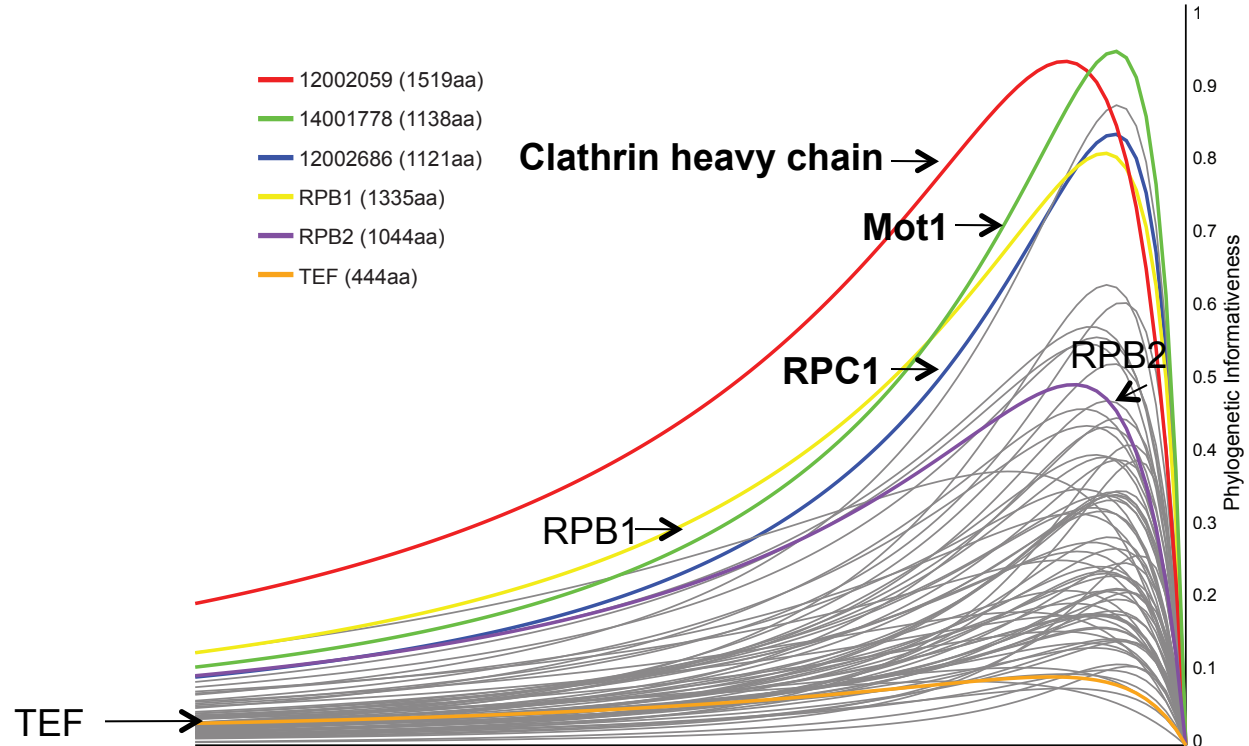
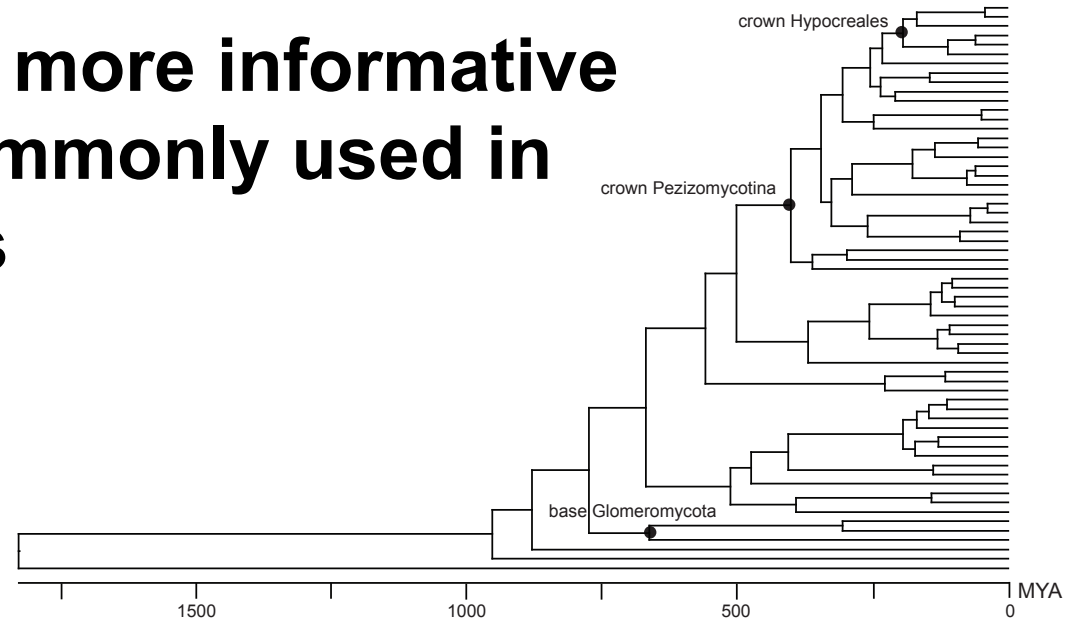
AFToL-2



There are potentially more informative genes than those commonly used in fungal phylogenetics



