IDENTIFICATION OF PROCESSED PSEUDOGENES INVOLVED IN RESISTANCE MECHANISM AGAINST BIOTIC STRESS IN *ORYZA SATIVA* USING RNA-SEQ DATA

by
Huriye Yanık

Submitted to Graduate School of Natural and Applied Sciences
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in
Biotechnology

Yeditepe University
2022

IDENTIFICATION OF PROCESSED PSEUDOGENES INVOLVED IN RESISTANCE MECHANISM AGAINST BIOTIC STRESS IN *ORYZA SATIVA* USING RNA-SEQ DATA

APPROVED BY:

Assoc. Prof. Dr. Emrah Nikerel ……
(Thesis Supervisor)
(Yeditepe University)

Assoc. Prof. Dr. Ali Özhan Aytekin ……
(Yeditepe University)

Assist. Prof. Dr. Bahar Soğutmaz Özdemir ……
(Yeditepe University)

Assoc. Prof. Dr. Tunahan Çakır ……
(Gebze Technical University)

Assist. Prof. Dr. Pınar Pir ……
(Gebze Technical University)

DATE OF APPROVAL: …./…./2022

I hereby declare that this thesis is my own work and that all information in this thesis has been obtained and presented in accordance with academic rules and ethical conduct. I have fully cited and referenced all material and results as required by these rules and conduct, and this thesis study does not contain any plagiarism. If any material used in the thesis requires copyright, the necessary permissions have been obtained. No material from this thesis has been used for the award of another degree.

I accept all kinds of legal liability that may arise in case contrary to these situations.

Name, Last name ……

Signature …..…….……..

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Assoc. Prof. Dr. Emrah Nikerel, for his guidance, support, and patience during my Ph.D. study. I would especially like to thank the thesis committee members of my thesis; Assoc. Prof. Dr. Tunahan Çakır, Assist. Prof. Dr. Bahar Soğutmaz Özdemir, Assist. Prof. Dr. Pınar Pir, and Assoc. Prof. Dr. Ali Özhan Aytekin for their critical questions, comments, and encouragement.

I would like to thank The Scientific and Technological Research Council of Turkey (TUBITAK) for their financial support provided by the TUBITAK Directorate of Science Fellowships and Grant Programmes (BIDEB) 2211-C Domestic Priority Doctoral Fellowship Program.

Furthermore, I am very grateful to all academicians who took part in the Biotechnology Doctorate Program at Yeditepe University. I sincerely thank Tülay Dalgıç, secretary of the Graduate School of Natural and Applied Sciences, ensuring correct and fast progress during official proceedings.

Special thanks to my dear friends Ezgi Tanıl, Sheyda Shakoory, Burcu Şirin, Derya Garip, Dr. Bahtir Hyseni, Özgür Yüksel and research assistant Neslihan Kayra for their emotional support. During the long, tiring, and action-packed doctoral process, each of you left special and sweet memories for me.

Dear family friends, Mrs. Selver & Mr. İsmail Uzun, I sincerely thank you for supporting me as part of being a guarantor in receiving a Ph.D. scholarship.

Most especially, I present my deepest thanks and express my gratitude to my mother, sister, brother, kizum, and father for their love, ongoing support, and patience during my academic life.

Lastly, I hope this work can offer a new perspective to all the people in the way of science working on pseudogenes. I will continue to work with all my strength to build a more fair world wherever I am.

ABSTRACT

IDENTIFICATION OF PROCESSED PSEUDOGENES INVOLVED IN RESISTANCE MECHANISM AGAINST BIOTIC STRESS IN *ORYZA SATIVA* USING RNA-SEQ DATA

Oryza sativa (rice), belonging to Poaceae, is the third most cultivated crop due to being rich in nutrients and energy sources. It is constantly exposed to biotic and abiotic stress factors. Biotic stress caused by living organisms such as fungi, nematodes, insects, bacteria, and viruses causes significant physiological and molecular changes, decreasing crop quality and yield. Plants have various defense mechanisms against these biotic stress factors. In particular, resistance (R) genes play an essential role in plant resistance mechanisms against biotic stress. R proteins interact with effector proteins that are products of avirulence (Avr) genes synthesized by pathogens. Pseudogenes are disabled genes, occurring with accumulated mutations as copies of protein-coding genes over time. These are typically classified into three groups processed (retrotransposed), duplicated (unprocessed), and unitary pseudogenes. Although it is debated whether these genes are functional or not, pseudogenes are known to play an essential role in the regulation of various genes. RNA-seq is an approach that enables the identification of genes expressed under specific conditions using sequencing methods. In this thesis, the role of R genes in the resistance of commercially important rice to biotic stress and its relation with pseudogenes be sought using RNA-seq data from the literature. As a result, two pseudogenes related to nematode stress and a pseudogene related to fungus stress were associated with biotic stress. 10, 34, and 15 R genes were found as differentially expressed genes (DEGs) in the nematode, fungus ZB13, and fungus ZH10 datasets. Using artificial neural networks (ANN), training and test data consisting of known pseudogenes and genes were correctly classified at 87 percent and 90 percent rates. The outcomes of this study need to be validated further in other cereals or even other plant species to find traits that are generally traceable for disease prevention.

ÖZET

ORYZA SATIVA' DA BİYOTİK STRESE KARŞI DİRENÇ MEKANİZMASINDA YER ALAN İŞLENMİŞ YALANCI GENLERİN RNA DİZİLEME VERİLERİNİ KULLANARAK TANIMLANMASI

Poaceae familyasına ait *Oryza sativa* (pirinç), besin ve enerji kaynakları açısından zengin olması nedeniyle en çok ekimi yapılan üçüncü üründür. Pirinç, sürekli olarak biyotik ve abiyotik stres faktörlerine maruz kalmaktadır. Mantarlar, yuvarlak solucanlar, böcekler, bakteriler ve virüsler gibi canlı organizmaların neden olduğu biyotik stres, tarımsal ürünlerde kalite ve verimi düşüren fizyolojik ve moleküler önemli değişikliklere neden olmaktadır. Bitkilerin stres faktörlerine karşı çeşitli savunma mekanizmaları vardır. Özellikle direnç (R) genleri, biyotik strese karşı bitki direnç mekanizmalarında önemli bir rol oynamaktadır. R proteinleri, patojenler tarafından sentezlenen avirülans (Avr) genlerinin ürünleri olan efektör proteinlerle etkileşime girer. Yalancı genler, protein kodlayan genlerin kopyaları olarak zamanla biriken mutasyonlarla ortaya çıkan, devre dışı bırakılmış genlerdir. Bunlar tipik olarak işlenmiş (retrotranspoze), kopyalanmış (işlenmemiş) ve üniter yalancı genler olarak üç gruba ayrılır. Bu genlerin işlevsel olup olmadığı tartışılsa da yalancı genlerin çeşitli genlerin düzenlenmesinde önemli rol oynadığı bilinmektedir. RNA dizileme (RNA-seq), dizileme yöntemleri kullanılarak belirli koşullar altında ifade edilen genlerin tanımlanmasını sağlayan bir yaklaşımdır. Bu tezde, ticari açıdan önemli pirincin biyotik strese direncinde R-genlerinin rolü ve yalancı genlerle ilişkisi literatürden RNA-dizileme verileri kullanılarak araştırılmıştır. Sonuç olarak, nematod stresi ile ilgili iki yalancı gen ve mantar stresi ile ilgili bir yalancı gen biyotik stres ile ilişkili bulundu. 10, 34 ve 15 R geni, sırasıyla nematod, mantar ZB13, mantar ZH10 veri setlerinde diferansiyel ifade edilmiş genler (DEG) olarak bulundu. Bilinen yalancı gen ve genlerden oluşan eğitim ve test veri setleri, yapay sinir ağları (YSA) kullanılarak yüzde 87 ve yüzde 90 oranlarında doğru olarak sınıflandırılmıştır. Bu çalışmanın sonuçlarının, hastalık önlemede genellikle izlenebilir özellikleri bulmak için diğer tahıllarda ve hatta diğer bitki türlerinde daha fazla doğrulanması gerekir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

ψ	Pseudogene
ANN	Artificial neural network
ARFF	Attribute-relation file format
Avr	Avirulence
BAM	Binary alignment/map
BED	Browser extensible data
BLAST	Basic local alignment search tool
bp	Base pair
CC	Coiled-coil
cDNA	Complementary-DNA
CDS	Coding sequence
DEG	Differentially expressed gene
dpi	Days post infection
ET	Ethylene
ETI	Effector-triggered immunity
ETS	Effector-triggered susceptibility
FC	Fold change
FDR	False discovery rate
FN	False negative
FP	False positive
F/RPKM	Fragments/Reads per kilobase of transcript per million mapped reads
FPR	False positive rate
GO	Gene ontology
HR	Hypersensitive response
IRGSP	International rice genome sequencing project
JA	Jasmonic acid
Ka	Nonsynonymous substitution ratio
Ks	Synonymous substitution ratio
KNN	K- nearest neighbor
LncRNA	Long non-coding RNA

LRR	Leucine-rich repeat
MAPK	Mitogen-activated protein kinases
miRNA	MicroRNA
MLP	Multilayer perceptron
NB	Nucleotide binding
NCBI	National center for biotechnology information
NGS	Next-generation sequencing
NLR	Nucleotide-binding leucine-rich repeat or NOD-like receptor
PAMP	Pathogen-associated molecular pattern
PCR	Polymerase chain reaction
PR	Pathogenesis-related
PRR	Pattern recognition receptor
PTENP1	Phosphatase and tensin homolog pseudogene 1
PTI	PAMP-triggered immunity
qRT-PCR	Quantitative real-time PCR
R	Resistance
RAP-DB	Rice annotation project database
RLK	Receptor-like kinase
RLP	Receptor-like protein
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
RT	Reverse transcriptase
SA	Salicylic acid
SAM	Sequence alignment/map
SAR	Systemic acquired resistance
SBL	Sequencing by ligation
SBS	Sequencing by synthesis
siRNA	Small interfering RNA
SMO	Sequential minimal optimization
SNP	Single nucleotide polymorphism
SRA	Sequence read archive
sRNA-seq	Small RNA sequencing
SVM	Support vector machines

T3SS	Type III secretion system
TGS	Third generation sequencing
TIR	Toll/interleukin-1 receptor
TN	True negative
TP	True positive
TPR	True positive rate
WEKA	Waikato environment for knowledge analysis
WGS	Whole genome sequencing

1. INTRODUCTION

1.1. CEREAL

Plants are essential photosynthetic eukaryotes that interact directly or indirectly with living organisms throughout their life. Because they are the primary source of carbon and oxygen for sustainable energy, plants are the core basis for life on earth. By providing the primary production of the food chain, they play an important role in transferring energy to other living systems. Additionally, they have benefits that make life easier from many perspectives, such as nutrition, medicine, clothing, shelter, and (bio)fuel. Due to the reasons mentioned above, they play a leading role in ensuring the energy required for the continuity of all ecosystems.

Cereals, economically important plants, have been classified in Poaceae (Gramineae) (Figure 1.1.). Cereal grains that are nutrient-rich in vitamins, minerals, and phytochemicals are the most cultivated agricultural products globally due to their importance to human nutrition. Particularly maize (*Zea mays* L.), wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), barley (*Hordeum vulgare* L.), oats (*Avena sativa* L.), and rye (*Secale cereale* L.) are the most planted cereals in the world, respectively [1].

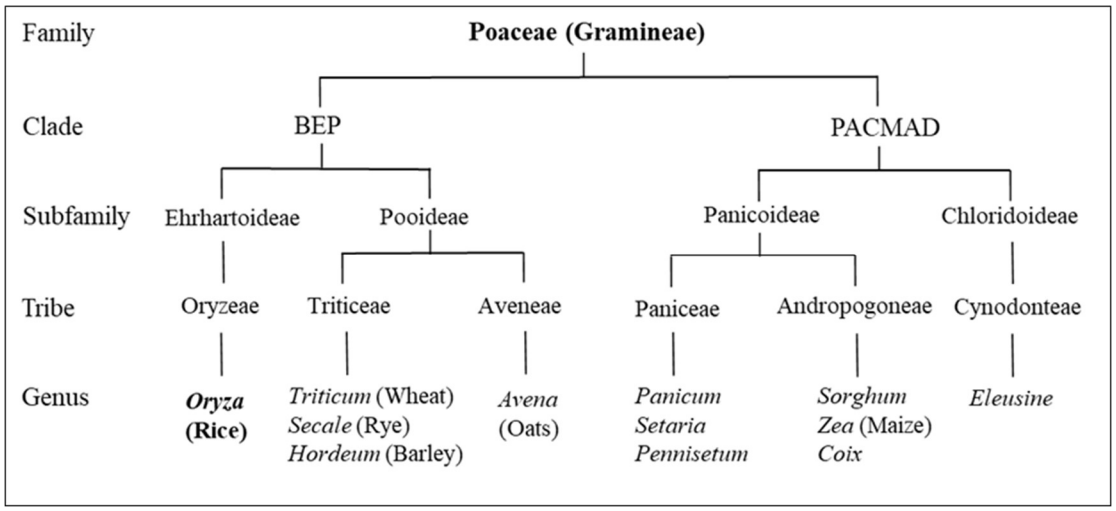

Figure 1.1. Taxonomy of Poaceae (Gramineae) [2–4]

Cereals have large and complex genomes due to repetitive sequences and polyploidy. They have different genome sizes ranging from 420 Mb to 17 Gb. As seen in Figure 1.2., genome sizes of rice (japonica, indica), maize, barley, bread wheat, rye, durum wheat were sequenced as almost 420 Mb [5], 466 Mb [6], 2.3 Gb [7], 5.1 Gb [8], 17 Gb [9], 7.9 Gb [10], 10.4 Gb [11], respectively.

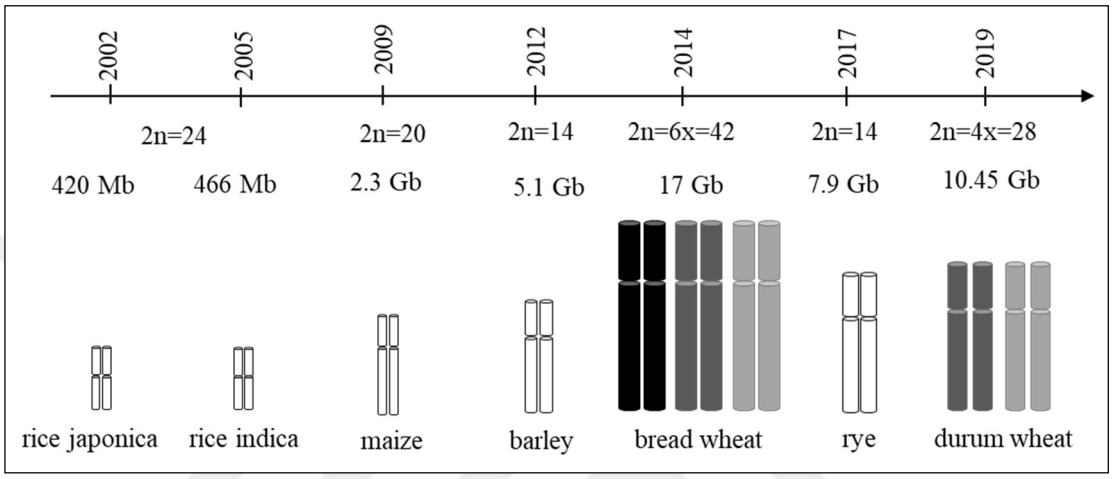

Figure 1.2. The genome size of economically important cereal genomes [12]

1.1.1. Oryza sativa L.

Oryza sativa (Rice) is an economically important cereal due to rich in nutrients and energy sources. It was the third most-produced grain globally, with 495.87 million metric tons in 2018/19 [1]. *Oryza sativa* L. (Asian rice) has almost 400 Mb genome size and a diploid AA genome consisting of 24 chromosomes [12,13]. It is an accepted model organism for monocot plants such as wheat, barley, maize, oat, rye.

Oryza has 22 wild and 2 cultivated species containing 11 genome types (diploid: AA, BB, CC, EE, FF, GG; tetraploid: BBCC, CCDD, HHJJ, HHKK, KKLL). *Oryza sativa* (Asian rice) and *Oryza glaberrima* (African rice) are the cultivated ones that have higher yields worldwide but lower yields in West Africa. Asian rice was domesticated from *Oryza rufipogon* ~10,000 BC, and African rice was domesticated from *Oryza longistaminata* ~3,000 BC [14–16]. *Oryza sativa* has two subspecies, including *Oryza sativa* ssp. *japonica* and *Oryza sativa* ssp. *indica* (Figure 1.3.).

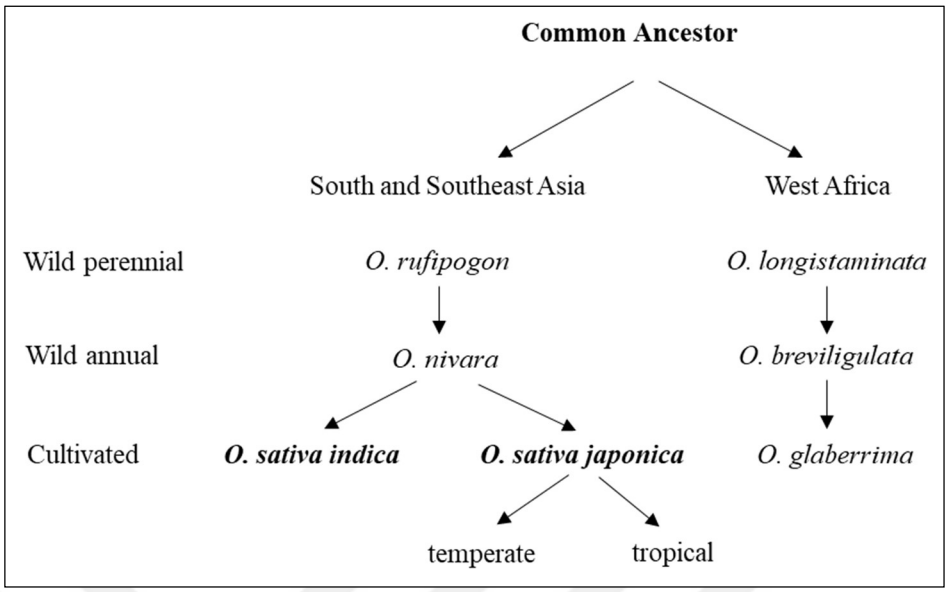

Figure 1.3. The domestication of two cultivated rice species [16]

1.2. BIOTIC STRESS

Stress is any kind of physical, chemical, or biological change that damages the plant. Being sessile organisms, plants are constantly exposed to various stress factors categorized as abiotic and biotic stress factors. Abiotic stress factors are the extreme value of water, salt, temperature, chemical, and radiation, while biotic stress factors are living organisms such as fungi, nematodes, bacteria, viruses, etc.

Each stress factor causes significant physiological and molecular changes, decreasing crop quality and yield. They are a significant threat by preventing the growth and reproduction, which are needed to keep the organism alive.

Among biotic stress factors, fungi, bacteria, and viruses are pathogen groups containing the largest number of species, respectively. Over 19000 fungal [17], 4100 nematode [18], 150 bacterial [19], 450 viral [20] species cause plant diseases so they are called as phytopathogens. To fight against biotic stress will reduce the use of pesticides, thereby positively contributing to human and environmental health in terms of biodiversity and food chain.

1.2.1. Fungi

Fungi are heterotrophic eukaryotic organisms consisting of various living groups from unicellular to multicellular. They have a complex life cycle involving sexual, asexual, or both reproduction. *Saccharomyces cerevisiae* (baker's yeast), *Candida albicans* (pathogen in humans), *Fusarium oxysporum* (in the plant), and *Aspergillus niger* (in the industry) are known model species of fungi [21].

Fungal pathogens are classified into necrotrophic, hemibiotrophic, and biotrophic pathogens, according to feeding mode. Necrotrophic pathogens obtain nutrients from dead cells to grow, but biotrophic pathogens feed within living cells [22]. Hemibiotrophic pathogens begin to live in living cells and continue their development in dead cells. Blight, spotting, rot, and wilting are the main symptoms in fungal diseases [23].

Magnaporthe oryzae and *Rhizoctonia solani* fungi are rice's major hemibiotrophic and necrotrophic pathogens. *Magnaporthe oryzae,* heterothallic ascomycete, asexual species of this is *Pyricularia oryzae*, and former name is *Magnaporthe grisea,* has almost 40 Mb genome size [24,25].

Kawahara *et al.* analyzed mixed transcriptome in *Oryza sativa* L. ssp. *japonica* cv. Nipponbare upon *Magnaporthe oryzae* field isolates P91-15B and Ina86-137. They infected leaves at 24 hpi, which is a time of primary infection. They found that pathogenesis-related (PR) and phytoalexin biosynthetic genes were upregulated in rice while glycosyl hydrolases, cutinases, and LysM domain-containing proteins were highly upregulated in fungus [26].

1.2.2. Nematodes

Nematodes, belonging to the Nematoda phylum, are roundworms that live as free or parasitic almost in all habitats [27]. They have bilaterally symmetrical and unsegmented bodies, ranging from 0.25 mm to 1 mm in length microorganism. The life cycle of a nematode is usually classified into six phases; the egg stage, four juvenile stages (J1-J4), and the adult stage [28]. *Caenorhabditis elegans* is the most well-known nematode species used as a model pathogen.

Out of 4100 plant-parasitic nematodes [18], 300 species are disease agents in rice [29]. Gall formation, cyst, stunting, chlorosis, necrosis caused by plant-parasitic nematodes reduce yield loss by 20 percent in rice [30].

Plant-parasitic nematodes (PPNs) are obligate biotrophic pathogens. Parasitic nematodes are classified as foliar and root parasites. Foliar parasites feed on leaves, flowers, or stems, while root parasites that make up the majority of parasites feed on the roots of plants. Root parasites are divided into sedentary endoparasites, migratory endoparasites, and ectoparasites. Sedentary endoparasitic nematodes that include root-knot and cyst nematodes penetrate the plant vascular system and stay permanently in the root. Migratory endoparasitic nematodes enter plant cells and migrate through plant cells, while ectoparasitic nematodes feed outside the plant tissue [31].

Hirschmanniella oryzae (root nematode) and *Meloidogyne graminicola* (root-knot nematode) are migratory and sedentary endoparasites [32]. *Meloidogyne graminicola* has an almost 30 Mb genome size, and it is in the first range on the top 10 lists of plant-parasitic nematodes [33,34].

Kyndt *et al.* investigated the systemic response of the *Oryza sativa* cv. Nipponbare upon *Hirschmanniella oryzae* nematode, which causes root infection. RNA sequencing was applied on shoots of inoculated and non-inoculated plants, which were collected at 3 and 7 days post-infection (dpi). Out of 195 genes detected at 3 dpi, 169 were repressed, and 26 were induced, while out of 17 genes detected at 7 dpi, 11 genes were repressed, of which 6 were induced. In the primary metabolism of the plant shoot, chlorophyll biosynthesis, brassinosteroid pathway, and amino acid production were suppressed, while isoprenoid and shikimate pathways were repressed in secondary metabolism [35].

1.2.3. Bacteria

Bacteria are prokaryotic organisms living as single or colonies, ranging between 0.15-700 μm in length [36]. According to the cell wall structure, they have been classified into two groups: gram-positive and gram-negative. Phytopathogenic bacteria usually occur from gram-negative bacteria. *Xanthomonas oryzae* pv. *oryzae* (Xoo), *Xanthomonas oryzae* pv. *oryzicola* (Xoc), and *Burkholderia glumae* are gram-negative pathogenic bacteria in rice,

and they have almost 5 Mb, 10 Mb, 7 Mb genome sizes, respectively [37]. Wilting, necrosis, chlorosis, rot, galls, and scab are the main symptoms of bacterial diseases [23].

Most phytopathogenic bacteria use a type III secretion system (T3SS) to secrete effectors into the plant cells. This macromolecular protein is encoded by hypersensitive response and pathogenicity (hrp) genes [38]. Toxoflavin, lipase, type III effector, extracellular polysaccharides (EPS), polygalacturonase are the most known virulence effectors in *Burkholderia glumae* [39].

Yu *et al.* revealed transcriptional responses of *Oryza sativa* L. cv. Nipponbare, which was infected by Xoo strain PXO99A (wild type-WT) and PXO99A relative to ΔfliC (flagellin deletion mutant derived from PXO99A). They identified that out of 1680 DEGs in rice inoculated with PXO99A relative to ΔfliC, 1159 genes were upregulated and 521 genes were downregulated. Moreover, 57 genes in WT and 21 genes in ΔfliC have specifically expressed genes (SEGs). Consequently, they showed that ΔfliC, a flagellin deletion mutant of PXO99A, enhanced disease in rice [40].

1.3. PLANT IMMUNE SYSTEM

Plants have an innate immune system to be protected against diseases caused by biotic stress factors. Unlike animals, plants do not have an adaptive immune system [41] and lack immune cells [42]. Two stages of the immune system play a role through plant-microbe interaction in plants (i) *PAMP (pathogen-associated molecular pattern)-triggered immunity (PTI)* and (ii) e*ffector-triggered immunity (ETI)*, which were formerly named horizontal and vertical resistances.

1.3.1. PAMP-Triggered Immunity (PTI)

PTI is the first defense layer in the plant resistance mechanism. PAMPs are microbial defense molecules, and pattern recognition receptors (PRRs) that are immune receptors on plant plasma membranes encountered in there. PAMPs are microbial molecules used as defense responses against plants by pathogens. Flagellin, peptidoglycan, elongation factor Tu, lipopolysaccharide, glucan, chitin, and xylanase are PAMPs. Bacterial *flagellin 22*

(flg22), bacterial *elongation factor thermo unstable (EF-Tu)*, fungal *ethylene-inducing xylanase (EIX)*, and oomycetes' *heptaglucan (HG)* are well-known PAMPs [43].

PRRs are immune receptors that are localized on the host plasma membrane. They are grouped with transmembrane receptor-like kinases (RLKs) and receptor-like proteins (RLPs) in plants [44]. PRRs have various domains, which are composed of leucine-rich repeats (LRRs), lysine motifs (LysMs), lectin motifs, or epidermal growth factor (EGF)-like domains. *Flagellin sensitive 2 (FLS2)*, *XA21*, and *brassinosteroid insensitive 1 (BRI1)-associated kinase 1 (BAK1)* are LRR-RLK PRRs. FLS2, which was identified in *Arabidopsis* against bacterial flg22 recognition [45], and XA21, was defined in *Oryza sativa* against bacterial blight pathogen *Xanthomonas oryzae* pv. *oryzae* [46] are well-known examples of LRR-RLK PRRs [42].

Chitin elicitor receptor kinase 1 (CERK1) is LysMs-RLK PRR, while *chitin elicitor binding protein (CEBiP)* is LysM-RLP PRR [44]. When PRRs recognize PAMPs, the first defense response is performed against the pathogen. Signaling pathways such as mitogen-activated protein kinases (MAPK), calcium-dependent protein kinase (CDPK), reactive oxygen species (ROS), and phytohormones (jasmonic acid (JA), salicylic acid (SA), ethylene (ET)) play a role in transcriptional reprogramming.

1.3.2. Effector-Triggered Immunity (ETI)

In comparison with PTI, ETI exhibits a more robust resistance mechanism [47]. In the second defense layer ETI, effector proteins that are produced by avirulence (Avr) genes in pathogen and resistance (R) proteins that are synthesized by R genes in plants, are interacted with each other [48]. Effector proteins are products of Avr genes expressed by pathogens during the infection process. Avr-Pita, AvrPtoB, Avr3a, and CP are fungal, bacterial, oomycete, and viral effector proteins, respectively [49]. Almost 83 effector proteins have been characterized by crop-infecting fungi and oomycetes [50].

