Linking Enzymes to Psychrophiles and how this Reflects Metabolic Network of Psychrophiles

Moon He, Alan Yu, Harvey Lee, Songtianze Huang, Peter Siu

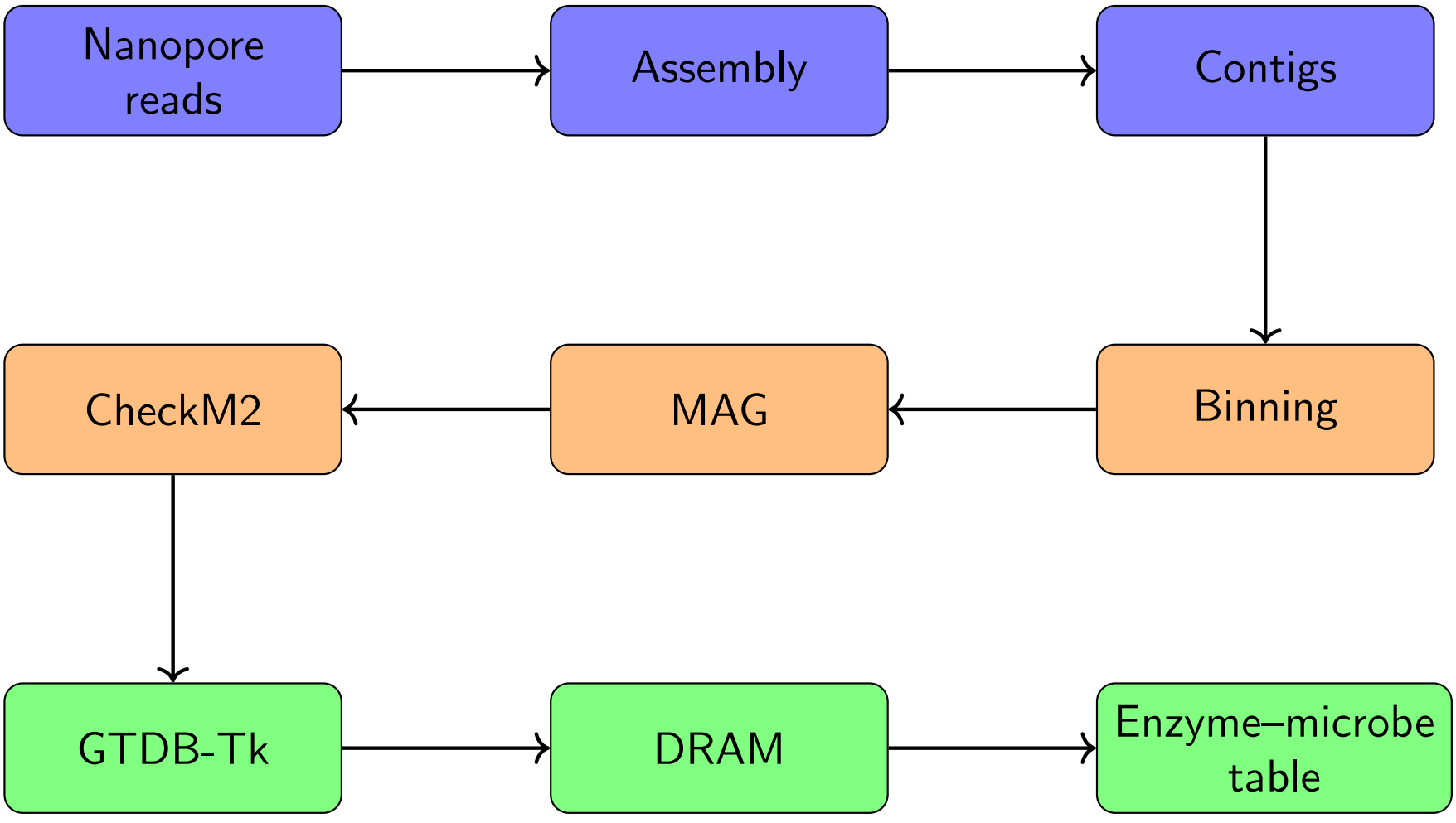
March 27, 2026

Introduction

Four facts:

- Pyschrophile: organism (microbe) live in cold environment.
- Pyschrophile produce enzymes to help them adapt to the temperature.
- Same enzyme may share among different psychrophiles.
- Some enzyme can help psychrophiles live/grow together.

