

# NaPDoS2: A versatile webtool for assessing biosynthetic diversity



Hans Wu Singh

03/13/2025

BB4F

R/V *Sally Ride*, Anacortes , Washington. Credit: Jeff Dillon.

# Talk outline

- **Introduction**
- **NaPDoS2: Genome mining webtool**
- **Metagenomic data reveals type I polyketide synthase distributions across biomes**
- **KS amplicons highlight polyketide biosynthetic potential across abyssal sediments**

# Talk outline

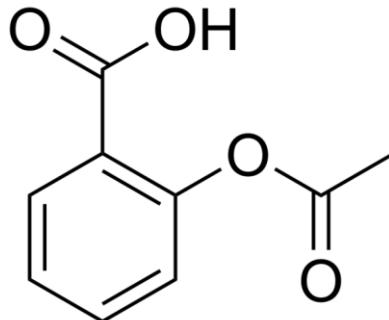
- **Introduction**
- **NaPDoS2: Genome mining webtool**
- **Metagenomic data reveals type I polyketide synthase distributions across biomes**
- **KS amplicons highlight polyketide biosynthetic potential across abyssal sediments**

# Natural products are crucial therapeutics



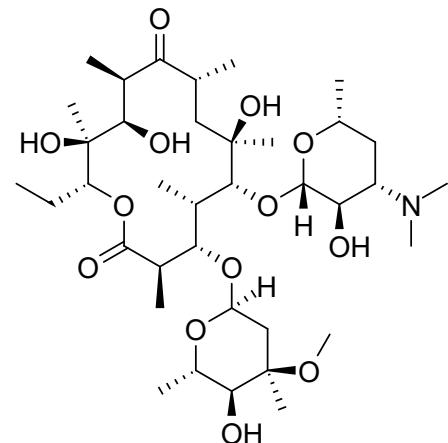
# Salicin > Aspirin

## Anti-inflammatory, willow tree



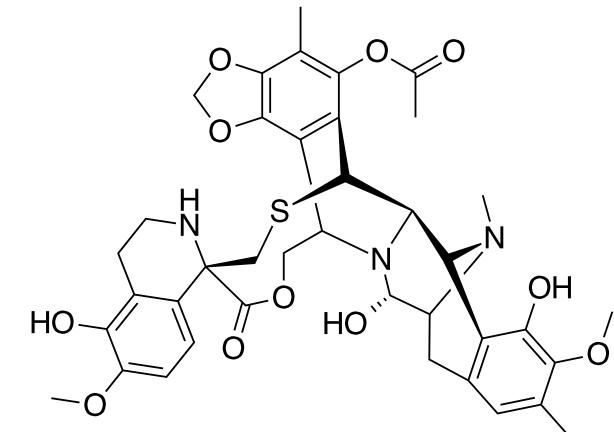
# Erythromycin

## Antibiotic, bacteria



## **Ecteinascidin-743 > Trabectedin**

### **Anticancer, tunicate**



# Marine bacteria produces notable therapeutics



*Salinispora*

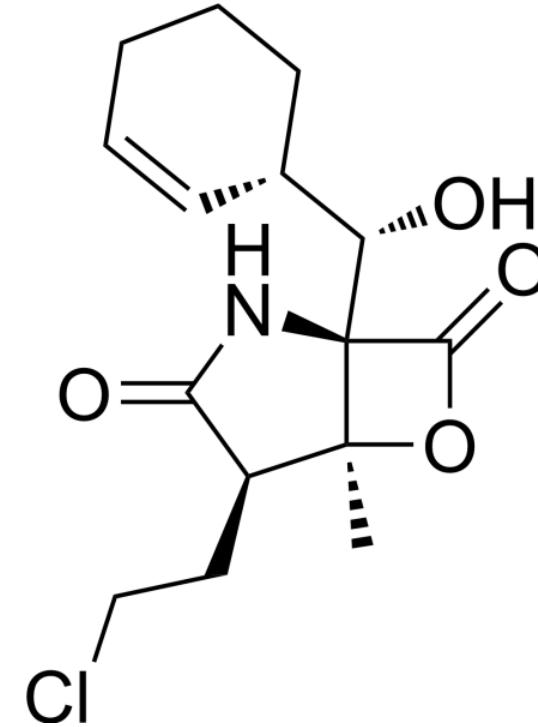
Marine actinomycete genus

# Marine bacteria produces notable therapeutics



*Salinispora*

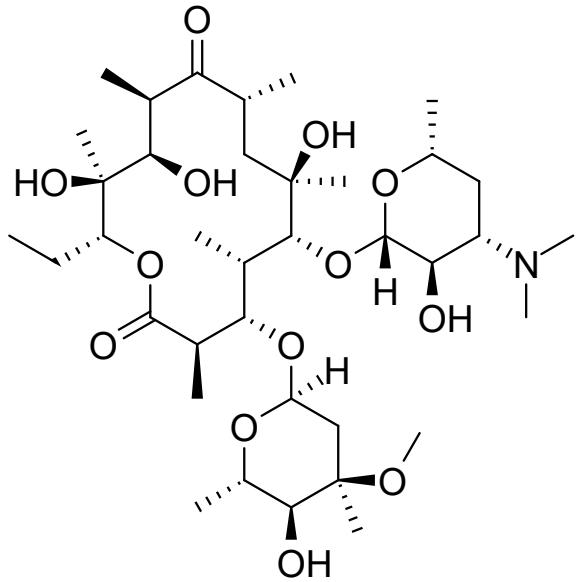
Marine actinomycete genus



Salinosporamide A

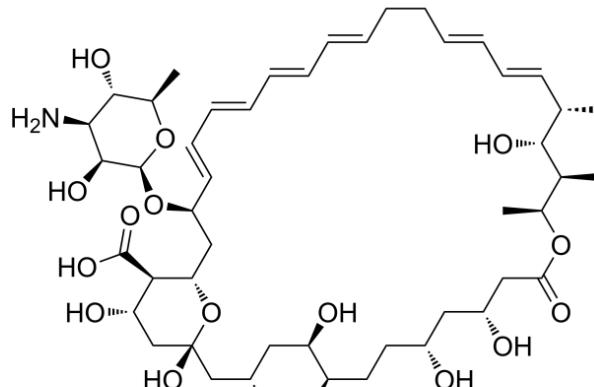
Can cross the blood-brain barrier

# Polyketides a large, diverse class of natural products



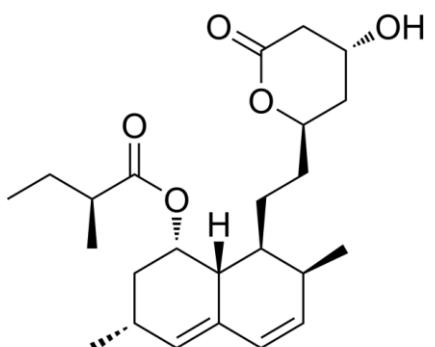
Erythromycin (antibiotic)

Producer: *Saccharopolyspora*



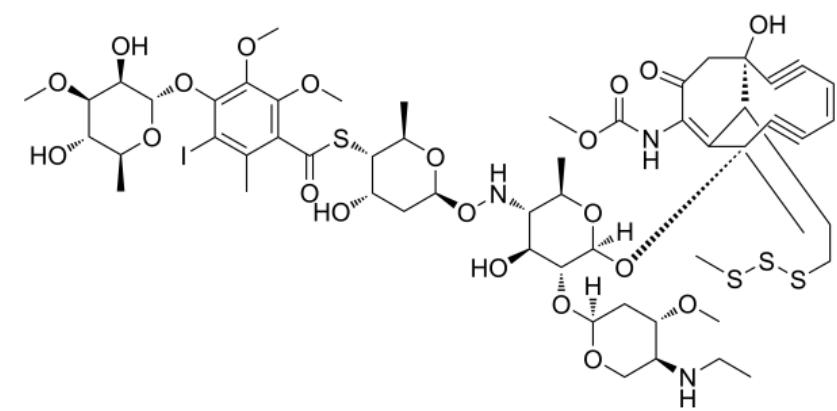
Amphotericin (antifungal)

Producer: *Streptomyces*



Lovastatin (lowers cholesterol)

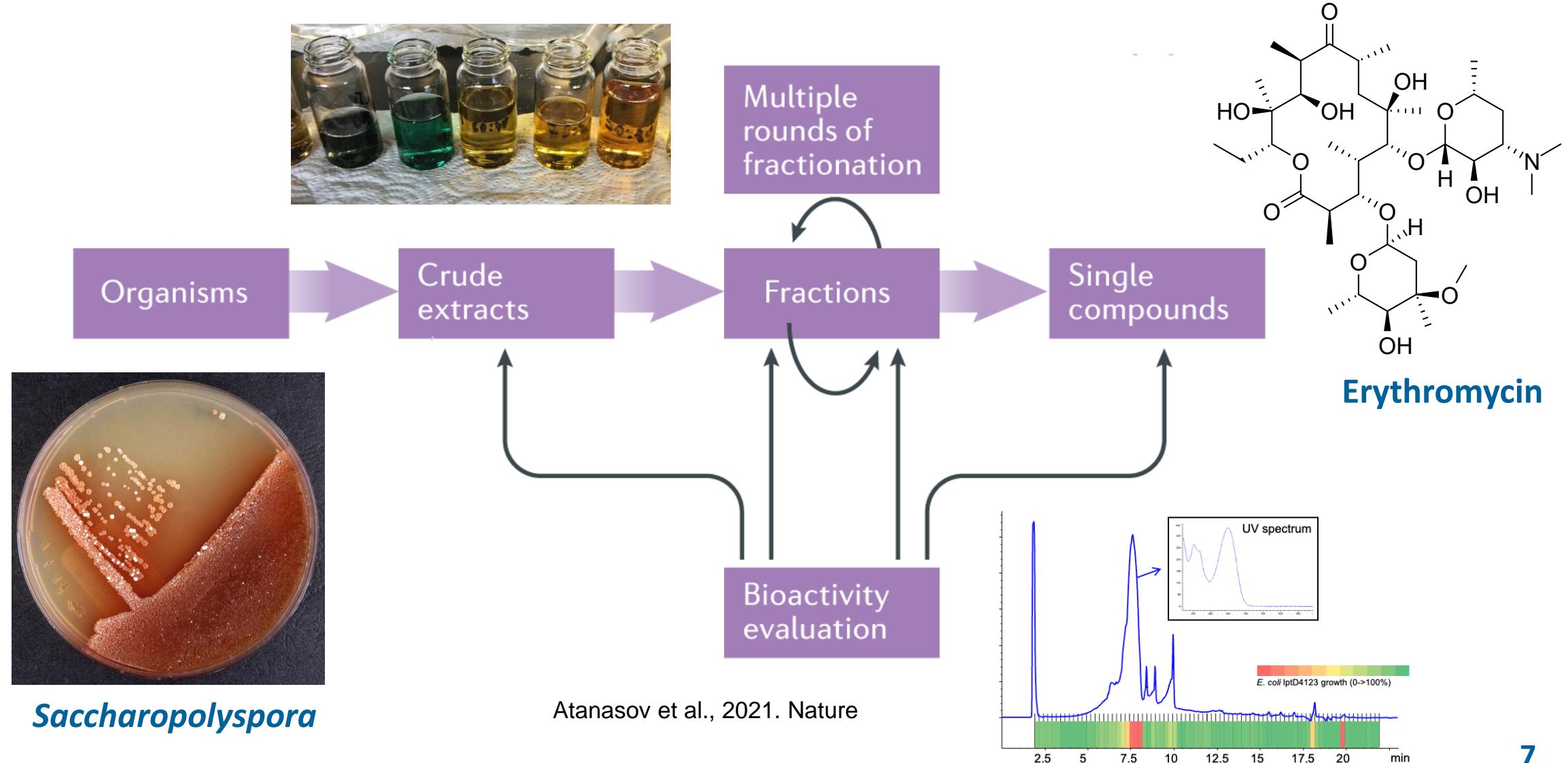
Producer: *Aspergillus*



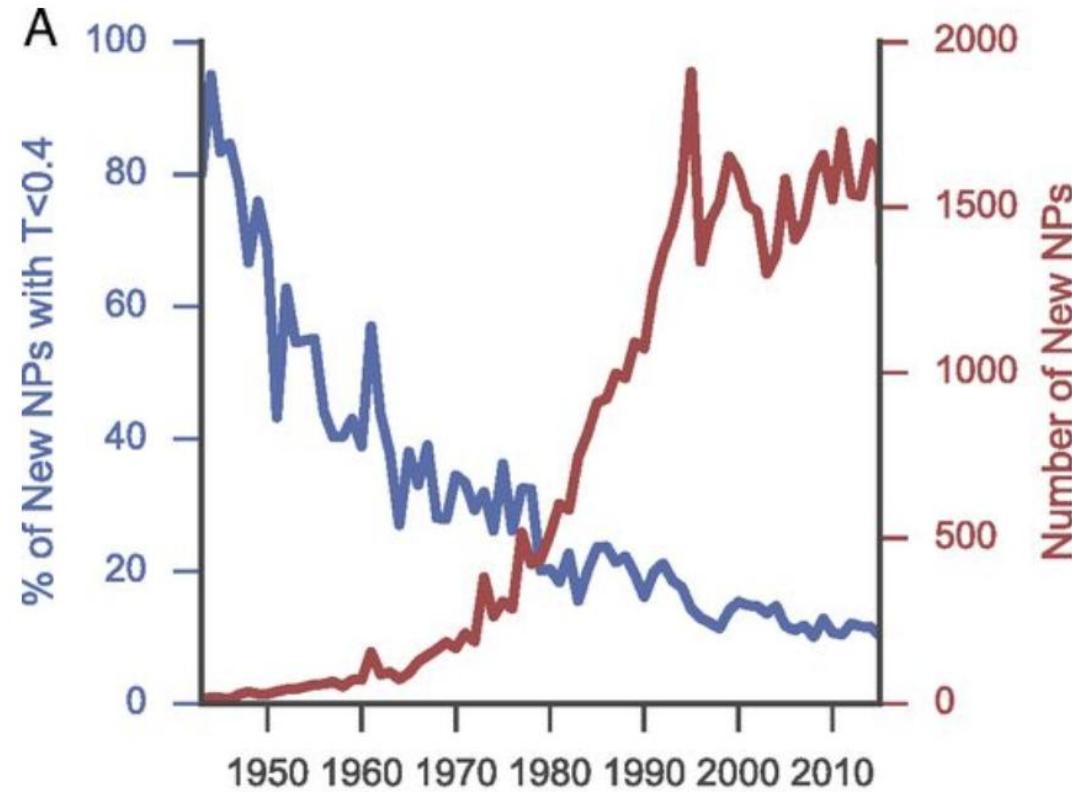
Calicheamicin (antitumor)

Producer: *Micromonospora*

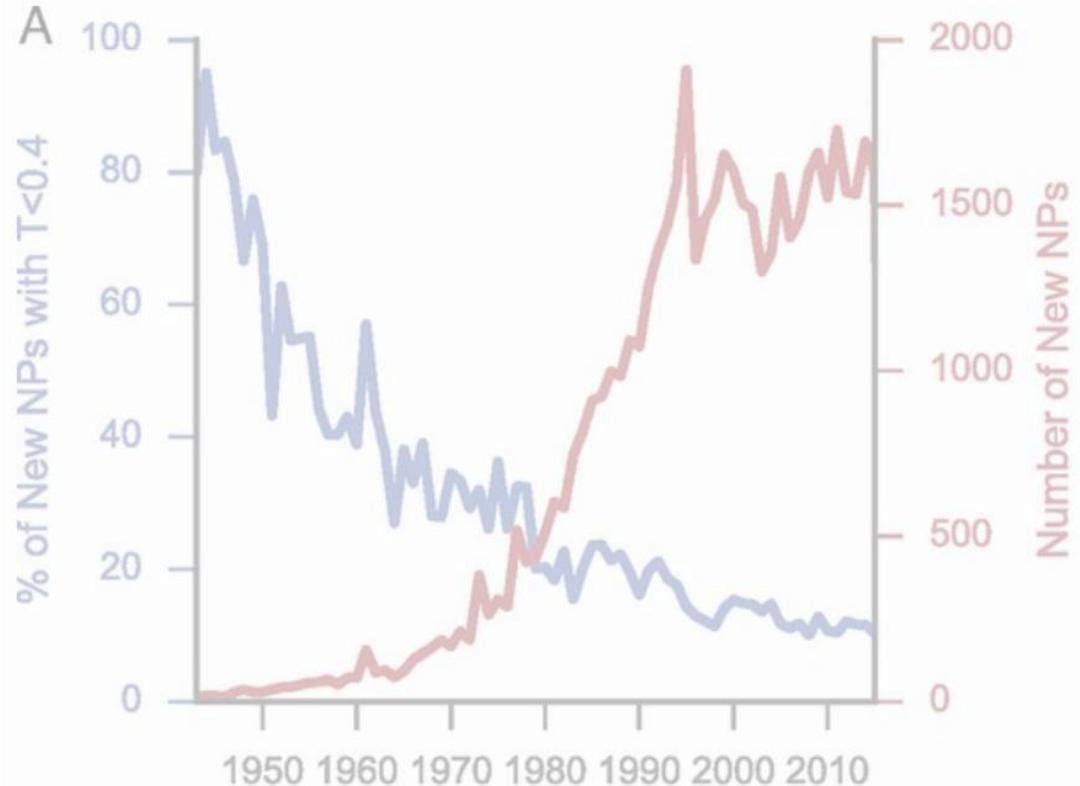
# Traditional natural products discovery pipeline



