A Comparative Analysis with Oxalis

Bret Larget

Department of Statistics
University of Wisconsin—Madison

April 19, 2011

Introduction

- Summary of comparative analysis of *Oxalis* data;
- Lessons in using R for comparative analysis;
- Bayesian approach to address tree uncertainty.

Reading the Data

```
> library(ape)
> library(picante)
> library(phylobase)
> library(lattice)
> oxalis.dat = read.csv("oxalis-data.csv")
> rownames(oxalis.dat) = oxalis.dat$species
> oxalis.dat = oxalis.dat[, -1]
```

The data as a table

```
> options(width = 120)
> oxalis.dat
```

	alt	precip	seasonality	scales.mm	lat	abslat	region
OadenophyllaADEPH1	2081	785	72	40	-40.1	40.1	Basal
ObrasiliensisBRAS2	79	1097	17	7	-30.5	30.5	SESouthAmerican
OperdicariaMV79	275	1262	63	64	-30.7	30.7	SESouthAmerican
OdebilisEE171	684	1613	62	10	-20.4	20.4	SESouthAmerican
OhispidulaMV44MV342	79	1211	31	8	-28.2	28.2	SESouthAmerican
ObipartitaMV59MV320	803	1510	28	5	-27.0	27.0	SESouthAmerican
OoreocharisEE583	3644	1154	48	5	-7.5	7.5	Andean
OtriangularisssppapilionaceaREG	1000	800	20	1	-20.0	20.0	Andean
OtrolliiEE281	3617	596	85	12	-18.6	18.6	Andean
OmacrocarpaAG49	1365	972	105	60	20.3	20.3	Andean
OpinguiculaceaWood22147	3691	718	78	34	-15.7	15.7	Andean
OalpinaSW976	2362	751	61	7	26.9	26.9	NorthAmerican
OdiscolorAG35	1441	820	97	8	18.7	18.7	NorthAmerican
OlatifoliaAG70	2363	1351	76	55	17.3	17.3	NorthAmerican
OlatifoliaEE756C	1897	580	76	13	-12.0	12.0	NorthAmerican
OlasiandraAG69	1965	782	90	18	17.2	17.2	NorthAmerican
OtetraphyllaTETRA2	1391	1270	101	35	18.3	18.3	NorthAmerican
OnelsoniiAG66	1824	1002	93	15	17.7	17.7	NorthAmerican
OcaeruleaAG47	2049	525	78	20	29.3	29.3	NorthAmerican
OprimaveraAG50	1409	954	110	17	20.8	20.8	NorthAmerican
OhernandesiiAG56	2055	859	98	52	20.5	20.5	NorthAmerican
OdivergensAG62	2057	1002	97	24	19.9	19.9	NorthAmerican
OdecaphyllaAR2610A	1872	792	100	26	20.7	20.7	NorthAmerican
OmorelosiiEPEREZ4856	2344	1081	95	30	19.6	19.6	NorthAmerican

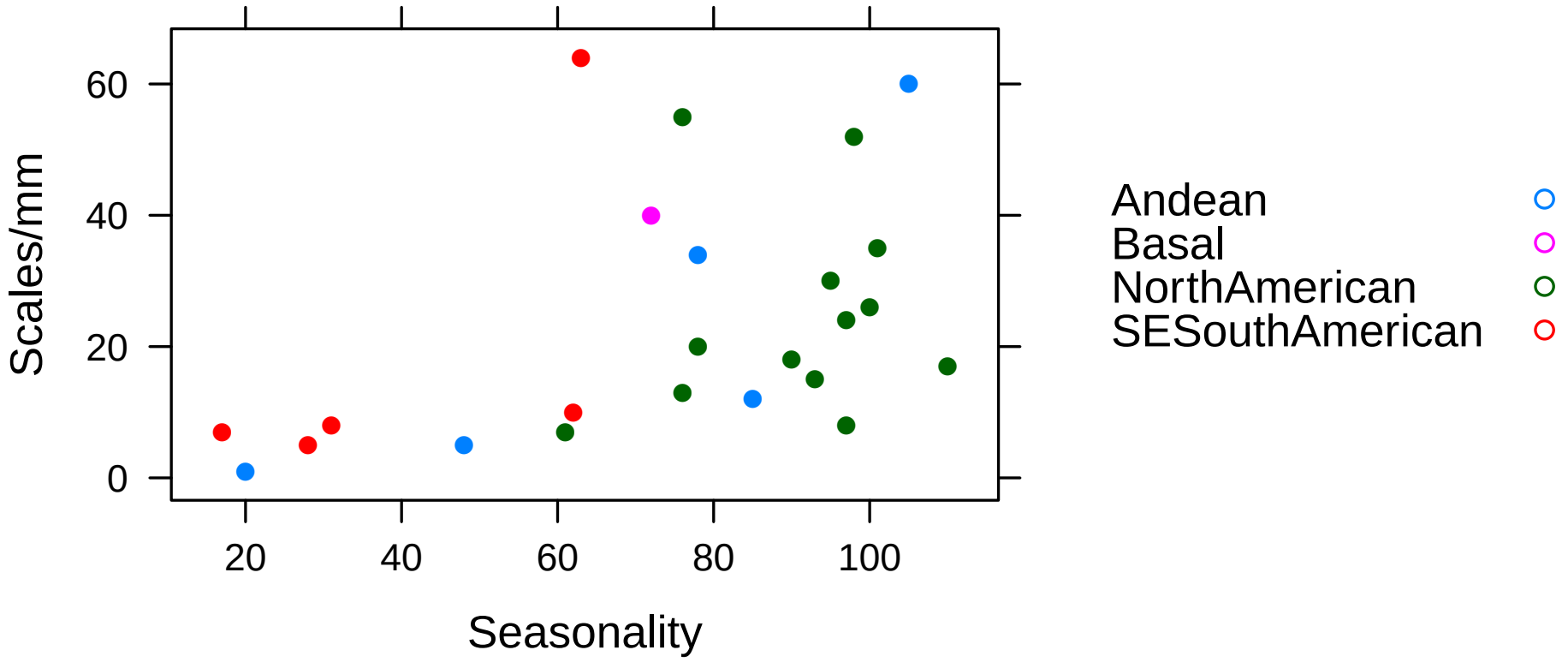
Examining Correlation

```
> print(round(cor(oxalis.dat[, -7]), 2))
```

