Mixture models of nucleotide sequence evolution, and the evolution of yeast genomes

Lars Jermiin | OCE Science Leader

8 November 2013

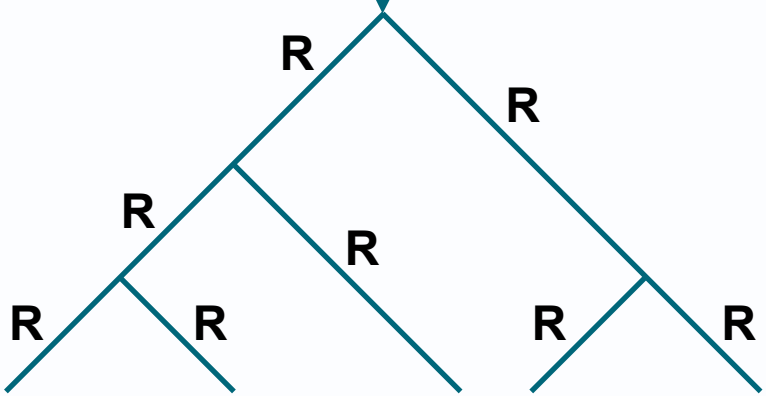
CSIRO ECOSYSTEM SCIENCES
www.csiro.au

"Common sense is actually rather uncommon…" — Christy McGeough (2006)

Common scenario

Multiple sequence alignment (**MSA**)

Phylogenetic method

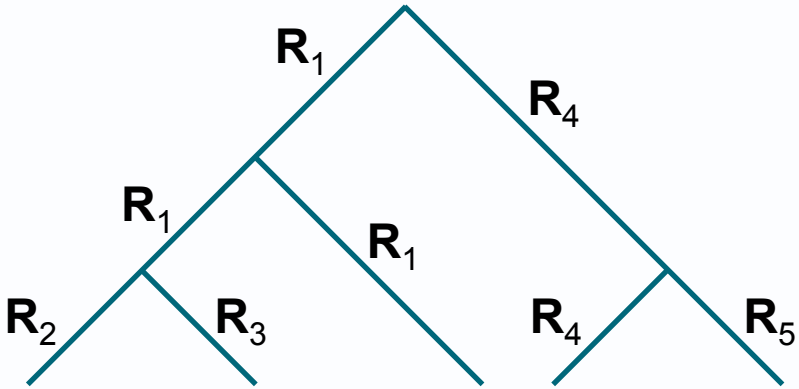


Common phylogenetic assumptions

- Evolutionary history **tree-like**
- Sites have evolved under **IID** conditions
- **Evolutionary process** can be modelled by a **time-reversible** Markov model, **R**

Reality check

- **Compositional heterogeneity** (CH) across the sequences is common
- CH across sequences implies that a more **complex model of evolution** is necessary



Modelling evolutionary processes

Nucleotides

$$\mathbf{R} = \begin{bmatrix} - & s_1 & s_2 & s_3 \\ s_1 & - & s_4 & s_5 \\ s_2 & s_4 & - & s_6 \\ s_3 & s_5 & s_6 & - \end{bmatrix} \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_G & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix}$$

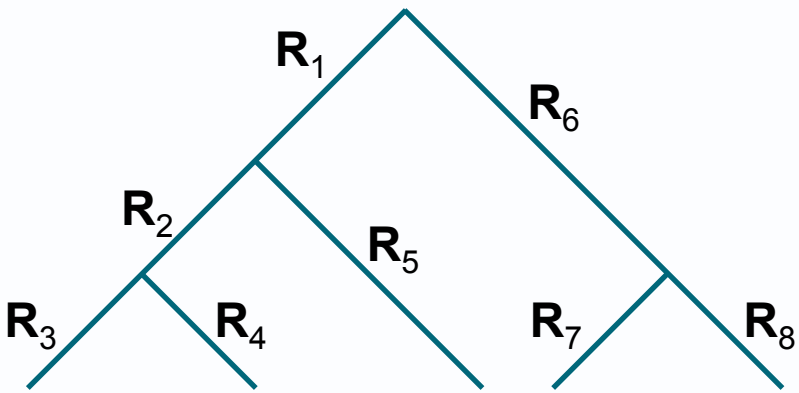
$$\mathbf{R} = \mathbf{S}\Pi$$

Amino acids

A similar formulation of **R** applies

Complex evolutionary models

- 15-20 papers since 1995 on complex models of evolution
- Several of these models assign a **unique rate matrix** to each edge



