From Phylogenetics to Phylogenomics: Mathematical and Statistical Challenges in the Era of Big Data

Oct 21 - 25, 2024

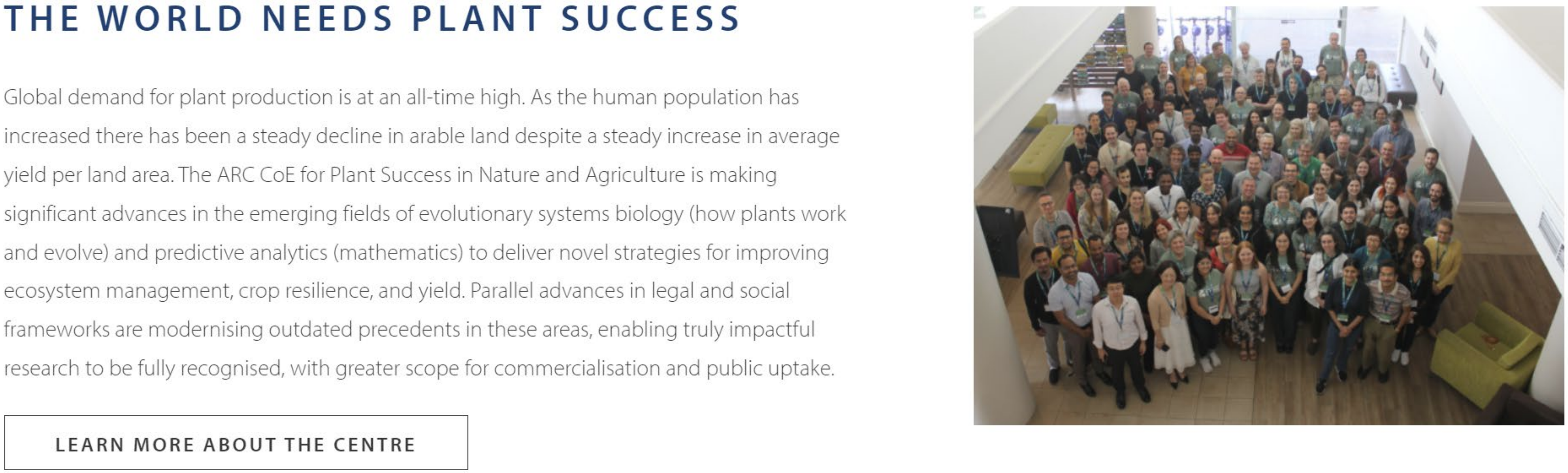
NSF | ICERM

ARC CENTRE OF EXCELLENCE FOR
PLANT SUCCESS
IN NATURE AND AGRICULTURE

Open problems in comparative phylogenomics

Barbara Holland

21/10/2024



THE WORLD NEEDS PLANT SUCCESS

Global demand for plant production is at an all-time high. As the human population has increased there has been a steady decline in arable land despite a steady increase in average yield per land area. The ARC CoE for Plant Success in Nature and Agriculture is making significant advances in the emerging fields of evolutionary systems biology (how plants work and evolve) and predictive analytics (mathematics) to deliver novel strategies for improving ecosystem management, crop resilience, and yield. Parallel advances in legal and social frameworks are modernising outdated precedents in these areas, enabling truly impactful research to be fully recognised, with greater scope for commercialisation and public uptake.

- Part of the CoE’s mission is to see if/how evolution can inform agriculture.
- We are interested in the evolution of tolerance to heat and drought.
- Can a phylogenetic perspective help?

Team Phylo @ the UTas Plant Success node

Barbara Holland

Ben Halliwell

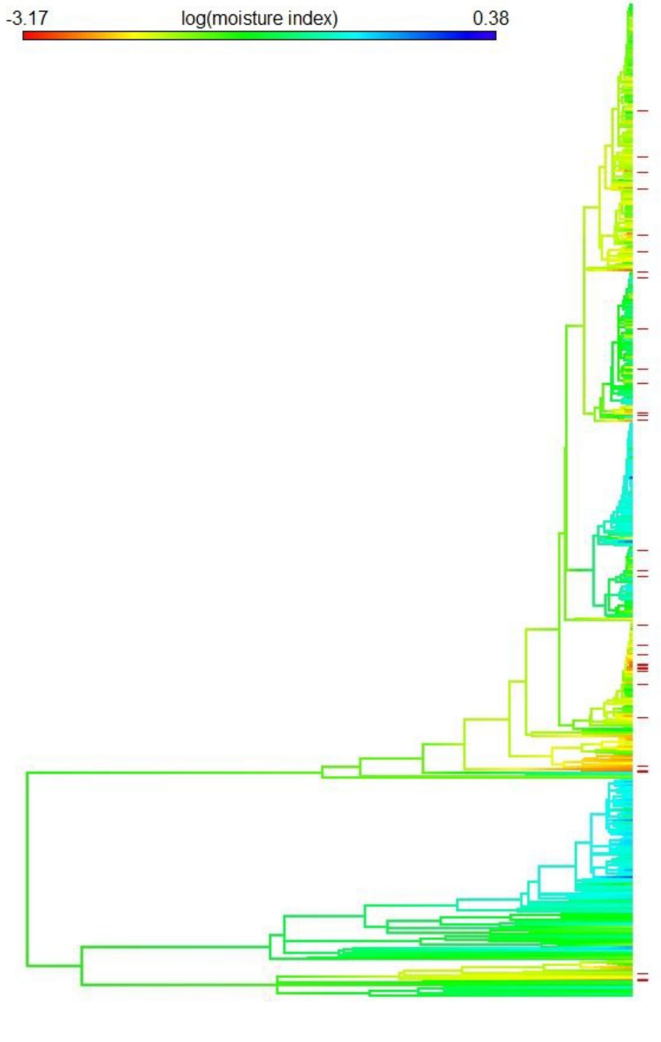
Luke Yates

Jonathan Mitchell

Arlie Macdonald

+ Claire Edwards

Evolution finds repeated solutions



Eucalyptus

- Approx. 800 species
- Mix of arid, semi-arid and mesic species
- Many, apparently independent, transitions into arid environments (<250mm/year) in different taxonomic sections
- Taxonomic sections are (mostly) reproductively isolated, discounting hybridization

Ben Halliwell

ARC CENTRE OF EXCELLENCE FOR PLANT SUCCESS IN NATURE AND AGRICULTURE

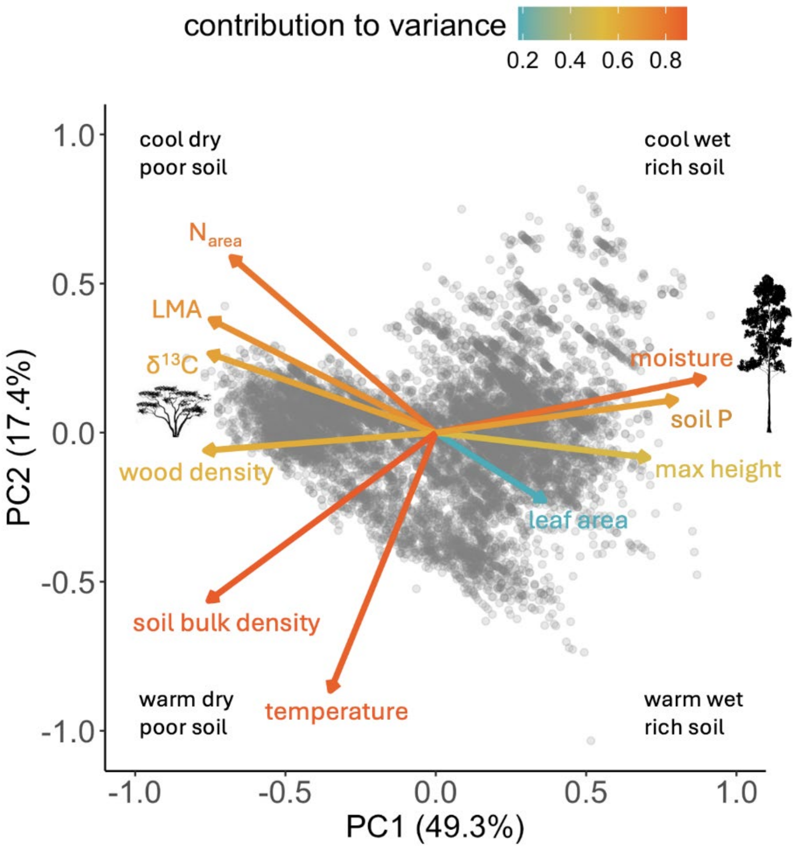
What traits are correlated with arid conditions?

