Experimental Objective in September : Organize all experimental data during the period and begin writing the final report and creating the wiki.

Week 1: 9.1-9.7

(A)Rationale and Experimental Design for Whole-Genome Sequencing (WGS)

1. Rationale

Through previous rounds of screening and engineering, we have successfully isolated strains—notably Strain 3-4 and Strain 10-7—that exhibit significantly enhanced production capabilities. While their superior phenotypes are well-characterized, the underlying genetic alterations responsible for these improvements remain unknown. To transition from empirical screening to rational metabolic engineering, it is imperative to elucidate the precise molecular changes within these strains. Therefore, we have opted to employ whole-genome re-sequencing, the gold standard for comprehensively identifying genomic variations, to create a detailed map of the genetic landscape of our evolved strains compared to the wild-type.

2. Experimental Design

To achieve our objective, a carefully designed re-sequencing experiment was planned as follows:

Sample Selection: Three key strains were selected for sequencing to enable a robust comparative analysis:

Wild-Type (WT): It is the parental strain, which *trpR* and *tnaAB* were knocked out.

Strain 3-4: The first high-performing mutant strain in third generation.

Strain 10-7: The second high-performing mutant strain in tenth generation.

Technology and Strategy: We selected the Illumina high-throughput sequencing platform for its high accuracy and depth of coverage. A Paired-End 150 bp (PE150) sequencing strategy was chosen to facilitate accurate read mapping, improve the detection of structural variants, and resolve repetitive genomic regions.

Collaborating Partner: The sequencing service and primary bioinformatics analysis will be professionally conducted by Sangon Biotech (Shanghai) Co., Ltd., ensuring high-quality data generation and processing according to industry best practices.

Week 2: 9.8 -9.14

(A) Assessment of Raw Sequencing Data

Upon completion of the sequencing run, we received the primary quality metrics for the raw data.

Table 1: Summary of Raw Sequencing Data Output

Sample ID	Total Reads	Total Bases (Gb)	Q30 Bases (%)	GC Content (%)
WT	6,447,452	0.97	94.82%	51.12%
3-4	6,936,556	1.04	94.78%	51.42%
10-7	6,328,836	0.95	94.97%	51.24%

The sequencing run generated a substantial volume of data for each sample. Critically, the Q30 base ratio, which represents the percentage of bases with a base-calling accuracy of 99.9% or higher, exceeded 94% for all samples, indicating a high quality of sequencing data..

(B) Assessment of Clean Data after Quality Control

To ensure maximum accuracy in variant calling, the raw data underwent a stringent quality control (QC) process.

1. Quality Control Procedure:

a.The Trimmomatic software was used to:

b.Remove Illumina adapter sequences.

c.Trim leading and trailing low-quality bases (Phred score < 20).

d.Scan the reads with a sliding window and trim when the average quality drops below a threshold.

e.Discard any reads that became shorter than 35 base pairs after trimming.

2. Post-QC Data Statistics:

This filtering step selectively removed low-quality data, resulting in a final, high-confidence "Clean Reads" dataset for analysis.

Table 2: Summary of Clean Data Statistics after QC

Sample ID	Total Reads	Total Bases (Gb)	Q30 Bases (%)	GC Content (%)
WT	5,964,902	0.88	98.06%	51.19%
3-4	6,409,472	0.94	97.97%	51.49%
10-7	5,877,234	0.87	98.05%	51.32%

The QC process successfully improved the overall quality of the dataset, with the Q30 ratio for all samples increasing.

(C) Rationale

While our experimental data clearly demonstrates the superior performance of the mutant AtCOMT, the structural basis for this enhancement is unknown. To understand why these mutations are beneficial, it is necessary to investigate their impact on the enzyme's three-dimensional structure and its ability to bind the substrate in a catalytically productive manner. Molecular docking is a powerful computational tool that predicts the preferred binding orientation and affinity of a ligand within a protein's active site. By comparing the docking results of the Wild-Type (WT) enzyme with our engineered mutant (Lys187Arg), we can generate a plausible, structure-based hypothesis for the observed increase in activity.

