

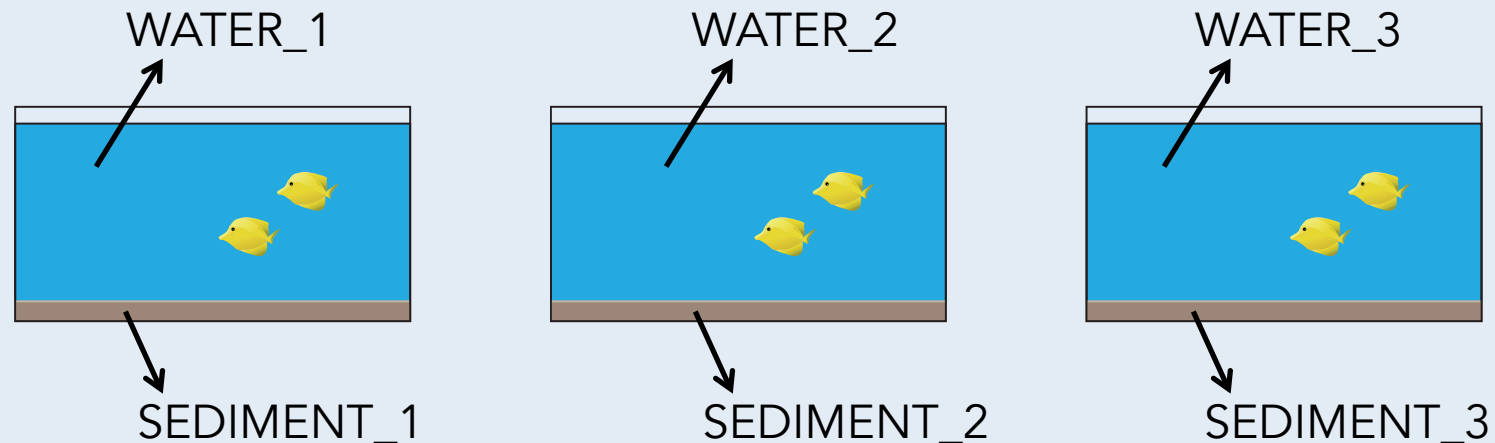
This talk

- How I'm framing this research
- Some results
- Advice, please.

Metagenomics Basics

Current metagenomics studies frequently:

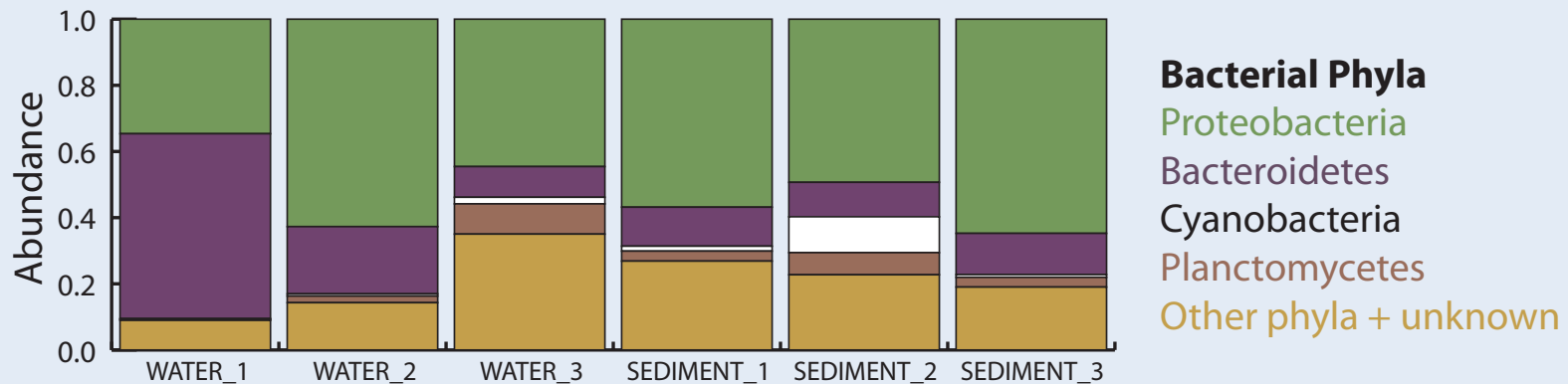
- collect environmental samples
- sequence
- use "elite" markers to estimate taxonomic composition of the sample
- use "regular" markers to estimate functional profile of the sample



Metagenomics Basics

Current metagenomics studies frequently:

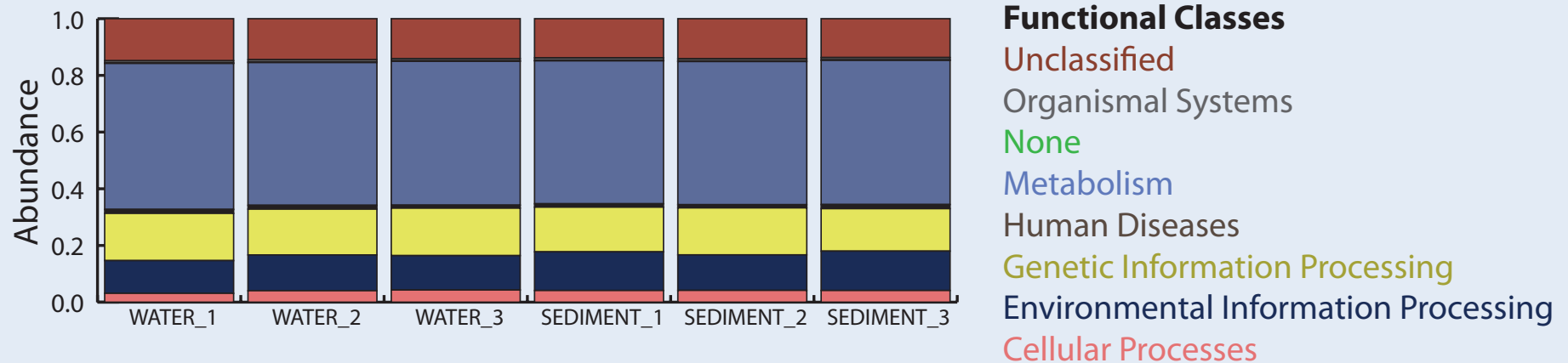
- collect environmental samples
- sequence
- use “**elite**” markers to estimate taxonomic composition of the sample
- use “regular” markers to estimate functional profile of the sample.



Metagenomics Basics

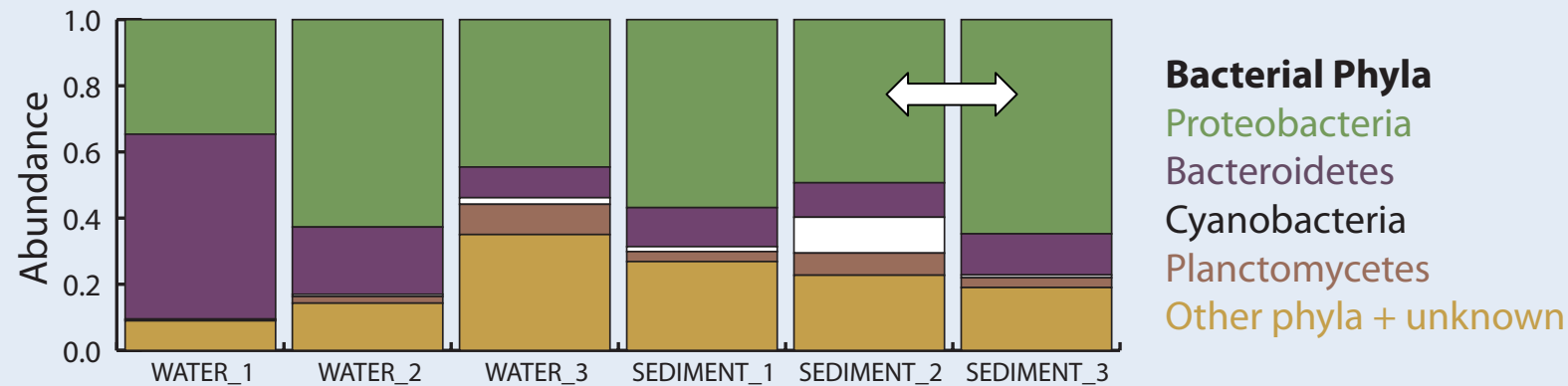
Current metagenomics studies frequently:

- collect environmental samples
- sequence
- use "elite" markers to estimate taxonomic composition of the sample
- use "regular" markers to estimate functional profile of the sample

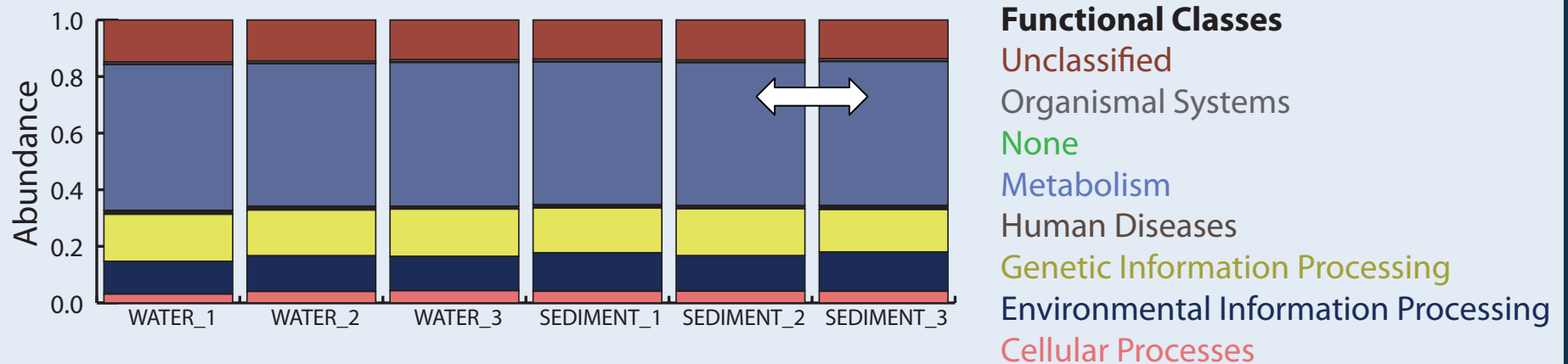


Limited Comparisons

Use "elite" markers to estimate taxonomy:

