

COMPUTING THE RICE GENOME

FUNDING BY THE SYNGENTA COMMUNITY GRANT PROGRAM



Developer:
Robert Gotwals
Reva Kumar
January 16, 2024

COMPUTING THE RICE GENOME. COPYRIGHT HELD BY THE NORTH CAROLINA SCHOOL OF SCIENCE AND MATH, JULY 20, 2023. ALL RIGHTS RESERVED.

CONTENTS

1	Introduction	4
1.1	About this Activity	4
1.2	Why rice?	6
1.2.1	Rice Fundamentals	6
1.2.2	The Rice Genome	9
2	Student Activities	11
2.1	Student Activity 1: Comparative Genomics	11
2.1.1	Background Reading	11
2.1.2	Student Activity	12
2.2	Student Activity 2: Quantitative Trait Loci (QTL) Analyses	14
2.2.1	Background Reading	14
2.2.2	Student Activity	17
2.3	Student Activity 3: Genome-Wide Association Studies / Association Mapping	18
2.3.1	Background Reading	18
2.3.2	Student Activity	21
3	Useful Resources	23
4	Teacher Notes	24
4.1	Using these materials	24
4.2	Meeting North Carolina Standards	25
4.2.1	Strand: Evolution and Genetics	25
4.2.2	Next Generation Science Standards - Biology	25
4.2.3	Additional Considerations	27
	References	29

Funding support for this activity was provided by the Syngenta Community Grant Program in the Research Triangle Park, Durham, North Carolina. Additional funding and administrative support was provided by Dr. Amy Sheck, Dean of Science, and Deann Barger, Science Department Program Associate, both of the North Carolina School of Science and Math, Durham, North Carolina. Appreciation is also extended to the NCSSM Online classes, *Introduction to Computational Science* (especially Peggy Chen, Lillian Churchwell, Eve Jenkins, Diya Menon, Srinithi Mohan, Ria Saheta, and Ava Scherer) and *Data Science for Scientists* (especially Pranav Nekkalaipudi) who served as the test students for this work. Finally, appreciation to Jennifer Williams, Chair of Science at NCSSM-Morganton, for her review of the teacher resources.

The materials were developed by Mr. Robert Gotwals (gotwals@ncssm.edu), NCSSM Computational Science Educator, and Reva Kumar, NCSSM Class of 2024.

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, 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, 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. Dr. Karl Broman, Professor of Biostatistics and Medical Informatics at the University of Wisconsin, for his help with the QTL dataset.
5. 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.
6. Dr. Robert A. Dietrich, retired senior research scientist in plant genomics at Syngenta, for his review of the proposal.

Cover graphic source: <https://www.naro.go.jp/english/index.html>. [1]

INTRODUCTION

1.1 ABOUT THIS ACTIVITY

This activity is designed to apply the technologies, techniques, and tools of computational science to the study of an interesting scientific event: the study of the genetic structure of rice, scientific name *Oryza sativa*. Computational science is an interdisciplinary discipline that comprises three domains: science (used generically to include physical, natural, life sciences along with the humanities, social sciences, etc.); mathematics; and computer science. Figure 1 captures the essence of computational science.

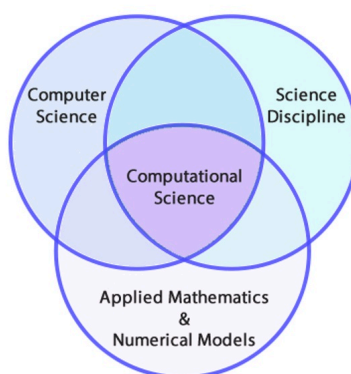


Figure 1: Computational Science: an interdisciplinary science

Computational science is considered to be the “third leg” of modern science, the others being observational/experimental science, and theoretical science. Computational science is sometimes called “modeling and simulation”, “scientific computing”, or “high-performance computing” (especially if a supercomputer is used).

In computational science, we can define a computational problem using a taxonomy of technologies, techniques, and tools. This taxonomy can be called **computational thinking**, and consists of these three elements:

1. **Technologies:** with technologies, we are defining the “X” in “computational X”. The “X” can be any discipline, such as chemistry, physics, biology, or environmental science. It can also be more specific, such as computational linguistics or computational art history. In this activity, our “X” is biology. **Computational biology** is a very broad and diverse field, focused on items such as protein structure, evolutionary relationships, the interactions of biological molecules, and the like. Under the category of computational biology is the sub-discipline of **bioinformatics**. Bioinformatics primarily deals with the analysis of biological data. In most cases, this data is genetic and/or genomic DNA data, such as base pairs (A, C, T, and

Taxonomy: a method of describing various approaches or models to a given task

G) or protein structures (consisting of amino acids). Much of this data is obtained through a variety of sequencing techniques. This activity uses a variety of resources that are of a general computational biology nature and those that are typical in the sub-discipline of bioinformatics.

2. **Techniques:** this refers to developing an understanding of the underlying science found in the topic. For computational biology/bioinformatics, it is most typically the case that a solid understanding of fundamental genetics (the structure and function of DNA, the Central Dogma – replication, translation, transcription – and the structure and function of proteins) is critical.
3. **Tools:** depending on the technology and techniques, one or more specific computational tools can be identified. For computational biology/bioinformatics, the use of the programming languages R or Python is the most common tool. Also, and it is certainly the case for this instructional package, the use of online databases and genomic browsers is of critical importance. Databases/browsers include such web-based tools PlantDB (<https://www.plantgdb.org/>) and Gramene (<https://www.gramene.org/>).

Figure 2 shows examples for a number of scientific problems and disciplines, including computational biology/bioinformatics.

Computational Thinking (CT): the cognitive processes necessary to engage with computational tools to solve problems.

Problem Statement	CT: Technology	CT: Techniques	CT: Tools
What are the properties of the water molecule?	Computational quantum chemistry	Schrodinger's Equation; use of molecular mechanics (MM), semi-empirical methods, ab initio methods, or density functional theory methods	Gaussian16 using WebMO interface
What genes are responsible for a phenotype such as high blood pressure?	Computational biology / bioinformatics	Quantitative trait loci (QTL) analyses	R, using the R/qtl package
How does a disease spread over time?	Computational Epidemiology / System Dynamics	SIR algorithm (Kermack/McKendrick, 1927); System of differential equations; solution using an integration approximation (Euler, Runge-Kutta, etc.)	STELLA; VenSim; <i>Mathematica</i>
What is the behavior of greenhouse gases (CO ₂ , N ₂ , H ₂ O) at various levels in the atmosphere?	Computational physical chemistry	Identification of appropriate gas laws: Ideal, van der Waals, Beattie-Bridgeman, Redlich-Kwong	Spreadsheet; Python; <i>Mathematica</i> ; R; Any procedural programming tool
How does a drug work to treat a medical condition?	Computational medicinal chemistry	Protein-ligand docking; determination of binding affinities (pK _i)	Molegro; Autodock Vina; Schrodinger Bioluminate

Figure 2: Computational Thinking: the "three T's"

1.2 WHY RICE?

There are a number of interesting and important plants that can be used for the study of genomics. The most typical plant is *Arabidopsis thaliana*, more commonly known as thale cress. This plant has a very short genome, consisting of five chromosomes and 157 mega-, or million, basepairs (Mbp), and was one of the first plants to have its genome sequenced [4]. This species is, however, generally considered to be a weed, and, as such, is not of any particular importance other than to geneticist.

Of more interest are plants that have value to humans, and, of course, the primary usefulness is plants that serve as food sources. Of the many plants that serve as crops, there are none quite as useful as rice. Rice is important for several reasons, and its significance can be seen from economic, cultural, nutritional, and agricultural perspectives. Here are some key reasons why rice is considered important:

1. **Rice as a food staple:** Rice is one of the primary food sources for a large population on Earth, with various estimates suggesting that rice is the main source of calories for 3/4 of the world's population.
2. **Rice nutrition:** Rice is a good source of carbohydrates, essential for energy production in the body. Rice also contains some essential vitamins, such as Vitamins A and B1. Efforts to genetically modify rice to contain some of these essential items is ongoing, but some are concerned about GMOs (genetically-modified organisms). Rice also has a fairly high caloric efficiency, meaning it produces a lot of calories per unit of farmland (land, water usage, etc.)
3. **Farming sustainability:** Rice can be grown in diverse environments, as evidenced by production reports from China, Japan, Africa, and others. Rice is typically grown in flooded paddies and upland areas.
4. **Economic Impact:** Rice is a highly traded commodity. Various market reports as of this writing (July 2023) Based on the current predict that the export price per kilogram of Rice from the US will remain at \$0.57 per kilogram (kg) in 2023 and 2024.
5. **By-products:** the by-products of rice production, such as rice straw, can be used for other commercially-viable products such as animal food, production of bioenergy, and compost.

The Wikipedia entry (<https://en.wikipedia.org/wiki/Rice>) is excellent, and the curious reader is encouraged to read this resource.

1.2.1 RICE FUNDAMENTALS

In this section, we describe the different species of rice. It is estimated that about 120,000 varieties of rice exist in the world.

There are two cultivated and twenty-one wild species of genus *Oryza* [8]. Generally, in talking about rice we are looking at the species *Oryza sativa*. *O. sativa* is the plant species most commonly referred to in English as rice. It is the type of farmed rice whose cultivars are most common globally, and was first domesticated in the Yangtze River basin in China 13,500 to 8,200 years ago.

Oryza sativa belongs to the genus *Oryza* of the grass family *Poaceae*. With a genome consisting of 430 Mbp across 12 chromosomes, it is renowned for being easy to genetically modify and is a model organism for the botany of cereals.[19]

Megabase-pair: A megabase pair, abbreviated Mbp, is a unit of length of nucleic acids, equal to one million base pairs. The term 'megabase' (or Mb) is commonly used interchangeably, although strictly this would refer to a single-stranded nucleic acid.

GMO: any organism, in this case rice, that has been genetically altered to produce a beneficial trait

1. Indica Rice (*Oryza sativa indica*):
 - (a) Adaptation: Indica rice varieties are well-suited to tropical and subtropical regions with high temperatures and relatively longer growing seasons.
 - (b) Morphology: Indica rice plants are generally taller with longer leaves and a tendency to lodge (fall over) due to their height. They have a higher tillering capacity, which means they produce more tillers or side shoots.
 - (c) Grains: Indica rice grains are typically long and slender, with a length-to-width ratio greater than 3:1. They tend to be less sticky when cooked.
 - (d) Examples: Basmati and Jasmine rice are well-known aromatic varieties of indica rice.
2. Japonica Rice (*Oryza sativa japonica*):
 - (a) Adaptation: Japonica rice is best suited for temperate regions with cooler climates and shorter growing seasons.
 - (b) Morphology: Japonica rice plants are shorter in height with shorter leaves, which makes them less susceptible to lodging. They generally have fewer tillers compared to indica varieties.
 - (c) Grains: Japonica rice grains are typically short and plump, with a length-to-width ratio around 2:1. They tend to be stickier when cooked, making them suitable for dishes like sushi.
 - (d) Examples: Sushi rice is a well-known variety of japonica rice.

It's important to note that within these two major subspecies, there is a vast array of rice varieties with diverse characteristics, including differences in grain size, color, aroma, cooking qualities, and disease resistance. Additionally, efforts in rice breeding have led to the development of hybrid varieties that combine the desirable traits of both indica and japonica rice, aiming to improve yield, quality, and adaptability to various environmental conditions.

Some specific species are described below.

