THESE ARE THE WORKSHEETS FROM DAY TWO OF OUR WORKSHOP. THE WORKSHEETS FOR DAY ONE CAN BE FOUND AT THIS LINK.

Q1 WORKSHEETS

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

ITS2 amplicon data could be in ASV or DIV form. Other data types: SNPs, MSATs, any molecular marker, right?

Ordination plot based on 'count table', each point represents the symbiont community of a distinct sample. Generated e.g. in R.
Possible underlying data: ITS2 amplicon data, normalized or non-normalized / subsampled, transformed, based on OTUs, ASVs, unique sequences, 'type profiles'

Have not generated this type of analysis via metabarcoding data myself, so won't speak directly to pre-processing. Assuming that generation is based on total algal diversity via conventional marker (ITS2), with some cutoff for similarity and OTU designation.

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

If ASV, I'd expect any intragenomic variants to cluster closely together, so perhaps ASV/DIV output wouldn't vary much.>>>

different community/symbiont/s/sequences between sites

higher sequence diversity in site A compared to site B

Samples from sites 1 & 2 have distinct symbiont communities. Sites are distinguished along axis 1 (what does it correlate with?). Samples at site 1 are less similar to each other (larger dispersion) than samples at site 2.

evidence for diversity difference across sites in one axis.

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

May first have to clean data set to remove noise (promotes nice grouping, but may lose some signal)

Past experience, haven't seen a lot of differences depending on input data type. Removing low abundance types doesn't really influence things that much provided you have enough read depth. Ordination plots are fairly robust.

Very important to have a lot of data to be confident in the clustering, so you need enough money to generate enough data.

Interpretation: strong separation of communities in colonies at different sites. If based on sequence data, implies symbiont populations are different at the different sites.

Good for same/different, but not for questions of e.g. species number differences

Q1. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

We started talking about the second graph:

-need to think about units of replication (for community studies, seeing more and more that you need replication at the reef scale)

-concerns about confounding (e.g. sites with depth, or different depths within a site)

-whereas ordination plot is robust, diversity plots are incredibly sensitive to methodology (sample selection, data preparation, choice of metric)--can produce very different outcomes

What drives diversity in phytoplankton? Our field doesn't necessarily address this question often enough. It would help us interpret some of the hyperdiversity better. There's a good reason we tend to have a dominant symbiont and it has everything to do with their ecology and physiological response to the environment they are in (both host and external)

But again, the question of the potential role of background symbionts is important to get to (e.g. bg/rare bacterial strains can be critical nitrogen fixers)

Need function!

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

B The figure legend indicates that it is metabarcoding data, so some type of universal marker (ITS2?) was applied to community samples that include a mix of different Symbiodiniaceae genotypes.

D: ITS2 amplicon sequencing and interpretation similar to our last session (or could be other methods e.g. DGGE). Each point is a sample (i.e. a symbiont community). Does not reflect differences in relative abundance (because these are not usually quantifiable) using these approaches. This is not an approach I have used myself much, so keen to hear from others!

C. You could generate with any marker (ITS2, LSU, cp23S, psbAncr) and then pre-process in SymPortal (and get DIVS or ITS2 type profiles) or with DADA2+LULU (and get clustered ASVs, which you then might further consolidate) or with just DADA2 (and no clustering and get raw ASVs or OTUS?) or with QIIME (OTUs), then do an nMDS or a PCA.

After aligning/trimming ITS2 data, assign each sequence to OTUs, then OTU counts are PCA inputs. (Not a sophisticated or even competent user!)

ES: For individual samples generate ITS - NGS or gel based data, use all the sequences/bands to generate a multivariate matrix, run ordination (whether or not abundance is taken into account for each sample would depend on which ordination/similarity is used).

Usual ways – sample, extract, sequence, annotate through symportal examine similarity assigned to source (location). It's been a long time since I personally have generated these sorts of data!

This looks like something that you might get from larvae rather than adult coral samples

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

A At a minimum it suggests population level differentiation between the two sites. The level of divergences is hard to scale without more information.
C. With ITS2 profiles or consolidated ASVs from Dada2+LULU, you might have something close to species, so this would be the most "correct" way of applying nMDS or PCA to these data, to look at community structure. With approaches that generate OTUs, it is more like community structre of intragenomic variants? Need ot be careful with interpretation. Site 1 is different from Site 2 on some level (either species or intragenomic)
ES: Given the data shown, I would then interpret this plot as groupings that identify a singe symbiont entity or depending on how close the clusters are different entities.If the data represents sites, then the community between sites look different
Ania: Site 1 has more variance than site 2
Both have outliers
The two sites appear to be quite different
Just looking at the grapth 1 don´t know what the PCs are

D: Both sites seem to have different symbiont communities

Sites are different, not sure if I'm looking at population or community-level differences, however.

Details of experimental design and level of filtering will impact ability to interpret this plot

We can say these sites are different, but what do we do with this otherwise (in the absence of knowing anything about biology in this example)
Interpretation can be influenced by biases. If you see a difference in (for example) physiology, then you may turn to this sort of approach and see whether any levels of filtering generate a difference like this)
This is easier to interpret if there are multiple coral species or multiple symbiont genera, then it is easier to be confident in this graph showing interspecific differences

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

--We know (based on a small number of types/papers) that copy number can vary significantly (esp across clades) but we generally don't take this into account in these types of analyses. Eg we know that ITS2 underdetects D relative to C. When we have intragenomic variation included in this sort of analysis, does it invalidate the assumptions of the approach?