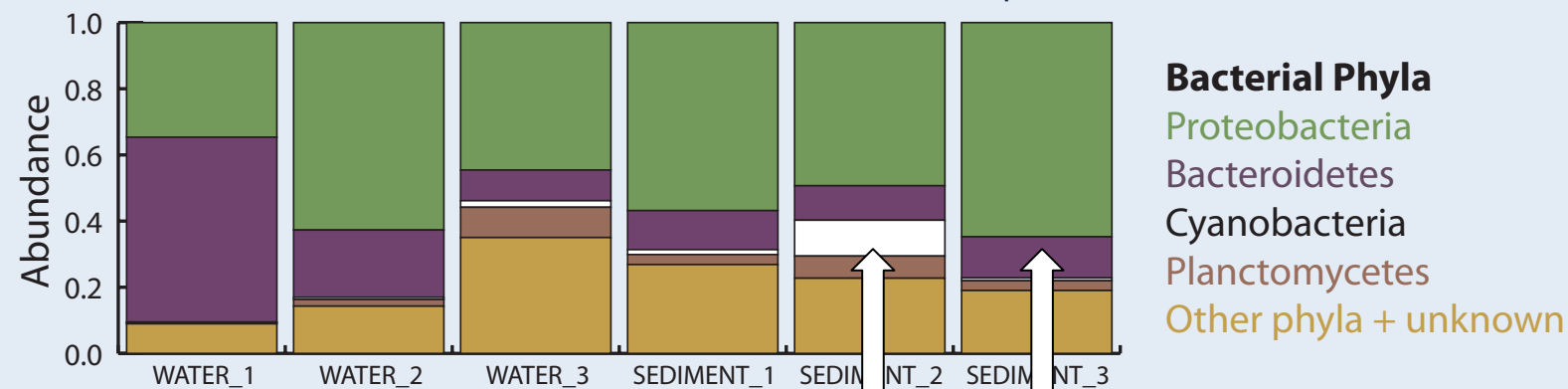


Use "regular" markers to assign function:



Limited Comparisons

Use "elite" (taxonomically consistent, single copy) markers to estimate taxonomy:



Bacterial Phyla

Proteobacteria

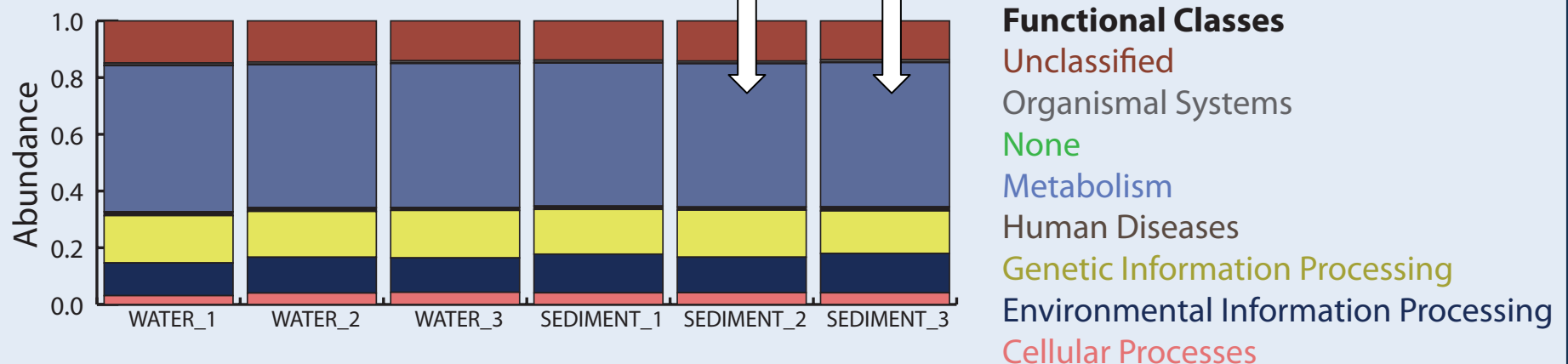
Bacteroidetes

Cyanobacteria

Planctomycetes

Other phyla + unknown

Use "regular" markers to assign function :



Functional Classes

Unclassified

Organismal Systems

None

Metabolism

Human Diseases

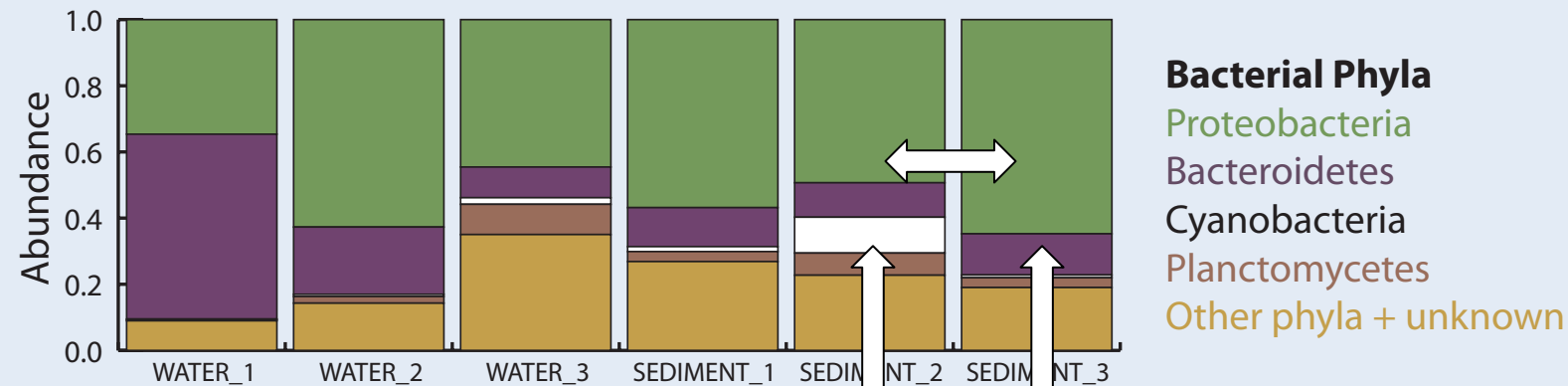
Genetic Information Processing

Environmental Information Processing

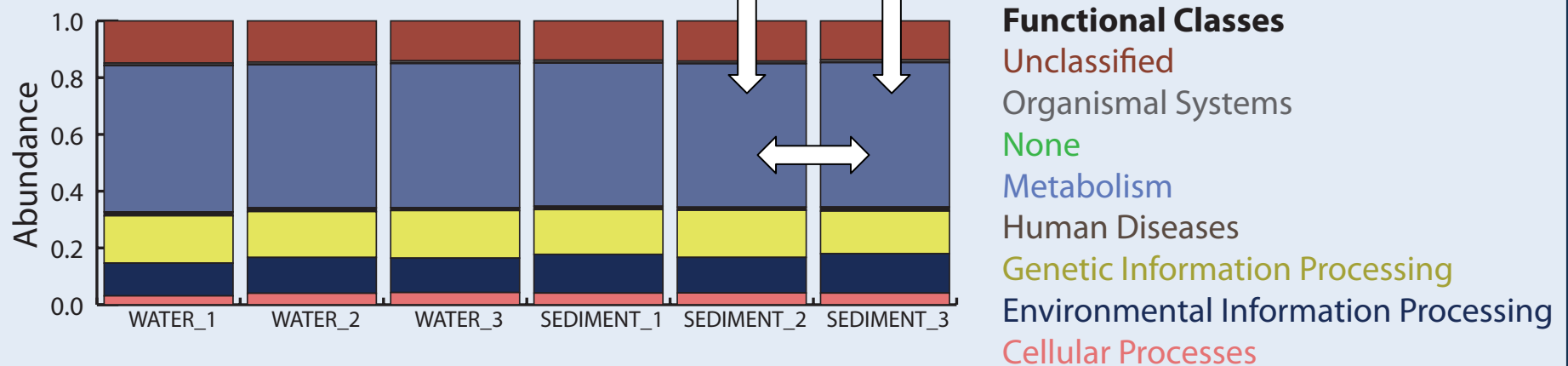
Cellular Processes

Limited Comparisons

Use "elite" (taxonomically consistent, single copy) markers to estimate taxonomy:

