Analysing amplicon data with DivNet

Amy Willis

Vignette Info

Mike Lee, a bioinformatics wizard, recently put together some fantastic tutorials on bioinformatics for analysing microbiome data at link. Mike has very kindly allowed us to distribute the phyloseq data object from this tutorial via DivNet. In this tutorial we're going to analyse this data set.

Preliminaries

Let's take a quick look at the Lee et al dataset.

```r
library(magrittr)
library(phyloseq)
library(breakaway)
library(DivNet)
data(Lee)
Lee
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1490 taxa and 16 samples ]
## sample_data() Sample Data:       [ 16 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 1490 taxa by 7 taxonomic ranks ]
```

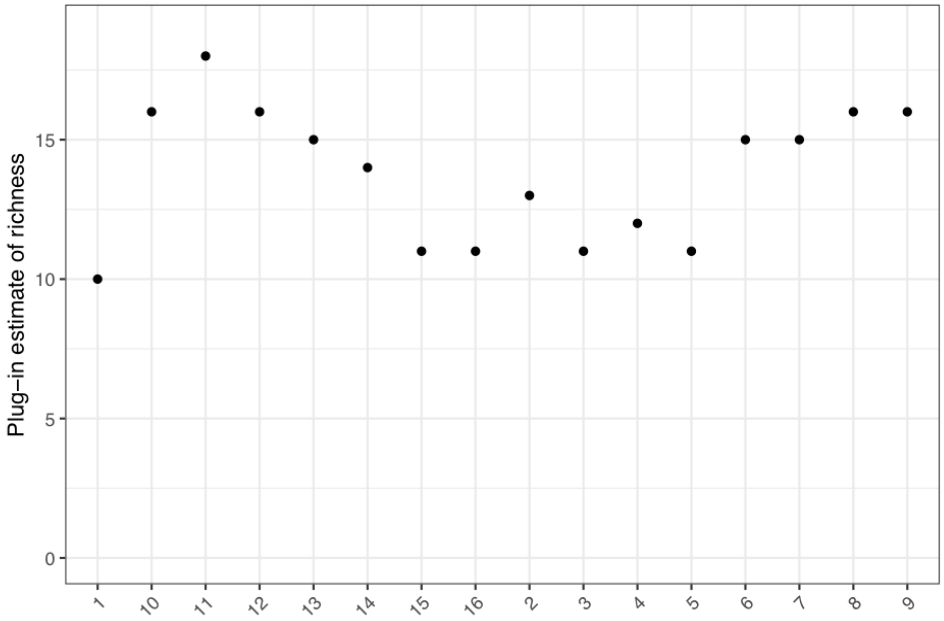
16 samples, 1490 ASVs, and taxonomy information. Fantastic!

What does DivNet do that I can't do already?

The idea behind DivNet, just like the idea behind breakaway, is that you don't care about your samples. You care about the population that your samples were drawn from. So Mike doesn't care about the X mL of sea scum that he scraped off these basalts, he cares about all of the microbes that live on those basalts. For example, if we knew the true relative abundances of all of the phyla living on seafloor basalts, we could calculate the true Shannon diversity at the phylum level of microbes living on seafloor basalt, and compare it to the, e.g., true Shannon diversity of microbes living on land basalts. In order to do this comparison, we need an estimate of the Shannon diversity of microbes living on seafloor basalts.

One approach to getting such an estimate is to consider the Shannon diversity of the samples that we took:

```r
tax_glom(Lee, taxrank="Phylum") %>%
  sample_richness %>%
  plot
```



These numbers tell us something about the X mL of microbial matter that we observed. However, while we hope that they are reflective of all microbial matter living on those rocks, the samples were inexhaustive (we didn't get all the microbes on the rocks into the MiSeq), and so the relative abundances that we observed are definitely not the relative abundances of all phyla on the rock. There are likely missing taxa (that are on the rocks but not in the sample), over sampled taxa (that were observed in greater proportion in the sample than live on the rock), and under sampled taxa (that were observed in lower proportion).

By taking advantage of biological replicates and covariate information, DivNet works to get a more complete picture of the diversity of the microbes on the rocks. Let's take a look to see how it works.

Using DivNet

It's tempting to jump right in and run `divnet(Lee)`, but DivNet gets more expensive as the number of taxa increases. For this reason, were going to analyse the data at a higher taxonomic level than ASVs. Let's look at the phylum level (just to illustrate!).

```
lee_phylum <- tax_glom(Lee, taxrank="Phylum")
lee_phylum
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 20 taxa and 16 samples ]
## sample_data() Sample Data:       [ 16 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 20 taxa by 7 taxonomic ranks ]
```

20 taxa is incredibly manageable! Let's go ahead and run divnet. My computer has 4 cores, so I am just going to run in parallel with the `ncores` argument.

```
divnet_phylum <- lee_phylum %>%
  divnet(ncores = 4)
```

Hopefully that didn't take too long! If you don't want to run in parallel, you can ignore the `ncores` argument.