	alt	precip	seasonality	scales.mm	lat	abslat
alt	1.00	-0.48	0.38	0.04	0.28	-0.49
precip	-0.48	1.00	-0.30	0.09	-0.25	0.05
seasonality	0.38	-0.30	1.00	0.46	0.71	-0.30
scales.mm	0.04	0.09	0.46	1.00	0.16	0.15
lat	0.28	-0.25	0.71	0.16	1.00	-0.36
abslat	-0.49	0.05	-0.30	0.15	-0.36	1.00

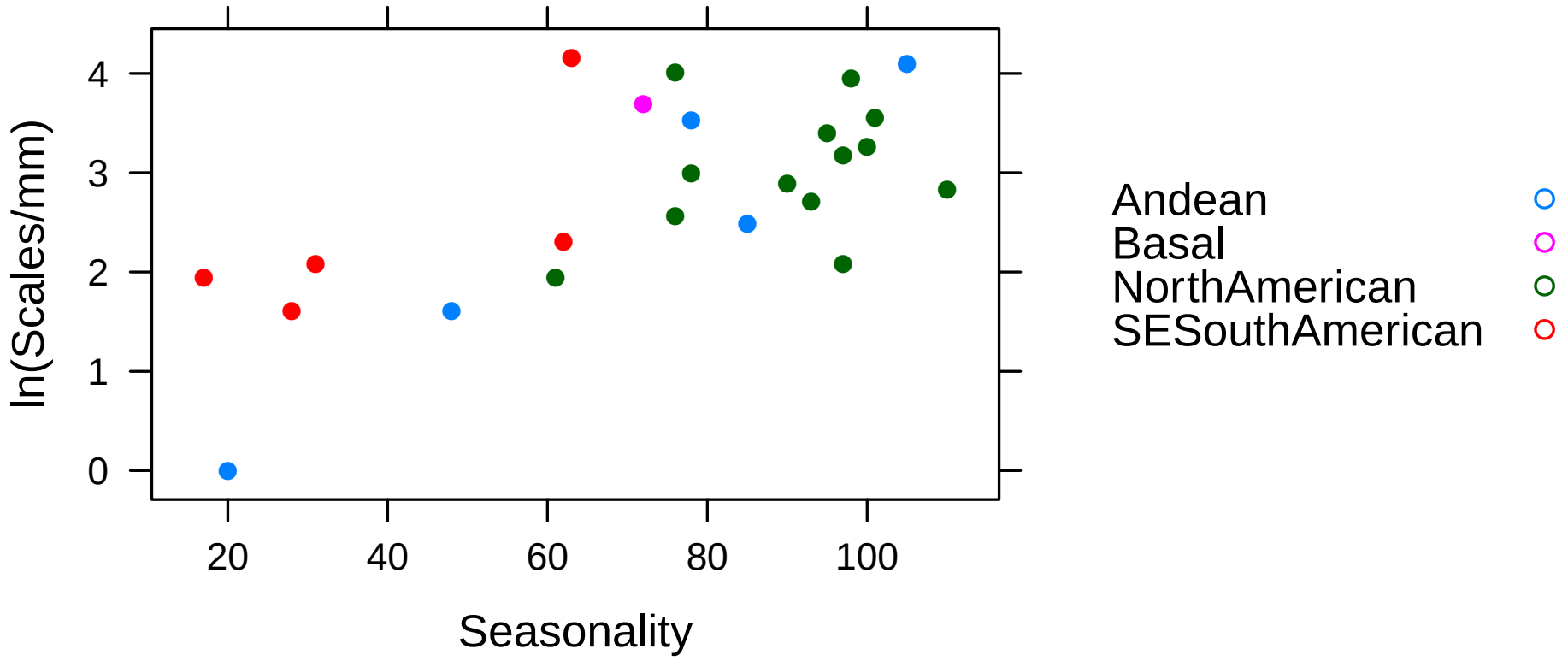
Plot of scales/mm versus seasonality

```
> plot(xyplot(scales.mm ~ seasonality, data = oxalis.dat,
+      pch = 16, xlab = "Seasonality", ylab = "Scales/mm",
+      groups = region, auto.key = list(space = "right")))
```