Parameters

DNA JC GTR RBH SBH BH

Parameters

Problem — Potentially **too many parameters** (over-parameterisation)

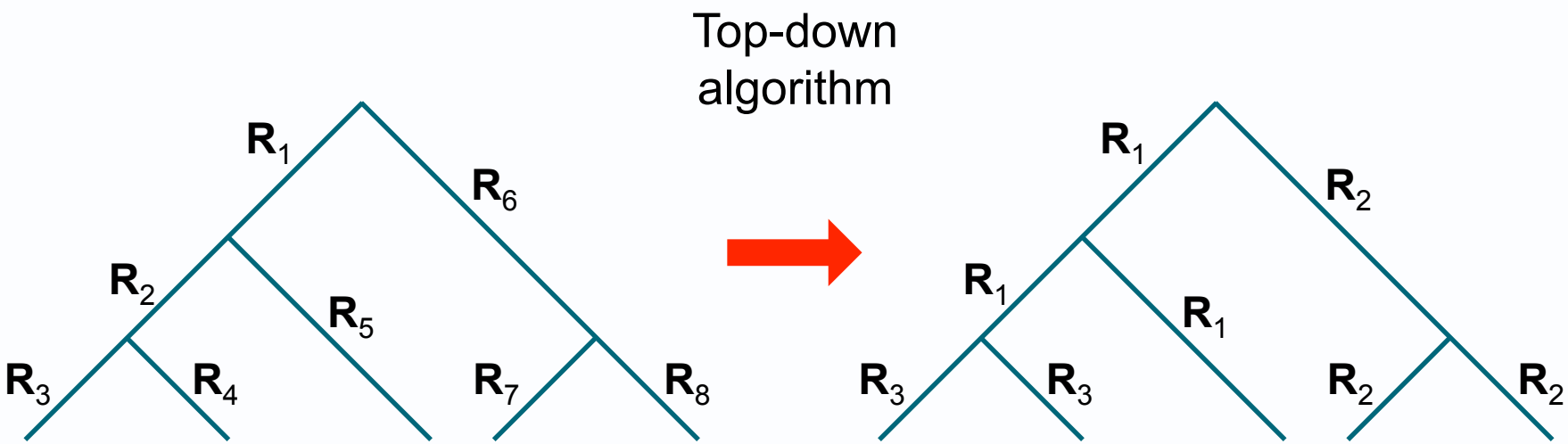
Source: Jayaswal, Robinson & Jermiin, **Syst. Biol.** 56, 155-162 [2007]; Jayaswal, Jermiin, Poladian & Robinson, **Syst. Biol.** 60, 74-86 [2011] .

Heterogeneity across lineages (HAL) • 1

Mol. Biol. Evol. 28(11):3045–3059. 2011

Reducing Model Complexity of the General Markov Model of Evolution

Vivek Jayaswal,[1,2] Faisal Ababneh,[3] Lars S Jermiin,[*,4,5,6] and John Robinson[1,2]



Problem — Parameters may be **non-identifiable** (**Syst. Biol**. 60, 872-875)

