Applied Genomics Extended Study Manual: Hybrid Genome Assembly and Multi-Omics Analysis of *Purpureocillium lilacinum* PLA-C1

Prepared for Oral Examination

Contents

1 Introduction and Scientific Rationale

Bioplastics such as polylactic acid (PLA) and polyhydroxyalkanoates (PHA) are widely presented as environmentally friendly alternatives to conventional petrochemical plastics. Their appeal lies in the concept of biodegradability, yet their actual environmental fate is often less satisfactory than expected. In industrial composting environments, PLA and PHA do not always degrade completely; instead, they can persist for extended periods and fragment into microplastics. This inconsistency between marketing claims and ecological outcomes represents a critical challenge in applied biotechnology and environmental genomics.

The rationale behind this project is to explore the enzymatic potential of fungi, which are well known for their ability to secrete extracellular hydrolytic enzymes. By sequencing, assembling, annotating, and analyzing the genome and transcriptome of a compost-derived isolate of *Purpureocillium lilacinum*, named PLA-C1, we can investigate whether this fungus possesses specialized enzymes—such as esterases, cutinases, and lipases—that may contribute to PLA or PHA degradation. The project thus integrates modern genomic approaches with applied goals in environmental sustainability and bioeconomy.

This manual does not merely summarize the project but expands it into a comprehensive theoretical guide. It connects experimental strategies with the theoretical frameworks taught in applied genomics courses, making it both a case study and a study tool for understanding how sequencing, assembly, annotation, transcriptomics, and comparative genomics function together to address real-world biotechnological questions.

—

2 Choice of Organism: Why *Purpureocillium lilacinum*?

Purpureocillium lilacinum is a filamentous ascomycete fungus with a cosmopolitan distribution, frequently isolated from soils, compost, and rhizosphere environments. Its ecological versatility is remarkable, as it can exist as a saprotroph, degrading decaying organic matter, or as a nematophagous parasite, attacking nematode eggs and larvae. It can also live as an endophyte inside plant tissues, suggesting a broad enzymatic arsenal that allows it to adapt to various substrates and ecological contexts.

From a genomic perspective, fungi with such multitrophic lifestyles often harbor large and diverse repertoires of carbohydrate-active enzymes (CAZymes) and other hydrolases. Indeed, *P. lilacinum* has been previously reported to secrete cutinases, esterases, and lipases, enzymes mechanistically capable of cleaving ester bonds similar to those found in

PLA and PHA polymers.

The isolate studied in this project, PLA-C1, was obtained directly from compost supplemented with PLA/PHA films. This selective environment increases the likelihood that surviving organisms have adapted to utilize these synthetic polymers as carbon sources. By sequencing and analyzing PLA-C1, the project aims to uncover whether this adaptation is reflected at the genomic and transcriptomic levels.

Alternative organisms could have been considered, such as bacteria like *Ideonella sakaiensis*, famous for degrading PET plastics. However, fungi present a unique advantage: their natural capacity for extracellular secretion makes them attractive for industrial enzyme production, where large-scale secretion is a prerequisite. This justifies the choice of a fungal system over a bacterial one for bioplastic degradation studies.

—

3 Sequencing Technologies: Theoretical Foundations and Project Choices

The success of any applied genomics project begins with the choice of sequencing technologies. Over the past decades, sequencing has evolved through three major generations. First-generation sequencing, represented by Sanger's chain-termination method, provided extremely accurate reads but at very limited throughput, making it unsuitable for whole-genome projects. Second-generation sequencing, dominated by Illumina and Ion Torrent platforms, introduced massive parallelization, producing billions of short reads with high accuracy at dramatically reduced costs. Third-generation sequencing technologies, such as Oxford Nanopore and Pacific Biosciences, moved toward single-molecule sequencing, providing long reads that can span repetitive regions and resolve structural complexity.

In this project, we used a hybrid strategy combining Illumina short reads with Oxford Nanopore long reads. The reasoning behind this choice is rooted in the theoretical strengths and weaknesses of each technology.