--Because this is ITS2 data, there are many ways of interpreting the graph. Because of the issues with ITS2, and unknown level of filtering, this graph becomes hard to interpret with certainty.

--Some differing views on the utility of fthis approach for population level questions
--can show that sites are different or that one site has more variation. Differences might be more easy to trust if you know that your different sample types contain multiple symbiont genera, because this puts a taxonomic "scale bar" onto the figure
-- might have more confidence in differentiation if you know that the groupings are different along some other axis (eg sediments vs coral). However, you have to watch out for confirmational thinking (esp. when you're comparing unknowns or sites)

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

A. Resolution of taxonomy might matter – species/strain/population/community

A. assuming this is amplicon data, I would identify variants, filter for those occurring in at least 3 unique samples and at sufficient depth to guard against sequencing/coverage errors, I would apply a transform to account for technical variation among samples (lower vs higher depth etc) then run the PCA. its too early for me to remember what distance I would use

B. either genetic variaiton within spp or different communities of spp

A ITS2 amplify using the SYM_VAR primer pair and standard protocol. Submit for Illumina PE sequencing. Receive demultiplexed fastq.gz files back.

Option 1: Submit to SymPortal. Choose one of the resultant PCoA ordinations to work with (either BrayCurtis- or Unifrac-based; sqrt transformed or not).

Option 2: Submit to SymPortal, then generate a Unifrac or Braycurtis distance matrix from post-MED seq data 'manually'. Then run PCOA. Then plot up.

Distance can be measured as distance to centroid.

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

A1. ITS2 metabarcoding, dada2, GeoSymBio database (2012, very dated), collapse co-correlated ASVs, plot PCA, perform adonis, I may also test for dispersion differences to test for differences in variation across sites

A2. ITS2 metabarcoding, SymPortal, plot PCA of both DIVs and other version, plot PCA, compare with dada2, perform adonis

B. If the Adonis was significant between site A and site B I would conclude that these corals host different communities of algae

B. PCA plot can look very differently depending on resolution

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

Might matter which genera are present, how do normalize the IGV?

What is actually being measured here? More granularity

Do DIVs link uniquely to spp? They do not necessarily, there are some that are characteristic of spp

We all agree that when you see these differences in the ordination, that it represent biological differences based on all of the caveats

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

>>> So, a lot of assumptions have to be made. The legend is vague, on purpose, so I am going to go out on a Friday night limb and assume:
1. each dot represents a community of Symbiodiniaceae from an individual zooxanthellate anthozoan colony of the same species.
2. the data were generated by some kind of NGS amplification of ITS2.
3. the data are grouped into clusters (black and gray circles) of similarity, maybe 70%? No idea.
4. each dot includes not only sequences but also the abundance of each sequences.
5. NGS data were QCed in a normal pipeline of some kind.

>>> qPCR relative abundance of symbiont genotypes

>>> I would expect that these data come from individual coral colonies of the same species sampled from two different sites.
-Preprocessing the data would include the usual steps of removing sequencing errors, running through a dada2 pipeline and then compared to a sym (ITS2) database.
-I suppose it could be either abundance (normalized) or presence/absence (jaccard dis. index) of particular ITS2 types, both might be interesting

>>> Uni/multi-loci (or other) data from individual samples/communities depicted. Pre-processing dependent on data type, but if sequence-based, then "typical" filtering of alignment of forward and reverse sequencing reads and removal of primers, as well as sequences with uncalled bases, highish quality score cut-off over at least 75% of the nucleotides in a read.

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

>>>Simple conclusions that could be made:

Site A and Site B are different (PC1)
Site A and B have a lot of variation
Site A and B have some outliers
Little symbiont community overlap between A and B
>>>sites have distinct populations though with some outliers; less variability at site 2; PC1 drives the majority of the pattern seen.

>>>I would be comfortable saying that there is something distinctive about the two communities from each site. 62% on the first axis is quite a lot of the variation.

- If it's based on abundance I think you would have to be a bit more careful with what exactly you say about how they are different.

- if it's P/A then I actually think you could say more definitively that differences in genotypes - ...strins/species (sorry Todd), are driving the differences. - real community turnove

>>> Most conservative interpretation of the plot is two entities with moderate/high variation within each. PC1 captures much of the variationr

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>> Con of ribosomal: Lots of data noise

>>> Pro of SNP/microsats: Better for population genetics

>>> All interpretations: scrutinize the data within your context, especially the outliers

>>> Highlight = agreement

>>>

>>>

>>>

Q1. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

>>> Community-level or presence/absence

>>> How do you interpret the outliers?

How different are they actually?

>>> Ideas on how to deal with recruits

Markers for each genera

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

>>> this plot looks like ASV level plotting (if from same coral host) or could be ITS2 profile level if encompassing different corals/environments/etc

ITS2 (or LSU) amplicon sequencing, processing through DADA2+LULU or Symportal, generating ordination plot based on a metric like bray Curtis calculated from symportal ITS2 profiles

Remove adapters, sequencing artifacts, merge sequences, blast sequences to databases, generate a matrix, perform multivariate analysis.

Symportal can do the input files for those plots

If these were multi-locus genotypes of symbionts FROM THE SAME SPECIES OF HOST, these would be different species if the sites were close by. If they were thousands of Km/miles away might be different populations of the same species. This could be verified with use of a few phylogenetic markers.

