Can phylogenomic data matrix end incongruence in the tree of life?

系统发育基因组数据能否解决生命之树中的冲突？

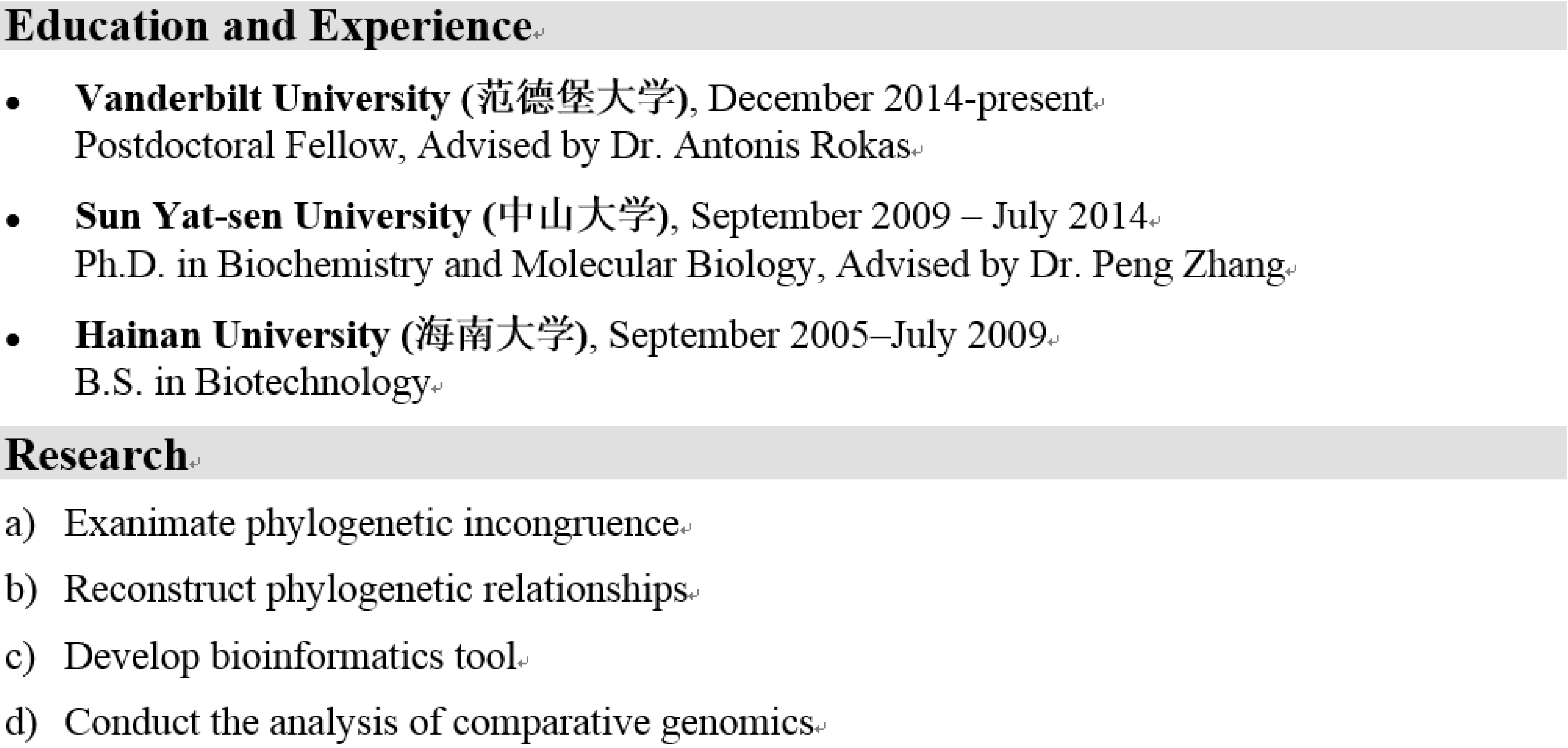
Xing-Xing Shen（沈星星）

https://xingxingshen.github.io/

Rokas Lab

Mar 2018

自我简介



Education and Experience

- **Vanderbilt University (范德堡大学)**, December 2014-present
  Postdoctoral Fellow, Advised by Dr. Antonis Rokas
- **Sun Yat-sen University (中山大学)**, September 2009 – July 2014
  Ph.D. in Biochemistry and Molecular Biology, Advised by Dr. Peng Zhang
- **Hainan University (海南大学)**, September 2005–July 2009
  B.S. in Biotechnology

Research

a) Exanimate phylogenetic incongruence

b) Reconstruct phylogenetic relationships

c) Develop bioinformatics tool

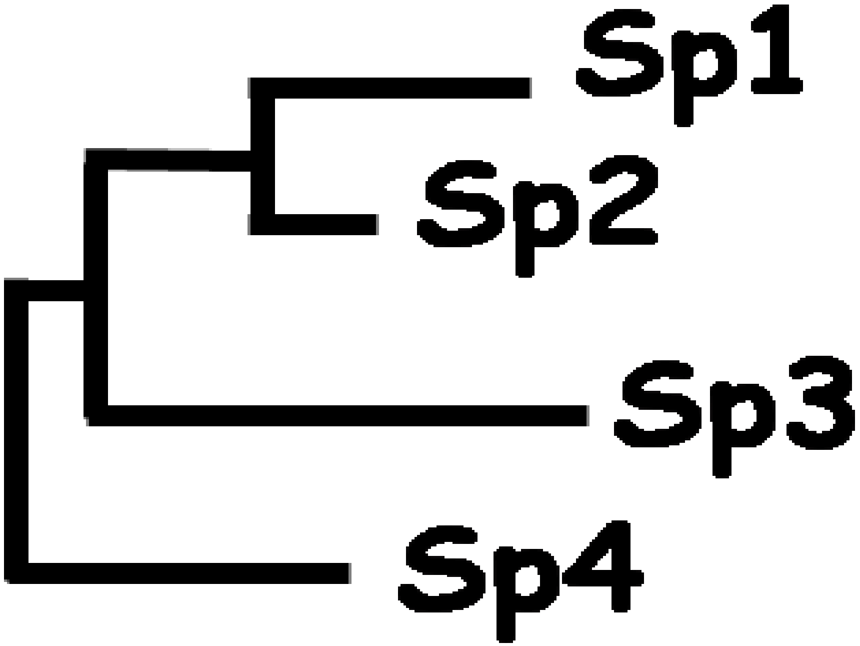
d) Conduct the analysis of comparative genomics

https://xingxingshen.github.io/

Course / workshop

- Workshop on Molecular Evolution (Since 1988)
- Workshop on Phylogenomics (Since 2017)

Phylogenetics



Utilities of phylogenetic tree

- Relationship
- Species delimitation
- Divergence time and Biogeography
- Evolutionary process (selections, ancestral state)