AFToL-2

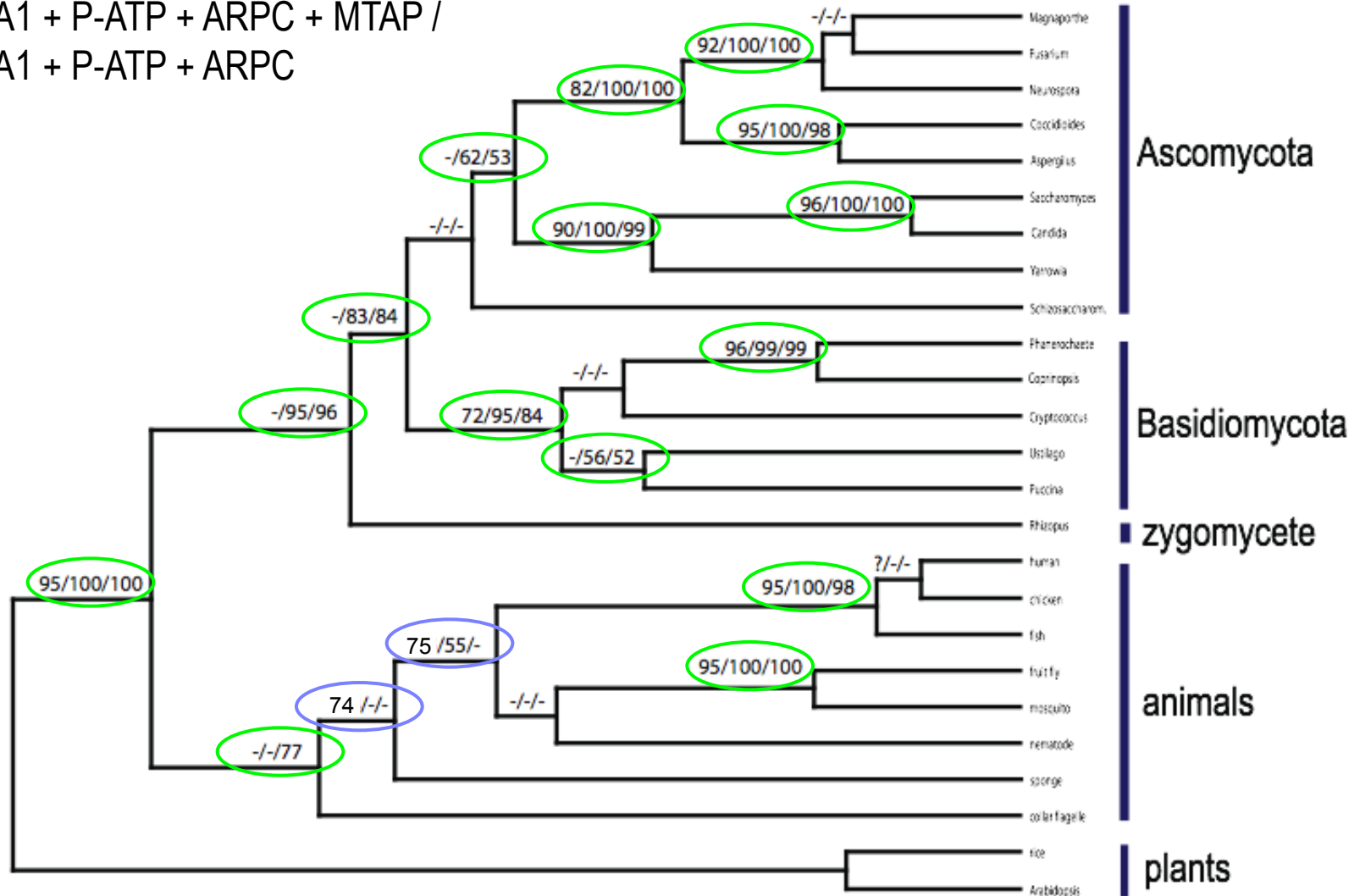


AFTOL 2 optimal marker selection


RPB1 (beginning, middle, and end) /

RPA1 + P-ATP + ARPC + MTAP /

RPA1 + P-ATP + ARPC



PhyDesign is an easy to use web app for obtaining PI profiles



- PhyDesign home
- Instructions
- FAQ
- Townsend lab

① Alignment file:

Upload Alignment

② Chronogram file:

Upload Tree

③ Choose program: ?

HyPhy

- OR -

○ Instead,
input a site rate file

PhyDesign

Profiling phylogenetic informativeness

1. INPUT DATA

2. RATES RESULTS

3. PI PROFILES

To estimate phylogenetic informativeness profiles (Townsend 2007), the PhyDesign web application is partitioned into of 3 components:

- a form to upload information and choose an application to calculate the substitution rates for each alignment site,
- a table listing the rate vectors produced by the analysis, including links for download, and
- a user-friendly graphical interface to plot phylogenetic profiles and calculate integration values.

To obtain the site rate distributions for each locus, we require for input **(1) an alignment in Nexus or Phylip format** of markers pruned to contain sequences only from taxa for which the tree topology is fairly well known, and **(2) an ultrametric tree in Newick (recommended) or nexus format** for those taxa. The ultrametric tree can be introduced as a chronogram (ultrametric tree with branch lengths proportional to time); if so, profiles will be aligned with respect to time.

1. Alignment: NONE UPLOADED
2. Tree: NONE UPLOADED

A link with the rate calculation results will be sent to the email address supplied below:

E-mail

Submit Job

When input the data, only 3 steps are needed

① Alignment file:

Upload Alignment

② Chronogram file:

Upload Tree

③ Choose program: ?

HyPhy

- OR -

○ Instead,
input a site rate file

PhyDesign
Profiling phylogenetic informativeness

1. INPUT DATA

2. RATES RESULTS

3. PI PROFILES

To estimate phylogenetic informativeness profiles (Townsend 2007) , the PhyDesign web application is partitioned into of 3 components:

- a form to upload information and choose an application to calculate the substitution rates for each alignment site,
- a table listing the rate vectors produced by the analysis, including links for download, and
- a user-friendly graphical interface to plot phylogenetic profiles and calculate integration values.

To obtain the site rate distributions for each locus, we require for input **(1) an alignment in Nexus or Phylip format** of markers pruned to contain sequences only from taxa for which the tree topology is fairly well known, and **(2) an ultrametric tree in Newick (recommended) or nexus format** for those taxa. The ultrametric tree can be introduced as a chronogram (ultrametric tree with branch lengths proportional to time); if so, profiles will be aligned with respect to time.

1. Alignment: NONE UPLOADED
2. Tree: NONE UPLOADED

A link with the rate calculation results will be sent to the email address supplied below:

E-mail

Submit Job

When input the data, only 3 steps are needed

① Alignment file:

Upload

② Chronogram file:

Upload Tree

③ Choose program: ?

HyPhy

- OR -

○ Instead,
input a site rate file

PhyDesign
Profiles

Alignment with format [nexus] uploaded. ×
Partitions named [genes] has been created.

1. INPUT DATA

To estimate phylogenetic informativeness profiles (Townsend 2007) , the PhyDesign web application is partitioned into of 3 components:

- a form to upload information and choose an application to calculate the substitution rates for each alignment site,
- a table listing the rate vectors produced by the analysis, including links for download, and
- a user-friendly graphical interface to plot phylogenetic profiles and calculate integration values.

To obtain the site rate distributions for each locus, we require for input **(1) an alignment in Nexus or Phylip format** of markers pruned to contain sequences only from taxa for which the tree topology is fairly well known, and **(2) an ultrametric tree in Newick (recommended) or nexus format** for those taxa. The ultrametric tree can be introduced as a chronogram (ultrametric tree with branch lengths proportional to time); if so, profiles will be aligned with respect to time.

1. Alignment: NONE UPLOADED
2. Tree: NONE UPLOADED

A link with the rate calculation results will be sent to the email address supplied below:

E-mail

Submit Job

When input the data, only 3 steps are needed

① Alignment file:

Upload

② Chronogram file:

Upload

③ Choose program: ?

HyPhy

- OR -

○ Instead,
input a site rate file

PhyDesign
Profiling phylogenetic informativeness

1. INPUT DATA

Tree format not recognised or branch length not present. We recommend to use newick format. ✕

To estimate phylogenetic informativeness, the application is partitioned into of 3 components:

- a form to upload information and choose an application to calculate the substitution rates for each alignment site,
- a table listing the rate vectors produced by the analysis, including links for download, and
- a user-friendly graphical interface to plot phylogenetic profiles and calculate integration values.

To obtain the site rate distributions for each locus, we require for input **(1) an alignment in Nexus or Phylip format** of markers pruned to contain sequences only from taxa for which the tree topology is fairly well known, and **(2) an ultrametric tree in Newick (recommended) or nexus format** for those taxa. The ultrametric tree can be introduced as a chronogram (ultrametric tree with branch lengths proportional to time); if so, profiles will be aligned with respect to time.

1. Alignment: NONE UPLOADED
2. Tree: NONE UPLOADED

A link with the rate calculation results will be sent to the email address supplied below:

E-mail

Submit Job

PhyDesign incorporates different substitution models to estimate the rates



- PhyDesign home
- Instructions
- FAQ
- Townsend lab

① Alignment file:

Upload

② Chronogram file:

Upload

③ Choose program: ?

HyPhy

Advanced options

- OR -

○ Instead,
input a site rate file

PhyDesign

Profiling phylogenetic informativeness

1. INPUT DATA

2. RATES RESULTS

3. PI PROFILES

To estimate phylogenetic informativeness, the application is partitioned into three main steps:

- a form to upload data for each alignment
- a table listing the substitution rates
- a user-friendly interface to select the values.

To obtain the site rates, the user must provide a

Nexus or Phylip format file

tree topology is fairly simple

or nexus format for

(ultrametric tree with

respect to time.

1. Alignment: 61

2. Tree: r8s_tree

HYPHY Advanced Options

- ☒ Base Frequencies

A: 0.25 C: 0.25 G: 0.25 T: 0.25

- Substitution rate matrix:

	A	C	G	T
A *				
C	1	*		
G	1	1	*	
T	1	1	1	*

For the **JC** model:

- set all matrices values to 1 and frequencies to 0.25.