Workflow overview



Field design

Comparing different extremes (gradient-based sampling):

- Temperature: Cold vs. relatively warmer spots;
- Oxygen: high oxygen vs. low oxygen zones;
- Habitat state: Newly thawed ice vs. established soil;
- Depth: Surface vs. underground layers.

Wet lab (physical workflow)

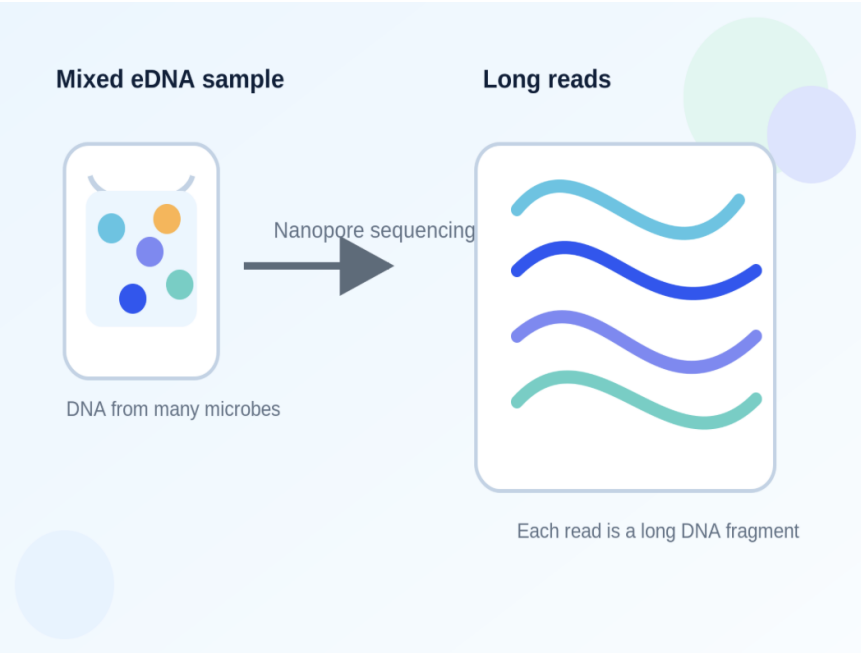
From soil to sequencer:

- DNA extraction: Extracting mixed environmental DNA (eDNA) directly from the samples;
- Sequencing: Running the DNA through the Nanopore long-read sequencer;

Output:
Translating physical biological samples into massive digital datasets.

Step 1 – collecting the "puzzle pieces" (the raw data)

- Starting point: Raw DNA sequences.
- The situation: Mixed and broken DNA extracted from all the different microbes living together in the Antarctic soil.
- Simple analogy: Like pouring hundreds of different puzzle boxes into a giant pool, creating a massive pile of mixed-up pieces.