1. *O. glaberrima*: *Oryza glaberrima*, commonly known as African rice, is one of the two domesticated rice species. It was first domesticated and grown in West Africa around 3,000 years ago. In agriculture, it has largely been replaced by higher-yielding Asian rice (*O. sativa*) and the number of varieties grown is declining. It still persists, making up an estimated 20% of rice grown in West Africa. It is now rarely sold in West African markets, having been replaced by Asian strains.[17]
2. *O. rufipogon*: also known as brownbeard rice, wild rice, and red rice, is an invasive species, and is considered to be a noxious weed in Alabama, California, Florida, Massachusetts, Minnesota, North Carolina, Oregon, South Carolina, and Vermont. It is native to East-, Southeast- and South- Asia. It has a close evolutionary relation to *Oryza sativa*, the plant grown as a major rice food crop throughout the world.[18]
3. *O. nivara*: *Oryza nivara* is a wild progenitor of the cultivated rice *Oryza sativa*. It is found growing in swampy areas, at edge of pond and tanks, beside streams, in ditches, in or around rice fields. Grows in shallow water up to 0.3 m, in seasonally dry and open habitats. It is an annual, short to intermediate height (usually <2 m) grass; panicles usually compact, rarely open; spikelets large, 6–10.4 mm long and 1.9-3.4 mm wide, with strong awn (4–10 cm long); anthers 1.5–3 mm long. Its distribution includes Bangladesh, Cambodia, China, India, Laos, Malaysia, Myanmar, Nepal, Sri Lanka, Thailand, and Vietnam.[15]
4. *O. longistaminata*: this is a perennial species of grass from the same genus as cultivated rice (*O. sativa*). It is native to most of sub-Saharan Africa and Madagascar. It has been introduced into the United States, where it is often regarded as a noxious weed. Its common names are longstamen rice and red rice.[14]

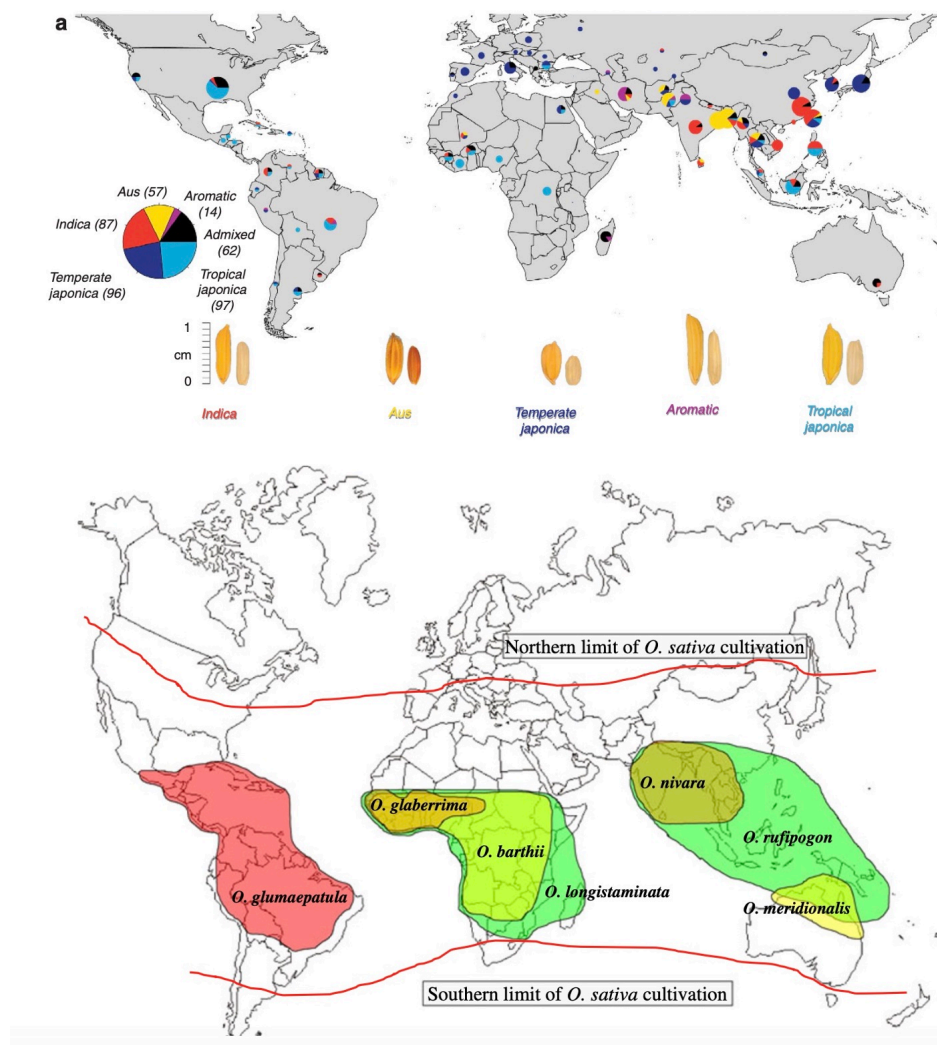


Figure 3: Maps showing where different species of rice are grown [13] [21]

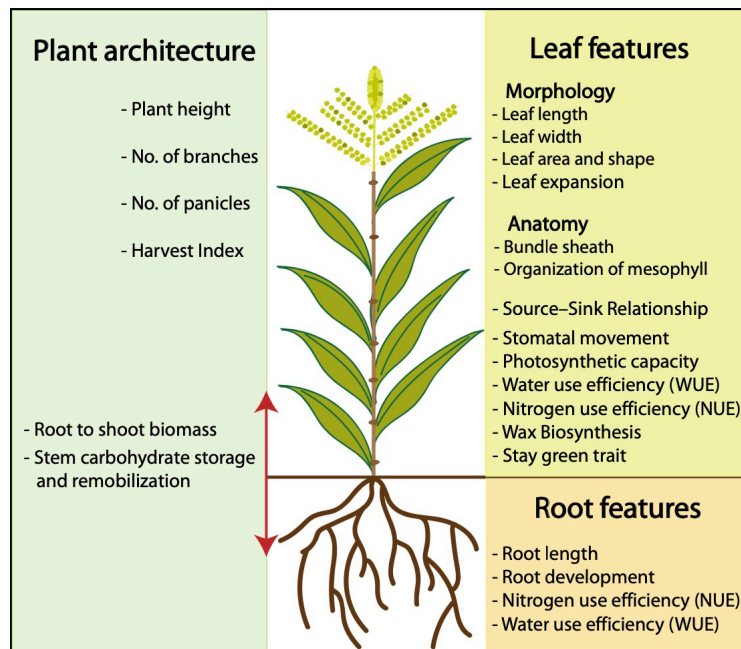


Figure 4: Architecture of a plant, showing traits relevant for high yield. [13]

1.2.2 THE RICE GENOME

Figure 5 shows an overview of the rice genome. As the figure shows, rice contains 12 chromosomes

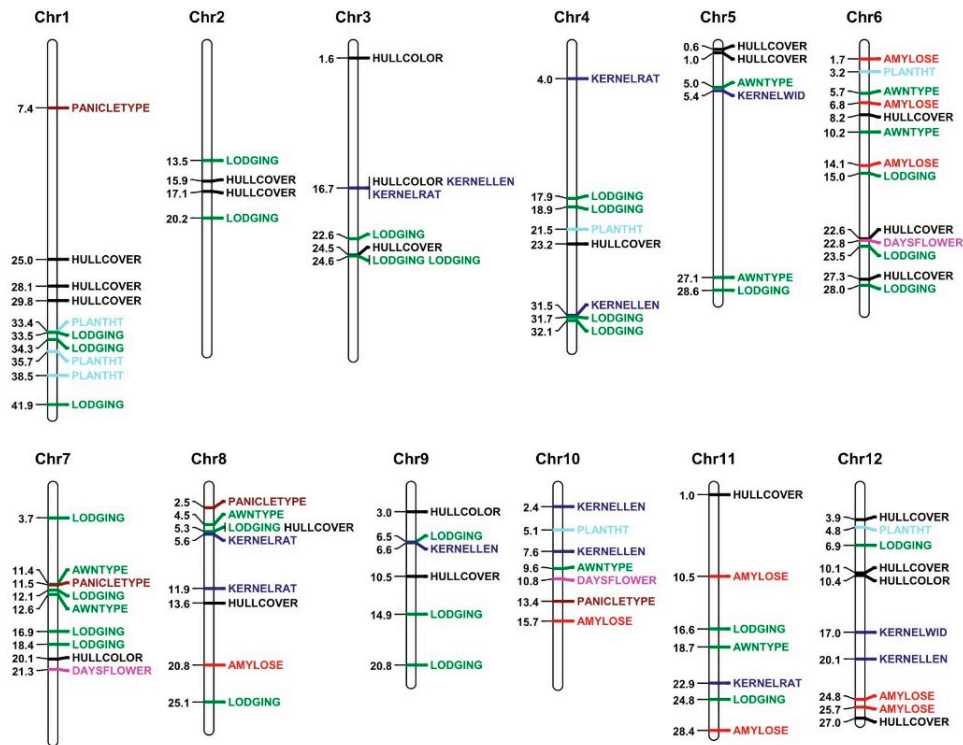


Figure 5: Genome structure of rice, showing chromosomes and major quantitative trait loci (QTLs) [11]

Examples of important genes in *O. sativa*

1. OsSPL16: Controls grain width and contributes to grain yield in rice.
2. OsSPL14: A gene involved in the regulation of grain size in rice.
3. OsNAC5: A NAC transcription factor associated with drought tolerance in rice.
4. OsDREB1A: A gene encoding a dehydration-responsive element-binding protein that enhances drought and salt tolerance in rice.
5. OsBADH2: This gene is related to submergence tolerance in rice and helps the plant survive under flooding condition
6. OsAGL71: Controls seed size and grain filling in rice.
7. OsSWEET11: A sugar transporter gene that plays a role in rice susceptibility to bacterial blight.
8. OsCIPK03: Involved in salt stress tolerance in rice by regulating ion homeostasis.
9. OsRFL: A gene involved in promoting root development and enhancing drought tolerance in rice.
10. OsMYB55: Regulates submergence tolerance in rice.

STUDENT ACTIVITIES

Note to students and teachers: These activities can be done independently as time and interest warrants. Each of the activities is a stand-alone project. The activities assume a basic knowledge of DNA and protein structure: that DNA consists of pairs of nucleotides (A, C, T, and G), and that these nucleotides code in triplets (for example, ATG) to form amino acids, which then are combined in amino acid, or peptide, sequences to form proteins.

Nucleotides: a unit of the DNA molecule which contains a sugar, phosphate group, and base

2.1 STUDENT ACTIVITY 1: COMPARATIVE GENOMICS

2.1.1 BACKGROUND READING

One of the most basic technique in computational biology/bioinformatics is **comparative genomics**. As the name suggests, comparative genomics allows researchers to compare the genomes between two or more organisms. This is important for being able to determine if these organisms are related genetically, descend from a common ancestor (and, if so, how long ago) or otherwise have some commonalities. A practical use of comparative genomics is in drug design and discovery: since it is difficult (and unethical) to test new drugs on humans, we want to be able to find animal models that have the same genetic – and hence, same phenotypic – characteristics as humans. Comparative genomics has pointed to the mouse (*Mus musculus*) as a viable animal model for drug testing, since it shares a significant amount of DNA with humans.

Genome: the complete set of genes or genetic material present in a cell or organism.

In this activity, the research question is: **How do six important cereal plants – two species of rice, wheat, maize (corn), oat, and foxtail millet – compare genetically.**

There are a number of techniques and tools for conducting comparative genomics analyses. Most comparative genomics experiments focus on one gene/one protein, and look to see how that gene/protein compare in terms of its genetic/genomic profile.

For this activity, we have chosen the protein *Rubisco*, a protein that serves as molecular machine – an enzyme – in an organism. Rubisco, also known as *ribulose biphosphate carboxylase/oxygenase*, is found in virtually every plant, and accounts for as much as 40% of the total protein content in plants. [6] Studies suggest that this protein emerged approximately four billion years ago in primordial metabolism prior to the presence of oxygen on earth [12]. Rubisco works to convert carbon dioxide (CO_2) into energy-rich molecules such as glucose. It is a part of the process of photosynthesis. The atom magnesium (Mg) plays an important role in the proper functioning of the rubisco enzyme.

Enzyme: a protein that serves to catalyze a chemical reaction in an organism

Figure 6 shows the secondary structure of the rubisco protein, with the magnesium atom shown as a red ball.

