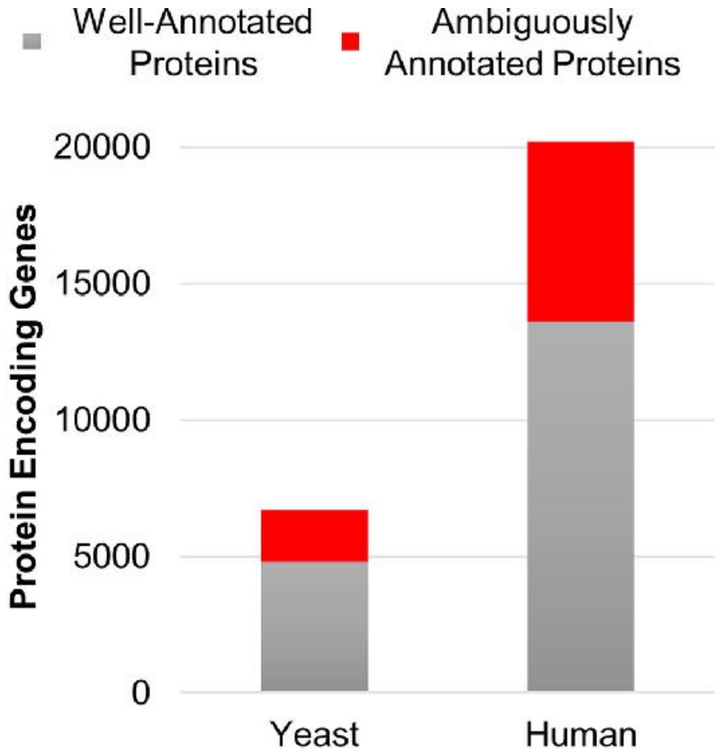
Max Planck Institute for Marine Microbiology
Microbial Genomics and Bioinformatics Group

Ecological implications and characterization of the functional unknown fraction of the marine environment

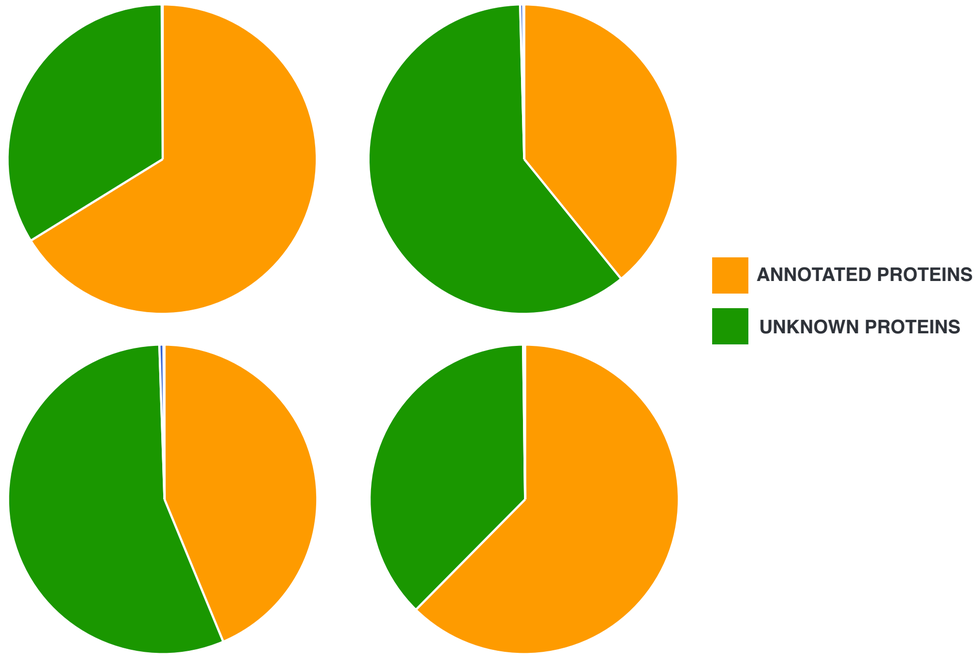
M. Sc. student: Matthew Schechter
email: mschecht@mpi-bremen.de
Supervisor: Dr. Antonio Fernandez-Guerra
Group leader: Prof. Dr. Frank Oliver Glöckner

Introduction to the Unknowns

Model organism genomes

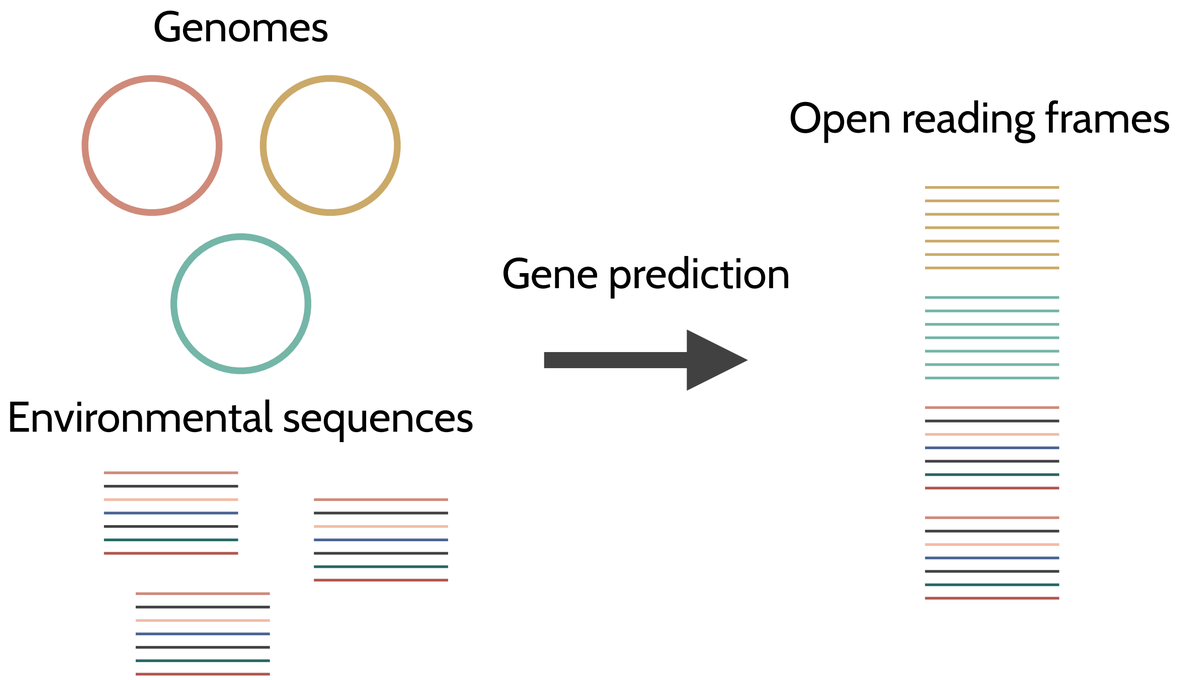


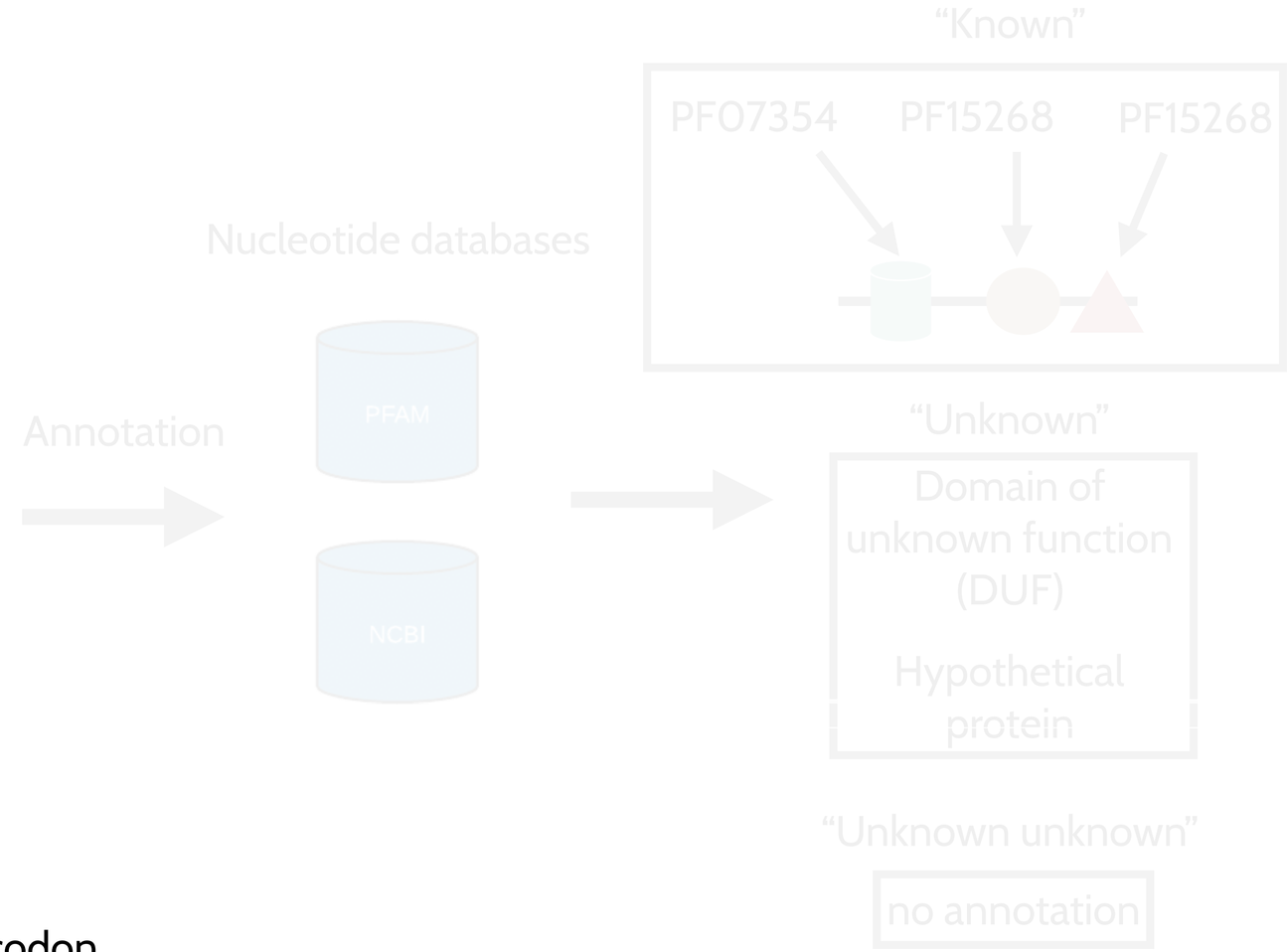
Environmental metagenomic data (MG-RAST)



Ellens, 2017

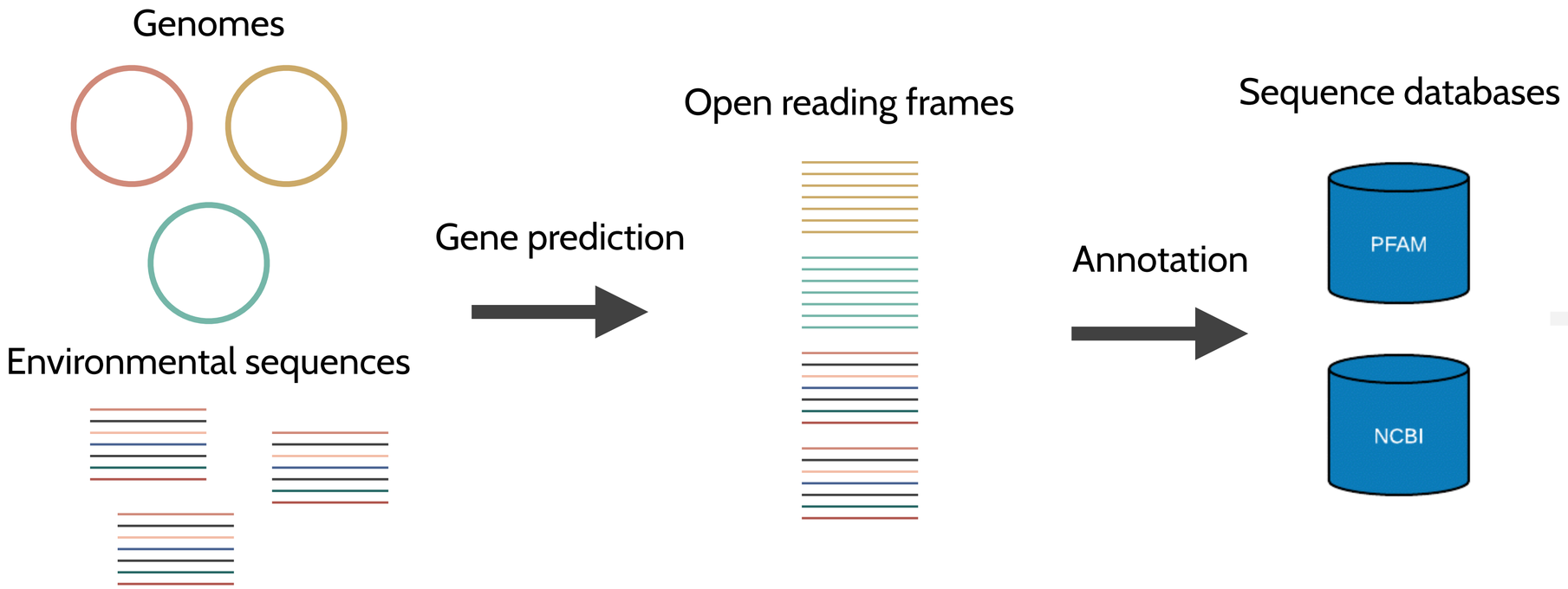
Pathway to unknowns





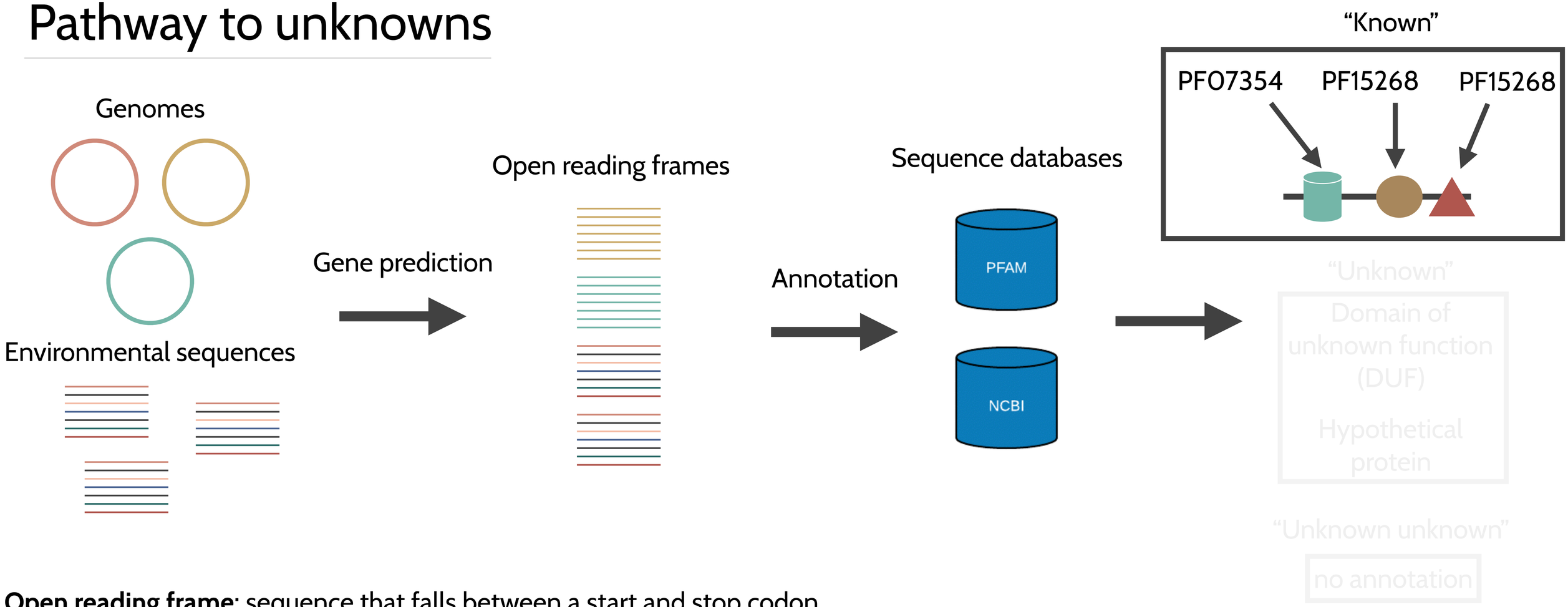
- **Open reading frame**: sequence that falls between a start and stop codon
- **Protein families**: groups of proteins that share an evolutionary/functional relationship (homologs)
- **Protein domains (Pfam)**: conserved functional/structural regions of proteins, usually responsible for overall role of protein
- **Pfam domain architecture**: the order of Pfam domain annotations on an open reading frame

Pathway to unknowns



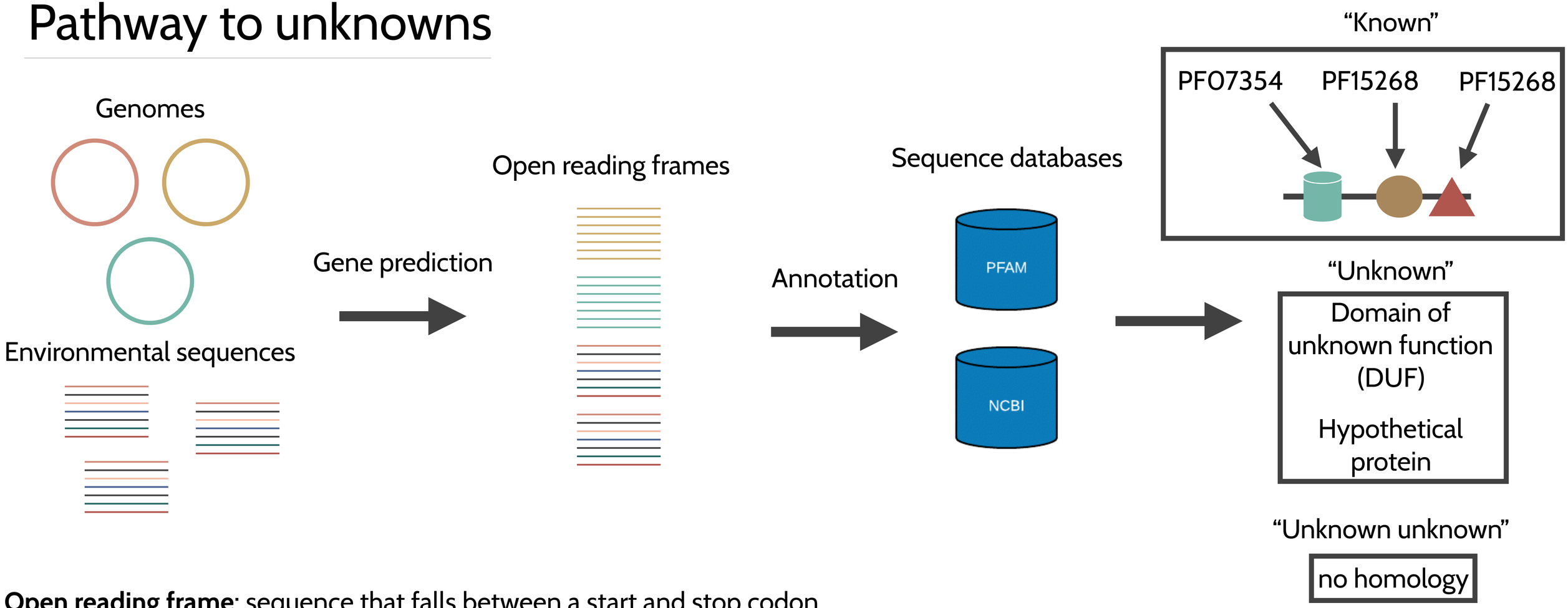
- **Open reading frame**: sequence that falls between a start and stop codon
- **Protein families**: groups of proteins that share an evolutionary/functional relationship (homologs)
- **Protein domains (Pfam)**: conserved functional/structural regions of proteins, usually responsible for overall role of protein
- **Pfam domain architecture**: the order of Pfam domain annotations on an open reading frame

Pathway to unknowns



- **Open reading frame**: sequence that falls between a start and stop codon
- **Protein families**: groups of proteins that share an evolutionary/function relationship (homologs)
- **Protein domains (Pfam)**: conserved functional/structural regions of proteins, usually responsible for overall role of protein
- **Pfam domain architecture**: the order of Pfam domain annotations on an open reading frame

Pathway to unknowns



- **Open reading frame**: sequence that falls between a start and stop codon
- **Protein families**: groups of proteins that share an evolutionary/function relationship (homologs)
- **Protein domains (Pfam)**: conserved functional/structural regions of proteins, usually responsible for overall role of protein
- **Pfam domain architecture**: the order of Pfam domain annotations on an open reading frame

How have the unknowns been approached?

Biochemistry and crystallography

Exploration of Uncharted Regions of the Protein Universe

Lukasz Jaroszewski[1], Zhanwen Li[2], S. Sri Krishna[1], Constantina Bakolitsa[1], John Wooley[3], Ashley M. Deacon[4], Ian A. Wilson[5], Adam Godzik[1,2,3]*

Ecological co-occurence

Ecogenomic Perspectives on Domains of Unknown Function: Correlation-Based Exploration of Marine Metagenomes

Pier Luigi Buttigieg[1,2]*, Wolfgang Hankeln[1], Ivaylo Kostadinov[3], Renzo Kottmann[1], Pelin Yilmaz[1], Melissa Beth Duhaime[4], Frank Oliver Glöckner[1,2]

Clustering approach

The *Sorcerer II* Global Ocean Sampling Expedition: Expanding the Universe of Protein Families

Shibu Yooseph[1*], Granger Sutton[1], Douglas B. Rusch[1], Aaron L. Halpern[1], Shannon J. Williamson[1], Karin Remington[1], Jonathan A. Eisen[1,2], Karla B. Heidelberg[1], Gerard Manning[3], Weizhong Li[4], Lukasz Jaroszewski[4], Piotr Cieplak[4], Christopher S. Miller[5], Huiying Li[5], Susan T. Mashiyama[6], Marcin P. Joachimiak[6], Christopher van Belle[6], John-Marc Chandonia[6,7], David A. Soergel[6], Yufeng Zhai[3], Kannan Natarajan[8], Shaun Lee[8], Benjamin J. Raphael[9], Vineet Bafna[8], Robert Friedman[1], Steven E. Brenner[6], Adam Godzik[4], David Eisenberg[5], Jack E. Dixon[8], Susan S. Taylor[8], Robert L. Strausberg[1], Marvin Frazier[1], J. Craig Venter[1]

Illuminating structural proteins in viral "dark matter" with metaproteomics

Jennifer R. Brum[a,1,2], J. Cesar Ignacio-Espinoza[b,1,3], Eun-Hae Kim[c,1,4], Gareth Trubl[c,2], Robert M. Jones[c,5], Simon Roux[a,2], Nathan C. VerBerkmoes[d,6], Virginia I. Rich[c,2,7], and Matthew B. Sullivan[a,b,c,2,7]

A most wanted list of conserved protein families with no known domains

Stacia K. Wyman, Aram Avila-Herrera, Stephen Nayfach, Katherine S. Pollard

Brum et al. 2016; Buttigieg et al. 2013; Jaroszewski et al. 2009; Wyman et al. 2017; Yooseph et al. 2007

How have the unknowns been approached?

Biochemistry and crystallography

Exploration of Uncharted Regions of the Protein Universe

Lukasz Jaroszewski[1], Zhanwen Li[2], S. Sri Krishna[1], Constantina Bakolitsa[1], John Wooley[3], Ashley M. Deacon[4], Ian A. Wilson[5], Adam Godzik[1,2,3]*

Ecological co-occurence

Ecogenomic Perspectives on Domains of Unknown Function: Correlation-Based Exploration of Marine Metagenomes

Pier Luigi Buttigieg[1,2]*, Wolfgang Hankeln[1], Ivaylo Kostadinov[3], Renzo Kottmann[1], Pelin Yilmaz[1], Melissa Beth Duhaime[4], Frank Oliver Glöckner[1,2]

Clustering approach

The *Sorcerer II* Global Ocean Sampling Expedition: Expanding the Universe of Protein Families

Shibu Yooseph[1*], Granger Sutton[1], Douglas B. Rusch[1], Aaron L. Halpern[1], Shannon J. Williamson[1], Karin Remington[1], Jonathan A. Eisen[1,2], Karla B. Heidelberg[1], Gerard Manning[3], Weizhong Li[4], Lukasz Jaroszewski[4], Piotr Cieplak[4], Christopher S. Miller[5], Huiying Li[5], Susan T. Mashiyama[6], Marcin P. Joachimiak[6], Christopher van Belle[6], John-Marc Chandonia[6,7], David A. Soergel[6], Yufeng Zhai[3], Kannan Natarajan[8], Shaun Lee[8], Benjamin J. Raphael[9], Vineet Bafna[8], Robert Friedman[1], Steven E. Brenner[6], Adam Godzik[4], David Eisenberg[5], Jack E. Dixon[8], Susan S. Taylor[8], Robert L. Strausberg[1], Marvin Frazier[1], J. Craig Venter[1]

Illuminating structural proteins in viral "dark matter" with metaproteomics

Jennifer R. Brum[a,1,2], J. Cesar Ignacio-Espinoza[b,1,3], Eun-Hae Kim[c,1,4], Gareth Trubl[c,2], Robert M. Jones[c,5], Simon Roux[a,2], Nathan C. VerBerkmoes[d,6], Virginia I. Rich[c,2,7], and Matthew B. Sullivan[a,b,c,2,7]

A most wanted list of conserved protein families with no known domains

Stacia K. Wyman, Aram Avila-Herrera, Stephen Nayfach, Katherine S. Pollard

Brum et al. 2016; Buttigieg et al. 2013; Jaroszewski et al. 2009; Wyman et al. 2017; Yooseph et al. 2007

How have the unknowns been approached?