Illumina sequencing works by first fragmenting DNA and ligating it to platform-specific adapters. These fragments are immobilized on a flowcell surface and undergo bridge amplification, producing dense clusters of identical molecules. Sequencing proceeds by synthesis: nucleotides carrying both a fluorescent label and a reversible terminator are incorporated one at a time by DNA polymerase. Each cycle produces an image of the incorporated bases, and the terminators are chemically removed before the next cycle. This process yields highly accurate reads, typically 150–300 bp in paired-end mode, with error rates below 0.1%. However, the short read length makes Illumina sequencing unsuitable for resolving long repeats or complex structural variants.

In contrast, Oxford Nanopore sequencing works by passing single-stranded DNA through a biological nanopore embedded in a membrane. A motor protein regulates the speed of translocation, and the ionic current across the pore is measured in real time. Each nucleotide produces a characteristic disruption of the current, allowing sequence inference. Nanopore reads can reach tens to hundreds of kilobases, with N50 often in the range of 10–20 kb. This capacity allows the resolution of repetitive regions and complex structural arrangements, something impossible with Illumina alone. However, Nanopore reads are prone to higher error rates (5–10%), particularly indels in homopolymeric stretches.

Pacific Biosciences HiFi sequencing represents another third-generation alternative. Its principle relies on Single-Molecule Real-Time (SMRT) sequencing, where DNA templates are circularized into SMRTbell constructs and repeatedly sequenced inside zero-mode waveguides. The multiple passes of the same molecule are then collapsed into a highly accurate consensus read (HiFi read), typically longer than 15 kb with >99.8% accuracy. PacBio HiFi thus provides the best of both worlds: long read length and high accuracy. Yet, it remains more costly and less accessible than Nanopore, which influenced the project design.

Ion Torrent, based on semiconductor sequencing, detects hydrogen ions released during nucleotide incorporation. While less expensive, it is prone to errors in homopolymer tracts and is now less commonly used for de novo fungal genome projects. Sanger sequencing, though still the most accurate method, remains limited to targeted sequencing and validation tasks.

By combining Nanopore and Illumina, the project exploits the complementary strengths of both: Nanopore long reads provide contiguity, while Illumina short reads correct the high error rate. This hybrid strategy is now widely considered the most cost-effective way to produce accurate fungal genome assemblies with reasonable budgets.

—

4 Genome Assembly: Algorithms, Challenges, and Quality Metrics

Genome assembly is the process of reconstructing a genome from short or long sequencing reads. Its theoretical foundations lie in graph-based algorithms designed to represent overlaps among sequences. Two major paradigms dominate assembly: the de Bruijn Graph (DBG) approach and the Overlap-Layout-Consensus (OLC) approach.

In de Bruijn graph assembly, reads are decomposed into fixed-length words called k-mers. Nodes in the graph represent k-mers, and edges represent overlaps of k-1 bases. This method is efficient for handling massive numbers of short reads, such as those produced by

Illumina. However, DBG assembly is sensitive to sequencing errors, which create spurious branches, and it struggles with repeats longer than the read length.

OLC assembly, on the other hand, is better suited to long reads. The first step computes all pairwise overlaps between reads. The layout stage organizes reads into a graph reflecting their overlaps, and the consensus stage derives the final sequence. OLC is computationally more demanding but handles long, noisy reads effectively. This is why OLC-based tools such as Flye are ideal for Nanopore data.

In this project, we employed Flye for long-read assembly, followed by polishing with Pilon using Illumina reads. Polishing corrects the high rate of insertion-deletion errors typical of Nanopore sequencing. Iterative rounds of polishing gradually increase consensus accuracy, approaching Illumina-level base quality.

Assembly quality is typically measured by several metrics. Coverage depth is defined by $C = \frac{L \times N}{G}$, where L is read length, N the number of reads, and G the genome size. Adequate coverage (usually >30X for long reads, 60–100X for short reads) is required for reliable assemblies. The N50 statistic measures contiguity: it is the length L such that 50% of the assembled genome lies in contigs of at least L. Higher N50 values reflect better continuity. Completeness is assessed by BUSCO, which searches for Benchmarking Universal Single-Copy Orthologs conserved in most fungi. BUSCO categories include Complete, Single-copy, Duplicated, Fragmented, and Missing. High BUSCO completeness (above 90%) is considered excellent; lower values indicate missing or fragmented genes.