# Natural product isolation challenged by high rediscovery rates

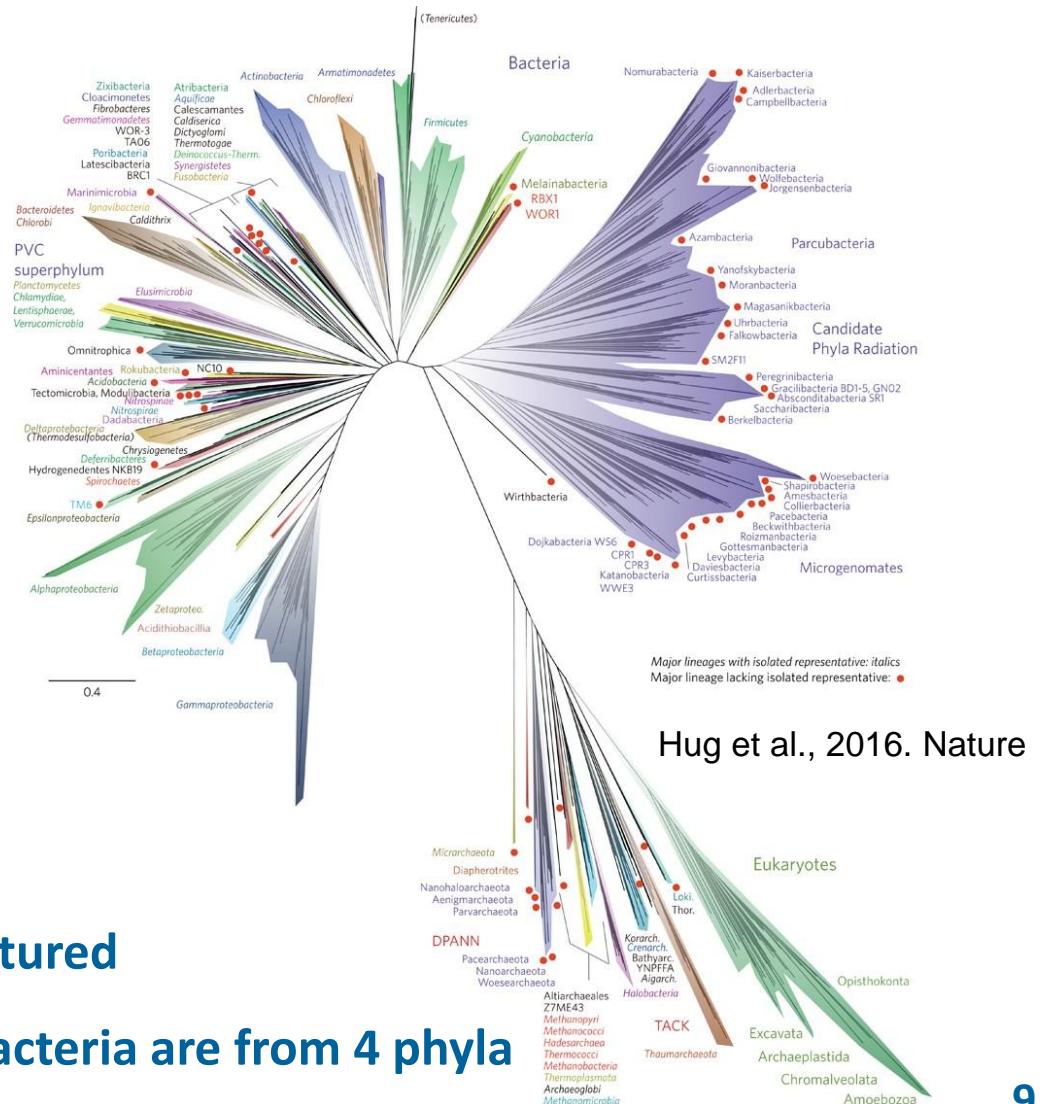


# Majority of bacteria remain uncultured



# <4% of species cultured

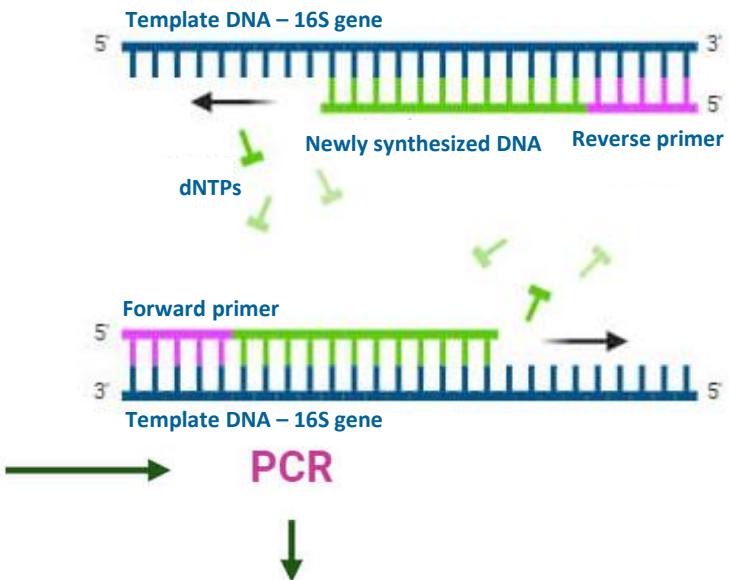
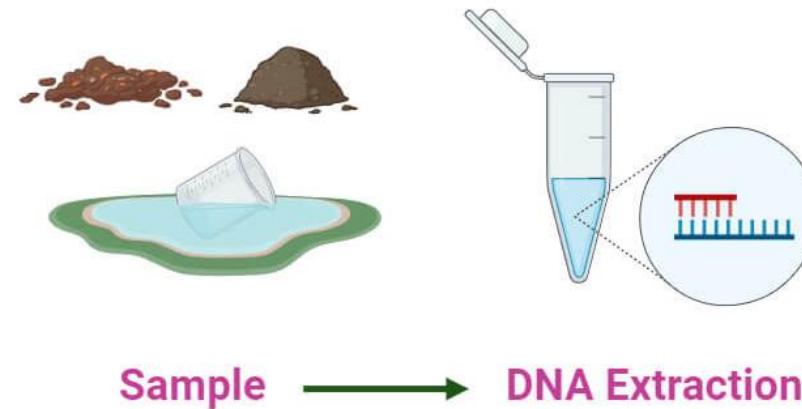
**97% of cultured bacteria are from 4 phyla**



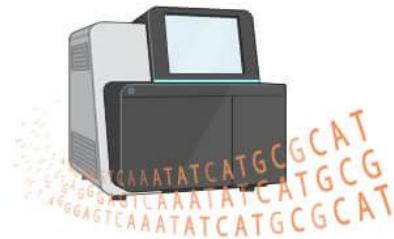
Pye et al., 2017. PNAS

# Amplicon sequencing: Culture-independent analysis of bacterial diversity

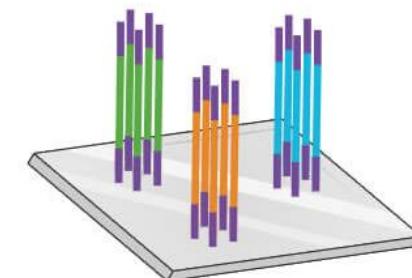
## 16S rRNA Gene Sequencing



Data Analysis



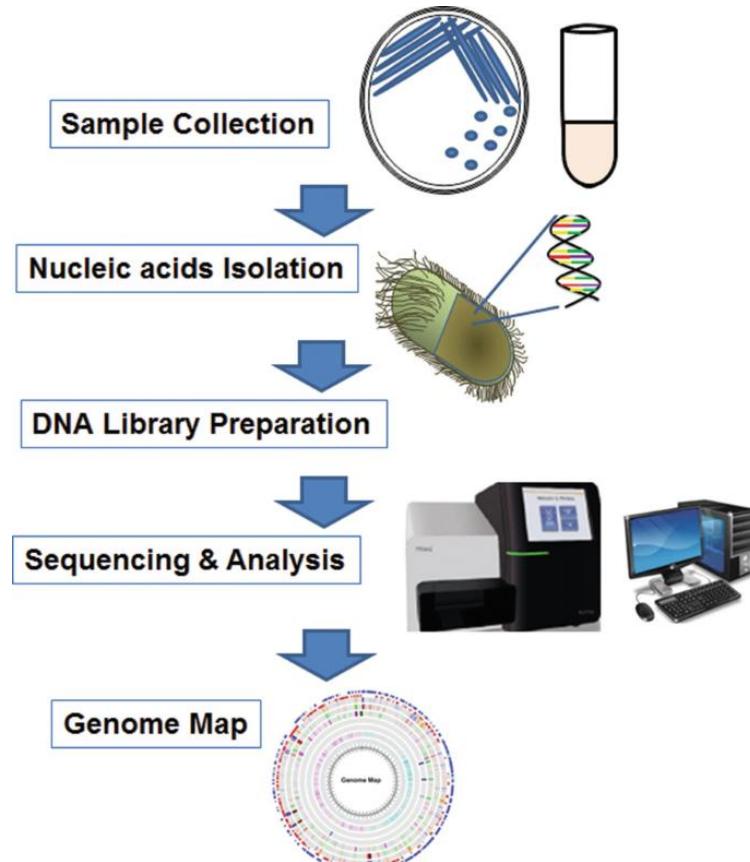
DNA Sequencing



DNA Library Construction

# Genome sequencing gives larger microbial context

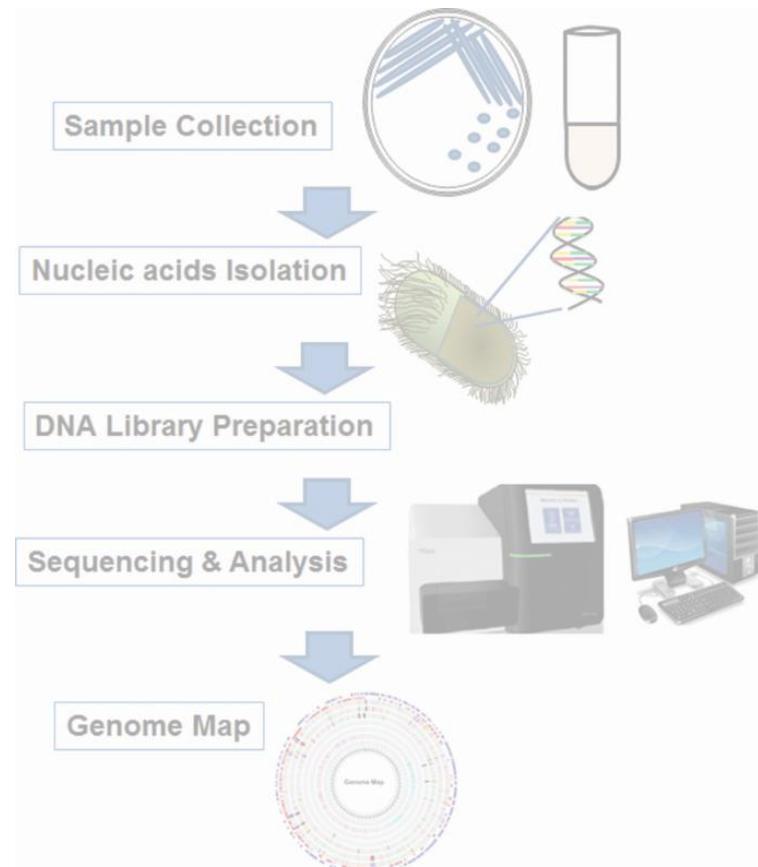
## Sequencing cultured isolates



Srivastav et al., 2019

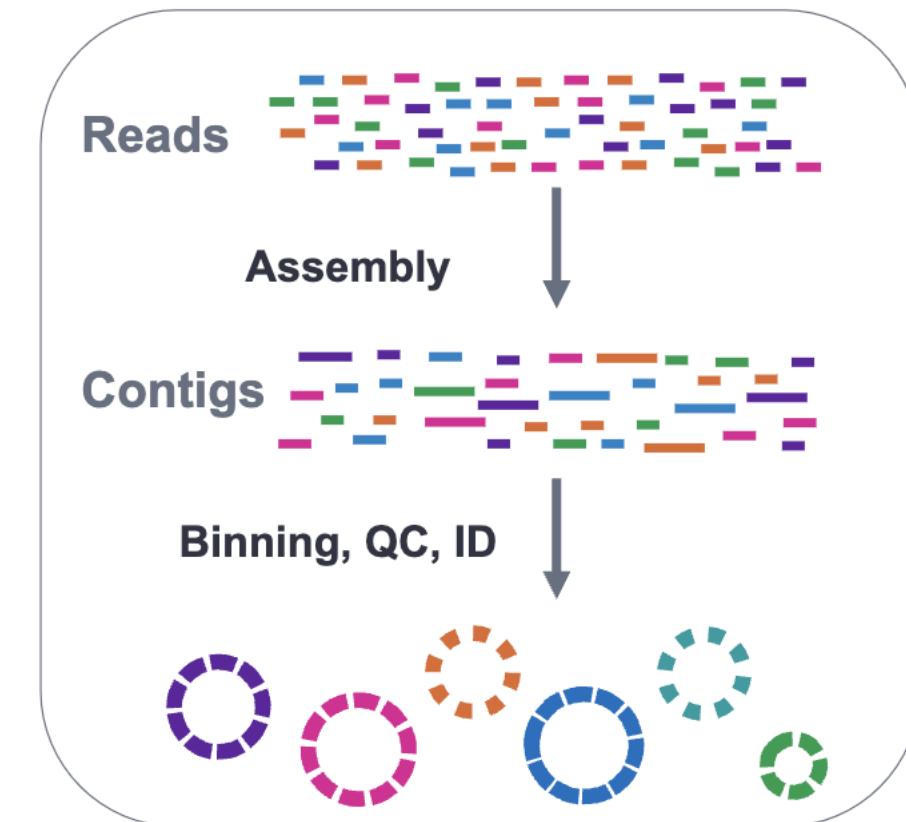
# Metagenomic sequencing allows access to unculturable majority

## Sequencing cultured isolates



Srivastav et al., 2019

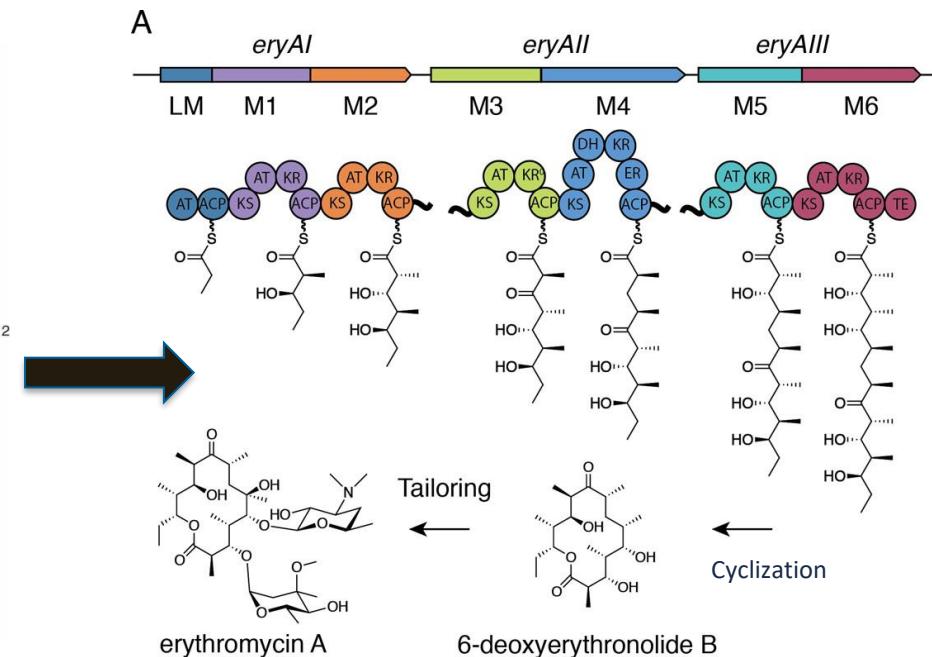
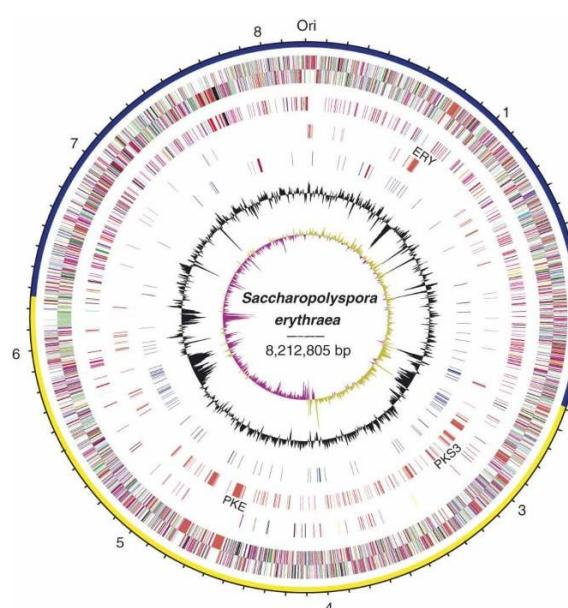
## Metagenomic sequencing



PacBio, 2024

Draft-quality MAGs

# Genome mining for natural product discovery



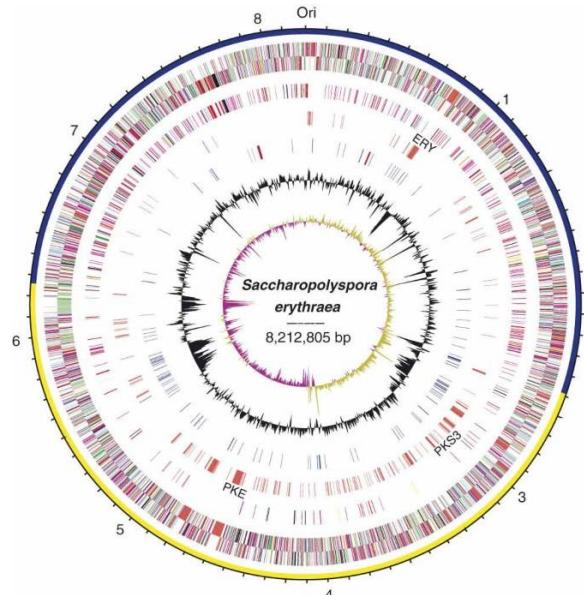
*Saccharopolyspora* genome

Erythromycin polyketide synthase (PKS)  
biosynthetic gene cluster (BGC)

