ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Μετα-ανάλυση δεδομένων μεγάλης κλίμακας αλληλούχισης RNA (RNA-seq) για τη μελέτη επίδρασης του φαινομένου της αλατότητας και των ανόργανων θρεπτικών συστατικών στο ρύζι (*Oryza sativa*)

Κωνσταντίνος Ανδρέας Τζοβάρας-Μήτρογλου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνη

Γεωργία Μπράλιου

Επίκουρος Καθηγήτρια

Λαμία,

Οκτώβριος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Μετα-ανάλυση δεδομένων μεγάλης κλίμακας αλληλούχισης RNA (RNA-seq) για τη μελέτη επίδρασης του φαινομένου της αλατότητας και των ανόργανων θρεπτικών συστατικών στο ρύζι (*Oryza sativa*)

Κωνσταντίνος Ανδρέας Τζοβάρας-Μήτρογλου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπουσα

Γεωργία Μπράλιου

Επίκουρος Καθηγήτρια

Λαμία,

Οκτώβριος έτος 2023

Μετα-ανάλυση δεδομένων μεγάλης κλίμακας αλληλούχισης RNA (RNA-seq) για τη μελέτη επίδρασης του φαινομένου της αλατότητας και των ανόργανων θρεπτικών συστατικών στο ρύζι (*Oryza sativa*)

Κωνσταντίνος Ανδρέας Τζοβάρας-Μήτρογλου

Τριμελής Επιτροπή:

Γεωργία Μπράλιου, Επίκουρος Καθηγήτρια (επιβλέπουσα), Πανεπιστήμιο Θεσσαλίας, Σχολή Θετικών Επιστημών, Τμήμα Πληροφορικής με Εφαρμογές στην Βιοϊατρική

Παντελής Μπάγκος, Καθηγητής, Πανεπιστήμιο Θεσσαλίας, Σχολή Θετικών επιστημών, Τμήμα Πληροφορικής με Εφαρμογές στην Βιοϊατρική

Παναγιώτα Κοντού, Επίκουρος Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας, Σχολή Θετικών Επιστημών, Τμήμα Μαθηματικών

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(Υπογραφή)

[1] «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά την επίκουρο καθηγήτρια του τμήματος Πληροφορικής με Εφαρμογές στη Βϊοιατρική, κα. Γεωργία Μπράλιου, για την εμπιστοσύνη που μου έδειξε να αναλάβω το θέμα της παρούσας πτυχιακής εργασίας και για τον πολύτιμο χρόνο που αφιέρωσε μέχρι το πέρας της. Θέλω να ευχαριστήσω θερμά την επίκουρο καθηγήτρια του τμήματος Μαθηματικών του Πανεπιστημίου Θεσσαλίας, κα. Κοντού Παναγιώτα, για τις πολυσήμαντες συμβουλές και σχόλια, καθώς και τον κ. Μπάγκο Παντελεήμων, καθηγητή του τμήματος Πληροφορικής με Εφαρμογές στη Βϊοιατρική, για τις υποδείξεις και την εποικοδομητική κριτική στη διαμόρφωση της εργασίας.

Θα κρινόταν αδύνατη η ολοκλήρωση αυτής της εργασίας χωρίς την απεριόριστη υποστήριξη της υποψήφιας διδάκτορος, Μαρίας Κάμπα και της τακτικής επικοινωνίας που διατηρούσαμε. Τέλος ευχαριστώ τους φίλους μου για την ψυχολογική υποστήριξη που μου παρείχαν καθ’όλη τη διάρκεια εκπόνησης της εργασίας και ιδιαίτερα την οικογένεια μου, στην οποία οφείλω όλα τα επιτεύγματά μου στο ακαδημαϊκό αυτό ταξίδι που ακόμα συνεχίζω.

ΠΡΟΛΟΓΟΣ

Η παρούσα πτυχιακή εργασία εκπονήθηκε στο πλαίσιο του προπτυχιακού προγράμματος σπουδών της σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας στο τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική κατά το χρονικό διάστημα 2021-2023.

Η εργασία πραγματοποιήθηκε υπό την επίβλεψη της επίκουρης καθηγήτριας Γεωργίας Μπράλιου.

Κύριο αντικείμενο της εργασίας είναι η εφαρμογή μεθοδολογιών μετα-ανάλυσης πάνω σε γονιδιωματικά δεδομένα του ρυζιού (*Oryza sativa*), με απώτερο σκοπό την εύρεση βιλογικών παραγόντων που επηρεάζουν την ανάπτυξη του φυτού υπό την συνθήκη της αλατότητας. Ο ερευνητικός χαρακτήρας της εργασίας αυτής μου έδωσε την ευκαιρία να εξοικοιωθώ περαιτέρω με τον τομέα της βιοπληροφορικής, συγκεκριμένα τη συλλογή και επεξεργασία δεδομένων μέσα από βιολογικές βάσεις, την εφαρμογή μεθόδων και διαδικασιών μετα-ανάλυσης σε γονιδιωματικά δεδομένα και τη διεξαγωγή πλούσιων συμπερασμάτων που αποκαλύπτουν τις συχετίσεις γονιδίων στόχων με την αντοχή της ανάπτυξης του ρυζιού στο στρες αλατότητας με χρήση εργαλειών βιοπληροφορικής.

ΠΕΡΙΛΗΨΗ

Το ρύζι είναι ένα από τα πιο ευρέως καλλιεργούμενα είδη σιτηρών παγκοσμίως και απέχει μεγάλη σημασία, αγροτική οικονομία, καθώς αποτελεί βασικό κομμάτι της διατροφής για το ήμισυ περίπου του πληθυσμού της γης. Η σημαντικότητά του φέρει την ανάγκη για βελτίωση της καλλιέργειάς του υπό ιδανικές συνθήκες και την εξάλειψη απειλών που μπορούν να αναστείλουν την ανάπτυξή του. Σημαντικός παράγοντας που στέκεται τροχοπέδη στην καλλιέργεια του ρυζιού είναι η αλατότητα του εδάφους.

Στην παρούσα πτυχιακή εργασία αρχικά συλλέχθηκαν δεδομένα πειραμάτων αλληλούχισης νέας γενιάς από δείγματα ρυζιού (*Oryza sativa*) μέσα από τη δημόσια βάση δεδομένων Gene Expression Omnibus σύμφωνα με το κριτήριο PRISMA για τη συστηματική ανασκόπηση. Συγκεκριμένα, χρησιμοποιήθηκαν κανονικοποιημένες τιμές πινάκων γονιδιακής έκφρασης εκείνων των δειγμάτων που αφορούν σε καταστάσεις παρουσίας και απουσίας αλατότητας για περεταίρω έρευνα. Συνολικά, εννιά μελέτες πληρούσαν τα κριτήρια επιλογής και τα δεδομένα τους χρησιμοποιήθηκαν για την μετα-ανάλυση, εκ των οποίων οι πέντε μελέτες συνιστούν την ομάδα των δειγμάτων για τον ιστό της ρίζας του ρυζιού και οι τέσσερις συνιστούν την ομάδα για τον ιστό των σπορόφυτων - βλαστών. Τα ευρήματα της μετα-ανάλυσης εξετάστηκαν στη συνέχεια μέσω της διαδικασίας της ανάλυσης εμπλουτισμού για τον εντοπισμό βασικών βιολογικών μονοπατιών.

Abstract

Rice is one of the most cultivated products globally and remains of great importance in the agricultural economy, as it is essential to the nutritional needs of almost half of the world's population. Its importance raises the need for cultivation improvement under non-ideal conditions and the prevention of hazards that may put rice yield at risk. A significant factor that hinders rice cultivation is the salinity concentration in the soil. In this thesis, next generation sequencing experimental data were collected regarding rice (*Oryza sativa*) samples from Gene Expression Omnibus public database. The total results after searching Gene Expression Omnibus were 410 studies. Transcripts Per Million normalized expression were used as expression level of samples with and without the presence of salt were used for further research. In total nine studies were integrated into the meta-analysis process, from which four are part of the group consisting of samples derived from the seedling/shoot tissue of rice and five belong to the root group. After testing the meta-analysis process for various FDR thresholds, the value of FDR = 0.0001 was considered suitable for enrichment analysis, since it was important to minimize the numbers of DEGs. For FDR = 0.0001 and |D|> 0.5, the seedling/shoot tissue meta-analysis results consisted of 902 DEGs and the root tissue meta-analysis results consisted of 251 DEGs. The DEGs for each tissue were separately used for enrichment analysis in the STRING, PANTHER, and g:Profiler databases. Statistically enriched pathways were only found for the seedling/shoot tissue and none for the root tissue, thus it was deduced that salinity stress signaling mechanisms were more prevalent in the seedling/shoot tissue compared to the root tissue. The statistically significant enriched GO terms that were recognized for the seedling/shoot tissue were: a)response to chemical signaling, b)protein folding and function of highly structured proteins, and c)abscisic acid-activated signaling pathway, seed germination and post-embryonic development of multicellular organisms.

Key words: *Oryza sativa*, Systematic review, Meta-analysis, Differential expression, Enrichment analysis

Contents

Chapter 1- Introduction

1.1 Origins of *Oryza sativa*

Oryza sativa, popularly known as Asian rice, and *Oryza glaberrima*, generally known as African rice, are the two kinds of rice; the first is native to South Asia and the latter to Africa. One fifth of the calories consumed worldwide are from rice, which is a significant component of our diet. Although there are 40,000 more, the two most prevalent types of *Oryza sativa* are indica and japonica. These two kinds exhibit some differences between them. Unlike japonica, which has a short grain and is sticky, indica has a long grain. In addition, depending on how it is processed, rice can be divided into white and brown varieties. Although rice was late to reach every place on earth, it is nowadays a product of huge economic importance. According to the USDA (United States Department of Agriculture), in 2008, over 430 million tons of rice were consumed worldwide [1].

Focusing on *Oryza sativa*, it is considered a descendant of weeds that were possibly being cultivated on the east Himalayan plains, while other theories state that it came from India and subsequently spread north towards the center of the country, reaching China in the end. It eventually made its way to Korea, the Philippines, and then Japan and Indonesia. In particular, rice is grown on more than 25% of India's territory. While the japonica variety is thought to be a descendent of wild rice from south China, the indica variety is said to have originated in the east Himalayan plains, Burma, Thailand, Laos, and south China. Since rice represents fertility and prosperity, it plays a significant role in the civilizations of many different nations, which is why guests during the ceremony of a wedding throw rice at the newlyweds. The oldest sample of rice was peeled rice, and it is said that it was found at Hastinapur, India, during 1000–750 B.C.

1.2 Features of *Oryza sativa*

While the average height of *Oryza sativa* is between 0.5 and 2 meters, certain varieties may grow up to 6 to 9 meters tall. The root and shoot systems of rice are distinct from one another. The root portion is made up of the seminal, the crown, and the lateral roots, whereas the shoot portion is made up of the culm, the leaves, the panicle, and the spikelet [2].

1.3 Nutritional value of *Oryza sativa*

Rice mostly consists of carbs and is considered a great source of energy. It serves a variety of purposes, including making up a large portion of meals, being used in breweries, as well as producing porcelain, glass, ceramics, and paper pulp. The type of rice and the conditions in which it is grown determine the nutrients that rice contains. It is noteworthy that the vitamins and minerals included in rice are diminished by the cooking methods we employ to prepare it. It is used to treat stomach and intestinal problems since it is known for being easily digested. Its applications as medicine can also be encountered in the treatment of diseases like indigestion, arthritis, paralysis, and epilepsy.

1.4 Hazards of rice cultivation and consequences for the environment

A few elements that affect rice productivity, diminish crops, and cause other catastrophes have been identified. Large tracts of land that are swept away by rivers, diseases, parasites like rodents and birds, and reptiles can all threaten rice production. Floods, nutrient-deficient soil, and repiquetes, however, are to blame for the majority of disasters. Repiquetes, which are brought on by differences in water levels downstream of a river, are shifts in the direction in which water flows [3]. They can result from manmade or natural occurrences, and although they don't happen frequently, they can have a serious impact on rice output. Methane emissions from flooded rice fields, which contribute to the greenhouse effect and global warming, are caused by microbes. Methane emissions equal those from energy production in terms of volume [4]. Additionally, rice farming uses a lot of water; therefore, peasants look for ways to avoid using it excessively. This can be accomplished by collecting rainwater or reusing it in the agricultural process.

1.5 Salt stress on rice

Drought and salt are the key environmental conditions that influence rice production. Usually, these conditions act cooperatively since a lot of drought causes the accumulation of large amounts of salt in the soil [5]. Around the world, salinity has an impact on 20% of the areas utilized for agriculture. The salted soil usually consists of sodium, magnesium, calcium, chloride, and sulphate [6].

