What controls the range of hosts a fish parasite infects?

Tad Dallas, Andrew Park, and John M. Drake

Knowledge gap

- What constrains the range of hosts that a parasite can infect? Is there a simple range of host functional traits that can determine the likelihood that a parasite infects a given host species? How well can we predict parasite occurrences given host life history traits? How about using solely information on parasite community structure?
- Does the importance of different host functional traits or parasite community information differ with parasite type? (supplement)
- Since geographic variables are important, what if we try to predict parasite niche breadth in a specific biogeographic region? (supplement)

Thesis paragraph

Here, I apply a series of predictive models in order to predict parasite occurrence across a range of potential host species for a large set of parasites of freshwater fish, using host functional traits, geographic location, and parasite community information. Parasite occurrence on a given host could be constrained by space (geographic location), patch quality (host characteristics), or through interactions with competing parasites (parasite community structure). It's important to note right up front that the importance of parasite community structure cannot be interpreted as evidence for community interactions, as parasites could infect hosts based on their traits, and the parasite community information could just be serving as a proxy for unmeasured host trait variation. However, predicting parasite occurrence based solely on parasite community information does remove some importance of the patch (host), and is easy to sell, as it may be possible to predict spillover of parasites, or the degree of biotic resistance a community offers to a potential invader, simply by having presence-absence data on parasite communities.

I examine the role of parasite group (Acanthocephala, Cestoda, Monogenea, Nematoda, Trematoda) on model performance, and on the relative contribution of different variables to prediction.

thorns in my side:

- Absence data aren't true absences. Should I even train on these data if the model treats them as true absences? I think I'm basically treating them as background data, much like MaxEnt.

Methods

Data and processing

We use an existing global database of fish-parasite associations (Strona et al. 2013) consisting of over 38000 parasite records spanning a large diversity of parasites (Acanthocephala, Cestoda, Monogenea, Nematoda, Trematoda). In order to allow for cross-validation and accurate prediction, we constrained our ananlyses to parasites with a minimum of 50 host records. In other words, we only examined parasites that had been recorded more than 50 times, but these occurrences could be on fewer than 50 host species. The inclusion of duplicate occurrences was only permitted if the parasite was recorded on a host in a different location, based on latitude and longitude values. Our response variable was parasite occurrence (binary), and was predicted using only host life history traits, and geographic location of host capture. Host trait information was obtained through the FishPest database (Strona and Lafferty 2012; Strona et al. 2013), and FishBase (Froese and Pauly 2010). Host traits descriptions are provided in Table 1

Model formulation

We trained a series of models in order to compare predictive performance of different techniques. Each model was trained on 70% of the data, and accuracy was determined from the remaining 30%. This process was repeated z times ($z = 20$). We generated background data by randomly sampling host species where parasite i was not recorded. To maintain proportional training data, the number of random samples was selected to be five times greater than the occurrence records.

Models used

discuss null predictions scenario, and then go into other algorithms used (brt, svm, lr, rf)

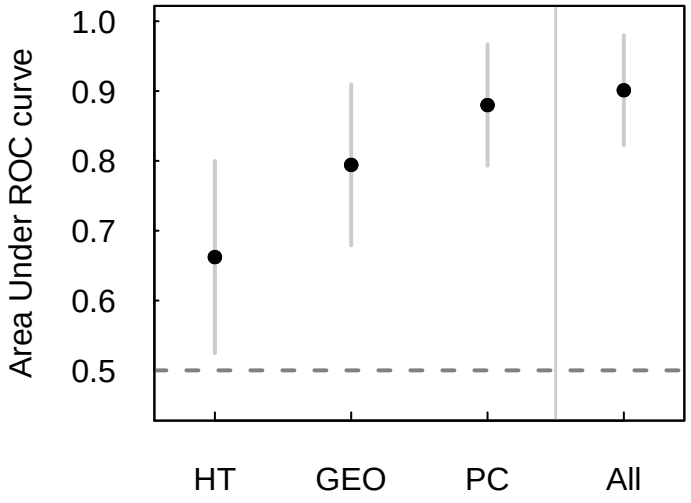
Training using only host trait data

Training using only geographic

Training using only parasite community data

```
testRet = apply(as.matrix(ret), 2, as.numeric)
testRetb = apply(as.matrix(ret.b), 2, as.numeric)
testRetc = apply(as.matrix(ret.c), 2, as.numeric)
testRetgeo = apply(as.matrix(ret.geo), 2, as.numeric)

meanAUC = colMeans(testRet)
meanSE = apply(testRet, 2, sd)/sqrt(nrow(testRet))
meanSD = apply(testRet, 2, sd)

meanAUC.b = colMeans(testRetb)
meanSE.b = apply(testRetb, 2, sd)/sqrt(nrow(testRetb))
meanSD.b = apply(testRetb, 2, sd)

meanAUC.c = colMeans(testRetc)
meanSE.c = apply(testRetc, 2, sd)/sqrt(nrow(testRetc))
meanSD.c = apply(testRetc, 2, sd)

meanAUC.geo = colMeans(testRetgeo)
meanSE.geo = apply(testRetgeo, 2, sd)/sqrt(nrow(testRetgeo))
meanSD.geo = apply(testRetgeo, 2, sd)


brt.meanAUC = c(meanAUC[3], meanAUC.geo[3], meanAUC.b[3], meanAUC.c[3])
brt.se = c(meanSE[3], meanSE.geo[3], meanSE.b[3], meanSE.c[3])
brt.sd = c(meanSD[3], meanSD.geo[3], meanSD.b[3], meanSD.c[3])


plot(1:4, brt.meanAUC, ylim = c(0.45, 1), xlim = c(0.7, 4.3), las = 1, pch = 16,
    tck = 0.01, xaxt = "n", xlab = "Predictive model used", ylab = "Area Under ROC curve")
```

```
abline(h = 0.5, col = grey(0.5), lty = 2, lwd = 2)
abline(v = 3.5, col = grey(0.8), lwd = 1.5)
segments(x0 = 1:4, y0 = brt.meanAUC + (brt.sd), y1 = brt.meanAUC - (brt.sd),
    col = grey(0.8), lwd = 2)

points(1:4, brt.meanAUC, pch = 16, cex = 1)
axis(1, at = 1:4, labels = c("HT", "GEO", "PC", "All"), tck = 0.02, lwd = 1)
box()
```



