Generalism data results

Josephine Walker

19 July 2016

## Summary of Metrics Calculated

Generalism, similar to diversity, can be measured in a number of ways. Poulin et al 2011 describe four facets of generalism: basic, phylogenetic, geographic, and ontogenetic. Basic metrics are calculated from numbers of hosts, phylogenetic metrics use genetic or taxonomic distances between hosts, geographic metrics look at specificity in space, and ontogenetic metrics look at specificity between life stages of the parasite. We don't expect these different measures to be independent and they tend to be correlated with number of hosts.

The nature of the data in the fish host-parasite database, which only includes associations and not abundances or prevalences, limits the metrics which we can calculate. In addition, due to only having information on fish hosts, we can't calculate an accurate measure of intermediate host specificity, so do not calculate ontogenetic measures. We have information on geography at a large regional scale, but only a small proportion of parasites have hosts in multiple locations - instead of calculating a geographic metric, we calculate basic and phylogenetic metrics both with and without parasite-host associations divided by geographic region.

The metrics we calculate are:

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Description | Facet | Source |
| degree | number of hosts | basic |  |
| G | binary measure, G=1 if degree > 1 | basic |  |
| SPD | mean pairwise phylogenetic distance between all hosts | phylogenetic | Poulin & Mouillot |
| Faith's PD | minimum total length of all the phylogenetic branches required to span all hosts on the phylogenetic tree | phylogenetic | Faith 1992 |

Both phylogenetic metrics, SPD and Faith's PD, are only calculated for parasites with more than 1 host (G=1).

## Analysis Methods

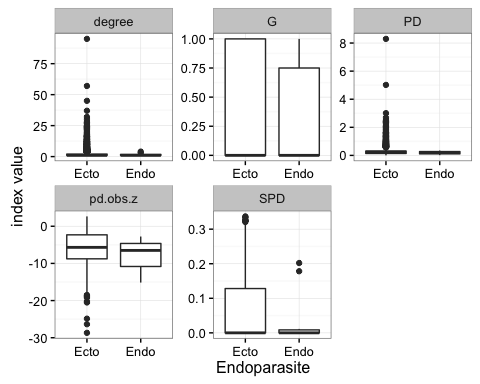
The generalism metrics for each parasite species were compared to measures of host and parasite characteristics (Table 1 in main document) for which theoretical predictions were made.

Metrics for parasites with direct or trophic life cycles were compared to summary measures (mean, maximum, and coefficient of variation (CV)) of the maximum length reported for each of their hosts. Note that CV of the host length is only calculated for parasites with more than one host. Endoparasites and ectoparasites were assessed separately, with an additional categorical length measure used for ectoparasites, whereby the length summary measures were divided into a categorical variable according to quartiles.

The effect of geographic region on parasites with direct or trophic life cycle was calculated for endoparasites and ectoparasites together. Regions were assessed as defined in Table 1 and also divided into two groups, where Antarctica, Nearctic, and Palearctic were assumed to be colder than Africa, Australia, Indopacific, and Neotropical regions. Some host-parasite associations were reported in more than one region, so for this analysis the generalism metrics were calculated separately for each region.

The generalism metrics each come from very different distributions, based on this GLMs were used as follows: for degree, negative binomial regression with a log link function (glm.nb() in R); for G, logistic regression (glm(family="binomial") in R); PD and SPD, linear regression (lm() in R), with PD log transformed. For lm(), assumptions of homoscedasticity and normality of residuals were assessed visually, and in some cases there was a bit of heteroscedasticity so there may be a bias in the standard error estimates, it has lower variance at higher values (but actually it has very few points at higher values, and so these have a lot of leverage in the estimates). [I considered doing bootstrapping to adjust p-values, or weighted least squares but not sure how to weight, but haven't done this...thoughts on better ways to do this analysis welcome!]