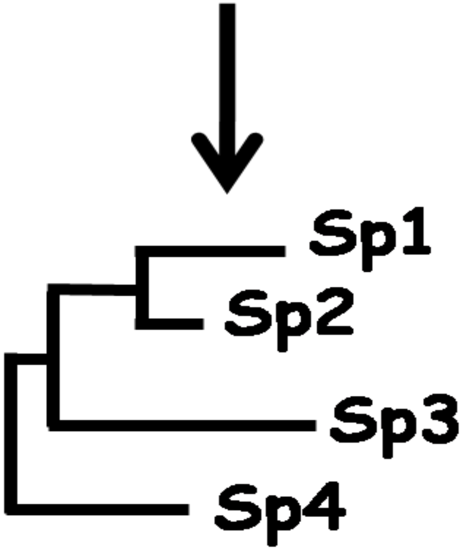
Molecular Phylogenetics

Sp1: TCTGT…AACTCTTT…GAATCGTT…GCC

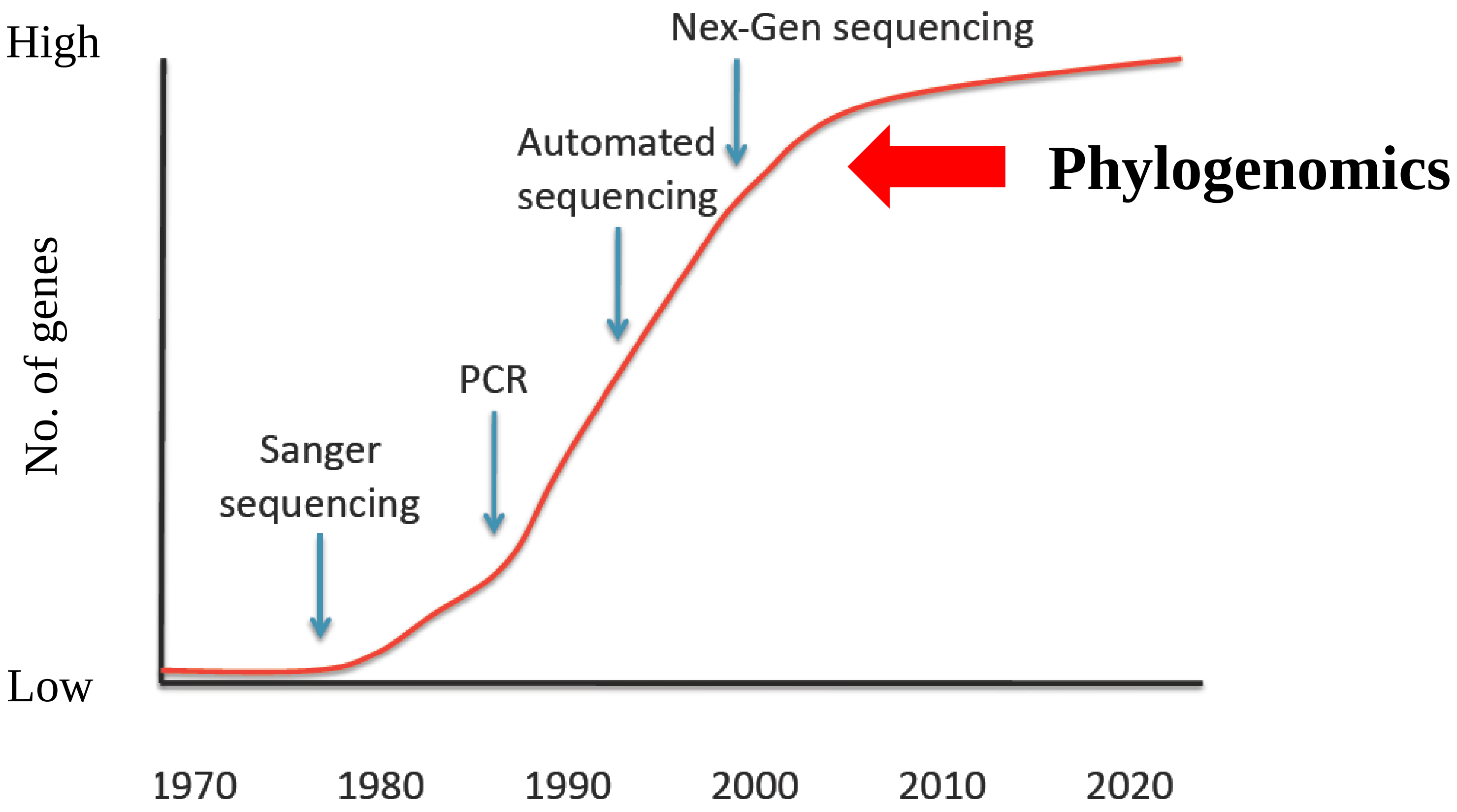
Sp2: TCTGC…GACTCGCT…GGAACGCT…CCC

Sp3: CTTAT…GATCTATT…GGAATATT…CGA

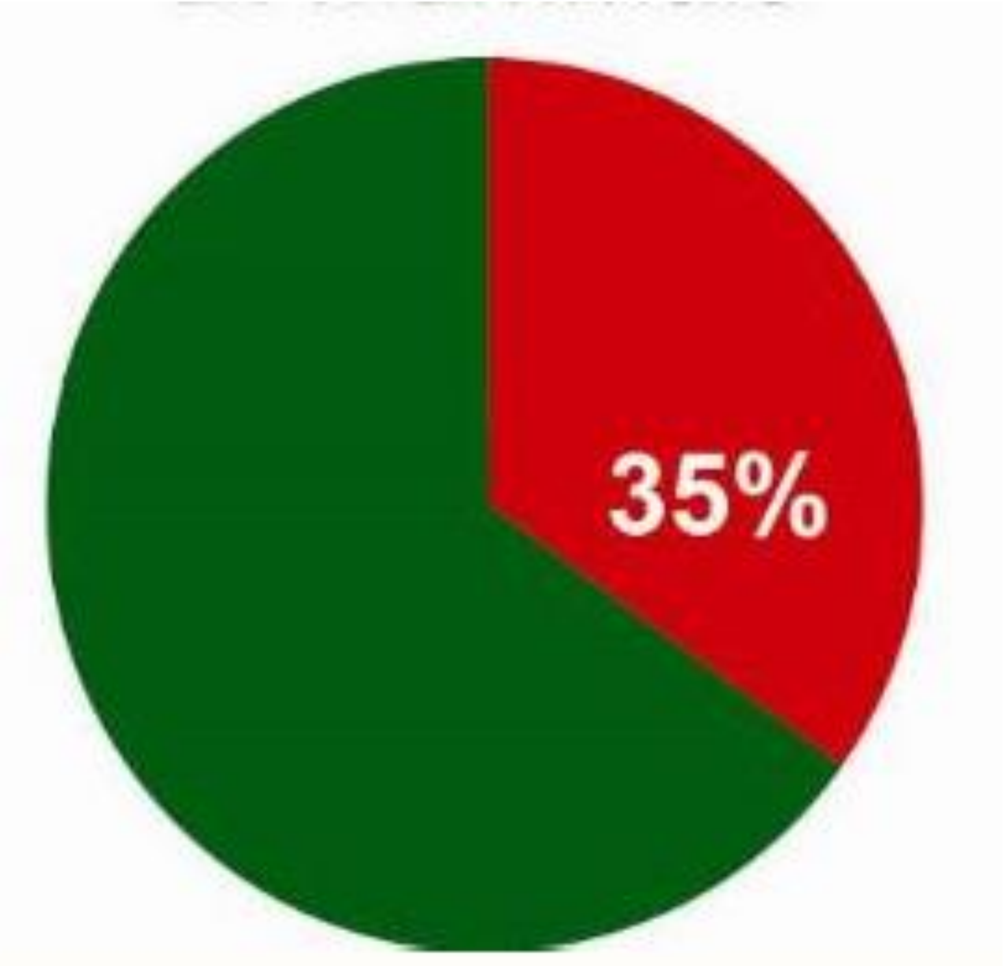
Sp4: CCTAT…GATCCATT…GGACCATT…CCA



Histories of data sampling

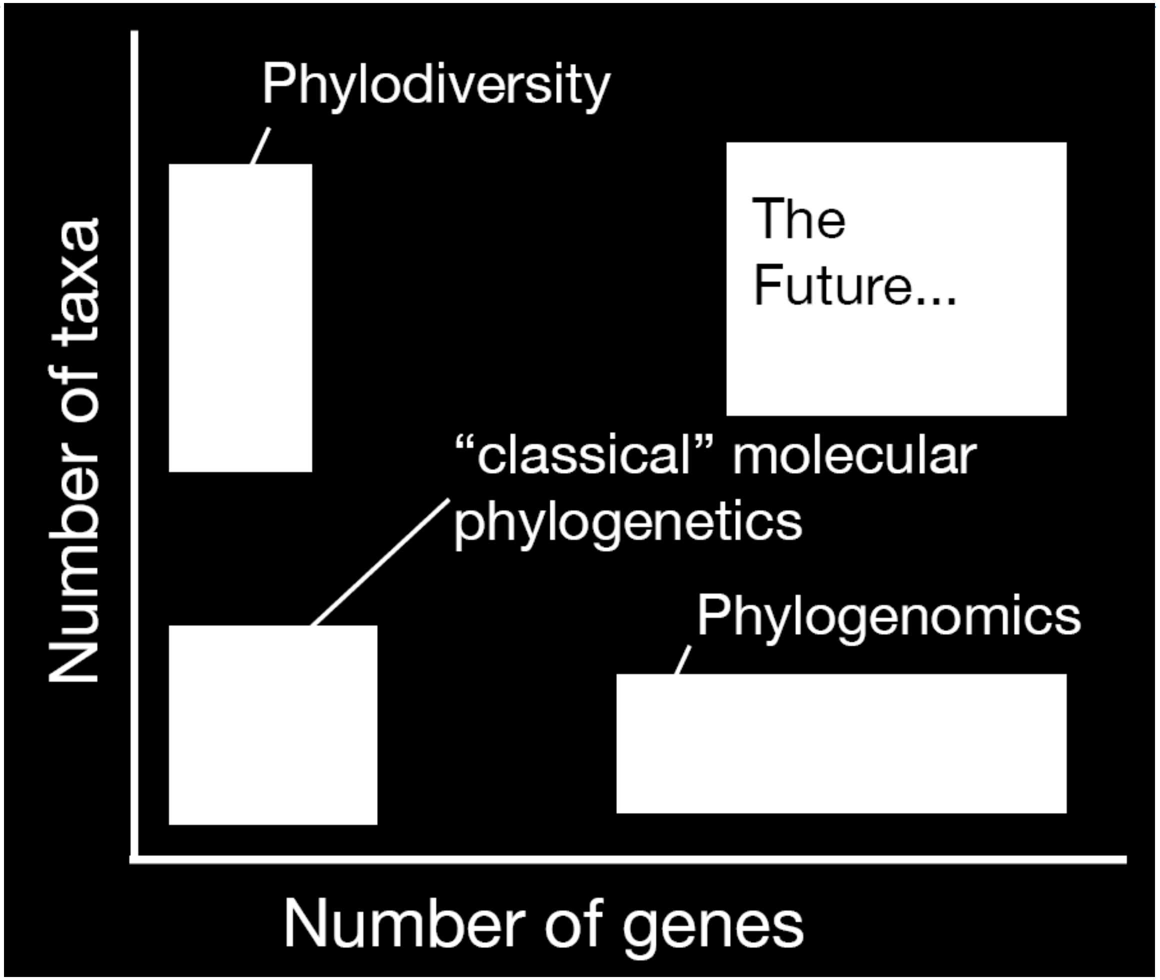


Why we need “many” genes?



Rokas & Chatzimanolis (2008) in Phylogenomics (W. J. Murphy, Ed.)

Histories of data sampling



(Slide from Casey Dunn)

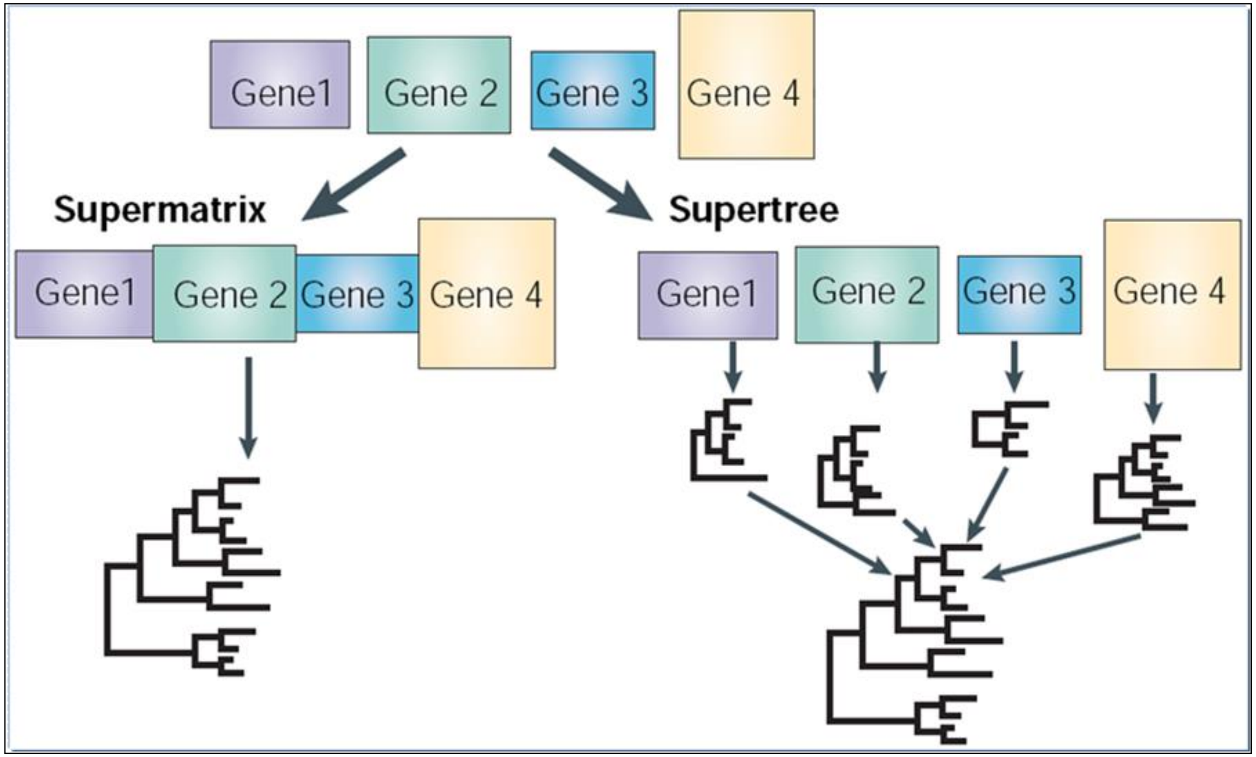
Phylogenomics

Phylogenomics: inference of species phylogeny with lots of genes

- Whole genome
- “Whole” transcriptome
- Targeted enrichment
- Directed PCR

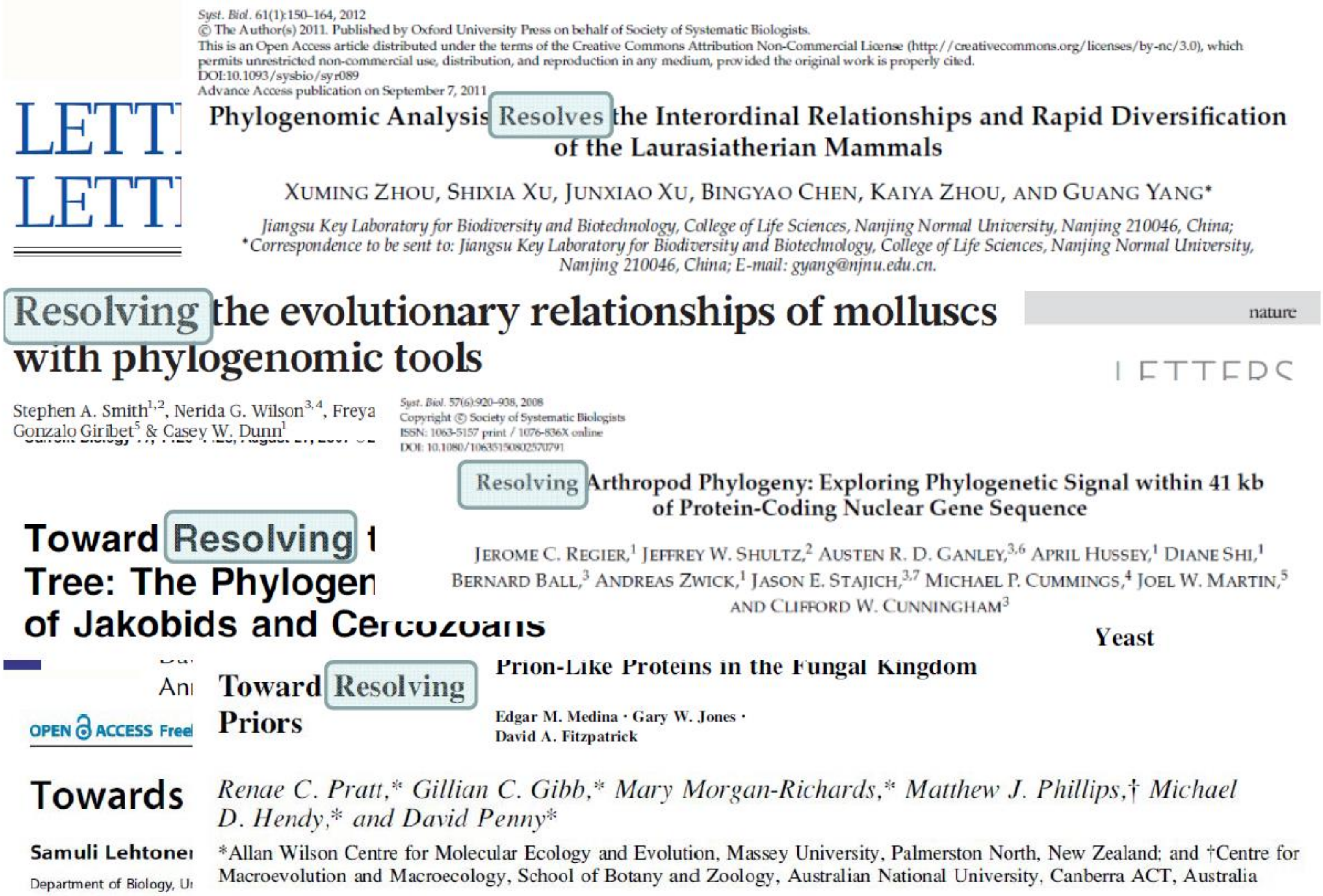
(Slide from Casey Dunn)

Phylogenomics



Delsuc et al. 2005

Phylogenomic power



Syst. Biol. 61(1):150–164, 2012
© The Author(s) 2011. Published by Oxford University Press on behalf of Society of Systematic Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syr089
Advance Access publication on September 7, 2011

LETT
LETT

Phylogenomic Analysis Resolves the Interordinal Relationships and Rapid Diversification of the Laurasiatherian Mammals

XUMING ZHOU, SHIXIA XU, JUNXIAO XU, BINGYAO CHEN, KAIYA ZHOU, AND GUANG YANG*

Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China;
**Correspondence to be sent to: Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China; E-mail: gyang@njnu.edu.cn.*

Resolving the evolutionary relationships of molluscs with phylogenomic tools

nature

LETTERS

Stephen A. Smith[1,2], Nerida G. Wilson[3,4], Freya
Gonzalo Giribet[5] & Casey W. Dunn[1]

Syst. Biol. 57(6):920–938, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802570791

Resolving Arthropod Phylogeny: Exploring Phylogenetic Signal within 41 kb of Protein-Coding Nuclear Gene Sequence

JEROME C. REGIER,[1] JEFFREY W. SHULTZ,[2] AUSTEN R. D. GANLEY,[3,6] APRIL HUSSEY,[1] DIANE SHI,[1] BERNARD BALL,[3] ANDREAS ZWICK,[1] JASON E. STAJICH,[3,7] MICHAEL P. CUMMINGS,[4] JOEL W. MARTIN,[5] AND CLIFFORD W. CUNNINGHAM[3]

Toward Resolving t
Tree: The Phylogen
of Jakobids and Cercozoans

Yeast

Prion-Like Proteins in the Fungal Kingdom

An

OPEN ACCESS Freel

Toward Resolving
Priors

Edgar M. Medina · Gary W. Jones ·
David A. Fitzpatrick

Towards

Renae C. Pratt, Gillian C. Gibb,* Mary Morgan-Richards,* Matthew J. Phillips,† Michael D. Hendy,* and David Penny**

*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and †Centre for Macroevolution and Macroecology, School of Botany and Zoology, Australian National University, Canberra ACT, Australia

Samuli Lehtone

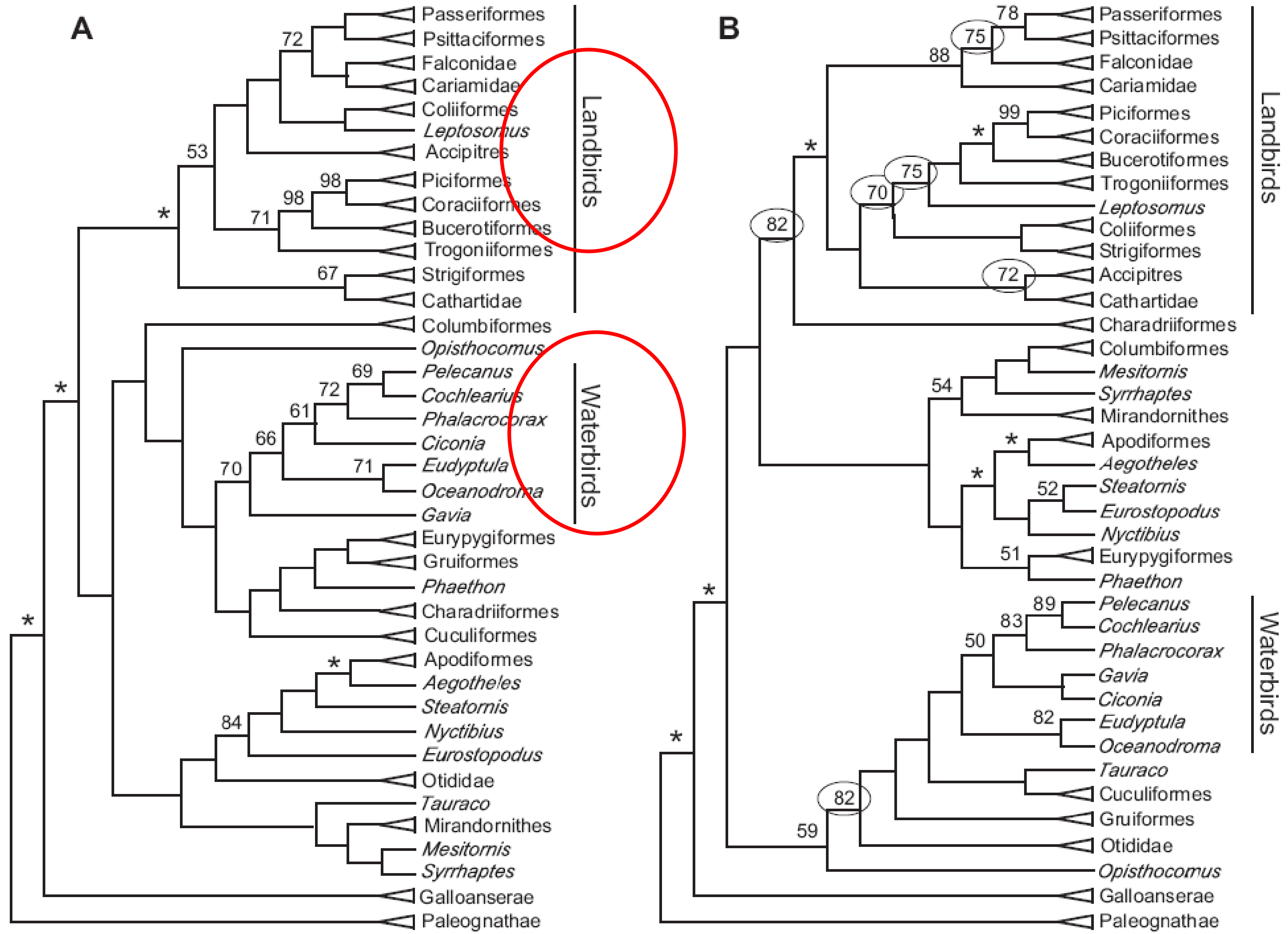
Department of Biology, U

(Slide from Antonis Rokas)