Due to the toxicity of sodium ions, osmotic and oxidative stressors, as well as ionic homeostasis imbalances brought on by potassium and calcium ions, this can have devastating effects on the crop [7]. Sodium and chloride also disrupt ion balance and cell metabolic functions [6]. During the early seedling age and reproductive phases, the majority of rice types are salt-susceptible, although there are a few exceptions. Salt tolerance is a complicated process that requires the overexpression of genes that are related to it and the metabolic pathways it takes part in. Some of these genes concern antioxidants, transcription factors, signal transmitters, homeostasis, ion transporters, and the regulation of osmotic potential [8].

According to research, environmental conditions like salt, cold, or drought cause epigenetic changes in rice, including DNA and chromatin alterations, differential gene expression, chromatin structural changes, and the generation of smallRNAs [9]. Changes in chromatin, in particular, are crucial in controlling the differential expression of genes associated with salt tolerance on the roots and seedlings of rice. There is also a difference in the impact of salt stress across rice varieties. Results from a study showed that the Nonabokra variety proved to be more salt-tolerant compared to the Pokkali variety under a long period of salinity stress [10].

1.6 RNA-seq

In 1990 with the invention of DNA microarrays, researchers were able to measure the expression of several genes derived from various organisms. Nevertheless, in the mid-2000s, the new technology of next generation sequencing (NGS) revolutionized the field of genetics. Before long, it was found out with the utilization of NGS that RNA sequencing provides more insight compared to DNA sequencing and a better understanding of the transcriptome and the events that are related to it [11].

RNA-seq, or RNA-sequencing, first appeared in 2010, and it is a powerful gene expression analysis technique. It relies on next generation sequencing (NGS) technologies and requires the sequencing of cDNA molecules, that derive from different samples. After mapping the reads collected downstream of the transcriptome, it is possible to quantify the transcript and the exons. Exon quantification can also be used to identify alternative splicing isomorphs. In RNA-seq, mutations and uncommon transcripts are more likely to be found the more reads there are. It provides the possibility of the exact localization of the transcript start position, while focusing on the localization of small non-coding RNA (sncRNA). However, this technique is expensive; hence, the repetitions of it are limited, and experimental mistakes may occur during the process [12].

As for the procedure of RNA-seq, there are several specified steps that should be followed. Initially, the samples for the differential expression analysis are selected. Afterwards, RNA is isolated, from which a cDNA library is constructed. Next, data gets collected, and next generation sequencing gets conducted. Following is the data analysis part. For clarity, the transcriptome reads are mapped, and the transcript units are put together. Finally, the findings are verified and published on databases to make it easier for scientists throughout the world to access and reuse this data for additional research. A handful of those databases are ENA(European Nucleotide Archive) [13], GEO(Gene Expression Omnibus), NCBI(National Center for Biotechnology Information) [14], SRA(Sequence Read Archive) [15], TCGA(The Cancer Genome Atlas), ArrayExpress of EBI(European Bioinformatics Institute) [16] and ENCODE(Encyclopedia of DNA Elements) [17].

There are a few techniques and platforms that can carry through next-generation sequencing. Some of those are the Illumina method, used by HiSeq technologies; pyrosequencing; ABI SOLiD sequencing, which has highly accurate results along with Complete Genomics; Ion Torrent, which relies on pH; and Pacific Biosciences. By means of next-generation sequencing development, the time and cost required to complete the procedure of sequencing have been drastically reduced compared to Sanger sequencing. Furthermore, the mass production of reads generated by next-generation sequencing is another notable achievement of this technology.

1.7 Systematic review

Meta-analysis and systematic review are two interconnected concepts, as the latter comes before the former in order for the results to be valid. The first meta-analysis was conducted by Karl Pearson in 1904 [18]. In a meta-analysis, data from studies is statistically analyzed with the goal of producing findings for a particular study issue.

A systematic review begins by formulating a clear research question that addresses the sample and the prevalence of the disease being studied. The studies that are appropriate for the study are then identified using the criteria for inclusion and exclusion, and their findings are then employed in the meta-analysis process. However, the bibliography shouldn't be missing as well, which must be systematic and analytical, including all the sources from studies and projects it has been referenced to.

1.8 Meta-analysis

Meta-analysis, the final phase of systematic review, yields an overall outcome by combining the findings of different studies, putting them into a single analysis, and describing them using the effect size method. It is important to decide on which effect size metric the meta-analysis results are going to be expressed in be comparable with other studies. Some common measures are the standardized mean difference, the odd ratio, and the correlation coefficient. Depending on the parameters of the effect sizes there are two models that should be implemented: the random effects model and the fixed effects model, where the first doesn't have identical effect size parameters across the studies and the latter does.

Initially, publication bias is being checked in order for to prevent the collection of non-random sample of studies, which may lead to erroneous results after the meta-analysis. For homogeneity to exist among the analyzed samples, the check for heterogeneity must be given importance. If the heterogeneity is greater than it is expected to be, then it may influence the credibility of the results. Finally, the meta-analysis can be concluded by estimating the overall outcome of the meta-analysis using a mathematical procedure [19].

Moreover, another method that can take place after obtaining the results of the meta-analysis is the sensitivity analysis. Normally, the first meta-analysis that was conducted used randomized studies. The sensitivity analysis is a second round of the meta-analysis, but this time non-randomized studies get integrated into the process to determine whether the results are similar or different from the results that occurred from the prior meta-analysis. Depending on the similarity of those results, it can be indicated that the type of study has little or big effect on the meta-analysis results [20].

1.9 Enrichment analysis

Enrichment analysis, evaluating the relationship between genes and proteins, biological pathways, and phenotypes, is the final step in many studies. The list of genes integrated for the process that follows, which consists of genes or proteins that are considered statistically significant, will next be identified and associated with the biological pathways they're related to. Coming after is the biological annotation and the matching of significant genes or proteins on the list with Gene Ontology (GO) terms [21] and ultimately detecting the most appeared terms in the list.

There are various well-known tools that easily implement enrichment analysis and deliver trustworthy findings. PANTHER [22] gives the possibility of exploitation of genes, ontologies, pathway functions, and statistical analysis tools. It contains 82 genomes, phylogenetic trees, and multiple sequence alignment functions. Information on protein interactions and their activities is provided and documented in the STRING [23] database. The possibility of designing and examining protein networks is another possibility in STRING. g:Profiler [24] is another tool that identifies genes according to the biological process in which they participate. It also allows for the comparison of these genes with ontologies and the conversion of gene IDs between different formats. It's worth noting that Ensembl [25] provides data to g:Profiler.

Chapter 2- Methods

2.1 Systematic review for studies associated with *Oryza sativa*

The Gene Expression Omnibus database (GEO) [26] was thoroughly searched for studies that comprise *Oryza sativa* samples generated from RNA-seq experiments and are associated with salinity stress. In order to gather all the necessary information, a combination of search terms and filters was used. Using the terms “(*Oryza sativa*) AND "*Oryza sativa*"[porgn:__txid4530]” in the search-bar and including additional filters such as “expression profiling by high throughput sequencing” and “non-coding RNA profiling by high throughput sequencing”. Nonetheless most of those studies were discarded since they weren’t related to the condition of interest. The systematic review was conducted by utilizing the PRISMA guidelines for systematic reviews [27]. Studies that provided samples for three distinct tissues, the root, the seedling, and the shoot, were categorized into two classes. The first class comprised of studies that referred to the seedling-shoot tissue, and the second class comprised of the studies that referred to the root tissue. All data comprised of case and control samples. Control refers to the wild type samples, whereas cases are the ones exposed to a condition (salinity stress in this study). Moreover, additional information was kept throughout each study, such as the rice cultivars for each sample, the year of publish of the study, the sequencing platform that was used for the RNA-seq experiment, as well as the total number of unique genes observed.

2.2 Data preprocessing

After collecting all the supplementary files needed from each study, the next step was to maintain the same gene ID format for all the gene lists throughout these files. Most gene identifiers in the supplementary files were provided in the MSU format according to the Rice Genome Annotation Project (RGAP) [MSU ID: LOC_OsXXgXXXXX] and were kept that way. Several others provided information according to the IRGSP gene annotation version. Those gene identifiers were that converted from the original RAP format [RAP ID: OsXXgXXXXXXX] to the MSU format using the ID converter tool of RAP-DB [28]. Surprisingly, not every gene ID had a corresponding one from RAP to MSU format. Finally, genes that weren't part of the *Oryza sativa* gene were excluded from the preprocessing step.

In order to make all data across all studies comparable, a universal format for the expression level values was needed to be used. Different metrics of expression levels were observed throughout the supplementary data files that were processed. The metrics that were encountered in these files were TPM (transcripts per kilobase million), and RPKM/FPKM (reads/fragments per kilobase million). In brief, these normalized units can quantify mRNA abundance and can be used for differential expression analysis. TPM is suitable as an alternative to RPKM/FPKM, since it is proportional to RPKM/FPKM [29]. Every RPKM/FPKM expression value of each data table was transformed in TPM metric using the equations below:

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j}\right) \cdot 10^6$$

$$TPM_i = \left(\frac{RPKM_i}{\sum_j RPKM_j}\right) \cdot 10^6$$

Another thing that came to consideration was the duplicate appearances of several genes in the same file, leading to multiple values concerning a single gene id. For every duplicate value across all studies the mean TPM value was calculated using the collapse command from Stata statistical software [30]. Additionally, genes that had a value of 0 across all samples in a study were excluded from the final table of each particular study.

2.3 Meta-analysis

Meta-analysis was conducted using the tool MAGE [31], a python package for gene expression data meta-analysis. Among its three main functions of probe converting to gene identifiers, meta-analysis, and enrichment analysis, the second was selected for the progress of this study. It includes operations for computing the standardized mean difference as the effect size in the meta-analysis process of gene expression studies as well as a Bayesian approach to the above method.

The standardized mean difference also known as Cohen's d is a mostly preferred method for meta-analysis, even though it tends to return biased values. That raises the need for a correction process, which can provide an unbiased measure. By converting the standardized mean difference into Hedge's g, the results are more reliable [32]. The formulas below showcase the calculation of this effect size as well as the conversion from Cohen's d to Hedge's g:

$$d = \frac{M_1 - M_2}{s_{pooled}}$$

$$s_{pooled} = \sqrt{\frac{(s_1^2 + s_2^2)}{2}}$$

$$g = \frac{M_1 - M_2}{SD_{pooled}^*}$$

$$SD_{pooled}^* = \sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-1}}[33]$$

Where:

M_1 and M_2 indicate the means of group 1 and group 2, respectively. s_1 and s_2 represent the variations as well as SD_1 and SD_2 and the standard deviations for group 1 and group 2, respectively. Lastly, n_1 and n_2 depict the sample sizes of each group.

Bayesian methods estimate results based on prior observed data. The model is decided by the combination of prior distributions of relevant studies and findings and a likelihood function, constructing a posterior density function. Consequently, this posterior distribution comprises the prior data for the next Bayesian approach that follows [34], [35]. The following is a model that includes within-study and between-study variances:

$$T_i \sim N(\mu, S_i^2 + \tau^2)$$

Where:

T_i refers to the study results, with μ means and S_i^2, τ^2 within-study variances and between-study variances, as mentioned before, respectively.

The tests ran individually for each study as well as for all the grouped studies across the same plant tissue, resulting in meta-analysis results. The parameter of false discovery rate (FDR), an error rate suitable for rejecting false true findings that would be considered significant in a different case, was tested for multiple thresholds particularly less than 0.05, 0.01, 0.005, 0.001, 0.0005 and 0.0001.

FDR as proposed from Benjamini and Hochberg is defined as:

$FDR = E(Q)[36]$

Where:

$Q \equiv \frac{V}{R}$ with R > 0. V is the number of false discoveries and R the number of significant results.

After compiling, the program indicates the most statistically significant genes that are differentially expressed and ready for enrichment analysis.

The measure of heterogeneity, I^2, was also taken into account in order to visualize the variability in the meta-analysis. I^2 is expressed mathematically as follows:

$$I^2 = \frac{\tau^2}{(\tau^2+\sigma^2)}[37]$$

Where:

τ^2 is the between trial heterogeneity and $\tau^2 + \sigma^2$ is the sum of the variation in the meta-analysis.

2.4 Enrichment analysis

After retrieving the results from the meta-analysis process, the list of significant genes is integrated into the PANTHER [22], STRING [23], and g:Profiler [24] tools to generate valuable results. All platforms provide tables of important proteins associated with either biological processes, molecular functions, or cellular components. Moreover, STRING creates a protein-protein interaction (PPI) network that depicts the relationship between proteins, where nodes represent the proteins and edges represent the way proteins are associated with each other. By observing a protein-protein interaction network, important nodes that have a high degree of connectivity can be detected, as well as nodes that are disconnected and are not of great significance.

Chapter 3 – Results

3.1 Studies of interest after the systematic review

From a total of 410 studies related to *Oryza sativa* and RNA-seq high throughput sequencing (that were gathered after searching in the GEO database on 22/12/2021) and after following the guidelines to fill the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram, 302 studies were discarded (Fig. 1). Seventeen studies associated with iron, cadmium, nitrogen, fluoride, zinc, alkaline, aluminum, arsenic and hydrogen peroxide absence had few studies for each study, so they were not further examined. Furthermore, studies using samples from tissues other than roots, shoots, or seeds were excluded since they were not relevant to the current topic of the meta-analysis.