# Endo vs Ecto (Direct only)

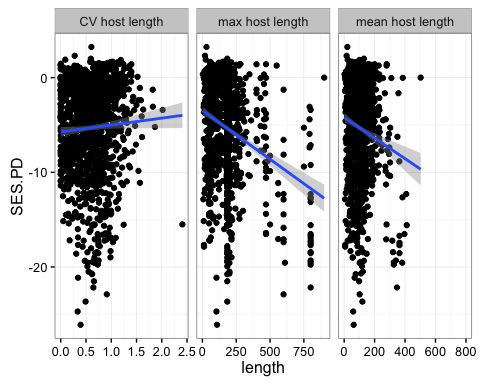
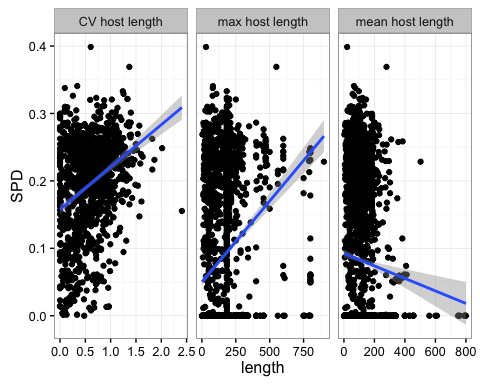
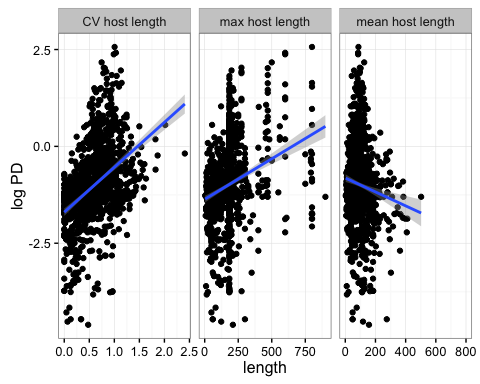
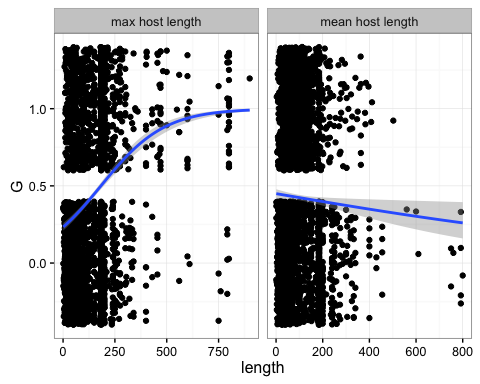
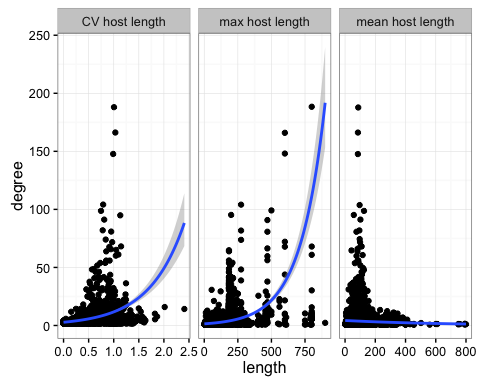
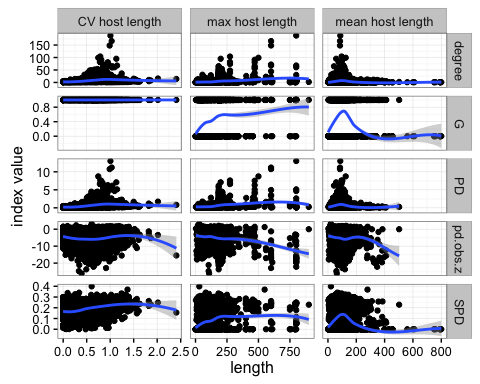


## Endoparasites & Host Body Size

Clay's model predicts that there should be a positive correlation among parasites’ generalism metrics and both the maximum host body size and the mean and CV body size across all host species.

For some measures we do see a positive correlation, and in particular there is a strong positive correlation between the coefficient of variation of host length and the mean phylogenetic distance between hosts (SPD).

In the plots, many of the relationships appear to be non-linear, but this is due to a few outlying points which don't have much impact on the fit. Plots are shown with loess ("locally weighted regression" smoothed lines to demonstrate the shape of the relationship).