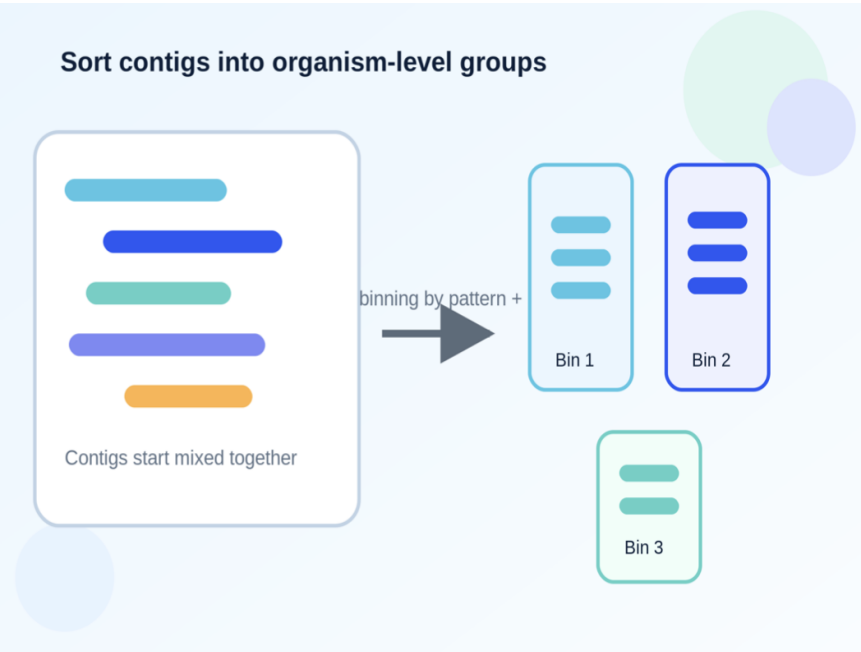


Step 2, 3 – building "large sections" (assembly)

- The process: Raw DNA $\rightarrow$ assembly $\rightarrow$ long DNA chunks (contigs).
- How it works: Finding overlapping edges to connect small pieces together.
- The long-read advantage: The larger the "puzzle pieces" you start with, the faster and more successful you are at building large sections.

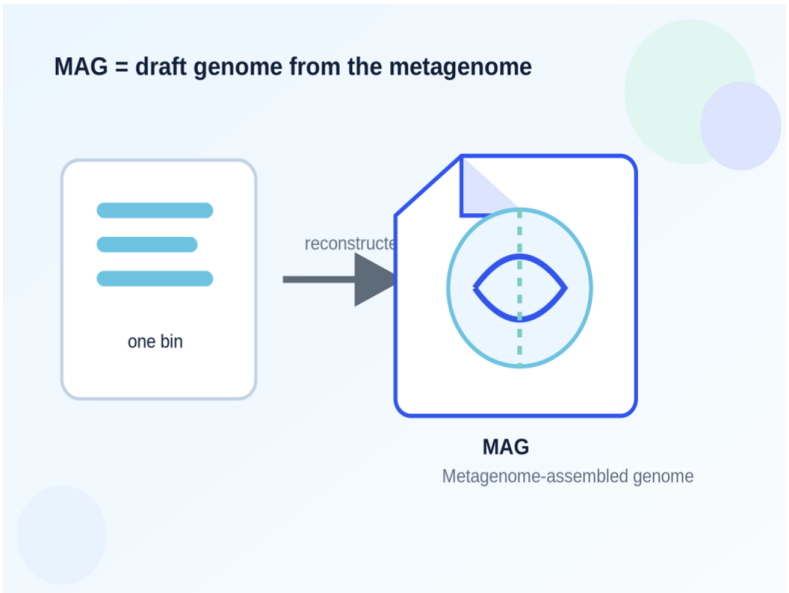
Step 4 – “sorting into boxes” (binning)

- The process: Long DNA chunks (contigs) → binning.
- The challenge: The assembled large sections are still a massive mix of all microbes.
- Simple analogy: Sorting pieces into specific boxes based on their “patterns” and “frequencies”, grouping pieces that belong to the same puzzle (or same microbe).



Step 5 – getting the "microbial blueprint" (getting the mategenome-assmebled genome, MAG)

- The process: Sorted bins → MAGs (metagenome-assembled genomes).
- The result: Each "box" becomes a draft genetic blueprint for a single Antarctic microbe.
- Moving forward: This individual blueprint provides the foundation to investigate "who it is" and "what functions it has."

