

A Computational Method for Predicting Tumour Cells in *Arachis Hypogaea* Root Nodules Through Differentially Expressed Genes

Surbhi Mistry¹, Shivani Saxena¹, Nilesh Patel¹, and Ahsan Rizvi^{1,*}

¹School of Computing and Technology Institute of Advanced Research Gandhinagar, Gujarat, India

Abstract.

Arachis hypogaea root nodules are specialized structures that entail complex plant-microbe interactions in processes such as stress responses and transcriptional regulation. The root nodules of *Arachis hypogaea* entail complex cellular proliferation and differentiation, and in certain stress conditions, the growth patterns may be similar to those of tumors. The molecular processes that mediate these changes are of equal importance in improving legume productivity as they are in the conceptual framework of abnormal cell proliferation in higher organisms. Despite some progress in the area of transcriptomics, the present bioinformatics tools for the analysis of differentially expressed genes (DEGs) and non-coding RNAs (ncRNAs) are still heavily reliant on static RNA-Seq data and do not take into account morphological data. This study aims to introduce a two-step approach to combine transcriptomic and morphological data effectively for improved prediction of abnormal growth patterns in peanut root nodules. Phase I of this study will include the analysis of RNA-Seq data obtained from root nodules under environmental stress and nodulation processes using standardized pipelines such as HISAT2 for alignment, StringTie for transcript assembly, and DESeq2 for differential expression analysis. Simultaneously, plant morphological characteristics will be assessed through imaging and sensor analysis to record growth. Phase II will concentrate on the establishment of a predictive computational model that integrates gene expression profiles with quantitative morphological parameters. Supervised machine learning algorithms will be trained on labeled data sets generated from transcriptomic profiles and quantitative morphological parameters to establish patterns linked with tumor-like growth patterns. Preliminary results obtained from Phase I suggest the existence of stress-mediated transcriptional processes involving genes linked with cell cycle regulation and signaling pathways. By integrating molecular and imaging data sets in a single analytical platform, this research aims to improve the detection of complex growth anomalies. The current research study contributes greatly to the comprehension of plant developmental regulation under stress conditions and provides a computational model that may provide additional insights into the mechanisms of abnormal cellular proliferation.

Keywords: *Arachis hypogaea*, root nodules, tumor-like growth, image-based phenotyping, bioinformatics.

1 Introduction

Arachis hypogaea (groundnut or peanut) is a legume that is of great nutritional value to humans and also increases soil fertility by fixing nitrogen in the root nodules through a symbiotic relationship with *Rhizobium* bacteria [1]. However, under certain biotic or abiotic stress conditions, these nodules can become tumor-like in nature. Conventional diagnostic methods, such as histological analysis and visual image analysis, are time-consuming, intrinsically subjective, and unsuitable for comprehensive monitoring [2]. Transcriptomic methods, like RNA sequencing, are able to identify differentially expressed genes (DEGs) associated with the onset of diseases but lack spatial or morphological information [3]. Conversely, image-based phenotyping is very efficient in detecting perceivable struc-

tural changes but can miss early molecular markers [4]. To overcome these limitations, we suggest a hybrid framework that combines DEG analysis with continuous image-based observation. This multi-modal framework takes advantage of the early detection capability of transcriptomics and structural information of image data to facilitate accurate, timely identification of tumor-like growths in peanut root nodules [5]. From sensor-recorded plant time-series data, measured at 1 day post-infection (1 DPI), 4 DPI, 8 DPI, and so on, we learn visual features of abnormal growth. Concurrently, RNA-Seq data are examined for DEGs exhibiting molecular changes. Temporal alignment of transcriptomic and phenotypic datasets enables more accurate characterization of abnormal cellular growth patterns and improves the biological interpretability of the predictive framework. Through the integration of molecular expression profiles and time-resolved morphological analyses, the model is able to incorporate the dy-

*e-mail: ahsan.rizvi@iar.ac.in

namics of tumor-like development. The approach provides a structured framework for computational analysis that could be used in future applications related to crop health monitoring. The overall workflow of the proposed methodology is illustrated in Figure 1.

2 Methodology

The present study employs a combination of transcriptome analysis and continuous image monitoring to predict the growth of tumor cells in *Arachis hypogaea* root nodules. The procedure started with the growth of peanut plants in a controlled environment and was categorized into two groups: a healthy control group and a treatment group exposed to infection. Both groups were continuously monitored by high-resolution image sensors, which took microscopic images of the plants at a predetermined time interval. The images were taken at critical times: 1, 4, 8, 12, and 28 days post-infection.