In our case, the assembly size was 38.6 Mb with an N50 of 5.3 Mb, demonstrating high contiguity. BUSCO completeness of 76.3% suggests that some genes remain missing or fragmented, possibly due to heterozygosity or unresolved repeats. While this is not perfect, it is acceptable for a first hybrid assembly of a non-model fungus.

—

5 Genome Annotation: From Gene Models to Function

Genome annotation transforms raw assemblies into biologically interpretable information. It involves two main steps: structural annotation, which identifies gene models, and functional annotation, which predicts gene functions.

Structural annotation can be performed using ab initio methods, which rely on intrinsic features such as codon usage bias, splice site motifs, and gene length distributions. Tools such as AUGUSTUS and GeneMark are examples of this approach. However, ab initio predictions alone often produce errors in intron-exon boundaries or miss short genes. Extrinsic methods, by contrast, use external evidence such as RNA-Seq alignments or homology to known proteins. Integrating both is considered best practice, which is why pipelines like MAKER3 combine ab initio predictors with RNA-Seq evidence.

Functional annotation assigns biological meaning to predicted gene models. The most common approach is sequence similarity searching via BLAST against protein databases, followed by domain identification using tools like InterProScan. For carbohydrate-active enzymes, a specialized approach is needed: dbCAN3 uses hidden Markov models to identify glycoside hydrolases, esterases, cutinases, carbohydrate esterases, and auxiliary enzymes. This is particularly relevant to our project because PLA and PHA degradation requires ester bond hydrolysis.

Annotation is not free from limitations. Homology-based annotation may propagate errors from poorly characterized proteins. Some predicted genes may appear to belong to CAZyme families but lack enzymatic activity in reality. Experimental validation is ultimately needed, but annotation provides a crucial first hypothesis.

—

6 Transcriptomics: RNA-Seq and Differential Expression

Transcriptomics investigates how gene expression changes under different conditions, providing insight into functional adaptation. RNA-Seq has become the standard technique, replacing microarrays.

In RNA-Seq, RNA is first extracted and converted into complementary DNA (cDNA). Libraries are constructed either by enriching mRNA through poly-A selection or by depleting rRNA to retain a broader transcript pool. Libraries are sequenced, and reads are then mapped or quasi-mapped back to the transcriptome or genome.

Expression quantification is commonly performed in terms of TPM (Transcripts Per Million) or RPKM (Reads Per Kilobase per Million). However, differential expression analysis requires statistical modeling of count data. DESeq2, a widely used tool, models raw counts as following a negative binomial distribution, which accommodates both the mean-variance relationship and overdispersion typical of RNA-Seq data. Dispersion parameters are estimated across genes, and normalization is performed using size factors to account for differences in sequencing depth. Multiple testing is addressed by controlling the False Discovery Rate (FDR), often using the Benjamini–Hochberg procedure. Standard thresholds are $\log_2$ Fold Change ≥ 2 and FDR <0.05.

In this project, Salmon was used for quantification because it implements fast quasi-mapping and bias correction. DESeq2 then identified differentially expressed genes between control, PLA, and PHA conditions. Results showed that PLA induced 84 genes, PHA induced 51, and 29 were common to both. Many of these were annotated as esterases, lipases, and cutinases, supporting the hypothesis that PLA-C1 responds specifically to polymer presence by inducing hydrolase genes.

Transcriptomics provides strong evidence of gene regulation, but expression does not guarantee enzymatic activity. Functional assays would be necessary for full validation.

—

7 Comparative Genomics and Evolutionary Insights

Comparative genomics situates the genome of interest within its evolutionary context. It can reveal conserved orthologous genes, lineage-specific expansions, and unique adaptations.

Orthology refers to genes diverged through speciation events, typically retaining similar functions across species. Paralogy refers to genes diverged through duplication within a genome, often giving rise to functional diversification. Tools like OrthoFinder group genes into orthogroups that reflect these relationships.

Synteny analysis, performed with tools like MCScanX, examines whether gene order is conserved between genomes. Conservation of synteny indicates evolutionary stability, whereas disruptions may signal rearrangements or adaptive changes.