Biochemistry and crystallography

Exploration of Uncharted Regions of the Protein Universe

Lukasz Jaroszewski[1], Zhanwen Li[2], S. Sri Krishna[1], Constantina Bakolitsa[1], John Wooley[3], Ashley M. Deacon[4], Ian A. Wilson[5], Adam Godzik[1,2,3]*

Ecological co-occurence

Ecogenomic Perspectives on Domains of Unknown Function: Correlation-Based Exploration of Marine Metagenomes

Pier Luigi Buttigieg[1,2]*, Wolfgang Hankeln[1], Ivaylo Kostadinov[3], Renzo Kottmann[1], Pelin Yilmaz[1], Melissa Beth Duhaime[4], Frank Oliver Glöckner[1,2]

Clustering approach

The *Sorcerer II* Global Ocean Sampling Expedition: Expanding the Universe of Protein Families

Shibu Yooseph[1*], Granger Sutton[1], Douglas B. Rusch[1], Aaron L. Halpern[1], Shannon J. Williamson[1], Karin Remington[1], Jonathan A. Eisen[1,2], Karla B. Heidelberg[1], Gerard Manning[3], Weizhong Li[4], Lukasz Jaroszewski[4], Piotr Cieplak[4], Christopher S. Miller[5], Huiying Li[5], Susan T. Mashiyama[6], Marcin P. Joachimiak[6], Christopher van Belle[6], John-Marc Chandonia[6,7], David A. Soergel[6], Yufeng Zhai[3], Kannan Natarajan[8], Shaun Lee[8], Benjamin J. Raphael[9], Vineet Bafna[8], Robert Friedman[1], Steven E. Brenner[6], Adam Godzik[4], David Eisenberg[5], Jack E. Dixon[8], Susan S. Taylor[8], Robert L. Strausberg[1], Marvin Frazier[1], J. Craig Venter[1]

Illuminating structural proteins in viral "dark matter" with metaproteomics

Jennifer R. Brum[a,1,2], J. Cesar Ignacio-Espinoza[b,1,3], Eun-Hae Kim[c,1,4], Gareth Trubl[c,2], Robert M. Jones[c,5], Simon Roux[a,2], Nathan C. VerBerkmoes[d,6], Virginia I. Rich[c,2,7], and Matthew B. Sullivan[a,b,c,2,7]

A most wanted list of conserved protein families with no known domains

Stacia K. Wyman, Aram Avila-Herrera, Stephen Nayfach, Katherine S. Pollard

Brum et al. 2016; Buttigieg et al. 2013; Jaroszewski et al. 2009; Wyman et al. 2017; Yooseph et al. 2007

How we are approaching the unknowns

Knowns

We know the function and the organism

Genomic Unknowns

We DO NOT know the function but we DO know the organism

Environmental Unknowns

We DO NOT know the function or the organism

Vanni et al. 2018 (in prep)

How we are approaching the unknowns

Knowns

We know the function and the organism

Genomic Unknowns

We DO NOT know the function but we DO know the organism

Environmental Unknowns

We DO NOT know the function or the organism

Vanni et al. 2018 (in prep)

Our dataset

Human Microbiome Project
Metagenomes: 1,249
ORFs: 162,687,295

TARA
Metagenomes: 242
ORFs: 111,903,261

Ocean Sampling Day
Metagenomes: 150
ORFs: 7,015,383

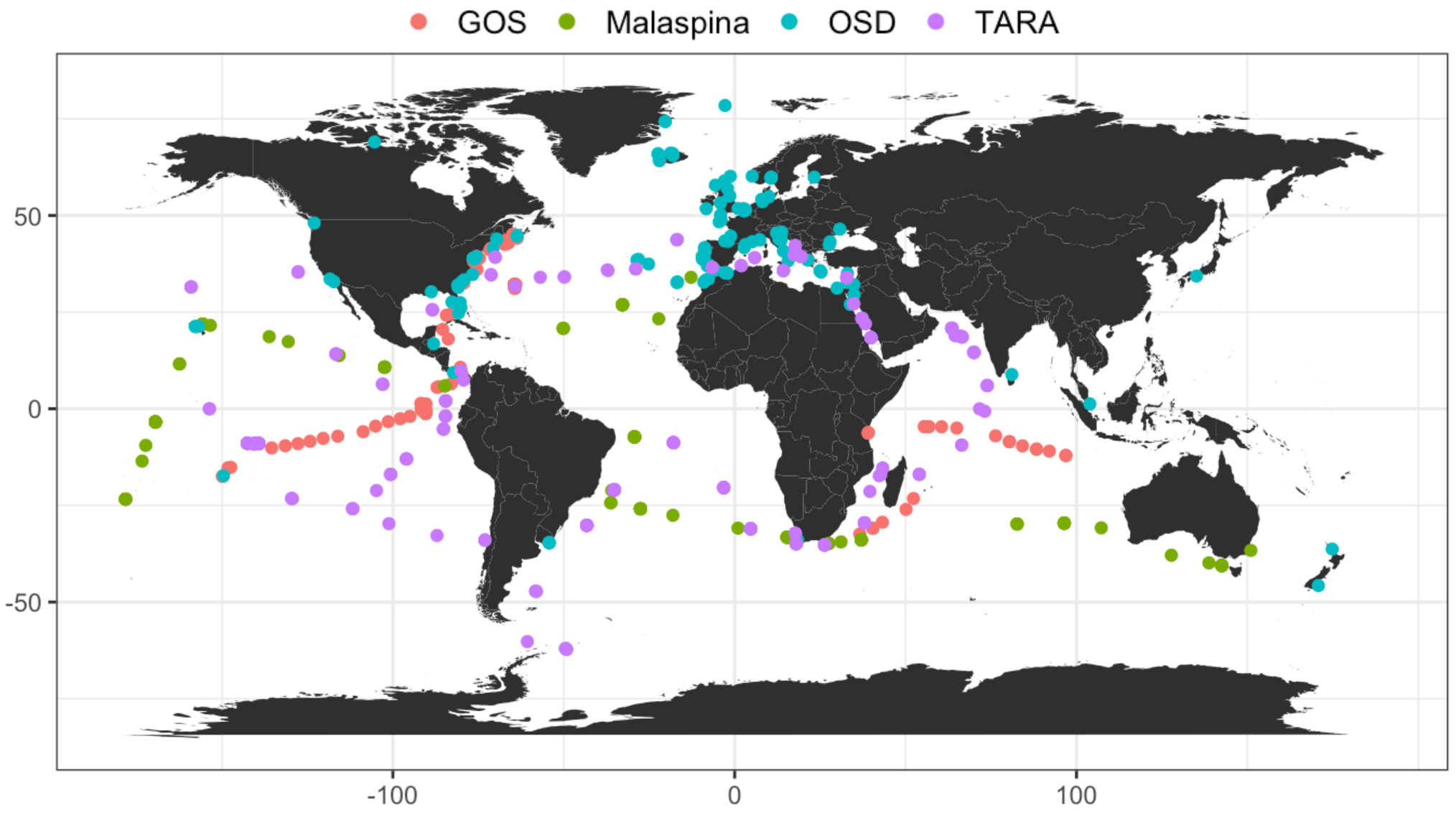
Global Ocean Sampling
Metagenomes: 80
ORFs: 20,068,580

Malaspina
Metagenomes: 116
ORFs: 20,574,033

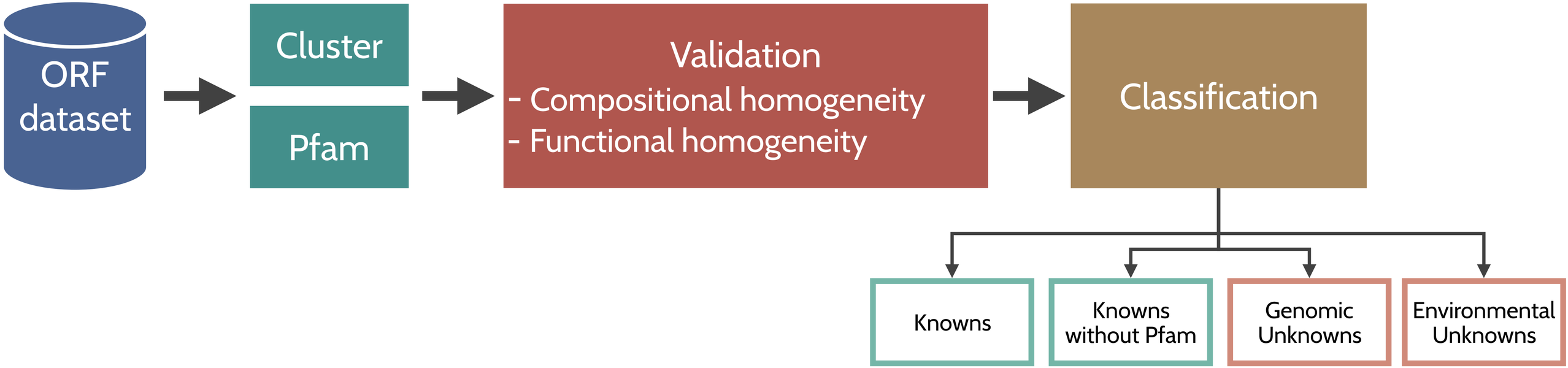
Total Metagenomes: **1,837**

Total ORFs: **322,248,552**

Good coverage of marine environment



How do we get our categories?



Vanni et al. 2018 (in prep)

Cluster aggregation into components

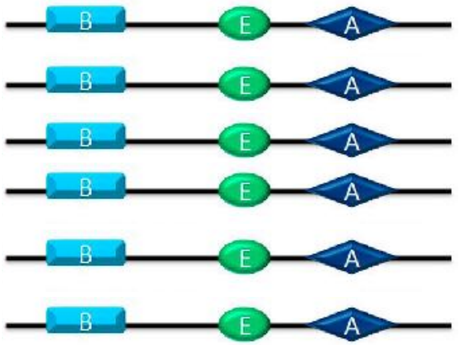
	Knowns	Genomic Unknowns	Environmental Unknowns	Knowns without Pfam	Total
ORFs	42,937,850	48,579,179	9,472,648	37,073,885	262,821,906
Clusters	914,932	894,280	388,624	756,067	2,953,903

Cluster aggregation into components

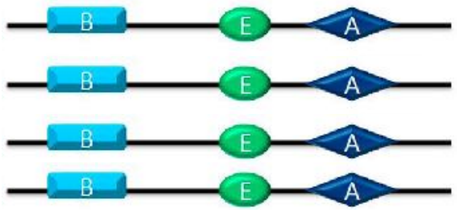
	Knowns	Genomic Unknowns	Environmental Unknowns	Knowns without Pfam	Total
ORFs	42,937,850	48,579,179	9,472,648	37,073,885	262,821,906
Clusters	914,932	894,280	388,624	756,067	2,953,903

Clustering

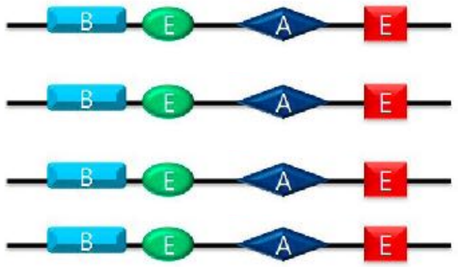
Cluster_26



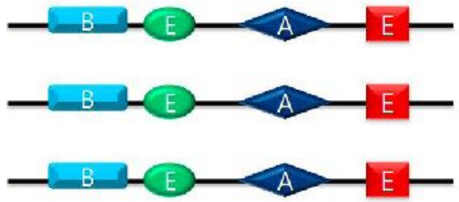
Cluster_783



Cluster_92



Cluster_3



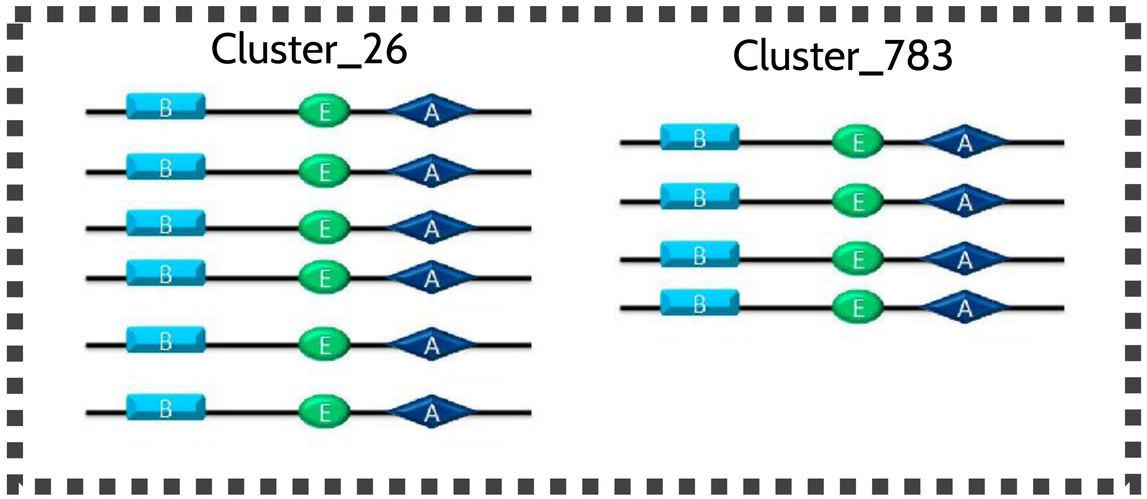
modified from Byungwook et al. 2009

Cluster aggregation into components

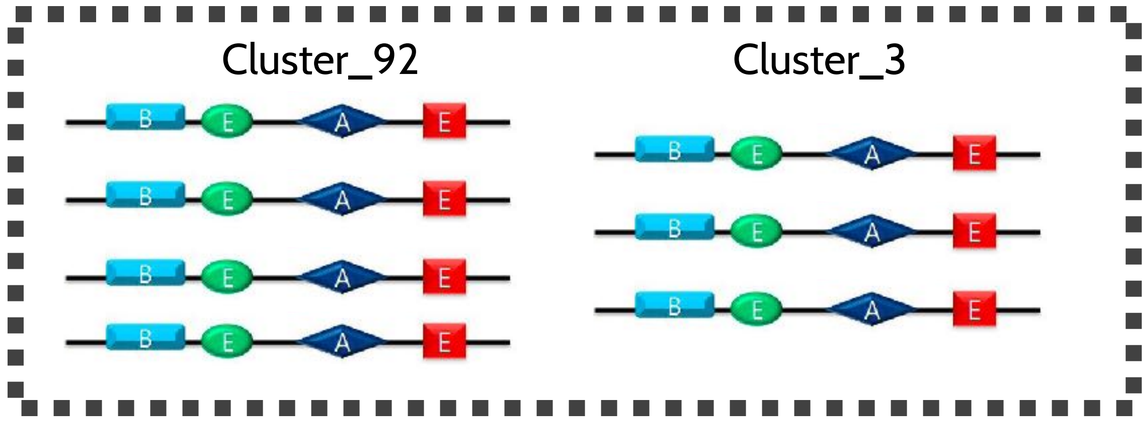
	Knowns	Genomic Unknowns	Environmental Unknowns	Knowns without Pfam	Total
ORFs	42,937,850	48,579,179	9,472,648	37,073,885	262,821,906
Clusters	914,932	894,280	388,624	756,067	2,953,903

Clustering

Component_17



Component_752



modified from Byungwook et al. 2009

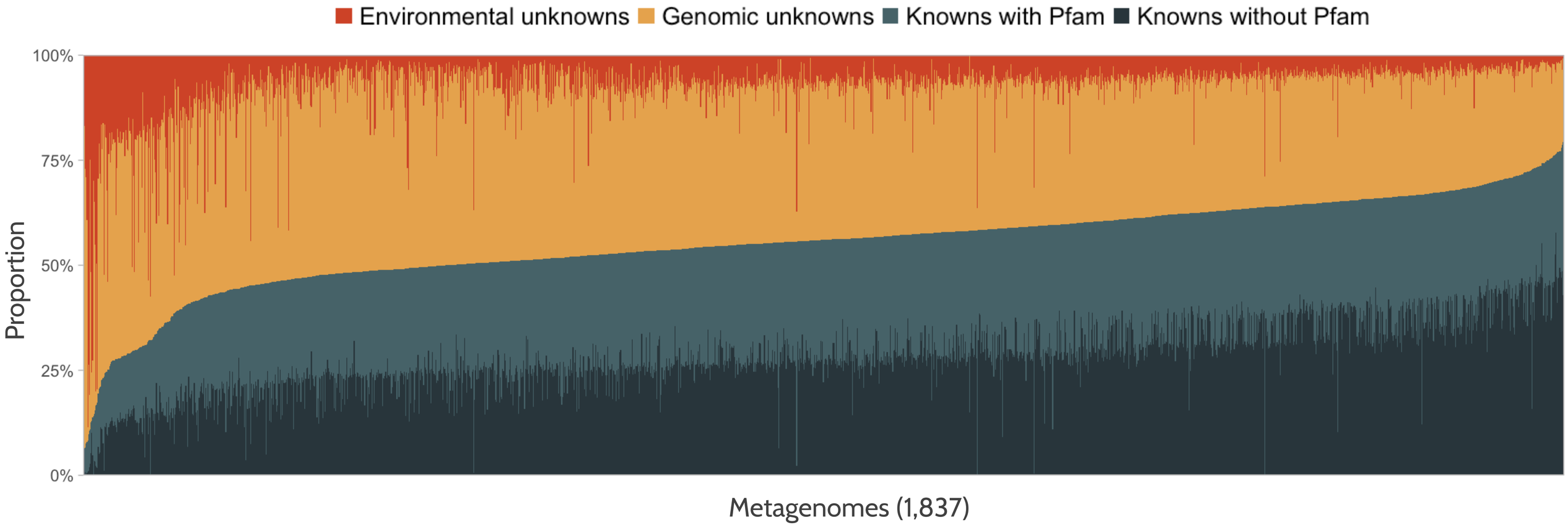
Output of Vanni et al. workflow

	Knowns	Genomic Unknowns	Environmental Unknowns	Knowns without Pfam	Total
ORFs	42,937,850	48,579,179	9,472,648	37,073,885	262,821,906
Clusters	914,932	894,280	388,624	756,067	2,953,903
Components	18,368	598,771	181,585	519,224	1,317,948

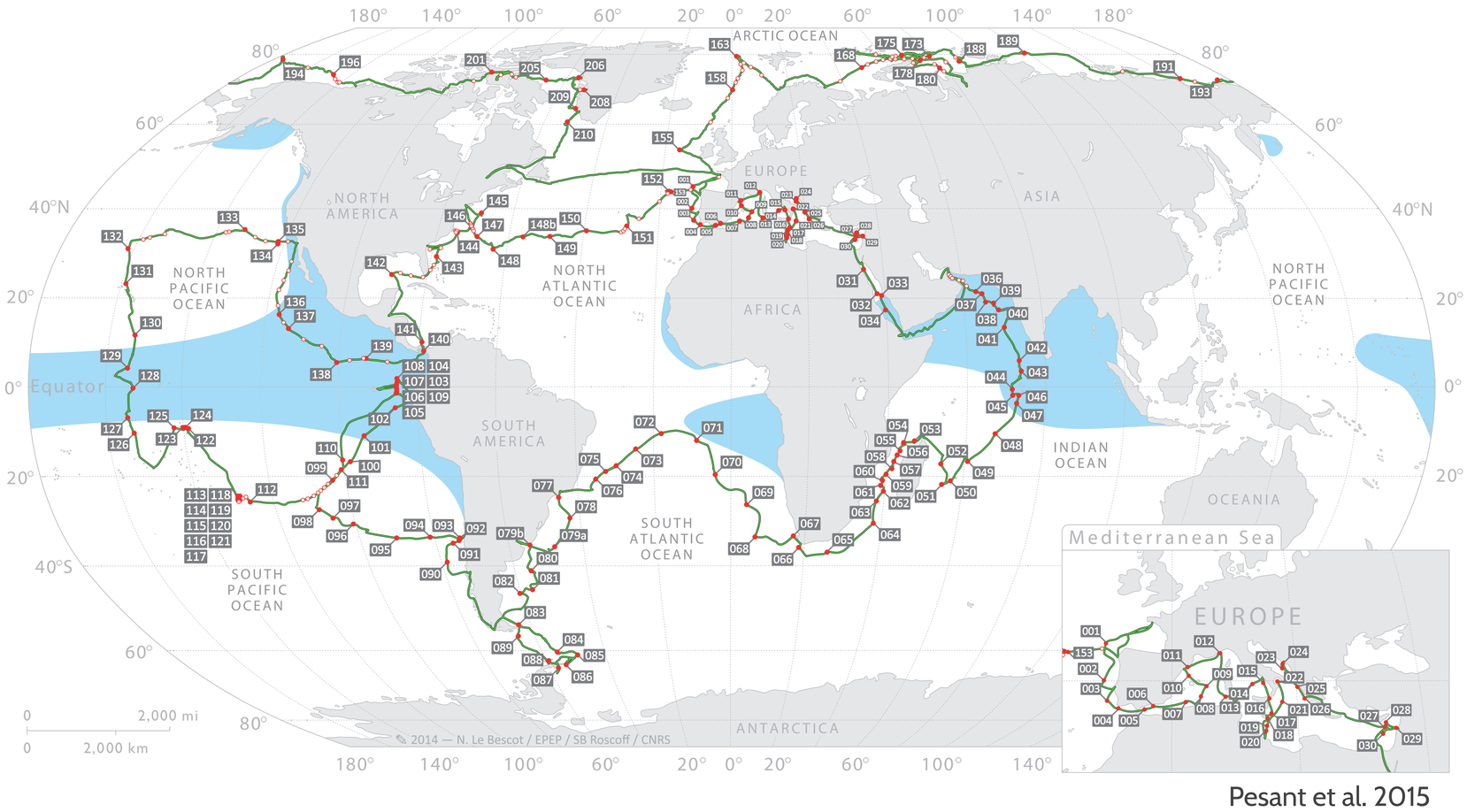
Clustering

Aggregation

How much is actually known?



Exploring unknowns in the TARA Oceans dataset



Pesant et al. 2015

Ocean microbial gene catalogue (OM-RGC)

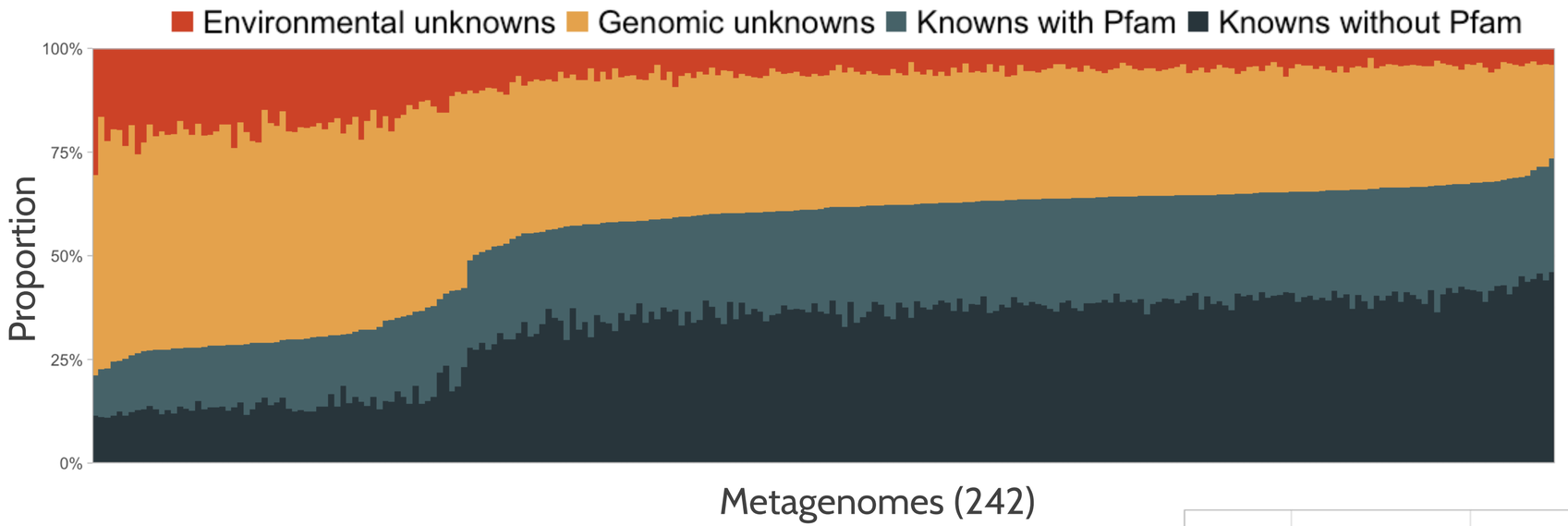
- 40% of ocean core genes were found to have unknown functions!