Oliynyk et al., 2007. Nature

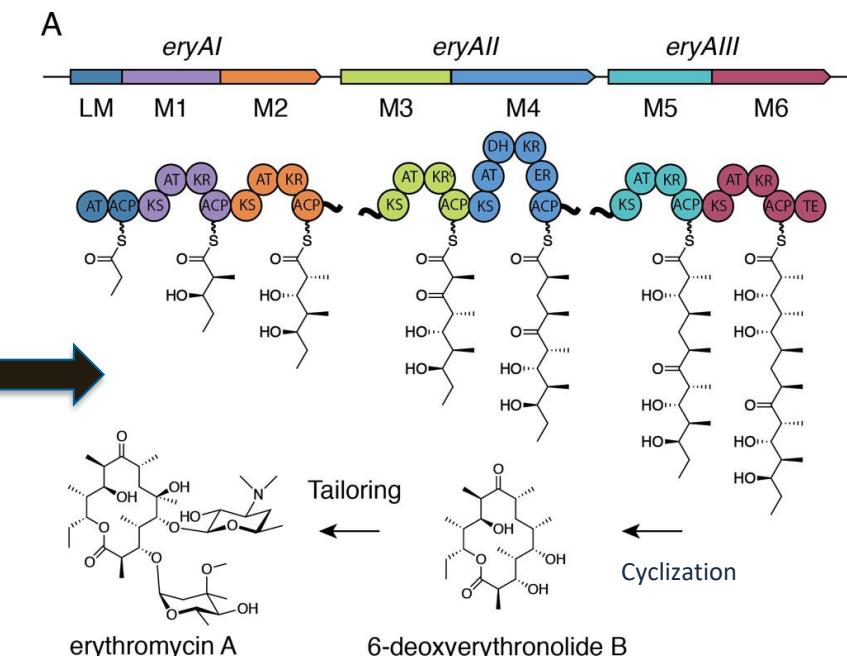
Nivina et al., 2019. ACS

# Genome mining for natural product discovery



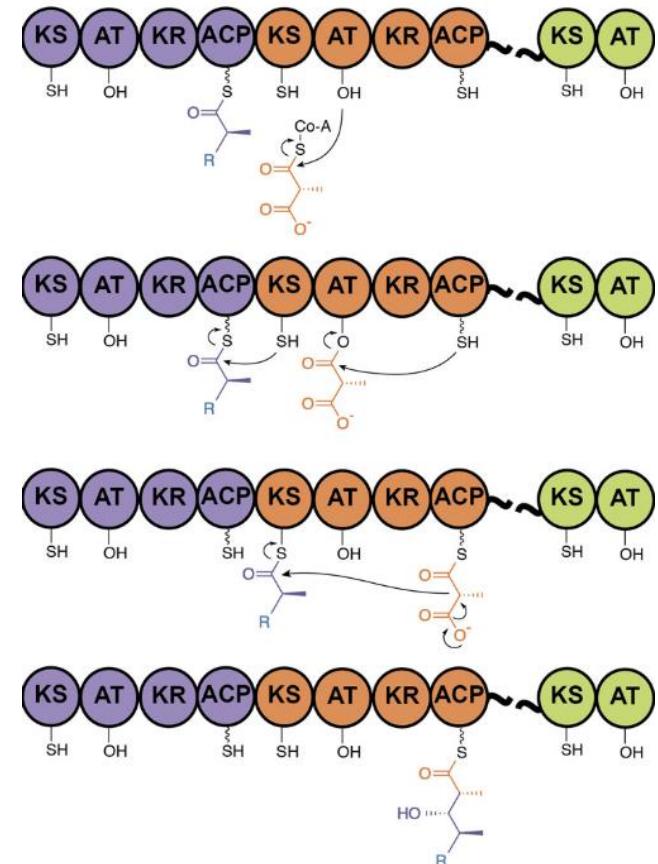
*Saccharopolyspora* genome

Oliynyk et al., 2007. Nature



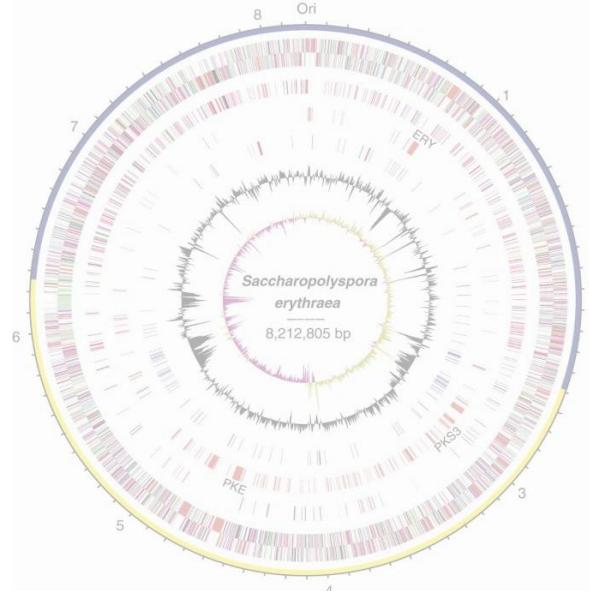
Erythromycin polyketide synthase (PKS)  
biosynthetic gene cluster (BGC)

Nivina et al., 2019. ACS

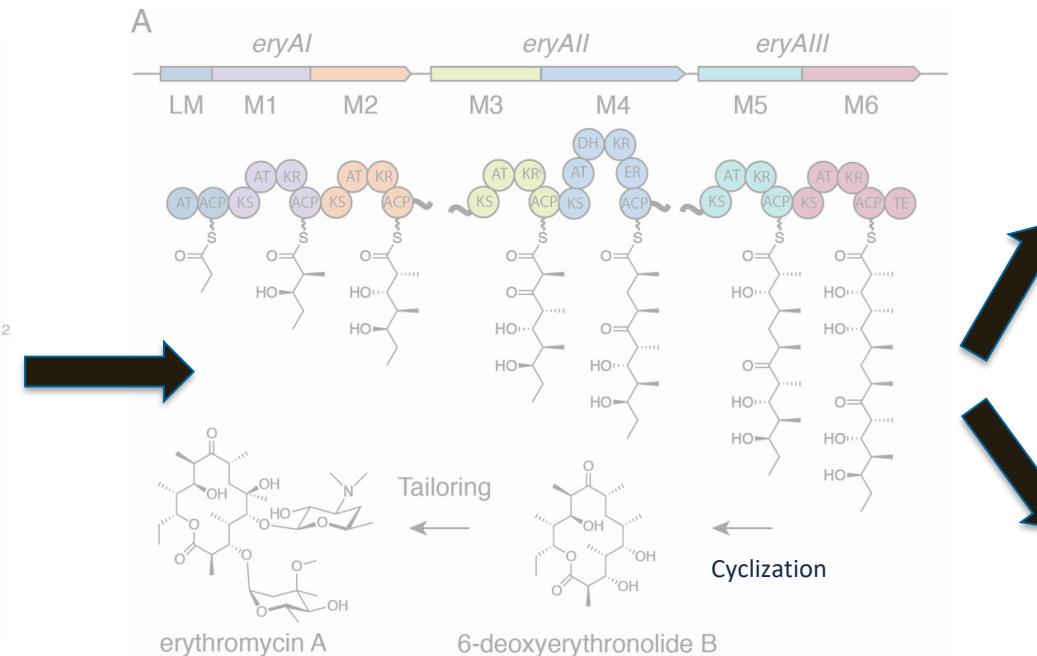


KS domain: catalyzes  
C-C bond formation

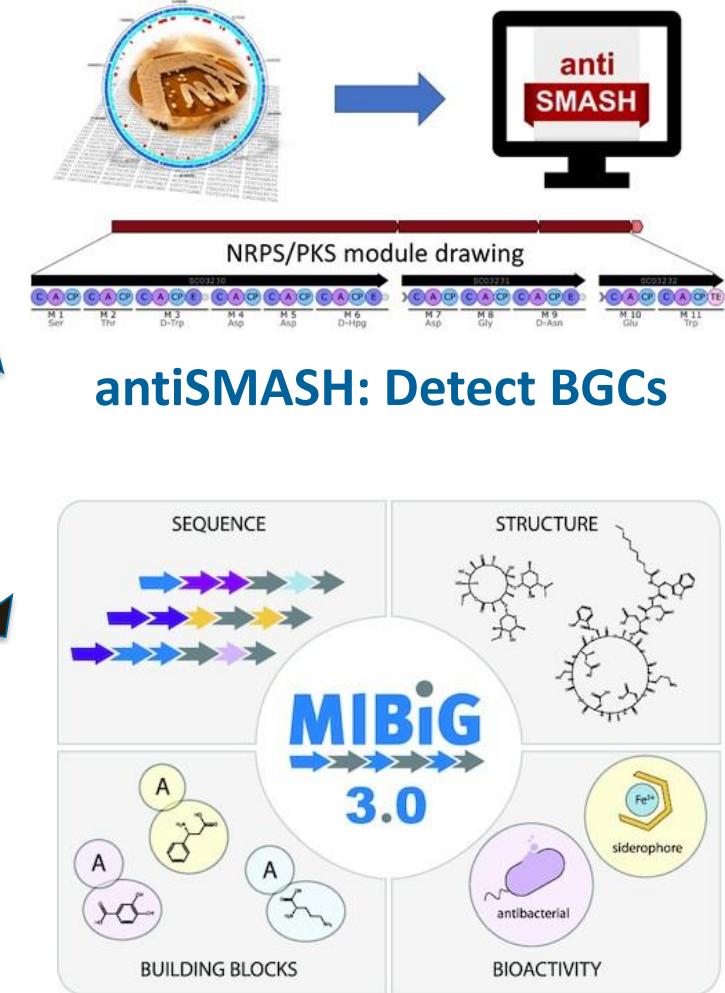
# Genome mining for natural product discovery



*Saccharopolyspora* genome

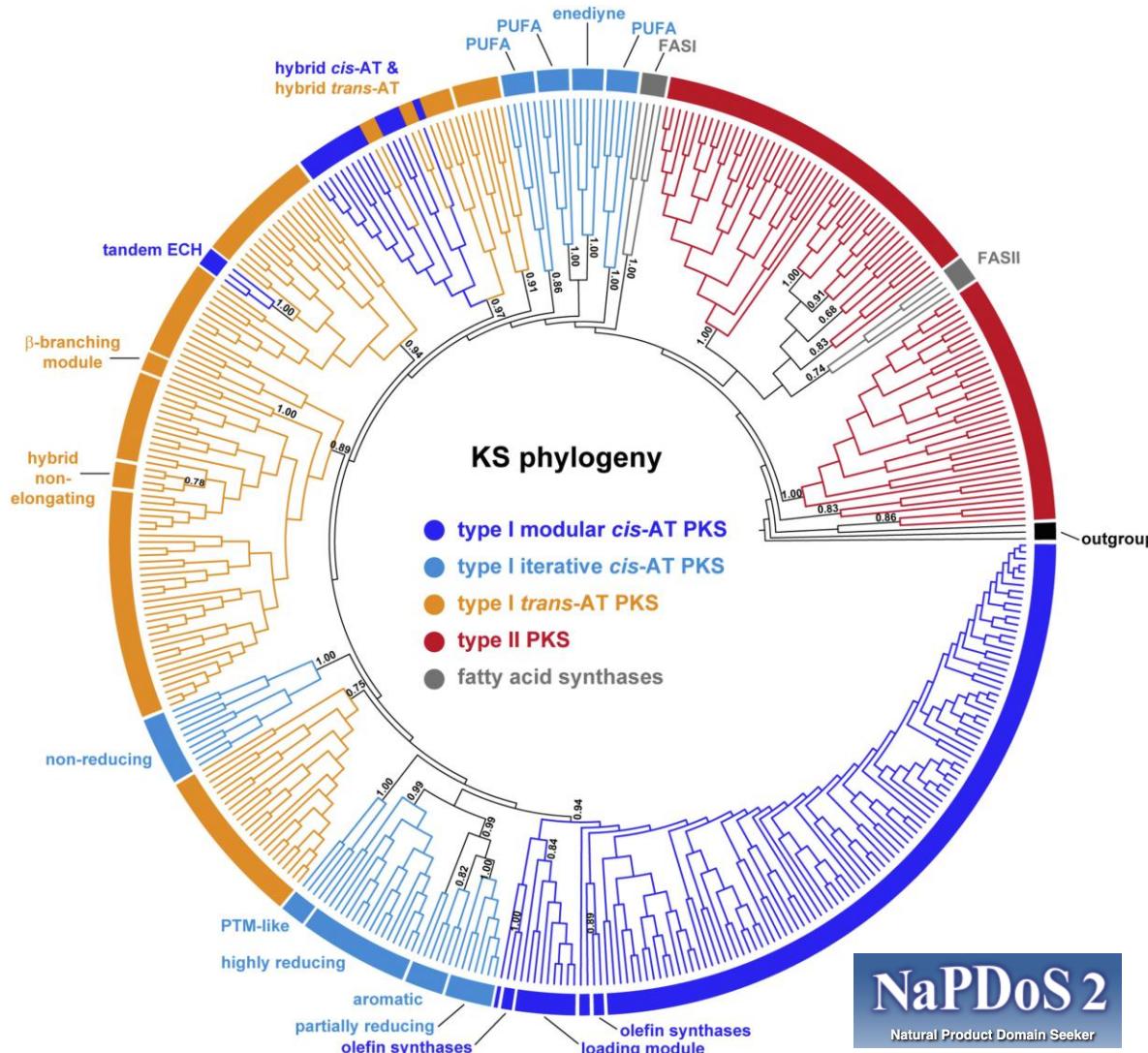


Erythromycin polyketide synthase (PKS)  
biosynthetic gene cluster (BGC)



**MiBiG:** BGC repository

# NaPDoS2: Webtool for polyketide genome mining



Phylogenetic conservation of KS domains

Classifies KS domains into 40+ subclasses

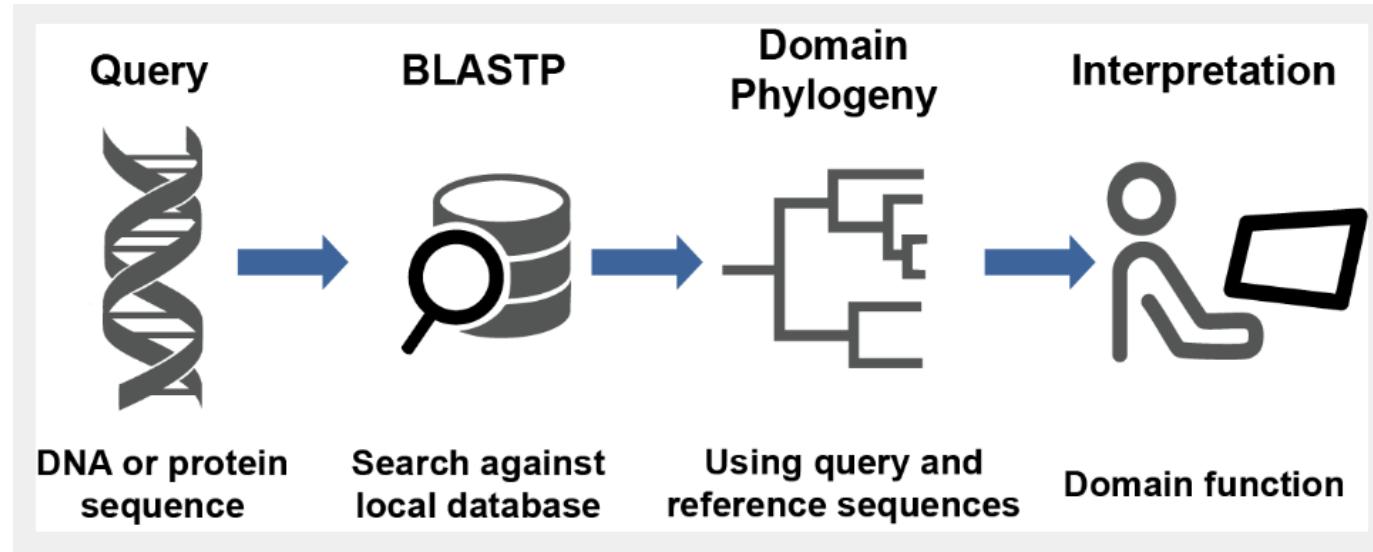
Excellent for short sequences

Klau, Podell, Creamer, Demko, Singh, Allen, Moore, Ziemert, Letzel, Jensen. 2022. Journal of Biological Chemistry

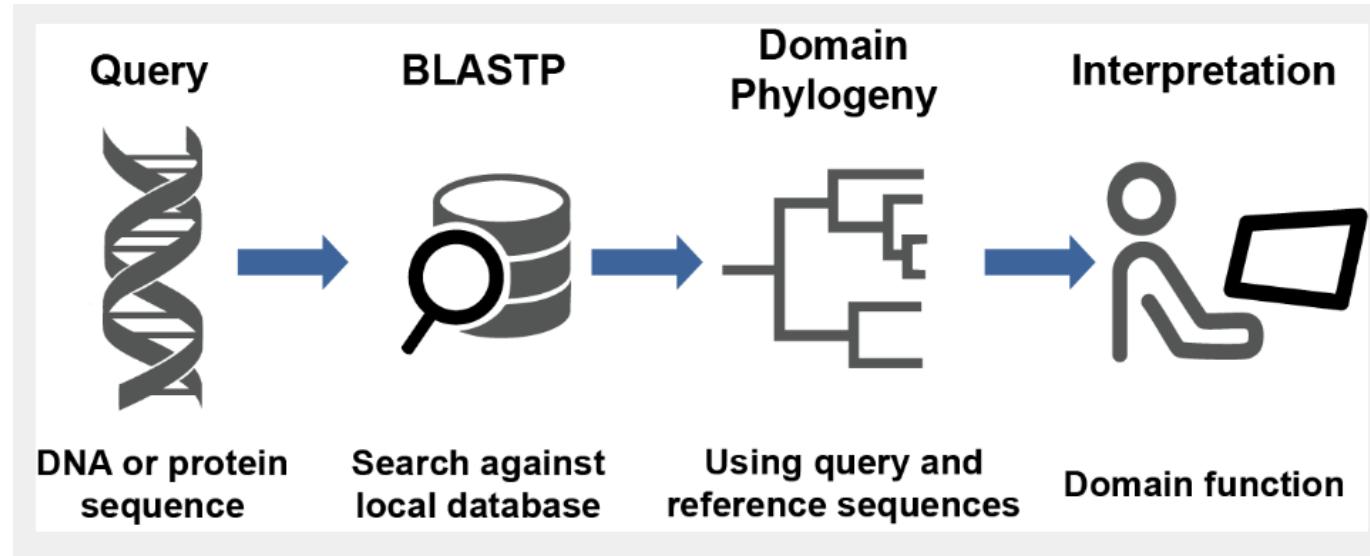
# Talk outline

- Introduction
- **NaPDoS2: Genome mining webtool**
- Metagenomic data reveals type I polyketide synthase distributions across biomes
- KS amplicons highlight polyketide biosynthetic potential across abyssal sediments

