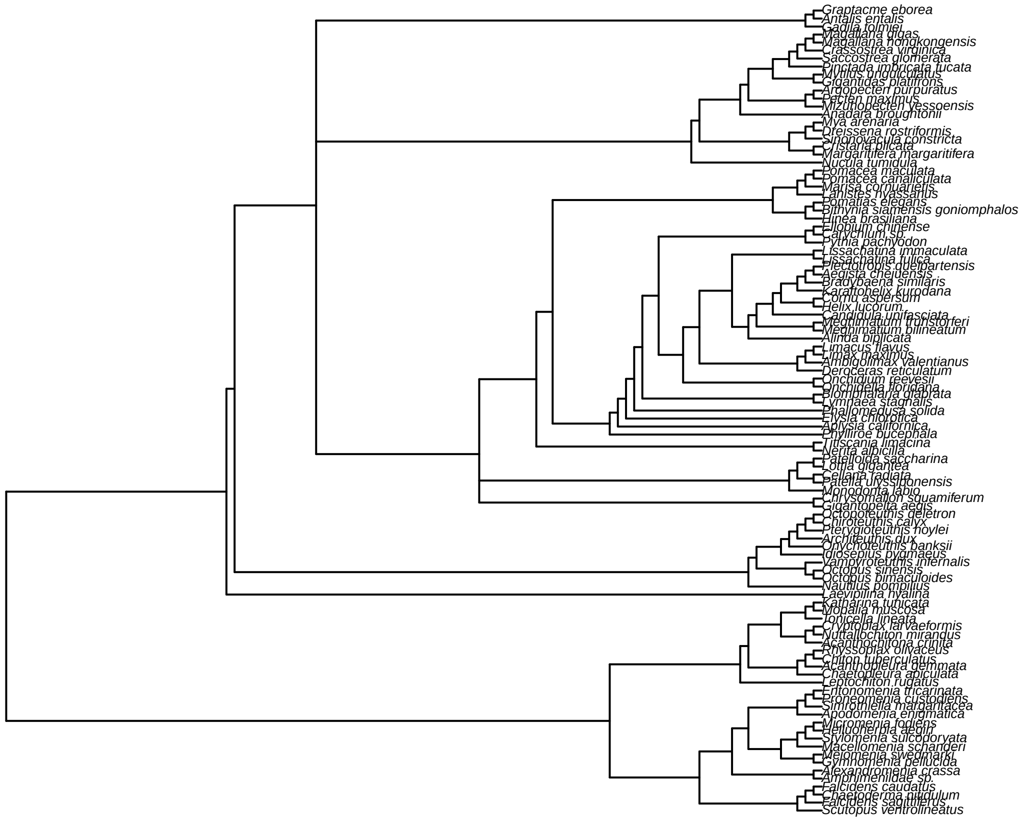
Study of expansions/contractions of AQP gene family under a Possion regression model in R

Set the Environment

```r
# Load packages
require(phylolm)
require(phytools)
require(reshape2)
require(phangorn)
require(ggplot2)
require(cowplot)
```

```r
# Load species tree in newick format
speciestree <- read.tree("speciestree_newik.txt")
plot(speciestree, cex=0.4, no.margin = T)
```



Non-parametric Bootstrapping for bracnch length simulation

We will simulate 1000 possible scenarios with different branch length for the species tree using the function “pb_edgelength” from [??????: L. Revell’s method (http://blog.phytools.org/2015/04/sampling-edge-lengths-under-yule-process.html)]. This function set branch length to a given species tree based on the Yule process.

```
# Set pb_edgelength function
pb_edgelength <- function(tree,b=1,plot=TRUE,...){
  ll <- rexp(n=Ntip(tree)-1,rate=2:Ntip(tree)*b)
  tree$edge.length <- rep(0,nrow(tree$edge))
  live.nodes <- Descendants(tree,Ntip(tree)+1,"children")
  tips <- vector()
  for(i in 1:length(ll)){
    tips <- c(tips,live.nodes[live.nodes<=Ntip(tree)])
    live.nodes <- setdiff(live.nodes,tips)
    ii <- which(tree$edge[,2]%in%c(live.nodes,tips))
    tree$edge.length[ii] <- tree$edge.length[ii]+ll[i]
    node <- if(length(live.nodes)<=1) live.nodes else
      sample(live.nodes,1) ## choose one node
    live.nodes <- c(setdiff(live.nodes,node),
                    Descendants(tree,node,"children"))
    if(plot) plotTree(tree,...)
  }
  tree
}
```

```
# Create 1000 replicates of the species tree
simulated_speciestrees <- rep(list(speciestree),1000)

# Use "pb_edgelength" to simulate different branch lengths for each replicate
simulated_speciestrees_bl <- lapply(simulated_speciestrees, pb_edgelength,plot=F)

# Change class to multiPhylo for further analyses
class(simulated_speciestrees_bl) <- "multiPhylo"
```

AQP1-like Analysis

Marine/Non-Marine classification

We will load data containing variables in the model (AQP1-like count and habitat classification per taxon) in tsv format. Then, we will compare the amount of AQP1-like copies between Marine and Non-Marine species considering their phylogenetic relationships. To do so, we will fit a Poisson regression model to the AQP1-like count data and the 1000 sampled trees using phyloglm()

```
# Load data
marine_data_1 <- read.table("AQP1_marinenonmarine_model.tsv", h=T, row.names = 1)

# Fit the model
marine_model_results_1 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP1_like~Habitat,
                                               marine_data_1, phy=x,
                                               method = "poisson_GEE")})
head(marine_model_results_1, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP1_like ~ Habitat, data = marine_data_1,
##     phy = x, method = "poisson_GEE")
##
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##       (Intercept) HabitatNon-Marine
##         1.3684456         0.3181062
```

Extract results for each of the 1000 scenarios

We will extract sampling means considering the log-scale relationships between the response (AQP1-like count) and the explanatory (Habitat) variables in the Poisson regression model. The mathematical function that explains this relationship looks like:

$$y = e^{\alpha+\beta(x)} = e^{\alpha} + e^{\beta * x}$$

```
# Extract sampling means
means_marine_1 <- exp(sapply(marine_model_results_1,
                             function(x){x$coefficients[1]}))
means_nonmarine_1 <- exp(sapply(marine_model_results_1,
                                function(x){x$coefficients[1]})
                         + sapply(marine_model_results_1,
                                  function(x){x$coefficients[2]}))
```

We will also extract p-values for each of the 1000 scenarios, which explain whether the differences between Marine and Non-Marine means are significant.

```
# Extract sampling p-values
marine_p_values_1 <- sapply(lapply(marine_model_results_1, summary),
                            function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

Mean value of sampling means for AQP1-like in marine and non-marine species.

```
# Phylogenetic mean
marine_means_habitat_1 = c("Marine" = mean(means_marine_1),
                    "Non-Marine" = mean(means_nonmarine_1))
```

	Mean
Marine	3.451217

	Mean
Non-Marine	5.005511

Median values of sampling means for AQP1-like in marine and non-marine species.

```
# Phylogenetic median
marine_medians_habitat_1 <- c("Marine"=median(means_marine_1),
                       "Non-Marine"=median(means_nonmarine_1))
```

	Median
Marine	3.427115
Non-Marine	5.011499

Compare model results to those without considering phylogenetic relationships.

```
# Non-phylogenetic mean
marine_nonphylogenetic_means_1 <- tapply(marine_data_1$AQP1_like,
                                         marine_data_1$Habitat, mean)
```

	Mean
Marine	3.800000
Non-Marine	6.666667

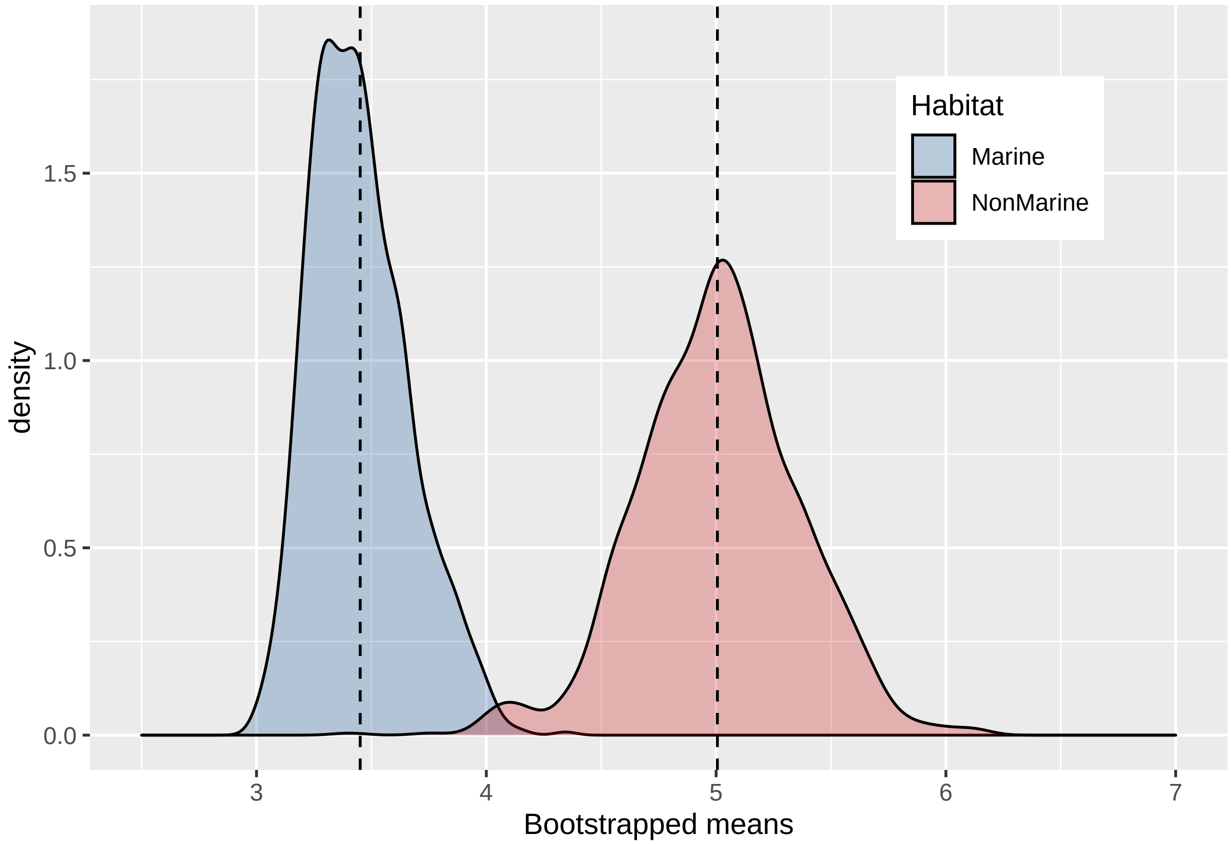
```
# Non-phylogenetic median
marine_nonphylogenetic_median_1 <- tapply(marine_data_1$AQP1_like,
                                          marine_data_1$Habitat, median)
```

	Median
Marine	3.5
Non-Marine	6.0

Finally, we will plot sampling distribution of Marine and Non-Marine means.

```
# Create dataframe
marine_df_means_1 <- data.frame(Marine = means_marine_1,
                                NonMarine = means_nonmarine_1)

# Melt Marine and Non-Marine means data
marine_melted_df_means_1 = melt(marine_df_means_1, value.name = "Mean",
                                variable.name = "Habitat")
```

```
# Plot means data
marine_p_means_1 <- ggplot(marine_melted_df_means_1, aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(2.5, 7) +
  geom_vline(data=marine_df_means_1, aes(xintercept=mean(Marine)),
             linetype="dashed") +
  geom_vline(data=marine_df_means_1, aes(xintercept=mean(NonMarine)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
marine_p_means_1
```



P-values

We will explore sampling distribution of p-values to conclude whether the differences in amount of AQP1-like are statistically significant between Marine and Non-Marine species, taking into account that only p-values smaller than 0.05 show statistical significance:

```
table(marine_p_values_1 < 0.05)
```

p-value < 0.05	count
FALSE	41
TRUE	959

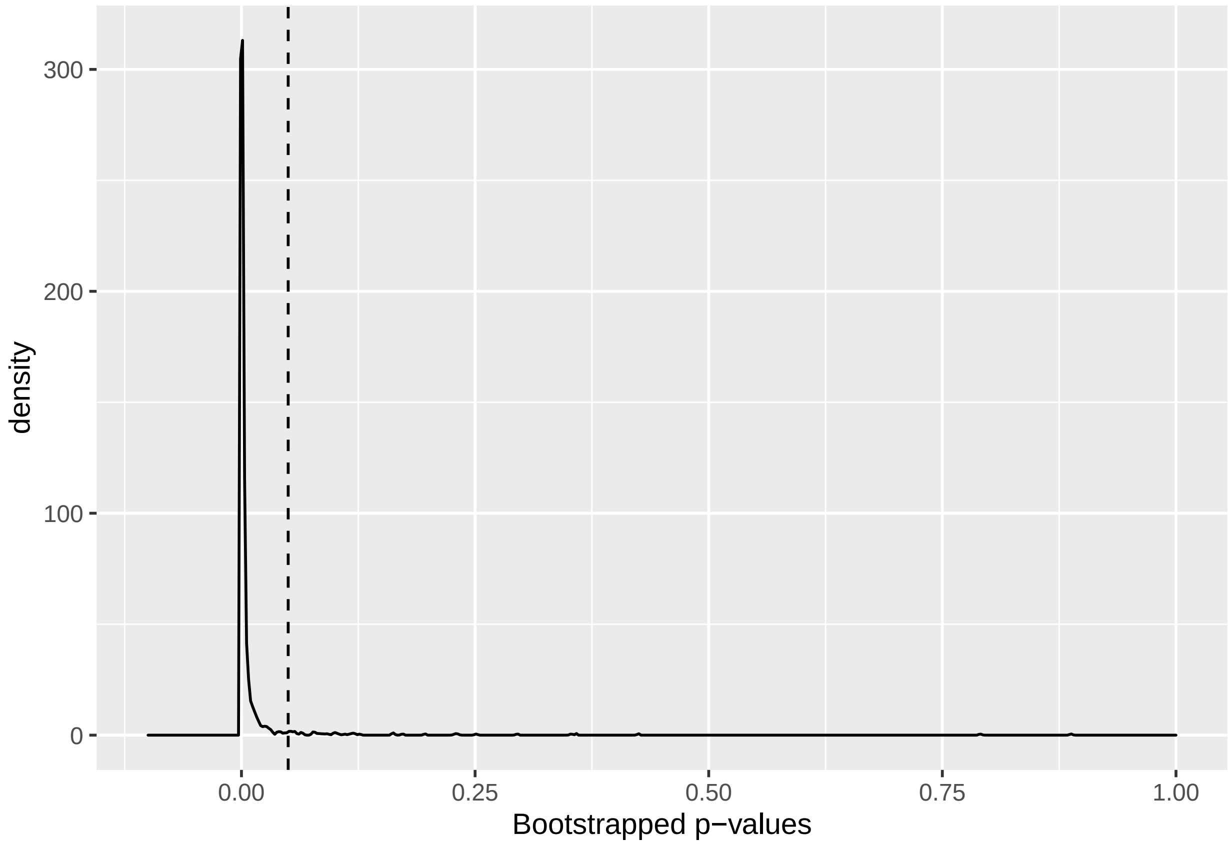
Since it seems that not all bootstrapped p-values are under 0.05, we need to summarize and visualize sampling distribution of p-values.