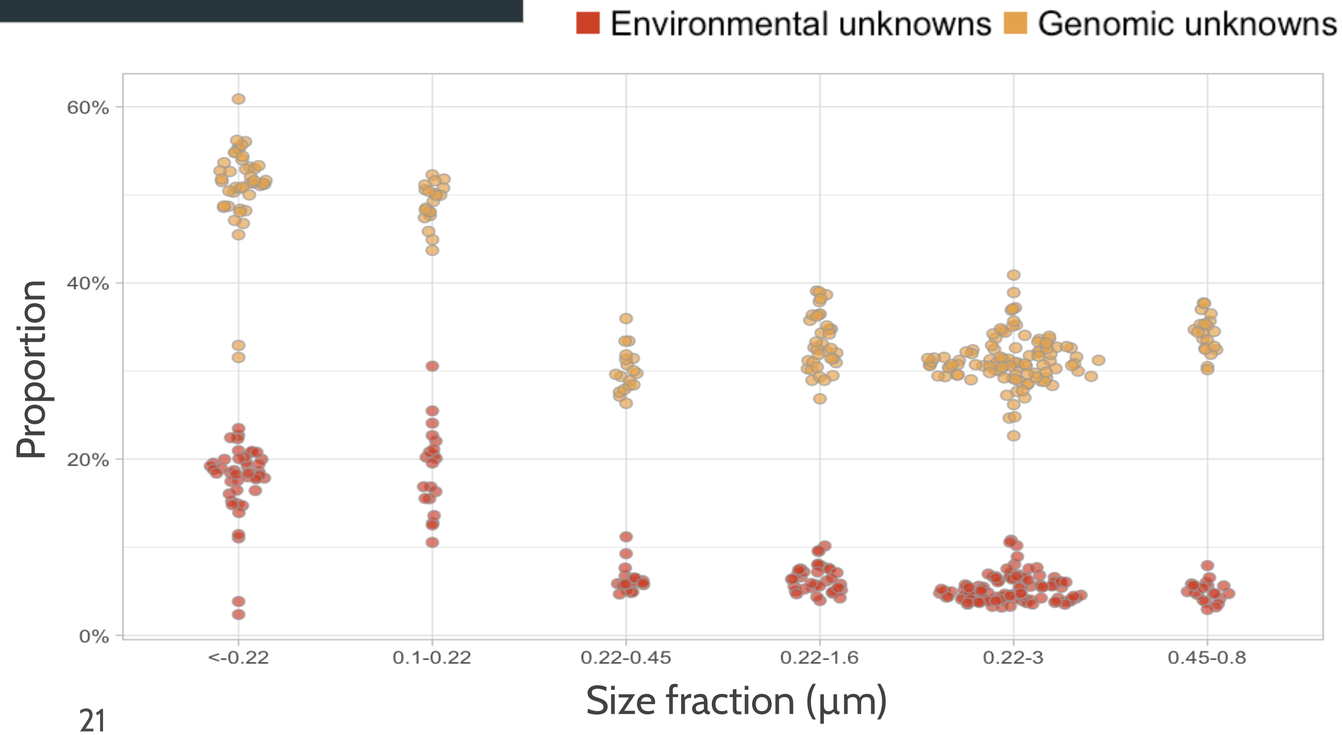
OCEAN PLANKTON

Structure and function of the global ocean microbiome

Sunagawa et al. 2015

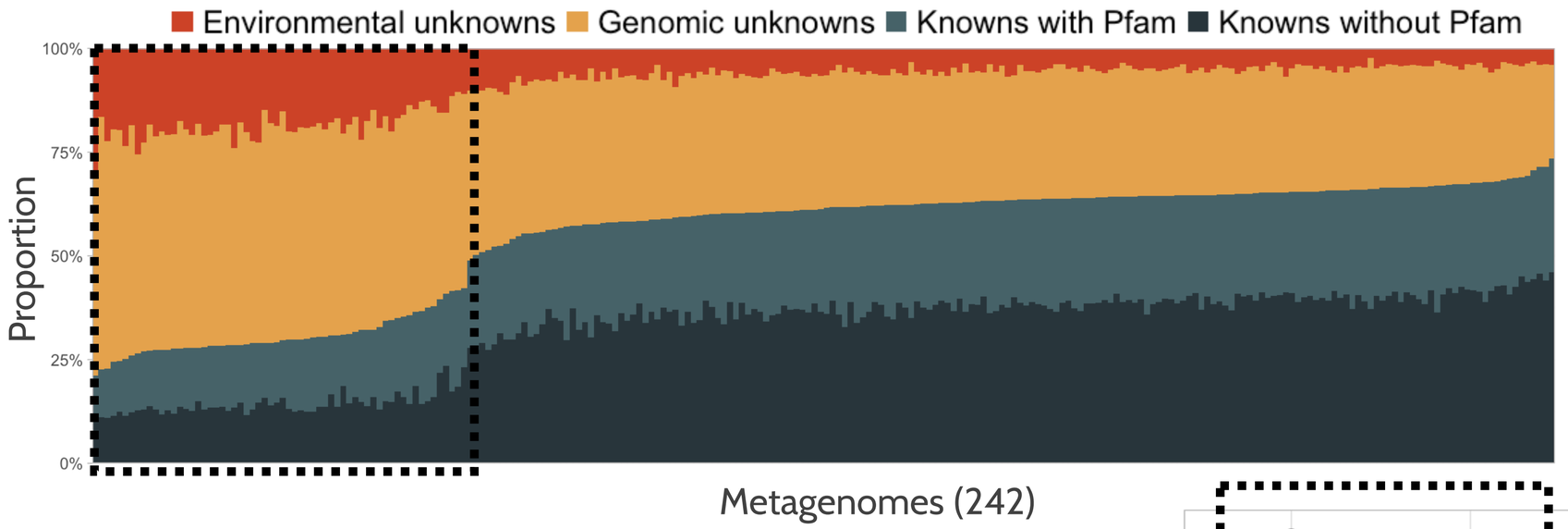
Exploring unknowns in the TARA Oceans dataset

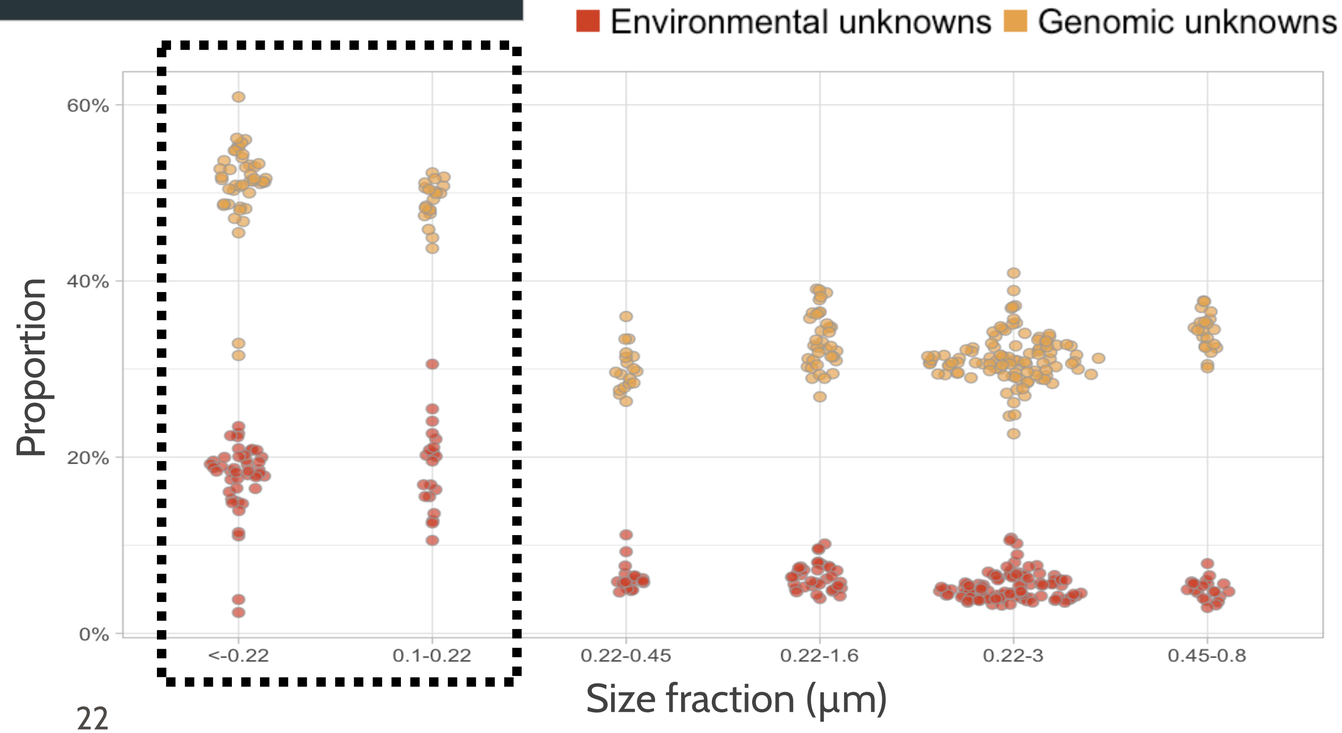




Vanni et al. 2018 (in prep)

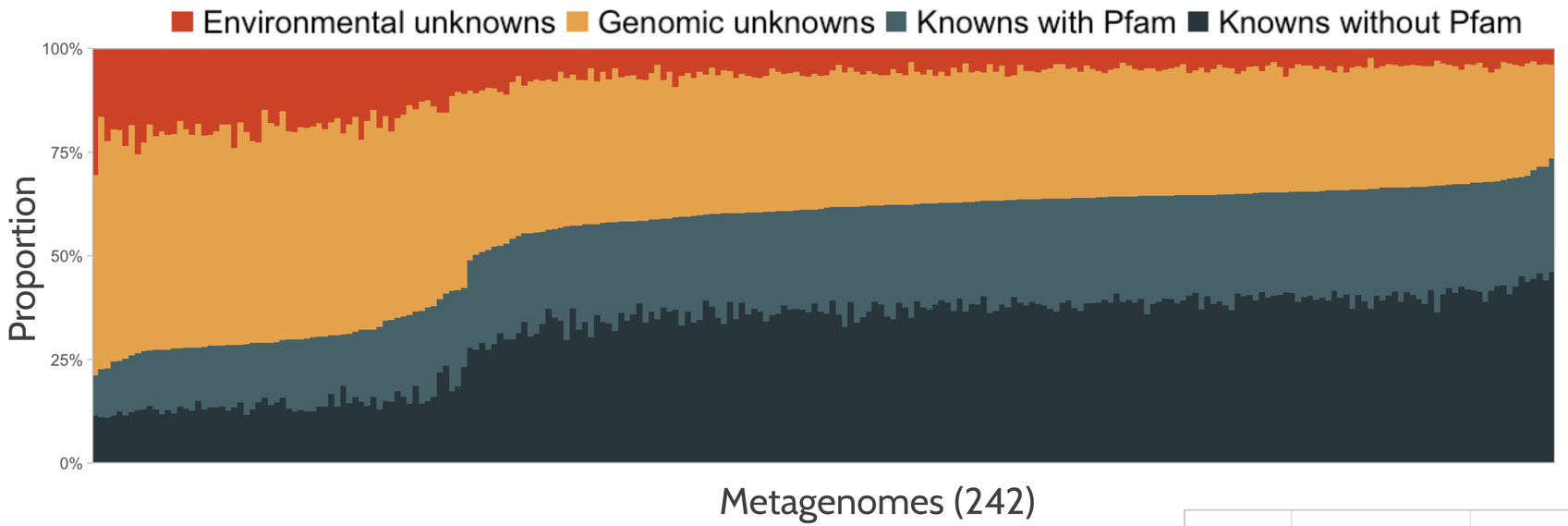
Exploring unknowns in the TARA Oceans dataset

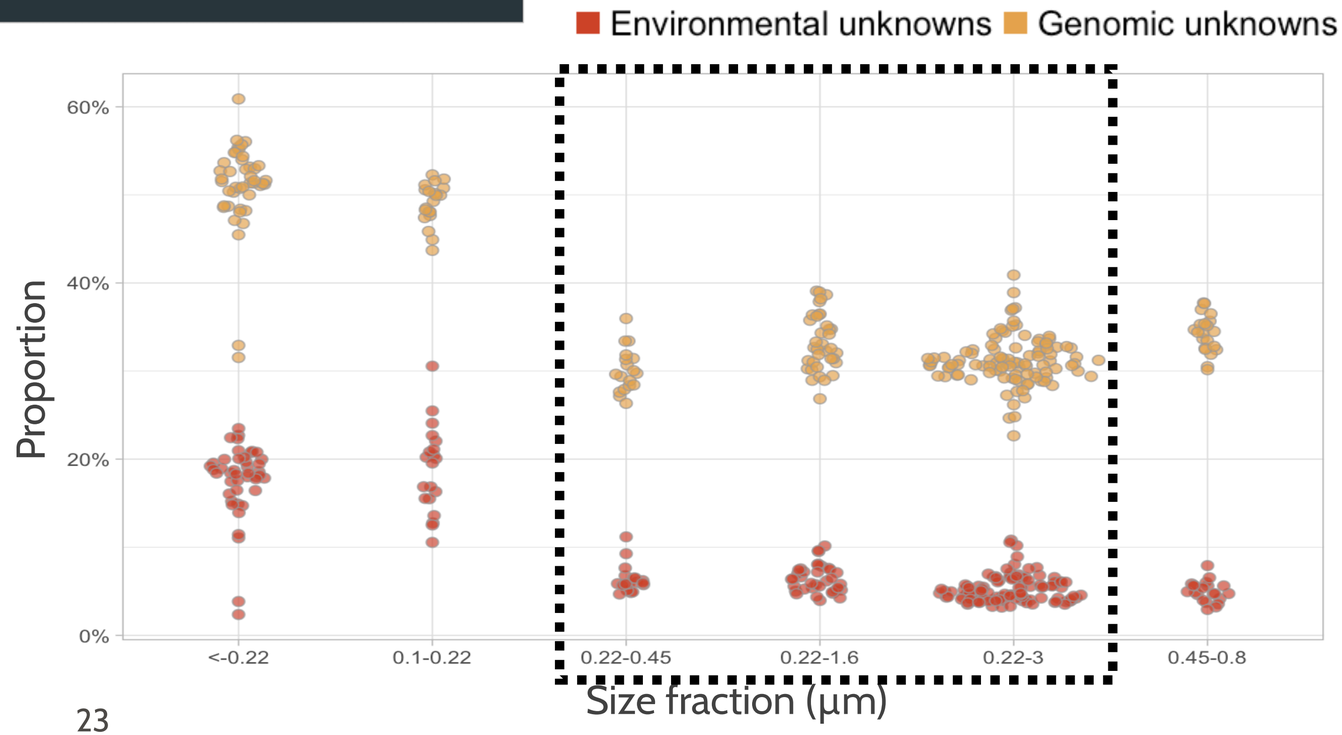




Vanni et al. 2018 (in prep)

Exploring unknowns in the TARA Oceans dataset





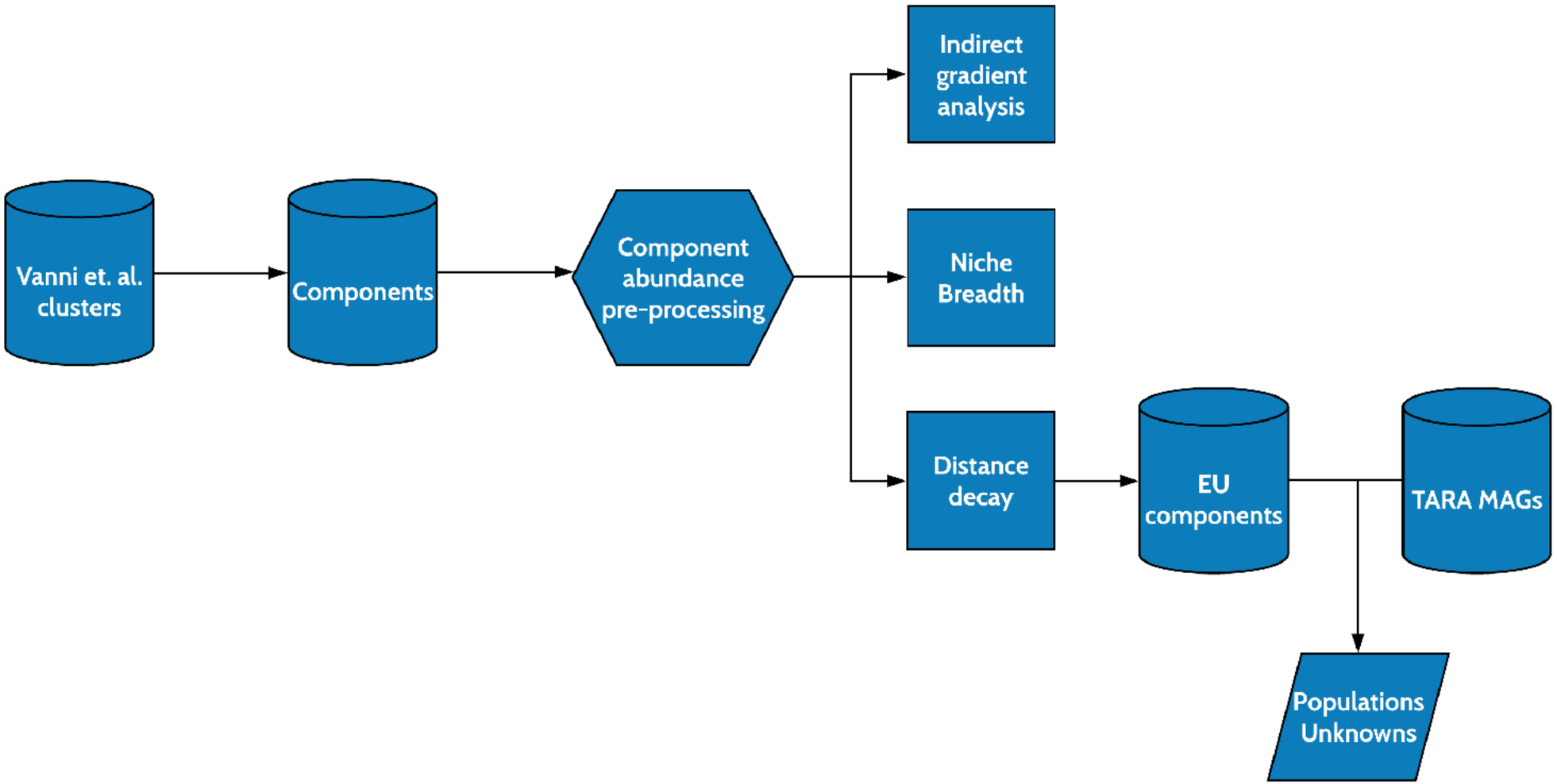
Vanni et al. 2018 (in prep)

TARA Oceans prokaryotic component subset

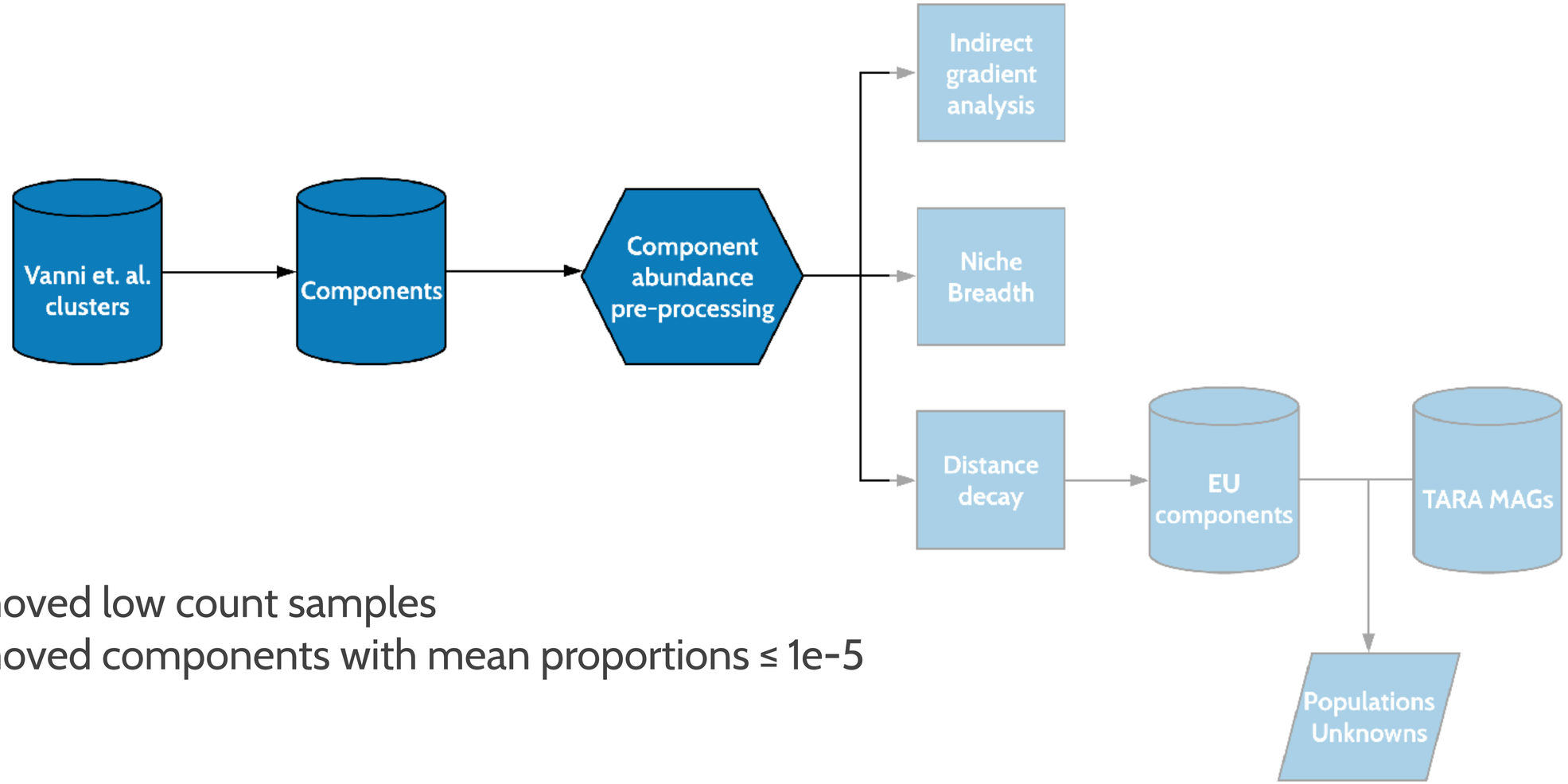
	Knowns	Genomic Unknowns	Environmental Unknowns	Knowns without Pfam	Total
ORFs	42,937,850	48,579,179	9,472,648	37,073,885	262,821,906
Clusters	914,932	894,280	388,624	756,067	2,953,903
Components	18,368	598,771	181,585	519,224	1,317,948
TARA	11,760	269,673	183,205	176,142	**640,780**

Can we utilize the unknowns to learn more about microbial ecology in the world's oceans?