The results of the systematic review led to two groups of studies classified according to the plant tissue they came from (Table 1). Specifically, four studies make up the seedling group, while the others are included in the root group. GSE132183 appears in both groups since it includes both root and shoot samples. The shoot ones were included in the seedling group as the shoot is the seedling of the rice when it reaches the surface. The studies collected were published from 2015 up to 2021, and the platforms used for the NGS were either Illumina Genome Analyzer II or Illumina HiSeq 2500. The supplementary files provided for every study gave valuable details about the genes of the rice genome as well as the expression values for every gene. The sum of the unique genes met in the studies varies between 24464 and 46224 among them.

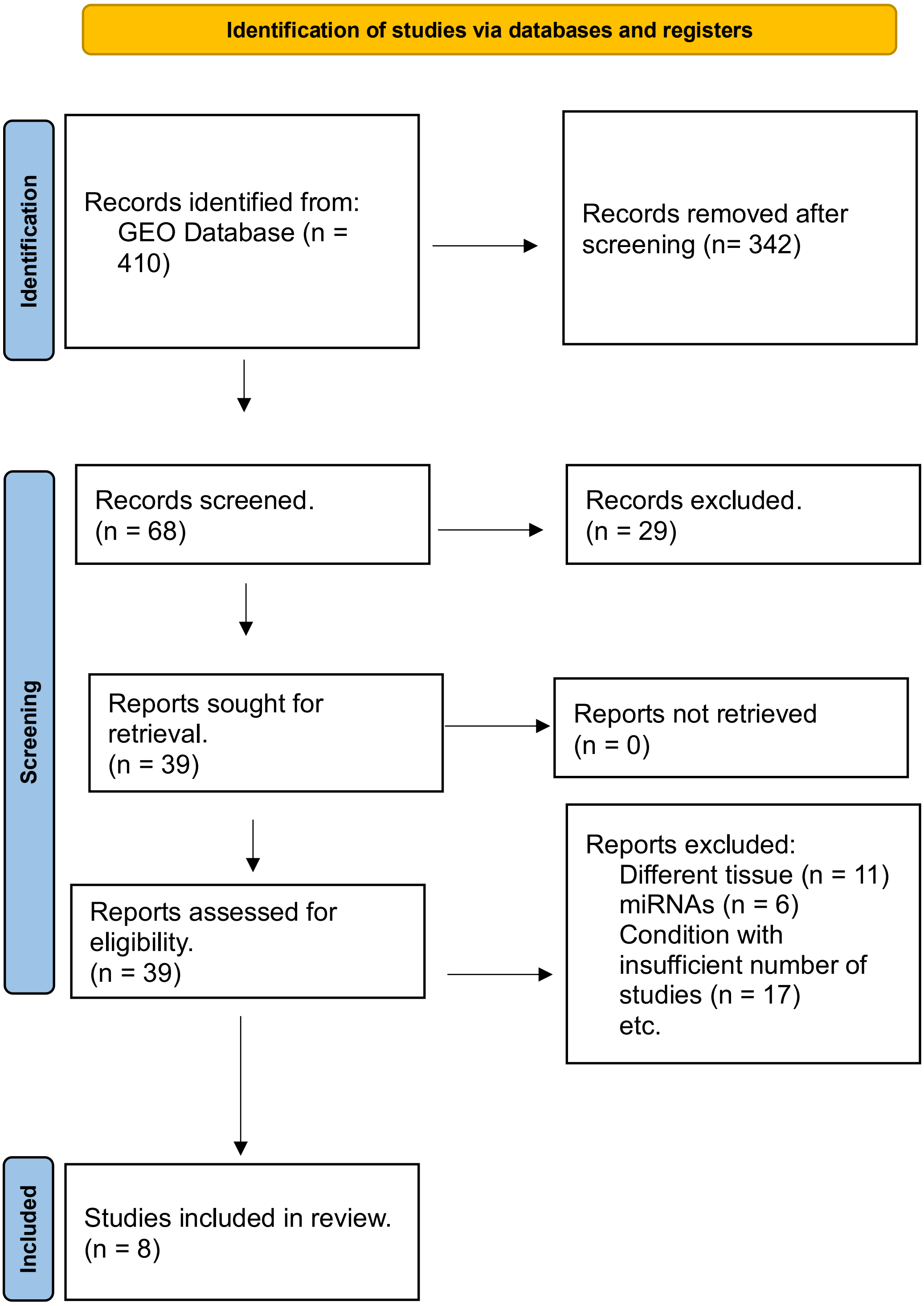
Identification of studies via databases and registers
Identification
Records identified from:
GEO Database (n = 410)
Records removed after screening (n= 342)
Screening
Records screened.
(n = 68)
Records excluded.
(n = 29)
Reports sought for retrieval.
(n = 39)
Reports not retrieved
(n = 0)
Reports assessed for eligibility.
(n = 39)
Reports excluded:
Different tissue (n = 11)
miRNAs (n = 6)
Condition with insufficient number of studies (n = 17)
etc.
Included
Studies included in review.
(n = 8)

Figure 1. PRISMA Flow Diagram. Detailed information on the process of systematic review and selecting of the studies most suited for the meta-analysis.

Table 1. Studies obtained from GEO for meta-analysis. A comprehensive array highlighting fundamental information for each study.

GEO Dataset	Year Published	Platform	Rice Tissue	Oryza sativa Subspecies and Cultivars	Number of Controls	Number of Cases	Number of Unique Genes
GSE60287	2015	Illumina Genome Analyzer II	Seedling	Pokkali, IR64	2	2	34828
GSE119720	2019	Illumina Genome Analyzer II	Seedling	Pokkali, IR64	2	2	46224
GSE132183	2019	Illumina HiSeq 2500	Shoot	Nipponbare	3	3	30634
GSE143922	2020	Illumina HiSeq 2500	Seedling	Nipponbare	3	3	24464
GSE86860	2021	Illumina HiSeq 2500	Root	IR29	3	3	26241
GSE80670	2021	Illumina HiSeq 2500	Root	Mulai	3	3	27557
GSE132183	2019	Illumina HiSeq 2500	Root	Nipponbare	3	3	32549
GSE102152	2017	Illumina HiSeq 2500	Root	Normal_Rice_R1	3	3	37660
GSE109617	2019	Illumina HiSeq 2500	Root	Nipponbare	2	2	33273

3.2 SMD and Bayesian meta-analysis

Going over two separate approaches for the meta-analysis procedure, it was expected for the outcomes to coincide, but that wasn’t the case. The number of differentially expressed genes was examined for FDR values of 0.05, 0.01, 0,005, and 0.001 and 0.0005, 0.0001 only for the SMD method (Table 2, Table 3). As a result, when compared to Bayesian results, the SMD method's outcomes are seen as being more reliable.

Table 2. Results of MAGE meta-analysis tool using the SMD method.

STUDY	0.05 FDR	0.01 FDR	0.005 FDR	0.001 FDR	0.0005 FDR	0.0001 FDR
GSE60287	14921	11466	10479	8648	7975	6454
GSE119720	5496	3821	3319	2470	2210	1978
GSE132183	10526	6248	4716	19	59	59
GSE143922	8186	4829	3551	10	10	10
GSE86860	4	3	3	3	3	3
GSE80670	497	0	0	0	0	0
GSE132183(shoot)	10526	6248	4716	19	19	19
GSE102152	15338	9594	7544	2541	112	112
GSE109617	19789	17298	16437	14546	13855	12036
Seedling meta-analysis	3893	1977	1598	1004	902	902
Root meta-analysis	1356	754	516	251	251	251

Table 3. Results of MAGE meta-analysis tool using the Bayesian method.

STUDY	0.05 FDR	0.01 FDR	0.005 FDR	0.001 FDR
GSE60287	8899	8453	7436	5977
GSE119720	3007	3007	3007	2495
GSE132183	11163	9940	8866	7014
GSE143922	8992	8016	7068	5617
GSE86860	2492	2492	2492	1985
GSE80670	3326	3326	3189	2326
GSE132183(shoot)	15492	13752	12635	10675
GSE102152	15603	13694	12220	9851
GSE109617	15227	13311	12112	10337
Seedling meta-analysis	2916	2916	2916	2593
Root meta-analysis	2593	2593	2593	2287

3.3 Effect size and Heterogeneity I squared

After retrieving the meta-analysis results, it was important to visualize the results in plots using variables such as effect size (d), which in this case is the standardized mean difference, and p-value. In the effect size plots, the effect size values can be observed between the ranges of -10 and 10 (Figure 2, Figure 3). The results for heterogeneity both in the seedling/shoot and root tissues depict a lot of values greater than 50% making the standardized mean difference suitable for this meta-analysis as belongs to the random-effects models(Figure 4). In the volcano plots, the distribution of the upregulated and downregulated genes is shown right and left, respectively (Figure 5). All the plots were produced using the tools provided by Microsoft Office Excel.

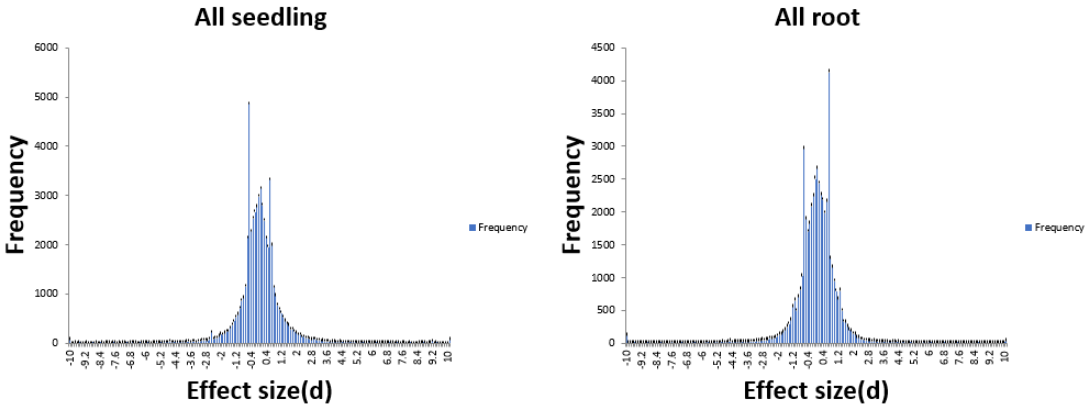

Figure 2. Effect size(d) plots for meta-analysis conducted separately on seedling and root group.

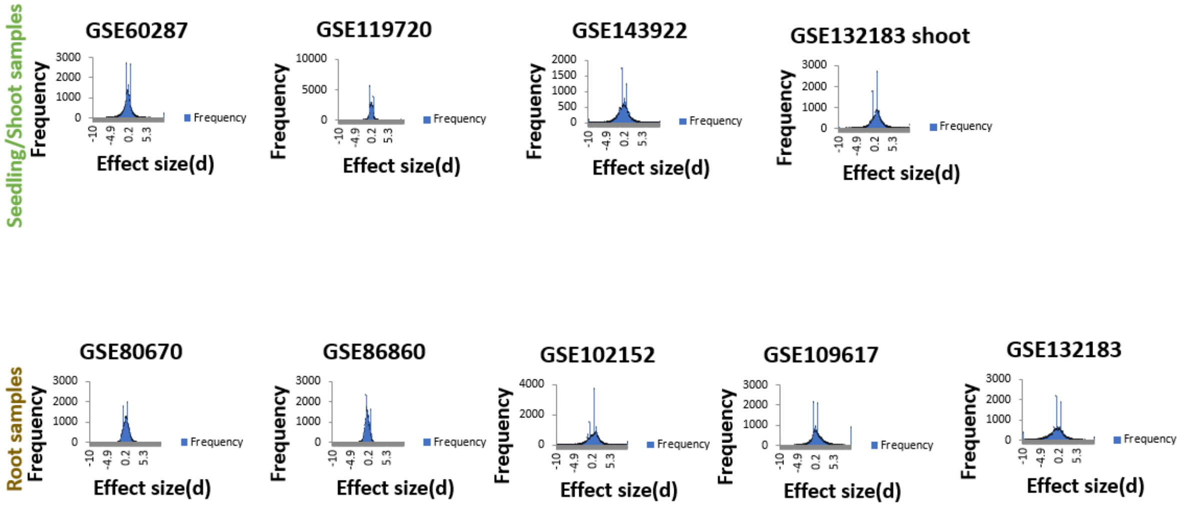

Figure 3. Effect size(d) plots for each sample.

