A computationally intensive approach to discover novel adaptation genes in extreme environments

Genomics of Gene Expression Lab

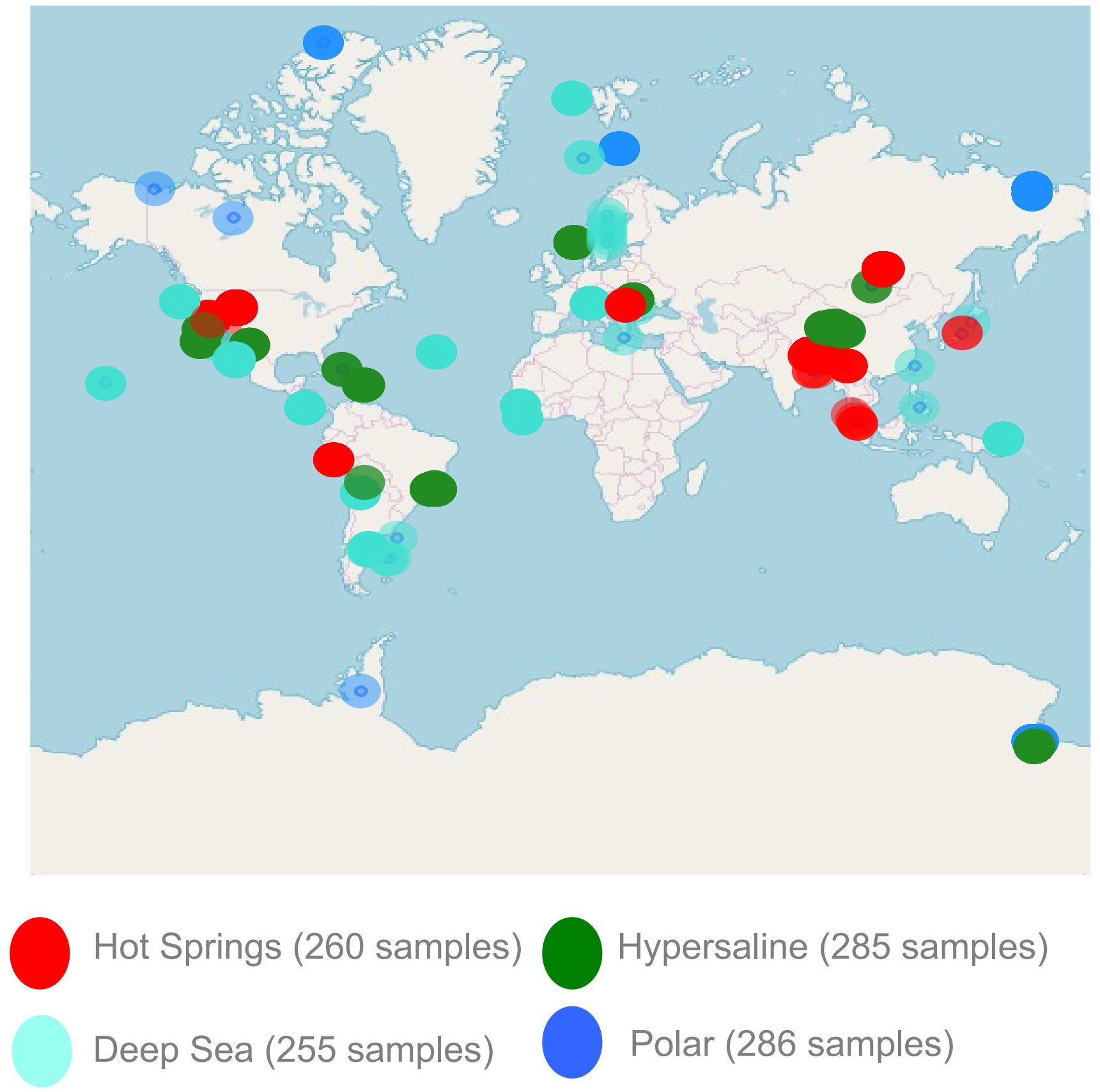
ISMB 2020 Submission #224

Tatyana Zamkovaya[1], Jamie S. Foster[2], Valérie de Crecy-Lágard[1,3], Ana Conesa[1,3]

