Phylogénie et la Biologie de la Conservation

Arne Mooers (SFU)
avec Dave Redding (UCL) & Florent Mazel (UJF)

with help and input from dozens, including Mike Steel, Klaas Hartmann, Iain Martyn, Gavin Thomas, Walter Jetz, Jeff Joy, Simon Ferrier, Marc Cadotte, Dan Faith, Sandrine Pavoine, Dan Rosauer and sPHY

with support from the sDiv working group sPHY (Leipzig, DL) & The Systematics Association (London,UK)

Basic premise (near as I can tell):

> "A system of priorities that reflects the value of taxonomic diversity can be achieved by setting priorities such that the subset of taxa that is protected has maximum underlying feature diversity. Such feature diversity of taxon subsets is difficult to estimate directly, but can be predicted by the cladistic/phylogenetic relationships among the taxa."

Faith, 1992; cited >700 times

NEWS FEATURE

NATURE|Vol 450|8 November 2007

WHAT TO LET GO

"The aim [of conservation] should be the preservation of the information content contained in the DNA of all the species on the Earth."

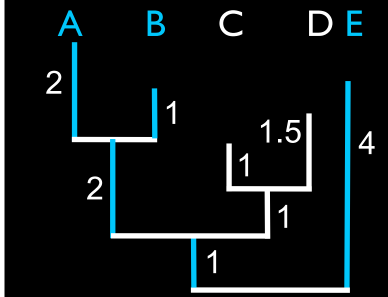
E.O.Wilson, 1992

Basic Metric

$$PD_S = \sum_{e \in S} \lambda_e$$

Phylogenetic Diversity (PD) is just sum of weights* of edges of the subtree defined by subset S of X, $|X|=n$

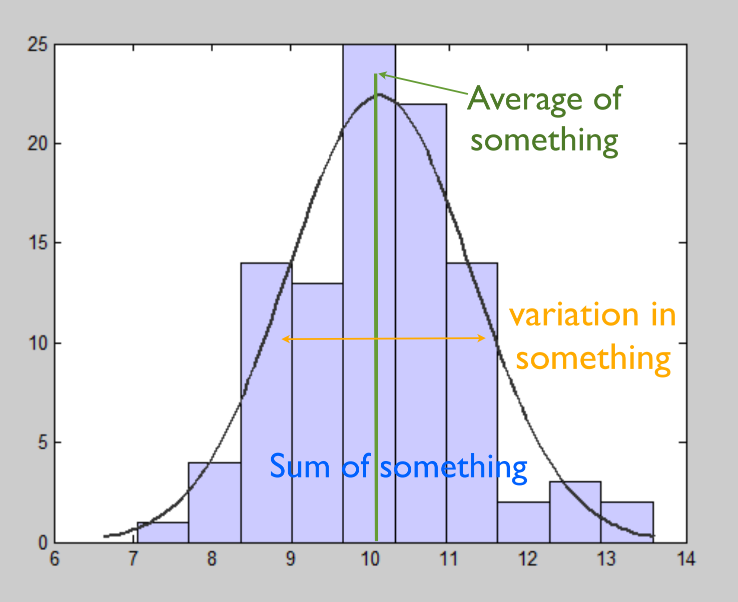
Here, $S = \{A,B,E\}$, $PD = 10$



* so these weights must reflect "underlying feature diversity"

Broad Recipe

1. Dimensions of biodiversity: $\alpha, \beta, \sum, \bar{x}, \mathrm{var}(x)$
2. Sum of edge weights on trees
3. Isolation as a surrogate for maximizing expected PD
4. Other dimensions of the tree and conservation
5. What does an edge weight represent?



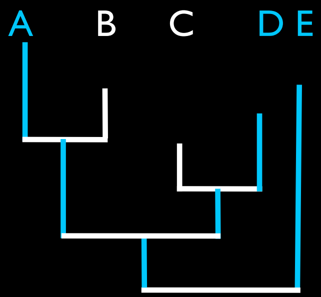
Dimensions of Phylogenetic Biodiversity

Dimension / Scale	Richness: How much Sum of...	Divergence: How different Mean of...	Regularity: How regular Variance of...
Diversity of sample(s) [alpha]	edge weights (PD)	mean(patr. dist.) mean(evol. isol.);	var(patr. dist.); tree shape, I_c
Differences among samples [beta]	sum of unshared edgeweights 'Unifrac'	?	?

following Pavoine & Bonsall, 2011, sPhy

$$PD_S = \sum_{e \in S} \lambda_e$$

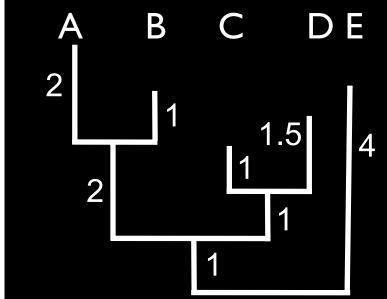
Phylogenetic Diversity (PD) is just sum of weights of edges of the subtree defined by subset S of X, |X|=*n*



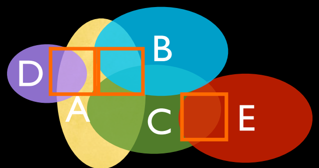
Original goal was one of optimization, generally to find maxPD given constraints, e.g. when choosing *k* of *n* tips: for *k*=3, max(PD) is reached for *S* = {*A,D,E*} (by inspection)

Maximizing PD ("max-sum") for *k of n* problem is possible with a greedy algorithm (Steel 2005; Pardi & Goldman 2005)

But, common PD setting is to choose among a defined set of *S* (e.g. representing conservation areas): this, and related problems are hard (Faller, Pardi & Steel, 2007)*



Given S = {A,D} {A,B,C} {C,E}
maxPD for 1 set is {A,B,C}
maxPD for 2 sets is {C,E} {A,D}



*but see Chernomor et al. poster 4

Recast *k* of *n* problem using "Expected PD"

$$E(PD) = \sum_{e \in T} \lambda_e \left(1 - \prod_{j=1}^{Ce} p(ext)_j\right)$$

Ce: # leaves that ultimately subtend edge *e*
p(ext)ⱼ : probability of extinction of leaf *j*

Edge weights are weighted by the probability that it persists; i.e, we build fictitious trees.

For *k* of *n*, we assign *p(ext)*=0 to leaves in *S* and *p(ext)*=1.0 to leaves not in *S*.