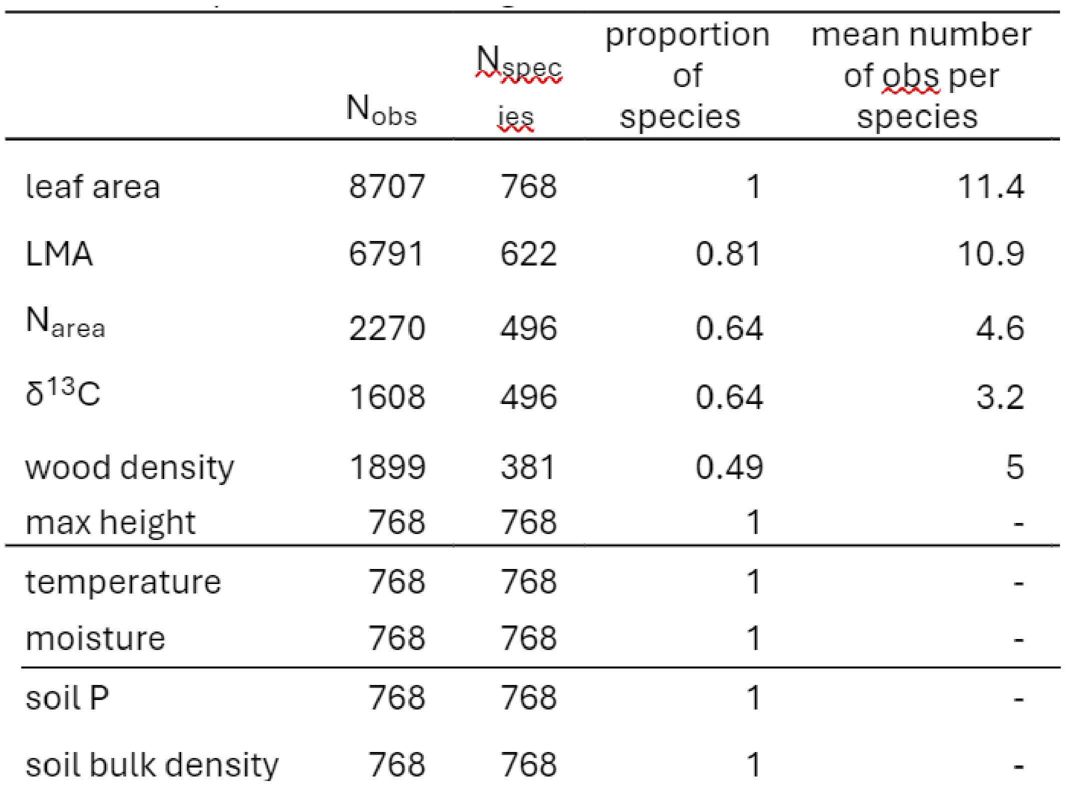
- Does the Anna Karenina principal apply? I.e. do we see the same combination of traits in all the arid adapted species?

- Seems like a question for comparative phylogenetic methods…

ARC CENTRE OF EXCELLENCE FOR PLANT SUCCESS IN NATURE AND AGRICULTURE

In Eucalypts there seems to be one way





	N_{obs}	$N_{species}$	proportion of species	mean number of obs per species
leaf area	8707	768	1	11.4
LMA	6791	622	0.81	10.9
N_{area}	2270	496	0.64	4.6
$\delta^{13}C$	1608	496	0.64	3.2
wood density	1899	381	0.49	5
max height	768	768	1	-
temperature	768	768	1	-
moisture	768	768	1	-
soil P	768	768	1	-
soil bulk density	768	768	1	-

Halliwell et al draft in prep

ARC CENTRE OF EXCELLENCE FOR PLANT SUCCESS IN NATURE AND AGRICULTURE

Phylogenies and the comparative method

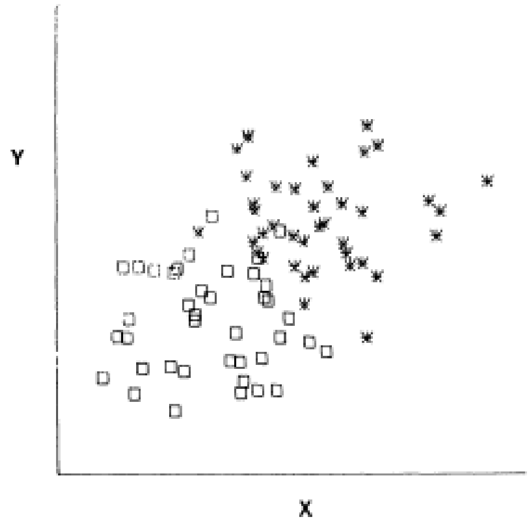


FIG. 7.—The same data set, with the points distinguished to show the members of the 2 monophyletic taxa. It can immediately be seen that the apparently significant relationship of fig. 6 is illusory.

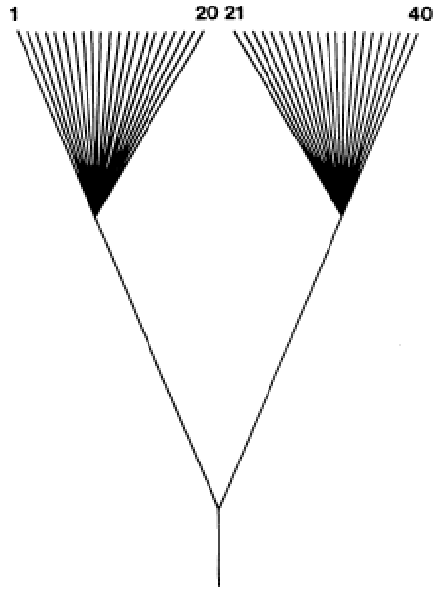


FIG. 5.—A "worst case" phylogeny for 40 species, in which there prove to be 2 groups each of 20 close relatives.

PHYLOGENIES AND THE COMPARATIVE METHOD

JOSEPH FELSENSTEIN

Department of Genetics SK-50, University of Washington, Seattle, Washington 98195

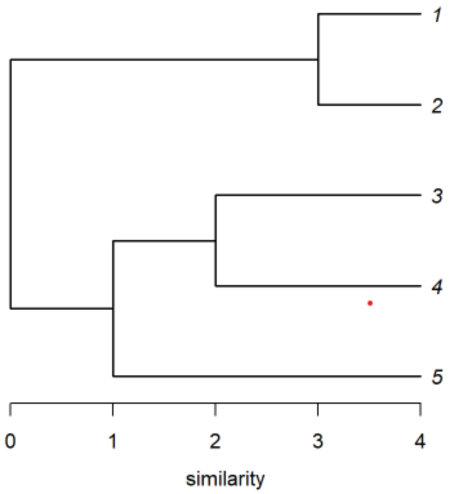
Submitted November 30, 1983; Accepted May 23, 1984

Things I didn’t know about PICs and PGLS

$$y = X\beta + \ \varepsilon$$
$$\varepsilon \sim MVN(0, \Sigma)$$
$$\Sigma = \ \lambda C + (1 - \lambda)I$$

PGLS and PIC are equivalent when λ = 1

PGLS assumes that there is NO phylogenetic signal in X



$$C = \begin{pmatrix} 4 & 3 & 0 & 0 & 0 \\ 3 & 4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 1 & 1 \\ 0 & 0 & 1 & 4 & 2 \\ 0 & 0 & 1 & 2 & 4 \end{pmatrix}$$