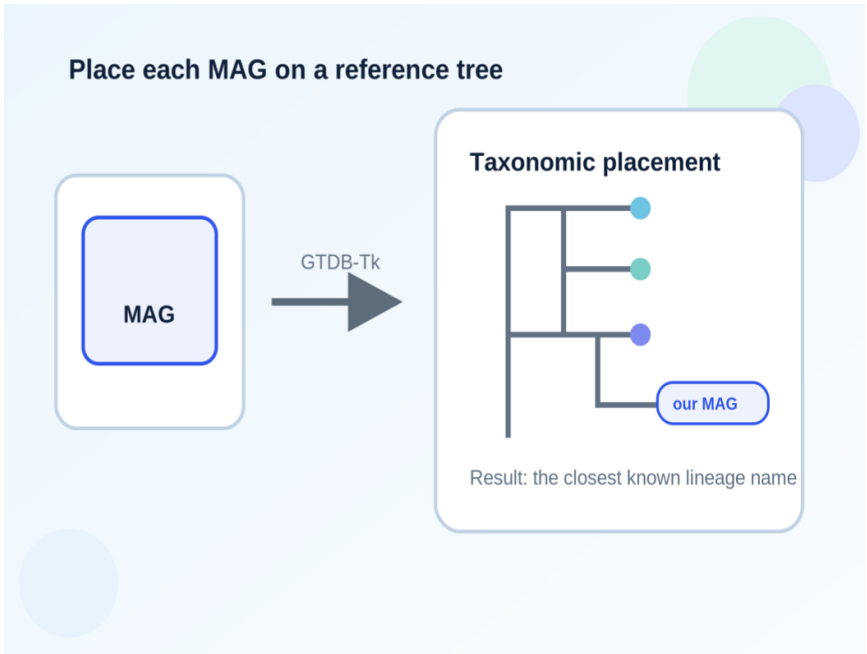


Step 6 – test whether each MAG is credible enough to trust (check M2)

- Estimate genome completeness: how much of the genome we recovered.
- Estimate contamination: whether DNA from different microbes got mixed together.
- Retain stronger MAGs for interpretation and downstream mapping.

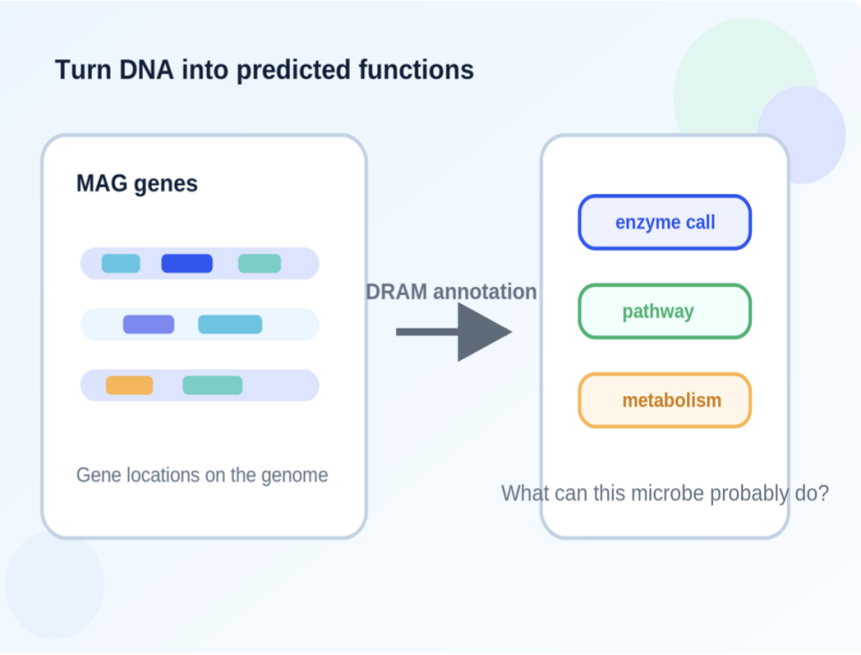
Step 7 – assign each MAG to the best available taxonomic lineage (GTDB-Tk)

- Place each MAG onto a reference genome tree.
- Find the closest bacterial or archaeal lineage.
- This gives us a taxonomic identity for each reconstructed genome.