Research workflow

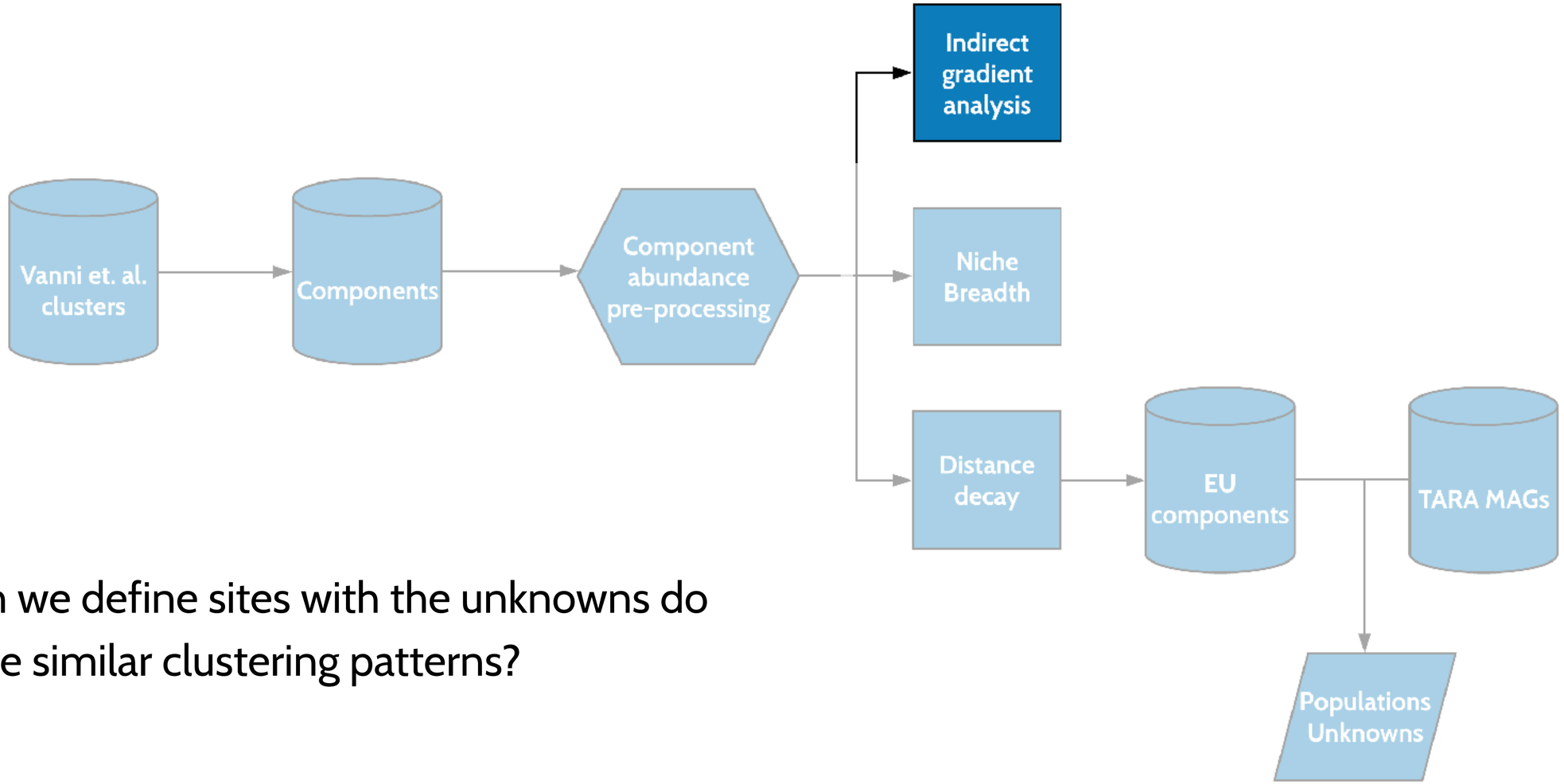


Research workflow - preprocessing data



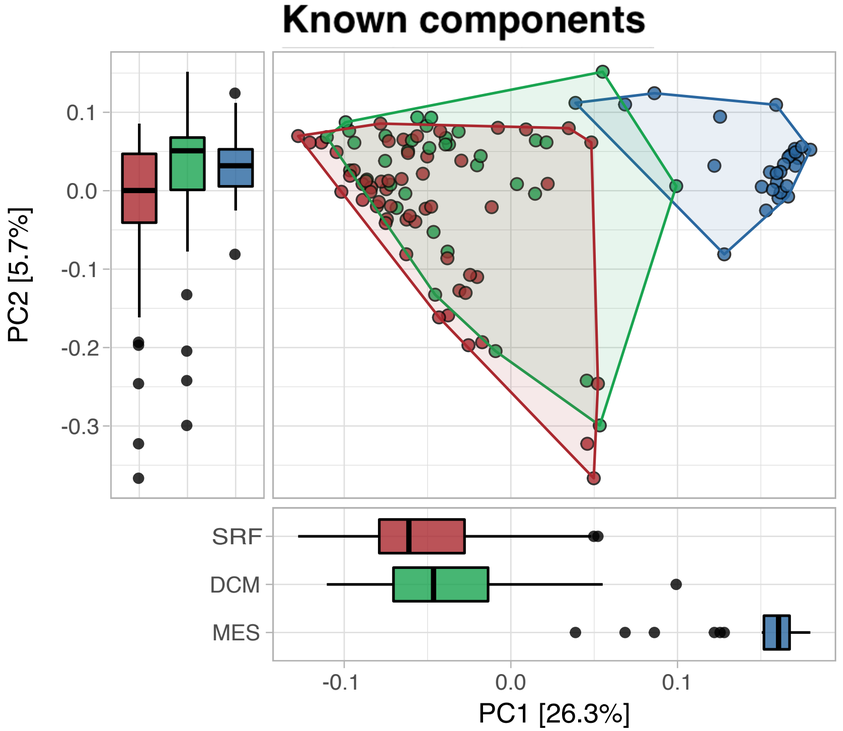
-Removed low count samples
-Removed components with mean proportions ≤ 1e-5

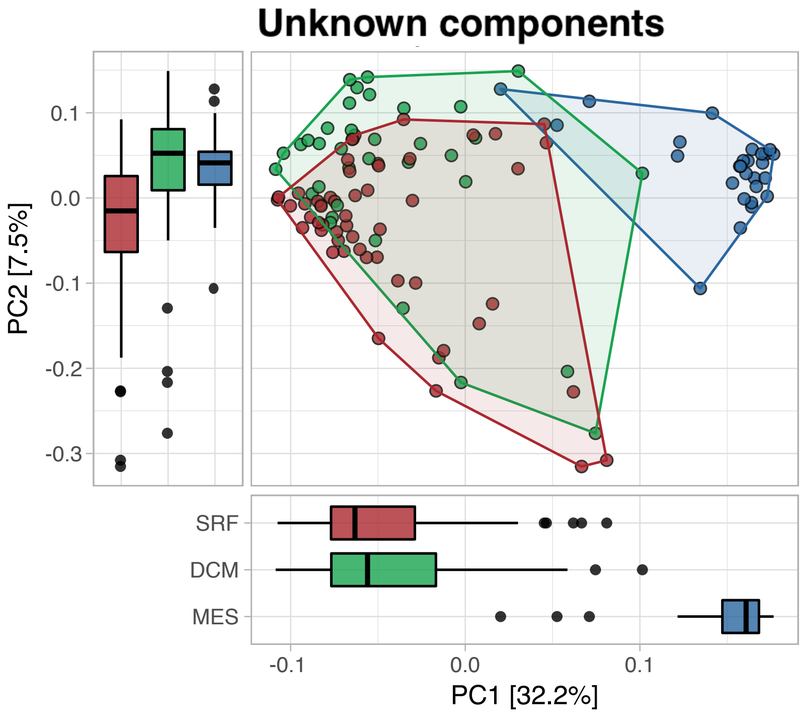
Research workflow - sample site ordinations

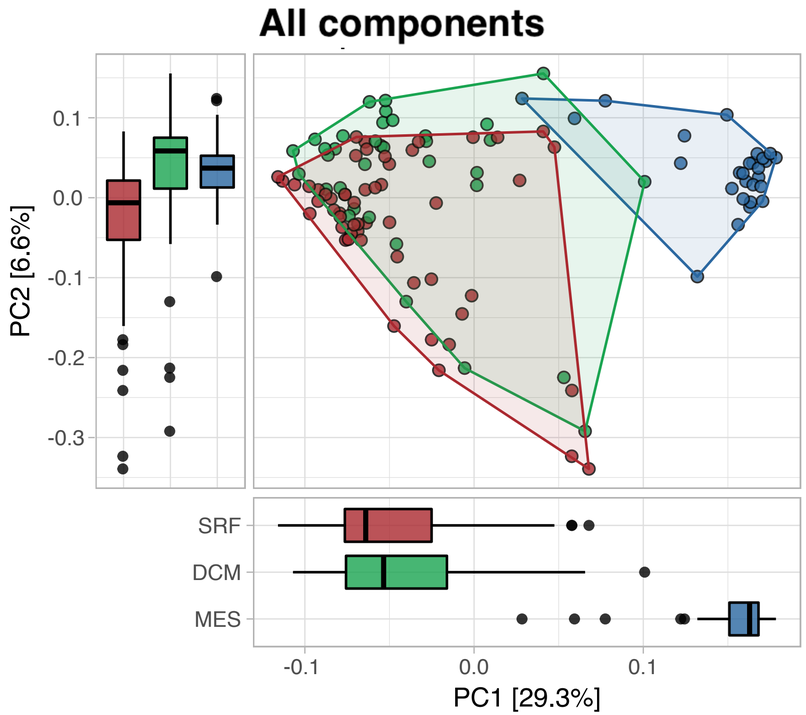


When we define sites with the unknowns do we see similar clustering patterns?

TARA Oceans sample site ordination - depth layers



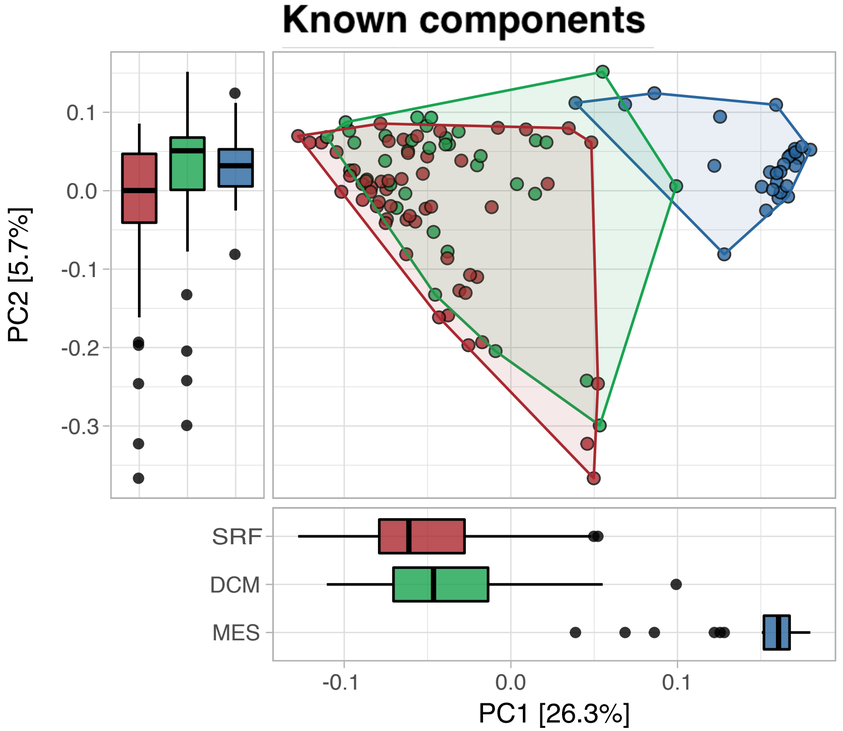


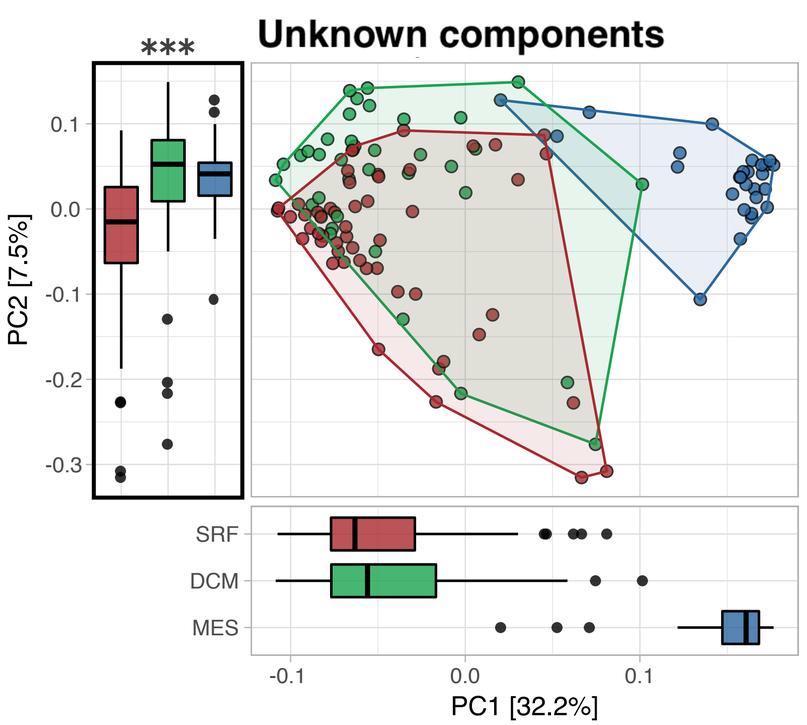


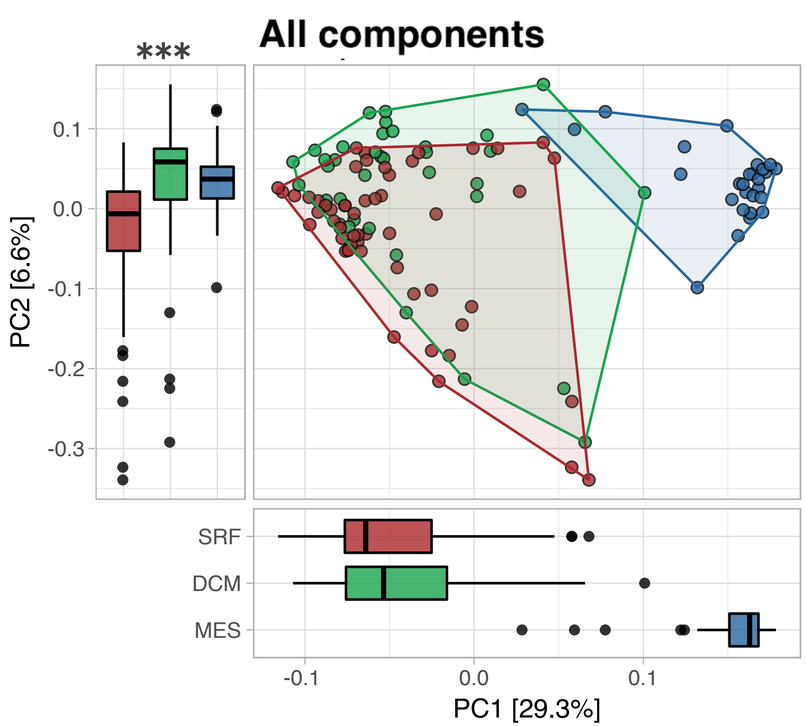
- Sampling site clustering by depth with each component category
- Increased variance when unknowns are included
- Principal component axis (PC)

SRF: surface **DCM**: deep chlorophyll maximum **MES**: mesopelagic

TARA Oceans sample site ordination - depth layers



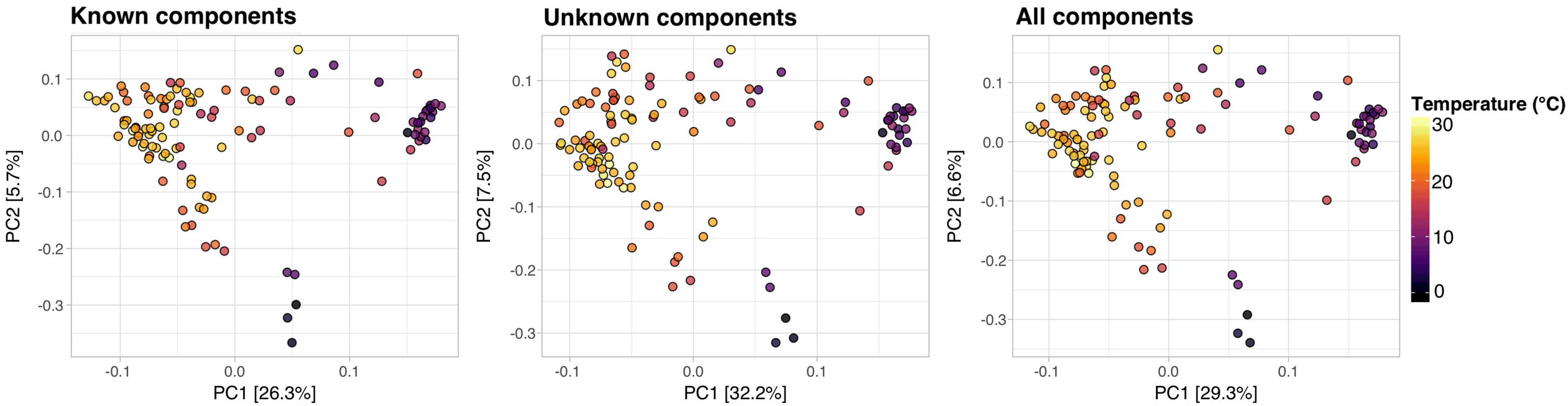




- Sampling site clustering by depth with each component category
- Increased variance when unknowns are included
- Principal component axis (PC)

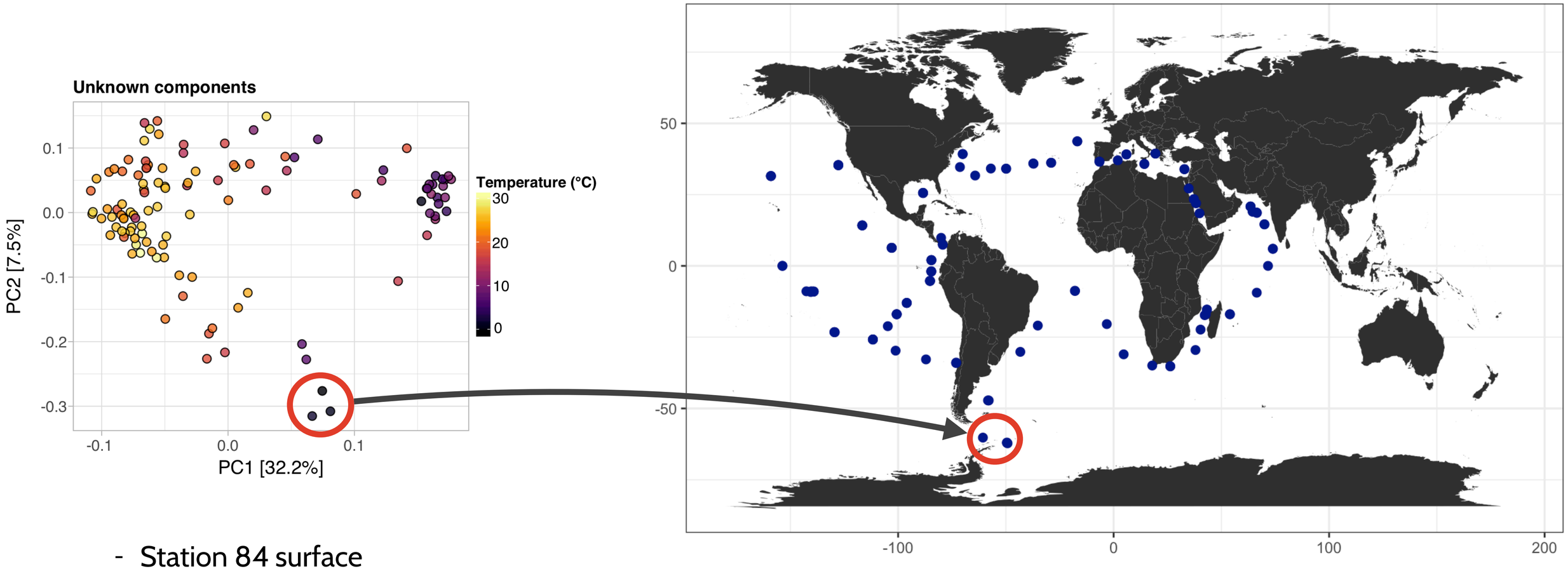
SRF: surface **DCM**: deep chlorophyll maximum **MES**: mesopelagic

TARA Oceans sample site ordination - temperature



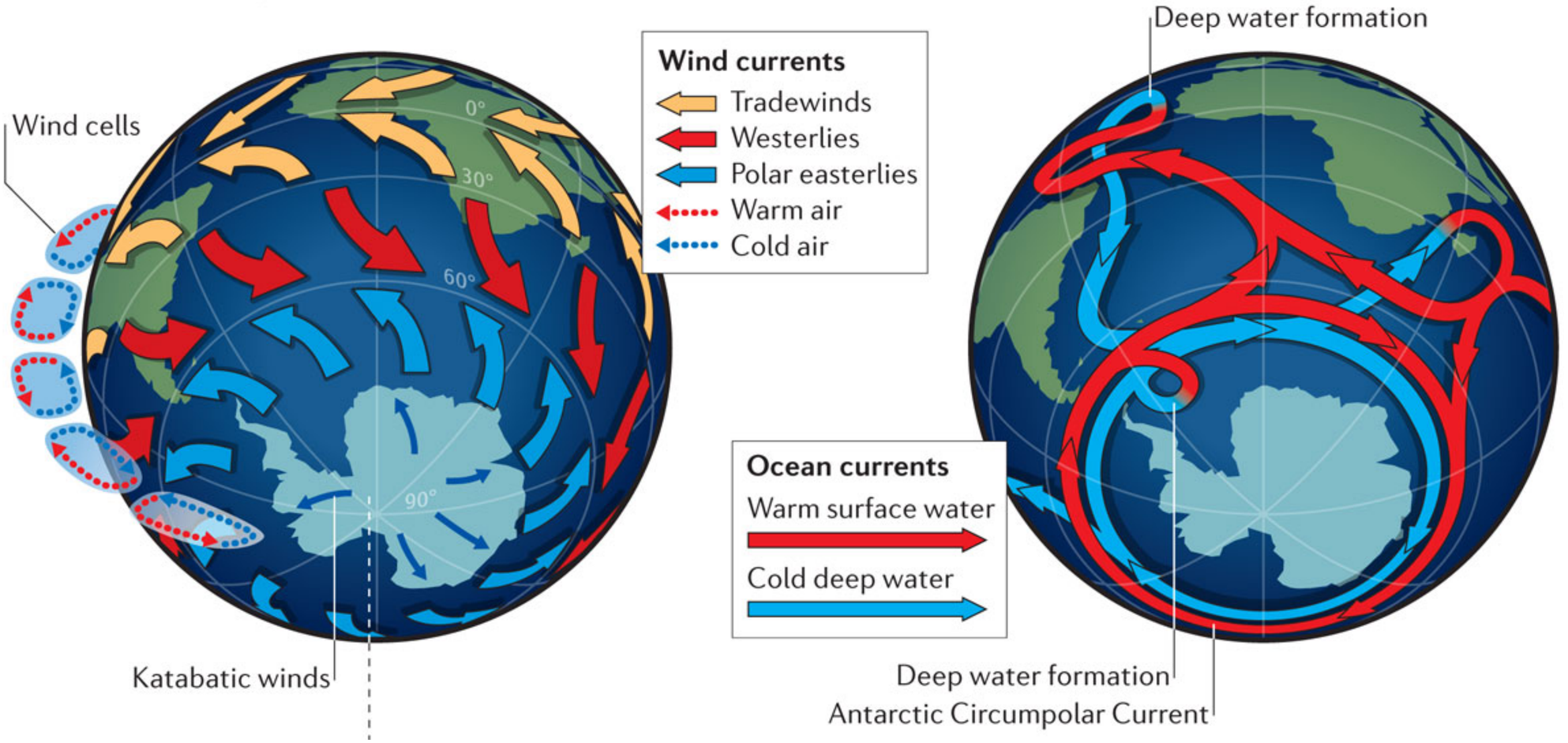
- Temperature gradient across principal component 1

Indirect gradient analysis - temperature



- Station 84 surface
- Station 85 surface and DCM

Antarctic circumpolar current

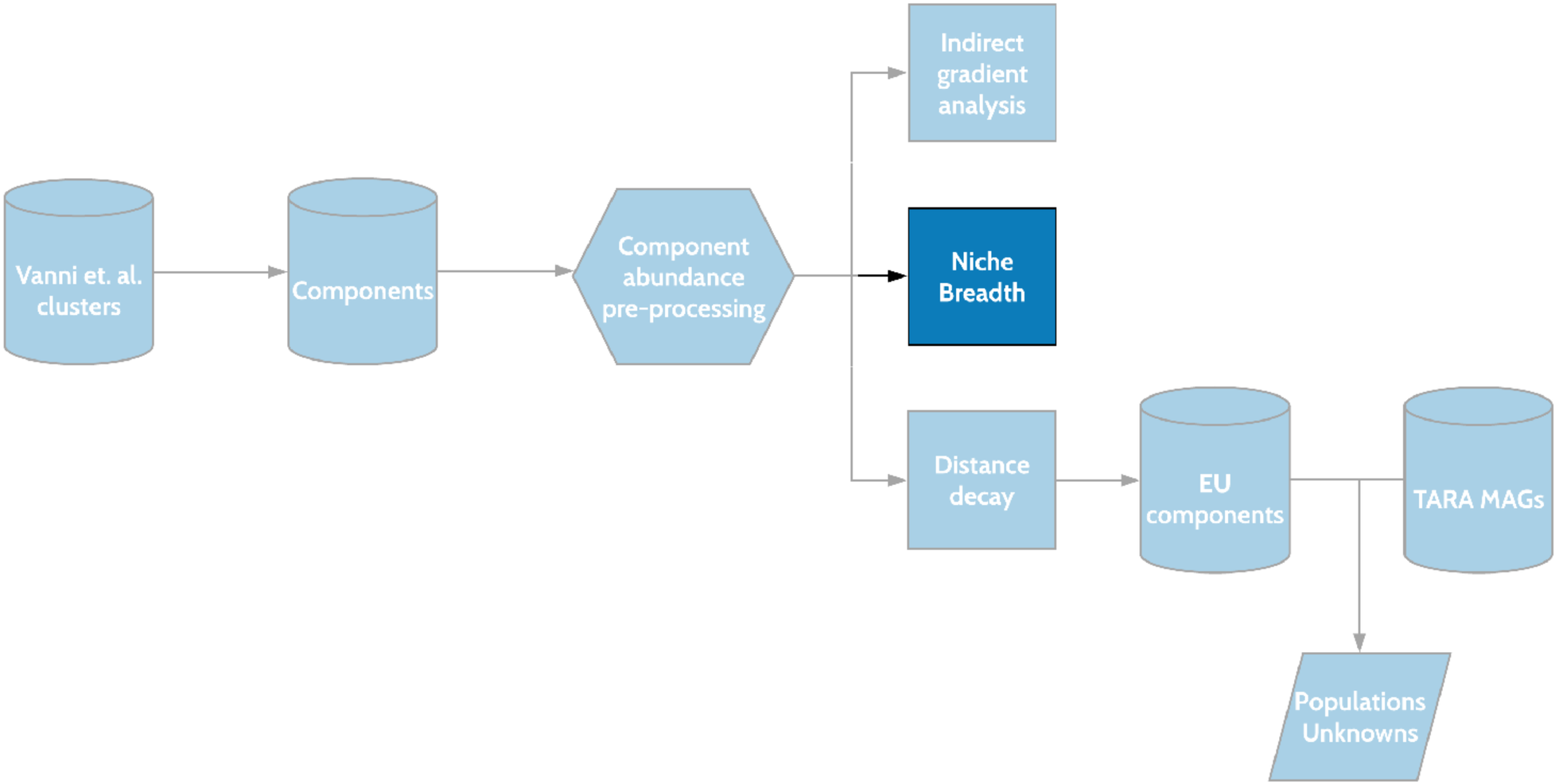


- Strong, isolated current
- High nutrient, low chlorophyll waters

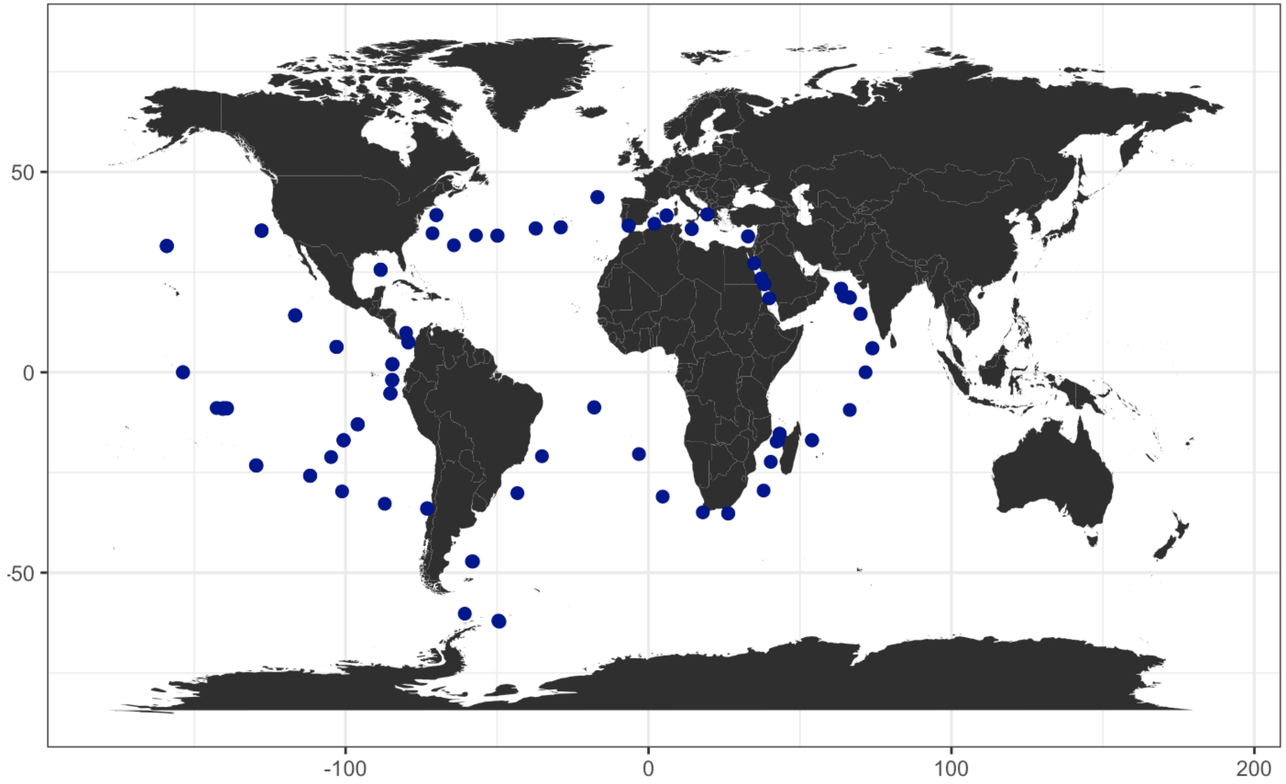
Modified from Cavicchioli, 2015

Can we identify distribution patterns in the unknowns?