```r
# Mean of sampling distribution of  p-values
marine_mean_pvalues_1 <-  mean(marine_p_values_1)

# Median of sampling distribution of  p-values
marine_median_pvalues_1 <-  median(marine_p_values_1)

# Extract the 95% confidence interval for sampling distribution of  p-values
marine_sorted_pvalues_1 <- sort(marine_p_values_1)
marine_lower_limit_1 <- marine_sorted_pvalues_1[26]
marine_upper_limit_1 <- marine_sorted_pvalues_1[975]
```

Mean	Median	Lower_limit	Upper_limit
0.009497	0.0004603	1.8e-06	0.0825906

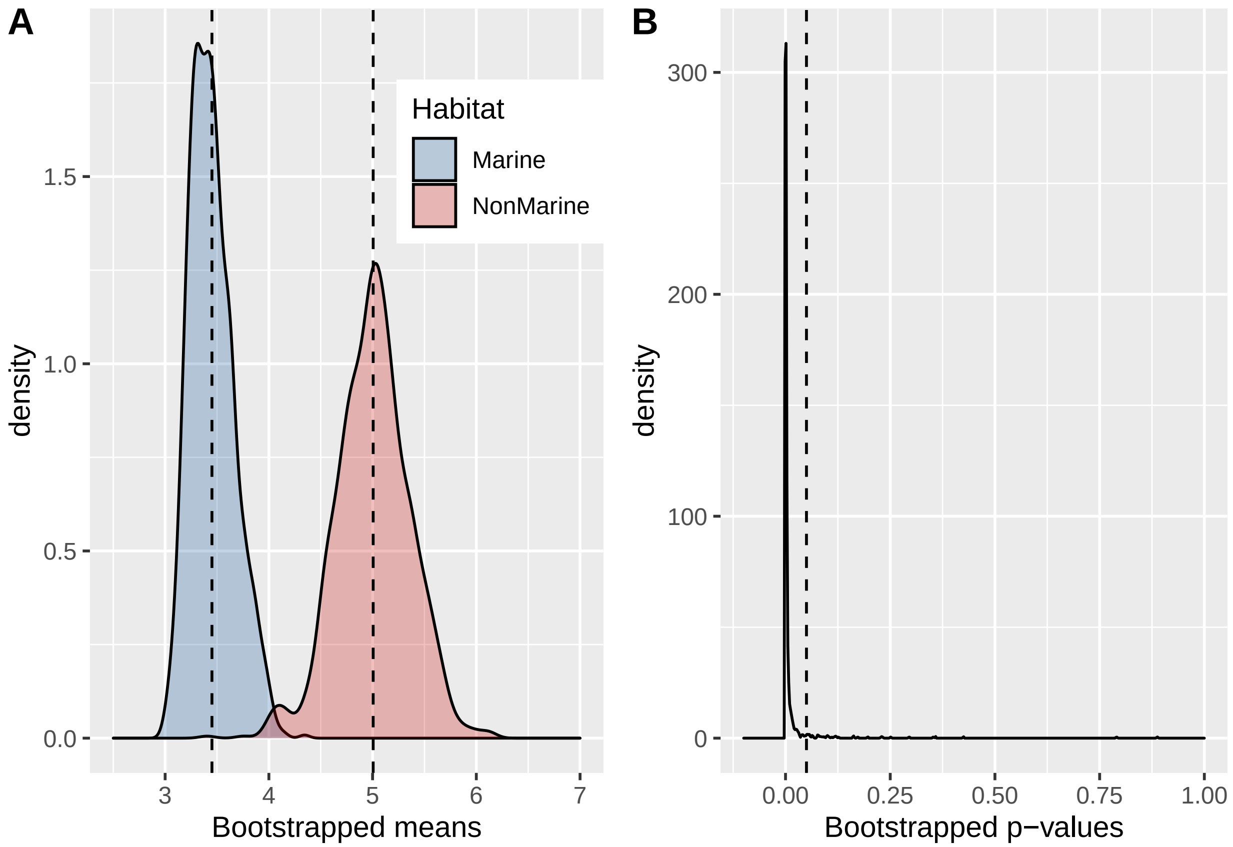
Finally, we will plot sampling distribution of p-values.

```r
# Create the dataframe
marine_df_pvalue_1 <- data.frame(Pvalue = marine_p_values_1)

# Plot bootstrapped p-values distribution
marine_p_pvalue_1 <- ggplot(marine_df_pvalue_1, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

marine_p_pvalue_1
```



Analyses summary:

```
plot_grid(marine_p_means_1, marine_p_pvalue_1, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```



Terrestrial/Non-Terrestrial classification

```
# Load data
terrestrial_data_1 <- read.table("AQP1_terrestrialnonterrestrial_model.tsv",
                                 h=T, row.names = 1)

# Fit the model
terrestrial_model_results_1 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP1_like~Habitat,
                                               terrestrial_data_1, phy=x,
                                               method = "poisson_GEE")})
head(terrestrial_model_results_1, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP1_like ~ Habitat, data = terrestrial_data_1,
##     phy = x, method = "poisson_GEE")
##
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##        (Intercept) HabitatTerrestrial
##         1.49117960         0.01820524
```

Extract results for each of the 1000 scenarios

```
# Extract sampling means
means_nonterrestrial_1 <- exp(sapply(terrestrial_model_results_1,
                                     function(x){x$coefficients[1]}))
means_terrestrial_1 <- exp(sapply(terrestrial_model_results_1,
                                  function(x){x$coefficients[1]})
                         + sapply(terrestrial_model_results_1,
                                  function(x){x$coefficients[2]}))
```

```
# Extract sampling p-values
terrestrial_p_values_1 <- sapply(lapply(terrestrial_model_results_1, summary),
                                 function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

```
# Phylogenetic mean
terrestrial_means_habitat_1 = c("Non-Terrestrial" = mean(means_nonterrestrial_1),
                    "Terrestrial" = mean(means_terrestrial_1))
```

	Mean
Non-Terrestrial	3.889458
Terrestrial	3.940548

```
# Phylogenetic median
terrestrial_medians_habitat_1 <- c("Non-Terrestrial"=median(means_nonterrestrial_1),
                       "Terrestrial"=median(means_terrestrial_1))
```

	Median
Non-Terrestrial	3.868995
Terrestrial	3.906468

```
# Non-phylogenetic mean
terrestrial_nonphylogenetic_means_1 <- tapply(terrestrial_data_1$AQP1_like,
                                              terrestrial_data_1$Habitat, mean)
```

	Mean
Aquatic	4.617284
Terrestrial	7.800000

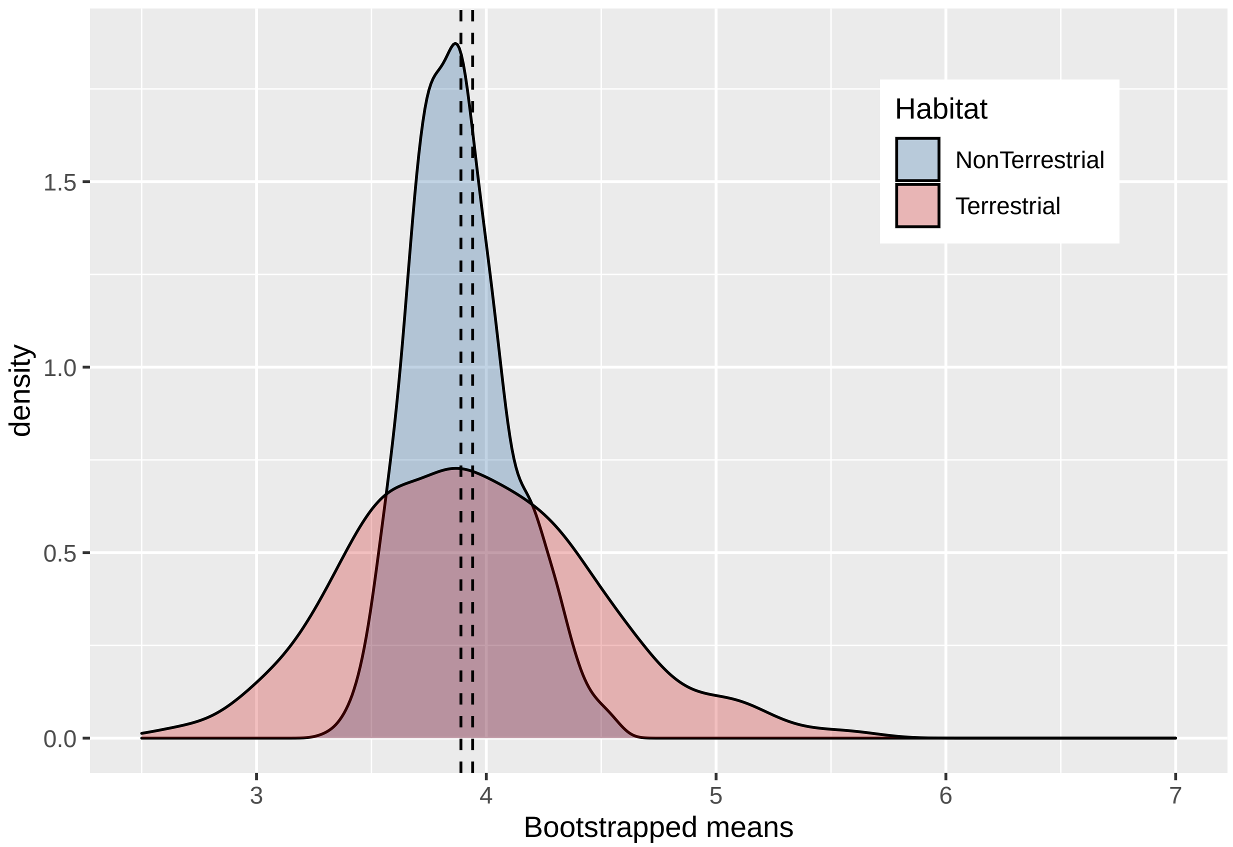
```
# Non-phylogenetic median
terrestrial_nonphylogenetic_median_1 <- tapply(terrestrial_data_1$AQP1_like,
                                               terrestrial_data_1$Habitat, median)
```

	Median
Aquatic	4.0
Terrestrial	8.5

```
# Create dataframe
terrestrial_df_means_1 <- data.frame(NonTerrestrial = means_nonterrestrial_1,
                                     Terrestrial = means_terrestrial_1)

# Melt Marine and Non-Marine means data
terrestrial_melted_df_means_1 = melt(terrestrial_df_means_1,
                                     value.name = "Mean", variable.name = "Habitat")
```

```
# Plot means data
terrestrial_p_means_1 <- ggplot(terrestrial_melted_df_means_1,
                                aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(2.5, 7) +
  geom_vline(data=terrestrial_df_means_1, aes(xintercept=mean(NonTerrestrial)),
             linetype="dashed") +
  geom_vline(data=terrestrial_df_means_1, aes(xintercept=mean(Terrestrial)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
terrestrial_p_means_1
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



P-values

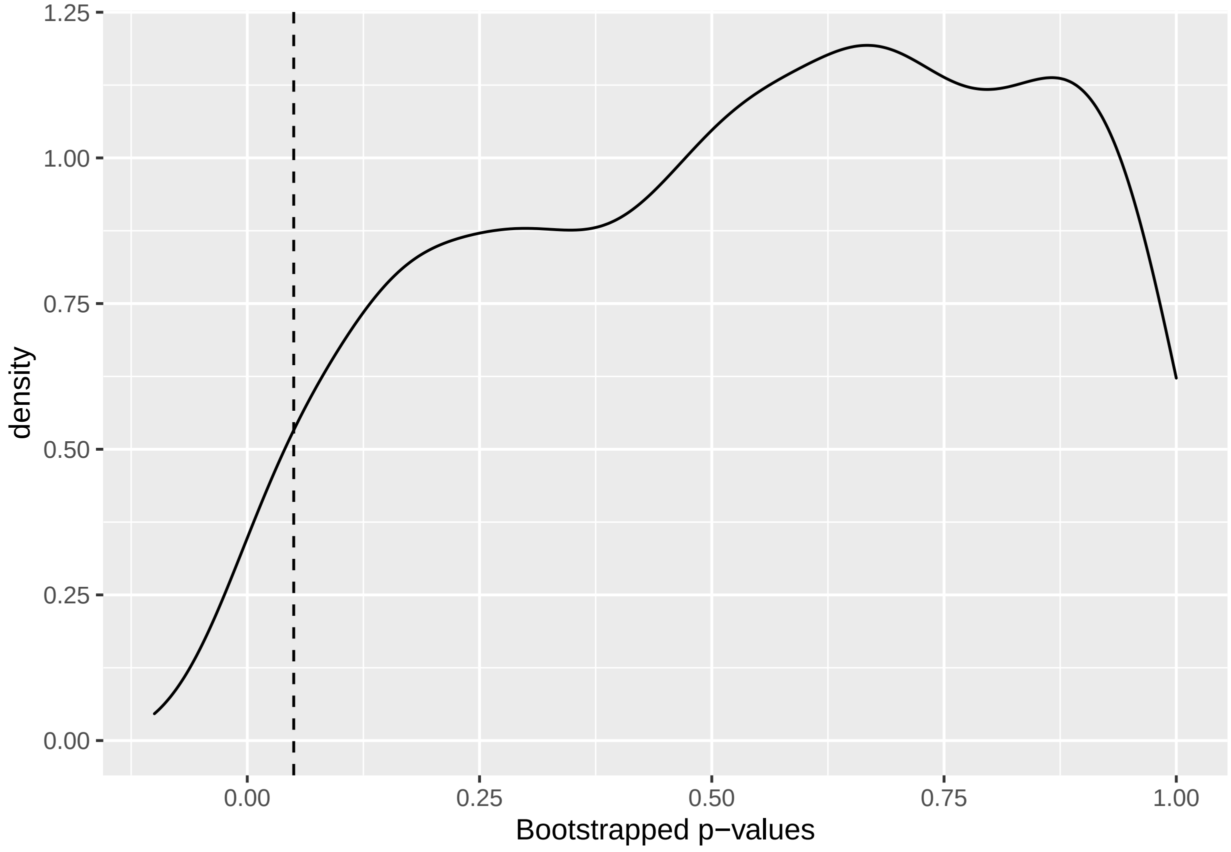
```
table(terrestrial_p_values_1 < 0.05)
```

p-value < 0.05	count
FALSE	964
TRUE	36

```
# Mean of sampling distribution of  p-values
terrestrial_mean_pvalues_1 <-  mean(terrestrial_p_values_1)

# Median of sampling distribution of  p-values
terrestrial_median_pvalues_1 <-  median(terrestrial_p_values_1)

# Extract the 95% confidence interval for sampling distribution of  p-values
terrestrial_sorted_pvalues_1 <- sort(terrestrial_p_values_1)
terrestrial_lower_limit_1 <- terrestrial_sorted_pvalues_1[26]
terrestrial_upper_limit_1 <- terrestrial_sorted_pvalues_1[975]
```