Effectors are secreted by T1SS, T2SS, T3SS, T4SS, T5SS, T6SS, and T7SS in bacteria [51]. Unlike bacteria, no TTSS has been found in fungi. Fungal effector proteins may be secreted from haustoria into plant apoplast [49]. Viral effector proteins entry occurs via physical injuries caused by biological vectors such as insects, nematodes, fungi [52]. However,

nematodes and insects enter the plant cell via stylet [53]. The overview of plant-pathogen interaction is illustrated in Figure 1.4. ETI induces hypersensitive response (HR), resulting in programmed cell death (PCD), usually resulting in a local necrotic lesion.

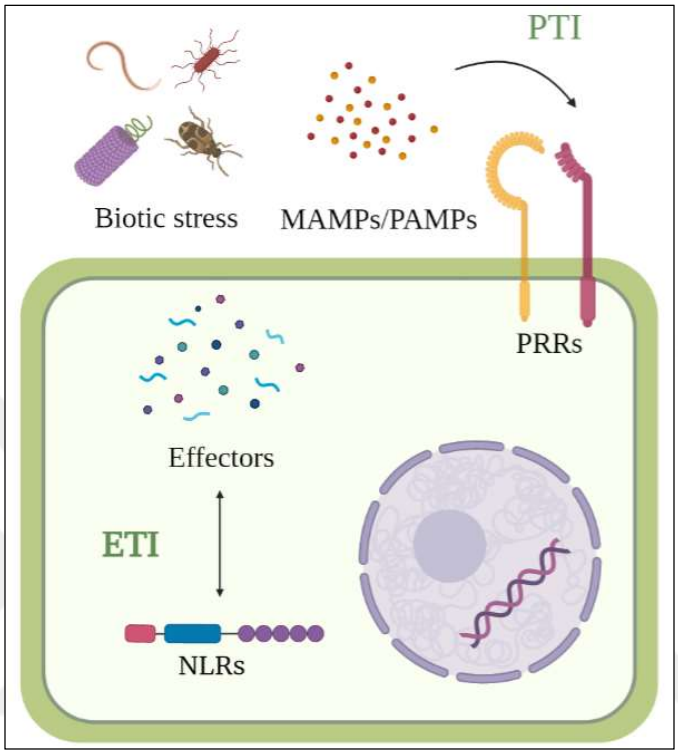

Figure 1.4. PTI and ETI systems [54]

1.3.2.1. *R Genes*

As a recognizer of effector proteins, R genes encode R proteins in plant genomes, which play a key role in plant disease resistance mechanisms. R genes have more than one classification due to their domain diversity. Nucleotide-binding leucine-rich repeat or NOD-like receptors (NLRs) that are the most of R genes are rapidly evolved both among family members in the same genome and individuals in the same species [55].

NLRs have major domains such as nucleotide-binding (NB), leucine-rich repeat (LRR), toll/interleukin-1 receptor (TIR), and coiled-coil (CC). They have been classified into 4 groups, including TIR-NB-LRR (TNL), TIR-NB-LRR-X, CC-NB-LRR (CNL), RPW8-NB-LRR (Figure 1.5.). X refers to the variable domain here. NB-LRR, the largest subfamily of R proteins, has ranged from almost 860 to 1900 amino acid residues [56]. NB domain binds and hydrolyzes ATP and GTP, while the LRR domain is repeated leucine amino acid

sequences involved in protein-protein interactions and ligand binding [57]. Several studies are available related to NB-LRR genes, both in monocot and dicots: out of dicots, 459 in grapevine [58], 333 in Medicago [59], 330 in poplar [58], 149 in Arabidopsis [60] and out of monocots, ~ 470 in rice [61], 175 in Brachypodium [62], 95 in maize [63].

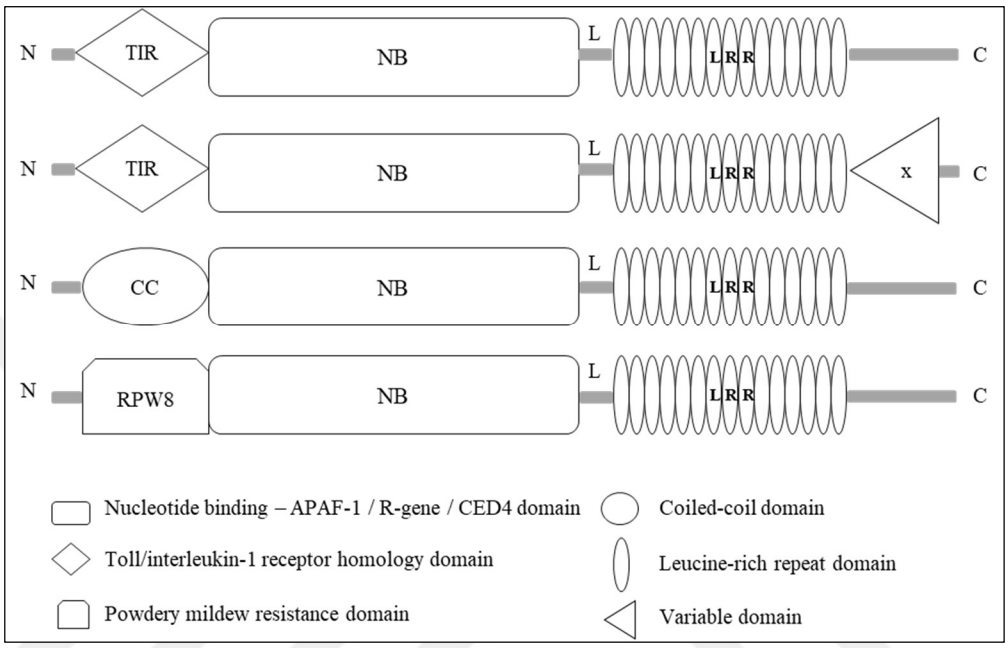

Figure 1.5. Four main classes of R proteins [64]

1.3.3. Systemic Immunity

Systemic acquired resistance (SAR) and induced systemic resistance (ISR) are systemic immunity mechanisms. SAR is induced with SA after a pathogen attack, while ISR is triggered with JA and ET after recruiting beneficial microbes such as plant growth-promoting rhizobacteria (PGPR) and plant growth-promoting fungi (PGPF) [65]. Additionally, pathogenesis-related (PR) proteins, characteristic in SAR, and non-expressor of PR genes1 (NPR1) play a role in long-distance signaling [66].

1.3.4. Plant-Pathogen Interaction Models

R and the corresponding effector proteins directly or indirectly interact. According to the first suggested gene-for-gene model as direct interaction, single plant R gene and single

pathogen Avr gene interact with one another. Flor [67] put forward that resistance (R) in plants and Avr in parasites are dominant, while susceptibility (r) in plants and virulence (avr) in parasites are recessive [68]. In the guard model, as an indirect interaction, R proteins detect modification of the effector's target proteins in a host, which is called "guardee" [69].

In the decoy model as another model with indirect interaction, decoy, duplicated copies of effector target proteins in the host, mimic effector target proteins [69,70]. Both guardee and decoy have been required for R protein function. Guardee proteins with a function trigger innate immunity in plant resistance mechanisms but decoy proteins are not due to functionless [69].

Jones and Dangl [61] have presented a "zigzag" model has four phases related to the plant immune system. According to the zigzag model, firstly, PAMPs are recognized by PRRs in PTI. In the second phase, successful pathogens secrete effectors, which can interfere with PTI, resulting in effector-triggered susceptibility (ETS). In the following stage, the effectors, either indirect or direct, are recognized by R proteins, resulting in ETI. Finally, in the fourth phase, natural selection drives the diversification of pathogens to avoid ETI, and this results in the interaction of new plant R proteins against new pathogen effectors. ETI defense mechanism is more robust than PTI due to co-evolution of R-Avr genes in ETI. The zigzag model is illustrated in Figure 1.6.

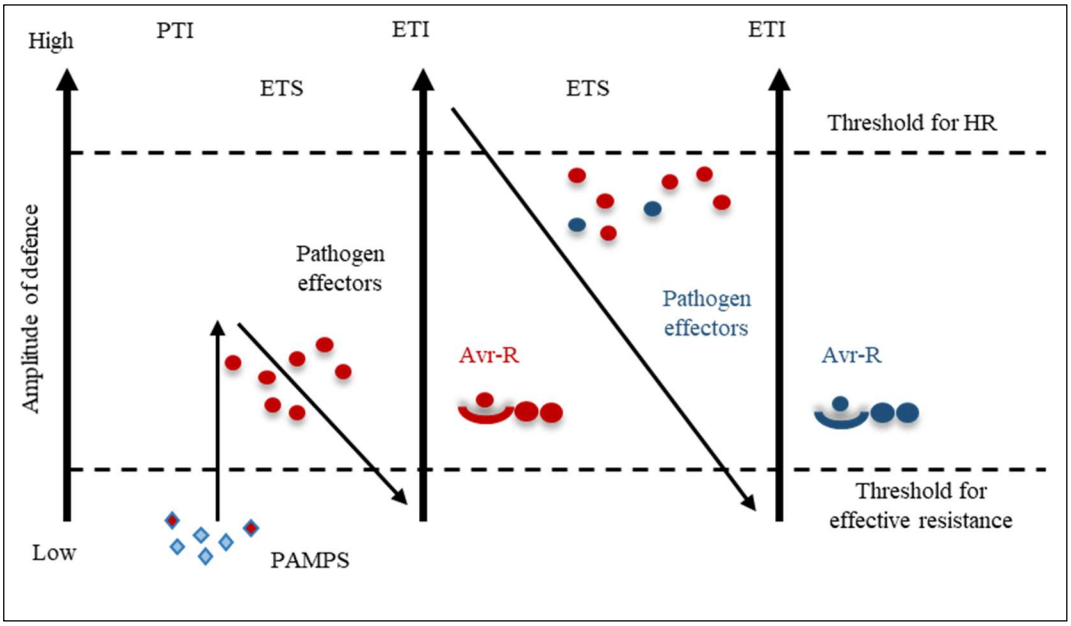

Figure 1.6. A zigzag model of the plant immune system [71]

1.4. PSEUDOGENES

Pseudogenes are disabled genomic sequences that usually occur with loss of function due to mutation accumulation in functional parental genes over time. They were first discovered in the 5S ribosomal DNA of the African claw frog (*Xenopus laevis)* oocyte [72,73]. Pseudogenes have been identified in different living organisms such as bacterium [74], yeast [75], insect [76], nematode [77], plant [78], and human [79]. The overall distribution of pseudogenes is not related to genome size or gene counts. For example, the human genome has about 100-fold, 50-fold, and 15-fold more pseudogenes than fly, zebrafish, and worm [80].

1.4.1. Pseudogene Formation

Pseudogenes have mutations involving premature stop codons, frameshift, insertion, or deletion (indels) that cause blocking the translation of a functional protein [72], as shown in Figure 1.7. Premature stop codons are found before their normal position in the DNA sequence, causing incomplete protein expression. Frameshift is a type of mutation that occurs due to the indel of one or more nucleotides.

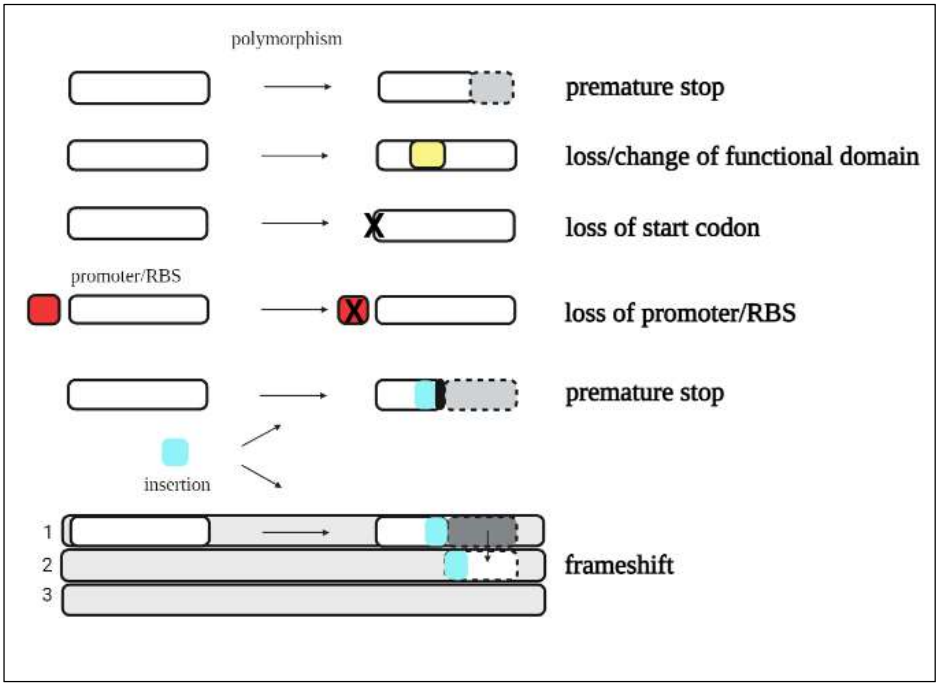

Figure 1.7. Pseudogene formation [81]

Genetic code is degenerate because multiple synonymous codons can code most amino acids except methionine (Met) and tryptophan (Trp). Length, location, time, expression level, translation efficiency, selective force, etc., of the gene may cause the usage of synonymous codons [82].

Point mutation is the base change in the DNA nucleotide classified as silent, missense, and nonsense. A silent mutation is a synonymous mutation that changes nucleotide sequence but does not change protein sequence. A missense mutation is a nonsynonymous mutation that changes both nucleotide and protein sequence. The nonsense mutation causes premature stop codons.

Single nucleotide polymorphism (SNP) is the base change in the DNA nucleotide occurring with 1 percent frequency in the population. Genome analysis toolkit (GATK) and SAMtools are mainly used for SNP calling [83]. SNP density in repeats, introns, and pseudogenes is higher than in the exons on the whole genome. This shows that functional genes are under selective pressure to maintain their functions while pseudogenes have not [84].

1.4.2. Predicting Pseudogenes with Ka/Ks Ratio

The mutation rate of pseudogenes can be measured by dividing the nonsynonymous substitution ratio (Ka) by the synonymous substitution ratio (Ks). Ka/Ks is a parameter that gives information about genes' the fast or slow evolution [85]. Ka results in amino acid change on a given polypeptide, while Ks does not result in a change of amino acid sequence. The Ka/Ks ratio demonstrates differences in amino acid sequences, with ranges of <1, $=1$, and >1 representing negative, neutral, and positive selection [86], respectively. Ka/Ks ratio is usually used to predict constraints on sequence evolution [87].

1.4.3. Pseudogene Classification

Pseudogenes are classified into three groups; i) processed, ii) duplicated, and iii) unitary pseudogenes (Figure 1.8.). mRNA of the parental gene is translated into complementary DNA (cDNA); of that, all or partial is inserted randomly into a new genomic location; that term is called a processed pseudogene. Processed pseudogenes do not contain promoter and

intron regions but poly-A tail. Duplicated pseudogenes accumulate mutations due to gene duplication, and unitary pseudogenes are mutated genes that occur with loss of function in functional genes [88].

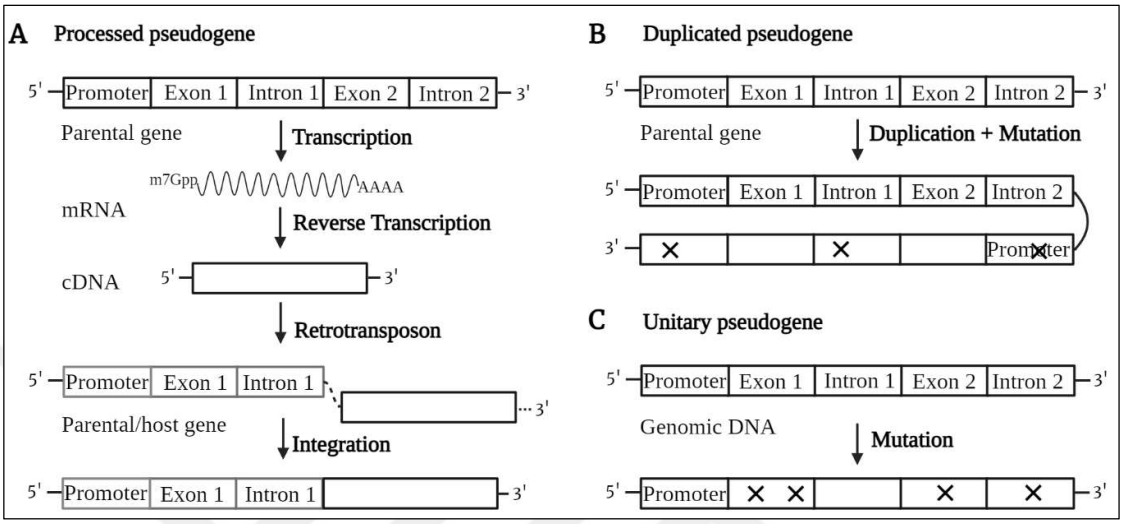

Figure 1.8. Pseudogene classification [89] (A: Processed pseudogene, B: Duplicated pseudogene, C: Unitary pseudogene)

Processed and unprocessed pseudogenes share high homology sequences with their parental genes, while unitary pseudogenes do not have any parental gene. The comparative properties of processed and duplicated pseudogenes are given in Table 1.1.

Table 1.1. Comparison between processed and duplicated pseudogenes [87]

Processed Pseudogenes	**Duplicated Pseudogenes**
Arise from mRNA that was reverse-transcribed and re-integrated into the genome	Arise from gene duplication
Lack of non-coding intervening sequences: introns and promoters	Possess promoters, exon-intron structure, and other regulatory sequences
Possess a poly-A tail at 3' end	No 3' poly-A tail
Possess flanking direct repeats associated with TE insertion sites	No flanking direct repeats
Mostly present at different loci from its parental genes	Some are present as a cluster with their parental gene
Have 3' or 5' truncations	Have 3' truncations
Generally shorter	Comparatively longer

1.4.4. Pseudogene Function

Pseudogenes are considered non-coding DNA, while they have been previously called junk DNA due to their functionless state [90]. Although it is debated whether these pseudogenes are functional or not, they have significant functional roles at the DNA level, such as random insertion and DNA sequence exchange. Pseudogene functions at the DNA level, as shown in Figure 1.9.

Pseudogenes also have functional roles at the RNA level, such as transcription as an antisense transcript, processing into small interfering RNAs (siRNAs), competition for microRNAs (miRNAs), production of long non-coding RNAs (lncRNAs), interaction with RNA-binding proteins (RBPs), and at the protein level such as encoding a protein or peptide [89], as shown in Figure 1.10. miRNAs are 20 to 22 nucleotides in length single-stranded sncRNAs discovered as a transcript of lin-4 miRNA precursor in *Caenorhabditis elegans* (nematode) by Lee in 1993 [91,92]. miRNAs consist of three strategies: pri-miRNA, pre-miRNA, and miRNA:miRNA*, respectively. Firstly pri-miRNAs are transcribed by Pol II enzymes from MIR genes, then they are converted to pri-miRNAs by dicer-like 1 (DCL1), hyponastic leaves1 (HYL1), serrate (SE) proteins in the nucleus. After that pri-miRNAs are transported to the cytoplasm by HASTY (HST). miRNA duplexes (miRNA:miRNA*) are formed by Dicer in RNA-induced silencing complex (RISC complex). From the double miRNA strands, the 5-end stable chain is selected as mature miRNA via Argonaute (AGO) proteins in the RISC complex, while the other chain is degraded as miRNA* [93]. Plant miRNAs are close to a perfect match with the 3'UTR of their target mRNA [94]. siRNAs are 21 to 24 nucleotides in length double-stranded sncRNA sourced from viruses, transposons [92]. They have one target gene with whole complementary and multiple DCL proteins compared to miRNA [95].

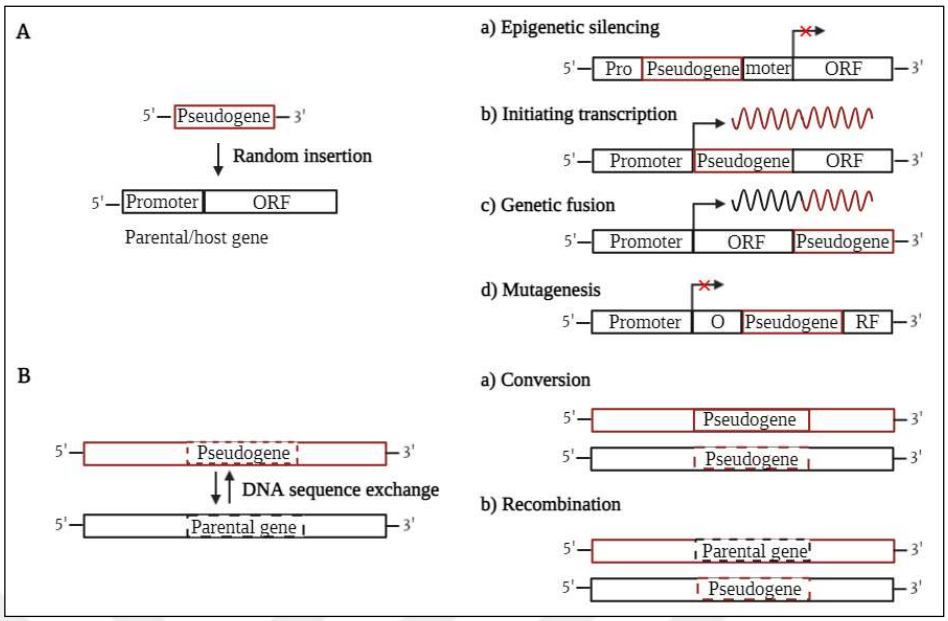

Figure 1.9. Pseudogene functions at the DNA level [89] (A: pseudogene can be randomly inserted into the parental gene, B: parental gene can be influenced by pseudogenes via sequence exchange)

Poliseno *et al.* revealed that pseudogenes have miRNA decoy function showing they are transcriptionally active [96]. Moreover, Kim *et al.* determined that over 200 peptides were translated from 140 pseudogenes [97]. Mostly human pseudogenes such as phosphatase and tensin homolog pseudogene 1 (PTENP1), B-raf proto-oncogene, serine/threonine kinase pseudogene 1 (BRAFP1) were listed together with functions [98].

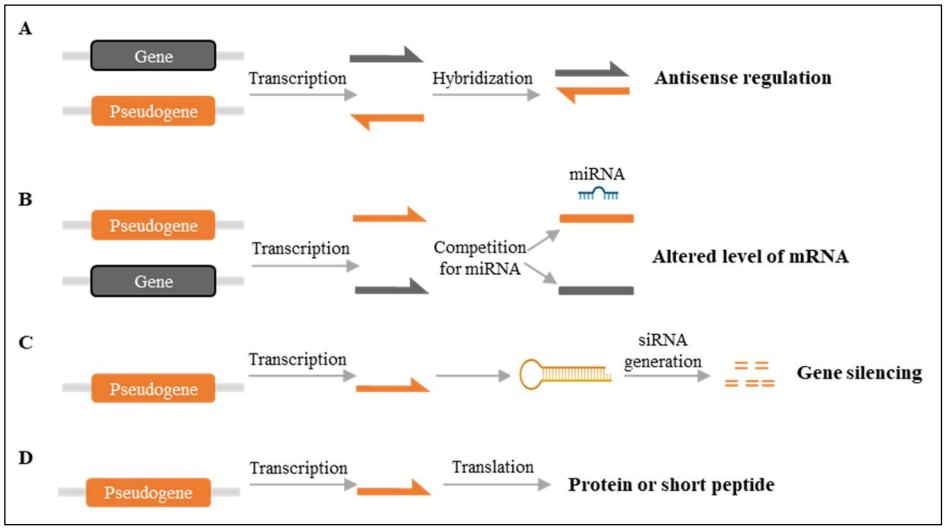

Figure 1.10. Pseudogene functions at the RNA level [99]

1.4.5. Pseudogenes on Literature

Human pseudogenes have been categorized in GENCODE [100], which has an annotation of gene features in the human after the completion of the encyclopedia of DNA elements (ENCODE) project [101], which was initially focused on almost 1 percent of the human genome. 14112 pseudogenes on the human genome have been estimated in GENCODE. Of these pseudogenes, they determined 8248 processed, 2127 duplicated pseudogenes. Moreover, they estimated that at least 9 percent of the pseudogenes in the human genome are actively transcribed. Among 863 transcribed human pseudogenes, 520 were processed and 343 duplicated pseudogenes were identified. Also, the conservation of 3.4 percent of processed pseudogenes and 28.1 percent of duplicated pseudogenes showed that processed and duplicated pseudogenes were differentially conserved [102].

Moreover, some studies related to identifying pseudogenes in rice have been available. Shang *et al.* performed a genome-wide analysis of NBS-LRR genes and their pseudogenes in *Oryza sativa* L. ssp. *japonica* cv. Nipponbare and *Oryza sativa* L. ssp. *indica* cv. 93-11. They identified Pid3, a new rice blast resistance gene, as pseudogenes due to a nonsense mutation in japonica varieties but not in indica varieties [103]. Luo *et al.* analyzed NB-LRR genes in the grass family, two *Oryza sativa* cultivars and one genotype from each of *Zea mays*, *Sorghum bicolor*, *Brachypodium distachyon*. 347 (55.7 percent) out of 623 R genes from rice cultivar Nipponbare and 345 (47.6 percent) out of 725 R genes from rice cultivar 93-11 were identified as pseudogenes. They recognized that most of the pseudogenes were caused by large deletions while the others were caused by nonsense point mutations or small frameshift indels [104].

Thibaud-Nissen *et al.* demonstrated that 1439 pseudogenes were characterized by frameshifts or premature stop codons in rice (*Oryza sativa* L. ssp. *japonica* cv. Nipponbare). Moreover, some protein families, such as F-box and cytochrome P450 contain pseudogenes considerably have been found [105]. Wang *et al.* reported that out of the 1235 retrogenes, 337 retropseudogenes (27 percent) include deletions in the rice genome. Additionally, they observed that retrogenes mostly take part euchromatic region in the genome while retropseudogenes are randomly located on chromosomes [106]. Zou *et al.* characterized that pseudogenes, compared with genes, have lower expression levels in Arabidopsis and rice. Furthermore, protein domain families related to stress responses have more pseudogenes

than transcription factors and receptor-like kinases, which are larger protein domain families [107].

Pseudogenes have been identified using RNA-seq data. Examples in the literature of pseudogene prediction using RNA-seq data; Tonner *et al.* detected 1709 processed and 79 duplicated ribosomal pseudogenes in sixteen different human tissues using RNA-seq data. To reduce FP, they used a strict mapping protocol with no mismatches and a single alignment location. RPL21 pseudogene with RPKM 170, was the highest expressed ribosomal protein pseudogene located intron of Thyroglobulin (TG) gene in the thyroid. Moreover, RPL7A pseudogene with RPKM 17, was found second highly expressed ribosomal protein pseudogene located in the intergenic region in white blood cells [108]. In another study, Guo *et al.* improved the computational methods to determine human pseudogenes using RNA-seq data. They found that expression levels of transcribed pseudogenes were 20 fold lower than protein-coding genes. Moreover, they suggested that transcribed pseudogenes can play essential roles in gene regulation [109]. Mascagni *et al.* characterized pseudogenes according to duplication modes consisting of whole genome, tandem, proximal, transposed, and dispersed in five plant species. They showed that duplicated pseudogenes were more than 10 fold more abundant than processed pseudogenes in *Populus trichocarpa*, *Phaseolus vulgaris*, *Arabidopsis thaliana,* and *Oryza sativa* 2 fold in *Vitis vinifera*. Moreover, transposed pseudogenes have been the shortest CDS coverage in investigated species [110].

Pseudogenes have also been identified using machine learning methods. Shein *et al.* investigated 3'UTR stem-loops in L1 and Alu transposons, mRNA, and processed pseudogenes in the human genome using machine learning models. Just sequence has been used in sequence based models while hydrophilicity and AT-rich loop positions have been the most critical parameters in structured based models. They have discovered that 62–68 percent of processed pseudogenes and mRNAs also have 3'-end stem-loops [111]. Liu *et al.* identified 125 pseudogenes for osteosarcoma using 94 osteosarcoma patients' RNA-seq data. Moreover, they determined 4 prognostic pseudogenes' signatures via COX-regression analysis [112]. Xing *et al.* identified 700 differential expressed pseudogenes using RNA-seq data obtained from the Cancer Genome Atlas. Furthermore, they determined 5 pseudogenes signatures related to immune regulation via COX-regression analysis [113].