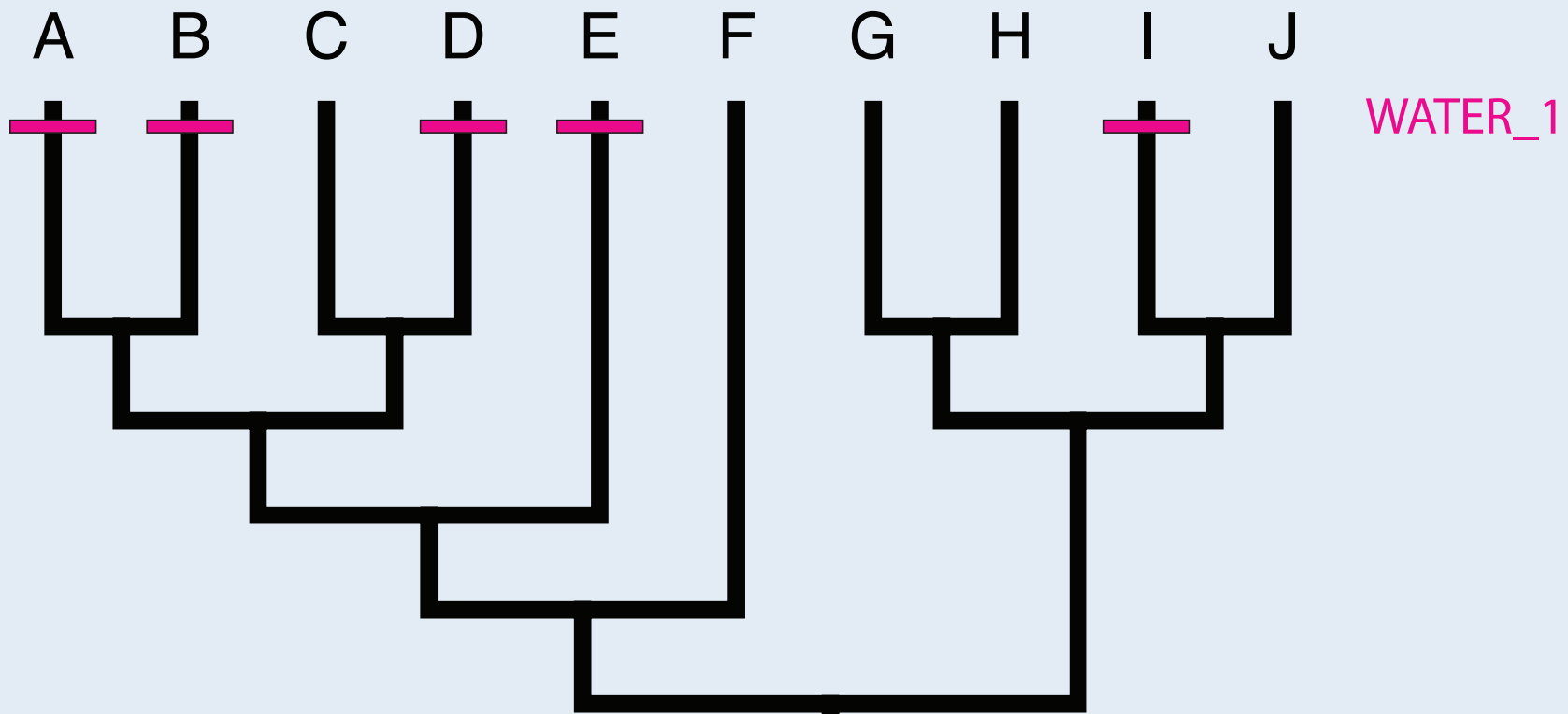


Use "regular" markers to assign function:



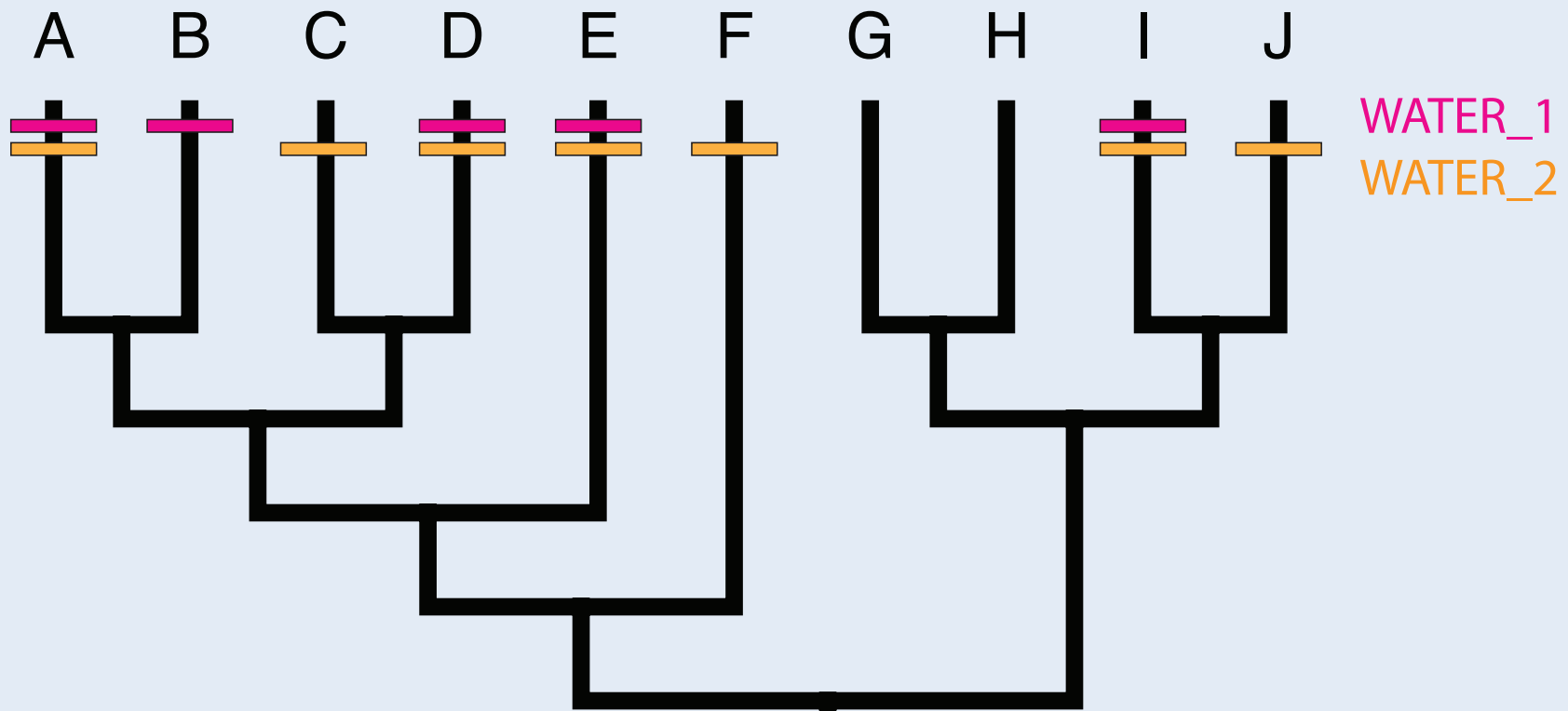
Phylogenetic Approach

With the same information we use to assign taxonomy to sequences, we can instead analyze phylogenetic trees which inherently contain more information than lists.



Phylogenetic Approach

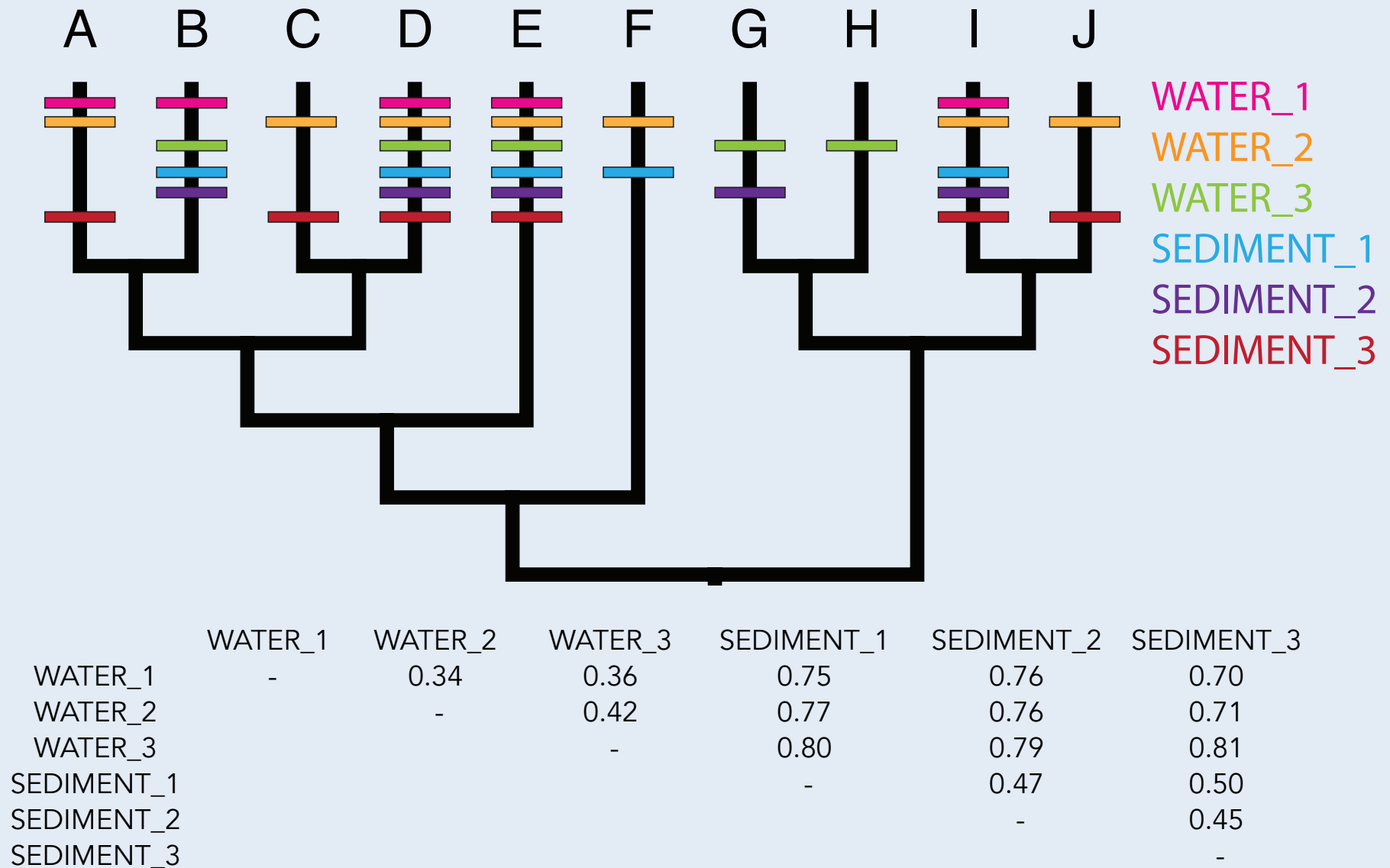
Calculate a distance between these two samples