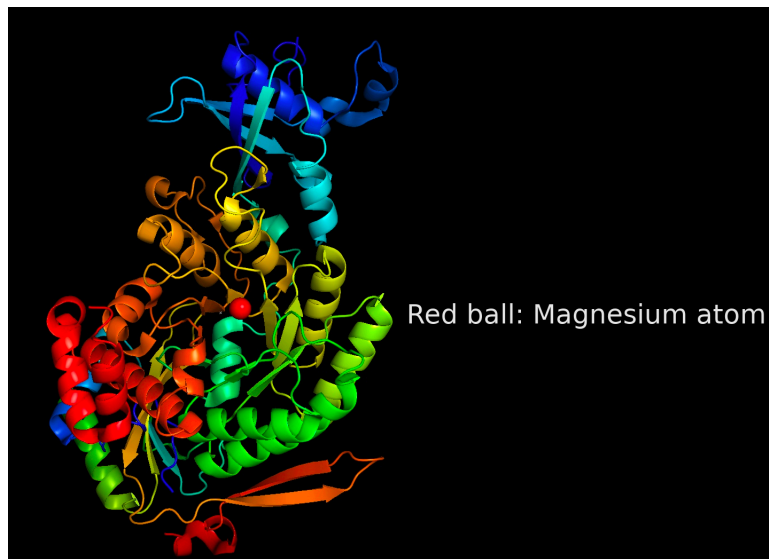


Figure 6: Protein secondary structure of rice, showing the location of the magnesium atom (red ball)) [11]

An excellent reading on Rubisco can be found on the Protein Data Bank's "Molecule of the Month" resource (<https://pdb101.rcsb.org/motm/11>).

2.1.2 STUDENT ACTIVITY

Research question: how does the Rubisco gene/protein compare in six plant cereals: two species of rice (*O. sativa japonica* and *O. sativa indica*), wheat (*Triticum aestivum*), maize (*Zea mays*), oat (*Avena sativa*), and foxtail millet (*Setaria italica*).

Procedure:

1. For this activity, the tool is a web-based resource, UniProt [2] (<https://www.uniprot.org/>). Open this site on your browser.
2. In the search box, look for "Rubisco".
3. You should get almost 250,000 hits! We need to reduce this number. Rubisco has two protein chains, one large and one small. The one of interest is the large chain. Do a search for "Ribulose biphosphate carboxylase large chain".
4. There are still a large number of hits, and this should be no surprise, given that every plant contains this protein. Notice that the "Entry ID" has the identifier "RBL" and then the name of the plant. In some cases, the plant name is the Latin name (RBL_ORYSI for *O. sativa indica*), and in some cases it's the general name of the plant (RBL_WHEAT).
5. Find the seven cereals and check the checkbox next to the names.
6. Do a search for RBL_ARATH (*Arabidopsis thaliana*, and add that.
7. Once you are done, look in the "basket" icon at the top right of the window. Clicking on that, you should see the seven cereals in your basket.
8. Let's start by comparing Rubisco in two plants, *Arabidopsis thaliana* and *O. sativa japonica*. In the basket, select these two plants. We are now going to use a tool called CLUSTAL [9] to align the amino acids in each of the two plants, click on the "Align". A new window opens. You might want to change the name of

Conserved: amino acids that are exactly similar. AAs can also be partially conserved, meaning the two AAs are similar but not exact.

your "job" at the bottom left, perhaps something like "Rice-Cress align". When ready, hit the "Run" button at the bottom right.

- the alignment takes a few minutes to run. When it's done, you should have the results listed as "Complete" and you can click to display your results. The first window shows the alignment of amino acids, shown in Figure 7.

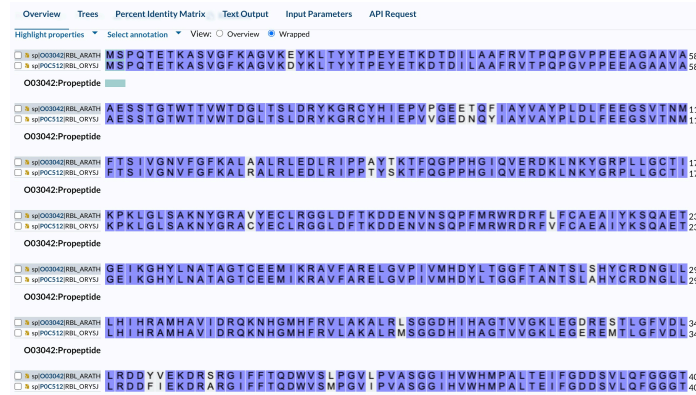


Figure 7: Comparative Genomics: *Arabidopsis thaliana* and *O. sativa japonica*

If the two amino acids are the same, we say they are **conserved**. If two amino acids are conserved, they are indicated with blue shading, and white if they are not. Partially conserved amino acids will appear as gray shading. Non-conserved amino acids will be indicated with a "-" symbol. Notice that most proteins start with the amino acid methionine, coded as "M". Methionine is also considered to be the "start reading the DNA" amino acid.

- A simple visual inspection will suggest that these two plants are very similar. We can, however, see exactly how similar they are by looking at the "Percent Similarity Matrix". Figure 8 shows the results, and we have added a table to help with interpretation. You can see that, for the rubisco gene/protein, the two plants are 93.71% similar. Thus, we could perform genetic experiments on *Arabidopsis thaliana* and have some degree of confidence that the results would also apply to *O. sativa japonica*.

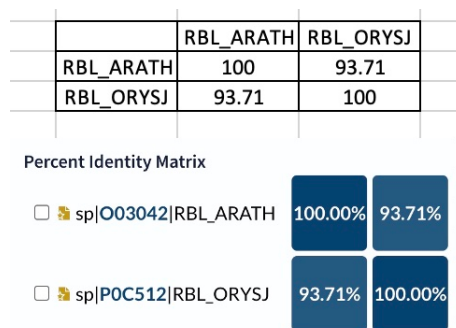


Figure 8: Comparative Genomics: Percent Identity Matrix

- Now you are ready to do a comparative genomics study on the seven cereal plants. In your basket, select all seven organisms, select "Align" and give your job a name such as "Seven cereals".
- Once your job is complete, look at your alignment. How similar are your seven cereals overall? 100%?

90%? You can write your observations in the text box below as desired.

Student Response:

13. Check your intuition by looking at the Percent Identity Matrix. You should notice that the two rice species are fundamentally 100% similar, and this should be no surprise. What about the other organisms? Other than the 100% plants, which two plants are the most similar? Which two plants are the least similar? You can jot your observations in the text box below.

Student Response:

Phylogenetic Tree: a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor.

14. When studying multiple organisms (more than two), another tool is a phylogenetic tree. A phylogenetic tree shows how a gene/protein changed over evolutionary time, typically measured in units of millions of years. Figure 9 shows a tree for our seven cereals. Evolutionary time starts on the far left, moving to present day on the far right. From the tree in Figure 9, we can see that wheat evolved first, followed by oat (RBL_AVEA). The two rice species evolved at the same time, again, no surprise. There are other types of trees, such as guide trees, and the curious student can explore these different types of trees. A short and simple description of phylogenetic trees can be found on YouTube (<https://www.youtube.com/watch?v=M7rqXEogkwU>).

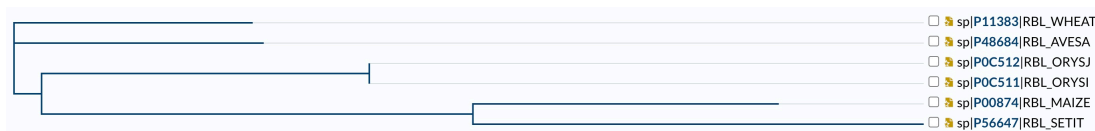


Figure 9: Comparative Genomics: Phylogenetic tree

15. You might try finding a plant gene that is common to a variety of plants. For example, I know that the gene/protein tubulin (TBA1) is important for the formation of exoskeletons for plants. A UniProt search found about 24 hits, and I selected examples such as rice, yeast, peas, green algae, maize, barley, and *Arabidopsis thaliana*. The results are interesting!

2.2 STUDENT ACTIVITY 2: QUANTITATIVE TRAIT LOCI (QTL) ANALYSES

2.2.1 BACKGROUND READING

Quantitative Trait Loci (QTLs) are a useful tool in genomic studies as they are similar to genes, being associated with phenotypic traits. However, while genes and QTLs can both be associated with certain traits, QTLs represent genetic regions that may involve multiple genes attributing to one phenotypic trait. Additionally, there is distinction between the types of trait indicated by both genetic regions. Quantitative traits represent phenotypes like height or weight, which are represented by a continuous set of possible quantities for a population.

In contrast, qualitative traits, like color, are discrete and can often be controlled by a single gene, without much overlap.

QTLs are useful for many applications, especially in plants. They can provide understanding of the basis of phenotypic traits, many of which may be hard to follow due to complicated genetic structure, as many genes and even environmental factors sometimes all contribute to one trait. Therefore, the mapping and identification of QTLs is crucial. QTL analysis works to map out and identify genetic structures associated with traits across the genome. Methods like linkage analysis and association studies look to find relationships between genetic markers and observable traits, to identify loci that are associated with a quantitative trait. For example, some region of a chromosome can involve a number of genes deciding the height of a plant organism. Typically, studies are conducted on backcrossed populations, or recombinant inbred lines, to guarantee high genetic diversity and therefore a more accurate QTL identification. During mapping, relationships between markers and traits are indicated through calculating strength of association, where high association between a genetic marker and an observed trait means there is high possibility of a QTL.

Data for conducting a QTL analysis comes from extracting DNA from the target organism, such as rice plants, mice, etc. Two different species, each with a different genotype, are bred, typically over multiple generations, to produce an organism that has a specific genetic/genomic configuration.

Figure 10 shows an example of breeding (in this case, in mice) to produce organisms with specific genotypes. Starting with two parents of mice that are homozygous (same) for some allele, we can produce mice with different genotype profiles. The figure shows three varieties: congenic strains, recombinant inbred (RI) strains, and consomic strains. This typically does not happen with the first group of baby mice (F₁ or filial, F₁), but requires multiple breeding using techniques such as backcrosses or intercrosses. Figure 11 shows some sample QTL data for rice, produced by the lab of Dr. Susan McCouch at Cornell University.

Genotype: Genotype is the genetic makeup of an individual cell or organism that determines or contributes to its phenotype. The contrasting terms genotype and phenotype are used to define the characteristics or traits of an organism.

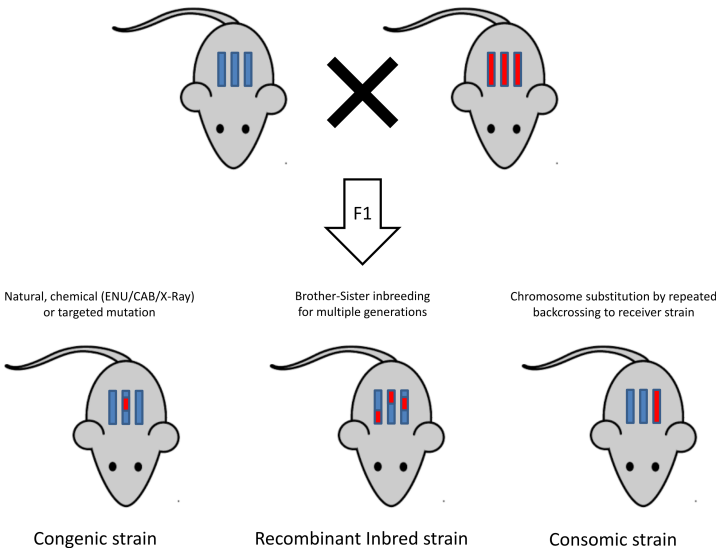


Figure 10: Graphic of a backcross in mice[20]

	A	B	C	D	E	F	G	H	I
1	Flow	Height	Tillers	Panicles	Pericarp	X8144	X19846	X20215	id1000955
2						1	1	1	1
3						1.2609	2.870808	2.918664	4.179784
4	81	97.6	94	82	White	AA	AA	AA	AA
5	80	122.4	75	60	Red	BB	BB	BB	BB
6	68	92.2	78	58	Red	AA	AA	AA	AA
7	85	93.4	76	70	White	BB	BB	BB	BB
8	77	110.8	68	62	White	AA	BB	BB	BB
9	79	104.2	70	62	Red	AA	AA	AA	AA
10	81	107	58	43	White	AA	AA	AA	AA
11	77	107.2	74	51	White	AA	AA	AA	AA
12	76	99.8	69	56	White	AA	AA	AA	AA
13	76	97.6	64	56	White	AA	AA	AA	AA

Figure 11: Sample of QTL data [10]

Once QTLs are mapped, they can be applied heavily in plant genomic studies. QTL Analysis is a helpful tool for computational genetic studies and is especially useful in marker-assisted selection (MAS). An example of a significant issue facing farmers is the inability to effectively choose breeding lines to optimize a particular trait for plants. For instance, a breeder may want to breed a population of rice that all have high drought-resistance. The convenience of utilizing QTLs is that they are located between genetic markers, and thus, breeders can use MAS to selectively breed organisms that have a genetic marker with a known QTL associated with a favorable trait. MAS can be used for several traits, like disease-resistance or salinity tolerance. Additionally, QTLs can be applied to developing genetic linkage maps, to provide greater insight into how genes are organized on a chromosome to therefore find other genetic elements that influence traits, like single-nucleotide polymorphisms. Further, QTLs can be used to comprehend trait development in several plant species, by performing comparative genomic studies to locate similar genomic regions across different species. Lastly, QTLs can be used to discover novel genes in plant species, and when incorporated with population study concepts like linkage disequilibrium, can optimize even further our knowledge about a populations genomics.

Examples of important Quantitative Trait Loci (QTLs) in *O. sativa*

1. SUB1 QTL: Associated with submergence tolerance. SUB1 QTL helps rice plants survive prolonged submergence by maintaining elongation growth.
2. qDTY QTL: qDTY QTLs are associated with drought tolerance. They help rice plants maintain better yields under water-limited conditions.
3. qSD1 QTL: Known as the "Semi-Dwarf 1" gene, it is associated with reduced plant height and increased lodging resistance, leading to improved yield stability.
4. qHD9/Heading Date 9 QTL: Influences flowering time and photoperiod sensitivity, affecting the time of heading (flowering) in rice.
5. qGW5/GRAIN WIDTH 5: This QTL regulates grain width and is associated with enhanced grain size and yield.
6. qP9/PANICLE NUMBER 9: Linked to panicle architecture and controls the number of panicles per plant, affecting overall grain production.
7. qRL7-1/QTL for Root Length 7-1: Associated with longer root length, which can enhance nutrient and water uptake in rice plants.
8. qWGW6/Wide and Heavy Grain 6: This QTL is related to grain size and weight, leading to increased grain yield.

9. qPH1/PANICLE LENGTH 1: Regulates panicle length, impacting the number of grains per panicle and overall yield.
10. qWLS1/WHITE LEAF SPOT 1: An important QTL associated with resistance to the devastating rice disease called "white leaf spot."

There are a number of computational tools that can be used to conduct QTL analyses. For this activity, we will be using a "package", sometimes call a library, in the programming language R, which uses an interface called RStudio. The package is called, appropriately, "qtl", and was developed by Dr. Karl Broman of the University of Wisconsin, primarily for mouse genomic studies. [5]

For this activity, we are going to use QTL data provided by Dr. Susan McCouch at Cornell University [10]. This data was generated by crossing, or breeding, two species of rice (*O.sativa: Orysa Curinga* and *Orysa Rufipogon*. *O. Curinga* is defined as a "a tropical japonica upland cultivar" [3]. A cultivar is a plant species that has been created by cultivation, or breeding. This list (https://en.wikipedia.org/wiki/List_of_rice_cultivars) on Wikipedia shows an extensive list of cultivars [16].

2.2.2 STUDENT ACTIVITY

For this activity, you will write a computer program (called a "script") using the programming language "R" and an interface (GUI, graphical user interface) called "RStudio". Both of these programs can be downloaded and installed on a computer, but if you are using a Chromebook, or do not wish to do so, you can use a version of R/RStudio "in the cloud" at <https://posit.cloud/>. You will be asked to create an account, but it is free.

The majority of the information needed to prepare an R script for this activity can be found in the document "Rice Quantitative Trait Loci (QTL): An Annotated Guide", found at the end of this document.

The *Annotated Guide* uses plant height as the example phenotype. You might want to use this phenotype to test your code, making sure you get the same results, but

1. **Flow:** number of days until the plant shows flowering.
2. **Tillers:** A tiller is a shoot that arises from the base of a grass plant. The term refers to all shoots that grow after the initial parent shoot grows from a seed. Tillers are segmented, each segment possessing its own two-part leaf.
3. **Panicles:** a panicle is a branched flower cluster (as of a lilac or some grasses) in which each branch from the main stem has one or more flowers.
4. **Pericarp:** the pericarp is the wall of a rice kernel, and is characterized by its color. For this lab, you don't want to pick this phenotype, since it is not a numerical value!

As a reminder, the **primary** goal of conducting a QTL analysis is to produce a mainscan, a graph that shows the chromosomes on the x-axis and the LOD score on the y-axis. As the previous margin note states, LOD scores of 3 or greater are considered significant.

Figure 12 shows the mainscan for the phenotype of plant height. In this mainscan, we see many peaks above three (3), but there is a substantial peak on Chromosome 2. Given that, we can make the statment that there exist a number of genes on Chromosome 2 that are factors for plant height. A breeder, if they wanted to genetically modify rice plants to get shorter or taller plants, would focus their efforts on Chromosome 2. It should be noted,

LOD score:
An LOD (short for "logarithm of the odds") score is a statistical estimate of the relative probability that two loci (e.g., a disease-associated gene and another sequence of interest, such as a variant or another gene) are located near each other on a chromosome and are therefore likely to be inherited together. Traditionally, a LOD score of more than 3 has been deemed to indicate significant linkage. A LOD score of 3 indicates that the odds that the loci are linked are 1,000 times greater than the odds that they are not.

however, that there are significant (LOD more than 3) peaks on other chromosomes, so a breeder would also have to take these into account. Plant height is what is considered to be a **complex trait**, meaning a phenotype that is controlled by more than one gene. Most phenotypes are complex traits.

Mainscan plot of height

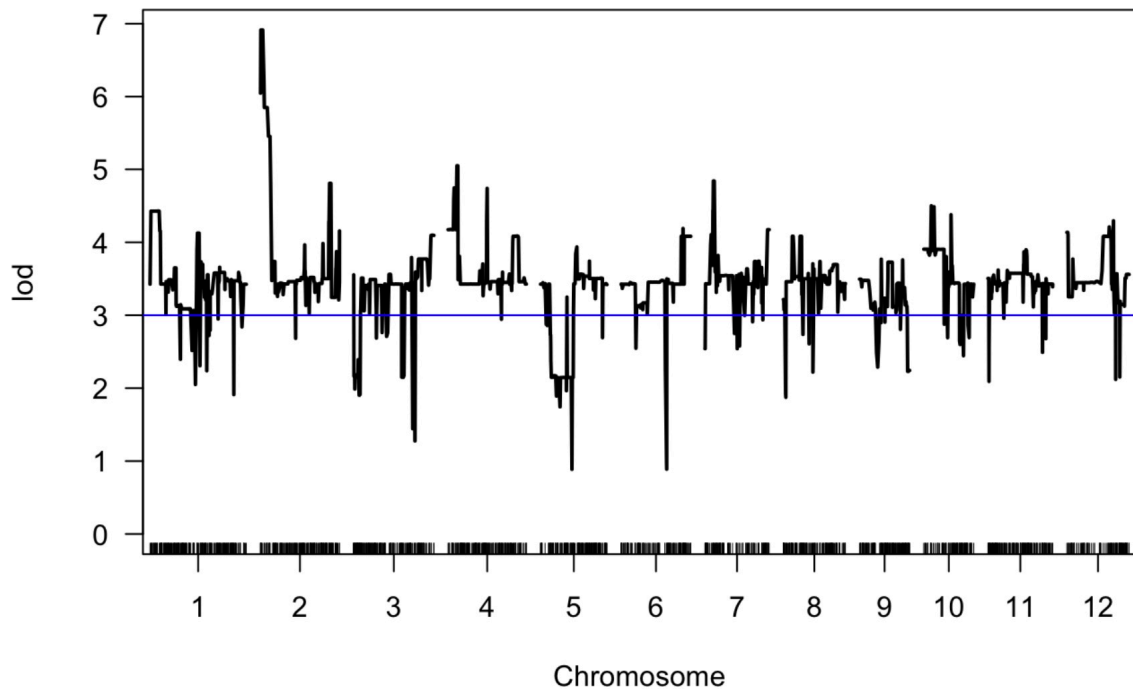


Figure 12: Mainscan for plant height

2.3 STUDENT ACTIVITY 3: GENOME-WIDE ASSOCIATION STUDIES / ASSOCIATION MAPPING

Materials for this activity were modified from materials developed by Dr. Julin Maloof for the course BIS180L "A Lab Course at UC Davis to introduce Genetics and Genomics majors to Bioinformatics"

2.3.1 BACKGROUND READING

In Student Activity 2, you learned how to use the statistical method of **Quantitative Trait Loci (QTL)** analysis to identify regions of the genome that are responsible for some phenotype, or trait, such as hair color or blood pressure.

Like QTL analyses, genome-wide association studies, or GWAS, is used to identify regions of the genome associated with complex traits, such as grain length or days to flowering. The major difference between a QTL

analysis and a GWAS study is that the data for a QTL study comes from a very controlled cross between two strains, in this case, of rice. GWAS, on the other hand, can examine the genomes of large, unrelated populations of, again in our case, rice plants.

Here's an overview of how GWAS works, generated with ChaptGPT:

1. **Study Population:** GWAS typically involves large study populations, often consisting of thousands or even tens of thousands of individuals. These individuals are usually divided into two groups: cases (individuals with the trait or disease of interest) and controls (individuals without the trait or disease).
2. **Genotyping:** Researchers collect DNA samples from the study participants and genotype them to identify millions of single nucleotide polymorphisms (SNPs) across the genome. SNPs are single-letter variations in the DNA code that can be used as genetic markers.
3. **Statistical Analysis:** The genotyping data is then subjected to statistical analysis to identify associations between specific SNPs and the trait or disease. This involves comparing the frequency of different genetic variants in cases versus controls.
4. **Genome-wide Significance Threshold:** To avoid false-positive results, a genome-wide significance threshold is established. Only associations that meet this threshold are considered statistically significant. This threshold accounts for the large number of tests conducted across the entire genome.
5. **Replication Studies:** Significant associations discovered in the initial GWAS are often subjected to replication in independent study populations to confirm the findings.
6. **Functional Annotation:** Once significant associations are identified, researchers may perform functional annotation to understand the biological mechanisms underlying these associations. This can involve exploring the potential impact of identified SNPs on gene expression or protein function.
7. **Biological Insights:** The results of GWAS can provide valuable insights into the genetic basis of traits and diseases. They can identify candidate genes and pathways that may be involved in the development of a particular condition.

GWAS have contributed to our understanding of the genetic underpinnings of various diseases and traits. However, it's essential to note that GWAS results provide statistical associations and do not necessarily establish causation. Additionally, the identified genetic variants often explain only a portion of the heritability of a trait or disease, and other factors, such as environmental influences, play a significant role.

GWAS have opened up new avenues for personalized medicine, risk prediction, and drug development. They continue to be a powerful tool in genetics and genomics research.

Figure 13 shows a very small sample of the rice genome genotype data being used for the GWAS activity described below.

	1_13147	1_73192	1_74969	1_75852	1_75953	1_91016	1_146625	1_149005	1_149754
NSFTV1	TT	TT	CC	GG	TT	AA	CC	TT	AA
NSFTV3	CC	CC	CC	AA	GG	0	CC	GG	TT
NSFTV4	CC	CC	CC	AA	GG	GG	CC	GG	TT
NSFTV5	CC	CC	TT	GG	GG	AA	TT	GG	TT
NSFTV6	CC	CC	CC	AA	GG	GG	CC	GG	TT
NSFTV7	TT	TT	CC	GG	TT	AA	CC	TT	AA
NSFTV8	TT	TT	CC	GG	TT	AA	CC	TT	AA
NSFTV9	TT	TT	CC	GG	TT	AA	CC	TT	AA
NSFTV10	TT	TT	CC	GG	TT	AA	CC	TT	AA

Figure 13: Sample of rice genome genotype data

The terms "NSFTV" in the first column refer to the accession identification number. NSF stands for the National Science Foundation, the federal organization that provides funding for much of the scientific research conducted in the United States. "TV" stands for "transgressive variations".

A rice accession refers to a specific variety or strain of rice (*Oryza sativa*) that is collected, preserved, and maintained in a gene bank or germplasm collection for research, breeding, and conservation purposes. These collections of rice accessions are critical resources for agricultural and scientific endeavors, as they represent the genetic diversity within the rice species.