I don't get the box and whisker plot. If it's based on ITS data one symbiont species could simply have more intra-genomic variants than the other. If these are population genetic data then one population is either larger or more genetically diverse than the other for many possible reasons (obviously)

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

- Site 1 and 2 are somewhat genetically different, but with high variability within each. However, most of the difference is captured by PC1.
- General questions: is this within the same species of cnidarian? Same depth?
- Sites 1 & 2 have different community compositions, Site 1 has more variability in community composition
- I have come to think that alpha and beta diversity are too abstract of concepts to be terribly useful, especially given the types of data we have. Beta diversity analyses don't add much beyond what differential abundance analysis would say; they essentially just summarize differential abundance, but using methods that don't control well for various different processes. I would interpret this PCA plot as saying there is at least one taxon that is different between site 1 and site 2, but with no context, it is difficult to know whether that is because the samples are all simply dominated by a single taxon (which is different), or if there are complex communities with many different taxa, or what.
- **Common agreement in group:**
    - **This plot is most likely encompassing multiple species/sites/etc and not from the same host in the same habitat (unless certain coral species that are known to contain multiple genera)**
    - **Would not see this much variation within a genus**
    - **Too hard to interpret biological relevance if this is plotted at ITS2 ASV level**
    - **Could be different environmental samples (could see this level of diversity from waterr etc)**
    - **These plots often over interpreted**

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>> ASV level plotting: more noise when most may be same strain & symportal ITS2 profile

>>>symportal profile plotting- this doesnt look like it could likely be generated from profile outputs (unless from many different corals etc)

>>ordination plot as display: con- can exaggerate differences that might just be due to absence of one taxon between groups etc, need to be very careful not to overinterpret

>>even if all samples generally dominated by same strain in the plot, how much variation could be biological vs artifactual? (differences at cellular level w ploidy, etc?)

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

>>> [have never worked with dna metabarcoding] - is this intragenomic variation in multivariate space? are low abundance OTUs removed? are pcr duplicates removed?

>>> Aligning, trimming, zOTUs or ASVs assignment, stats and PCA generate it. I assume it is ITS2?

For ITS2 analyses, this could be used to compare intragenomic variant (all/DIVs) composition or the composition of ITS2 type profiles.

Distance metrics? Point of discussion

Stats always needed alongside ordination plots. Need to consider dispersion for PERMANOVA analyses with unbalanced design.

PCA from proportion of seq reads assigned to OTUs / type profiles after filtering

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

Different interpretations (in my opinion!)

1) For intragenomic variant based ordination, samples between sites show some differences in their IGV composition along pc1

2) For ITS type profiles, samples between sites show differences in their symbiont community composition along pc1

Distinct community composition between sites with more similarity among samples within site 2

Different differencial populations between the two sites, replicates/samples are more different among each other in site 1. These differences are based on the marker used, so it is hard to tell how accurate such conclusions are.

Important considerations

- **Whether multiple genera are present**
  - **Consider plotting them separately**
- **Sequencing depth/methods**
- **Which distance metric is applied**
- **Dispersion/unbalanced design messes up the PERMANOVA results**

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

- **Type profiles**
- **Sequence counts**
  - **qPCR of known culture concentrations to determine actual relative abundances of genera**
- **^Good if the two above give you similar results, but would be a point of contention if they don't match**
  - **Could go to a different marker for clarity**

- **Evaluation of distance metrics:**
  - **Is it abundance based?**
  - **Is it phylogenetic distance based?**
  - **^These may impact your outcome.. Is there consensus on which is most applied and/or most appropriate?**

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

>>> filter and QC sequences, collapse into profiles (or OTUs/ASVs), generate count data, potentially transform, calculate distance metric, perform ordination.

>>>-remove barcodes, inspect the raw reads, -clean/filter the reads, -variance normalize w/ preferred method,-run PCoA analysis method

>>>-Calculations of diversity are generally done on raw reads, do not want to variance normalize, boxplot

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

>>>differences between sites, but could be entirely different symbionts, or different relative abundance of same symbionts. depends on what kind of counts were used (all sequences vs. collapsed 'taxa'), intragenomic architecture (# variants) and distance metrics and transformations used (e.g., square-root transformation increases effect of low-abundance taxa/counts)

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>>ordination using all sequence data appropriate for some 'genetic diversity' questions (within species), but not species diversity

>>> how is copy number considered in the analysis?

Q1. INTERPRET ITS2 ORDINATION DATA

Clustering of samples using PCA (or some sort of ordination) and measuring dispersion

A. How might you generate and pre-process data to generate this plot?

>>> ITS2 metabarcoding data.

Generated through 2 x 250 bp MiSeq amplicon seq.

Pre-process through either symportal or DADA2.

If analysing ITS2 ASVs, Hellinger transform abundance.

Unifrac-distances using a non-alignment based phylogenetic tree (e.g. using k-mers).

PCA/PCoA/NMDS.

Generated by sequencing, perhaps ITS2 amplicon seq. Pre-processed by QC, filtering, profile or cluster generation, some distance matrix and ordination

I interpreted this plot to be the result of next-gen sequencing of ITS2 "defining intragenomic variants" that was used to create ITS2 profiles and then plotted with PCA

Data may originate from an OTU table. Preprocessing would likely involve data QC, clustering, removing very low abundance otus, etc.

Generate sequencing data, QC + trimming, remove uninformative markers (i.e., markers which are low confidence or absent in some samples), assembly relevant metainfo about each sample so it can be used to annotate the points.

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

PCA graphic: site 1 and 2 have different communities. Site 1 has greater dispersion

The sites appear to be well-separated by PC1, with a large fraction of support. The loading of this axis may be interesting to interpret. Clustering / collapsing of OTUs may have a great impact on the ordination result. That is, the ordination COULD be driven by noise/errors, or it could very well be signal.

Two clusters. Depending on what has gone into it (ASVs, profiles) etc will change interpretation. Two symbiont species, or two distinct communities?

