A Problem Based Learning Exercise on Food Security: Understanding the Role of Genomic Variation and Plant Breeding

Student Handout

Jolie Wax[1], Zhu Zhuo[2], Anna Bower[1], Jessica Cooper[1], Susan Gachara[1], Isaac Kamweru[1], Terrance Mhora[1], Sabari Nath Neerukonda[2], Danielle Novick[1], Julia Winkeler[1], Todd Yoder[1], and Randall J. Wisser[1,*]

[1]Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716
[2]Department of Animal and Food Sciences, University of Delaware, Newark, DE 19716

[*]**Corresponding Author:** Randall J. Wisser, 531 S. College Ave, 152 Townsend Hall, Newark, DE 19716, 302-831-1356, rjw@udel.edu

PBL Stage I: Food Security

With over 200,000 people added to the world each day, the global population is projected to reach 9.7 billion by 2050 (United Nations, 2015). This growth, coupled with effects due to climate change and reductions in arable land, places a greater demand on global food production. Currently, food insecurity and malnutrition affect nearly 795 million people worldwide (FAO, IFAD & WFP, 2015). Climate change exacerbates food insecurity through frequent extreme weather events in the form of droughts and floods, as well as increased salinity for coastal agriculture through gradual changes in sea-level rise. Increasing temperatures alone are projected to reduce global crop production by more than 10% by 2050 (Tai *et al*., 2014). Additionally, altered weather patterns can increase crop susceptibility to pathogen infection, pest infestations and competition from rapidly growing weeds.

Plants are able to alter their growth and development when confronted with changing environmental conditions and have developed adaptive mechanisms in order to survive and reproduce. For instance, in drought-prone environments, some plants have evolved various mechanisms of avoidance, tolerance, escape and recovery. In habitats with extreme temperature changes, antifreeze (extreme cold) and heat shock (extreme heat) proteins can protect plants from damage by maintaining the structural integrity of tissues and enzymatic proteins required for cellular function. In water-saturated soils, some plants may develop specialized tissues that allow for gas exchange between shoots and roots.

Rice (*Oryza* spp.) is the second most important food crop worldwide, with China and India being the two largest producers (Mohanty, 2013). Over 22 million hectares of rice production is prone to sudden and prolonged flooding (Bailey-Serres, 2010). Most current commercial rice varieties in Asia are intolerant to inundation and respond to submergence by elongating their leaves and increasing the

length of their stems. This coping mechanism is appropriate if the flooding is either shallow or temporary, but results in death of the plant if the ability to photosynthesize and respire are inhibited for an extended period of time (Bailey-Serres *et al.*, 2010). Natural **landraces** of rice, originally collected and stored in a gene bank in the 1950's, were experimentally screened for submergence tolerance in the 1970's. Astonishingly, one of these landraces demonstrated extreme tolerance to inundation — 100% of ten-day old seedlings survived seven days of complete submergence — but it lacked other important agronomic attributes (Bailey-Serres et al., 2010).

The above example illustrates natural adaptive responses of plants to environmental stress, a phenomenon known as **plasticity**. However, adaptational **phenotypes** may have different observable states within a species such that individuals may differ in the way they respond to the environment. This variation in a phenotype among individuals is conditioned by variation in the corresponding gene or **genotype**, the environment in which the individuals grow, and the interaction between genotypes and the environments. Moreover, variation in phenotypes is the basis for evolution by natural selection and human-assisted artificial selection, where plants bearing phenotypes of interest are bred to contribute to subsequent generations.

As alluded to above, crop production is critical to economic stability and food security, and the famers who produce these crops are sometimes confronted with extraordinary challenges. For example, Nekkanti Subba Rao is a rice farmer affected by frequent and prolonged flood events along the eastern coast of India. Flooding not only impacted the wellbeing of farmers in this region; the failure to produce a crop has a rippling effect in society and may cause substantial economic loss.

Landrace
A traditional plant varieties that have been developed through informal (farmer-based) breeding. Landraces are adapted to local conditions and are unique due to isolation from other populations within the same species.

Plasticity
The ability of an organism to change its phenotype in response to the environment.

Phenotype
The observable state of a given characteristic or combination of characteristics.

Genotype
The genetic makeup (DNA sequence) of an organism.