Distance
WATER_1 : WATER_2 = 0.342

Phylogenetic Approach Creates Distance Matrix

Generate a pairwise distance matrix

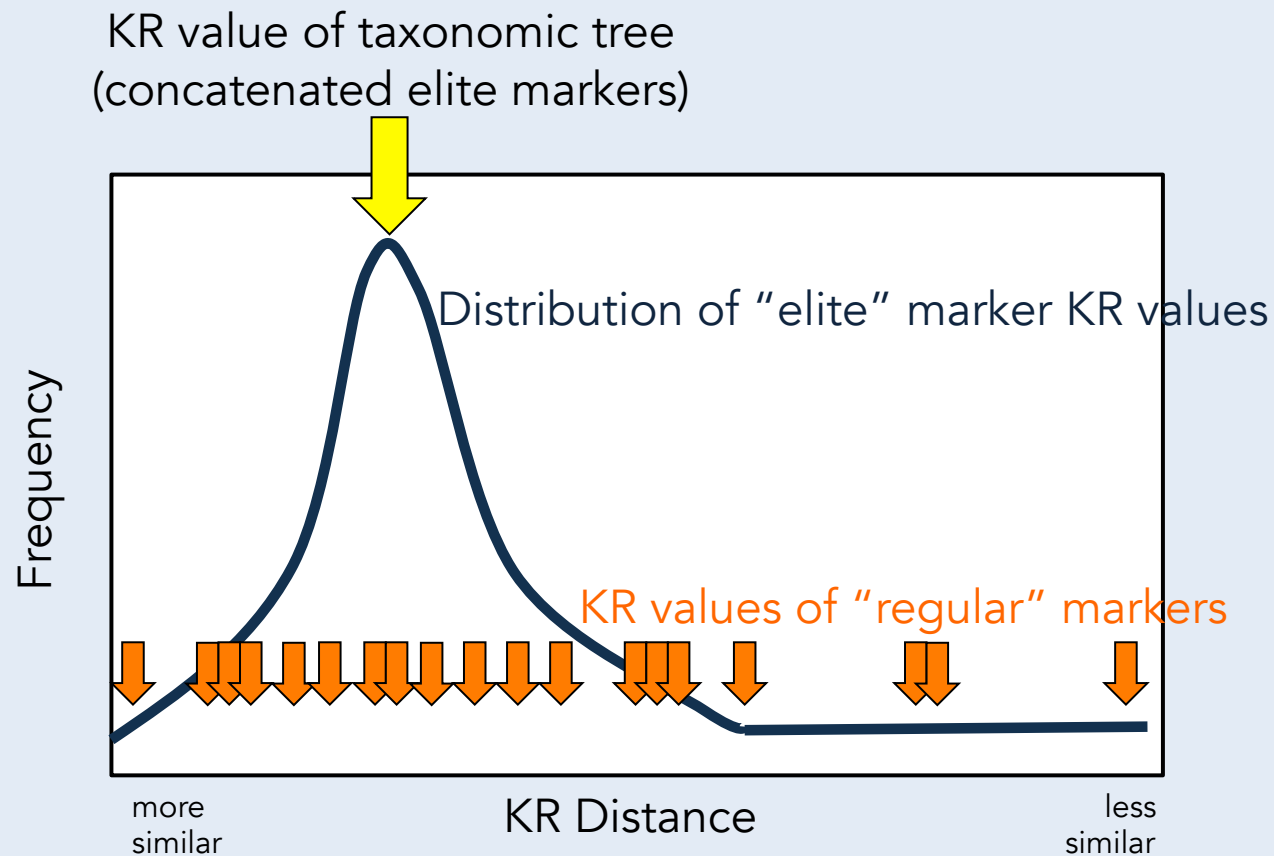


Phylogenetic Approach Creates Distance Matrix

		Markers							
		concat	elite1	elite2	elite3	regular1	regular2	regular3	regular4
Pairwise Comparisons	W1_W2	0.296	0.329	0.353	0.233	0.344	1.029	0.367	2.173
	W1_W3	0.362	0.466	0.558	0.656	0.473	-	0.478	2.587
	W2_W3	0.209	0.208	0.476	0.459	0.481	-	0.460	2.271
	W1_S1	0.172	0.445	0.130	0.364	0.339	0.392	0.546	2.332
	W1_S2	0.384	0.336	0.308	0.216	0.440	1.034	0.368	2.243
	W1_S3	1.661	0.940	1.331	1.888	1.991	1.748	1.937	3.334
	W2_S1	1.372	0.922	1.396	1.935	1.823	2.137	1.823	3.313
	W2_S2	0.348	0.396	0.655	0.569	0.399	0.725	0.500	2.593
	W2_S3	0.247	0.238	0.409	0.400	0.324	1.051	0.244	2.159
	W3_S1	0.305	0.340	0.313	0.352	0.161	-	0.367	2.283
	W3_S2	0.579	0.168	0.170	0.248	0.167	-	0.270	2.148
	W3_S3	1.754	1.007	1.191	1.896	1.895	-	1.924	3.272
	S1_S2	1.465	1.050	1.358	1.901	1.694	1.931	1.768	3.250
	S1_S3	0.365	0.382	0.933	0.581	0.387	0.761	0.527	2.625
	S2_S3	0.417	0.326	0.541	0.731	0.367	1.000	0.589	2.556

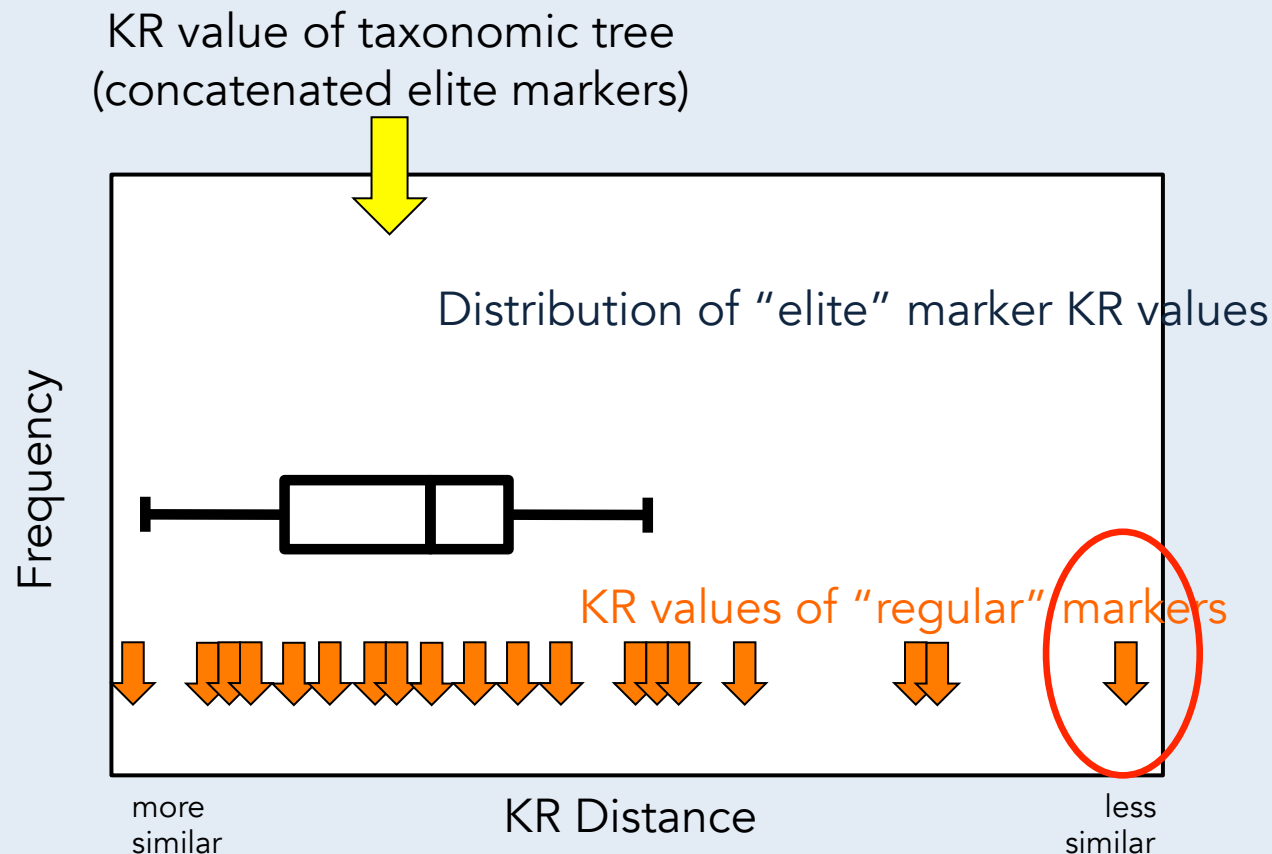
Phylogenetic Approach Yields Expectations

For each pairwise sample comparison, we've generated a distribution of "elite" distances and point estimates for all "regular" markers



Phylogenetic Approach Yields Expectations

For each pairwise sample comparison, we've generated a distribution of "elite" distances and point estimates for all "regular" markers