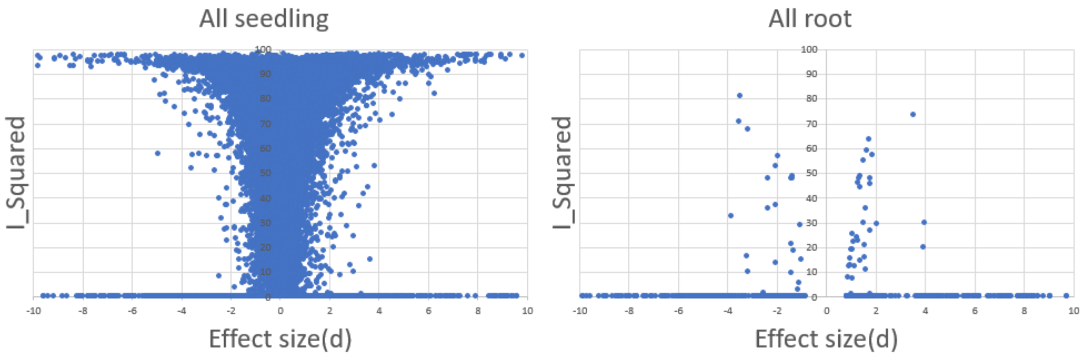

Figure 4. I squared plots for meta-analysis conducted separately on seedling and root group.

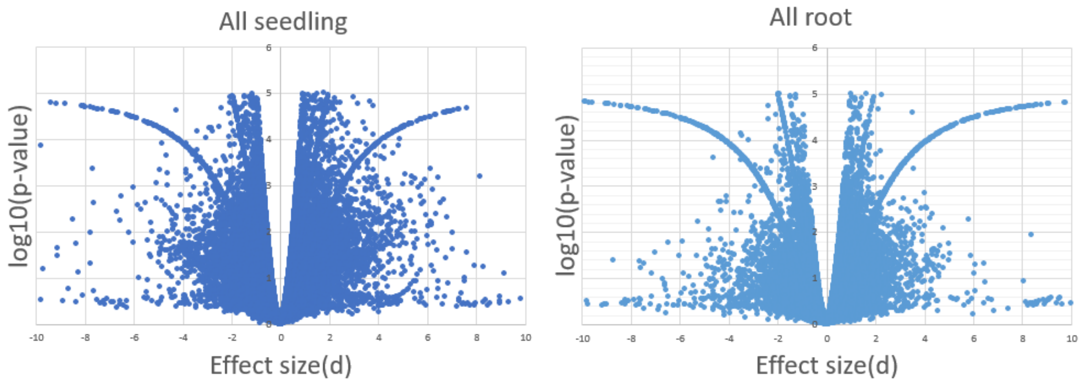

Figure 5. Volcano plots for meta-analysis conducted separately on seedling and root group.

3.4 Enrichment analysis for seedling/shoot tissue

The gene list for the seedling/shoot tissue that was used for the enrichment analysis tools consists of 902 genes, for FDR<0.0001 and -0.5<SMD<0.5 (Table 4). A STRING [23] network was created with a medium interaction score of 0.4 including 478 DEGs from the initial 902 and contains interactions predicted by gene neighborhood and verified from experiment, and database sources (Figure 6). Purple edges refer to protein interactions that are experimentally determined, blue edges refer to protein interactions from curated databases and green edges are predicted by gene neighborhood. Numerous nodes demonstrated high connectivity to others, rendering them important as central nodes. A table was exported from STRING in order to showcase the proteins with the highest centrality, along with their entry identifiers from UniprotKB, their biological processes, molecular functions and cellular components (Table 5, Table 6). Three tables were also acquired from STRING, which highlighted enriched biological processes, molecular functions, and cellular components (Table 7, Table 8, Table 9). In these tables the first column includes the GO identifier of each term, the second column enumerates the count of proteins observed in the STRING network, the third column notes the total number of annotated proteins associated to the term, the fourth column indicates the effect of the enrichment analysis, and the fifth column is comprised of false discovery rates, highlighting the statistical significance of the term.

A table of enriched terms was also created after using the g:Profiler [24] tool (Table 10). Some important information that are noted in the table are the types of terms, their GO identifiers, their adjusted p values, the total number of documented proteins associated with a term, as well as the number of proteins present in this query.

Finally, the enriched pathways that were acquired from the PANTHER [22] database, were results associated with biological processes and a molecular function along with their GO identifiers and false discovery rates (Table 11, Table 12).

The results indicate that three major biological process pathways were enriched, referring to GO terms: a) genes involved in response to various kinds of chemicals; b) genes involved in protein folding, oligomerization, and binding; and c) genes involved in the abscisic acid-activated signaling pathway, seed germination, and post-embryonic development of multicellular organisms. Additionally, regarding the STRING PPI network, the proteins noted with the biggest centralities in Table 5 and Table 6 belong to major protein families such as AMP-binding enzymes, guanylate kinase 2, N-terminal domain-containing, FACT complex, TATA-binding, heat stress transcription factor, and heat shock proteins. Some of those like the ubiquitin and SUMO proteins have protein modification functions, along with transcription factor binding to DNA and chromosome remodeling are included in the second GO term category that was mentioned about the protein folding and binding.

Table 4. Statistically significant differentially expressed seedling/shoot genes for FDR<0.0005 and |D|>0.5.

1	Os01g0565650		36	Os01g0558100
2	Os01g0105800		37	Os01g0583100
3	Os01g0136200		38	Os01g0592100
4	Os01g0149800		39	Os01g0607600
5	Os01g0153950		40	Os01g0631900
6	Os01g0163600		41	Os01g0633400
7	Os01g0166700		42	Os01g0638600
8	Os01g0176700		43	Os01g0656200
9	Os01g0183600		44	Os01g0672166
10	Os01g0184100		45	Os01g0695700
11	Os01g0204900		46	Os01g0699100
12	Os01g0221900		47	Os01g0702450
13	Os01g0222001		48	Os01g0705700
14	Os01g0225600		49	Os01g0712000
15	Os01g0226400		50	Os01g0733801
16	Os01g0232000		51	Os01g0759550
17	Os01g0237366		52	Os01g0759800
18	Os01g0240500		53	Os01g0767600
19	Os01g0243400		54	Os01g0771300
20	Os01g0243700		55	Os01g0780800
21	Os01g0256500		56	Os01g0794700
22	Os01g0262900		57	Os01g0800500
23	Os01g0276300		58	Os01g0806101
24	Os01g0308600		59	Os01g0818900
25	Os01g0316500		60	Os01g0819000
26	Os01g0320100		61	Os01g0837400
27	Os01g0324300		62	Os01g0856800
28	Os01g0363300		63	Os01g0862600
29	Os01g0368000		64	Os01g0867300
30	Os01g0369432		65	Os01g0868800
31	Os01g0370600		66	Os01g0881900
32	Os01g0501000		67	Os01g0884700
33	Os01g0502700		68	Os01g0885800
34	Os01g0505400		69	Os01g0915600
35	Os01g0524850		70	Os01g0918400

71	Os01g0962700		106	Os02g0757600
72	Os01g0973200		107	Os02g0768400
73	Os02g0126000		108	Os02g0782300
74	Os02g0129200		109	Os02g0782500
75	Os02g0132200		110	Os02g0785600
76	Os02g0137800		111	Os02g0799600
77	Os02g0140800		112	Os02g0799700
78	Os02g0142400		113	Os02g0810200
79	Os02g0157700		114	Os02g0811400
80	Os02g0178400		115	Os02g0816100
81	Os02g0227900		116	Os02g0828200
82	Os02g0233200		117	Os03g0101600
83	Os02g0259700		118	Os03g0113700
84	Os02g0274600		119	Os03g0118450
85	Os02g0288925		120	Os03g0125900
86	Os02g0296700		121	Os03g0127700
87	Os02g0323000		122	Os03g0132200
88	Os02g0487900		123	Os03g0137400
89	Os02g0508100		124	Os03g0147200
90	Os02g0513100		125	Os03g0156000
91	Os02g0527200		126	Os03g0158600
92	Os02g0531900		127	Os03g0161200
93	Os02g0555600		128	Os03g0166100
94	Os02g0559500		129	Os03g0168200
95	Os02g0602500		130	Os03g0173800
96	Os02g0607500		131	Os03g0179700
97	Os02g0619500		132	Os03g0192700
98	Os02g0642300		133	Os03g0232250
99	Os02g0655750		134	Os03g0237500
100	Os02g0671100		135	Os03g0249700
101	Os02g0686100		136	Os03g0266300
102	Os02g0715000		137	Os03g0266500
103	Os02g0734900		138	Os03g0277300
104	Os02g0735900		139	Os03g0286900
105	Os02g0753800		140	Os03g0298300

141	Os03g0301600		176	Os03g0804100
142	Os03g0304500		177	Os03g0806700
143	Os03g0305100		178	Os03g0813500
144	Os03g0305400		179	Os03g0827900
145	Os03g0307000		180	Os03g0843300
146	Os03g0314500		181	Os03g0852000
147	Os03g0320900		182	Os03g0857500
148	Os03g0322100		183	Os03g0862200
149	Os03g0322900		184	Os04g0101300
150	Os03g0325000		185	Os04g0107900
151	Os03g0325700		186	Os04g0111100
152	Os03g0330200		187	Os04g0136700
153	Os03g0340100		188	Os04g0137600
154	Os03g0340700		189	Os04g0162500
155	Os03g0341300		190	Os04g0163150
156	Os03g0351100		191	Os04g0165100
157	Os03g0366900		192	Os04g0244201
158	Os03g0367900		193	Os04g0244800
159	Os03g0380700		194	Os04g0306400
160	Os03g0393375		195	Os04g0327100
161	Os03g0436600		196	Os04g0341650
162	Os03g0449200		197	Os04g0351300
163	Os03g0587700		198	Os04g0375300
164	Os03g0600500		199	Os04g0409900
165	Os03g0619850		200	Os04g0415800
166	Os03g0623100		201	Os04g0453200
167	Os03g0633500		202	Os04g0472600
168	Os03g0663300		203	Os04g0476400
169	Os03g0695000		204	Os04g0481800
170	Os03g0702700		205	Os04g0486300
171	Os03g0717500		206	Os04g0486900
172	Os03g0750900		207	Os04g0500100
173	Os03g0760100		208	Os04g0504100
174	Os03g0783800		209	Os04g0530500
175	Os03g0784900		210	Os04g0539100

211	Os04g0558700		246	Os05g0425700
212	Os04g0569300		247	Os05g0447700
213	Os04g0571200		248	Os05g0460000
214	Os04g0576800		249	Os05g0466200
215	Os04g0578800		250	Os05g0477900
216	Os04g0580300		251	Os05g0508900
217	Os04g0580400		252	Os05g0541200
218	Os04g0592400		253	Os05g0545200
219	Os04g0607500		254	Os05g0557100
220	Os04g0632000		255	Os05g0562300
221	Os04g0635500		256	Os05g0571800
222	Os04g0635650		257	Os05g0571900
223	Os04g0654600		258	Os06g0104800
224	Os04g0659100		259	Os06g0110200
225	Os04g0666100		260	Os06g0120466
226	Os04g0692750		261	Os06g0129900
227	Os05g0111100		262	Os06g0131700
228	Os05g0149500		263	Os06g0134800
229	Os05g0162500		264	Os06g0135460
230	Os05g0179950		265	Os06g0139800
231	Os05g0182600		266	Os06g0140400
232	Os05g0235800		267	Os06g0140700
233	Os05g0254800		268	Os06g0142200
234	Os05g0267800		269	Os06g0142350
235	Os05g0296398		270	Os06g0146700
236	Os05g0305150		271	Os06g0154200
237	Os05g0333500		272	Os06g0157500
238	Os05g0344200		273	Os06g0166400
239	Os05g0349800		274	Os06g0169700
240	Os05g0358200		275	Os06g0180200
241	Os05g0389400		276	Os06g0191000
242	Os05g0394401		277	Os06g0191700
243	Os05g0407500		278	Os06g0193700

244	Os05g0409300		279	Os06g0211200
245	Os05g0414600		280	Os06g0213551

281	Os06g0224000		316	Os07g0203400
282	Os06g0261600		317	Os07g0204100
283	Os06g0266266		318	Os07g0224200
284	Os06g0269100		319	Os07g0249800
285	Os06g0277200		320	Os07g0421932
286	Os06g0300500		321	Os07g0448150
287	Os06g0302000		322	Os07g0507500
288	Os06g0498800		323	Os07g0515900
289	Os06g0499600		324	Os07g0516700
290	Os06g0517700		325	Os07g0533201
291	Os06g0551500		326	Os07g0558300
292	Os06g0553100		327	Os07g0575700
293	Os06g0565300		328	Os07g0597100
294	Os06g0592000		329	Os07g0633200
295	Os06g0609700		330	Os07g0642800
296	Os06g0635200		331	Os07g0661600
297	Os06g0659750		332	Os07g0669300
298	Os06g0660625		333	Os07g0673400
299	Os06g0665200		334	Os07g0675300
300	Os06g0671600		335	Os07g0685900
301	Os06g0698400		336	Os08g0105100
302	Os06g0705000		337	Os08g0105800
303	Os06g0707000		338	Os08g0106501
304	Os06g0710300		339	Os08g0120500
305	Os06g0713900		340	Os08g0125850
306	Os06g0714600		341	Os08g0141600
307	Os06g0728000		342	Os08g0144000
308	Os07g0101000		343	Os08g0155800
309	Os07g0113450		344	Os08g0157600
310	Os07g0114000		345	Os08g0360800
311	Os07g0130200		346	Os08g0371608
312	Os07g0146300		347	Os08g0372700
313	Os07g0160300		348	Os08g0376300
314	Os07g0173601		349	Os08g0376700
315	Os07g0183050		350	Os08g0408500