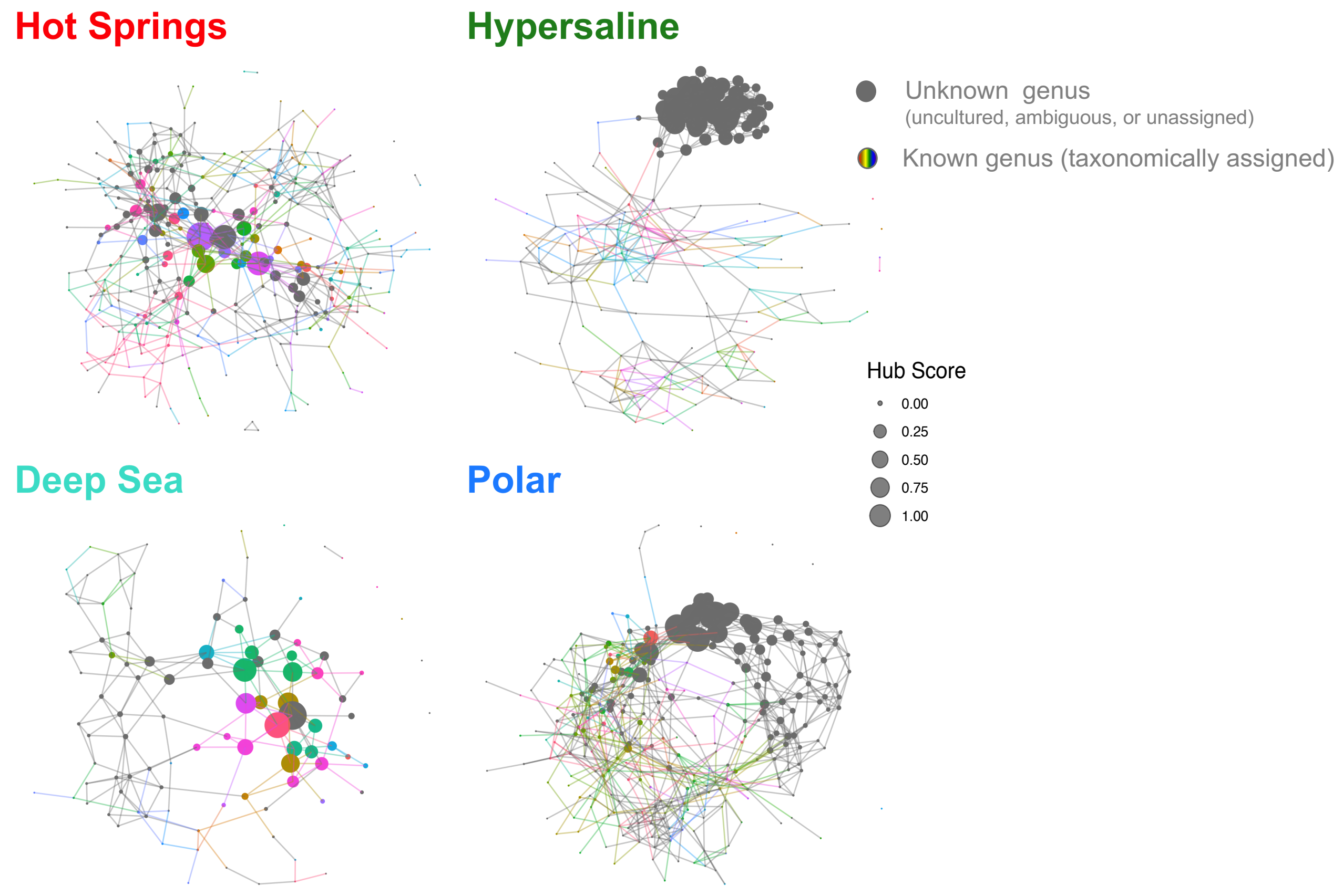
[1] Microbiology and Cell Science Department, Institute for Food and Agricultural Sciences, University of Florida, Gainesville, FL, USA
[2] Department of Microbiology and Cell Science, Space Life Sciences Lab, Merritt Island, FL, USA
[3] Genetics and Genomics Institute, University of Florida, Gainesville, FL, USA

Introduction

Unknown taxa act as top hubs in extreme environmental networks, yet have unknown functional roles



A) Map of samples collected.