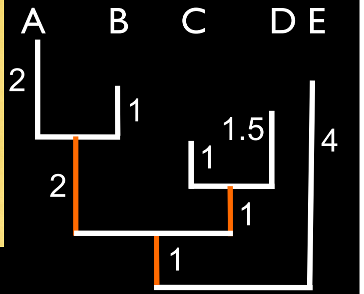
Witting & Loeschcke, 1995
Hartmann & Steel, 2006

Expected PD

p(ext) = 0.9 0.7 0.5 0.5 0.6

$$E(PD) = \sum_{e \in T} \lambda_e \left(1 - \prod_{j=1}^{Ce} p(ext)_j\right)$$

Ce: # leaves that ultimately subtend edge *e*
$p(ext)_j$: probability of extinction of leaf *j*



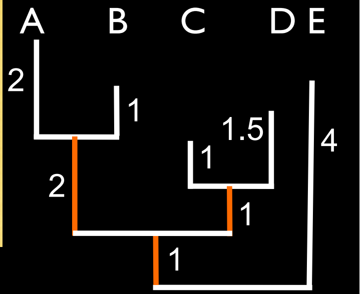
E(PD) = .2+.3+.5+.75+1.6+ .74+.75+.8425= 5.6825

Expected PD

p(ext) = 0.9 0.7 0.5 0.5 0.6

$$E(PD) = \sum_{e \in T} \lambda_e \left(1 - \prod_{j=1}^{Ce} p(ext)_j\right)$$

Ce: # leaves that ultimately subtend edge *e*
$p(ext)_j$: probability of extinction of leaf *j*



E(PD) = .2+.3+.5+.75+1.6+ .74+.75+.8425= 5.6825

If we could halve the p(extinction) of a single species, which should it be?

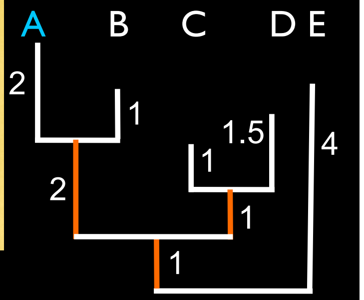
Expected PD

$p(ext) = 0.9 \quad 0.7 \quad 0.5 \quad 0.5 \; 0.6$

$$E(PD) = \sum_{e \in T} \lambda_e (1 - \prod_{j=1}^{Ce} p(ext)_j)$$

Ce: # leaves that ultimately subtend edge *e*

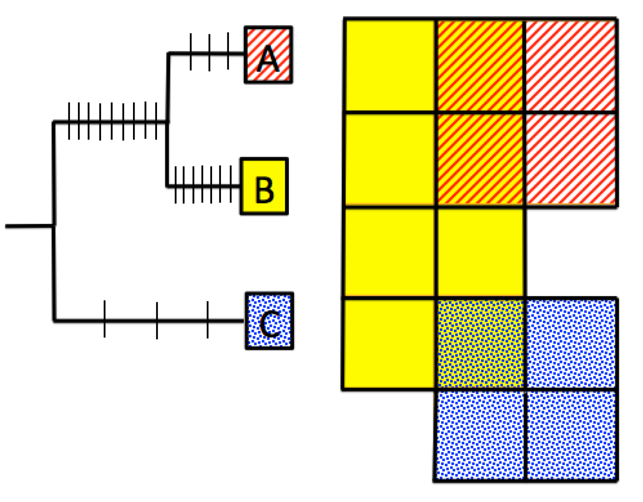
$p(ext)_j$: probability of extinction of leaf *j*



$E(PD) = .2+.3+.5+.75+1.6+ .74+.75+.8425= 5.6825$

If we could halve the p(extinction) of a single species, which should it be? (answer: A)

Can map PD on a landscape

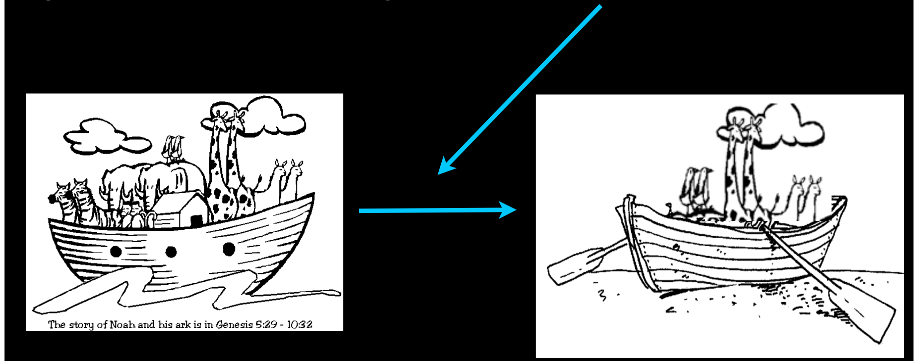


Rosauer et al., 2009

Can also produce the mirror of E(PD) tree (the E(Loss tree) and, e.g., map that on the landscape

Magnuson-Ford et al., 2010
Gudde et al., 2013

Maximizing E(PD) when changing *p(ext)* of leaves at specified costs and budget: "Noah's Ark Problem"*



*Weitzman, 1998: cited >150 times

Noah's Ark Problem formulation

$$E(PD|S) = \sum_{e\in T} \lambda_e \Big(1 - \prod_{j\notin S}^{Ce} p(ext)_j \prod_{k\in S}^{Ce} p(ext)'_k\Big)$$

Find S that maximizes $E(PD)$, given all k members of S now have $p(ext)' < p(ext)$ (are conserved).
Each $delta(p(ext))_k$ has a cost c_k, and $\sum_{k\in S} c_k \le B$, the overall budget.

Again, generally a hard problem
(Hartmann & Steel 2006, Fardi & Goldman 2007, Billionnet 2013)

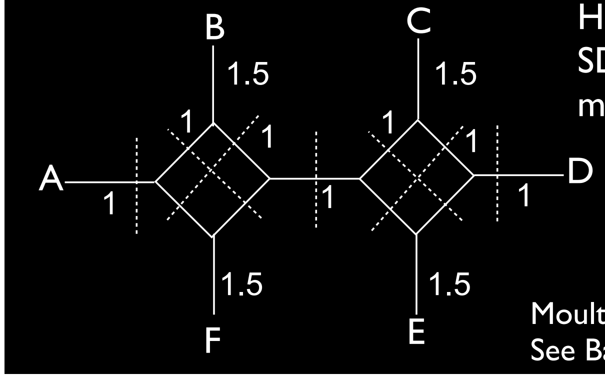
Hartmann & Steel 2006

What about networks?

$$SD_S = \sum_e \lambda_e (X|Y)$$

$X|Y$ is a set of splits;
X is in S, and Y is in S

Consider edges on a network as defining "splits" and sum of splits is total SD. We can visualise with a circular network



Here, for S = {A,D},
SD = 7
max(SD) for k=2

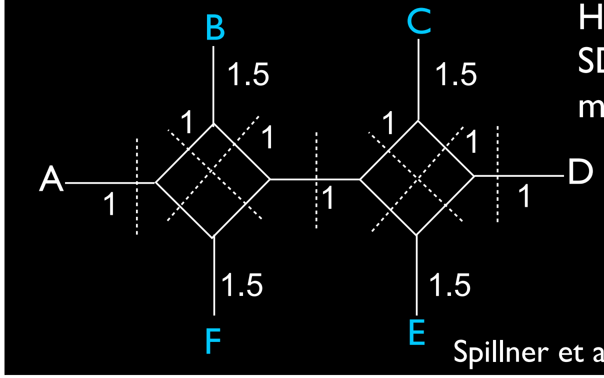
Moulton et al., 2007
See Bastkowski et al. poster