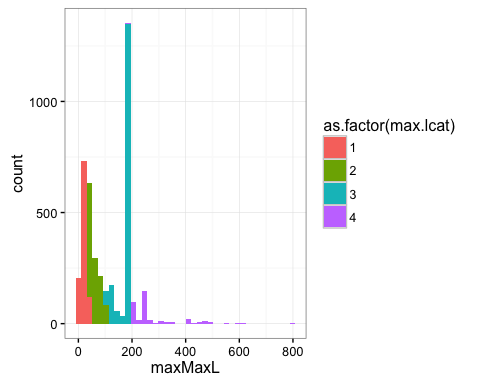
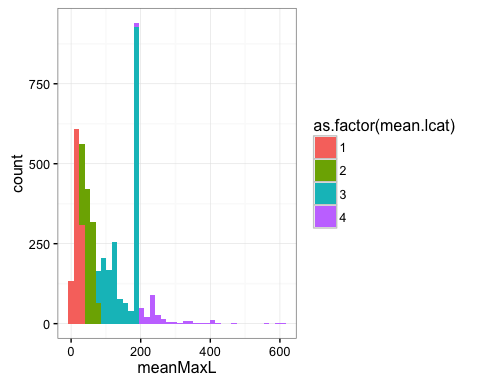
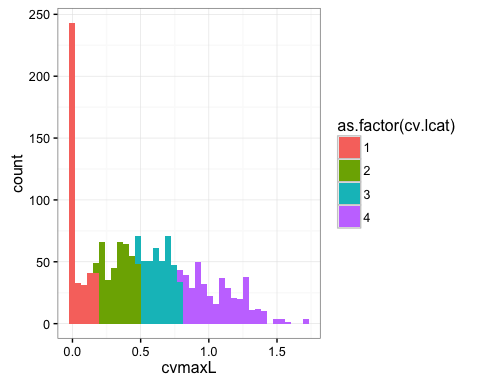
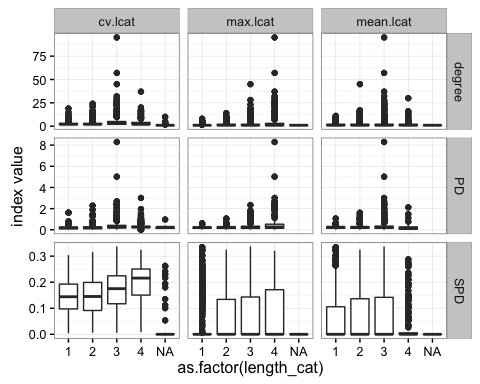
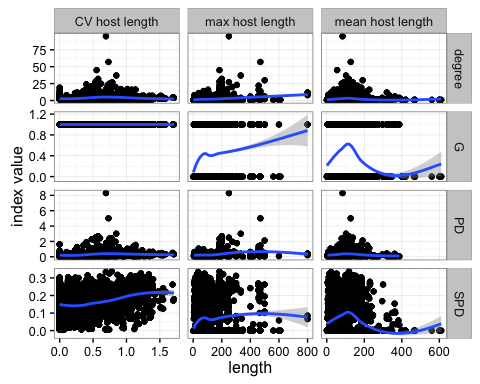


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Length Measure | Model Coefficient | Std Error | Conf Int |
| degree | mean | -0.0015239 |  | -0.0021114, -9.292210^{-4} |
| degree | max | 0.0054466 | 1.504510^{-4} | 0.0050744, 0.0058274 |
| degree | CV | 1.435891 | 0.0719312 | 1.2712456, 1.6013438 |
| G | mean | -0.0010537 | 4.487310^{-4} | -0.0019432, -1.825210^{-4} |
| G | max | 0.006522 | 4.422810^{-4} | 0.005669, 0.0074023 |
| PD | mean | -0.0017986 | 4.297310^{-4} | -0.0026416, -9.555710^{-4} |
| PD | max | 0.0020919 | 2.007810^{-4} | 0.001698, 0.0024858 |
| PD | CV | 1.1602908 | 0.0685215 | 1.0258664, 1.2947153 |
| SPD | mean | -9.28410^{-5} | 2.30310^{-5} | -1.3810^{-4}, -4.76710^{-5} |
| SPD | max | 2.404210^{-4} | 1.57810^{-5} | 2.094810^{-4}, 2.713610^{-4} |
| SPD | CV | 0.0627589 | 0.0050283 | 0.0528944, 0.0726234 |
| SESPD | mean | -0.0108949 | 0.0021264 | -0.0150664, -0.0067235 |
| SESPD | max | -0.0102986 | 9.974710^{-4} | -0.0122554, -0.0083417 |
| SESPD | CV | 0.7280691 | 0.3752025 | -0.0079981, 1.4641362 |

## Ectoparasites & Host Body Size

Clay's model predicts that for ectoparasites, there should be few generalist parasites of either very small bodied or very large bodied hosts.