Training using all data

```
dev.copy(pdf, width = 3.6, height = 3.6, "../Manuscript/Figures/brtAccuracy.pdf")
```

```
## pdf
##   3
```

```
dev.off()
```

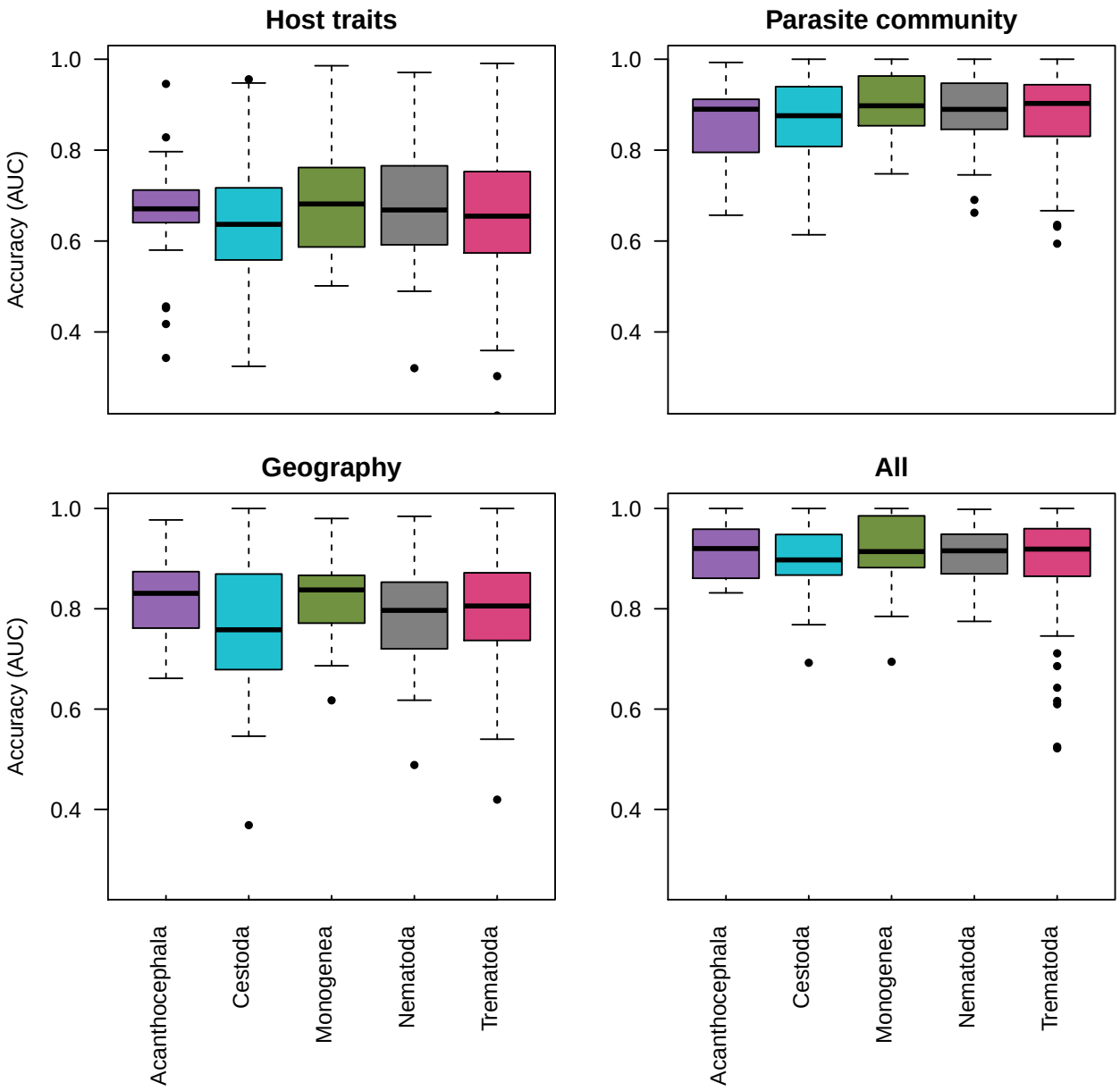
```
## pdf
##   2
```

Does predictive power differ systematically among parasites?

```
rownames(testRet) = parPest[1:238]

#Number of hosts each parasite infects
hostNum = summary(pest[,'P_SP'])[1:238]

#Beating the data up to get parasite 'type'
parType = unique(pest[which(pest[,'P_SP']%in% parPest[1:238]), 2:4], MARGIN=1)
parType[,1] = gsub('[*]','',parType[,1])
parType = unique(parType, MARGIN=1)
parType = parType[order(parType[,3]),]
```

```r
# Hand-wavey solution to the problem of classifying parasites as two different classes of parasites
# (C and T, or A and M).

stupid=names(summary(parType[,3], maxsum=242)[1:3])
for(i in 1:length(stupid)){
  parType=parType[-which(parType[,3] %in% stupid[i])[1],]
}
```

```r
layout(matrix(c(1,2,3,4), ncol=2), heights=c(1,1.5))

hostCol=c('#9367B1', '#20C0D0','#719240', grey(0.5), '#D9437E')
par(mar=c(1,4,2,0.25))
plot(as.factor(parType[,1]), testRet[1:238,3], col=hostCol, tck=0.01, pch=16, las=1, ylab='Accuracy (AU

par(mar=c(8,4,2,0.25))
plot(as.factor(parType[,1]), testRetgeo[1:238,3], col=hostCol, tck=0.01, pch=16, las=1, ylab='Accuracy
parGroups=c('Acanthocephala', 'Cestoda', 'Monogenea', 'Nematoda', 'Trematoda')