Mean	Median	Lower_limit	Upper_limit
0.5498899	0.5727568	0.0265037	0.9810499

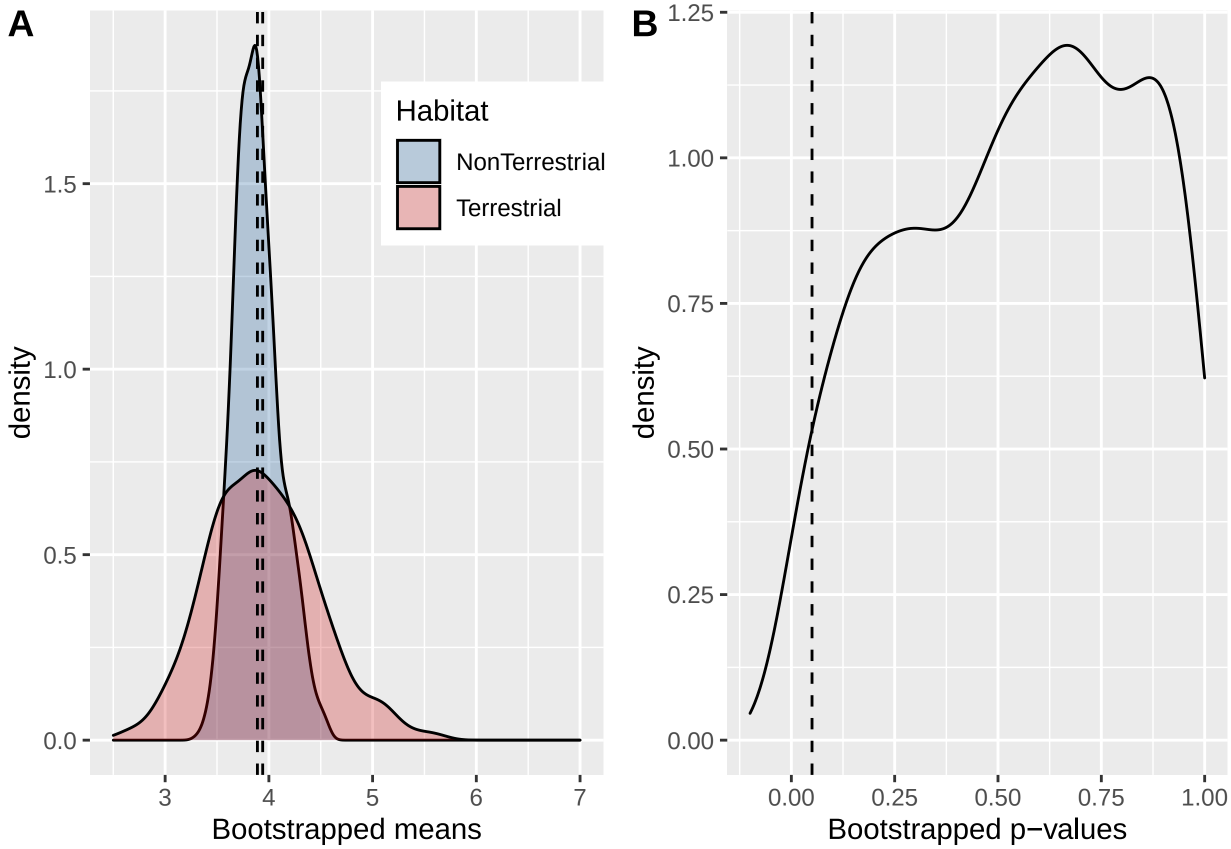
```
# Create the dataframe
terrestrial_df_pvalue_1 <- data.frame(Pvalue = terrestrial_p_values_1)

# Plot bootstrapped p-values distribution
terrestrial_p_pvalue_1 <- ggplot(terrestrial_df_pvalue_1, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

terrestrial_p_pvalue_1
```



Analyses summary:

```
plot_grid(terrestrial_p_means_1, terrestrial_p_pvalue_1, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



AQP3-like Analysis

Marine/Non-Marine classification

```
# Load data
marine_data_3 <- read.table("AQP3_marinenonmarine_model.tsv", h=T, row.names = 1)

# Fit the model
marine_model_results_3 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP3_like~Habitat,
                                               marine_data_3, phy=x,
                                               method = "poisson_GEE")})
head(marine_model_results_3, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP3_like ~ Habitat, data = marine_data_3,
##     phy = x, method = "poisson_GEE")
##
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##       (Intercept) HabitatNon-Marine
##        0.90340917        0.04990589
```

Extract results for each of the 1000 scenarios

```
# Extract sampling means
means_marine_3 <- exp(sapply(marine_model_results_3,
                             function(x){x$coefficients[1]}))
means_nonmarine_3 <- exp(sapply(marine_model_results_3,
                                function(x){x$coefficients[1]})
                         + sapply(marine_model_results_3,
                                  function(x){x$coefficients[2]}))
```

```
# Extract sampling p-values
marine_p_values_3 <- sapply(lapply(marine_model_results_3, summary),
                            function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

```
# Phylogenetic mean
marine_means_habitat_3 = c("Marine" = mean(means_marine_3),
                    "Non-Marine" = mean(means_nonmarine_3))
```

	Mean
Marine	2.690793
Non-Marine	2.697608

```
# Phylogenetic median
marine_medians_habitat_3 <- c("Marine"=median(means_marine_3),
                       "Non-Marine"=median(means_nonmarine_3))
```

	Median
Marine	2.692615
Non-Marine	2.682093

```
# Non-phylogenetic mean
marine_nonphylogenetic_means_3 <- tapply(marine_data_3$AQP3_like,
                                         marine_data_3$Habitat, mean)
```

	Mean
Marine	2.380000
Non-Marine	2.294118

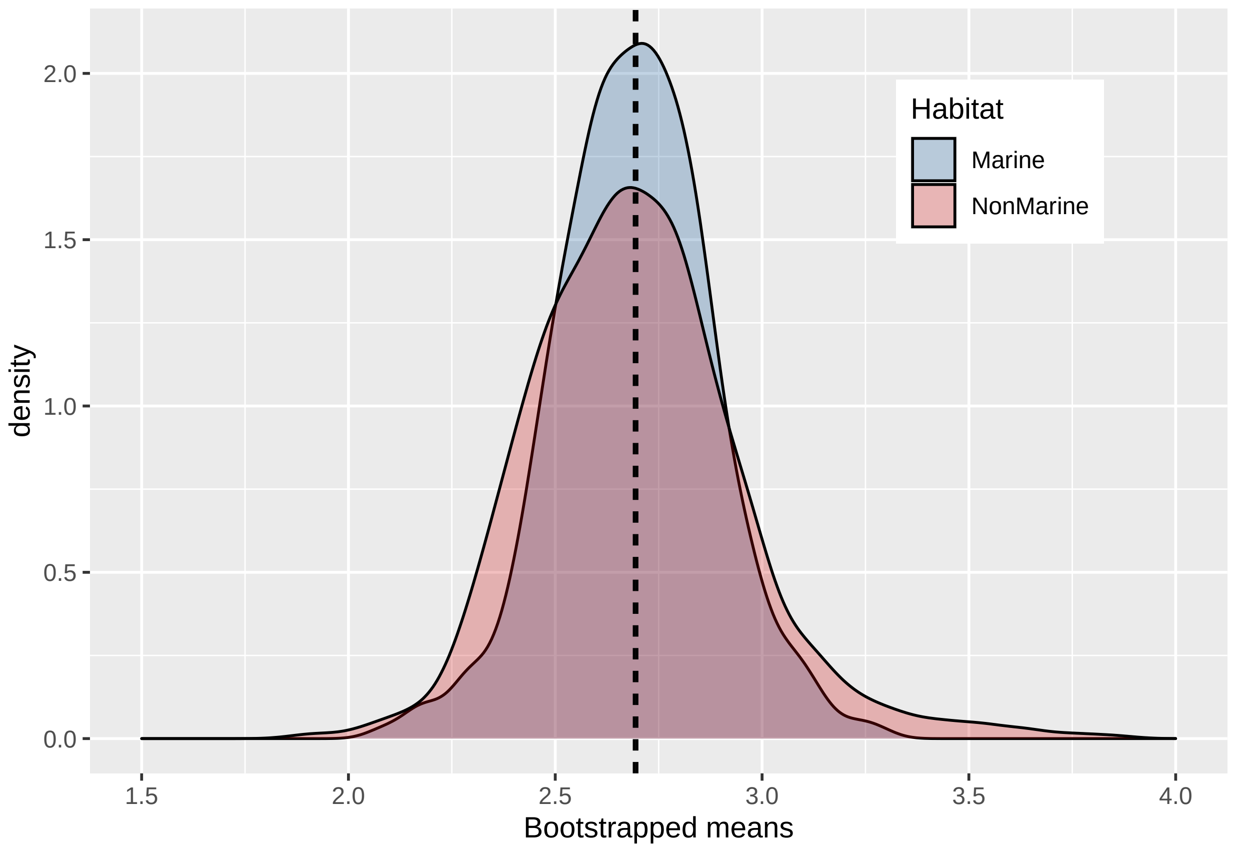
```r
# Non-phylogenetic median
marine_nonphylogenetic_median_3 <- tapply(marine_data_3$AQP3_like,
                                          marine_data_3$Habitat, median)
```

	Median
Marine	2
Non-Marine	2

```r
# Create dataframe
marine_df_means_3 <- data.frame(Marine = means_marine_3,
                                NonMarine = means_nonmarine_3)

# Melt Marine and Non-Marine means data
marine_melted_df_means_3 = melt(marine_df_means_3, value.name = "Mean",
                                variable.name = "Habitat")
```

```r
# Plot means data
marine_p_means_3 <- ggplot(marine_melted_df_means_3, aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(1.5, 4) +
  geom_vline(data=marine_df_means_3, aes(xintercept=mean(Marine)),
             linetype="dashed") +
  geom_vline(data=marine_df_means_3, aes(xintercept=mean(NonMarine)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
marine_p_means_3
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



P-values

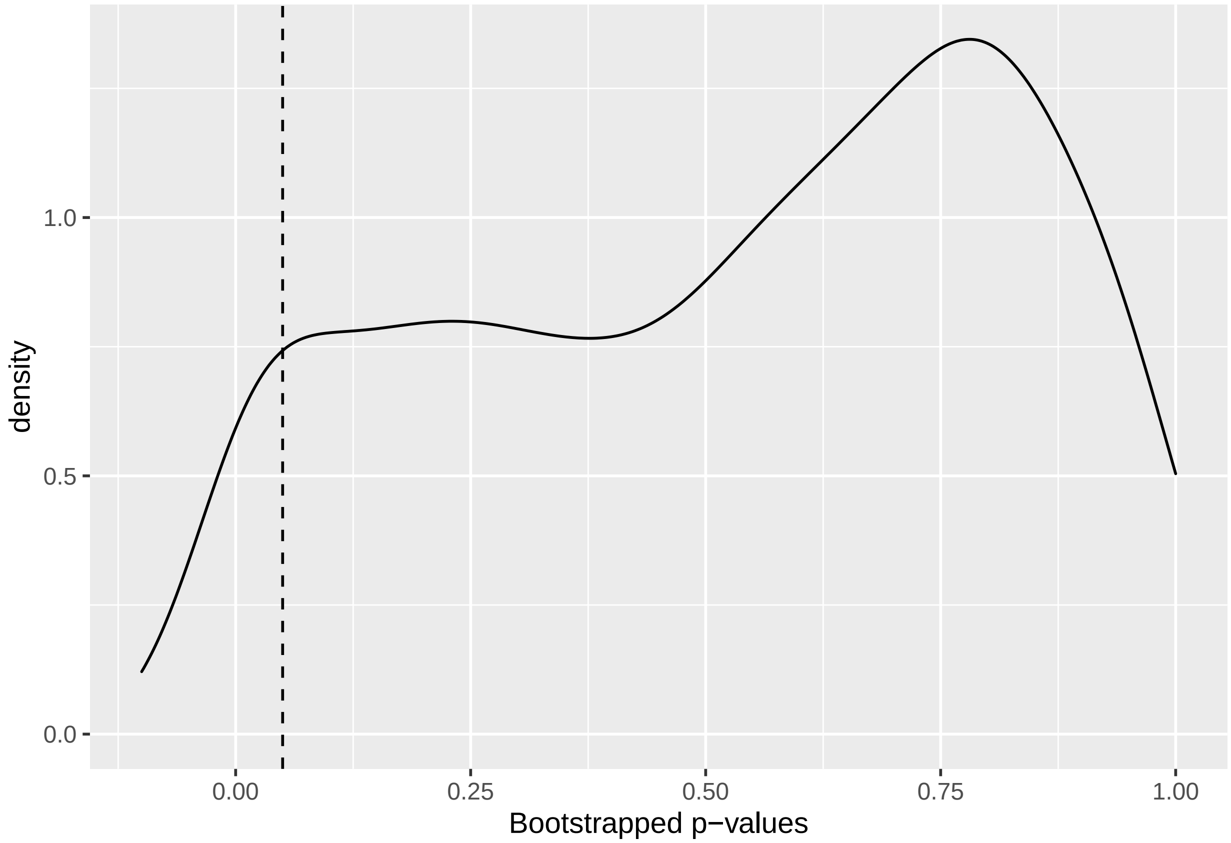
```
table(marine_p_values_3 < 0.05)
```

p-value < 0.05	count
FALSE	922
TRUE	78

```
# Mean of sampling distribution of  p-values
marine_mean_pvalues_3 <-  mean(marine_p_values_3)