(D) Homology Modeling and Structure Preparation

1. Homology Modeling of Wild-Type (WT) AtCOMT:

The foundation of any docking study is an accurate protein structure. In the absence of an experimentally determined crystal structure for AtCOMT, a high-fidelity 3D model was generated using the AlphaFold2 deep learning network. This state-of-the-art tool is renowned for its exceptional accuracy in predicting protein structures from their primary amino acid sequence, providing a reliable template for our wild-type enzyme.

Figure 1. The aligned result of 3'-O-methyltransferase (PDB ID:1kyz) and homologous constructed wild-type protein

2. Mutant Structure Generation and Refinement:

The identified beneficial mutation was introduced into the wild-type AtCOMT model in silico using standard molecular modeling software. To ensure the resulting structure was physically realistic, it was subjected to energy minimization. This crucial refinement step relaxes any potential steric clashes introduced by the mutation and allows the surrounding amino acid side chains to reorient into their most energetically favorable positions, yielding a stable and accurate conformation of the mutant enzyme for subsequent docking studies.

Week 3: 9.15 - 9.21, 2025

(A) Global Analysis of Variant Loci

The primary analysis identified all single nucleotide polymorphisms (SNPs) and short insertions/deletions (Indels) in each strain relative to the reference genome. The results revealed a dramatic difference in the mutational landscape between the wild-type and the evolved strains.

The wild-type strain contains only 7 variants, representing a stable genetic background. In stark contrast, Strain 3-4 and Strain 10-7 accumulated 62 and 186 variants, respectively. This significant increase in the number of mutations is a key finding, providing direct evidence that our engineering and/or screening processes induced substantial genomic evolution.

(B) Screening for Strain-Specific Mutations

To isolate the mutations most likely responsible for the improved phenotypes, we performed a comparative analysis to identify variants unique to each high-performing strain.

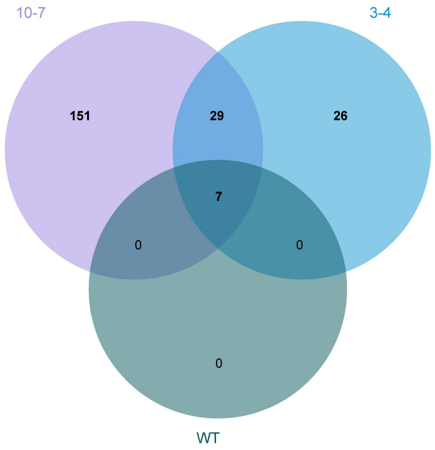

Figure 1. Venn diagram illustrating the distribution of shared and unique variants among the WT, Strain 3-4, and Strain 10-7 genomes.

Key Findings from Comparative Analysis:

Shared Background: A core set of 7 variants is shared among all three strains, representing the baseline genotype of our parental WT strain.

Unique Candidate Mutations: The analysis successfully filtered out the background mutations, revealing the most compelling candidates for further study. Strain 4 possesses 55 unique variants not present in the WT, while Strain 7 possesses 179 unique variants. This catalog of strain-specific mutations forms the primary dataset for our subsequent functional analysis.

(C) Molecular Docking Simulation and Results

1. Docking Procedure:

Both the wild-type and the energy-minimized mutant protein structures were prepared for docking by adding polar hydrogens and assigning charges. The substrate molecule was then docked into the putative active site of each enzyme using AutoDock Vina. Given that the enzyme's catalytic mechanism is known to be dependent on a critical Histidine (His) residue, our analysis of the resulting docking poses was specifically focused on conformations that placed the substrate in close proximity to this catalytic residue.

2. Docking Results and Comparative Analysis:

The docking simulations yielded distinct, low-energy binding poses for the substrate in the WT versus the mutant enzyme, suggesting a tangible structural impact of the mutation. The results are visualized below.