Rice accessions can vary in terms of characteristics such as grain type (e.g., long-grain, short-grain, aromatic), growth habits, resistance to diseases and pests, tolerance to environmental conditions, and nutritional content. Scientists and plant breeders use these accessions to develop new rice varieties with desired traits, such as increased yield, improved taste, or enhanced resistance to specific stresses.

Preserving a wide range of rice accessions is important for ensuring genetic diversity, which can be critical for food security and adapting to changing environmental conditions. These accessions are typically stored under controlled conditions to maintain their viability and genetic integrity for future research and breeding efforts.

The values in the first column refer to a single nucleotide polymorphism, or SNP, located at 13147 base pairs on Chromosome 1.

Transgressive variation:

refers to the production of offspring with phenotypic traits or characteristics (such as flowering time or number of seeds per plant) that exceed those of the parents. This generally results from cooperation or interaction between the genes present in the two parental types.

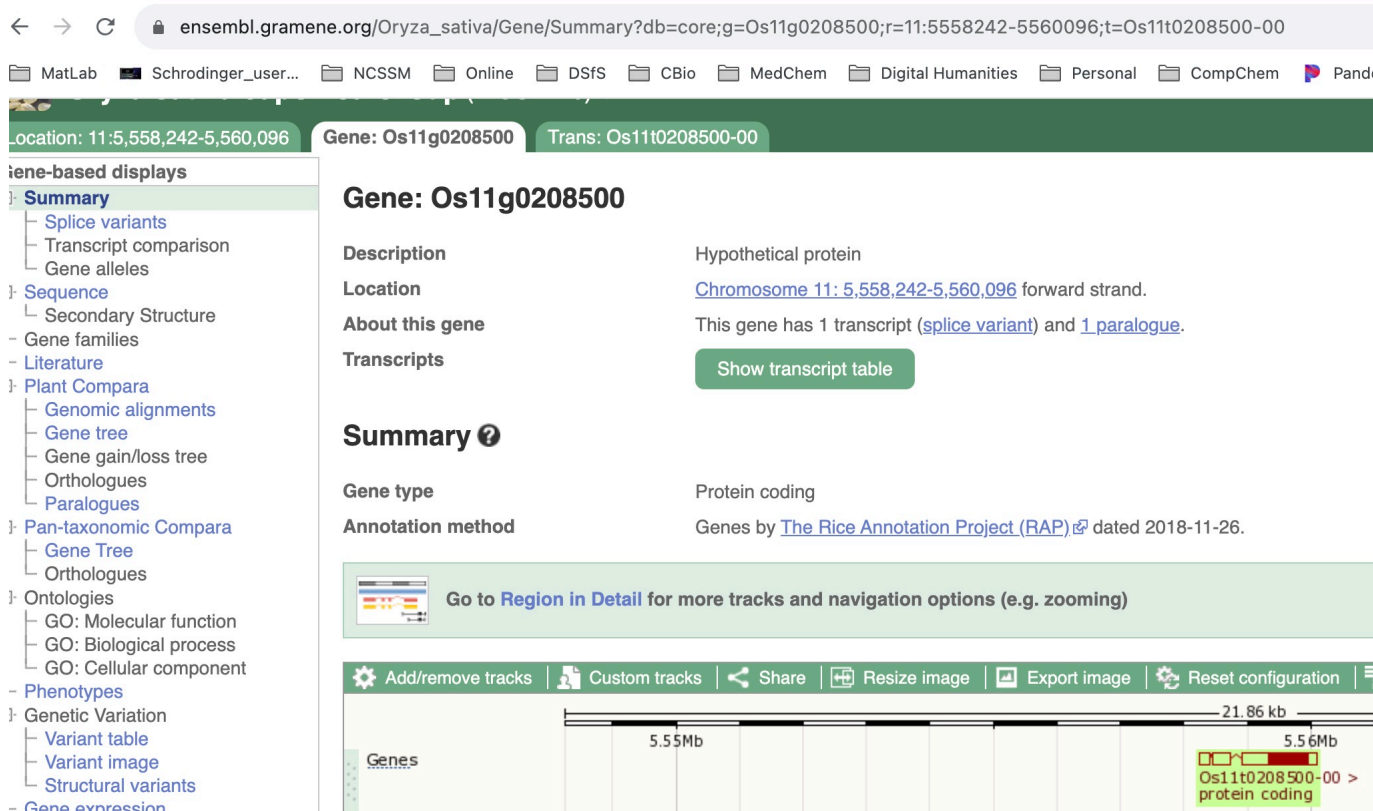


Figure 14: Genes located on Chromosome 1 at location 13147 basepairs

Figure 14 shows that there is a gene, Os11g0208500 (LOC_Os11g0208500) located in that region. This gene is a hypothetical protein, defined as a protein that is predicted to be expressed in an organism, but no evidence of its existence is known.

2.3.2 STUDENT ACTIVITY

For this activity, you are going to investigate a large dataset of 413 different strains of rice. Your task is to evaluate this dataset for one of a large number of phenotypes:

1. Alu.Tol: tolerance to aluminum
2. Flowering.time.at.Arkansas
3. Flowering.time.at.Faridpur
4. Flowering.time.at.Aberdeen
5. FT.ratio.of.Arkansas.Aberdeen
6. FT.ratio.of.Faridpur.Aberdeen
7. Culm.habit: A culm is the upright stem in the middle of a shoot (or tiller). The habit describes the growth form of a plant, comprising its size, shape, texture and orientation.
8. Leaf.pubescence: the onset of soft down or fine short hairs on the leaves and stems of plants. Many plants have pubescence designed to provide a tiny bit of shade to reduce the temperature of the leaves and stems and protect the leaves from losing too much water from transpiration.
9. Flag.leaf.length: the flag leaf is the last leaf to emerge in plant growth.

10. Flag.leaf.width
11. Awn.presence: an awn is a stiff bristle, especially one of those growing from the ear or flower of barley, rye, rice, and many grasses.
12. Panicle.number.per.plant: a panicle is a loose branching cluster of flowers, as in oats.
13. Plant.height
14. Panicle.length
15. Primary.panicle.branch.number
16. Seed.number.per.panicle
17. Florets.per.panicle: a floret is one of the small flowers making up a composite flower head. Most people know a floret from a head of broccoli.
18. Panicle.fertility
19. Seed.length
20. Seed.width
21. Seed.volume
22. Seed.surface.area
23. Brown.rice.seed.length
24. Brown.rice.seed.width
25. Brown.rice.surface.area
26. Brown.rice.volume
27. Seed.length.width.ratio
28. Brown.rice.length.width.ratio
29. Amylose.content: amylose is the crystallizable form of starch, consisting of long unbranched polysaccharide chains.
30. Alkali.spreading.value
31. Protein.content

For this activity, you are provided with two documents:

1. Rice Genome Wide Area Association Studies (GWAS): An Annotated Guide
2. Rice Genome Wide Area Association Studies (GWAS): Data Analysis

The first document provides fully functional R code, written in RMarkdown, with the phenotype "Seed.length.width.ratio" as the featured target. The Data Analysis document provides working code – your task is to pick one of the phenotypes from the list above, substitute it for Seed.length.width.ratio as appropriate, and "knit" the RMarkdown code. A short podcast is available to help you if you are new to RMarkdown. (<https://youtu.be/CP1x1o6q7RY>).

[NOTE! This is a VERY large dataset, and when you "knit" your RMarkdown file, you should expect it to take 5-10 minutes, perhaps longer, depending on your machine!]

In your Data Analysis document, you are also expected to write some narrative about your findings. What does the data tell you?

USEFUL RESOURCES

Below is a list of useful resources, primarily websites, for use in studying plant genomics, with a particular focus on rice. These resources are in no particular order.

1. **Gramene:** comparative plant genomics for crops and model organisms (<https://www.gramene.org/>)
 - (a) **Gramene archive:** older version of Gramene, with additional resources such as QTLs (<https://archive.gramene.org/>)
2. **PlantDB:** tools and resources for plant genomics (<https://www.plantgdb.org/>)
3. **Oryzabase:** comprehensive database of rice information, developed in Japan (<https://shigen.nig.ac.jp/rice/oryzabase/>)
4. **Rice Genome Annotation Project:** rice genome annotation project, funded by the National Science Foundation (<http://rice.uga.edu/>)
5. **Rice Genome Hub:** a collection of genomes, datasets, and various tools (<https://rice-genome-hub.southgreen.fr/>)
6. **UniProt:** general-purpose database for conducting studies on proteins (<https://www.uniprot.org/>)
7. Tutorial on QTL mapping from Sue McClatchy at the Jackson Lab (mouse genomics) <https://smcclatchy.github.io/mapping/>

TEACHER NOTES

NOTE! Answer keys for the various activities are available by emailing Robert Gotwals at gotwals@ncssm.edu

4.1 USING THESE MATERIALS

These materials can be used in a variety of ways. Each of the three student activities – Comparative Genomics, Quantitative Trait Loci analyses, and Genome-Wide Association Studies – are independent of each other, although they can build on each other.

All of these materials assume that the student knows some basic concepts from genetics, including the structure and role of DNA, the Central Dogma of Biology (replication, transcription, translation), and how genes contain alleles that come from parental DNA. It is helpful if students have done activities such as using/creating Punnet Squares and/or looking at pedigree charts.

Some estimates of required class time are described below:

1. **Comparative Genomics:** this activity requires 2-4 class periods, and assumes that at least one of these is on a block schedule and/or is a lab block.

There are a number of excellent, and relatively short, YouTube videos on relevant topics. A short list is provided below, but a simple YouTube search will yield a large number of options from which you can choose.

- (a) What is Genomics?: <https://www.youtube.com/watch?v=mmgIClg0Y1k> (6:19 minutes)
- (b) What is Comparative Genomics: <https://www.youtube.com/watch?v=Irs6SGAu1YM> (2:29 minutes)
- (c) Rice as a model crop for Comparative Genomics: <https://www.youtube.com/watch?v=qTKMga21e4A> (2:16 minutes)
- (d) **Quantitative Trait Loci:** For this activity, students need to write code using a programming language called "R", and the interface to R, RStudio. R/RStudio can be downloaded and installed on individual machines (Google "download R" and "download RStudio" to find the appropriate resource), but it is easiest to use the RStudio Cloud resource, located at <https://posit.cloud/>. Students do need to create an account, but this is free.

A short video on how to start the Rice QTL programming activity is found on the Canvas resources page, at this link: <https://youtu.be/WhzfxAlu2jo>

For students new to programming, this activity might take a little longer, but the Annotated Guide should allow a student to do the coding part of this activity in one lab block (90 minutes) or less. Introductory instruction is best done by using the YouTube videos listed here:

- i. Quantitative vs Qualitative - Tales from the Genome: https://www.youtube.com/watch?v=sww_gzxjUtg (1:46 minutes)
- ii. What is Gene Mapping?: https://www.youtube.com/watch?v=cAFAjYq_68I (5:42 minutes)
- iii. Multiple loci (S), quantitative trait loci (QTL): <https://www.youtube.com/watch?v=1FuP-kiPuxk> (14:11 minutes)

(e) **Genome-Wide Association Studies:**

4.2 MEETING NORTH CAROLINA STANDARDS

These lab materials can be used to directly and/or indirectly address several North Carolina Standard Course of Study requirements for biology. Ones of particular note are highlighted in **bold**.

4.2.1 STRAND: EVOLUTION AND GENETICS

1. **NCES.Bio.3.1 - Explain how traits are determined by the structure and function of DNA.**
 - (a) NCES.Bio.3.1.1 - Explain the double-stranded, complementary nature of DNA as related to its function in the cell.
 - (b) NCES.Bio.3.1.2 - Explain how DNA and RNA code for proteins and determine traits.
 - (c) **NCES.Bio.3.1.3 - Explain how mutations in DNA that result from interactions with the environment (i.e. radiation and chemicals) or new combinations in existing genes lead to changes in function and phenotype.**
2. NCES.Bio.3.2 - Understand how the environment, and/or the interaction of alleles, influences the expression of genetic traits.
 - (a) NCES.Bio.3.2.1 - Explain the role of meiosis in sexual reproduction and genetic variation.
 - (b) NCES.Bio.3.2.2 - Predict offspring ratios based on a variety of inheritance patterns (including: dominance, co-dominance, incomplete dominance, multiple alleles, and sex-linked traits).
 - (c) NCES.Bio.3.2.3 - Explain how the environment can influence the expression of genetic traits.
3. NCES.Bio.3.3 - Understand the application of DNA technology.
 - (a) NCES.Bio.3.3.1 - **Interpret how DNA is used for comparison and identification of organisms.**
 - (b) NCES.Bio.3.3.2 - Summarize how transgenic organisms are engineered to benefit society.
 - (c) NCES.Bio.3.3.3 - Evaluate some of the ethical issues surrounding the use of DNA technology (including: cloning, genetically modified organisms, stem cell research, and Human Genome Project).