What about networks?

$$SD_S = \sum_e \lambda_e (X|Y)$$

$X|Y$ is a set of splits;
X is in S, and Y is in S

Consider edges on a network as defining "splits" - can visualise with a circular network



Here, for S = {A,D},
SD = 7
max(SD) for k=2

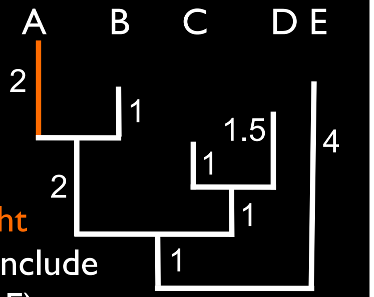
S = {B,C,E,F}
SD = 11
max(SD) for k=4

Moulton et al. 2007
Spillner et al. 2008 Minh et al. 2009

PD = sum of edge weights
Expected(PD) =sum or weighted edge weights
PD complementarity = a leaf's contribution to a tree (ie to the sum of edge weights)

$$PDcomp_x = PD_T - PD_{T-x}$$

T is tree
T-x is subtree without leaf x



This is just the pendant edge weight or PE (for a rooted tree, need to include a zero-length outgroup leaf to get E)

Faith 1992

PD complementarity and evolutionary isolation

One can define other sets that a leaf can complement, e.g. its contribution to the fictitious E(PD) of the tree

$$E(PD)comp_x = E(PD)_{p(ext)_x=0} - E(PD)_{p(ext)_x=1}$$

Christened "heightened evolutionary distinctness" by Steel et al. 2007. This formulation by Faith 2008.

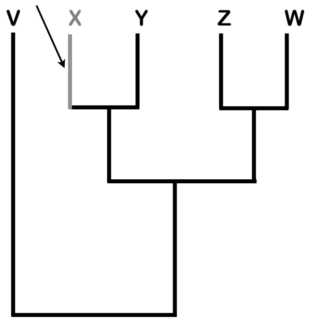
The original motivation for this measure comes from:

1. Ad-hoc measures of "evolutionary isolation"
2. A specific measure of complementarity (Shapley):

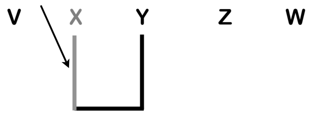
$$\psi_x^{sh} = \underbrace{\frac{1}{n!}\sum_{S\subseteq X, x\subseteq S}(|S|-1)!(|X|-|S|)!}_{\text{Averaged over all equally-weighted subsets!}}\underbrace{(PD_S - PD_{S-x})}_{\text{PD complementarity of } x \text{ to subset } S}$$

Haake et al. 2005, 2008

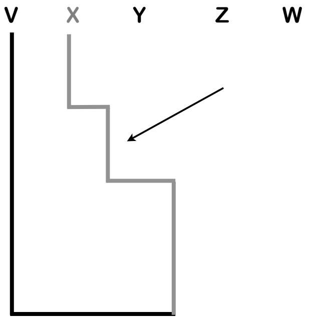
X can complement
S'={*V*,*Y*,*Z*,*W*}



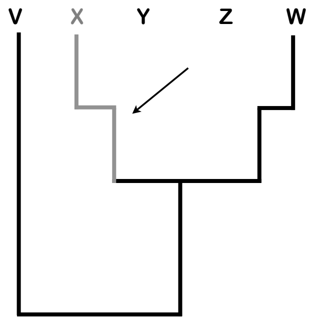
X can complement
S'={*Y*}



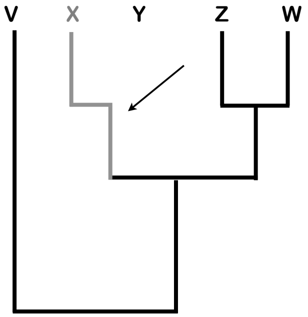
X can complement
S'={*V*}



X can complement
S'={*V*,*W*}



X can complement
S'={*V*,*Z*,*W*}



5 of 12 total possible *S*', *all set sizes* equally likely

The original motivation for considering E(PD)comp$_x$:

1. Ad-hoc measures of "evolutionary isolation"
2. A specific measure of complementarity:

$$FP_x = \sum_{e\in s(T,x,r)}^{r} \frac{\lambda_e}{c_e}$$

s: set of edges from *x* to root *r*
C_e: # leaves descending from *e*

Fair proportion divides the tree up among its leaves

Redding 2003, Redding et al. 2008

EDGE EVOLUTIONARILY DISTINCT & GLOBALLY ENDANGERED

ZSL LIVING CONSERVATION

Home About Species Conservation Blog Get Involved Donate

GET OUR NEWSLETTER

Enter your email address SIGN UP

CONNECT WITH EDGE

Explore the bird tree of life with OneZoom

9,993 species all with their EDGE scores

DISCOVER

Meet some of the World's most extraordinary species on the edge of existence

OUR COMMUNITY

Join an exciting community wanting to raise awareness of EDGE species

TAKE ACTION

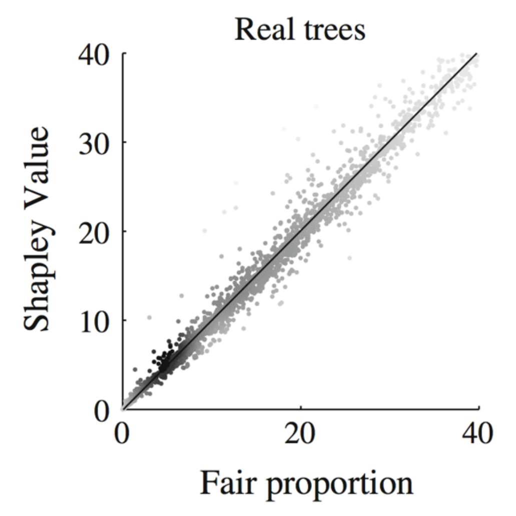
Help support EDGE species today. See how you can help save them from extinction

About EDGE

News & Blogs

Instant WILD

edgeofexistence.org



Shapley and FP are the same in expectation for large *n*, and also in practise

Hartmann 2008, 2012

Can prioritize populations using splits on networks

$$\psi_x^{sh}(T) = \sum_{e\in E} \frac{|\bar{S}_e(x)|}{|X||S_e(x)|}\lambda(e)$$

Can use Shapley if no p(ext) or E(SD)comp$_x$ if we have p(ext)

NeighborNet using F_{ST} based on 8 microsats, 13 pops

Volkmann, (Spillner, Moulton) 2014

$$FP_x = \sum_{e\in s(T,x,r)}^{r} \frac{\lambda_e}{c_e}$$

s: set of edges from *x* to root *r*

C_e: # leaves descending from *e*

FP assigns all of the pendant edge to a leaf, and then some diminishing proportion of "deeper" edges.