# NaPDoS2 overview



# NaPDoS2 overview

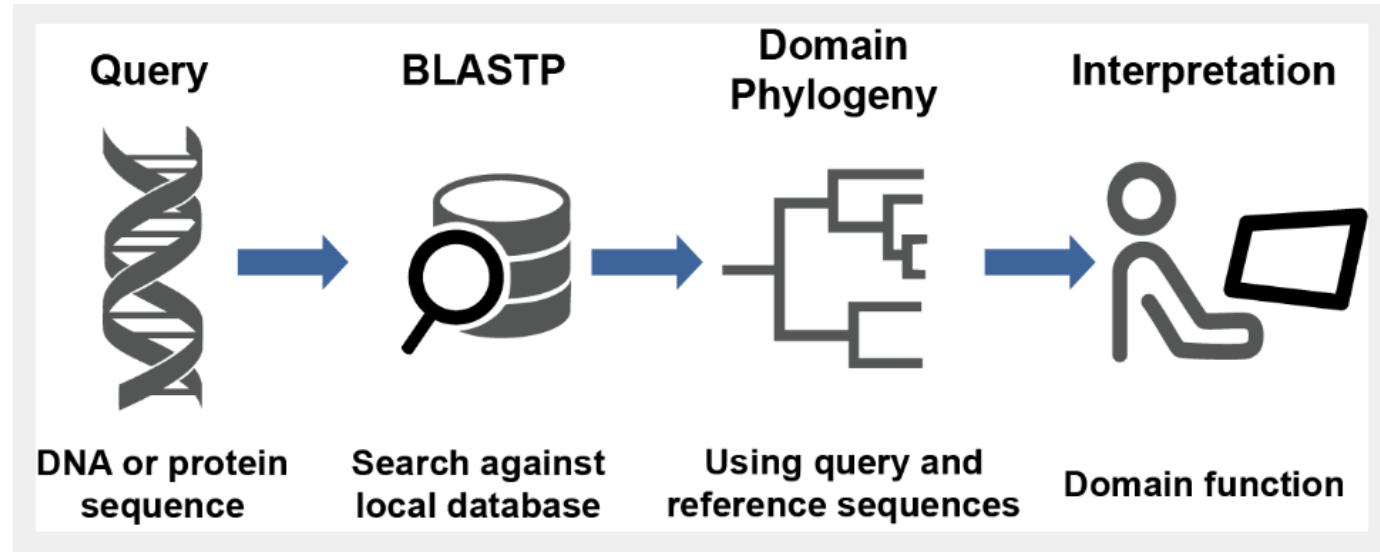


- Speed

- Specificity

- Size

# NaPDoS2 overview



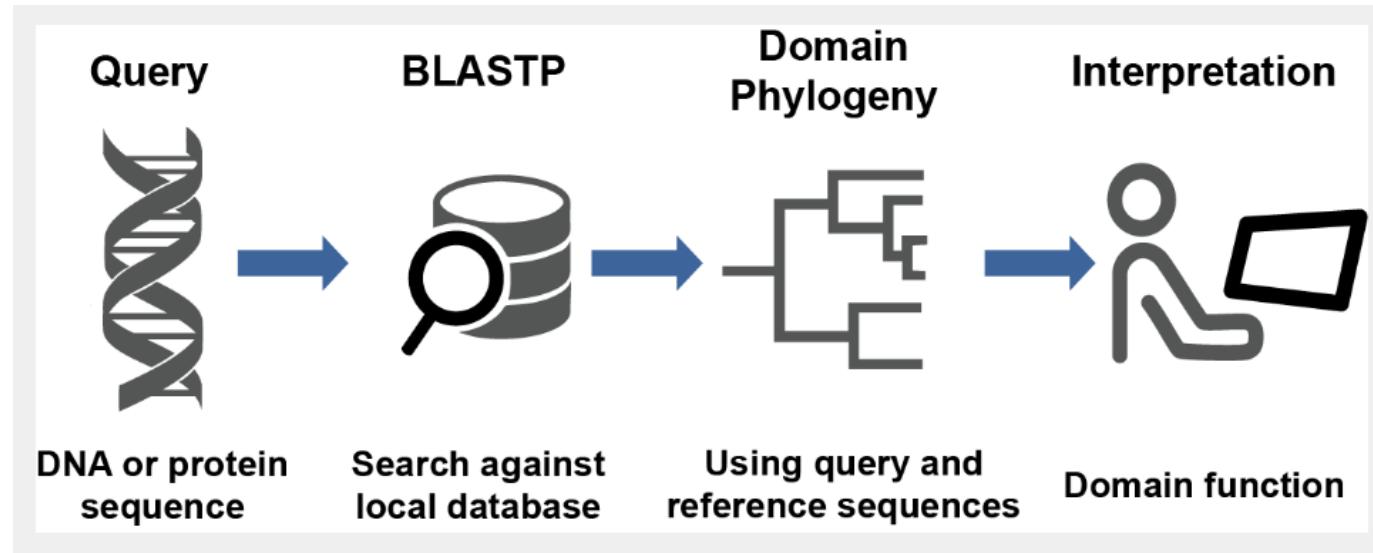
- Speed

- 10x faster

- Specificity

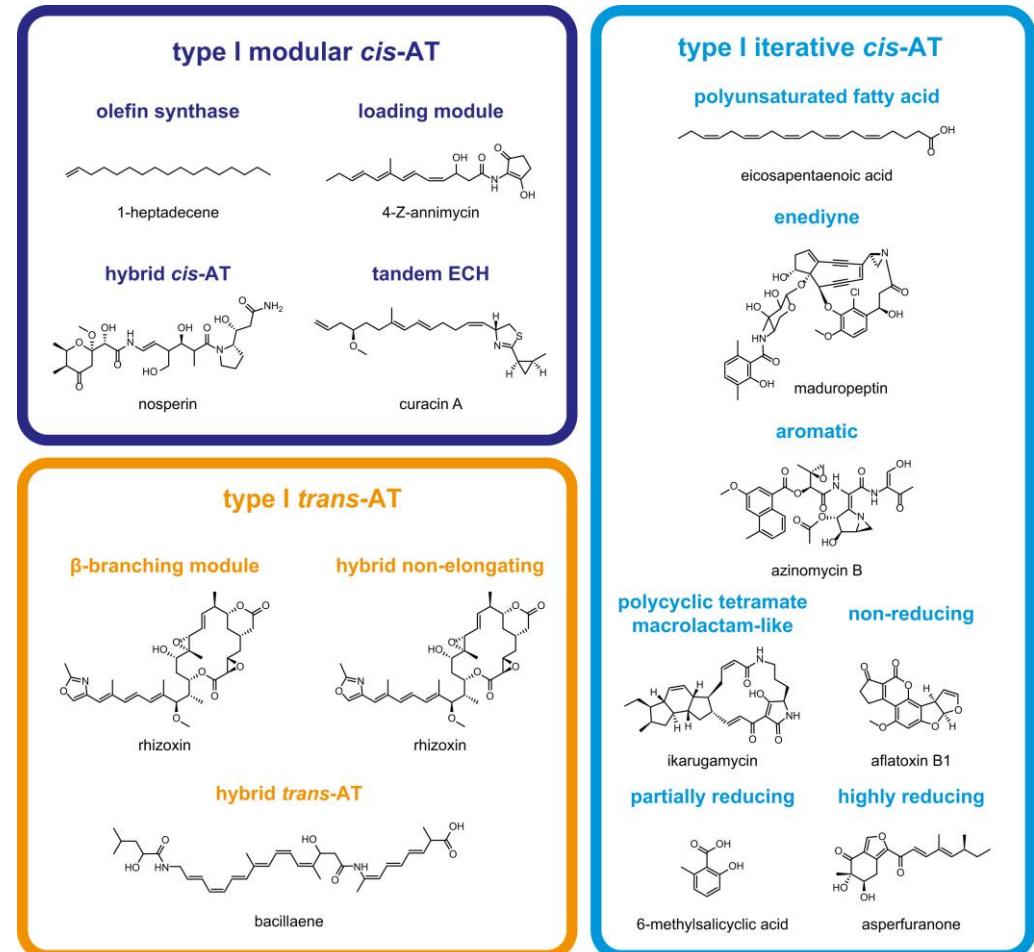
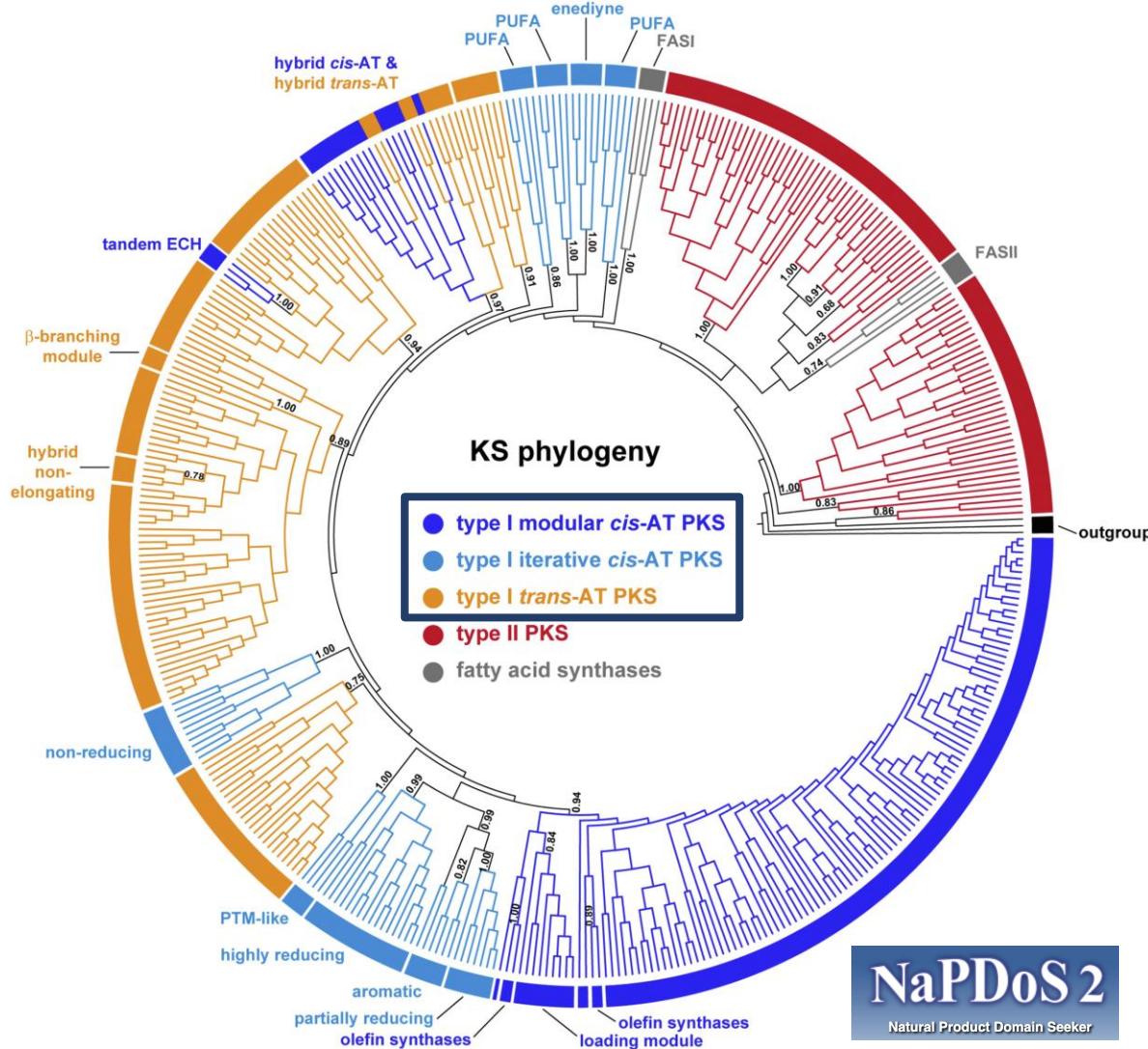
- Size

# NaPDoS2 overview



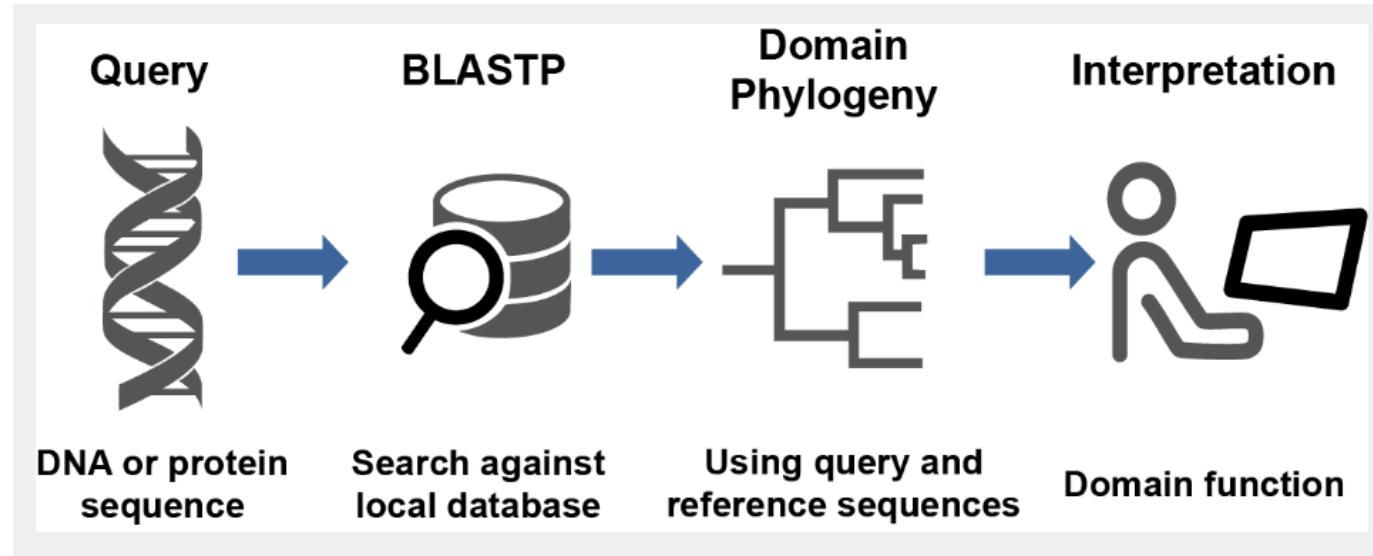
- Speed
- 10x faster
- Specificity
- 40+ KS subclasses
- Size

# NaPDoS2: Webtool for polyketide genome mining



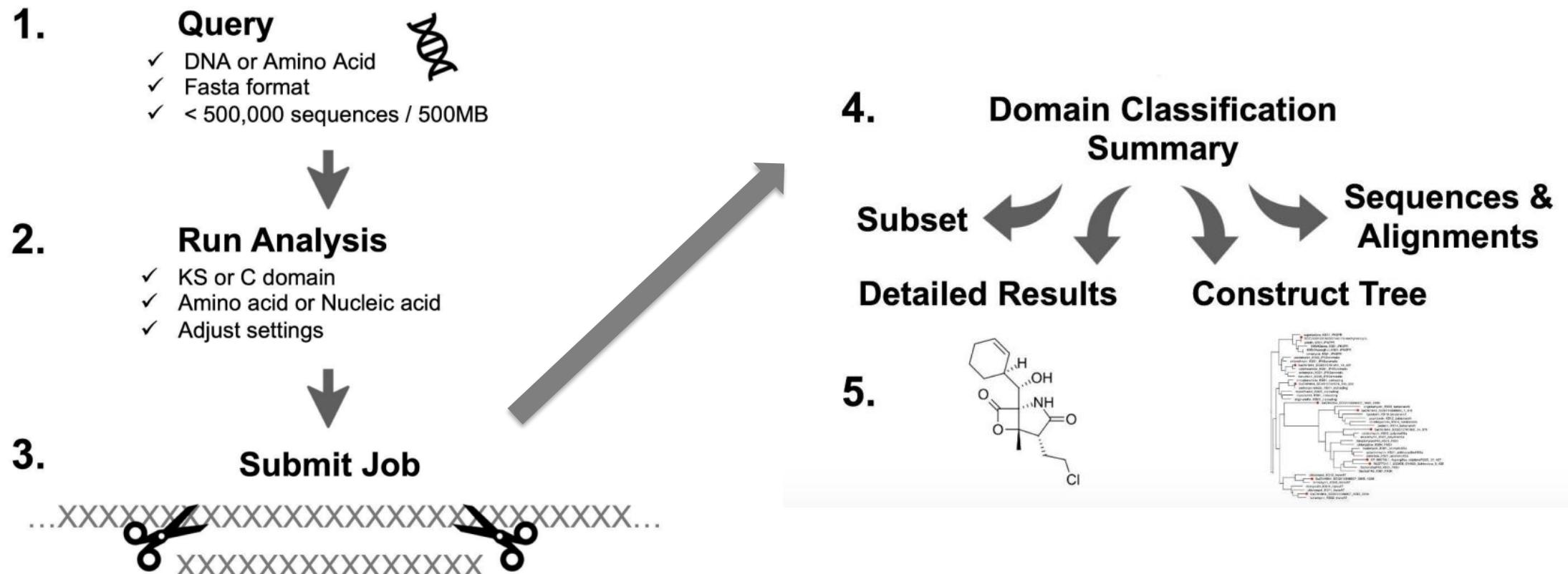
Klau, Podell, Creamer, Demko, Singh, Allen, Moore, Ziemert, Letzel, Jensen. 2022. Journal of Biological Chemistry

# NaPDoS2 overview



- Speed
- 10x faster
- Specificity
- 40+ KS subclasses
- Size
- Can handle amplicons

# NaPDoS2 detailed overview



# Finding NaPDoS2

Google napdos2

All Images Videos Shopping Short videos News Forums More Tools

**NaPDoS**  
https://npdomainseeker.sdsc.edu

**NaPDoS | Home**  
Natural Product Domain Seeker NaPDoS is a bioinformatic tool for the rapid detection and analysis of secondary metabolite genes.  
[Run Analysis](#) · [Overview](#) · [Tutorial](#) · [Pathways](#)

**JBC** Journal of Biological Chemistry  
https://www.jbc.org › article › fulltext

**The Natural Product Domain Seeker version 2 (NaPDoS2) ...**  
by LJ Klau · 2022 · Cited by 47 — NaPDoS is particularly useful for the analysis of incomplete biosynthetic genes or gene clusters, as are often observed in poorly assembled ...

Images :

Omics Data Platform (60) and down...  
Amplicons. KS amplicon seq lengths of 50 aa in NaPDoS2 unless references are listed in Table S3. Wt were obtained using custom p...  
Supporting Information: Additional

NTNU Open: The Natural Prod... NTNU Open

Figures S1-S13: NaPDoS2 wor... R\* ResearchGate

The Natural Product Domai... Journal of Biological Che...

Show more images ▾

# Finding NaPDoS2

**UPGRADE NOTICE:**

This version of NaPDoS has been superceded by a newer version called [NaPDoS2](#). We highly recommend that you switch to the new version, which is faster, more accurate, and includes a wider range of reference products:

<https://npdomainseeker.sdsc.edu/napdos2>

A complete description of the upgrade is provided in the following publication:

Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. *Journal of Biological Chemistry*, 2022, 102480, ISSN 0021-9258, <https://doi.org/10.1016/j.jbc.2022.102480>.

# NaPDoS2: Homepage

**NaPDoS2**  
Natural Product Domain Seeker

Home QuickStart Run Analysis Classification BGCs Contact Us

## Natural Product Domain Seeker

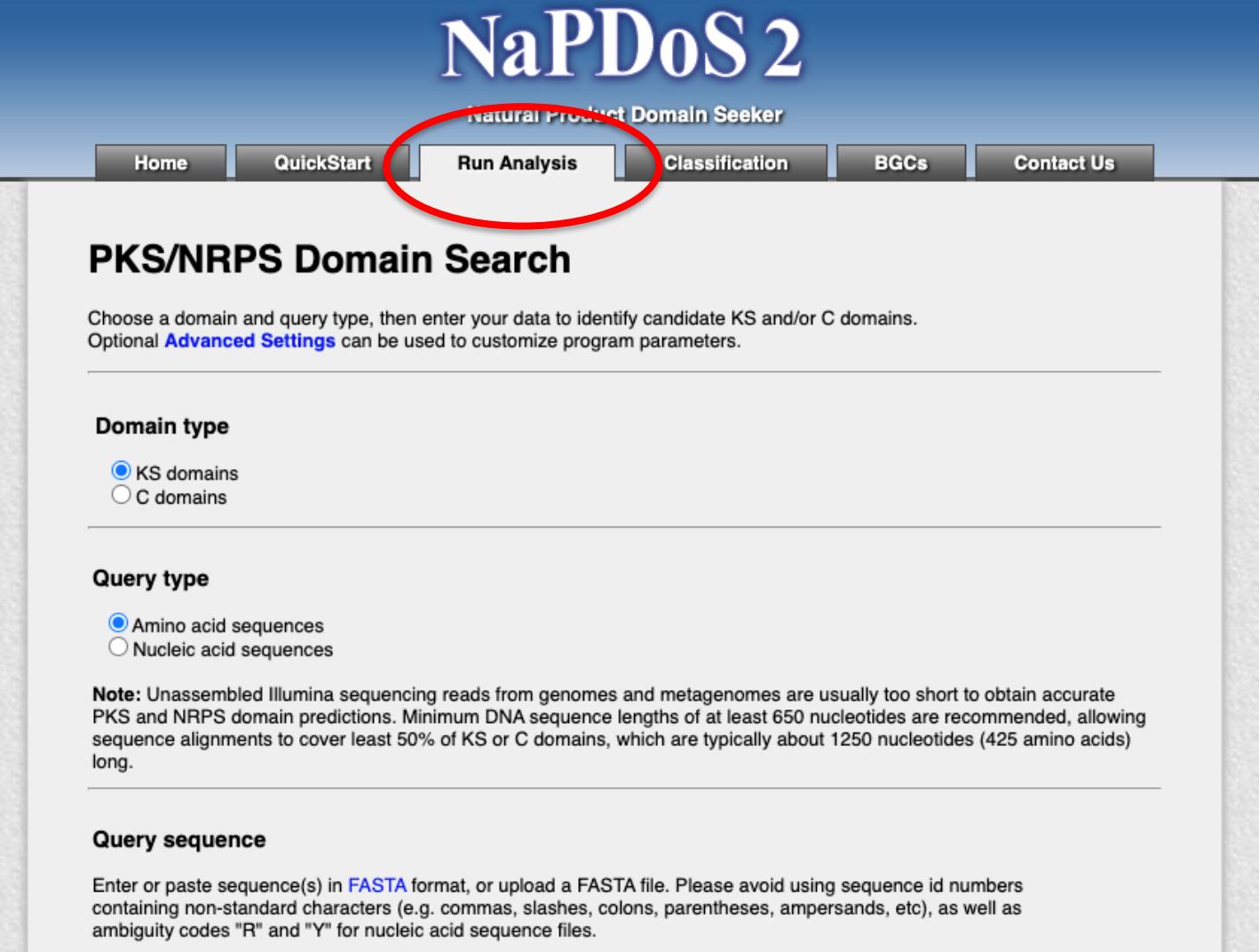
The diagram illustrates the workflow of NaPDoS2. It starts with a DNA or protein sequence (Query), which is then used to search against a local database (BLASTP). This leads to the creation of a domain phylogeny tree. Finally, the interpretation phase involves using query and reference sequences to determine the domain function.