But ……

Incongruences still exist in phylogenomic studies

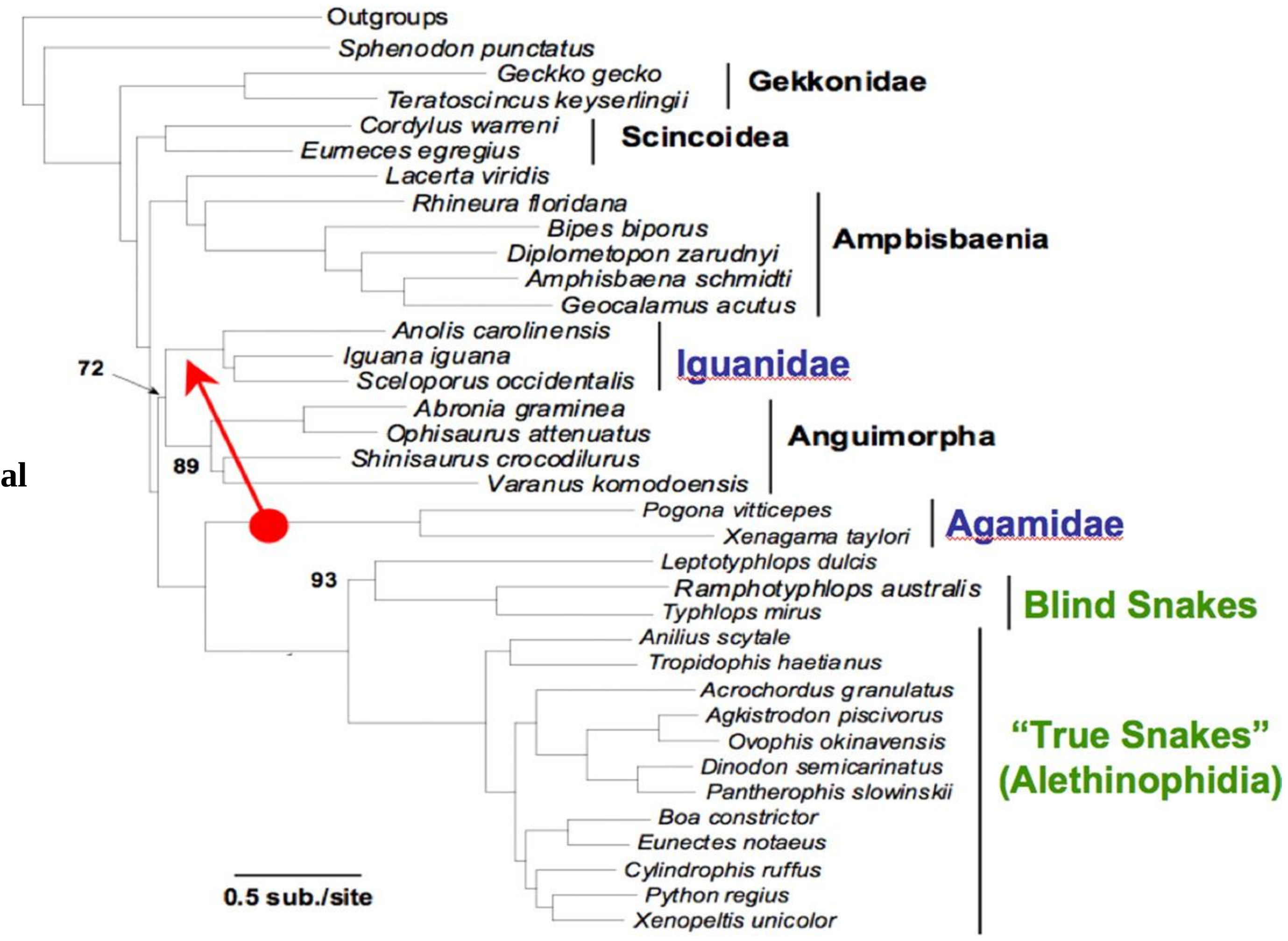
Incongruence: avian



Kimball et al, MPE 2013

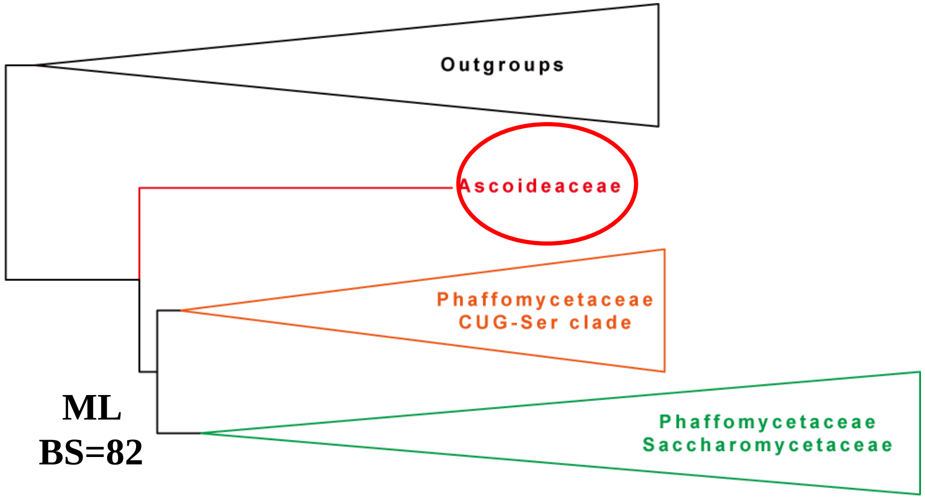
Incongruence: squamates

Mitochondrial genome



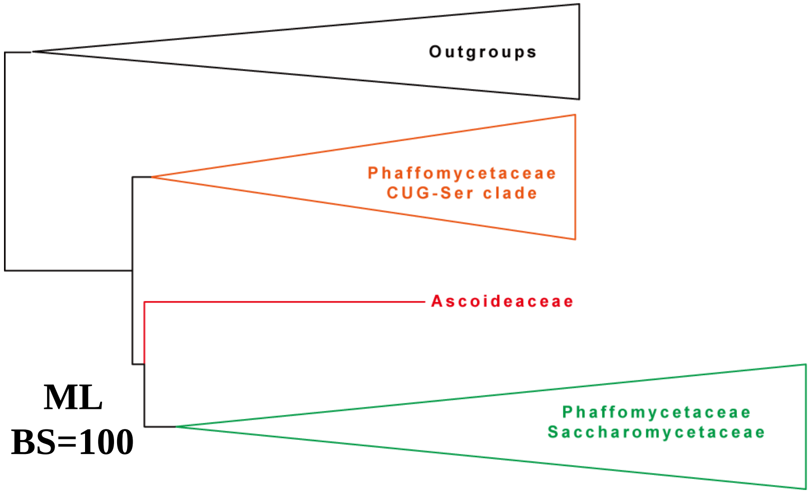
Castoe et al. PNAS 2009

Incongruence: Ascoideaceae (酱霉科)



1,559-gene and 38-taxon

Riley et al. PNAS 2016



1,233-gene and 86 yeasts

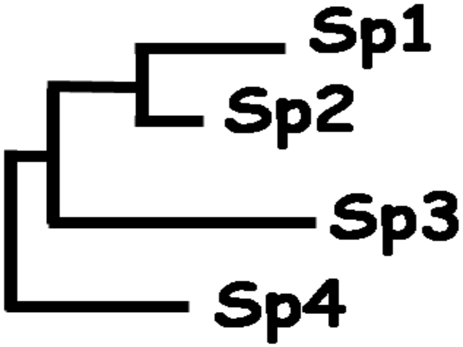
Shen et al. G3 2016

How could we quantify phylogenetic signal for each gene?

Maximum Likelihood (最大似然法)

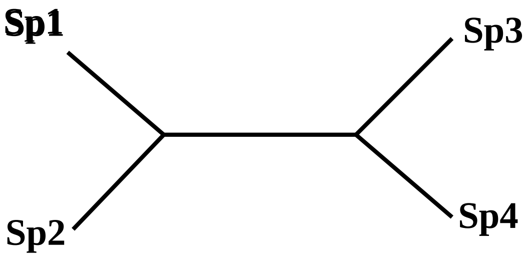
```
Sp1:  TCTGT…AACTCTTT…GAATCGTT…GCC
Sp2:  TCTGC…GACTCGCT…GGAACGCT…CCC
Sp3:  CTTAT…GATCTATT…GGAATATT…CGA
Sp4:  CCTAT…GATCCATT…GGACCATT…CCA
```

Substitution model

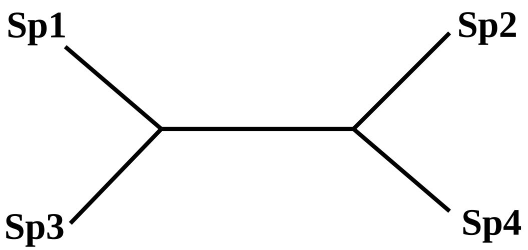


Maximum **L**ikelihood (最大似然法)

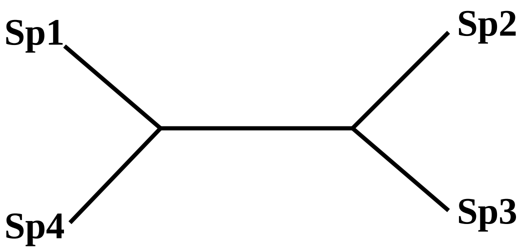
```
Sp1:  TCTGT...AACTCTTT...GAATCGTT.
Sp2:  TCTGC...GACTCGCT...GGAACGCT.
Sp3:  CTTAT...GATCTATT...GGAATATT.
Sp4:  CCTAT...GATCCATT...GGACCATT.
```



***ln*L= - 2000**

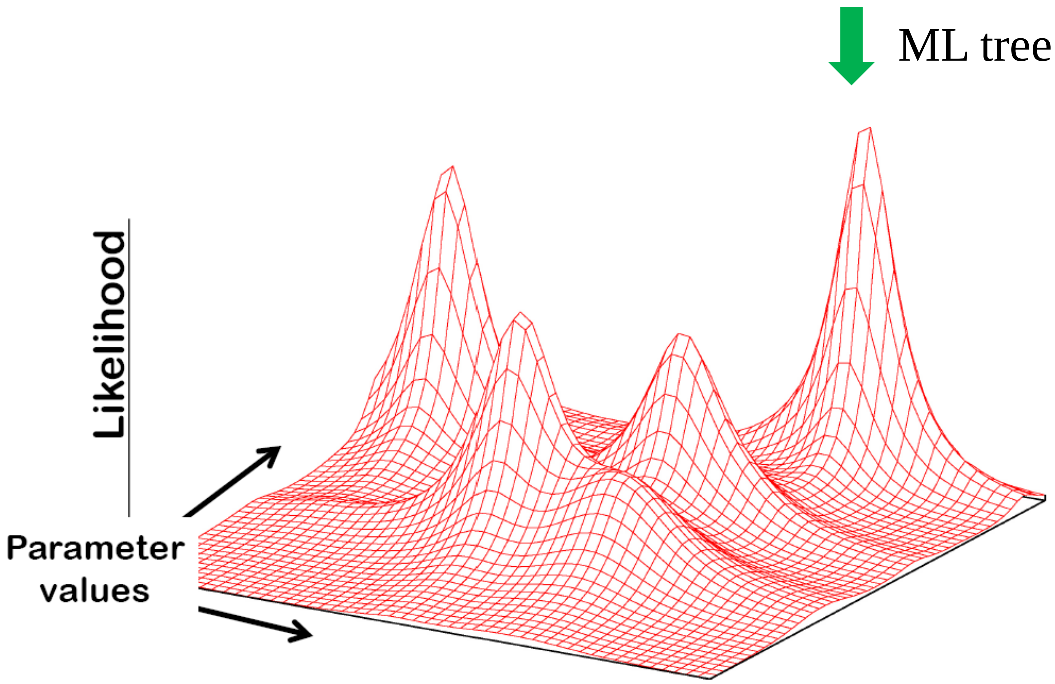


***ln*L= - 1500**



***ln*L= - 1000**

Maximum **L**ikelihood (**ML**)