Current Research

Methods & Expectations

The Data and Pipeline

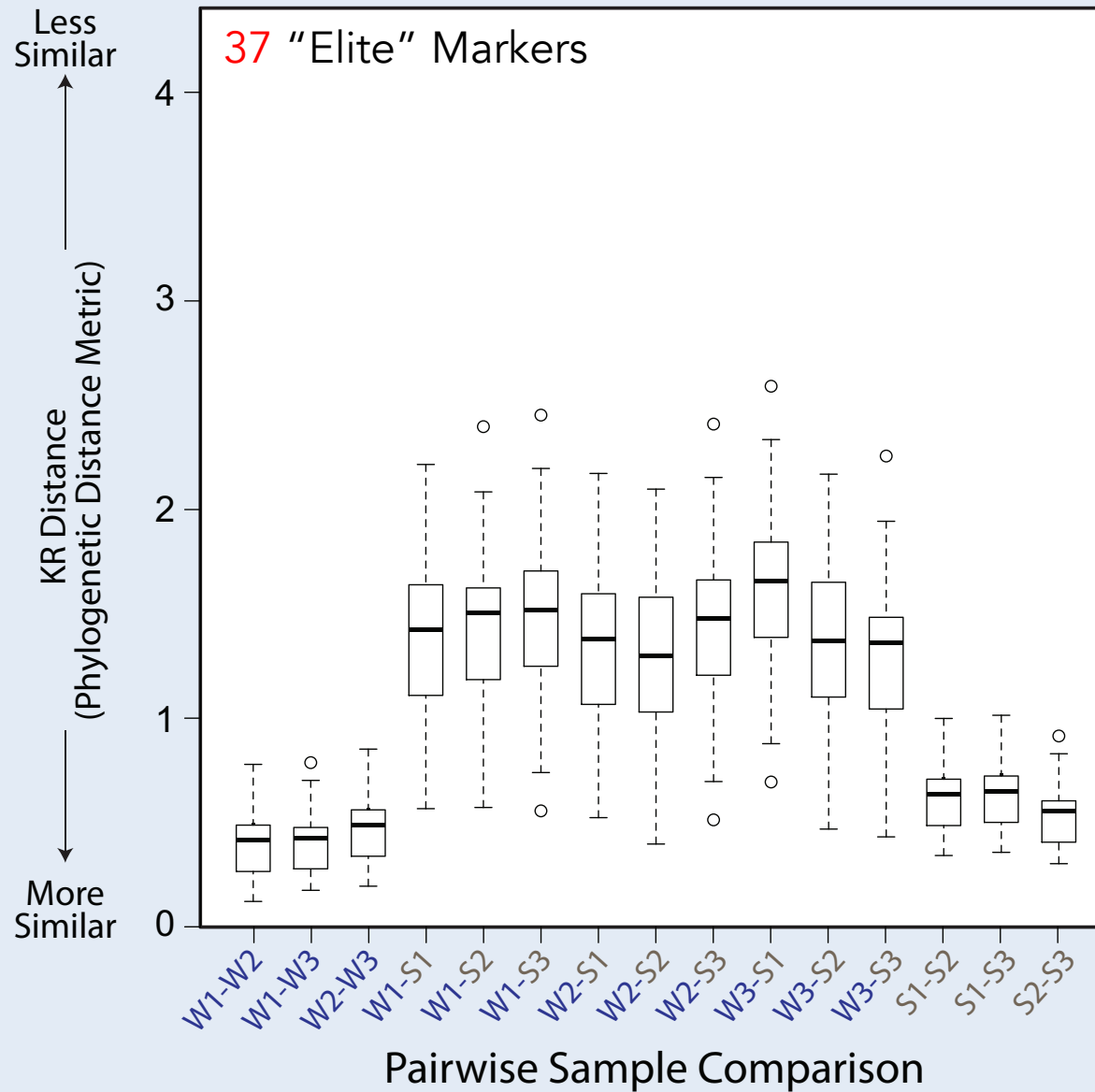
INPUT: Metagenomes



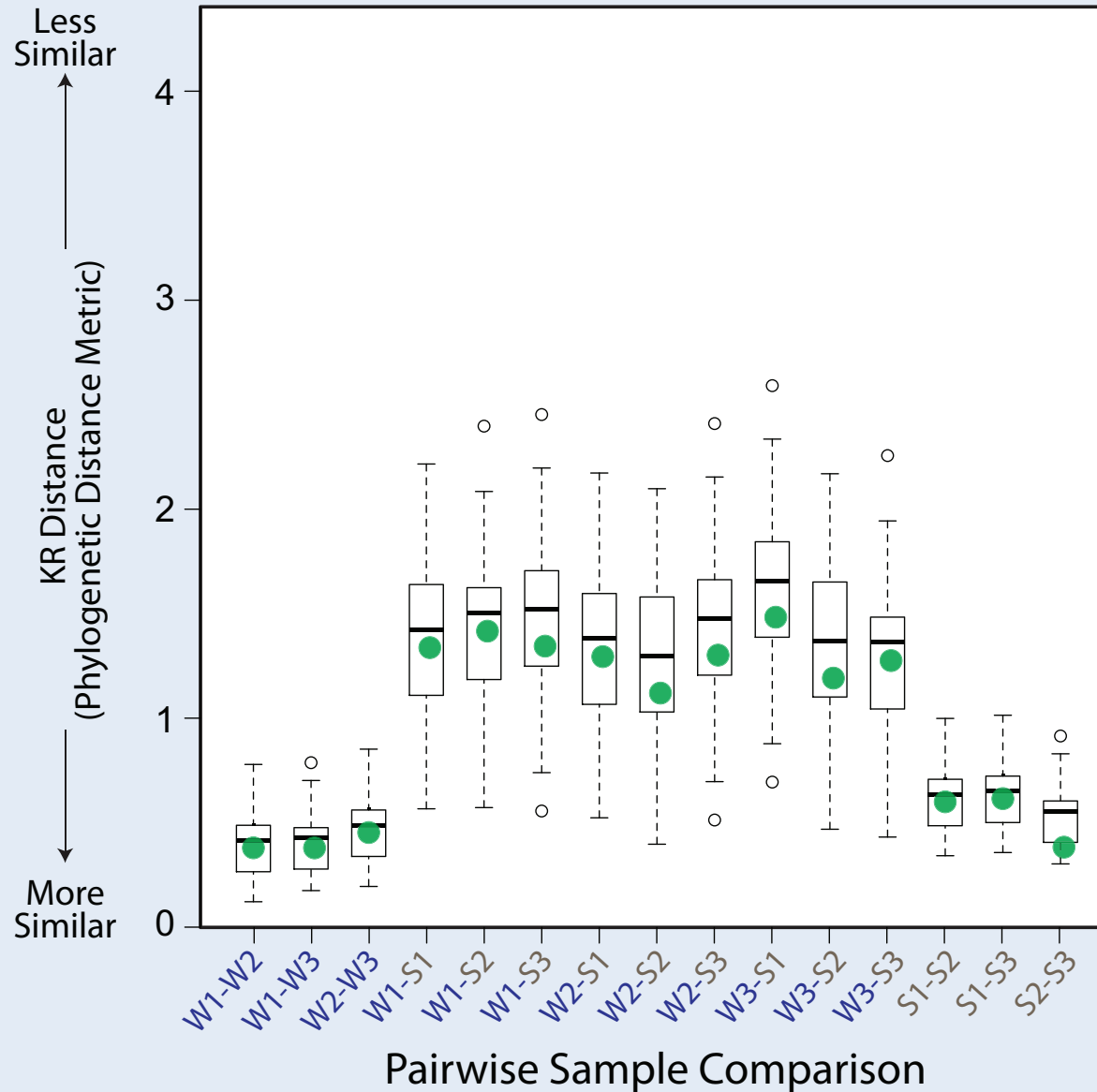
***Kantorovich-Rubenstein (KR) Distance:**
Generalized version of the more familiar weighted UniFrac, accounts for uncertainty in phylogenetic placements