**NaPDoS2 Features**

- NaPDoS2, the second-generation Natural Product Domain Seeker, rapidly detects and classifies **ketosynthase (KS)** and **condensation (C)** domains from genomic, metagenomic, or PCR amplicon sequence data.
- Biosynthetic potential is assessed using short KS and C domain sequence tags, enabling rapid analysis of large datasets without the need for complete, fully assembled biosynthetic gene clusters (BGCs).
- A phylogeny-based classification scheme is used to make broader predictions about the polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes in which these domains are found, including specific product [classification](#) assignments.
- NaPDoS2 provides major speed and accuracy improvements compared to legacy NaPDoS ([version 1](#)), along with greatly expanded taxonomic and functional diversity in KS classification categories ([Klau et al. 2022](#)).

Please check out our [quick start](#) page for instructions on how to [run analyses](#) and [interpret results](#). More details and a complete tutorial are provided in our full [documentation](#) file (right-click link to download).

# NaPDoS2: Run analysis



The screenshot shows the NaPDoS2 web application interface. At the top, the title "NaPDoS 2" is displayed in a large serif font, with "Natural Product Domain Seeker" in a smaller sans-serif font below it. A navigation bar contains links for "Home", "QuickStart", "Run Analysis" (which is circled in red), "Classification", "BGCs", and "Contact Us". The main content area is titled "PKS/NRPS Domain Search". It instructs users to choose a domain and query type, then enter their data to identify candidate KS and/or C domains. It also mentions optional "Advanced Settings". Below this, there are two sections: "Domain type" (radio buttons for "KS domains" and "C domains", with "KS domains" selected) and "Query type" (radio buttons for "Amino acid sequences" and "Nucleic acid sequences", with "Amino acid sequences" selected). A note at the bottom states that unassembled Illumina sequencing reads from genomes and metagenomes are usually too short to obtain accurate PKS and NRPS domain predictions, and recommends minimum DNA sequence lengths of at least 650 nucleotides. The final section, "Query sequence", asks users to enter or paste sequence(s) in FASTA format or upload a FASTA file.

NaPDoS 2  
Natural Product Domain Seeker

Home QuickStart Run Analysis Classification BGCs Contact Us

## PKS/NRPS Domain Search

Choose a domain and query type, then enter your data to identify candidate KS and/or C domains. Optional [Advanced Settings](#) can be used to customize program parameters.

---

**Domain type**

KS domains  
 C domains

---

**Query type**

Amino acid sequences  
 Nucleic acid sequences

**Note:** Unassembled Illumina sequencing reads from genomes and metagenomes are usually too short to obtain accurate PKS and NRPS domain predictions. Minimum DNA sequence lengths of at least 650 nucleotides are recommended, allowing sequence alignments to cover least 50% of KS or C domains, which are typically about 1250 nucleotides (425 amino acids) long.

---

**Query sequence**

Enter or paste sequence(s) in **FASTA** format, or upload a FASTA file. Please avoid using sequence id numbers containing non-standard characters (e.g. commas, slashes, colons, parentheses, ampersands, etc), as well as ambiguity codes "R" and "Y" for nucleic acid sequence files.

# NaPDoS2: Run analysis

## Query sequence

Enter or paste sequence(s) in [FASTA](#) format, or upload a FASTA file. Please avoid using sequence id numbers containing non-standard characters (e.g. commas, slashes, colons, parentheses, ampersands, etc), as well as ambiguity codes "R" and "Y" for nucleic acid sequence files.

Note: Current size limit for **genomic** and **metagenomic** files is 500 MB. Amplicon data queries are currently restricted to < 50,000 sequences. For larger files, please consider our suggestions for [file size management](#) before submission. Large nucleotide sequence files may take several minutes to load. We are working on solutions to enable analysis of larger amplicon data sets, but do not yet have a projected availability date.

Upload a file:  No file chosen

**SEEK**

**RESET**

# NaPDoS2: Run analysis



## Advanced Settings



Sensitivity may be increased by choosing higher e-values and shorter minimum match lengths, but these adjustments may also increase false positives. Conversely, selectivity can be improved with lower e-values and longer minimum match lengths, at the cost of decreased sensitivity.

**BLASTP**

**e-value cutoff**

1e-8 (default) ▾

**Minimum align length**

200 aa (600 nt) (default) ▾

**SEEK**

**RESET**

# NaPDoS2: Run analysis

The screenshot shows the NaPDoS 2 web application. At the top, the logo "NaPDoS 2" is displayed in large blue letters, with "Natural Product Domain Seeker" in smaller text below it. A navigation bar with links for Home, QuickStart, Run Analysis, Classification, BGCs, and Contact Us is visible. The main content area displays a job submission form. It includes a bolded "Job ID 30956717" followed by a horizontal line. Below this, the text "Estimated processing time: ~22 seconds." is shown. A blue "SUBMIT JOB" button is present. A section titled "Search Parameters" contains a bulleted list of search criteria:

- query type = aa
- number of input sequences = 1250
- domain type = KS
- minimum match length = 200 aa
- min e-value for db match = 1e-8
- output table matches per query = 1
- reference sequences = all\_KS\_191020\_1877.faa
- database version = NaPDoS2\_v13b

# NaPDoS2: Analysis outputs

**Domain Classification Summary**

- 28 KS domains were identified from 4530 input sequences.
- Click on buttons below to view a detailed table of results, download domain sequences in fasta format, or build comparative trees.

**VIEW ALL MATCHES**

---

**Individual Domain Classes**

- Select one or more categories below to view a subset of matches.

**Select All**

Select	Class	Subclass	Num matches
<input type="checkbox"/>	type I modular cis-AT	no subclass	12
<input type="checkbox"/>	type II polyene	KSa	3
<input type="checkbox"/>	type II polyene	KSb	2
<input type="checkbox"/>	type II FAS	no subclass	2
<input type="checkbox"/>	type I modular cis-AT	hybrid KS	2
<input type="checkbox"/>	type I iterative cis-AT	enediyne	2
<input type="checkbox"/>	type II aromatic	pentangular polyphenol KSa	1
<input type="checkbox"/>	type I modular cis-AT	loading module	1
<input type="checkbox"/>	type II aromatic	pentangular polyphenol KSb	1
<input type="checkbox"/>	type II aromatic	angucycline II KSb	1
<input type="checkbox"/>	type II aromatic	angucycline II KSa	1

Right-click to **DOWNLOAD** this table in tab-delimited format.

**VIEW A SUBSET**

# NaPDoS2: Analysis outputs

## Database Search Results

28 KS domains were identified from 4530 input sequences.

Use check boxes to select candidates for [further analysis](#) below.

(options may take a few seconds to load for large match tables) Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Select All

	cand_id	database match	percent identity	align length	e-value	BGC match
<input type="checkbox"/>	WP_012013922.1_3_459	sporolide_KS01_iPKSenediyne	100	457	1.6e-270	sporolide
<input type="checkbox"/>	WP_148217347.1_129_556	salinilactam_KS01_cisAT	100	428	1.4e-252	salinilactam
<input type="checkbox"/>	WP_148217347.1_1867_2294	salinilactam_KS02_cisAT	100	428	6.7e-250	salinilactam
<input type="checkbox"/>	WP_148217347.1_5338_5765	salinilactam_KS04_cisAT	100	428	4.4e-249	salinilactam
<input type="checkbox"/>	WP_148217347.1_3599_4027	salinilactam_KS03_cisAT	100	429	1.8e-247	salinilactam
<input type="checkbox"/>	WP_012014001.1_34_461	salinilactam_KS05_cisAT	100	428	3.1e-251	salinilactam
<input type="checkbox"/>	WP_012014001.1_1742_2153	salinilactam_KS06_cisAT	100	412	2.9e-241	salinilactam
<input type="checkbox"/>	WP_012014004.1_34_459	salinilactam_KS10_cisAT	100	426	1.0e-248	salinilactam

# NaPDoS2: Analysis outputs

## Database Search Results

28 KS domains were identified from 4530 input sequences.

Use check boxes to select candidates for [further analysis](#) below.

(options may take a few seconds to load for large match tables) Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Select All

	cand_id	database match	percent identity	align length	e-value	BGC match
<input type="checkbox"/>	WP_012013922.1_3_459	sporolide_KS01_iPKSenediyne	100	457	1.6e-270	sporolide
<input type="checkbox"/>	WP_148217347.1_129_556	salinilactam_KS01_cisAT	100	428	1.4e-252	salinilactam
<input type="checkbox"/>	WP_148217347.1_1867_2294	salinilactam_KS02_cisAT	100	428	6.7e-250	salinilactam
<input type="checkbox"/>	WP_148217347.1_5338_5765	salinilactam_KS04_cisAT	100	428	4.4e-249	salinilactam
<input type="checkbox"/>	WP_148217347.1_3599_4027	salinilactam_KS03_cisAT	100	429	1.8e-247	salinilactam
<input type="checkbox"/>	WP_012014001.1_34_461	salinilactam_KS05_cisAT	100	428	3.1e-251	salinilactam
<input type="checkbox"/>	WP_012014001.1_1742_2153	salinilactam_KS06_cisAT	100	412	2.9e-241	salinilactam
<input type="checkbox"/>	WP_012014004.1_34_459	salinilactam_KS10_cisAT	100	426	1.0e-248	salinilactam

Right-click to [DOWNLOAD](#) this table in tab-delimited format.

### Options

Output selected sequences in fasta format.

**GET RESULTS**

# NaPDoS2: Analysis outputs

## Database Search Results

28 KS domains were identified from 4530 input sequences.

Use check boxes to select candidates for [further analysis](#) below.

(options may take a few seconds to load for large match tables) Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Click on column headers to sort (multiple clicks toggle between ascending and descending order).

Select All

	cand_id	database match	percent identity	align length	e-value	BGC match
<input type="checkbox"/>	WP_012013922.1_3_459	sporolide_KS01_iPKSenediyne	100	457	1.6e-270	sporolide
<input type="checkbox"/>	WP_148217347.1_129_556	salinilactam_KS01_cisAT	100	428	1.4e-252	salinilactam
<input type="checkbox"/>	WP_148217347.1_1867_2294	salinilactam_KS02_cisAT	100	428	6.7e-250	salinilactam
<input type="checkbox"/>	WP_148217347.1_5338_5765	salinilactam_KS04_cisAT	100	428	4.4e-249	salinilactam
<input type="checkbox"/>	WP_148217347.1_3599_4027	salinilactam_KS03_cisAT	100	429	1.8e-247	salinilactam
<input type="checkbox"/>	WP_012014001.1_34_461	salinilactam_KS05_cisAT	100	428	3.1e-251	salinilactam
<input type="checkbox"/>	WP_012014001.1_1742_2153	salinilactam_KS06_cisAT	100	412	2.9e-241	salinilactam
<input type="checkbox"/>	WP_012014004.1_34_459	salinilactam_KS10_cisAT	100	426	1.0e-248	salinilactam

Right-click to [DOWNLOAD](#) this table in tab-delimited format.

### Options

Output selected sequences in fasta format.

**GET RESULTS**

# NaPDoS2: Analysis outputs

## Options

---

- Output selected sequences** in fasta format
- Output Alignment** with closest database matches  
Select alignment format:
- Construct tree** (candidate domains + blast matches + reference domains)

**GET RESULTS**

# Talk outline

- Introduction
- NaPDoS2: Genome mining webtool
- **Metagenomic data reveals type I polyketide synthase distributions across biomes**
- KS amplicons highlight polyketide biosynthetic potential across abyssal sediments

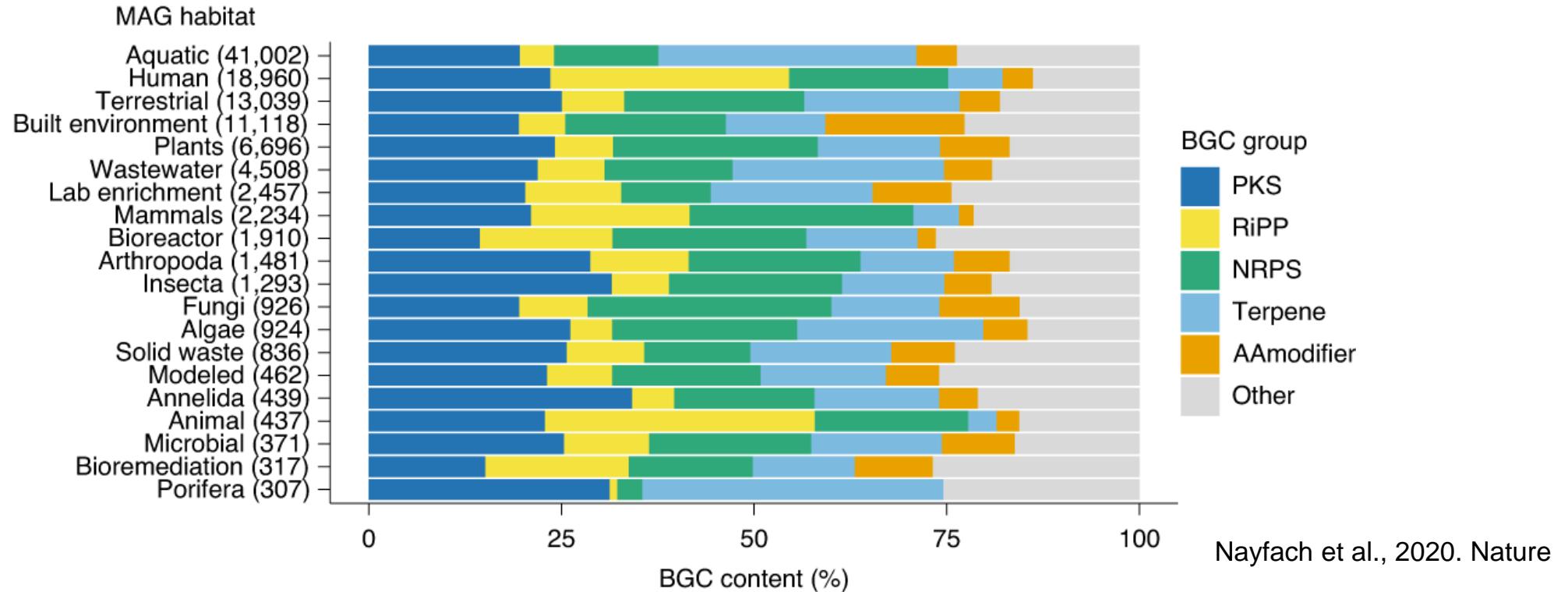
# PKS distribution across biomes

## Objectives:

- Are biomes enriched in different classes of PKSs?
- Are metagenome-extracted KSs different from those in known pathways?
- Are there biome-specific KS clades?