For the **F81** model:

- let the program calculate base frequencies from the data matrix unchecking the

A link with the rate calculation results will be sent to the email address supplied below:

E-mail

Submit Job

Analysis of the alignment and chronogram produces site rates for each locus



- PhyDesign home
- Instructions
- FAQ
- Townsend lab

Get Profiles

1.INPUT DATA

2.RATES RESULTS

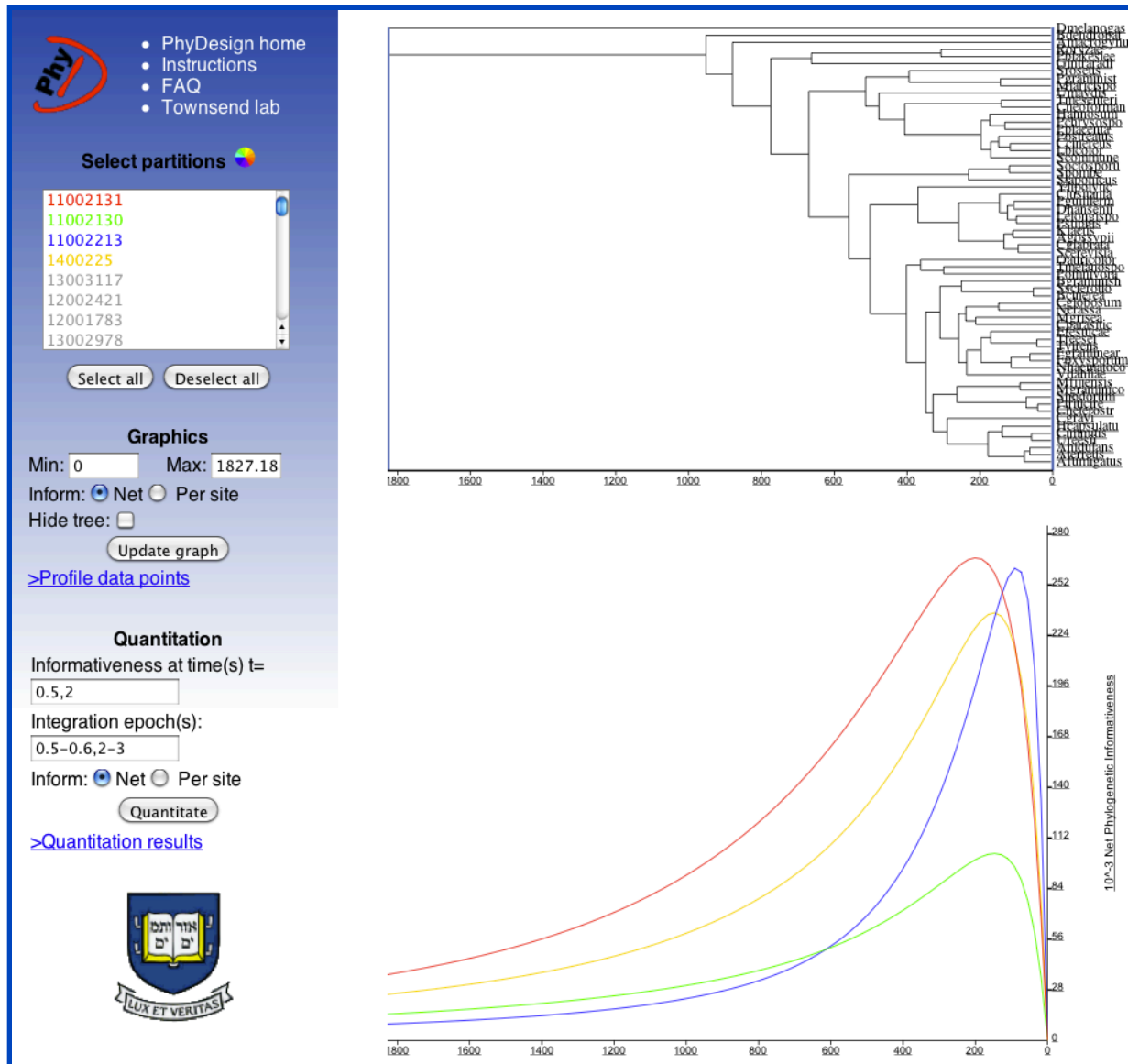
3.PI PROFILES

This table presents the rate estimation results for each site. Site rates are reported partitioned as in the rows that you defined in the input alignment. There is link to download the original output files produced by the rate estimation program for all partitions. A file with all the rate data, entitled "[site rate file](#)" is also created. It can be downloaded using the links in the side bar. Number of profiles and color for each one can be modified here or after viewing profiles.

To obtain the profiles, just click the "Get Profiles" button.

#	Partition	Site rates	#Sites	#Rates	#Undef.	Profile all	Assign Colors
1	11002131	Allrates	326	326	0	<input checked="" type="checkbox"/>	#FF0000
2	11002130	Allrates	138	138	0	<input checked="" type="checkbox"/>	#00FF00
3	11002213	Allrates	135	135	0	<input checked="" type="checkbox"/>	#0000FF
4	1400225	Allrates	303	303	0	<input checked="" type="checkbox"/>	#FFCC00
5	13003117	Allrates	163	163	0	<input type="checkbox"/>	#999999
6	12002421	Allrates	635	635	0	<input type="checkbox"/>	#999999
7	12001783	Allrates	542	542	0	<input type="checkbox"/>	#999999
8	13002978	Allrates	114	114	0	<input type="checkbox"/>	#999999
9	15002822	Allrates	150	150	0	<input type="checkbox"/>	#999999
10	12002831	Allrates	456	456	0	<input type="checkbox"/>	#999999
11	12002908	Allrates	289	289	0	<input type="checkbox"/>	#999999

PhyDesign displays PI profiles based on the inferred rates.





- PhyDesign home
- Instructions
- FAQ
- Townsend lab

Select partitions



11002131
11002130
11002213
1400225
13003117
12002421
12001783
13002978

Select all

Deselect all

Graphics

Min: 0 Max: 1827.18

Inform: ☒ Net ☐ Per site

Hide tree: ☐

Update graph

[>Profile data points](#)

Quantitation

Informativeness at time(s) t=

0.5,2

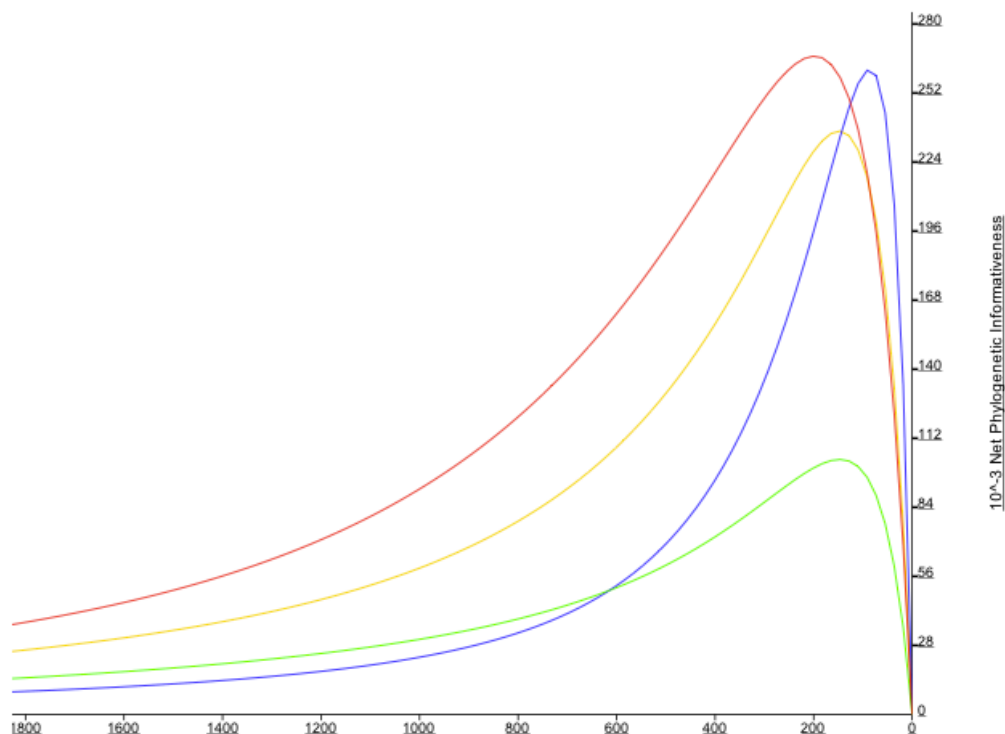
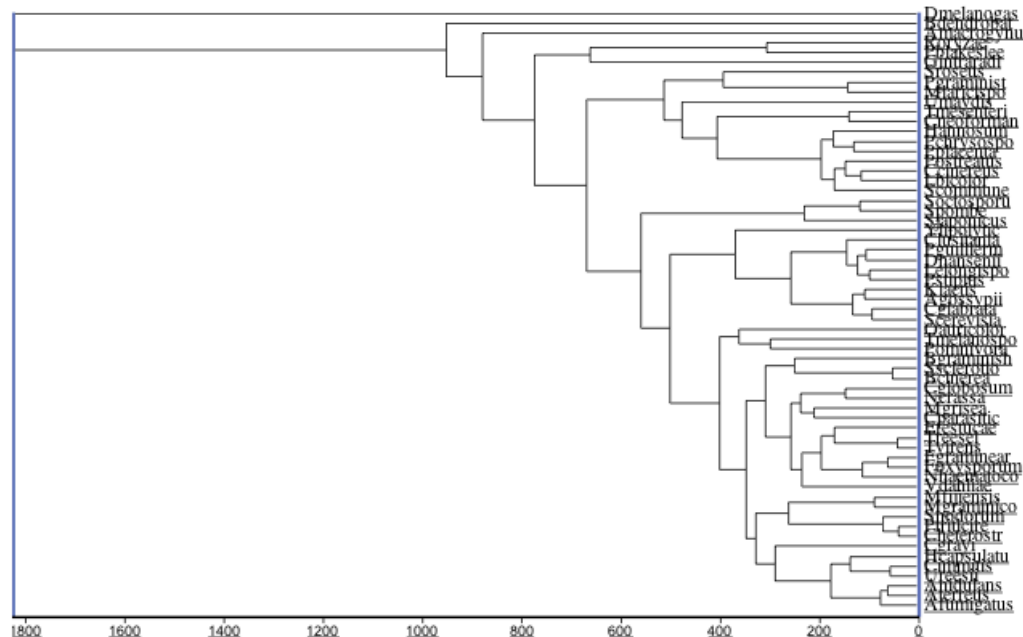
Integration epoch(s):