FP is weighted towards local tree shape nearest the leaves, and so is strongly correlated with PE

$$FP_x = \sum_{e \in s(T,x,r)}^{r} \frac{\lambda_e}{C_e}$$

s: set of edges from x to root r
C_e: # leaves descending from e

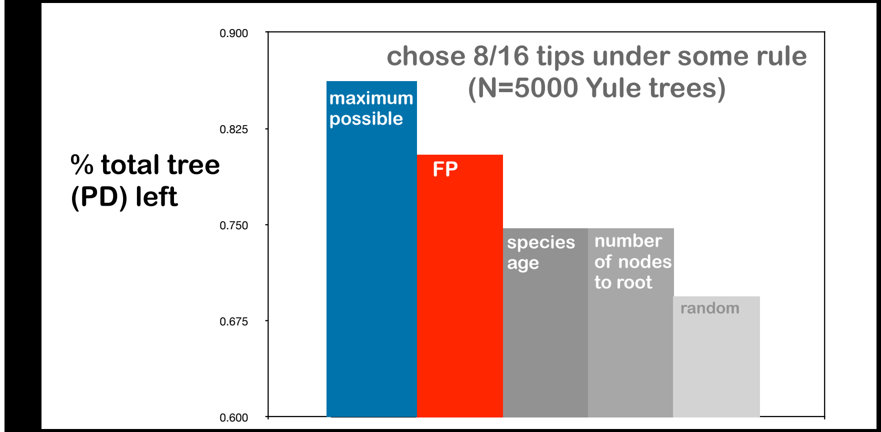
FP assigns all of the pendant edge to a leaf, and then some diminishing proportion of "deeper" edges.

FP is weighted towards local tree shape nearest the leaves, and so is strongly correlated with PE

Quickly asymptotes as we move deeper in the tree and C_e increases. This makes absolute values broadly comparable across large groups (e.g mammals vs. birds vs. plants).

Redding et al submitted

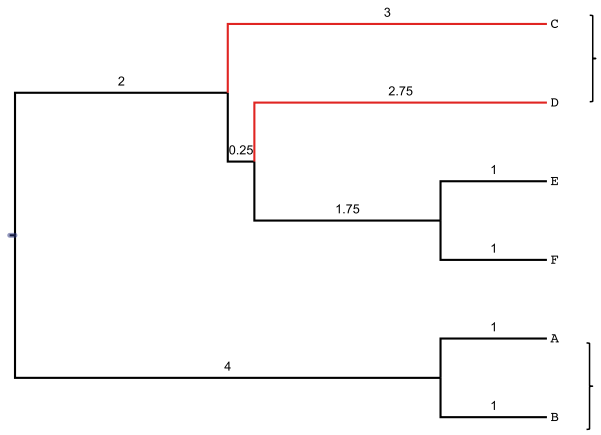
Motivation for identifying isolated species is that they are (i) uniquely identifiable for conservation (ii) sets of isolated species spans much of the tree.



But this assumes p(ext)=1 for all other species!

Redding et al. 2008

Sets of isolated species may have high PD, but not high PDcomp (or high exp(PD)comp

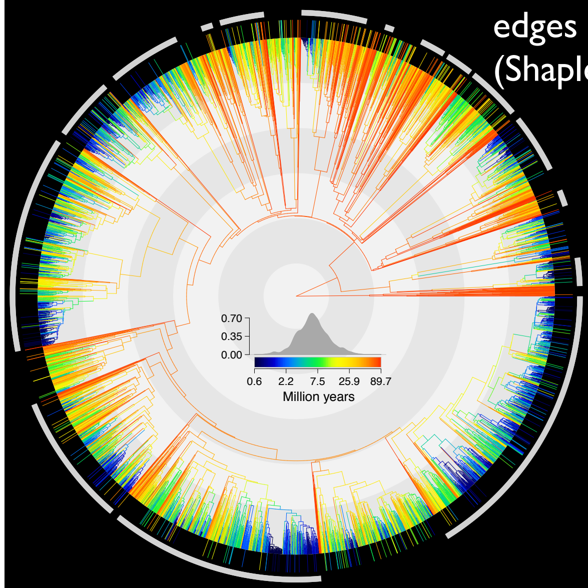


highest FP species sample much of tree but not complementary to rest of tree

This pair offers highest complementarity to rest of tree

Shane McInnis

One tree of all 9993 birds = 77 * 10^9 years



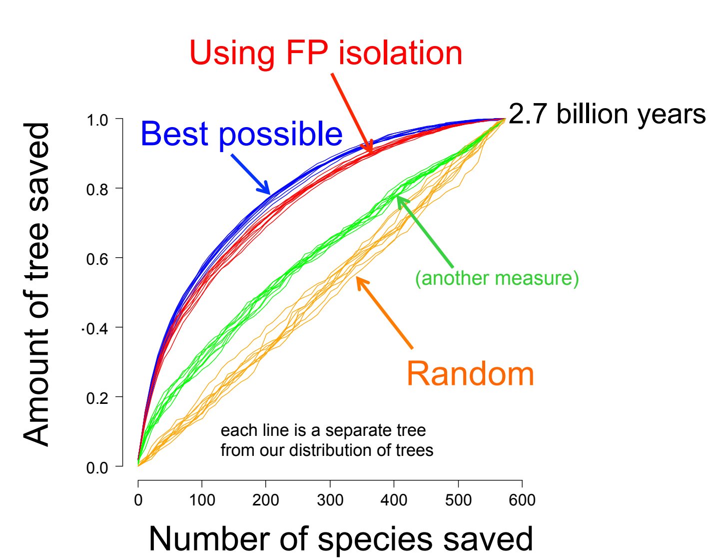
edges coloured by FP score (Shapley complementarity)

575 colored tips are imperiled

= 2.7 * 10^9 years

Jetz et al. 2014

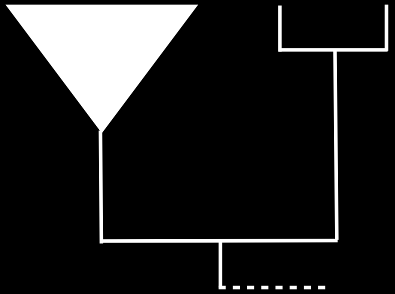
"greedy" E(PD)comp measure chooses near-optimal set



Jetz et al. 2014

For this to be generally the case, high FP species are rarely close relatives.

Under what diversification models is this not rare (Morlon, Etienne)*?