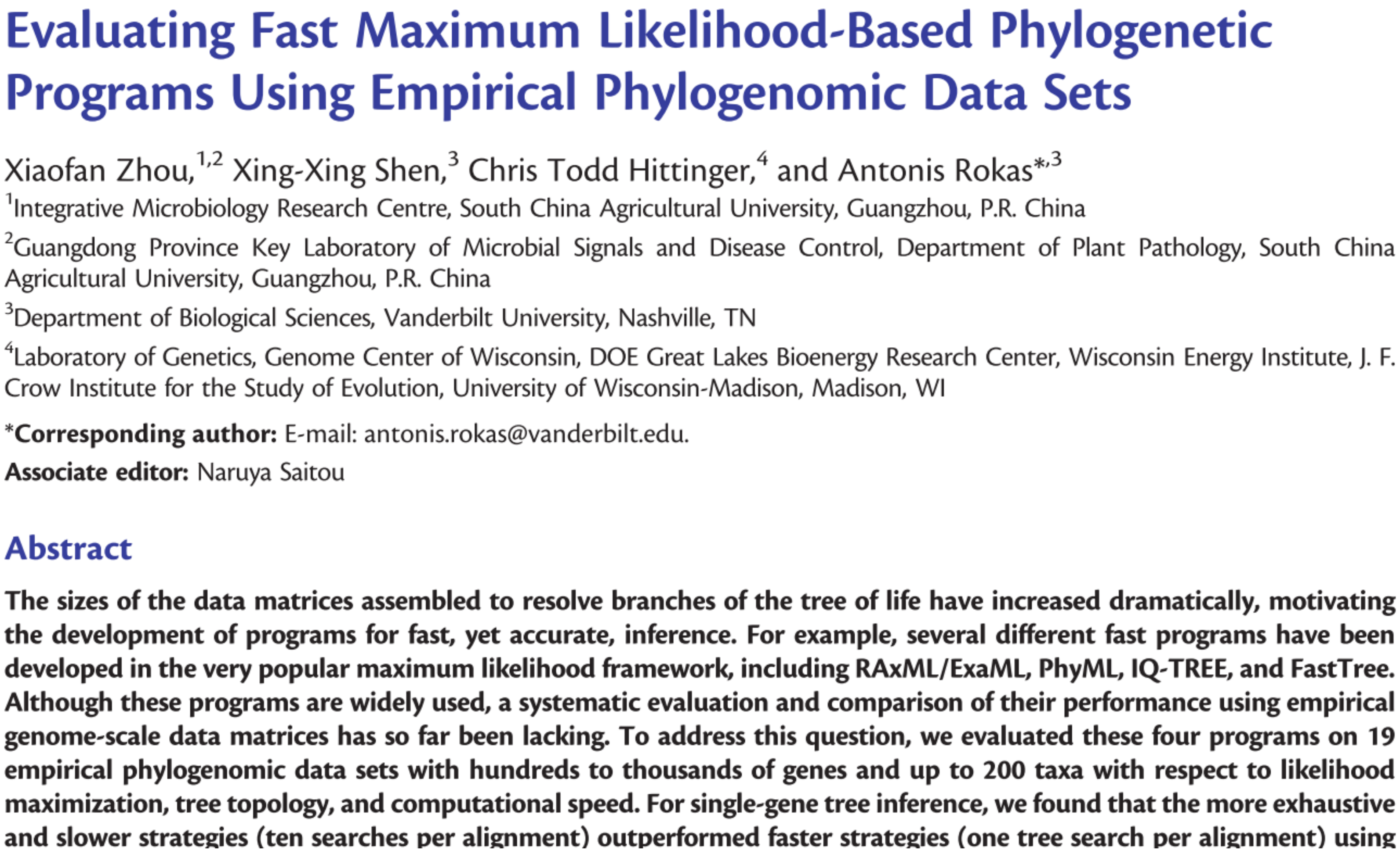


ML tree

$$\mathbf{ln}\boldsymbol{L}\mathbf{=} \sum_{\mathbf{i\,=\,1}}^{\boldsymbol{N}} \mathbf{log(L_i|Tree)}$$

(Slide from Derrick Zwickl)

Maximum Likelihood (最大似然法)



Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets

Xiaofan Zhou,[1,2] Xing-Xing Shen,[3] Chris Todd Hittinger,[4] and Antonis Rokas*,[3]

[1]Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou, P.R. China

[2]Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Department of Plant Pathology, South China Agricultural University, Guangzhou, P.R. China

[3]Department of Biological Sciences, Vanderbilt University, Nashville, TN

[4]Laboratory of Genetics, Genome Center of Wisconsin, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI

*__Corresponding author:__ E-mail: antonis.rokas@vanderbilt.edu.

Associate editor: Naruya Saitou

Abstract

The sizes of the data matrices assembled to resolve branches of the tree of life have increased dramatically, motivating the development of programs for fast, yet accurate, inference. For example, several different fast programs have been developed in the very popular maximum likelihood framework, including RAxML/ExaML, PhyML, IQ-TREE, and FastTree. Although these programs are widely used, a systematic evaluation and comparison of their performance using empirical genome-scale data matrices has so far been lacking. To address this question, we evaluated these four programs on 19 empirical phylogenomic data sets with hundreds to thousands of genes and up to 200 taxa with respect to likelihood maximization, tree topology, and computational speed. For single-gene tree inference, we found that the more exhaustive and slower strategies (ten searches per alignment) outperformed faster strategies (one tree search per alignment) using

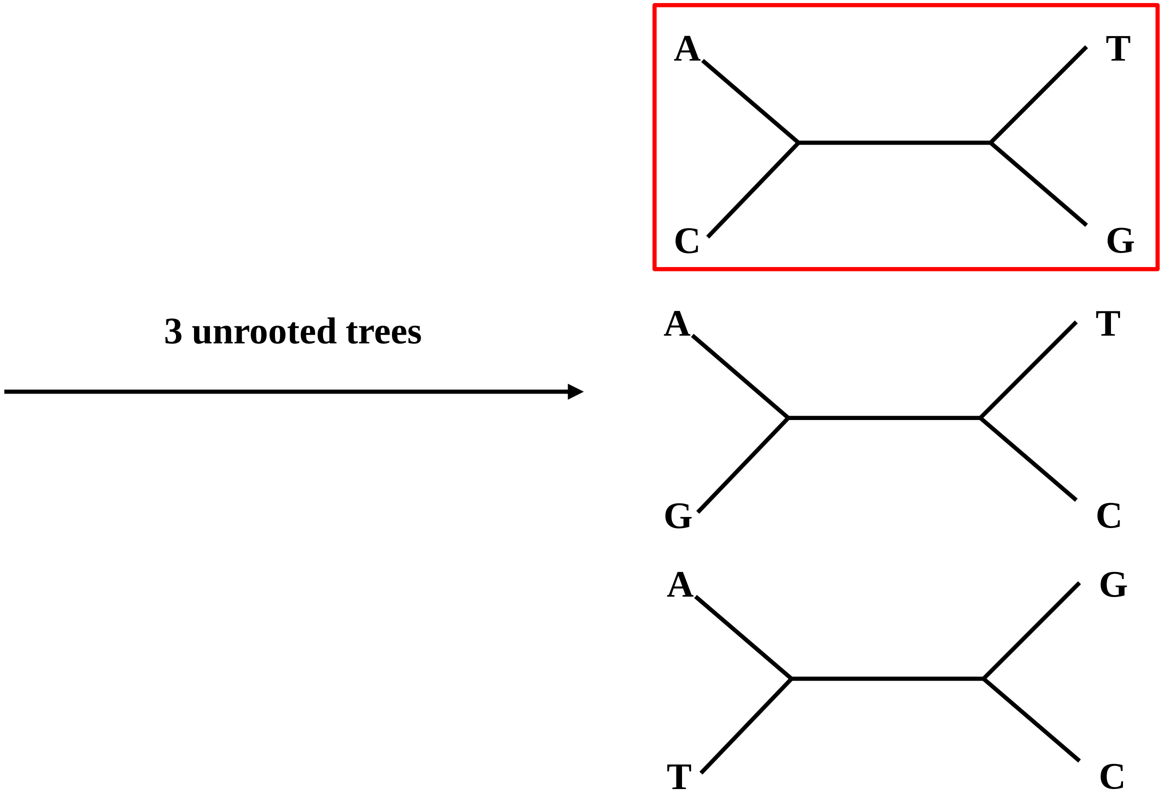
Maximum Likelihood (最大似然法)

Sp1=A

Sp2=T

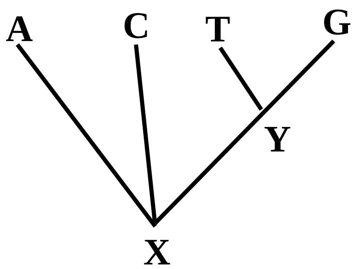
Sp3=C

Sp4=G

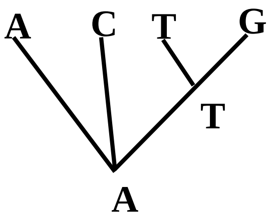


3 unrooted trees

Likelihood (似然值)



(1)

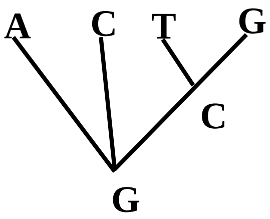


$$L_{(1)} = \pi_A * P_{AA} * P_{AC} * P_{AT} * P_{TT} * P_{TG}$$

…

…

(16)



$$L_{(16)} = \pi_G * P_{GA} * P_{GC} * P_{GC} * P_{CT} * P_{CG}$$

$$L(site1|tree) = L_{(1)} + L_{(2)} + \ldots\ldots + L_{16}$$

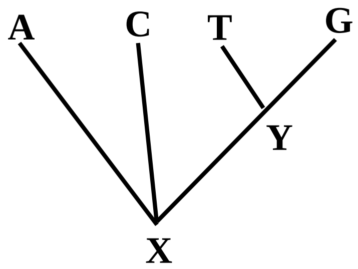
Likelihood (似然值)

Sp1=A C …

Sp2=T T …

Sp3=C C …

Sp4=C C …

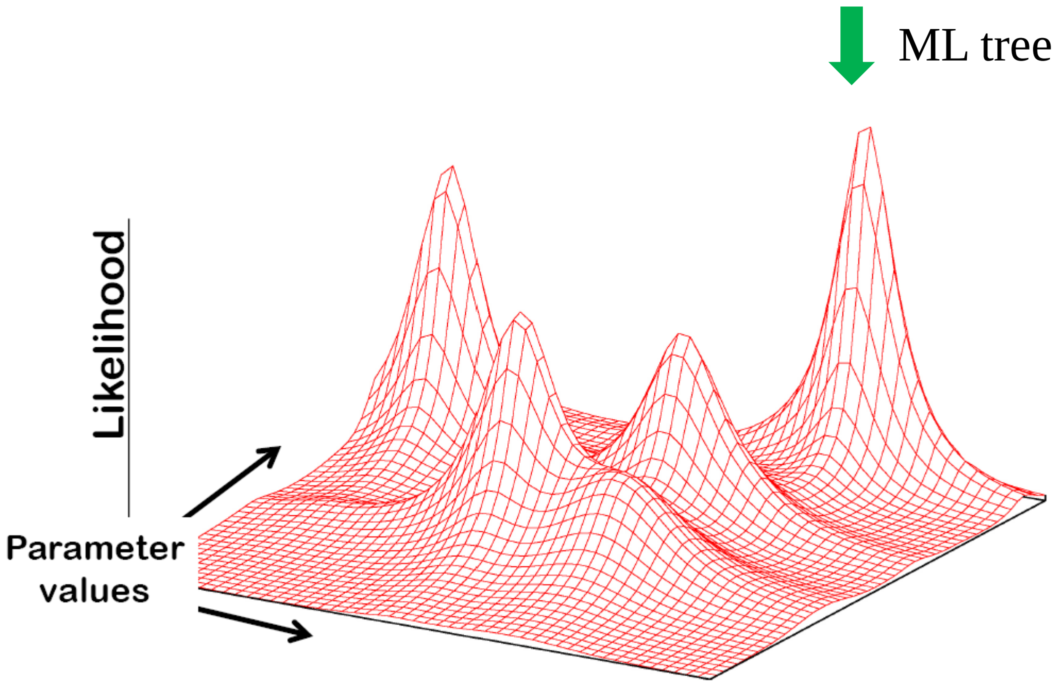


L(Data|tree) = L(site1|tree) * L(site2|tree) …* L(siten|tree)

Log L(Data|tree) = Log L(site1|tree) + Log L(site2|tree) ..+Log L(siten|tree)

$$\ln L = \sum_{i=1}^{N} \mathrm{LnL}_i$$

Maximum **L**ikelihood (**ML**)



ML tree

$$\mathbf{ln}\boldsymbol{L} = \sum_{\mathbf{i}=\mathbf{1}}^{N} \mathbf{log}(\mathbf{L_i}|\mathbf{Tree})$$

(Slide from Derrick Zwickl)

Maximum **L**ikelihood (**ML**):gap/missing/ambiguous characters

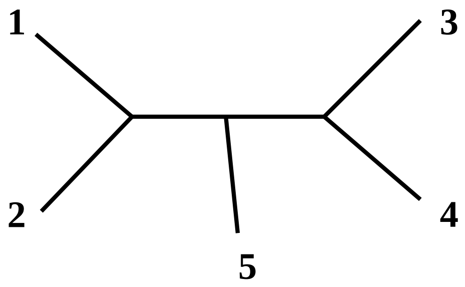
Sp1= A

Sp2= G

Sp3= G

Sp4= C

Sp5= G



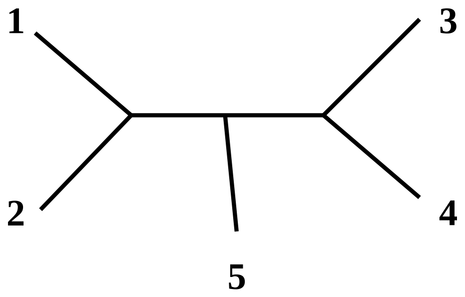
Sp1= A **A**

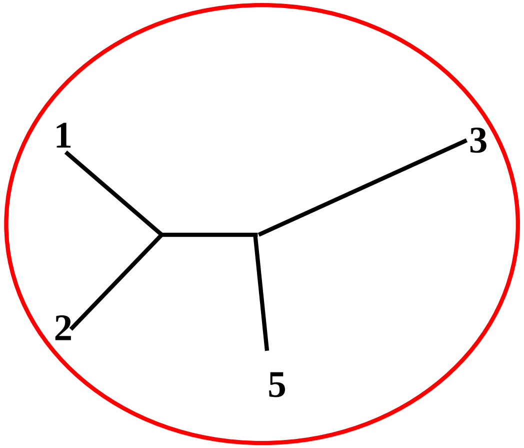
Sp2= G **T**

Sp3= G **C**

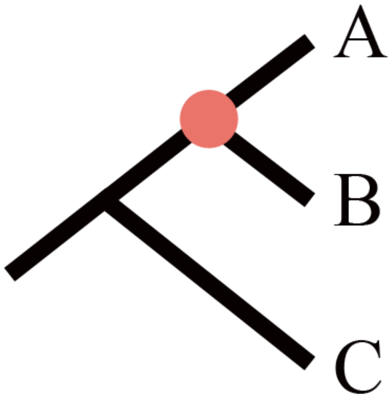
Sp4= C **-**

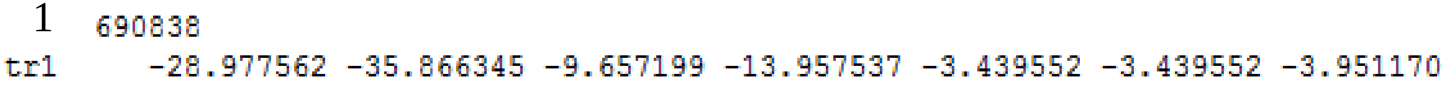
Sp5= G **C**





Site-wise log likelihood support (SLS)