The image data were preprocessed using noise removal and normalization to enhance clarity. Advanced image processing techniques were then used to outline the nodules and to extract useful features. The features extracted were shape, texture, and structural abnormalities, which were measured using descriptors like the Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) [6]. The features extracted provided a close approximation of the cellular changes with time.

Simultaneously, RNA was extracted from the same nodule samples at the same time points and later processed by RNA-Seq. Raw reads were quality-checked and trimmed using FastQC [11] and Trimmomatic [12]. Later, the reads were then mapped to the reference genome using HISAT2 [13], and transcript assembly was done using StringTie [7]. These transcripts are fed to the modified pipelines (refer Figure 1). Differential gene expression analysis was conducted to identify the genes that were seen to be significantly upregulated or downregulated in infected samples compared to healthy controls. Genes with Log_2 Fold Change ≥ 1 and p -value < 0.05 were considered as differentially expressed and were later annotated to identify their functional roles in cell proliferation, defense, and tumor-like activities.

One of the major contributions of this study is the integration of molecular and image data. For every time point indicated, gene expression profiles were integrated with image features to form a dense, multimodal dataset. Integration allowed correlation of specific patterns of gene activity with morphological alterations seen in nodules. With this augmented dataset, machine learning models such as Support Vector Machines (SVM) [9], Random Forest [10], and Neural Networks [14] were used to classify samples as tumor or non-tumor. Expert annotations were utilized to supply the gold standard truth labels, and model performance was evaluated using cross-validation and standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

3 Result and Discussion

In the current phase of this study, transcriptomic analysis has been successfully performed using RNA-Seq data collected from infected and control root nodules at different time points. Data are collected from the NCBI database with project accession number GSE98997 (18 libraries) [8]. DE analysis identified a set of genes that were significantly upregulated or downregulated in response to infection. A total of 7179 ncRNAs were detected using the proposed pipeline. The DEG analysis of *Arachis hypogaea* across various developmental stages of infection and nodulation reveals dynamic transcriptional changes is shown in figure 2. In the initial comparison between the controlled (UI) stage and 1DPI, 209 genes were upregulated while 410 were downregulated, indicating an early suppression of gene activity in response to infection. Between 1DPI and 4DPI, 175 genes were upregulated and 215 downregulated, suggesting a continued yet moderate adjustment in gene expression. A sharp increase in gene activation was observed between 4DPI and 8DPI, with 576 upregulated genes and only 82 downregulated, highlighting a strong transcriptional response likely related to nodule initiation or development. From 8DPI to 12DPI, 194 genes were upregulated and 40 downregulated, showing a reduced but ongoing shift in gene activity.

The transition between 12 days post-infection (12 DPI) and the mature nodule (NOD) stage showed the largest shift in gene expression, with 1,346 transcripts significantly upregulated and 2,232 downregulated. These expression changes reflect a significant transcriptional reprogramming process during the course of progression towards functional nodule maturation and symbiosis stabilization. The expression patterns revealed in this study suggest a coordinated regulation of gene expression pathways involved in cellular differentiation, metabolic reprogramming, and developmental regulation. Some of the genes that have been identified in this study are known to be involved in cell cycle regulation, hormonal signaling, and nodule development, which affect localized cell proliferation. For instance, *CYCD3;1* and *CDKA;1* are key regulators of cell cycle progression and are involved in regulated cell division during organ development. On the other hand, Kip-related proteins (KRPs) function as negative regulators of cyclin-dependent kinases and play a role in regulating balanced cell proliferation.

Genes involved in auxin biosynthesis and transport, such as *YUCCA*, *TIR1*, and *PIN1*, also demonstrated differential expression patterns. Auxin gradients are essential for the initiation of nodules and tissue differentiation, and any disruption in their regulation could affect localized growth responses. Cytokinin signaling pathway genes, such as type-A and type-B response regulators (ARRs), are also crucial for regulating infection responses and nodule organogenesis. *ENOD40*, an early nodulin gene that has been well characterized, was also identified among the transcripts that are associated with developmental transitions. *ENOD40* has