axis(1, 1:5, parGroups, las=2, tck=0.01, padj=0)

par(mar=c(1,4,2,0.25))
plot(as.factor(parType[,1]), testRetb[1:238,3], col=hostCol, tck=0.01,pch=16, las=1, ylab='', xaxt='n',

par(mar=c(8,4,2,0.25))
plot(as.factor(parType[,1]), testRetc[1:238,3], col=hostCol, pch=16, tck=0.01, las=1, ylab='', xaxt='n'
parGroups=c('Acanthocephala', 'Cestoda', 'Monogenea', 'Nematoda', 'Trematoda')
axis(1, 1:5, parGroups, las=2, tck=0.01, padj=0)
```



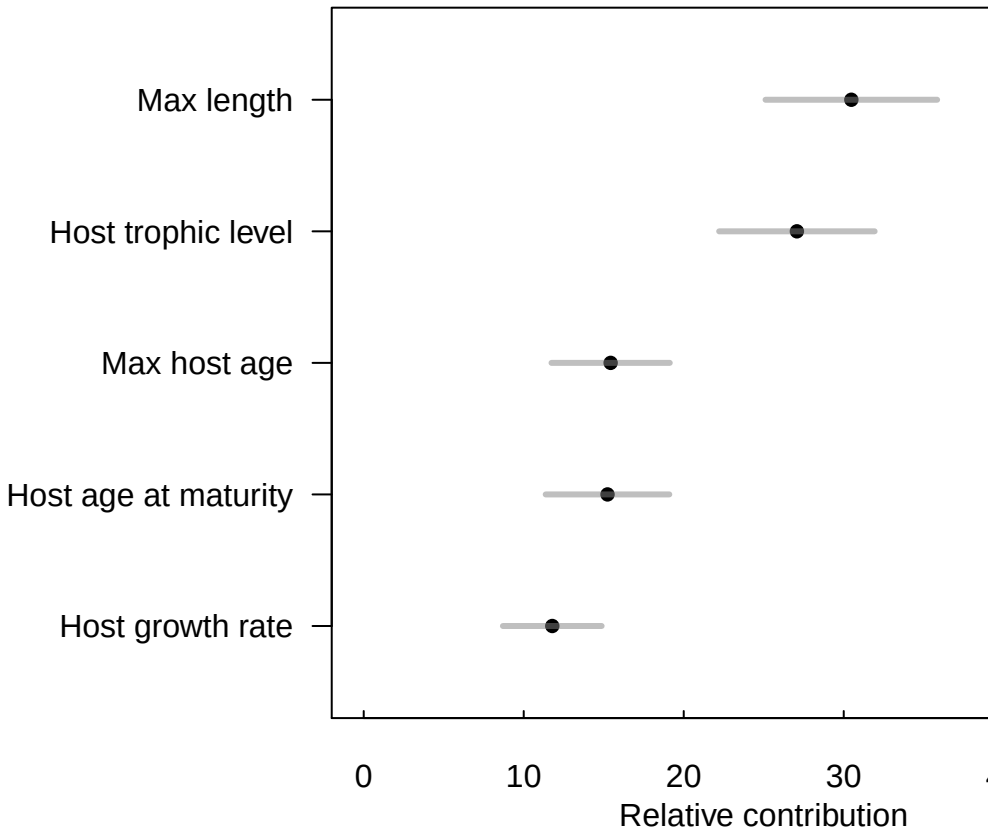

```
dev.copy(pdf, width=5, height=5, '../Manuscript/Figures/parAccuracy.pdf'); dev.off()
```

```
## pdf
##   2
```

Are parasite distributions shaped by different factors?

```
#varNames=c('Geographic region', 'Area of occupancy', 'Latitude', 'Max length', 'Host trophic level', '

brtMean = rowMeans(brtResults.traits)
brtSE = apply(brtResults.traits, 1, sd) / sqrt(ncol(brtResults.traits))
brtSD = apply(brtResults.traits, 1, sd)
```

```
inds.a=order(brtMean)
par(mar=c(3,9,1,1))
plot(brtMean[inds.a], 1:5, pch=16, xlim=c(0,50), ylim=c(0.5,5.5), las=1, xlab='', ylab='', yaxt='n', col
segments(x0=brtMean[inds.a] - brtSD[inds.a], x1=brtMean[inds.a] + brtSD[inds.a], y0=1:5, col=grey(0.5,0
mtext("Relative contribution", side=1, line=2)
axis(2, at=1:5, labels=c('Host growth rate', 'Host age at maturity', 'Max host age', 'Host trophic level
```