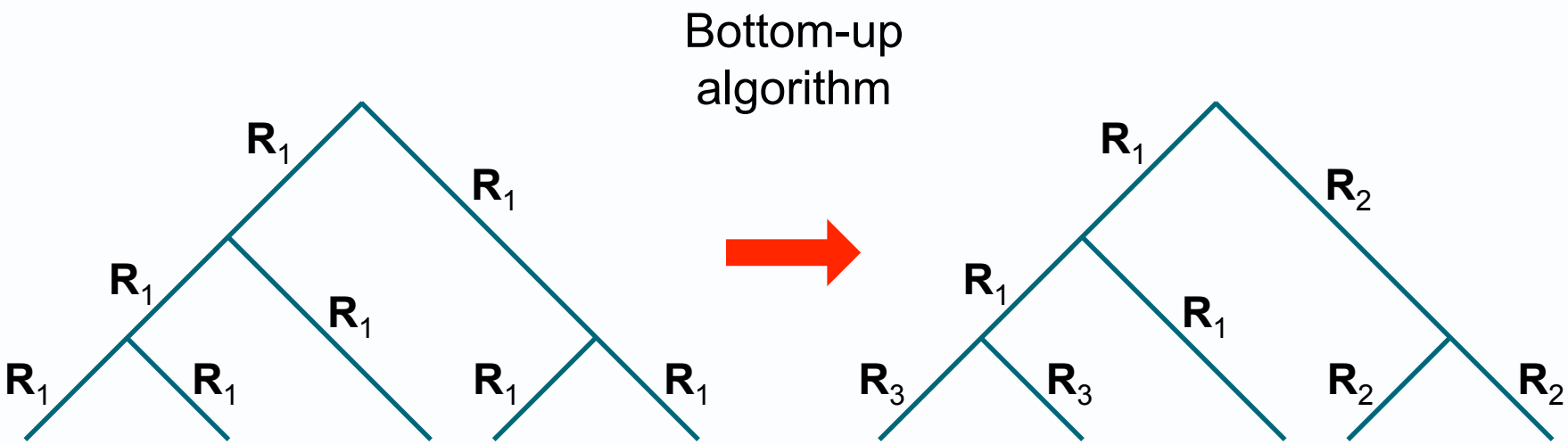
Source: Jayaswal, Ababneh, Jermiin & Robinson, **Mol. Biol. Evol.** 28, 3045-3059 [2011].

Heterogeneity across lineages (HAL) • 2

Syst. Biol. (accepted pending major revision)

Mixture Models of Nucleotide Sequence Evolution that account for Rate-heterogeneity Across Sites and Across Lineages

Vivek Jayaswal[1,2], Thomas KF Wong[3], John Robinson[2,4], Leon Poladian[2,4], Lars S. Jermiin[3]



Note — The Top-down algorithm may then be used to reduce complexity

Source: Jayaswal, Wong, Robinson, Poladian & Jermiin, **Syst. Biol.** [2014].

Heterogeneity across sites (HAS) • 1

Common models

$$I \quad (pI) \qquad \underbrace{\Gamma_k \quad (\alpha) \qquad I + \Gamma_k \quad (pI,\ \alpha)}$$

Probability a site belongs to the *i*-th rate category

$$a_1 = a_2 = a_i = \cdots a_k = \frac{1}{k}$$

User defined

Our model

Invariable sites	Variable sites

$$I \quad (pI) \qquad a_1 + a_2 + a_i + \cdots + a_k = 1$$

Inferred from data

Heterogeneity across sites (HAS) • 2

Sites belonging to **different rate categories** have…

Model	Common $\mathbf{f_0}$	Common **S**	Common **Π**	Scalar edge lengths
HAS_1	No	No	No	No
HAS_2	No	Yes	No	No
HAS_3	No	Yes	Yes	No
HAS_4	Yes	Yes	Yes	No
HAS_5	Yes	Yes	Yes	Yes