MR-PMMs put it all on the LHS

Received: 27 December 2022 | Accepted: 28 May 2023

DOI: 10.1111/1365-2745.14150

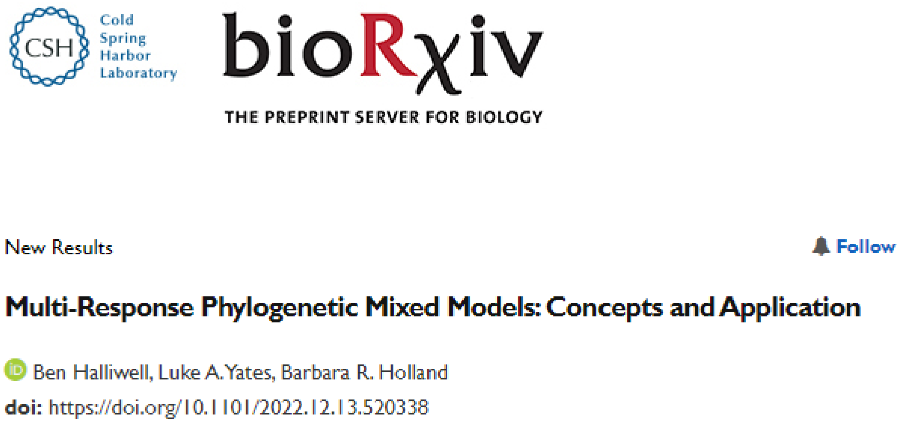
Journal of Ecology BRITISH ECOLOGICAL SOCIETY

REVIEW

Grime Review: Phil Grime's Impact on the Present and Future of Plant Ecology

Phylogenetically conservative trait correlation: Quantification and interpretation

Mark Westoby[1] | Luke Yates[2] | Barbara Holland[3] | Ben Halliwell[2]



CSH Cold Spring Harbor Laboratory

bioRχiv THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow

Multi-Response Phylogenetic Mixed Models: Concepts and Application

Ben Halliwell, Luke A. Yates, Barbara R. Holland

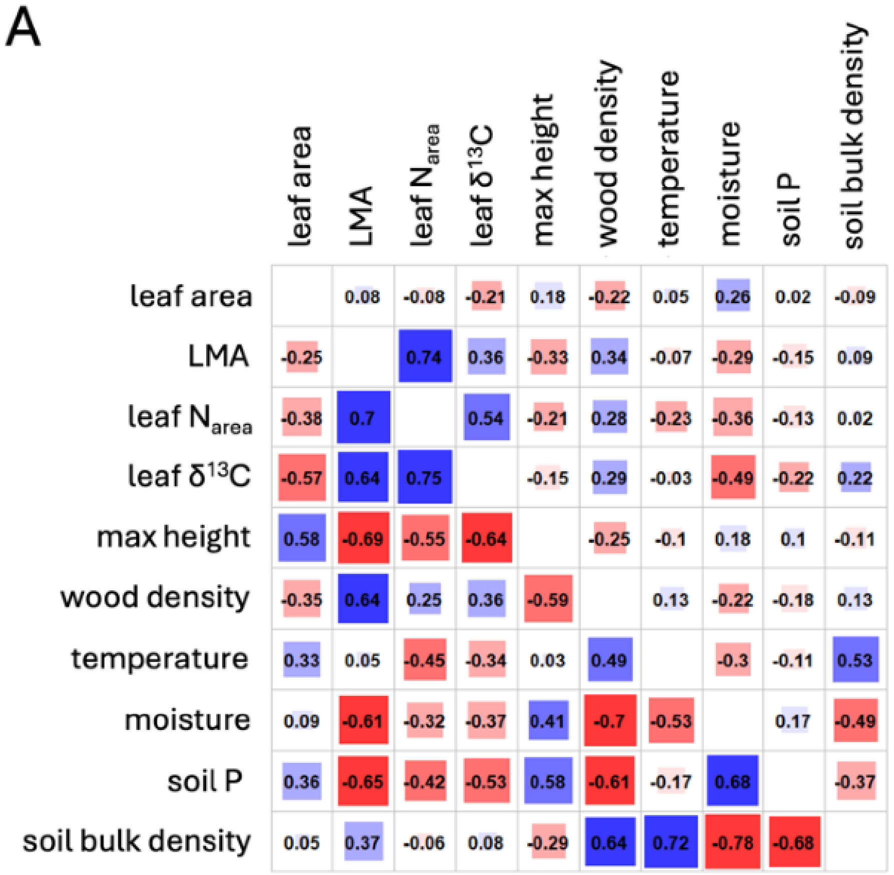
doi: https://doi.org/10.1101/2022.12.13.520338

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1 + \mathbf{b}_1 + \mathbf{e}_1 \\ \boldsymbol{\mu}_2 + \mathbf{b}_2 + \mathbf{e}_2 \end{pmatrix}$$

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \begin{pmatrix} \beta_{0,1}\mathbf{1} + \beta_{1,1}\boldsymbol{X}_{1,1} + \ldots + \beta_{k,1}\boldsymbol{X}_{k,1} \\ \beta_{0,2}\mathbf{1} + \beta_{1,2}\boldsymbol{X}_{1,2} + \ldots + \beta_{k,2}\boldsymbol{X}_{k,2} \end{pmatrix}$$

$$(\mathbf{b}_1, \mathbf{b}_2)^{\mathrm{T}} \sim \mathrm{MVN}(0, \Sigma^{\mathrm{b}} \otimes C)$$
$$(\mathbf{e}_1, \mathbf{e}_2)^{\mathrm{T}} \sim \mathrm{MVN}(0, \Sigma^{\mathrm{e}} \otimes I)$$

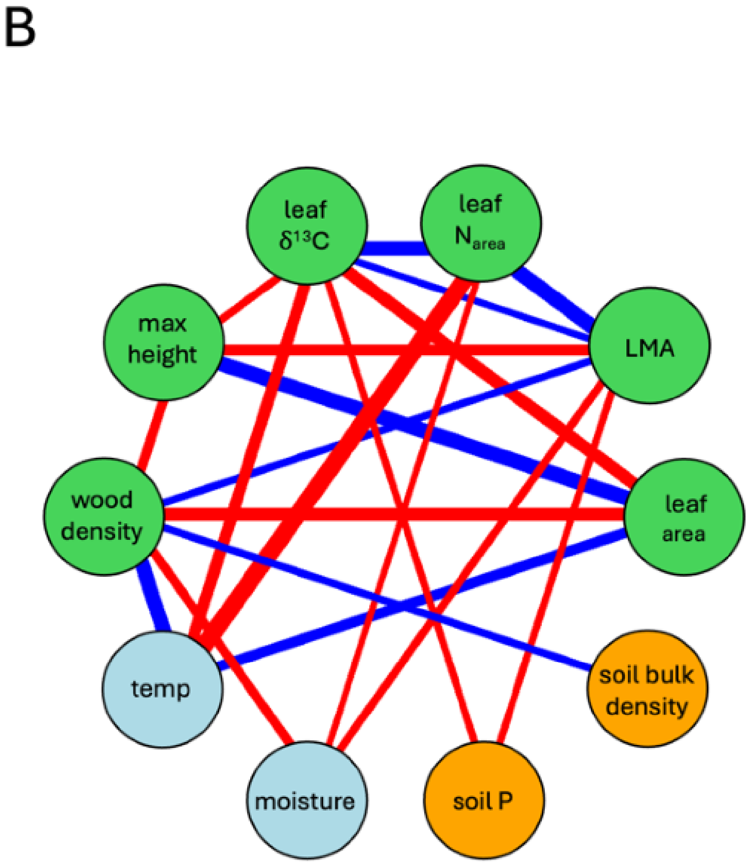
$$\Sigma = \Sigma^{\mathrm{b}} \otimes C + \Sigma^{\mathrm{e}} \otimes I$$



