Imputation and the viral richness ~ citation count relationship

Setup

```
library(ape)
library(dplyr)
library(caper)

# data
clover <- read.csv("../data/clover.csv")
tref <- read.csv("../artifacts/trefle.csv")
cites <- read.csv("../data/PubMed_HostCounts_Total_CLOVER.csv") # from Rory Gibb
phylo_trans <- read.csv("../data/mammal_phylo_translations.csv")
phy <- read.nexus("../data/upham_tree.nex")
phy$tip.label <- gsub("_"," ", phy$tip.label)

# match names to phylogeny
lookup <- setNames(phylo_trans$Host_Upham, phylo_trans$Host)
clover$Host <- lookup[clover$Host]
tref$host <- lookup[tref$host]
cites$Host <- lookup[cites$Host]

# Calculate viral richness per host (aka parasite species richness, aka PSR)
PSR_tref <- colSums(table(tref))
PSR_clover <- colSums(table(clover[,c("Virus","Host")]))

# join into dataframe with citation counts
PSR_tref <- data.frame(Host=names(PSR_tref), PSR_tref=PSR_tref)
PSR_clover <- data.frame(Host=names(PSR_clover), PSR_clover=PSR_clover)
dat <- left_join(cites, PSR_clover)
dat <- left_join(dat, PSR_tref)

# Because of the name merger, some species have more than one citation count estimate
# Removing duplicates, but keeping estimates with highest citation count
dat <- dat[order(dat$Host, -abs(dat$Pubs_All) ),]
dat <- dat[!duplicated(dat$Host),]

# log transformations
dat$PSR_clover <- log10(dat$PSR_clover)
dat$PSR_tref <- log10(dat$PSR_tref)
dat$Pubs_All <- log10(dat$Pubs_All+1)
dat$Pubs_VirusRelated <- log10(dat$Pubs_VirusRelated+1)

comp.data <- comparative.data(phy, dat, names.col="Host", warn.dropped=TRUE)
```

```
## Warning in comparative.data(phy, dat, names.col = "Host", warn.dropped = TRUE):
## Data dropped in compiling comparative data object
```

Models using all publications

```
if (!file.exists("m1.rds")){

    m1 <- pgls(PSR_clover ~ Pubs_All, data=comp.data, lambda="ML")
    saveRDS(m1, "m1.rds")

} else { m1 <- readRDS("m1.rds") }
summary(m1)
```

```
##
## Call:
## pgls(formula = PSR_clover ~ Pubs_All, data = comp.data, lambda = "ML")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.140524 -0.024738  0.002813  0.030095  0.136912
##
## Branch length transformations:
##
## kappa  [Fix]  : 1.000
## lambda [ ML]  : 0.588
##    lower bound : 0.000, p = < 2.22e-16
##    upper bound : 1.000, p = < 2.22e-16
##    95.0% CI   : (0.472, 0.689)
## delta  [Fix]  : 1.000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.555742   0.223766 -2.4836  0.01316 *
## Pubs_All     0.527684   0.017691 29.8274  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04172 on 1062 degrees of freedom
## Multiple R-squared: 0.4559,  Adjusted R-squared: 0.4553
## F-statistic: 889.7 on 1 and 1062 DF,  p-value: < 2.2e-16
```

```
if (!file.exists("m2.rds")){

    m2 <- pgls(PSR_tref ~ Pubs_All, data=comp.data, lambda="ML")
    saveRDS(m2, "m2.rds")

} else { m2 <- readRDS("m2.rds") }
summary(m2)
```

```
##
## Call:
## pgls(formula = PSR_tref ~ Pubs_All, data = comp.data, lambda = "ML")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.175023 -0.034481 -0.000087  0.029784  0.176216
##
## Branch length transformations:
```

```
##
## kappa  [Fix]  : 1.000
## lambda [ ML]  : 0.595
##    lower bound : 0.000, p = < 2.22e-16
##    upper bound : 1.000, p = < 2.22e-16
##    95.0% CI   : (0.451, 0.716)
## delta  [Fix]  : 1.000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.659038   0.282697  2.3313  0.01993 *
## Pubs_All    0.392682   0.022127 17.7468  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0525 on 1062 degrees of freedom
## Multiple R-squared: 0.2287,  Adjusted R-squared: 0.228
## F-statistic: 314.9 on 1 and 1062 DF,  p-value: < 2.2e-16
```

Sensitivity analyses: citation counts using only "virus" related publications

```
# sensitivity analyses with Pubs_VirusRelated

if (!file.exists("m1_2.rds")){

    m1.2 <- pgls(PSR_clover ~ Pubs_VirusRelated, data=comp.data, lambda="ML")
    saveRDS(m1.2, "m1_2.rds")

} else { m1.2 <- readRDS("m1_2.rds") }
summary(m1.2)
```

```
##
## Call:
## pgls(formula = PSR_clover ~ Pubs_VirusRelated, data = comp.data,
##     lambda = "ML")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.127562 -0.022709 -0.001772  0.022894  0.117716
##
## Branch length transformations:
##
## kappa  [Fix]  : 1.000
## lambda [ ML]  : 0.450
##    lower bound : 0.000, p = < 2.22e-16
##    upper bound : 1.000, p = < 2.22e-16
##    95.0% CI   : (0.309, 0.584)
## delta  [Fix]  : 1.000
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.066188   0.162654  0.4069   0.6841
```

```
## Pubs_VirusRelated 0.705349   0.019974 35.3130   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03407 on 1062 degrees of freedom
## Multiple R-squared: 0.5401,  Adjusted R-squared: 0.5396
## F-statistic:  1247 on 1 and 1062 DF,  p-value: < 2.2e-16
```

```
if (!file.exists("m2_2.rds")){
    m2.2 <- pgls(PSR_tref ~ Pubs_VirusRelated, data=comp.data, lambda="ML")
    saveRDS(m2.2, "m2_2.rds")

} else { m2.2 <- readRDS("m2_2.rds") }
summary(m2.2)
```

```
##
## Call:
## pgls(formula = PSR_tref ~ Pubs_VirusRelated, data = comp.data,
##     lambda = "ML")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15818 -0.03348 -0.00233  0.02880  0.21350
##
## Branch length transformations:
##
## kappa  [Fix]  : 1.000
## lambda [ ML]  : 0.595
##    lower bound : 0.000, p = < 2.22e-16
##    upper bound : 1.000, p = < 2.22e-16
##    95.0% CI   : (0.457, 0.714)
## delta  [Fix]  : 1.000
##
## Coefficients:
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       1.129314   0.282240  4.0013 6.738e-05 ***
## Pubs_VirusRelated 0.474243   0.027385 17.3178 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05283 on 1062 degrees of freedom
## Multiple R-squared: 0.2202,  Adjusted R-squared: 0.2195
## F-statistic: 299.9 on 1 and 1062 DF,  p-value: < 2.2e-16
```

Summary Table

Response	Predictor	Slope	Std. Error	R Squared	Lambda	Lambda 95% CI
Viral Richness (clover)	# pubs	0.53	0.02	0.46	0.59	0.47 - 0.69
Viral Richness (trefle)	# pubs	0.39	0.02	0.23	0.59	0.45 - 0.72
Viral Richness (clover)	# virus related pubs	0.71	0.02	0.54	0.45	0.31 - 0.58
Viral Richness (trefle)	# virus related pubs	0.47	0.03	0.22	0.60	0.46 - 0.71

Table 1: Estimated parameters for models of viral richness per host. Models are fit using phylogenetic generalized least squares (PGLS) via the R pagkage caper (Orme et al 2013). Response and predictor variables were log10 transformed prior to model fitting.