Plot of log-transformed data

```
> plot(xyplot(log(scales.mm) ~ seasonality, data = oxalis.dat,
+      pch = 16, xlab = "Seasonality", ylab = "ln(Scales/mm)",
+      groups = region, auto.key = list(space = "right")))
```



Regression Ignoring Phylogeny

```
> fit.0 = lm(log(scales.mm) ~ seasonality, data = oxalis.dat)
> print(summary(fit.0))

Call:
lm(formula = log(scales.mm) ~ seasonality, data = oxalis.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4892 -0.5252 -0.1082  0.5392  1.6408

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.010727   0.444698   2.273 0.033155 *
seasonality 0.023926   0.005626   4.252 0.000326 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7498 on 22 degrees of freedom
Multiple R-squared: 0.4511, Adjusted R-squared: 0.4262
F-statistic: 18.08 on 1 and 22 DF,  p-value: 0.0003258
```

Finding a Tree

- Used DNA sequence data (a plastid sequence and an ITS sequence) and an analysis in MrBayes to sample trees assuming:
  - Separate GTR models with different rates and gamma-distributed rate variation among sites for each part of the data partition;
  - A clock tree with a uniform prior distribution;
  - Trees are proportional for each part of the data partition.
- MCMC:
  - Two independent runs, each with two chains (one heated);
  - 600,000 generations per run;
  - Discard first 100,000 generations;
  - Sample every 100th tree from last 500,000 generations;
  - Total sample of 10,000 trees from both runs.