0.5-0.6,2-3

Inform: ☒ Net ☐ Per site

Quantitate

[>Quantitation results](#)



1) Uploading information for analysis

In order to provide an analysis of phylogenetic informativeness, PhyDesign requires the site rate distribution for each locus. To obtain the rates, the user will need to input two files: (1) an alignment in Nexus or Phylip format of loci of interest pruned to contain a set of taxa for which the tree topology is fairly well known, and (2) an ultrametric tree for those taxa in Newick (recommended) or Nexus format. The ultrametric tree can be either a chronogram -- an ultrametric tree with branch lengths proportional to time -- or it can be in unspecified molecular evolutionary units. After uploading the alignment and the tree files, the user can choose a program from the drop-down menu with which to obtain the substitution rates. Once a program has been chosen, it is possible to access to advance options where the user will be offered with different evolutionary models and parameters. For DNA sequences, we recommend use of HyPhy, for which a HyPhy batch file was created to implement all time-reversible models. Unlike DNARates, HyPhy also accepts multifurcating trees. For amino acid sequences, rate4site is provided. To facilitate extensive analyses of large datasets, the user is also asked to provide an email address where a link to the rates result will be sent.

Alternatively, if the rate distribution for each locus is known, the user can input a file that includes these rates in a specific format for the markers to be analyzed. (For more information on accepted file types and formats, check the [FAQ](#).) If an alignment and a tree are uploaded, the site rate vectors will first be calculated. A file containing these rates with the proper format can be downloaded for future use in the site rate form, eliminating the need to repeat this calculation step.

PhyDesign home
• Instructions
• FAQ
• Townsend lab

- Alignment file:**
- Chronogram file:**
- Choose program:**
HyPhy

>Instead, input a site rate file

PhyDesign home
• Instructions
• FAQ
• Townsend lab

- Site rate file:**
- Chronogram file (optional):**

>Back to alignment/tree input

2) Downloading and choosing rates for profiling

By clicking the emailed link, the user will be taken to a page similar to one in the image. The table presents the rate estimation results for each alignment site divided by the partitions/genes (rows) that you defined in the input alignment. There is a link for each partition to the original output produced by the rate estimation program. Please check this file, just in case PhyDesign is missing some error message produced by the rate estimation program. This table also summarizes the number of sites for which a mutation rate could be calculated (#Rates), and the number of faulty sites for which this calculation was not possible (#Undef.). The last two columns of the table are used to specify the partitions to be profiled and the colors for representing them. Profiles and color for each one can be modified here or later.

Additionally, if the rates were estimated with PhyDesign, two results files are offered as downloads in the left panel: (1) a compressed file containing individual rate files for each locus, and (2) a single file containing site rate vectors for all loci. The latter can be downloaded for future uploading in the site rate form, eliminating the need to repeat the rate calculation. To obtain the profiles from the partitions selected, just click in the "Get Profiles" button.

PhyDesign home
• Instructions
• FAQ
• Townsend lab

Download all files
Rate vectors file
Get Profiles

#	Partition	Rates by prog.	#Sites	#Rates	#Undef.	Profile all	Assign Colors
1	EN5000000083782_EPYC_sh	Rates by hyphy	969	969	0	✓	#FF0000
2	EN50000000140382_HMG20A_sh	Rates by hyphy	1044	1044	0	✓	#00FF00
3	EN50000000124810_OPN5_sh	Rates by hyphy	1039	1039	0	✓	#0000FF
4	EN50000000166510_CCD68_sh	Rates by hyphy	991	991	0	✓	#FFCC00
5	EN50000000136404_TH6SF1_sh	Rates by hyphy	1026	1026	0	☐	#999999
6	EN50000000133800_LIVE1_sh	Rates by hyphy	951	951	0	☐	#999999
7	EN50000000085365_SCAMP1_sh	Rates by hyphy	1006	1006	0	☐	#999999
8	EN50000000144644_GADL1_sh	Rates by hyphy	1014	1014	0	☐	#999999
9	EN50000000102383_ZDHHC15_sh	Rates by hyphy	1014	1014	0	☐	#999999

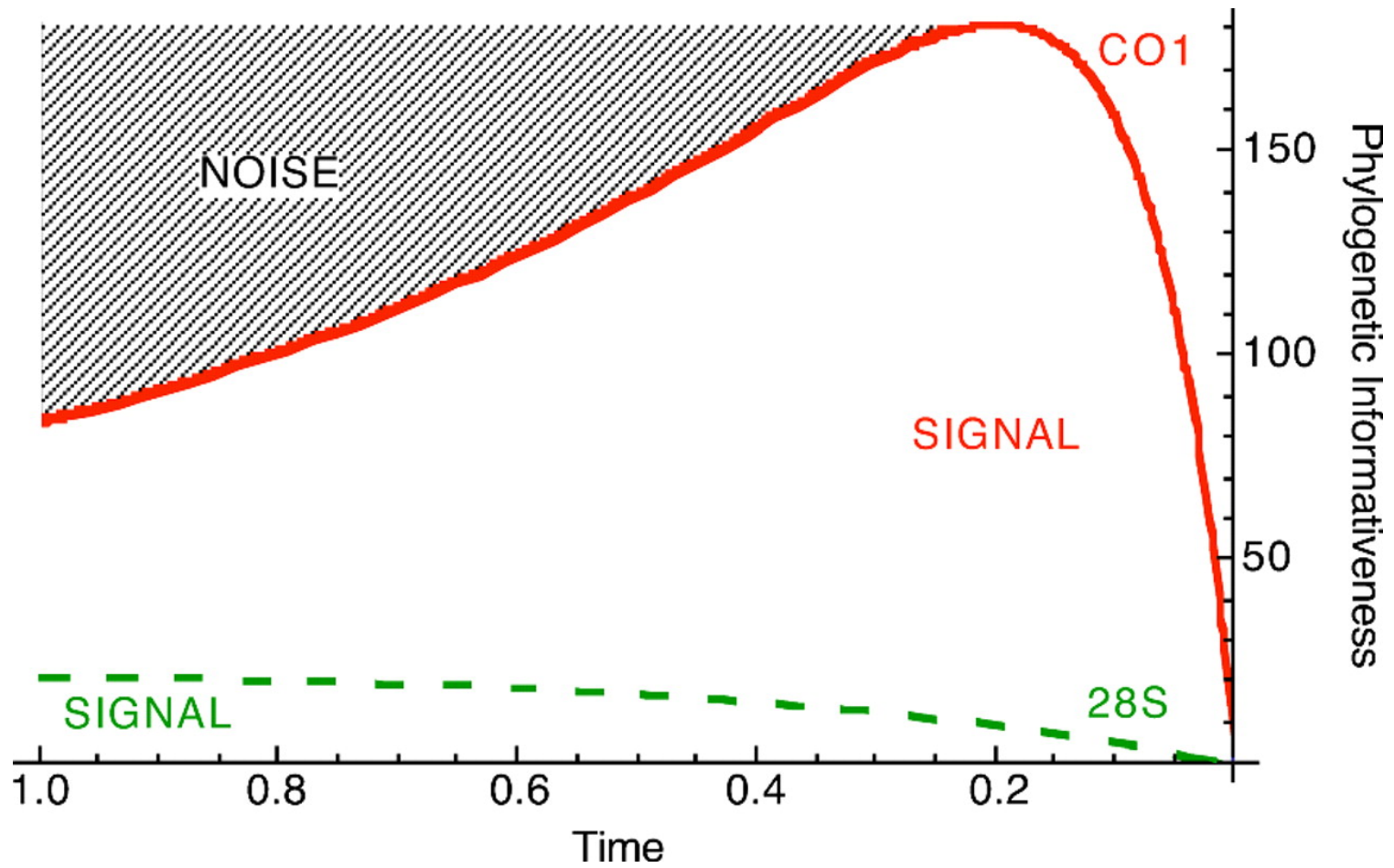
This table presents the rate estimation results for each alignment site divided by the partitions/genes (rows) that you defined in the input alignment. There is a link for each partition to the original output produced by the rate estimation program. A file with [site rate file](#) is also created and it can be downloaded using the links in the side bar. Number of profiles and color for each one can be modified here or later (recommended).