B) Bacterial networks at genus level, with nodes resized by hub score.

Hypothesis and Aim

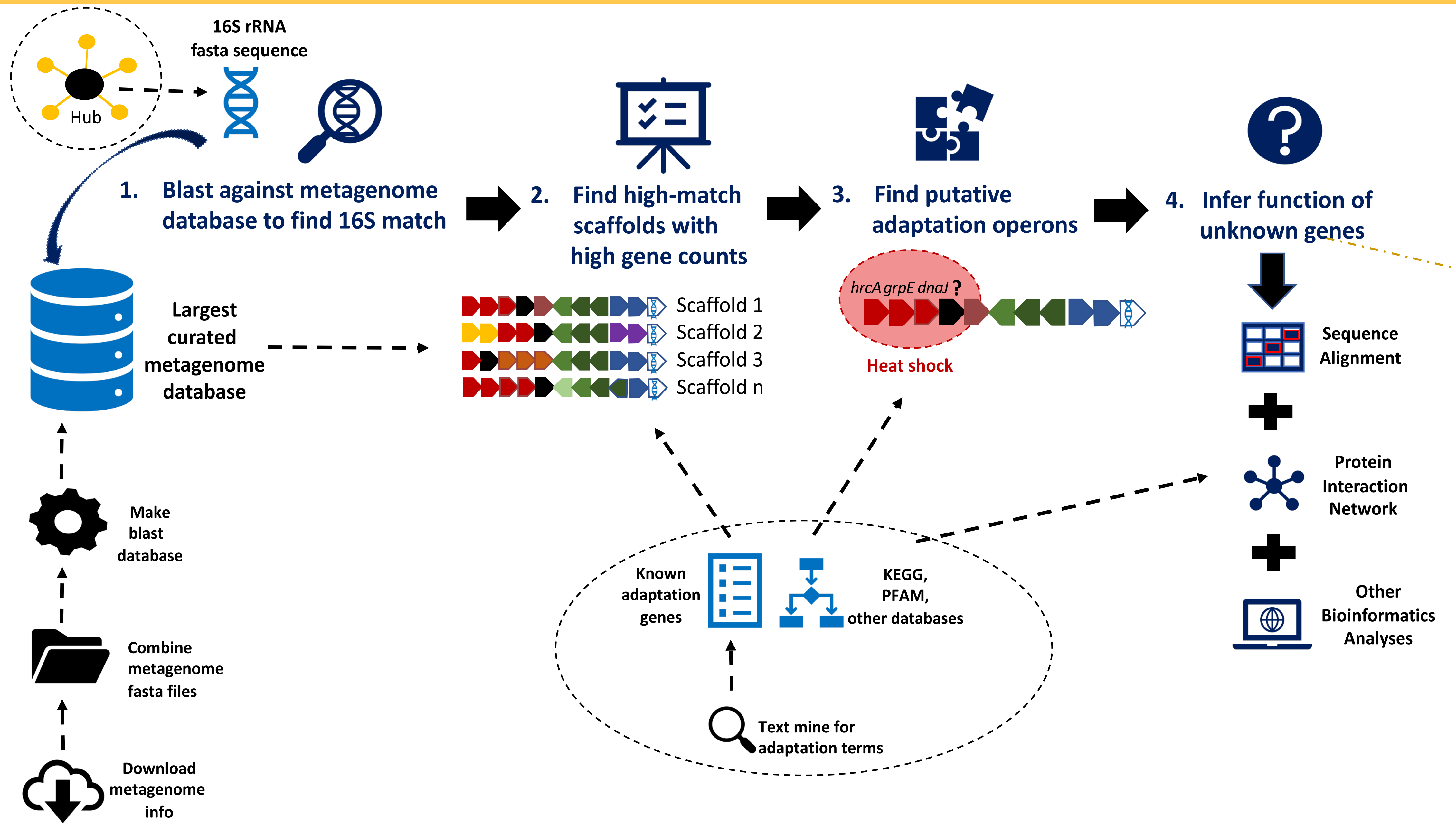
The key to understand how life evolves and adapts to extreme conditions, within and beyond Earth, lies in the uncharacterized, unknown taxa, called Microbial Dark Matter (MDM), that dominate even the most extreme of environments. We were able to identify unknown hubs of extreme environments using a network-based approach on amplicon data. We found that unknowns are highly prevalent and connected and wanted to know more about how and why these unknowns were able to adapt to such harsh conditions. Building on these results, we now propose a computational approach to identify adaptation-related genes in these unknown organisms.