Guiding Questions

1. Based on the information provided, what could be done to help Nekkanti Subba Rao and other farmers in this region to mitigate crop loss due to flood events?

2. List and discuss some of the factors that affect food security.

3. What is phenotypic variation, and how can it be useful for breeding and for food security?

4. What are the factors that influence phenotypic variation, and which do you think is of greatest value to plant breeders? Why?

References

Bailey-Serres, J., T. Fukao, P. Ronald, A. Ismail, S. Heuer, and D. Mackill (2010) Submergence tolerant rice: *SUB1*'s journey from landrace to modern cultivar. *Rice* 3:138-147.

FAO, IFAD and WFP (2015) The State of Food Insecurity in the World 2015. Meeting the 2015 international hunger targets: taking stock of uneven progress. Rome, FAO.

Mohanty, S. (2013) Game changers in the global rice market. *Rice Today*; International Rice Research Institute. 12(3):44-45.

Tai, A. P., M. V. Martin, and C. L. Heald (2014) Threat to future global food security from climate change and ozone air pollution. *Nature Climate Change* 4(9):817-821.

United Nations, Department of Economic and Social Affairs, Population Division (2015) World Population Prospects: The 2015 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP.241.

Stage II: Genetic Diversity

Farmer Nekkanti Subba Rao of Andhra Pradesh, India, was one of the first to adopt the Swarna-*SUB1* variety of rice in his community. After seeing its high tolerance to flooding (compared to the Swarna landrace lacking the *SUB1* **allele**), he distributed seeds he reproduced from Swarna-*SUB1* to other farmers, which provided flood protection for nearly 1,000 hectares across his village and nearby areas during the wet season of 2009 (Hettel, 2015).

The Swarna-*SUB1* story exemplifies the world of crop breeding and the importance of genetics for food production (Bailey-Serres *et al*., 2010). After its discovery in an exotic landrace of rice, an allele at the *SUB1* **locus** associated with flood tolerance was **introgressed** into otherwise locally adapted varieties grown by farmers such as Mr. Subba Rao. This example highlights the importance of programs that maintain and provide access to genetically diverse **germplasm**, such as the traditional landrace that harbored the favorable allele at *SUB1*. The Convention on Biological Diversity recognized the need to protect and conserve plant species, and crop scientists have asserted the importance of accelerated collection and characterization of genetic diversity (United Nations Environment Programme, 1992). As a result, gene banks that maintain germplasm collections were established at regional, national and international levels. Breeders and geneticists can also look for diverse germplasm in situ and in other ex situ collections such as breeding programs, botanical gardens and arboreta. However, it is not adequate to simply maintain or provide access to a collection; it is also necessary to measure and document the variation among individuals or populations in a collection so that breeders can make decisions about which to use for the development of new varieties.

Many phenotypes of interest such as submergence tolerance in rice, and others including disease resistance, nutrient content and yield are quantitative in nature. Such traits display a spectrum of

Allele
Variant form of a gene which gives rise to genetic variation.

Locus
Specific region of a chromosome, often encompassing one or more genes.

Introgression
The substitution of an allele of interest from one parent genome into the background of another genome via backcross breeding.

Germplasm
Living material from which new plants can be grown such as seeds or tissues which are maintained for the purpose of preservation, breeding, and other research uses.

variation and can be difficult to dissect into their genetic basis because they are influenced by multiple loci and environmental factors. However, over the last quarter-century, **quantitative geneticists** have developed techniques to map regions of the genome underlying variation in quantitative traits — these techniques were applied to determine that the *SUB1* locus was associated with variation in submergence tolerance. These mapped regions of the genome are referred to as quantitative trait loci (**QTL**). Typically, multiple QTL throughout the genome underlie variation in a quantitative trait, with each QTL contributing to a fraction of the total variation. Finer mapping and dissection of QTL has led to the understanding that different types of variation in the DNA sequence are causal of the variation in quantitative traits, including single nucleotide polymorphisms (**SNPs**), insertion-deletion variants (**Indels**), and copy number variants (**CNVs**). In the *SUB1* example, scientists discovered a QTL that was associated with 70% of the total variation in flood tolerance. Using a technique called positional cloning, a CNV at the *SUB1* locus was found to underlie this tolerance (Bailey-Serres 2010). Leveraging knowledge of the position of this CNV in the rice genome, plant breeders used molecular markers linked to the CNV to facilitate introgression of the favorable allele into otherwise flood sensitive varieties. This allowed them to create improved varieties like Swarna-*SUB1*, which maintains the preferred characteristics of the Swarna variety while having the added benefit of flood tolerance.

