Comparing taxonomic and functional diversity in metagenomic samples

Sarah Hird – 24 September 2014

WHY? Genes make up organisms but not (necessarily) taxa:

Nearly identical genes can be found in disparate taxa and non-overlapping gene sets are found within species.

GOALS **Identify** biologically interesting communities and genes

Estimate **diversity**

Determine **function**

Infer evolutionary **processes**

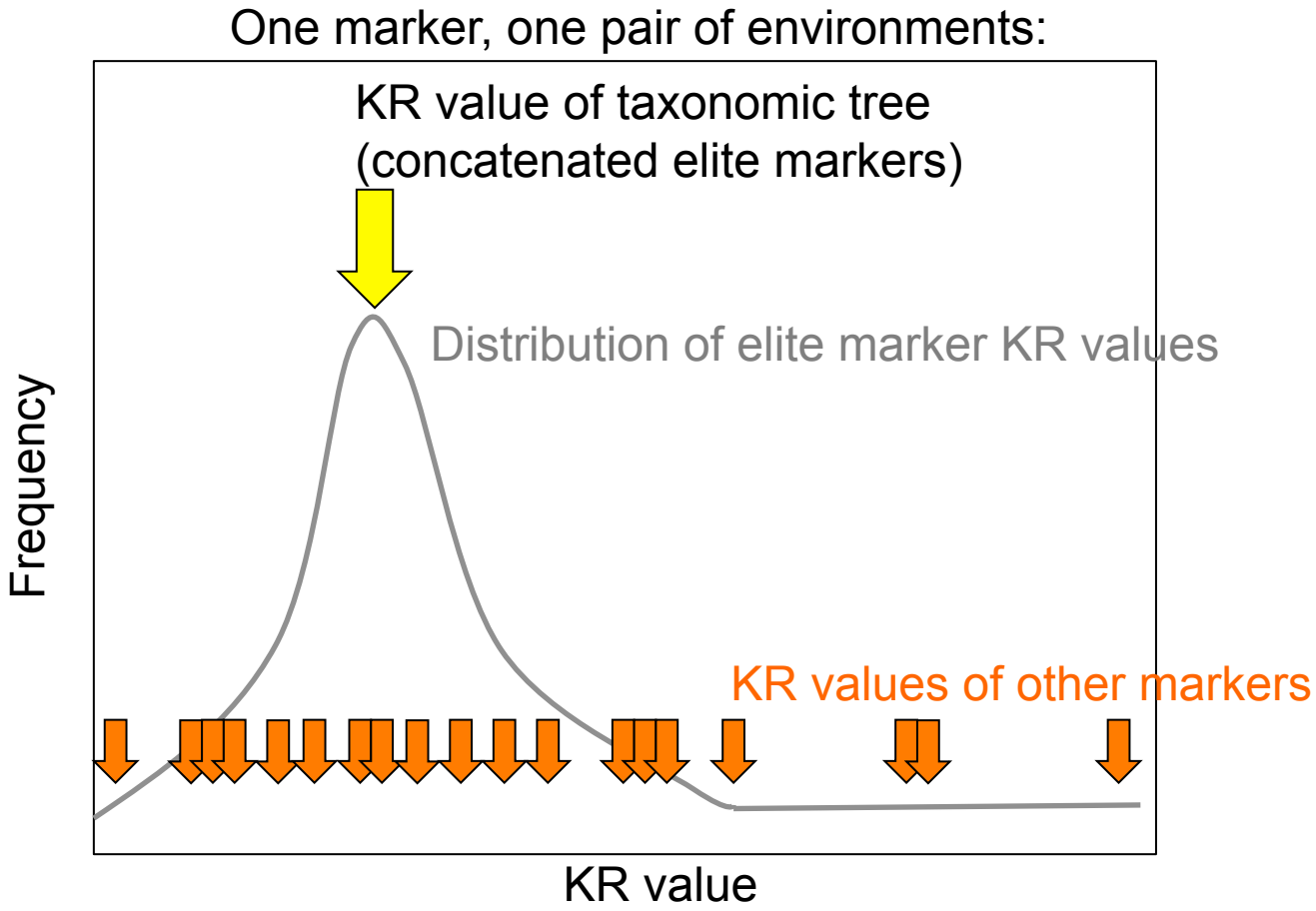
PhyloSift

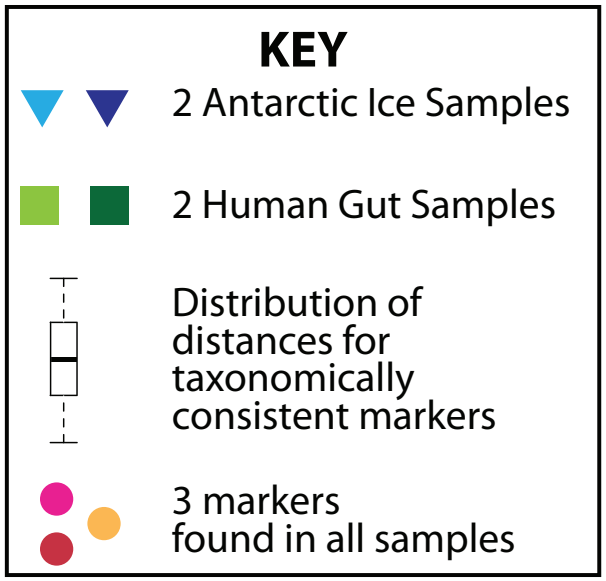
- Input=metagenomes
- Outputs = (1) trees for a given set of reference markers, (2) estimates the taxonomic composition of a sample from a phylogenetic tree made by concatenating 37 “elite” markers and placing them on a large reference tree

- I calculate pairwise diversity metrics between all the samples for all the markers – elite markers comprise a distribution of values that are “taxonomically consistent”, i.e., the diversity found within the elite markers could be attributable to the taxonomic diversity of the sample.

For each pairwise comparison of samples and for each marker, calculate the KR distance

KR DISTANCE: How much "work" it takes to make reads from one environment match reads from second environment (aka "earth mover's distance")