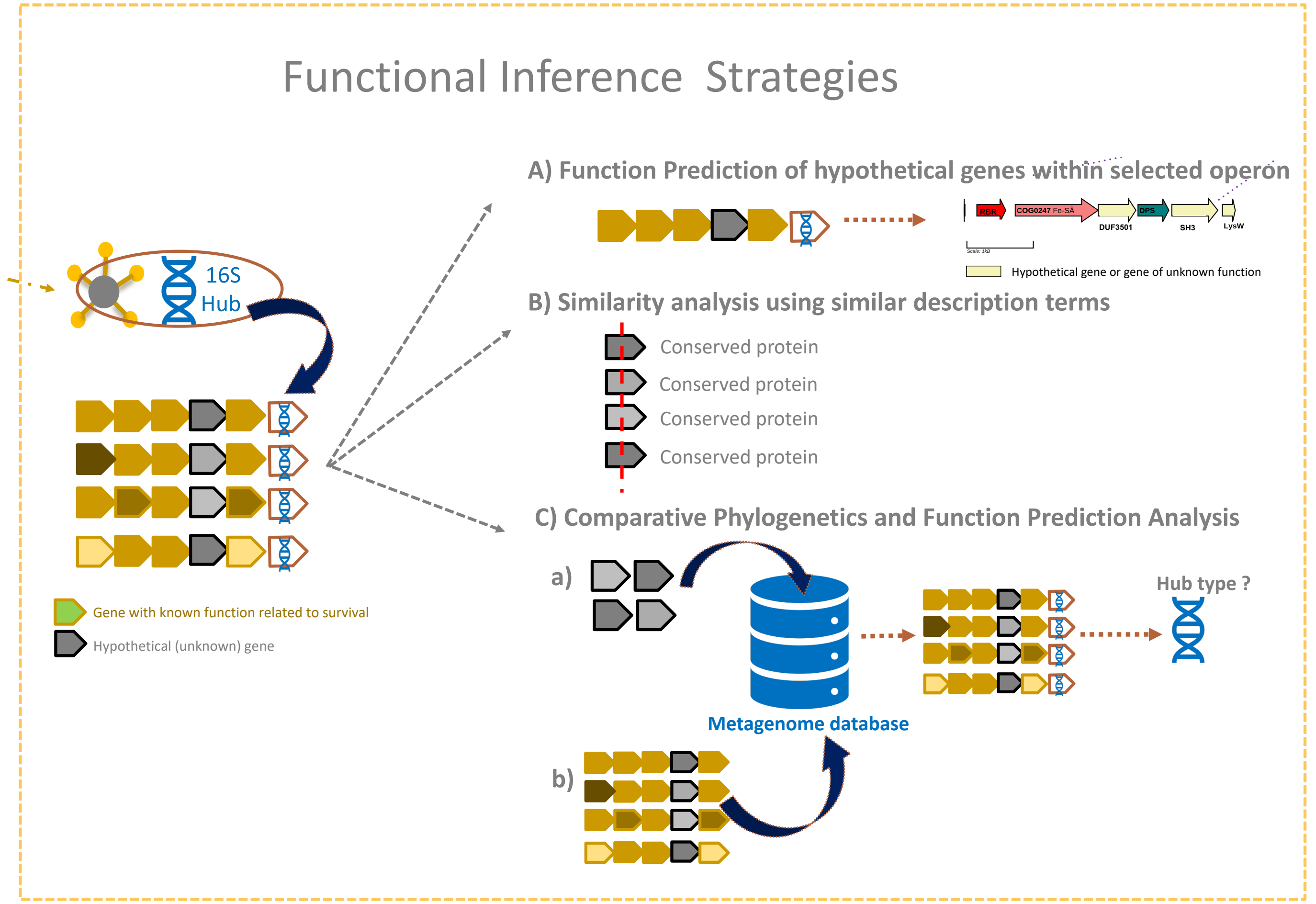
Due to their integral network position, these extreme environmental hubs may possess optimized adaptation strategies and therefore are ideal candidates to search for novel biological pathways of survival.