A challenge for plant breeders is determining which germplasm might harbor favorable alleles like the *SUB1* CNV. Phenotypic screens are a useful way to survey plant germplasm, but this can be costly or infeasible when there is a large number of accessions to screen, or when the target environmental conditions are difficult to reproduce (e.g. creating a screen for flood tolerance or testing for disease resistance). Fortunately, due to tremendous advances in genome sequencing technology, it is now feasible to rapidly survey whole genomes for SNPs, Indels and CNVs. Although determining the specific subset of nucleotide variant(s) that underlie quantitative trait variation is still challenging, the DNA sequence information can be used to compare individuals belonging to different populations across

Quantitative Genetics
Study of the inheritance of characteristics that vary quantitatively and tend to be controlled by multiple genes.

QTL
A gene or chromosomal region that contributes to the expression of quantitative characteristics.

SNP
Single base pair differences in the DNA sequence between individual members of a species.

Indel
Insertions and deletions in the DNA sequence between individual members of a species.

CNV
Differences in the number of copies of a DNA sequence (often a gene) between individual members of a species.

their whole genomes. This information can be summarized using different metrics that help to describe how the sequence variation is structured or distributed among germplasm prior to embarking on efforts to breed improved varieties or genetically dissect quantitative traits.

Commonly used measures of genetic diversity based on DNA sequence data include **nucleotide diversity**, **coefficient of inbreeding** and $\boldsymbol{F_{ST}}$. For instance, a recent study by Belemkar *et al*. (2016) measured nucleotide diversity and inbreeding in a collection of 52 accessions of *Apios americana*, a semi-domesticated crop for which this information had not been previously known. These data, combined with phenotypic data, allowed the researchers to determine that certain phenotypes, such as tuber width, were positively correlated with inbreeding, while tuber number was negatively correlated with inbreeding. Based on the same genotypic data, the researchers also found that the accessions could be clustered into multiple groups or subpopulations, some of which may have lower levels of inbreeding and therefore be enriched with individuals that have few, large tubers. In lieu of phenotype data crop scientists may want to look more closely at the relatedness of the germplasm in a species collection. F_{ST} is a measure of differentiation between pre-defined groups of individuals, such as those collected from different geographies or ecological zones. If F_{ST} is low, this indicates genotypic variation is shared among subpopulations and there was some mechanism(s) for genetic exchange between them (e.g. if a species reproduces by outcrossing). If F_{ST} is high, this indicates genotypic variation is distributed across subpopulations, and favorable alleles would be more likely to be present only in specific subpopulations.

Nucleotide Diversity
Average number of nucleotide differences between two randomly chosen homologous sequences; a measure of the degree of variation among alleles within a population.

Coefficient of Inbreeding
A measure indicating the probability that genes at a randomly chosen location in the genome are identical by descent.

F_{ST}
A measure of genetic differentiation between subpopulations. More technically, it is a measure of the proportion of genetic variance within a subpopulation relative to the total genetic variance among subpopulations. It ranges from 0 to 1, where 1 indicates maximum differentiation.

Guiding Questions

1. What can plant scientists do to avoid or minimize the loss of natural diversity in plant species?

2. How can scientists quantify genetic diversity among individuals?

3. How might estimates of inbreeding for *A*. *americana* individuals be useful to breeders?

4. Without phenotypic data, how can scientists use measures of F_{ST} for managing germplasm collections?

References

Bailey-Serres, J., T. Fukao, P. Ronald, A. Ismail, S. Heuer, and D. Mackill (2010) Submergence tolerant rice: *SUB1*'s journey from landrace to modern cultivar. *Rice* 3:138-147.

Belamkar, V., A. D. Farmer, N. T. Weeks, S. R. Kalberer, W. J. Blackmon, and S. B. Cannon (2016) Genomics-assisted characterization of a breeding collection of *Apios americana*, an edible tuberous legume. *Scientific Reports* 6:34908.

Hettel, G. (2015) Indian farmer kick-starts two green revolutions. *Rice Today*; International Rice Research Institute. 14(2):10-11.