4.2.2 NEXT GENERATION SCIENCE STANDARDS - BIOLOGY

These materials can also be used to address a number of standards under the Next Generation Science Standards for Life Sciences. A link to the full document is provided below. Some of the highlights for Inheritance and Variation of Traits are shown below.

Next Generation Science Standards full document: <https://www.nextgenscience.org/sites/default/files/HS%20LS%20topics%20combined%206.13.13.pdf>

It should be noted that the NGSS also addresses a number of skills and competencies that accompany the factual description of what students should know. Figure 15 shows the Science and Engineering Practices that accompany the section on Inheritance and Variation of Traits.

Science and Engineering Practices
<p>Asking Questions and Defining Problems Asking questions and defining problems in 9-12 builds on K-8 experiences and progresses to formulating, refining, and evaluating empirically testable questions and design problems using models and simulations.</p> <ul style="list-style-type: none"> Ask questions that arise from examining models or a theory to clarify relationships. (HS-LS3-1)
<p>Developing and Using Models Modeling in 9-12 builds on K-8 experiences and progresses to using, synthesizing, and developing models to predict and show relationships among variables between systems and their components in the natural and designed worlds.</p> <ul style="list-style-type: none"> Use a model based on evidence to illustrate the relationships between systems or between components of a system. (HS-LS1-4)
<p>Analyzing and Interpreting Data Analyzing data in 9-12 builds on K-8 experiences and progresses to introducing more detailed statistical analysis, the comparison of data sets for consistency, and the use of models to generate and analyze data.</p> <ul style="list-style-type: none"> Apply concepts of statistics and probability (including determining function fits to data, slope, intercept, and correlation coefficient for linear fits) to scientific and engineering questions and problems, using digital tools when feasible. (HS-LS3-3)
<p>Engaging in Argument from Evidence Engaging in argument from evidence in 9-12 builds on K-8 experiences and progresses to using appropriate and sufficient evidence and scientific reasoning to defend and critique claims and explanations about the natural and designed world(s). Arguments may also come from current scientific or historical episodes in science.</p> <ul style="list-style-type: none"> Make and defend a claim based on evidence about the natural world that reflects scientific knowledge, and student-generated evidence. (HS-LS3-2)

Figure 15: Science and Engineering Practices [7]

1. HS-LS1-4. Use a model to illustrate the role of cellular division (mitosis) and differentiation in producing and maintaining complex organisms. [Assessment Boundary : Assessment does not include specific gene control mechanisms or rote memorization of the steps of mitosis.]
2. HS-LS3-1. Ask questions to clarify relationships about the role of DNA and chromosomes in coding the instructions for characteristic traits passed from parents to offspring. [Assessment Boundary : Assessment does not include the phases of meiosis or the biochemical mechanism of specific steps in the process.]
3. HS-LS3-2. Make and defend a claim based on evidence that inheritable genetic variations may result from: (1) new genetic combinations through meiosis, (2) viable errors occurring during replication, and/or (3) mutations caused by environmental factors. [Clarification Statement: Emphasis is on using data to support arguments for the way variation occurs.] [Assessment Boundary : Assessment does not include the phases of meiosis or the biochemical mechanism of specific steps in the process.]
4. HS-LS3-3. Apply concepts of statistics and probability to explain the variation and distribution of expressed traits in a population.

4.2.3 ADDITIONAL CONSIDERATIONS

Figure 16 shows an excellent schematic of how some of the activities in this module fit into the larger scenario. These activities are taking areas such as genomics and, to some extent, phenomics, and integrating that data with the eventual goal of being able to improve "varieties with higher yield and nutritious quality as well as enhanced tolerance to biotic and abiotic stress." [13]. These activities specifically address the gray ovals, "Identification of Genes/QTLs/markers" and "Trait identification". We also are having students engage in "Prediction of QTLs and natural allelic variation."

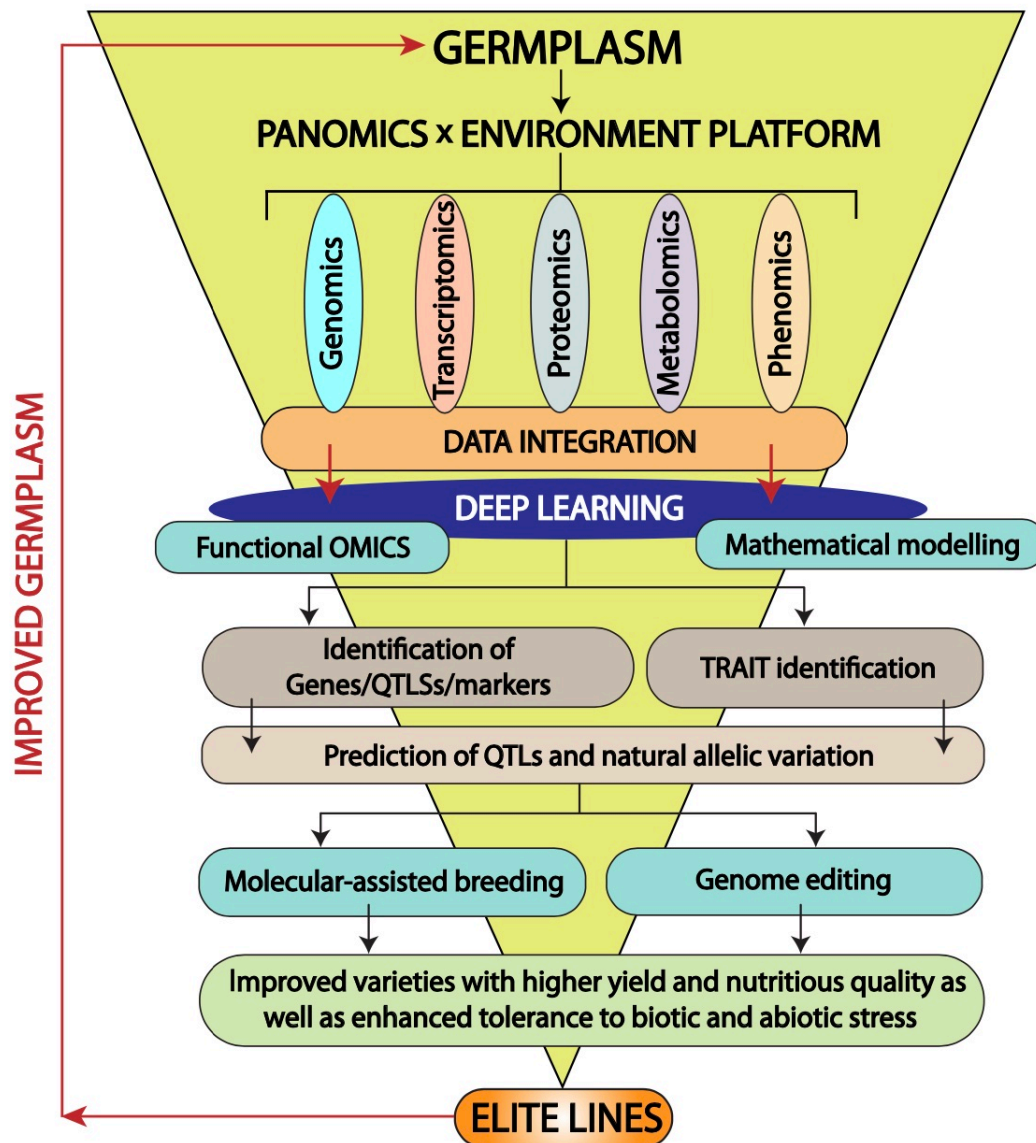


Figure 16: Panomics to improved germplasm [13]

Armed with these skills, students could engage in more detailed research into areas such as molecular-assisted breeding and genome editing. These are ideal topics for longer-term individual and group projects, all of which

might be competitive at local, regional, state, national, or international science competitions.

For students interested in medicine and health care, ventures into the role of panomics in personalized medicine might be of interest. Articles such as *Panomics for Precision Medicine* are worth reading! [13] For example, one of the authors of this document (RRG) had his DNA sequenced by the UNC Eshleman School of Pharmacy. Based on the genetic markers, QTLs, and other data found, they were able to provide a report showing most of the more commonly used drugs in medicine. Based on the sequenced markers, the report – a pharmacogenomics, PGx report – shows that this person does not have the correct genetic profile to be able to take the pain-relieving drug acetaminophen.

PGx Report - Pain Management

Type: Anti-inflammatory Agent, Analgesic, Antipyretic

Drug Class	Generic	Primary Mechanism Involved	Other Mechanisms Involved	May Have Decreased Efficacy	Used As Directed	Increased Adverse Outcomes
The Nonsteroidal Antiinflammatory Drugs (NSAIDs)						
Acetic acid derivatives	Indomethacin	CYP2C9	CYP2C19			
Enolic acid (Oxicam) derivatives	Meloxicam	CYP2C9	CYP3A4, CYP3A5		✓	
	Piroxicam	CYP2C9	CYP3A4, CYP3A5		✓	
	Tenoxicam	CYP2C9			✓	
	Lornoxicam	CYP2C9			✓	
Selective COX-2 inhibitors (Coxibs)	Etoricoxib	CYP3A4	CYP3A5, CYP2C9, CYP2D6		✓	
	Parecoxib	CYP2C9	CYP3A4, CYP3A5		✓	
	Celecoxib	CYP2C9	CYP2C19		✓	
Propionic acid derivatives	Ibuprofen	CYP2C9	CYP2C19		✓	
	Flurbiprofen	CYP2C9			✓	
	Ketoprofen	CYP3A4	CYP2C9, CYP3A5		✓	
	Fenoprofen	CYP2C9			✓	
	Vicoprofen	CYP2D6	CYP3A4		✓	
	Naproxen	CYP2C9			✓	
Anthranilic acid derivatives (Fenamates)	Metenamic acid	CYP2C9			✓	
The Non-NSAIDs Analgesic	Acetaminophen	UGT1A1, SULT1A1, GSHs	CYP3A4, CYP3A5, CYP2D6			

Figure 17: Sample entry of a pharmacogenomics report

REFERENCES

- [1] English | National Agriculture and Food Research Organization — naro.go.jp. <https://www.naro.go.jp/english/index.html>. [Accessed 27-07-2023].
- [2] Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.
- [3] Juan D Arbelaez, Laura T Moreno, Namrata Singh, Chih-Wei Tung, Lyza G Maron, Yolima Ospina, César P Martinez, Cécile Grenier, Mathias Lorieux, and Susan McCouch. Development and gbs-genotyping of introgression lines (ils) using two wild species of rice, *O. meridionalis* and *O. rufipogon*, in a common recurrent parent, *O. sativa* cv. *curinga*. *Molecular Breeding*, 35:1–18, 2015.
- [4] Michael D Bennett, Ilia J Leitch, H James Price, and J Spencer Johnston. Comparisons with *Caenorhabditis* (approximately 100 mb) and *Drosophila* (approximately 175 mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 mb and thus approximately 25% larger than the *Arabidopsis* genome initiative estimate of approximately 125 mb. *Annals of botany*, 91(5):547–557, 2003.
- [5] Karl W Broman and Saunak Sen. *A Guide to QTL Mapping with R/qtl*, volume 46. Springer, 2009.
- [6] Geoffrey M Cooper and RE Hausman. The chloroplast genome. *The Cell: A Molecular Approach*, 2000.
- [7] National Research Council et al. Next generation science standards: For states, by states. 2013.
- [8] Gurdev S Khush. Origin, dispersal, cultivation and variation of rice. *Plant molecular biology*, 35:25–34, 1997.
- [9] Mark A Larkin, Gordon Blackshields, Nigel P Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal w and clustal x version 2.0. *bioinformatics*, 23(21):2947–2948, 2007.
- [10] Susan McCouch. Personal Communication, 2023.
- [11] Jianzong Nan, Xiaomin Feng, Chen Wang, Xiaohui Zhang, Rongsheng Wang, Jiaxin Liu, Qingbo Yuan, Guoqiang Jiang, and Shaoyang Lin. Improving rice grain length through updating the *gs3* locus of an elite variety kongyu 131. *Rice*, 11(1):21, 2018.
- [12] Luca Schulz, Zhijun Guo, Jan Zarzycki, Wieland Steinchen, Jan M Schuller, Thomas Heimerl, Simone Prinz, Oliver Mueller-Cajar, Tobias J Erb, and Georg KA Hochberg. Evolution of increased complexity and specificity at the dawn of form i rubiscos. *Science*, 378(6616):155–160, 2022.
- [13] Wolfram Weckwerth, Arindam Ghatak, Anke Bellaire, Palak Chaturvedi, and Rajeev K Varshney. Panomics meets germplasm. *Plant Biotechnology Journal*, 18(7):1507–1525, 2020.
- [14] Wikipedia contributors. *Oryza longistaminata* — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Oryza_longistaminata&oldid=1126716401, 2022. [Online; accessed 8-August-2023].