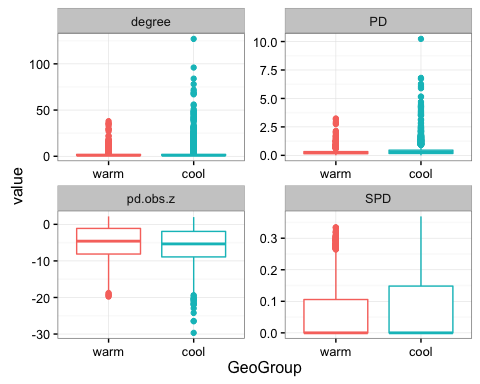
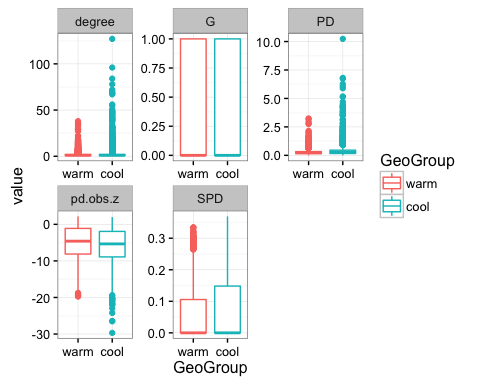
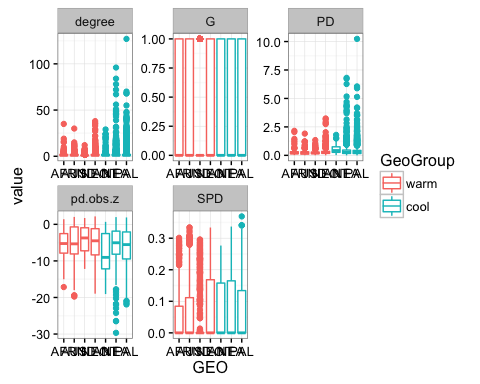
Based on the plots, this seems to be the case for degree, PD, and PS, but not so clear for SPD. However, I categorized by quartiles and due to the skewed distribution and long tail, in some cases relatively low values for length are in the 4th quartile, for example for max length.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric | Length Measure | Quartile (ref=1) | Model Coefficient | Std Error | Conf Int |
| degree | mean | 2 | 0.2235999 | 0.0414848 | 0.1423384, 0.3049629 |
| degree | mean | 3 | 0.406262 | 0.0365418 | 0.334761, 0.4780085 |
| degree | mean | 4 | 0.0995918 | 0.0660292 | -0.0302969, 0.2285828 |
| degree | max | 2 | 0.2568647 | 0.0421825 | 0.1742667, 0.339628 |
| degree | max | 3 | 0.5386372 | 0.037757 | 0.4648116, 0.6128243 |
| degree | max | 4 | 1.019527 | 0.05033 | 0.9210028, 1.1183049 |
| degree | CV | 2 | 0.1901761 | 0.0558179 | 0.0808269, 0.2996433 |
| degree | CV | 3 | 0.674365 | 0.0529644 | 0.5707044, 0.7783339 |
| degree | CV | 4 | 0.2631751 | 0.0552824 | 0.1548948, 0.3716116 |
| G | mean | 2 | 0.3274065 | 0.0919143 | 0.1475052, 0.5078821 |
| G | mean | 3 | 0.466491 | 0.0817604 | 0.3068503, 0.6274147 |
| G | mean | 4 | -0.2616669 | 0.1556438 | -0.5719861, 0.0389363 |
| G | max | 2 | 0.7788821 | 0.0963581 | 0.590882, 0.968715 |
| G | max | 3 | 1.0020415 | 0.0888333 | 0.8291696, 1.1774999 |
| G | max | 4 | 1.0857277 | 0.1285982 | 0.8338062, 1.3381528 |
| PD | mean | 2 | 0.0292508 | 0.0599103 | -0.0882624, 0.1467641 |
| PD | mean | 3 | 0.0802839 | 0.053254 | -0.0241731, 0.1847409 |
| PD | mean | 4 | -0.4862968 | 0.1073407 | -0.6968441, -0.2757497 |
| PD | max | 2 | 0.0143388 | 0.0655307 | -0.1141987, 0.1428764 |
| PD | max | 3 | 0.110301 | 0.0604254 | -0.0082226, 0.2288246 |
| PD | max | 4 | 0.3628297 | 0.0816213 | 0.2027306, 0.5229288 |
| PD | CV | 2 | 0.11333 | 0.0560856 | 0.0033182, 0.2233418 |
| PD | CV | 3 | 0.5111788 | 0.0560133 | 0.4013089, 0.6210488 |
| PD | CV | 4 | 0.5073291 | 0.0560494 | 0.3973884, 0.6172698 |
| SPD | mean | 2 | 0.010031 | 0.0039731 | 0.0022415, 0.0178204 |
| SPD | mean | 3 | 0.012963 | 0.003527 | 0.0060482, 0.0198778 |
| SPD | mean | 4 | -0.0244058 | 0.0062699 | -0.0366981, -0.0121134 |
| SPD | max | 2 | 0.0255724 | 0.0039045 | 0.0179175, 0.0332273 |
| SPD | max | 3 | 0.0291506 | 0.0035662 | 0.0221589, 0.0361422 |
| SPD | max | 4 | 0.0384315 | 0.0055195 | 0.0276104, 0.0492526 |
| SPD | CV | 2 | -0.0019701 | 0.0051343 | -0.012041, 0.0081007 |
| SPD | CV | 3 | 0.0210478 | 0.0051276 | 0.01099, 0.0311057 |
| SPD | CV | 4 | 0.0550058 | 0.0051309 | 0.0449414, 0.0650701 |
| SESPD | mean | 2 | -0.9680191 | 0.32625 | -1.6079538, -0.3280843 |
| SESPD | mean | 3 | -1.4574457 | 0.290002 | -2.0262805, -0.8886109 |
| SESPD | mean | 4 | -3.6327163 | 0.5845383 | -4.77928, -2.4861527 |
| SESPD | max | 2 | -0.8936799 | 0.3545775 | -1.5891787, -0.1981812 |
| SESPD | max | 3 | -2.2616954 | 0.3269535 | -2.90301, -1.6203809 |
| SESPD | max | 4 | -2.3225065 | 0.4416416 | -3.1887804, -1.4562327 |
| SESPD | CV | 2 | -0.7854338 | 0.3104452 | -1.394371, -0.1764966 |
| SESPD | CV | 3 | -0.4861553 | 0.3100449 | -1.0943073, 0.1219967 |
| SESPD | CV | 4 | 1.9586667 | 0.3102446 | 1.350123, 2.5672105 |