1.5. LOW AND HIGH THROUGHPUT METHODS

Pseudogenes may be determined with both low throughput and high throughput methods. According to the size of produced data, polymerase chain reaction (PCR), quantitative real-time PCR (qRT-PCR), and first generation sequencing are low throughput methods, while microarray, second and third generation sequencing technologies are high throughput methods.

PCR is based on amplification in vitro of DNA for multiple copies of the desired gene. Although easy, cheap, and rapid, it has some disadvantages such as the need for gene sequence to design primer, dimer formation, high contamination risk, self-annealing of PCR product, and plateau effect. qRT-PCR depends on the detection and quantification of fluorescence emitted from DNA to analyze gene expression. In this technique, cDNA converted from RNA through reverse transcriptase (RT) enzyme is amplified by PCR and quantified synchronously. Except for its high price level, it has several advantages such as high sensitivity, specificity, reproductivity, reliability, and wide dynamic quantification compared to PCR [114]. Accurately measuring pseudogenes is challenging due to PCR/qRT-PCR bias based on sequence homology, RNA quality, and primer design. RHD pseudogene (RHDψ) was detected using PCR and real-time PCR [115,116].

As for high throughput methods, microarray or DNA chip is a method that depends on the hybridization of fluorescently labeled DNA to their complementary single strand DNA sequences on a solid surface. It has two principal types: oligonucleotide and cDNA microarrays. Microarray system has applications including analysis of gene expression, diagnostic and clinical research, drug discovery, toxicological research, agricultural applications, food quality, microbial detection [117]. Compared to low throughput methods, it provides simultaneous detection between different treatments on a chip. Along with being more advantageous, there are limiting factors in microarray technology. It is complex, relatively insensitive [118], and is more expensive. It has also not been used for gene discovery due to limited known or expressed sequences.

Sequencing is the method used to determine each base involved in any sequence sourced from genome or transcriptome. Sanger sequencing has been involved in first generation sequencing technology. The first genome sequenced through the Sanger method based on

dideoxy chain termination was bacteriophage φX174 [119,120]. Sequencing of the first human genome was performed by this method between 1990 and 2003 [121]. In this method, DNA is fragmented by either supersonic or enzyme digestion, and these DNA fragments are cloned into plasmids for amplification. Then the amplified fragments are sequenced using ddNTPs [122]. This first generation sequencing has an average read length of 700 base pairs (bp) with high accuracy, but it also has several disadvantages, such as including toxic reagents due to radioactive ^{32}P molecules, low production, high cost, and time-consuming.

In 2005, second/next generation sequencing called NGS was discovered. NGS is also known as massively parallel sequencing. The technologies are Roche/454, Illumina/Genome Analyzer and HiSeq-MiSeq, Life Technologies/sequencing by oligonucleotide ligation and detection (SOLiD), Life Technologies/IonTorrent. NGS technologies differ in both sequencing chemistry and clonal amplification. Sequencing chemistry has variance: pyrosequencing in 454; reversible terminator chemistry in Illumina; ligation based sequencing in SOLiD; ion semiconductor sequencing in IonTorrent. 454, Illumina, and IonTorrent are based on "sequencing by synthesis (SBS)" principles, whereas SOLiD relies on "sequencing by ligation (SBL)" [123]. Clonal amplification of sequencing template is performed by emulsion PCR (emPCR) in 454, SOLiD and Ion Torrent [124], and bridge PCR in Illumina [125].

NGS has many applications: whole-genome sequencing (WGS), whole exome sequencing (WES), RNA sequencing (RNA-seq), metagenomics, targeted analysis sequencing (TAS), chromatin immunoprecipitation sequencing (ChIP-seq), and small RNA sequencing (sRNA-seq). WGS, WES, metagenomics, TAS, ChIP-seq are genome analyses, whereas RNA-seq and sRNA-seq are transcriptome analyses. Compared with first generation Sanger sequencing, NGS is reliable, has high speed, throughput, low cost, and is directly detected without electrophoresis [126]. They also do not require any prior knowledge, such as probe selection in microarray, to enable novel sequence discovery. Van Dijk *et al.* reported that Illumina comes to the forefront with the highest throughput and the lowest per-base cost among NGS platforms [127].

Finally, third generation sequencing (TGS) was developed around 2011. Pacific Biosciences (PacBio), Oxford Nanopore/NanoPore, and Helicos Biosciences/HeliScope use third generation technology that depends on single molecule real-time (SMRT), nanopore, and single molecule fluorescent sequencing methods, respectively. Different from NGS, these

systems do not need pre-amplification steps because they have been designed to be sensitive enough to detect a single molecule. Therefore these are called "single molecule" sequencers. This has advantages such as higher throughput, longer read length, higher accuracy, single molecule, low cost [128]. The first, second, and third generation sequencing methods are presented in Table 1.2.

Table 1.2. Comparison of sequencing platforms [129]

	First Generation	**Second Generation**					**Third Generation**	
Platform	Sanger	454	SOLiD	Ion Torrent	Illumina		PacBio	Nanopore
Sequencing method	Chain terminator	Pyrosequencing	Sequencing by ligation	Ion semiconductor	Sequencing by synthesis		Single-molecule real-time (SMRT)	Nanopore sequencing
Instrument	3730xl	GS FLX+	5500 SOLiD	Ion PGM 318	HiSeq 2500	MiSeq	PacBio RSII	MinION
Read Length	400–900 bp	Up to 700 bp	35–50 bp	100–200 bp	50–300 bp	50–300 bp	10–15 kb	100 bp–10 kb
Accuracy	99.90%	99.90%	99.90%	98.00%	98.00%	98.00%	87% or >99.9%	60–80% or >99%
Throughput	N/A	Up to 1 million	1.0–1.5 billion	4–5.5 million	0.6–4 billion	20–30 million	50,000	10,000–50,000
Run time	2 h	20 h	1–2 weeks	2 h	6 h to 11 days	6–40 h	2 h	6 h
Instrument cost	++	++	++++	++	++++	++	++++	+
Sequencing cost	++++	+++	+	++	+	+	+	?
Key advantages	Long reads; fast run times	Long reads; fast run times	Low cost per base	Fast run times	Highest yield; low cost per base	FDA cleared; low cost per base	Long read	Real-time sequencing; portable; long reads
Key disadvantages	Lowest throughput	Low throughput; homopolymer errors	Very short reads; slow	Homopolymer errors	Instrumentation expensive	Lower throughput	Instrumentation expensive	High error rate; low throughput

1.5.1. RNA-seq

RNA-seq is an approach that enables the identification of genes expressed under specific conditions using sequencing methods or microarray. It aims to determine and quantify transcriptomics (mRNA, ncRNA) data. Moreover, this reveals novel transcript isoforms and unknown genes. Experimental design, pre-processing, mapping, transcript assembly, abundance estimation, differential expression analysis, and visualization are usually found in the flow chart of RNA-seq [130].

RNA-seq datasets related to biotic stress in the rice have available in the sequence read archive (SRA) at the national center for biotechnology information (NCBI), a publicly available database. Table 1.3. provides detailed information about RNA-seq datasets related to biotic stress in rice, such as SRA number, host, pathogen, tissue, instrument, library layout, and byte. As it can be seen from the table, the most commonly used instrument is Illumina HiSeq 2000, the library layout is paired, the sample is the leaf, the most studied pathogens are fungus and bacteria. *Fusarium* and *Magnaporthe* among the fungi genera, *Blumeria,* and *Xanthomonas* among bacteria genera are noteworthy pathogens.

Table 1.3. List of RNA-seq datasets related to biotic stress in rice

	SRA number	**Pathogen**	**Tissue**	**Instrument**	**Library Layout**	**Run**	**Bytes**
Fungus	SRP344444	*Rhizoctonia solani*	Leaf	Illumina NovaSeq 6000	Paired	10	19.50 Gb
	SRP322296	*Ustilaginoidea virens*	Spikelet	Illumina HiSeq 2500	Single	12	48.97 Gb
	SRP222559	*Magnaporthe oryzae, (Pyricularia oryzae)*	Leaf	Illumina HiSeq 2500	Paired	12	14.28 Gb
	SRP212323	*Rhizoctonia solani*	Leaf	Illumina HiSeq 2500	Paired	12	47.08 Gb
	SRP186638	*Pyricularia oryzae*	Leaf	Illumina HiSeq 2000	Paired	18	46.75 Gb
	SRP165938	*Tilletia horrida*	Kernel	Illumina HiSeq 4000	Paired	36	120.48 Gb
	SRP127211	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 2500	Paired	22	143.93 Gb
	SRP095383	*Magnaporthe oryzae (Pyricularia oryzae)*	Leaf	Illumina HiSeq 2500	Paired	30	39.24 Gb
	SRP068995	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 2000	Single	10	1.85 Gb

Fungus	SRP068861	*Rhizoctonia solani*	Leaf	Illumina Genome Analyzer IIx	Single	1	12.25 Gb
	DRP000568	*Magnaporthe oryzae*	Leaf	Illumina Genome Analyzer IIx	Single	13	5.09 Gb
	SRP324816	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 2000	Paired	16	103.34 Gb
	SRP227501	*Rhizoctonia solani*	Steam	Illumina HiSeq 2000	Single	4	379.9 Mb
	SRP199834	*Magnaporthe oryzae*	Leaf	Illumina HiSeq X Ten, NextSeq 500	Paired	18	267.23 Gb
	SRP297630	*Rhizoctonia solani*	Leaf	Illumina HiSeq 2500	Paired	8	26.15 Gb
	SRP200361	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 2500	Single	40	3.52 Gb
	SRP131078	*Magnaporthe oryzae, Fusarium fujikuroi*	Leaf	Illumina HiSeq 2500	Single	8	2.69 Gb
	SRP049444	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 2000	Paired	48	57.74 Gb
	SRP262611	*Magnaporthe oryzae*	Leaf	NextSeq 550	Single	12	4.49 Gb
	SRP153207	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 1000	Paired	48	68.39 Gb
	SRP075722	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 1000	Paired	24	32.55 Gb
	SRP309552	*Magnaporthe oryzae*	Leaf	Illumina HiSeq 2500	Paired	6	16.76 Gb
	SRP150958	*Magnaporthe oryzae*	Root	HiSeq X Ten	Paired	12	35.90 Gb
	SRP111367	*Magnaporthe grisea*	Root	HiSeq X Ten	Paired	13	40.11 Gb
	SRP326551	*blast fungus*	Leaf	Illumina HiSeq 2000	Paired	16	39.70 Gb
	SRP321240	*Piriformospora indica*	Root	BGISEQ-500	Single	12	7.30 Gb
	SRP263499	*Pyricularia oryzae*	Leaf	Illumina HiSeq 2000	Paired	18	36.04 Gb
	SRP064943	*Rhizoctonia solani*	Sheath	Illumina HiSeq 2000	Paired	8	11.22 Gb
	SRP079683	*Magnaporthe oryzae*	Shoot	Illumina HiSeq 2000	Single	18	9.17 Gb
	SRP168030		Leaf	Illumina HiSeq 2500	Paired	8	8.89 Gb
	SRP017770	*Magnaporthe oryzae*	Mycelium	Illumina Genome Analyzer IIx	Single	8	2.65 Gb
Bacteria	SRP321505	*Xanthomonas oryze* pv. *oryzicola* (Xoc)	Leaf	Illumina NovaSeq 6000	Paired	18	39.16 Gb
	SRP261005	Xoc	Leaf	Illumina HiSeq 4000	Paired	30	103.13 Gb
	SRP231412	Xoc	Leaf	Illumina HiSeq 2500	Single	16	16.72 Gb
	ERP110296	Xoc	Leaf	Illumina HiSeq 2500	Single	12	12.63 Gb

Bacteria	SRP154928	Xoc	Leaf	Illumina HiSeq 2500	Paired	9	26.24 Gb
	ERP114552	*Paraburkholderia kururiensis, Burkholderia vietnamiensis, Paraburkholderia phytofirmans*	Leaf, Root	Illumina HiSeq 2500	Single	24	28.19 Gb
	SRP149593	*Gluconoacetobacter diazotrophicus, Bradyrhizobium japonicum*	Root	Illumina HiSeq 2500	Paired	12	52.62 Gb
	SRP071314	*Xanthomonas oryzae* pv. *oryzae* (Xoo)	Leaf	HiSeq 2000	Single	48	116.51 Gb
	SRP015433	*Burkholderia glumae*	Leaf	Illumina HiSeq 2000	Single	36	13.54 Gb
	SRP101342	*Xanthomonas oryzae* (Xo)	Leaf	Illumina HiSeq 2500	Single	28	19.83 Gb
	SRP132257	Xo	Leaf	Illumina HiSeq 2500	Paired	10	30.16 Gb
	SRP066357	Xo	Leaf	Illumina HiSeq 2000	Single	3	3.11 Gb
	SRP057260	Xo	Leaf	Illumina HiSeq 2000	Single	6	12.39 Gb
	SRP049040	Xo	Leaf	Illumina HiSeq 2000	Paired	18	24.06 Gb
	SRP320302	*Azoarcus olearius*	Root	NextSeq 500	Paired	60	82.75 Gb
	SRP187838	*Azospirillum brasilense, Herbaspirillum seropedicae*	Root	Illumina HiSeq 4000	Paired	12	51.65 Gb
	SRP127526	Xoo	Leaf	Illumina HiSeq 2000	Paired	6	9.42 Gb
	SRP071037	Xoo	Leaf	Illumina HiSeq 2000	Paired	9	31.73 Gb
Nematode	SRP267874	*Meloidogyne graminicola*	Root, Gall	NextSeq 500	Single	6	13.27 Gb
	SRP042017	*Hirschmanniella oryzae*	Shoot	Illumina Genome Analyzer IIx	Paired	8	3.85 Gb
	SRP092466	*Meloidogyne graminicola*	Shoot	Illumina Genome Analyzer IIx	Paired	4	1.87 Gb
	SRP161893	*Hirschmanniella sp.*	Root	Illumina HiSeq 2000	Paired	4	16.36 Gb
	SRP292916	*Rice root-knot*	Root	HiSeq X Ten	Paired	6	8.73 Gb
	SRP010960	*Meloidogyne graminicola, Hirschmanniella oryzae*	Root	Illumina Genome Analyzer IIx	Single	21	4.21 Gb
	SRP181124	*Meloidogyne graminicola*	Root, Gall	NextSeq 500	Single	48	5.12 Gb

Virus	SRP278002	*Rice black-streaked dwarf virus (RBSDV)*	Leaf	Illumina HiSeq 2500	Paired	6	27.76 Gb
	SRP115030	*Rice dwarf virus (RDV)*	Leaf	Illumina HiSeq 4000, HiSeq X Ten	Paired	17	25.80 Gb
	SRP090211	*Rice stripe virus (RSV)*	Leaf	Illumina HiSeq 2000	Single	6	3.23 Gb
	SRP028592	*RSV-Laodelphax striatellus*	Leaf	Illumina Genome Analyzer IIx	Paired	4	6.49 Gb
	SRP261278	*RBSDV*	Leaf	Illumina HiSeq 4000	Paired	12	39.29 Gb
	SRP065503	*RSV*	Leaf	Illumina HiSeq 2000, NextSeq 500	Single	6	2.72 Gb

1.6. BIOINFORMATICS AND MACHINE LEARNING

Bioinformatics is an interdisciplinary science that covers the main branches of biology, biochemistry, computer science, mathematics, and statistics enabling the processing of large volumes of biological data in silico, the term coined by Hogeweg in the 1970s [131]. Bioinformatics provides convenience in evaluating data that is too complex for the human brain to handle and too large to be calculated with paper and pencil. It is used to acquire, analyze, interpret, apply, integrate, and store biological data.

Data are unorganized facts that can be form of numbers, texts, images, audio, and videos. Biological data can be sequences of nucleotide/amino acid, 3D structures of nucleotide/protein, biological pathways, gene expression, molecules' biological activity, and proteins' interactions. Information is the meaningful form of organized, processed, structured, and presented data. Knowledge is an integrated form of information obtained from the interpretation of data.

Various databases and tools enable data to be interpreted and transformed into information. Database occurring collection of organized data has divided into two categories primary and secondary. The list of frequently used databases in bioinformatics is given in Table 1.4.

Table 1.4. List of essential biological databases

	Database	Content	URL
Primary	National Center for Biotechnology Information (NCBI)	Submit, develop, analyze, research services	https://www.ncbi.nlm.nih.gov/
	GenBank	Sequence data	https://www.ncbi.nlm.nih.gov/genbank/
	European Nucleotide Archive (ENA)		https://www.ebi.ac.uk/ena/browser/home
	Data Bank of Japan (DDBJ)		https://www.ddbj.nig.ac.jp/index-e.html
Secondary	Ensembl	Genome browser	https://www.ensembl.org/index.html
	UniProt	Sequence and function of protein	https://www.uniprot.org/
	Protein Data Bank (PDB)	3D shapes of proteins, nucleic acids	https://www.rcsb.org/
	STRING	Protein-protein interactions	https://string-db.org/
	Kyoto Encyclopedia of Genes and Genomes (KEGG)	Process of large-scale molecular datasets	https://www.genome.jp/kegg/

Bioinformatics tools have been developed to manage and analyze biological data. The important bioinformatics tools for RNA-seq are listed in Table 1.5. according to their categories.

Table 1.5. List of necessary bioinformatics tools for RNA-seq

Category	Tool	Content	Reference
Sequence search	BLAST	Comparison nucleotide or protein sequences to sequence databases	[132]
Mapping	Bowtie2	Aligning sequencing reads to reference genome	[133]
Quantification	Cufflinks	Assembling transcripts and analyzing differential expressed genes	[134]
Data storage	SRA Toolkit	Storing raw sequence data from NGS	[135]
NGS manipulation	SAMtools	Utility for SAM/BAM/CRAM format	[136]
Genome arithmetic	BEDTools	Toolset for genome arithmetic	[137]
Variant	VarScan	Platform-independent mutation caller	[138]
Pre-processing	FastQC	Quality control tool for NGS data	[139]
Pre-processing	Trimmomatic	Trimming of reads	[140]
Browser	Artemis	Visualization of sequence features	[141]

Omics approaches have been used in bioinformatics. Genomics is used for structural and functional characterization of genes, transcriptomics is for regulation and expression profiling of genes, proteomics is for identifying proteins and protein-protein interactions, and metabolomics is for identifying metabolites and hormone profiling, etc [142].

Bioinformatics tools (e.g., mapping, assembly, annotation, visualization) are necessary for processing large-volume data, while machine learning tools are required to extract meaning or information from this large-volume complex data. Although machine learning has attracted the attention of scientists and practitioners for a long time, it has come to the fore in the evaluation of large-volume data, especially with the increase in computer capacity in recent years.

Machine learning is broadly evaluated in a range from linear regression to artificial neural networks (ANN), from random forest to support vector machines. Its purpose is to "learn" information from data or extract patterns. The main disadvantage of these methods is that they require large amounts of data [143].

Data mining aims to extract useful knowledge from data using pre-processing, classification, regression, clustering, association, and visualization tasks. Machine learning uses various

algorithms for these tasks. It has two approaches consist of supervised and unsupervised learning. Supervised learning is based on predicting a model with labeled data, while unsupervised learning is based on the grouping of data. Classification and regression are supervised learning methods, and clustering is an unsupervised learning method [144]. Support vector machines (SVM), k-nearest neighbor (KNN), discriminant analysis are often used as classification methods. Linear regression, decision trees are used for regression. K-means, hidden markov algorithms are used for clustering. ANNs are used for classification, regression, and clustering [145]. A comparison of four non-parametric algorithms was presented by Boateng [146]. Parametric methods, like classification, fit a parametric model to the training data and interpolate to classify test data, while nonparametric methods use alternative means to determine classifications.

ANN is a nonlinear, parametric classification algorithm with high accuracy. Inspired by the biological neural network in humans, it consist of an input layer, one or more hidden layers, and an output layer. It is classified in two ways: feedforward and feedback (recurrent) ANN, given in Figure 1.11.

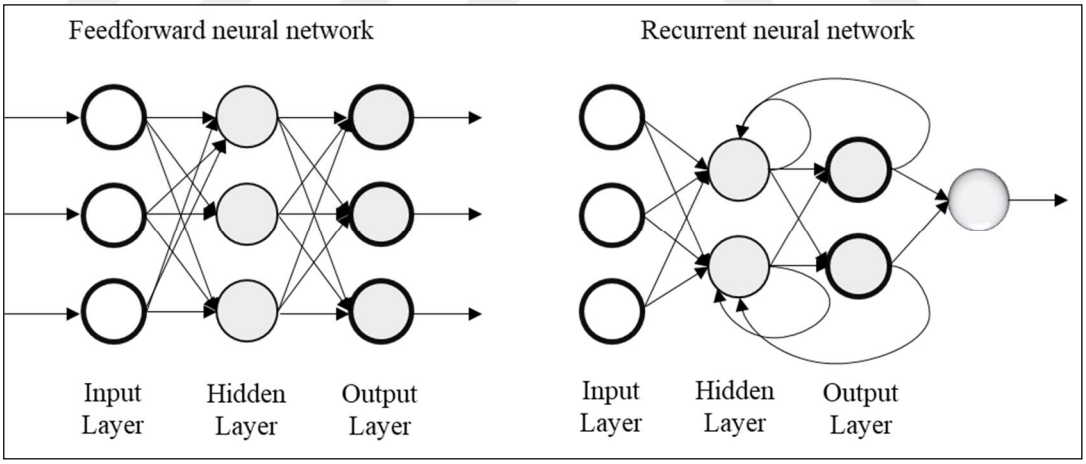

Figure 1.11. Feedforward ANN and recurrent ANN [147]

Multilayer perceptron (MLP), a class of ANN, is a nonlinear and parametric algorithm used for both classification and regression. It consists of artificial neurons that are also named nodes. Synapses are modeled by the set of inputs. The cell body is modeled by an activation function. Weighted inputs are collected and filtered in this. The model of an artificial neuron is shown in Figure 1.12.

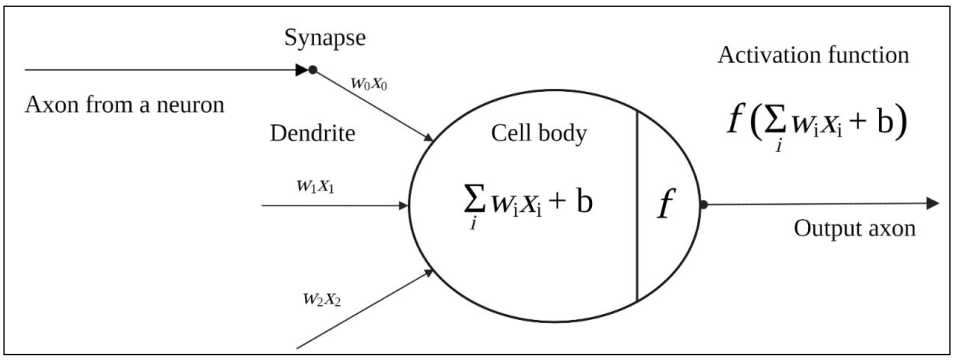

Figure 1.12. The artificial neuron [148]

Sigmoid, tanh or hyperbolic tangent, rectified linear unit (ReLu), softplus are commonly used as activation functions in ANN. ReLu is used as an activation function in a hidden layer in MLP. Activation function formulas in neural networks are shown in Table 1.6.

Table 1.6. Activation function formulas [149]

Activation Function	**Equation**
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
Tanh	$f(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$
ReLu	$f(x) = \max(0, x)$
Softplus	$f(x) = \ln(1+e^{x})$

Evaluation is essential in measuring the effectiveness of algorithms. A confusion matrix describes the complete performance of the models. According to the confusion matrix, true positive (TP) and true negative (TN) represent values correctly classified in the model. False positive (FP) or type I error indicate negatives in fact, while a false negative (FN) or type II error sign positives as such (Table 1.7.).

Performance metrics produced from the confusion matrix are used to evaluate the model. Accuracy is correct predictions over the total number of instances. Specificity is correct negative predictions to the total negative instances. Sensitivity (recall) is positive predictions to the total positive instances. Precision is correct positive predictions to the total predicted positives. F-measure is the harmonic mean between recall and precision values.

Table 1.7. Confusion matrix

		Actually	
		Positive	**Negative**
Predicted	**Positive**	True Positive (TP)	False Positive (FP) **Type I Error**
	Negative	False Negative (FN) **Type II Error**	True Negative (TN)

2. AIM OF THE THESIS

This thesis aims to identify pseudogenes related to biotic stress in rice that play a role in resistance mechanisms using the relevant RNA-seq data from the literature. The underlying hypothesis is that pseudogenes associated with biotic stress can be biomarkers for disease resistance mechanisms in rice.

Although several studies related to understanding resistance mechanism with biotic stress treatments in rice from a molecular perspective are available in the literature, how pseudogenes play a role against these stress factors is still unknown. Moreover, pseudogenes can be misannotated due to the similarity to functional genes, causing complexity in evolutionary genomic studies. It has been recently determined that pseudogenes, previously considered junk genes, play an important role in gene regulation. Furthermore, world rice production has increased by 3.3 percent in the last five years [150], while the human population in the world has increasingly reached around 7.8 billion, and is estimated to reach 10 billion by 2057 [151]. As long as the rate of rice production continues to move so slowly and the worldwide population increases at a faster rate, scarcity of rice will be inevitable. Therefore, all factors which reduce rice production must be reduced as soon as possible.

Within this scope, the role of pseudogenes in disease resistance and which pseudogenes will be related to various biotic stress factors as markers in rice was studied. In this study, pseudogenes play in resistance mechanism against biotic stress in rice were identified using various bioinformatics and machine learning tools. Furthermore, new workflows consisting of DEG analysis, the discovery of pseudogenes, and functional annotation of pseudogenes are generated within this thesis.

3. MATERIALS AND METHODS

3.1. DATASETS

RNA-seq datasets related to biotic stress in rice were downloaded from SRA in NCBI, a publicly available database. SRP010960 dataset related to nematode *Meloidogyne graminicola* stress in rice and SRP049444 dataset related to fungus *Magnaporthe oryzae* ZB13 and ZH10 stress in rice were used in this study.

A total of 21 sample files belonging to *Oryza sativa* ssp. *japonica* cv. Nipponbare (GSOR-100) with single-read library sequenced on Illumina Genome Analyzer IIx instrument were used for SRP010960 dataset. These samples are; mock (SRR408729, SRR408730, SRR408731) and *Meloidogyne graminicola*-infected (SRR408739, SRR408740, SRR408741) at 3 dpi, mock (SRR408732, SRR408733, SRR408734) and *Meloidogyne graminicola*-infected (SRR408742, SRR408743, SRR408744) at 7 dpi, mock (SRR408735, SRR408736) and *Hirschmanniella oryzae*-infected (SRR408745, SRR408746, SRR408747) at 3 dpi, mock (SRR408737, SRR408738) and *Hirschmanniella oryzae*-infected (SRR408748, SRR408749) at 7 dpi.

A total of 48 sample files belonging to *Oryza sativa* ssp. *japonica* cv. Nipponbare (GSOR-100) with paired read library sequenced on Illumina Hiseq 2000 instrument were used for SRP049444 dataset. These samples are; mock samples at 0 h (SRR1636813, SRR1636814), at 8 h (SRR1636825, SRR1636826), at 16 h (SRR1636837, SRR1636838), at 24 h (SRR1636849, SRR1636850), and *Magnaporthe oryzae* ZB13-infected samples at 0 h (SRR1636815, SRR1636816), at 8 h (SRR1636827, SRR1636828), at 16 h (SRR1636839, SRR1636840), at 24 h (SRR1636851, SRR1636852), and *Magnaporthe oryzae* ZH10-infected samples at 0 h (SRR1636817, SRR1636818), at 8 h (SRR1636829, SRR1636830), at 16 h (SRR1636841, SRR1636842), at 24 h (SRR1636853, SRR1636854) for Pid3 rice race, mock samples at 0 h (SRR1636819, SRR1636820), at 8 h (SRR1636831, SRR1636832), at 16 h (SRR1636843, SRR1636844), at 24 h (SRR1636855, SRR1636856), and *Magnaporthe oryzae* ZB13-infected samples at 0 h (SRR1636821, SRR1636822), at 8 h (SRR1636833, SRR1636834), at 16 h (SRR1636845, SRR1636846), at 24 h

(SRR1636857, SRR1636858), and *Magnaporthe oryzae* ZH10-infected samples at 0 h (SRR1636823, SRR1636824), at 8 h (SRR1636835, SRR1636836), at 16 h (SRR1636847, SRR1636848), at 24 h (SRR1636859, SRR1636860) for Pi9 rice race.