Step 8 – DRAM: annotate genes, enzymes, and pathways on each MAG

- Read the genes carried by each MAG.
- Predict enzyme families, pathways, and metabolic functions.
- Translate raw DNA sequence into biological meaning.



Step 9 – enzyme-microbe correspondence table

- Combine enzyme annotation with contig, MAG, taxonomy, and sample context.
- For each enzyme, record which reconstructed microbe carries it.
- This becomes the biological input for the downstream metabolic network model.

Table: Example: enzyme-microbe correspondence table

Enzyme	**Contig**	**MAG**	**Taxon**	**Sample / condition**
glycosidase	contig_12	MAG_3	Flavobacteriaceae	cold, oxic
peptidase	contig_41	MAG_7	Pseudoalteromonas	low O_2
lipase	contig_58	MAG_9	Marinobacter	warmer site

Problem description

From the former experiments, we should know different types of psychrophiles and enzymes they have. In this part, we want to construct a metabolic network of psychrophiles, i.e., if physchrophile p_1 increases a certain amount, how would this change be reflected in the a group P of different pcychrophiles.

Problem setup: notations and hypothesis

We define the following sets:

- The set of psychrophiles: $P = \{p_1, p_2, \ldots\}$.
- The set of enzymes: $E = \{e_1, e_2, \ldots\}$.
- The set of subtrates / metabolites: $S = \{s_1, s_2, \ldots\}$.

Hypothesis:

- Each enzyme enduces a transformation:

$$\varphi : E \to S \times S,\ \varphi(e_i) = (s_{\text{in}}, s_{\text{out}}).$$

- Metabolites produced by one organism are availble to all others.
- Ignore spatial separation.
- Growth depends on availability of required substrates.

Problem setup: psychrophile-enzyme incidence matrix

Definition

$$A \in \{0,1\}^{n \times m},\ A_{i\alpha} = \begin{cases} 1 & \text{if } p_i \text{ posesses } e_\alpha \\ 0 & \text{otherwise} \end{cases}.$$

Problem setup: production and requirement sets

Define:
For each p_i,

$$\mathcal{O}_i = \{s_{\text{out}} : \exists e_\alpha \text{ with } A_{i\alpha} = 1\},$$

i.e., all the subtrates produced by p_i; and

$$\mathcal{I}_i = \{s_{\text{in}} : \exists e_\alpha \text{ with } A_{i\alpha} = 1\},$$

i.e., all the subtrates required by p_i (for its growth).

Problem setup: dependency relation

Define a weighted directed graph on P:

$$W_{i,j} = |\mathcal{O}_i \cap \mathcal{I}_j| .$$

We say

$$p_i \to p_j$$

if $W_{i,j} > 0$. This should be quite obvious; it means the outcome of p_i is also a requirement of p_j. Thus we say p_i and p_j are related.

Remark

Normalized W:

$$\overline{W_{i,j}} = \frac{W_{i,j}}{\sum_{k=1}^{n} W_{i,k}}.$$

Structural theorems: existence of a graph describing psychrophile interdependency

Theorem

Given (P, E, S, φ, A), there exists a canonical directed weighted graph:

$$G_{PP} = (P, W)$$

encoding all metabolic interdependencies.

Structural theorems: dependency & mutual dependency

Theorem (Dependency)

There exists a path $p_i \to p_j$ iff there exists a path

$$p_i \to e_\alpha \to s \to e_\beta \to p_j$$

in the tripartite structure.

Remark

Note that the tripartite structure here is $P \cap E \cap S$.

Theorem (Mutual dependency)

Two psychrophiles are mutually dependent if they lie in a directed cycle of G_{PP}.