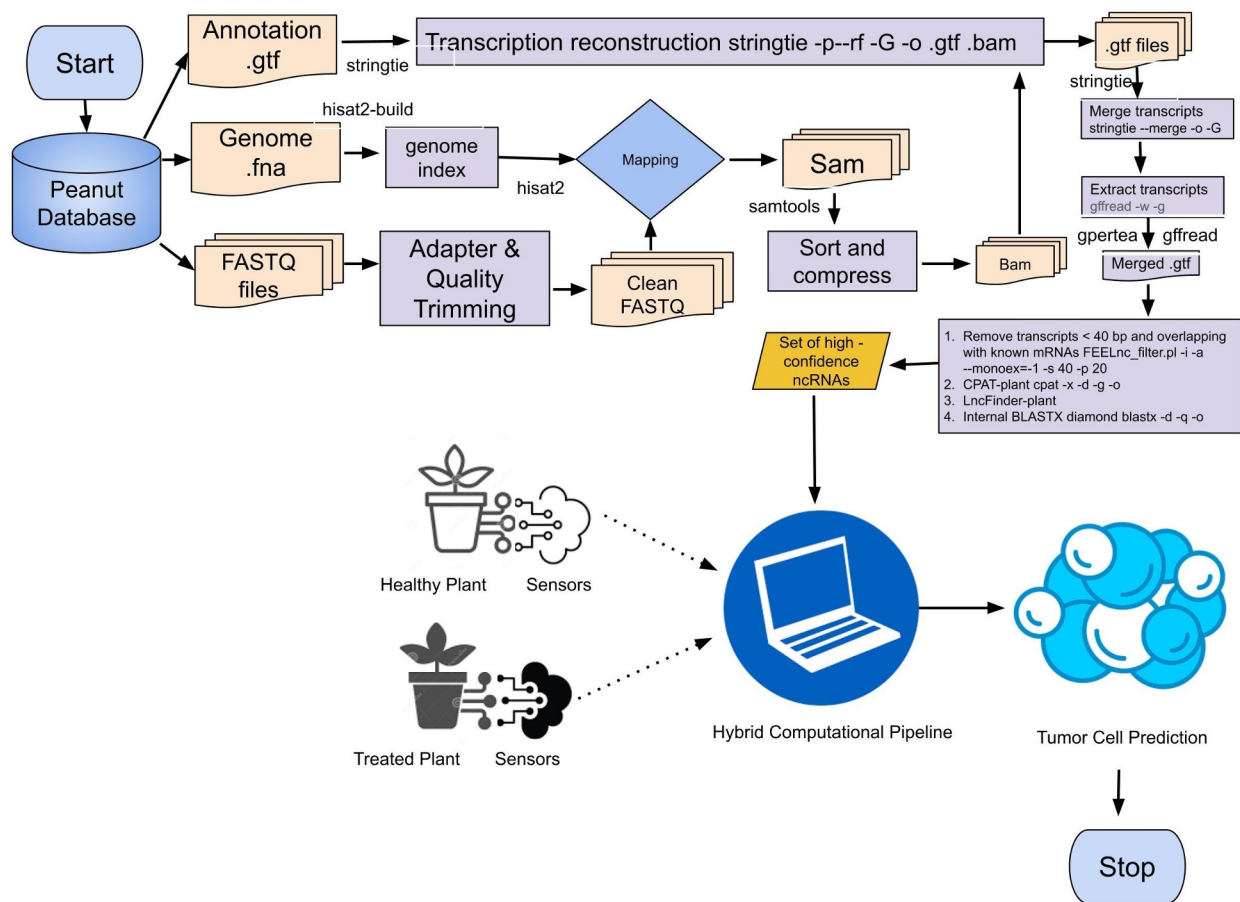


Figure 1: Proposed Methodology for predicting Tumor Cell in Arachis hypogaea Root Nodules.

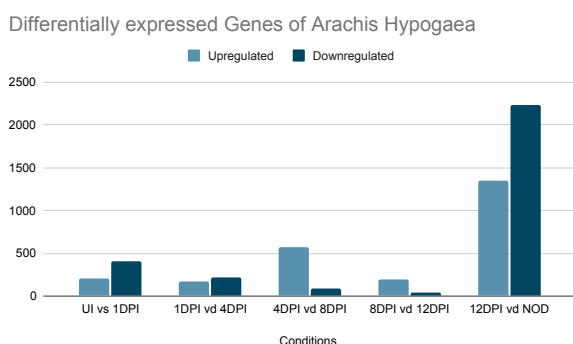


Figure 2: Differentially expressed genes analysis of Arachis hypogaea across various developmental stages of infection and nodulation.

been shown to play a role in the initiation of cortical cell divisions during nodule development. Taken together, these genes are important regulatory genes that are involved in cell proliferation and developmental signaling in peanut root nodules.