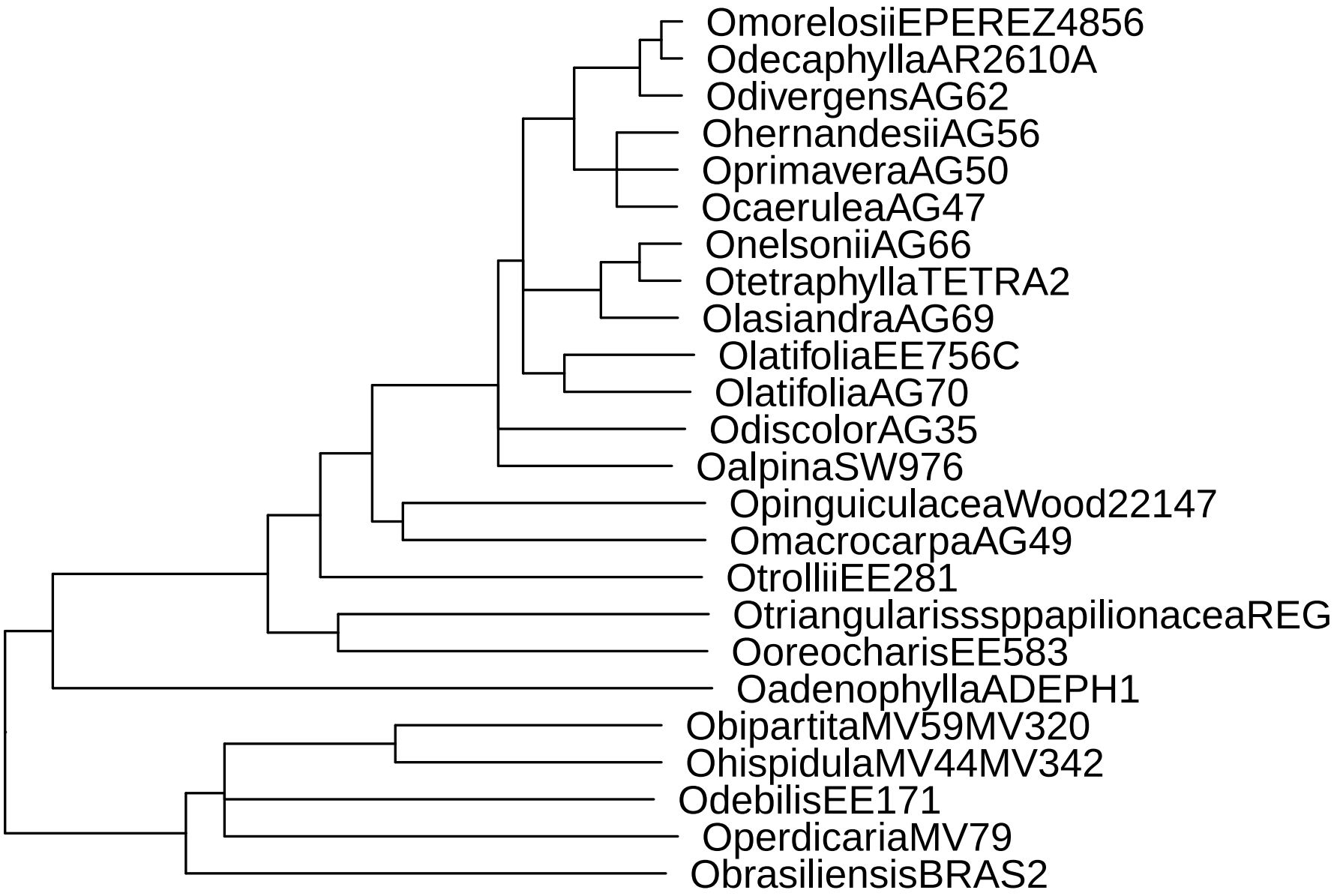
Consensus Tree

- Even though all individual sampled trees are ultrametric, the majority rule consensus tree need not be.
- Examine analysis using only this tree.
- Compare results with sample of ten trees to see robustness to tree uncertainty.
- Note that ran MCMC with 24 species with tip data, but more taxa were available with sequence data.
- In retrospect, it would have been better to use trees from larger data set, pruned to those with tip data.
- The extra outgroup data are helpful to better root the ingroup.

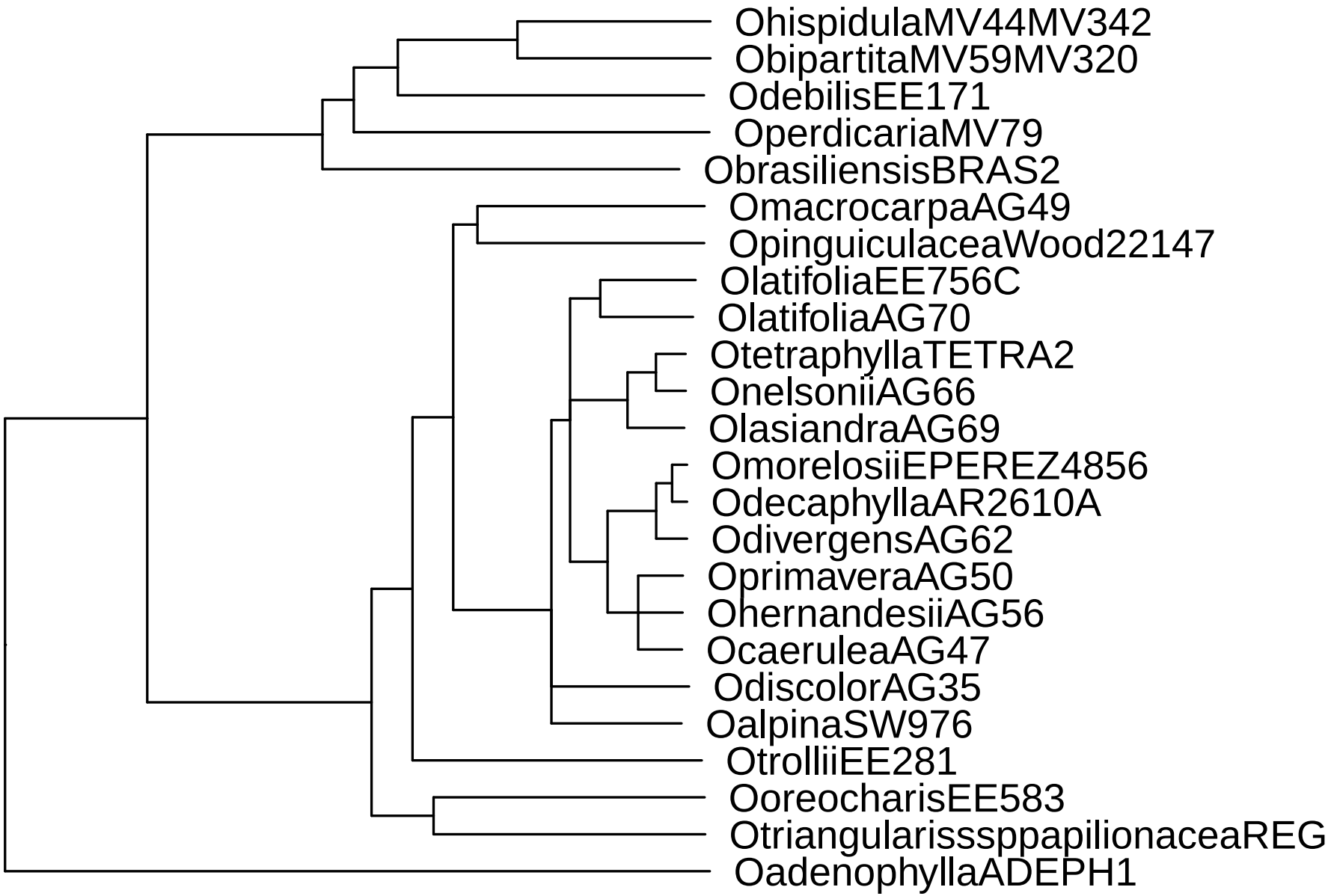
Read in the Trees

```
> trees = read.tree("reduced-sample-names.tre")
> consensus = read.tree("reduced.con")
> full = read.tree("oxalis.tre")
> full.phy = phylo4(full)
> full.phy.pruned = prune(full.phy, tips.exclude = c(1:11,
+     36:37))
```