Note – 11 other HAS models not yet considered

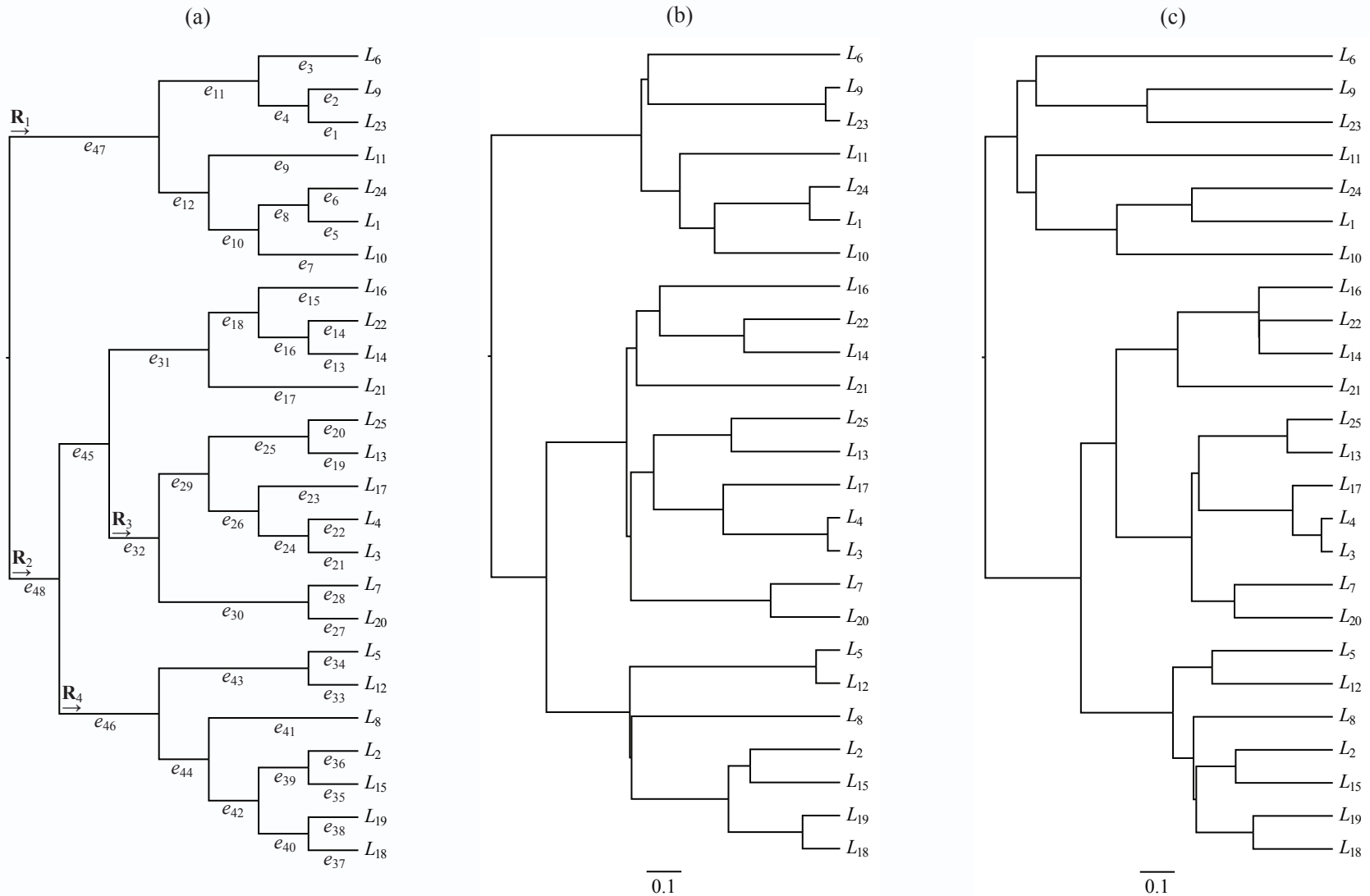
Testing the HAL-HAS model • 1

Ancestral sequence

10,000 sites

Categories	$\beta = 0.50$	$\alpha_1 = 0.35$	$\alpha_2 = 0.15$
Composition	$f_A = 0.31$	$f_A = 0.23$	$f_A = 0.48$
	$f_C = 0.17$	$f_C = 0.22$	$f_C = 0.16$
	$f_G = 0.14$	$f_G = 0.28$	$f_G = 0.16$
	$f_T = 0.39$	$f_T = 0.27$	$f_T = 0.20$