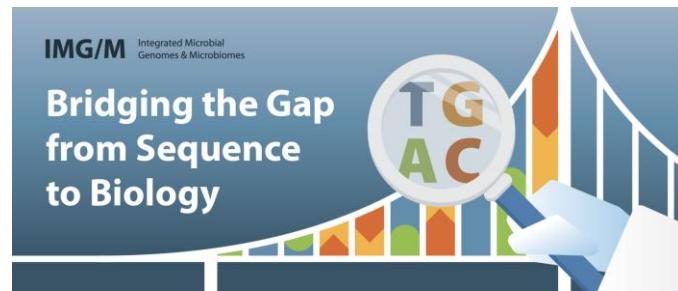


## Previous study: BGCs seen at similar rates across biomes



“BGC composition not limited or amplified by environment”

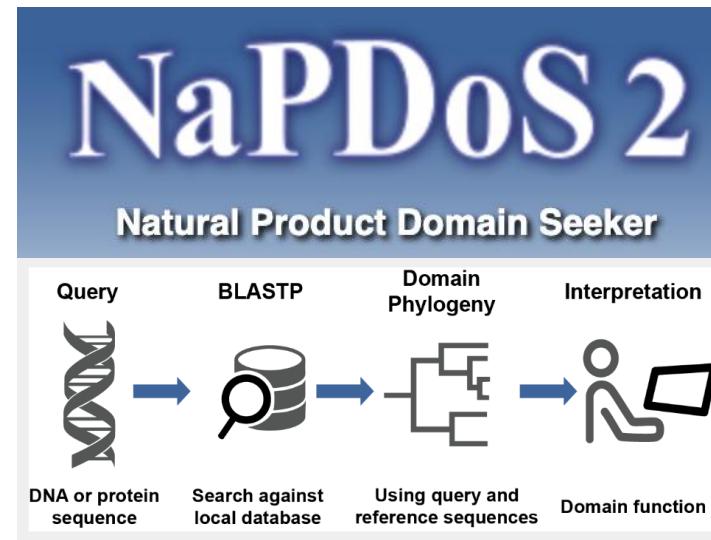
# Workflow – Metagenomic data



137 metagenomes

8 biomes

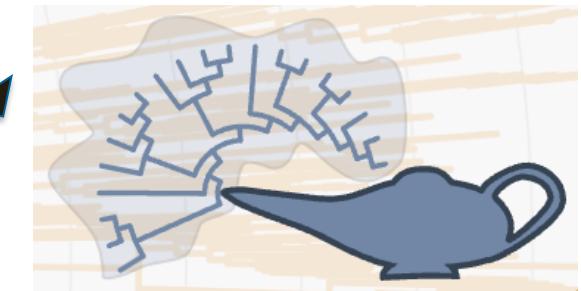
240 Gbps of data



>35,000 KS domains extracted

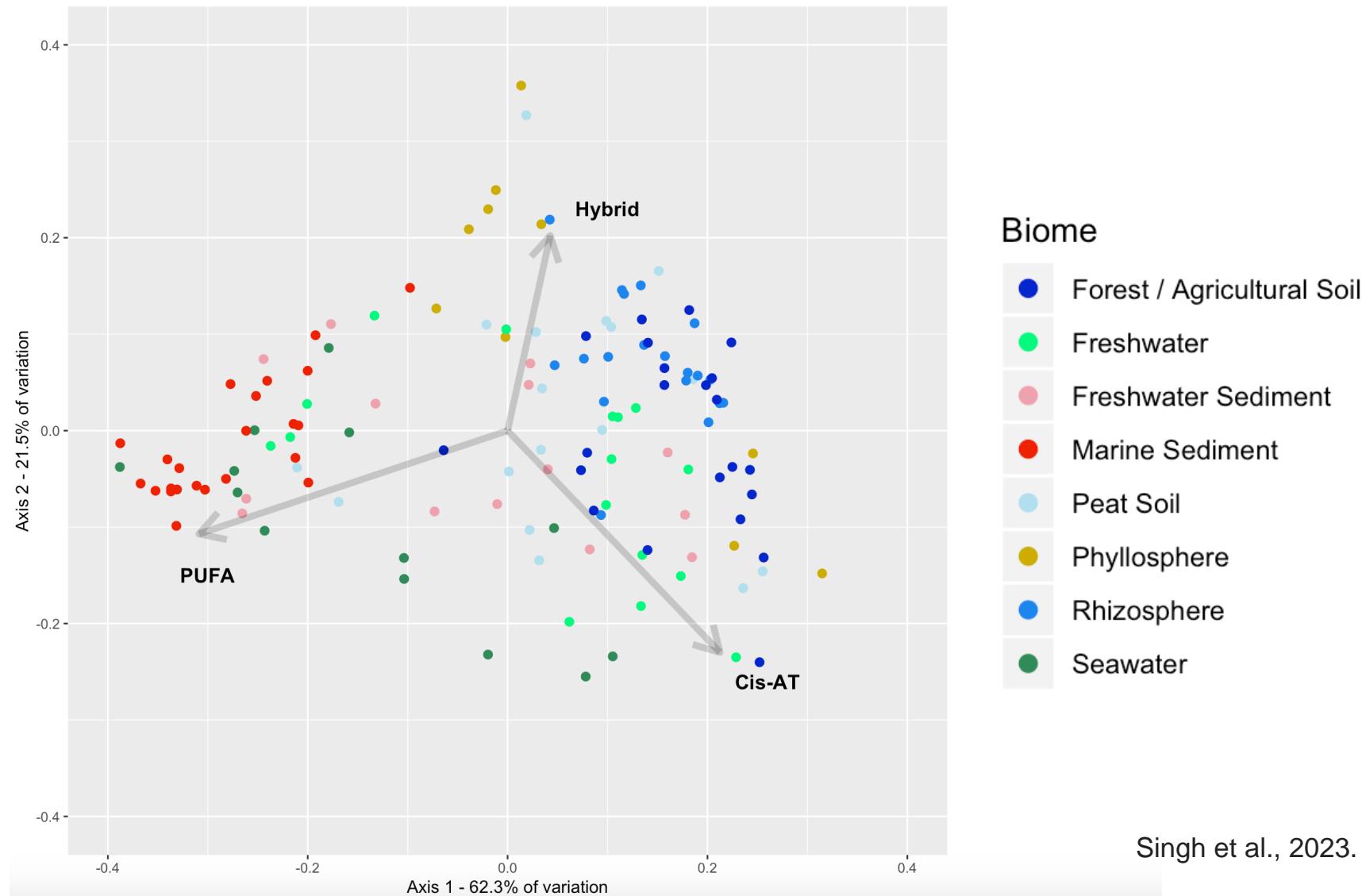


Compare to known BGCs

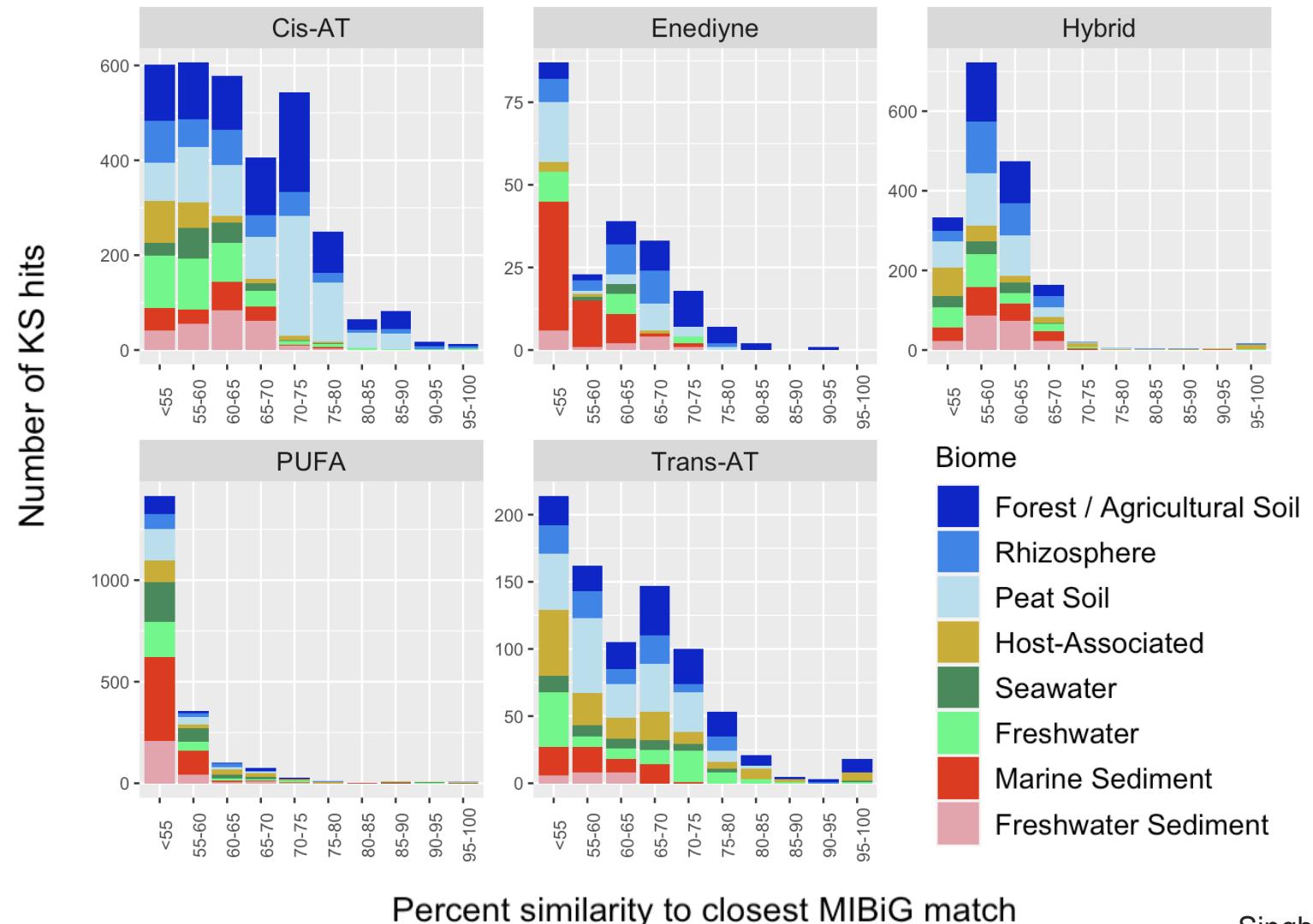


KS subclass phylogenies

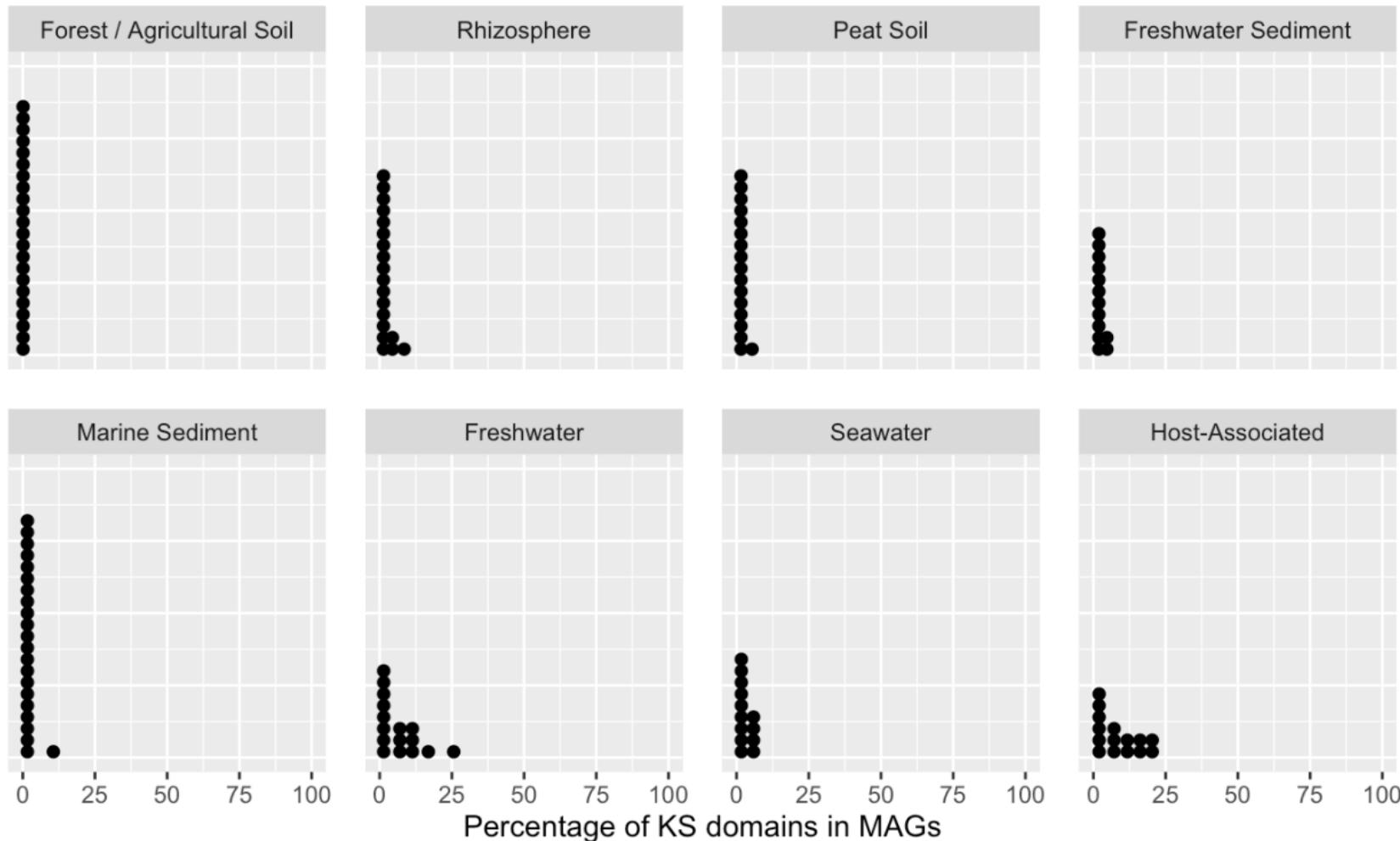
# Biomes differ in KS composition



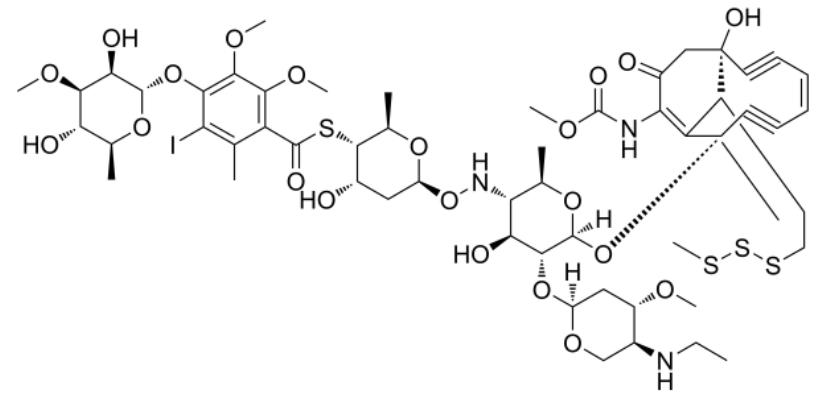
# Most metagenome-extracted KS domains differ from known BGCs



# Most KS domains fall outside of MAGs



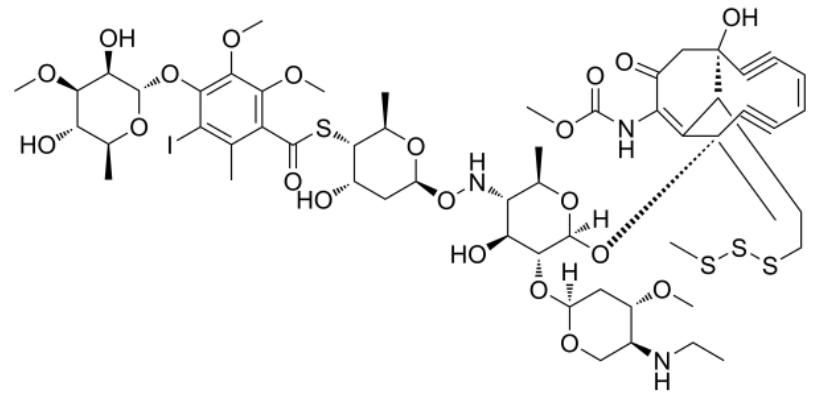
# Enediyne KS domains phylogeny



**Calicheamicin**

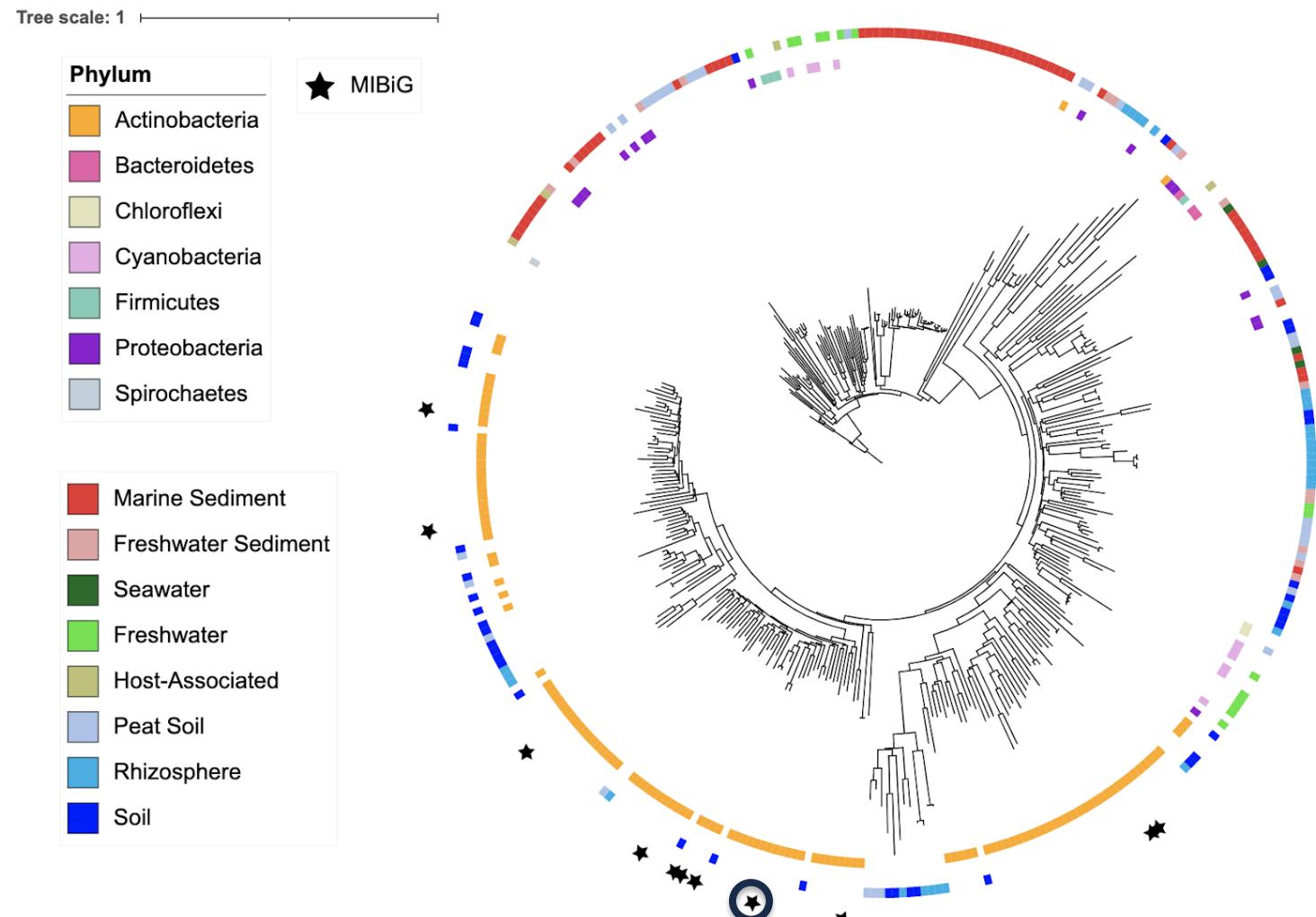
**Antitumor agent**

# Enediyne KS domains clade by biome and taxonomy



Calicheamicin

Antitumor agent



Singh et al., 2023. mSystems

## Conclusions from metagenomic data

- **Biomes have distinct type I KS compositions**
- **Most metagenome-extracted KSs are different from characterized BGCs**
- **Enediyne KS domains clade by biome and taxonomy, hint at novel enediyne discovery opportunities**