To obtain the profiles just click in the "Get Profiles" button.

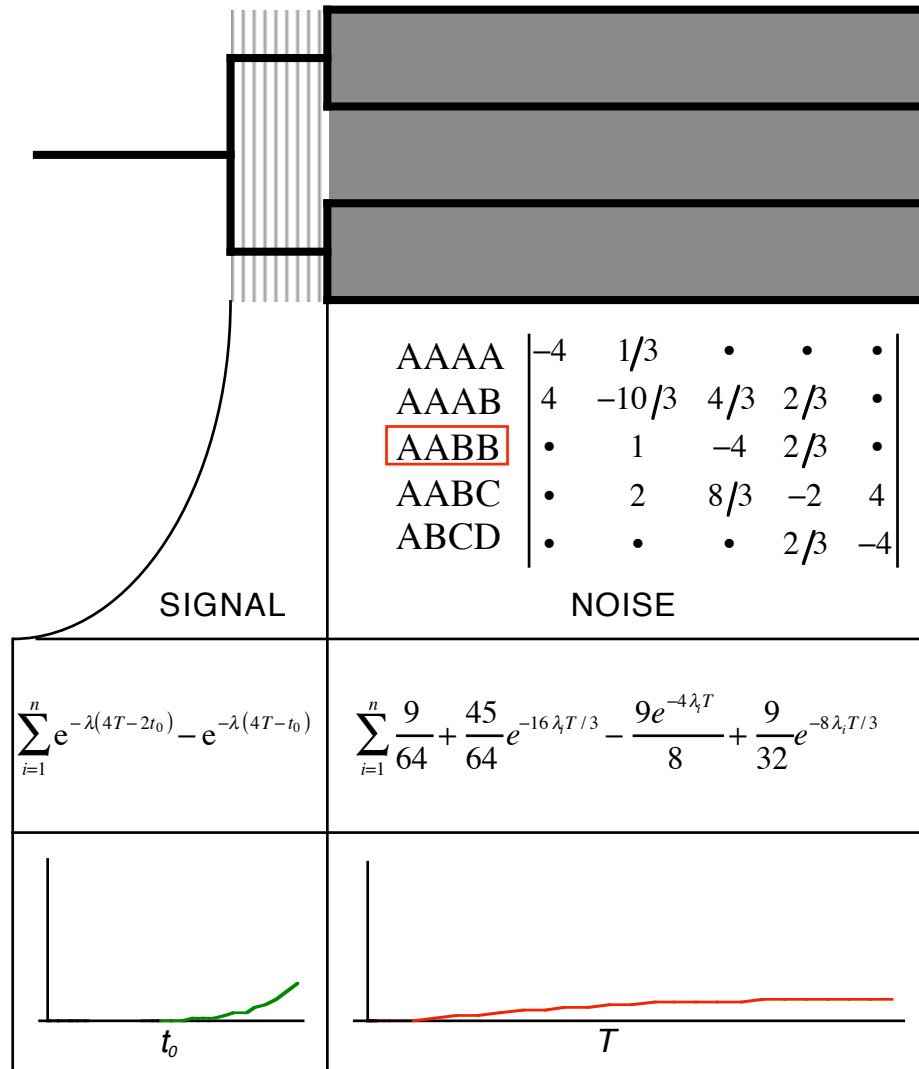
3) Manipulating profile: graph and quantitation