# Median of sampling distribution of  p-values
marine_median_pvalues_3 <-  median(marine_p_values_3)

# Extract the 95% confidence interval for sampling distribution of  p-values
marine_sorted_pvalues_3 <- sort(marine_p_values_3)
marine_lower_limit_3 <- marine_sorted_pvalues_3[26]
marine_upper_limit_3 <- marine_sorted_pvalues_3[975]
```

Mean	Median	Lower_limit	Upper_limit
0.5320279	0.5809523	0.0026383	0.9729916

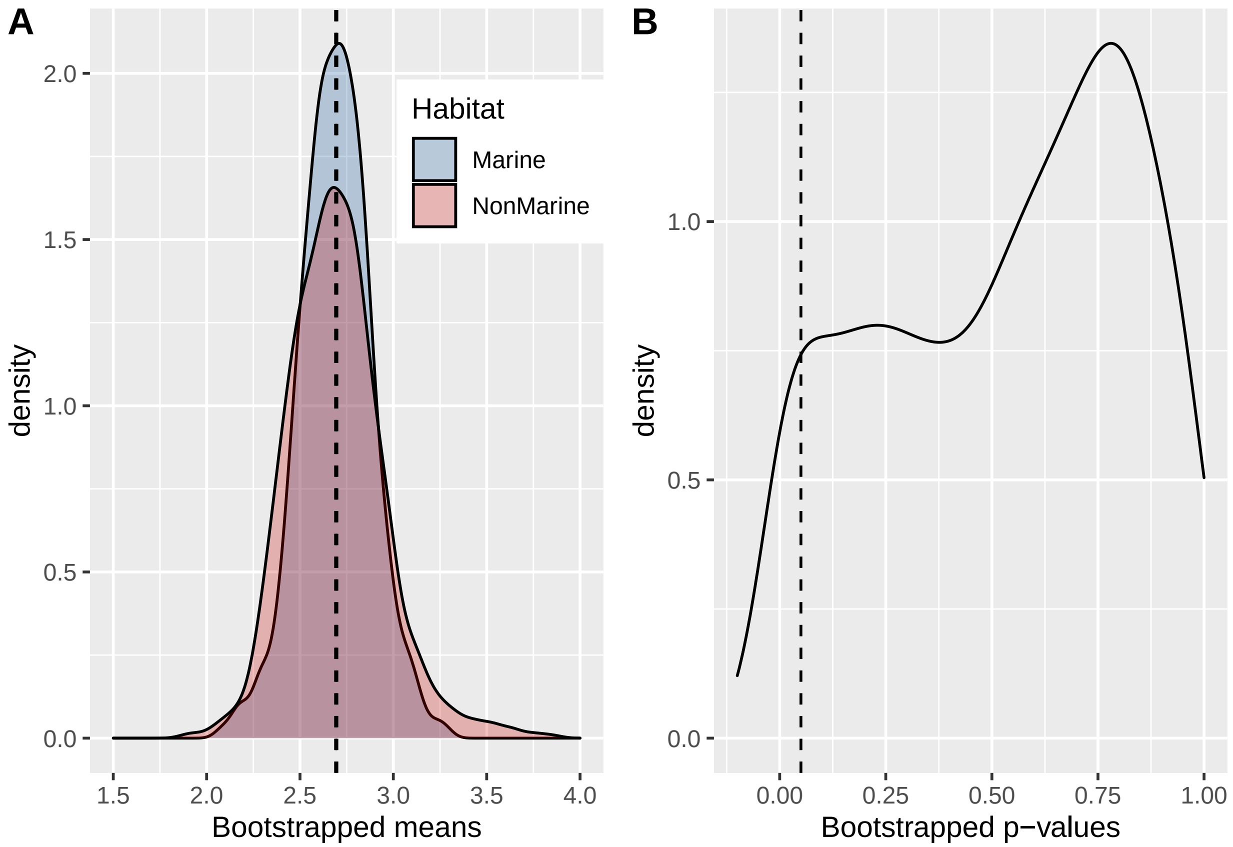
```
# Create the dataframe
marine_df_pvalue_3 <- data.frame(Pvalue = marine_p_values_3)

# Plot bootstrapped p-values distribution
marine_p_pvalue_3 <- ggplot(marine_df_pvalue_3, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

marine_p_pvalue_3
```



Analyses summary:

```
plot_grid(marine_p_means_3, marine_p_pvalue_3, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



Terrestrial/Non-Terrestrial classification

```
# Load data
terrestrial_data_3 <- read.table("AQP3_terrestrialnonterrestrial_model.tsv",
                                 h=T, row.names = 1)

# Fit the model
terrestrial_model_results_3 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP3_like~Habitat,
                                               terrestrial_data_3, phy=x,
                                               method = "poisson_GEE")})
head(terrestrial_model_results_3, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP3_like ~ Habitat, data = terrestrial_data_3,
##     phy = x, method = "poisson_GEE")
##
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##        (Intercept) HabitatTerrestrial
##          0.9221440         -0.1299878
```

Extract results for each of the 1000 scenarios

```
# Extract sampling means
means_nonterrestrial_3 <- exp(sapply(terrestrial_model_results_3,
                                     function(x){x$coefficients[1]}))
means_terrestrial_3 <- exp(sapply(terrestrial_model_results_3,
                                  function(x){x$coefficients[1]})
                         + sapply(terrestrial_model_results_3,
                                  function(x){x$coefficients[2]}))
```

```
# Extract sampling p-values
terrestrial_p_values_3 <- sapply(lapply(terrestrial_model_results_3, summary),
                                 function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

```
# Phylogenetic mean
terrestrial_means_habitat_3 = c("Non-Terrestrial" = mean(means_nonterrestrial_3),
                    "Terrestrial" = mean(means_terrestrial_3))
```

	Mean
Non-Terrestrial	2.690692
Terrestrial	2.195769

```
# Phylogenetic median
terrestrial_medians_habitat_3 <- c("Non-Terrestrial"=median(means_nonterrestrial_3),
                       "Terrestrial"=median(means_terrestrial_3))
```

	Median
Non-Terrestrial	2.692673
Terrestrial	2.196321

```
# Non-phylogenetic mean
terrestrial_nonphylogenetic_means_3 <- tapply(terrestrial_data_3$AQP3_like,
                                              terrestrial_data_3$Habitat, mean)
```

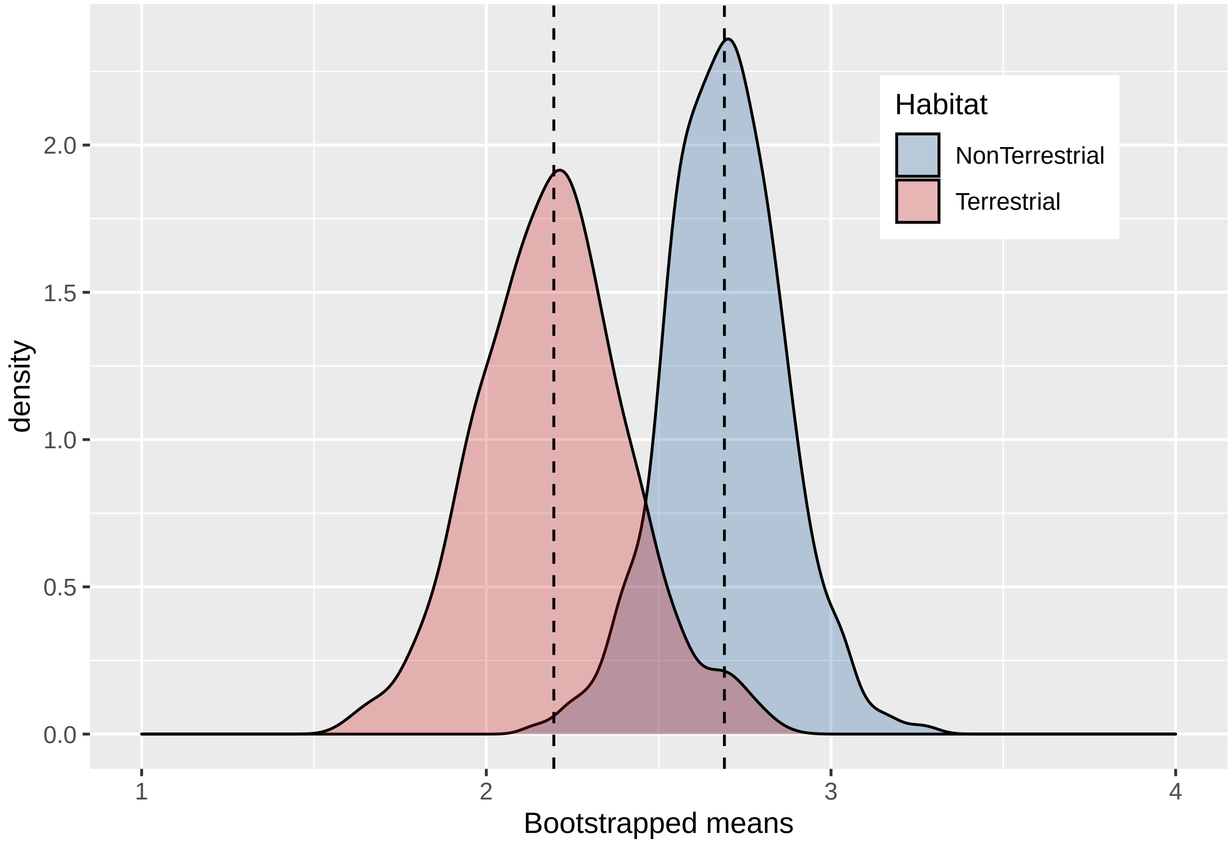
	Mean
Aquatic	2.567901
Terrestrial	1.400000

```
# Non-phylogenetic median
terrestrial_nonphylogenetic_median_3 <- tapply(terrestrial_data_3$AQP3_like,
                                               terrestrial_data_3$Habitat, median)
```

	Median
Aquatic	2
Terrestrial	1

```r
# Create dataframe
terrestrial_df_means_3 <- data.frame(NonTerrestrial = means_nonterrestrial_3,
                                     Terrestrial = means_terrestrial_3)

# Melt Marine and Non-Marine means data
terrestrial_melted_df_means_3 = melt(terrestrial_df_means_3,
                                     value.name = "Mean", variable.name = "Habitat")
```

```r
# Plot means data
terrestrial_p_means_3 <- ggplot(terrestrial_melted_df_means_3,
                                aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(1, 4) +
  geom_vline(data=terrestrial_df_means_3, aes(xintercept=mean(NonTerrestrial)),
             linetype="dashed") +
  geom_vline(data=terrestrial_df_means_3, aes(xintercept=mean(Terrestrial)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
terrestrial_p_means_3
```



P-values

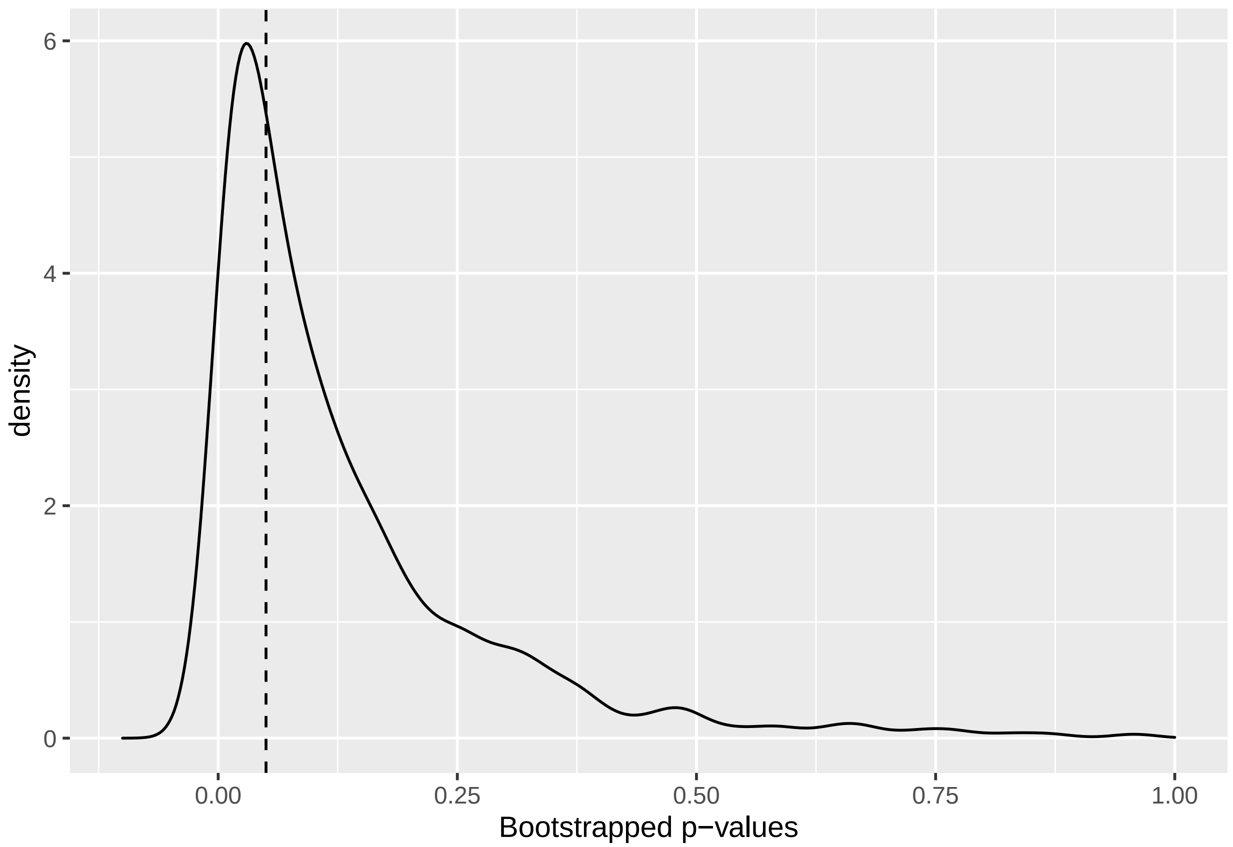
```
table(terrestrial_p_values_3 < 0.05)
```

p-value < 0.05	count
FALSE	634
TRUE	366

```
# Mean of sampling distribution of  p-values
terrestrial_mean_pvalues_3 <-  mean(terrestrial_p_values_3)