PCA:

- **Assuming each point is a separate coral individual**
- **Obvious difference between sites (assuming this was not a constrained ordination). Run adonis or RDA to compute significance but it is pretty clear.**
- **Incredible amount of variation explained considering N. Want to see the screeplot!**
- **Is scaling correct? To keep the point-cloud shape intact, during plotting PCs are supposed to be scaled by the sqroot of lambda/sum(lambdas) – in vegan, scores(..., scaling="sites") . Saying this because I feel the Y-axis should be narrower range.**

Sites appear to separate well along PC1, Site 1 has higher dispersion, and both samples have a single outlier but they are still clearly separated (so probably not sample mix-up).

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>>

It all depends on the hypothesis (question of interest).

Pre-processing, "clustering", and distance measure and

Unifrac-distances - Pros - Helps to separates distinct genera & communities.. Cons - sequences from within a single genome can be more diverse than sequences across genomes (!)

Q2 WORKSHEETS

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

>>>

filter and QC sequence data, collapse into profiles (or OTUs/ASVs), generate count data, potentially transform, choose diversity metric, calculate and plot… or calculate a phylogenetic diversity metric from sequence data

Usual ways – sample, extract, sequence, annotate through symportal? It's not clear how diversity is then resolved – relative abundance?

I would expect that these data points come from individual coral colonies of the same species sampled from 2 sites

- Preprocessing the data would include the usual steps of removing sequencing errors, running through a dada2 pipeline and then compared to a sym (ITS2) database.
- to attempt something like diversity you would definitely have to run through SymPortal and even then... diversity is very tricky if this is ITS2.
- Another thing to do would be to make some kind of co-occurrence adjustment to deal with intragenomic ITS2 variants.

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

>>>site A has higher 'diversity' than site B, but depending on underlying data, this could represent just sequence diversity, or something closer to taxonomic diversity (e.g., if sequence variants collapsed into type profiles or something similar).

Sites exhibited different median diversity but substantial variance to both with different distributions

Just realizing that it doesn't say what level of diversity but I guess - alpha # species within a colony.

- I want to be able to make this type of analysis at intracolony community level but I don't think you can get there with much confidence based on ITS2 alone unless there is a high level of divergence between your types

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

Crux of issue (as always): how do you collapse seq data into units (e.g. OTUs)

So many options at each step and it really affects interpretation

PCA can account for more variables, but diversity collapses to one idea and harder to accommodate many sources of variation

In review, using ITS2 data often seen as problematic

Standardization/pre-processing to deal with sequencing depth and sequence collapsing can be very hard because most methods require raw reads

Unless you can apply a copy number standardization before the collapsing step, any diversity metric is going to be quite biased

With PCA, can plot several different ways (e.g. ITS2 abundance, symportal output, presence/absence), then see how they compare. Could we do something similar here?

Abundance vs. richness: yikes

~~Could potentially make more informative if you sequence multiple markers~~

Potential recommendation:

Q2. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

Vertically transmitted symbionts: may prune IGV, so degree of symbiosis in lifestyle may influence genomic variation. But that's a pretty 16S/prokaryotic idea, and we might not be there yet with Sym given their incredibly large genomes.

PAM measurements of mixed communities: how do you interpret it?

May measure and then analyze community diversity and find good correlation, but in a simple system with 2 species. Less straightforward for very diverse communities.

Potential strategies for getting around sequence abundance not being equal to cell abundance

Flow cytometry is great for relative abundance provided you have good markers for your system, which take time to develop (currently genus-level tags)

- **Igepal--magic detergent for opening holes in Sym cell wall**
- **Autofluorescence is an issue, but new FACS can get between the fluorescence peaks with the availability of many lasers**

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

A. I interpreted this plot to be the result of next-gen sequencing of ITS2 that was used to create intragenomic variant profiles, which were then assessed via a diversity metric and plotted. Unsure which diversity metric, as there are many, and is critical for interpreting this figure.

R. This plot is likely raw ASVs, not profiles, so one way to produce it is by running raw reads through DADA2 and then calculating simple shannon diversity for each sample. Although I don't find abstract alpha diversity very useful in general, it would at least be better to try to reduce the influence of intragenomic copy variation by running the data through something like SymPortal or apparently lulu, then calculating shannon diversity based on numbers of profiles rather than numbers of unique ITS2 variants. Other diversity metrics may be more useful depending on the question, such as Faith's phylogenetic diversity

DK Use ITS profiles that have been screened or collapsed for intragenomic variation.

DK The difference in Symbiodiniaceae genera will greatly influence how "diversity" is generated. For example, a colony with Breviolum could be much different than a colony with Cladocopium

E. You could generate with any marker (ITS2, LSU, cp23S, psbAncr) and then pre-process in SymPortal (and get DIVS or ITS2 type profiles) or with DADA2+LULU (and get clustered ASVs, which you then might further consolidate) or with just DADA2 (and no clustering and get raw ASVs or OTUS?) or with QIIME (OTUs), then do shannon's diversity, or simpsons D, or species richness.

- Amplicon sequencing -> ASVs or ITS2 profiles delineation -> estimation of diversity -RG

C. Assume it's a measure of beta diversity (?), given distinction between two sites. Check for uniformity of species richness. The diversity index used matters, as some are more robust to variability in richness.