Phylogenomics uses multiple conserved genes to reconstruct evolutionary histories. In this project, RAxML was used to infer a maximum likelihood phylogeny based on single-copy orthologs, confirming the placement of PLA-C1 within the *Purpureocillium* clade.

Results revealed that PLA-C1 shared 10,312 orthogroups with related fungi but also contained 314 unique genes, potentially reflecting niche adaptation to polymer-rich compost. Such isolate-specific genes may encode novel enzymes relevant to PLA/PHA degradation.

—

8 Functional Genomics and Multi-Omics Integration

While genome and transcriptome analysis provide predictions, functional genomics integrates multiple omics layers to validate and expand these insights.

Proteomics, typically based on LC-MS/MS, measures the actual proteins expressed and can confirm whether predicted enzymes are secreted under PLA or PHA conditions. Metabolomics, using NMR or LC-MS, tracks metabolic intermediates, potentially revealing breakdown products of PLA or PHA. Integration of genomics, transcriptomics, proteomics, and metabolomics—so-called multi-omics—offers the most complete picture of fungal adaptation and enzymatic activity.

Although this project focused on genome and transcriptome data, future directions could include proteomic validation of secreted hydrolases and metabolomic analysis of polymer degradation products.

—

9 Applications in Biotechnology and Bioeconomy

The discovery of new fungal enzymes capable of degrading synthetic polymers has significant applications. Lipases and cutinases are already widely used in detergents, textiles, and food processing. Extending their use to bioplastic degradation could improve recycling and composting. Patents already exist for fungal cutinases used to depolymerize PET, indicating clear industrial interest.

From a policy perspective, this work aligns with the European Green Deal and the EU Circular Economy Action Plan, both of which emphasize the development of sustainable biotechnologies to reduce plastic pollution. Funding initiatives under Horizon Europe target microbial and enzymatic solutions for plastic recycling, placing fungal genomics directly in line with European bioeconomic priorities.

—

10 Results and Critical Interpretation

The hybrid assembly of PLA-C1 yielded a 38.6 Mb genome with an N50 of 5.3 Mb, demonstrating strong contiguity. BUSCO completeness of 76.3% suggests further improvements are possible with higher coverage or complementary technologies like PacBio HiFi or Hi-C scaffolding.

Annotation identified 272 CAZymes, many induced under PLA conditions. Transcriptomics revealed condition-specific induction of hydrolases, supporting the ecological relevance of PLA-C1 in polymer degradation. Comparative genomics highlighted both shared core genes and isolate-specific content, suggesting niche adaptation.

Limitations include reliance on homology-based annotation, incomplete genome representation, and the lack of biochemical validation. Nevertheless, the project demonstrates the feasibility of combining hybrid assembly with multi-omics to uncover enzymatic potential in non-model fungi.

—

11 Extended Glossary and Key Formulas

PHRED score $Q = -10 \log_{10} P$. Q20 corresponds to 99% accuracy (1 error in 100 bases), Q30 to 99.9% accuracy (1 error in 1000 bases).

Coverage (C) $C = \frac{L \times N}{G}$, where L = read length, N = number of reads, G = genome size. At least 30X coverage is required for basic assemblies, 60–100X recommended.

N50 Contig length L such that 50% of the assembly is contained in contigs $\geq L$. Higher N50 indicates greater contiguity.

BUSCO Benchmarking Universal Single-Copy Orthologs; measures genome completeness relative to conserved orthologs. >90% excellent, 70–90% acceptable.

Orthology Genes diverged through speciation, usually retaining function.

Paralogy Genes diverged through duplication, may acquire new functions.

Pan-genome The union of core and accessory genes across multiple strains.

Synteny Conservation of gene order across species, reflecting evolutionary constraints.

Negative Binomial Probability distribution used by DESeq2 to model RNA-Seq count data with overdispersion.

FDR False Discovery Rate, probability of type I error across multiple tests; standard cutoff <0.05.

TPM/RPKM/FPKM Methods for normalizing RNA-Seq expression values; TPM is most comparable across samples.

Michaelis–Menten $v = \frac{V_{max}[S]}{K_m + [S]}$, describes enzyme kinetics.