A

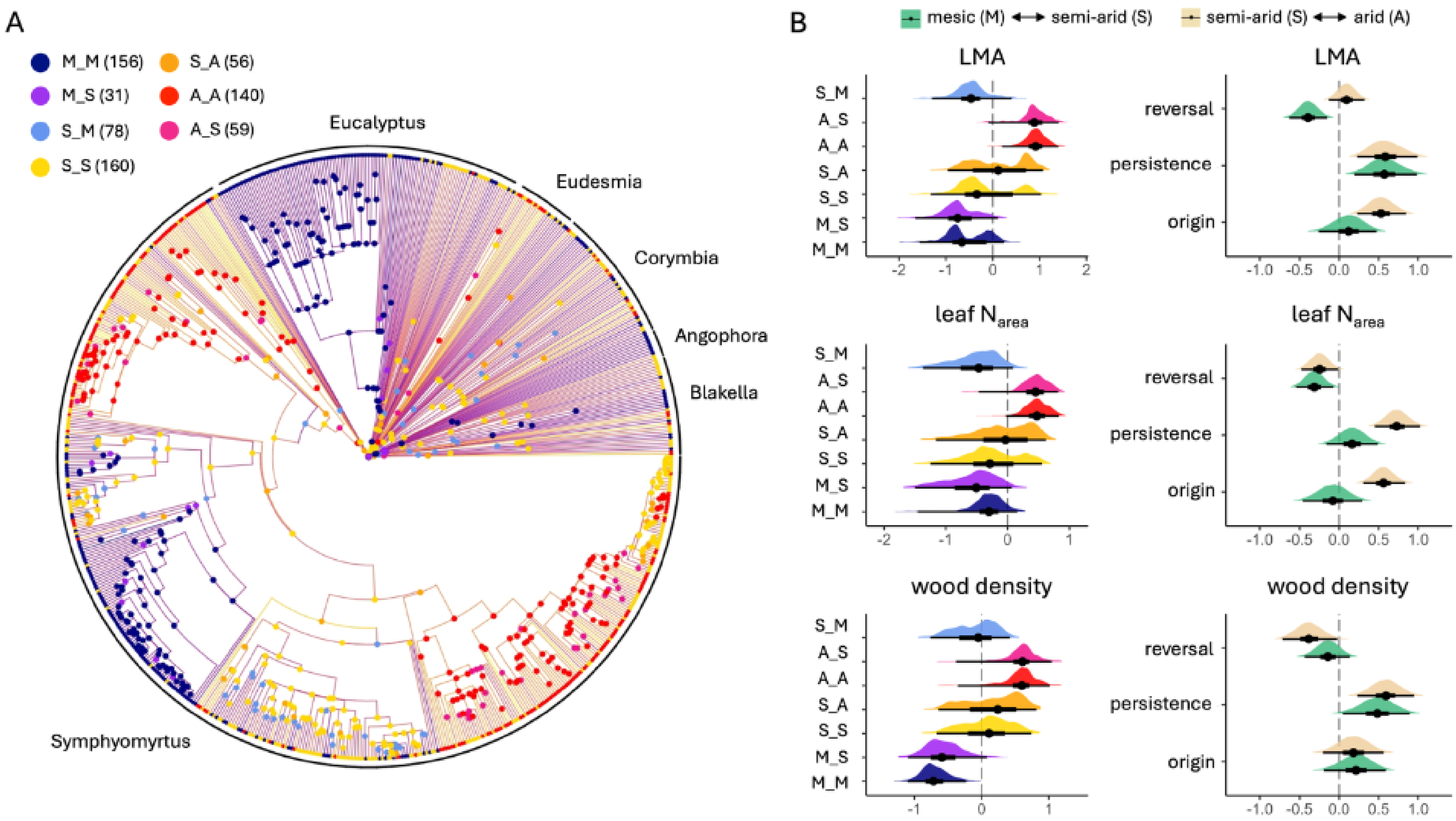
	leaf area	LMA	leaf N_{area}	leaf $\delta^{13}C$	max height	wood density	temperature	moisture	soil P	soil bulk density
leaf area		0.08	-0.08	-0.21	0.18	-0.22	0.05	0.26	0.02	-0.09
LMA	-0.25		0.74	0.36	-0.33	0.34	-0.07	-0.29	-0.15	0.09
leaf N_{area}	-0.38	0.7		0.54	-0.21	0.28	-0.23	-0.36	-0.13	0.02
leaf $\delta^{13}C$	-0.57	0.64	0.75		-0.15	0.29	-0.03	-0.49	-0.22	0.22
max height	0.58	-0.69	-0.55	-0.64		-0.25	-0.1	0.18	0.1	-0.11
wood density	-0.35	0.64	0.25	0.36	-0.59		0.13	-0.22	-0.18	0.13
temperature	0.33	0.05	-0.45	-0.34	0.03	0.49		-0.3	-0.11	0.53
moisture	0.09	-0.61	-0.32	-0.37	0.41	-0.7	-0.53		0.17	-0.49
soil P	0.36	-0.65	-0.42	-0.53	0.58	-0.61	-0.17	0.68		-0.37
soil bulk density	0.05	0.37	-0.06	0.08	-0.29	0.64	0.72	-0.78	-0.68	

Below diagonal: phylogenetic correlations
Above diagonal: residual correlations



Partial Phylogenetic Correlations

Halliwell et al draft in prep



Halliwell et al draft in prep

Open protoproblem[1] #1

- Most (all?) methods of ancestral state reconstruction assume neutral evolution, i.e. no directional selection

- Is there a statistically sound way to do ASR if there is directional selection?

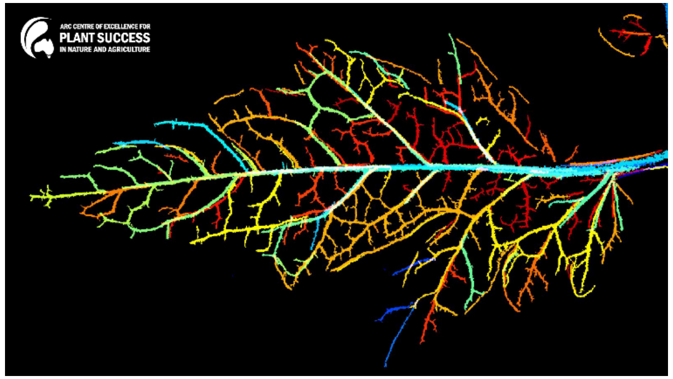
[1] An open protoproblem is a poorly formed open problem

Modelling extinction risk

Rachel Gallagher

Suz Everingham

- Rachel and Suz are interested in modelling extinction risk due to drought in plants
- A question of how best to integrate
  - trait data (AusTraits, TRY)
  - geographic range data
  - climate data associated with that range
  - projected climate associated with that range
  - phylogenies
- Some traits (p_{50}, g_{min}, T_p) are more relevant than others, but they are often harder to measure



Protoproblems #2 & #3

- If you can afford to measure hard traits (e.g. p50) in a subset of your species and soft traits (e.g. wood density) in a broader group, how should you optimally collect data to reduce uncertainty in a phylogenetic imputation?

- What are the expectations about loss of phylogenetic diversity (PD) (or feature diversity) when propensity to go extinct depends on suites of correlated traits? I.e NOT the field of bullets model?

Protoproblem #4

- Can we find breakpoints on a tree where the association between traits and environment alters?

Within Eucs there seems to be a common strategy plants use to tolerate arid environments, but if we looked at broader taxonomic groups, we'd expect the see different strategies and hence different trait-trait and trait-environment correlations