*because it does happen, e.g. tailed frogs

Ok, why are we concerned with a sum of edge weights? (anyway)

Ok, why are we concerned with a sum of edge weights? (anyway)

1. Wilson's exhortation to preserve "information"
2. Practical extension to ecosystem services

David Tilman's Cedar Creek plots

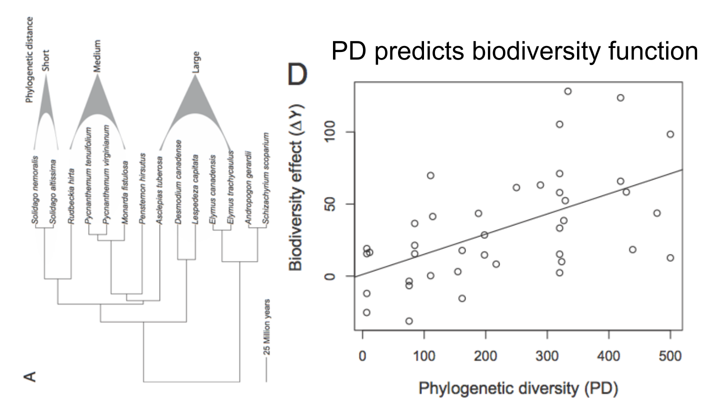
A practical reason to conserve PD (sum of edge weights)

1. Across 29 global experiments, PD > 20 AIC units better predictor of productivity than species richness or functional group representation.

2. Across a set of famous experimental plots testing biodiversity and ecosystem functioning, PD > 10 AIC units better predictor of productivity than 13 other measures, including species richness.

3. Experiment designed to test whether PD predicts productivity confirm these other analyses.

Cadotte et al. 2008, 2009; others; Cadotte 2013

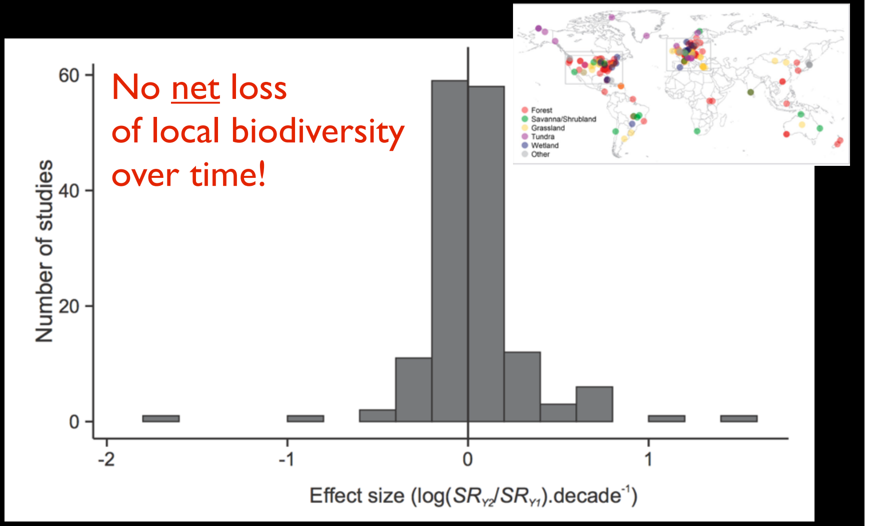


Cadotte 2013 PNAS

"The ecological consequences of biodiversity loss can be predicted from evolutionary history."

Emerging trend 4 of 4, Cardinale et al. 2012 Nature

Most controversial conservation paper of 2013?

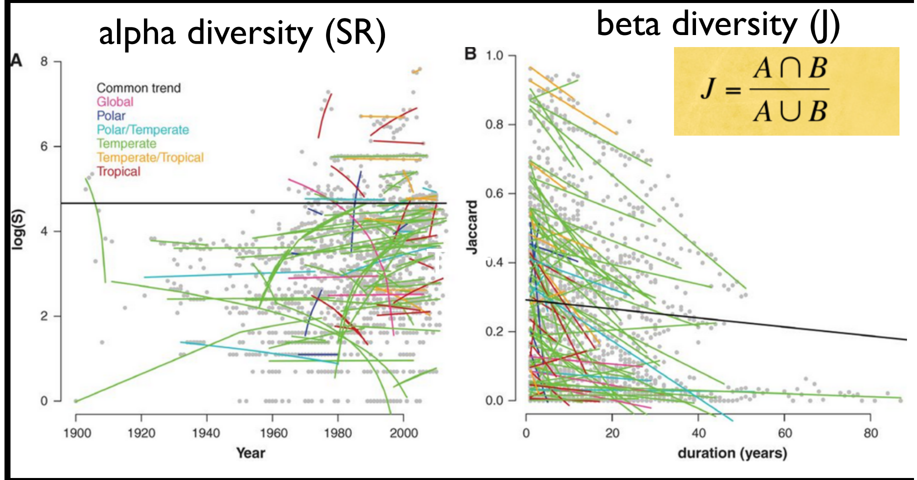

No <u>net</u> loss
of local biodiversity
over time!

Vellend et al. 2013 PNAS

Controversial conservation paper from 2014

alpha diversity (SR)

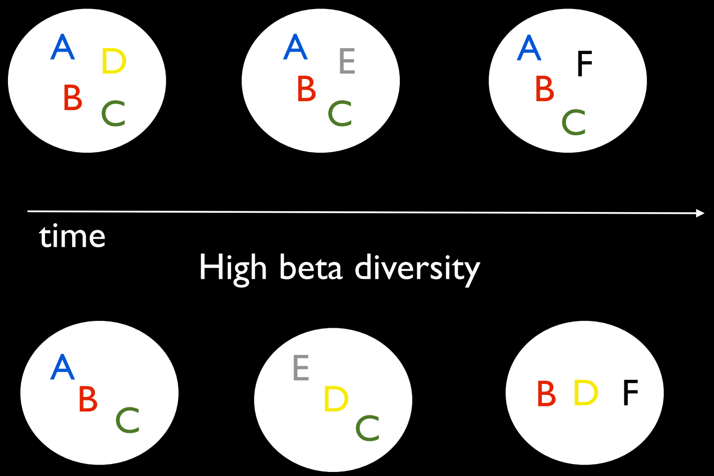
beta diversity (J)



$$J = \frac{A \cap B}{A \cup B}$$

Domelas et al. 2014 Science

Low beta diversity



High beta diversity

Dimensions of Phylogenetic Biodiversity

Dimension Scale	Richness: How much Sum of...	Divergence: How different Mean of...	Regularity: How regular Variance of...
Diversity of sample(s) [alpha]	edge weights (PD)	mean(patr. dist.) mean(evol. isol.);	var(patr. dist.); tree shape, I_c
Differences among samples [beta]	sum of unshared edgeweights 'Unifrac'	?	?