Lack of well-annotated reference genomes prevents existing tools from predicting gene function of unknown taxa.

Here, we circumvent this problem by using **16S hubs as probes** to **find novel genes** within **functionally annotated metagenome scaffolds.**

Hub Blast Pipeline


Hub
16S rRNA fasta sequence
1. Blast against metagenome database to find 16S match
2. Find high-match scaffolds with high gene counts
3. Find putative adaptation operons
4. Infer function of unknown genes
Largest curated metagenome database
Scaffold 1
Scaffold 2
Scaffold 3
Scaffold n
hrcA grpE dnaJ ?
Heat shock
Sequence Alignment
Protein Interaction Network
Other Bioinformatics Analyses
Make blast database
Combine metagenome fasta files
Download metagenome info
Known adaptation genes
KEGG, PFAM, other databases
Text mine for adaptation terms


Functional Inference Strategies
16S Hub
Gene with known function related to survival
Hypothetical (unknown) gene
A) Function Prediction of hypothetical genes within selected operon
RBR
COG0247 Fe-SÂ
DUF3501
DPS
SH3
LysW
Hypothetical gene or gene of unknown function
B) Similarity analysis using similar description terms
Conserved protein
Conserved protein
Conserved protein
Conserved protein
C) Comparative Phylogenetics and Function Prediction Analysis
a)
b)
Metagenome database
Hub type ?

Results

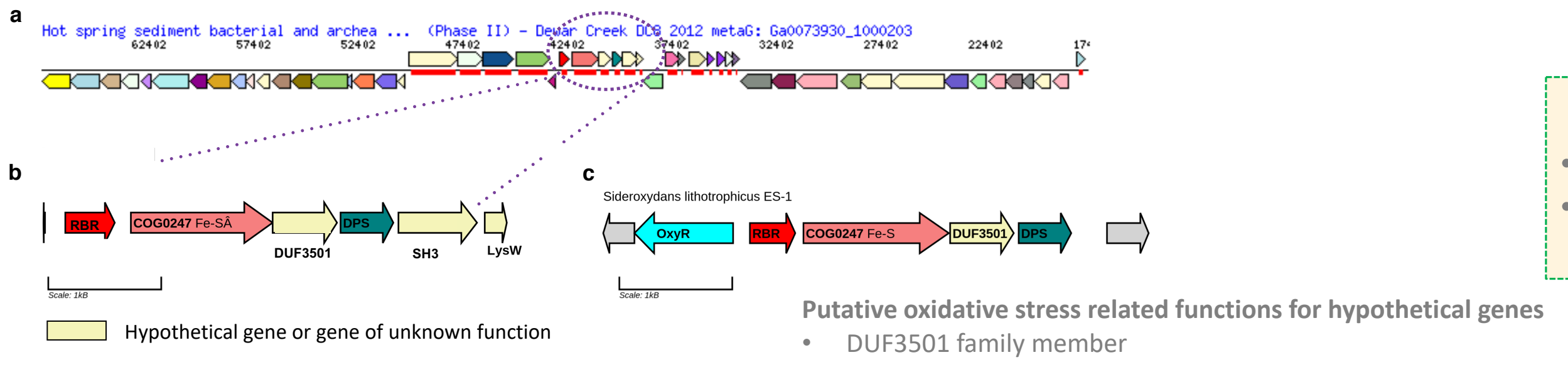
Approach successfully returns high-confidence gene-rich scaffolds for known and unknown hubs belonging to different environments and enables prediction of novel adaptation genes