Hardy–Weinberg $p^2 + 2pq + q^2 = 1$, describes expected genotype frequencies under random mating.

Hi-C scaffolding Uses chromatin conformation capture to order and orient contigs into chromosome-scale scaffolds.

Circular Economy EU policy promoting sustainable design, recycling, and biotechnology for plastic waste.

Formula Recap

- PHRED: $Q = -10 \log_{10} P$
- Coverage: $C = \frac{L \times N}{G}$
- N50 definition

- Hardy–Weinberg equilibrium: $p^2 + 2pq + q^2 = 1$
- Michaelis–Menten: $v = \frac{V_{max}[S]}{K_m+[S]}$
- $\log_2$ Fold Change: $\log_2 \frac{Expr_{cond1}}{Expr_{cond2}}$

Appendix: Comparative Tables

Table 1: Sequencing Technologies Comparison

Technology	Read Length	Accuracy	Throughput	Main Limitation
Sanger	800–1000 bp	>99.9%	Very low	Not scalable for WGS
Illumina	150–300 bp PE	99.9%	Very high (100s Gb/run)	Short reads, cannot span repeats
Ion Torrent	200–400 bp	98–99%	Moderate	Homopolymer errors
Nanopore	10–100 kb+ (N50 ∼20 kb)	90–95% (raw)	Moderate–high	High error rate (indels)
PacBio HiFi	>15 kb	99.8%	High (20–30 Gb/run)	Cost, lower access

—

Table 2: Genome Assembly Tools Comparison

Assembler	Algorithm	Input	Strengths	Limitations
SPAdes	de Bruijn Graph	Illumina	Good for bacterial genomes	Fails with complex repeats
Flye	OLC	Long reads (Nanopore/PacBio)	Robust to noisy data	May misassemble repeats
Canu	OLC + correction	Long reads	Strong error correction	Computationally expensive
MaSuRCA	Hybrid (DBG+OLC)	Illumina + long reads	Flexible hybrid assemblies	Requires high coverage

Hi-C scaffolding	Graph-based ordering	Assembled contigs + Hi-C	Chromosome-scale scaffolds	Needs extra data

—

Table 3: RNA-Seq Normalization Methods

Method	Normalization Principle	Strengths and Weaknesses
RPKM/FPKM	Reads per kilobase per million	Simple, length-corrected; not comparable across samples
TPM	Transcripts per million	Comparable across samples; standard in modern studies
DESeq2 size factors	Median ratio normalization	Robust to outliers, accounts for depth differences
TMM (edgeR)	Trimmed mean of M-values	Effective for unequal library sizes; sensitive to composition bias

—

Table 4: Key QC Metrics and Thresholds

Metric	Formula / Definition	Rule of Thumb / Threshold
PHRED score	$Q = -10 \log_{10} P$	Q30 = 1 error/1000 bases (gold standard)
Coverage	$C = \frac{L \times N}{G}$	≥30X min, 60–100X ideal
N50	Contig length containing 50% of assembly	≥1 Mb = good fungal assembly
BUSCO completeness	Conserved single-copy orthologs	>90% excellent, 70–90% acceptable
FDR (DEG analysis)	Adjusted p-value	<0.05 standard cutoff
$\log_2$FC	log fold change in expression	≥2 for significant induction

Appendix B: Project-Specific Tables

Table 5: Sequencing Data Overview for PLA-C1

Platform	Data Generated	Notes
Illumina NovaSeq	∼5 Gb paired-end reads (150 bp)	High accuracy; used for polishing and transcriptomics
Oxford Nanopore MinION	∼10 Gb long reads (N50 ∼20 kb)	Provided assembly contiguity; higher error rate
RNA-Seq (Illumina)	∼50M reads per condition (Control, PLA, PHA)	Used for quantification and differential expression

—

Table 6: Assembly Statistics for PLA-C1

Metric	Value	Interpretation
Genome size	38.6 Mb	Typical size for ascomycete fungi
Number of contigs	10	High contiguity for fungal genome
N50	5.3 Mb	Very good continuity for hybrid assembly
GC content	∼47%	Consistent with related fungi
BUSCO completeness	76.3%	Partial completeness; suggests gaps or collapsed repeats