## Temperature (Region) and Generalism

Clay's model predicts that generalists are more likely in cooler temperatures. Our measure of temperature based on region here is very, very rough. We see higher degree in cool regions, but slightly lower phylogenetic generalism. Could this be because of confounding with available hosts and higher host diversity in warm regions?

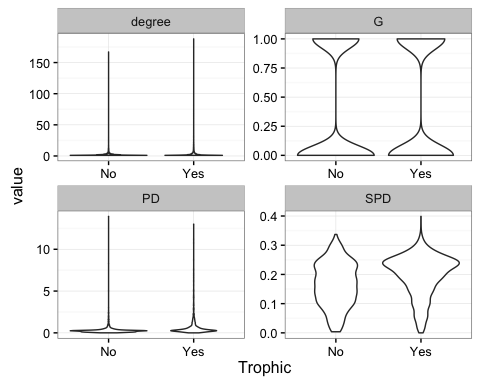


|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Model Coefficient (ref="warm") | Std Error | Conf Int |
| degree | 0.4052638 | 0.0224507 | 0.361268, 0.4492741 |
| G | 0.2775012 | 0.0459381 | 0.187575, 0.3676597 |
| PD | 0.2072072 | 0.0338973 | 0.1407433, 0.2736711 |
| SPD | 0.0119899 | 0.0020659 | 0.0079402, 0.0160396 |
| SESPD | -0.7006804 | 0.1726378 | -1.0391793, -0.3621815 |