Testing the HAL-HAS model • 2



Performance of HAL-BU

- **Correct model — 4 unique rate matrices over 48 edges**
- Optimal model — **identified after comparing 2400 ± 218 models** (out of a total of 6.3×10^{44} models)
- Optimal model — **always had 4 unique rate matrices**
- Optimal model — **correct in 75% of cases**
- Number of incorrectly assigned rate matrices — **never more than 3 for a given data set**
- Average rate matrix assignment success rate — **99.25%**

Performance of HAS

- **Correct model — HAS_3 with $k = 2$**
- Optimal model — **correct in 98% of cases**
- Incorrect optimal model — **in both cases HAS_4 with $k = 2$** (implying a slight tendency to under-parameterise the data)

Accuracy of the HAL-HAS model

Value	β	α_1	α_2
Actual	0.4967	0.3547	0.1485
Inferred	0.4967 ± 0.0001	0.3562 ± 0.0088	0.1471 ± 0.0088

Type	Value	A	C	G	T
$\boldsymbol{\pi}^{inv}$	Actual	0.3055	0.1666	0.1352	0.3927
	Inferred	0.3056 ± 0.0001	0.1665 ± 0.0000	0.1353 ± 0.0000	0.3926 ± 0.0001
$\mathbf{f}_0^1$	Actual	0.2318	0.2152	0.2780	0.2751
	Inferred	0.2289 ± 0.0195	0.2140 ± 0.0140	0.2827 ± 0.0202	0.2744 ± 0.0105
$\mathbf{f}_0^2$	Actual	0.4837	0.1592	0.1589	0.1983
	Inferred	0.4806 ± 0.0287	0.1554 ± 0.0186	0.1646 ± 0.0284	0.1993 ± 0.0136

Note – Similar results were obtained for $R_1, \ldots, R_4$

Take-home message

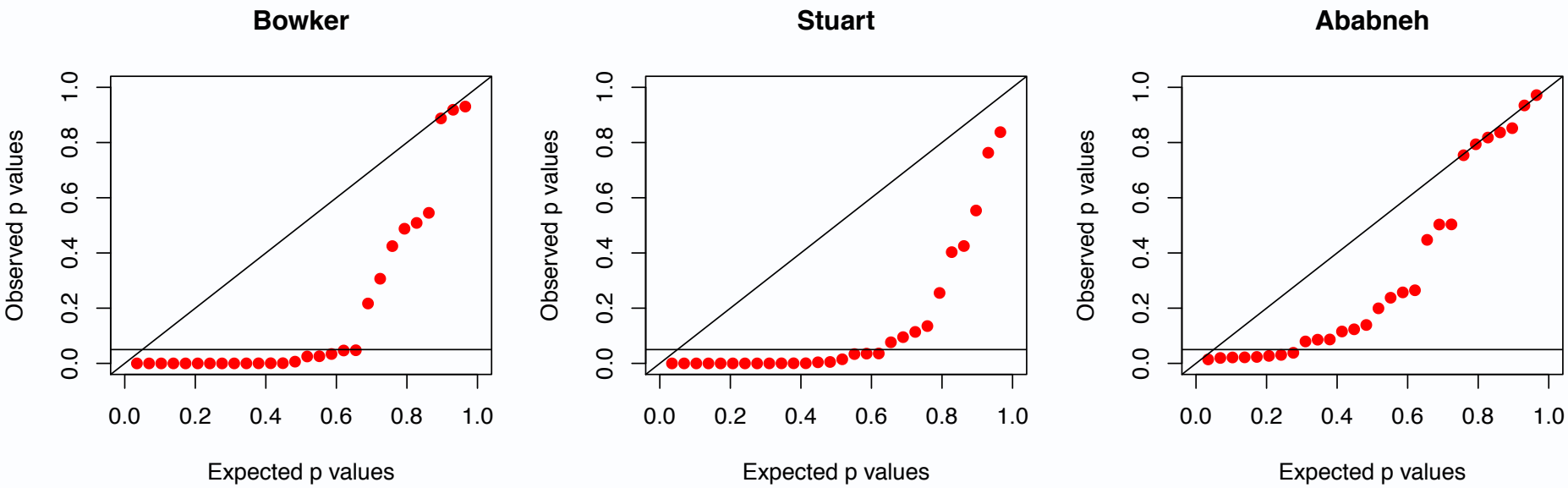
The new RAL-RAS mixture model is **efficient**, **accurate**, and **precise**

Example – evolution of 8 yeast genomes • 1

Question Could the genomes have evolved under non-SRH conditions?