United Nations Environment Programme. "UNEP: Convention on Biological Diversity." In UNEP: Convention on Biological Diversity, 1992. Ed. Stephen Tully. Cheltenham, UK, Edward Elgar Publishing, 2005.

Stage III: High-Throughput Sequencing

It has become routine for scientists like Belamkar *et al*. (2016) to use **high-throughput sequencing** (HTS) technologies in plant breeding. HTS—also referred to as next-generation sequencing (NGS)—can rapidly produce massive amounts of sequence data at a low cost per base pair for the evaluation of genomic diversity. Table 1 compares four current HTS platforms to the first-generation Sanger sequencing platform. Although Sanger sequencing has high quality **reads**, due to its low throughput it is now used primarily for small scale experiments or for validating HTS results. In contrast, HTS platforms can produce tens of thousands to hundreds of millions of reads. This is achieved by sequencing molecules in parallel (originally called "massively parallel sequencing"), but parallelization comes at the cost of having substantially higher error rates than Sanger sequencing. Nevertheless, to some degree error can be corrected by implementing bioinformatics error-correction methods during data analysis. A key advantage to the massively parallel sequencing paradigm is the ability to perform **multiplex** sequencing, where DNA from different individuals are labeled with **molecular barcodes** and sequenced at the same time. This allows for efficient analysis of genomic diversity in large samples. When embarking a new study using HTS, it is important to be familiar with the performance attributes of different sequencing platforms (Table 1) in order to choose the most appropriate platform for a given project.

High-Throughput Sequencing
Technology that has the capacity to rapidly generate massive volumes of sequence data.

Read
A contiguous sequence of DNA produced by a sequencing machine.

Multiplexing
A sequencing approach in which samples tagged by molecular barcodes are pooled and sequenced simultaneously.

Molecular Barcode
A unique, short DNA sequence that distinguishes the sequences belonging to different samples.

Table 1. Summary of Sequencing Platforms.

Platform	Read length	Reads	Throughput	Runtime	Error rate	Cost (US$/Gb)
ABI 3730xl Sanger	400-900 bp	96	38-86 kb	2-3 hrs	0.001%	5.5-12.5 M
Illumina MiSeq v3	2 × 300 bp	44-50 M	13.2-15 Gb	21-56 hrs	0.1%	110
Illumina HiSeq 2500 v2	2 × 250 bp	600 M	125-150 Gb	60 hrs	0.1%	40
Pacific BioSciences RSII	Max: ≈60 kb	~55,000	500 Mb - 1 Gb	30 min - 6 hrs	13%	400-800
Oxford Nanopore MK1 MinION	Max: ≈200 kb	>100,000	10-20 Gb	1 min - 48 hrs	12%	50-100

Table adapted from Rhoads and Au (2015) and Goodwin *et al*. (2016) and manufacturer's data.

The advantages of HTS also present challenges. The vast amount of data generated by HTS requires high-performance computing systems to process and analyze the data. Intensive computation is required to perform quality control steps to filter poor quality sequences, to de-multiplex the sequences into sample-specific sets based on the molecular barcodes, and to align the reads to identify nucleotide variation at shared positions in the genome among the different samples. SNPs discovered from this analysis can be used to generate a **genotype matrix,** with SNP sites on one axis and individual samples on the other axis such that each cell of the matrix indicates the genotypic state of an individual at a particular SNP site (e.g. A/A, A/C, or C/C)**.** The genotype matrix can then be used

Genotype Matrix
A matrix with the genotypic state of each marker for each individual.

to estimate heterozygosity, inbreeding, and F_{ST}, in order to evaluate genetic diversity and population structure. In addition, the relationships among individuals can be computed from the genotype matrix and visualized in a **phylogenetic tree**, with nodes indicating points of genetic divergence and branch lengths representing the genetic distance between individuals split by the node.