3.2. BIOINFORMATICS ANALYSIS

3.2.1. Overview of Workflow

The DEG analysis followed by pseudogene discovery workflow is shown in Figure 3.1. and Figure 3.2.

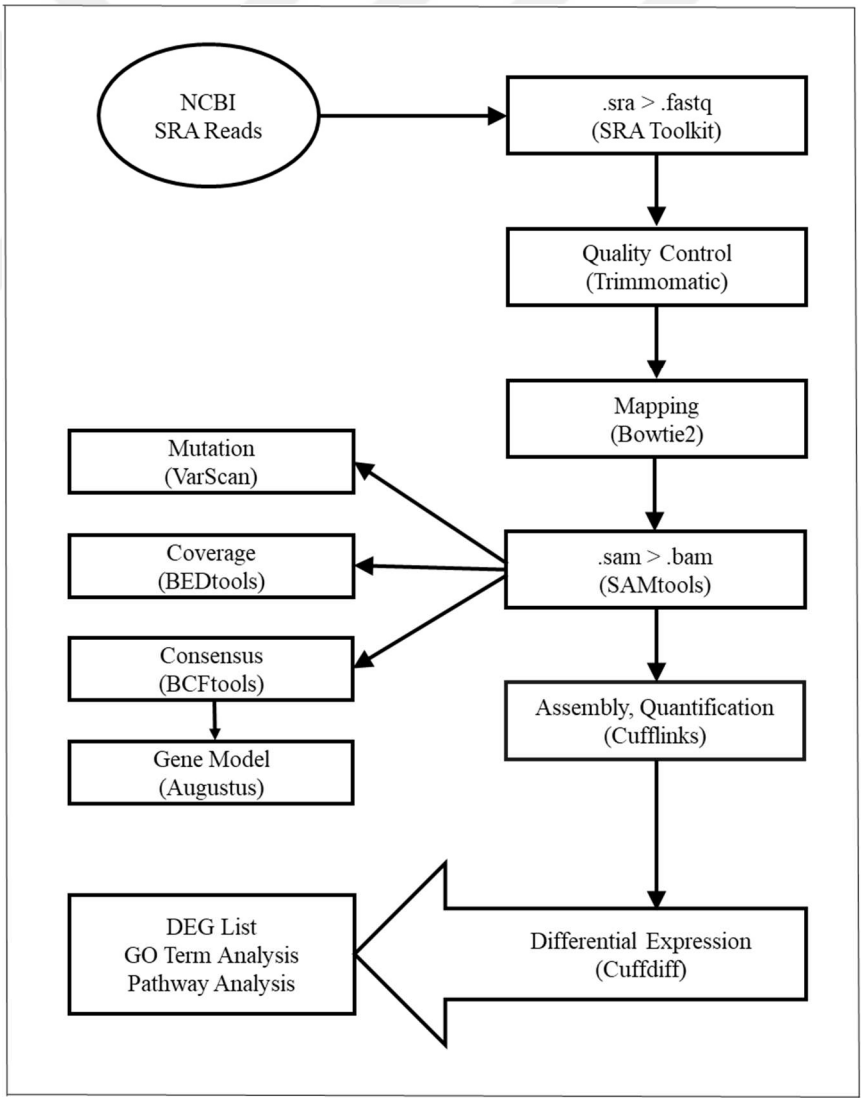

Figure 3.1. Workflow for DEG analysis using NGS data

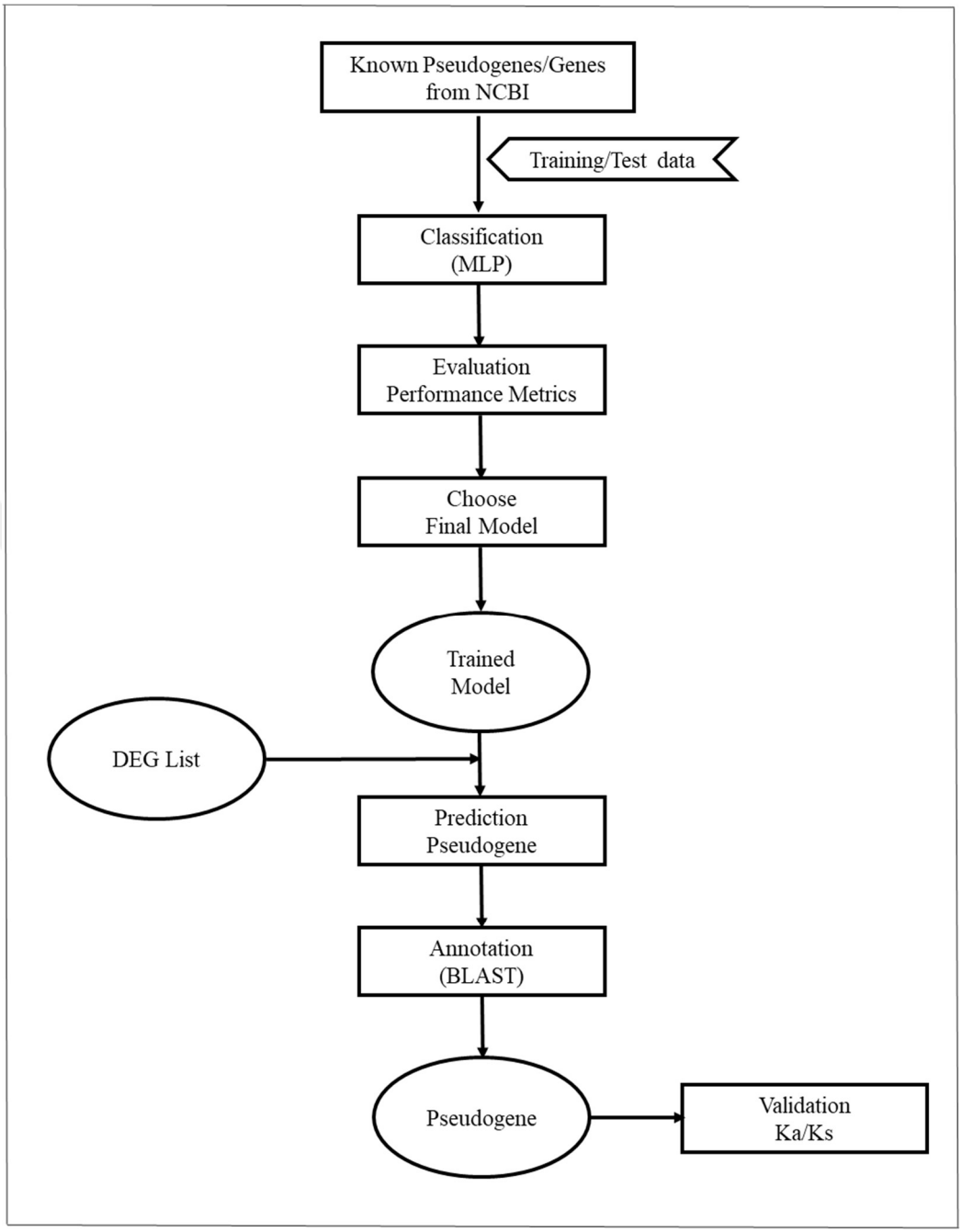

Figure 3.2. Workflow for pseudogene discovery

3.2.2. Pre-Processing

SRA files that contain raw sequence data were converted to FASTQ files via SRA Toolkit [135]. After quality control of FASTQ files was checked via FastQC [139], reads were trimmed and filtered via Trimmomatic [140] if the quality score was less than 20.

3.2.3. Mapping

Mapping is the critical step that informs the locations of the reads in the genome. "IRGSP-1.0_genome.fasta" was used as a reference genome file for *Oryza sativa* subsp. *japonica* and was indexed by Bowtie2. All high quality FASTQ files were aligned to the reference genome via Bowtie2 [133] and were reported as sequence alignment/map (SAM) files.

3.2.4. Assembly

SAM files were converted to binary alignment/map (BAM) files and sorted via SAMtools [136]. The sorted BAM files were visualized on genome browsers such as Artemis [141]. Then they were fed to assembly Cufflinks [134] package consisting of Cufflinks, Cuffmerge, and Cuffdiff. Reads of sorted BAM files were assembled into transcripts on Cufflinks, transcript GTF files were merged on Cuffmerge, and fragments per kilobase of transcript per million mapped reads (FPKM) values of transcripts were normalized on Cuffdiff.

3.2.5. Differential Expression

Transcripts with statistical significance were calculated to identify DEGs between two groups via Cuffdiff. Fold change (FC), t-test, p-value, and q-value were calculated using FPKM values. According to Benjamini Hochberg's correction, transcripts signed with "yes" were used as DEGs. Volcano plots of DEGs were generated using fold change on the x-axis and the -log10 of the p-value on the y-axis in RStudio 4.1.2 [152].

3.2.6. Identification of DEGs

The genomic locations of DEGs were obtained from sorted BAM files via SAMtools and then converted to a browser extensible data (BED) file containing tab-delimited three-columns chromosome, start, end. And then, BED files were matched to the locations in the transcripts.gff file downloaded from the rice annotation project database (RAP-DB) [153] via BEDTools [137].

3.2.7. Coverage Plots

The coverage plot gives information about how many genes are expressed in which exon regions are expressed. Depth of coverage at each genomic location was found through SAMtools. Lastly, coverage plots were created with genomic locations on the x-axis and the numbers of reads overlapping in the genome on the y-axis.

3.2.8. Gene Ontology (GO)

GO analysis was performed to classify the biological functionality of DEGs using GO terms in three categories: molecular function, biological process, cellular component. GO terms related to biotic stress in rice were found using g:Profiler web server [154].

3.2.9. KEGG

KEGG is a knowledge base that consists of sixteen databases (e.g. PATHWAY, KEGG ONTOLOGY (KO), COMPOUND, GLYCAN, ENZYME, VARIANT) [155]. It gives information about metabolism (carbohydrate, energy, lipid, nucleotide, amino acid, vitamins, cofactors, macromolecules, etc.), cell process (membrane transport, signal transduction, ligand-receptor interaction, etc.), and cell organization [156]. KEGG was used to understand the functional annotation of DEGs at the molecular level by processing large-scale datasets.

3.2.10. Variant Analysis

Gene mutation analysis was performed to understand whether they accumulated the mutation or not, which is an important criterion for pseudogenes. Mutations in genomic locations were detected via SAMtools and VarScan 2 [138]. Read depth and mapping quality greater than 20 were filtered out to avoid mismatch detection due to sequencing error. Calling variants and getting consensus sequences were performed using call and consensus parameters on BCFtools, respectively.

3.2.11. Gene Model

Gene models contain a prediction of exon-intron structures on genomic sequence. AUGUSTUS [157] is a tool for finding the transcript, coding sequence (CDS), intron, exon structures on the sequence. The gene models were obtained from the consensus sequences using "partial" parameter on AUGUSTUS.

3.2.12. Ka/Ks Ratio

Ka/Ks ratio was calculated to determine the rate of amino acid change in pseudogenes. Several methods such as Nei-Gojobori 1986 [158], Li-Wu-Luo 1985 [159], Pamilo-Bianchi-Li 1993 [160], and PAML [86] are available for Ka/Ks calculation. dnds function on MATLAB [161], which estimates synonymous and nonsynonymous substitution rates, was used for Ka/Ks calculation. The Ka/Ks ratios of the sequences were calculated by comparing the sequences on codon-based alignment.

3.3. MACHINE LEARNING ANALYSIS

3.3.1. Collection of Training/Test Data

For the training and test data, pseudogenes and corresponding genes belonging to cereal were downloaded from NCBI Nucleotide sections as a result of an advanced search by filtering (Oryza/Triticum/Hordeum/Zea/Avena/Secale/Sorghum/Brachypodium/Poa/Setaria, etc. [Organism] AND pseudogene/gene [Title]). The data were randomly split into: 85 percent as training data and 15 percent as test data. Pseudogenes and genes shorter than 3000 nt and larger than 200 nt were filtered.

3.3.2. Construction of the MLP Classifier

Construction of MLP was performed using training and test data in the Waikato environment for knowledge analysis (WEKA). Attribute-relation file format (ARFF) is used as a file format in WEKA.

To generate an ARFF file, optimum nucleotide lengths of sequences were found. Then totally, 64 codon counts were calculated with the codoncount function on MATLAB. Only four codons AAA, stop codons (TAA, TAG, TGA) were used as attributes in the ARFF file. Nucleotides in the sequences were converted to numerical values to obtain high accuracy in classification. And then, sequences were labeled with class names as pseudogene or gene. ARFF file consists of 205 features because 200 nucleotides are taken separately. ARFF file containing Length, PolyA, StopTAA, StopTAG, StopTGA, 200 nt length sequences as numeric and Class as nominal attributes were fed in WEKA, as given in Table 3.1.

Table 3.1. Attributes used for ARFF file in WEKA

Attribute	**Description**	**Type**
Length	Length of the sequence	Numeric
PolyA	AAA codon numbers of the sequence	Numeric
StopTAA	TAA stop codon numbers of the sequence	Numeric
StopTAG	TAG stop codon numbers of the sequence	Numeric
StopTGA	TGA stop codon numbers of the sequence	Numeric
$S_1 \ldots S_{200}$	Numeric values of the sequence	Numeric

10-folds cross-validation was applied to determine the predictive accuracy of the model. The calculation was made in 10 iterations with 90 percent training and 10 percent test data.

3.3.3. Evaluation of Models

The model obtained for pseudogene prediction was evaluated with performance metrics as given in Table 3.2.

Table 3.2. Performance metrics used for MLP

Performance Metric	**Formula**
Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$
Specificity	$\frac{TN}{TN + FP}$
Sensitivity (Recall)	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
F-measure	$\frac{2 * Precision * Recall}{Precision + Recall}$

The receiver operating characteristic (ROC) curve is a performance metric for binary classification algorithms. It is plotted with the true positive rate (TPR) against the false positive rate (FPR). Classifiers that give curves closer to the top-left corner indicate better performance.

4. RESULTS

4.1. PROCESSING RNA-SEQ DATA

RNA-seq pipeline consisting of quality control of reads, aligning reads to reference genome, differential gene expression, and annotation of DEGs was performed using RNA-seq reads.

4.1.1. Quality Control

SRA files related to biotic stress in rice were converted to FASTQ files via SRA Toolkit. Reads were primarily checked via FastQC using several parameters such as sequence quality, GC content, N content, length, duplication levels, etc. Adapter sequences, as well as low quality reads (Q<20) have been discarded through Trimmomatic.

Sequencing data downloaded as a single read (76 bp) produced by Genome Analyzer IIx is 4.20 GB. FASTQ files belonging to 21 samples are 24 GB before trimming of reads. After cleaning low reads, adapters, etc., with Trimmomatic, the size of the data files resulted in 13 GB. For example, 37.40 percent out of 4831121 reads in mock SRR408729 sample and 36.95 percent out of 3855375 reads in *Hirschmanniella oryzae*-infected SRR408747 sample were dropped for trimming. Sequence quality graph per base for samples mock SRR408729, and *Hirschmanniella oryzae*-infected SRR408747 are shown in Figure 4.1.

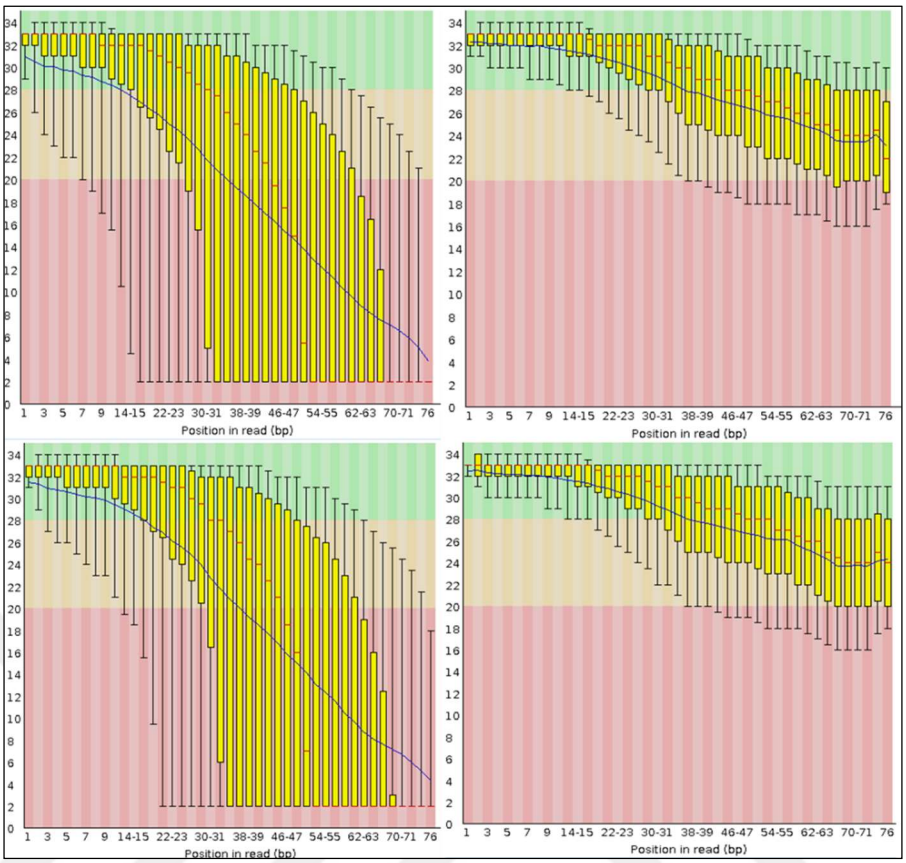

Figure 4.1. Before and after quality control of reads in SRR408729 and SRR408747

Sequencing data downloaded as a paired read (200 bp) produced by Illumina Hiseq 2000 is 43.5 GB. FASTQ files belonging to 48 samples are 151 GB. The trimming process was not performed because the samples' sequence quality was good. Sequence quality graph per base for samples mock SRR1636825 and *Magnaporthe oryzae* ZB13-infected SRR1636827 samples are shown in Figure 4.2.

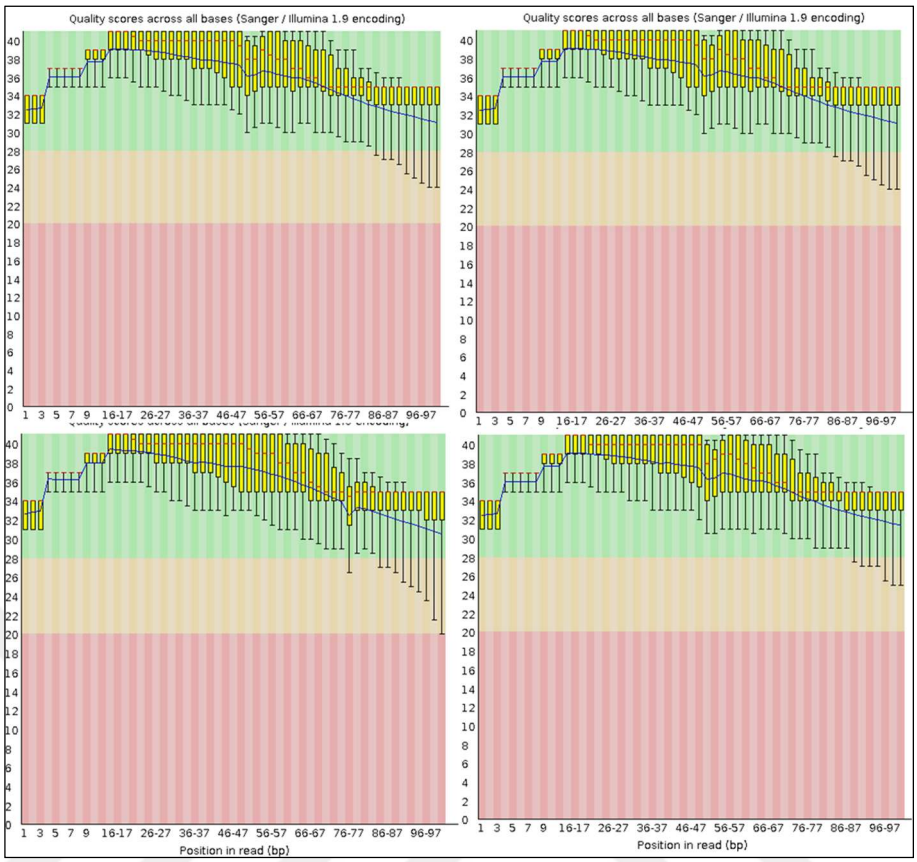

Figure 4.2. Quality control of reads in SRR1636825 and SRR1636827

4.1.2. Mapping Reads to Reference Genome

"IRGSP-1.0_genome.fasta" belongs to *Oryza sativa* subsp. *japonica*, downloaded from RAP-DB was used as a reference genome file and indexed. Sequencing reads were aligned to the reference genome via Bowtie2 [133]. Given that the data source comes from the plant-pathogen interaction, the "--local" parameter was used to avoid alignments obtained from similar sequence motifs of different sequence sources.

SAM files that were obtained from alignment were converted to BAM files and sorted for faster manipulation of files via SAMtools [136]. Furthermore, the depth of reads on genomic location was found through SAMtools. After sorted BAM files were visualized on Artemis genome browser [141], coverage plots of transcripts were plotted.

The total alignment rate of the reads belonging to 21 samples was found to be 88.95 percent, 79.87 percent, 73.68 percent, 39.39 percent, 91.13 percent, 77.49 percent, 76.53 percent,

45.40 percent, 75.21 percent, 76.62 percent, 86.52 percent, 59.10 percent, 64.81 percent, 76.77 percent, 62.70 percent, 74.48 percent, 71.22 percent, 79.68 percent, 40.25 percent, 75.33 percent, 77.57 percent, respectively. The average of alignments is 71.08 percent, this overall low alignment rate was achieved due to reads, sources from the nematode genome. The depth of mapped reads is shown on the coverage plot. The coverage plot of Os12t0623950-00 transcript that translates hypothetical protein in 194 bp length, which is located between the 26648135-26648328 genomic location on chromosome 12 is plotted (Figure 4.3.).

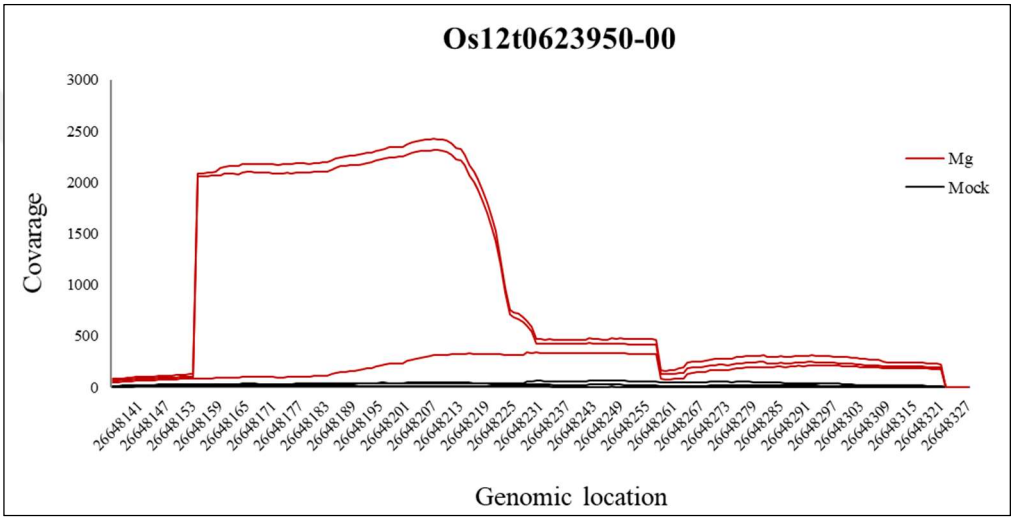

Figure 4.3. Coverage plot of Os12t0623950-00 transcript

According to the SRP049444 dataset, the average of alignments is 90 percent. Rice genome and 4834 known cereal pseudogene sequences were combined into a single fasta file to determine whether the RNA-seq reads belong to the known pseudogene transcript. Reads were then aligned to this combined genome. Two pseudogenes NR_165376.1 and NR_160899.1 were found as differentially expressed in the nematode dataset. NR_165376.1 that is *Oryza sativa* Japonica Group ribosomal protein eL42 pseudogene has decreased 1.26 fold in *Meloidogyne graminicola*-infected samples compared to mock. NR_160899.1 is *Zea mays* ADP-ribosylation factor pseudogene has increased 2.35 fold in *Meloidogyne graminicola*-infected samples compared to mock. The coverage plots of NR_165376.1 and NR_160899.1 pseudogene transcripts are plotted (Figure 4.4. and Figure 4.5.).

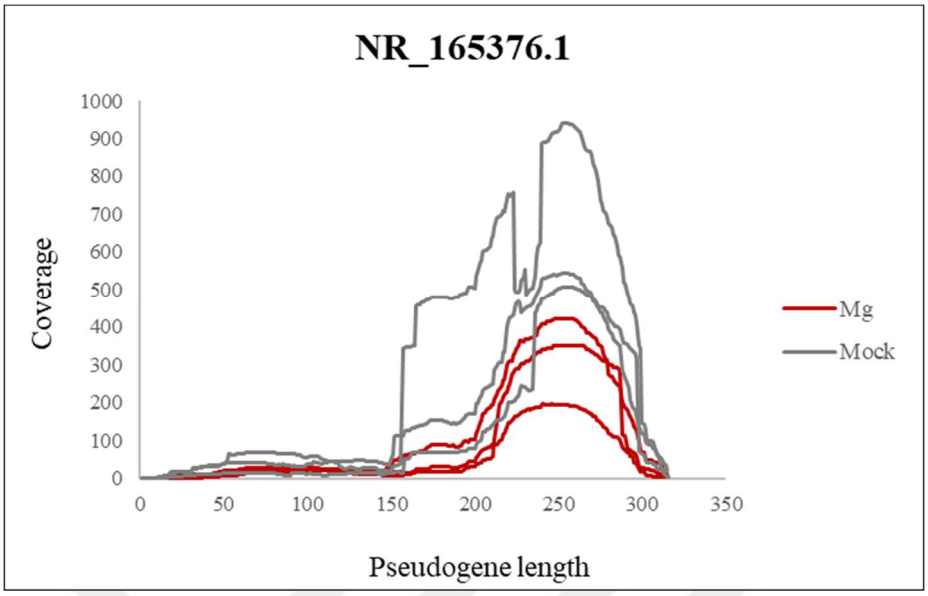

Figure 4.4. Coverage plot of NR_165376.1 pseudogene

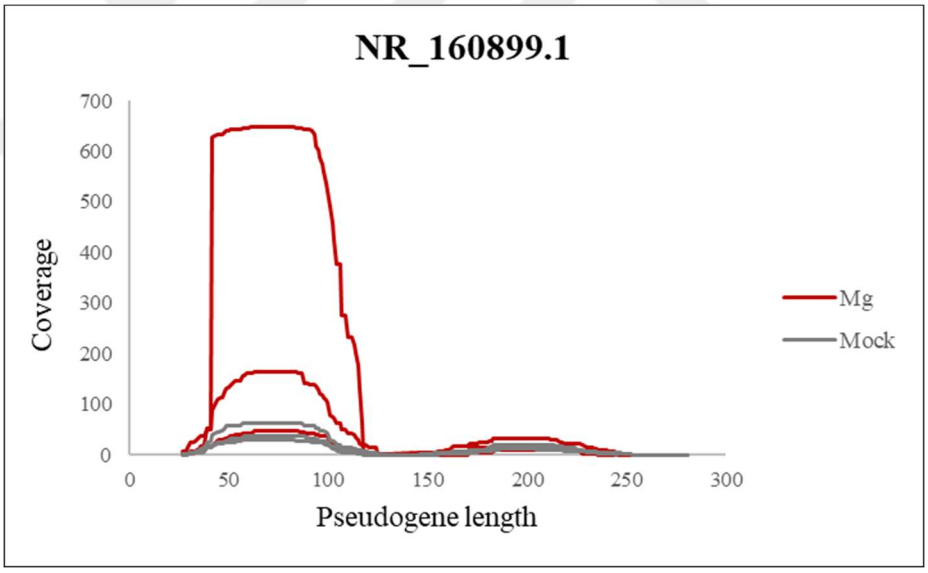

Figure 4.5. Coverage plot of NR_160899.1 pseudogene

One pseudogene NR_163866.1 was found as differentially expressed in the fungus dataset. NR_163866.1 is *Zea mays* phenylalanine/tyrosine ammonia-lyase pseudogene has increased 0.72 and 0.44 folds in *Magnaporthe oryzae* ZB13 and *Magnaporthe oryzae* ZH10-infected samples compared to mock. The coverage plot of NR_163866.1 pseudogene transcript is plotted (Figure 4.6).