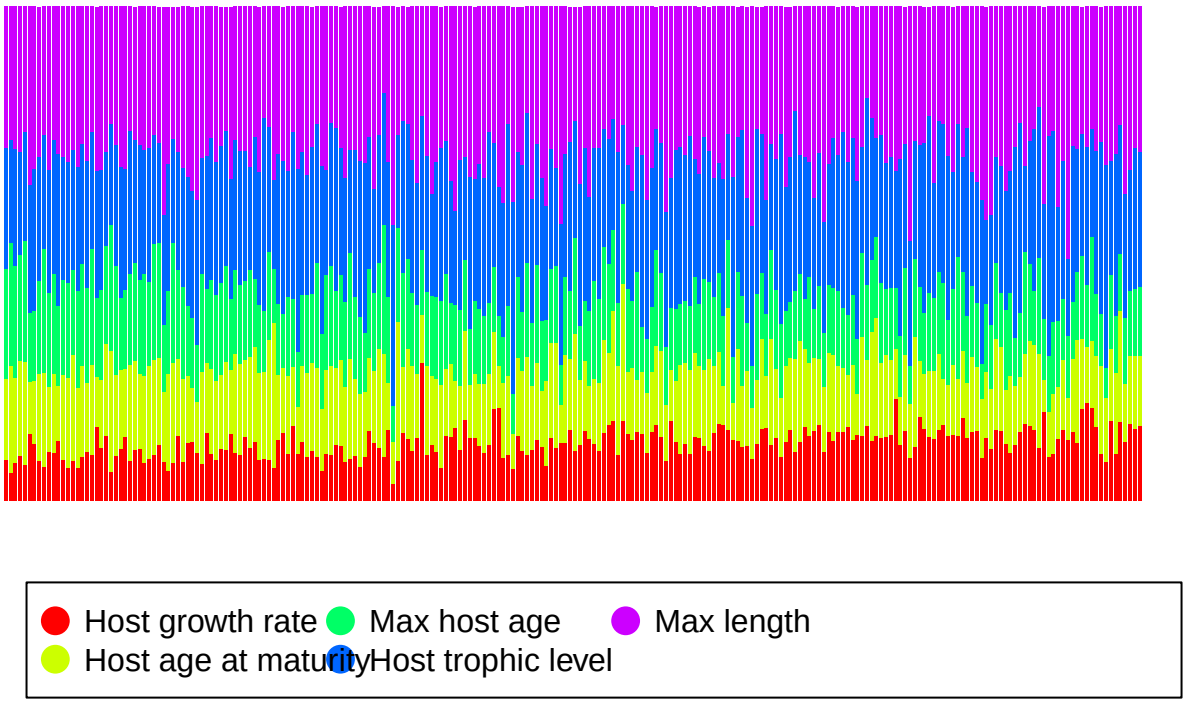


BRT model using only host traits

```
dev.copy(pdf, width=6, height=5, '../Manuscript/Figures/hostTraits.pdf'); dev.off()
```

```
## pdf
##   2
```

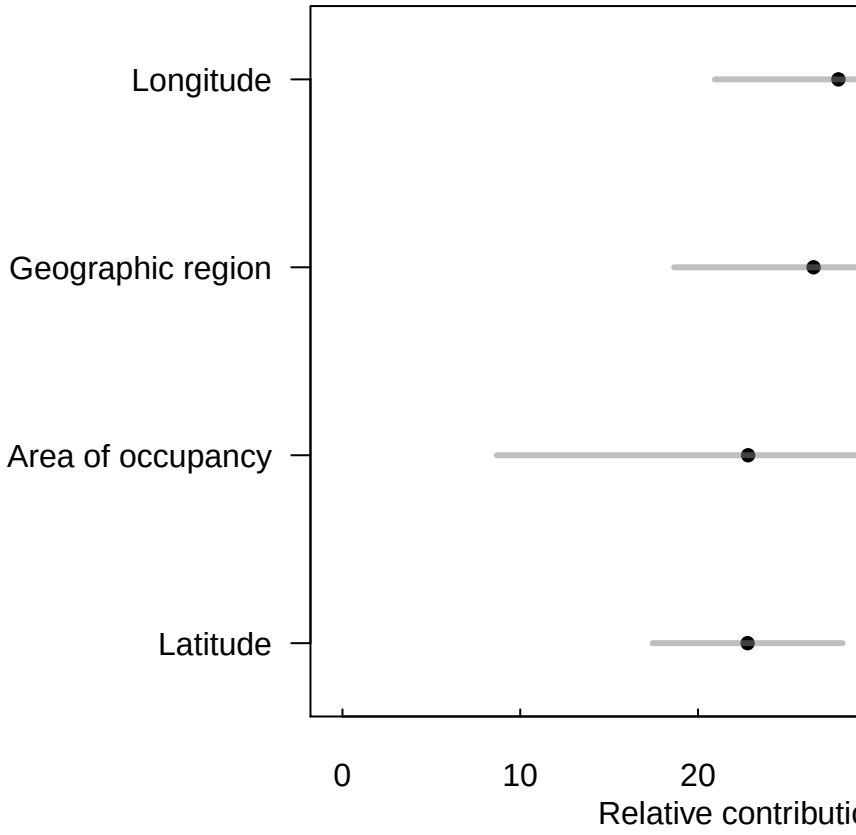
```
layout(matrix(c(1,2),ncol=1), heights=c(1.5, 1))
par(mar=c(0.25,0.25,0.25,0.25))
barplot(brtResults.traits[inds.a,], col=rainbow(5), xaxt='n', axes=FALSE, border = NA)
plot.new()
legend(0.02,0.85, c('Host growth rate', 'Host age at maturity', 'Max host age', 'Host trophic level', '
```




```
dev.copy(pdf, width=7, height=5, '../Manuscript/Figures/hostTraitsColor.pdf'); dev.off()
```

```
## pdf
##   2
```

```
varNames=c('Geographic region', 'Area of occupancy', 'Latitude', 'Max length', 'Host trophic level', 'L
#rownames(brtResults[rev(inds),])]

brtMean = rowMeans(brtResults.geo)
brtSE = apply(brtResults.geo, 1, sd) / sqrt(ncol(brtResults.geo))
brtSD = apply(brtResults.geo, 1, sd)

inds.a=order(brtMean)
par(mar=c(3,9,1,1))
plot(brtMean[inds.a], 1:4, pch=16, xlim=c(0,45), ylim=c(0.75,4.25), las=1, xlab='', ylab='', yaxt='n',
segments(x0=brtMean[inds.a] - brtSD[inds.a], x1=brtMean[inds.a] + brtSD[inds.a], y0=1:4, col=grey(0.5,0
mtext("Relative contribution", side=1, line=2)
axis(2, at=1:4, labels=c('Latitude', 'Area of occupancy', 'Geographic region', 'Longitude'), las=1)
```