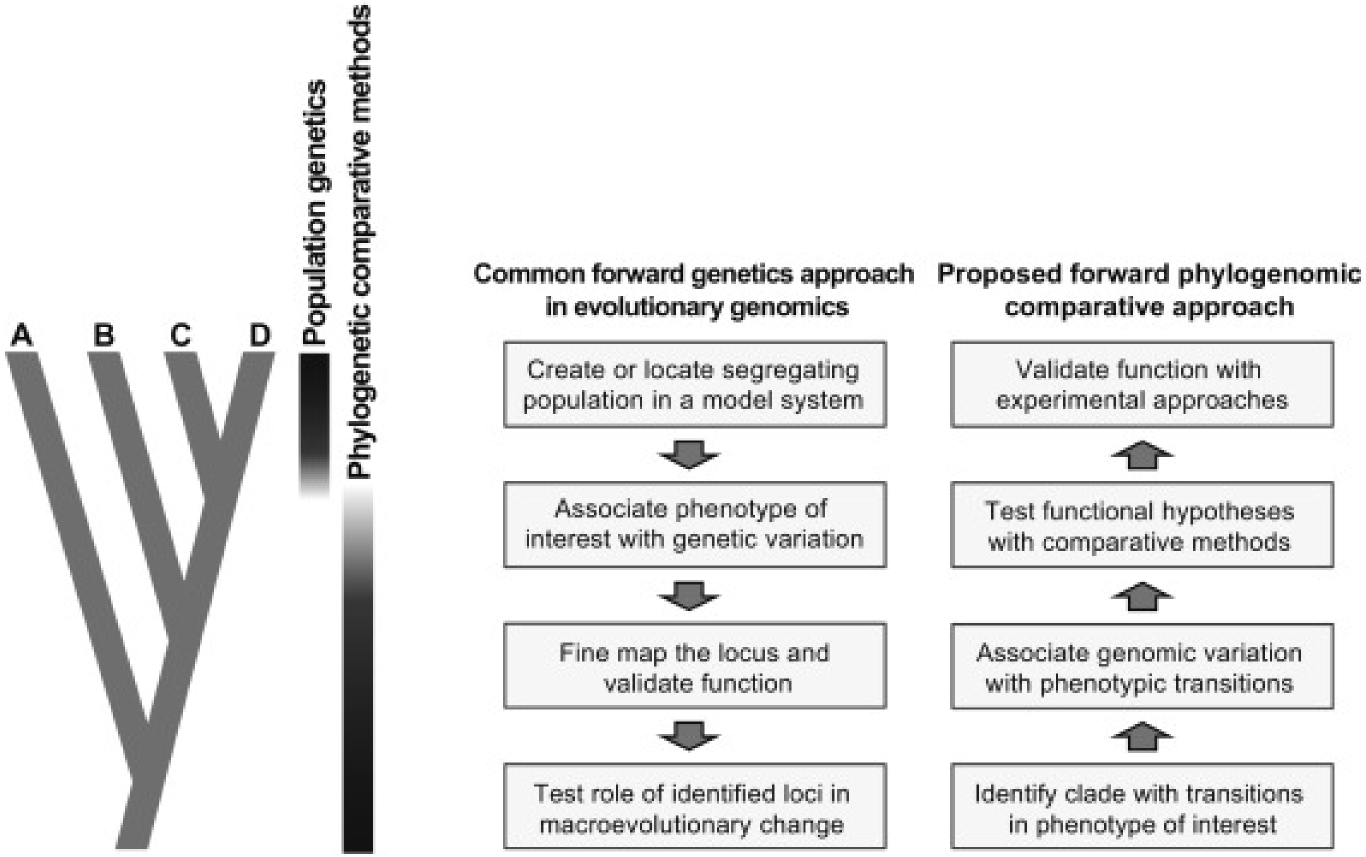
There seem to be several methods that look at changes in mean, e.g. different OU processes in different parts of the tree, but I haven't found anything that looks for a change in association…

PhyloGWAS / PhyloG2P

- Assuming that evolution has found the same solution, has it used the same genes/ genomic regions?
- Can we find what they are?

- This area of research, aimed at identifying genomic regions associated with traits of interest, is called PhyloG2P or PhyloGWAS
  - Look for SNPS that match the presence/absence of the trait
  - Look for evidence of accelerated branch lengths in species with/without the trait

PhyloG2P: Smith, S. D., Pennell, M. W., Dunn, C. W., & Edwards, S. V. (2020). Phylogenetics is the new genetics (for most of biodiversity). *Trends in Ecology & Evolution*, *35*(5), 415-425.



Common forward genetics approach in evolutionary genomics

Create or locate segregating population in a model system

Associate phenotype of interest with genetic variation

Fine map the locus and validate function

Test role of identified loci in macroevolutionary change

Proposed forward phylogenomic comparative approach

Validate function with experimental approaches

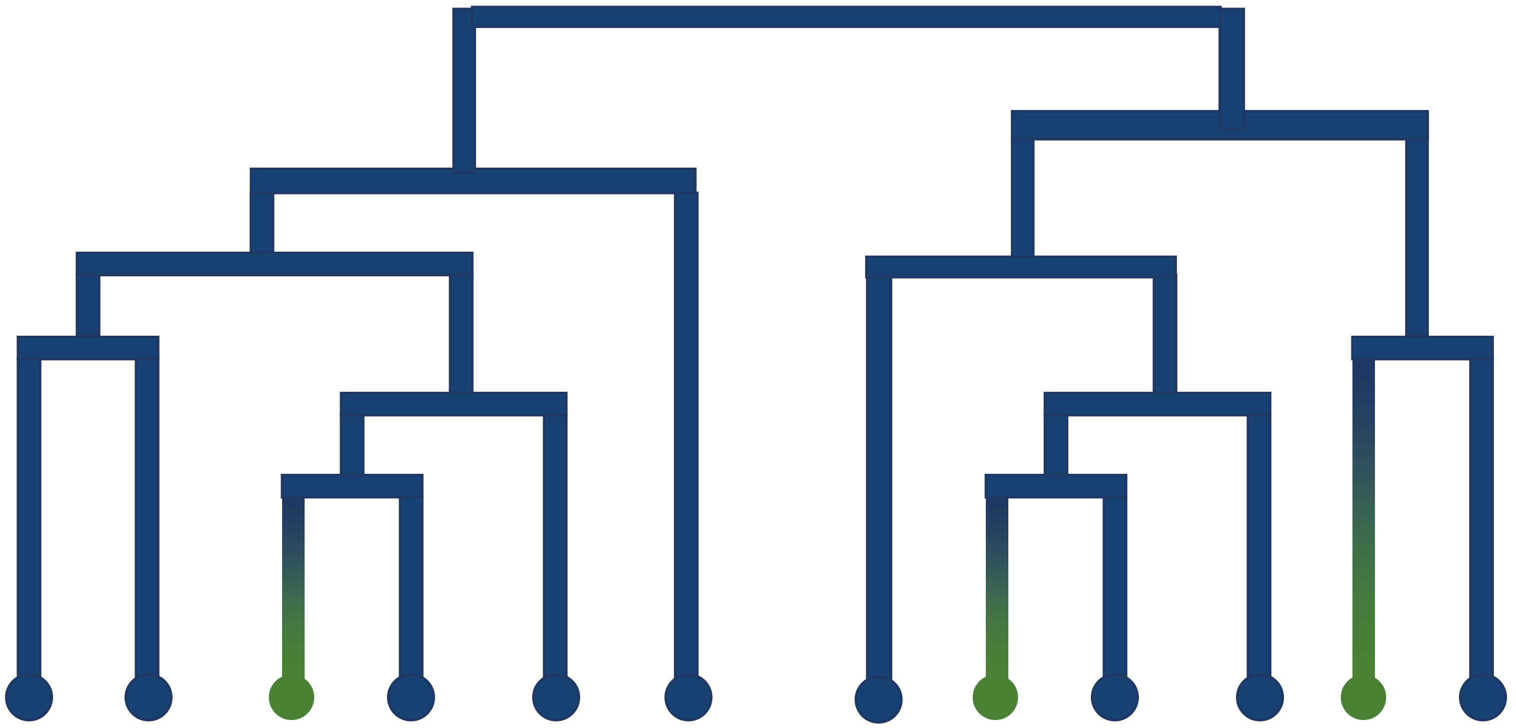
Test functional hypotheses with comparative methods