Imagine your team works at the *Global Institute for Germplasm Diversity* where you are commissioned to assist breeders with capitalizing on plant genetic diversity. Potato production is in decline due to an emerging disease pandemic, and there is interest in developing alternative food crops. *Apios americana* shows promise as a new crop, as it not only produces edible tubers with high protein content (approximately 3X that of potato), but it also symbiotically fixes nitrogen and can be grown in marginal soils. Furthermore, starting in the 1980s, Blackmon and Reynolds (1984) began domesticating and breeding varieties of *A*. *americana*. However, Blackmon and Reynolds' founding population was based on wild accessions primarily from the southeastern U.S. (Belamkar et al, 2015). Out of concerns of a **genetic bottleneck** that may limit progress in adapting and further developing new breeds, your team decides to use HTS to characterize genetic diversity of additional wild populations of *A. americana* across the eastern U.S. Your team will need to decide which HTS platform you would use to assays SNPs across the genome (see question 1).

Using the SNP data, your team calculated F_{ST} between all pairs of the population samples collected from each location (Table 2) and generated an unrooted phylogenetic tree (Figure 1).

Phylogenetic Tree
A diagram that shows the relationships among individuals. Rooted trees describe evolutionary history, while unrooted trees describe the hierarchical relationships among the samples.

Genetic bottleneck
A drastic reduction in the size of a population which reduces variation in the gene pool.

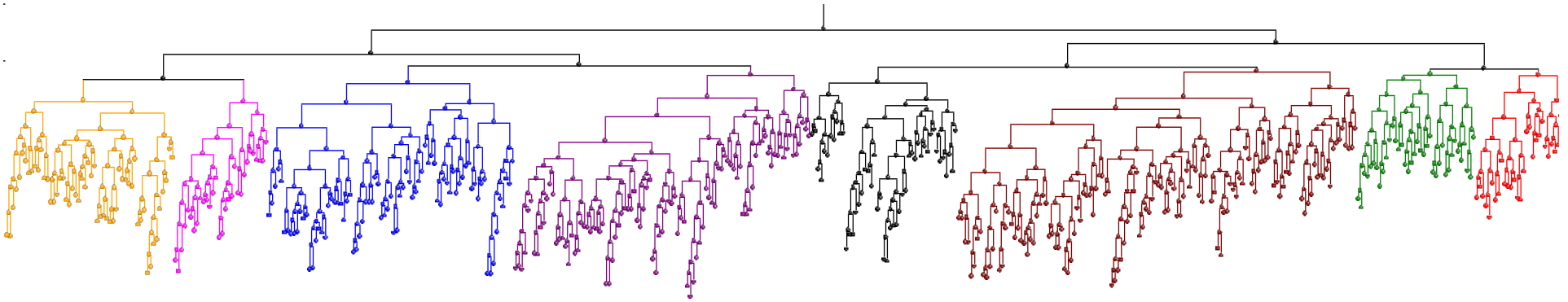

Figure 1. Unrooted phylogenetic tree of *Apios americana* across the Eastern U.S. Accessions are color-coded according tc geographic origin: Maine (orange); Massachusetts (fuchsia); New Jersey (blue); Delaware (purple); North Carolina (black); Soutl Carolina (maroon); Georgia (green); Florida (red).

Table 2. Matrix of pairwise F_{ST} values between populations of *Apios americana* collected in the U.S.

	ME	MA	NJ	DE	NC	SC	GA	FL
ME	-	0.07	0.12	0.10	0.20	0.18	0.18	0.18
MA	-	-	0.12	0.10	0.20	0.18	0.18	0.19
NJ	-	-	-	0.04	0.16	0.14	0.15	0.15
DE	-	-	-	-	0.14	0.12	0.13	0.13
NC	-	-	-	-	-	0.02	0.11	0.11
SC	-	-	-	-	-	-	0.09	0.09
GA	-	-	-	-	-	-	-	0.01
FL	-	-	-	-	-	-	-	-

After completing the project, your team needs to advise breeders on how they can use your results to facilitate ongoing breeding efforts to develop new varieties and expand the production range into potato growing areas of the world. Keep in mind that a diversified breeding pool provides access to a wide range of phenotypic variation that maximizes the discovery potential for unique phenotypic states and sustains the development of improved varieties in the long term. With conservation of germplasm and comprehensive evaluation of genetic diversity, DNA sequence variants associated with favorable phenotypes can be discovered in such managed resources and introgressed into modern varieties, just like the Swarna-*SUB1* story. However, selection and breeding for desirable phenotypes can result in genetic bottlenecks that reduce variation and limit the potential to develop unique varieties that address new issues, such as a new disease and stressors associated with changes in climate. