Identification of Single-Nucleotide Polymorphisms related to the phenotypic expression of drought tolerance in *Oryza sativa*

Reva Kumar
Teacher/Mentor: Mr. Robert Gotwals
Research in Computational Science
North Carolina School of Science and Math
18 April 2024

Abstract

Environmental stressors have contributed to the depletion of crop yield for major crops like rice due to climate change. Some characteristics, such as drought tolerance, are critical indicators of stress. I identified genes associated with drought tolerance using several quantitative techniques. These include quantitative trait loci (QTL) analyses and genome-wide association studies (GWAS). The QTL analyses used a cross between *Curinga* and *O. rufipogon*. The GWAS study looked at a dataset of 413 rice varieties and approximately 44,000 single nucleotide polymorphisms (SNPs) on each of the species. The dataset contained 37 phenotypes. The goal was to use a kinship matrix as a correction for a GWAS analysis on a specified phenotype and to identify the most significant SNPs. I selected four major phenotypes possibly correlated with drought tolerance: 1. plant height; 2. seed length to width ratio; 3. amylose content; 4. protein content. The analyses identified several genes local to the SNP region using a genome browser. Current work involves relating phenotype expression under environmental stimuli to orthologous genes in model plants like *Arabidopsis thaliana* and *Zea mays*. This study aims to use these genes in understanding the relationship between drought tolerance and gene presence/expression in rice. This study will continue to identify species with genetic expressions correlating to high drought tolerance, with further applications to finding species that can be cross-bred using marker-assisted selection to produce drought tolerant species of rice.

1. Introduction

Drought occurs during an extended imbalance between precipitation and evaporation. With the increasing severity of climate change, historically dry areas are likely to have increased drought occurrences in the coming years. From 2015-2021, extreme dry and wet events in the U.S. occurred four times per year, versus three times in the 15 preceding years. It is predicted that over 50% of the world's arable land will be affected by drought in the year 2050 [1]. This is a major concern for agriculture. Rice is considered a key crop in global food security, since it is the primary nutrient source for more than three billion people [2]. However, rice is one of the most drought-susceptible plants because it cannot take up much water due to its small root system [3].

As the global population continues to grow, the demand for technologies that enhance crop yield has risen. Environmental stressors, including drought, have contributed to the depletion of crop yield. For rice, there was a 25.4% decline from 1980 to 2017 [4]. Critical mitigation of rice drought tolerance is needed to maintain global food security.

Because of the drastic effects drought has on plant growth, understanding the genetic expression of rice during drought is crucial for the mitigation of drought stress. Drought tolerance is composed of morphological adaptations, so this study focuses on the phenotypic expression of some morphological features.

Rice is cultivated globally and has a genome with 12 chromosomes encompassing 430 Mbp (mega base pairs). Past literature has identified genes in rice correlated with drought tolerance, but this approach seeks to use several quantitative techniques, including Quantitative Trait Loci (QTL) analyses and Genome-Wide Association Studies (GWAS), to identify genes associated with phenotypes related to crop yield. This paper also uses these two quantitative techniques to compare genes observed for plant height, and this is a unique approach compared to past studies [6].

Drought stress is the result of a long-term period of low soil moisture content accompanied by a continuous loss of water through evaporation and transpiration [7]. Thus, to understand drought stress, it is imperative to understand the morphological, physiological, and biochemical responses of rice to stress. Since phenotypes are mainly observed through morphological responses [5], this paper observes how decreased plant height and a reduced number of tillers, also known as shoots, could denote poor yield, which is determined by attributes such as impaired assimilate partitioning, reduced grain filling, grain weight and size, and death of the plant. Figure 1 shows how drought stress could contribute to morphological responses and correspond to poor yield. Though physiological and biochemical responses were not as heavily studied, all three response types could result in poor yield attributes [5]. Furthermore, morphological responses could be related to shoot development and characteristics during early development.

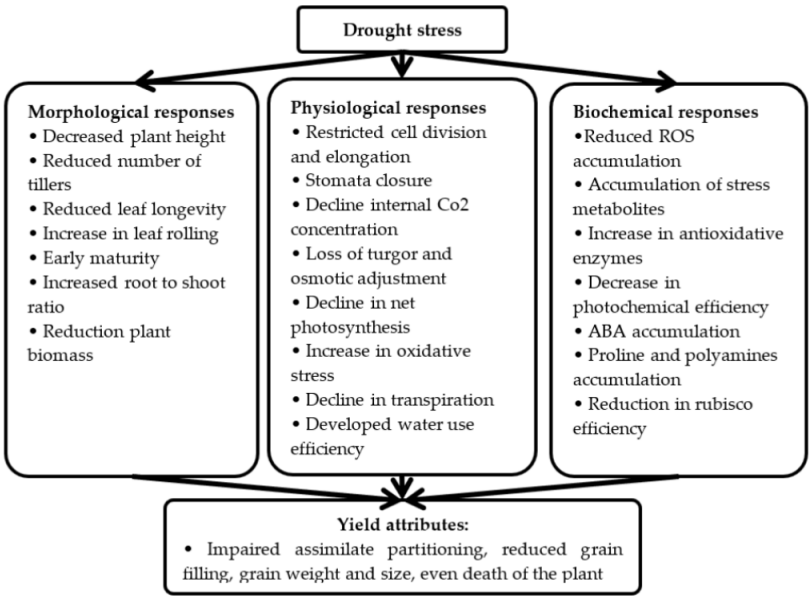

Figure 1: Drought stress influence on morphological, physiological, and biochemical responses of rice [5]

Various studies have found that other than morphological attributes denoting poor yield, there could be traits such as amylose and protein content that indicate poor yield. Insufficient water supply could lead to the reduction of carbohydrate synthesis in crops and lower grain and protein yield [8].

Overall, this study analyzes phenotypes with known correlations to environmental stimuli to observe how morphological changes due to drought tolerance may be the result of gene functions. Selected genes with quantitative statistical significance were observed using genome browsers to find gene functions and drought tolerant traits. This study responded to three main research questions:

1. Are there genes expressed in rice that are influenced by phenotypes related to drought?
2. Can the functions of these genes correlate to drought tolerance?
3. How can these genes be used to identify orthologous genes in other species and draw conclusions for drought tolerance?

2. Computational Approach/Methods

This study used two main quantitative techniques for identifying loci. The first technique is a Genome-Wide Association Study (GWAS). Generally, GWAS work to find genes that are associated with phenotypes across the whole genome [9]. Association mapping in this study identifies single-nucleotide polymorphisms (SNPs), or genomic variants at a single location within one nucleotide [10]. If there is a non-zero slope along the association between observed data for the same trait, there is an association between a SNP allele and a phenotype. An example of this association is shown in Figure 2.

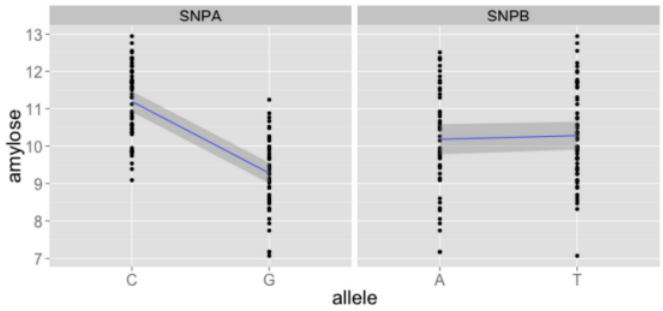

Figure 2: The non-zero slope for SNPA shows that it is significant because there is a correlation between the allele present and the phenotype observed. The flat correlation between the alleles on SNPB show that it is an insignifcant SNP. The y-axis corresponds to relative quantities of amylose, rather than specified quantities with units. In this case, two SNPs for amylose were portrayed, with SNPA showing the presence of C versus G yielded higher amylose content.

I used a GWAS model with a dataset including 413 rice varieties from 82 countries and 44,000 genetic mutations, or SNPs on each type of rice. Although there are around 500,000 SNPs among rice varieties, due to linkage disequilibirum, closely linked SNPs have high correlation. Thus, the majority of known SNPs are not necessary for the study. The goal of the GWAS was to identify loci that may indicate the expression of a particular trait. The dataset includes 37 phenotypes, but those that were most significantly linked to drought tolerance were studied, including: 1. plant height; 2. seed length to width ratio; 3. amylose content; 4. protein content.

Since the dataset was so large, a kinship matrix was used to adjust for population structure, assuming kinship from the observed genotypic data. A kinship matrix displays the genetic relatedness of organisms, with values representing genetic similarity ranging from 0 to 1 (1 as identical organisms, 0 as unrelated organisms). The kinship matrix in this method is estimated from the SNPs and data given, since the pedigree of the species is unknown. A LOD-score, also known as logarithm of odds, was used to estimate the genetic relatedness of a marker and an expressed trait. The LOD-score threshold for this data was 5.85, which is significant for plants [11]. After including a kinship matrix in the GWAS, significant SNPs were chosen. The three SNPs with the highest LOD scores at that location were identified as most significant. In this study a specific type of LOD score was used that used a -log(P) to normalize the dataset. This calculated probability was used to represent association, and is often synonymous with a LOD score, so in this paper LOD and -log(P) are used synonymously. There is a 95% certainty in this study, with a -log(P) used to normalize the data for association rather than probability of association by chance [12].

The second major quantitative technique that was used is a Quantitative Trait Loci (QTL) Analysis. QTL analyses work to identify molecular markers associated with phenotypes, rather than genes or purely genetic loci [6]. The QTL analysis was conducted on a cross-inbred population of *Curinga* x *Oryza rufipogon*. For the QTL data, the dataset was rather small, incorporating phenotypes of flow, height, tillers, panicles,

and pericarp. Panicles refer to branching clusters of rice and perciparp refers to the layer surrounding the ovary wall of a plant's seed. Based on the category "Morphological Responses" from Figure 1, tillers and height were chosen as proxies for drought.

Using publicly available rice and plant genome browsers, genes were identified that were nearest to the identified loci. The Rice Genome Annotation Project (RGAP) Browser from the University of Georgia shows all 12 chromosomes with markers along all base pairs. Rice loci were identified, as well as the best orthologous genes/proteins in *Arabidopsis thaliana* (thale cress) and *Zea mays* (maize). These were two plants that had genetic similarities of 93.71% and 94.95% to rice, respectively, as seen in Figure 3. Although *Triticum aestivum* (wheat) had a higher genetic relatedness, it was easier to find orthologous genes in *Arabidopsis thaliana* and *Zea Mays* because their genomes have been mapped extensively (especially *Arabidopsis thaliana*). The RGAP individual gene pages show gene ontologies (GO) and their accessions. These include the biological processes that are known as gene functions.

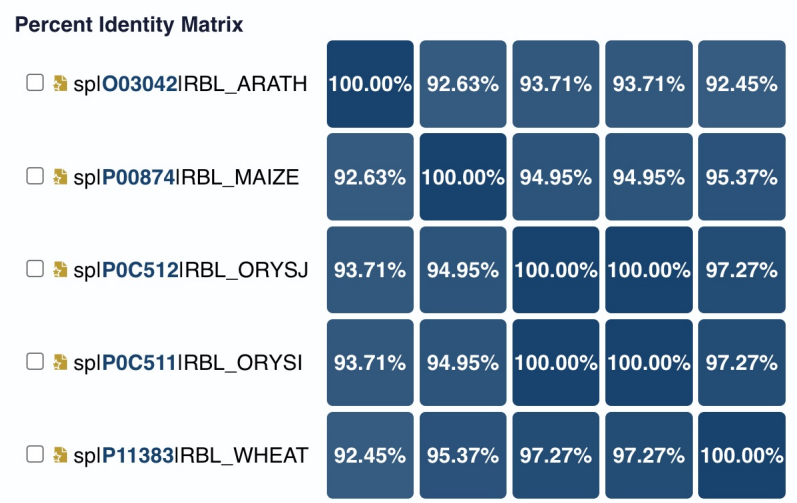

Figure 3: The percent genetic relatedness between *Arabidopsis thaliana*, *Zea mays*, *Oryza sativa japonica*, *Oryza sativa indica* and *Triticum aestivum*. These are listed vertically, where the *japonica* and *indica* species have 100% relatedness. This image was produced by creating a percent identity matrix on UniProt.

While the RGAP browser did not have extensive details on the ontologies, UniProt, a web-accessed resource, did, and was used to search for known genes from the QTL and GWAS studies and identify details about the proteins encoded in those genes and their functions. The main ontologies that aligned with the goals of the study were those related to biological processes marking responses to environmental external stimuli, as seen in Figure 4. Response to general environmental stress and heat/temperature stimuli ontologies were found.

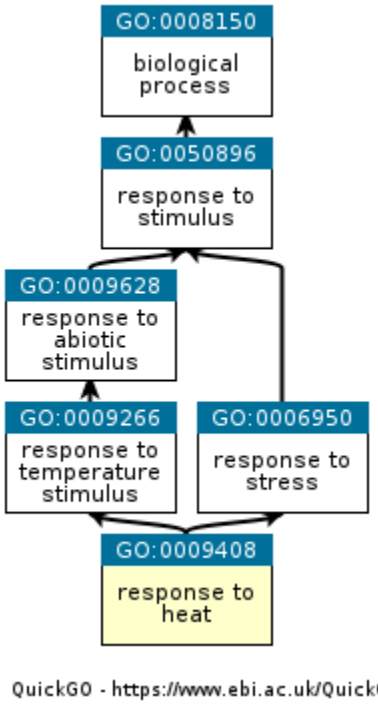

Figure 4: Gene Ontology accessions with responses to heat stimuli. This image shows blue boxes as biological processes with the black arrows pointing to possible processes that are influenced. The GO accession numbers refer to different possible processes in the rice genes. Each specific number classification is not significant, but the processes they entail are. It was produced by the European Bioinformatics Institute's gene ontology browser.

While not every gene had extensive information about it, general conclusions were drawn from gene ontologies on how those genes could be affected by external stimuli, with a focus on drought and heat.

3. Results and Discussion

For the plant height phenotype studied using a GWAS, one significant SNP was found, with a LOD score of 5.995572 and location of 38111539 bp on Chromosome 1, as shown in Figure 5.

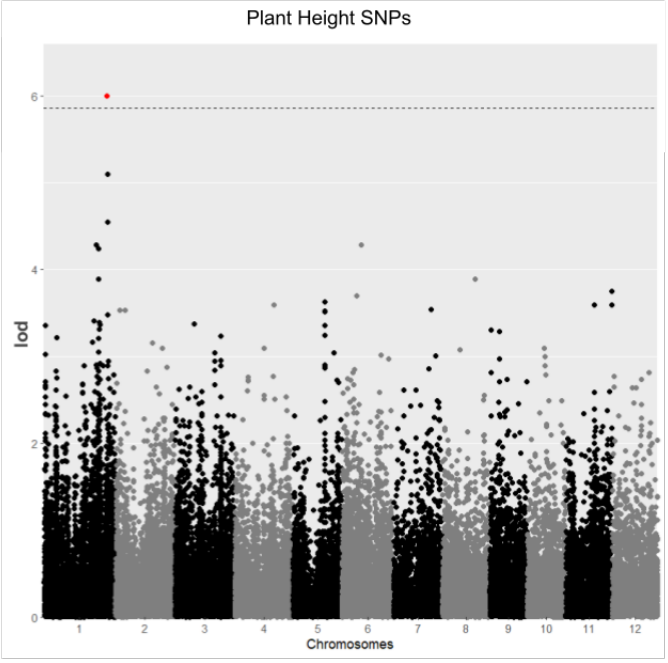

Figure 5: This shows the GWAS analysis adjusted with a kinship matrix for plant height, identifying one significant SNP as shown in red.

The three nearest loci were LOC_Os01g65650, LOC_Os01g65640, and LOC_Os01g65660, as seen in Figure 6. LOC refers to an identified loci, Os refers to *Oryza sativa*, the number 01 refers to the chromosome, and the remaining numbers denote the number gene it is.

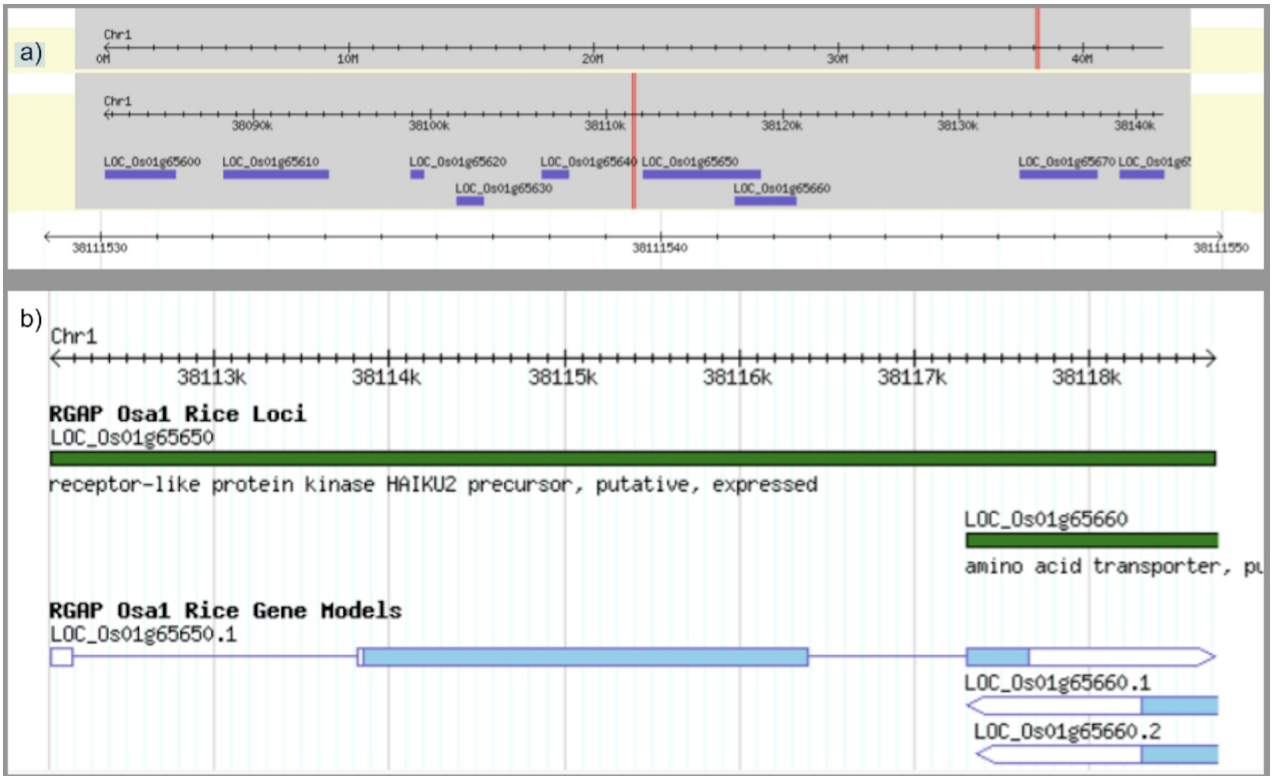

Figure 6: a) Genome Browser showing SNP on the genome with nearby loci. b) LOC_Os01g65650 shown on the genome.

LOC_Os01g65650 was observed as a protein-coding gene but little information was available. An orthologous *Arabidopsis thaliana* gene was identified as AT1G72180. The gene ontology showed the purpose of the gene is to mediate nitrate uptake and to regulate the shoot system development and response to osmotic stress. It is possible that this result with plant height correlated with shoot system regulation could be used as a proxy to potential results with tillers. This gene is also expressed heavily during the growth stages of the plant. Growth stages involving the root system may be correlated to plant height expression. Since the small root system of rice contributes to its high susceptibility to drought, this gene could be studied more extensively in the future to understand correlations between plant root development and height. Furthermore, osmotic stress could directly be related to drought influence, due to the imbalance in salinity and important ions in the cell [13]. There is an emphasis on this gene's functionality with shoot development, a key indicator of healthy plant growth under sufficient hydration.

For observing the plant height phenotype using QTL analysis, a mainscan plot was produced, as shown in Figure 7. The most significant QTL identified was one at Chromosome 2 and position 3.12 cM, with a LOD score of 6.91. 3.12 is the location in cM (centimorgans), so conversion to base pairs found the location as 842400 bp. This was observed in the rice genome browsers (Figure 8).

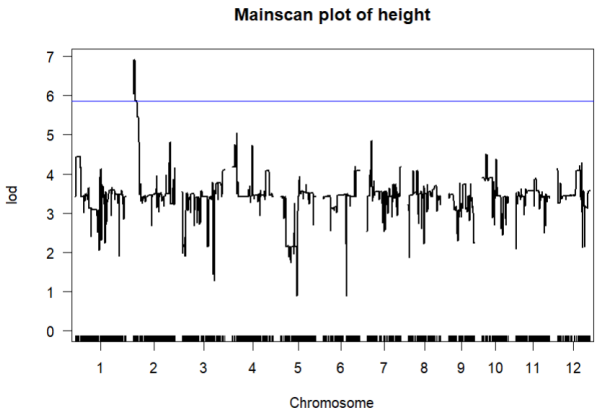

Figure 7: Mainscan plot showing the QTLs for height along the genome, with a peak at Chromosome 2 indicating a high LOD score and therefore high statistical significance. The blue line represents the LOD threshold of 5.85.

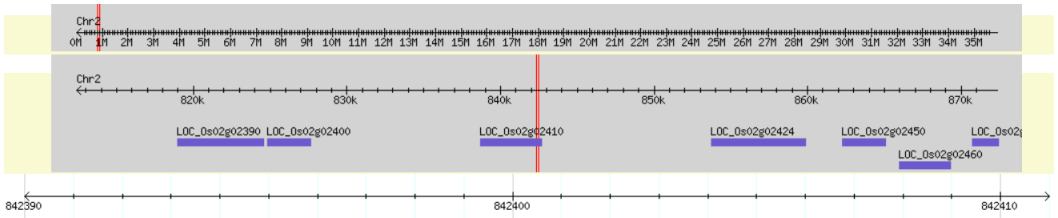

Figure 8: Rice Genome Browser showing location of Chromosome 2 at 842 kbp

The QTL was observed in gene LOC_Os02g02410, a DnaK family protein with an orthologous gene in *Arabidopsis thaliana* as AT5G42020. A DnaK protein expressed in chromosome 3 was also found, with possible proteins encoding for heat stress. It acts as a protein binder in response to heat stress, and there is a possibility that this DnaK protein could be used to target other environmental stressors, like drought. Other GO accessions not identified in the genes so far were 0009628 and 0006950, which could correspond to drought stress response. Further research may include identifying these accessions in QTLs and finding more phenotypes associated with these biological processes. Furthermore, a GO accession of GO:0009408 was present, which is related to heat stimuli response. Using the rice expression database, UniProt was used to identify another protein in the gene, Q6Z7B0, a kDa protein involved in the control of seed storage proteins during seed maturation. Due to the nature of the GWAS-identified plant height gene as a protein-coding gene, it is assumed that this gene has a purpose in coding for endoplasmic reticulum stress and therefore is crucial to the seed development. Additionally, both are expressed after flowering during seed development, not in mature seeds. Drought can decrease the quality of seeds during early development, thus contributing to poor plant growth [14]. No QTLs were identified for tillers, an additional possible indicator of drought tolerance. This is possibly due to the lack of data for this phenotype.

Other phenotypes were observed using GWAS. Seed length to width ratio was observed having a

significant SNP at Chromosome 5 and location 5425317 bp, with LOD score 12.592110, as seen in Figure 9. LOC_Os05g09550 was the gene that contained this SNP, and is a Der-1-like family conain containing protein. It is involved in the degradation of proteins in the endoplasmic reticulum, which is similar to the traits observed from the plant height loci.

Figure 9: This shows the GWAS analysis adjusted with a kinship matrix for seed length to width ratio. There is a peak in significant SNPs at chromosome 5, and these genes are all close to each other.

The amylose content phenotype GWAS analysis suggested the gene LOC_Os06g04200, a starch synthase with molecular functions in protein binding and metabolism. Since these are traits mainly correlated with plant regulation and the synthesis of vital nutrients, these could be compromised under drought stress, like they have been in sweet potato [15].

Finally, the protein content phenotype GWAS analysis suggested the gene LOC_Os06g09880, which contributes to major reproductive and embryonic development. Though this is different from the protein binding and metabolism processes seen with the other phenotypes, it still deals with seed development, suggesting that traits dealing with drought tolerance often are expressed during seed development. Future applications of this research could be to observe these traits as expressed during development, not after plant growth, and in response to environmental stimuli.

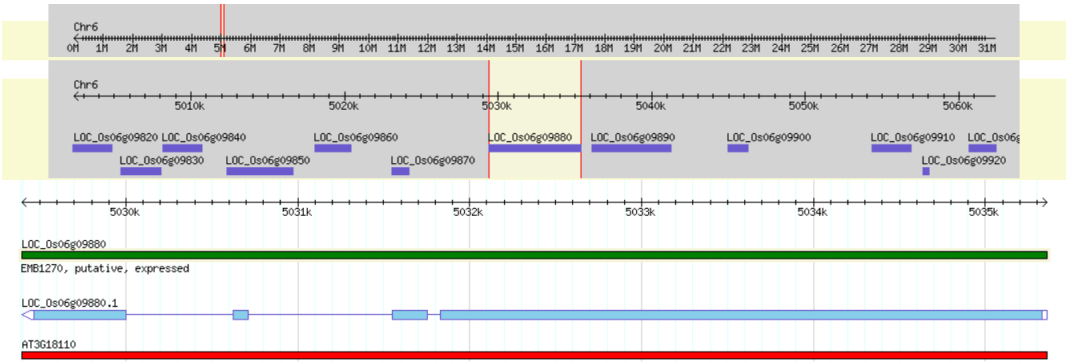

Figure 10: Genome browser showing SNPs for Protein Content

4. Conclusion

From both the QTL and GWAS techniques, genes associated with traits that are often responsive to drought stress were found to show high functions in protein content and development as well as seed development and early stages of plant growth. This is supported by orthologous genes and GO accessions for genes which these were available for. Generally, this study shows that there are genes that could be correlated with drought tolerance, and supports the hypothesis that their genetic functions are directly related to responding to environmental stress. Further applications of this study are extensive. More QTLs and SNPs with information available regarding gene ontologies could be studied to support the conclusions of this study. I will continue to identify genes that correspond to morphological processes that are influenced by drought. Additionally, the use of genome browsers to directly identify the genes that correlate with drought tolerance could support the conclusions of this study. Once genes with conclusive applications to drought tolerance are observed, these genes could be isolated in populations of rice and populations could be bred using marker-assisted selection to produce drought-tolerant species of rice.

5. Acknowledgements

This study is supported by the North Carolina School of Science and Mathematics. I would like to thank Mr. Robert Gotwals and Dr. Amy Sheck for making the program available and for their ongoing support. Special thanks is given to these individuals:

1. Dr. Susan McCouch, Professor at the School of Integrative Plant Science Plant Breeding and Genetics Section and Professor of Computational Biology at Cornell University, for help with data acquisition and curation.
2. Dr. Juan Velez, Post-Doctoral Fellow at the School of Integrative Plant Science Plant Breeding and Genetics Section at Cornell University, for help with data acquisition and curation.
3. Dr. Julin N. Maloof, Professor of Plant Biology, University of California Davis, for the use of his lab

activities on GWAS.

4. Drs. Eli Hornstein, Elysia Creative Biology, and Bri Edwards, Research Assistant, Alonso-Stepanova Lab, Plant & Microbial Biology, North Carolina State University, for guidance on navigating rice genome browsers.