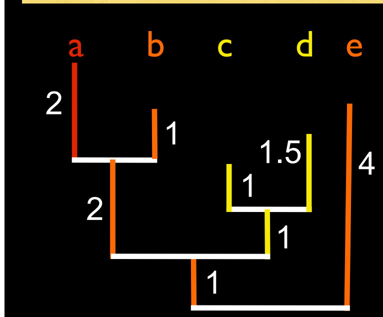
following Pavoine & Bonsall, 2011, sPhy

This suggests conservation biologists might consider effects of "beta diversity" (e.g. between samples A & B)*

*various speakers, CSEE meeting 2014

$$Unifrac = \frac{2 * PD_{A \cup B} - PD_A - PD_B}{PD_{A \cup B}}$$

proportion of entire tree found in only one sample {0,1} (=1-J)



Is this the best measure? does not, e.g. differentiate how shared edges are organized

Luzupone & Knight 2005
also Ferrier et al. 2007
Bryant (Morlon) et al. 2008

Dimensions of Phylogenetic Biodiversity

Dimension Scale	Richness: How much? Sum of...	Divergence: How different? Mean of...	Regularity: How regular? Variance of...
Diversity of sample(s) [alpha]	edge weights (PD)	mean(patr. dist.) mean(evol. isol.);	var(patr. dist.); tree shape, I_c
Differences among samples [beta]	sum of unshared edgeweights 'Unifrac'	?	?

?

following Pavoine & Bonsall, 2011, sPhy

Broad Recipe

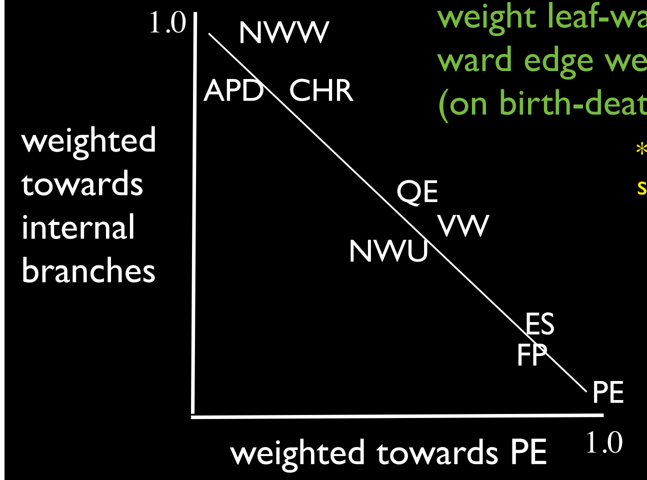
1. Dimensions of biodiversity: $\alpha, \beta, \sum, \bar{x}, \mathrm{var}(x)$
2. Sum of edge weights on trees
3. Isolation as a surrogate for maximizing expected PD
4. Other dimensions of the tree and conservation
5. What does an edge weight represent?

All the dimensions and scales use edge weights

"...underlying feature diversity of taxon subsets...can be predicted by...phylogenetic relationships among the taxa." (Faith 1992)

for discussion, see, e.g. Diniz-Filho et al. 2013

Observation: there are 13 published measures of evolutionary isolation (some redundant like FP&Shapley)



Acronyms don't matter; weight leaf-ward and root-ward edge weights differently (on birth-death trees*)

*so, could use some maths...

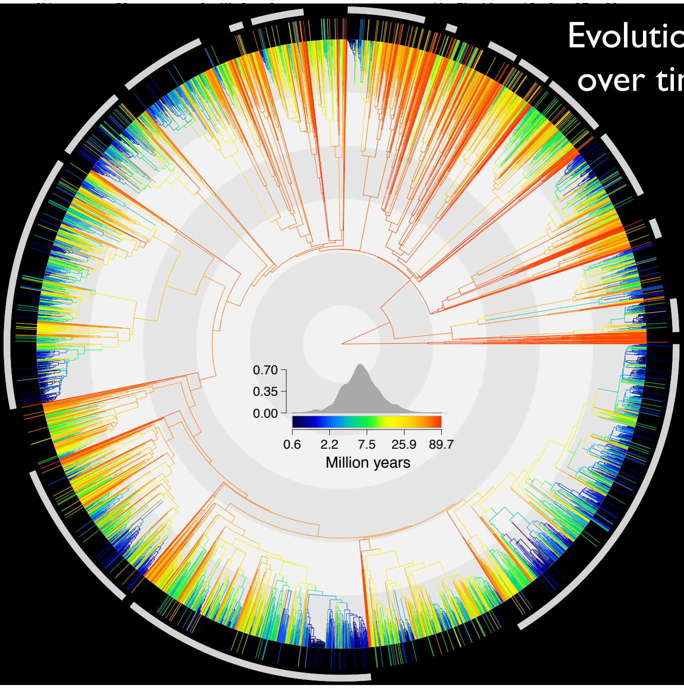
Redding et al. submitted

Obviously, the score you use matters...top EDGE spp

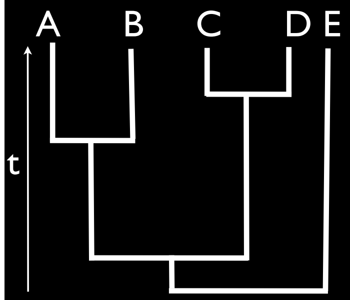
	Mammals (n=4920)			Amphibians (n=5713)		
Metric	Shared Species	Rank Similarity	Unique Scores	Shared Species	Rank Similarity	Unique Scores
ED/FP	Top 100 species		2326	Top 100 species		1271
SV	100	1	2327	97	1	1271
ES	79	0.564	2291	92	0.907	1270
PE	78	0.667	544	90	0.894	1140
MVW	55	0.401	124	79	0.757	139
VW	55	0.401	124	79	0.757	139
QE	49	0.51	2055	59	0.435	1367
APD	47	0.447	2310	57	0.528	1271
CHR	47	0.44	2339	57	0.594	1271
NWW	44	0.146	419	57	0.515	368
NWU	53	0.237	119	41	0.494	148
Genus	50	0.42	187	15	0.667	278

Redding et al. submitted

Evolution happens over time

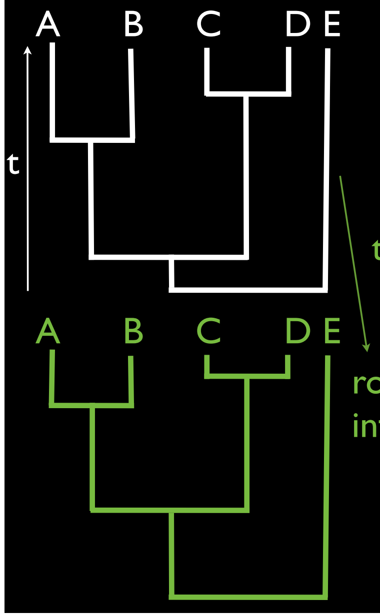


"Actual" phylogeny



This tree will capture "feature diversity" under one very common model: if many features evolve under independent Brownian motion processes.