351	Os08g0423600		386	Os09g0482000
352	Os08g0425000		387	Os09g0484300
353	Os08g0429800		388	Os09g0491644
354	Os08g0460000		389	Os09g0528200
355	Os08g0471401		390	Os09g0530700
356	Os08g0543900		391	Os09g0534700
357	Os08g0558300		392	Os09g0541100
358	Os08g0563200		393	Os09g0557500
359	Os09g0108600		394	Os10g0118800
360	Os09g0118666		395	Os10g0135100
361	Os09g0135100		396	Os10g0140700
362	Os09g0268100		397	Os10g0144900
363	Os09g0274900		398	Os10g0147400
364	Os09g0278000		399	Os10g0155900
365	Os09g0305900		400	Os10g0181200
366	Os09g0312400		401	Os10g0181600
367	Os09g0315000		402	Os10g0182200
368	Os09g0320400		403	Os10g0337700
369	Os09g0320600		404	Os10g0371700
370	Os09g0330000		405	Os10g0389100
371	Os09g0338400		406	Os10g0389200
372	Os09g0341100		407	Os10g0392400
373	Os09g0342751		408	Os10g0413500
374	Os09g0376400		409	Os10g0415300
375	Os09g0378700		410	Os10g0428900
376	Os09g0379900		411	Os10g0430600
377	Os09g0397300		412	Os10g0438100
378	Os09g0400700		413	Os10g0440500
379	Os09g0411675		414	Os10g0444850
380	Os09g0416600		415	Os10g0487100
381	Os09g0427125		416	Os10g0508700
382	Os09g0441050		417	Os10g0518200
383	Os09g0441700		418	Os10g0528900
384	Os09g0447000		419	Os10g0529300
385	Os09g0481800		420	Os10g0542400

421	Os10g0545000		456	Os11g0703900
422	Os10g0546400		457	Os12g0104400
423	Os10g0555900		458	Os12g0114100
424	Os11g0106900		459	Os12g0138200
425	Os11g0110200		460	Os12g0140600
426	Os11g0145200		461	Os12g0146625
427	Os11g0146200		462	Os12g0175000
428	Os11g0149200		463	Os12g0182900
429	Os11g0151800		464	Os12g0187800
430	Os11g0177400		465	Os12g0229000
431	Os11g0195600		466	Os12g0236250
432	Os11g0198900		467	Os12g0242500
433	Os11g0202350		468	Os12g0444500
434	Os11g0215600		469	Os12g0467700
435	Os11g0220900		470	Os12g0468000
436	Os11g0239400		471	Os12g0486700
437	Os11g0282100		472	Os12g0504900
438	Os11g0284400		473	Os12g0506200
439	Os11g0417400		474	Os12g0507800
440	Os11g0435100		475	Os12g0508266
441	Os11g0453900		476	Os12g0543701
442	Os11g0485000		477	Os12g0568700
443	Os11g0490300		478	Os12g0569900
444	Os11g0493200		479	Os12g0576600
445	Os11g0522900		480	Os12g0580300
446	Os11g0523200		481	Os12g0583000
447	Os11g0529500		482	Os12g0590500
448	Os11g0543100		483	Os12g0592700
449	Os11g0556600		484	Os12g0607800
450	Os11g0558100		485	Os12g0612300
451	Os11g0582300		486	Os12g0626901
452	Os11g0596633		487	Os12g0637300
453	Os11g0609880		488	LOC_Os01g03452
454	Os11g0641800		489	LOC_Os01g08240
455	Os11g0696400		490	LOC_Os01g10380

491	LOC_Os01g10650		526	LOC_Os02g13250
492	LOC_Os01g13500		527	LOC_Os02g15830
493	LOC_Os01g17980		528	LOC_Os02g18260
494	LOC_Os01g20080		529	LOC_Os02g19670
495	LOC_Os01g21080		530	LOC_Os02g21950
496	LOC_Os01g21280		531	LOC_Os02g22080
497	LOC_Os01g21542		532	LOC_Os02g22150
498	LOC_Os01g23800		533	LOC_Os02g22170
499	LOC_Os01g26190		534	LOC_Os02g22700
500	LOC_Os01g28060		535	LOC_Os02g22750
501	LOC_Os01g29130		536	LOC_Os02g24480
502	LOC_Os01g31340		537	LOC_Os02g24550
503	LOC_Os01g32400		538	LOC_Os02g24940
504	LOC_Os01g32430		539	LOC_Os02g25210
505	LOC_Os01g32900		540	LOC_Os02g25260
506	LOC_Os01g33900		541	LOC_Os02g25340
507	LOC_Os01g34290		542	LOC_Os02g28400
508	LOC_Os01g34440		543	LOC_Os02g29580
509	LOC_Os01g35380		544	LOC_Os02g35420
510	LOC_Os01g41070		545	LOC_Os02g36470
511	LOC_Os01g41690		546	LOC_Os02g38330
512	LOC_Os01g46110		547	LOC_Os02g38960
513	LOC_Os01g47240		548	LOC_Os02g42980
514	LOC_Os01g47319		549	LOC_Os02g44050
515	LOC_Os01g48150		550	LOC_Os02g44390
516	LOC_Os01g48560		551	LOC_Os02g44440
517	LOC_Os01g54260		552	LOC_Os02g45550
518	LOC_Os01g56440		553	LOC_Os02g46589
519	LOC_Os01g62680		554	LOC_Os02g46890
520	LOC_Os02g01840		555	LOC_Os02g48430
521	LOC_Os02g05860		556	LOC_Os02g50720
522	LOC_Os02g06350		557	LOC_Os02g54170
523	LOC_Os02g11730		558	LOC_Os02g54380
524	LOC_Os02g11880		559	LOC_Os03g11280
525	LOC_Os02g12330		560	LOC_Os03g13984

561	LOC_Os03g14110		596	LOC_Os04g07810
562	LOC_Os03g18670		597	LOC_Os04g08530
563	LOC_Os03g22090		598	LOC_Os04g08560
564	LOC_Os03g22980		599	LOC_Os04g08670
565	LOC_Os03g23080		600	LOC_Os04g11320
566	LOC_Os03g24230		601	LOC_Os04g14600
567	LOC_Os03g25304		602	LOC_Os04g14850
568	LOC_Os03g30810		603	LOC_Os04g16190
569	LOC_Os03g31810		604	LOC_Os04g17180
570	LOC_Os03g32460		605	LOC_Os04g18220
571	LOC_Os03g33710		606	LOC_Os04g18690
572	LOC_Os03g39200		607	LOC_Os04g19230
573	LOC_Os03g39790		608	LOC_Os04g19250
574	LOC_Os03g40630		609	LOC_Os04g19520
575	LOC_Os03g42360		610	LOC_Os04g19940
576	LOC_Os03g45690		611	LOC_Os04g20090
577	LOC_Os03g46830		612	LOC_Os04g20290
578	LOC_Os03g48520		613	LOC_Os04g21430
579	LOC_Os03g49840		614	LOC_Os04g21680
580	LOC_Os03g52350		615	LOC_Os04g22400
581	LOC_Os03g53419		616	LOC_Os04g24210
582	LOC_Os03g56080		617	LOC_Os04g25630
583	LOC_Os03g61249		618	LOC_Os04g28660
584	LOC_Os03g61319		619	LOC_Os04g29540
585	LOC_Os03g63050		620	LOC_Os04g31650
586	LOC_Os03g63820		621	LOC_Os04g31840
587	LOC_Os04g02220		622	LOC_Os04g32130
588	LOC_Os04g03480		623	LOC_Os04g34690
589	LOC_Os04g04109		624	LOC_Os04g36550
590	LOC_Os04g05380		625	LOC_Os04g37680
591	LOC_Os04g05920		626	LOC_Os04g41599
592	LOC_Os04g06429		627	LOC_Os04g43180
593	LOC_Os04g06460		628	LOC_Os04g46430
594	LOC_Os04g06860		629	LOC_Os04g46639
595	LOC_Os04g06950		630	LOC_Os04g48920

631	LOC_Os04g49770		666	LOC_Os05g26710
632	LOC_Os04g51230		667	LOC_Os05g27430
633	LOC_Os04g52152		668	LOC_Os05g27600
634	LOC_Os04g52220		669	LOC_Os05g27740
635	LOC_Os04g55820		670	LOC_Os05g28110
636	LOC_Os04g55870		671	LOC_Os05g31650
637	LOC_Os04g59270		672	LOC_Os05g35700
638	LOC_Os05g01160		673	LOC_Os05g37180
639	LOC_Os05g02270		674	LOC_Os05g38240
640	LOC_Os05g02910		675	LOC_Os05g39570
641	LOC_Os05g07150		676	LOC_Os05g41720
642	LOC_Os05g07320		677	LOC_Os05g44690
643	LOC_Os05g08200		678	LOC_Os05g45540
644	LOC_Os05g10260		679	LOC_Os05g48350
645	LOC_Os05g10290		680	LOC_Os05g49110
646	LOC_Os05g10700		681	LOC_Os06g02304
647	LOC_Os05g11000		682	LOC_Os06g04770
648	LOC_Os05g11800		683	LOC_Os06g17760
649	LOC_Os05g13550		684	LOC_Os06g18160
650	LOC_Os05g15440		685	LOC_Os06g18810
651	LOC_Os05g16220		686	LOC_Os06g22350
652	LOC_Os05g16510		687	LOC_Os06g22780
653	LOC_Os05g17390		688	LOC_Os06g23170
654	LOC_Os05g19930		689	LOC_Os06g24330
655	LOC_Os05g20110		690	LOC_Os06g29464
656	LOC_Os05g20330		691	LOC_Os06g29560
657	LOC_Os05g20360		692	LOC_Os06g29820
658	LOC_Os05g20830		693	LOC_Os06g30050
659	LOC_Os05g22550		694	LOC_Os06g30350
660	LOC_Os05g22750		695	LOC_Os06g31170
661	LOC_Os05g23280		696	LOC_Os06g31920
662	LOC_Os05g23290		697	LOC_Os06g32550
663	LOC_Os05g24720		698	LOC_Os06g32580
664	LOC_Os05g24740		699	LOC_Os06g35100
665	LOC_Os05g25250		700	LOC_Os06g35110

701	LOC_Os06g35610		736	LOC_Os07g25090
702	LOC_Os06g36020		737	LOC_Os07g25484
703	LOC_Os06g36620		738	LOC_Os07g25660
704	LOC_Os06g40660		739	LOC_Os07g26120
705	LOC_Os06g41520		740	LOC_Os07g26280
706	LOC_Os06g42010		741	LOC_Os07g27570
707	LOC_Os06g45700		742	LOC_Os07g27630
708	LOC_Os07g01680		743	LOC_Os07g28190
709	LOC_Os07g04470		744	LOC_Os07g31220
710	LOC_Os07g04770		745	LOC_Os07g31570
711	LOC_Os07g10010		746	LOC_Os07g32640
712	LOC_Os07g11200		747	LOC_Os07g37590
713	LOC_Os07g11690		748	LOC_Os07g38500
714	LOC_Os07g12090		749	LOC_Os07g45039
715	LOC_Os07g12500		750	LOC_Os07g46050
716	LOC_Os07g13660		751	LOC_Os07g47060
717	LOC_Os07g15170		752	LOC_Os08g01540
718	LOC_Os07g15300		753	LOC_Os08g06750
719	LOC_Os07g16370		754	LOC_Os08g08490
720	LOC_Os07g16470		755	LOC_Os08g09540
721	LOC_Os07g17670		756	LOC_Os08g11020
722	LOC_Os07g18060		757	LOC_Os08g11350
723	LOC_Os07g18390		758	LOC_Os08g11630
724	LOC_Os07g18730		759	LOC_Os08g11880
725	LOC_Os07g19170		760	LOC_Os08g12140
726	LOC_Os07g20380		761	LOC_Os08g12590
727	LOC_Os07g20710		762	LOC_Os08g13650
728	LOC_Os07g20750		763	LOC_Os08g13770
729	LOC_Os07g20860		764	LOC_Os08g15860
730	LOC_Os07g22080		765	LOC_Os08g17220
731	LOC_Os07g22440		766	LOC_Os08g17350
732	LOC_Os07g24690		767	LOC_Os08g19124
733	LOC_Os07g24870		768	LOC_Os08g19390
734	LOC_Os07g24950		769	LOC_Os08g20280
735	LOC_Os07g25040		770	LOC_Os08g22030