Concurrently with the transcriptomic analysis, an image acquisition system has also been set up to track

the progress of root nodule formation. High-resolution microscopic images are being systematically acquired at specific time points to track morphological development related to infection and nodule development. These images serve as temporal data points that relate to the stages analyzed in the RNA-Seq analysis. Currently, data from the transcriptomic and imaging analyses are being generated and analyzed separately. The next step of the analysis will be to integrate the data. This will involve the extraction of quantitative features from microscopic images and the alignment of morphological features with gene expression profiles at specific stages. Computational tools will be used to investigate correlations between molecular and morphological features. The aim of this integrated analysis is to develop a framework of classification that can distinguish between normal transitions of development and abnormal or excessive proliferation. By integrating molecular and morphological data into a single framework of analysis, it is hoped that a better understanding of growth regulation in peanut root nodules can be achieved. This integration could help in better surveillance of nodule health and stress-induced developmental transitions.

4 Conclusion

The current research offers a hybrid analytical model that seeks to combine transcriptome profiling with quantitative image analysis to investigate the abnormal growth patterns of *Arachis hypogaea* root nodules. By integrating the profiles of differentially expressed genes with time-course imaging data from sensors, the hybrid model is able to identify both molecular and morphological patterns of nodule development in infection and stress responses. The results of the analysis show that a number of differentially expressed genes associated with cell cycle regulation, defense response, and hormonal regulation have stage-specific expression patterns that match the observable structural differences in the infected nodules.

The machine learning models trained on the merged dataset showed robust classification accuracy in identifying typical developmental patterns and abnormal growth patterns. These findings validate the effectiveness of merging multi-modal datasets for enhancing pattern recognition in complex biological systems. Instead of focusing on transcriptomic patterns or visual interpretation, the merged model offers a more holistic view of growth patterns. The proposed approach helps to enhance computational analysis of plant developmental responses and offers a systematic framework for relating gene expression to phenotypic data. This method could be of great use in developing systematic tools for monitoring the health of nodules and stress-related modifications. Future work will focus on improving the framework for analysis by using advanced feature extraction techniques and developing deep learning models for image analysis. The proposed approach will be extended to other plant species and conditions to test its generalizability to other plant systems.

References

1. Beshah, A., et al. "Isolation and Screening of Effective Rhizobium Species Nodulating and Fixing Atmospheric Nitrogen on Groundnut (*Arachis hypogaea*) Grown in Some Important Regions of Ethiopia." *Eurasian Soil Science* 57.Suppl 1 (2024): S78-S91.
2. Lv, Zhenghao, et al. "Identification of candidate genes associated with peanut pod length by combined analysis of QTL-seq and RNA-seq." *Genomics* 116.3 (2024): 110835.
3. Tatmiya, Ritisha N., et al. "Comparative transcriptome profiling of resistant and susceptible groundnut (*Arachis hypogaea*) genotypes in response to stem rot infection caused by *Sclerotium rolfsii*." *Plant Pathology* 73.9 (2024): 2500-2515.
4. Murphy, Katherine M., et al. "Deep learning in image-based plant phenotyping." *Annual Review of Plant Biology* 75 (2024).
5. Sangeetha, S. K. B., et al. "An Empirical Analysis of Transformer-Based and Convolutional Neural Network Approaches for Early Detection and Diagnosis of Cancer Using Multimodal Imaging and Genomic Data." *IEEE Access* (2024).
6. Ahmad, Wakeel, Syed M. Adnan, and Aun Irtaza. "Local triangular-ternary pattern: a novel feature descriptor for plant leaf disease detection." *Multimedia Tools and Applications* 83.7 (2024): 20215-20241.
7. Rosati, Diletta, et al. "Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: a review." *Computational and structural biotechnology journal* (2024).
8. Geer, Lewis Y., et al. "The NCBI biosystems database." *Nucleic acids research* 38.suppl_1 (2010): D492-D496.
9. Hearst, Marti A., et al. "Support vector machines." *IEEE Intelligent Systems and their applications* 13.4 (1998): 18-28.
10. Salman, Hasan Ahmed, Ali Kalakech, and Amani Steiti. "Random forest algorithm overview." *Babylonian Journal of Machine Learning* 2024 (2024): 69-79.
11. Brown, Joseph, Meg Pirrung, and Lee Ann McCue. "FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool." *Bioinformatics* 33.19 (2017): 3137-3139.
12. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.
13. Kim, Daehwan, et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype." *Nature biotechnology* 37.8 (2019): 907-915.
14. Abdi, Hervé, Dominique Valentin, and Betty Edelman. *Neural networks*. No. 124. Sage, 1999.