Detailed instructions are available on the website

PI predicts phylogenetic signal for a given gene, but does not quantify noise

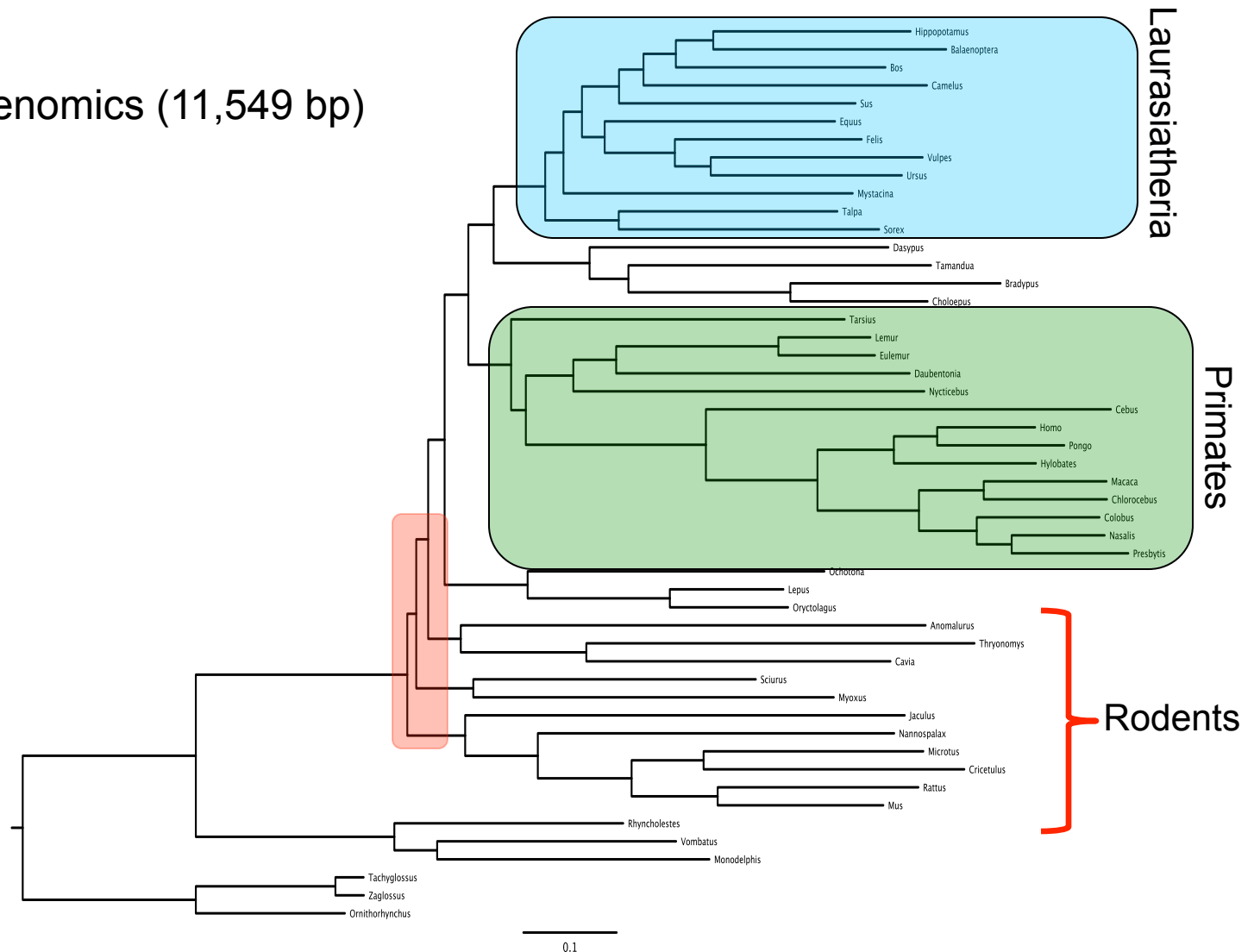


Theory can quantify the effect of parallelism and convergence

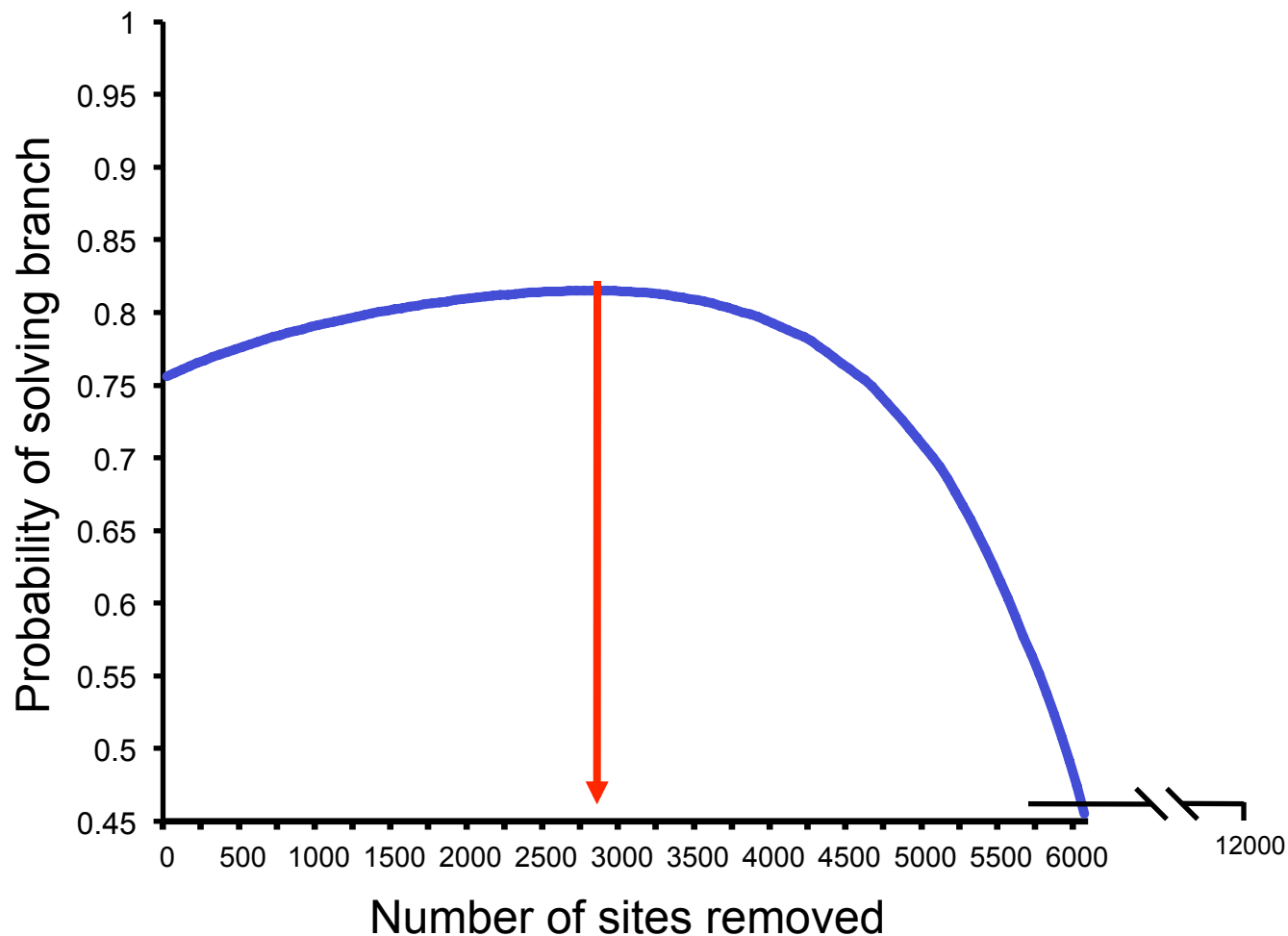


We are implementing this approach to remove noisy sites and, thus, improve resolution

Mammal mitogenomics (11,549 bp)

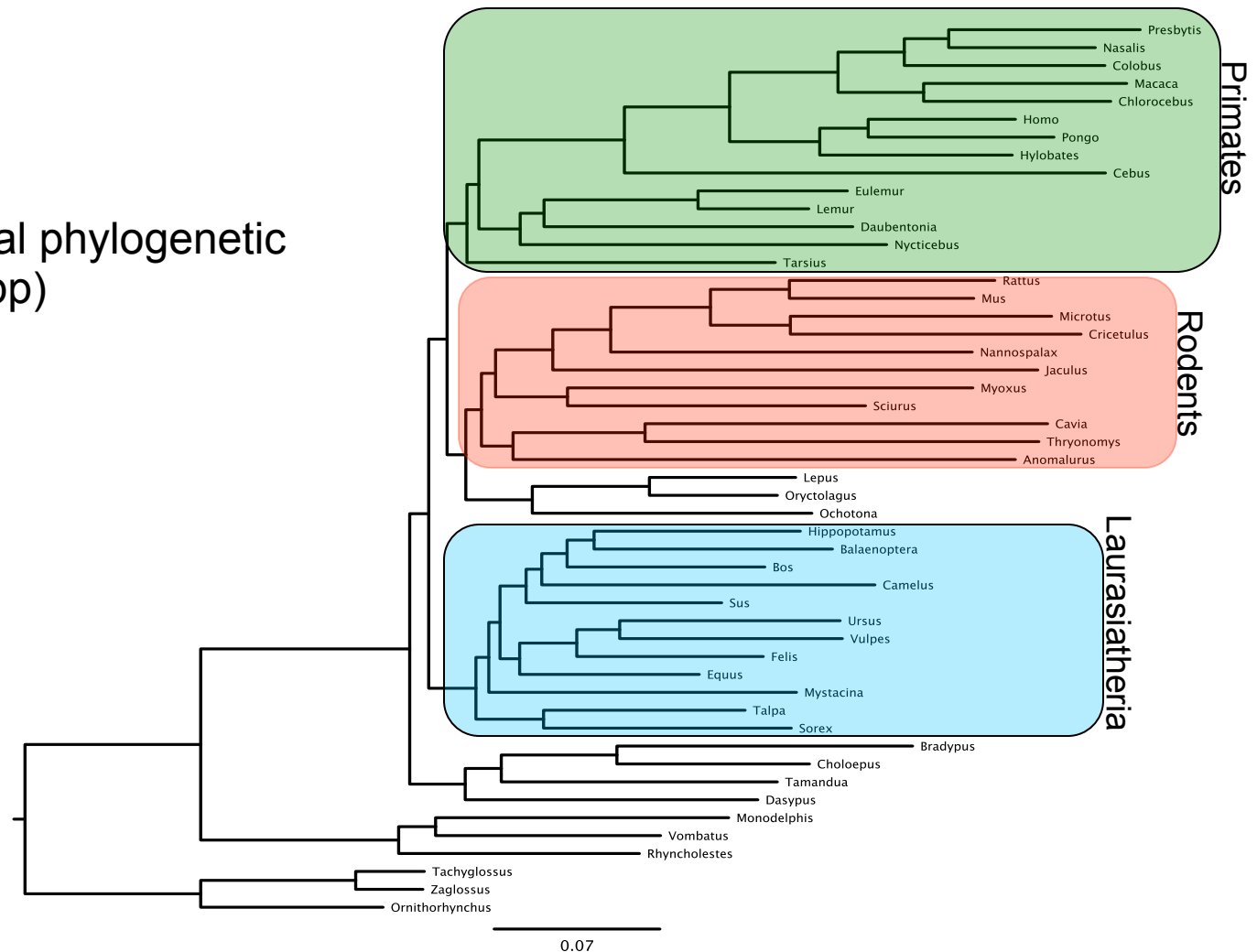


Fast evolving sites can be removed to maximize probability of resolution



Removal of the noisy sites improved resolution

New mammal phylogenetic tree (9,349 bp)



Conclusions

- PI predicts gene performance and it can be used to prioritize gene selection for specific phylogenetic inferences.
- PhyDesign is an easy to use web application for marker selection.
<http://phydesign.townsend.yale.edu>
- Signal and Noise theory provides a way to quantify probability of resolution that incorporates the effect of noise. It is incorporated to the PhyDesign web application.