Let's take a look at what the output of DivNet is: a list of diversity indices and some variances.

```
divnet_phylum %>% names
```

```
## [1] "shannon"                "simpson"                "bray-curtis"
## [4] "euclidean"              "shannon-variance"       "simpson-variance"
## [7] "bray-curtis-variance"   "euclidean-variance"     "X"
```

For each of the 4 diversity indices mentioned above, we have an estimate of the diversity index of the population from which that sample was drawn. So the estimated Shannon index is

```
divnet_phylum$shannon %>% head
```

```
## $BW1
## Estimate of shannon from method DivNet:
##   Estimate is 1.1 with standard error 0.01
##   Confidence interval: (1.08, 1.11)
##
##
## $BW2
## Estimate of shannon from method DivNet:
##   Estimate is 1.08 with standard error 0.14
##   Confidence interval: (0.8, 1.37)
##
##
## $R10
## Estimate of shannon from method DivNet:
##   Estimate is 1.11 with standard error 0.07
##   Confidence interval: (0.98, 1.25)
##
##
## $R11
## Estimate of shannon from method DivNet:
##   Estimate is 1.01 with standard error 0.06
##   Confidence interval: (0.88, 1.14)
##
##
## $R11BF
## Estimate of shannon from method DivNet:
##   Estimate is 0.44 with standard error 0.02
##   Confidence interval: (0.4, 0.47)
##
##
## $R12
## Estimate of shannon from method DivNet:
##   Estimate is 1.42 with standard error 0.04
##   Confidence interval: (1.34, 1.49)
```

and the variance of the estimate is also shown.

Why are the estimates all different? We didn't tell DivNet about any covariate information, so it just assumes that all the samples are from different populations. But we have information about the samples and the conditions under which they observed:

```
lee_phylum %>% sample_data
```

```
##        temp    type      char      color
## BW1     2.0   water     water       blue
## BW2     2.0   water     water       blue
## R10    13.7    rock    glassy      black
## R11     7.3    rock    glassy      black
## R11BF   7.3 biofilm   biofilm  darkgreen
## R12    <NA>    rock   altered chocolate4
## R1A     8.6    rock   altered chocolate4
## R1B     8.6    rock   altered chocolate4
## R2      8.6    rock   altered chocolate4
## R3     12.7    rock   altered chocolate4
## R4     12.7    rock   altered chocolate4
## R5     12.7    rock   altered chocolate4
## R6     12.7    rock   altered chocolate4
## R7     <NA>    rock carbonate  darkkhaki
## R8     13.5    rock    glassy      black
## R9     13.7    rock    glassy      black
```

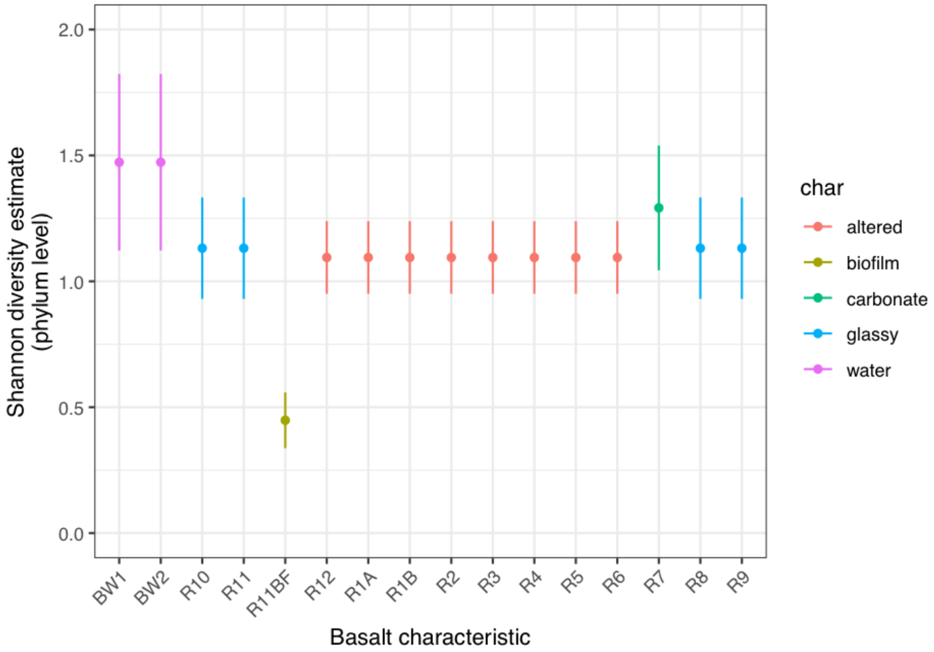
Let's use this to estimate the Shannon diversity of basalts of different characteristics (`char`) using DivNet.

```
divnet_phylum_char <- lee_phylum %>%
  divnet(X = "char", ncores = 4)
```