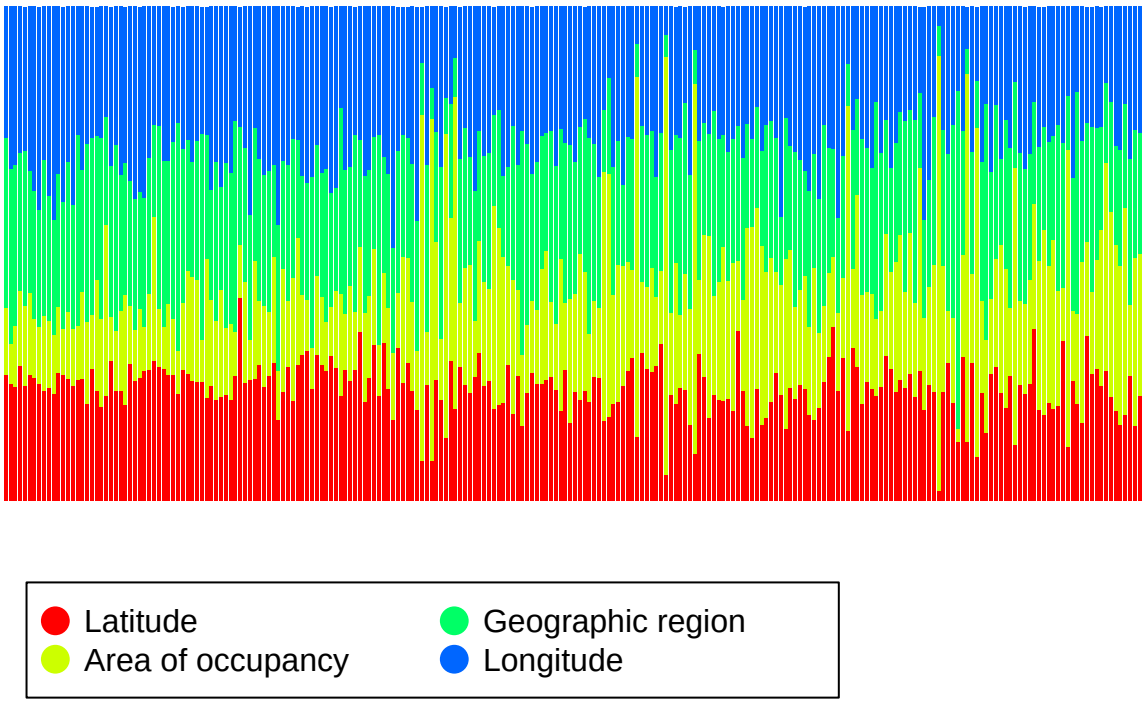


BRT model using only geographic variables

```
dev.copy(pdf, width=6, height=5, '../Manuscript/Figures/geography.pdf'); dev.off()
```

```
## pdf
##   2
```

```
layout(matrix(c(1,2),ncol=1), heights=c(1.5, 1))
par(mar=c(0.25,0.25,0.25,0.25))
barplot(brtResults.geo[inds.a,], col=rainbow(5), xaxt='n', axes=FALSE, border = NA)
plot.new()
legend(0.02,0.85, c('Latitude', 'Area of occupancy', 'Geographic region', 'Longitude'), pch=16, col=rai
```

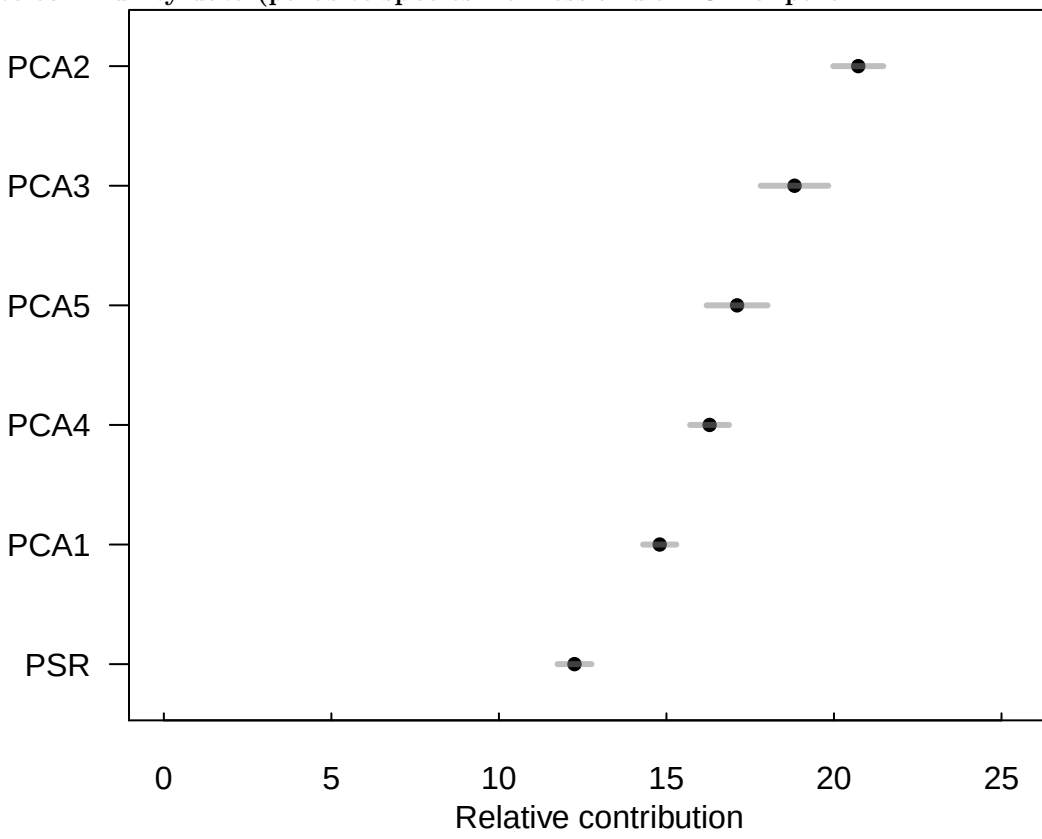



```
dev.copy(pdf, width=7, height=5, '../Manuscript/Figures/geographyColor.pdf'); dev.off()
```

```
## pdf
##   2
```

```
brtMean.b = rowMeans(brtResults.pars)
brtSE.b = apply(brtResults.pars, 1, sd) / sqrt(ncol(brtResults.pars))
brtSD.b = apply(brtResults.pars, 1, sd)

inds=order(brtMean.b)
par(mar=c(3,7,1,1))
plot(brtMean.b[inds], 1:6, pch=16, xlim=c(0,26), ylim=c(0.75,6.25), las=1, xlab='', ylab='', yaxt='n',
segments(x0=brtMean.b[inds] - brtSE.b[inds], x1=brtMean.b[inds] + brtSE.b[inds], y0=1:6, col=grey(0.5,
mtext("Relative contribution", side=1, line=2)
axis(2, at=1:6, labels=rev(c('PCA2', 'PCA3', 'PCA5', 'PCA4', 'PCA1', 'PSR')), las=1)
```