Research workflow - component spatial distribution



Levins niche breadth analysis

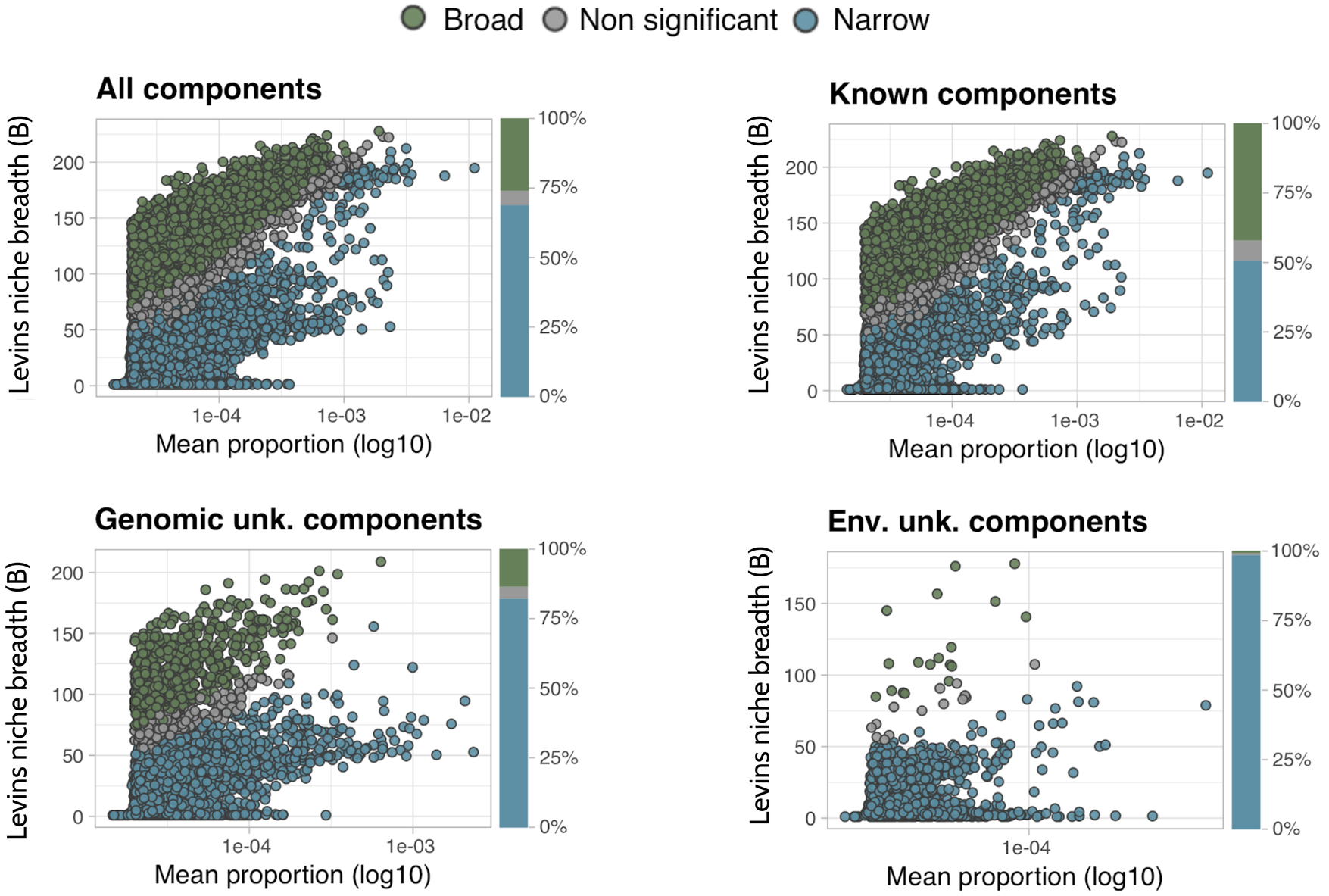


What is the distribution of the components within the different categories?

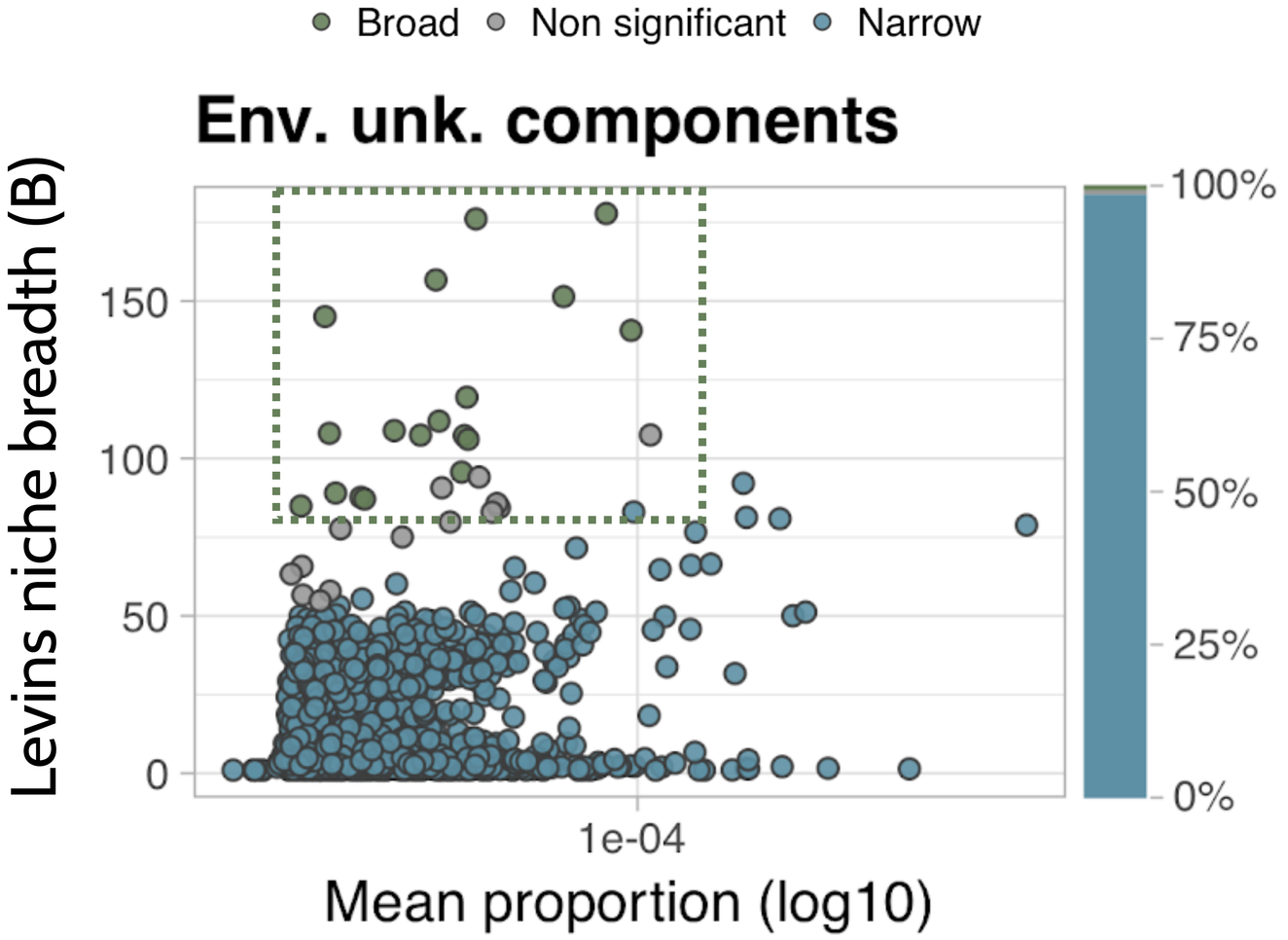
$$B = \sum_{1}^{n} 1/P^2ij$$

Using Levins niche breadth we can assign components as having a **broad** or **narrow** distribution.

Levins niche breadth

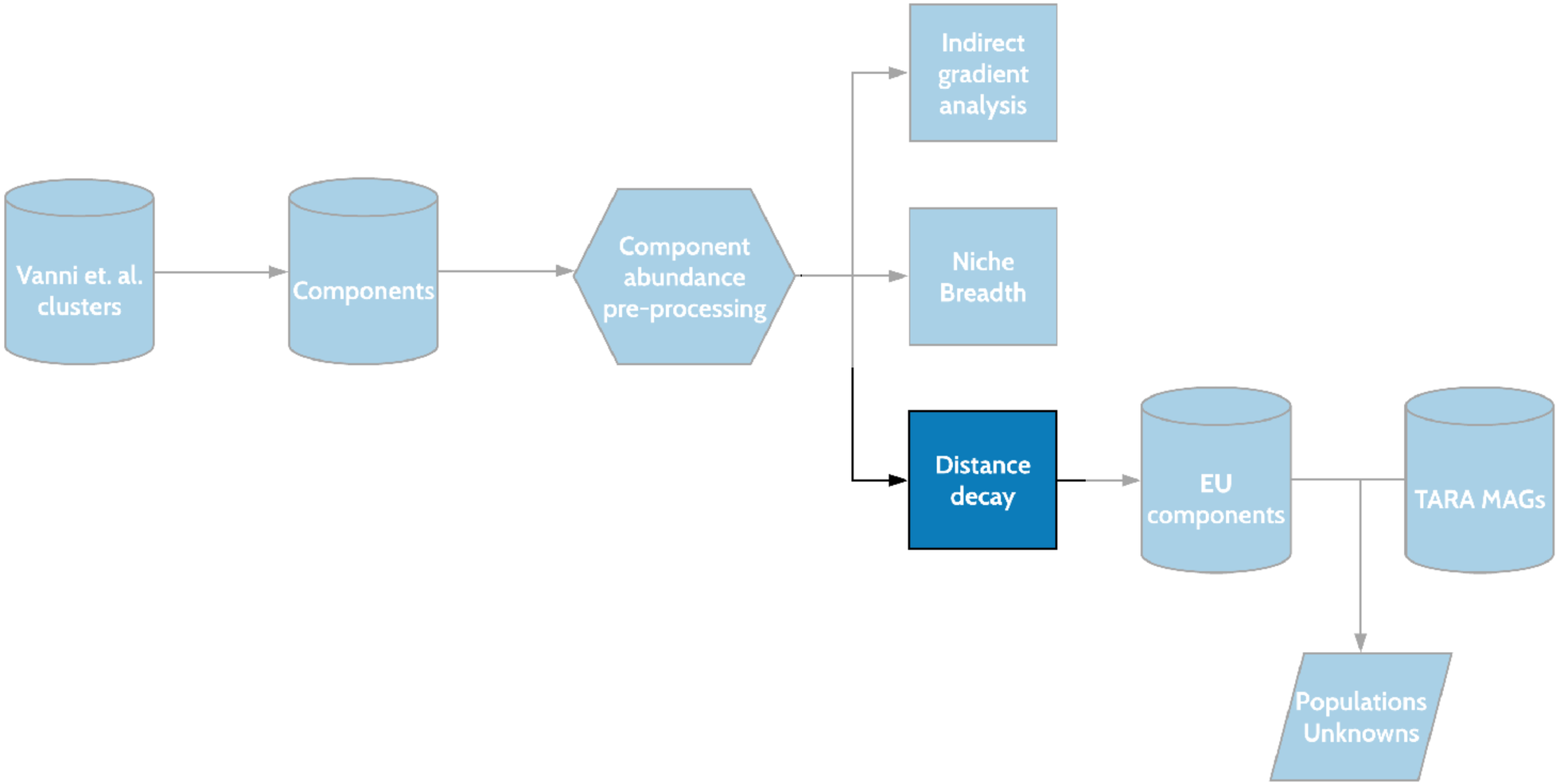


Levins niche breadth

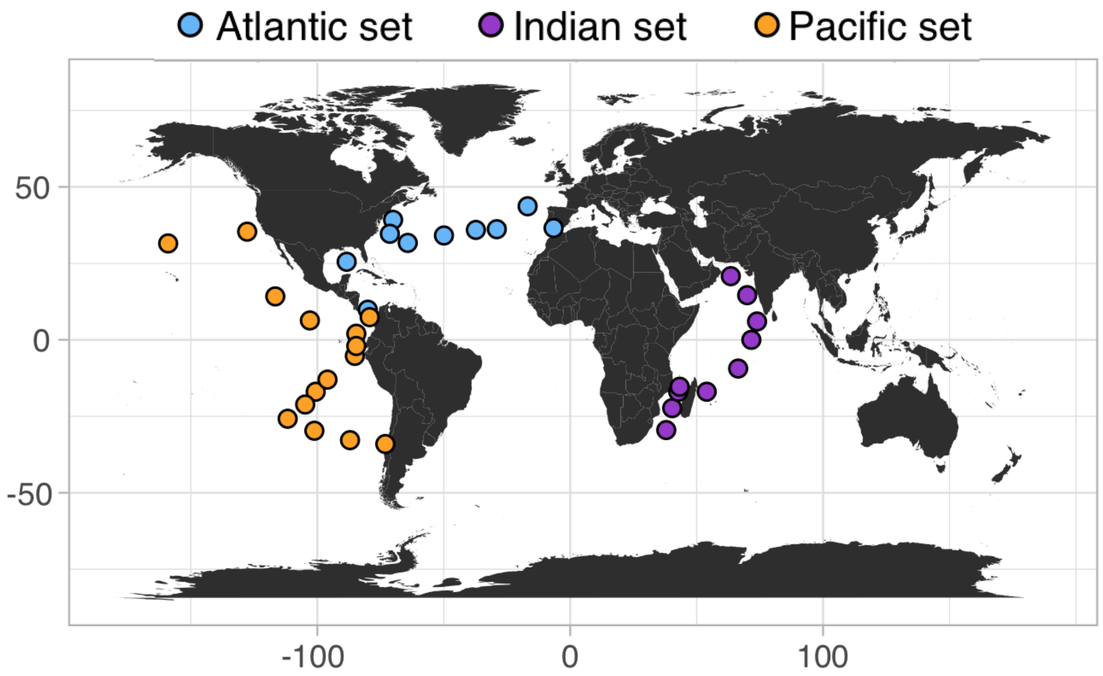


Is there a relationship between geographical distance and functional dissimilarity in the unknowns?

Research workflow - biogeography

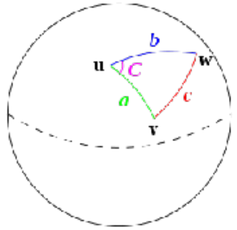


Functional dissim. vs. geographic distance

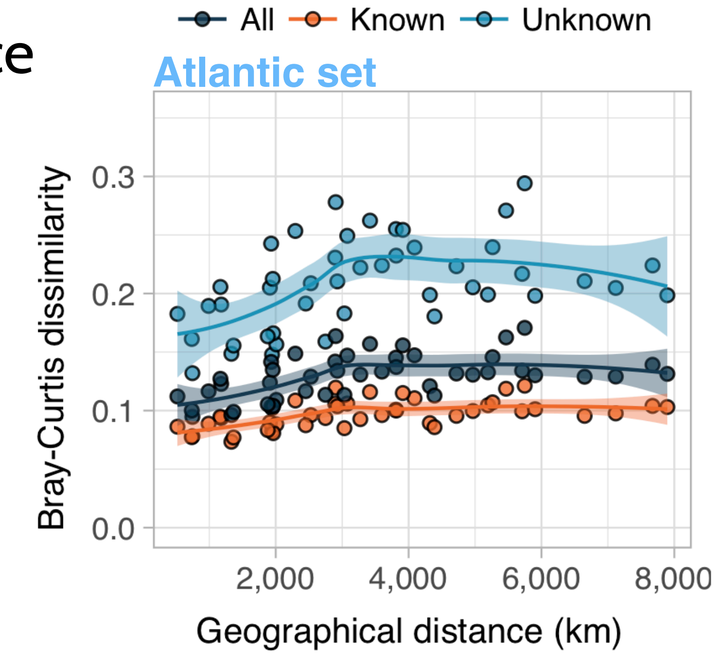


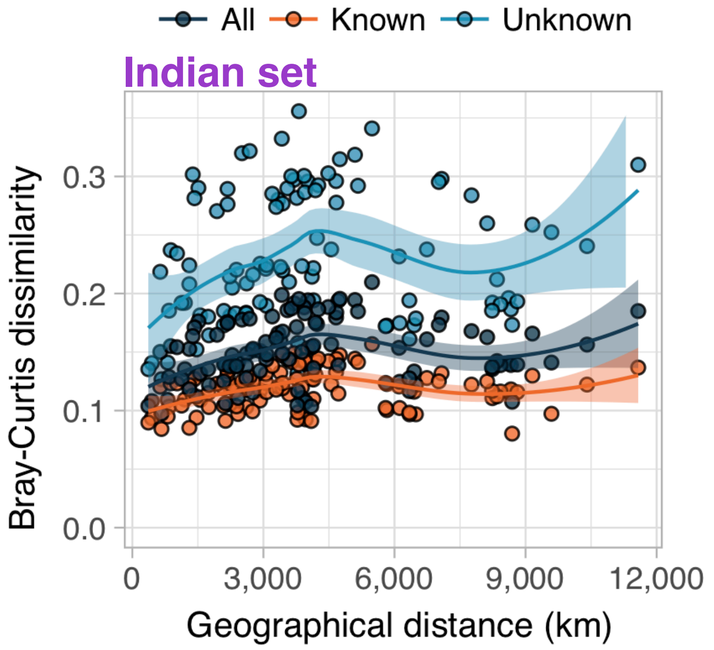
Delmont et al. 2017

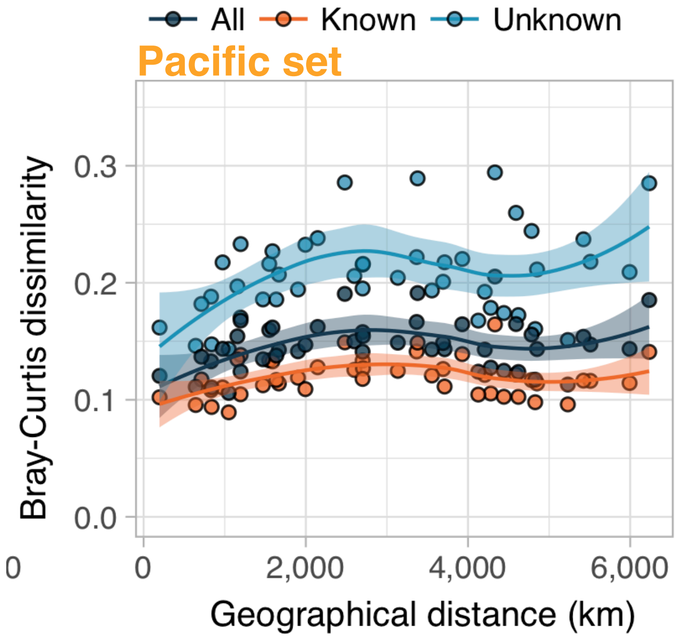
Haversine distance



Haversine distance https://en.wikipedia.org/wiki/Haversine_formula

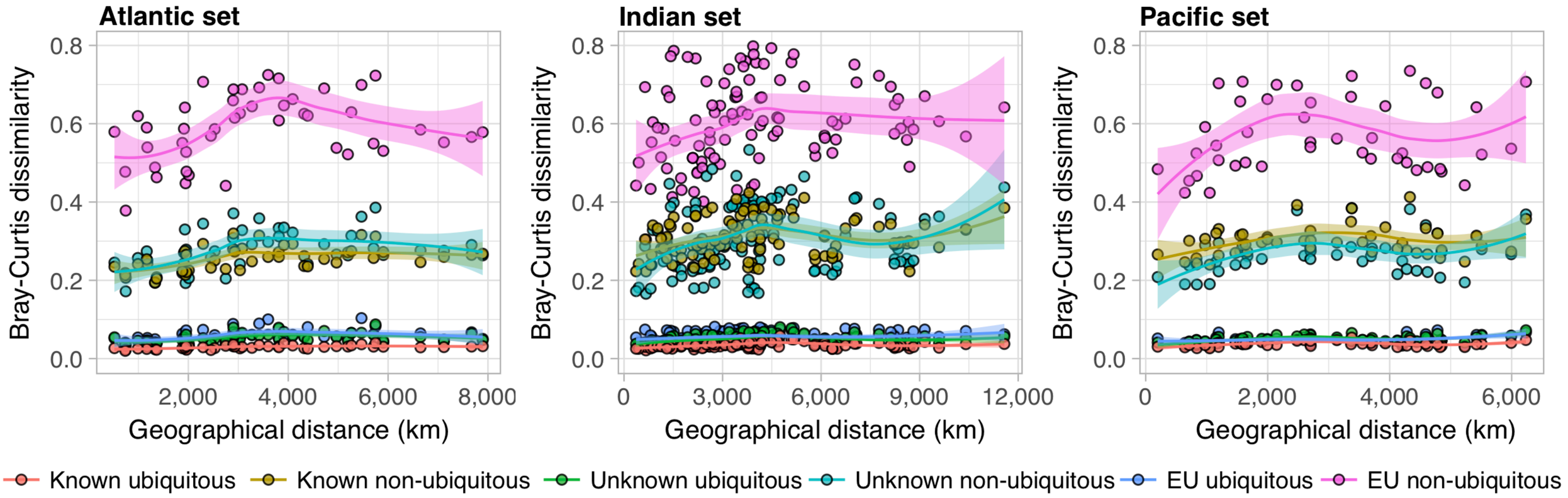






Mantel test: not sig.