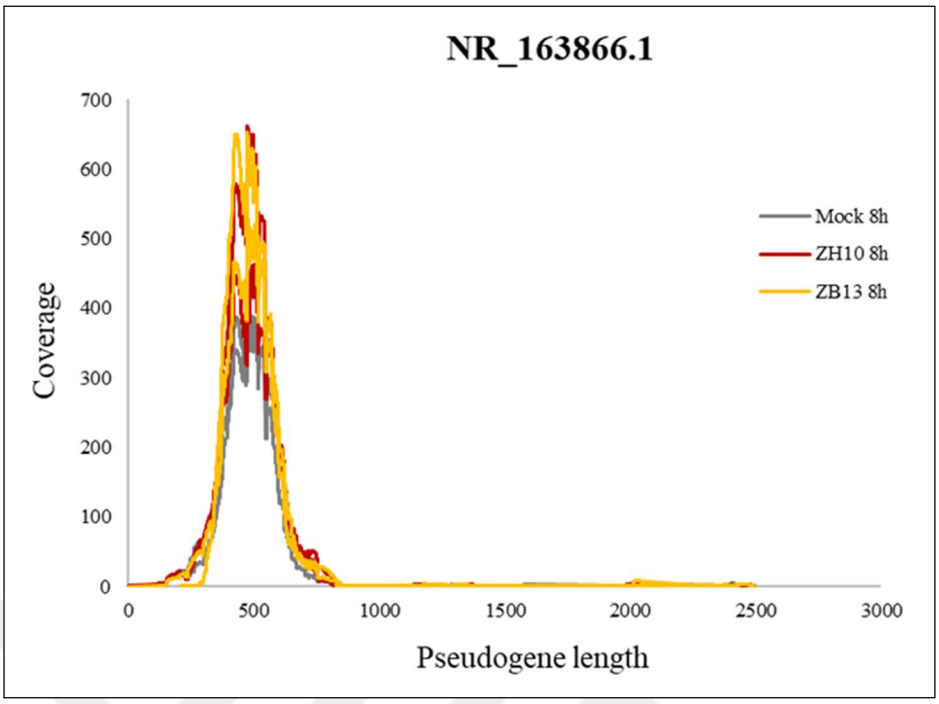

Figure 4.6. Coverage plot of NR_163866.1 pseudogene

4.1.3. Assessing Differential Expression

Differential expression analysis consists of normalization, statistical modeling, and gene expression test [162]. Transcripts were assembled using coverage of mapped reads in sorted BAM files via Cufflinks [134]. After the FPKM values of the transcripts were normalized in Cufflinks, transcript assemblies of replicate samples for each condition were combined with Cuffmerge. Then, assembled transcripts merged for each condition were compared to identify DEGs with statistical significance via Cuffdiff.

According to Cuffdiff results, FC, t-test, p-value, q-value were calculated for differential expression using the FPKM value of each gene. Student t-test, with the null hypothesis that there is no difference among groups, is used to determine significant differences between two groups. The cut-off for adjusted p-value is used as 0.05.

According to the significant tag in the "gene_exp.diff" output file of Cuffdiff, 367 transcripts with "yes" and 9856 transcripts with "no" out of the 10223, were determined. Out of 367 transcripts with "yes" 359 and 8 belong to rice and nematode. Within comparison of maximum FPKM values, *Meloidogyne graminicola*-infected samples, counted as 312342,

were detected more than mock samples, counted as 271145. The mean of FPKM values in *Meloidogyne graminicola*-infected samples ($\bar{x}$ = 3271.12) was higher than the mean of FPKM values in mock samples ($\bar{x}$ = 2910.89). Out of transcripts, 83 were not tested because they were calculated in small amounts. A screenshot of the DEG list in the "gene_exp.diff" file is given in Figure 4.7.

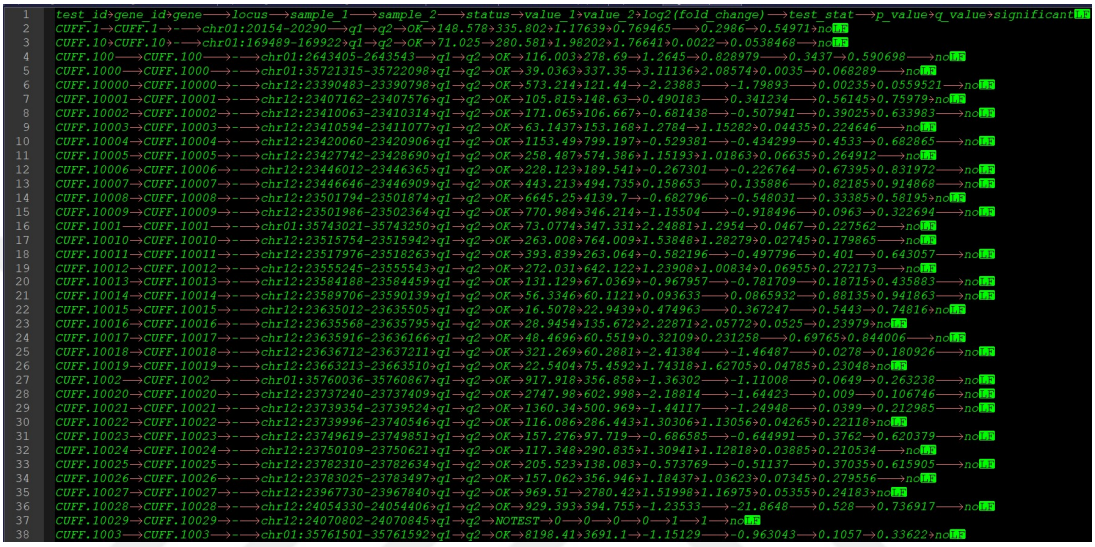

Figure 4.7. Screenshot of DEG list in SRP010960 dataset

Furthermore, the lowly expressed genes were also discarded before the assessment of DEGs. To assess the DEGs, volcano plots, a scatter plot that shows significantly expressed genes is used. This is drawn using –log10 of the q-value on the y-axis against the $\log_2$FC on the x-axis. Colored genes with green and red passing the thresholds for –log10 (q-value) and $\log_2$FC show the most upregulated genes are located in the upper right and the most downregulated genes in the upper left.

DEG analysis of the SRP010960 dataset was performed according to the comparison between mock (SRR408732/33/34) samples and nematode *Meloidogyne graminicola*-infected samples collected at 7 dpi (SRR408742/43/44). As a result of filtering, 362 out of 10223 transcripts had a q value less than 0.05 and an absolute value of $\log_2$FC greater than 2. The volcano plot of 362 DEGs in the SRP010960 dataset is drawn using RStudio 4.1.2 (Figure 4.8.).

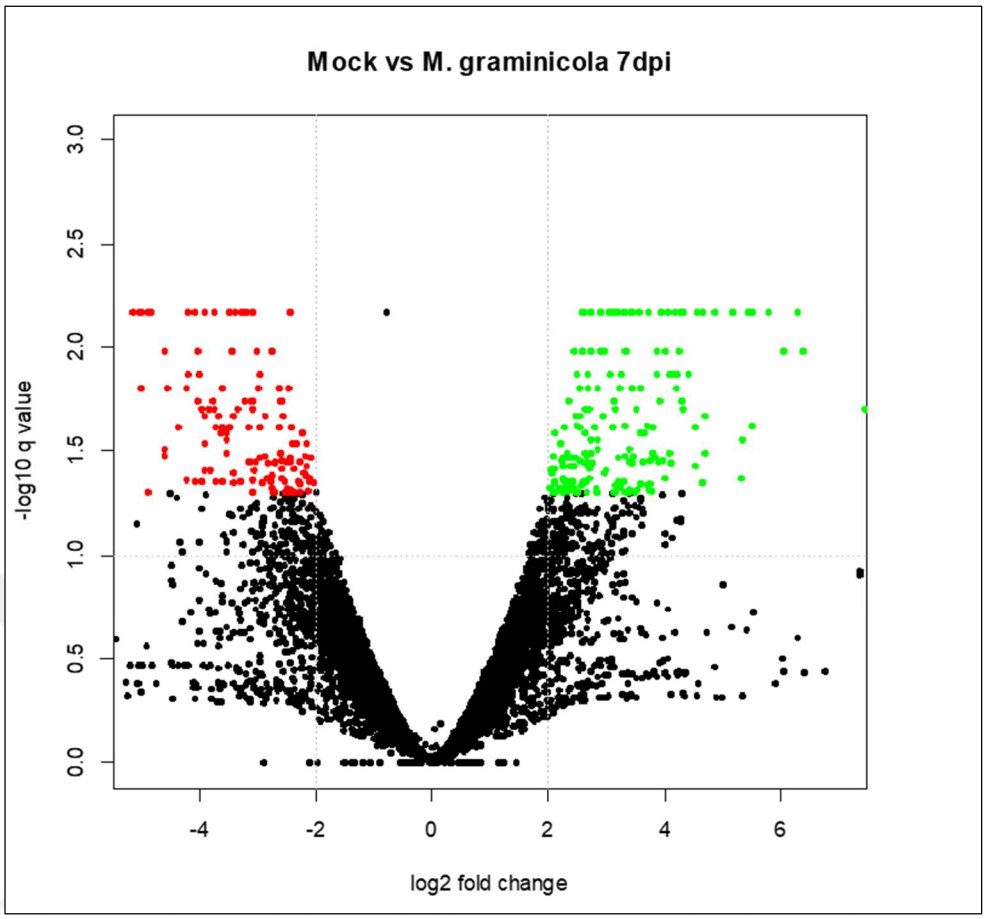

Figure 4.8. Volcano plot of 362 DEGs in SRP010960

DEG analyses of the SRP049444 dataset were performed according to comparison among mock (SRR1636825/26), fungus *Magnaporthe oryzae* ZB13-infected (SRR1636827/28), fungus *Magnaporthe oryzae* ZH10-infected (SRR163689/30) samples that were collected at 8 h. As a result of filtering, 1268 out of 34401 transcripts had a q value less than 0.05 and an absolute value of log_2FC greater than 2. The volcano plot of 1268 DEGs in the SRP049444 dataset is drawn using RStudio 4.1.2 (Figure 4.9.).

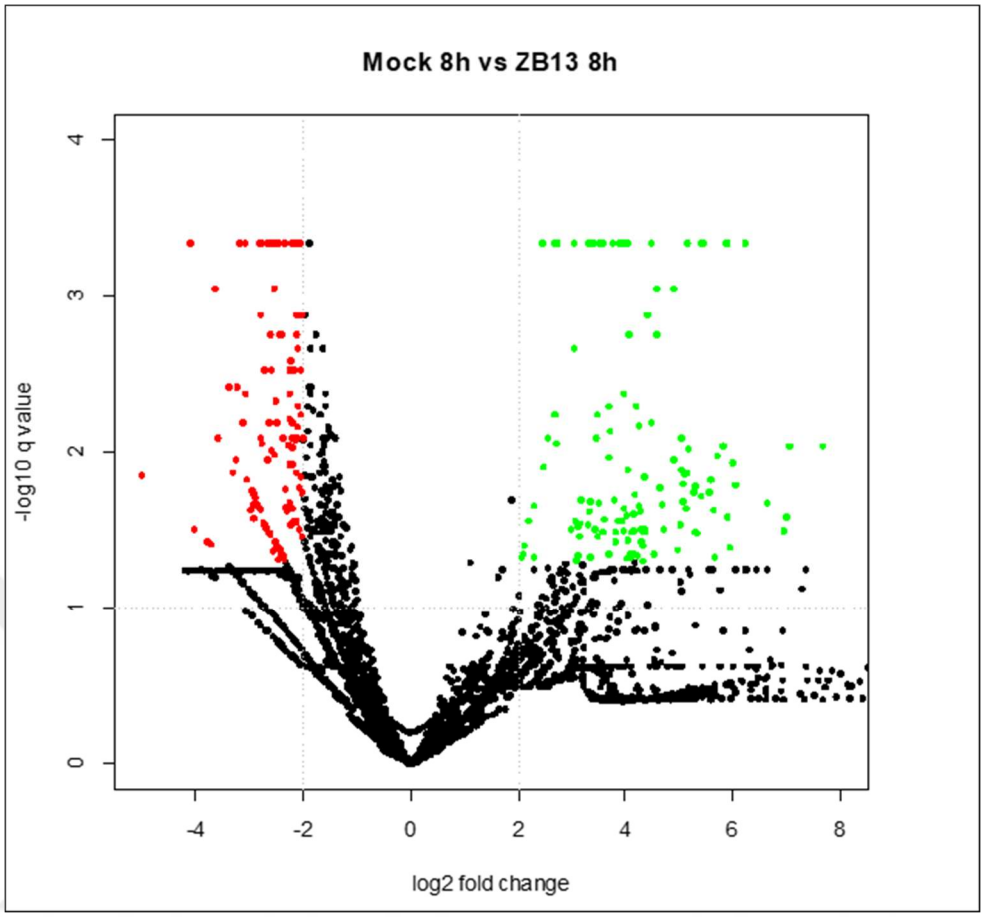

Figure 4.9. Volcano plot of 1268 DEGs in SRP049444

As a result of filtering, 236 out of 34726 transcripts had a q value less than 0.05 and an absolute value of log_2FC greater than 2. The volcano plot of 236 DEGs in the SRP049444 dataset is drawn using RStudio 4.1.2 (Figure 4.10.).

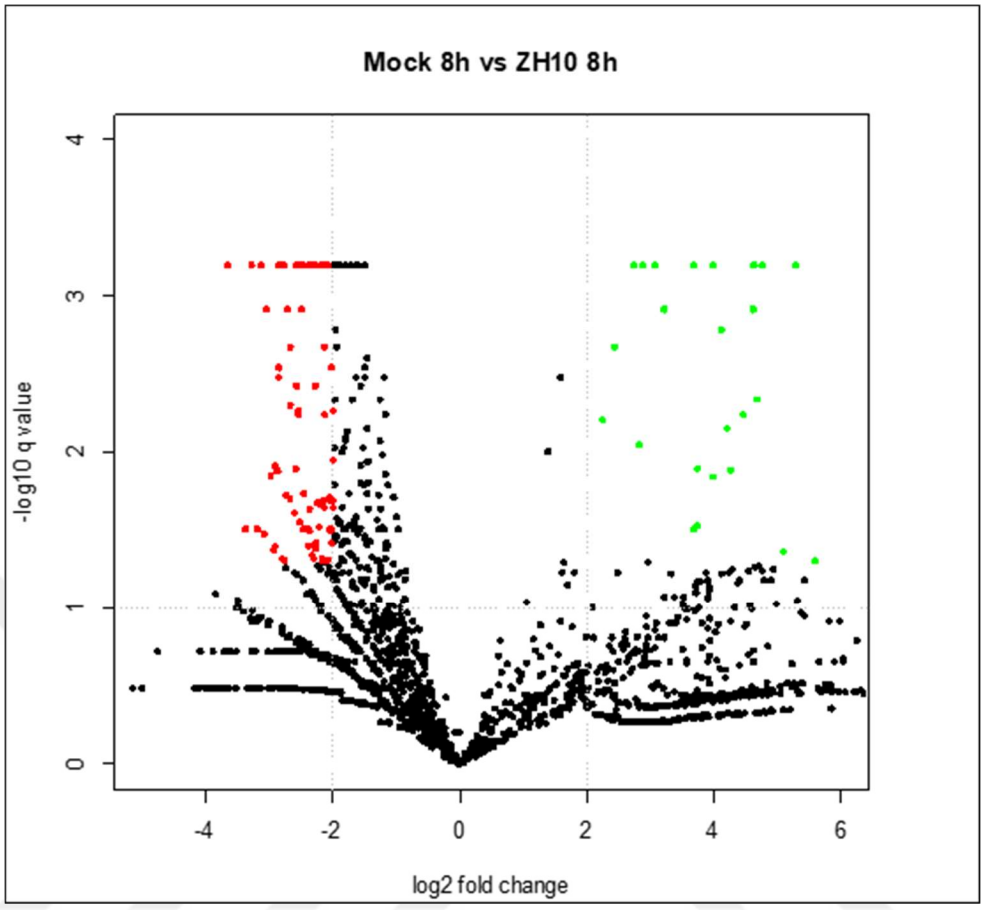

Figure 4.10. Volcano plot of 236 DEGs in SRP049444

4.1.4. Annotation of DEGs

The Cuffdiff tool provides a list of all transcripts with the corresponding FC and q-value combination. The next step is to filter DEGs and locate those in the genome annotation allowing identification and the genomic location of each DEG.

BED file consists of the genomic locations of DEGs, obtained from using SAMtools, which were matched to the locations in the "transcripts.gff" file downloaded from RAP-DB [153] via BEDTools intersect parameter [137].

According to the significant tag in the "gene_exp.diff" output file of Cuffdiff compared to mock and nematode *Meloidogyne graminicola*-infected samples, out of 367 transcripts, 231 were upregulated and 136 were downregulated. The top 20 DEGs related to nematodes stress are given in Table 4.1.

Table 4.1. List of top 20 DEGs, sorted based on q-value

Gene	Description	Up/Down
Os01g0253300	Importin alpha-1a subunit	Up
Os03g0712700, Os03g0712733	Cytosolic phosphoglucomutase, Hypothetical gene	Up
Os09g0279000, Os09g0279050	Similar to r-interacting factor1, Non-protein coding transcript	Up
Os08g0191100, Os08g0191150	Aconitase, Response to heat stress, Non-protein coding transcript	Up
Os03g0147400, Os03g0147550	Similar to cDNA clone J023091E03, full insert sequence, Hypothetical protein	Up
Os01g0908200	Member of the Bric-a-Brac/Tramtrack/Broad (BTB) family, BT1/BT2 ortholog, Negative regulation of nitrate uptake and nitrogen use efficiency	Up
Os08g0195400	Protein of unknown function DUF3778 domain containing protein	Up
Os08g0130600	BTB domain containing protein	Up
Os08g0130700	Conserved hypothetical protein	Up
Os07g0230400	Conserved hypothetical protein	Up
Os02g0220500	Elongation factor 1-gamma	Up
Os02g0489400	Similar to 40S ribosomal protein S8	Down
Os01g0282800, Os01g0282866	Similar to Tubulin beta-5 chain (Beta-5 tubulin), Hypothetical protein	Down
Os07g0568700	Polygalacturonase-inhibiting protein, Inhibitor of fungal polygalacturonase, Regulation of floral organ number	Down
Os02g0629800	Defensin-like protein, Antimicrobial peptide, Positive regulation of Cd accumulation in rice leaves, Mediation of Cd efflux from cytosol into extracellular spaces via chelation, Pathogen defense	Down
Os05g0486200	Protein of unknown function KRTCAP2 domain containing protein	Down
Os07g0645000	Allergen V5/Tpx-1 related family protein	Down
Os09g0482100	Heat shock protein 82, A member of heat shock protein 90 (Hsp90) chaperone family	Up
Os03g0565100	OST3/OST6 family protein	Down
Os03g0240800	Similar to AGL157Cp	Down

According to the significant tag in the "gene_exp.diff" output file of Cuffdiff compared to mock and fungus *Magnaporthe oryzae* ZB13-infected samples, out of 1385 transcripts, 1141 were upregulated and 244 were downregulated. The top 20 DEGs related to ZB13 group fungus stress are given in Table 4.2.

Table 4.2. List of top 20 DEGs related to ZB13 infection in SRP049444

Gene	Description	Up/Down
Os02g0478050	Similar to predicted protein	Up
Os05g0411600	Similar to nucleic acid binding protein	Down
Os06g0698711, Os06g0698748	Conserved hypothetical protein, Oxidative stress 3-like 7	Down
Os08g0119800, Os08g0119900	Photosystem II subunit, Hypothetical protein	Down
Os03g0118400, Os03g0118425	Cell division control protein 2 homolog 1, Hypothetical gene	Up
Os04g0566400, Os04g0566450	Indeterminate domain 10, Hypothetical gene	Down
Os05g0230700	Aux/IAA protein 17	Down
Os04g0617800	Imidazoleglycerol-phosphate dehydratase	Up
Os04g0687900	Tubby-like protein 3	Down
Os07g0663800	Short-chain alcohol dehydrogenase/reductase 110C-MI4	Down
Os07g0684100	Similar to Thioredoxin-like 1	Down
Os08g0482700	Uclacyanin-like protein 30	Down
Os11g0573200	Similar to Zinc knuckle family protein	Up
Os03g0411300	EF-Hand type domain containing protein	Down
Os05g0333200	Dwarf 1, Daikoku dwarf, rice G protein alpha subunit 1	Up
Os02g0598400, Os02g0598500	Lipovitellin, superhelical domain containing protein, Protein phosphatase 2C family protein	Up
Os10g0525000, Os10g0525101	Cytochrome P450 704A5, Hypothetical gene	Down
Os07g0155600	Ethylene-insensitive protein 2	Up
Os06g0270900	Aldehyde dehydrogenase 2B1	Down
Os10g0356000	Ribulose bisphosphate carboxylase large chain 1	Up

According to the significant tag in the "gene_exp.diff" output file of Cuffdiff compared to mock and fungus *Magnaporthe oryzae* ZH10-infected samples, out of 330 transcripts, 153 were upregulated and 177 were downregulated. The top 20 DEGs related to ZH10 group fungus stress are given in Table 4.3.

Table 4.3. List of top 20 DEGs related to ZH10 infection in SRP049444

Gene	Description	Up/Down
Os05g0151400	AIG1 domain containing protein, Similar to TOC159	Down
Os09g0452300	Similar to cDNA	Up
Os09g0481200, Os09g0481300	Photosystem I subunit, Similar to Rwp34	Down
Os08g0504700	Stress associated protein gene 11	Down
Os04g0437600, Os04g0437700	Similar to H0315A08.7 protein, Mitogen activated protein kinase kinase kinase 3 domain containing protein, Hypothetical protein	Down
Os02g0224300	HMW glutenin family protein, Hypothetical conserved gene	Down
Os09g0287000	Submergence 1B	Down
Os11g0704500	Metallothionein 1e	Down
Os01g0769000	Topoisomerase II-associated protein PAT1 domain containing protein, Similar to predicted protein	Down
Os08g0366000, Os08g0366050	Phosphoenolpyruvate carboxylase 2a, Hypothetical protein	Down
Os06g0130400	Substandard starch grain 6	Down
Os09g0425900	Senescence associated protein, Tetraspanin 13	Down
Os03g0390600, Os03g0390633	Non-protein coding transcript, Hypothetical gene	Down
Os07g0669650, Os07g0669675, Os07g0669700	Similar to isoform 2 of Potassium transporter 7, Hypothetical protein, High affinity potassium (K^+) transporter 7	Down
Os06g0199200	Smr protein/MutS2 C-terminal domain containing protein, Smr protein/MutS2 C-terminal domain containing protein, Similar to smr domain containing protein	Down
Os11g0465200, Os11g0465750	Similar to Bx2-like protein, Similar to Cytochrome P450 family protein, Hypothetical protein	Down
Os07g0687200, Os07g0687100	Calmodulin 1-2, Calmodulin B, Protein of unknown function DUF341 domain containing protein	Down
Os03g0616400	Ca^{2+} P-Type ATPase 3	Down
Os07g0173700	Ribosomal protein S18a	Up
Os03g0806600	Conserved hypothetical protein	Down

4.2. GO ANALYSIS

GO analysis categorizes biological characteristics of gene/gene products into three branches molecular function, biological process, and cellular component. So it explains how it functions in which cellular location, in which biological activity. Enrichment analysis of GO

terms reveals which GO terms are over-or under-presented for a set of genes that are significantly expressed under certain conditions.

4.2.1. GO Terms of DEGs

GO terms associated with nematode stress in rice were identified through g:Profiler. Gene products of DEGs were systematically categorized in molecular functions, including such as peroxidase activity, oxidoreductase activity, acting on peroxide as acceptor, antioxidant activity. GO analysis results showed that processes such as reactive oxygen species metabolic process, hydrogen peroxide catabolic process, hydrogen peroxide metabolic process, cellular oxidant detoxification, cellular detoxification, cellular response to toxic substance, response to oxidative stress, detoxification were significantly enriched in biological process. GO terms related to plasmodesma, cell junction, anchoring junction, cell-cell junction, symplast, cell wall, external encapsulating structure, cytosol, plant-type cell wall (Table 4.4.).

Table 4.4. GO annotation of DEGs in SRP010960 dataset

Category	GO ID	GO Term	q-value	Number of unigenes in subgroup	Number of DEGs
MF	GO:0004601	peroxidase activity	4.13E-08	179	17
	GO:0016684	oxidoreductase activity, acting on peroxide as acceptor	4.92E-08	181	17
	GO:0016209	antioxidant activity	3.98E-07	207	17
BP	GO:0072593	reactive oxygen species metabolic process	1.28E-08	169	18
	GO:0042744	hydrogen peroxide catabolic process	6.01E-08	140	16
	GO:0042743	hydrogen peroxide metabolic process	1.02E-07	145	16
	GO:0098869	cellular oxidant detoxification	2.0245E-06	203	17
	GO:1990748	cellular detoxification	4.835E-06	215	17
	GO:0097237	cellular response to toxic substance	4.835E-06	215	17
CC	GO:0009506	plasmodesma	1.72E-07	264	19
	GO:0030054	cell junction	1.72E-07	264	19
	GO:0070161	anchoring junction	1.72E-07	264	19
	GO:0005911	cell-cell junction	1.72E-07	264	19
	GO:0055044	symplast	1.72E-07	264	19
	GO:0005618	cell wall	9.6086E-06	338	19

GO terms associated with fungus ZB13 stress in rice were identified through g:Profiler. Numbers of GO terms were found as 48, 99, 53 in molecular function, biological process, cellular component. Gene products of DEGs were systematically categorized in molecular functions, consisting of binding such as heterocyclic compound, organic cyclic compound, RNA, ion, small molecule, nucleic acid, nucleoside phosphate, carbohydrate derivative and activity such as nucleoside-triphosphatase, hydrolase, ATP hydrolysis, pyrophosphatase, translation factor, acting on acid anhydrides, acting on acid anhydrides, in phosphorus-containing anhydrides. GO analysis results showed that process such as cellular metabolic, cellular amide metabolic, organonitrogen compound metabolic, amide biosynthetic, peptide metabolic, peptide biosynthetic, translation, organonitrogen compound biosynthetic, nitrogen compound metabolic, organic substance metabolic, generation of precursor metabolites and energy, primary metabolic, cellular nitrogen compound metabolic, photosynthesis, gene expression, cellular protein metabolic, protein-chromophore linkage were significantly enriched in the biological process. GO terms related to cellular components were intracellular anatomical structure, cytoplasm, intracellular organelle, intracellular membrane-bounded organelle, plastid, chloroplast, protein-containing complex, cellular component, cellular anatomical entity, photosystem, organelle envelope, thylakoid, photosynthetic membrane (Table 4.5.).