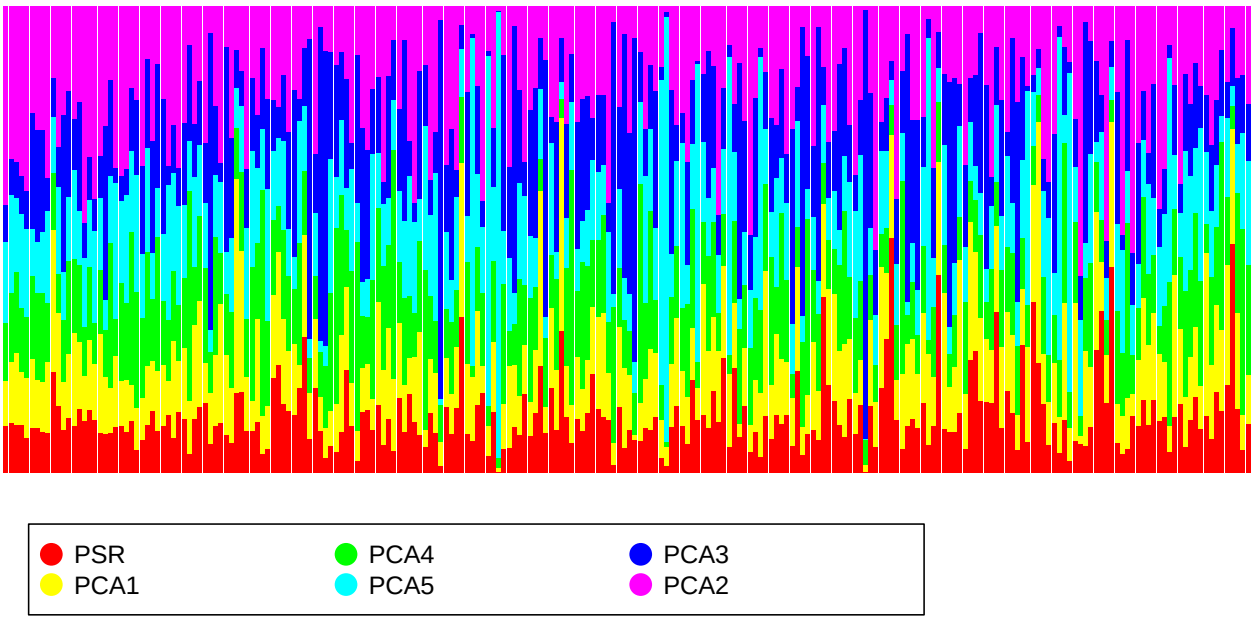
BRT models using only parasite community data (parasite species richness and a PCA of parasite community composition)




```
dev.copy(pdf, width=6, height=5, '../Manuscript/Figures/parasiteCommunity.pdf'); dev.off()
```

```
## pdf
##   2
```