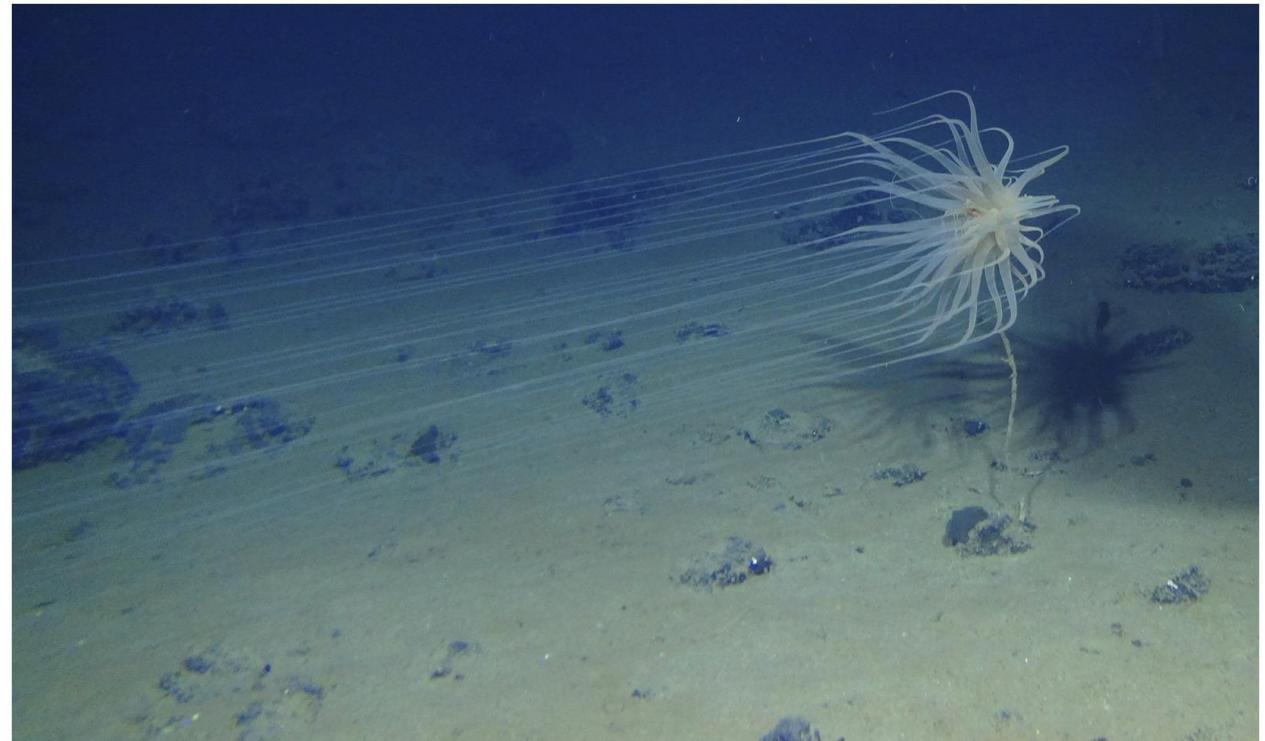
# Talk outline

- Introduction
- NaPDoS2: Genome mining webtool
- Metagenomic data reveals type I polyketide synthase distributions across biomes
- **KS amplicons highlight polyketide biosynthetic potential across abyssal sediments**

# Amplicons reveal biosynthetic potential of abyssal sediments

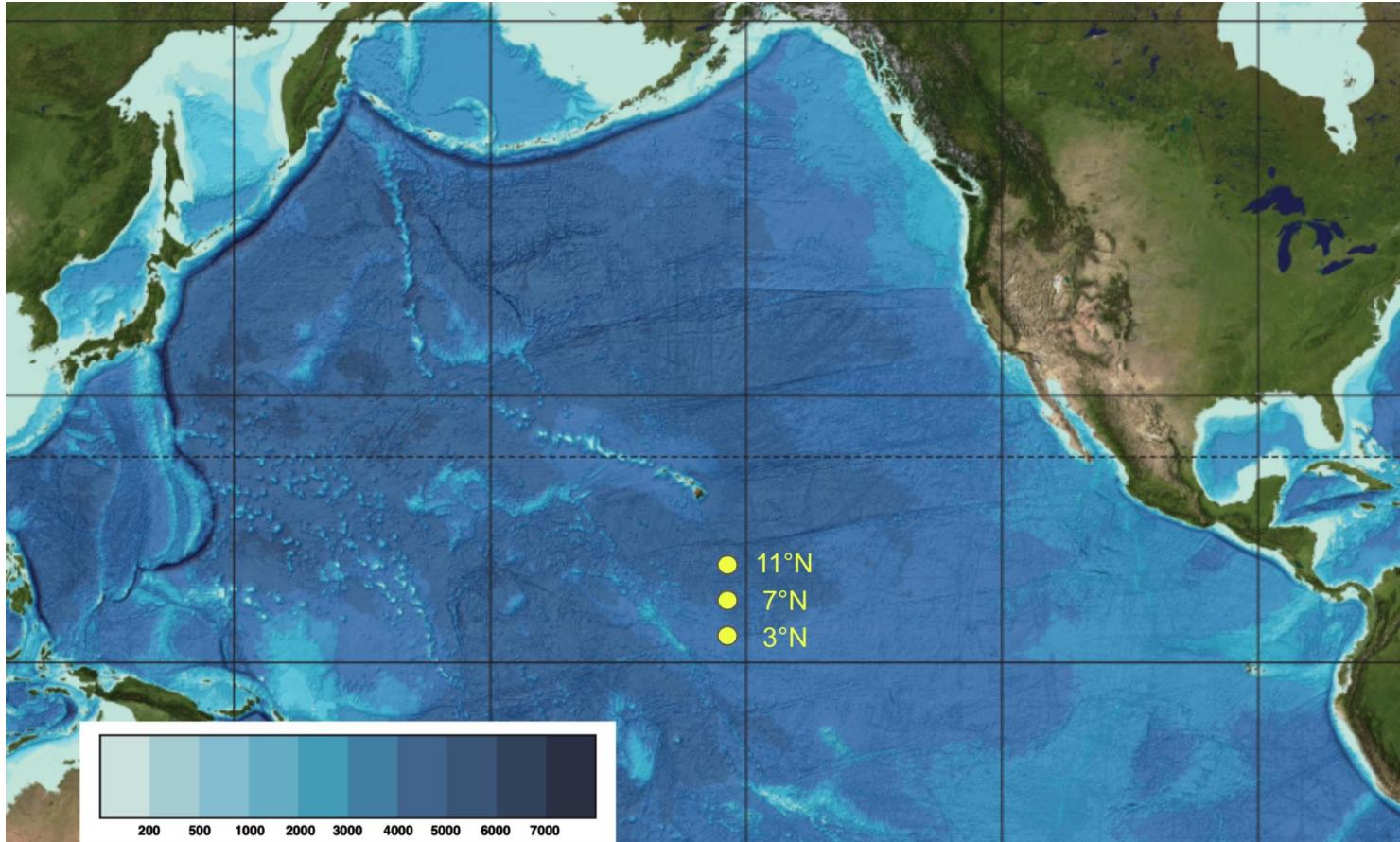
## Objectives:

- **What is the polyketide biosynthetic potential within abyssal sediments?**
- **How do different strategies (KS amplicon sequencing, metagenomics) compare?**



Collaborators: Dr. Johanna Gutleben, Dr. Alexander Chase, Dr. Alex Bogdanov

# Sampling of abyssal sediments

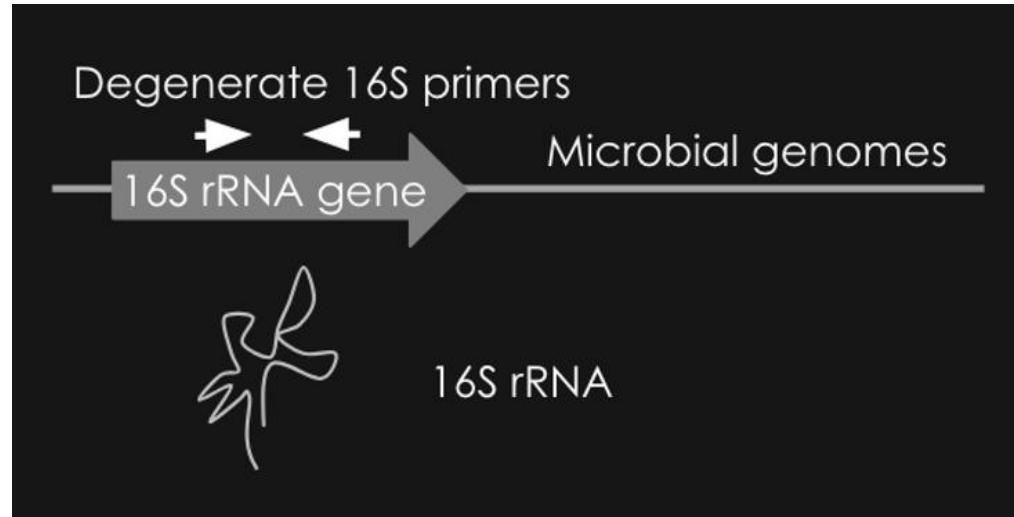


Abyssal sediments  
>80% of ocean floor

NSF cruise led by Dr. Brian Haley

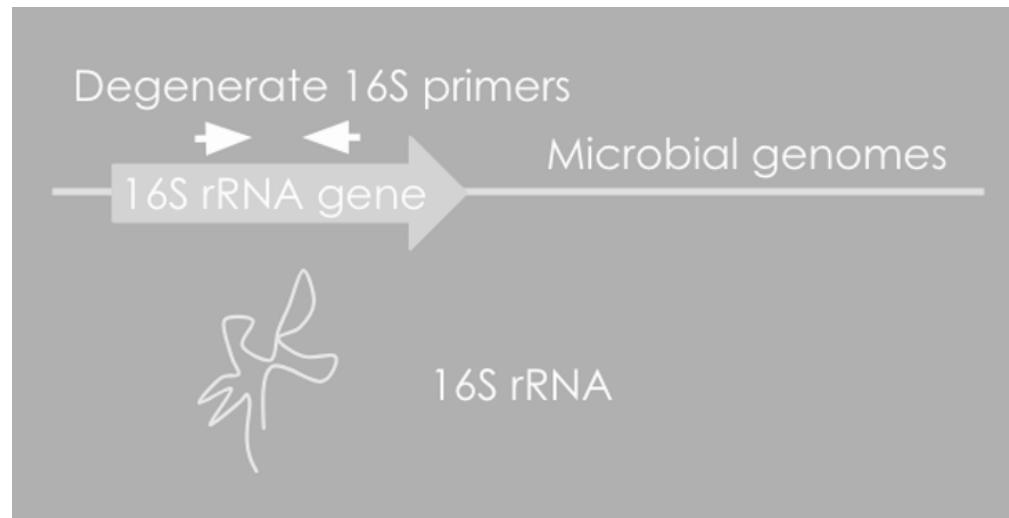
All 3 sites at depths > 5000 m

# KS amplicon sequencing analogous to 16S amplicon sequencing

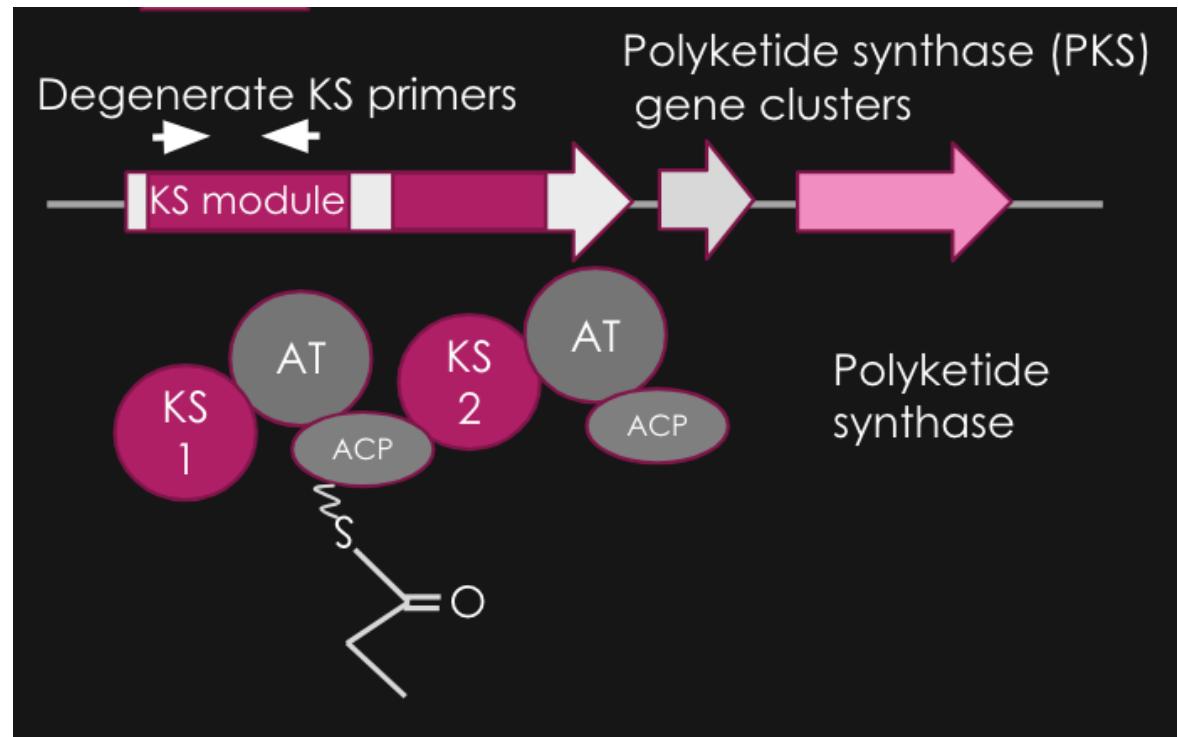


Gutleben, 2023

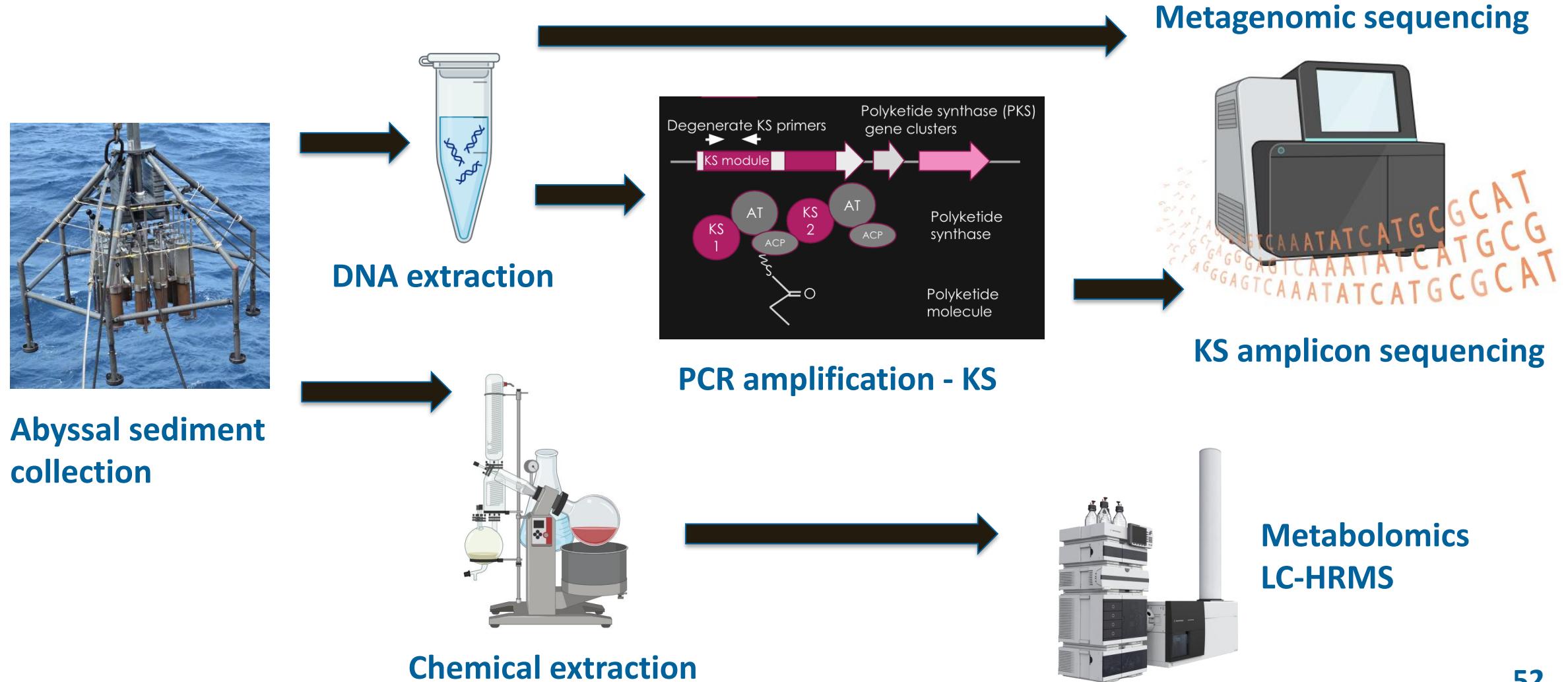
# KS amplicon sequencing analogous to 16S amplicon sequencing