Few remarks

Remark (Self-sustaining metabolic consortium)

Each strongly connected component of G_{PP} corresponds to a self-sustaining metabolic consortium.

Remark

If p_i maximizes

$$\sum_j W_{i,j},$$

then p_i is a maximal provider.

Dynamical system

Definition (population dynamics)

Let $x_i(t)$ be the population of p_i at time t. Define

$$\text{growth rate} = \frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i \left(r_i + \sum_j W_{j,i} x_i - \sum_j C_{i,j} x_i \right)$$

where r_i is the intrinsic growth rate for p_i, $C_{i,j}$ is the competition relation.

Refinement of the weight

From the growth rate expression, the effect of p_i on p_j is

$$\frac{\partial}{\partial x_i}\left(\frac{1}{x_j}\frac{\mathrm{d}x_i}{\mathrm{d}t}\right) = W_{i,j}.$$

This will lead to

$$W_{i,j} = \frac{x_i \sum_{s\in\mathcal{O}_i\cap\mathcal{I}_j} w_s}{\sum_k x_k \sum_{s\in\mathcal{O}_k\cap\mathcal{I}_j} w_s}$$

where w_s is the importance of each metabolic. The measurement and definition can be discussed.

Further refinements

Not all three parts of this model will be known. Constructing steps will be based on the workflow outcome.

Introduction to PageRank Optimization

Let $\mathcal{W} = \{W_1, W_2, \ldots, W_N\}$ be a set of matrices, where each $W_i \in \mathbb{R}^{d \times n}$ (or $\mathbb{R}^{n \times n}$). We seek to minimize the following objective over a matrix M:

$$\mathcal{L}(M) = \sum_{i=1}^{N} \sum_{j=1}^{N} \|MW_i - W_j\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

What are W and M in our context

- W represents the metabolic network betweeen different psychrophiles.
- M is what we're trying to find. It's a transformation—a mathematical operation that, when applied to one microbe's enzyme profile, predicts another microbe's enzyme profile.

What are we trying to minimize

Our goal is to find the M that makes this transformation as accurate as possible across our entire dataset. We measure accuracy using $\mathcal{L}(M)$ in PageRank Optimization.

In plain language:
For every pair of microbes in our collection, we apply M to the first microbe's enzyme profile, compare it to the second microbe's actual enzyme profile, calculate the difference, add it up across all possible pairs, and then ask:

Which M makes this total difference as small as possible?

Mathematical derivation

We have

$$\|MW_i - W_j\|^2 = \operatorname{tr}(W_i^\top M^\top M W_i) - 2\operatorname{tr}(W_j^\top M W_i) + \operatorname{tr}(W_j^\top W_j).$$

When we sum over i and j, the first term on the right hand side gives

$$N \operatorname{tr}\left(M^\top M \sum_i W_i W_i^\top\right).$$

The second term gives

$$-2\operatorname{tr}\left(M \sum_{i,j} W_i W_j^\top\right).$$

The third term is constant with respect to M, so it doesn't affect the minimizer.

Mathematical derivation

Then we define

$$S_1 = \sum_i W_i W_i^\top$$

$$S_2 = \sum_{i,j} W_i W_j^\top = A A^\top \text{ where } A = \sum_i W_i.$$

Taking the derivative and setting it to zero yields:

$$2NMS_1 - 2S_2^\top = 0 \quad \Rightarrow \quad M = \frac{1}{N} S_2^\top S_1^{-1}$$

If S_1 is invertible, this gives us the optimal transformation.

Expectation

- Functional variation: Seeing how the microbes' "toolkits" change as the environment changes.
- Psychrophile network: Finding evidence that one microbe's waste is another microbe's food (complementarity).
- Community survival: Showing how these microbes rely on each other to survive the extreme cold.

Significance

- Linking specific functions to specific organisms.
- Revealing true ecological network.
- The big picture: Moving beyond just making a list of “who is there,” to actually understanding “how they survive together.”