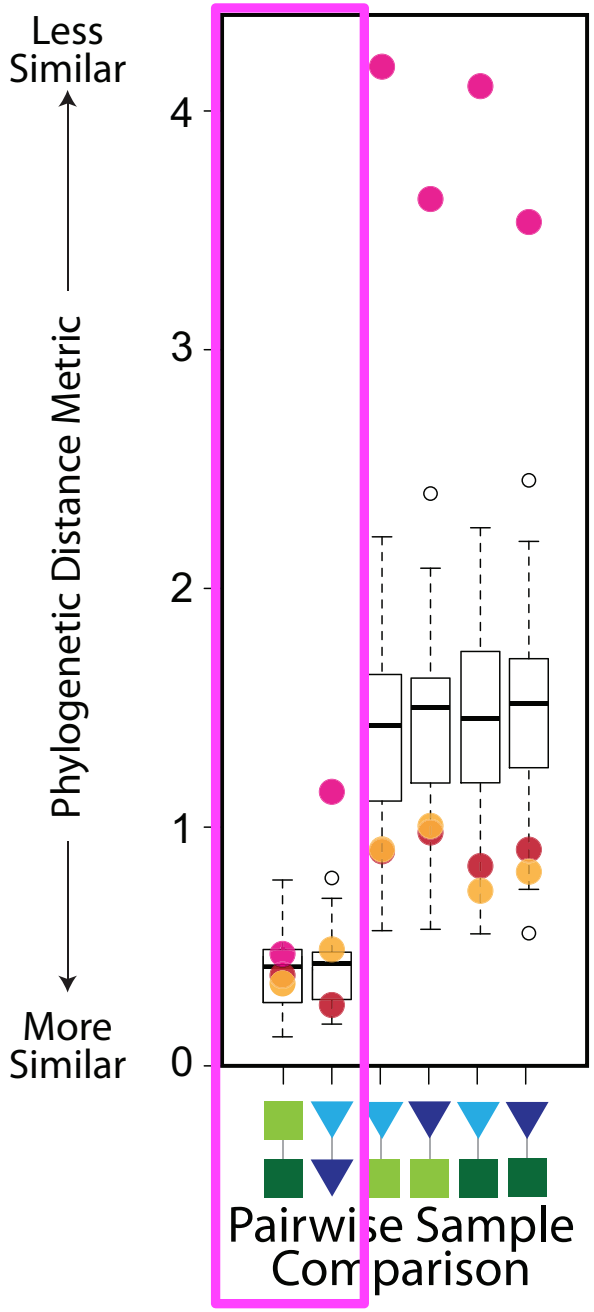


4 samples = 6 pairwise comparisons

Like samples have lower distances