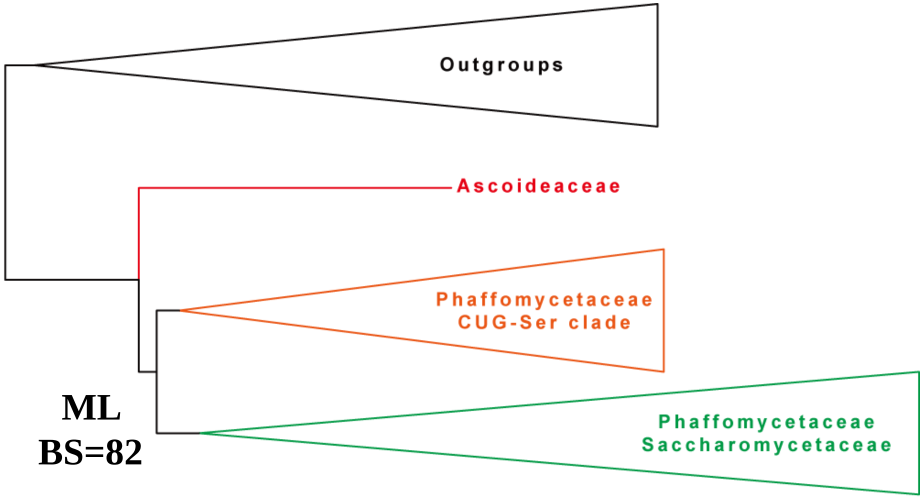




```
 1   690838
tr1      -28.977562 -35.866345 -9.657199 -13.957537 -3.439552 -3.439552 -3.951170  ......
```

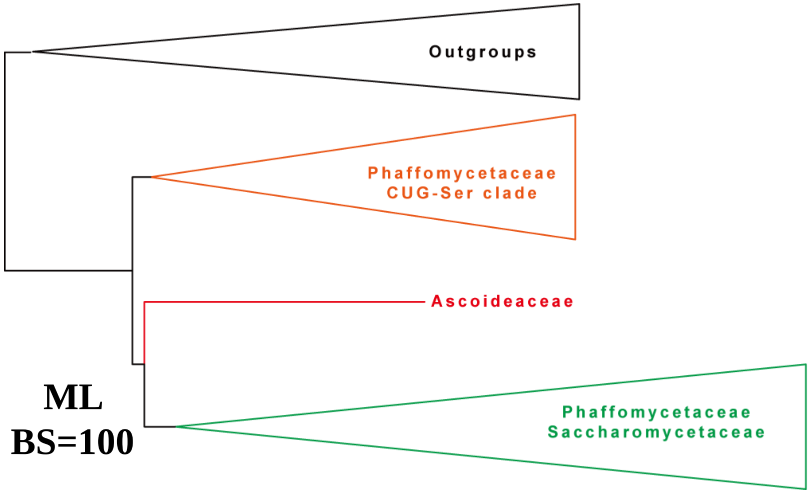
$$\ln L = \sum_{i=1}^{N} \mathrm{SLS}_i$$

Incongruence: Ascoideaceae (酱霉科)



1,559-gene and 38-taxon

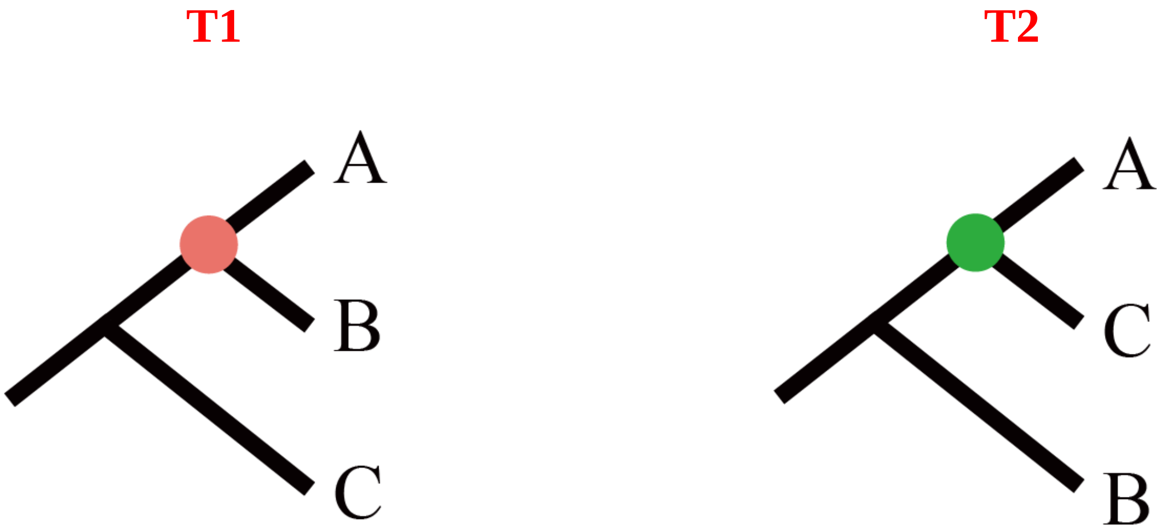
Riley et al. PNAS 2016

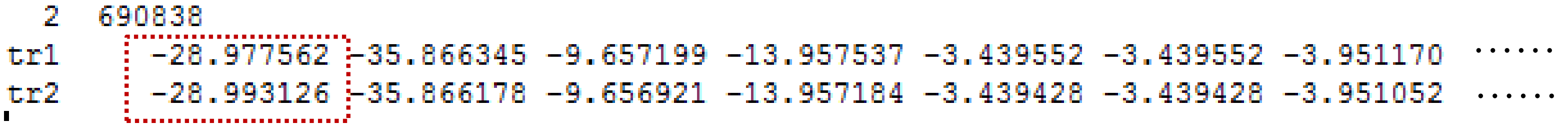


1,233-gene and 86 yeasts

Shen et al. G3 2016

Site-wise log likelihood support (SLS)

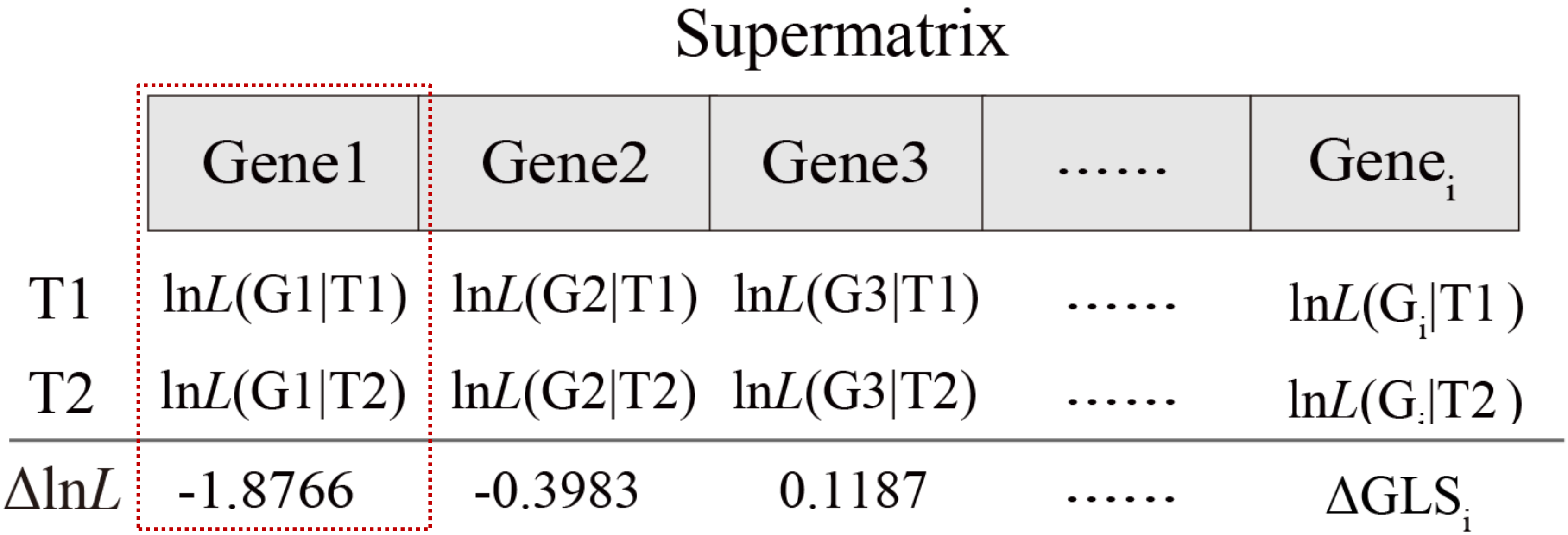




```
  2   690838
tr1     -28.977562 -35.866345 -9.657199 -13.957537 -3.439552 -3.439552 -3.951170  ······
tr2     -28.993126 -35.866178 -9.656921 -13.957184 -3.439428 -3.439428 -3.951052  ······
```

$$\triangle \mathbf{SLS_1=log(P_1|T1) - log(P_1|T2)=0.01556}$$

Gene-wise log likelihood support (GLS)

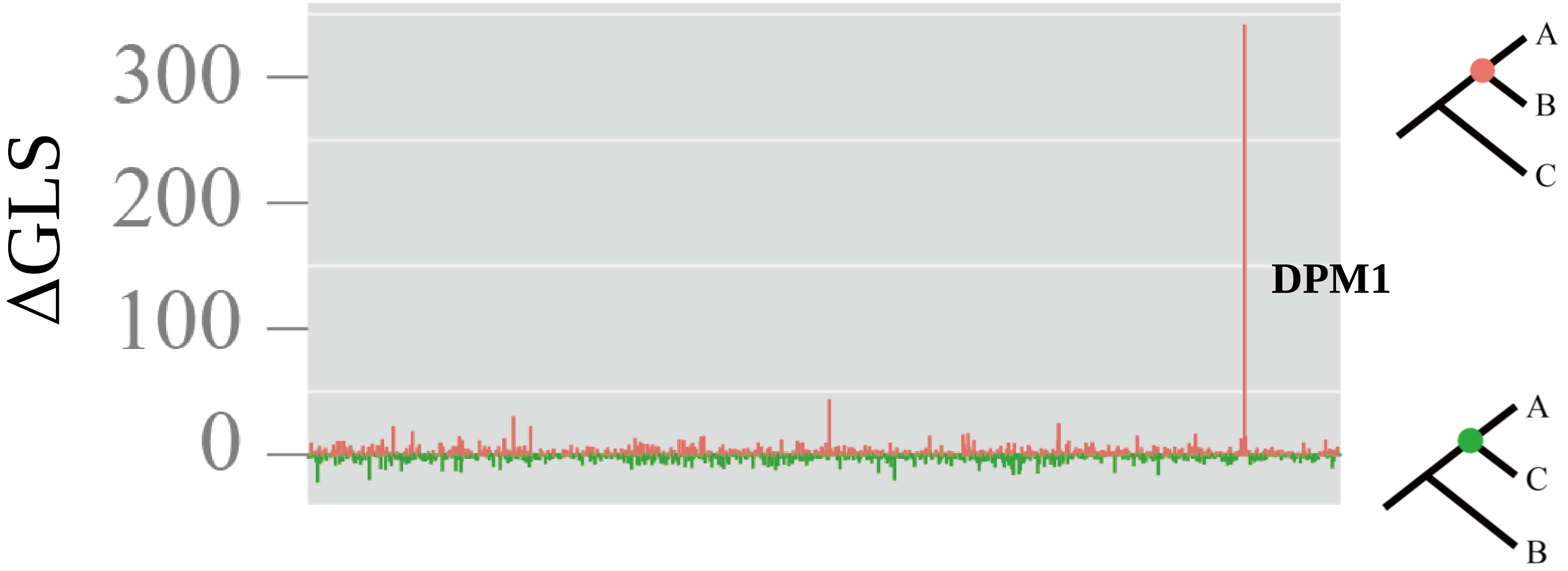


Supermatrix

	Gene1	Gene2	Gene3	……	Gene_i
T1	$\ln L(\text{G1}\|\text{T1})$	$\ln L(\text{G2}\|\text{T1})$	$\ln L(\text{G3}\|\text{T1})$	……	$\ln L(\text{G}_i\|\text{T1})$
T2	$\ln L(\text{G1}\|\text{T2})$	$\ln L(\text{G2}\|\text{T2})$	$\ln L(\text{G3}\|\text{T2})$	……	$\ln L(\text{G}_i\|\text{T2})$
$\Delta\ln L$	-1.8766	-0.3983	0.1187	……	ΔGLS_i

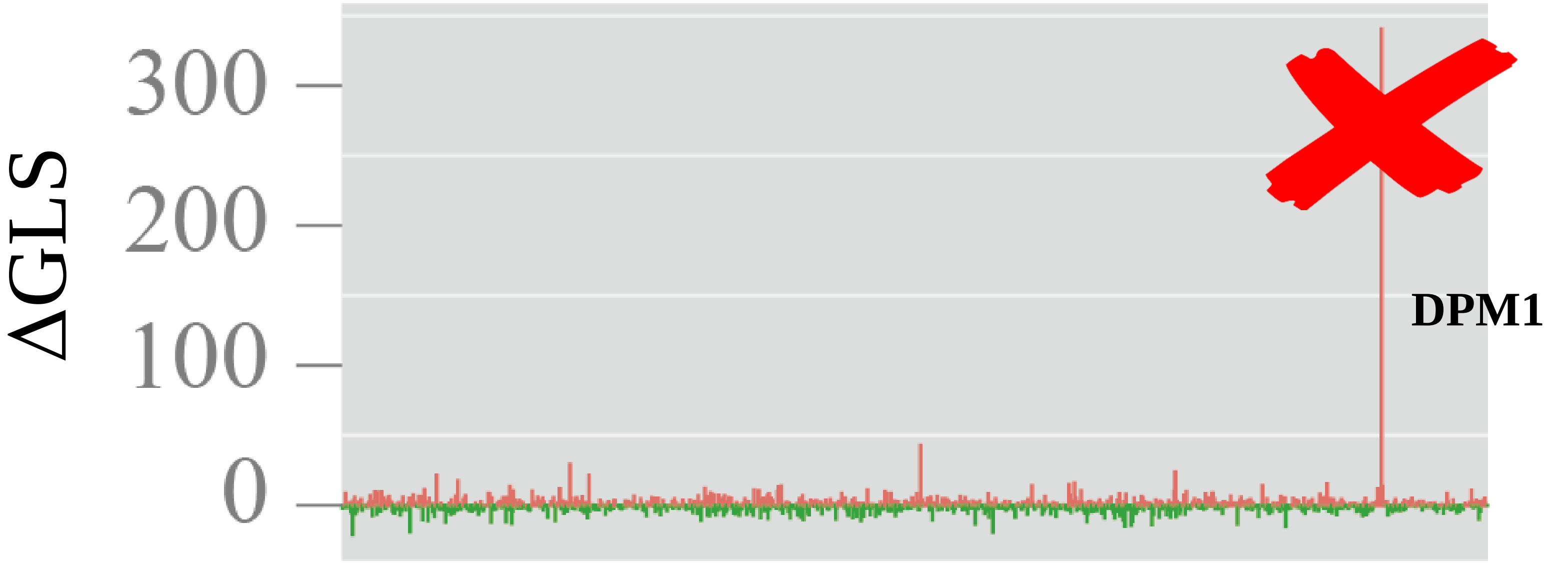
$$\triangle\mathbf{GLS_1=ln}\boldsymbol{L}\mathbf{(G_1|T1)\ -\ ln}\boldsymbol{L}\mathbf{(G_1|T2)=\ -1.8766}$$