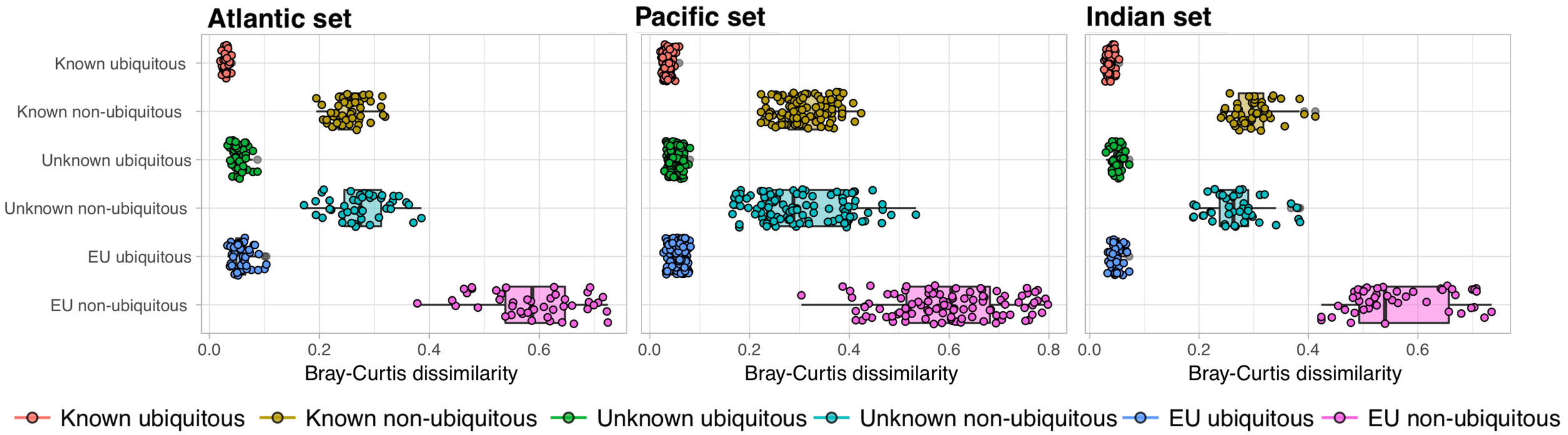
Ubiquitous vs. non-ubiquitous components



- Ubiquitous: components found in every TARA sample

Ubiquitous components

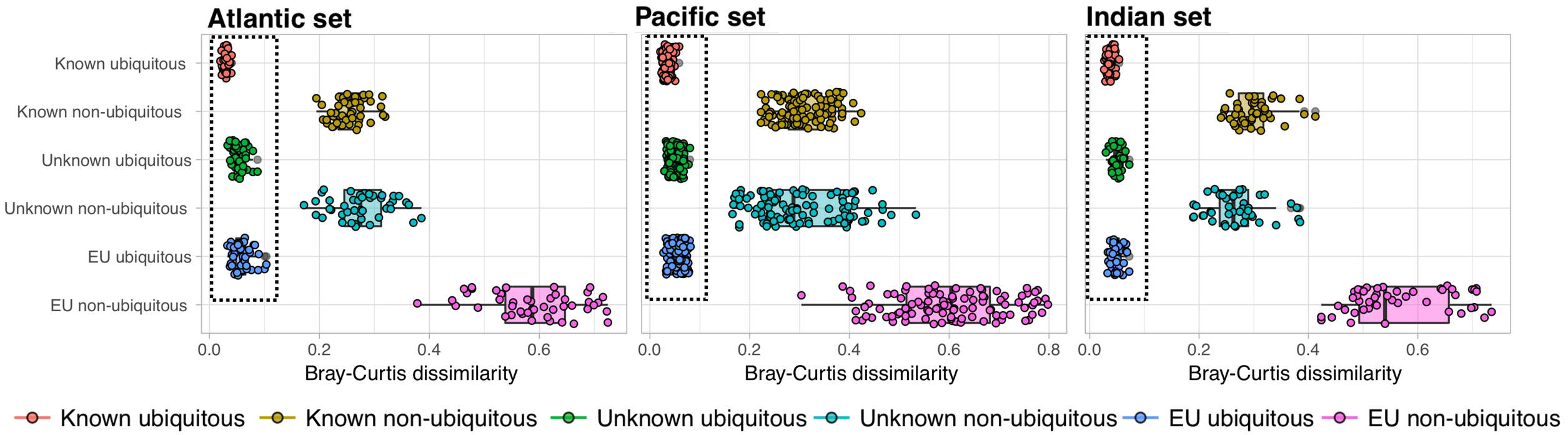
All ubiquitous components follow the same trend



Ubiquitous EUs **MAY BE** a new subset of microbial essential genes similar to “housekeeping” genes

Ubiquitous components

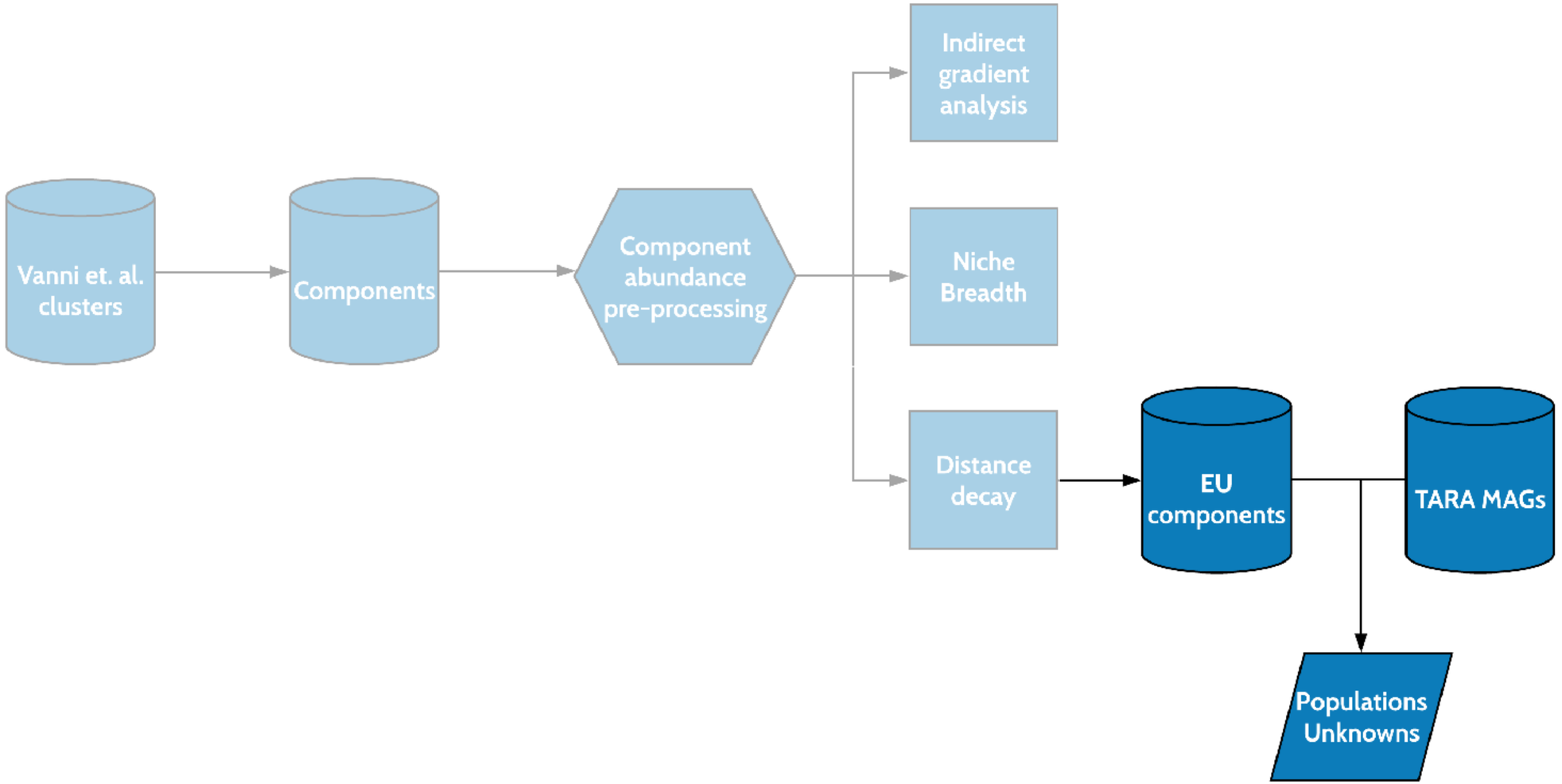
All ubiquitous components follow the same trend



Ubiquitous EUs **MAY BE** a new subset of microbial essential genes similar to “housekeeping” genes

Can we use metagenomic assembled genomes (MAGs) to give context to the ubiquitous EUs?

Research workflow - zooming in on contigs



Functional Categories

Knowns

Pfam domains of known function, present in sequenced genomes

Genomic Unknowns

Unknown function, present in sequenced genomes

Environmental Unknowns

Unknown function, NO distant homology detected to databases

Population Unknowns

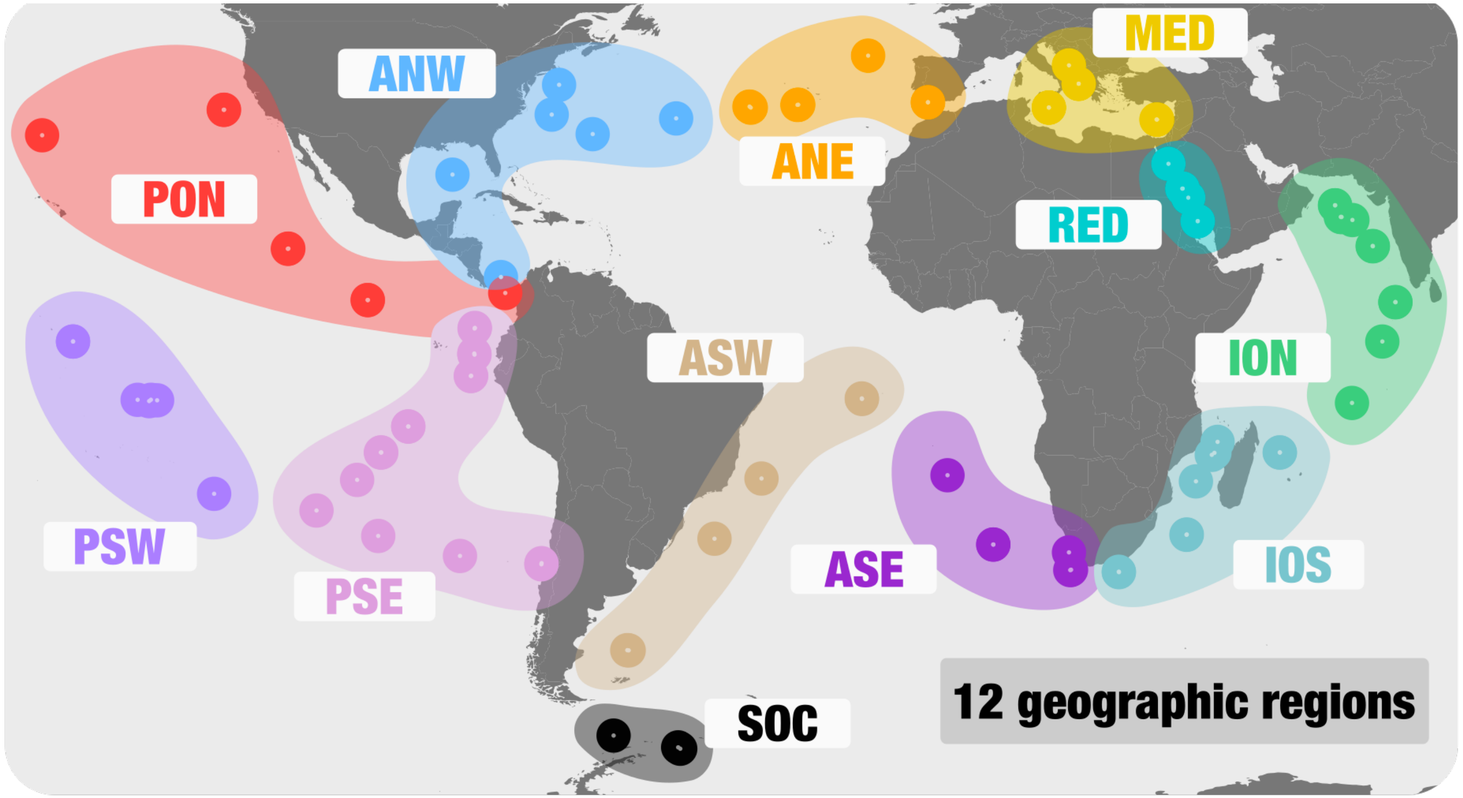
Unknown function, present in metagenomic assembled genomes

TARA Ocean MAGs

New Results

Nitrogen-Fixing Populations Of Planctomycetes And Proteobacteria Are Abundant In The Surface Ocean

Tom O. Delmont, Christopher Quince, Alon Shaiber, Ozcan C. Esen, Sonny T. M. Lee, Sebastian Lucker, A. Murat Eren

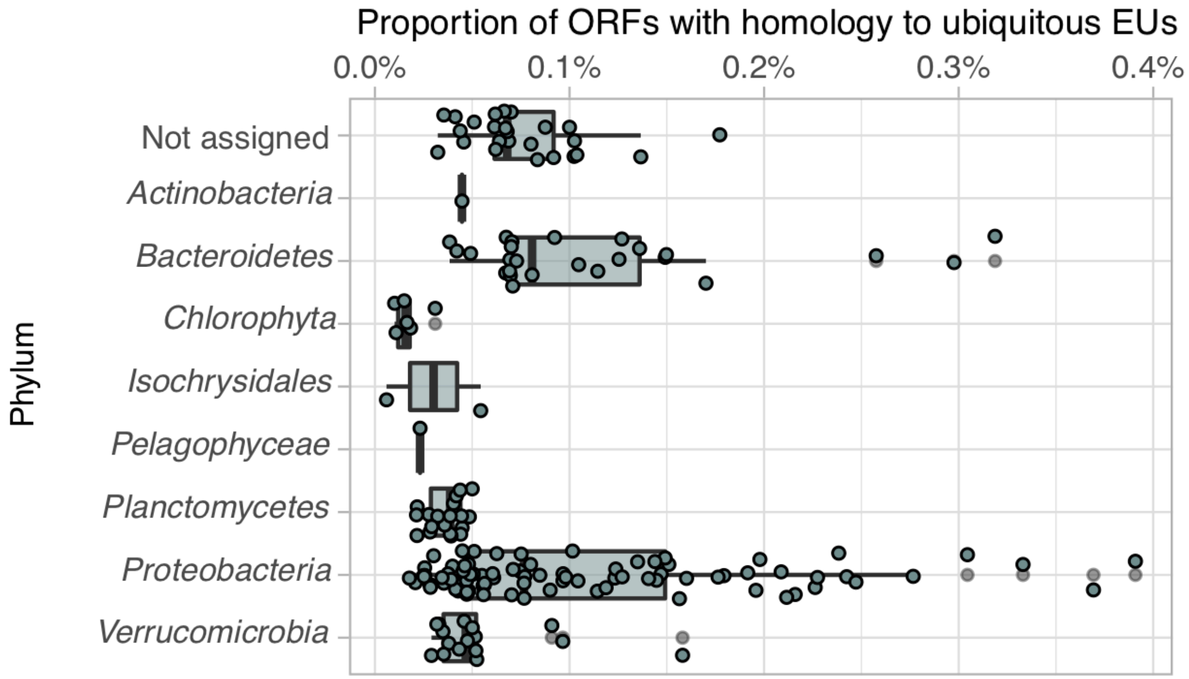


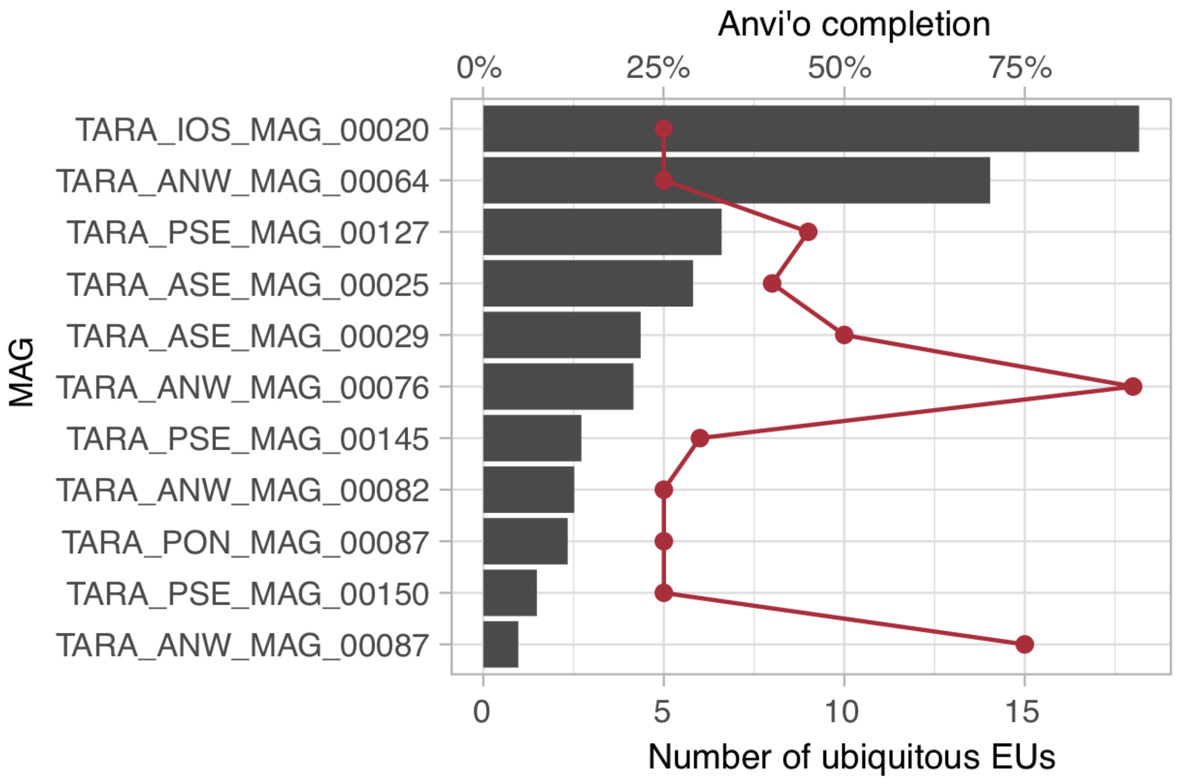
Delmont et al. 2017

Mapping potential EUs to TARA MAGs

-13,972 (91%) of EUs mapped to manually curated TARA MAGs

-Picked MAGs with lowest completion but highest amount of ubiquitous EUs

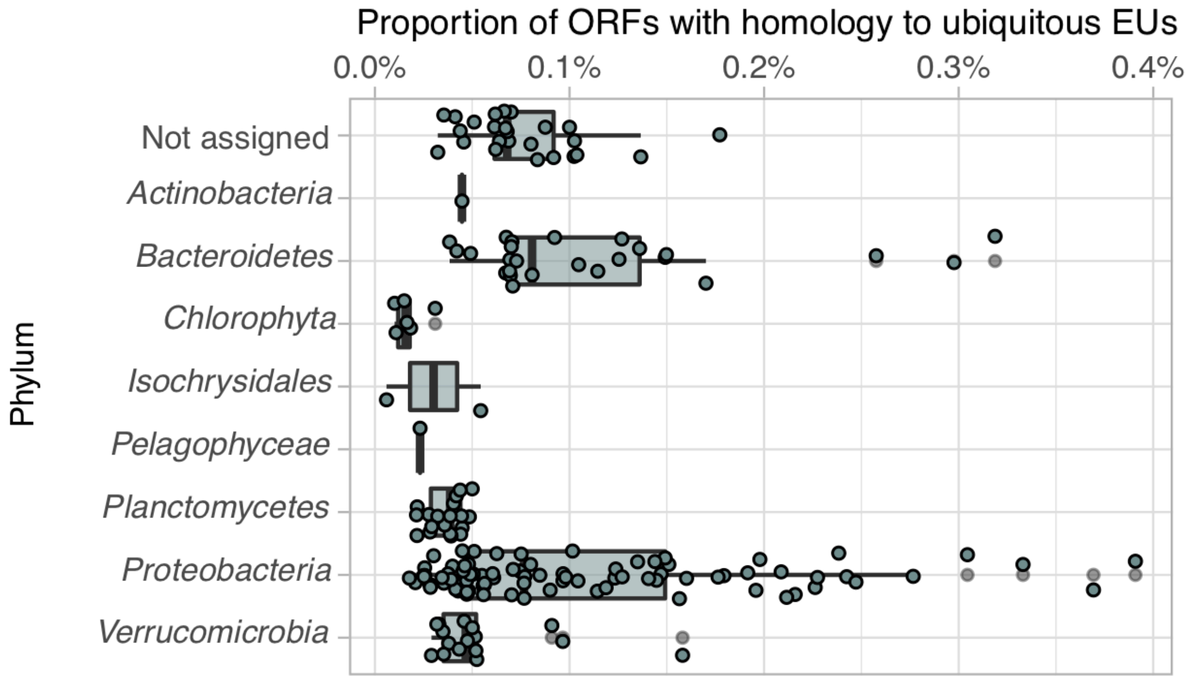


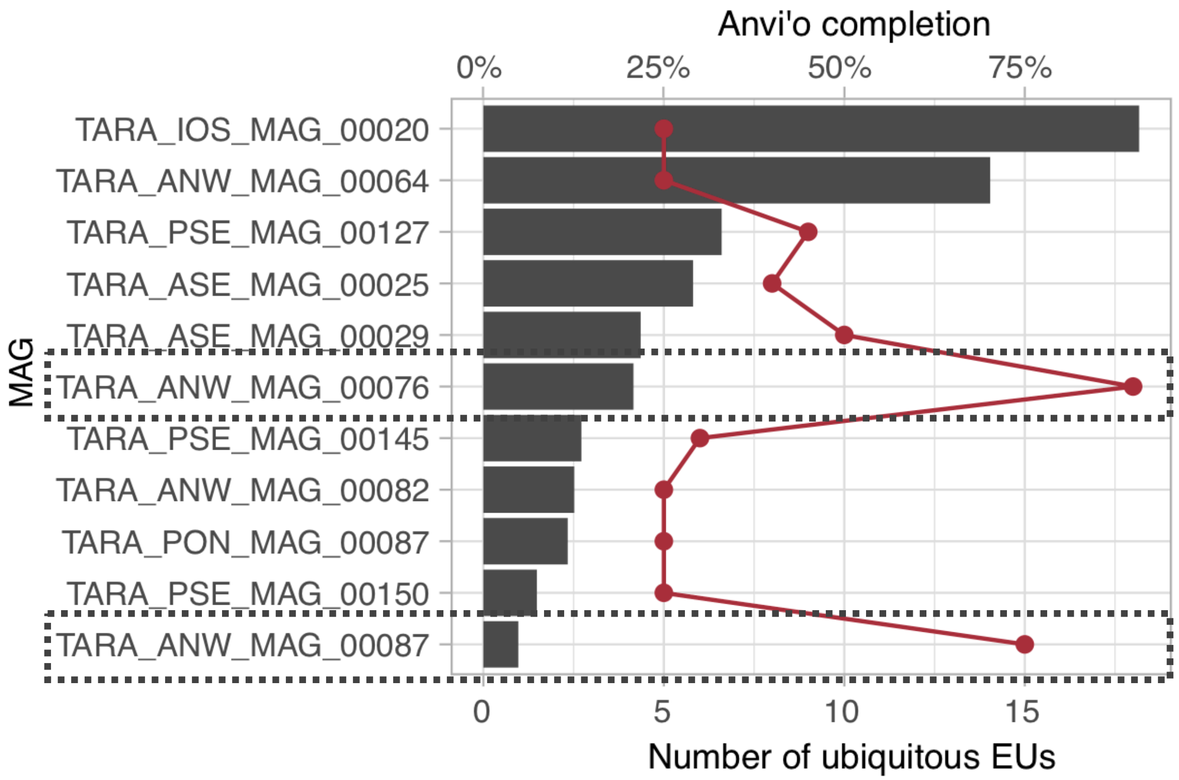


Mapping potential EUs to TARA MAGs

-13,972 (91%) of EUs mapped to the TARA MAGs

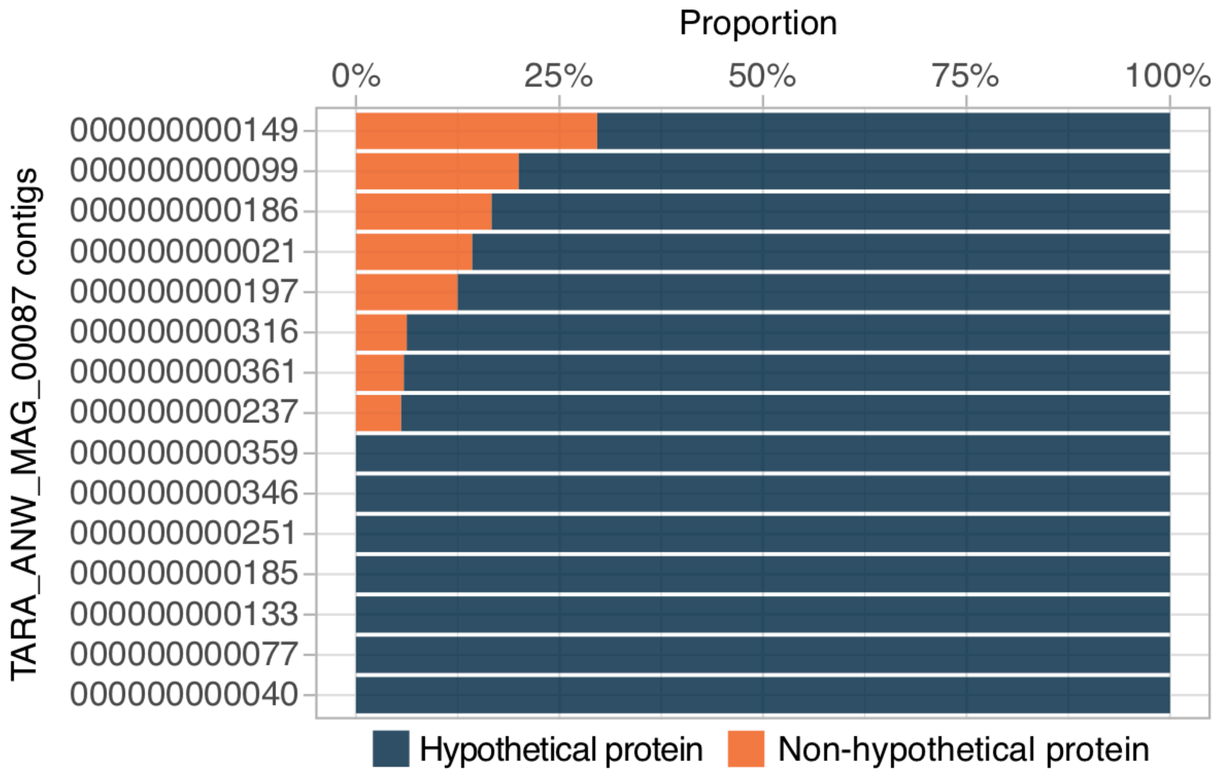
-Picked MAGs with lowest completion but highest amount of ubiquitous EUs