# Median of sampling distribution of  p-values
terrestrial_median_pvalues_3 <-  median(terrestrial_p_values_3)

# Extract the 95% confidence interval for sampling distribution of  p-values
terrestrial_sorted_pvalues_3 <- sort(terrestrial_p_values_3)
terrestrial_lower_limit_3 <- terrestrial_sorted_pvalues_3[26]
terrestrial_upper_limit_3 <- terrestrial_sorted_pvalues_3[975]
```

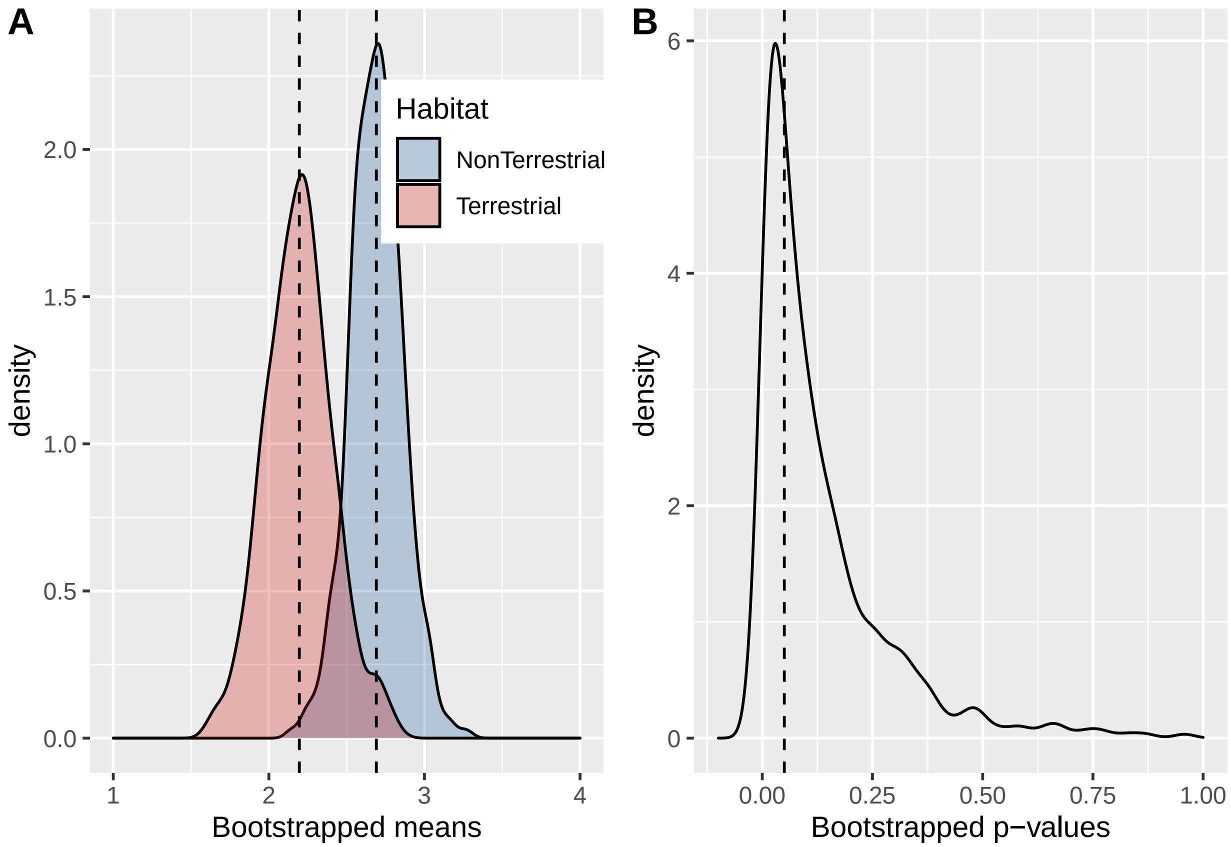
Mean	Median	Lower_limit	Upper_limit
0.1313033	0.0796738	0.0032968	0.5822021

```
# Create the dataframe
terrestrial_df_pvalue_3 <- data.frame(Pvalue = terrestrial_p_values_3)

# Plot bootstrapped p-values distribution
terrestrial_p_pvalue_3 <- ggplot(terrestrial_df_pvalue_3, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

terrestrial_p_pvalue_3
```



Analyses summary:

```
plot_grid(terrestrial_p_means_3, terrestrial_p_pvalue_3, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```



AQP8-like Analyses

Marine/Non-Marine classification

```
# Load data
marine_data_8 <- read.table("AQP8_marinenonmarine_model.tsv", h=T, row.names = 1)

# Fit the model
marine_model_results_8 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP8_like~Habitat,
                                               marine_data_8, phy=x,
                                               method = "poisson_GEE")})
```

```
## Warning in phyloglm(AQP8_like ~ Habitat, marine_data_8, phy = x, method = "poisson_GEE"): phyloglm fa
```

```
head(marine_model_results_8, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP8_like ~ Habitat, data = marine_data_8,
##     phy = x, method = "poisson_GEE")
##
```

```
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##       (Intercept) HabitatNon-Marine
##        0.89447486        0.08448404
```

Extract results for each of the 1000 scenarios

```
# Extract sampling means
means_marine_8 <- exp(sapply(marine_model_results_8,
                             function(x){x$coefficients[1]}))
means_nonmarine_8 <- exp(sapply(marine_model_results_8,
                                function(x){x$coefficients[1]})
                         + sapply(marine_model_results_8,
                                  function(x){x$coefficients[2]}))
```

```
# Extract sampling p-values
marine_p_values_8 <- sapply(lapply(marine_model_results_8, summary),
                            function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

```
# Phylogenetic mean
marine_means_habitat_8 = c("Marine" = mean(means_marine_8),
                    "Non-Marine" = mean(means_nonmarine_8))
```

	Mean
Marine	2.171854
Non-Marine	2.571623

```
# Phylogenetic median
marine_medians_habitat_8 <- c("Marine"=median(means_marine_8),
                       "Non-Marine"=median(means_nonmarine_8))
```

	Median
Marine	2.165946
Non-Marine	2.564994

```
# Non-phylogenetic mean
marine_nonphylogenetic_means_8 <- tapply(marine_data_8$AQP8_like,
                                         marine_data_8$Habitat, mean)
```

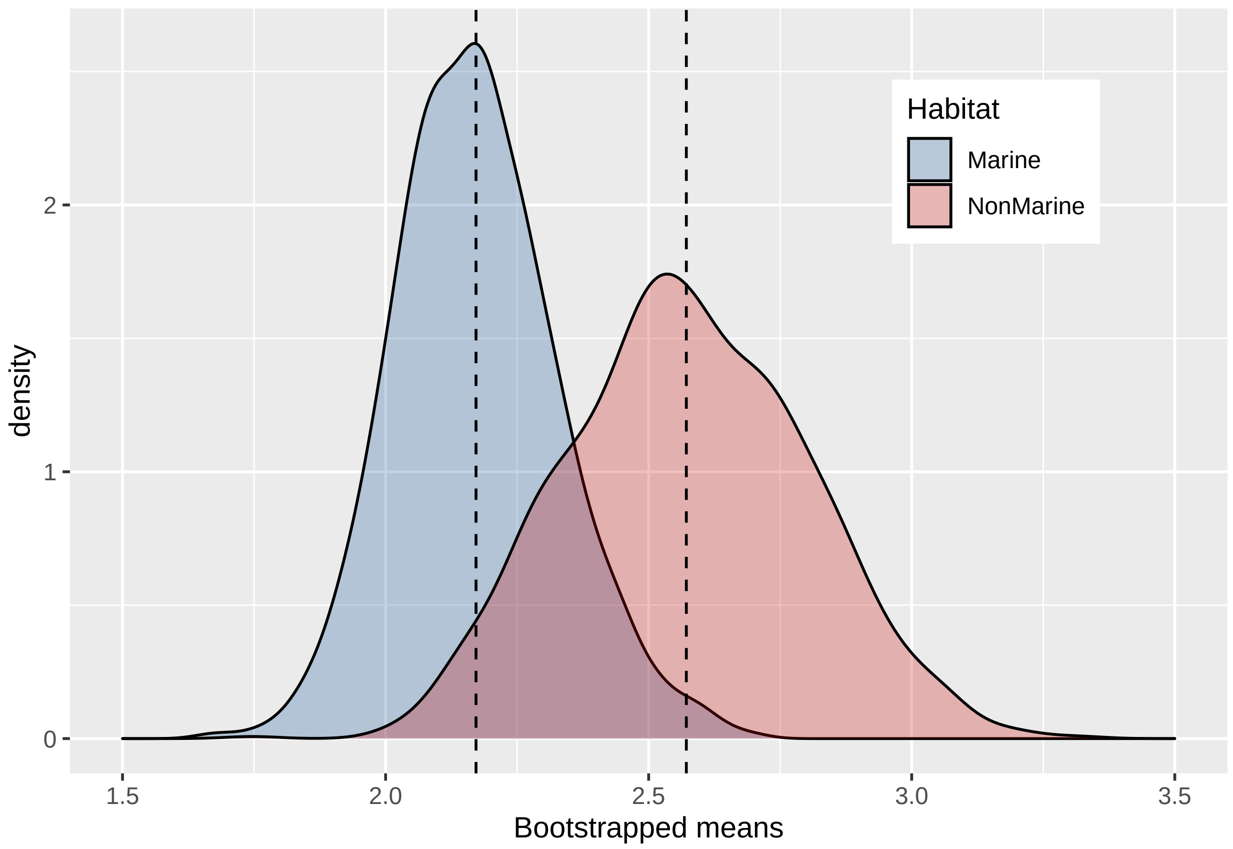
	Mean
Marine	2.720000
Non-Marine	3.078431

```r
# Non-phylogenetic median
marine_nonphylogenetic_median_8 <- tapply(marine_data_8$AQP8_like,
                                          marine_data_8$Habitat, median)
```

	Median
Marine	2
Non-Marine	3

```r
# Create dataframe
marine_df_means_8 <- data.frame(Marine = means_marine_8,
                                NonMarine = means_nonmarine_8)

# Melt Marine and Non-Marine means data
marine_melted_df_means_8 = melt(marine_df_means_8, value.name = "Mean",
                                variable.name = "Habitat")
```

```r
# Plot means data
marine_p_means_8 <- ggplot(marine_melted_df_means_8, aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(1.5, 3.5) +
  geom_vline(data=marine_df_means_8, aes(xintercept=mean(Marine)),
             linetype="dashed") +
  geom_vline(data=marine_df_means_8, aes(xintercept=mean(NonMarine)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
marine_p_means_8
```



P-values

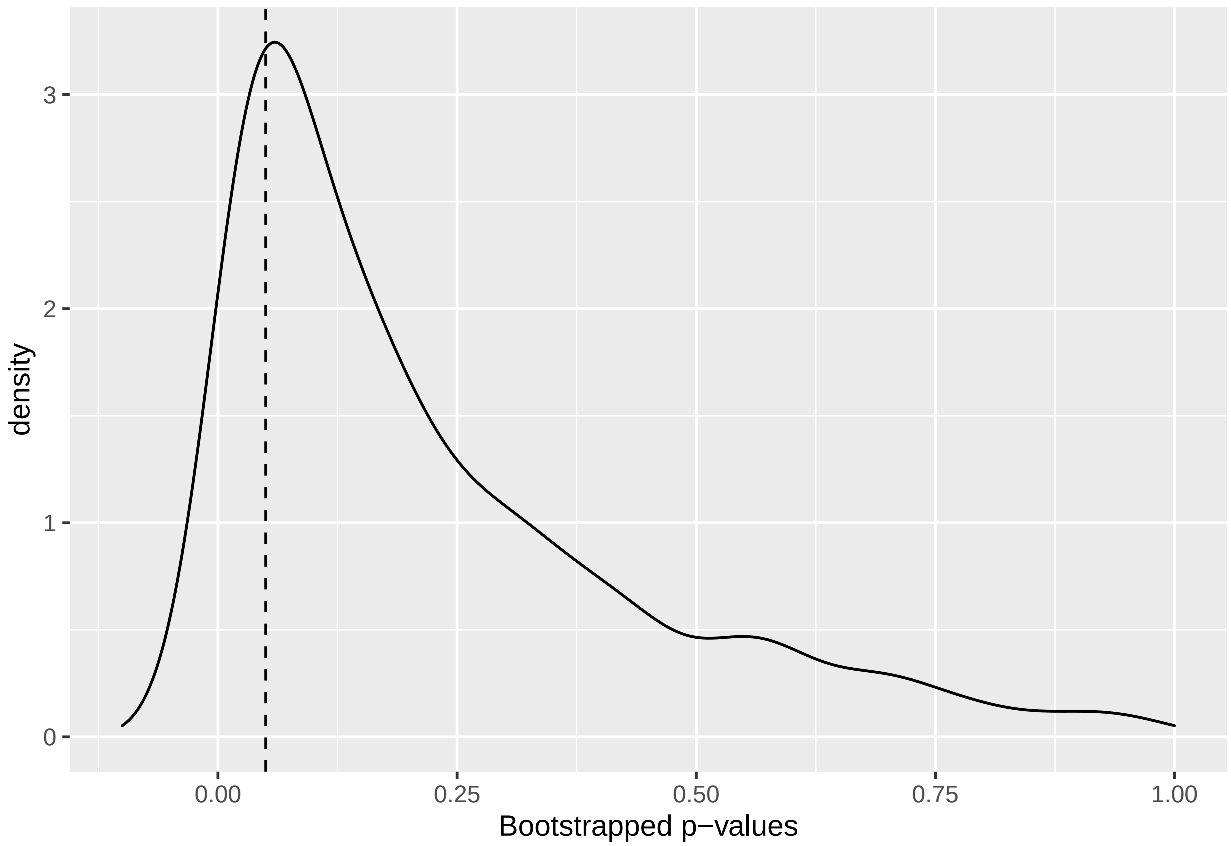
```
table(marine_p_values_8 < 0.05)
```

p-value < 0.05	count
FALSE	784
TRUE	216

```
# Mean of sampling distribution of  p-values
marine_mean_pvalues_8 <-  mean(marine_p_values_8)

# Median of sampling distribution of  p-values
marine_median_pvalues_8 <-  median(marine_p_values_8)

# Extract the 95% confidence interval for sampling distribution of  p-values
marine_sorted_pvalues_8 <- sort(marine_p_values_8)
marine_lower_limit_8 <- marine_sorted_pvalues_8[26]
marine_upper_limit_8 <- marine_sorted_pvalues_8[975]
```

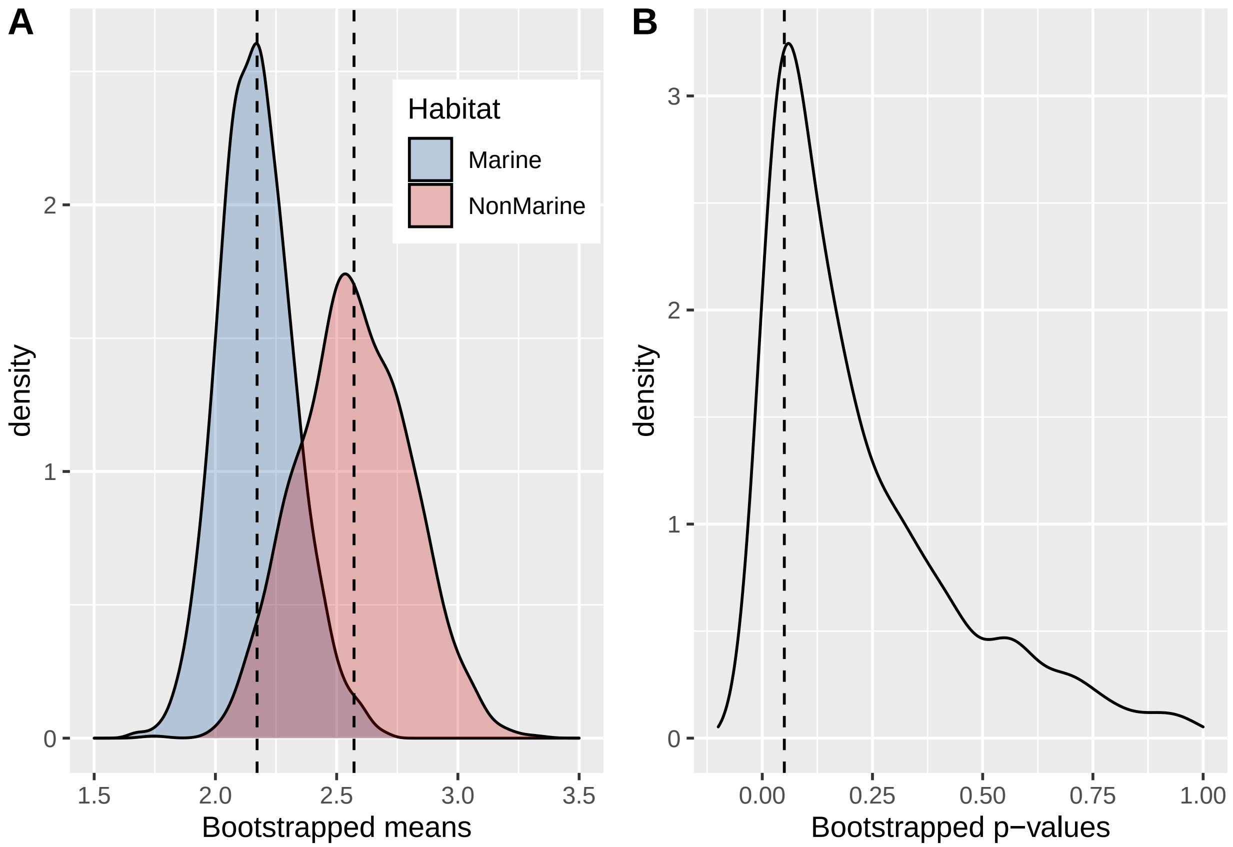
Mean	Median	Lower_limit	Upper_limit
0.2215673	0.1453152	0.0048216	0.7835163