between the elite markers

Extended markers have unique patterns

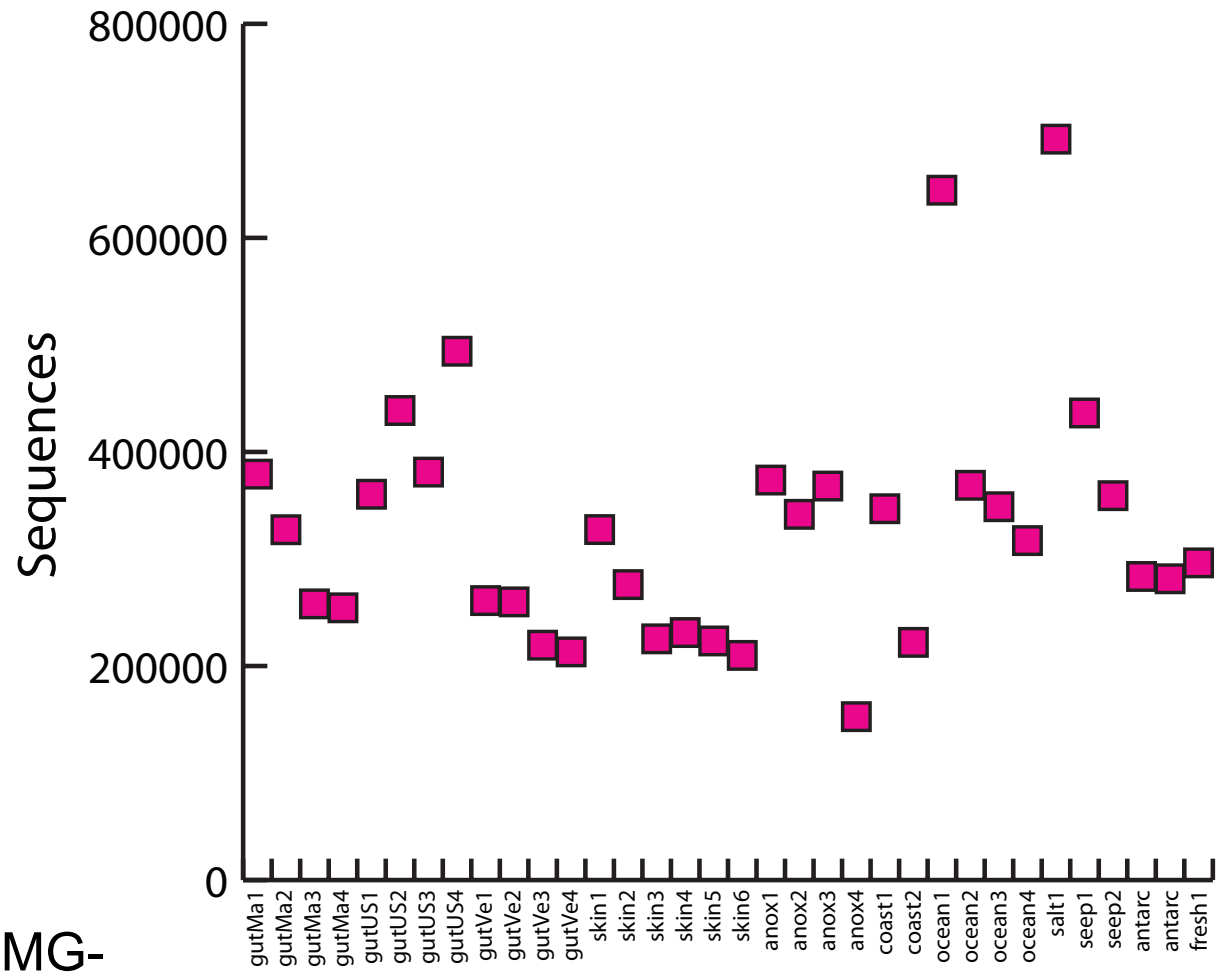
Figure?



Dataset34

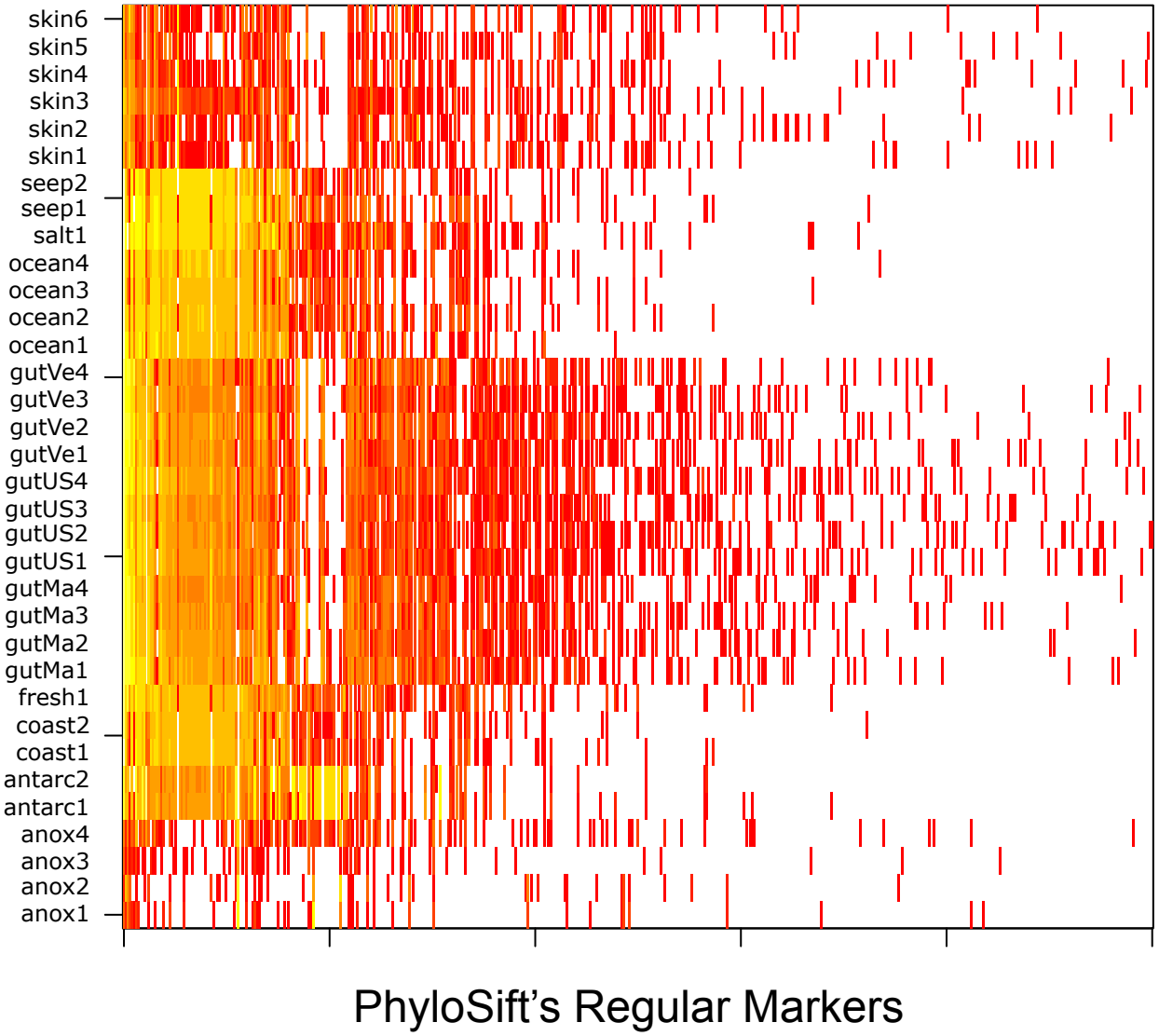
- **6 skin**
- **12 gut**
  - 4 Malawi
  - 4 USA
  - 4 Venezuela
- **13 marine**
  - 4 open ocean
  - 1 hypersaline
  - 2 cold seep
  - 2 coast
  - 4 anoxic
- **3 freshwater**
  - 2 Antarctic
  - 1 open water

- All downloaded from MG-RAST

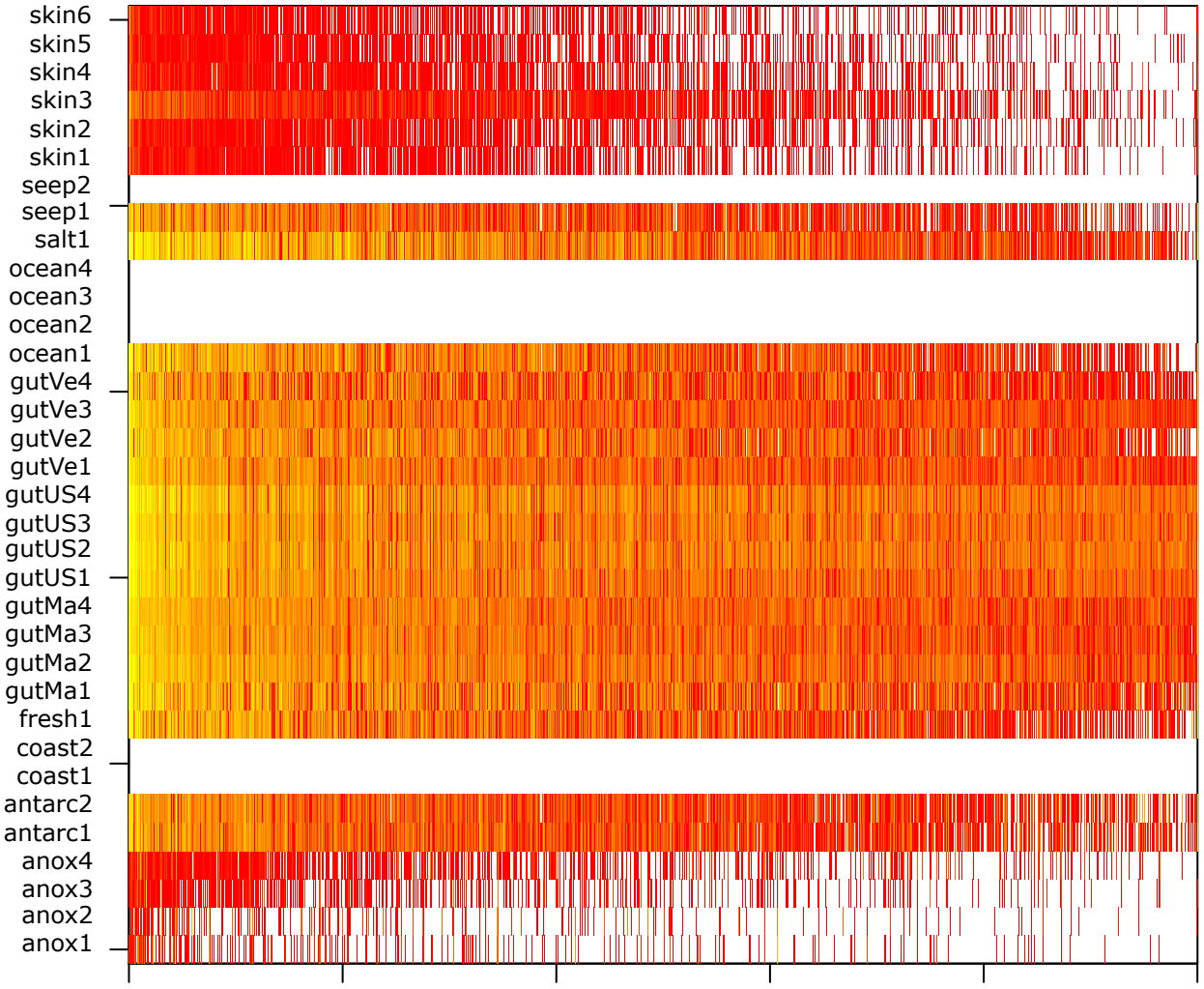


Sufficient diversity of samples? Too diverse?
Should these have been rarefied?

How many reads from each sample hit to the markers



How many reads from each sample hit to the markers



How to account for uneven sampling?