# Not including this:

## Trophic Transmission

Amy H's model predicts that parasites with trophic transmission are less likely to be generalists, but we see the opposite pattern in this data.



|  |  |  |
| --- | --- | --- |
| Metric | Model Coefficient (ref="No") | Std Error |
| degree |  | 0.0248951 |
| G | 0.201761 | 0.0482196 |
| PD | 0.5203141 | 0.0339004 |
| SPD | 0.0316267 | 0.0027228 |

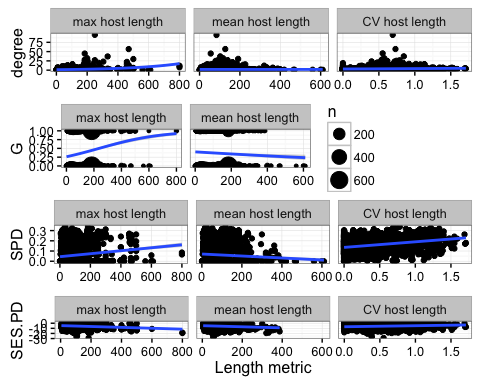
# New plots 19 July 2016: Combine endo and ecto and look at difference between direct and trophic

## Direct Life Cycles and body size metrics

The first plot where all the metrics are together has a loess smoothed line, the others have model fit lines that match the statistical analysis.

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

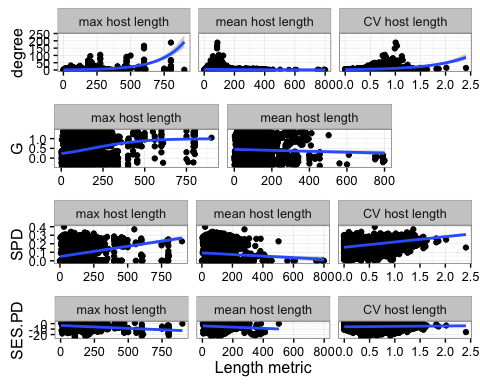


## quartz\_off\_screen   
## 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric | Length Measure | Model Coefficient | Std Error | Conf Int | Z-score |
| degree | mean | 4.71810^{-5} |  | -3.26910^{-4}, 4.20710^{-4} | 0.2582041 |
| degree | max | 0.0032361 | 1.452410^{-4} | 0.0029455, 0.0035283 | 22.2818545 |
| degree | CV | 0.2765134 | 0.0476664 | 0.1776969, 0.375511 | 5.8010138 |
| G | mean | -0.0012707 | 4.189510^{-4} | -0.0020955, -4.528710^{-4} | -3.0330028 |
| G | max | 0.0043619 | 3.799710^{-4} | 0.0036218, 0.0051115 | 11.4796355 |
| SPD | mean | -9.67710^{-5} | 1.8210^{-5} | -1.324510^{-4}, -6.10810^{-5} | -5.3163116 |
| SPD | max | 1.456710^{-4} | 1.59110^{-5} | 1.144810^{-4}, 1.768610^{-4} | 9.1552947 |
| SPD | CV | 0.0547228 | 0.0045534 | 0.0457914, 0.0636543 | 12.0180131 |
| SESPD | mean | -0.0106756 | 0.0016812 | -0.0139733, -0.007378 | -6.3499775 |
| SESPD | max | -0.0086726 | 0.0012072 | -0.0110404, -0.0063048 | -7.1843743 |
| SESPD | CV | 2.0028076 | 0.2777804 | 1.4579438, 2.5476715 | 7.2100403 |

## Trophic Life Cycles and body size metrics

The first plot where all the metrics are together has a loess smoothed line, the others have model fit lines that match the statistical analysis.