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

>>> R. Extremely difficult to interpret. This level of variation suggests the data are raw ASVs, not profiles, and variation in copy numbers is more likely to affect these plots than variation in species, strains, anything useful biologically

- While this suggest very different diversity in the two sample sites, it is not something you would typically obtain from ITS2 profiles of symbiotic Symbiodiniaceae. Maybe something like this could be found when comparing environmental samples (i.e. sediments)? Still, more information is required to interpret this plot.-RG

E. With ITS2 profiles or consolidated ASVs from Dada2+LULU, you might have something close to species, so then this would be a reasonable analysis. However, if you're using OTUs or data that contains lots of intragenomic variants, then the results in this graph are very difficult to interpret, and might break assumptions of approach. This is probably easier applied to/interpreted from an experimental design that involves very different hosts (and are therefore more likely to contain different symbiont species). If this is sites, maybe you can use this to say Site A differs from Site B, but what does that mean? Maybe nothing about species diversity…

DK Unclear what is used to define a "species" are each unique ITS profile considered a unique genotype?

DK What diversity estimate is used? Simpson, Shannon, Chao1, etc. All will mean different things.

RR different community/symbiont/s/sequences between sites, variance and metrics look Weird, high variance higher sequence diversity in site A compared to site B . Number of copies and metrics also influence the results immensely.

C. Not using appropriate diversity index may skew data. Doesn't appear typical of Symbiodiniaceae(?).

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>> R. For alpha div, using raw ASVs or OTUs is rarely appropriate.

Copy number is a definite issue for this. Some folks are saying that they just completely ignore these graphs in bacteria or Symbiodiniaceae papers. Instead of doing these types of analyses, it could be better to use linear models to parameterize this.

How much does phylogenetic information help interpretation here? Parallels with challenges in bacterial community. Eg phylogenetic signal often present in bleaching susceptibility. Can test for this and hten decide whether to include

--Concern expressed about potential for "pvalue hacking"… ie output tons of metric results from pipelines and you'll find SOMETHING that is significant

--matters which symbiont genera you're working with… if your sample is dominated by Cladocopium, there will likely be a ton of intragenomic variants, which will impact this.

Alpha diversity of a reef (across conspecific colonies)… is that more useful?

Importance in shifting ideas over time, as we learn more about system. Importance of transparency in how we're processing and interpreting things.

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

D: ITS2 amplicon sequencing (and associated interpretation/filtering!) similar to our last session (or could be other methods, e.g., DGGE). Each point is a sample (i.e., a symbiont community).

A. ITS2 metabarcoding, dada2, collapse co-correlated ASVs, estimate diversity using inverse simpson/shannon,check for equal variances and run aov, with Tukeys (if more than two levels)
B. If the aov was sig I would say that one site has higher ITS2 diversity than another site, in this case I would say it is a proxy for community diversity but I am not sure that this is correct

A. assuming its amplicon (ITS2) I would identify sequence variants accounting for errors, identity to symbiodiniaceae etc, then filter to retain only variants occurring in at least 3 samples at sufficient depth, then to think about diversity I would probably try to identify intragenomic variants and collapse ASVs into representative seqs and then use shannon or simpson
B. Caveat is that I generally don't do this bc I think interpretation is very fraught. If you haven't collapsed intragenomic variation this could be a very misleading plot as you could have a very similar diversity just more intragenomic variants, so I think I would require some filter for that

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

>>>

There MAY be higher algal diversity in site A vs. site B, but I would want to know more about how the data were generated and what the cut-off was for defining 'types' within each site.

D: Seems to be more diversity at A. But one problem here is that the ITS2 data produced by these methods does not accurately reflect differences in relative abundance of taxa (even though we often assume that it does). So this plot is perhaps better viewed as a “richness” plot vs a diversity plot per se. Another discussion point is that is the unknown level of filtering

1. There is greater diversity - on average - at Site A

2. Numerous outliers suggest that there is might be insufficient sampling

General questions: is this within the same species of cnidarian or between? Same depth? Same light levels?

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

- Can’t really say much, for me PCA / RDA / envfit / ordisurf are more informative at least at the data exploration point. Boxplot are good to make a point once you figured out what is going on, but I would not start there.
- Also, betadisper in vegan will test whether points are more or less spread out for specific groups (just to avoid various diversity indices and keep everything in the PCA framework)
- Can there be a bias, based on depth of sequencing for example? Or, unequal number of samples across populations?

-has anyone explored how much variation is there within a colony
-throw everything into SymPortal and get DIV profiles, then estimate richness-richness vs eveness, how much confidence do you have in estimated this
-taxonomic scale matters
-Diversity is a particularly hard thing to estimate
-a lot of problems with ITS2, but new markerse have not been properly developed
General consensus: the read of the entire room was that this sort of analysis is generally not great, but if you do it you just need to be as transparent. The info that is going in (ITS2 data) is maybe not the most appropriate. Need to be very specific about what went in to generate it.
Need to be very careful in the discussion of the paper and not to go too crazy in interpretation.

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

I would need to know what genetic data are being generated and how before processing. AND what hypotheses were being tested.

Diversity plot based on 'count table', each point represents the community diversity (what diversity metric?) of a distinct sample. Generated e.g. in R.
Possible underlying data: ITS2 amplicon data, normalized or non-normalized / subsampled, transformed, based on OTUs, ASVs, unique sequences, 'type profiles', exclusion of low abundance entities/singletons

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

The box and whisker plot can be interpreted very differently depending on the data being analyzed and the locals were specimens were collected. If it's based on ITS data one symbiont species could simply have more intra-genomic variants than the other.

If these are population genetic data (even better) then one population is either larger or more genetically diverse than the other for MANY possible reasons (obviously).
>>>

Samples from sites A & B have different 'diversity' based on whatever was used as input data (e.g. different (intragenomic) sequence diversity, different symbiont type/species diversity)

Too little data to really say one way or another. The obvious assumption is that site A has higher diversity than site B, but that may not necessarily hold true, depending on copy number variations across species, etc.