—

Table 7: Annotation Summary

Feature	Count	Notes
Predicted protein-coding genes	∼10,000	Within expected fungal range
CAZymes identified (dbCAN3)	272	Strong enzymatic potential
Cutinases	9 (5 PLA-induced)	Candidate PLA degraders
Esterases	45 (12 PLA-induced)	Core ester bond hydrolases
Lipases	27 (7 PLA-induced)	Potential polymer degraders

PHA depolymerase-like	4 (2 PLA-induced)	Indicate adaptation to PHA

—

Table 8: Differential Expression Results

Comparison	DEGs Identified	Key Findings
PLA vs Control	84	Enrichment of esterases, cutinases, lipases
PHA vs Control	51	Upregulation of hydrolases with PHA
PLA & PHA shared	29	General stress/esterase response

—

Table 9: Comparative Genomics Results

Result	Value	Interpretation
Orthogroups shared with relatives	10,312	Conserved fungal core functions
PLA-C1 specific genes	314	Potential niche adaptation to polymers
Phylogenetic placement	Within *Purpureocillium* clade	Confirms expected taxonomy

—

Table 10: Bioinformatics Tools and Their Role

Tool	Purpose	Rationale
Flye	Long-read assembly	OLC algorithm robust to Nanopore error
Pilon	Polishing with Illumina	Corrected indels and homopolymers
MAKER3	Structural annotation	Combined ab initio + RNA-Seq evidence
dbCAN3	Functional annotation	Specific for CAZymes detection
Salmon	Transcript quantification	Fast, bias-aware quasi-mapping

DESeq2	Differential expression	Negative binomial model, FDR control
OrthoFinder	Orthology inference	Identified orthogroups for phylogenomics
MCScanX	Synteny analysis	Detected conserved gene order
RAxML	Phylogeny reconstruction	Maximum likelihood, robust statistics

Appendix C: Useful Pills and Exam Skills

Quick Skills Acquired During the Project

- Ability to design a hybrid sequencing strategy balancing cost, accuracy, and contiguity.
- Knowledge of assembly algorithms (DBG vs OLC) and when to apply them.
- Experience in genome quality control metrics: N50, BUSCO, coverage depth.
- Structural and functional annotation, integrating ab initio predictions and experimental evidence.
- Application of transcriptomics tools (Salmon, DESeq2) with understanding of statistical foundations.
- Comparative genomics and phylogenomic inference, including orthology, synteny, and niche adaptation analysis.
- Critical reflection on limitations of genomics-only approaches and importance of functional validation.

—

Critical Thinking Pills

- **Sequencing choice:** Illumina ensures accuracy, Nanopore ensures contiguity; hybrid design maximizes both.
- **Assembly pitfalls:** collapsed repeats, heterozygosity, contamination; mitigated with high coverage and polishing.
- **Annotation inflation:** homology-based prediction may mislabel genes; careful curation and experimental validation are required.

- **Transcriptomics caveat:** differential expression $\neq$ protein activity; proteomics/metabolomics essential for validation.
- **Comparative genomics:** isolate-specific genes can indicate niche adaptation; orthology/paralogy distinctions are essential.
- **Budget scenario:** with unlimited resources, PacBio HiFi and Hi-C would yield chromosome-level assemblies.

—

Exam-Oriented Reminders

- Always link methods to **rationale**: why this tool/technology, what alternative, what limitation.
- Mention **theory behind algorithms**: OLC vs DBG, negative binomial distribution, FDR correction.
- Highlight **numerical thresholds**: Q30, $\geq$30X coverage, BUSCO >90%, $\log_2$FC $\geq$2 with FDR <0.05.
- Stress **applied impact**: fungal hydrolases align with EU bioeconomy and industrial enzyme markets.
- If asked about **future improvements**, cite multi-omics integration and functional assays.

—

Connections Across Theory and Project

- Sequencing errors → assembly accuracy → annotation completeness → interpretation of DEGs.
- Library preparation biases → transcript quantification noise → normalization methods (TPM, DESeq2 size factors).
- Gene duplication (paralogy) → enzyme family expansion → potential neofunctionalization for bioplastic degradation.
- Pan-genome concepts → why isolate-specific genes matter in adaptation.