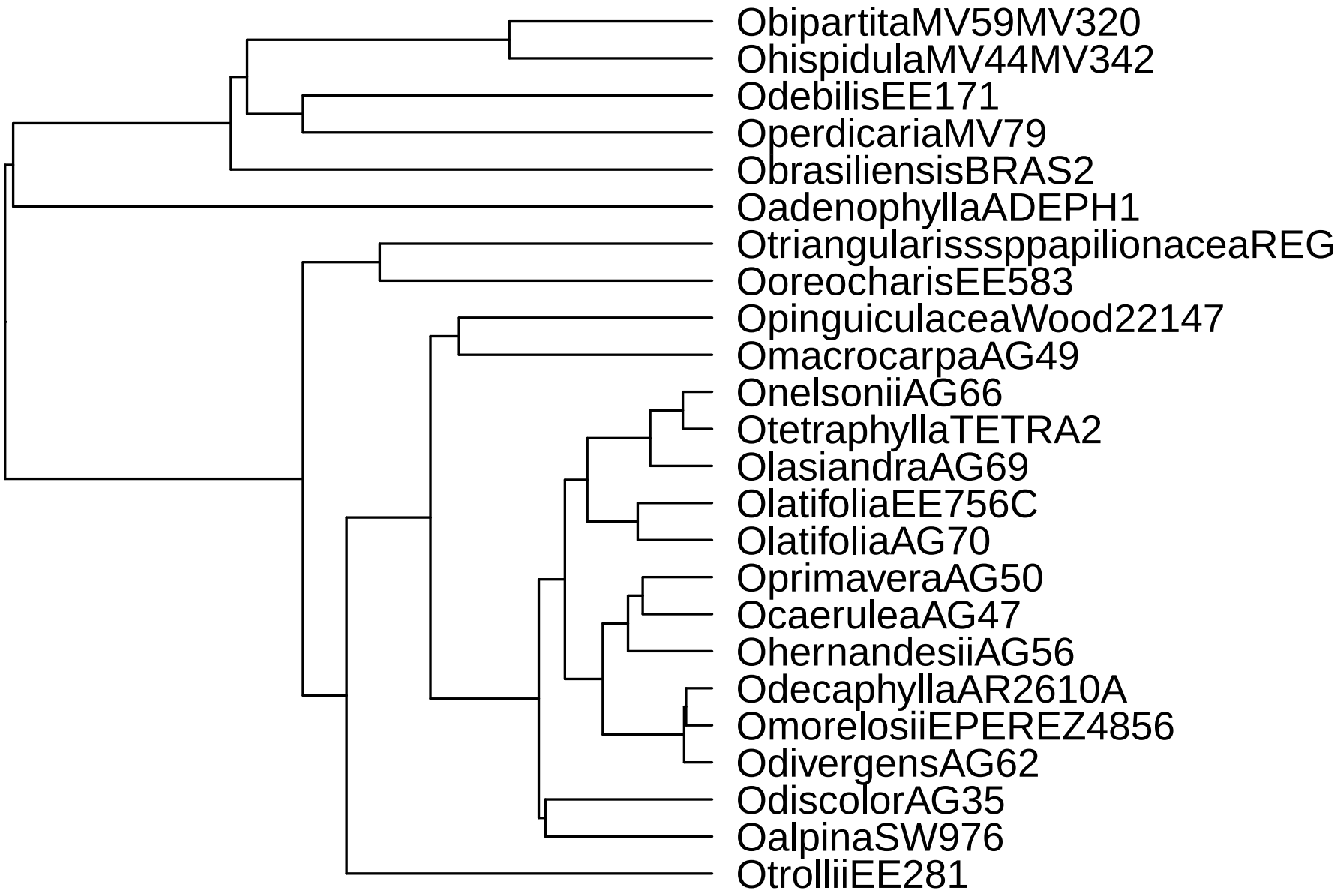
Tree from Reduced Data Set



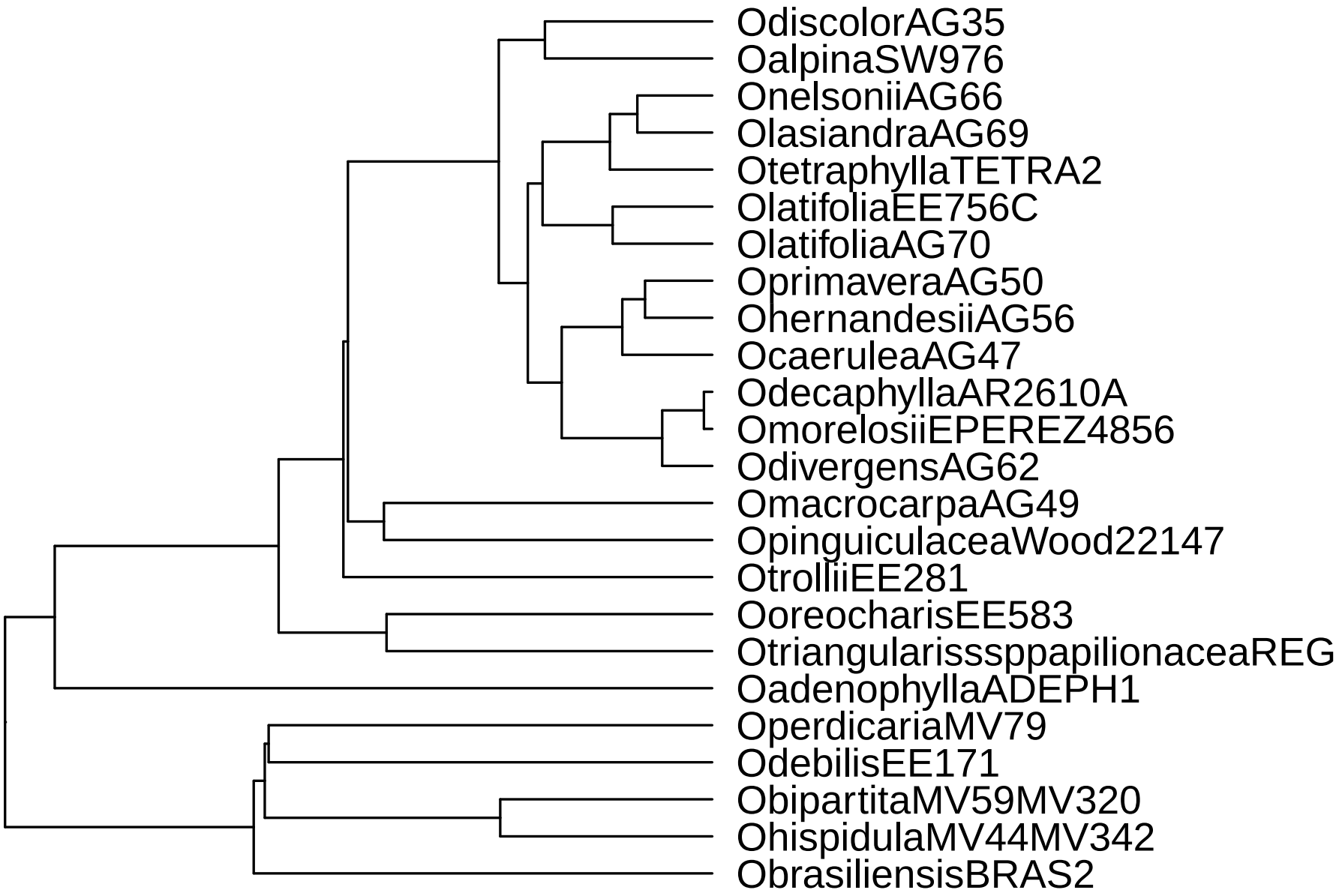
Pruned Tree from Full Data Set



One Sampled Tree



A Second Sampled Tree

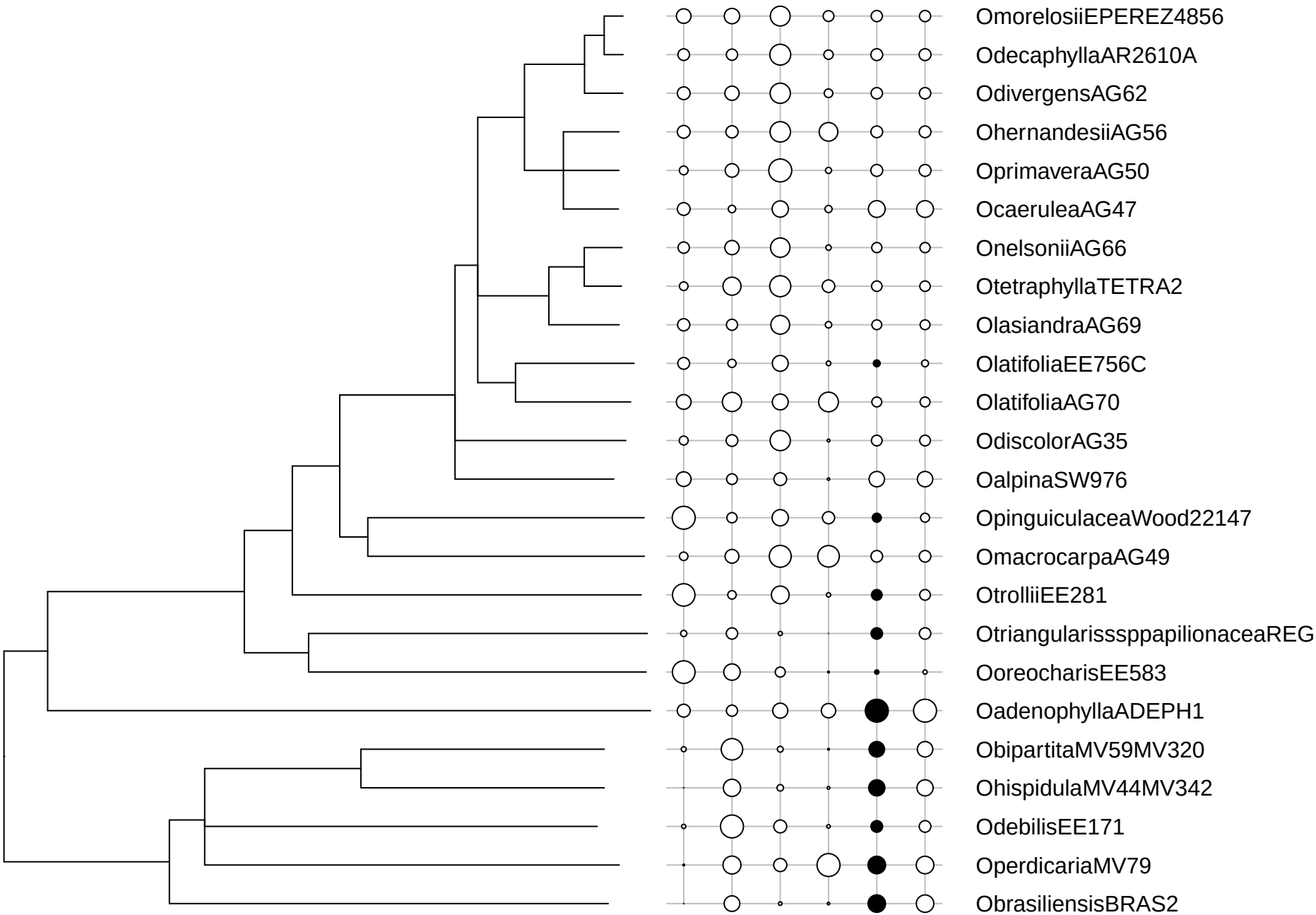


Bubble Plot of Data with Phylogeny

- The following R code shows the tree with bubbles for the data for each variable.
- Nice!

```
> happy.0 = match.phylo.data(consensus, oxalis.dat[,
+     -7])
> bub = phylo4d(happy.0$phy, happy.0$data)
> plot(bub)
```

Bubble Plot


OmorelosiiEPEREZ4856
OdecaphyllaAR2610A
OdivergensAG62
OhernandesiiAG56
OprimaveraAG50
OcaeruleaAG47
OnelsoniiAG66
OtetraphyllaTETRA2
OlasiandraAG69
OlatifoliaEE756C
OlatifoliaAG70
OdiscolorAG35
OalpinaSW976
OpinguiculaceaWood22147
OmacrocarpaAG49
OtrolliiEE281
OtriangularisssppapilionaceaREG
OoreocharisEE583
OadenophyllaADEPH1
ObipartitaMV59MV320
OhispidulaMV44MV342
OdebilisEE171
OperdicariaMV79
ObrasiliensisBRAS2

Do the regression for the consensus tree

```
> happy = match.phylo.data(consensus, oxalis.dat[, -7])