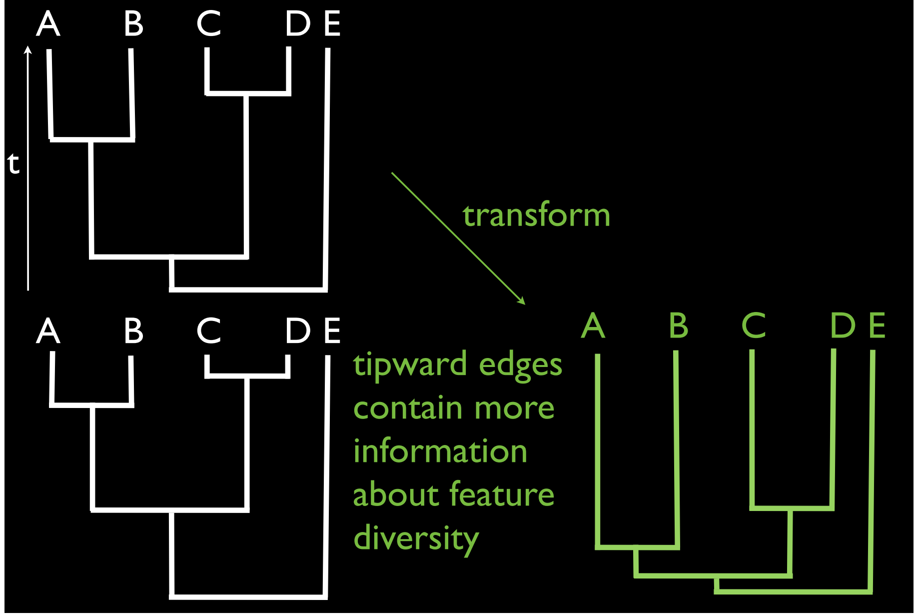
"Actual" phylogeny



transform

rootward edges contain more information on feature diversity

"Actual" phylogeny

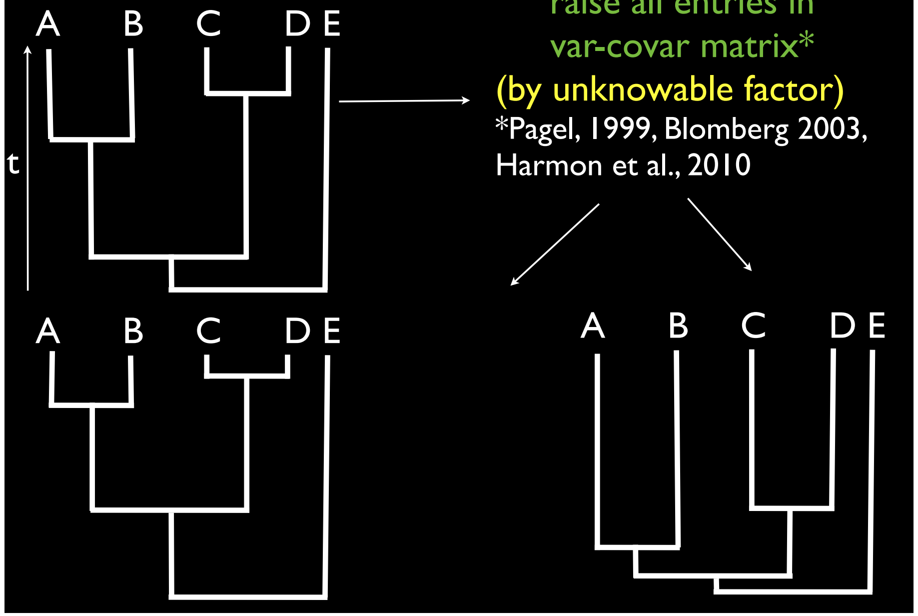


transform

tipward edges contain more information about feature diversity

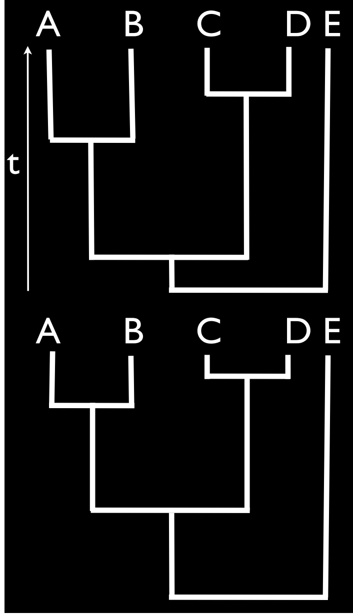
---

"Actual" phylogeny

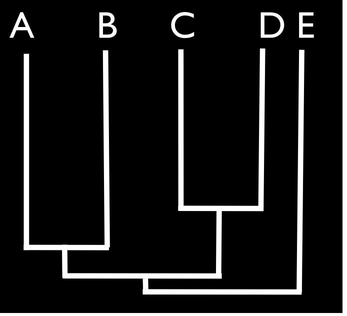


raise all entries in var-covar matrix*
(by unknowable factor)
*Pagel, 1999, Blomberg 2003, Harmon et al., 2010

"Actual" phylogeny

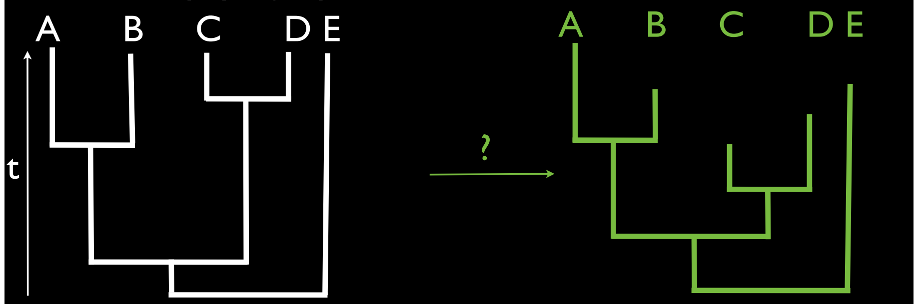


All PD calculations flow directly, and several problems go from hard to less hard.