771	LOC_Os08g24070		806	LOC_Os10g06240
772	LOC_Os08g25950		807	LOC_Os10g07300
773	LOC_Os08g26300		808	LOC_Os10g08080
774	LOC_Os08g26780		809	LOC_Os10g09590
775	LOC_Os08g28590		810	LOC_Os10g10090
776	LOC_Os08g29640		811	LOC_Os10g11420
777	LOC_Os08g34670		812	LOC_Os10g12320
778	LOC_Os08g36880		813	LOC_Os10g17240
779	LOC_Os08g37420		814	LOC_Os10g17370
780	LOC_Os09g02970		815	LOC_Os10g17470
781	LOC_Os09g06610		816	LOC_Os10g17510
782	LOC_Os09g07600		817	LOC_Os10g18250
783	LOC_Os09g07709		818	LOC_Os10g19150
784	LOC_Os09g07840		819	LOC_Os10g19310
785	LOC_Os09g08170		820	LOC_Os10g20300
786	LOC_Os09g08510		821	LOC_Os10g20580
787	LOC_Os09g12410		822	LOC_Os10g22920
788	LOC_Os09g12860		823	LOC_Os10g24190
789	LOC_Os09g14120		824	LOC_Os10g24820
790	LOC_Os09g14940		825	LOC_Os10g25789
791	LOC_Os09g15000		826	LOC_Os10g26040
792	LOC_Os09g15650		827	LOC_Os10g29230
793	LOC_Os09g15900		828	LOC_Os10g29630
794	LOC_Os09g16310		829	LOC_Os10g30740
795	LOC_Os09g17820		830	LOC_Os10g31750
796	LOC_Os09g20160		831	LOC_Os10g35880
797	LOC_Os09g21490		832	LOC_Os10g42600
798	LOC_Os09g25450		833	LOC_Os11g01109
799	LOC_Os09g26270		834	LOC_Os11g02030
800	LOC_Os09g27630		835	LOC_Os11g02340
801	LOC_Os09g32390		836	LOC_Os11g02940
802	LOC_Os09g40022		837	LOC_Os11g03010
803	LOC_Os09g40075		838	LOC_Os11g03020
804	LOC_Os10g01900		839	LOC_Os11g08810
805	LOC_Os10g04790		840	LOC_Os11g08840

841	LOC_Os11g09120		876	LOC_Os12g02780
842	LOC_Os11g09190		877	LOC_Os12g05030
843	LOC_Os11g10080		878	LOC_Os12g06690
844	LOC_Os11g11460		879	LOC_Os12g06860
845	LOC_Os11g12830		880	LOC_Os12g09380
846	LOC_Os11g13480		881	LOC_Os12g09680
847	LOC_Os11g14690		882	LOC_Os12g10300
848	LOC_Os11g14710		883	LOC_Os12g11230
849	LOC_Os11g15660		884	LOC_Os12g13430
850	LOC_Os11g16490		885	LOC_Os12g16020
851	LOC_Os11g16650		886	LOC_Os12g17650
852	LOC_Os11g16750		887	LOC_Os12g22080
853	LOC_Os11g17460		888	LOC_Os12g22360
854	LOC_Os11g17700		889	LOC_Os12g22410
855	LOC_Os11g18410		890	LOC_Os12g22480
856	LOC_Os11g19380		891	LOC_Os12g24210
857	LOC_Os11g19550		892	LOC_Os12g24400
858	LOC_Os11g19640		893	LOC_Os12g24555
859	LOC_Os11g20530		894	LOC_Os12g24740
860	LOC_Os11g22190		895	LOC_Os12g26020
861	LOC_Os11g22250		896	LOC_Os12g29640
862	LOC_Os11g23940		897	LOC_Os12g29900
863	LOC_Os11g24680		898	LOC_Os12g30300
864	LOC_Os11g24920		899	LOC_Os12g31080
865	LOC_Os11g25350		900	LOC_Os12g34700
866	LOC_Os11g25690		901	LOC_Os12g38510
867	LOC_Os11g29950		902	LOC_Os12g40250
868	LOC_Os11g30670			
869	LOC_Os11g35610			
870	LOC_Os11g43260			
871	LOC_Os11g43650			
872	LOC_Os11g44140			
873	LOC_Os11g45490			
874	LOC_Os11g45650			
875	LOC_Os11g47494			

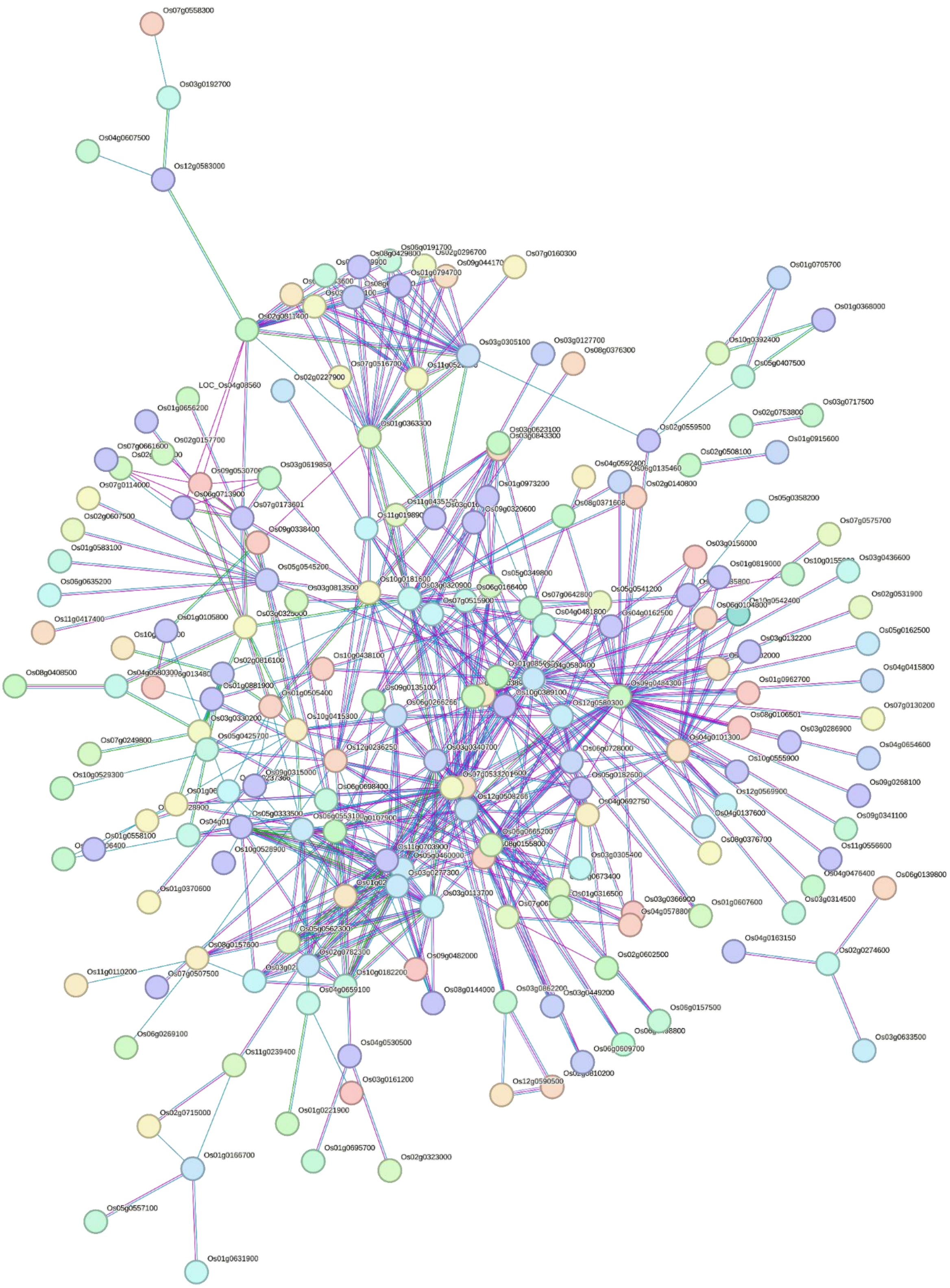

Figure 6. STRING network for seedling/shoot genes depicting interactions between proteins. The color of the proteins is related to their query names. Edges are categorized according to their color: a) blue edges are confirmed interactions by curated databases, b) purple edges are confirmed interactions experimentally determined, c) green edges are predicted interactions by gene neighborhood.

Table 5. Table with nodes-proteins, that display the highest centrality.

#node	UniprotKB Entry Identifier	Node Degree
Os09g0484300	A0A0N7KR07	44
Os04g0580400	Q0JAS6	37
Os03g0320900	Q10M74	29
Os04g0107900	A0A0P0W604	21
Os06g0714600	Q5Z9Q7	19
Os07g0533201	A0A0P0X6W3	19
Os01g0363300	Q9ARX2	17
Os12g0236250	A0A0P0Y8F5	17
Os02g0811400	Q6K5W5	15
Os03g0277300	Q10NA1	15
Os05g0460000	Q6L509	15
Os11g0703900	Q53NM9	15
Os01g0226400	Q5NAF6	14
Os03g0113700	Q10SR3	14
Os03g0305100	Q10MK9	14
Os04g0101300	A0A0N7KIF9	14
Os03g0340700	Q10LP1	13
Os06g0665200	A0A0P0WZM4	13
Os08g0155800	Q6ZDA7	13
Os10g0415300	A0A0P0XUU0	12
Os11g0529500	Q2R3A8	12
Os12g0508266	B9GDC7	12
Os12g0580300	Q2QN41	12
Os05g0182600	Q65WY8	11
Os05g0545200	A0A0N7KL66	11
Os06g0728000	Q5Z7N3	11
Os06g0553100	Q0DBL6	10

Table 6. Table with proteins that had high centrality, indicating their biological processes, molecular functions, and cellular components.

#node	Biological Process	Molecular Function	Cellular Component
Os09g0484300	protein polyubiquitination	ubiquitin protein ligase activity	cytoplasm
Os04g0580400	protein sumoylation	SUMO conjugating enzyme activity	nucleus
Os03g0320900	phosphorylation/regulation of developmental growth	ATP binding/guanylate kinase activity	mitochondrion/cytosol/chloroplast
Os04g0107900	none	ATP binding/ATP hydrolysis activity/ATP-dependent protein folding chaperone/unfolded protein binding	none
Os06g0714600	none	GTP binding/GTPase activity	none
Os07g0533201	regulation of developmental process	DNA binding/DNA-binding transcription factor activity	nucleus
Os01g0363300	none	none	plasma membrane
Os12g0236250	ubiquitin-dependent protein catabolic process	ubiquitin protein ligase binding	none
Os02g0811400	lignin biosynthetic process	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	none
Os03g0277300	protein refolding/chaperone cofactor-dependent protein folding	ATP binding/ATP hydrolysis activity/ATP-dependent protein folding chaperone/heat shock protein binding/protein folding chaperone	cytoplasm
Os05g0460000	protein refolding/chaperone cofactor-dependent protein folding	ATP binding/ATP hydrolysis activity/ATP-dependent protein folding chaperone/heat shock protein binding/protein folding chaperone	cytoplasm
Os11g0703900	protein refolding/chaperone cofactor-dependent protein folding	ATP binding/ATP hydrolysis activity/ATP-dependent protein folding chaperone/heat shock protein binding/protein folding chaperone	cytoplasm
Os01g0226400	none	ATP binding/ATP hydrolysis activity	mitochondrial outer membrane
Os03g0113700	protein refolding/chaperone cofactor-dependent protein folding	ATP binding/ATP hydrolysis activity/ATP-dependent protein folding chaperone/heat shock protein binding/protein folding chaperone/unfolder protein bidning	cytoplasm/mitochondrion
Os03g0305100	phenylpropanoid metabolic process	4-coumarate-CoA ligase activity/trans-cinnamate-CoA ligase activity	none

Os04g0101300	chromatin organization/histone acetylation/regulation of DNA-templated transcription	none	histone acetyltransferase complex/membrane/nucleus
Os03g0340700	none	none	intracellular organelle
Os06g0665200	none	none	plasmodesma
Os08g0155800	none	none	none
Os10g0415300	cell redox homeostasis	flavin adenine dinucleotide binding/thioredoxin-disulfide reductase activity	cytoplasm
Os11g0529500	flavonoid biosynthetic process/polyketide biosynthetic process	acyltransferase activity, transferring groups other than amino-acyl groups	none
Os12g0508266	none	zinc ion binding	none
Os12g0580300	DNA-templated transcription initiation	DNA binding/RNA polymerase II general transcription initiation factor activity	nucleus
Os05g0182600	DNA repair/DNA replication	DNA binding/histone binding/nucleosome binding	FACT complex
Os05g0545200	none	metal ion binding	none
Os06g0728000	none	DNA binding	nucleus
Os06g0553100	cellular response to heat/regulation of transcription by RNA polymerase	DNA-binding transcription factor activity/RNA polymerase II cis-regulatory region sequence-specific DNA binding	nucleus

Table 7. STRING annotated biological process terms for DEGs in seedling shoot tissue.