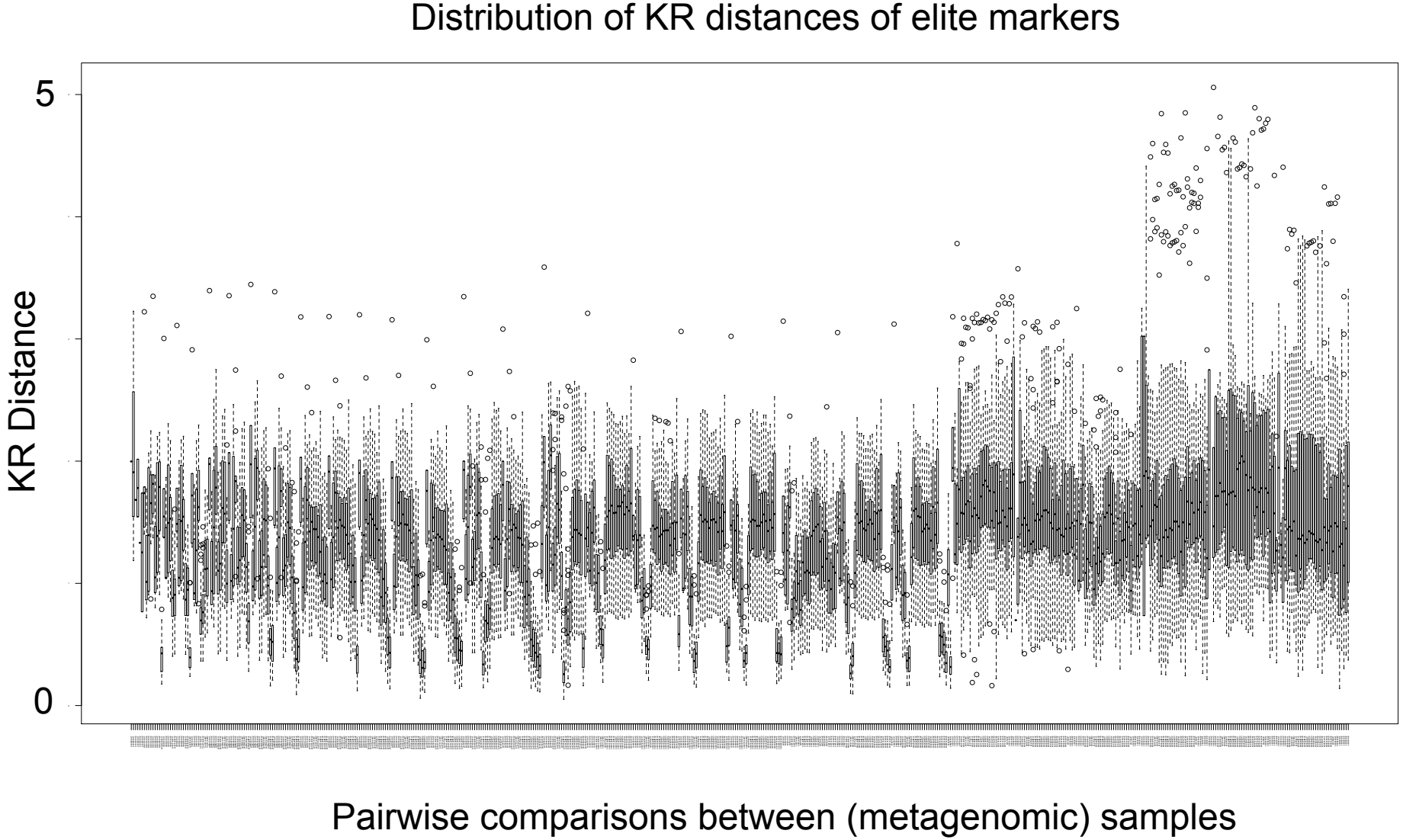
Current Data

- Ran metagenomes through PhyloSift (regular markers and 5000 extended markers)
  - 5385 trees with >3,000,000 placed reads (I think)