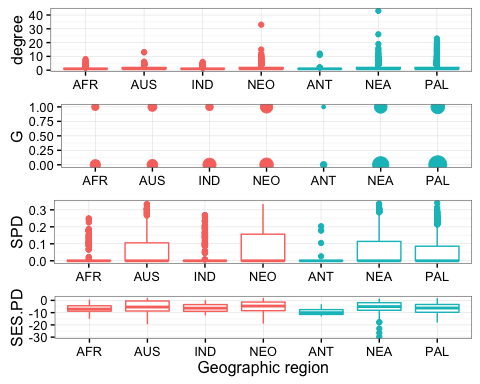


## quartz\_off\_screen   
## 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric | Length Measure | Model Coefficient | Std Error | Conf Int | Z-score |
| degree | mean | -0.0015228 |  | -0.0021116, -9.268510^{-4} | -5.7657453 |
| degree | max | 0.0054518 | 1.505610^{-4} | 0.0050788, 0.0058335 | 36.2096424 |
| degree | CV | 1.4296516 | 0.0720461 | 1.2643902, 1.5957262 | 19.8435685 |
| G | mean | -0.0010366 | 4.491110^{-4} | -0.0019269, -1.646310^{-4} | -2.3082268 |
| G | max | 0.0065672 | 4.437410^{-4} | 0.0057114, 0.0074505 | 14.79969 |
| SPD | mean | -9.30910^{-5} | 2.30910^{-5} | -1.383610^{-4}, -4.78110^{-5} | -4.0313014 |
| SPD | max | 2.411110^{-4} | 1.5810^{-5} | 2.101310^{-4}, 2.720810^{-4} | 15.2602011 |
| SPD | CV | 0.062533 | 0.0050249 | 0.0526751, 0.0723909 | 12.4445203 |
| SESPD | mean | -0.01115 | 0.0021286 | -0.0153259, -0.0069742 | -5.2382138 |
| SESPD | max | -0.0103663 | 9.969410^{-4} | -0.0123221, -0.0084105 | -10.3981061 |
| SESPD | CV | 0.7223738 | 0.3752382 | -0.0137644, 1.458512 | 1.9251074 |

## Direct life cycle and geographic region

## Warning: Removed 3280 rows containing non-finite values (stat\_boxplot).



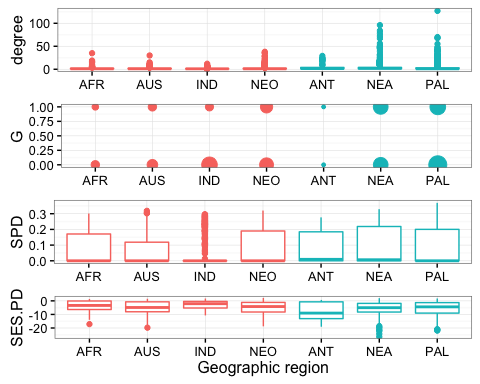
## Warning: Removed 3280 rows containing non-finite values (stat\_boxplot).

## quartz\_off\_screen   
## 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Model Coefficient (ref="warm") | Std Error | Conf Int | Z-score |
| degree | 0.0884202 | 0.0249142 | 0.0396377, 0.1373022 | 3.5489972 |
| G | 0.0341042 | 0.0623994 | -0.0880585, 0.1565706 | 0.5465479 |
| SPD | -1.865410^{-4} | 0.0024935 | -0.005075, 0.0047019 | -0.0748103 |
| SESPD | -0.4243263 | 0.2257372 | -0.8671087, 0.0184561 | -1.8797356 |

## Trophic life cycle and geographic region

## Warning: Removed 2287 rows containing non-finite values (stat\_boxplot).



## Warning: Removed 2287 rows containing non-finite values (stat\_boxplot).

## quartz\_off\_screen   
## 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Model Coefficient (ref="warm") | Std Error | Conf Int | Z-score |
| degree | 0.6627308 | 0.0364354 | 0.591304, 0.7341342 | 18.1891753 |
| G | 0.5811795 | 0.0684301 | 0.4473767, 0.7156506 | 8.4930403 |
| SPD | 0.0285904 | 0.0033896 | 0.0219448, 0.0352359 | 8.4347864 |
| SESPD | -1.0868156 | 0.2633792 | -1.6034515, -0.5701797 | -4.1264295 |