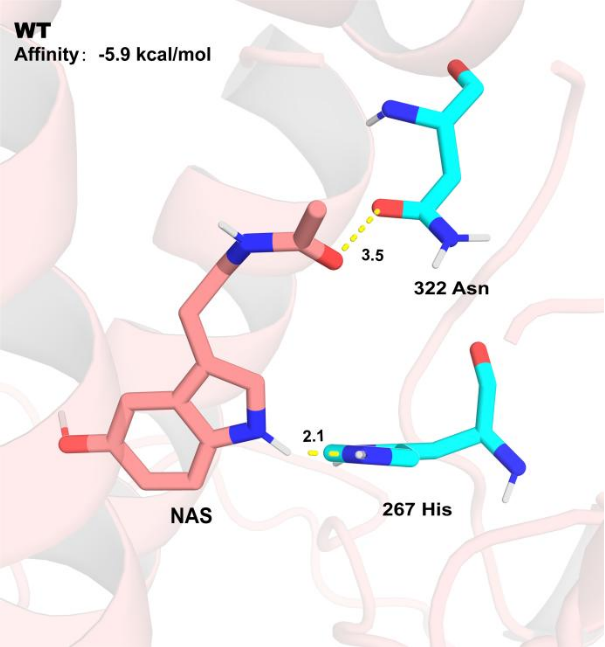

Figure 1. Molecular docking of the substrate into the active site of the Wild-Type (WT) AtCOMT enzyme. The figure displays the predicted lowest-energy binding pose of the substrate within the catalytic pocket of the WT enzyme.

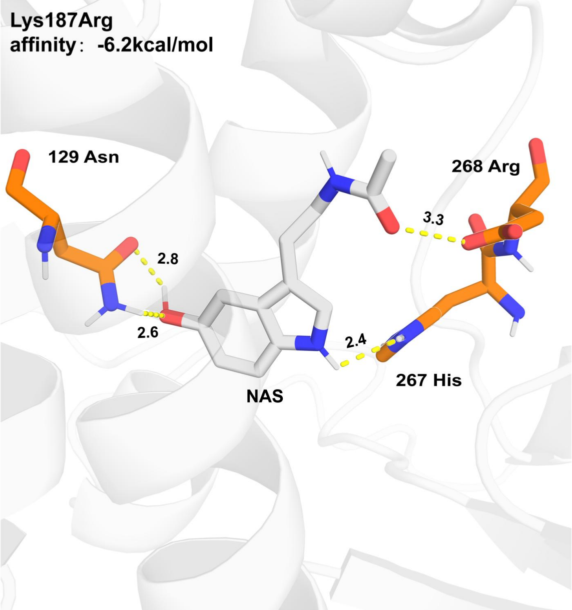

Figure 2. Molecular docking of the substrate into the active site of the mutant (Lys187Arg) AtCOMT enzyme. The simulation predicts a noticeably different binding orientation for the substrate compared to the WT, likely influenced by the altered geometry and electrostatics of the active site.

Week 4:9.22 - 9.28

(A) Functional Enrichment Analysis of Mutated Genes

To understand if the identified mutations were randomly distributed across the genome or concentrated in specific biological systems, we performed a Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) functional enrichment analysis on the list of genes containing non-synonymous variants.

(B)Prioritization of Key Mutations for Tryptophan Overproduction

By cross-referencing our list of unique, high-impact variants with established literature and the KEGG pathway database, we have distilled the extensive list of mutations into a high-priority set of candidate genes. These genes are categorized based on their known roles in tryptophan biosynthesis and related pathways.

Table 3 Summary of Filtered Sequencing Data Statistics

Sample ID	Total Reads	Total Bases (Gb)	Q30 Bases (%)	GC Content (%)
WT	5,964,902	0.88	98.06%	51.19%
3-4	6,409,472	0.94	97.97%	51.49%
10-7	5,877,234	0.87	98.05%	51.32%