ΔGLS plot

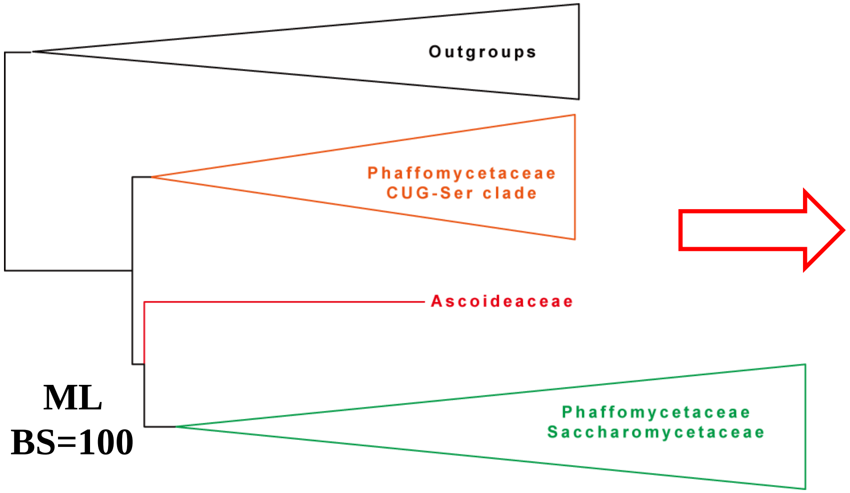


A single gene displays very strong difference in gene-wise log likelihood scores for T1 against T2.

Removal

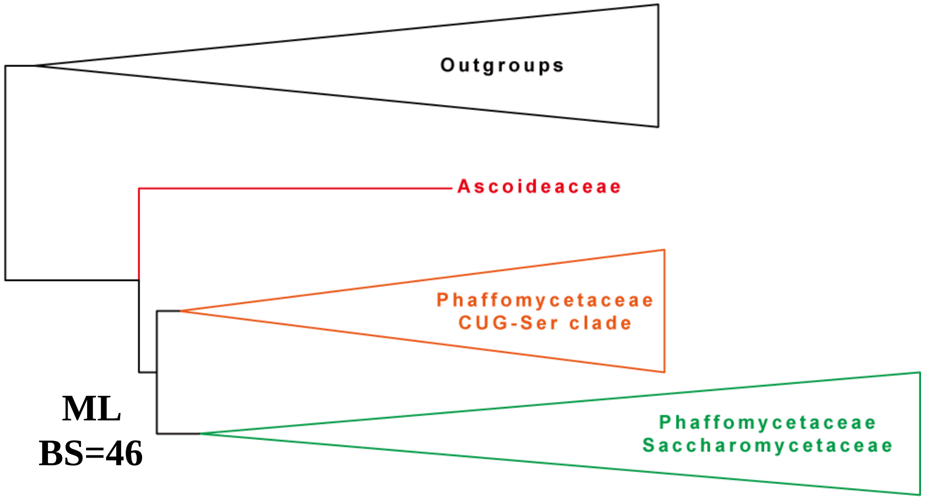


Removal of the strongest gene



1,233-gene (T1)

Shen et al. G3 2016



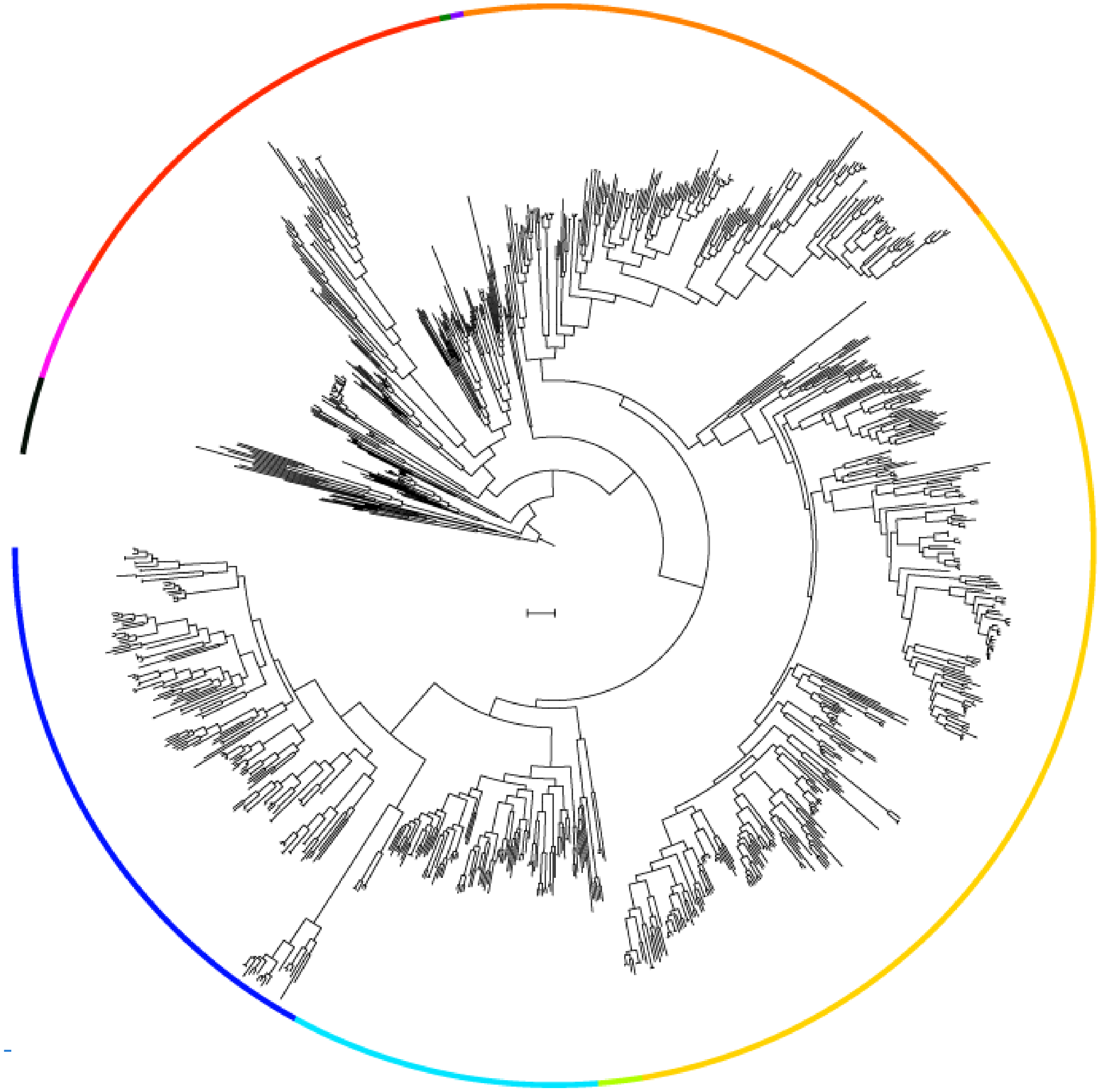
1,232-gene (T2)

Shen et al. G3 2016

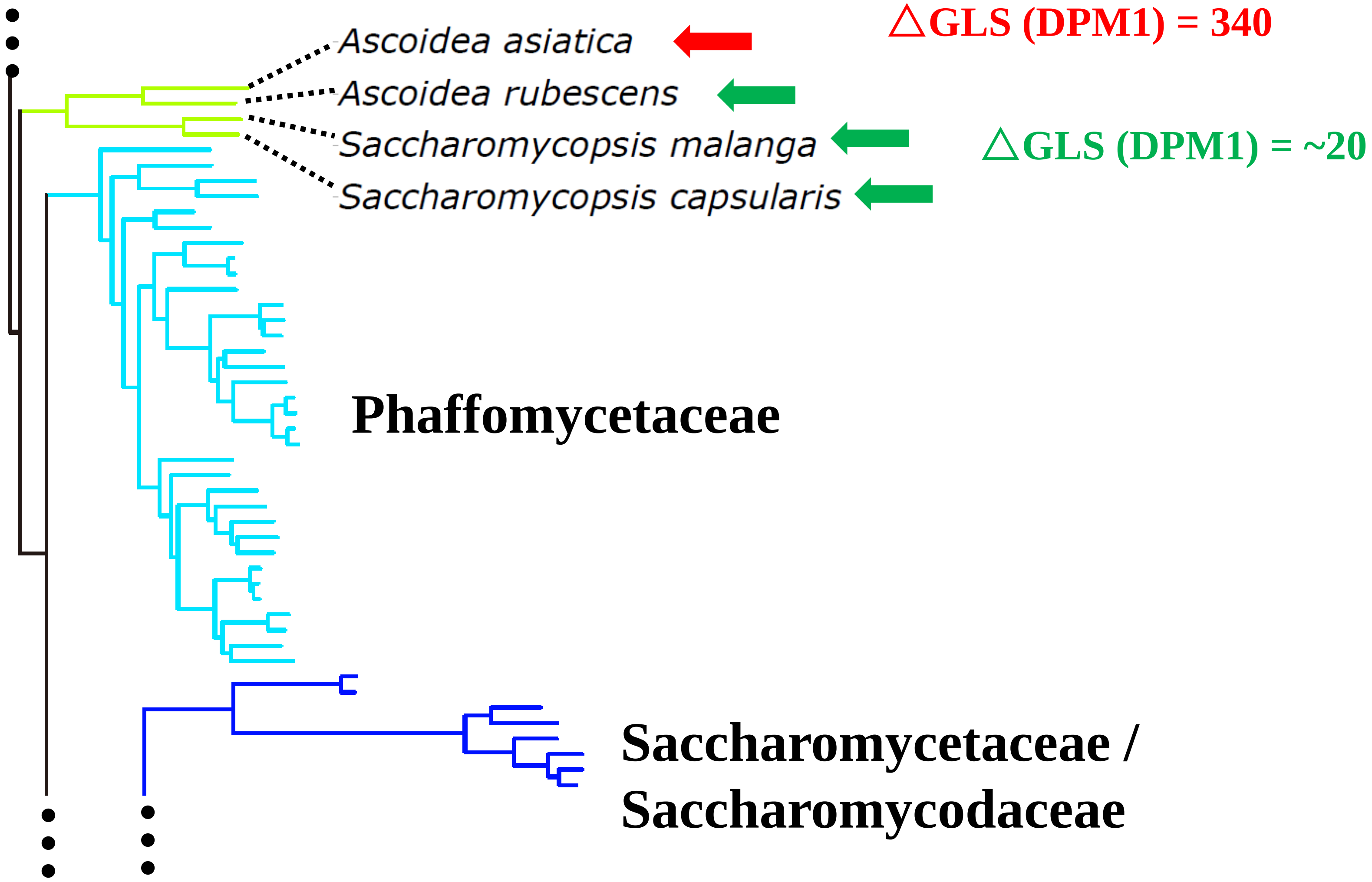
Removal of the strongest gene witched the ML tree’s support from T1 to T2

Contentious relationships in phylogenomic studies can be driven by a handful of genes