> fit.1 = gls(log(scales.mm) ~ seasonality, data = happy$data,
+     correlation = corBrownian(1, happy$phy))
```

Examine Summary

```
> summary(fit.1)

Generalized least squares fit by REML
  Model: log(scales.mm) ~ seasonality
  Data: happy$data
       AIC      BIC    logLik
  68.96661 72.23974 -31.48331

Correlation Structure: corBrownian
 Formula: ~1
 Parameter estimate(s):
numeric(0)

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 1.3146924 0.7793614 1.686884  0.1058
seasonality 0.0227868 0.0089614 2.542770  0.0185

 Correlation:
            (Intr)
seasonality -0.645

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.3473804 -0.5586487 -0.2582067  0.2128215  1.0720311

Residual standard error: 1.313978
Degrees of freedom: 24 total; 22 residual
```

Regressions for all ten trees

```
> fits = list()
> for (i in 1:10) {
+     happy = match.phylo.data(trees[[i]], oxalis.dat[, -7])
+     fit = gls(log(scales.mm) ~ seasonality, data = happy$data
+         correlation = corBrownian(1, happy$phy))
+     fits[[length(fits) + 1]] = fit
+ }
```

Summary of one analysis

```
> print(summary(fits[[1]]))

Generalized least squares fit by REML
  Model: log(scales.mm) ~ seasonality
  Data: happy$data
       AIC      BIC    logLik
  69.22243 72.49556 -31.61122

Correlation Structure: corBrownian
 Formula: ~1
 Parameter estimate(s):
numeric(0)

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 1.2425994 0.8318035 1.493862  0.1494
seasonality 0.0239182 0.0093493 2.558290  0.0179

 Correlation:
            (Intr)
seasonality -0.639

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.2307329 -0.5410106 -0.2428587  0.2199854  1.0079491

Residual standard error: 1.398324
Degrees of freedom: 24 total; 22 residual
```

Collect all estimates

```
> slopes = rep(NA, 10)
> ses = rep(NA, 10)
> p.values = rep(NA, 10)
> for (i in 1:10) {
+     tab = summary(fits[[i]])$tTable
+     slopes[i] = tab[2, 1]
+     ses[i] = tab[2, 2]
+     p.values[i] = tab[2, 4]
+ }
```