Data 42,337 second codon sites (no gaps or ambiguous characters)

Method We carried out the matched-pairs test of symmetry, marginal symmetry, and internal symmetry (using **SymTest**)



Source. Rokas et al.. **Nature** 425, 798-804 [2003].

Example – evolution of 8 yeast genomes • 2

Question Which genomes have evolve under non-SRH conditions?

Data *p*-values obtained from the matched-pairs tests of symmetry

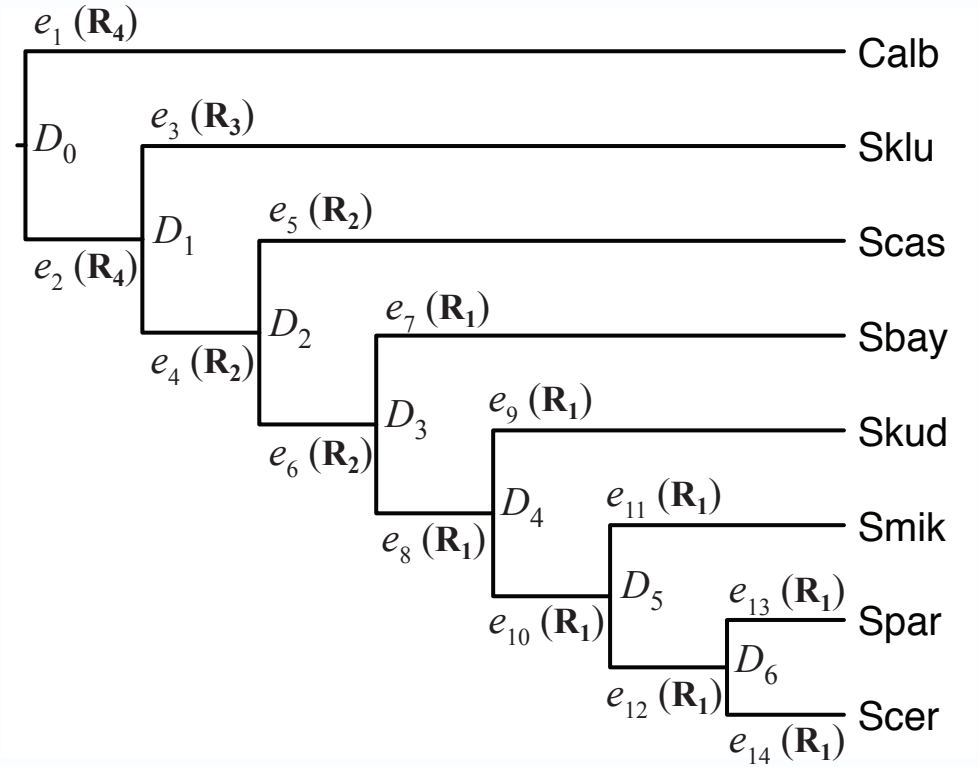
Method Evaluate Holm-Bonferroni-corrected *p*-values in a heat map

	Scer	Spar	Smik	Skud	Sbay	Scas	Sklu	Calb
Scer	—	0.5088	0.8875	0.2170	0.9187	0.0467	0.0003	0.0000
Spar	0.5088	—	0.3067	0.0476	0.5453	0.0251	0.0001	0.0000
Smik	0.8875	0.3067	—	0.4248	0.9304	0.0340	0.0000	0.0000
Skud	0.2170	0.0476	0.4248	—	0.4878	0.0063	0.0007	0.0000
Sbay	0.9187	0.5453	0.9304	0.4878	—	0.0259	0.0006	0.0000
Scas	0.0467	0.0251	0.0340	0.0063	0.0259	—	0.0000	0.0000
Sklu	0.0003	0.0001	0.0000	0.0007	0.0006	0.0000	—	0.0000
Calb	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	—

Example – evolution of 8 yeast genomes • 3

Question How complex is the evolutionary process given a ‘correct’ tree?