Improved sampling



Sampling of 3 Additional Taxa “Breaks” the Long Branch



Ascoidea asiatica

Ascoidea rubescens

Saccharomycopsis malanga

Saccharomycopsis capsularis

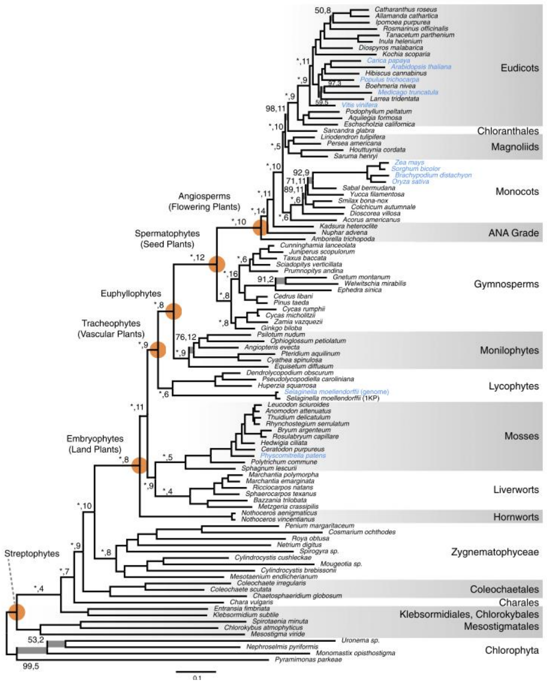
△GLS (DPM1) = 340

△GLS (DPM1) = ~20

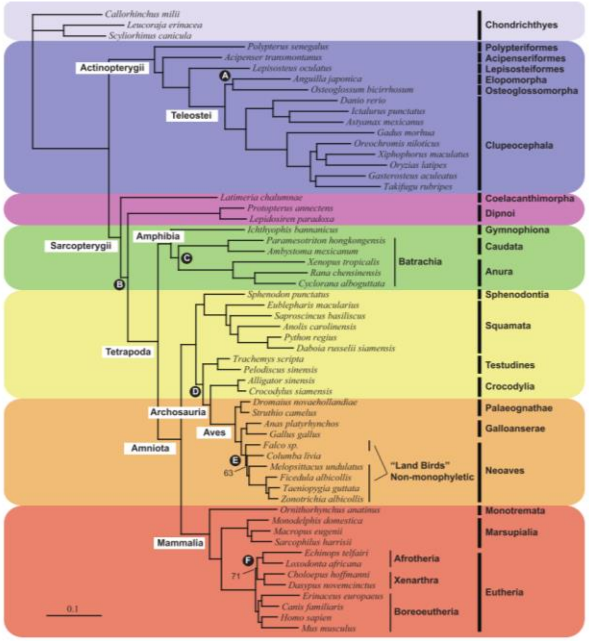
Phaffomycetaceae

Saccharomycetaceae / Saccharomycodaceae

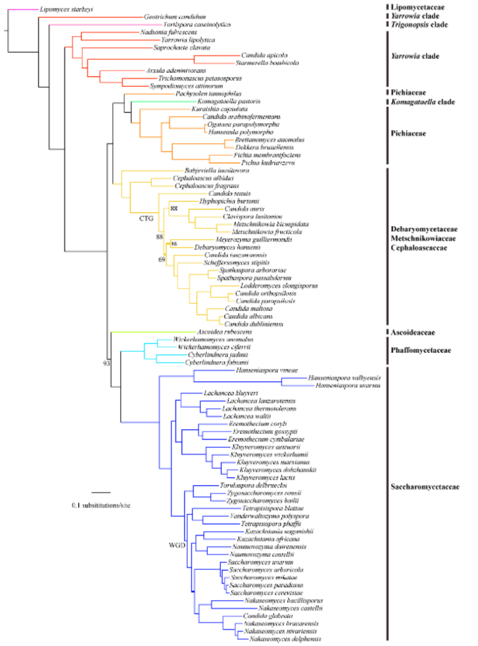
Three large phylogenomic data matrices



Plant:103 taxa and 674 gens
Wickett et al. PANS 2014

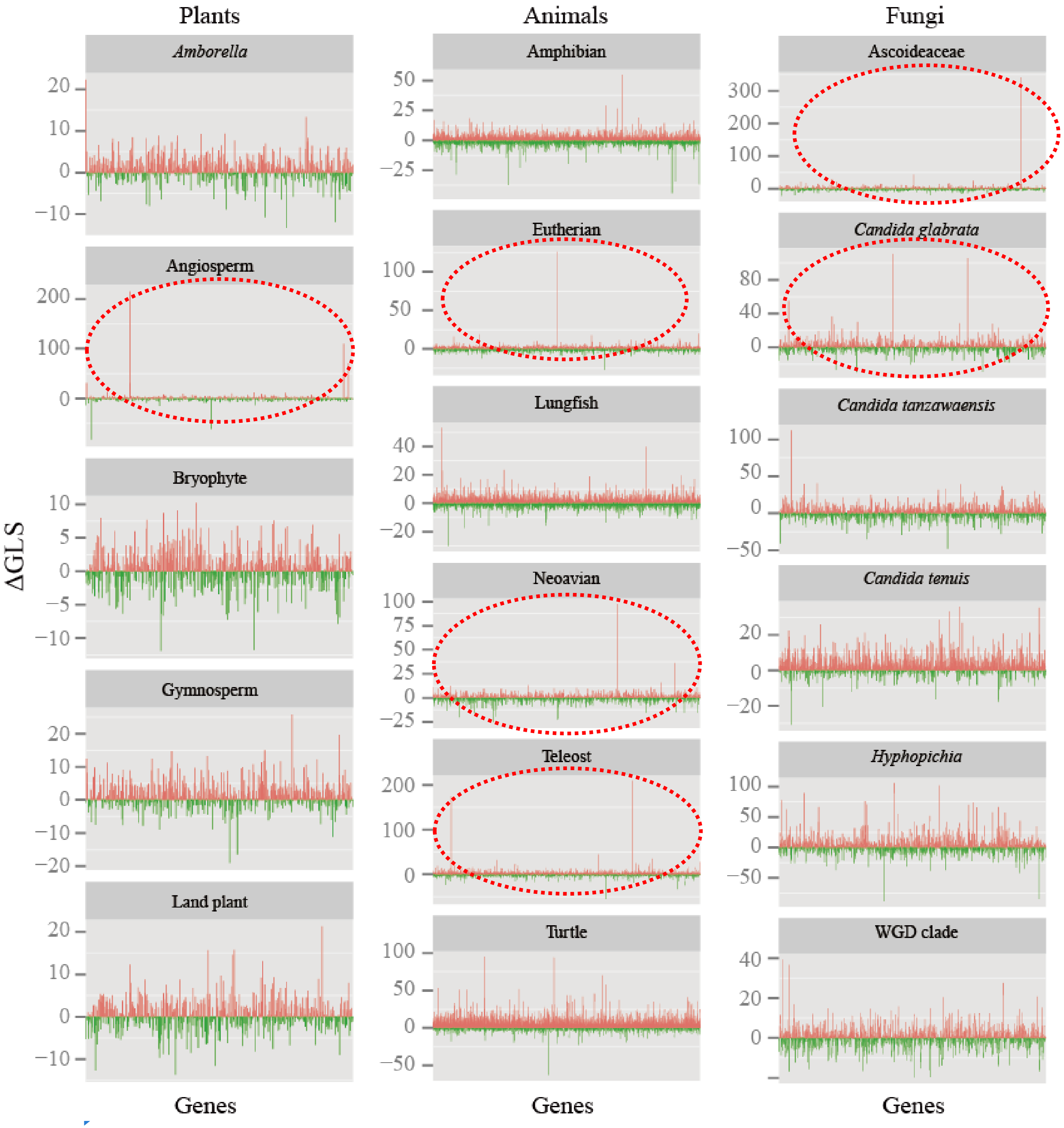


Animal: 58 taxa and 4682 genes
Chen et al. Syst Biol 2015



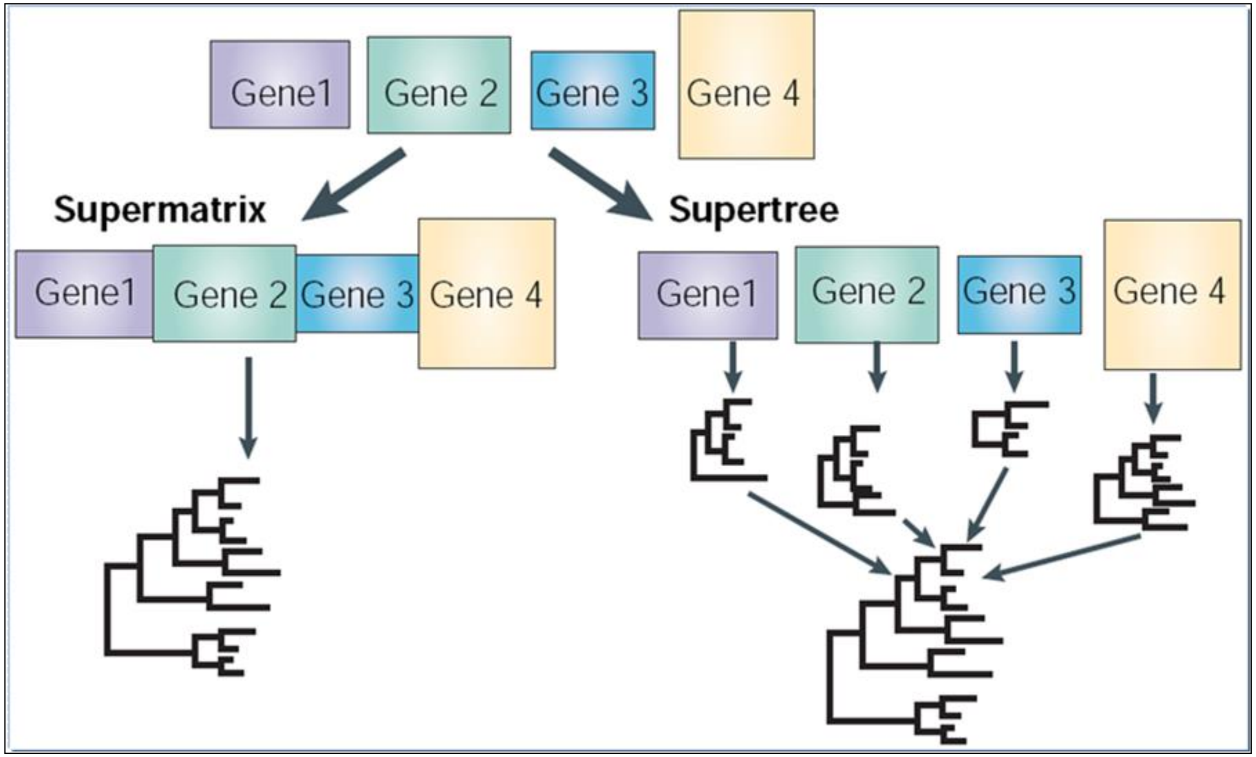
Fungi: 86 yeasts and 1233 genes
Shen et al. G3 2016

ΔGLS plot (17 branches)



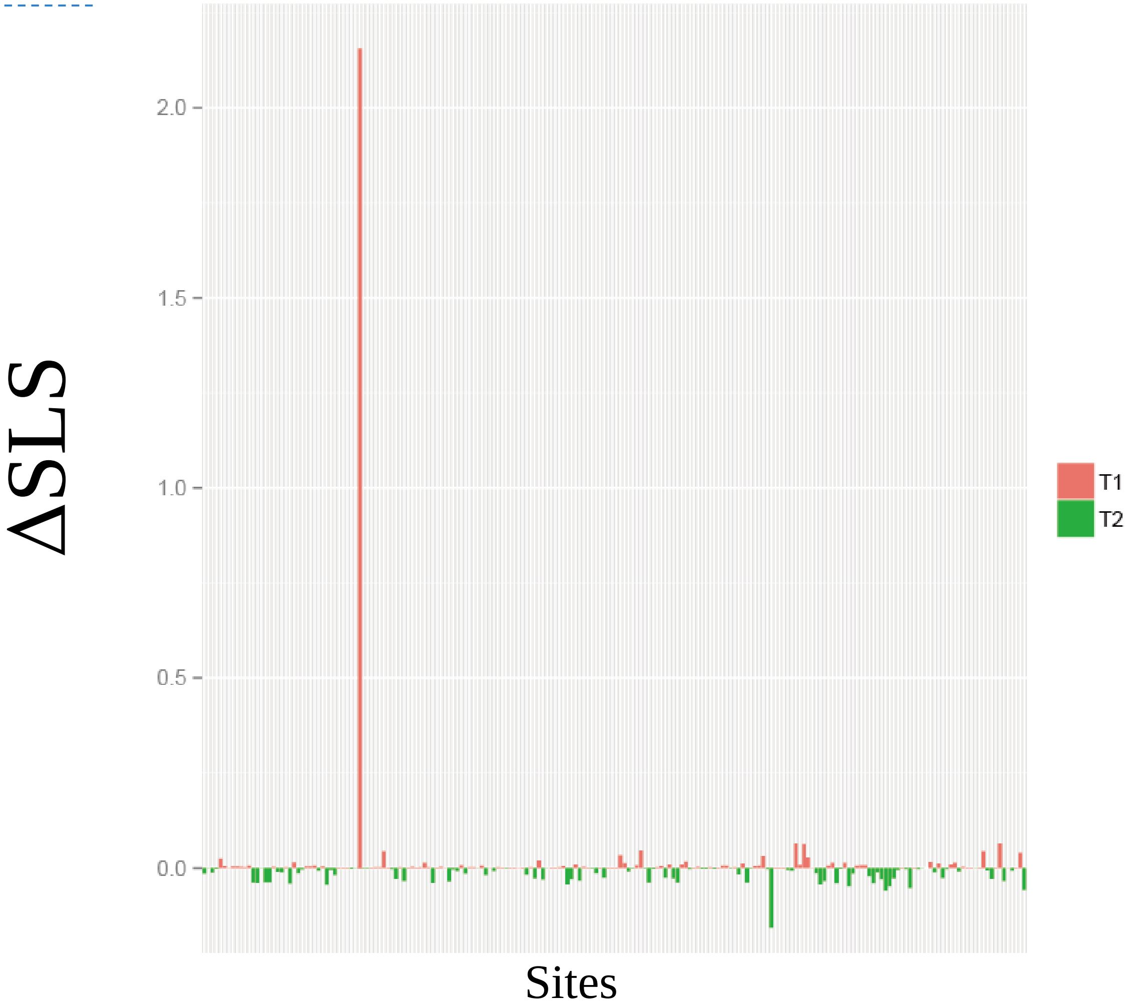
6 branches show a single or a handful of genes display very strong ΔGLS

Phylogenomics



Delsuc et al. 2005

Difference in Site-wise log likelihood support (Δ SLS) within individual gene



Conclusions

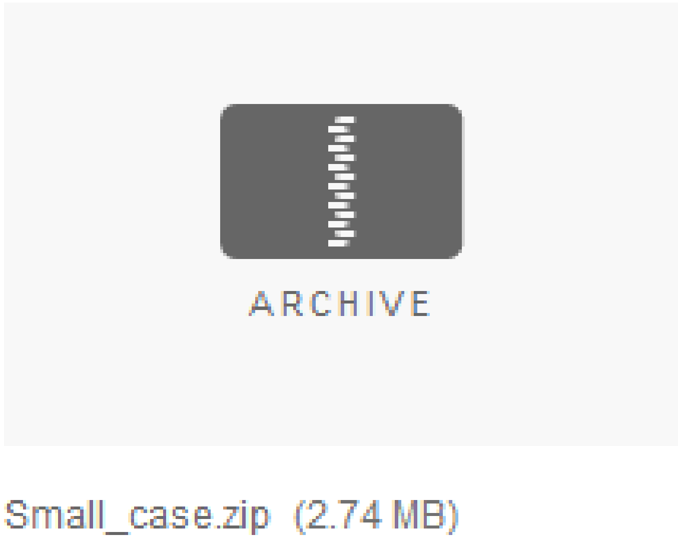
- A tiny amount of data in very large phylogenomic data matrix can drive the resolution of specific internodes

  Why it happens? Wrong sequence, paralogs, HGTs

- Distribution of phylogenetic signal for each of the main alternative hypotheses

  Majority of genes support your results?