Total Gene, Hypothetical Gene, and Adaptation Operon Counts

	met_hub_type	avg_gene_count	avg_percent_hyp_genes	avg_percent_adapt_genes	total_gene_count	total_hyp_count	total_adapt_count	sum_operon_counts	mean_operon_counts
1	HS_Known	238.1429	17.78571	8.585714	3334	683	254	75	2.678571
2	HS_Unknown	345.6596	17.09362	8.212766	16246	2985	1396	344	3.659574
3	HY_Known	102.9524	15.41429	8.552381	2162	335	201	52	1.238095
4	HY_Unknown	34.0000	11.80000	8.800000	34	4	3	2	1.000000
5	DS_Known	236.8462	19.77308	7.980769	6158	1191	520	174	3.346154
6	DS_Unknown	144.8750	15.92500	8.318750	2318	406	190	72	2.250000
7	PO_Known	198.6000	19.85000	8.050000	1986	367	170	56	2.800000
8	PO_Unknown	70.5000	28.80000	7.100000	141	41	8	2	0.500000

High # of hypothetical genes, adaptation genes, and adaptation-related operons were found

Functional Inference of hypothetical genes from Hot Springs scaffold suggests oxidative-stress related roles


a
Hot spring sediment bacterial and archea ... (Phase II) - Dewar Creek DC8 2012 metaG: Ga0073930_1000203
62402
57402
52402
47402
42402
37402
32402
27402
22402
b
RBR
COG0247 Fe-SÂ
DUF3501
DPS
SH3
LysW
Scale: 1kB
c
Sideroxydans lithotrophicus ES-1
OxyR
RBR
COG0247 Fe-S
DUF3501
DPS
Scale: 1kB
Hypothetical gene or gene of unknown function

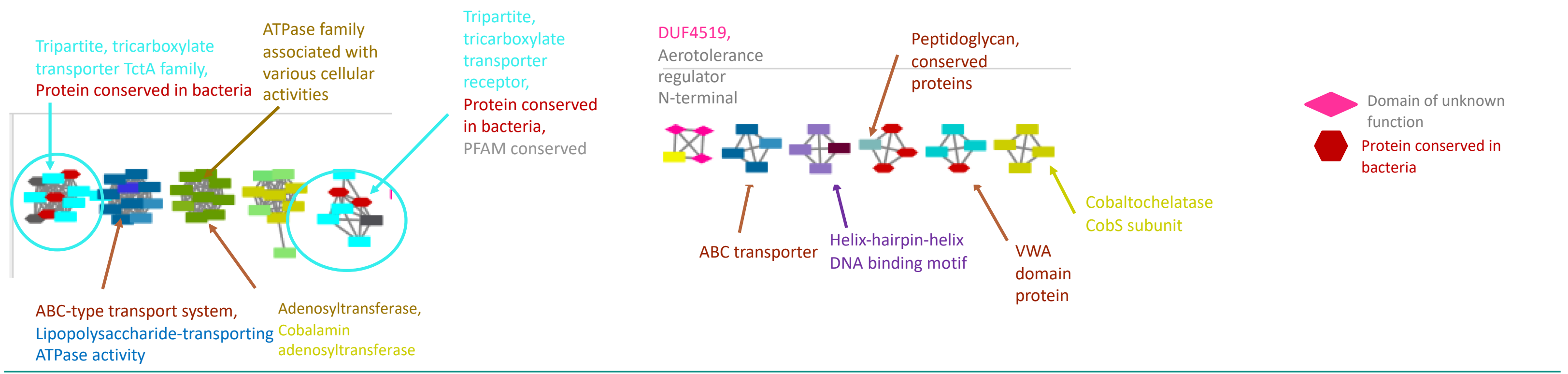
Putative oxidative stress related functions for hypothetical genes