#term ID	Term Description	Observed Gene Count	Background Gene Count	Strength	False Discovery Rate
GO:0042221	Response to chemical	73	2969	0.35	0.000000633
GO:0050896	Response to stimulus	122	6863	0.21	0.0000642
GO:1901700	Response to oxygen-containing compound	39	1362	0.42	0.00015
GO:0009628	Response to abiotic stimulus	40	1604	0.36	0.0023
GO:0010035	Response to inorganic substance	20	554	0.52	0.0052
GO:0010033	Response to organic substance	42	1849	0.32	0.0077
GO:0097305	Response to alcohol	16	408	0.55	0.0121
GO:0009737	Response to abscisic acid	14	326	0.59	0.0135
GO:0009738	Abscisic acid-activated signaling pathway	9	135	0.78	0.0144
GO:1902584	Positive regulation of response to water deprivation	5	29	1.2	0.0144
GO:0009414	Response to water deprivation	14	343	0.57	0.0162
GO:0070887	Cellular response to chemical stimulus	39	1784	0.3	0.0167
GO:0033993	Response to lipid	21	709	0.43	0.0173
GO:0010029	Regulation of seed germination	5	35	1.12	0.018
GO:2000026	Regulation of multicellular organismal development	12	275	0.6	0.0195
GO:0071310	Cellular response to organic substance	29	1203	0.34	0.0218
GO:0051085	Chaperone cofactor-dependent protein refolding	6	63	0.94	0.0229
GO:0009266	Response to temperature stimulus	17	531	0.47	0.0232
GO:0006950	Response to stress	67	3911	0.19	0.0319
GO:0048580	Regulation of post-embryonic development	11	259	0.59	0.0319
GO:0006457	Protein folding	12	310	0.55	0.0352
GO:0009725	Response to hormone	32	1464	0.3	0.0352
GO:0035970	Peptidyl-threonine dephosphorylation	5	49	0.97	0.0419
GO:0009408	Response to heat	11	278	0.56	0.0481

Table 8. STRING annotated molecular function term for DEGs in seedling shoot tissue.

#term ID	Term Description	Observed Gene Count	Background Gene Count	Strength	False Discovery Rate
GO:0051082	Unfolded protein binding	12	190	0.76	0.0053

Table 9. STRING annotated cellular component term for DEGs in seedling shoot tissue.

#term ID	Term Description	Observed Gene Count	Background Gene Count	Strength	False Discovery Rate
GO:0110165	Cellular anatomical entity	348	26859	0.07	0.00013

Table 10. g:Profiler annotated terms for DEGs in seedling shoot tissue.

Source	Term name	Term Id	Adjusted P Value	Term Size	Query Size
GO:BP	response to organic substance	GO:0010033	3.11E-06	1002	211
GO:BP	response to oxygen-containing compound	GO:1901700	0.000214	648	211
GO:BP	response to heat	GO:0009408	0.000214	152	211
GO:BP	response to inorganic substance	GO:0010035	0.000244	294	211
GO:BP	response to chemical	GO:0042221	0.000257	1487	211
GO:BP	response to temperature stimulus	GO:0009266	0.001537	273	211
GO:BP	response to osmotic stress	GO:0006970	0.001864	176	211
GO:BP	response to water deprivation	GO:0009414	0.001864	116	211
GO:BP	response to water	GO:0009415	0.001902	118	211
GO:BP	response to abiotic stimulus	GO:0009628	0.002115	800	211
GO:BP	response to salt stress	GO:0009651	0.002115	152	211
GO:BP	response to acid chemical	GO:0001101	0.002115	123	211
GO:BP	'De novo' post-translational protein folding	GO:0051084	0.002571	51	211
GO:BP	chaperone cofactor-dependent protein refolding	GO:0051085	0.002571	51	211
GO:BP	response to desiccation	GO:0009269	0.002571	6	211
GO:BP	response to hydrogen peroxide	GO:0042542	0.004153	35	211
GO:BP	cellular response to unfolded protein	GO:0034620	0.004929	37	211
GO:BP	cellular response to topologically incorrect protein	GO:0035967	0.004929	59	211
GO:BP	response to unfolded protein	GO:0006986	0.005243	38	211
GO:BP	response to topologically incorrect protein	GO:0035966	0.005353	61	211
GO:BP	cellular response to heat	GO:0034605	0.006943	65	211
GO:BP	'de novo' protein folding	GO:0006458	0.006943	65	211
GO:BP	protein folding	GO:0006457	0.009351	239	211
GO:BP	response to salt	GO:1902074	0.011644	170	211
GO:BP	response to stress	GO:0006950	0.011644	2256	211
GO:BP	chaperone-mediated protein folding	GO:0061077	0.013858	76	211
GO:BP	positive regulation of response to water deprivation	GO:1902584	0.017391	13	211
GO:BP	protein complex oligomerization	GO:0051259	0.019746	32	211
GO:BP	iron-sulfur cluster assembly	GO:0016226	0.019746	32	211
GO:BP	response to lipid	GO:0033993	0.019746	359	211
GO:BP	metallo-sulfur cluster assembly	GO:0031163	0.019746	32	211
GO:BP	protein refolding	GO:0042026	0.02705	60	211
GO:BP	response to hormone	GO:0009725	0.028765	734	211
GO:BP	regulation of response to water deprivation	GO:2000070	0.031633	17	211
GO:BP	response to endogenous stimulus	GO:0009719	0.031924	744	211
GO:BP	positive regulation of abscisic acid-activated signaling pathway	GO:0009789	0.038999	5	211
GO:BP	cellular response to organic substance	GO:0071310	0.043316	604	211
GO:MF	unfolded protein binding	GO:0051082	0.001566	153	237

Table 11. PANTHER annotated biological process terms for DEGs in seedling shoot tissue.

#term ID	GO biological process complete	false discovery rate
GO:0051085	chaperone cofactor-dependent protein refolding	4.08E-02
GO:0051084	de novo' post-translational protein folding	2.72E-02
GO:0006458	de novo' protein folding	2.76E-02
GO:0010035	response to inorganic substance	2.56E-02
GO:0006457	protein folding	2.24E-02
GO:1901700	response to oxygen-containing compound	5.26E-02
GO:0010033	response to organic substance	2.32E-02

Table 12. PANTHER annotated molecular function terms for DEGs in seedling shoot tissue.

#term ID	GO molecular function complete	false discovery rate
GO:0051082	unfolded protein binding	4.88E-02

Table 13. Statistically significant differentially expressed root genes for FDR<0.0005 and |D|>0.5.

1	Os01g0277900		36	Os05g0508800
2	Os01g0155300		37	Os05g0546800
3	Os01g0182832		38	Os05g0548500
4	Os01g0201600		39	Os06g0216550
5	Os01g0299150		40	Os06g0506100
6	Os01g0531300		41	Os06g0534700
7	Os01g0540400		42	Os06g0669075
8	Os01g0577600		43	Os07g0409500
9	Os01g0622000		44	Os07g0414000
10	Os01g0759000		45	Os07g0422100
11	Os01g0759100		46	Os07g0441300
12	Os01g0789200		47	Os07g0513400
13	Os02g0113300		48	Os07g0595400
14	Os02g0147500		49	Os07g0609700
15	Os02g0163500		50	Os07g0664400
16	Os02g0202000		51	Os08g0141000
17	Os02g0202250		52	Os08g0302000
18	Os02g0219800		53	Os08g0367625
19	Os02g0470400		54	Os09g0354600
20	Os02g0496400		55	Os09g0423800
21	Os02g0608550		56	Os10g0435600
22	Os02g0805400		57	Os10g0538500
23	Os02g0832400		58	Os10g0580100
24	Os03g0281700		59	Os11g0264200
25	Os03g0392400		60	Os11g0300700
26	Os03g0665000		61	Os11g0440200
27	Os03g0703700		62	Os11g0549000
28	Os04g0135400		63	Os11g0578200
29	Os04g0278100		64	Os11g0582801
30	Os04g0330200		65	Os11g0586600
31	Os04g0488400		66	Os11g0618400
32	Os04g0588000		67	Os11g0681100
33	Os05g0190300		68	Os11g0686250
34	Os05g0324500		69	Os12g0132800
35	Os05g0463500		70	Os12g0501133

71	Os12g0516000		106	LOC_Os02g39110
72	Os12g0534700		107	LOC_Os02g39900
73	Os12g0538066		108	LOC_Os02g53930
74	Os12g0599700		109	LOC_Os02g57920
75	Os12g0599800		110	LOC_Os03g23830
76	LOC_Os01g02230		111	LOC_Os03g34190
77	LOC_Os01g03240		112	LOC_Os03g37377
78	LOC_Os01g03780		113	LOC_Os03g51730
79	LOC_Os01g07830		114	LOC_Os03g55000
80	LOC_Os01g17460		115	LOC_Os04g05600
81	LOC_Os01g18390		116	LOC_Os04g07700
82	LOC_Os01g18570		117	LOC_Os04g08210
83	LOC_Os01g23000		118	LOC_Os04g10190
84	LOC_Os01g24540		119	LOC_Os04g12350
85	LOC_Os01g27910		120	LOC_Os04g13610
86	LOC_Os01g38070		121	LOC_Os04g13700
87	LOC_Os01g38550		122	LOC_Os04g16712
88	LOC_Os01g38820		123	LOC_Os04g16824
89	LOC_Os01g46100		124	LOC_Os04g18110
90	LOC_Os01g47700		125	LOC_Os04g23470
91	LOC_Os01g60090		126	LOC_Os04g27220
92	LOC_Os01g64610		127	LOC_Os04g37860
93	LOC_Os02g04220		128	LOC_Os04g44970
94	LOC_Os02g07540		129	LOC_Os04g48320
95	LOC_Os02g14330		130	LOC_Os04g57980
96	LOC_Os02g14350		131	LOC_Os05g04290
97	LOC_Os02g15490		132	LOC_Os05g05050
98	LOC_Os02g19410		133	LOC_Os05g05880
99	LOC_Os02g22570		134	LOC_Os05g08190
100	LOC_Os02g24634		135	LOC_Os05g09040
101	LOC_Os02g26930		136	LOC_Os05g12550
102	LOC_Os02g28300		137	LOC_Os05g13730
103	LOC_Os02g30680		138	LOC_Os05g15130
104	LOC_Os02g35270		139	LOC_Os05g15490
105	LOC_Os02g37350		140	LOC_Os05g15990

141	LOC_Os05g20390		176	LOC_Os07g30540
142	LOC_Os05g20450		177	LOC_Os07g35170
143	LOC_Os05g23770		178	LOC_Os07g45660
144	LOC_Os05g27210		179	LOC_Os07g45830
145	LOC_Os05g27270		180	LOC_Os08g02790
146	LOC_Os05g27470		181	LOC_Os08g05180
147	LOC_Os05g27600		182	LOC_Os08g05990
148	LOC_Os05g33490		183	LOC_Os08g11430
149	LOC_Os05g40870		184	LOC_Os08g13600
150	LOC_Os05g43730		185	LOC_Os08g13650
151	LOC_Os05g46680		186	LOC_Os08g15236
152	LOC_Os05g50520		187	LOC_Os08g15288
153	LOC_Os06g07370		188	LOC_Os08g20486
154	LOC_Os06g11880		189	LOC_Os08g23330
155	LOC_Os06g12890		190	LOC_Os08g23900
156	LOC_Os06g14940		191	LOC_Os08g26080
157	LOC_Os06g15959		192	LOC_Os08g26280
158	LOC_Os06g16520		193	LOC_Os08g32820
159	LOC_Os06g16690		194	LOC_Os08g40480
160	LOC_Os06g16840		195	LOC_Os08g43310
161	LOC_Os06g16870		196	LOC_Os09g03050
162	LOC_Os06g19900		197	LOC_Os09g04950
163	LOC_Os06g24350		198	LOC_Os09g07180
164	LOC_Os06g24810		199	LOC_Os09g08630
165	LOC_Os06g30150		200	LOC_Os09g09860
166	LOC_Os06g33240		201	LOC_Os09g12140
167	LOC_Os06g33890		202	LOC_Os09g13340
168	LOC_Os06g34510		203	LOC_Os09g13954
169	LOC_Os06g34550		204	LOC_Os09g14270
170	LOC_Os06g35710		205	LOC_Os09g15920
171	LOC_Os06g38609		206	LOC_Os09g19300
172	LOC_Os06g48090		207	LOC_Os09g19990
173	LOC_Os07g04770		208	LOC_Os09g21610
174	LOC_Os07g18080		209	LOC_Os09g21990
175	LOC_Os07g28860		210	LOC_Os09g26640