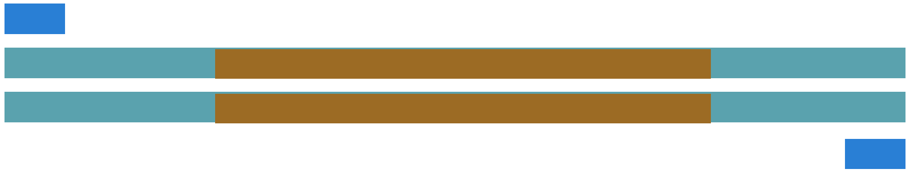
Small case on the Figshare



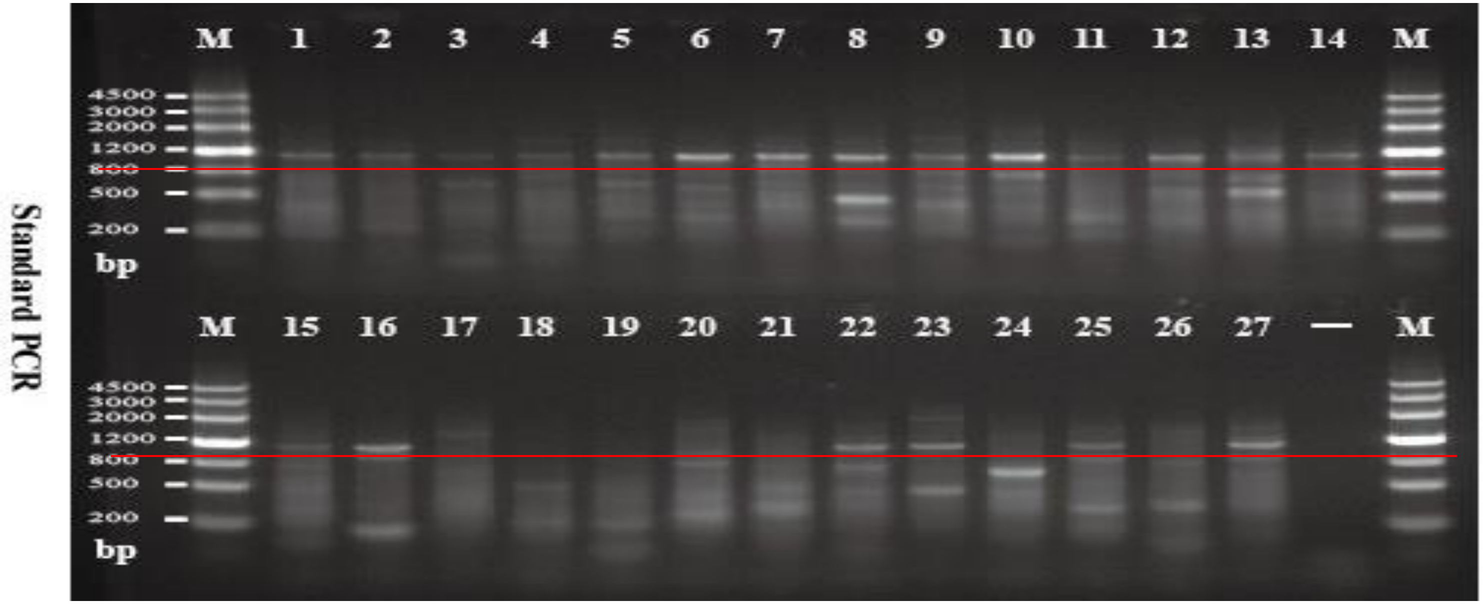
Small_case.zip (2.74 MB)

常规PCR

🟦 简并引物



标准PCR结果


Standard PCR
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 M
4500
3000
2000
1200
800
500
200
bp
M 15 16 17 18 19 20 21 22 23 24 25 26 27 — M
4500
3000
2000
1200
800
500
200
bp

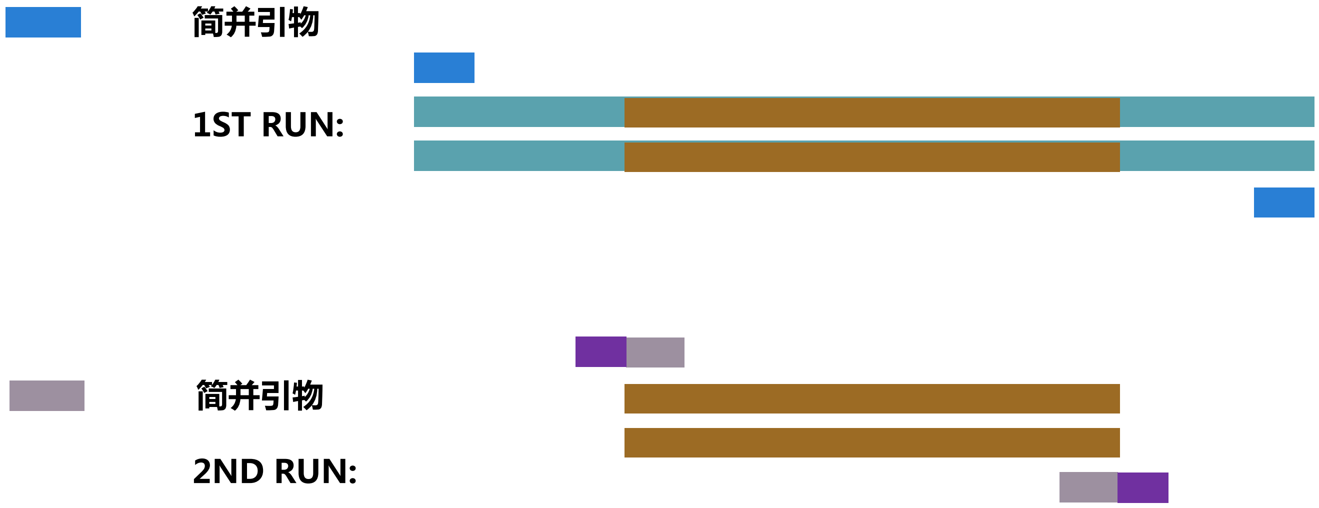
测定非特异性扩增PCR产物需要许多额外工作

PCR product → Gel cutting → Cloning → Cleanup → Sequencing

Shen et al. 2011 (Mol Biol Evol)

PCR

高效的巢式PCR



简并引物

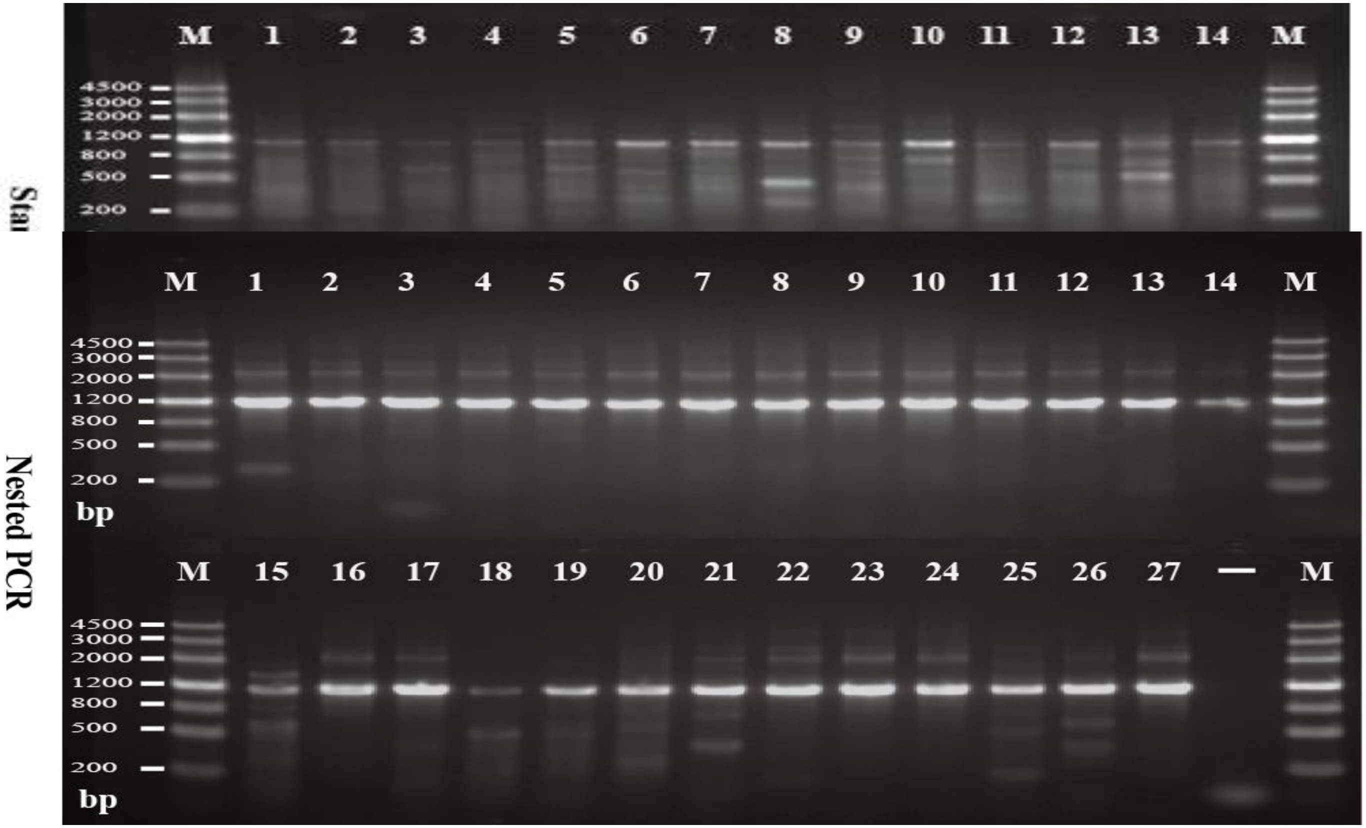
1ST RUN:

简并引物

2ND RUN:

Shen et al. 2011 (PLoS One)

比较标准PCR和巢式PCR扩增效果


Stan
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 M
4500
3000
2000
1200
800
500
200
Nested PCR
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 M
4500
3000
2000
1200
800
500
200
bp
M 15 16 17 18 19 20 21 22 23 24 25 26 27 — M
4500
3000
2000
1200
800
500
200
bp

巢式PCR显著提高NPCL扩增成功率和易产生单一且亮的目标条带。

Shen et al. 2011 (PLoS One)

两栖动物：有尾目

Hynobiidae
（3）

Cryptobranchidae
（1）

Sirenidae
（2）

Ambystomatidae
（1）

Dicamptodontidae
（1）

Salamandridae
（3）

Proteidae
（2）

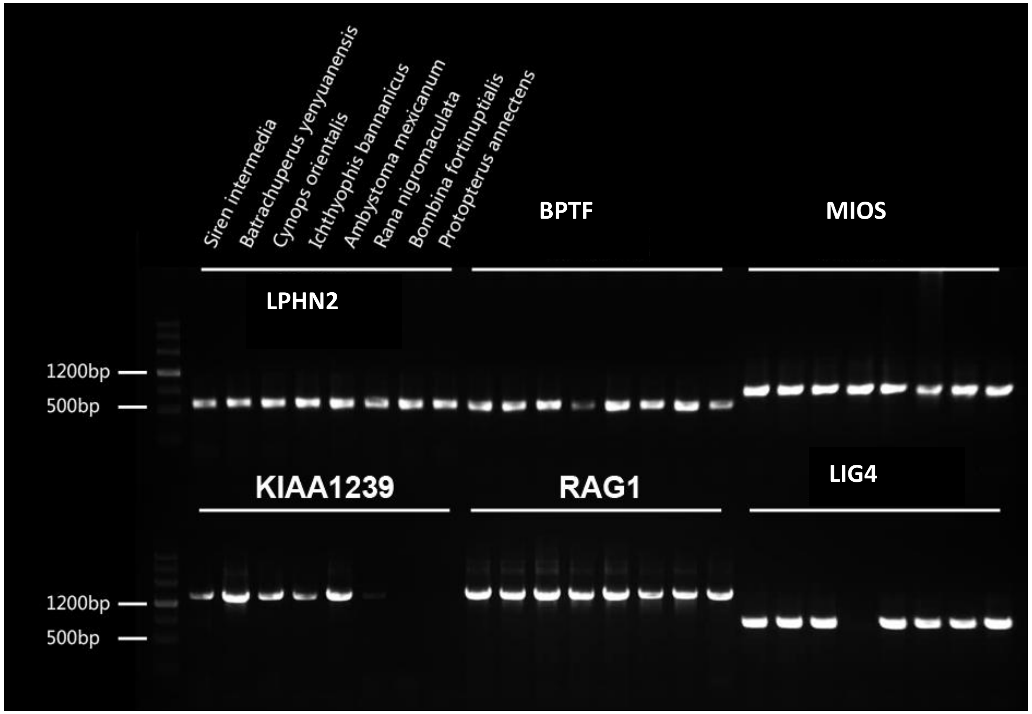
Rhyacotritonidae
（1）

Plethodontidae
（4）

Amphiumidae
（1）

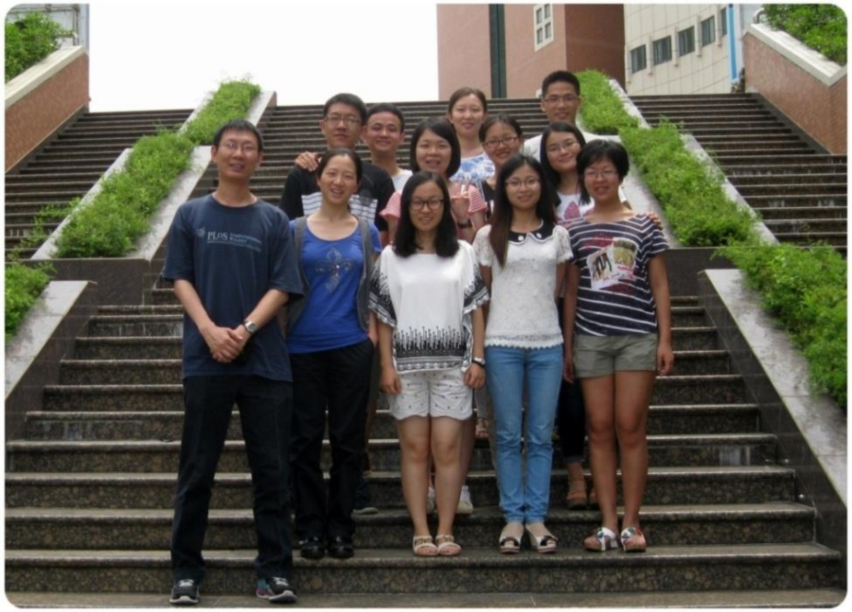
Shen et al. 2013 (MBE)

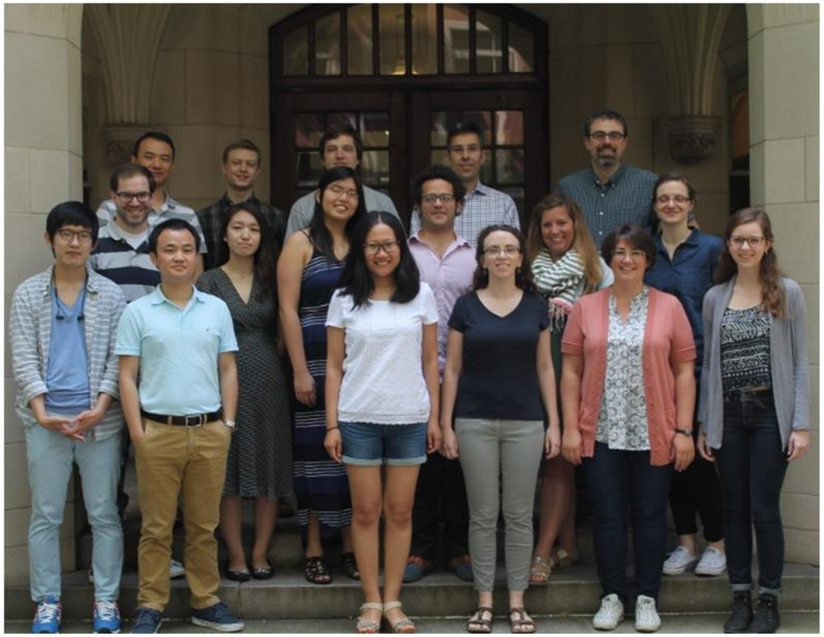
分子标记PCR结果



Shen et al. 2013 (Mol Biol Evol)

Thank you!





https://xingxingshen.github.io/