- [15] Wikipedia contributors. Oryza nivara — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Oryza_nivara&oldid=1064075851, 2022. [Online; accessed 8-August-2023].
- [16] Wikipedia contributors. List of rice cultivars — Wikipedia, the free encyclopedia, 2023. [Online; accessed 9-October-2023].
- [17] Wikipedia contributors. Oryza glaberrima — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Oryza_glaberrima&oldid=1163670434, 2023. [Online; accessed 8-August-2023].
- [18] Wikipedia contributors. Oryza rufipogon — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Oryza_rufipogon&oldid=1166153657, 2023. [Online; accessed 8-August-2023].
- [19] Wikipedia contributors. Oryza sativa — Wikipedia, the free encyclopedia, 2023. [Online; accessed 8-August-2023].
- [20] Fereshteh T Yazdi, Susanne M Clee, and David Meyre. Obesity genetics in mouse and human: back and forth, and back again. *PeerJ*, 3:e856, 2015.
- [21] Keyan Zhao, Chih-Wei Tung, Georgia C Eizenga, Mark H Wright, M Liakat Ali, Adam H Price, Gareth J Norton, M Rafiqul Islam, Andy Reynolds, Jason Mezey, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nature communications*, 2(1):467, 2011.

Rice Quantitative Trait Loci (QTL): An Annotated Guide

Robert Gotwals, Computational Science Educator, NCSSM

Last compiled on October 08, 2023

About this *Guide*

The purpose of this guide is to provide a line-by-line, code chunk-by-code chunk description of the R code needed to perform a quantitative trait loci (QTL) analysis of breeding data between two species of rice (*Oryza sativa*): *Curinga* x *O. rufipogon*. This *Guide* does not provide a description of QTL analyses; that is found in the curricular materials. This *Guide* will annotate the required code for conducting the analyses.

This *Guide* assumes that you have downloaded and installed the R software (<https://cran.rstudio.com/> (<https://cran.rstudio.com/>)) and the interface to R, RStudio (<https://posit.co/download/rstudio-desktop/> (<https://posit.co/download/rstudio-desktop/>)). It also assumes that, in R, you have installed a “package” called “qtl”. To install this, run the command “install.packages(“qtl”) at the console command line. You only have to do this once.

Initial setup

There are three steps here:

1. Load the “qtl” package/library. This assumes that you have already installed the package, a one-time process.
2. Remove any previous items in memory. The “Global Environment” window on the top righthand side of RStudio should be empty after running this command.
3. The QTL dataset for this analyses is large by R standards, but small by genomics standards. Regardless, you need to set the system environment with enough memory to handle the data.

You should also add some documentation to your file by using an asterisk. Well-documented code is really important if you want to keep your job as a programmer/data scientist!

```
# Your name
# Today's date
# Rice QTL analysis
#
setwd("/Users/gotwals/Desktop/SygentaPlants")
library(qtl)
rm(list=ls())
Sys.setenv(VROOM_CONNECTION_SIZE="500000")
#
```


Load Data

Now we are ready to load the data. The data has been upload to a server maintained by the Department of Chemistry at NCSSM, as a “comma-separated values” (csv) file. The command **read.cross** states that you are reading a CSV file located at <http://chemistry.ncssm.edu> (<http://chemistry.ncssm.edu>), in the data/gwas directory. The file name is CuRUFCSL_QTL.csv. Then, the code states that there are three genotypes coded in

the data: AA, H (heterozygous), and BB. Some of the genotypes are missing, and those are indicated by the **na.strings** command, with a space between the quotes. Finally, the code states that there are two different alleles, A and B. It is likely that you will receive a warning, but this will not have any impact on your analyses. You will also receive a brief summary of the cross.

	A	B	C	D	E	F	G	H	I
1	Flow	Height	Tillers	Panicles	Pericarp	X8144	X19846	X20215	id1000955
2						1	1	1	1
3						1.2609	2.870808	2.918664	4.179784
4	81	97.6	94	82	White	AA	AA	AA	AA
5	80	122.4	75	60	Red	BB	BB	BB	BB
6	68	92.2	78	58	Red	AA	AA	AA	AA
7	85	93.4	76	70	White	BB	BB	BB	BB
8	77	110.8	68	62	White	AA	BB	BB	BB
9	79	104.2	70	62	Red	AA	AA	AA	AA
10	81	107	58	43	White	AA	AA	AA	AA
11	77	107.2	74	51	White	AA	AA	AA	AA
12	76	99.8	69	56	White	AA	AA	AA	AA
13	76	97.6	64	56	White	AA	AA	AA	AA

Screenshot of phenotypes.

 Screenshot of QTL data.

Screenshot of QTL data.

```
cross <- read.cross("csv", file="http://chemistry.ncssm.edu/data/gwas/CuRUFCSL_QTL.csv",
  genotypes = c("AA","H", "BB"), alleles = c("A", "B"))
```

```
## --Read the following data:
## 256 individuals
## 1769 markers
## 5 phenotypes
## --Cross type: f2
```

It is also helpful to display the names of the phenotypes (pay attention to case!). Finally, it's instructive to request a summary of the cross data, which provides a detailed analyses of the number of crosses, descriptions of the phenotypes and genotypes, etc.

```
names(cross$pheno)
```

```
## [1] "Flow"      "Height"    "Tillers"   "Panicles" "Pericarp"
```

```
summary(cross)
```



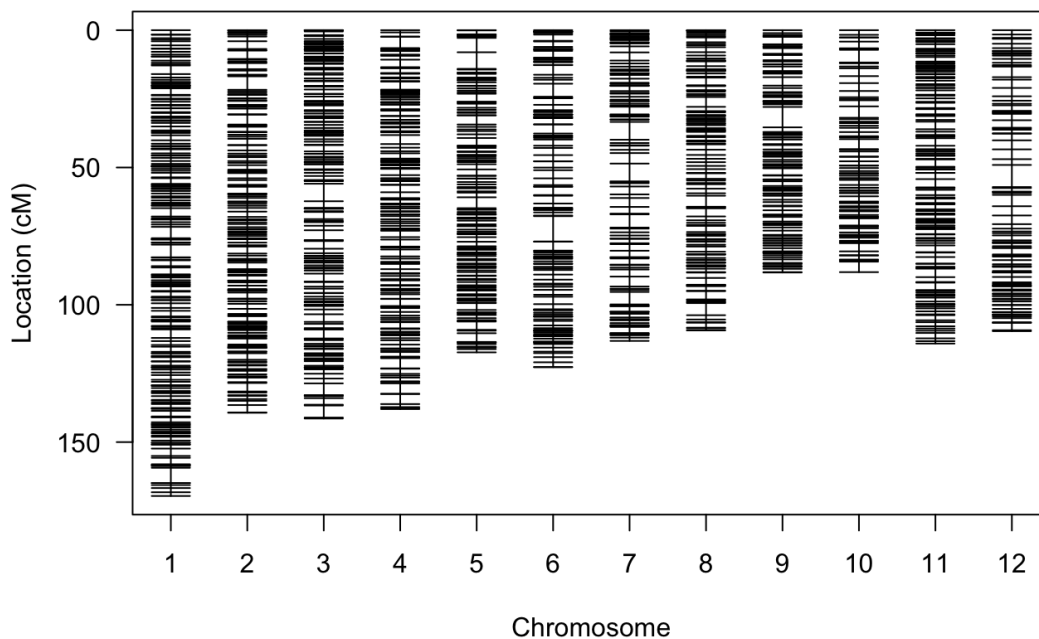
```
##      F2 intercross
##
##      No. individuals:    256
##
##      No. phenotypes:     5
##      Percent phenotyped: 100 100 100 100 100
##
##      No. chromosomes:    12
##      Autosomes:          1 2 3 4 5 6 7 8 9 10 11 12
##
##      Total markers:       1769
##      No. markers:         213 183 185 174 158 133 111 141 127 89 152 103
##      Percent genotyped:   27.2
##      Genotypes (%):       AA:86.9 AB:0.7 BB:12.4 not BB:0.0 not AA:0.0
```

Genetic Markers

Next, we want to see a map of the genetic markers, and, even though this is a relatively small dataset, there are almost 1800 markers. The **plot.map** command shows all of the markers by chromosome. Notice that they are not evenly spaced!

```
plot.map(cross)
```

Genetic map



Running preliminary files

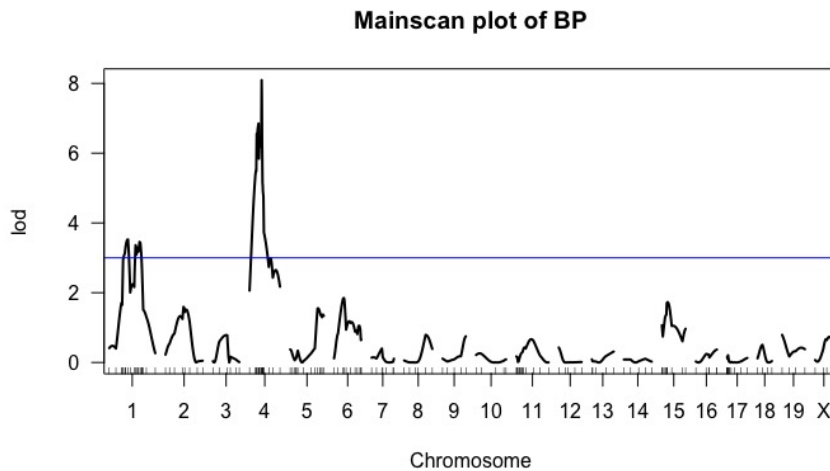
Two preliminary files need to be generated before performing the main analysis – the “mainscan” – of your data. Both of these use a machine learning method known as Hidden Markov to check the actual data against what the actual data should be, a check on genotyping errors. The options for both are the same, and are described below for the curious reader! These notes come from the QTL package documentation.

1. `calc.genoprob`: Uses the hidden Markov model technology to calculate the probabilities of the true underlying genotypes given the observed multipoint marker data, with possible allowance for genotyping errors.
 - a. `step`: Maximum distance (in cM) between positions at which the simulated genotypes will be drawn, though for `step=0`, genotypes are drawn only at the marker locations.
 - b. `off.end`: Distance (in cM) past the terminal markers on each chromosome to which the genotype simulations will be carried.
 - c. `error.prob`: Assumed genotyping error rate used in the calculation of the penetrance $\Pr(\text{observed genotype} \mid \text{true genotype})$.
 - d. `map.function`: indicates whether to use the Haldane, Kosambi, Carter-Falconer, or Morgan map function when converting genetic distances into recombination fractions.
 - e. `stepwidth`: indicates whether the intermediate points should with fixed or variable step sizes
2. `sim.genoprob`: Uses the hidden Markov model technology to simulate from the joint distribution $\Pr(g \mid O)$ where g is the underlying genotype vector and O is the observed multipoint marker data, with possible allowance for genotyping errors. The option “`n.draws`” describes the number of simulated probabilities to calculate.

```
cross <- calc.genoprob(cross, step=2.0, off.end=0.0, error.prob=1.0e-4, map.function= "haldane", stepwidth = "fixed")
cross <- sim.geno(cross, step=2.0, off.end=0.0, error.prob=1.0e-4, map.function= "haldane", stepwidth = "fixed", n.draws=16)
```

Running a mainscan for plant height

Now we are ready for the main goal of the analyses: to find where on one or more chromosomes there might be genes that are responsible for a specific trait, or phenotype. The graphic below comes from mouse data, and we are looking to see where genes that control blood pressure (BP) might be located. On the x-axis, we see the 19 chromosomes of a mouse, as well as the X-chromosome. On the y-axis is a LOD score. From the National Human Genome Research Institute (<https://www.genome.gov/>): “A LOD (short for “logarithm of the odds”) score is a statistical estimate of the relative probability that two loci (e.g., a disease-associated gene and another sequence of interest, such as a variant or another gene) are located near each other on a chromosome and are therefore likely to be inherited together.”