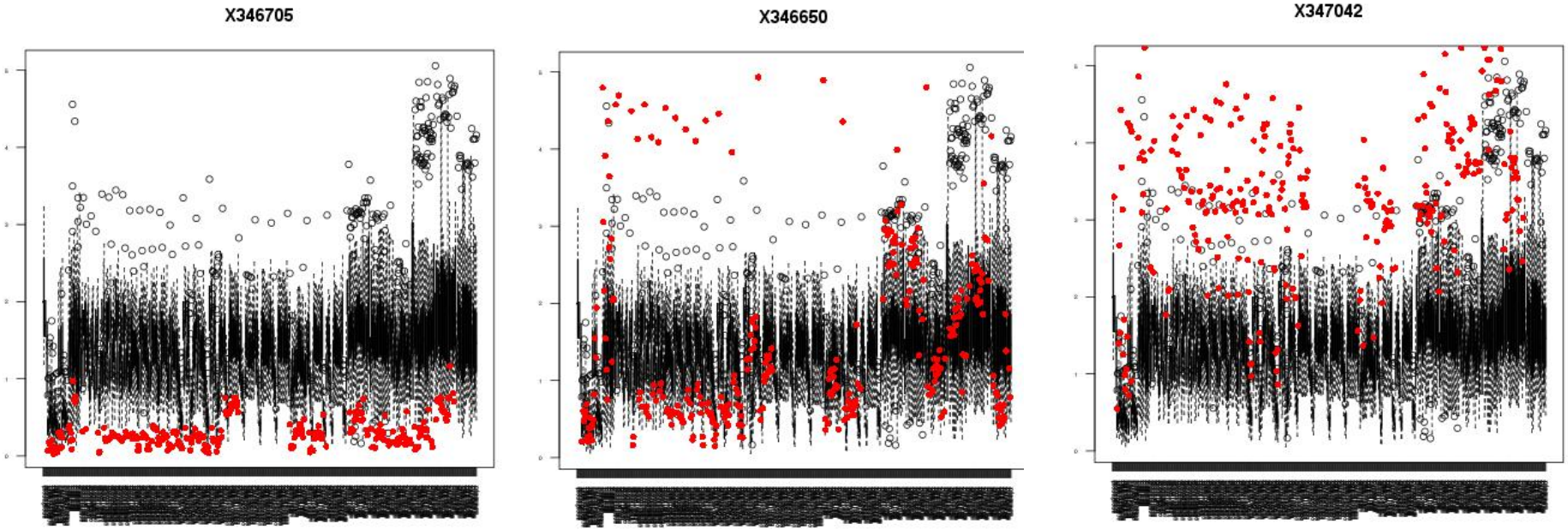
- Calculated for each marker:
  - Pairwise* KR distances (might be better for this type of data)
  - Pairwise* UniFrac distances (more intuitive and more widely used – is it worthwhile to compare KR/unifrac results in a systematic way?)
  - Alpha diversity metrics for each marker (default in guppy): phylogenetic entropy (Allen 2009), quadratic entropy (Rao 1982, Warwick and Clark 1995) phylogenetic diversity (Faith 1992), phylogenetic diversity which only requires distal mass (this is as oppposed to pd requiring both distal and proximal mass), and a new diversity metric generalizing PD to incorporate abundance: balance-weighted phylogenetic diversity (are any of these appropriate/useful?)
  - Performed squash clustering (allows tree distance metrics with samples as tips, if I can find an appropriate test.)

- Working on: Writing Perl and R scripts to parse these gigantic distance matrices