References

[1] Li, B., & Cawdrey, K. (2023, March 20). *Warming Makes Droughts, Extreme Wet Events More Frequent, Intense – GRACE-FO*. GRACE-FO. Retrieved January 9, 2024, from https://gracefo.jpl.nasa.gov/news/220/warming-makes-droughts-extreme-wet-events-more-frequent-intense/

[2] Janakiraman, A. (2021, September 8). Rice crop: A vital cog in ensuring food security. Open Access Government. Retrieved January 9, 2024, from https://www.openaccessgovernment.org/ensuring-food-security/119387/

[3] Sahebi, M., Hanafi, M. M., Rafii, M. Y., Mahmud, T. M. M., Azizi, P., Osman, M., Abiri, R., Taheri, S., Kalhori, N., Shabanimofrad, M., Miah, G., & Atabaki, N. (2018). Improvement of Drought Tolerance in Rice (*Oryza sativa* L.): Genetics, Genomic Tools, and the WRKY Gene Family. *BioMed research international*, *2018*, 3158474. https://doi.org/10.1155/2018/3158474

[4] Zhang J, Zhang S, Cheng M, Jiang H, Zhang X, Peng C, Lu X, Zhang M, Jin J. Effect of Drought on Agronomic Traits of Rice and Wheat: A Meta-Analysis. Int J Environ Res Public Health. 2018 Apr 24;15(5):839. doi: 10.3390/ijerph15050839. PMID: 29695095; PMCID: PMC5981878.

[5] Oladosu, Y., Rafii, M. Y., Samuel, C., Fatai, A., Magaji, U., Kareem, I., Kamarudin, Z. S., Muhammad, I., & Kolapo, K. (2019). Drought Resistance in Rice from Conventional to Molecular Breeding: A Review. In International Journal of Molecular Sciences (Vol. 20, Issue 14, p. 3519). MDPI AG. https://doi.org/10.3390/ijms20143519

[6] Zheng, B. S., Yang, L., Mao, C. Z., Huang, Y. J., & Wu, P. (2008). Mapping QTLs for morphological traits under two water supply conditions at the young seedling stage in rice. In Plant Science (Vol. 175, Issue 6, pp. 767–776). Elsevier BV. https://doi.org/10.1016/j.plantsci.2008.07.012

[7] Climate Change Indicators: Drought — US EPA. (2023, November 1). Environmental Protection Agency (EPA). Retrieved January 10, 2024, from https://www.epa.gov/climate-indicators/climate-change-indicators-drought

[8] Wan, C., Dang, P., Gao, L., Wang, J., Tao, J., Qin, X., Feng, B., & Gao, J. (2022). How Does the Environment Affect Wheat Yield and Protein Content Response to Drought? A Meta-Analysis. In Frontiers in Plant Science (Vol. 13). Frontiers Media SA. https://doi.org/10.3389/fpls.2022.896985

[9] Al-Chalabi, A. (2009). Genome-Wide Association Studies. In Cold Spring Harbor Protocols (Vol. 2009, Issue 12, p. pdb.top66). Cold Spring Harbor Laboratory. https://doi.org/10.1101/pdb.top66

[10] Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., Mezey, J., McClung, A. M., Bustamante, C. D., & McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. In Nature Communications (Vol. 2, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/ncomms1467

[11] Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., & Kearsey, M. (2007). Gene expression quantitative trait locus analysis of 16000 barley genes reveals a complex pattern of genome-wide transcrip-

tional regulation. In The Plant Journal (Vol. 53, Issue 1, pp. 90–101). Wiley. https://doi.org/10.1111/j.1365-313x.2007.03315.x

[12] Qu, H. Q., Tien, M., & Polychronakos, C. (2010). Statistical significance in genetic association studies. Clinical and investigative medicine. Medecine clinique et experimentale, 33(5), E266–E270. https://doi.org/10.25011/cim.v33i5.14351

[13] Ma, Y., Dias, M. C., & Freitas, H. (2020). Drought and Salinity Stress Responses and Microbe-Induced Tolerance in Plants. In Frontiers in Plant Science (Vol. 11). Frontiers Media SA. https://doi.org/10.3389/fpls.2020.591911

[14] Abdul Rahman SM, Ellis RH. Seed quality in rice is most sensitive to drought and high temperature in early seed development. Seed Science Research. 2019;29(4):238-249. doi:10.1017/S0960258519000217

[15] Zhou, Z., Tang, J., Cao, Q., Li, Z., & Ma, D. (2022). Differential response of physiology and metabolic response to drought stress in different sweetpotato cultivars. In S. Dai (Ed.), PLOS ONE (Vol. 17, Issue 3, p. e0264847). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0264847