Let's now compare the plug-in Shannon index with the divnet estimates

```
library(ggplot2)
divnet_phylum_char$shannon %>%
  plot(lee_phylum, color = "char") +
  xlab("Basalt characteristic") +
  ylab("Shannon diversity estimate\n(phylum level)") +
  coord_cartesian(ylim = c(0,2))
```

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one
```



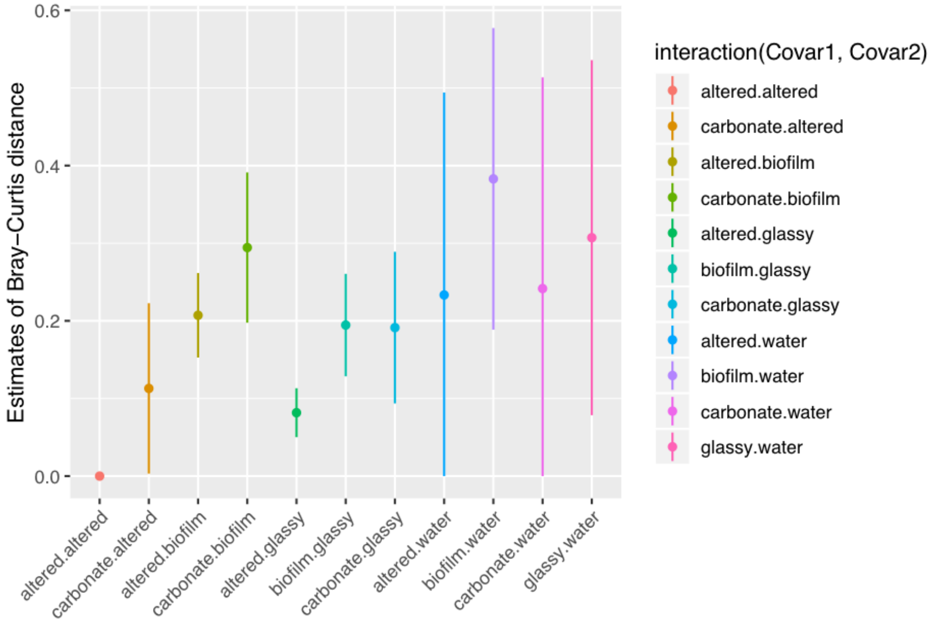
You will notice that the plug-in estimates of Shannon diversity are different for each sample, but there is only a single DivNet estimate for each characteristic (along with error bars). For characteristics for which many samples were observed, there are smaller error bars than for samples for which there was only one sample (seems reasonable – we had less data).

For glassy and water samples, the estimated diversity is in the middle of the individual diversity estimates, while for altered samples the estimated diversity is lower. That's because the water and glassy samples are much more similar to each other in terms of relative abundances than the altered samples, as we can see from the distribution of Bray-Curtis distances between the samples:

```
simplifyBeta(divnet_phylum_char, lee_phylum, "bray-curtis", "char")
```

```
##       Covar1  Covar2     beta_est     beta_var        lower        upper
## 1     glassy   water 3.071559e-01 0.0130761459 7.845394e-02 5.358578e-01
## 2    biofilm   water 3.828949e-01 0.0094346450 1.886307e-01 5.771591e-01
## 3    altered   water 2.333394e-01 0.0170044307 0.000000e+00 4.941415e-01
## 4  carbonate   water 2.415649e-01 0.0185051083 0.000000e+00 5.136318e-01
## 5    biofilm  glassy 1.945580e-01 0.0010890072 1.285578e-01 2.605582e-01
## 6    altered  glassy 8.166048e-02 0.0002484701 5.013462e-02 1.131864e-01
## 7  carbonate  glassy 1.912748e-01 0.0023852200 9.359733e-02 2.889522e-01
## 8    altered biofilm 2.071801e-01 0.0007405695 1.527533e-01 2.616070e-01
## 9  carbonate biofilm 2.944033e-01 0.0023428167 1.975980e-01 3.912086e-01
## 10   altered altered 1.110223e-16 0.0000000000 1.110223e-16 1.110223e-16
## 11 carbonate altered 1.129445e-01 0.0030109808 3.199652e-03 2.226893e-01
```