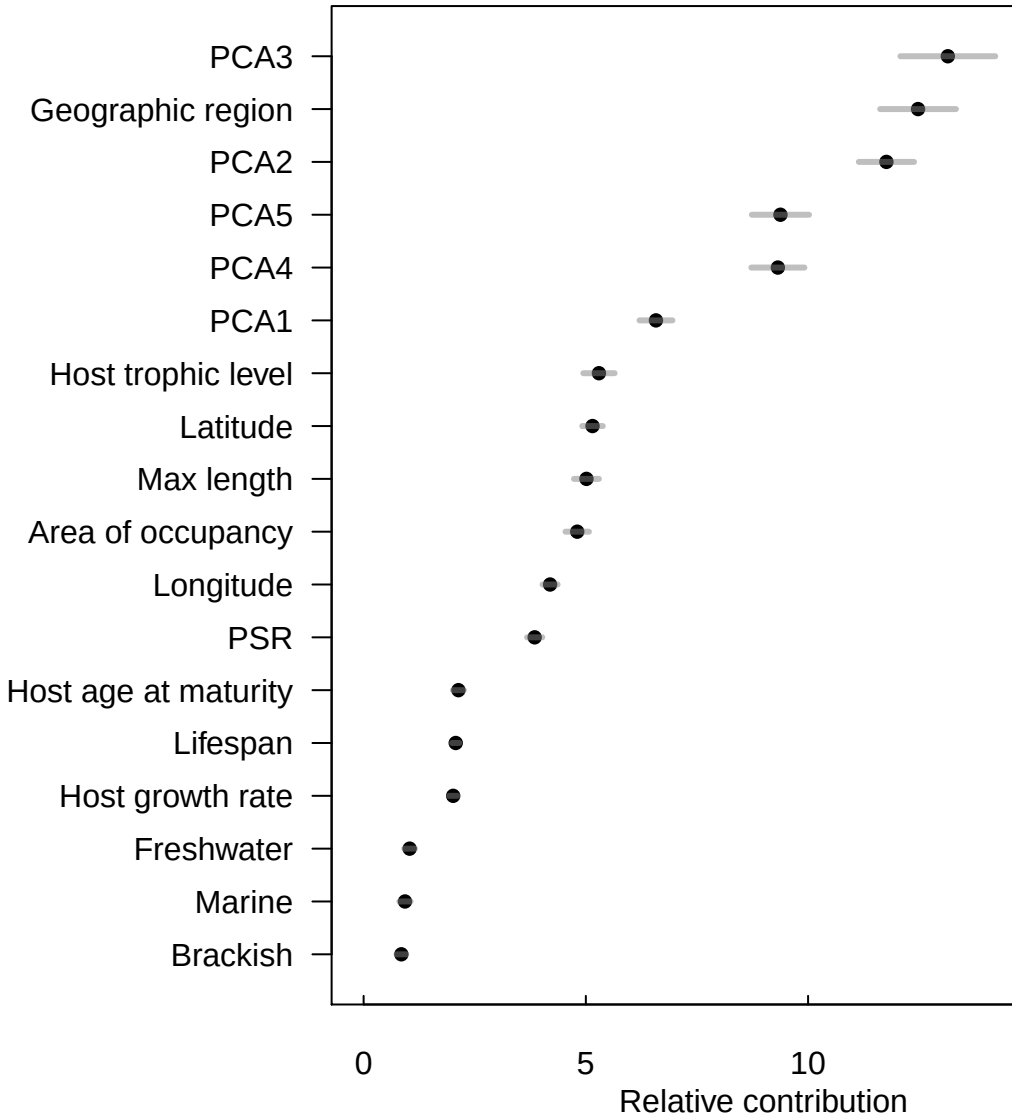
```
layout(matrix(c(1,2),ncol=1), heights=c(2.5, 1))
par(mar=c(0.25,0.25,0.25,0.25))
barplot(brtResults.pars[inds,], col=rainbow(6), xaxt='n', axes=FALSE, border = NA)
plot.new()
legend(0.02,0.85, rev(c('PCA2', 'PCA3', 'PCA5', 'PCA4', 'PCA1', 'PSR')), pch=16, col=rainbow(6), ncol=3
```




```
dev.copy(pdf, width=7, height=5, '../Manuscript/Figures/parasiteCommunityColor.pdf'); dev.off()
```

```
## pdf
##   2
```

```
brtMean.c = rowMeans(brtResults.all)
brtSE.c = apply(brtResults.all, 1, sd) / sqrt(ncol(brtResults.all))
brtSD.c = apply(brtResults.all, 1, sd)

inds=order(brtMean.c)
par(mar=c(3,9,1,1))
plot(brtMean.c[inds], 1:18, pch=16, xlim=c(0,18), ylim=c(0.75,18.25), las=1, xlab='', ylab='', yaxt='n'
segments(x0=brtMean.c[inds] - brtSE.c[inds], x1=brtMean.c[inds] + brtSE.c[inds], y0=1:18, col=grey(0.5,
mtext("Relative contribution", side=1, line=2)
axis(2, at=1:18, labels=c('Brackish', 'Marine', 'Freshwater', 'Host growth rate', 'Lifespan', 'Host age
```