## Distribution of host lengths by region

##   
## Call:  
## lm(formula = value ~ centermaxL \* GeoGroup, data = geoTlong[geoTlong$index ==   
## "SES.PD", ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.6856 -3.0434 0.8182 3.7477 11.3619   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.6382 0.2051 -22.618 < 2e-16 \*\*\*  
## centermaxL 0.4378 0.2198 1.992 0.0466 \*   
## GeoGroupcool -0.4310 0.2615 -1.648 0.0995 .   
## centermaxL:GeoGroupcool -1.7817 0.2539 -7.016 3.46e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.686 on 1479 degrees of freedom  
## (2716 observations deleted due to missingness)  
## Multiple R-squared: 0.08303, Adjusted R-squared: 0.08117   
## F-statistic: 44.64 on 3 and 1479 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = value ~ centermaxL \* GEO, data = geoTlong[geoTlong$index ==   
## "SES.PD", ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.3843 -2.9968 0.6342 3.5238 12.0075   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.0100 0.6042 -6.636 4.51e-11 \*\*\*  
## centermaxL -1.7350 1.1010 -1.576 0.1153   
## GEOAUS -1.2334 0.7575 -1.628 0.1037   
## GEOIND 1.0030 0.7837 1.280 0.2008   
## GEONEO -1.1788 0.6677 -1.765 0.0777 .   
## GEOANT -4.1643 1.0551 -3.947 8.30e-05 \*\*\*  
## GEONEA -1.2032 0.6508 -1.849 0.0647 .   
## GEOPAL -0.7697 0.6441 -1.195 0.2322   
## centermaxL:GEOAUS 1.9640 1.1656 1.685 0.0922 .   
## centermaxL:GEOIND 1.9749 1.1986 1.648 0.0996 .   
## centermaxL:GEONEO 2.4938 1.1525 2.164 0.0306 \*   
## centermaxL:GEOANT 0.9269 1.6781 0.552 0.5808   
## centermaxL:GEONEA 1.0511 1.1280 0.932 0.3516   
## centermaxL:GEOPAL 0.1311 1.1108 0.118 0.9061   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.623 on 1469 degrees of freedom  
## (2716 observations deleted due to missingness)  
## Multiple R-squared: 0.1138, Adjusted R-squared: 0.106   
## F-statistic: 14.52 on 13 and 1469 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = value ~ centermaxL \* GeoGroup, data = geoDlong[geoDlong$index ==   
## "SES.PD", ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.9948 -2.8861 0.1935 3.5523 8.7162   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.61546 0.17060 -32.917 < 2e-16 \*\*\*  
## centermaxL -0.60686 0.17490 -3.470 0.000535 \*\*\*  
## GeoGroupcool -0.12437 0.23135 -0.538 0.590960   
## centermaxL:GeoGroupcool -0.08424 0.22906 -0.368 0.713110   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.343 on 1549 degrees of freedom  
## (3856 observations deleted due to missingness)  
## Multiple R-squared: 0.02362, Adjusted R-squared: 0.02173   
## F-statistic: 12.49 on 3 and 1549 DF, p-value: 4.511e-08

##   
## Call:  
## lm(formula = value ~ centermaxL \* GEO, data = geoDlong[geoDlong$index ==   
## "SES.PD", ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.6274 -2.9029 0.1961 3.4208 9.4146   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.02572 0.55771 -12.598 < 2e-16 \*\*\*  
## centermaxL -0.04476 0.66621 -0.067 0.94645   
## GEOAUS 1.92036 0.67727 2.835 0.00464 \*\*   
## GEOIND 1.10756 0.76055 1.456 0.14553   
## GEONEO 1.62860 0.60271 2.702 0.00697 \*\*   
## GEOANT -1.62681 2.23430 -0.728 0.46666   
## GEONEA 1.92598 0.60763 3.170 0.00156 \*\*   
## GEOPAL 0.79831 0.59411 1.344 0.17924   
## centermaxL:GEOAUS -1.08794 0.73980 -1.471 0.14160   
## centermaxL:GEOIND -0.39641 0.80897 -0.490 0.62419   
## centermaxL:GEONEO -0.41637 0.71600 -0.582 0.56098   
## centermaxL:GEOANT -1.00468 2.75782 -0.364 0.71568   
## centermaxL:GEONEA -0.56019 0.70732 -0.792 0.42849   
## centermaxL:GEOPAL -0.75845 0.69213 -1.096 0.27333   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.31 on 1539 degrees of freedom  
## (3856 observations deleted due to missingness)  
## Multiple R-squared: 0.04486, Adjusted R-squared: 0.03679   
## F-statistic: 5.56 on 13 and 1539 DF, p-value: 5.523e-10