- DUF3501 family member
- SH3-domain-containing protein
- Lysine synthesis protein W

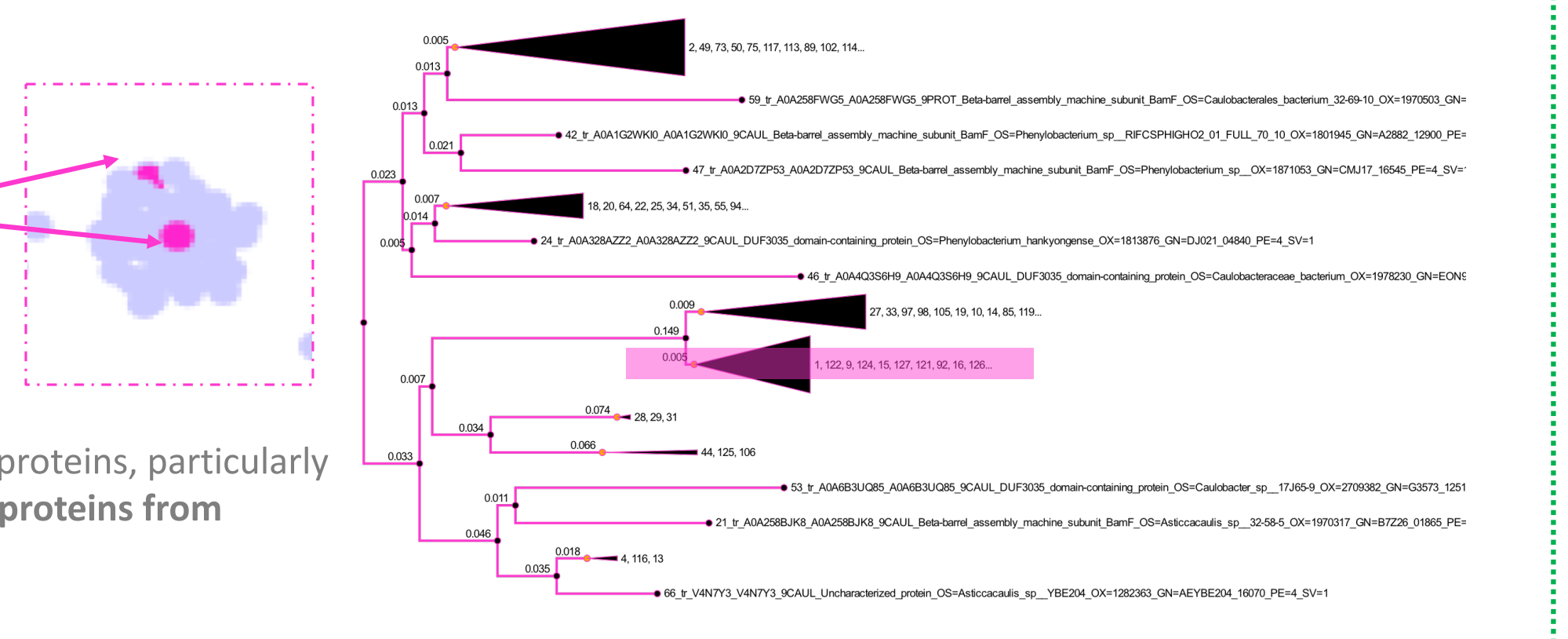
Function Prediction was possible & successful

- Examining select hypothetical genes in scaffolds
- All (~2600) Deep Sea and Hot Springs hypothetical genes by SSN/ all-vs-all BLAST

Sequence Similarity Network (SNN) of All-vs-All BLAST of hypothetical proteins


Tripartite, tricarboxylate transporter TctA family,
Protein conserved in bacteria
ATPase family associated with various cellular activities
Tripartite, tricarboxylate transporter receptor,
Protein conserved in bacteria,
PFAM conserved
ABC-type transport system,
Lipopolysaccharide-transporting ATPase activity
Adenosyltransferase,
Cobalamin adenosyltransferase
DUF4519,
Aerotolerance regulator N-terminal
Peptidoglycan, conserved proteins
ABC transporter
Helix-hairpin-helix DNA binding motif
VWA domain protein
Cobaltochelatase CobS subunit
Domain of unknown function
Protein conserved in bacteria