"Actual" phylogeny

Additive tree

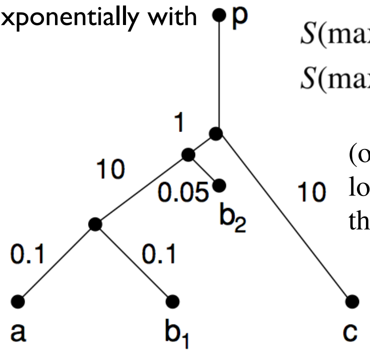


idiosyncratic transformation

This is the general case (additive tree) considered by Faith and others since. Because the edge lengths are idiosyncratic, there is no "transformation" possible. Usually, single or few traits used to produce additive tree

Under particular models of feature evolution on additive trees, PD will not capture maximum feature diversity

features arise linearly with time and are lost exponentially with rate lambda



$$S(\max PD_{|S|=3}) = \{a, b_1, c\}$$

$$S(\max(features)_{|S|=3} = \{a, b_2, c\}$$

(on this tree, if characters are lost a rate = 2000x slower than they arise up to 10x faster)

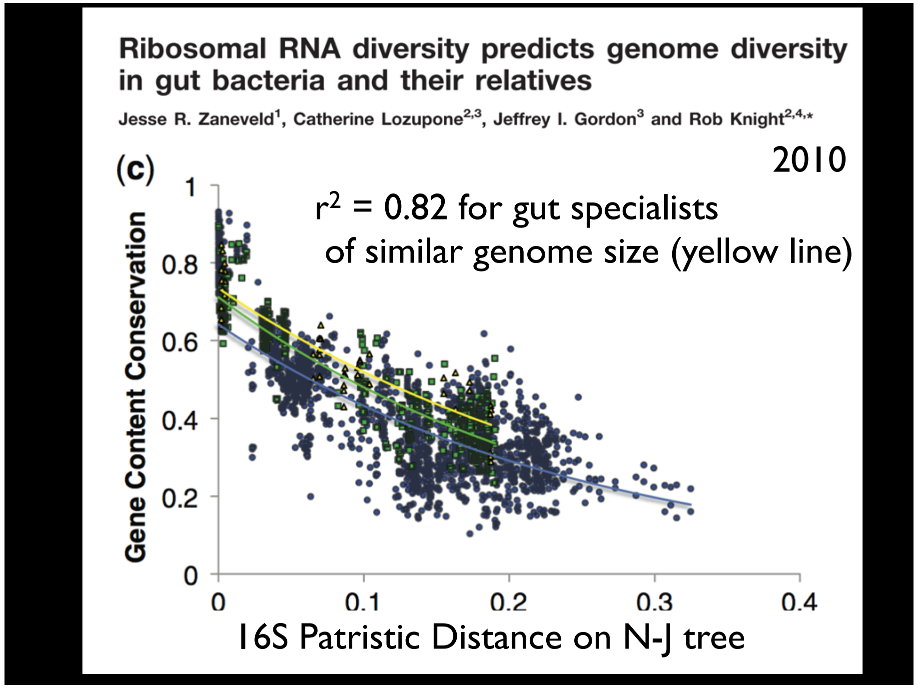
$$M(S) = \min\{PD_{x,y} : x, y \in S\}$$

maximizing $M(S)$ for $|S| = k$ as alternate function

Bordewich et al. 2008 (see also Moulton et al. 2007)

FIGURE 1. Reproduction of Woese's (Woese, 1987) small-subunit ribosomal RNA tree showing the subtree subtended by three EUs chosen by (a) minimizing PD and by (b) maximizing the minimum distance. We constructed this tree using small-subunit ribosomal RNA sequences

Bordewich's motivating tree, based on 16s RNA sequences

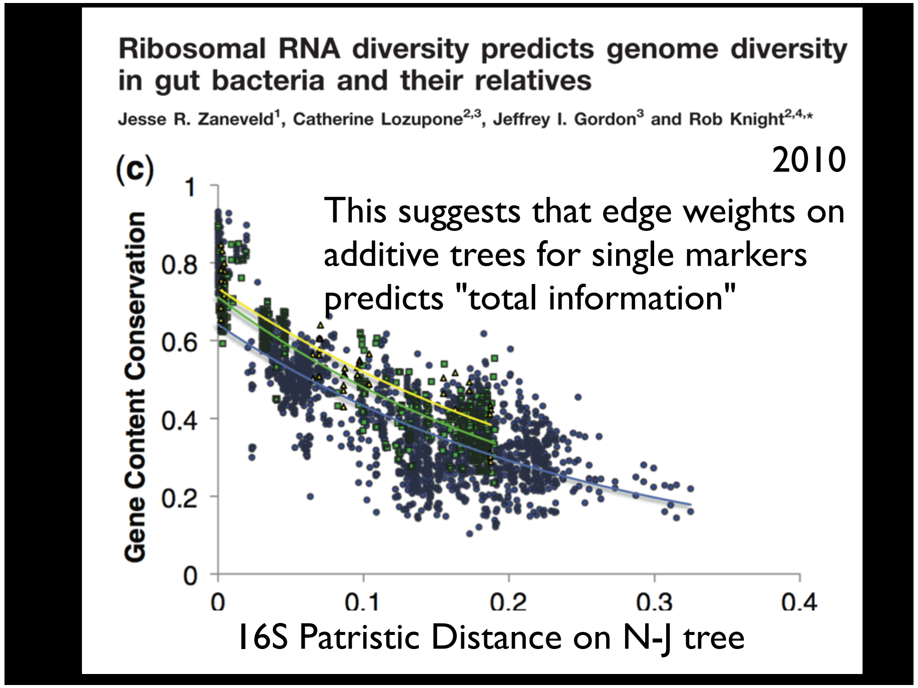


Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives

Jesse R. Zaneveld[1], Catherine Lozupone[2,3], Jeffrey I. Gordon[3] and Rob Knight[2,4,*]

2010

$r^2 = 0.82$ for gut specialists of similar genome size (yellow line)



Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives

Jesse R. Zaneveld[1], Catherine Lozupone[2,3], Jeffrey I. Gordon[3] and Rob Knight[2,4,*]

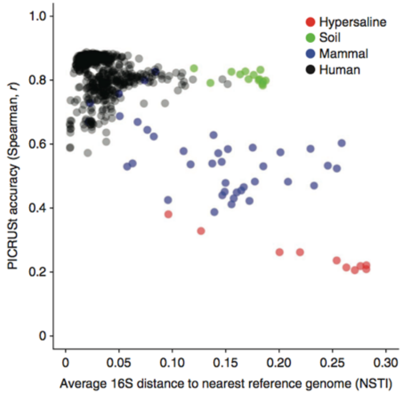
2010

This suggests that edge weights on additive trees for single markers predicts "total information"

Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences

Morgan G I Langille[1,14], Jesse Zaneveld[2,14], J Gregory Caporaso[3,4], Daniel McDonald[5,6], Dan Knights[7,8], Joshua A Reyes[9], Jose C Clemente[10], Deron E Burkepile[11], Rebecca L Vega Thurber[2], Rob Knight[10,12], Robert G Beiko[1] & Curtis Huttenhower[9,13]

VOLUME 31 NUMBER 9 SEPTEMBER 2013 NATURE BIOTECHNOLOGY



Across a broader array of communities, 16S genetic distance predicts (with suitable preparation of data) functional components of genomes

Is there *any* way to measure the shared and unique "total information content" of a genome* (or feature diversity of an organism)?

With this, we could see how well additive (or, better-still, ultrametric) trees (or splits on networks) predicted the difference in information/features between leaves.

*using, e.g. compression algorithms? (I. Martyn, pers. comm.)