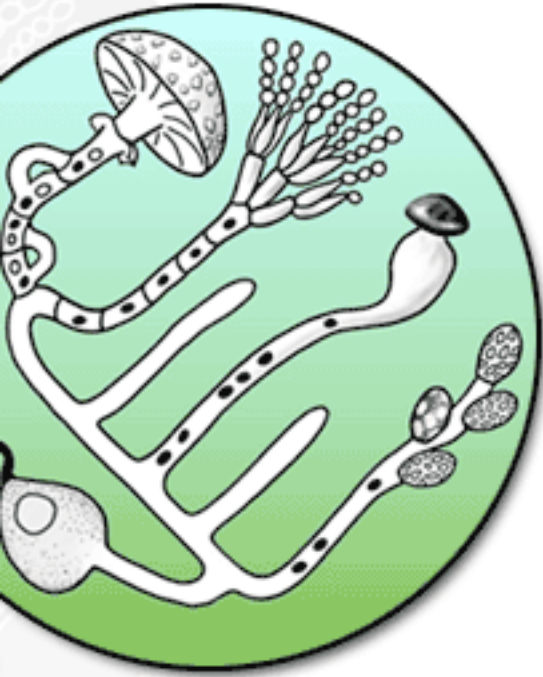
<http://phydesign.townsend.yale.edu/>

HPC Cluster at Yale



AFTOL 1

Assembling the Fungal Tree of Life



12 initial candidates.

6 final genes:

RPB1 - RNA polymerase II subunit A

RPB2 - RNA polymerase II subunit B

TEF - Translation elongation factor

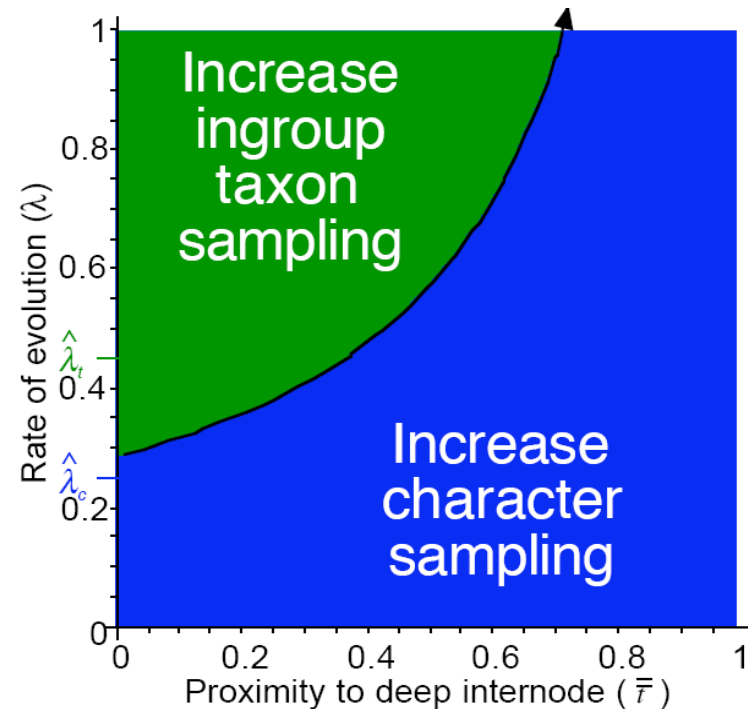
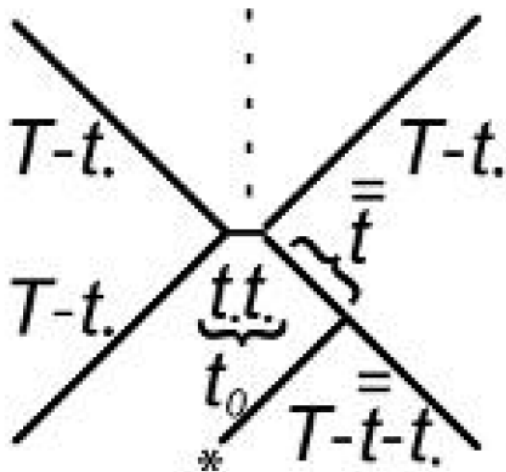
SSU - Small subunit rRNA

LSU - Large subunit rRNA

MSSU - mitochondria SSU rRNA

We have expanded the PI to quantify the relative utility of increased taxonomic versus character sampling

$$\text{PITA} = \sum_{i=1}^n e^{-2(T_1+T_2)\lambda_i} (1 - e^{-t_0\lambda_i}) (1 - e^{-(T_1 - \bar{t})\lambda_i})$$