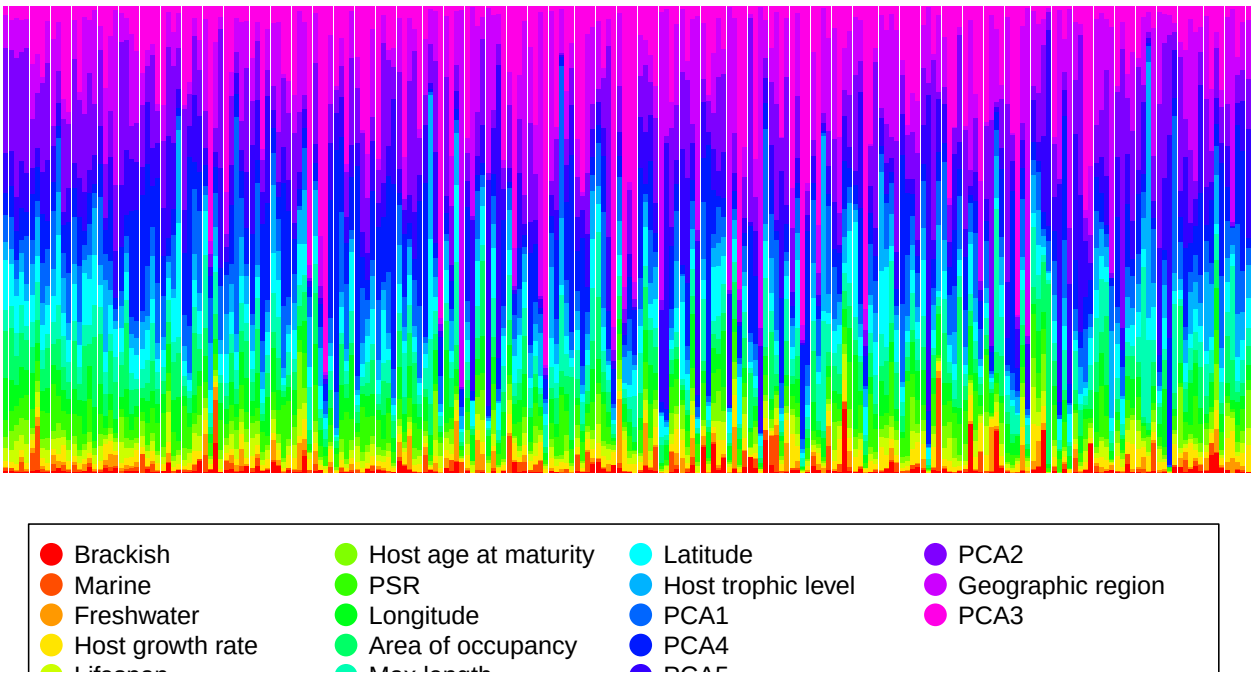


BRT models trained on all data

```
dev.copy(pdf, width=6, height=5, '../Manuscript/Figures/allData.pdf'); dev.off()
```

```
## pdf
##   2
```

```
layout(matrix(c(1,2),ncol=1), heights=c(2.5, 1))
par(mar=c(0.25,0.25,0.25,0.25))
barplot(brtResults.all[inds,], col=rainbow(20), xaxt='n', axes=FALSE, border = NA)
plot.new()
legend(0.02,0.85, c('Brackish', 'Marine', 'Freshwater', 'Host growth rate', 'Lifespan', 'Host age at ma
```

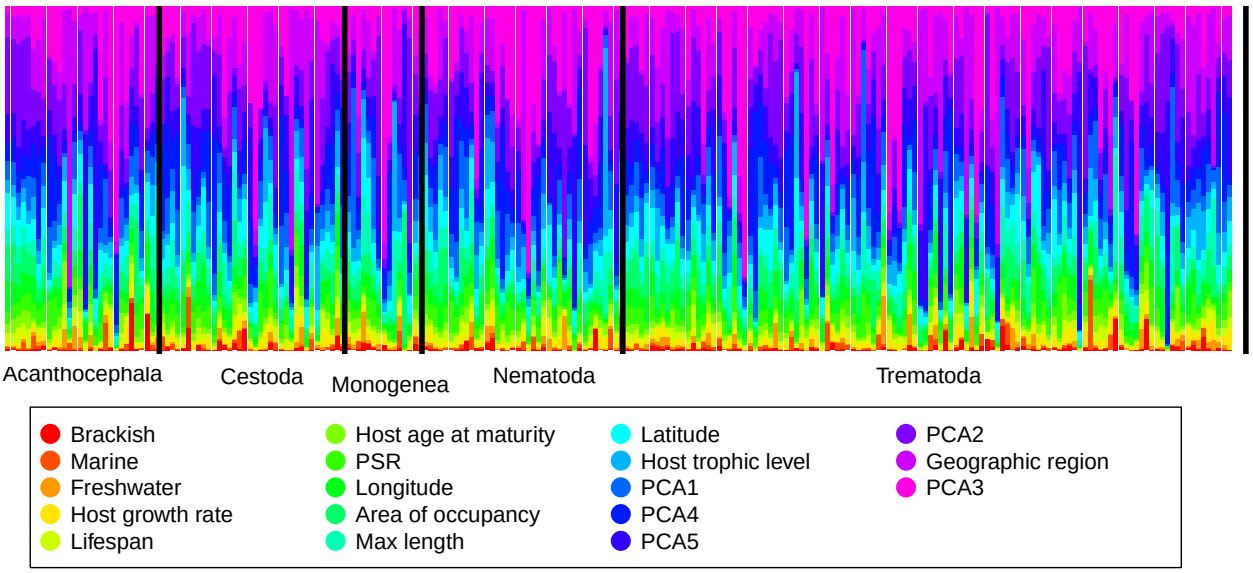



```
dev.copy(pdf, width=7, height=5, '../Manuscript/Figures/allDataColor.pdf'); dev.off()
```

```
## pdf
##   2
```

Is this variation attributable to parasite type?

```
layout(matrix(c(1,2),ncol=1), heights=c(1.5, 1))
par(mar=c(0.25,0.25,0.25,0.25))

brtResults.ParType=brtResults.all[inds,order(parType[,1])]

barplot(brtResults.ParType, col=rainbow(20), xaxt='n', axes=FALSE, border = NA)
abline(v=c(30,66,81,120,241)*1.199, lwd=4)
plot.new()
text(15/239, 1 , parGroups[1],srt=0)
text(50/239, 1, parGroups[2], srt=0)
text(75/239, 0.95, parGroups[3], srt=0)
text(105/239, 1, parGroups[4], srt=0)
text(180/239, 1, parGroups[5], srt=0)


legend(0.02,0.85, c('Brackish', 'Marine', 'Freshwater', 'Host growth rate', 'Lifespan', 'Host age at ma
```




```
dev.copy(pdf, width=7, height=5, '../Manuscript/Figures/parTypeColor.pdf'); dev.off()
```

```
## pdf
##   2
```

References

Froese, Rainer, and Daniel Pauly. 2010. "FishBase." International Center for Living Aquatic Resources Management.

Strona, Giovanni, and Kevin D Lafferty. 2012. "FishPEST: an Innovative Software Suite for Fish Parasitologists." *Trends in Parasitology* 28 (4): 123.

Strona, Giovanni, Maria Lourdes D Palomares, Nicolas Bailly, Paolo Galli, and Kevin D Lafferty. 2013. "Host Range, Host Ecology, and Distribution of More Than 11 800 Fish Parasite Species: Ecological Archives E094-045." *Ecology* 94 (2): 544–544.