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

- **Is this analysis appropriate given intragenomic variation within a species?**
- **Have to be careful that there aren't artefacts or environmental contamination**
    - **Type profiles should be less influenced by low abundance sequences than ASVs**

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

qPCR, microsats – depends on the question and level of resolution needed!

Do you analyze intragenomic variants (DIVs) or ITS2 type profiles? How does this impact your interpretation

ITS2 amplicon sequencing, processing through DADA2+LULU or Symportal, alpha diversity metrics

Vegan or Phyloseq (R packages)

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

-if diversity is calculated at ITS2 profile or genera level, site A has higher diversity than site B

Colonies at Site A more diverse than Site B, though some overlap and some outliers; more variability at Site A

This is completey dependent on the diversity metric at hand. If we are talking about diversity of sequence richness across samples, there are problems related to interpreting richness from samples with varying sequence depth that are not straightforward to rectify.

Common themes:

>sequence diversity should not be displayed as boxplot; should be at species level

>sequence diversity more appropriately plotted as the PCOA

>symportal profiles could be used to compare alpha diversity if you make the assumption that a symportal profile is its own genetic entity

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>>illustrate biological differences → boxplot; genetic differences -> ordination plot

>need much more context to interpret plots, need more attention to appropriate ways to use statistical tests

Q2. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

>>> if you want to know # of species, need more than ITS2

>>>can still make meaningful comparisons using same marker across multiple samples

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

>>> Amplification with universal ITS2 primers, high throughput sequencing, and then processing the data via SymPortal

>>> Next gen seq w Sym Portal, followed by diversity estimate that hopefully accounts for sample size

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

>>> This is pretty tough to interpret in this contextualization. It looks like Site A has a greater variety of symbionts than B, but without more information it is hard to say.

>>> Diversity at site A > B but what kind of diversity is this? Need separation/understanding of intra- vs inter-genomic to interpret

>>> Ben [under the assumption that this has been generated from some ITS2 data] This plot puts me in 'red-flag mode'. That's because I'm used to seeing people trying to draw inference on species diversity from NGS ITS2 diversity. There are so many caveats involved with doing this, that its almost not worth doing/misleading to do.

But there again, maybe this plot is generated from a different kind of data J.

>>> Commmunity 1 is more diverse than community 2
Both communities have outliers
The two algal communities are quite different

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>> Con: Many caveats with diversity and ITS2, especially when considering intra/intergenomic variation

>>> ITS2 can be used at certain levels (genus) as long as the context and processing are well-established (sequence-level beware)

>>>

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

Generate sequencing data, QC + trimming, remove uninformative markers (i.e., markers which are low confidence or absent in some samples), generate diversity values using your favorite/appropriate package/statistic.

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

Sites A appears to have higher diversity than site B but there are some outliers, especially for site B (check this is not sample mix-up or problems with data).

E: The interpretation of this boxplot is very dependant the origin of the data generated. If these are ITS2 generated sequences from a singe coral species for example and diversity is estimated for each sample then averaged across samples in these boxplot, I would never use a diversity index - it would be showing you the level of variability between many co-dominant repeats in the ribosomal array, and all it tells you that the genome of species at site. If this data were derived at the symbiont species level, then it would be a good way to show the differences in species diversity between sites with higher diversity at site 1.

Difficult to draw distinct conclusions without knowing how this data was generated. There are differences in diversity across the two sites, but unclear what this means and whether it's functionally relevant.

If these plots are at the species level, then the question is again of how to get from e.g. ITS2 data to a species level.

Interpretation depends on whether the plot shows diversity or whether it is also a measure of proportional abundances.

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

>>>

Q2. COMMUNITY DIVERSITY ESTIMATES

A. How might you generate and pre-process data to generate this plot?

>>>

Generated by sequencing, pre-processed by QC, filtering, profile or cluster generation, some measure of diversity

Some sort of clustering/grouping/denoising of sequence data to estimate diversity

Uni/multi-loci (or other) data from individual samples/communities depicted. Pre-processing dependent on data type, but if sequence-based from a multi-copy molecule, then "typical" filtering of alignment of forward and reverse sequencing reads and removal of primers, as well as sequences with uncalled bases, highish quality score cut-off over at least 75% of the nucleotides in

Once again, a lot of assumptions have to be made. The legend is vague, on purpose, so I am going to assume:
1. Each dot represents a community of Symbiodiniaceae from an individual zooxanthellate anthozoan colony of the same species.
2. The data were generated by a diversity index derived from some kind of NGS amplification of ITS2 dataset. Clustering or ?
3. We would hope that NGS data were QCed in a normal pipeline of some kind.
4. The data are grouped means and outliers for each site.
5. What does each dot include/entail? (abundances too or not; no idea).

B. Considering the variety of approaches to (1), what are the different ways you might interpret this plot?

Site A has higher diversity, but it is unclear what diversity means here, so interpretation would be question or hypothesis specific, as well as based on what level of diversity is being assessed here.

Unsure if this represents inter- or intragenomic variation

I'm not convinced that these sites are different (though they probably are), and I don't know what "Diversity" means in this context. Precision may be insufficient to detect differences.

There is higher diversity in site A – it could mean that there is a higher diversity of symbionts in site A, or, that there is greater intragenomic variation in symbionts in site A that has remained in the dataset.