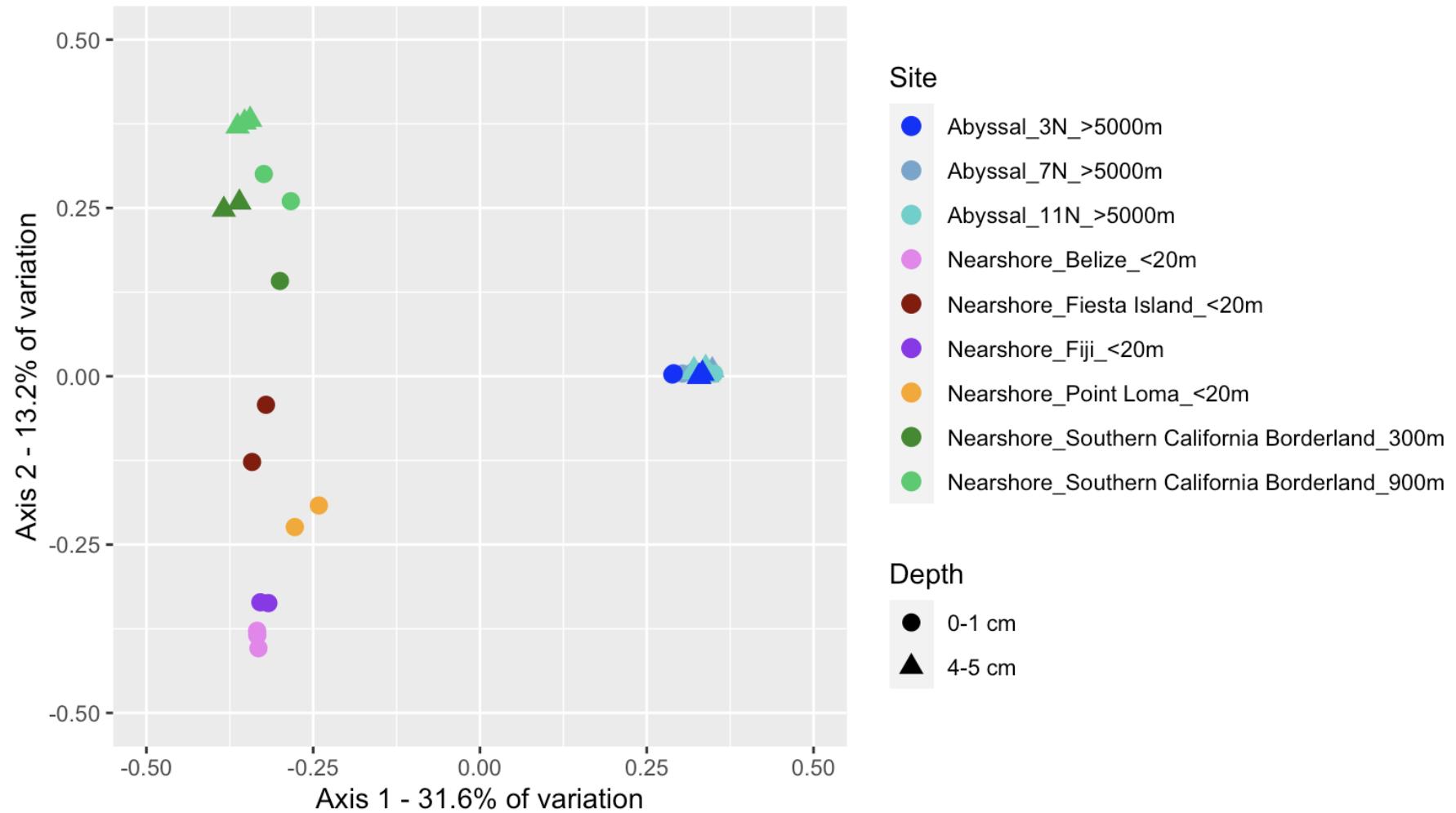
Gutleben, 2023



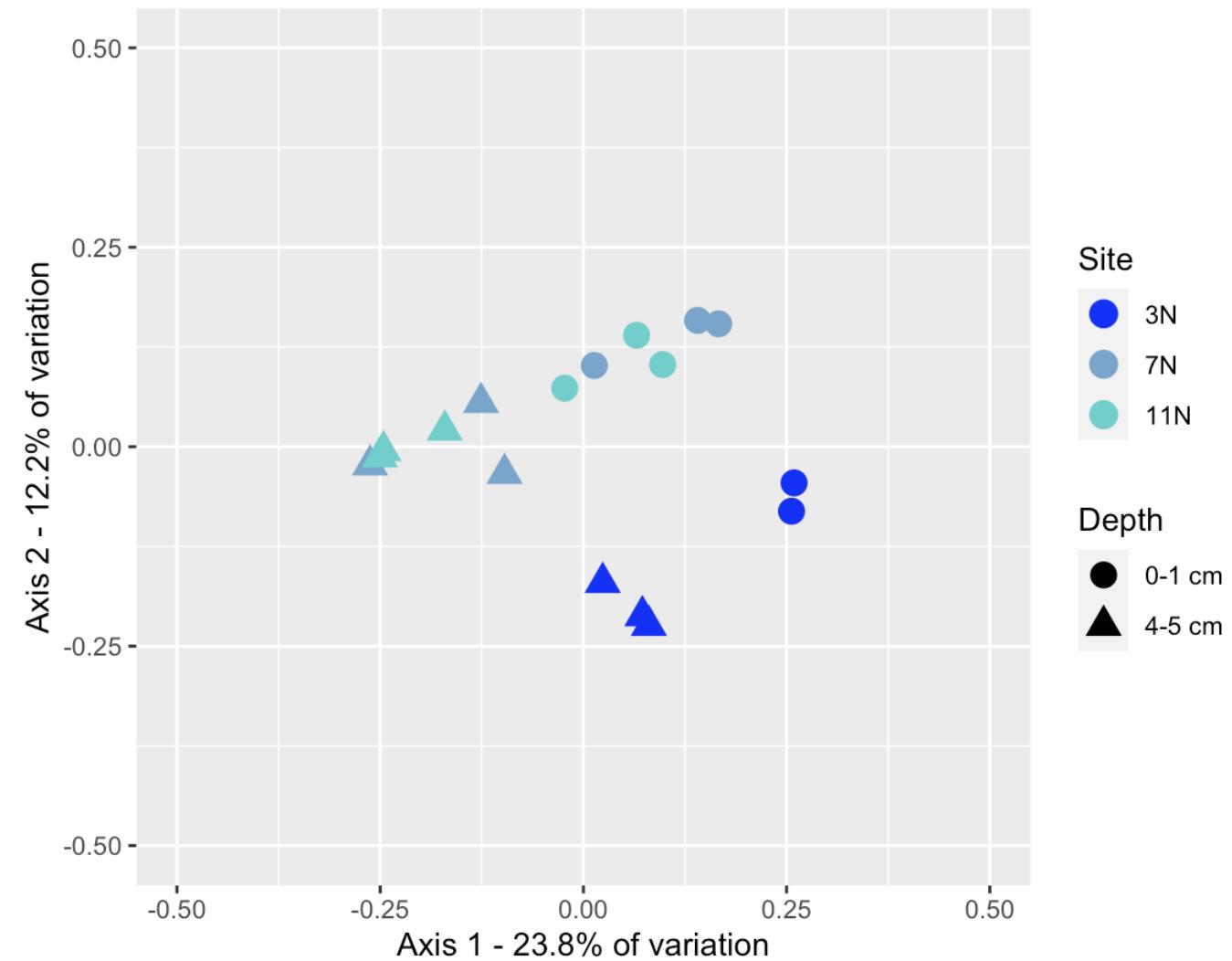
# Multi-omics workflow



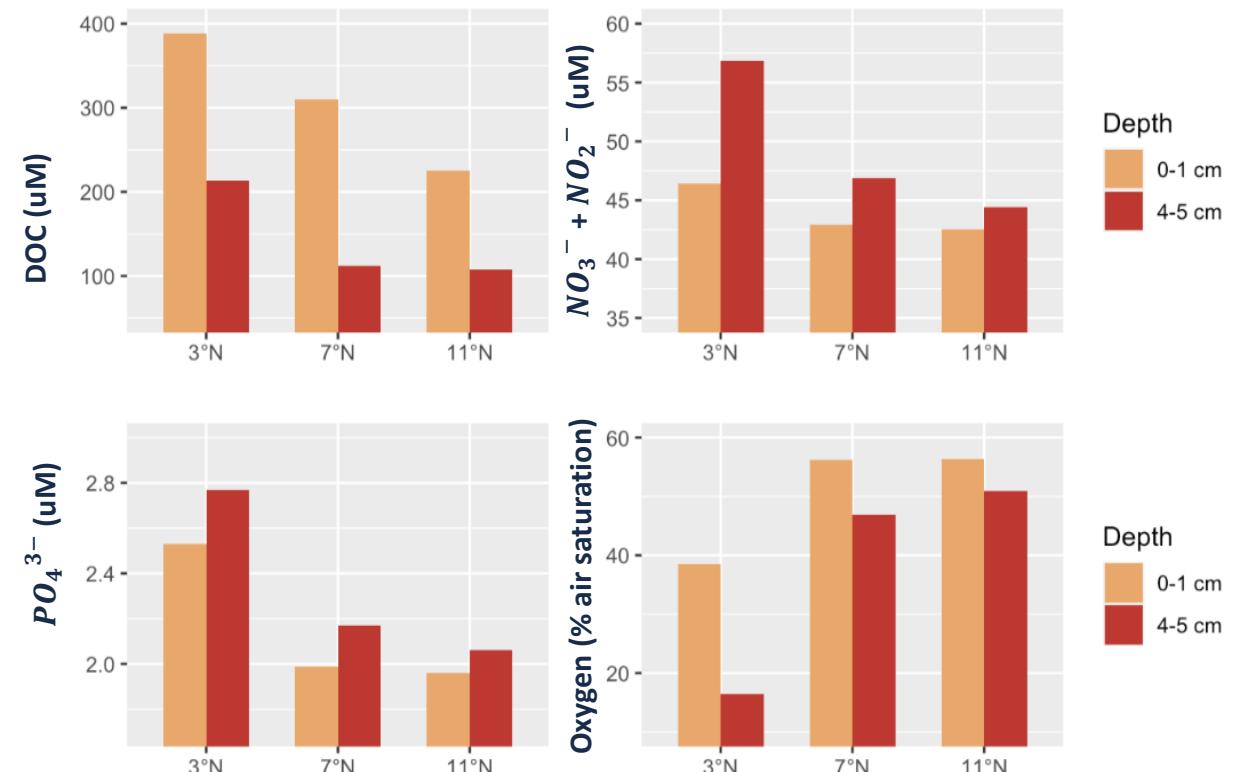
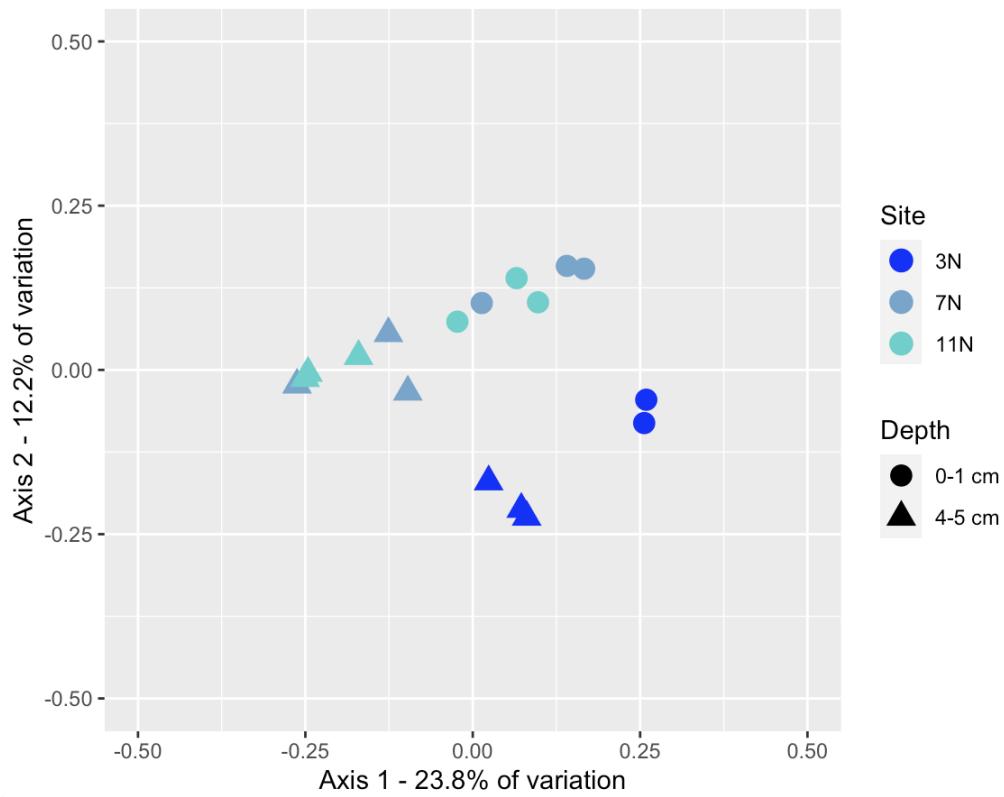
# Abyssal sediments have unique KS amplicon communities



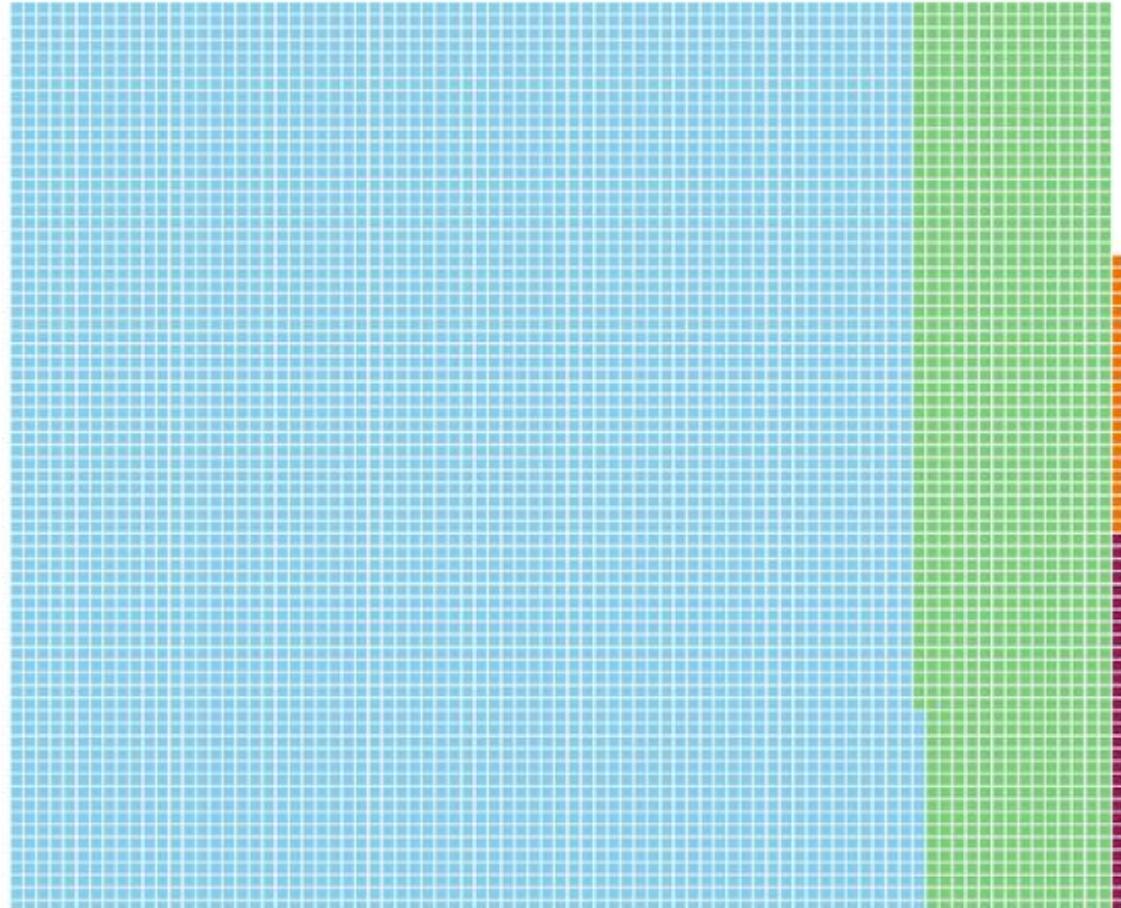
# Abyssal KS amplicon communities differ by site, sediment layer



# KS community differences align with biogeochemical differences



# Abyssal sediments contain unknown KS domains



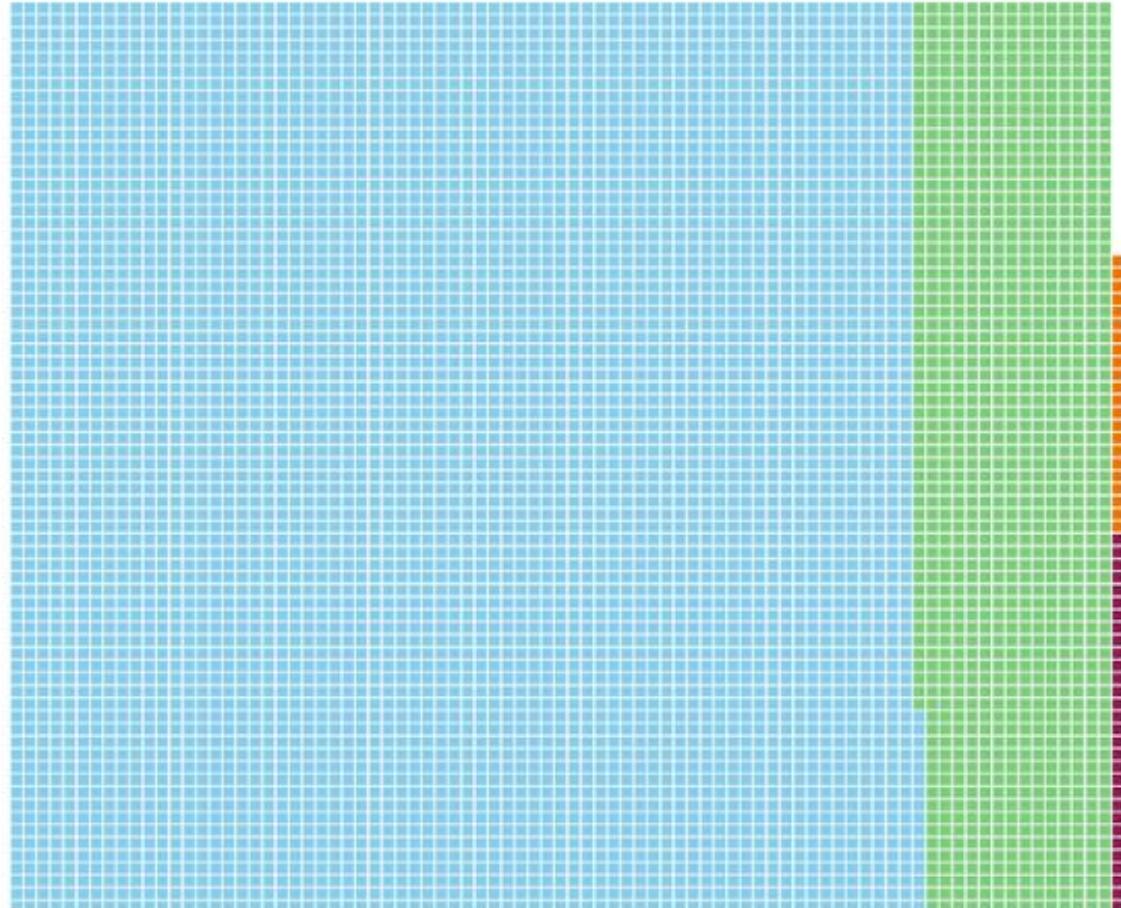
## OBU composition

- Unique to abyssal sediments
- Cluster with nearshore sediment KS amplicons
- Cluster with MIBiG
- Cluster with MIBiG & nearshore sediment KS amplicons

**6,000+ abyssal sediment KS amplicon OBUs**

**Less than 1% cluster with MIBiG**

# KS amplicon sequencing more effective than metagenomics



## OBU composition

- Unique to abyssal sediments
- Cluster with nearshore sediment KS amplicons
- Cluster with MIBiG
- Cluster with MIBiG, nearshore sediment KS amplicons

**6,000+ abyssal sediment KS amplicon OBUs**

**Less than 1% cluster with MIBiG**

**Metagenomes only contain 1 KS OBU**

## Conclusions from KS amplicon data

- **Abyssal KS amplicon communities distinct from nearshore sediments and experimentally characterized PKS BGCs**
- **KS amplicon sequencing recovered orders of magnitude more KS domains compared to metagenomes**

# Overall conclusions

## Overall conclusions

- NaPDoS2 an effective way to assess PKS potential
  - Speed, specificity, size

## Overall conclusions

- NaPDoS2 an effective way to assess PKS potential
  - Speed, specificity, size
- **Metagenomic data show that biomes differ in KS composition and diversity**

# Overall conclusions

- **NaPDoS2 an effective way to assess PKS potential**
  - **Speed, specificity, size**
- **Metagenomic data show that biomes differ in KS composition and diversity**
- **KS amplicons highlight that abyssal sediments carry unique PKS potential**

## Overall conclusions

- NaPDoS2 an effective way to assess PKS potential
  - Speed, specificity, size
- Metagenomic data show that biomes differ in KS composition and diversity
- **KS amplicons highlight that abyssal sediments carry unique PKS potential**

# THANKS TO :

- Dr. Paul Jensen
- Dr. Johanna Gutleben
- Dr. Alex Chase
- Dr. Kaitlin Creamer
- Dr. Gabriel Castro
- Dr. Dulce Guillen-Matus
- Dr. Doug Sweeney
- Dr. Alyssa Demko
- Dr. Sheila Podell
  
- Dr. Maggie Yuan
- Dr. Eoin Brodie
- Dr. Alison Sherwood
- Dr. Heejung Cho
- Dr. Pat Sorenson
- Dr. Rachael Wade