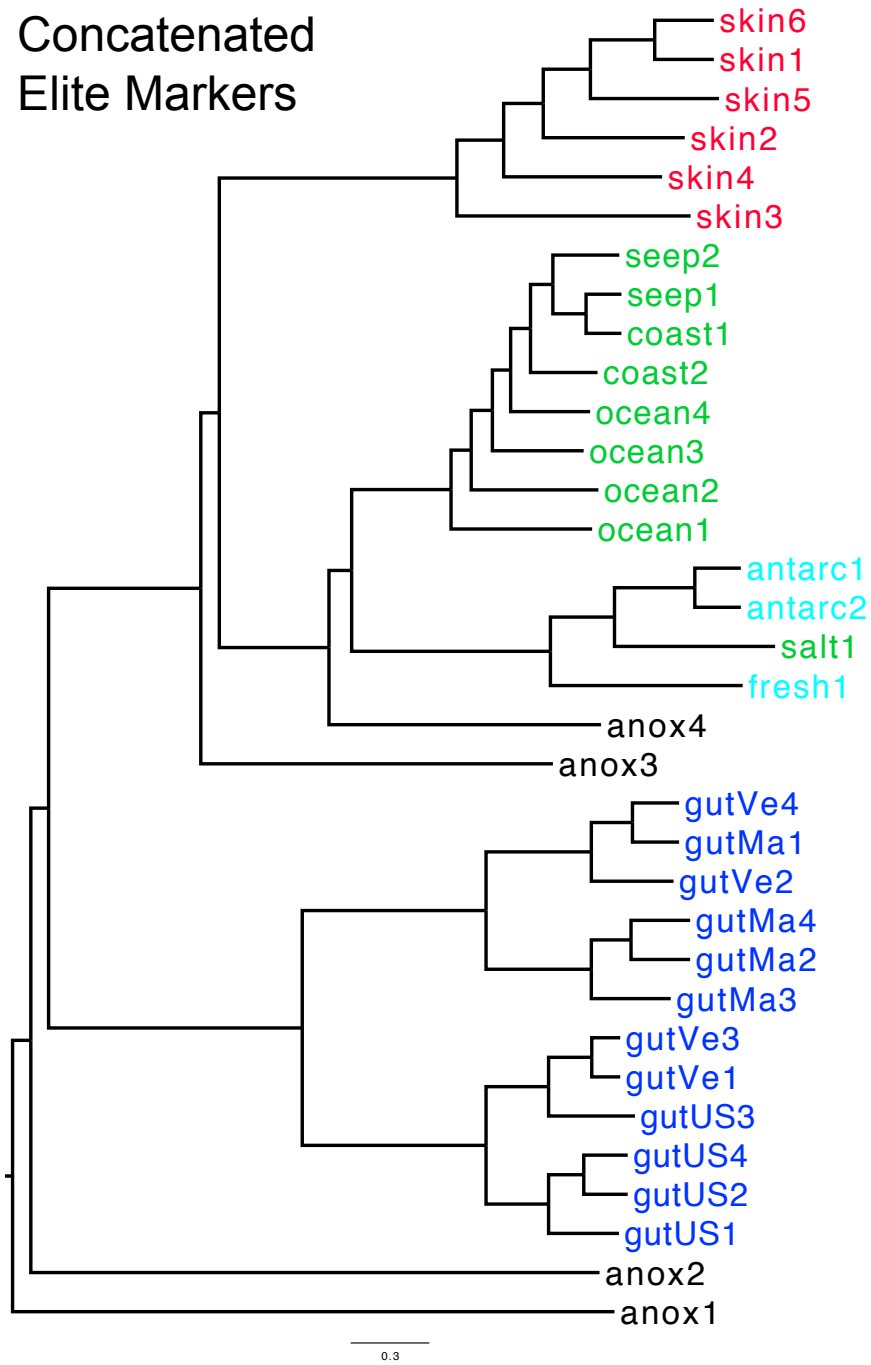
34 datasets = 561 pairwise comparisons



Pairwise KR distances (red dots) for 3 extended markers plotted against elite boxplots



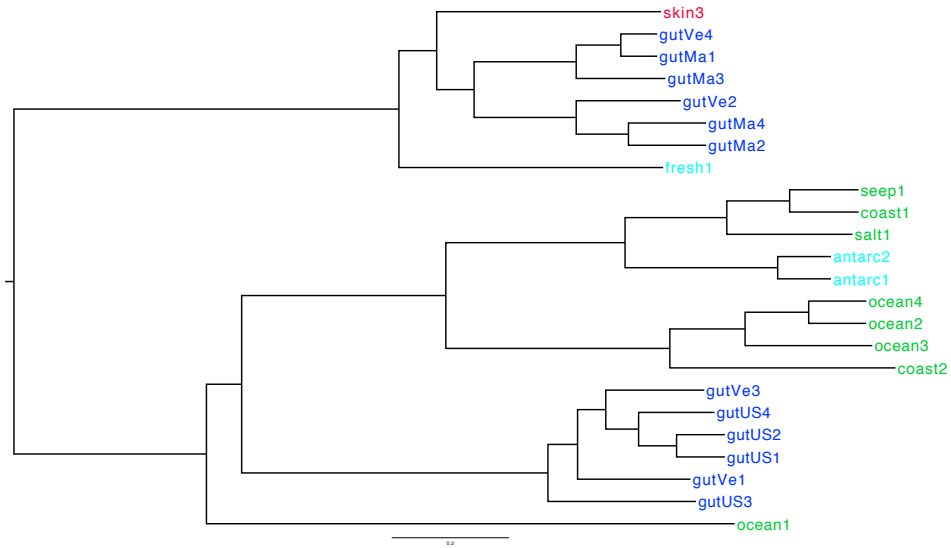
Concatenated
Elite Markers



Squash clustering may allow tree comparison tests as a way to compare communities

Variable communities = problem?

Tree1977



My Current Challenges

- Finding the true outliers (markers, environments, taxa) – multiple comparisons…
- Best practices for assigning functions to protein families
- Accounting for uncertainty, bias, error
  - When a marker is absent in a sample…MISSING DATA
- Are there properties of metagenomes that might “mislead” diversity metrics – low coverage?

Assumptions/Expectations:

1. For a single marker, KR distance between two similar environments will be less than KR distance between two more different environments

2. 37 elite markers span the KR distances that can be expected to be taxonomically consistent

3. Markers well above or below the elite range might be special – especially when you look at their value across pairwise comparisons.