Table 4.5. GO annotation of DEGs related to ZB13 infection in SRP049444 dataset

Category	GO ID	GO Term	q-value	Number of unigenes in subgroup	Number of DEGs
MF	GO:0005488	binding	4.00E-21	12586	685
	GO:1901363	heterocyclic compound binding	4.28E-15	7488	433
	GO:0097159	organic cyclic compound binding	5.42E-15	7498	433
	GO:0003723	RNA binding	6.35E-12	1370	116
	GO:0043167	ion binding	1.21E-11	6244	360
	GO:0003674	molecular_function	1.86E-11	18744	889
	GO:0017111	nucleoside-triphosphatase activity	1.55E-10	530	60
	GO:0016462	pyrophosphatase activity	3.10E-09	599	62
	GO:0016817	hydrolase activity, acting on acid anhydrides	3.78E-09	617	63
	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	6.23E-09	609	62

MF	GO:0036094	small molecule binding	3.67E-08	3576	218
	GO:0003676	nucleic acid binding	7.50E-08	3688	222
	GO:0016887	ATP hydrolysis activity	1.25E-07	341	41
	GO:0008135	translation factor activity, RNA binding	3.93E-07	138	24
	GO:0043168	anion binding	3.99E-07	3487	209
	GO:0000166	nucleotide binding	7.52E-07	3368	202
	GO:1901265	nucleoside phosphate binding	7.52E-07	3368	202
	GO:0090079	translation regulator activity, nucleic acid binding	1.46E-06	147	24
	GO:0097367	carbohydrate derivative binding	1.58E-06	3088	187
	GO:0045182	translation regulator activity	1.68E-06	148	24
BP	GO:0009987	cellular process	1.52E-21	14377	764
	GO:0044237	cellular metabolic process	4.70E-21	10666	603
	GO:0043603	cellular amide metabolic process	1.16E-16	998	105
	GO:0008152	metabolic process	7.25E-16	11938	635
	GO:1901564	organonitrogen compound metabolic process	1.20E-15	5673	352
	GO:0043604	amide biosynthetic process	2.82E-15	832	91
	GO:0006518	peptide metabolic process	4.98E-15	884	94
	GO:0008150	biological_process	9.12E-15	16739	829
	GO:0043043	peptide biosynthetic process	1.82E-14	768	85
	GO:0006412	translation	2.06E-14	755	84
	GO:1901566	organonitrogen compound biosynthetic process	5.75E-14	1551	133
	GO:0015979	photosynthesis	2.09E-13	203	39
	GO:0006807	nitrogen compound metabolic process	5.17E-13	8999	493
	GO:0071704	organic substance metabolic process	5.74E-13	11088	584
	GO:0006091	generation of precursor metabolites and energy	8.72E-13	437	58
	GO:0044238	primary metabolic process	1.93E-11	10418	547
	GO:0034641	cellular nitrogen compound metabolic process	4.48E-11	5025	302
	GO:0010467	gene expression	5.51E-11	3772	241
	GO:0044267	cellular protein metabolic process	1.29E-10	4253	263
	GO:0018298	protein-chromophore linkage	2.18E-10	32	15
CC	GO:0005622	intracellular anatomical structure	3.38E-36	11287	684
	GO:0005737	cytoplasm	4.31E-33	7334	491
	GO:0043229	intracellular organelle	5.07E-30	9731	593
	GO:0043226	organelle	7.69E-30	9745	593
	GO:0043231	intracellular membrane-bounded organelle	1.04E-27	9144	558

CC	GO:0043227	membrane-bounded organelle	1.17E-27	9170	559
	GO:0009536	plastid	2.23E-23	1474	149
	GO:0009507	chloroplast	5.31E-23	1409	144
	GO:0032991	protein-containing complex	9.76E-23	2884	231
	GO:0005575	cellular_component	1.40E-19	17576	883
	GO:0110165	cellular anatomical entity	1.38E-17	17411	867
	GO:0009522	photosystem I	2.81E-13	29	16
	GO:0009521	photosystem	3.29E-13	85	25
	GO:0031967	organelle envelope	7.64E-13	885	88
	GO:0031975	envelope	7.64E-13	885	88
	GO:0009579	thylakoid	7.43E-12	355	49
	GO:0031976	plastid thylakoid	1.19E-11	273	42
	GO:0009534	chloroplast thylakoid	1.19E-11	273	42
	GO:0042651	thylakoid membrane	6.01E-11	262	40
	GO:0034357	photosynthetic membrane	1.78E-10	283	41

GO terms associated with fungus ZH10 stress in rice were identified through g:Profiler. Numbers of GO terms were found as 8, 26, 38 in molecular function, biological process, cellular component. Gene products of DEGs were systematically categorized in molecular functions, consisting binding such as chlorophyll, heterocyclic compound, organic cyclic compound, protein domain specific and activity such as nucleoside-triphosphatase, carbon-carbon lyase, pyrophosphatase, glutamate-cysteine ligase. GO analysis results showed that photosynthesis, photosynthesis, light harvesting, photosynthesis, light harvesting in photosystem I, protein-chromophore linkage, response to abiotic stimulus, generation of precursor metabolites and energy, photosynthesis, light reaction, carbon fixation, ethylene-activated signaling pathway, cellular response to ethylene stimulus, peptide biosynthetic process were significantly enriched in the biological process. GO terms related to cellular components were photosystem I, photosystem, chloroplast, plastid, cytoplasm, thylakoid membrane, photosystem II, photosynthetic membrane, plastid thylakoid membrane (Table 4.6.).

Table 4.6. GO annotation of DEGs related to ZH10 infection in SRP049444 dataset

Category	GO ID	GO Term	q-value	Number of unigenes in subgroup	Number of DEGs
MF	GO:0016168	chlorophyll binding	6.29E-11	19	9
	GO:1901363	heterocyclic compound binding	0.000125	7488	123
	GO:0097159	organic cyclic compound binding	0.000135	7498	123
	GO:0019904	protein domain specific binding	0.00046	18	5
	GO:0017111	nucleoside-triphosphatase activity	0.031072	530	17
	GO:0016830	carbon-carbon lyase activity	0.043398	133	8
	GO:0016462	pyrophosphatase activity	0.043738	599	18
	GO:0004357	glutamate-cysteine ligase activity	0.049883	2	2
BP	GO:0015979	photosynthesis	2.11E-16	203	25
	GO:0009765	photosynthesis, light harvesting	1.36E-12	27	11
	GO:0009768	photosynthesis, light harvesting in photosystem I	5.22E-12	15	9
	GO:0018298	protein-chromophore linkage	1.28E-11	32	11
	GO:0009628	response to abiotic stimulus	8.35E-09	798	34
	GO:0006091	generation of precursor metabolites and energy	1.05E-08	437	25
	GO:0019684	photosynthesis, light reaction	1.58E-08	93	13
	GO:0015977	carbon fixation	0.000181	26	6
	GO:0009873	ethylene-activated signaling pathway	0.000495	69	8
	GO:0071369	cellular response to ethylene stimulus	0.000619	71	8
	GO:0043043	peptide biosynthetic process	0.00074	768	25
CC	GO:0009522	photosystem I	2.11E-18	29	14
	GO:0009521	photosystem	2.25E-16	85	18
	GO:0009507	chloroplast	6.25E-12	1409	51
	GO:0009536	plastid	9.36E-12	1474	52
	GO:0005737	cytoplasm	1.66E-11	7334	142
	GO:0042651	thylakoid membrane	2.15E-11	262	22
	GO:0009523	photosystem II	8.27E-11	70	13
	GO:0034357	photosynthetic membrane	1.02E-10	283	22
	GO:0055035	plastid thylakoid membrane	1.24E-10	228	20

4.3. KEGG PATHWAY ANALYSIS

KEGG pathway analysis provides to understanding functional annotation of DEGs at the molecular level. It reveals which genes can interact and which pathways are most impacted under certain conditions.

4.3.1. KEGG Pathways of DEGs

KEGG pathway analysis results revealed that DEGs were enriched in phenylpropanoid biosynthesis, biosynthesis of amino acids, mRNA surveillance pathway, ubiquitin mediated proteolysis (Table 4.7.).

Table 4.7. KEGG pathways of DEGs related to nematode stress in SRP010960 dataset

KEGG ID	KEGG Pathway	q-value	Number of unigenes in subgroup	Number of DEGs
KEGG:00940	Phenylpropanoid biosynthesis	0.063031	136	10
KEGG:01230	Biosynthesis of amino acids	0.155091	179	11
KEGG:03015	mRNA surveillance pathway	0.663225	87	6
KEGG:04120	Ubiquitin mediated proteolysis	0.697523	88	6

KEGG pathway analysis results revealed that DEGs were significantly enriched in KEGG root term, metabolic pathways, photosynthesis, carbon metabolism, photosynthesis - antenna proteins, biosynthesis of secondary metabolites, protein processing in endoplasmic reticulum, RNA degradation, carbon fixation in photosynthetic organisms (Table 4.8.).

Table 4.8. KEGG pathways of DEGs related to fungus ZB13 stress in SRP049444 dataset

KEGG ID	KEGG Pathway	q-value	Number of unigenes in subgroup	Number of DEGs
KEGG:00000	KEGG root term	4.43E-38	3442	300
KEGG:01100	Metabolic pathways	7.99E-18	1693	148
KEGG:00195	Photosynthesis	8.94E-10	41	15
KEGG:01200	Carbon metabolism	1.48E-09	216	33
KEGG:00196	Photosynthesis - antenna proteins	3.98E-08	15	9

KEGG pathway analysis results revealed that DEGs were significantly enriched in KEGG root term, photosynthesis - antenna proteins, metabolic pathways, photosynthesis, carbon metabolism (Table 4.9.).

Table 4.9. KEGG pathways of DEGs related to fungus ZH10 stress in SRP049444 dataset

KEGG ID	KEGG Pathway	q-value	Number of unigenes in subgroup	Number of DEGs
KEGG:00000	KEGG root term	5.53E-15	3442	92
KEGG:00196	Photosynthesis - antenna proteins	2.50E-13	15	9
KEGG:01100	Metabolic pathways	3.63E-10	1693	52
KEGG:00195	Photosynthesis	1.37E-08	41	9
KEGG:01200	Carbon metabolism	2.1965E-05	216	13

4.4. VARIANT CALLING

Variant calling was performed to detect mutations as one of the markers on pseudogenes. Firstly, genotype likelihoods were generated on genomic position using coverages of nucleotides, with mpileup parameter via SAMtools. SNPs were identified from multisample files using the mpileup2snp parameter on VarScan. VCF files containing SNPs were converted to BED files for multiple comparisons. To obtain SNP annotation, intersect parameter was used on BEDTools.

Total 81 SNVs were observed in three comparisons among mock and infected samples (Figure 4.11.). Out of 81 SNVs, 34 were annotated. SNV in 34912882nd position in Os02g0814700 gene that translates to ribosomal protein 117 was randomly selected from annotated SNV file. Multiple sequence alignment was done among reference genome sequence and consensus sequences belonging to each sample for certain Os02g0814700 gene on CLUSTAL. When the genomes of mock and infected samples were compared with the reference genome, it was determined that Cytosine (C) nucleotide in the reference genome was converted to Adenine (A) nucleotide (Figure 4.12.).

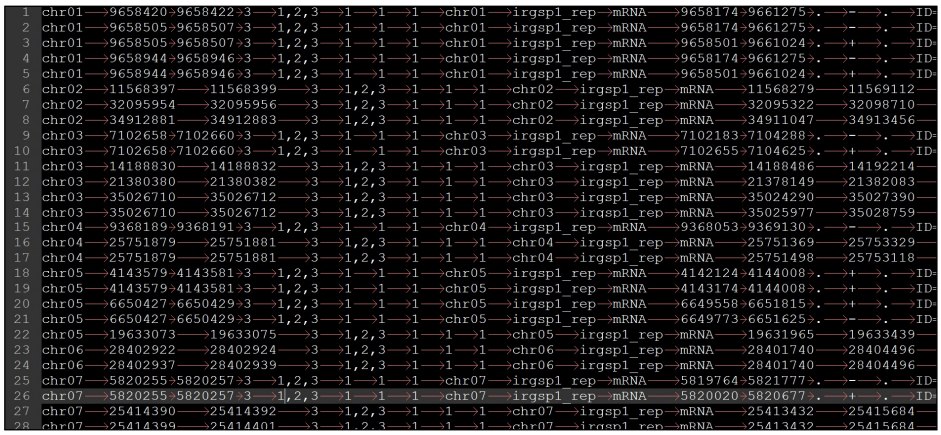

Figure 4.11. Comparison of sample files containing SNV

Figure 4.12. SNV identification in Os02g0814700 gene

Another example, SNVs were found in 5771144th and 5771835th positions in the Os01g0205500 gene that translates to non-protein coding transcript (Figure 4.13.).

Figure 4.13. SNV identification in Os01g0205500 gene

4.5. GENE MODEL

Consensus sequences were obtained from sorted BAM files for each sample using mpileup, call, filter, and consensus parameters on BCFtools, respectively. Firstly, actual calls were made from genotype likelihoods with mpileup and call parameters, respectively. Then SNPs in VCF files obtained from variant calls were filtered with a quality score over 20 and read depth lower than 10. Consensus sequences were separately obtained from VCF files that belong to each sample. Then sequences intervals defined with DEGs' location in a BED file were extracted from the consensus sequence file previously obtained for each sample, with the getfasta parameter on BEDTools.

Gene models were constructed from consensus sequence files obtained in the previous step, with "partial" parameter on AUGUSTUS. Predicted parts containing gene, transcript, tss, exon, CDS, intron, start codon, stop codon, tts were obtained on AUGUSTUS. After intron locations were filtered, both depth and breadth of mapped reads for each sample in sorted BAM file on the features in BED file containing intron locations were computed, with coverage parameter via BEDTools. According to intron location, coverages of introns that predicted from gene models were plotted.

Longer than 200 nt expressed sequences were determined as five sequences in mock and eleven sequences in infected samples. The highest mean of coverage with 125 was found in 449 nt length expressed sequence in the infected sample (Figure 4.14.). This gene (Os03g0158500) translates YT521-B-like protein family protein, hypothetical gene.

Os06g0654900 gene translates similar to constans-like protein CO6 was found in the intronic region in fungus *Magnaporthe oryzae* ZB13-infected samples. Os12g0292301 hypothetical gene was found in fungus *Magnaporthe oryzae* ZH10-infected samples. (Figure 4.15. and Figure 4.16.). Interestingly, these expressed sequences are derived from predicted intron regions on AUGUSTUS.

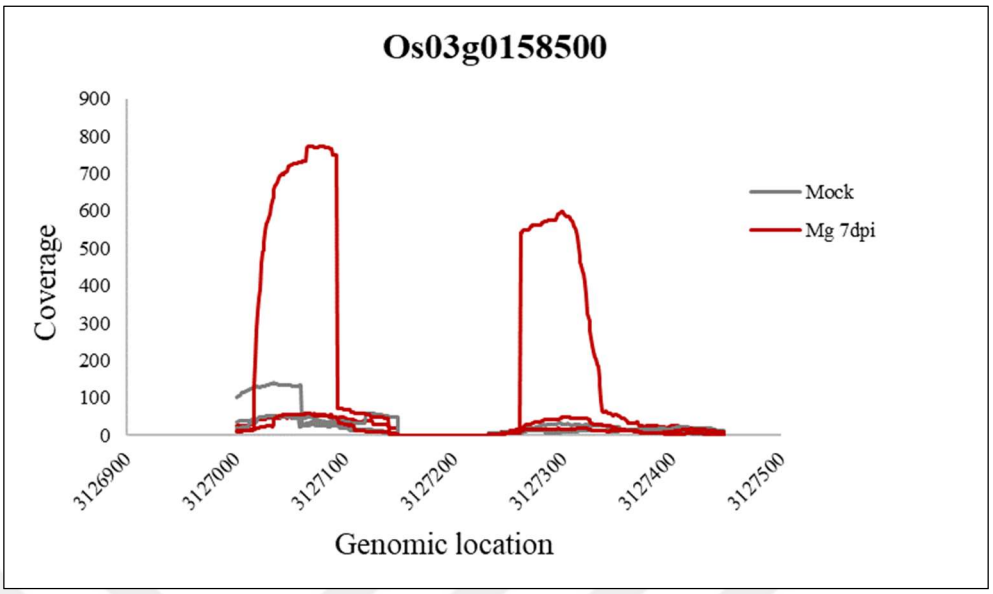

Figure 4.14. Intronic read coverage for Os03g0158500

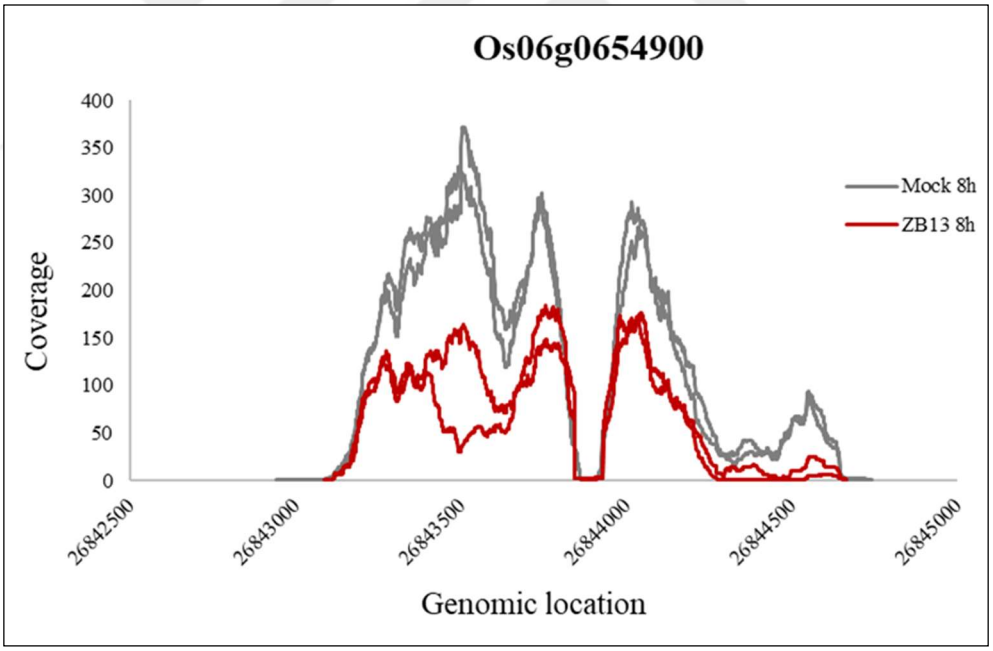

Figure 4.15. Intronic read coverage for Os06g0654900

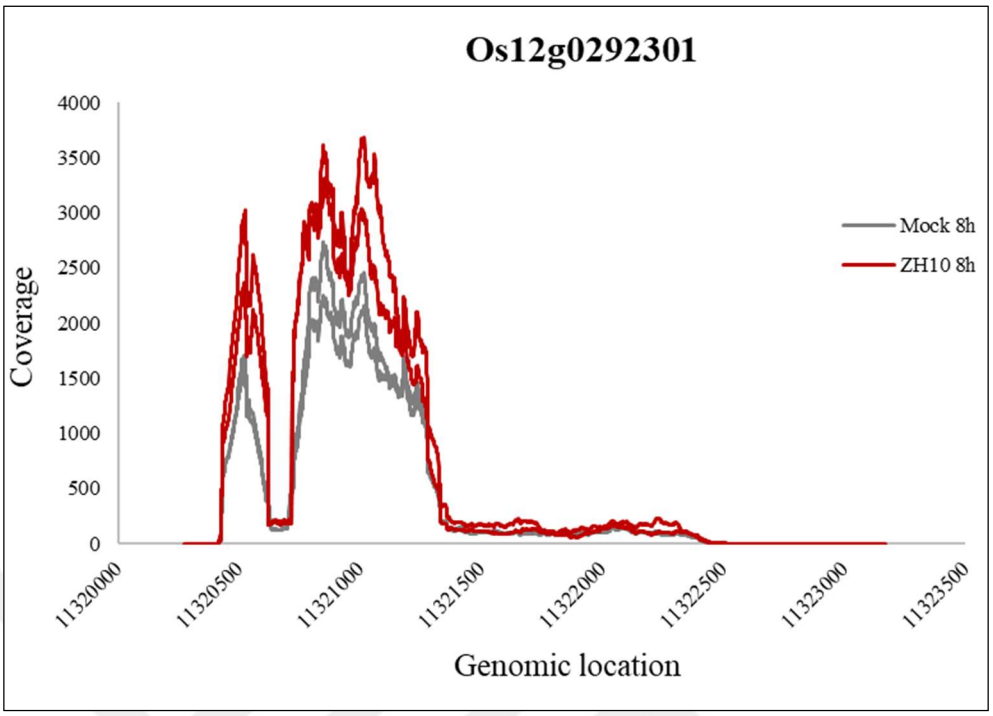

Figure 4.16. Intronic read coverage for Os12g0292301

4.6. KA/KS ANALYSIS

The Ka/Ks ratio was calculated between the consensus sequences obtained from DEGs and the corresponding sequences in the reference genome. The 2410 nt-length Os02g0814700 gene tends to have a positive selection in its first 500 nt, 1000-1500 nt, and 1800-2200 nt-length regions. Ka/Ks ratios of the Os02g0814700 gene that translates to ribosomal protein 117 are shown in Figure 4.17.

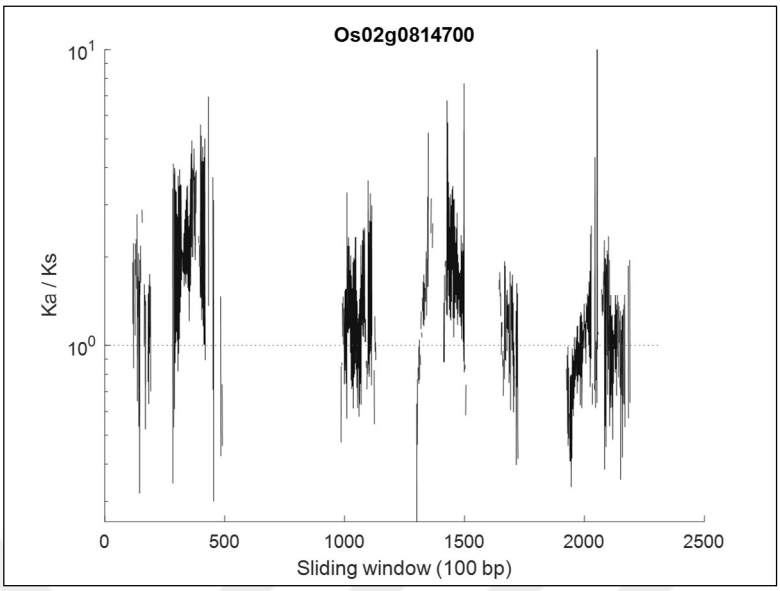

Figure 4.17. Sliding window for Ka/Ks ratio of Os02g0814700 gene

Pseudogenes and genes belonging to Os11gRGAC in rice and EFL1 in humans were downloaded NCBI for comparison Ka/Ks ratio. While there is a significant relationship between Ka/Ks ratio and protein alignment score in human pseudogenes, it was not detected in rice (Figure 4.18. and Figure 4.19.). In human pseudogenes, as the protein alignment score increases, the value of the Ka/Ks ratio increases. These findings will contribute to the overall pseudogene prediction from sequence alone.

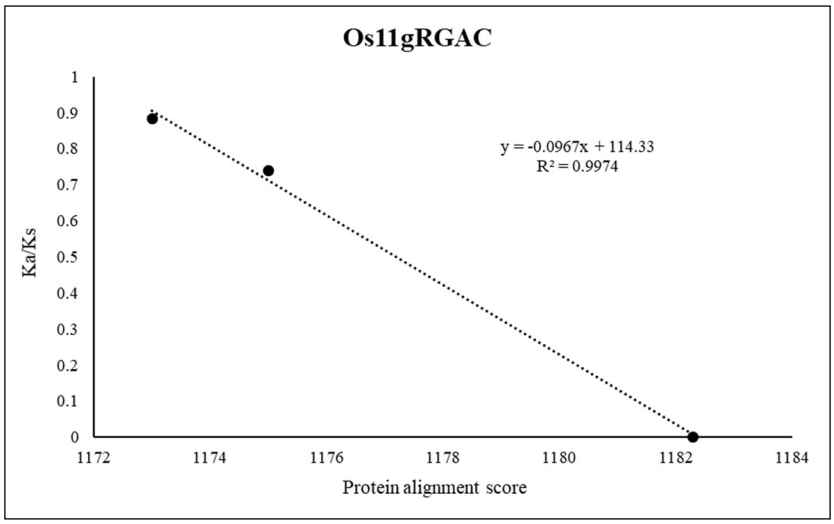

Figure 4.18. Ka/Ks ratios of Os11gRGAC gene/pseudogene on rice

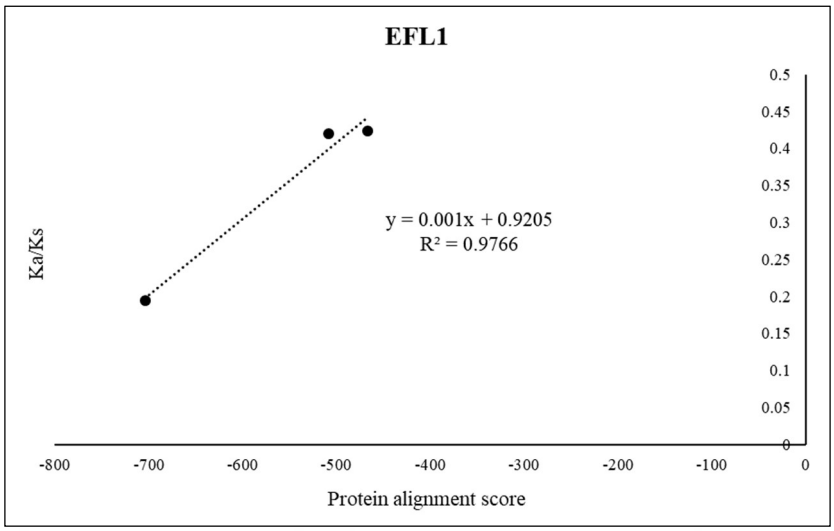

Figure 4.19. Ka/Ks ratios of EFL1 gene/pseudogene on human

4.7. DISCOVERY OF PSEUDOGENES USING ML METHODS

4.7.1. MLP Analysis

4.7.1.1. *Characteristics of the Training Data*

Pseudogenes and corresponding genes belonging to cereal genera such as Triticum, Hordeum, Zea, Oryza, Avena, Sorghum, Brachypodium, etc. were downloaded on NCBI to make training and test data. Sequences shorter than 3000 and longer than 200 of 5107 genes and 4477 pseudogenes were filtered out. While there are 5684 genes and 3777 pseudogenes in the training data, there are 800 genes and 700 pseudogenes in the test data. The test data constitute approximately 15 percent of the training data.

The mean lengths of pseudogenes and genes are 1036.3 nt and 1556.1 nt in the training data. Lengths of known pseudogenes and genes in cereal downloaded for training data are shown below Figure 4.20.