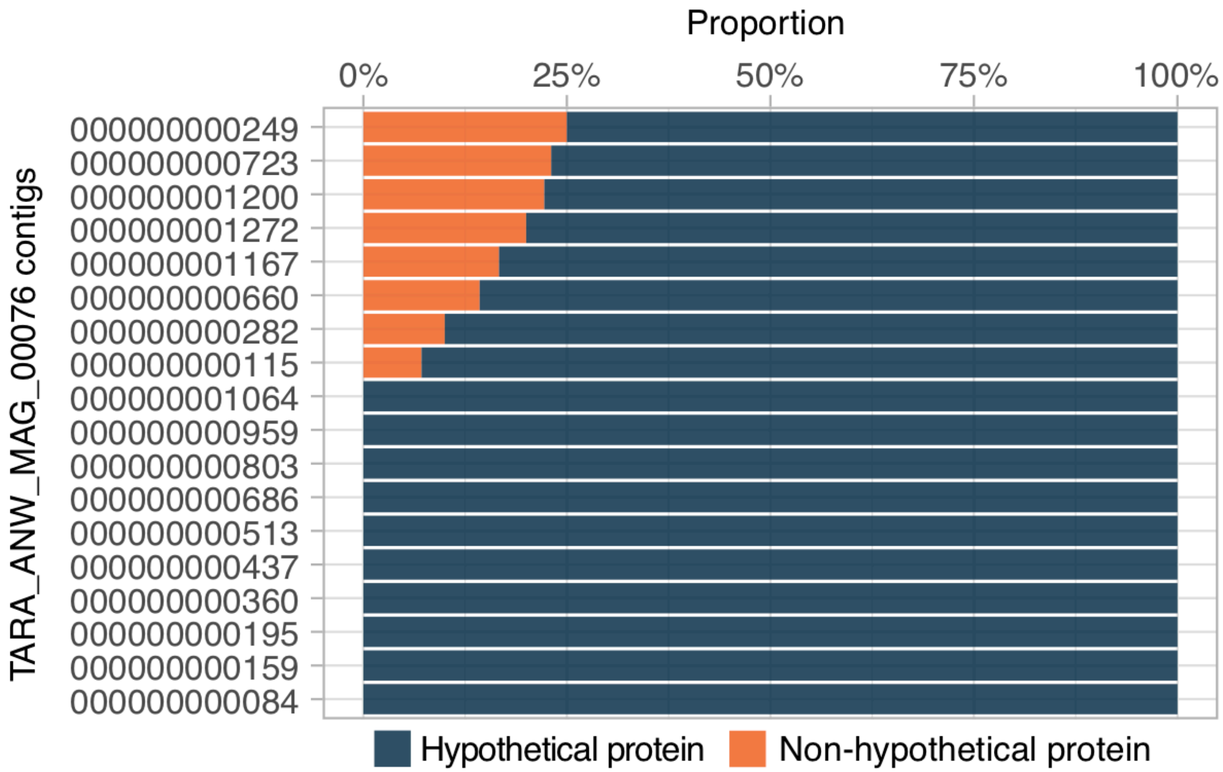


MAG contig protein composition

Gammaproteobacteria
Completion: 4.84%
No. of ubiquitous EUs: 15

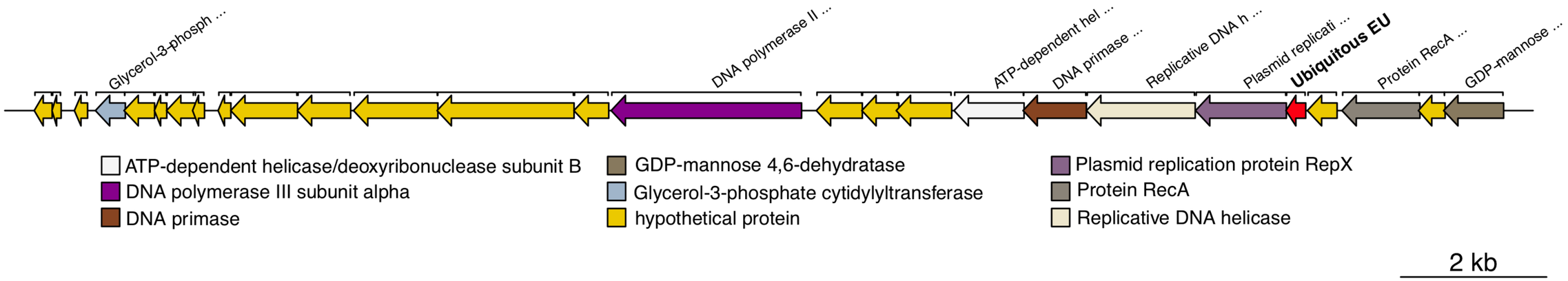


Flavobacteria
Completion: 20.8%
No. of ubiquitous EUs: 21



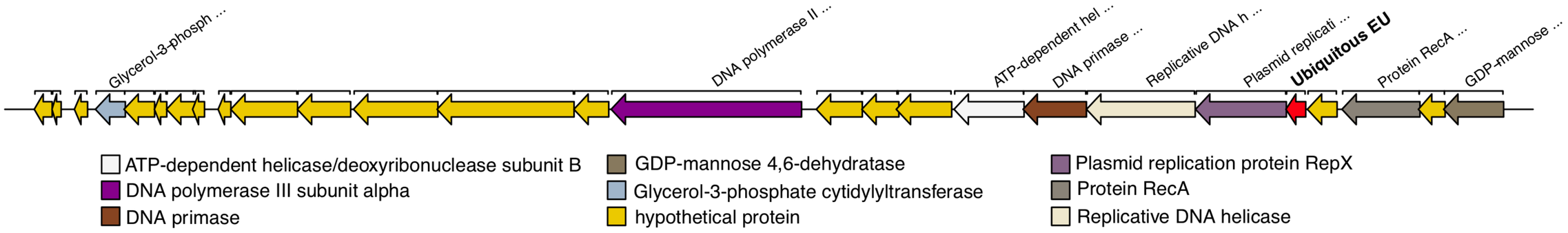
Contextualizing potential EU components

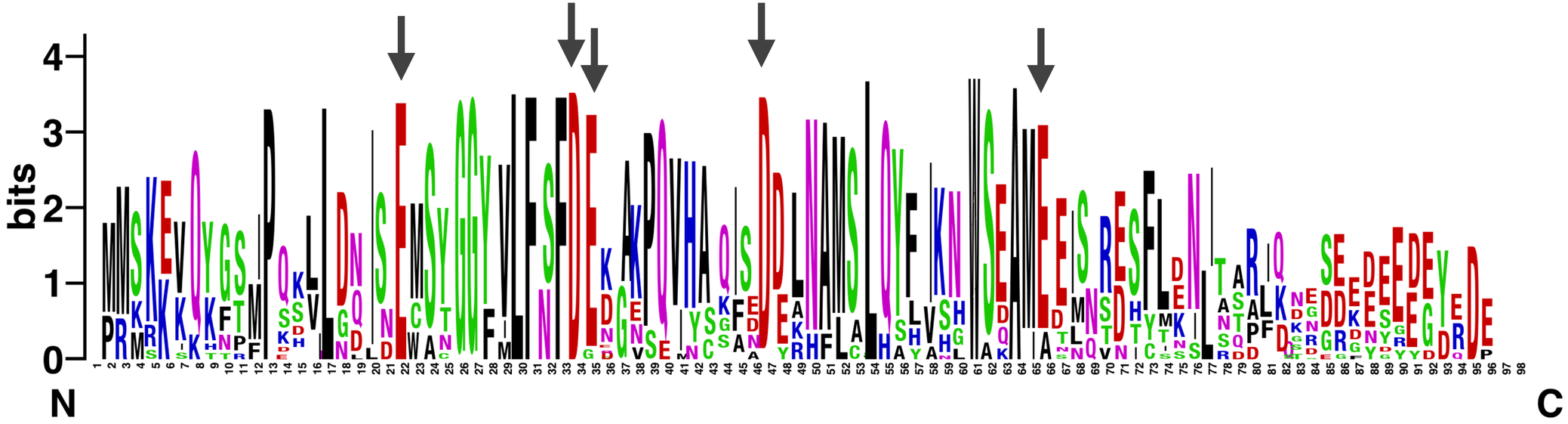
Gammaproteobacteria; contig: 0000000000149



Contextualizing potential EU components

Gammaproteobacteria; Completion: 4.84%; contig: 0000000000149

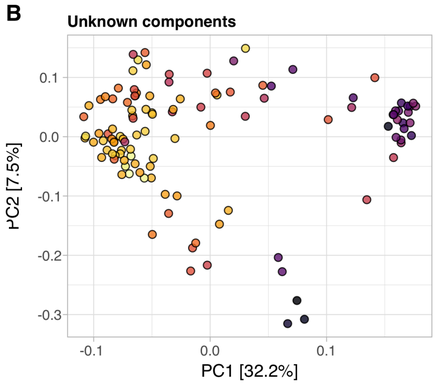




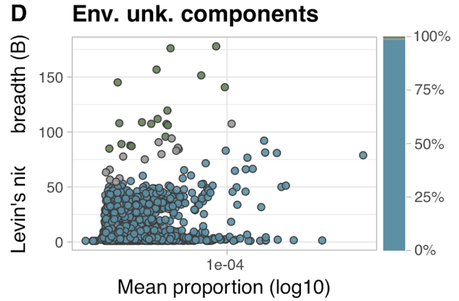
-consensus residues ≥ 2 active residues

Conclusions

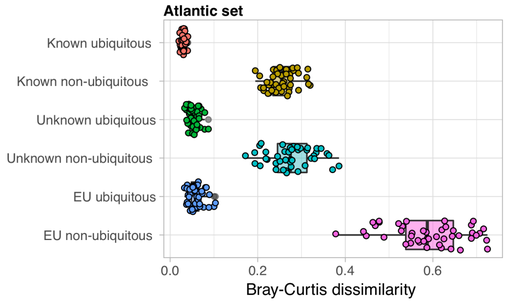
- The inclusion of the unknown fraction increases sample site resolution in ordinations



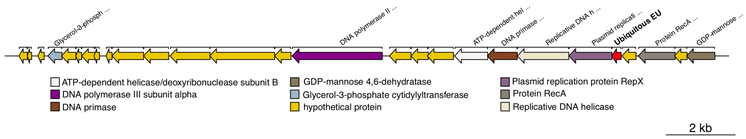
- The unknowns show distinct ecological patterns and may be enriched in adaptive proteins



- First hints of a ubiquitous, unknown functional fraction that needs to be explored



- Functional hypotheses may be inferred by investigating genomic context on MAGs



Code and data availability

https://github.com/mschecht/Unknown_unknowns

DOI - 10.6084/m9.figshare.5979658

Acknowledgments

Reviewers: Prof. Dr. Frank Oliver Glöckner, Dr. Pier Luigi Buttigieg

Supervisor: Dr. Antonio Fernandez-Guerra

Chiara Vanni

Lab rotation supervisors: Dr. David Probant, Dr. Gunter Wegener, Dr. Manuel Liebeke

Christiane and Karl-Heinz

The MarMic class of 2021

Ben, David B., Alaina, Bledina, David P., Cora, Alex, Henny, Kai, and MGG - Thanks for you feedback!

Coffee amigos: David, Cora, and Andrea

Lot of love to my family and friends!

Max Planck Institute for Marine Microbiology

marmic

Citations

Buttigieg, P. L., Hankeln, W., Kostadinov, I., Kottmann, R., Yilmaz, P., Duhaime, M. B. & Glöckner, F. O. (2013). Ecogenomic Perspectives on Domains of Unknown Function: Correlation-Based Exploration of Marine Metagenomes. PLoS ONE, 8(3), e50869. URL https://doi.org/10.1371%2Fjournal.pone.0050869.

Byungwook Lee & Doheon Lee (2009). Protein comparison at the domain architecture level. BMC Bioinformatics, 10, S5. URL https://doi.org/10.1186/1471-2105-10-S15-S5

Cavicchioli, R. (2015). Microbial ecology of Antarctic aquatic systems. Nature Reviews Microbiology, 13(11), 691–706. URL https://doi.org/10.1038%2Fnrmicro3549.

Ellens, K. W., Christian, N., Singh, C., Satagopam, V. P., May, P. & Linster, C. L. (2017). Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Research, 45(20), 11495–11514. URL https://doi.org/10.1093%2Fnar% 2Fgkx937.

Jaroszewski, L., Li, Z., Krishna, S. S., Bakolitsa, C., Wooley, J., Deacon, A. M., Wilson, I. A. & Godzik, A. (2009). Exploration of Uncharted Regions of the Protein Universe. PLoS Biology, 7(9), e1000205. URL https://doi.org/10.1371% 2Fjournal.pbio.1000205.

Jennifer R. Brum, J. Cesar Ignacio-Espinoza, Eun-Hae Kim, Gareth Trubl, Robert M. Jones, Simon Roux, Nathan C. VerBerkmoes, Virginia I. Rich & Matthew B. Sullivan (2016). Illuminating structural proteins in viral \textquotedblleftdark matter\textquotedblright with metaproteomics. *Proceedings of the National Academy of Sciences, 113*, 2436-2441.

Wyman, S. K., Avila-Herrera, A., Nayfach, S. & Pollard, K. S. (2017). A most wanted list of conserved protein families with no known domains. URL https: //doi.org/10.1101%2F207985.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J. et al. (2007). The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. PLoS Biology, 5(3), e16. URL https://doi.org/10.1371% 2Fjournal.pbio.0050016.

Supplemental Figures

Cluster aggregation into components

	Knowns	Genomic Unknowns	Environmental Unknowns	Knowns without Pfam	Total
ORFs	42,937,850	48,579,179	9,472,648	37,073,885	262,821,906
Clusters	914,932	894,280	388,624	756,067	2,953,903

DomArch_1

cluster_1

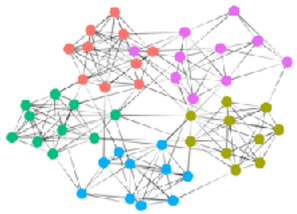
cluster_326

cluster_78

…

Consensus →

SSN
overlap 80%
similarity 20%



DomArch_2

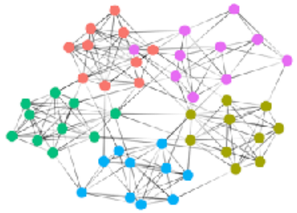
cluster_26

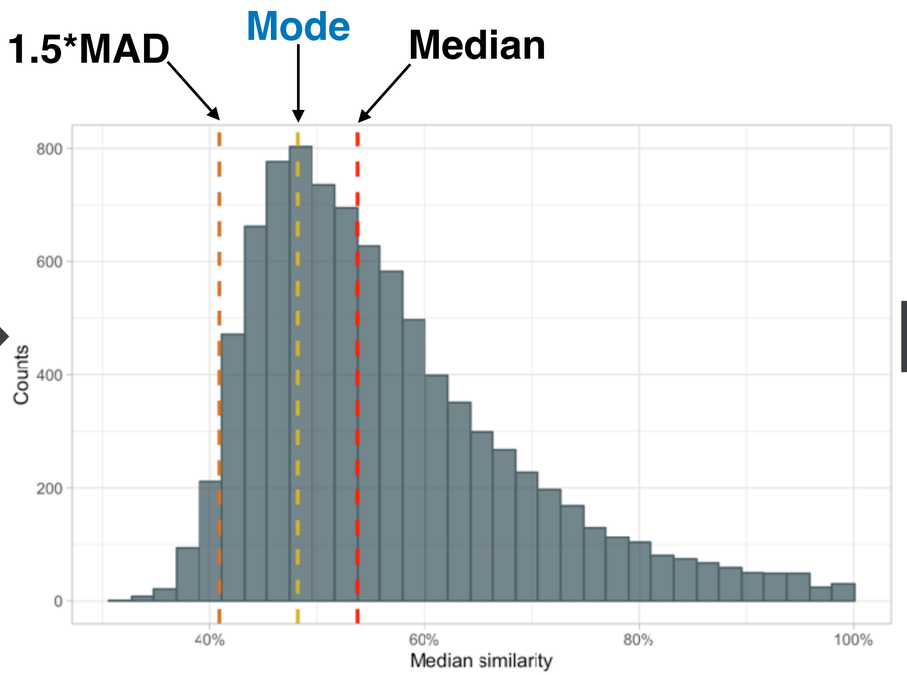
cluster_98

cluster_56

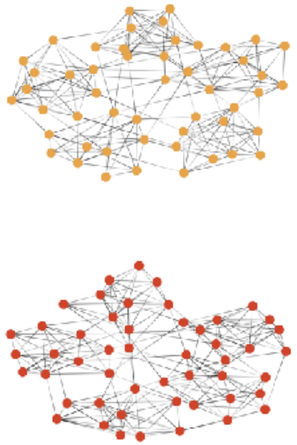
…

Consensus →





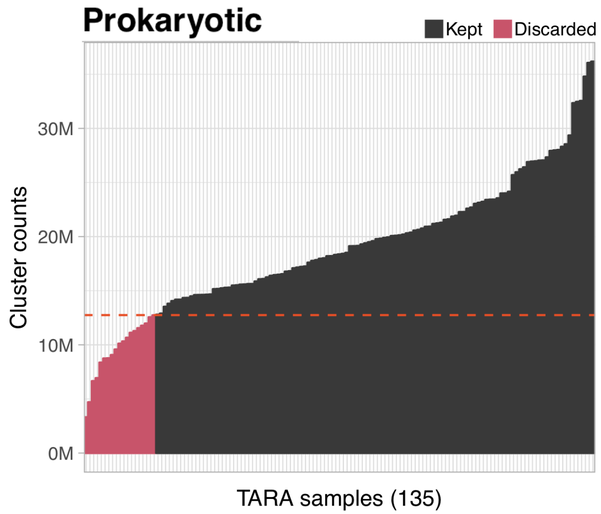
Cutoff threshold

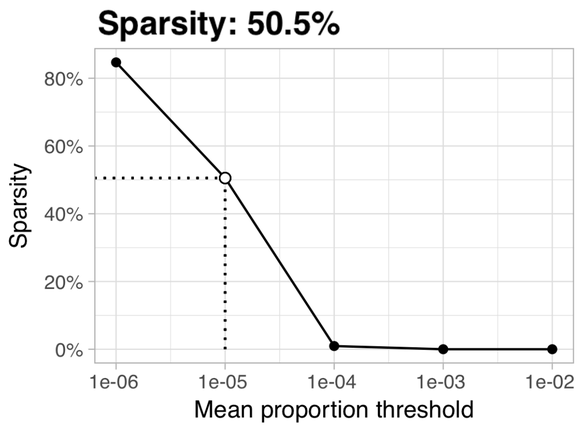


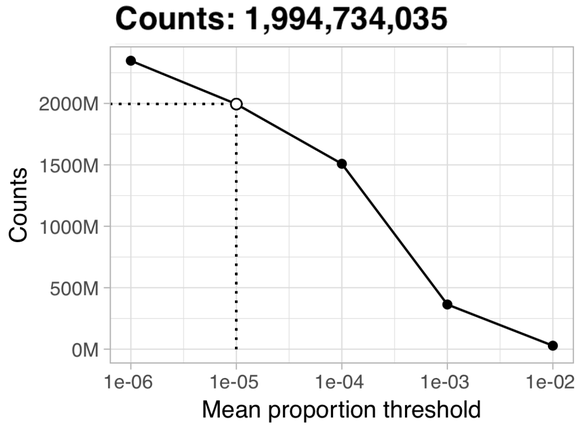
SSN: Sequence Similarity network with Parasail (Daily, 2016)

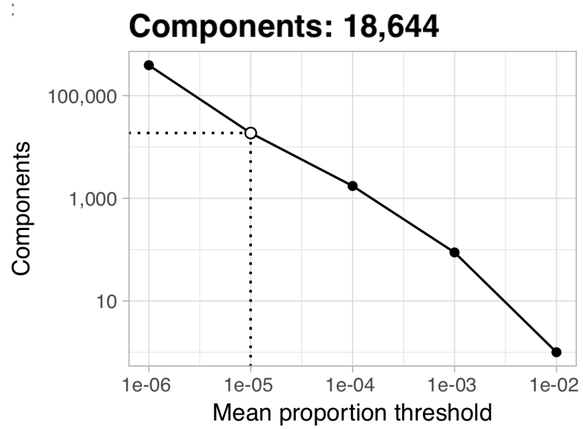
Pre-processing component abundance data

- Low count samples
- Low proportion components





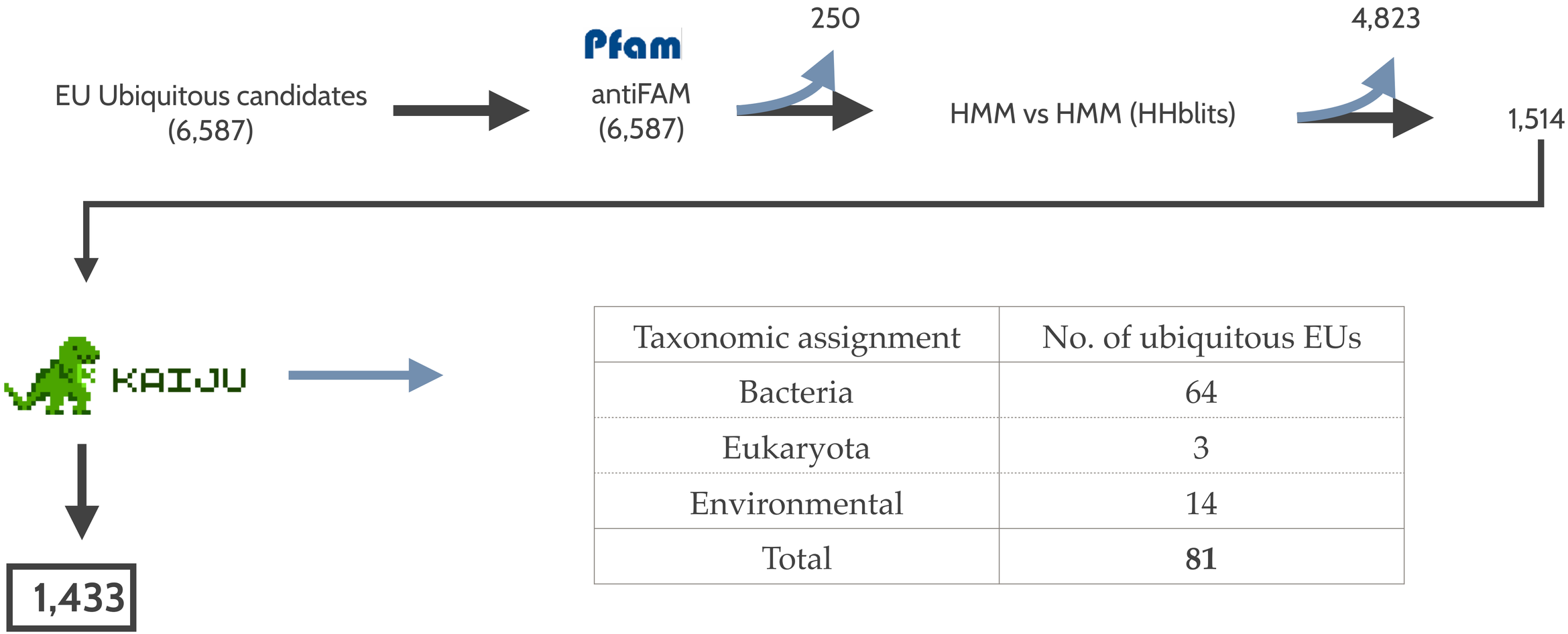




Final TARA component subset

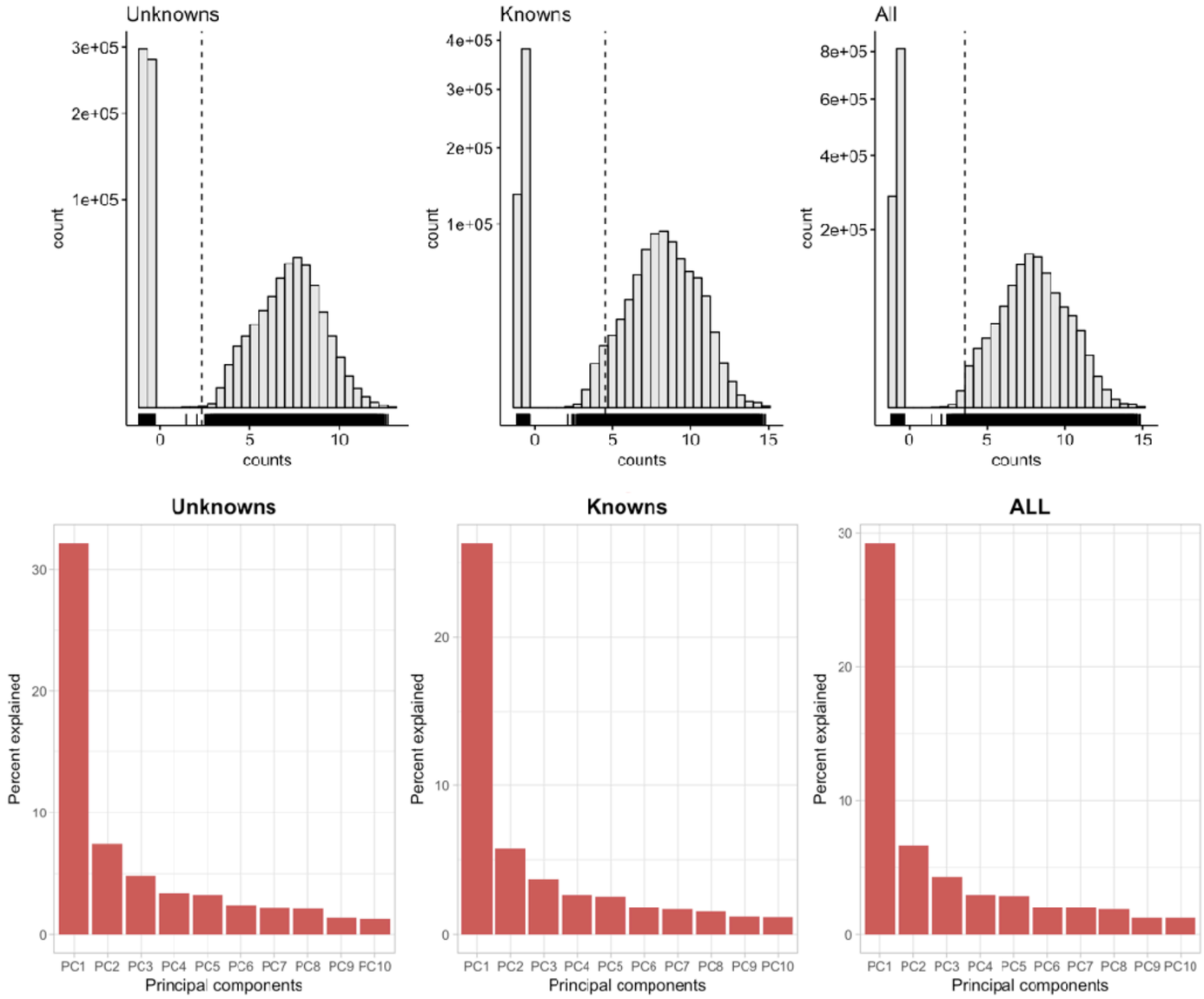
	Depths	Knowns	Genomic Unknowns	Environmental Unknowns	Knowns without Pfam	Total
Prokaryotic	Surface	3593	4997	1287	5476	15,353
	All	3997	6497	1534	6616	18,644