Data Output from our RAL-RAS mixture model



Example – evolution of 8 yeast genomes • 4

Question What are the characteristics of the inferred ancestral sequence?

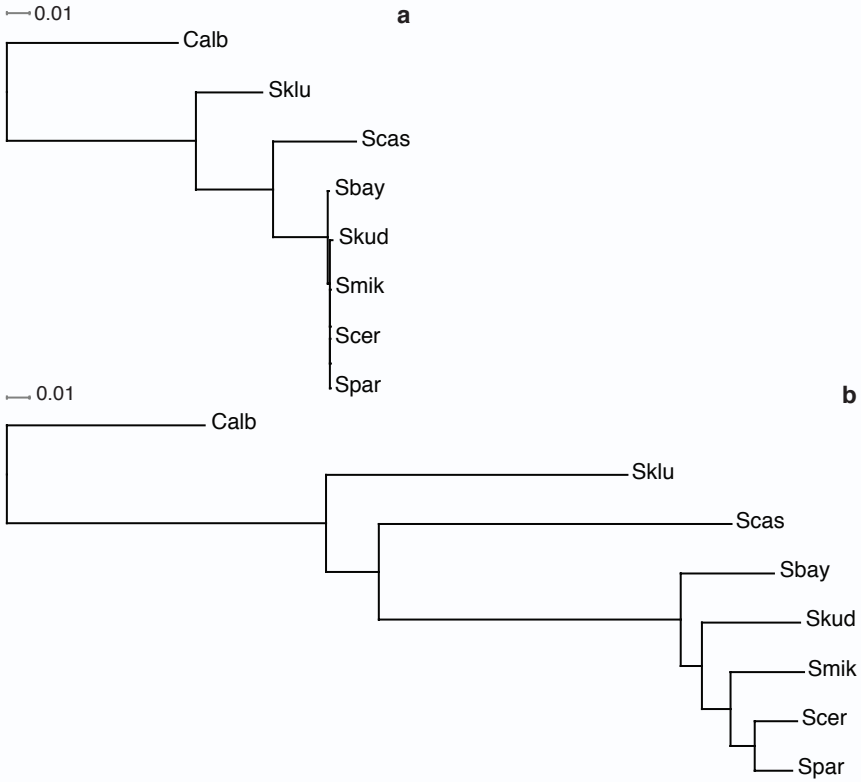
Data Output from our RAL-RAS mixture model

Invariable (50%)		ν_1 (35%)		ν_2 (15%)	
f_A	0.31	f_A	0.23	f_A	0.48
f_C	0.17	f_C	0.22	f_C	0.16
f_G	0.14	f_G	0.28	f_G	0.16
f_T	0.39	f_T	0.28	f_T	0.20

Example – evolution of 8 yeast genomes • 5

Question How much have the variable sites (ν_1 & ν_2) evolved?

Data Output from our RAL-RAS mixture model



Take-home message

The yeast genome data are inconsistent with evolution under commonly assumed phylogenetic assumptions

Thank you

Lars Jermiin
OCE Science Leader
Bioinformatics & Phylogenomics Team
t +61 2 6246 4043
e Lars.Jermiin@csiro.au
w www.csiro.au/people/Lars.Jermiin.html

Collaborators

Faisal Ababneh (Al-Hussein Bin Talal University)
Vivek Jayaswal (Queensland University of Technology)
Leon Poladian (University of Sydney)
John Robinson (University of Sydney)
Thomas Wong (CSIRO Ecosystem Sciences)