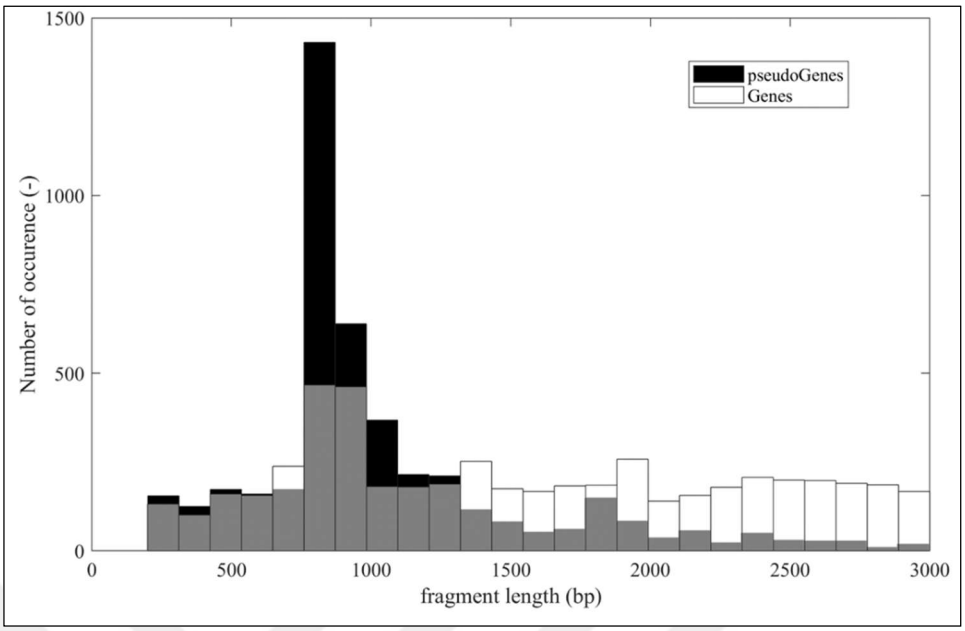

Figure 4.20. Length distribution of known pseudogenes and genes in the training/test data

4.7.1.2. *Characteristics of the Test Data*

The test data occurred consensus sequences obtained from DEGs were used. Consensus sequences were obtained from sorted BAM files for each sample using mpileup, call, filter, and consensus parameters on BCFtools. After consensus sequences were obtained, sequences longer than 200 and shorter than 3000 were filtered out. Then first 200 nucleotides of the sequences were taken, they were converted to a single-line fasta file.

MLP was implemented to predict pseudogenes using the WEKA tool. Parameters, number of hidden layers, number of hidden neurons are important factors that ensure correct results from the model.

Different sequence lengths were obtained to understand which part of the sequence will give the most accurate prediction in classification (Table 4.10.). Therefore, MLP was performed using the sequences' first 10, 50, 100, 500, 1000, 1500 nucleotides. As the lengths of sequences increase, the classification accuracy increases. However, it was observed that the ROC area decreased when 1500 nt was used. Between 100 nt and 500 nt can be selected as optimum length.

Table 4.10. Comparison of selected nucleotide length as an attribute

Nucleotide length	Accuracy	F-Measure	ROC Area
10	90.25	0.91	0.94
50	91.32	0.92	0.96
100	91.55	0.92	0.94
500	91.09	0.91	0.95
1000	93.30	0.93	0.95
1500	93.83	0.94	0.88

Two different classification algorithms, MLP and sequential minimal optimization (SMO) were compared (Table 4.11.). 50 nt and 350 nt were get from the middle of the sequences. The accuracy rate of the MLP analysis was found to be higher than SMO.

Table 4.11. Comparison of classification methods in WEKA

Classification method	Nucleotide length	Accuracy	F-Measure	ROC Area	Class
MLP	50	88.58	0.89	0.94	Pseudogene
			0.88	0.94	Gene
SMO	50	81.72	0.82	0.82	Pseudogene
			0.82	0.82	Gene
MLP	350	92.58	0.93	0.95	Pseudogene
			0.93	0.95	Gene
SMO	350	91.12	0.92	0.91	Pseudogene
			0.91	0.91	Gene

MLP neural network used in this study has an input layer with 205 inputs, 1 hidden layer with 104 hidden neurons, and an output layer with 2 neurons (Figure 4.21. and Figure 4.22.).

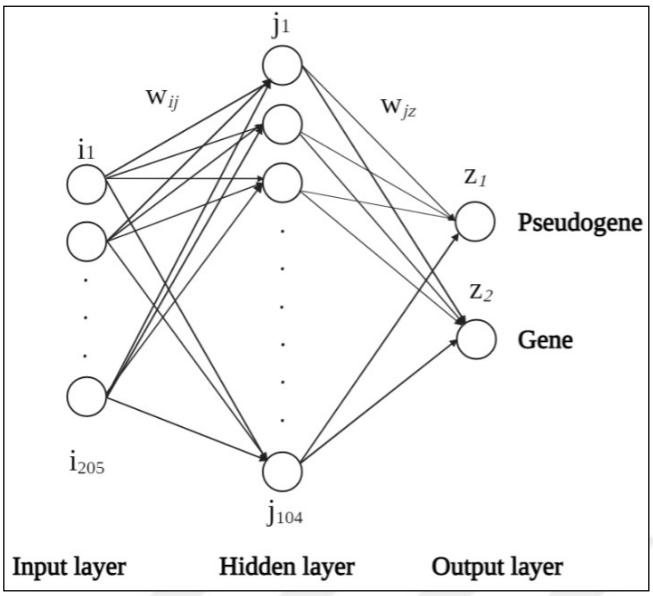

Figure 4.21. Structure of MLP neural network

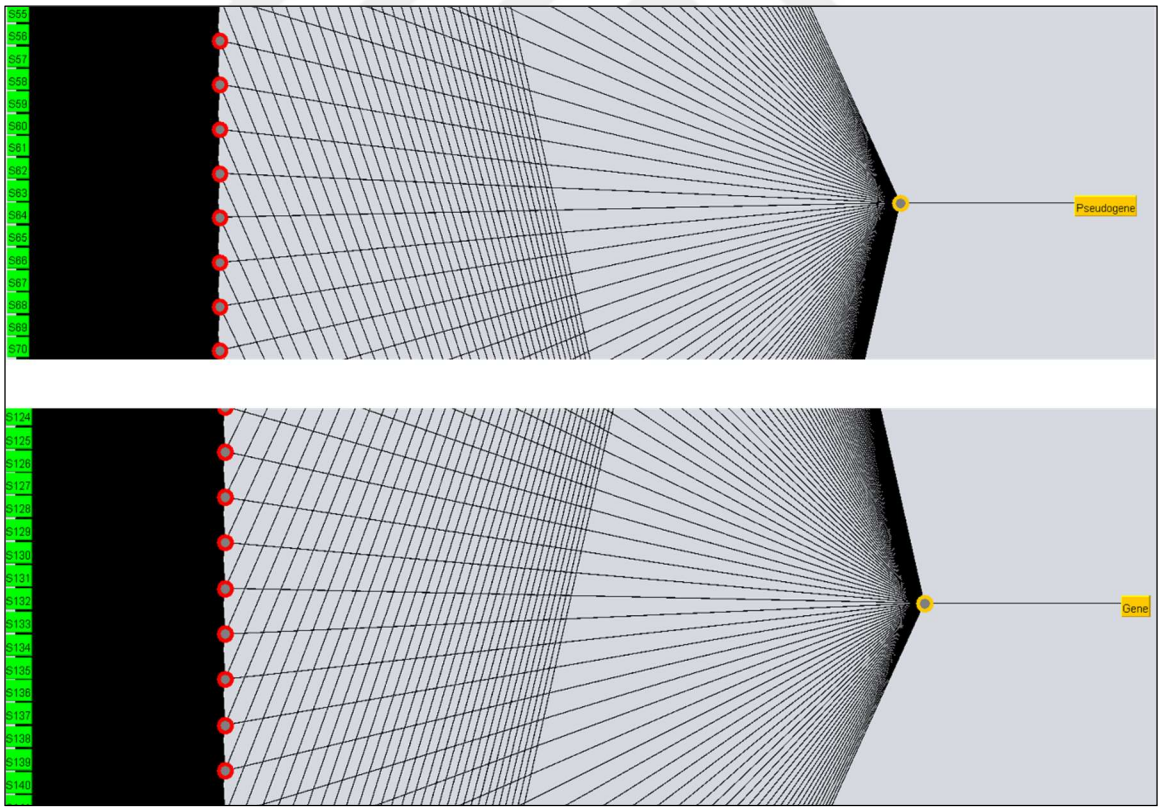

Figure 4.22. MLP neural network with 205 input attributes

4.7.1.3. *Evaluation of Model*

To evaluate the performance of the model, performance metrics obtained from the confusion matrix and area under the curve (AUC) obtained from the ROC curve were used. The confusion matrix describes the complete performance of the models. Training data that contains 7124 sequences were used to develop a more accurate model with various hidden neuron numbers such as 25, 50, 100, 200 in a hidden layer for MLP in WEKA.

According to the confusion matrix in the first model, 3188 pseudogenes and 2785 genes were correctly classified as TP and TN, respectively. 538 genes were classified as pseudogenes, this is FP, type 1 error. 613 pseudogenes were classified as genes, this is FN, type 2 error. Confusion matrix in the second model, 3197 pseudogenes and 2830 genes were correctly classified as TP and TN, respectively. 493 genes were classified as pseudogenes, this is FP, type 1 error. 604 pseudogenes were classified as genes, this is FN, type 2 error. Confusion matrix in the third model, 3210 pseudogenes and 2800 genes were correctly classified as TP and TN, respectively. 523 genes were classified as pseudogenes, this is FP, type 1 error. 591 pseudogenes were classified as genes, this is FN, type 2 error. Confusion matrix in the last model, 3207 pseudogenes and 2853 genes were correctly classified as TP and TN, respectively. 470 genes were classified as pseudogenes, this is FP, type 1 error. 594 pseudogenes were classified as genes, this is FN, type 2 error.

Confusion matrices were compared for MLP models with four different numbers of hidden neurons in Table 4.12.

Table 4.12. Confusion matrix comparison for MLP models

	Prediction					
Actual	I	**Pseudogene**	**Gene**	II	**Pseudogene**	**Gene**
	Pseudogene	3188	613	**Pseudogene**	3197	604
	Gene	538	2785	**Gene**	493	2830
	III	**Pseudogene**	**Gene**	IV	**Pseudogene**	**Gene**
	Pseudogene	3210	591	**Pseudogene**	3207	594
	Gene	523	2800	**Gene**	470	2853

The ROC curve is a performance metric for binary classification algorithms. The AUC is used to measure the prediction accuracy of the classifier. A high AUC above 90 percent indicates the high discrimination ability of the model.

The highest AUC value (0.9214) was observed in a model that performed with 200 hidden neurons. However, the AUC values of all models have over 90 percent. Therefore, the number of hidden neurons was chosen as the default value for further analysis. Prediction accuracy of MLP classifier is indicated with ROC curve, as given in Figure 4.23.

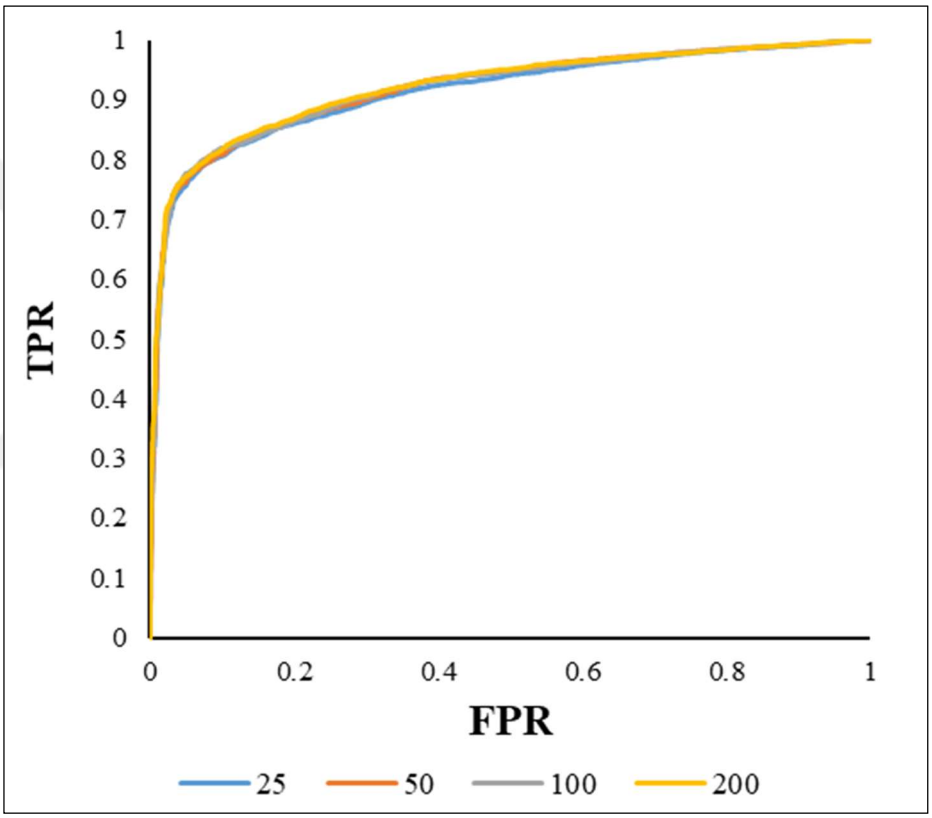

Figure 4.23. ROC curve comparison of MLP classifier

Training and test data were performed for MLP in Weka. AUC values were found as 0.87 and 0.90 in training and test data. AUC values between 0.80 and 0.90 have been accepted as excellent. ROC curves for the training and test data are shown in Figure 4.24 and Figure 4.25.

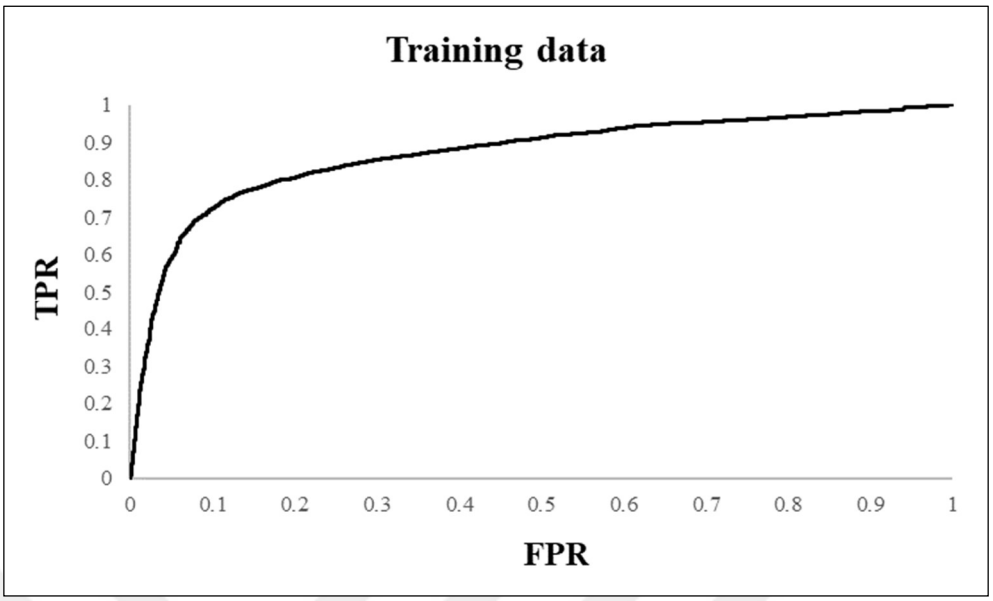

Figure 4.24. ROC curve of training data

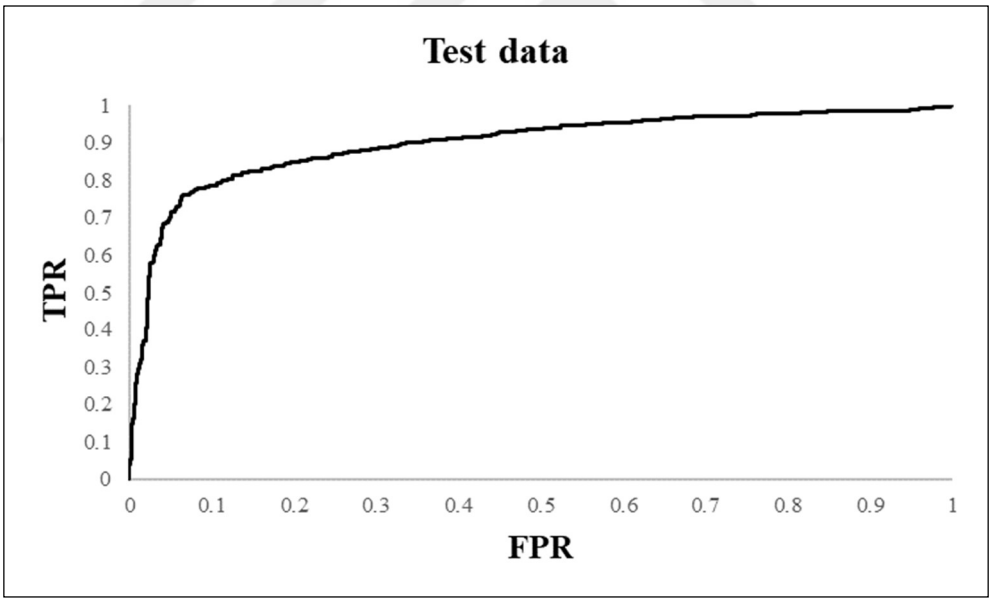

Figure 4.25. ROC curve of test data

Predicted pseudogenes were obtained from the model were annotated with BLASTN and TBLASTN. While pseudogene could not be found with BLASTN, it was determined that TBLASTN showed homology with pseudogenes belonging to cereals. Annotation of predicted pseudogenes is shown in Table 4.13.

Table 4.13. Annotation of predicted pseudogenes

Description	Max Score	Total Score	Query Cover	E value	Percent Identity	Accession Length	Accession
Oryza sativa Japonica Group ribosomal protein eL42 pseudogene (LOC4343415), non-coding RNA	136	136	87%	1.00E-39	58.02	433	NR_165376.1
Zea mays WD40 repeat superfamily pseudogene (LOC103627790), non-coding RNA	104	104	100%	8.00E-28	53.76	512	NR_161252.1
Triticum aestivum clone TaeDHS1b_3D deoxyhypusine synthase (DHS) pseudogene mRNA, complete sequence	121	121	69%	3.00E-31	61	1194	MG188732.1
Arundo donax clone AdoDHS2 deoxyhypusine synthase (DHS) pseudogene mRNA, partial sequence	77.4	77.4	70%	4.00E-15	49.02	1215	MG188688.1

5. DISCUSSION

Pseudogenes have long been accepted as junk DNA in the genome because they lost coding potential due to accumulated mutations. With the GENCODE project, the annotation of lncRNA and pseudogenes began to gain importance. Recent studies that show pseudogenes can produce proteins and play a role in gene regulation have made them increasingly essential.

Pseudogene discovery is a significant challenge due to their high sequence similarity to parental genes; therefore computational approaches are useful. This thesis focuses on pseudogene discovery using data or information from DEGs, intronic regions from available gene models, variant analysis, and MLP models, using bioinformatics and machine learning methods. These methods were originally designed in this thesis.

It was hypothesized that pseudogenes related to biotic stress can be biomarkers for disease resistance mechanisms in rice inspired by Poliseno's studies. Poliseno *et al.* found that PTENP1 and KRAS proto-oncogene, GTPase pseudogene 1 (KRASP1) play an active role in the regulation of PTEN and KRAS tumor suppressor genes in prostate cancer, respectively. They also showed that PTENP1 competes as miRNA decoy due to miRNA binding sites [96]. Moreover, Poliseno *et al.* reported that several pseudogenes had been determined as prognostic or diagnostic markers in cancer diseases. For example, PTENP1 can be a marker in prostate and breast cancer types [163].

Pseudogenes have also been listed as ncRNA, such as rRNA, tRNA, lncRNA [164]. On the contrary, Yu *et al.* reported that lncRNAs had been related to disease and miRNA decoy, like pseudogenes. However, they categorized the pseudogenes separately from lncRNA. They studied that over 10 percent (73 out of 567) of lncRNAs interact with 39 JA-related genes Xoo-infected leaves in rice. 96.91 percent putative lncRNAs had single-exon, and the median length of peptides was 51 aa. ALEX1 lncRNA has been identified as highly expressed in JA-pathway [165]. In that respect, pseudogenes are known to have accumulated mutations as the most important feature and longer transcripts, unlike ncRNA.

In RNA-seq analysis, repeats, genetic variants, and sequencing errors cause false mapping. In addition, the presence of polyploids in some crop plants poses a great difficulty in

alignment. Paired-end reads sequencing suggests reducing sequencing errors as much as possible [166].

Repeated regions such as repetitive sequences among chromosomes, rDNA units, telomeres, satellites are also main challenges for assembly [167]. Cufflinks assembler was used in this study because it accurately assembles transcripts, estimates their abundance, and tests differential expression [168].

The quality of reads was generally found poor in infected samples consisting of rice and pathogen genome. This is because the mapping is only to the rice reference genome. The low quality reading causes the alignment score to be low.

The accuracy of the alignment is the first step in reliable downstream analysis. Musich *et al.* compared six aligners such as Bowtie2, Burrows–Wheeler-Alignment (BWA), Hierarchical indexing for spliced alignment of transcripts 2 (HISAT2), MUMmer4, Spliced transcripts alignment to a reference (STAR), and TopHat2 to find the best performance for mapping in differential expression. It was presented that BWA and Bowtie2 with local mode had the highest alignment score with an average of 87 percent. BWA, Bowtie2 (local mode), Bowtie2 (end-to-end mode) were found with high coverage 97.8 percent, 97.1 percent, and 95.6 percent, respectively. Moreover, unmapped reads were obtained with the lowest number in BWA. Unmapped reads can be an important source for novel genes and genetic variety [169]. Normalization detecting and calibrating biases is one of the factors affecting differential expression. Wu *et al.* investigated normalization methods such as RPKM, trimmed mean of m values (TMM), relative log expression (RLE), and quantiles using microarray and RNA-seq datasets. They reported that TMM is the most robust procedure, as it accurately calculates DEGs even in various RNA-seq distributions [170]. Transcripts per million (TPM) was suggested for normalization compare to RPKM [171]. DESeq and TMM are robust methods used in both different library sizes and compositions [172]. TMM could be used instead of the RPKM normalization method used in this study.

Coverage is related to read counts on transcripts, while sequence depth means the count of total reads in RNA-seq. Coverage plots showed a high expression in exons and low or no expression in introns. Expression of two pseudogenes related to nematode stress and one pseudogene related to fungus stress was shown with coverage plot. According to this, pseudogenes have been less expressed than genes.

According to DEG analysis, 16 common DEGs were found as common in each dataset. Of these, importin alpha 1a which acts protein transporter activity was found upregulated, Class III peroxidase 59 and Rice ubiquitin 2 were found downregulated in each dataset. Numbers of unique DEGs were found as 329, 1230, 120 in the nematode, fungus ZB13, and fungus ZH10 datasets. As rice is susceptible to fungus ZB13 race, the most DEG was obtained in this comparison.

GO terms of gene or gene products were found related to peroxidase, oxidoreductase, and antioxidant activities in the nematode dataset, while binding such as RNA, ion, chlorophyll were found as molecular functions in the fungus dataset. As biological process, ROS, hydrogen peroxide were found in the nematode dataset while cellular metabolic process, organonitrogen compound metabolic process and photosynthesis, response to abiotic stimulus were found in the fungus ZB13 and ZH10 datasets. GO terms related to cellular components such as plasmodesma, cytosol were found in the nematode dataset, intracellular anatomical structure, intracellular organelle and photosystem were found in the fungus datasets.

According to KEGG pathway analysis, DEGs were significantly enriched in phenylpropanoid biosynthesis, photosynthesis, metabolic pathways. Dong *et al.* determined that phenylpropanoid metabolism supplies lignin to plants for mechanical support, water and nutrient transportation, physical barriers to water loss, and pathogen invasion [173].

To make biological sense of the effect of biotic stress on rice, DEGs were determined between mock and infected samples. Several DEG tools were compared, and Cuffdiff was chosen as the most suitable. Costa-Silva *et al.* have compared nine DEG tools except for Cuffdiff for differential expression. NOIseq, DESeq2, limma+voom were found with an accuracy of 90 percent, 89 percent, 88 percent [174]. Trapnell *et al.* have reported that DEGs were found in DESeq and edgeR more than Cuffdiff 2, due to small fold changes or low sequencing depth. Moreover, they presented Cuffdiff 2 can be performed robust differential expression at both gene and transcript levels in a single analysis [175].

That sequencing depth increases improve the high quality base, uniform coverage and reduces false discovery rate (FDR) for variant calling [176]. Wu *et al.* have compared variant calling tools such as SAMtools alignment-based and GATK HaplotypeCaller assembly-based with two different aligners BWA-MEM and Bowtie2-tuned. They observed that

GATK-HC and SAMtools with BWA-MEM aligner determined filtered SNPs 13.6 M, 10.4 M in domesticated tomatoes and 91.5 M, 78.7 M in wild tomatoes, respectively. GATK-HC and SAMtools with Bowtie2-tuned aligner identified filtered SNPs 8.3 M, 8.8 M in domesticated tomatoes and 28.2 M, 44.6 M in wild tomatoes, respectively. There was a significant difference between 30 wild and 52 domesticated tomatoes species [177]. In this thesis, variant analysis was done limited to the individual level.

To call SNVs, VCF files should be created for each sample, especially they should not merge. Because there can be different nucleotides that are source from sequencing error or difference one of biological replicates in sample files. Reliable SNVs should be filtered without merging files. Moreover, read length, assembler, SNP caller were main parameters for SNP calling cited by Zhao *et al.* [83]. They compared five assemblers with Trinity among them and two SNP callers such as GATK and GBS to predict reliable SNPs using different combinations. Reliable SNPs were identified in 150 bp paired-end read lengths more than 125 bp. Balasubramanian *et al.* found that SNPs are higher in repeat and intron regions than exon regions in human chromosomes 21 and 22. Moreover, they determined both Ka/Ks ratio and SNP density were higher in pseudogenes compared to genes. That Ka/Ks ratio is almost 0.5 in exons and 1 in pseudogenes has indicated that functional genes have under selective pressure but not pseudogenes [84].

To predict pseudogenes from the gene model, according to both conserved GT-AG and non-canonical GC-AG, AT-AC splice sites, introns were predicted to discover duplicated pseudogenes in intron regions predicted from consensus sequences, on AUGUSTUS.

Ka/Ks ratio was used as an important parameter for pseudogene identification, but not alone. To support this, Nekrutenko *et al.* calculated Ka/Ks ratios of protein-coding genes in humans and mice. They reported that mean lengths of initial, internal, and terminal exons were found 222 nt, 132 nt, 203 nt, respectively. Ka/Ks ratios of exons longer 300 bp were calculated as smaller than 1 [178]. In parallel, Xu *et al.* showed that translated pseudogenes in human are functional or not. They found 1464 nontranscribed, 656 transcribed, and 34 translated common pseudogenes between human and macaque ortholog genes. Moreover, they measured that translational pseudogenes have a smaller Ka/Ks ratio (<1) than transcribed or non-transcribed pseudogenes [179].

Ka/Ks ratio in the sliding window was generated to identify under positive selection regions in this study. On the contrary, Xia *et al.* recommended that the Ka/Ks peaks may be due to the increased frequency of polar amino acids and that this cannot be evidence for positive selection [180].

The relation of Ka/Ks ratio with nucleotide and protein alignment scores was investigated. A positive correlation was found between protein alignment score and Ka/Ks ratio in humans and in rice. It may have given clearer results because the number of pseudogenes detected in humans is high.

To construct an ARFF file for MLP in WEKA, length, stop codons, and numerical values of nucleotide sequences were used as attributes. Nucleotide sequences of different lengths were tried to find the optimum part and length of the consensus sequence in predicting processed pseudogenes in WEKA. The optimum length was selected as 200 bp.

Nucleotide sequences could not get as a string for the ARFF file because many machine learning algorithms use numerical value instead of nominal value. So sequences were converted to numerical values. For this, 1, 2, 3, 4 numbers were given for A, Guanine (G), Timin (T), C, respectively.

Training and test data were performed for MLP in Weka. According to the ROC curve, AUC values were found to be 0.87 and 0.90 in training and test data. 10 fold cross validation was performed to overcome overfitting.