SSN of All-vs-All BLAST of hypothetical/UNIPROT DUF3035 genes


Hypothetical DUF3035 proteins
All DUF3035 proteins from UNIPROT
0.005
2, 49, 73, 50, 75, 117, 113, 89, 102, 114...
0.013
59_tr_A0A258FWG5_A0A258FWG5_9PROT_Beta-barrel_assembly_machine_subunit_BamF_OS=Caulobacterales_bacterium_32-69-10_OX=1970503_GN=
0.013
0.021
42_tr_A0A1G2WKI0_A0A1G2WKI0_9CAUL_Beta-barrel_assembly_machine_subunit_BamF_OS=Phenylobacterium_sp._RIFCSPHIGHO2_01_FULL_70_10_OX=1801945_GN=A2882_12900_PE=
47_tr_A0A2D7ZP53_A0A2D7ZP53_9CAUL_Beta-barrel_assembly_machine_subunit_BamF_OS=Phenylobacterium_sp._OX=1871053_GN=CMJ17_16545_PE=4_SV=
0.023
0.007
18, 20, 64, 22, 25, 34, 51, 35, 55, 94...
0.014
24_tr_A0A328AZZ2_A0A328AZZ2_9CAUL_DUF3035_domain-containing_protein_OS=Phenylobacterium_hankyongense_OX=1813876_GN=DJ021_04840_PE=4_SV=1
0.005
46_tr_A0A4Q3S6H9_A0A4Q3S6H9_9CAUL_DUF3035_domain-containing_protein_OS=Caulobacteraceae_bacterium_OX=1978230_GN=EON
0.009
27, 33, 97, 98, 105, 19, 10, 14, 85, 119...
0.149
0.005
1, 122, 9, 124, 15, 127, 121, 92, 16, 126...
0.007
0.074
28, 29, 31
0.034
0.066
44, 125, 106
0.033
0.011
53_tr_A0A6B3UQ85_A0A6B3UQ85_9CAUL_DUF3035_domain-containing_protein_OS=Caulobacter_sp._17J65-9_OX=2709382_GN=G3573_1251
21_tr_A0A258BJK8_A0A258BJK8_9CAUL_Beta-barrel_assembly_machine_subunit_BamF_OS=Asticcacaulis_sp._32-58-5_OX=1970317_GN=B7Z26_01865_PE=
0.046
0.018
4, 116, 13
0.035
66_tr_V4N7Y3_V4N7Y3_9CAUL_Uncharacterized_protein_OS=Asticcacaulis_sp._YBE204_OX=1282363_GN=AEYBE204_16070_PE=4_SV=1

Hypothetical DUF3035 proteins cluster with **high similarity/phylogeny** to known DUF3035 proteins, particularly **Beta-barrel assembly machine subunit BamF proteins from Brevundimonas sp**

Conclusions

- **Our approach enables identification and prediction of novel genes**
- Our pipeline **accommodates the computational power and memory required of large-scale datasets**
- Function prediction of select operons shows conserved roles and similar functional categories among neighboring genes
- Large-scale function prediction of hypothetical genes via SSN shows high similarity and potential conservation among genes with similar Gene Ontology annotations, even among genes originating from different environmental metagenomes.

Future Work

Similarity analysis of hypothetical genes and adaptation operons may help identify novel but well-conserved survival/ stress genes

Acknowledgements

This study was funded in part by National Aeronautics and Space Administration through the University of Central Florida's NASA Florida Space Grant Consortium and Space Florida (NX15AI10H). This work was also funded in part by the NASA Exobiology program (80NSSC20K0612) awarded to AC.

References

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., … Kyrpides, N. C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research*, *47*(D1), D666–D677. https://doi.org/10.1093/nar/gky901

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610