OUTPUT: Table of pairwise distances for 5800 protein families

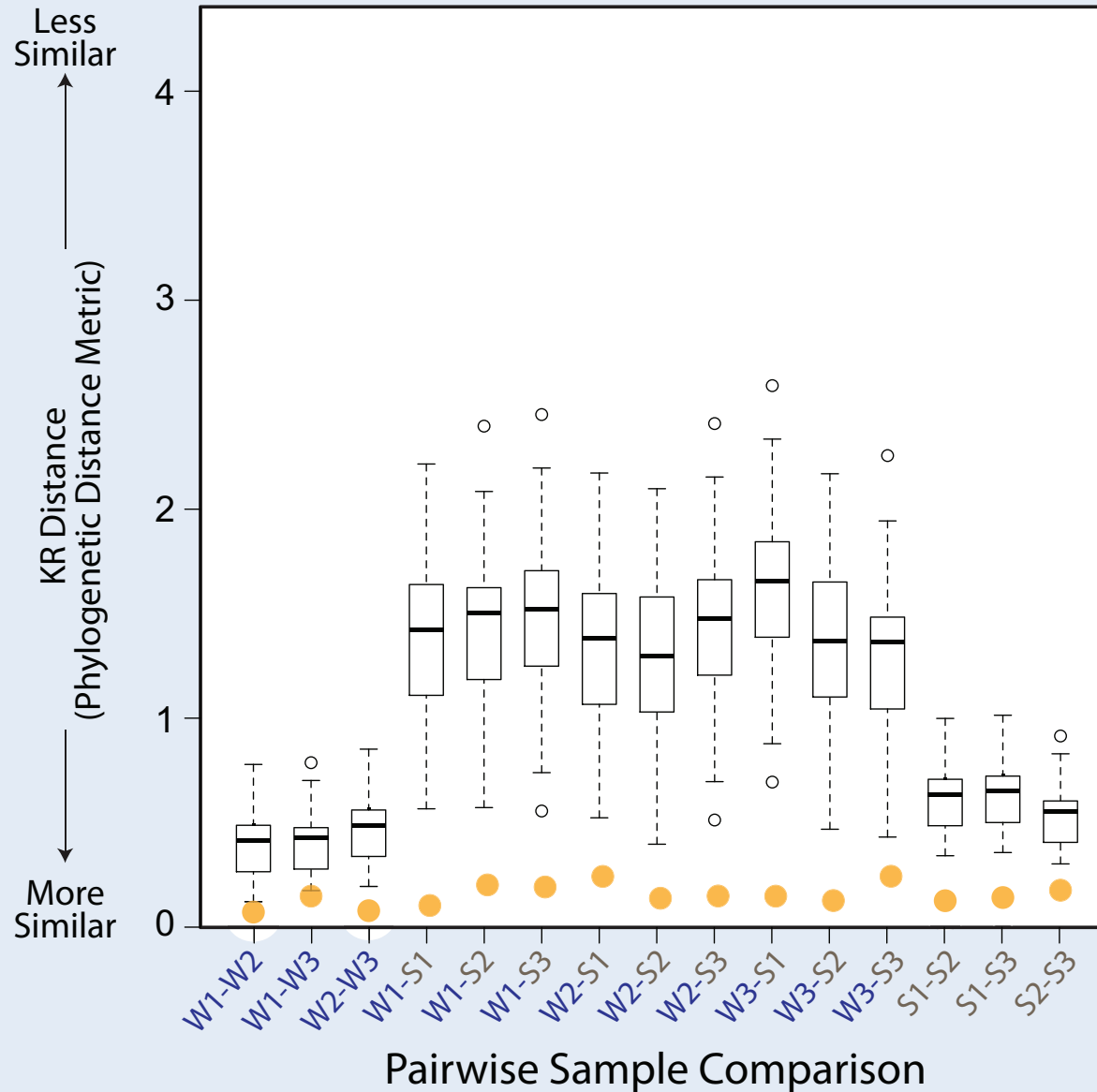
Aquarium Example Continued



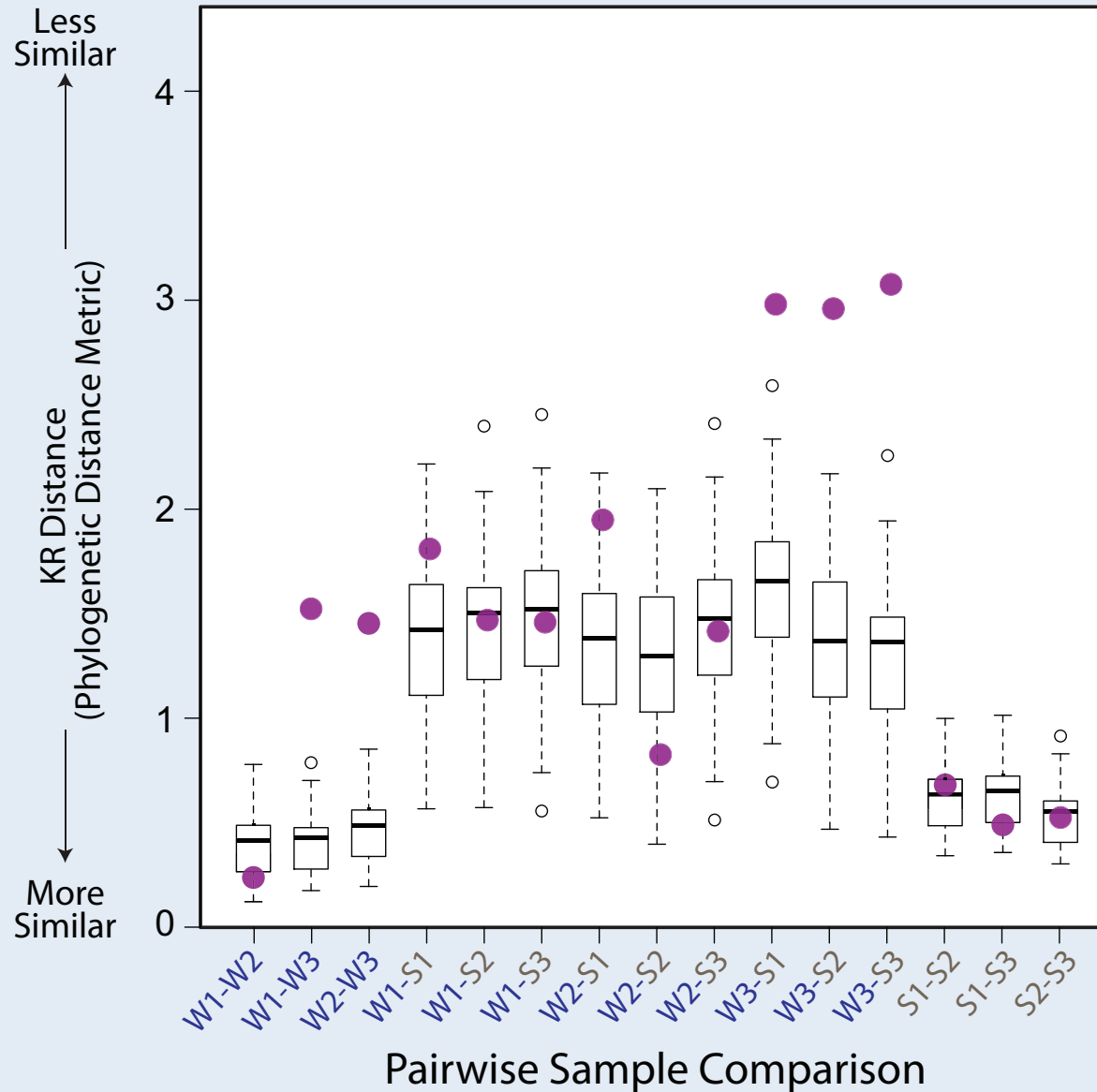
Aquarium Example Continued: The Green Gene



Aquarium Example Continued: The Orange Gene



Aquarium Example Continued: The Purple Gene



Current Research

Results & Discussion

Datasets

Completed:

Dataset 34

- 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

In Progress:

Global Gut Dataset (Yatsunenkeno et al. 2012)

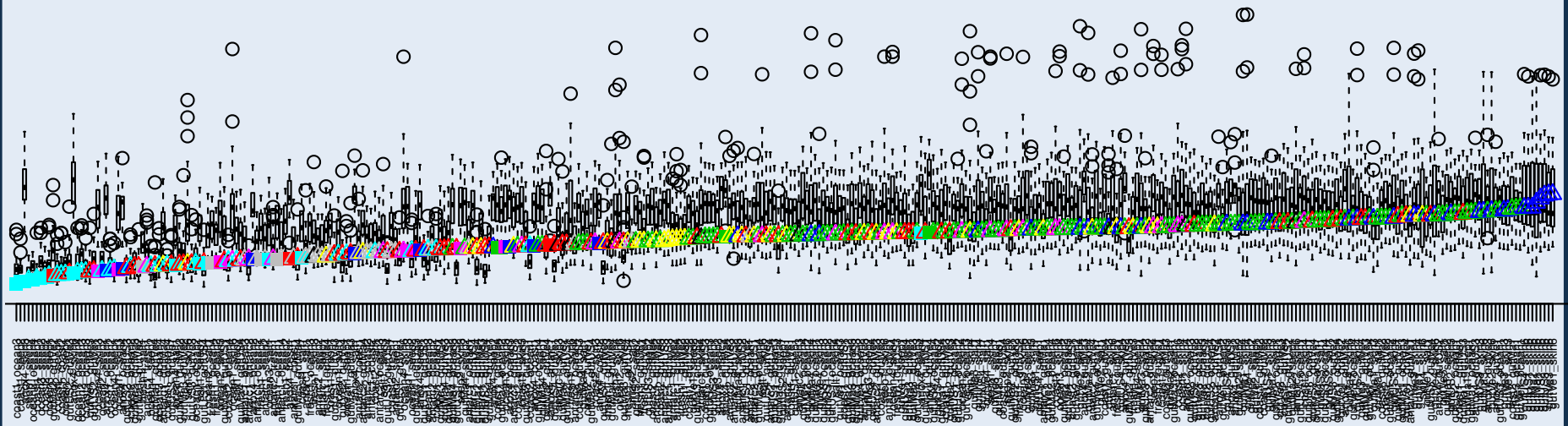
- 23 Malawi (1 month – 27 years old)
- 22 Venezuela (3 months – 53 years old)
- 66 USA (1 month – 50 years old)

Simulation Study



Family 10827 is a typical result...

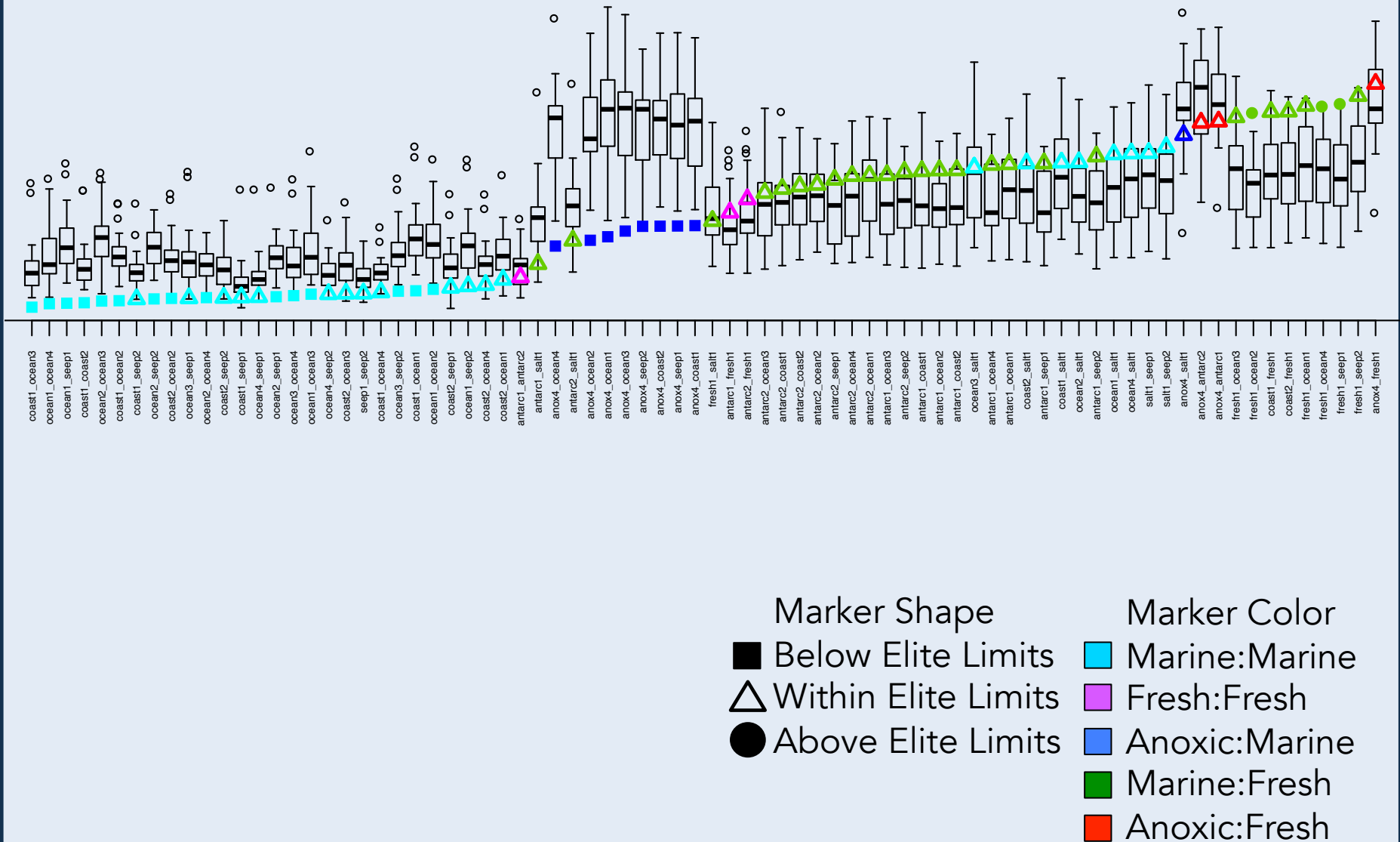
...nearly all the comparisons fall within the range of “elite” values.