```
# You can plot this easily
simplifyBeta(divnet_phylum_char, lee_phylum, "bray-curtis", "char") %>%
  ggplot(aes(x = interaction(Covar1, Covar2),
             y = beta_est,
             col = interaction(Covar1, Covar2))) +
  geom_point() +
  geom_linerange(aes(ymin = lower, ymax = upper)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("") + ylab("Estimates of Bray-Curtis distance")
```



As a result, when information from the glassy samples is amalgamated using DivNet, there is similar information from the samples, and so the diversity is sort of averaged. In contrast, there is conflicting information from altered samples, giving the different picture overall. It's somewhat similar to the result that we get when we amalgamate across characteristics and then estimate the diversity:

```
merge_samples(lee_phylum, "char") %>%
  sample_shannon
```

```
## Warning in asMethod(object): NAs introduced by coercion
```

```
## A collection of 5 alpha diversity estimates:
##
## Estimate of shannon from method Plug-in:
##   Estimate is 1.37 with standard error 0
##   Confidence interval: (NA, NA)
##
## Estimate of shannon from method Plug-in:
##   Estimate is 0.43 with standard error 0
##   Confidence interval: (NA, NA)
##
## Estimate of shannon from method Plug-in:
##   Estimate is 1.29 with standard error 0
##   Confidence interval: (NA, NA)
```

```
##
## Estimate of shannon from method Plug-in:
##   Estimate is 1.22 with standard error 0
##   Confidence interval: (NA, NA)
##
## Estimate of shannon from method Plug-in:
##   Estimate is 1.44 with standard error 0
##   Confidence interval: (NA, NA)
```

Where do the variance estimates come from?

When taxa cluster together (spatially), you can imagine that your samples are more likely to be different from each other. For example, you may happen upon a patch of Microbe A in your first sample, but get a patch of Microbe B in your second sample. The statistics word for this is variance, and it is well studied that dependence structures (like spatial organisation) lead to greater variance. Plug-in estimates of diversity (the ones you're familiar with: where you just take the sample data and calculate the diversity index on it) ignore any dependence structure. This leads to understatements of variance (samples are more varying than the model that underpins the estimate allows), and exaggerated risk of concluding significant differences when none exist (p values are smaller than they should be, leading to an increased Type 1 error rate). DivNet addresses this by explicitly estimating those dependence structures, and then using them to come up with variance estimates (useful for the error bars). Let's see that in action.

Hypothesis testing with DivNet

I think the most common way ecologists currently test hypotheses about diversity is with a t-test. For example, if biological replicates are available, they would take the estimated index for one group and compared to the estimated index of the second group (Don't stop reading! You shouldn't do this!):

```
plugin <- tax_glom(Lee, taxrank="Phylum") %>%
            estimate_richness(measures = "Shannon") %$% Shannon
char <- Lee %>% sample_data %$% char
t.test(plugin[char == "altered"],
       plugin[char == "glassy"])$p.value
```

```
## [1] 0.1160722
```

Underpinning this approach is a major problem: the plug-in estimates of diversity are not very good because they don't account for missing taxa or undersampling or oversampling of taxa. For the Shannon index, estimated diversity is too low, with a known negative bias. They are also based on the multinomial model, which ignores taxon-taxon interactions and the additional variance attributable to them.

DivNet addresses these problems by accounting for oversampling and undersampling, and modelling these interactions. This gives us better variance estimates with which to do hypothesis testing.

Let's say we want to compare the Shannon index across the different characteristics. The package `breakaway` provides an implementation for statistical inference for alpha diversity called `betta`. We are going to use this function here for inference on the Shannon index.

To set this up, we need a vector of our alpha diversity estimates, a vector of the standard errors in these estimates, and a design matrix

```
estimates <- divnet_phylum_char$shannon %>% summary %$% estimate
ses <- sqrt(divnet_phylum_char$`shannon-variance`)
X <- breakaway::make_design_matrix(lee_phylum, "char")
betta(estimates, ses, X)$table
```

```
##                        Estimates Standard Errors p-values
## (Intercept)           1.09462448      0.02043387    0.000
## predictorsbiofilm    -0.64661299      0.05533746    0.000
## predictorscarbonate   0.19707983      0.12394339    0.112
## predictorsglassy      0.03686509      0.05032660    0.464
## predictorswater       0.37832517      0.12397480    0.002
```

The intercept term is our altered basalts (they are the only ones that aren't listed), and so all comparisons are made to this baseline. We see that there are no significant differences between altered and glassy, as we saw before with the t-test, but now the p value is 0.53 not 0.12. I think that this is the great thing about DivNet – hypothesis testing now reflects instability in microbiomes better than plug-in estimates.

On the other hand, we do see some differences in alpha diversity across the different groups: biofilm and water basalts have significantly different diversity (at the phylum level) than altered basalts. If we want to do the global test of whether or not there are differences across categories, we can get the p value as follows:

```
betta(estimates, ses, X)$global[2]
```

```
## [1] 0
```