Associate genomic variation with phenotypic transitions

Identify clade with transitions in phenotype of interest

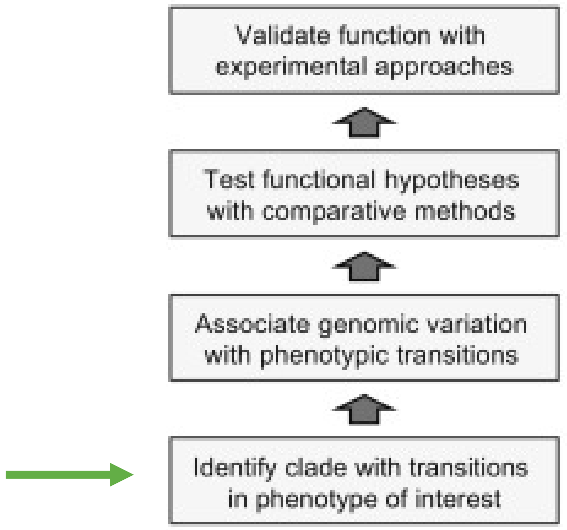
Trends in Ecology & Evolution

E.g. Gene Presence/Absence data



Gene A	1	1	1	1	1	1	0	0	0	0	0	0
Gene B	1	1	1	1	1	1	1	1	1	0	0	0
Gene C	0	0	1	1	1	1	0	0	0	0	0	0
Gene D	1	1	1	1	0	1	0	0	0	0	1	0
Gene E	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**0**
Gene F	*0*	*0*	*1*	*1*	*0*	*0*	*0*	*1*	*1*	*0*	*1*	*0*

Proposed forward phylogenomic comparative approach



Validate function with experimental approaches

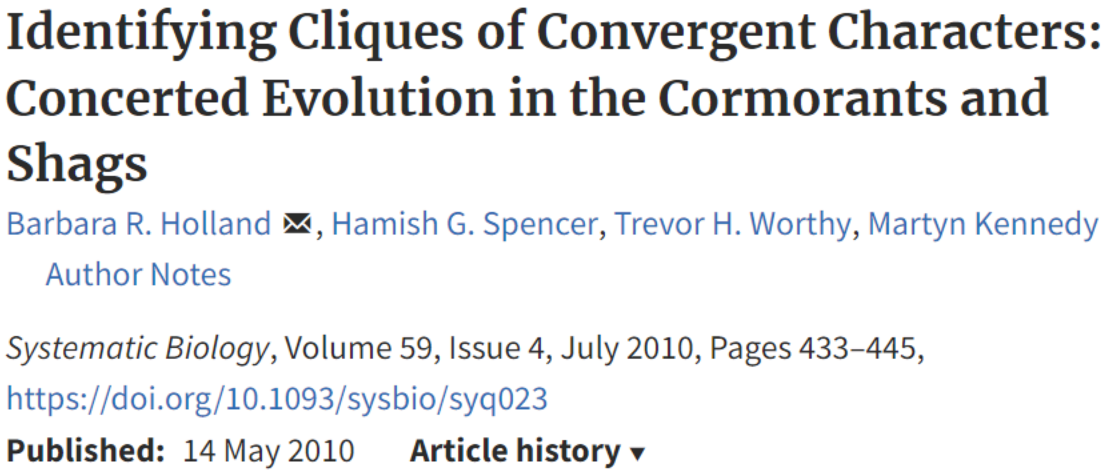
Test functional hypotheses with comparative methods

Associate genomic variation with phenotypic transitions

Identify clade with transitions in phenotype of interest

Moving beyond one SNP at a time

Protoproblem #5: Given a phylogeny and a character matrix. How unusual is it to find a subset of n mutually compatible characters given that their excess (additional mutations) is k?



Identifying Cliques of Convergent Characters: Concerted Evolution in the Cormorants and Shags

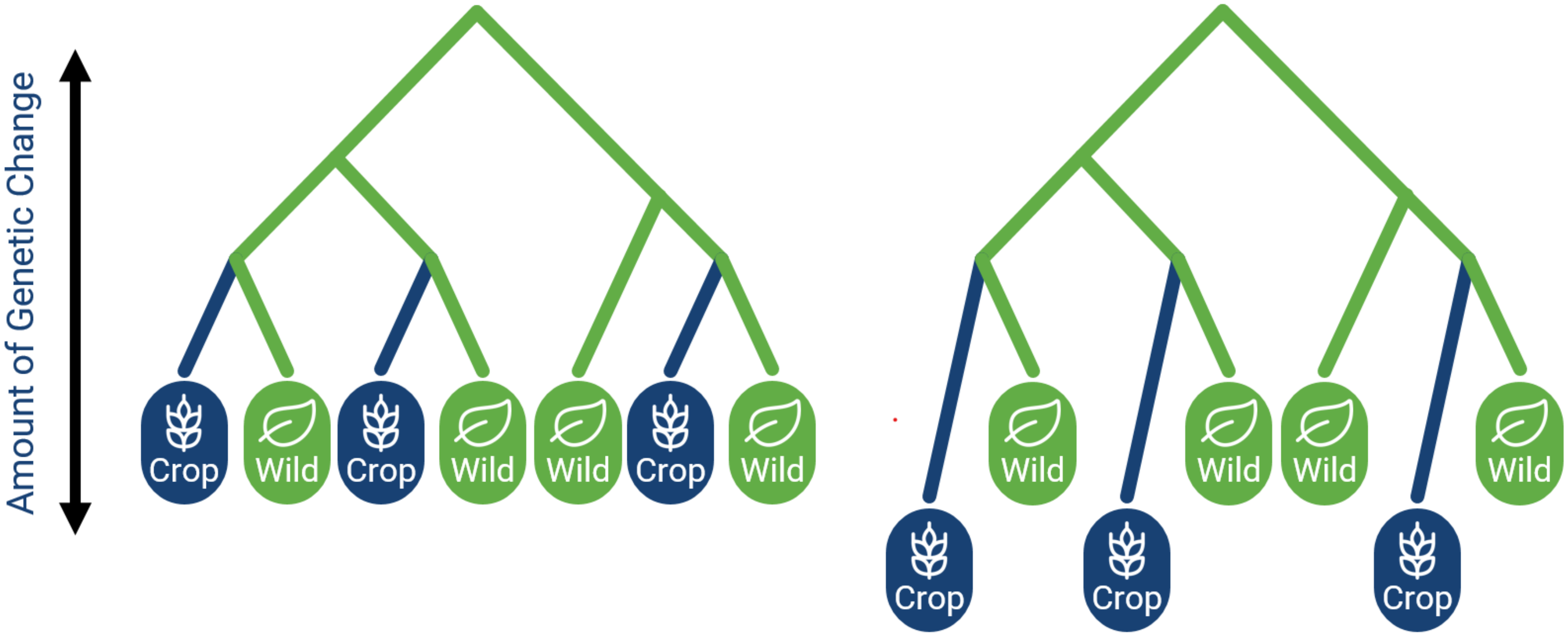
Barbara R. Holland ✉, Hamish G. Spencer, Trevor H. Worthy, Martyn Kennedy

Author Notes

Systematic Biology, Volume 59, Issue 4, July 2010, Pages 433–445, https://doi.org/10.1093/sysbio/syq023

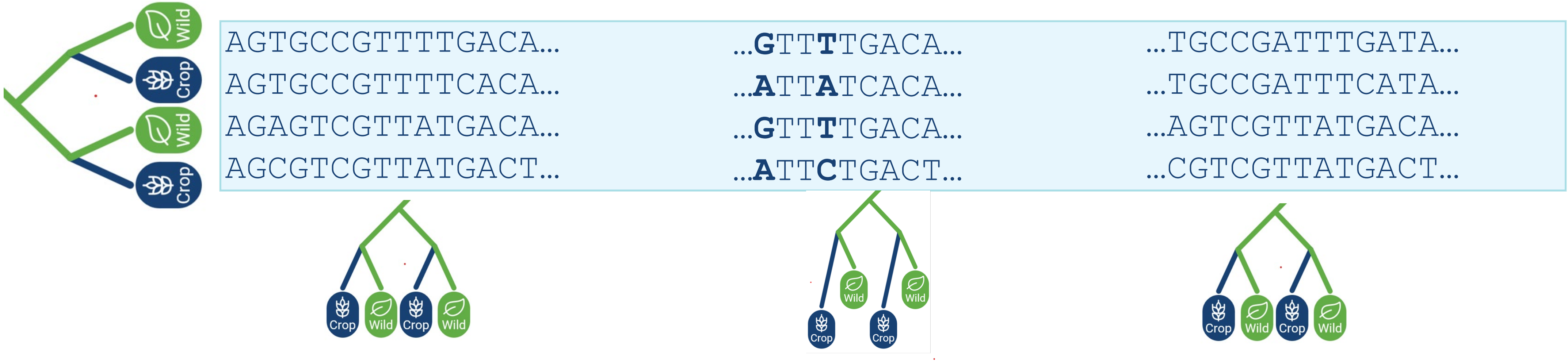
Published: 14 May 2010 **Article history** ▾