- Marker Shape
- Below Elite Limits
 - △ Within Elite Limits
 - Above Elite Limits

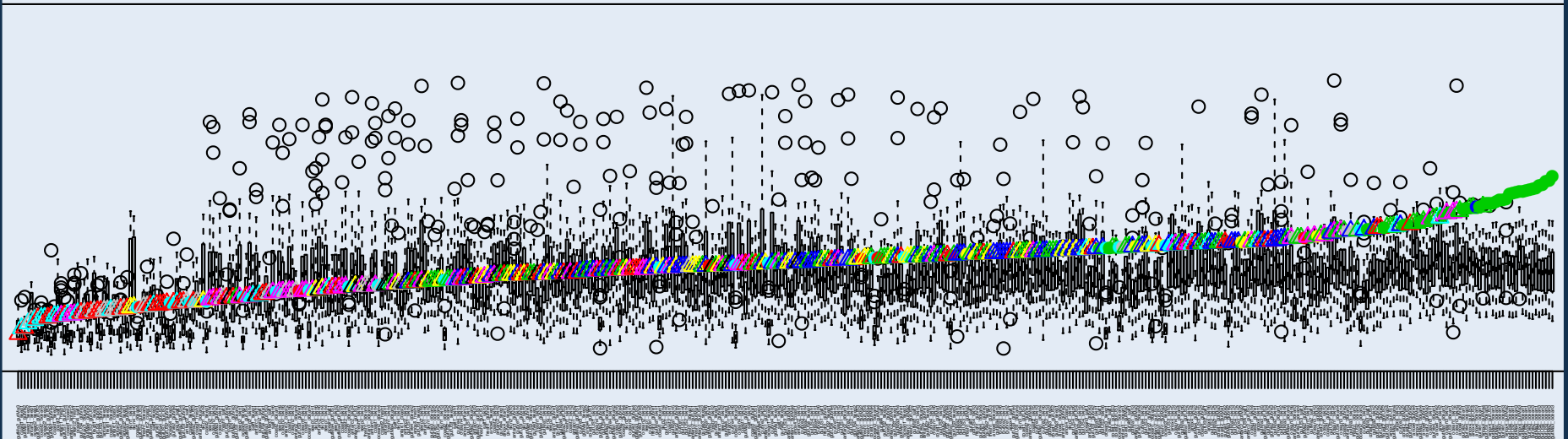
*The y-axis is KR Distance

Proteorhodopsin behaves as expected



*The y-axis is KR Distance

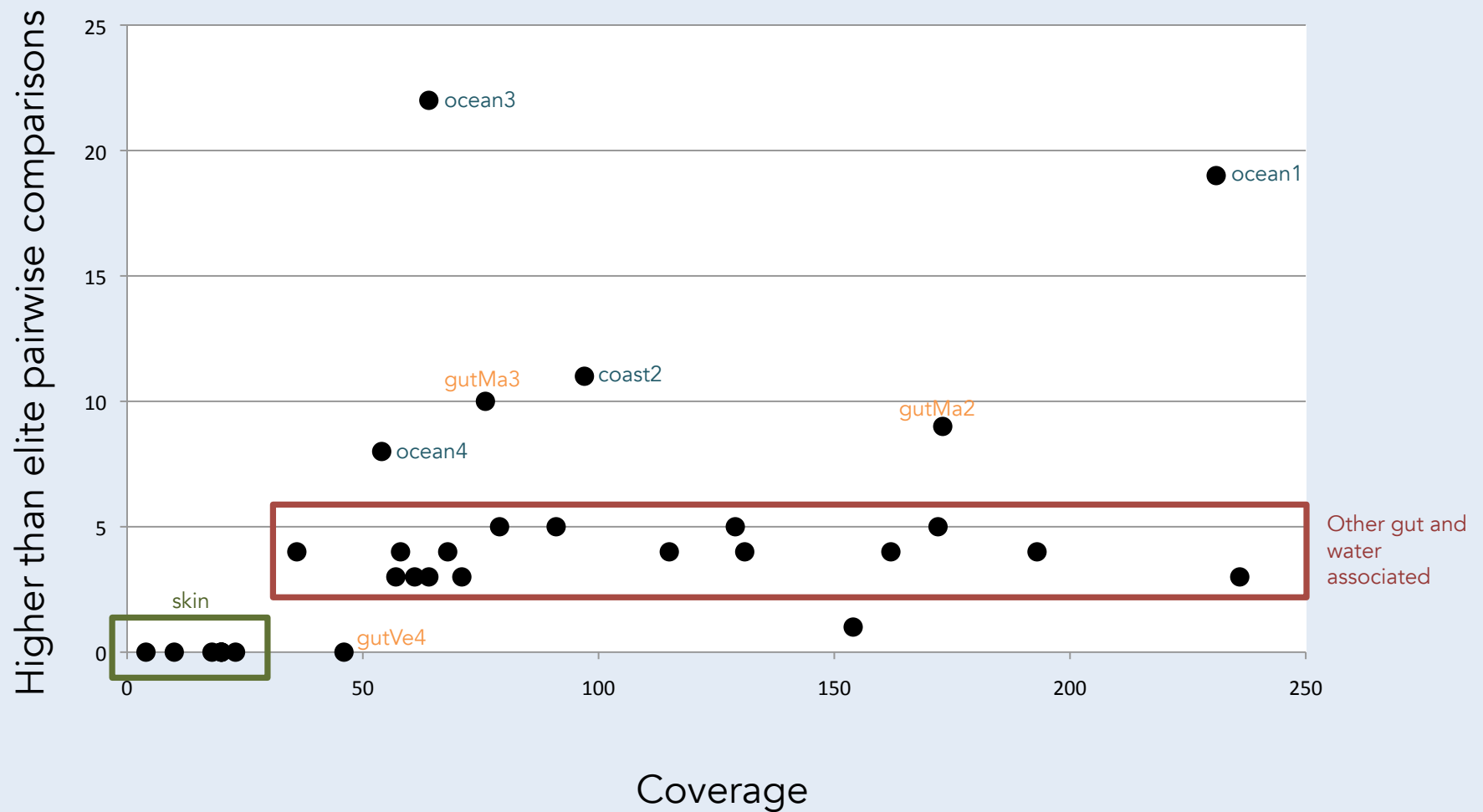
Family 1905 is promising. (?)