211	LOC_Os09g32560		246	LOC_Os12g28970
212	LOC_Os09g40018		247	LOC_Os12g30420
213	LOC_Os10g04630		248	LOC_Os12g31540
214	LOC_Os10g12000		249	LOC_Os12g33870
215	LOC_Os10g16740		250	LOC_Os12g36480
216	LOC_Os10g19130		251	LOC_Os12g39740
217	LOC_Os10g19980			
218	LOC_Os10g21192			
219	LOC_Os10g21212			
220	LOC_Os10g21236			
221	LOC_Os10g21312			
222	LOC_Os10g22550			
223	LOC_Os10g24290			
224	LOC_Os10g31680			
225	LOC_Os10g33220			
226	LOC_Os10g40806			
227	LOC_Os11g08850			
228	LOC_Os11g15290			
229	LOC_Os11g17690			
230	LOC_Os11g22310			
231	LOC_Os11g24090			
232	LOC_Os11g36370			
233	LOC_Os11g41750			
234	LOC_Os11g43430			
235	LOC_Os11g43660			
236	LOC_Os11g46830			
237	LOC_Os11g47430			
238	LOC_Os12g03000			
239	LOC_Os12g12830			
240	LOC_Os12g14000			
241	LOC_Os12g14870			
242	LOC_Os12g16100			
243	LOC_Os12g23910			
244	LOC_Os12g24310			
245	LOC_Os12g26250			

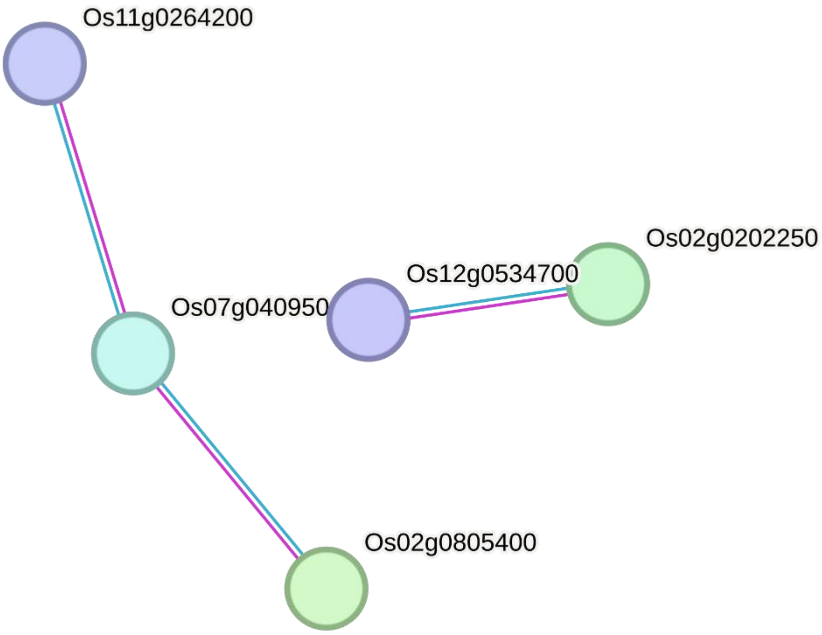

Figure 7. STRING network for root genes depicting interactions between proteins. The color of the proteins is related to their query names. Edges are categorized according to their color: a) blue edges are confirmed interactions by curated databases, b) purple edges are confirmed interactions experimentally determined.

3.5 Enrichment analysis for root tissue

The gene list for the root tissue that was used for the enrichment analysis tools consists of 251 genes, for FDR<0.0001 and -0.5<SMD<0.5 (Table 13). A STRING [23] network was created with a medium interaction score of 0.4 including 74 DEGs from the initial 251, but only a few of them appeared to interact with each other and contains interactions verified from experiment, and database sources (Figure 7). Purple edges refer to protein interactions that are experimentally determined, whilst blue edges refer to protein interactions from curated databases. However, the enrichment analysis didn't result in any enriched pathways. Subsequently, after using the g:Profiler [24] and PANTHER [22] tools , none statistically significant results were acquired.

Chapter 4 – Discussion

For this thesis, it was initially required to obtain a large number of studies that were collected from the GEO [26] database associated with salt stress experiments. Of the 410 studies, nine were further studied after following the PRISMA [27] and went through preprocessing in order to get them ready for the meta-analysis. The nine studies were divided into the root group, including five of them, and the seedling/shoot group, including the rest of the four. After encountering different metrics such as RPKM, FPKM, and TPM, it was eventually decided to use the TPM unit due to its computational advantage compared to the rest. The different formats of gene identifiers should also be converted to the same type across all the studies. The RAP-DB [28] provided valuable documentation about the rice genome as well as a converter tool for the gene identifiers.

Eventually, the meta-analysis was conducted using the MAGE [31] tool for various FDR boundaries. After gathering all the resulting numbers of DEGs for each FDR, the most strict threshold of FDR = 0.0001 was selected for enrichment analysis in order to limit the number of DEGs that were going to be integrated into the databases of PANTHER [22], STRING [23], and g:Profiler [24]. With the completion of the enrichment analysis that took place after the meta-analysis process, it was assumed that the stress response to salinity in rice is more prevalent in the seedling/shoot tissue compared to the root tissue. As a result, no significant biological pathways were detected associated with the root tissue, leaving room for further investigation. In contrast, genes from the seedling/shoot tissue were able to be recognized for their involvement in chemical signaling, protein folding, and large protein complexes, and those regulating the development of multicellular organisms.

Hopefully, this study will provide valuable insight for future research and a better understanding of the signaling mechanisms during salinity-stress conditions in rice. The research on a salt-tolerant rice cultivar proves to be necessary, considering climate change, the greenhouse effect, and other environmental factors that can put its production at stake. The results of this study about response to chemical signaling confirms that this approach was correct and the indication of the enriched pathways highlights the statistically significant genes that are involved in the process of salt tolerance. Progressively this can pave the way for the establishment of salt tolerant genotypes.

References

[1] J. P. Sahoo and V. Sharma, “Impact of LOD score and recombination frequencies on the microsatellite marker based linkage map for drought tolerance in kharif rice of Assam,” *Int. J. Curr. Microbiol. App. Sci*, vol. 7, no. 8, pp. 3299–3304, 2018.

[2] D. H. Seo, S. Seomun, Y. Do Choi, and G. Jang, “Root development and stress tolerance in rice: the key to improving stress tolerance without yield penalties,” *Int J Mol Sci*, vol. 21, no. 5, p. 1807, 2020.

[3] G. List and O. T. Coomes, “Natural hazards and risk in rice cultivation along the upper Amazon River,” *Natural Hazards*, vol. 87, pp. 165–184, 2017.

[4] M. S. de Miranda, M. L. Fonseca, A. Lima, T. F. de Moraes, and F. A. Rodrigues, “Environmental impacts of rice cultivation,” *Am J Plant Sci*, vol. 6, no. 12, p. 2009, 2015.

[5] R. Shankar, A. Bhattacharjee, and M. Jain, “Transcriptome analysis in different rice cultivars provides novel insights into desiccation and salinity stress responses,” *Sci Rep*, vol. 6, no. 1, p. 23719, 2016.

[6] S. Hussain *et al.*, “Effects of salt stress on rice growth, development characteristics, and the regulating ways: A review,” *J Integr Agric*, vol. 16, no. 11, pp. 2357–2374, 2017.

[7] Y.-F. Li *et al.*, “Comparative transcriptome and translatome analysis in contrasting rice genotypes reveals differential mRNA translation in salt-tolerant Pokkali under salt stress,” *BMC Genomics*, vol. 19, pp. 95–113, 2018.

[8] R. Chen *et al.*, “Whole genome sequencing and comparative transcriptome analysis of a novel seawater adapted, salt-resistant rice cultivar–sea rice 86,” *BMC Genomics*, vol. 18, pp. 1–11, 2017.

[9] D. Zheng *et al.*, “Salt-responsive genes are differentially regulated at the chromatin levels between seedlings and roots in rice,” *Plant Cell Physiol*, vol. 60, no. 8, pp. 1790–1803, 2019.

[10] N. Ghosh, M. K. Adak, P. D. Ghosh, S. Gupta, D. N. Sen Gupta, and C. Mandal, “Differential responses of two rice varieties to salt stress,” *Plant Biotechnol Rep*, vol. 5, pp. 89–103, 2011.

[11] A. Gondane and H. M. Itkonen, “Revealing the history and mystery of RNA-seq,” *Curr Issues Mol Biol*, vol. 45, no. 3, pp. 1860–1874, 2023.

[12] J. Pevsner, *Bioinformatics and functional genomics*. John Wiley & Sons, 2015.

[13] R. Leinonen *et al.*, “The European nucleotide archive,” *Nucleic Acids Res*, vol. 39, no. suppl_1, pp. D28–D31, 2010.

[14] C. L. Schoch *et al.*, “NCBI Taxonomy: a comprehensive update on curation, resources and tools,” *Database*, vol. 2020, p. baaa062, 2020.

[15] R. Leinonen, H. Sugawara, M. Shumway, and I. N. S. D. Collaboration, “The sequence read archive,” *Nucleic Acids Res*, vol. 39, no. suppl_1, pp. D19–D21, 2010.

[16] D. B. Emmert, P. J. Stoehr, G. Stoesser, and G. N. Cameron, “The European bioinformatics institute (EBI) databases,” *Nucleic Acids Res*, vol. 22, no. 17, pp. 3445–3449, 1994.

[17] N. de Souza, “The ENCODE project,” *Nat Methods*, vol. 9, no. 11, p. 1046, 2012.

[18] Ε. Πατελάρου and Η. Μπροκαλάκη, “Μεθοδολογία της συστηματικής ανασκόπησης και μετα-ανάλυσης,” *Νοσηλευτική*, vol. 49, no. 2, pp. 122–130, 2010.

[19] P. Galanis, “Systematic review and meta-analysis,” *Arch. Greek Med*, vol. 26, pp. 826–841, 2009.

[20] K. A. L’ABBÉ, A. S. Detsky, and K. O’ROURKE, “Meta-analysis in clinical research,” *Ann Intern Med*, vol. 107, no. 2, pp. 224–233, 1987.

[21] E. I. Boyle *et al.*, “GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes,” *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[22] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, “Large-scale gene function analysis with the PANTHER classification system,” *Nat Protoc*, vol. 8, no. 8, pp. 1551–1566, 2013.

[23] D. Szklarczyk *et al.*, “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest,” *Nucleic Acids Res*, vol. 51, no. D1, pp. D638–D646, 2023.

[24] U. Raudvere *et al.*, “g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update),” *Nucleic Acids Res*, vol. 47, no. W1, pp. W191–W198, 2019.

[25] E. Birney *et al.*, “An overview of Ensembl,” *Genome Res*, vol. 14, no. 5, pp. 925–928, 2004.

[26] R. Edgar, M. Domrachev, and A. E. Lash, “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository,” *Nucleic Acids Res*, vol. 30, no. 1, pp. 207–210, 2002.

[27] M. J. Page *et al.*, “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *International journal of surgery*, vol. 88, p. 105906, 2021.

[28] R. A. Project, “The rice annotation project database (RAP-DB): 2008 update,” *Nucleic Acids Res*, vol. 36, no. suppl_1, pp. D1028–D1033, 2007.

[29] S. Zhao, Z. Ye, and R. Stanton, “Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols,” *Rna*, vol. 26, no. 8, pp. 903–909, 2020.

[30] R. G. Gutierrez, “Stata,” *Wiley Interdiscip Rev Comput Stat*, vol. 2, no. 6, pp. 728–733, 2010.

[31] I. A. Tamposis, G. A. Manios, T. Charitou, K. E. Vennou, P. I. Kontou, and P. G. Bagos, “MAGE: An Open-Source Tool for Meta-Analysis of Gene Expression Studies,” *Biology (Basel)*, vol. 11, no. 6, p. 895, 2022.

[32] P. I. Kontou, A. Pavlopoulou, and P. G. Bagos, “Methods of analysis and meta-analysis for identifying differentially expressed genes,” *Genetic Epidemiology: Methods and Protocols*, pp. 183–210, 2018.

[33] D. Enzmann, “Notes on effect size measures for the difference of means from two independent groups: The case of Cohen’sd and Hedges’g,” *January*, vol. 12, no. 2015, 2015.

[34] C. H. Schmid and K. Mengersen, “Bayesian meta-analysis,” *The handbook of meta-analysis in ecology and evolution*, pp. 145–173, 2013.

[35] A. J. Sutton and K. R. Abrams, “Bayesian methods in meta-analysis and evidence synthesis,” *Stat Methods Med Res*, vol. 10, no. 4, pp. 277–303, 2001.

[36] S. Pounds and C. Cheng, “Improving false discovery rate estimation,” *Bioinformatics*, vol. 20, no. 11, pp. 1737–1745, 2004.

[37] J. P. T. Higgins and S. G. Thompson, “Quantifying heterogeneity in a meta-analysis,” *Stat Med*, vol. 21, no. 11, pp. 1539–1558, 2002.