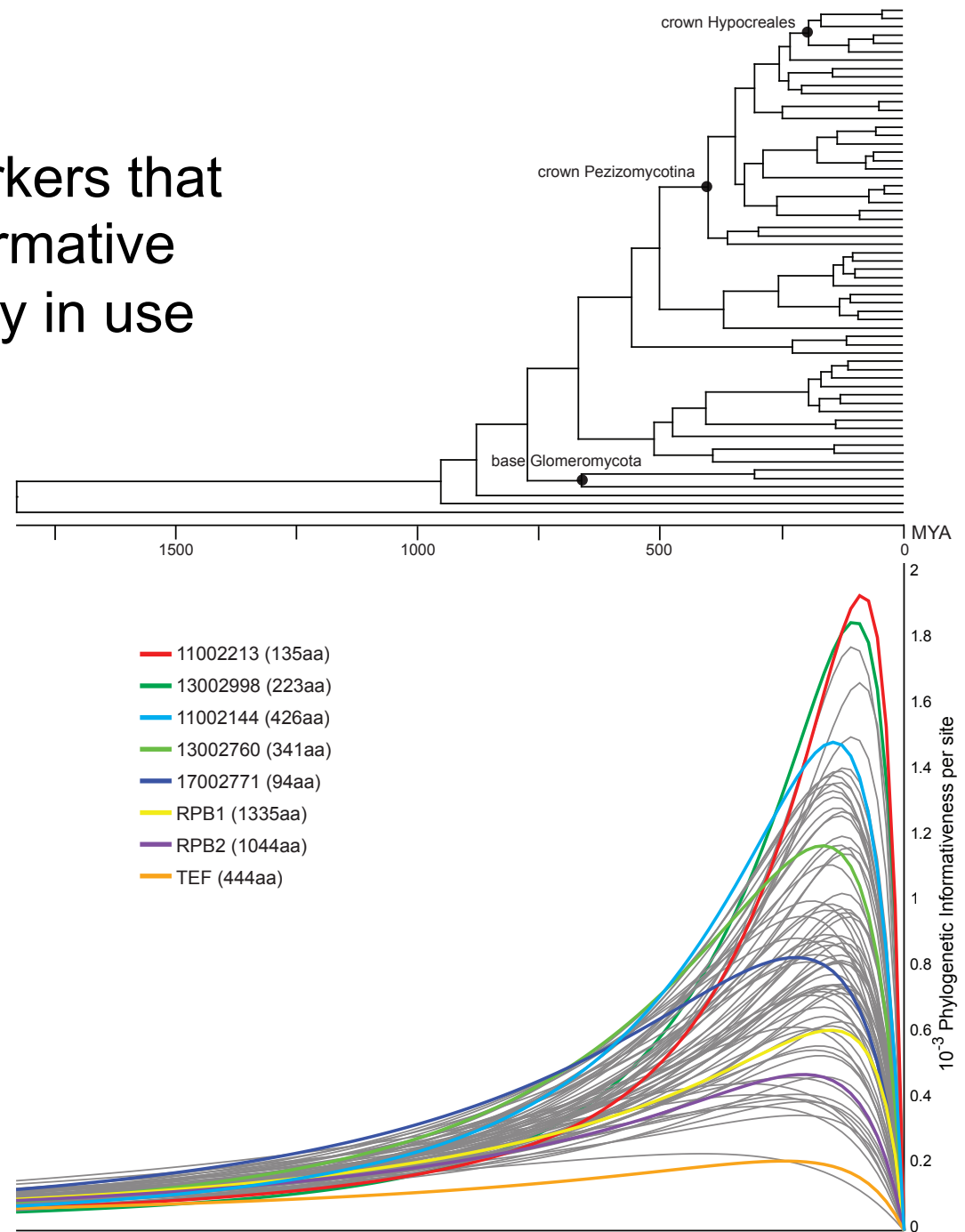


To quantitatively evaluate phylogenetic informativeness as a procedure for selecting loci

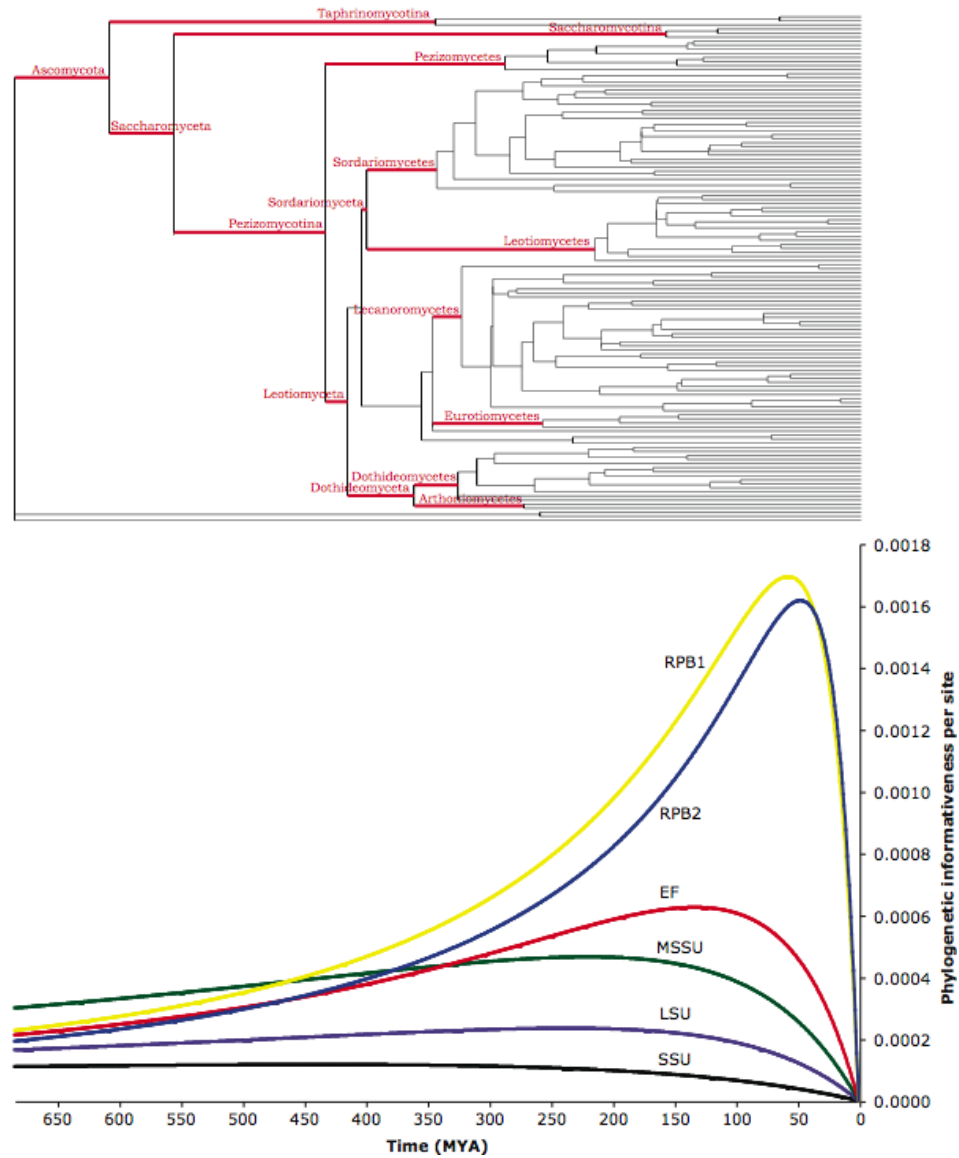


Genes were ranked by phylogenetic informativeness
Vs
the accuracy and robustness of recovering the reference tree

AFTOL-2 orthologous markers that
are significantly more informative
than standard loci currently in use
in fungal phylogenetics



Phylogenetic Informativenesses of AFToL 1 genes



Schoch et al., 2009

PhyDesign

Web application for profiling phylogenetic information content of genes

Create a web page with intuitive user friendly GUI to generate the profiles and quantitations.

- Multi-locus analysis
- Integration over multiple epochs
- Tree aligned with profiles in time scale
- Graphics in SVG

Web Browser

Javascript / Ajax



Server site

CGI / PERL

PhyDesign

Profiling phylogenetic informativeness

1. INPUT DATA	2. RATES RESULTS	3. PI PROFILES
---------------	------------------	----------------

To estimate phylogenetic informativeness profiles (Townsend 2007), the PhyDesign web application is partitioned into of 3 components:

- a form to upload information and choose an application to calculate the substitution rates for each alignment site,
- a table listing the rate vectors produced by the analysis, including links for download, and
- a user-friendly graphical interface to plot phylogenetic profiles and calculate integration values.

To obtain the site rate distributions for each locus, we require for input **(1) an alignment in Nexus or Phylip format** of markers pruned to contain sequences only from taxa for which the tree topology is fairly well known, and **(2) an ultrametric tree in Newick (recommended) or nexus format** for those taxa. The ultrametric tree can be introduced as a chronogram (ultrametric tree with branch lengths proportional to time); if so, profiles will be aligned with respect to time.

1. Alignment: NONE UPLOADED
2. Tree: NONE UPLOADED

A link with the rate calculation results will be sent to the email address supplied below:

E-mail

① Alignment file:

② Chronogram file:

③ Choose program: ?

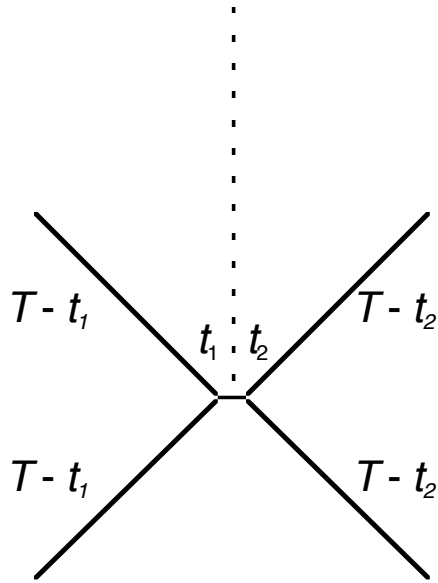
HyPhy

- OR -

☐ Instead,
input a site rate file



Profiling PI allow us to identify genes with high phylogenetic information content during experimental design



Calculate the probability of a site being informative, given a rate of evolution, λ , and letting $t_1 + t_2 \rightarrow 0$.

$$\rho(T; \lambda_1, \dots, \lambda_n) = \sum_{i=1}^n 16 \lambda_i^2 T e^{-4 \lambda_i T}$$