PhyloG2P



ARC CENTRE OF EXCELLENCE FOR PLANT SUCCESS IN NATURE AND AGRICULTURE

Phylogenetic models of rate variation

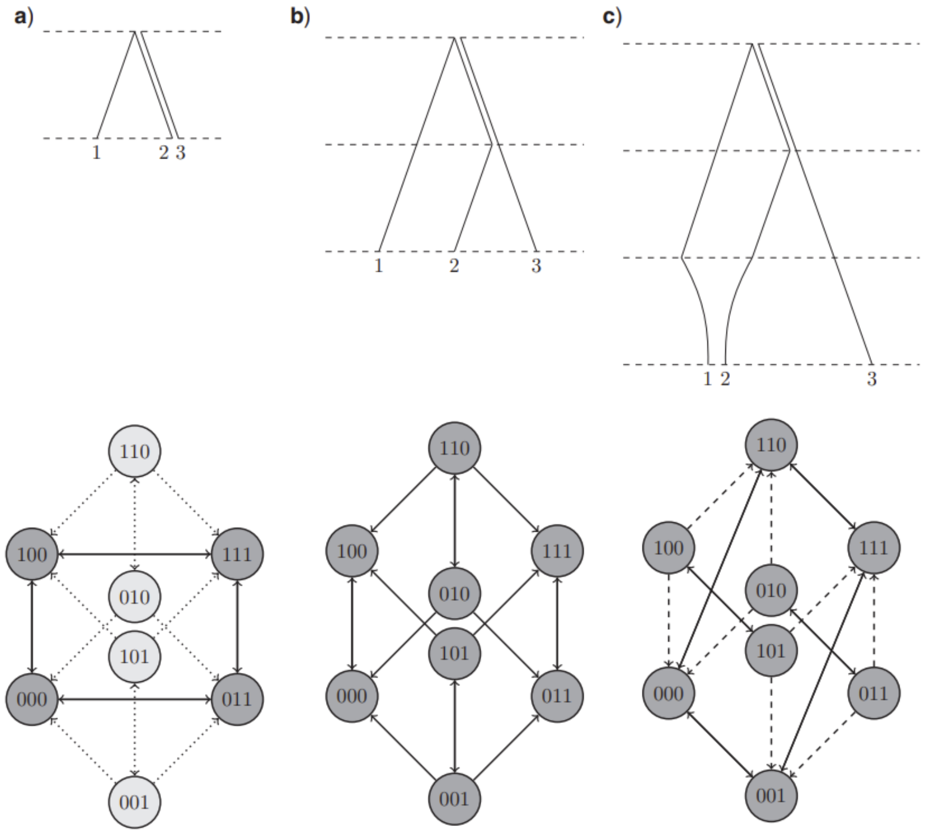


AGTGCCGTTTTGACA…	…**G**TTT**T**GACA…	…TGCCGATTTGATA…
AGTGCCGTTTTCACA…	…**A**TT**A**TCACA…	…TGCCGATTTCATA…
AGAGTCGTTATGACA…	…**G**TTT**T**GACA…	…AGTCGTTATGACA…
AGCGTCGTTATGACT…	…**A**TT**C**TGACT…	…CGTCGTTATGACT…

Partition models – let different partitions (genes) have different edge lengths
Mixture Models – fit all trees to all sites

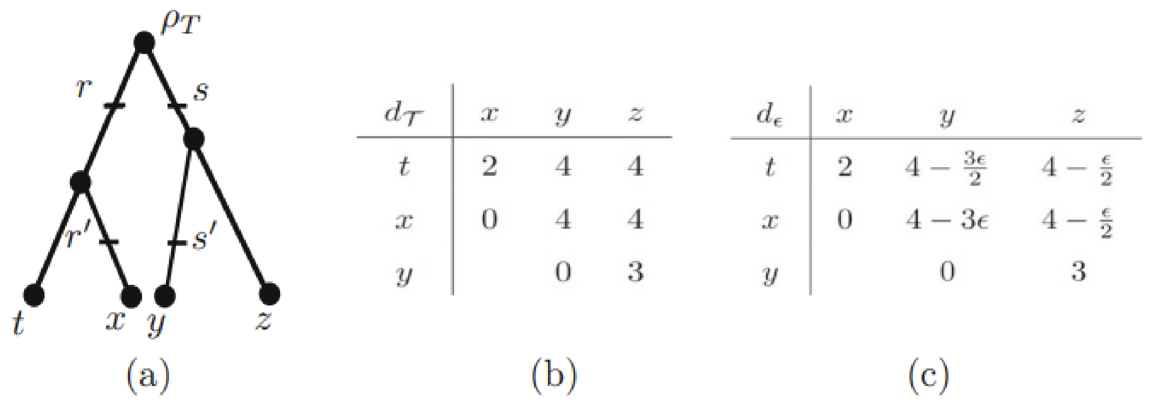
PhyloHMMs – best of both worlds?

Convergence – models in need of algorithms…



Distinguishing Between Convergent Evolution and Violation of the Molecular Clock for Three Taxa

Jonathan D. Mitchell[1,2,*], Jeremy G. Sumner[1], and Barbara R. Holland[1]



$d_\mathcal{T}$	x	y	z
t	2	4	4
x	0	4	4
y		0	3

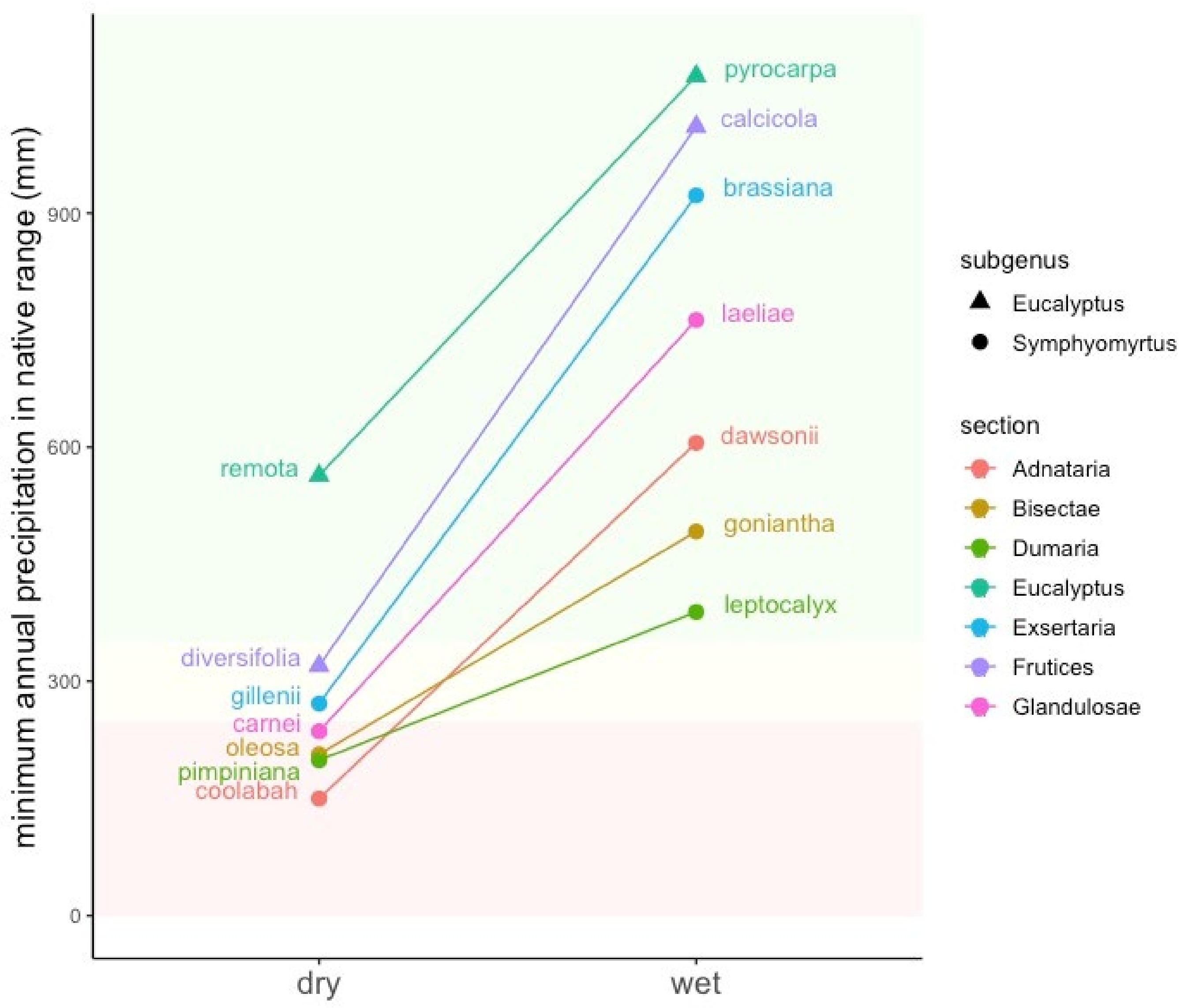
d_ϵ	x	y	z
t	2	$4-\frac{3\epsilon}{2}$	$4-\frac{\epsilon}{2}$
x	0	$4-3\epsilon$	$4-\frac{\epsilon}{2}$
y		0	3