According to tBLASTn mode results of BLAST, ribosomal protein eL42 pseudogene, WD40 repeat superfamily pseudogene, DHS pseudogene, and hypothetical proteins were found as predicted pseudogenes obtained from model. Kong *et al.* reported that wheat WD40 repeat-containing protein acts as a positive regulator of plant response to abiotic stress in crop plants [181]. Woriedh *et al.* showed that wheat DHS expression is upregulated to fungus *Fusarium graminearum* [182]. Ijaq *et al.* indicated using SVM, decision tree, naive bayes, and perceptron classifiers that most hypothetical proteins are produced from pseudogenes [183].

6. CONCLUSION

In this study, a novel method was developed to identify pseudogenes. This method occurs the integration of bioinformatics and machine learning analyzes. Moreover, pseudogenes identified from computational methods should be verified by further experimental studies.

As a result, two pseudogenes *Zea mays* ADP-ribosylation factor pseudogene and *Oryza sativa* Japonica Group ribosomal protein eL42 pseudogene were found in nematode *Meloidogyne graminicola*-infected samples. One pseudogene *Zea mays* phenylalanine/tyrosine ammonia-lyase pseudogene was found in fungus *Magnaporthe oryzae*-infected samples. 10, 34, and 15 R genes (DEG) were determined in the nematode, fungus ZB13, fungus ZH10 datasets. Functional annotation of the DEGs was performed. Expression was found in predicted intron regions obtained from the gene model. MLP was found in high prediction accuracy compared to SMO. AUC values were 87% and 90% for training and test data.

The workflow generated in this thesis will help for further discovery of novel biomarkers and will pave the way for early detection of diseases for improved disease-fighting mechanisms. The use of advanced molecular tools, combined with an artificial intelligence toolbox, will enable improved crop management. The outcomes of this study need to be validated further in other cereals or even other plant species to find traits that are generally traceable for disease prevention.

REFERENCES

1. Grain production worldwide by type, 2018/19 | Statista. [cited 2020 Oct 30]. Available from: https://www.statista.com/statistics/263977/world-grain-production-by-type/

2. Shewry PR, Tatham AS, Kasarda DD. Cereal proteins and coeliac disease. *Coeliac Dis*. 1992;305–48.

3. McKevith B. Nutritional aspects of cereals. *Nutr Bull*. 2004;29(2):111–42.

4. Tetlow IJ, Emes MJ. Starch biosynthesis in the developing endosperms of grasses and cereals. *Agronomy*. 2017;7(4).

5. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*. 2002;296(5565):92–100.

6. Yu J, Hu S, Wang J, Wong GKS, Li S, Liu B, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*. 2002;296(5565):79–92.

7. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*. 2012;326(June):1112–5.

8. Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, et al. A physical, genetic and functional sequence assembly of the barley genome. *Nature*. 2012;491(7426):711–6.

9. Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*. 2012;491(7426):705–10.

10. Bauer E, Schmutzer T, Barilar I, Mascher M, Gundlach H, Martis MM, et al. Towards a whole-genome sequence for rye (Secale cereale L.). *Plant J*. 2017;89(5):853–69.

11. Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, et al. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet*. 2019;51(5):885–95.

12. Haberer G, Mayer KFX, Spannagl M. The big five of the monocot genomes. *Curr Opin Plant Biol*. 2016;30:33–40.

13. Kumagai M, Tanaka T, Ohyanagi H, Hsing YIC, Itoh T. Genome sequences of Oryza species. In: Sasaki T, Ashikari M, editors. *Rice Genomics, Genetics and Breeding*. Springer Singapore; 2018:1–20.

14. Jena KK. The species of the genus Oryza and transfer of useful genes from wild species into cultivated rice, O. sativa. *Breed Sci*. 2010;60(5):518–23.

15. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat Genet*. 2018;50(2):285–96.

16. Khush GS. Origin, dispersal, cultivation and variation of rice. *Plant Mol Biol*. 1997;35(1–2):25–34.

17. Jain A, Sarsaiya S, Wu Q, Lu Y, Shi J. A review of plant leaf fungal diseases and its environment speciation. *Bioengineered*. 2019;10(1):409–24.

18. Decraemer W, Hunt DJ. Structure and classification. In: Perry RN, Moens M, editors. *Plant Nematology*. 2006:3–32.

19. Buonaurio R. Infection and plant defense responses during plant- bacterial interaction. *Plant-Microbe Interact*. 2008;661(2):169–97.

20. Soosaar JLM, Burch-Smith TM, Dinesh-Kumar SP. Mechanisms of plant resistance to viruses. *Nat Rev Microbiol*. 2005;3(10):789–98.

21. Nieuwenhuis BPS, James TY. The frequency of sex in fungi. *Philos Trans R Soc B Biol Sci*. 2016;371(1706).

22. Talbot NJ. Living the sweet life: How does a plant pathogenic fungus acquire sugar from plants? *PLoS Biol*. 2010;8(2):2–4.

23. Nazarov PA, Baleev DN, Ivanova MI, Sokolova LM, Karakozova M V. Infectious Plant Diseases: Etiology, Current Status, Problems and Prospects in Plant Protection. *Acta Naturae*. 2020;12(3):46–59.

24. Zhang H, Zheng X, Zhang Z. The Magnaporthe grisea species complex and plant pathogenesis. *Mol Plant Pathol*. 2016;17(6):796–804.

25. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, et al. The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature*. 2005;434(7036):980–6.

26. Kawahara Y, Oono Y, Kanamori H, Matsumoto T, Itoh T, Minami E. Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction. *PLoS One*. 2012;7(11):1–15.

27. Perry RN, Moens M. Introduction to plant-parasitic nematodes; modes of parasitism. In: Jones J, Gheysen G, Fenoll C, editors. *Genomics and Molecular Genetics of Plant-Nematode Interactions*. Springer Netherlands; 2011:3–20.

28. Coyne DL NJ and C-CB. Basic biology of plant parasitic nematodes. In: *Practical plant nematology: A field and laboratory guide*. Ibadan, Nigeria: International Institute of Tropical Agriculture (IITA); 2007:3–9.

29. Biswal AK, Shamim M, Cruzado K, Soriano G, Ghatak A, Toleco MR, et al. Role of biotechnology in rice production. In: Chauhan BS, Jabran K, Mahajan G, editors. *Rice Production Worldwide*. Springer International Publishing; 2017:487–547.

30. Mantelin S, Bellafiore S, Kyndt T. Meloidogyne graminicola: a major threat to rice agriculture. *Mol Plant Pathol*. 2017;18(1):3–15.

31. Vieira P, Gleason C. Plant-parasitic nematode effectors — insights into their diversity and new tools for their identification. *Curr Opin Plant Biol*. 2019;50:37–43.

32. Nicol JM, Turner SJ, Coyne DL, Nijs L den, Hockland S, Maafi ZT. Current nematode threats to world agriculture. In: Jones J, Gheysen G, Fenoll C, editors. *Genomics and Molecular Genetics of Plant-Nematode Interactions*. Springer Netherlands; 2011:21–43.

33. Somvanshi VS, Tathode M, Shukla RN, Rao U. Nematode genome announcement: A draft genome for rice root-knot nematode, Meloidogyne graminicola. *J Nematol*. 2018;50(2):111–6.

34. Jones JT, Haegeman A, Danchin EGJ, Gaur HS, Helder J, Jones MGK, et al. Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol Plant Pathol*. 2013;14(9):946–61.

35. Kyndt T, Denil S, Bauters L, Van Criekinge W, De Meyer T. Systemic suppression of the shoot metabolism upon rice root nematode infection. *PLoS One*. 2014;9(9):1–11.

36. Weiser JN. The battle with the host over microbial size. *Curr Opin Microbiol*. 2013;16(1):59–62.

37. Liu W, Liu J, Triplett L, Leach JE, Wang GL. Novel insights into rice innate immunity against bacterial and fungal pathogens. *Annu Rev Phytopathol*. 2014;52(5):213–41.

38. Leonard S, Hommais F, Nasser W, Reverchon S. Plant–phytopathogen interactions: bacterial responses to environmental and plant stimuli. *Environ Microbiol*. 2017;19(5):1689–716.

39. Ham JH, Melanson RA, Rush MC. Burkholderia glumae: Next major pathogen of rice? *Mol Plant Pathol*. 2011;12(4):329–39.

40. Yu C, Chen H, Tian F, Leach JE, He C. Differentially-expressed genes in rice infected by Xanthomonas oryzae pv. oryzae relative to a flagellin-deficient mutant reveal potential functions of flagellin in host–pathogen interactions. *Rice*. 2014;7(1):1–11.

41. Zipfel C. Pattern-recognition receptors in plant innate immunity. *Curr Opin Immunol*. 2008;20(1):10–6.

42. Spoel SH, Dong X. How do plants achieve immunity? Defence without specialized immune cells. *Nat Rev Immunol*. 2012;12(2):89–100.

43. Newman MA, Sundelin T, Nielsen JT, Erbs G. MAMP (microbe-associated molecular pattern) triggered immunity in plants. *Front Plant Sci*. 2013;4(139):1–14.

44. Macho AP, Zipfel C. Plant PRRs and the activation of innate immune signaling. *Mol Cell*. 2014;54(2):263–72.

45. Gómez-Gómez L, Boller T. FLS2: An LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Mol Cell*. 2000;5(6):1003–

11.

46. Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, et al. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*. 1995;270(5243):1804.

47. Thomma BPHJ, Nürnberger T, Joosten MHAJ. Of PAMPs and effectors: The blurred PTI-ETI dichotomy. *Plant Cell*. 2011;23(1):4–15.

48. Zipfel C. Plant pattern-recognition receptors. *Trends Immunol*. 2014;35(7):345–51.

49. Chisholm ST, Coaker G, Day B, Staskawicz BJ. Host-microbe interactions: Shaping the evolution of the plant immune response. *Cell*. 2006;124(4):803–14.

50. Selin C, de Kievit TR, Belmonte MF, Fernando WGD. Elucidating the role of effectors in plant-fungal interactions: Progress and challenges. *Front Microbiol*. 2016;7(4):600.

51. Tseng TT, Tyler BM, Setubal JC. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol*. 2009;9(1):1–9.

52. Pallas V, García JA. How do plant viruses induce disease? Interactions and interference with host components. *J Gen Virol*. 2011;92(12):2691–705.

53. Dangl JL, Horvath DM, Staskawicz BJ. Pivoting the plant immune system from dissection to deployment. *Science*. 2013;341(6147):746–51.

54. BioRender Templates. [cited 2021 Dec 31]. Available from: https://app.biorender.com/biorender-templates

55. Barragan AC, Weigel D. Plant NLR diversity: The known unknowns of pan-NLRomes. *Plant Cell*. 2021;33(4):814–31.

56. McHale L, Tan X, Koehl P, Michelmore RW. Plant NBS-LRR proteins: Adaptable guards. *Genome Biol*. 2006;7(4):212.

57. Yang S, Li J, Zhang X, Zhang Q, Huang J, Chen JQ, et al. Rapidly evolving R genes in diverse grass species confer resistance to rice blast disease. *Proc Natl Acad Sci U*

S A. 2013;110(46):18572–7.

58. Yang S, Zhang X, Yue JX, Tian D, Chen JQ. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol Genet Genomics*. 2008;280(3):187–98.

59. Ameline-Torregrosa C, Wang BB, O'Bleness MS, Deshpande S, Zhu H, Roe B, et al. Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant Medicago truncatula. *Plant Physiol*. 2008;146(1):5–21.

60. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell*. 2003;15(4):809–34.

61. Tsuda K, Katagiri F. Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. *Curr Opin Plant Biol*. 2010;13(4):459–65.

62. Tan S, Wu S. Genome wide analysis of nucleotide-binding site disease resistance genes in Brachypodium distachyon. *Comp Funct Genomics*. 2012;2012:1–12.

63. Li J, Ding J, Zhang W, Zhang Y, Tang P, Chen JQ, et al. Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. *Mol Genet Genomics*. 2010;283(5):427–38.

64. Hofberger JA, Zhou B, Tang H, Jones JDG, Schranz ME. A novel approach for multi-domain and multi-gene family identification provides insights into evolutionary dynamics of disease resistance genes in core eudicot plants. *BMC Genomics*. 2014;15(1).

65. Pieterse CMJ, Zamioudis C, Berendsen RL, Weller DM, Van Wees SCM, Bakker PAHM. Induced systemic resistance by beneficial microbes. *Annu Rev Phytopathol*. 2014;52:347–75.

66. Zehra A, Raytekar NA, Meena M, Swapnil P. Efficiency of microbial bio-agents as elicitors in plant defense mechanism under biotic stress: A review. *Curr Res Microb Sci*. 2021;2.

67. Flor HH. Host-parasite interactions in flax rust-its genetics and other implications. *Phytopathology*. 1955;45:680–5.

68. Harris MO, Friesen TL, Xu SS, Chen MS, Giron D, Stuart JJ. Pivoting from Arabidopsis to wheat to understand how agricultural plants integrate responses to biotic stress. *J Exp Bot*. 2015;66(2):513–31.

69. Van Der Hoorn RAL, Kamoun S. From guard to decoy: A new model for perception of plant pathogen effectors. *Plant Cell*. 2008;20(8):2009–17.

70. Ispolatov I, Doebeli M. On the evolution of decoys in plant immune systems. *Biol Theory*. 2010;5(3):256–63.

71. Jones JDG, Dangl JL. The plant immune system. *Nature*. 2006;444(7117):323–9.

72. Xiao J, Sekhwal M, Li P, Ragupathy R, Cloutier S, Wang X, et al. Pseudogenes and Their Genome-Wide Prediction in Plants. *Int J Mol Sci*. 2016;17(12):1991.

73. Jacq C, Miller JR, Brownlee GG. A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*. 1977;12(1):109–20.

74. Lerat E, Ochman H. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res*. 2005;33(10):3125–32.

75. Lafontaine I, Dujon B. Origin and fate of pseudogenes in Hemiascomycetes: A comparative analysis. *BMC Genomics*. 2010;11(1).

76. Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. Identification of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Res*. 2003;31(3):1033–7.

77. Mitrovich QM, Anderson P. mRNA surveillance of expressed pseudogenes in C. elegans. *Curr Biol*. 2005;15(10):963–7.

78. Wicker T, Mayer KFX, Gundlach H, Martis M, Steuernagel B, Scholz U, et al. Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell*. 2011;23(5):1706–18.

79. Johnsson P, Ackley A, Vidarsdottir L, Lui WO, Corcoran M, Grandér D, et al. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nat Struct Mol Biol*. 2013;20(4):440–6.

80. Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, et al. Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A*. 2014;111(37):13361–6.

81. Goodhead I, Darby AC. Taking the pseudo out of pseudogenes. *Curr Opin Microbiol*. 2015;23:102–9.

82. Salim HMW, Cavalcanti ARO. Factors influencing codon usage bias in genomes. *J Braz Chem Soc*. 2008;19(2):257–62.

83. Zhao Y, Wang K, Wang WL, Yin TT, Dong WQ, Xu CJ. A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics*. 2019;20(1):1–10.

84. Balasubramanian S, Harrison P, Hegyi H, Bertone P, Luscombe N, Echols N, et al. SNPs on human chromosomes 21 and 22 – analysis in terms of protein features and pseudogenes. *Pharmacogenomics*. 2002;3(3):393–402.

85. Wang D, Liu F, Wang L, Huang S, Yu J. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biol Direct*. 2011;6:1–17.

86. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586–91.

87. Xiao-Jie L, Ai-Mei G, Li-Juan J, Jiang X. Pseudogene in cancer: Real functions and promising signature. *J Med Genet*. 2015;52(1):17–24.

88. Poliseno L. Pseudogenes: Newly discovered players in human cancer. *Sci Signal*. 2012;5(242):re5–re5.

89. Chen X, Wan L, Wang W, Xi WJ, Yang AG, Wang T. Re-recognition of pseudogenes: From molecular to clinical applications. *Theranostics*. 2020;10(4):1479–99.

90. Borger P, Schmidtgall B. Free science from dogma. or: How the “Pseudogene” hampered scientific progress opinion. Vol. 8, AJBSR. 2020. Available from: https://biomedgrid.com/fulltext/volume8/free-science-from-dogma-or-how-the-pseudogene-hampered-scientific-progress.001239.php

91. Lee, R. C., Feinbaum, R. L., & Ambros V. The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to lin-14. *Cell*. 1993;75(5):843–54.

92. Khraiwesh B, Zhu JK, Zhu J. Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochim Biophys Acta - Gene Regul Mech*. 2012;1819(2):137–48.

93. Wahid F, Shehzad A, Khan T, Kim YY. MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. *Biochim Biophys Acta - Mol Cell Res*. 2010;1803(11):1231–43.

94. Bartel B. MicroRNAs directing siRNA biogenesis. *Nat Struct Mol Biol*. 2005;12(7):569–71.

95. Lam JKW, Chow MYT, Zhang Y, Leung SWS. siRNA versus miRNA as therapeutics for gene silencing. *Mol Ther - Nucleic Acids*. 2015;4(9):e252.

96. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*. 2010;465(7301):1033–8.

97. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. *Nature*. 2014;509(7502):575–81.

98. Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat Rev Genet*. 2020;21(3):191–201.

99. Li W, Yang W, Wang XJ. Pseudogenes: Pseudo or Real Functional Elements? *J Genet Genomics*. 2013;40(4):171–7.

100. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res*. 2012;22(9):1760–74.

101. Feingold EA, Good PJ, Guyer MS, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*. 2004;306(5696):636–40.

102. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, et al. The GENCODE pseudogene resource. *Genome Biol*. 2012;13(9).

103. Shang J, Tao Y, Chen X, Zou Y, Lei C, Wang J, et al. Identification of a new rice blast resistance gene, Pid3, by genomewide comparison of paired nucleotide-binding site-leucine-rich repeat genes and their pseudogene alleles between the two sequenced rice genomes. *Genetics*. 2009;182(4):1303–11.

104. Luo S, Zhang Y, Hu Q, Chen J, Li K, Lu C, et al. Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. *Plant Physiol*. 2012;159(1):197–210.

105. Thibaud-Nissen F, Ouyang S, Buell CR. Identification and characterization of pseudogenes in the rice gene complement. *BMC Genomics*. 2009;10:1–13.

106. Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, et al. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*. 2006;18(8):1791–802.

107. Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol*. 2009;151(1):3–15.

108. Tonner P, Srinivasasainagendra V, Zhang S, Zhi D. Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics*. 2012;13(1).

109. Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. Characterization of human pseudogene-derived non-coding RNAs for functional potential. *PLoS One*. 2014;9(4).

110. Mascagni F, Usai G, Cavallini A, Porceddu A. Structural characterization and duplication modes of pseudogenes in plants. *Sci Rep*. 2021;11(1):1–14.

111. Shein A, Zaikin A, Poptsova M. Recognition of 3′-end L1, Alu, processed pseudogenes, and mRNA stem-loops in the human genome using sequence-based and structure-based machine-learning models. *Sci Rep*. 2019;9(1):1–16.

112. Liu F, Xing L, Zhang X, Zhang X. A four-pseudogene classifier identified by machine learning serves as a novel prognostic marker for survival of osteosarcoma. *Genes*

(Basel). 2019;10(6).

113. Xing L, Zhang X, Guo M, Zhang X, Liu F. Application of Machine Learning in Developing a Novelty Five-Pseudogene Signature to Predict Prognosis of Head and Neck Squamous Cell Carcinoma: A New Aspect of "junk Genes" in Biomedical Practice. *DNA Cell Biol*. 2020;39(4):709–23.

114. Nolan T, Hands RE, Bustin SA. Quantification of mRNA using real-time RT-PCR. *Nat Protoc*. 2006;1(3):1559–82.

115. Singleton BK, Green CA, Avent ND, Martin PG, Smart E, Daka A, et al. The presence of an RHD pseudogene containing a 37 base pair duplication and a nonsense mutation in africans with the Rh D-negative blood group phenotype. *Blood*. 2000;95(1):12–8.

116. Silva-Malta MCF, Santos CCS, Gonçalves PC, Schmidt LC, Martins ML. Molecular analysis of the RHD pseudogene by duplex real-time polymerase chain reaction. *Transfus Med*. 2019;29(2):116–20.

117. Hadidi A, Czosnek H, Barba M. DNA microarrays and their potential applications for the detection of plant viruses, viroids, and phytoplasmas. *J Plant Pathol*. 2004;86:97–104.

118. Boonham N, Glover R, Tomlinson J, Mumford R. Exploiting generic platform technologies for the detection and identification of plant pathogens. *Eur J Plant Pathol*. 2008;121(3):355–63.

119. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., ... & Smith M. Nucleotide sequence of bacteriophage φX174 DNA. *Nature*. 1977;265(5596):687–95.

120. Ari Ş, Arikan M. Next Generation Sequencing: Advantages, Disadvantages and Future. In: Hakeem K, Tombuloglu H, Tombuloglu G, editors. *Plant Omics: Trends and Applications*. Springer, Cham; 2016:109–35.

121. Voelkerding K V., Dames SA, Durtschi JD. Next-generation sequencing:from basic research to diagnostics. *Clin Chem*. 2009;55(4):641–58.

122. Wu Q, Ding SW, Zhang Y, Zhu S. Identification of Viruses and Viroids by Next-

Generation Sequencing and Homology-Dependent and Homology-Independent Algorithms. *Annu Rev Phytopathol*. 2015;53(May):425–44.

123. Garrido-Cardenas JA, Garcia-Maroto F, Alvarez-Bermejo JA, Manzano-Agugliaro F. DNA sequencing sensors: An overview. *Sensors (Switzerland)*. 2017;17(3):1–15.

124. Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N. Application of next-generation sequencing technologies in virology. *J Gen Virol*. 2012;93(9):1853–68.

125. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*. 2008;24(3):133–41.

126. Kchouk M, Gibrat JF, Elloumi M. Generations of Sequencing Technologies: From First to Next Generation. *Biol Med*. 2017;09(03).

127. Van Dijk E, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet*. 2014;30(9):418–26.

128. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet*. 2010;19(R2):227–40.

129. Chiu C, Miller S, Next-generation OOF, Methods S. Next-Generation Sequencing. 2016;68–79.

130. Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLoS Comput Biol*. 2015;11(8):1–20.

131. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol*. 2011;7(3):1–5.

132. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.

133. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9.

134. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene

and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562–78.

135. Toolkit Documentation : Software : Sequence Read Archive : NCBI/NLM/NIH. [cited 2021 Dec 31]. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=prefetch

136. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.

137. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.

138. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568–76.

139. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

140. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20.

141. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: Sequence visualization and annotation. *Bioinformatics*. 2000;16(10):944–5.

142. Kaur B, Sandhu KS, Kamal R, Kaur K, Singh J, Röder MS & MQ. Omics for the Improvement of Abiotic, Biotic, and Agronomic Traits in Major Cereal Crops: Applications, Challenges, and Prospects. *Plants*. 2021;10(10):1989.

143. Arf C. Makine Düşünebilir Mi ve Nasıl Düşünebilir? *Atatürk Üniversitesi*. 1958;1959.

144. Bunker RP, Thabtah F. A machine learning framework for sport result prediction. *Appl Comput Informatics*. 2019;15(1):27–33.

145. Le Duc T, Leiva RG, Casari P, Östberg PO. Machine learning methods for reliable resource provisioning in edge-cloud computing: A survey. *ACM Comput Surv*. 2019;52(5):1–35.

146. Boateng EY, Otoo J, Abaye DA. Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *J Data Anal Inf Process*. 2020;08(04):341–57.

147. Pekel E, Kara S. A Comprehensive Review for Artifical Neural Network Application to Public Transportation. *Sigma J Eng Nat Sci*. 2017;35:157–79.

148. Roffo G. Ranking to Learn and Learning to Rank: On the Role of Ranking in Pattern Recognition Applications. Doctorate Thesis. University of Verona; 2017.

149. Wang Y, Li Y, Song Y, Rong X. The influence of the activation function in a convolution neural network model of facial expression recognition. *Appl Sci*. 2020;10(5).

150. Production volume of milled rice worldwide 2020 | Statista. [cited 2020 Oct 30]. Available from: https://www.statista.com/statistics/271972/world-husked-rice-production-volume-since-2008/

151. World population projections - Worldometer. [cited 2020 Oct 30]. Available from: https://www.worldometers.info/world-population/world-population-projections/

152. Ihaka R, Gentleman R. R: A language for data analysis and graphics. *J Comput Graph Stat*. 1996;5(3):299–314.

153. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice annotation project database (RAP-DB): An integrative and interactive database for rice genomics. *Plant Cell Physiol*. 2013;54(2):1–11.

154. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47(W1):W191–8.

155. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci*. 2021;(August):1–7.

156. Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi HB, Kanehisa* and M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 1999;27(1):29–34.

157. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006;34(Web Ser.):435–9.

158. Nei M, Gojoborit T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 1986;3(5):418–26.

159. Li, W. H., Wu, C. I., & Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*. 1985;2(2):150–74.

160. Pamilo P, Bianchi NO. Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes. *Mol Biol Evol*. 1993;10(2):271–81.

161. Analyzing synonymous and nonsynonymous substitution rates - MATLAB & Simulink. [cited 2020 Nov 5]. Available from: https://www.mathworks.com/help/bioinfo/ug/analyzing-synonymous-and-nonsynonymous-substitution-rates.html

162. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*. 2013;14(9).

163. Poliseno L, Marranci A, Pandolfi PP. Pseudogenes in human cancer. *Front Med*. 2015;2(September):1–8.

164. Wang J, Samuels DC, Zhao S, Xiang Y, Zhao YY, Guo Y. Current research on non-coding ribonucleic acid (RNA). *Genes (Basel)*. 2017;8(12).

165. Yu Y, Zhou YF, Feng YZ, He H, Lian JP, Yang YW, et al. Transcriptional landscape of pathogen-responsive lncRNAs in rice unveils the role of ALEX1 in jasmonate pathway and disease resistance. *Plant Biotechnol J*. 2020;18(3):679–90.

166. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. *Brief Funct Genomics*. 2015;14(2):130–42.

167. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-

Pozo N. Why assembling plant genome sequences is so challenging. *Biology (Basel)*. 2012;1(2):439–59.

168. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–5.

169. Musich R, Cadle-Davidson L, Osier M V. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Front Plant Sci*. 2021;12(April).

170. Wu PY, Phan JH, Zhou F, Wang MD. Evaluation of normalization methods for RNA-Seq gene expression estimation. *EEE Int Conf Bioinforma Biomed Work BIBMW*. 2011;(1):50–7.

171. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131(4):281–5.

172. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*. 2013;14(6):671–83.

173. Dong NQ, Lin HX. Contribution of phenylpropanoid metabolism to plant development and plant–environment interactions. *J Integr Plant Biol*. 2021;63(1):180–209.

174. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One*. 2017;12(12):1–18.

175. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 2013;31(1):46–53.

176. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat Rev Genet*. 2014;15(2):121–32.

177. Wu X, Heffelfinger C, Zhao H, Dellaporta SL. Benchmarking variant identification

tools for plant diversity discovery. *BMC Genomics*. 2019;20(1):1–15.

178. Nekrutenko A, Makova KD, Li WH. The KA/KS ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. *Genome Res*. 2002;12(1):198–202.

179. Xu J, Zhang J. Are human translated pseudogenes functional'. *Mol Biol Evol*. 2016;33(3):755–60.

180. Xia X, Kumar S. Codon-based detection of positive selection can be biased by heterogeneous distribution of polar amino acids along protein sequences. *Comput Syst Bioinforma*. 2006;335–40.

181. Kong D, Li M, Dong Z, Ji H, Li X. Identification of TaWD40D, a wheat WD40 repeat-containing protein that is associated with plant tolerance to abiotic stresses. *Plant Cell Rep*. 2015;34(3):395–410.

182. Woriedh M, Hauber I, Martinez-Rocha AL, Voigt C, Maier FJ, Schröder M, et al. Preventing fusarium head blight of wheat and cob rot of maize by inhibition of fungal deoxyhypusine synthase. *Mol Plant-Microbe Interact*. 2011;24(5):619–27.

183. Ijaq J, Malik G, Kumar A, Das PS, Meena N, Bethi N, et al. A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. *BMC Bioinformatics*. 2019;20(1):1–8.