Interpretation of plot will vary on resolution of marker

Simple conclusions that could be made:
The diversity at Site A and Site B are different, probably significantly (from a stats point of view), B is lower than A, but there is that crazy outlier.
Site A and B have a lot of variation, with many outliers, some extreme.
As the famous Scott Santos once said, "What is up with these outliers?"

Most conservative interpretation of the plot is two sites with moderate/high variation within each (Site A higher relative to Site B). "Relatively more" diversity among individual samples/communities at Site A. Outliers present at each site, with reasoning for these potentially ranging from artifacts to real variation between sites.

Evaluate the pros and cons of each approach; when might each approach be the most appropriate?

- **What Marker and copy number impact diversity?**
- **How is diversity defined (a, B, index)?**
- **It all depends on the question!**

Q3 WORKSHEETS

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

>>>ITS2 is still a fantastic marker when you want to know the dominant symbiont, well tested, matches with DDGE, there is convergence, ITS2 will tell you genera presence/absence, ITS2 is very general- works across genera, also super easy to amp because it has so much IGVs

-as a community we might need to set limits that we don't cross, we should allow for intellectual freedom, but give boundaries.

-how people use the marker is more detrimental than ITS2 itself

-cheaper, lower activation energy

-legacy research, how can we incorporate this data that can help contextualize

What are the limitations of single marker studies?

>>>-problems with how to interpret IGV? Rae ITS2 does not represent diversity in the sample, cannot translate ITS2 diversity to species diversity

-depends on your question

How can single marker designs incorporate additional markers (within reason)?

>>>if you know lots of IGV then maybe add a second locus. If you're interested in variation within a genus for example, then maybe incorporate a different marker or set of markers, but again depends on research Q

-additional single markers can help clarify that you indeed have only a single species

Low coverage whole genome sequencing should be explored (1-2x)

Q3. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

>>>flow chart in manuscript - what are your interests, question determines the marker

-sometimes a feeling that the community requires the fanciest methods

-need to get more honest with ourselves as a field

-on our consensus paper - need to have a place where we determine

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

--if a single marker works well, it is easier and cheaper than having to use multiple markers. Especially if that is not the main focus of your study. And more time effective.

--may be enough if you're working in an established system (eg you know what symbionts will be in there and the single marker works)

--there isn't really a scientific advantage to having a single marker (unless the study is about that specific gene?)... it is really about practicality. So 1) is the marker reliable, and 2) does the interpretation from the marker match biological question?

What are the limitations of single marker studies?

--not always clear when another marker is going to be necessary

--for a population level question, you'll need to do pre-screening first to make sure your single marker is going to work.

--this becomes more challenging when you have hosts that are co-infected etc

--does not work for taxonomy

As part of consensus road map, we could include a set of specific ideas for what to do if you are trying to apply an additional marker and it isn't working with your system.

Idea to help get additional markers being used more commonly: have something like a community slack (but that is used globally) where folks can ping others re challenges or questions (or successes!) that they may be encountering with a marker

How can single marker designs incorporate additional markers (within reason)?

--need to be careful that we don't make requirements for multiple markers that are overburdensome. If you don't have multiple markers, you may need ot limit your interpretation

--we should be explicit in the products from this workshop about when a second marker is needed, and which additional markers (specifically) would be the priority.

--need to prioritize generation of additional markers in areas where they are needed

--for something like Porites lobata, you could either sanger sequence

--To minimize burden w/ something predictable w/ ITS2 like Porites lobata, you could start with "end marker" (aka highest resolution) and then go back and do more broad markers (ie spot check a few of the normal samples and then also do the ones that had an unexpected result). In contrast, for something Acropora muricata (which has different symbionts in diff environmental zones) you might want to start out with a more general marker (eg ITS2) and then zoom in w subsequent markers

--could start by pre-screening with RFLP and then decide which markers to apply

--include links to existing curated databases for non-ITS2 markers in the products of this workshop (so if someone wants to run these markers, they can find sequences to compare them to)

--if you focus on developing a local system, then it makes investing in primer design (eg to get psbAncr working for your system) have a better payoff. During primer optimization, it can be helpful to make sure you have a good positive control (eg something in culture or a sample from someone else that contains symbionts that are known to amplify with the primer. Or you can design primers from a genome, if available)

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

Don't want to throw away all the previous ITS2 data because there's so much of it

Don't really care what marker it is as long as it's functional and can link back to the older marker (ITS2)

What are the limitations of single marker studies?

Really need to know your system first.

ITS is terrible--can we move away from it? (multicopy, IGV, etc)

How can single marker designs incorporate additional markers (within reason)?

cp23S may be a better tool than ITS2 for answering similar questions (but not as good for Cladocopium)

Hierarchical approach: perhaps start with ITS2, then use another marker to drill down into the diversity within a genus with genus-specific markers

Cross-validation: confirm DIV patterns with additional marker (not necessarily full amplicon sequencing but targeted/cheap)

Can we have people from outside our field come tell us about which markers work best in their field? We should be seeking out help more. HAB/free-living dinos (earlier recognition of ITS IGV in that field, so they avoided it)

Potentially have a modern "working" workshop (lab workshop, but not really) where groups get together (virtually?) where community can work toward identifying/developing/testing additional markers, potential single-copy markers (do extra science, needs funding)--EAGER?

With new genome mega dataset coming out, even before complete assemblies, could target individual markers first to see how they hold up across genera/lineages

Q3. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

Look into meta-genome assemblies for functional studies of coral colony Sym communities

But taxonomy is important!

Can we get a better marker? And what do we do in the meantime?