Screenshot of a mainscan graphic.

For a LOD score to be significant, we typically use a cutoff of 3. There is an obvious peak on Chromosome 4, so someone hunting for the BP gene would focus most of their attention there. There is also some activity on Chromosome 1, so blood pressure is a polygenetic – more than one gene – trait. Attention would also need to be paid to Chromosome 1.

For this analyses, recall that there are four numerical phenotypes, one of them being plant height. This example demonstrates that analyses. You will then have the opportunity to analyze the other three.

The command to do a mainscan is `**scanone*`. We are scanning the cross data, and the phenotype of interest is in Column 2 of the dataset. We are using a simple “normal” model and the “expectation-maximization” (em) method. There are, as you might suspect, different algorithms that could be applied to this analyses, but EM is the most common.

There is a second analyses we can do, called a **permutation** test. This test basically tears the data apart and puts it back together. In other words, “a permutation tests shuffles genotypes and phenotypes, essentially breaking the relationship between the two.” (<https://smcclatchy.github.io/mapping/06-perform-perm-test/>) (<https://smcclatchy.github.io/mapping/06-perform-perm-test/>). We specify the number of permutations to run, in this case 100. For a more thorough analyses, one would run 500, 1000, or more permutations. 100 is enough for this particular dataset.

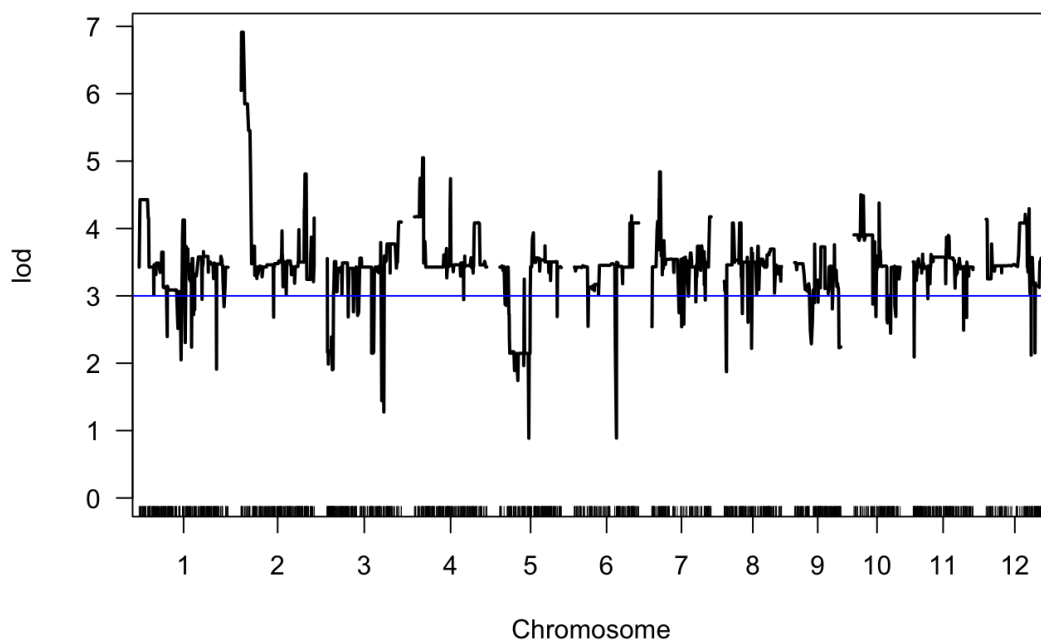
Once we run the scanones, we can plot the results. A simple `plot` command does the trick, and we can add a graph title using the `main` command. We might also want to add a threshold line at a LOD value of 3 If there are no significant QTLs, no line will be plotted as there are no peaks above that score.

```
cross.scanheight <- scanone(cross, pheno.col=2, model = "normal", method="em")
cross.scanheight.perm <- scanone(cross, pheno.col=2, model="normal", method="em", n.perm
=100)
```

```
## Permutation 5
## Permutation 10
## Permutation 15
## Permutation 20
## Permutation 25
## Permutation 30
## Permutation 35
## Permutation 40
## Permutation 45
## Permutation 50
## Permutation 55
## Permutation 60
## Permutation 65
## Permutation 70
## Permutation 75
## Permutation 80
## Permutation 85
## Permutation 90
## Permutation 95
## Permutation 100
```

```
plot(cross.scanheight, main="Mainscan plot of height")
lodline <- -3
abline(h=lodline, col="blue")
```

Mainscan plot of height



If it is the case that you have one or more significant QTLs – those with a LOD score of 3 or greater – you might want to look at effect plots. To do that, you first need to look at a summary of your QTLs. The **summary** command will show you that. The **alpha** option says only look at the QTLs that are 95% significant. You might need to change this number to 90% (0.10) or lower.

The **summary** command will show you the ID of the closest marker, the chromosome number, the location in centiMorgans (cM), and the LOD score. You will need that information for the next step.

```
summary(cross.scanheight, perm=cross.scanheight.perm, alpha=0.05)
```

```
##          chr    pos  lod
## c1.loc14    1  15.26 4.43
## c2.loc2     2   3.12 6.91
## c3.loc138   3 139.88 4.10
## c4.loc16    4  16.24 5.05
## c5.loc64    5  65.02 3.94
## X6851172    6 110.02 4.19
## c7.loc16    7  18.62 4.84
## c8.loc16    8  17.47 4.08
## X9469699    9  39.65 3.77
## X10099158  10  17.25 4.50
## X11465012  11  68.91 3.90
## X12852964  12  82.58 4.29
```

```
summary(cross.scanheight, perm=cross.scanheight.perm)
```

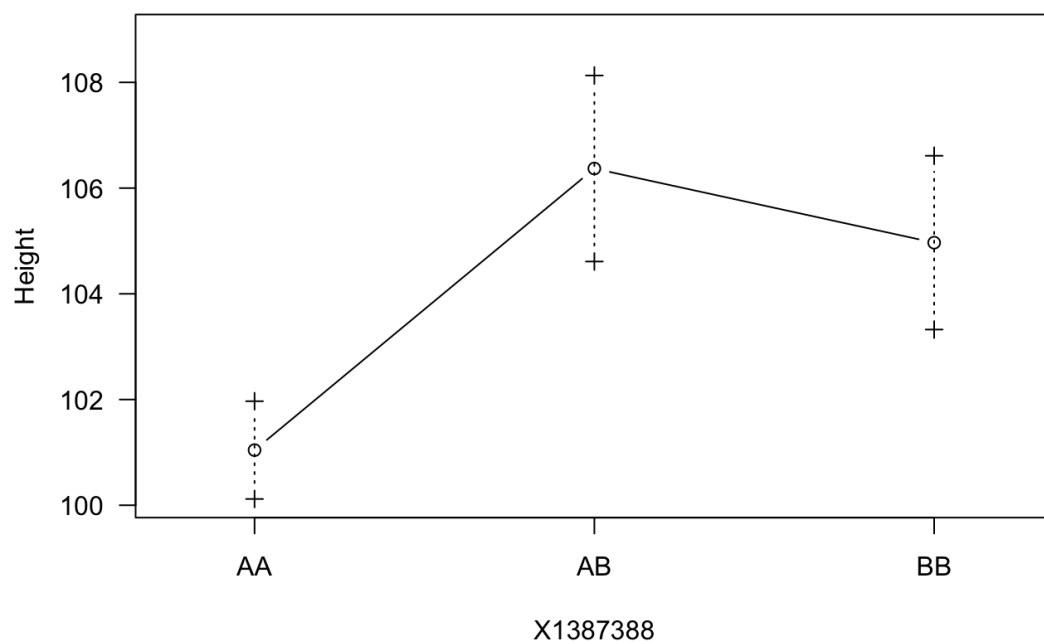
```
##          chr    pos  lod
## c1.loc14    1  15.26 4.43
## c2.loc2     2   3.12 6.91
## c3.loc138   3 139.88 4.10
## c4.loc16    4  16.24 5.05
## c5.loc64    5  65.02 3.94
## X6851172    6 110.02 4.19
## c7.loc16    7  18.62 4.84
## c8.loc16    8  17.47 4.08
## X9469699    9  39.65 3.77
## X10099158  10  17.25 4.50
## X11465012  11  68.91 3.90
## X12852964  12  82.58 4.29
```

Running an effect plot for plant height

Based on your QTL information, you modify the code below to reflect the QTL data. We found two significant QTLs, one on Chromosome 2 at 3.12 cM and one on Chromosome 4 at 16.4 cM. We use **find.marker** to identify those, give them a name (height1 and height2), then use **effectplot** to see the results. Note that you have to specify the column number for the phenotype, in our case height is in Column 2. You might notice that plants with the BB genome tend to be taller! AA plants are almost 4 centimeters shorter than BB plants.

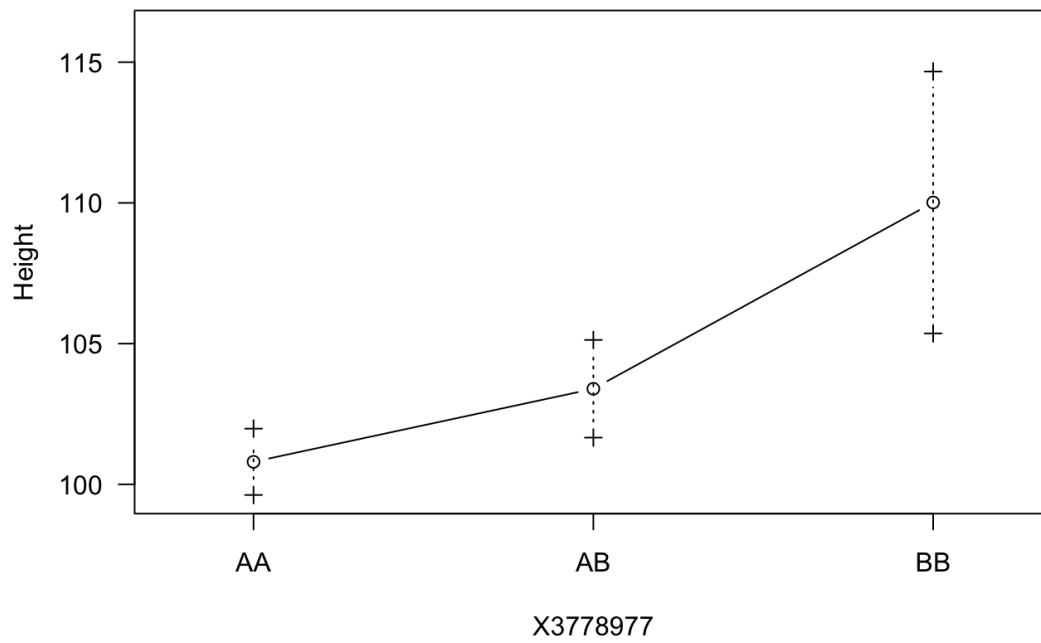
```
height1 <- find.marker(cross, chr=2, pos=3.12)
effectplot(cross, pheno.col=2, mname1=height1, main="Effect plot for height1")
```

Effect plot for height1



```
height2 <- find.marker(cross, chr=4, pos=16.24)
effectplot(cross, pheno.col=2, mname1=height2, main="Effect plot for height2")
```

Effect plot for height2



Now you are ready to try this on your own. It is recommended that you do a mainscan (and subsequent effect plots) for height to ensure that you get the same results. NOTE! Because some of these calculations use random number techniques such as Hidden Markov methods, your results may not be exactly the same! Then, consider modifying your code to find QTLs for "day to flow" (flow), "tillers", and "panicles". Note that there might not be QTLs for one or more of these phenotypes.

A skeleton "starter code" is available to guide your coding work.