Fig. 2 **a** Example of a convergence scenario ($\mathcal{T}=(T,w), R, \epsilon$) on $X=\{x,y,z,t\}$, where T is the depicted phylogenetic tree on X, $h(\rho_T)=2$, $h(lca_T(t,x))=1$, $h(lca_T(y,z))=\frac{3}{2}$, $\alpha=\frac{1}{4}$, $\beta=\frac{7}{4}$, and $0<\epsilon<\frac{4}{3}$. **b** The distance matrix for $d_\mathcal{T}$. **c** The distance matrix d_ϵ. Note that d_ϵ is a metric, but not a tree metric

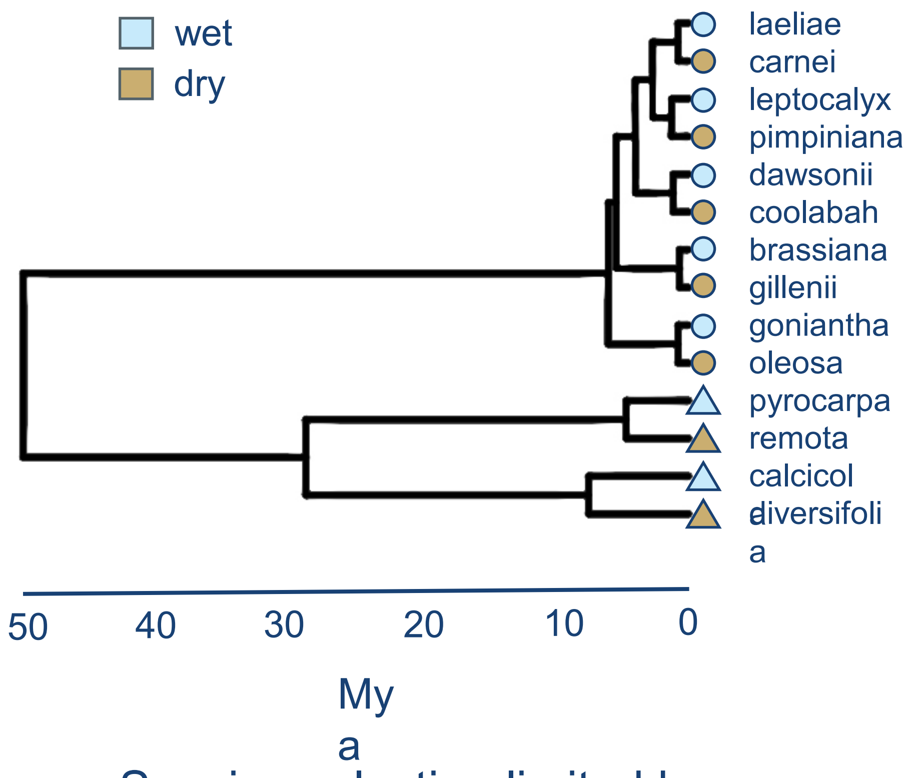
A distance-based model for convergent evolution

Barbara Holland[1] · Katharina T. Huber[2] · Vincent Moulton[2]

Comparative transcriptomics of drought



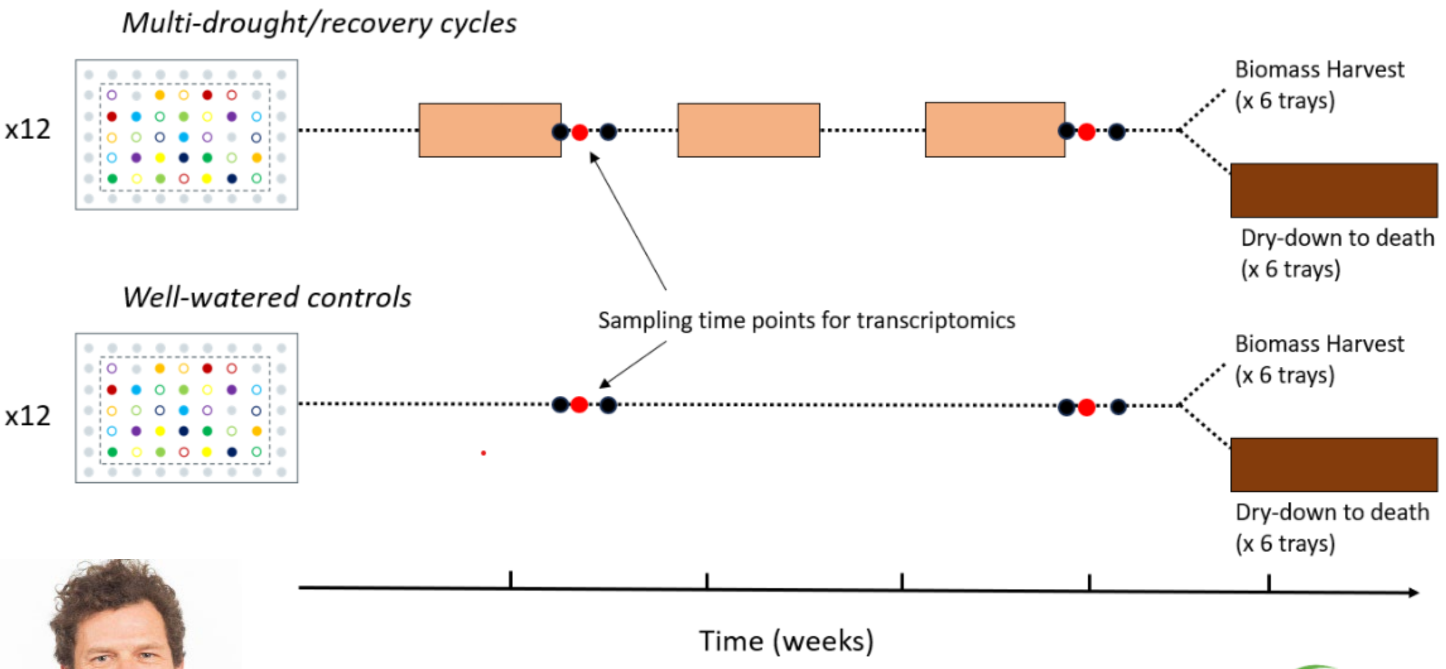
Symphyomyrtus & Eucalyptus



Species selection limited by

- Included in phylogeny
- Seed availability
- Magnitude of contrast

Comparative transcriptomics of drought



Not possible to fit a full MR-PMM to 20,000 genes.

How best should the models we fit for different genes “learn” from each other?

Chris Blackman

$$Gene\ Expression \sim Treatment + Type + Treatment{:}Type + \ \varepsilon$$

$$\varepsilon \sim MVN(0, \Sigma)$$
$$\Sigma = \ \lambda C + (1-\lambda)I$$

ARC CENTRE OF EXCELLENCE FOR PLANT SUCCESS IN NATURE AND AGRICULTURE