Steps to validate the potential unknown

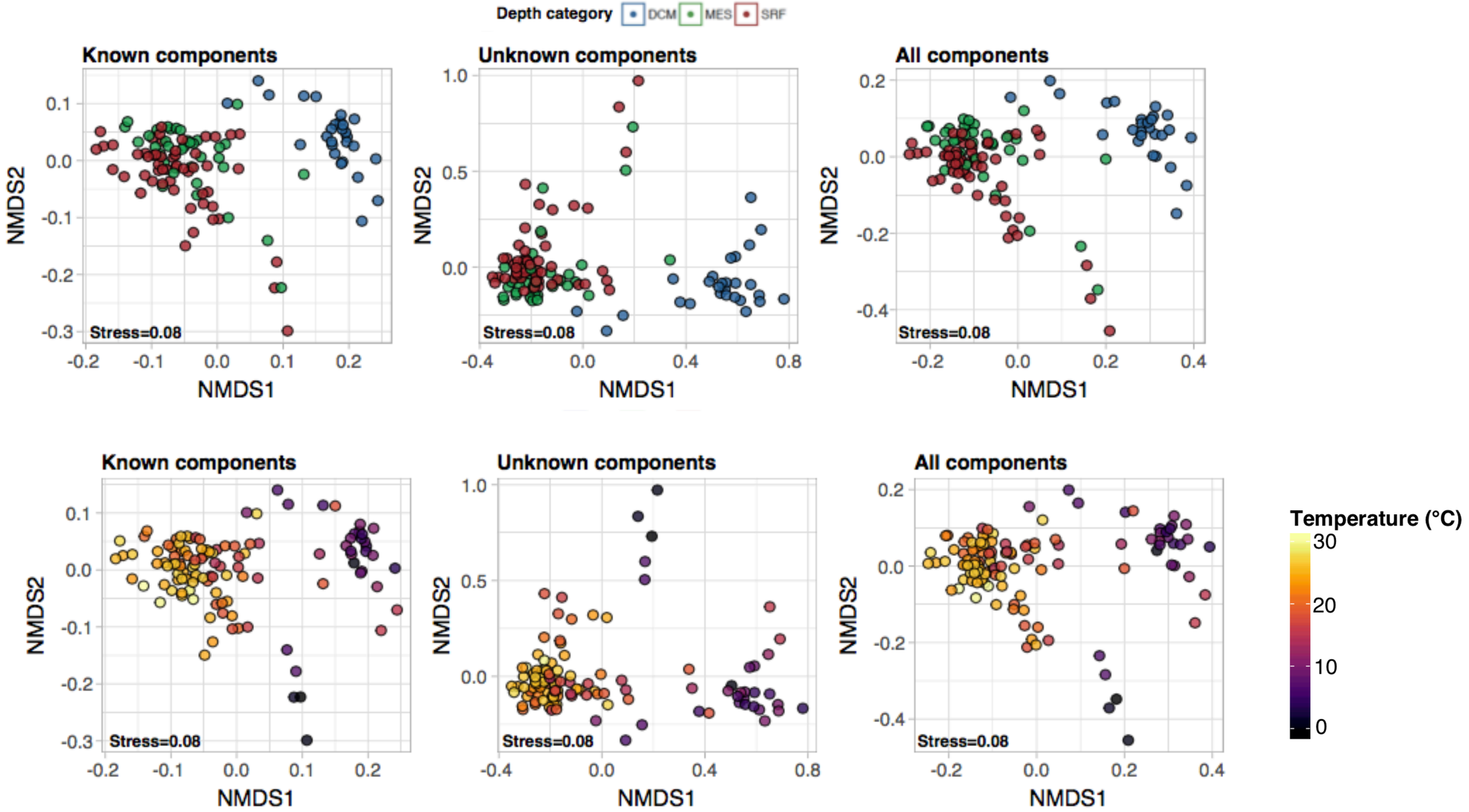


Taxonomic assignment	No. of ubiquitous EUs
Bacteria	64
Eukaryota	3
Environmental	14
Total	**81**

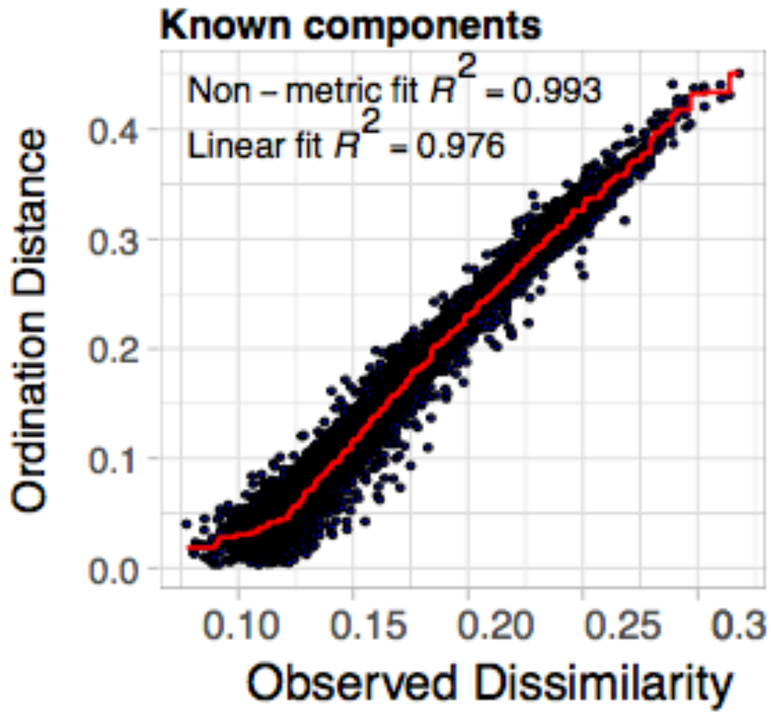
PCA residuals and scree plot

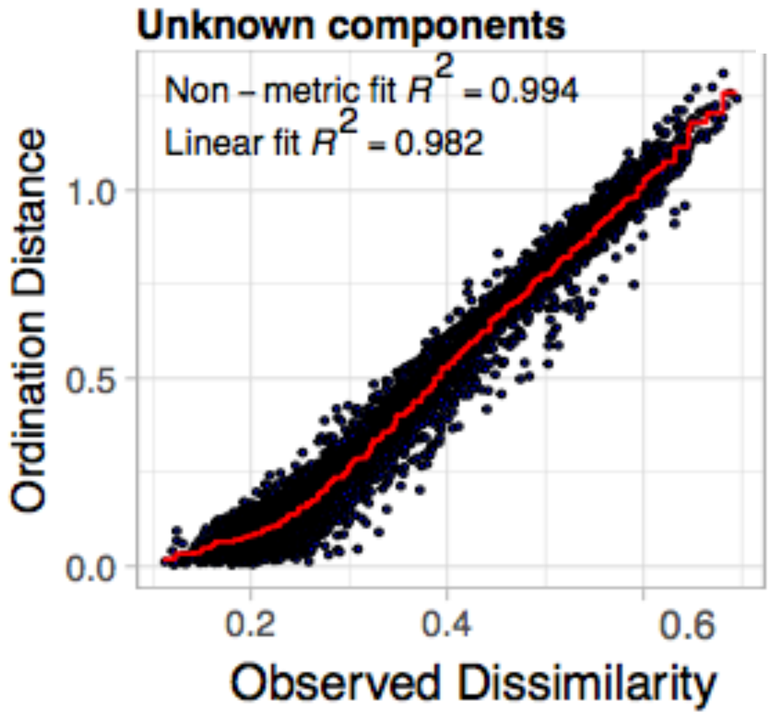


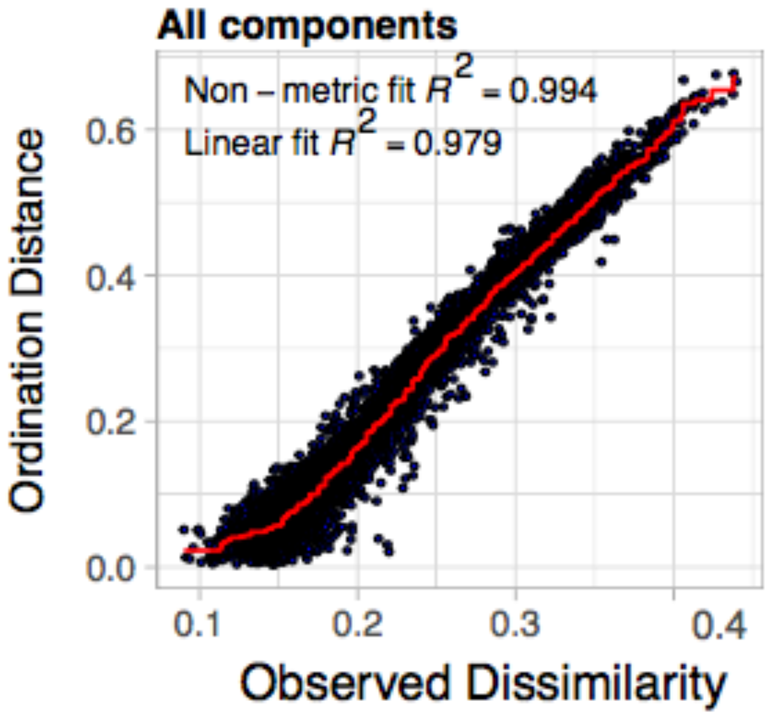
NMDS plots


Depth category
DCM
MES
SRF
Known components
Unknown components
All components
NMDS1
NMDS2
Stress=0.08
Temperature (°C)
30
20
10
0

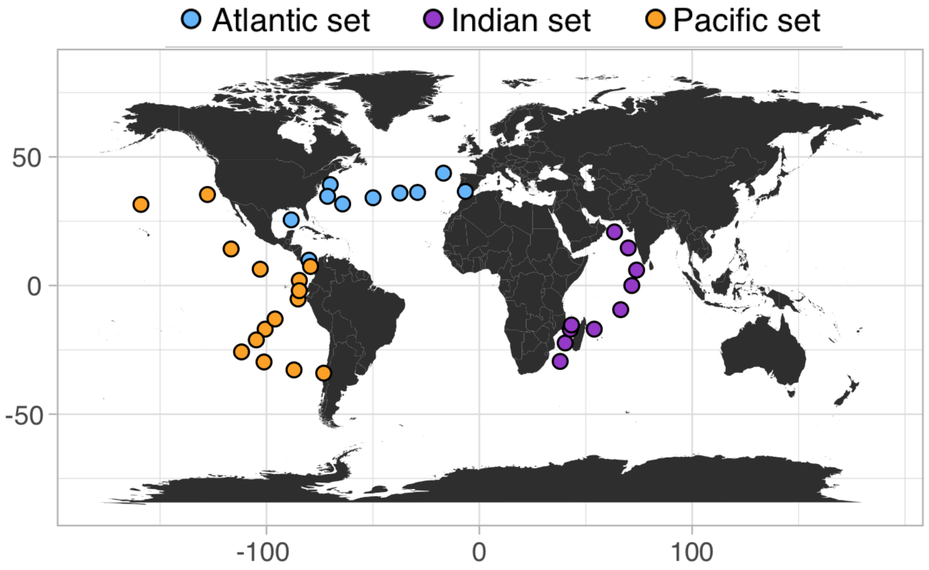
NMDS Shepard plot

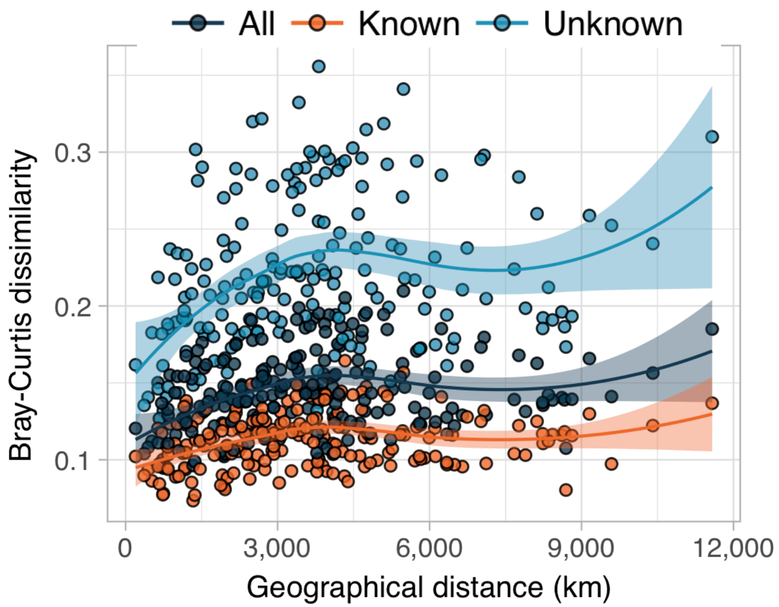


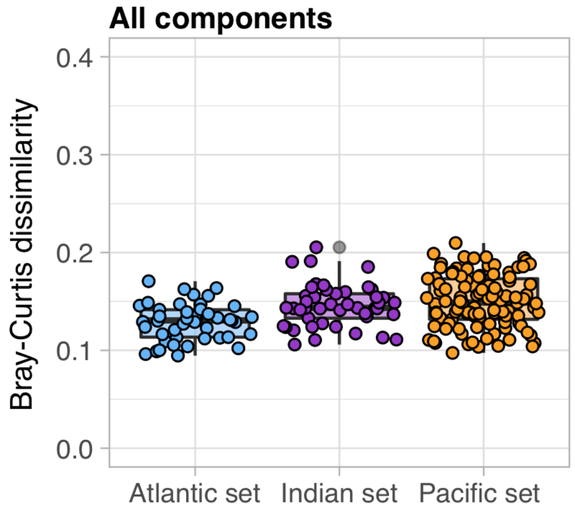


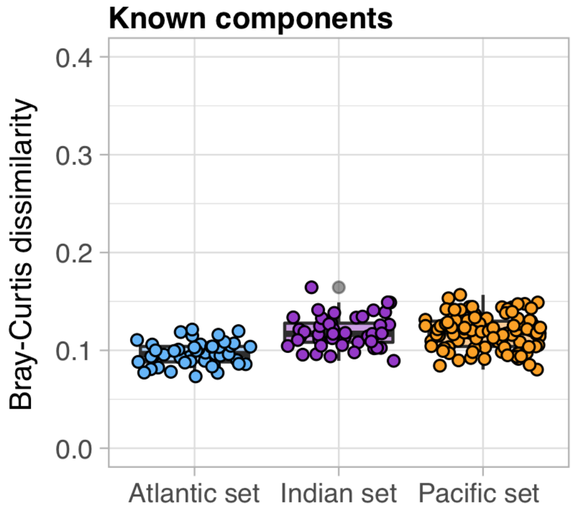


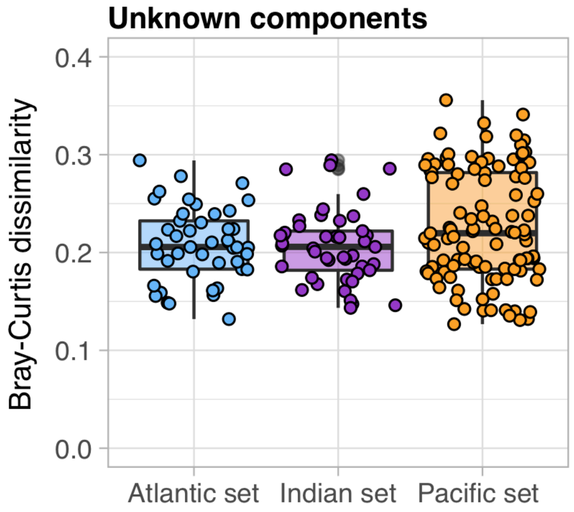
Genetic vs. geographic distance











Statistical tests

	PERMANOVA		Beta-dispersion
	R^2	p-value	p-value
Unknowns (EUs + GUs)	0.27237	0.001	0.073
Knowns (K + Kwp)	0.22649	0.001	0.121
All combined	0.25064	0.001	0.117

Depth category

	PERMANOVA	
	R^2	p-value
Unknowns (EUs + GUs)	0.27237	0.001
Knowns (K + Kwp)	0.22649	0.001
All combined	0.25064	0.001

Temperature

	Mantel test		partial-Mantel test	
	ρ	p-value	ρ	p-value
Unknowns (EUs + GUs)	0.966	0.001	0.512	0.001
Knowns (K + Kwp)	0.968	0.001	0.509	0.001
All combined	0.965	0.001	0.515	0.001

	Mantel test		partial-Mantel test	
	ρ	p-value	ρ	p-value
Unknowns (Atlantic)	0.385	0.0201	0.0201	0.0240
Knowns (Atlantic)	0.475	0.0066	0.387	0.0240
All (Atlantic)	0.433	0.0120	0.417	0.0173
Unknowns (Pacific)	0.128	0.2528	0.086	0.3160
Knowns (Pacific)	0.072	0.3384	0.046	0.3818
All (Pacific)	0.111	0.2630	0.076	0.3317
Unknowns (Indian)	0.068	0.3189	0.068	0.3070
Knowns (Indian)	-0.196	0.8849	-0.177	0.8595
All (Indian)	-0.013	0.5159	-0.013	0.5159

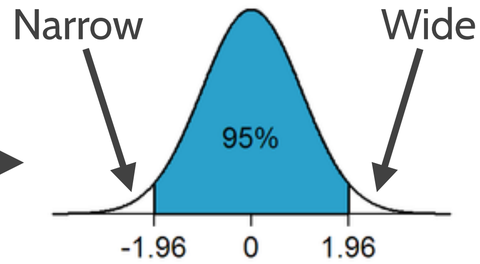
Niche breadth

Calculate Niche Breadth scores → Create component score distribution (quasiswap) (Miklós & Podani 2004) → Calculate 95% confidence interval → See where original score lies and classify

	Component_1	Component_2	Component_3
Site_1	4	3	0
Site_2	6	2	2
Site_3	4	0	7
Site_4	2	3	0

	Component_1	Component_2	Component_3
Site_1	6	3	0
Site_2	2	2	2
Site_3	4	0	7
Site_4	2	3	0

	Component_1	Component_2	Component_3
Site_1	4	3	0
Site_2	6	7	0
Site_3	4	2	2
Site_4	2	3	0

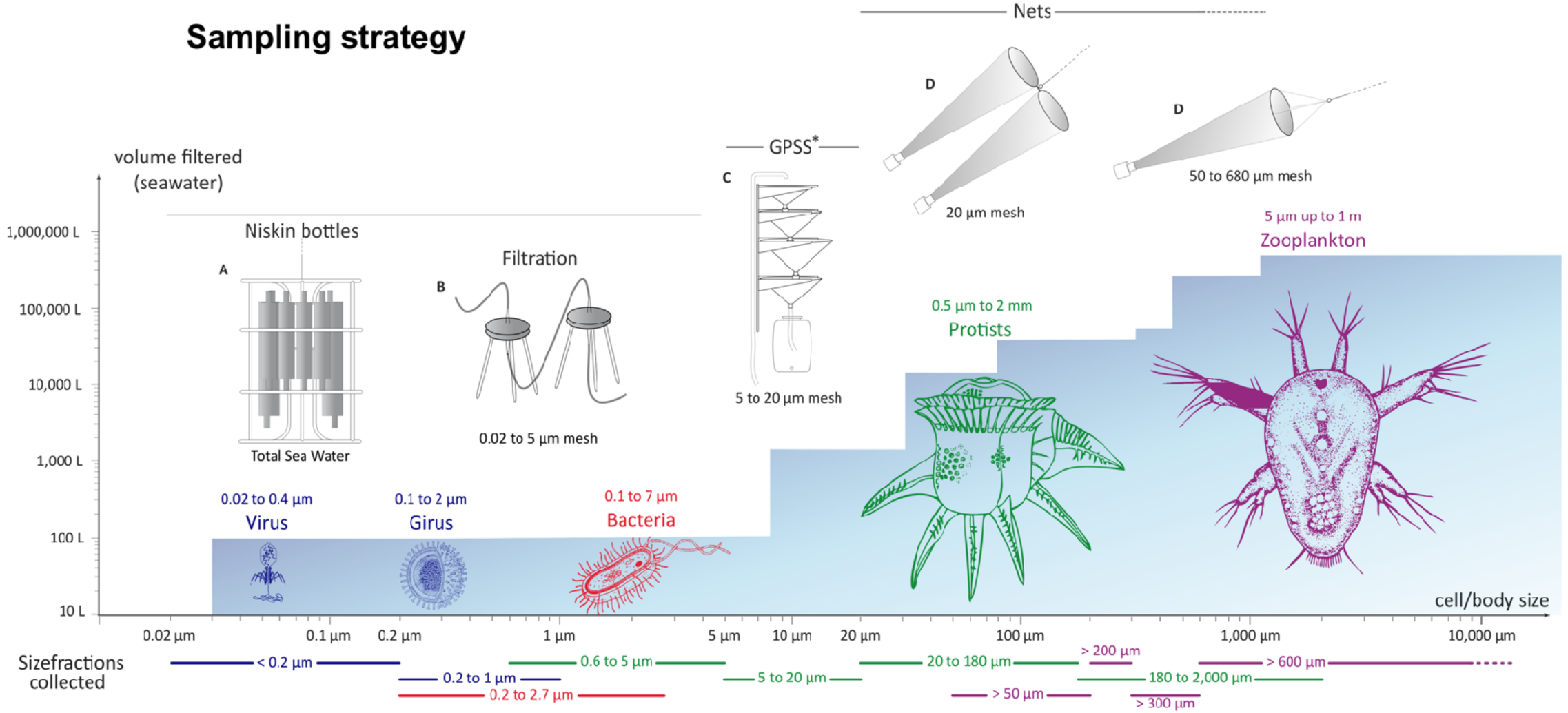


- Quasiswap: maintains column and row sums

... 100 x

https://en.wikipedia.org/wiki/1.96

TARA Ocean filter fractions



Karsenti et al. 2011

Contextualizing potential EU components

ORF	Prokka annotation	Manual annotation
TARA_ANW_MAG_00087_00079	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00080	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00081	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00082	Glycerol-3-phosphate cytidylyltransferase	Glycerol-3-phosphate cytidylyltransferase
TARA_ANW_MAG_00087_00083	hypothetical protein	endonuclease
TARA_ANW_MAG_00087_00084	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00085	hypothetical protein	nuclease
TARA_ANW_MAG_00087_00086	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00087	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00088	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00089	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00090	hypothetical protein	Glycosil transferase family 9
TARA_ANW_MAG_00087_00091	hypothetical protein	Glycosyl transferase family 1

Contextualizing potential EU components

ORF	Prokka annotation	Manual annotation
TARA_ANW_MAG_00087_00092	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00093	DNA polymerase III subunit alpha	DNA polymerase III, alpha subunit
TARA_ANW_MAG_00087_00094	hypothetical protein	DNA polymerase III, epsilon subunit
TARA_ANW_MAG_00087_00095	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00096	hypothetical protein	DNA-directed DNA polymerase
TARA_ANW_MAG_00087_00097	ATP-dependent helicase/deoxyribonuclease subunit B	ATP-dependent helicase/deoxyribonuclease subunit B
TARA_ANW_MAG_00087_00098	DNA primase	DNA primase
TARA_ANW_MAG_00087_00099	Replicative DNA helicase	Replicative DNA helicase
TARA_ANW_MAG_00087_00100	Plasmid replication protein RepX	Cell division protein FtsZ
TARA_ANW_MAG_00087_00101	**hypothetical protein**	**hypothetical protein**
TARA_ANW_MAG_00087_00102	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00103	Protein RecA	Protein RecA
TARA_ANW_MAG_00087_00104	hypothetical protein	hypothetical protein
TARA_ANW_MAG_00087_00105	GDP-mannose 4,6-dehydratase	UDP-glucose 4-epimerase