Summary

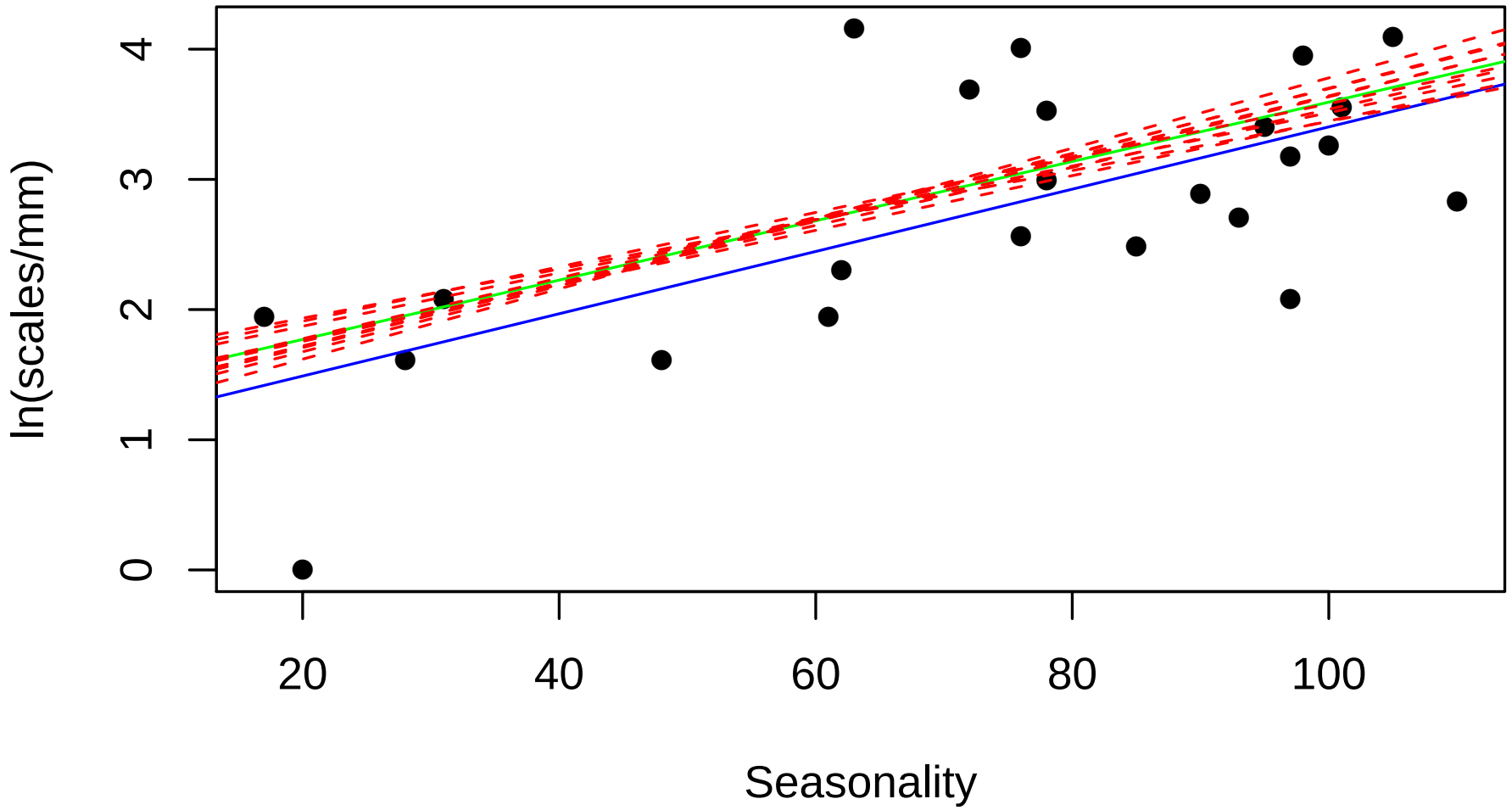
```
> print(cbind(slopes, ses, p.values))
```

```
           slopes          ses    p.values
 [1,] 0.02391817 0.009349280 0.017922967
 [2,] 0.02050531 0.009153915 0.035514849
 [3,] 0.02228628 0.009410289 0.027076348
 [4,] 0.02530943 0.009500933 0.014179973
 [5,] 0.02485925 0.009057706 0.011831342
 [6,] 0.02099803 0.009325286 0.034657461
 [7,] 0.02088805 0.008742418 0.025885626
 [8,] 0.02332012 0.009104641 0.017802804
 [9,] 0.02701006 0.009536820 0.009698778
[10,] 0.01890160 0.008728548 0.041463919
```

Plotted Regression Lines

```
> with(oxalis.dat, plot(seasonality, log(scales.mm), pch = 16,
+     xlab = "Seasonality", ylab = "ln(scales/mm)"))
> abline(fit.0, col = "blue")
> abline(fit.1, col = "green")
> for (i in 1:10) {
+     abline(fits[[i]], col = "red", lty = 2)
+ }
```

Plotted Regression Lines



Summary

- There are many people developing good (and bad) software packages for comparative analyses using R.
- The package *phylobase* is especially good.
- The package *ape* is essential.
- It is not difficult (and, in fact, I find it easier) to incorporate tree uncertainty into a comparative analysis by examining results from several trees from the posterior distribution.
- The average slope estimate is nearly identical to the single estimate from the consensus tree.
- The average p-values are a bit larger (as makes sense).
- Uncertainty in the regression lines can be displayed graphically.