```
# Create the dataframe
marine_df_pvalue_8 <- data.frame(Pvalue = marine_p_values_8)

# Plot bootstrapped p-values distribution
marine_p_pvalue_8 <- ggplot(marine_df_pvalue_8, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

marine_p_pvalue_8
```



Analyses summary:

```
plot_grid(marine_p_means_8, marine_p_pvalue_8, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```



Terrestrial/Non-Terrestrial classification

```
# Load data
terrestrial_data_8 <- read.table("AQP8_terrestrialnonterrestrial_model.tsv",
                                h=T, row.names = 1)

# Fit the model
terrestrial_model_results_8 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP8_like~Habitat,
                                              terrestrial_data_8, phy=x,
                                              method = "poisson_GEE")})
head(terrestrial_model_results_8, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP8_like ~ Habitat, data = terrestrial_data_8,
##     phy = x, method = "poisson_GEE")
##
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##        (Intercept) HabitatTerrestrial
##          0.9161201          0.5517799
```

Extract results for each of the 1000 scenarios

```
# Extract sampling means
means_nonterrestrial_8 <- exp(sapply(terrestrial_model_results_8,
                                     function(x){x$coefficients[1]}))
means_terrestrial_8 <- exp(sapply(terrestrial_model_results_8,
                                  function(x){x$coefficients[1]})
                         + sapply(terrestrial_model_results_8,
                                  function(x){x$coefficients[2]}))
```

```
# Extract sampling p-values
terrestrial_p_values_8 <- sapply(lapply(terrestrial_model_results_8, summary),
                                 function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

```
# Phylogenetic mean
terrestrial_means_habitat_8 = c("Non-Terrestrial" = mean(means_nonterrestrial_8),
                    "Terrestrial" = mean(means_terrestrial_8))
```

	Mean
Non-Terrestrial	2.273694
Terrestrial	3.820235

```
# Phylogenetic median
terrestrial_medians_habitat_8 <- c("Non-Terrestrial"=median(means_nonterrestrial_8),
                       "Terrestrial"=median(means_terrestrial_8))
```

	Median
Non-Terrestrial	2.268447
Terrestrial	3.813666

```
# Non-phylogenetic mean
terrestrial_nonphylogenetic_means_8 <- tapply(terrestrial_data_8$AQP8_like,
                                              terrestrial_data_8$Habitat, mean)
```

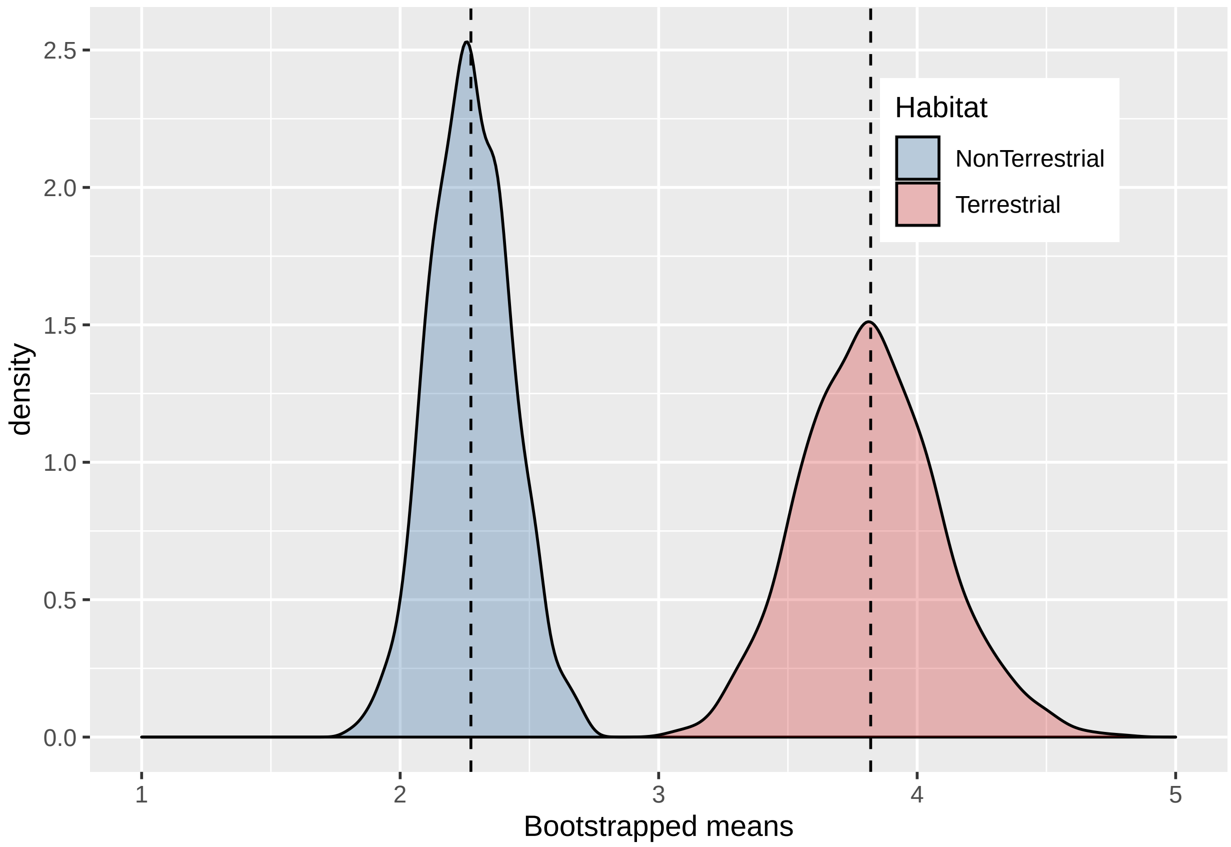
	Mean
Aquatic	2.604938
Terrestrial	4.100000

```
# Non-phylogenetic median
terrestrial_nonphylogenetic_median_8 <- tapply(terrestrial_data_8$AQP8_like,
                                               terrestrial_data_8$Habitat, median)
```

	Median
Aquatic	2
Terrestrial	5

```
# Create dataframe
terrestrial_df_means_8 <- data.frame(NonTerrestrial = means_nonterrestrial_8,
                                     Terrestrial = means_terrestrial_8)

# Melt Marine and Non-Marine means data
terrestrial_melted_df_means_8 = melt(terrestrial_df_means_8,
                                     value.name = "Mean", variable.name = "Habitat")
```

```
# Plot means data
terrestrial_p_means_8 <- ggplot(terrestrial_melted_df_means_8,
                                aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(1, 5) +
  geom_vline(data=terrestrial_df_means_8, aes(xintercept=mean(NonTerrestrial)),
             linetype="dashed") +
  geom_vline(data=terrestrial_df_means_8, aes(xintercept=mean(Terrestrial)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
terrestrial_p_means_8
```



P-values

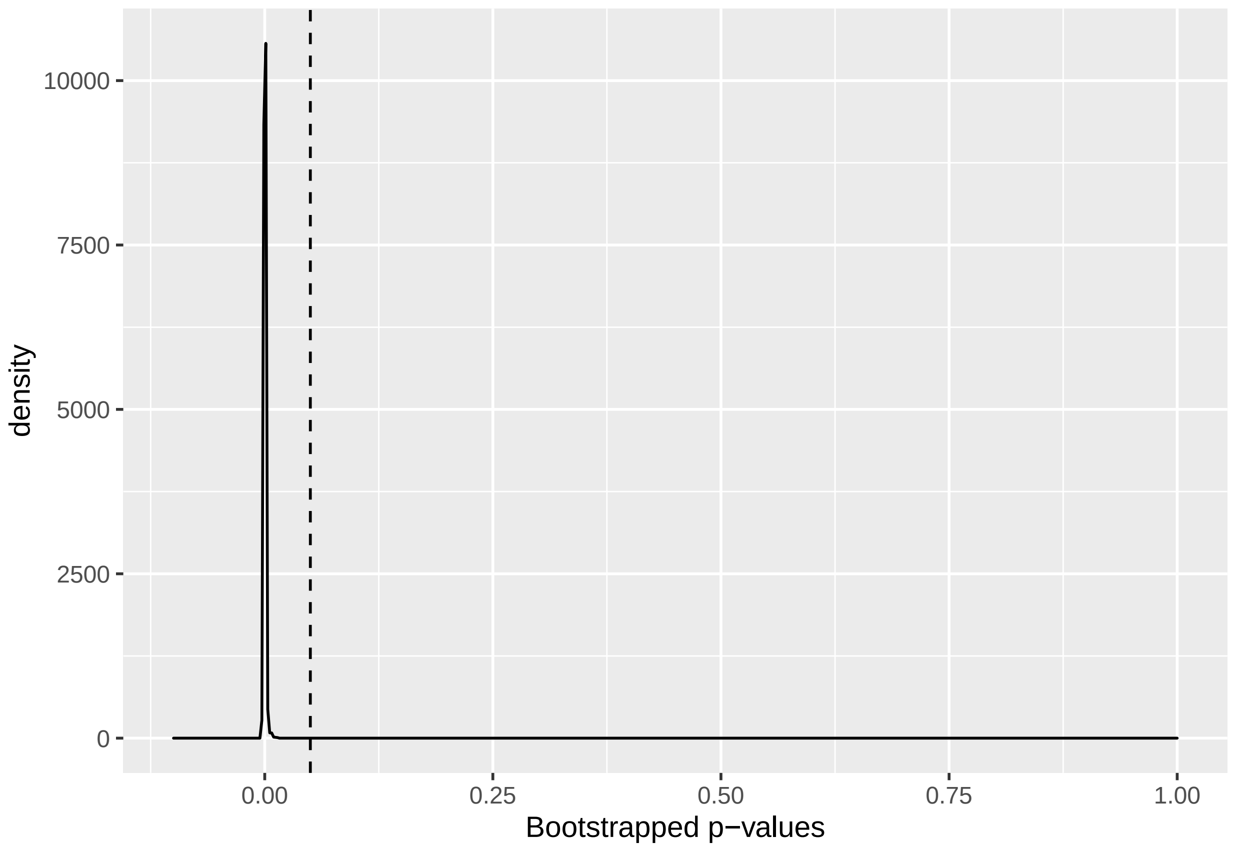
```r
table(terrestrial_p_values_8 < 0.05)
```

p-value < 0.05	count
TRUE	1000

```r
# Mean of sampling distribution of  p-values
terrestrial_mean_pvalues_8 <-  mean(terrestrial_p_values_8)

# Median of sampling distribution of  p-values
terrestrial_median_pvalues_8 <-  median(terrestrial_p_values_8)

# Extract the 95% confidence interval for sampling distribution of  p-values
terrestrial_sorted_pvalues_8 <- sort(terrestrial_p_values_8)
terrestrial_lower_limit_8 <- terrestrial_sorted_pvalues_8[26]
terrestrial_upper_limit_8 <- terrestrial_sorted_pvalues_8[975]
```

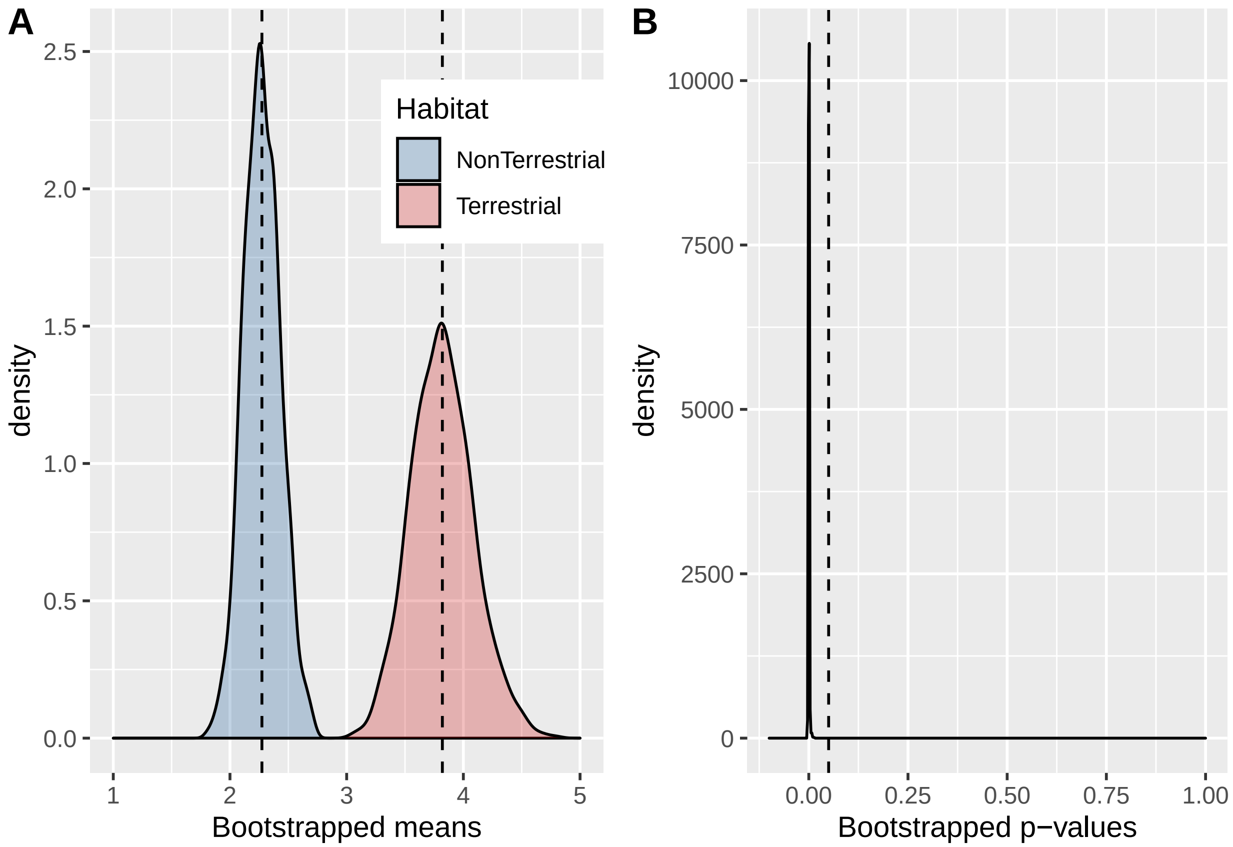
Mean	Median	Lower_limit	Upper_limit
0.0002574	2.23e-05	2e-07	0.002289

