Project Integration Report: Applied Genomics Pipeline for Bioplastic-Degrading Fungi

Martina Castellucci
University of Bologna

1 Introduction and Integration Rationale

This project simulates the design of a multi-omics workflow to identify fungal isolates from PLA/PHA-enriched compost with potential for bioplastic degradation. Each methodological step is motivated by theoretical principles of genomics and biotechnology, considering alternative options and trade-offs. The purpose is not only to describe the workflow but also to demonstrate critical reasoning behind each technical choice.

2 Sampling and Screening

Choice: Compost with visible PLA/PHA residues; screening on PDA + PLA + Rhodamine B. **Rationale:** Environments enriched with target polymers are more likely to host adapted organisms. Rhodamine B provides a direct assay for extracellular esterase activity, allowing rapid screening. **Alternative:** Metagenomics without cultivation; however, this would prevent downstream functional assays.

3 DNA and RNA Extraction

Choice: CTAB-based DNA extraction for long-read compatibility; RNA extracted under Control, PLA, PHA conditions (replicates). **Rationale:** Long-read sequencing requires intact DNA fragments (>20 kb). RNA under induced vs control conditions enables expression analysis of degradative genes. **Alternative:** Commercial kits (e.g. Qiagen) – faster but more expensive.

4 Sequencing Strategy

Choice: Illumina NovaSeq PE150 (high accuracy) + Nanopore GridION (long reads). **Rationale:** Illumina ensures base-level accuracy; Nanopore resolves repeats and structural variants. Hybrid sequencing balances contiguity and accuracy. **Alternatives:** - PacBio HiFi – highly accurate long reads, but costly. - Illumina-only – cheaper, but fragmented assemblies. - Nanopore-only – more contiguous, but high error rates.

5 Genome Assembly and QC

Choice: Flye (OLC algorithm for Nanopore) + Pilon (Illumina polishing). QC via QUAST and BUSCO. **Rationale:** Flye is optimized for noisy long reads. Pilon corrects systematic indels. QUAST and BUSCO are standard metrics for contiguity and completeness. **Alternatives:** Canu (robust but slower); MaSuRCA (hybrid, requires high coverage).

6 Genome Annotation

Choice: MAKER3 (integrating Augustus, GeneMark-ES, RNA-Seq). Functional annotation with dbCAN3 (CAZymes). Manual curation in Geneious. **Rationale:** MAKER3 integrates ab initio prediction with transcript evidence for accuracy. dbCAN3 identifies CAZymes critical in polyester hydrolysis. **Alternatives:** BRAKER2 (faster, RNA-Seq guided only).

7 Transcriptomic Profiling

Choice: RNA-Seq (12 libraries, 3 conditions $\times$ 4 replicates). Quantification with Salmon, DEG analysis with DESeq2. **Rationale:** Replicates ensure statistical robustness. Salmon provides fast quasi-mapping with bias correction. DESeq2 uses a negative binomial model, controlling for false discovery. **Alternatives:** edgeR (effective for small replicates); limma-voom (for high replication).

8 Comparative Genomics and Phylogenomics

Choice: OrthoFinder (orthogroups), MCScanX (synteny), RAxML (phylogeny). **Rationale:** OrthoFinder robustly defines core vs accessory genes. MCScanX identifies

conserved collinear blocks. RAxML reconstructs phylogenies via maximum likelihood. **Alternatives:** OrthoMCL, IQ-TREE.

9 Integration and Expected Insights

Multi-omics integration ensures that predictions (genomic content) are supported by expression (RNA-Seq) and contextualized by evolution (comparative genomics). This reflects the applied genomics workflow: from DNA sequence → functional evidence → biotechnological applications.

10 Strengths and Limitations

Strengths: - Hybrid sequencing ensures both contiguity and accuracy. - Multi-omics integration provides functional validation. - Comparative context supports ecological adaptation.

Limitations: - Reliance on homology-based annotation (possible overestimation). - Moderate genome completeness in hybrid assemblies. - No direct enzymatic validation yet.

—

11 Summary Tables

Table 1: Tools and Theoretical Basis

Tool	Purpose	Theoretical Basis	Alternatives
Flye	Long-read assembly	OLC algorithm	Canu, MaSuRCA
Pilon	Assembly polishing	Short-read consensus correction	Racon, Medaka
QUAST	Assembly QC	Contiguity metrics (N50, L50)	None widely used
BUSCO	Completeness	Conserved single-copy orthologs	CEGMA (obsolete)
MAKER3	Structural annotation	Ab initio + evidence integration	BRAKER2
dbCAN3	CAZyme annotation	HMM profiles for enzyme families	Hotpep

Salmon	Quantification	Lightweight quasi-mapping	Kallisto
DESeq2	DEG analysis	Negative binomial model	edgeR, limma
OrthoFinder	Orthology inference	Graph clustering of proteins	OrthoMCL
MCScanX	Synteny	Gene order conservation	Mauve
RAxML	Phylogeny	Maximum likelihood inference	IQ-TREE, MrBayes

Table 2: Sequencing Strategies Comparison

Platform	**Strengths**	**Weaknesses**	**Best Use**
Illumina	High accuracy, cheap	Short reads, fragmented assemblies	Polishing, RNA-Seq
Nanopore	Long reads, real-time	Higher error rate	Structural assembly
PacBio HiFi	Long + accurate	Expensive, less accessible	Gold-standard assemblies
Sanger	Extremely accurate	Low throughput	Validation only

—

12 Possible Oral Questions and How to Answer

General

- Why use a hybrid sequencing approach instead of only Illumina? *Because Illumina is accurate but short, leading to fragmented assemblies, while Nanopore is long but error-prone. Combining them gives the best of both.*
- Why fungi and not bacteria? *Fungi naturally secrete extracellular enzymes in large quantities, making them better candidates for industrial biodegradation.*

Assembly and QC

- What does N50 mean and why is it important? *It measures assembly contiguity: half of the genome lies in contigs at least N50 long. A higher N50 = more contiguous assembly.*
- Why is BUSCO a gold standard for completeness? *It checks for the presence of highly conserved orthologs expected in all fungi. Missing BUSCOs suggest gaps or misassemblies.*

Annotation and Transcriptomics

- How does dbCAN3 detect CAZymes? *It uses HMM profiles built from curated enzyme families, detecting conserved motifs even with low sequence identity.*
- Why use DESeq2 and not simply fold change? *Because RNA-Seq data are count-based and overdispersed; DESeq2 models variance with a negative binomial distribution and controls FDR.*

Comparative and Evolutionary

- Why is synteny analysis important? *It reveals conservation of genome structure, helping to detect rearrangements linked to adaptation.*
- What is the difference between orthologs and paralogs? *Orthologs diverged by speciation, usually same function; paralogs by duplication, often divergent function.*

Applied Aspects

- How does this align with EU bioeconomy goals? *Because enzymes from fungi could improve bioplastic composting, reducing microplastic accumulation and supporting circular economy.*
- What would you improve with more budget? *Use PacBio HiFi for chromosome-level assemblies, add proteomics and metabolomics for functional validation.*