Having a “stable” of geneticists or are “on call” for assistance/guidance--as a service

Online course could potentially accomplish this

Follow model of PAM course? (15-20 people, exponential sharing of knowledge)

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

- Simplicity in analysis, in terms of generation
- Lower cost
- Big databases available if they've been studied for a while
- ITS2:
    - Can span community levels

What are the limitations of single marker studies?

- Reliability across taxa?
    - Might evolve at different rates across taxa - can't draw inferences across taxa in this case
- Good to see if you get the same answer across multiple markers
- Hard to compare across studies if everyone is using different markers
-

How can single marker designs incorporate additional markers (within reason)?

- We could use a hierarchical design with multiple markers, rather than using one marker or the other
- Use markers appropriate for your question
- We could look to other fields that are classifying new markers, identifying housekeeping genes
    - We can use genome sequencing data to mine these
    - Can mine SNPs

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

>low cost

>more accessible/ ease of use

>don't always need highest level of resolution of symbiont ID, depends on study/question, sometimes single marker can provide important/sufficient insights

>in well studied/characterized systems, single marker could be sufficient (ex Orbicella)

>don't always need to get to species level (might need multiple markers for species level, but sometimes this can be overkill)

>

What are the limitations of single marker studies?

>>>

>>in less well characterized systems, sometimes incorrect assumptions can be made about symbionts being the same w single marker studies

>sometimes do not provide sufficient resolution

>single marker can help compare across genera but copy number challenges- we need to figure out copy numbers to quantitatively compare

How can single marker designs incorporate additional markers (within reason)?

>>> Hierarchical manner/design to start more conserved and then get more specific

>generate online molecular keys- when you might need additional markers to get higher confidence

>generate different SOPs for each genera

>at least highlight/make very accessible limitations with certain makers & certain symbiont genera (ex ITS2 variation can be super high in cladocopium- high copy number, be warned!)

>build databases with ribosomal copy numbers for different symbiont taxa

>need to develop better databases, especially cladocopium, for different markers → will help with accessibility

>

Q3. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

>>>adding additional makers will be important and will require cross referencing to other markers esp ITS2 since we have so much ITS2 data

LS1 (LSU?) also good but not for high throughput amplicon sequencing since over 600bp

>are copy numbers generally similar in genera? Or will we have to figure this out based on species.

- **Between genera def sig variation in copy number, likely variation within genus as well, lot of work needs to be done to determine this**
- **Copy number variation within cladocopium- C1, C3, C15 radiations**

>cladocopium has so much variation, can ID many cladocopium ITS2 profiles from South Pacific that are then collapsed into a single type w psbA sequencing

>to lower barrier to entry, can have guides to help ppl narrow it down (ie working in a certain part of the world→ most likely to encounter XX and XX symbionts → use XX SOPs)

>

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

>>> With rationale and the appropriate questions/scope, markers are useful, no matter how "outdated" they may seem

"If it ain't broke, don't fix it."

Great for genera

What are the limitations of single marker studies?

>>> Single markers go through fads

>>> Single markers are not geared towards diversity/population genetics → need more markers to define communities

>>> NGS data pipelines may not fit our purposes (you shouldn't use a lab mouse-based pipeline to answer questions about algae)

>>> NGS downstream choices create a lot of headache for the researcher and the reviewer: the "correct" choice is unclear. Most researchers want to be an end user to just be able to answer their ecological questions

>>> ITS2 does not cover what some researchers want to look at (Octocorals)

How can single marker designs incorporate additional markers (within reason)?

>>> Even with a suite of markers, they may not be able to handle communities since they do not across all genera

>>> Microsats/flanks clear up some resolution

>>> More accessible methods (PCR-based; limit NGS and pipelines)

>>> ITS2 is a great tool for certain questions, but there are plenty of other useful tools, some of which are easier and less expensive, for other questions

>>> Cheat sheet for what data are equivalent across markers ("this ITS2 type = this psbA")

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

>>>

Easily accessible to a bunch of different labs.

What are the limitations of single marker studies?

>>>

Does it assay all the diversity? Can it pull out all the diversity that is there?

How can single marker designs incorporate additional markers (within reason)?

>>>

Q3. SCRATCH PAD

CAPTURE ANY ADDITIONAL NOTES OR IDEAS YOU WANT TO PRESERVE.

>>>

Clearly we need more single markers. We suggest that there needs to be more genomics, perhaps single cell whole genome sequencing of all the different genera that can be used to mine useful markers. There is a need for more traditional population genetics studies in Sym. There is a need to benchmark all of these different potential markers otherwise people won't use them. . PSVA is a marker than is useful for Cladocopium. Potential markers could be functional genes - look into other systems that have similar markers (potato!). Surely there are museum or freezer samples that are benchmarked by an expert to help set up a sequence database.

Markers must be aligned otherwise there is no good way forward. The question is how to do it most efficiently.

Q3. MOLECULAR MARKERS

What are the advantages of single marker studies?

>>>

It is specific to the marker chosen

Cost, time, higher throughput, ease and interpretation

Ease of cross comparison between studies and with “historical” data

What are the limitations of single marker studies?

>>>

Limited number of types of inferences you can make

Lack of consensus on interpretation from a single marker

Single or limited level of resolution (taxonomic breadth) and symbiont of interest

Some markers work better for some groups than others. Specificity of markers across groups can vary.

How can single marker designs incorporate additional markers (within reason)?

>>>

Is 2 markers really double the cost? Ways forward:

- Homemade reagents or reduced reaction sizes. Determine the depth needed for each marker.
- Creating resources for the community for efficiency (protocols.io, github)
- Creating a multiplexed PCR
- Target follow up samples
- Search for new markers that give both breadth and depth