```r
# Create the dataframe
terrestrial_df_pvalue_8 <- data.frame(Pvalue = terrestrial_p_values_8)

# Plot bootstrapped p-values distribution
terrestrial_p_pvalue_8 <- ggplot(terrestrial_df_pvalue_8, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

terrestrial_p_pvalue_8
```



Analyses summary:

```
plot_grid(terrestrial_p_means_8, terrestrial_p_pvalue_8, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```



AQP11-like Analyses

Marine/Non-Marine classification

```
# Load data
marine_data_11 <- read.table("AQP11_marinenonmarine_model.tsv", h=T, row.names = 1)

# Fit the model
marine_model_results_11 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP11_like~Habitat,
                                             marine_data_11, phy=x,
                                             method = "poisson_GEE")})
head(marine_model_results_11, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP11_like ~ Habitat, data = marine_data_11,
##     phy = x, method = "poisson_GEE")
##
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##       (Intercept) HabitatNon-Marine
##         0.4390796           0.1539790
```

Extract results for each of the 1000 scenarios

```r
# Extract sampling means
means_marine_11 <- exp(sapply(marine_model_results_11,
                             function(x){x$coefficients[1]}))
means_nonmarine_11 <- exp(sapply(marine_model_results_11,
                                function(x){x$coefficients[1]})
                         + sapply(marine_model_results_11,
                                  function(x){x$coefficients[2]}))
```

```r
# Extract sampling p-values
marine_p_values_11 <- sapply(lapply(marine_model_results_11, summary),
                            function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

```r
# Phylogenetic mean
marine_means_habitat_11 = c("Marine" = mean(means_marine_11),
                    "Non-Marine" = mean(means_nonmarine_11))
```

	Mean
Marine	1.615659
Non-Marine	1.861902

```r
# Phylogenetic median
marine_medians_habitat_11 <- c("Marine"=median(means_marine_11),
                       "Non-Marine"=median(means_nonmarine_11))
```

	Median
Marine	1.613049
Non-Marine	1.855204

```r
# Non-phylogenetic mean
marine_nonphylogenetic_means_11 <- tapply(marine_data_11$AQP11_like,
                                         marine_data_11$Habitat, mean)
```

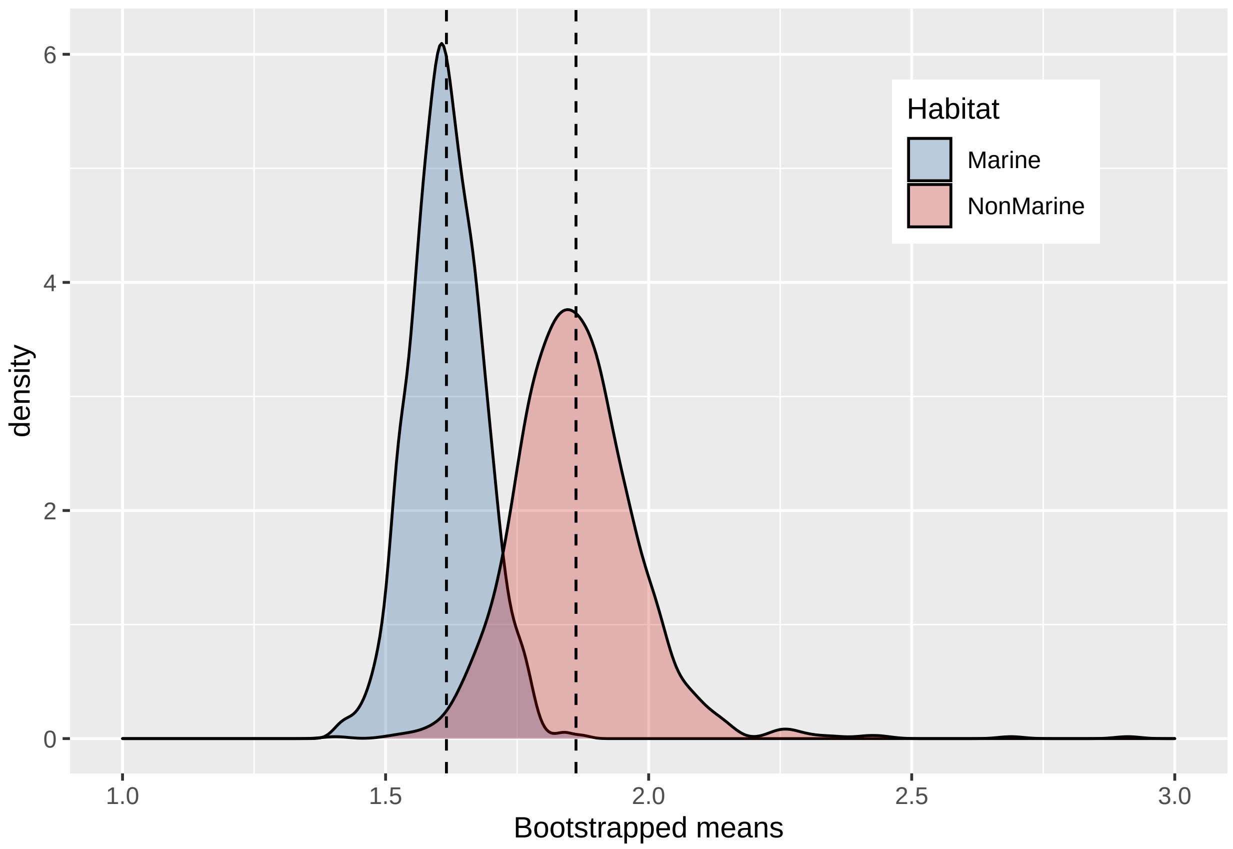
	Mean
Marine	1.960000
Non-Marine	1.254902

```r
# Non-phylogenetic median
marine_nonphylogenetic_median_11 <- tapply(marine_data_11$AQP11_like,
                                           marine_data_11$Habitat, median)
```

	Median
Marine	2
Non-Marine	1

```r
# Create dataframe
marine_df_means_11 <- data.frame(Marine = means_marine_11,
                                 NonMarine = means_nonmarine_11)

# Melt Marine and Non-Marine means data
marine_melted_df_means_11 = melt(marine_df_means_11, value.name = "Mean",
                                 variable.name = "Habitat")
```

```r
# Plot means data
marine_p_means_11 <- ggplot(marine_melted_df_means_11, aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(1, 3) +
  geom_vline(data=marine_df_means_11, aes(xintercept=mean(Marine)),
             linetype="dashed") +
  geom_vline(data=marine_df_means_11, aes(xintercept=mean(NonMarine)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
marine_p_means_11
```



P-values

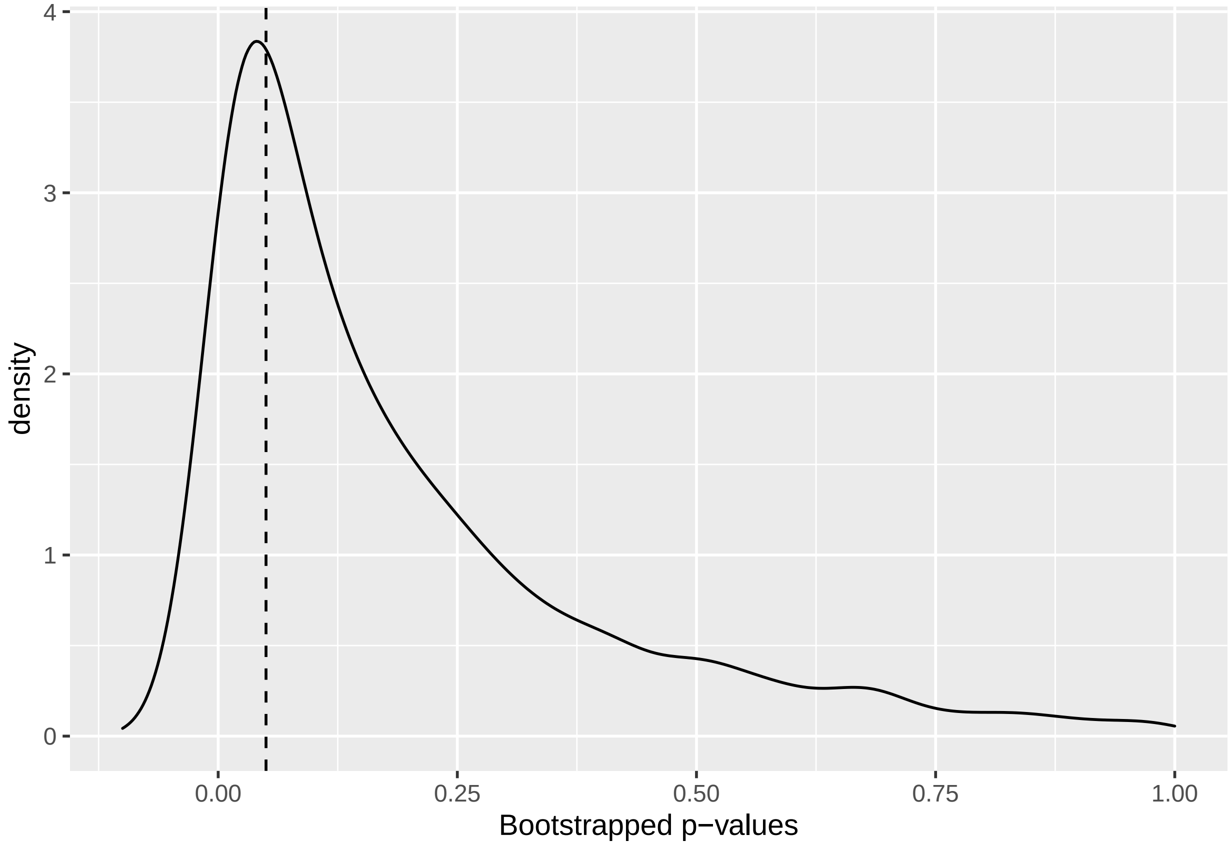
```
table(marine_p_values_11 < 0.05)
```

p-value < 0.05	count
FALSE	700
TRUE	300

```
# Mean of sampling distribution of  p-values
marine_mean_pvalues_11 <-  mean(marine_p_values_11)

# Median of sampling distribution of  p-values
marine_median_pvalues_11 <-  median(marine_p_values_11)

# Extract the 95% confidence interval for sampling distribution of  p-values
marine_sorted_pvalues_11 <- sort(marine_p_values_11)
marine_lower_limit_11 <- marine_sorted_pvalues_11[26]
marine_upper_limit_11 <- marine_sorted_pvalues_11[975]
```

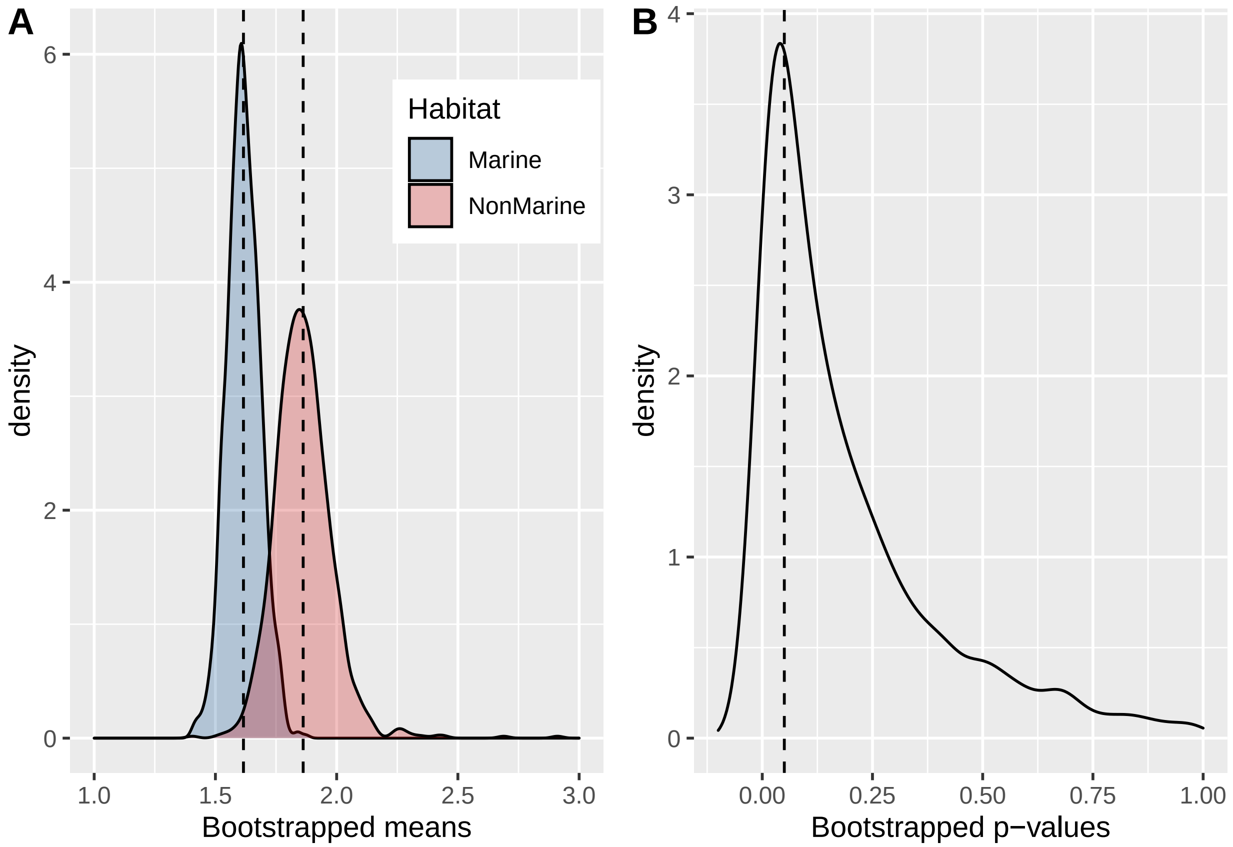
Mean	Median	Lower_limit	Upper_limit
0.190361	0.116134	0.0005974	0.7706079

```r
# Create the dataframe
marine_df_pvalue_11 <- data.frame(Pvalue = marine_p_values_11)

# Plot bootstrapped p-values distribution
marine_p_pvalue_11 <- ggplot(marine_df_pvalue_11, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

marine_p_pvalue_11
```



Analyses summary:

```r
plot_grid(marine_p_means_11, marine_p_pvalue_11, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```