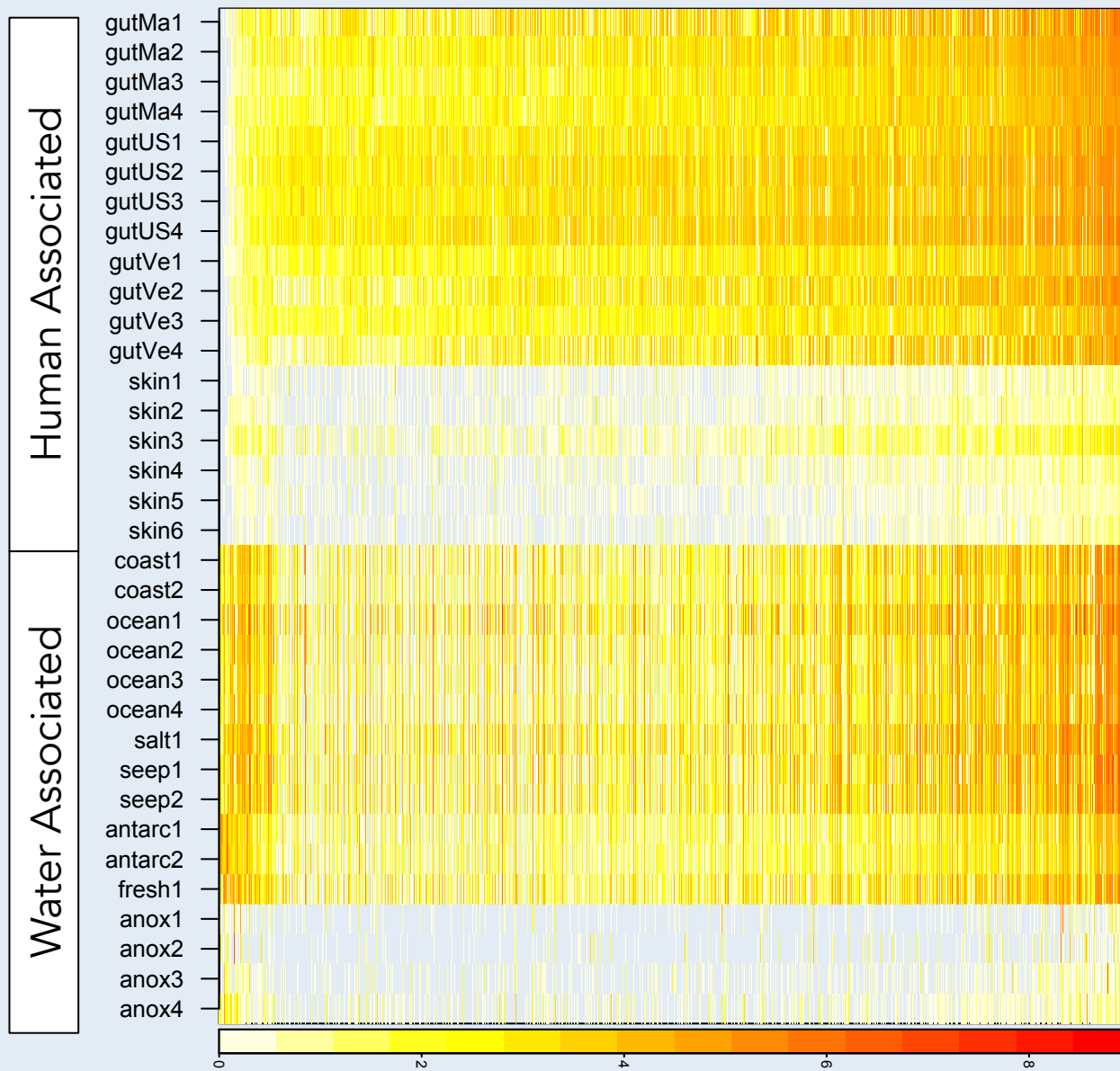
Marker Shape

- Below Elite Limits
- △ Within Elite Limits
- Above Elite Limits

Family 1905 is promising. (?)

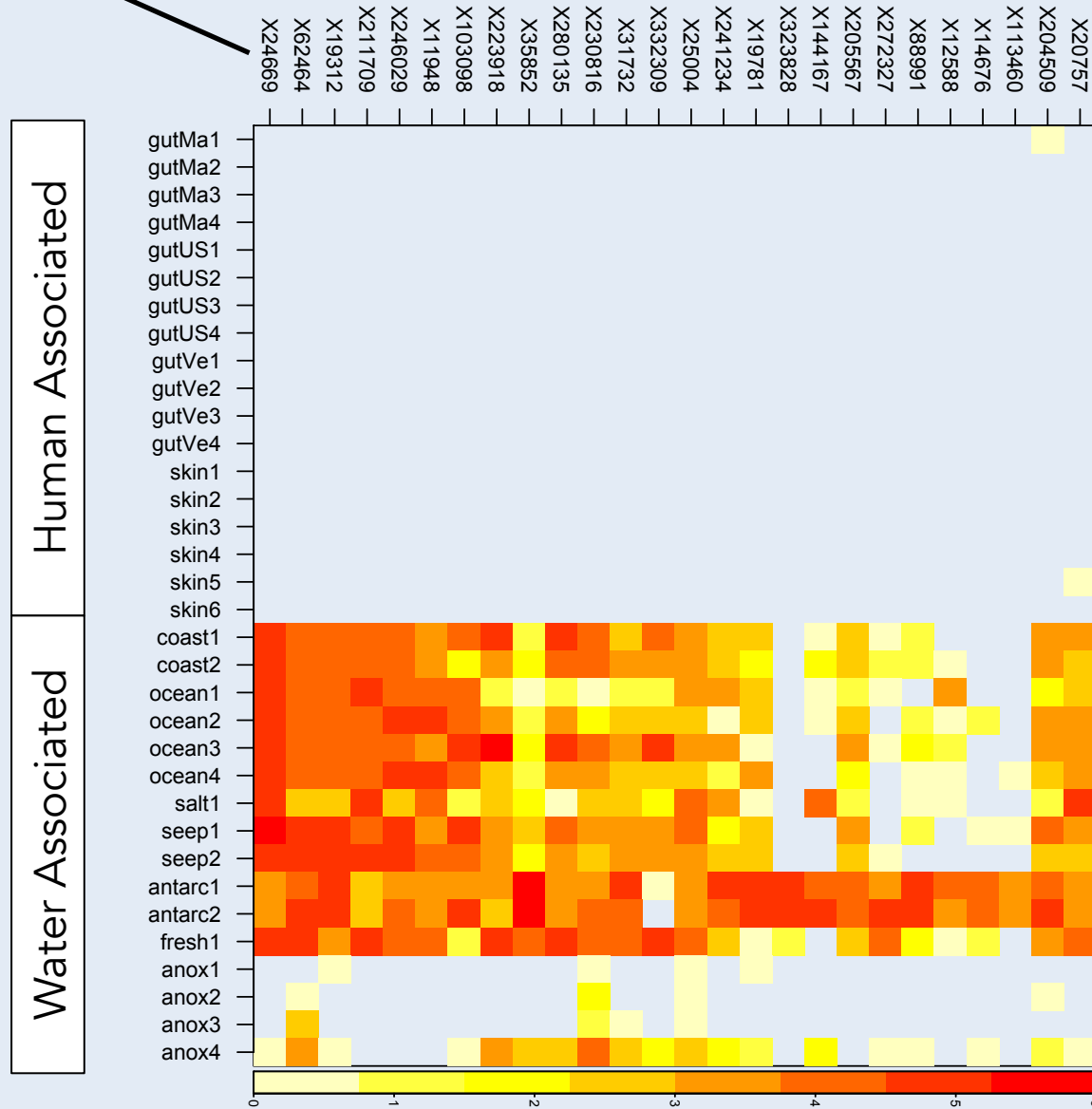


The coverage issue

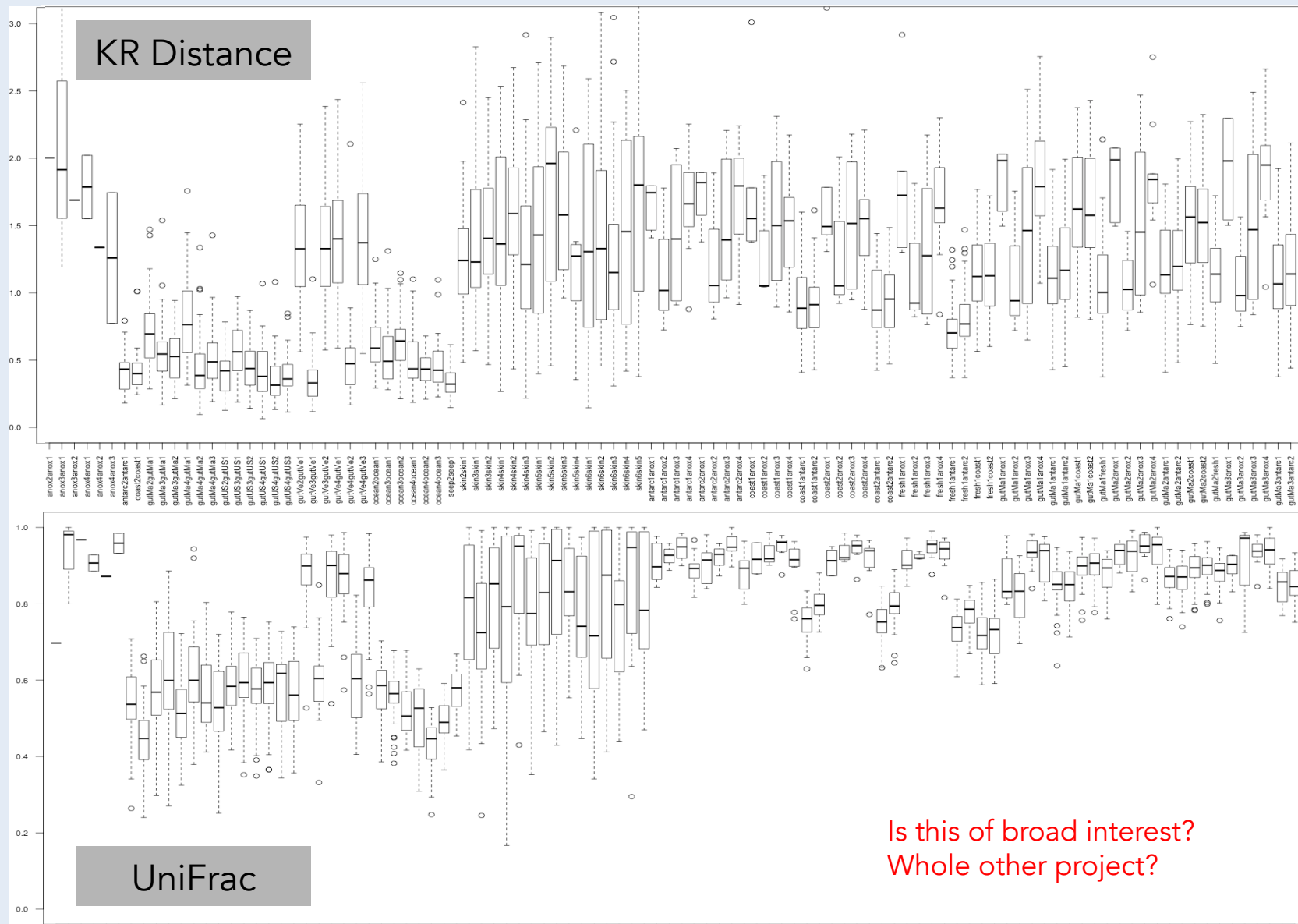


Presence/absence/coverage may be informative

Proteorhodopsin



Similarities & differences between KR & UniFrac



What's next

Path of least resistance to publication

Is a convincing simulation the key?

Assigning function to Sfams = difficult.

Positive controls (other than proteorhodopsin?)