Terrestrial/Non-Terrestrial classification

```
# Load data
terrestrial_data_11 <- read.table("AQP11_terrestrialnonterrestrial_model.tsv",
                                  h=T, row.names = 1)

# Fit the model
terrestrial_model_results_11 <- lapply(simulated_speciestrees_bl,
                          function(x){phyloglm(AQP11_like~Habitat,
                                               terrestrial_data_11, phy=x,
                                               method = "poisson_GEE")})
head(terrestrial_model_results_11, n=1)
```

```
## [[1]]
## Call:
## phyloglm(formula = AQP11_like ~ Habitat, data = terrestrial_data_11,
##     phy = x, method = "poisson_GEE")
##
## Parameter estimate(s) from poisson_GEE:
##
## Coefficients:
##        (Intercept) HabitatTerrestrial
##          0.4889922          0.1001441
```

Extract results for each of the 1000 scenarios

```r
# Extract sampling means
means_nonterrestrial_11 <- exp(sapply(terrestrial_model_results_11,
                                     function(x){x$coefficients[1]}))
means_terrestrial_11 <- exp(sapply(terrestrial_model_results_11,
                                  function(x){x$coefficients[1]})
                           + sapply(terrestrial_model_results_11,
                                  function(x){x$coefficients[2]}))
```

```r
# Extract sampling p-values
terrestrial_p_values_11 <- sapply(lapply(terrestrial_model_results_11, summary),
                                 function(x){x$coefficients[2,4]})
```

Statistical analysis of bootstrapped model results

Mean and Median

```r
# Phylogenetic mean
terrestrial_means_habitat_11 = c("Non-Terrestrial" = mean(means_nonterrestrial_11),
                    "Terrestrial" = mean(means_terrestrial_11))
```

	Mean
Non-Terrestrial	1.680427
Terrestrial	1.793592

```r
# Phylogenetic median
terrestrial_medians_habitat_11 <- c("Non-Terrestrial"=median(means_nonterrestrial_11),
                       "Terrestrial"=median(means_terrestrial_11))
```

	Median
Non-Terrestrial	1.677680
Terrestrial	1.797702

```r
# Non-phylogenetic mean
terrestrial_nonphylogenetic_means_11 <- tapply(terrestrial_data_11$AQP11_like,
                                              terrestrial_data_11$Habitat, mean)
```

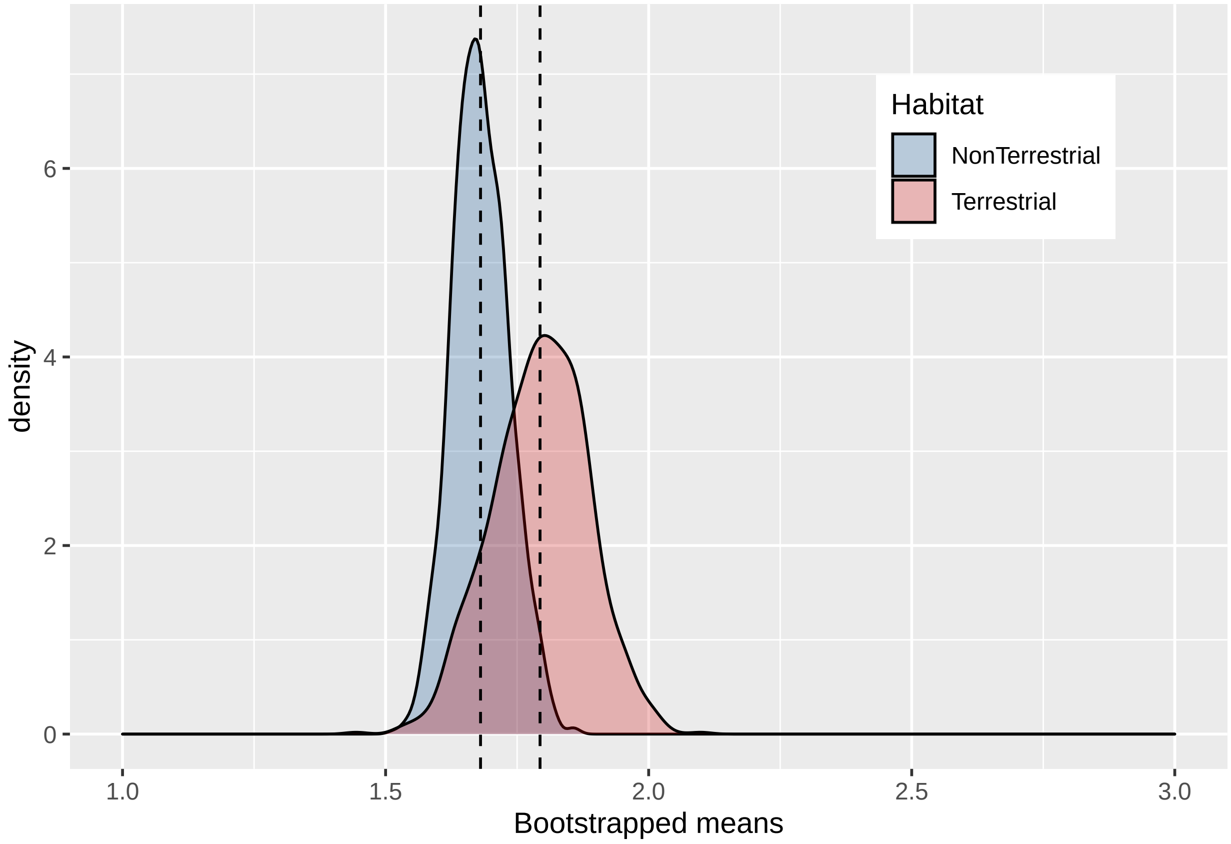
	Mean
Aquatic	1.82716
Terrestrial	0.70000

```r
# Non-phylogenetic median
terrestrial_nonphylogenetic_median_11 <- tapply(terrestrial_data_11$AQP11_like,
                                               terrestrial_data_11$Habitat, median)
```

	Median
Aquatic	2
Terrestrial	1

```
# Create dataframe
terrestrial_df_means_11 <- data.frame(NonTerrestrial = means_nonterrestrial_11,
                                      Terrestrial = means_terrestrial_11)

# Melt Marine and Non-Marine means data
terrestrial_melted_df_means_11 = melt(terrestrial_df_means_11,
                                      value.name = "Mean", variable.name = "Habitat")
```

```
# Plot means data
terrestrial_p_means_11 <- ggplot(terrestrial_melted_df_means_11,
                                 aes(x=Mean, fill=Habitat)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#0b5394", "#cc0000")) +
  xlim(1, 3) +
  geom_vline(data=terrestrial_df_means_11, aes(xintercept=mean(NonTerrestrial)),
             linetype="dashed") +
  geom_vline(data=terrestrial_df_means_11, aes(xintercept=mean(Terrestrial)),
             linetype="dashed") +
  xlab("Bootstrapped means") +
  theme(legend.position = c(0.8, 0.8))
terrestrial_p_means_11
```



P-values

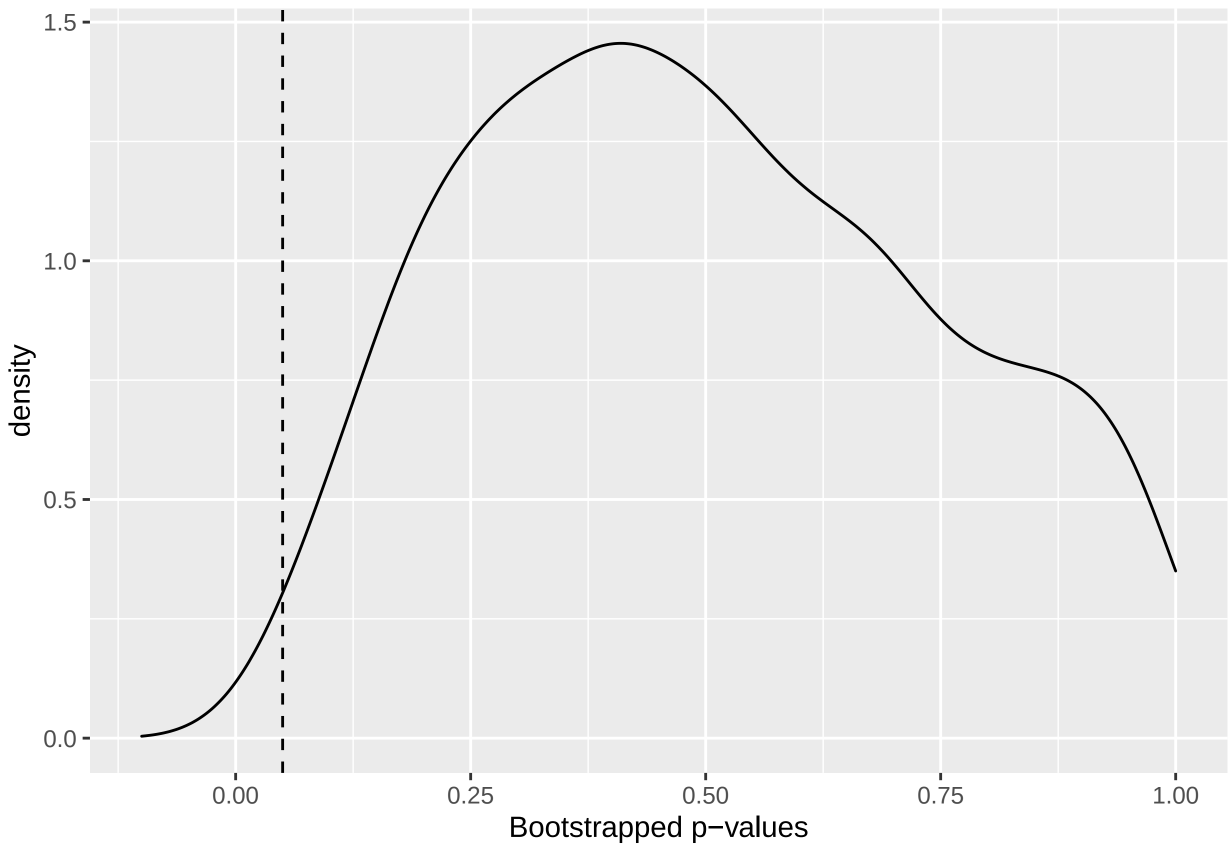
```
table(terrestrial_p_values_11 < 0.05)
```

p-value < 0.05	count
FALSE	993
TRUE	7

```
# Mean of sampling distribution of  p-values
terrestrial_mean_pvalues_11 <-  mean(terrestrial_p_values_11)

# Median of sampling distribution of  p-values
terrestrial_median_pvalues_11 <-  median(terrestrial_p_values_11)

# Extract the 95% confidence interval for sampling distribution of  p-values
terrestrial_sorted_pvalues_11 <- sort(terrestrial_p_values_11)
terrestrial_lower_limit_11 <- terrestrial_sorted_pvalues_11[26]
terrestrial_upper_limit_11 <- terrestrial_sorted_pvalues_11[975]
```

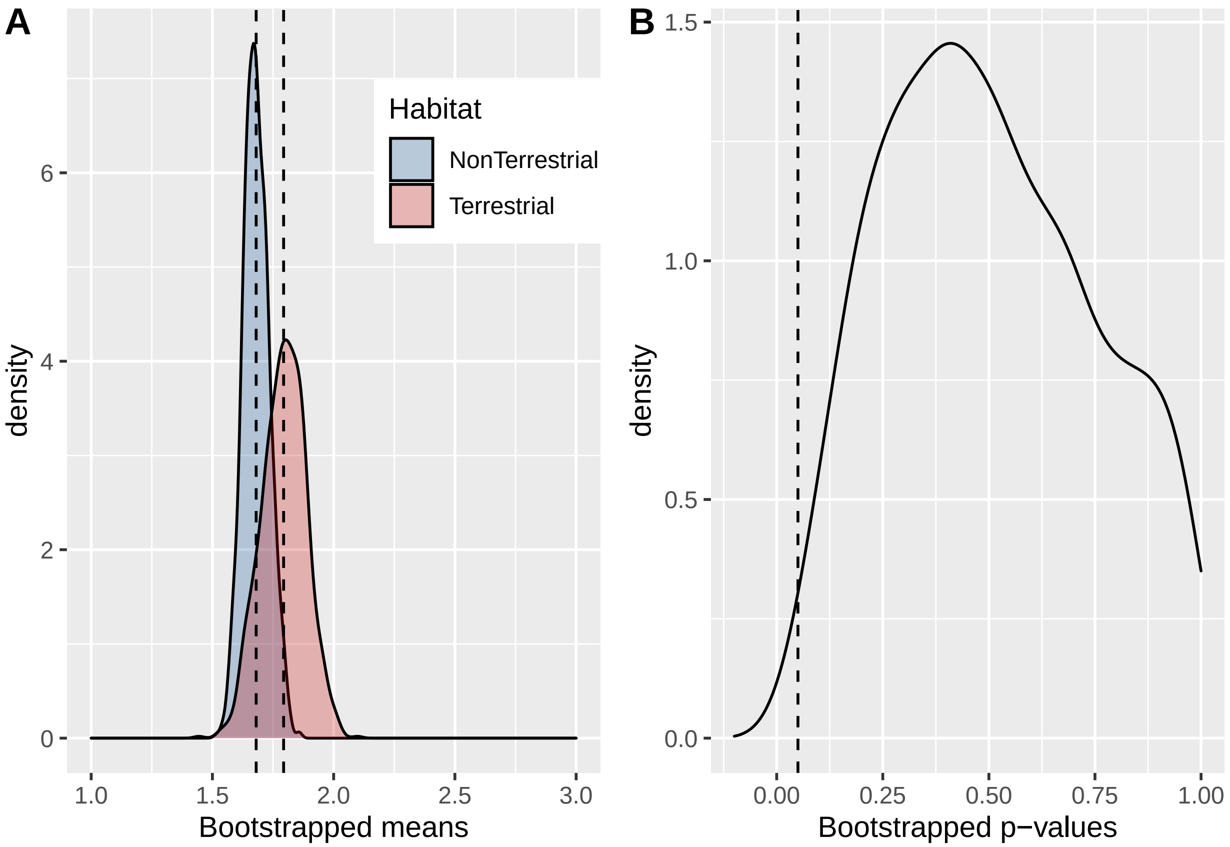
Mean	Median	Lower_limit	Upper_limit
0.5008527	0.479314	0.0970311	0.9623558

```
# Create the dataframe
terrestrial_df_pvalue_11 <- data.frame(Pvalue = terrestrial_p_values_11)

# Plot bootstrapped p-values distribution
terrestrial_p_pvalue_11 <- ggplot(terrestrial_df_pvalue_11, aes(x=Pvalue)) +
  geom_density(alpha=0.25) +
  scale_fill_manual(values=c("#8fce00")) +
  xlim(-0.1, 1) +
  geom_vline(xintercept = 0.05, linetype="dashed") +
  xlab("Bootstrapped p-values")

terrestrial_p_pvalue_11
```



Analyses summary:

```
plot_grid(terrestrial_p_means_11, terrestrial_p_pvalue_11, labels = c("A", "B"),
          ncol = 2, nrow = 1)
```


A
B
density
Bootstrapped means
Bootstrapped p−values
Habitat
NonTerrestrial
Terrestrial
0
2
4
6
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
0.00
0.25
0.50
0.75
1.00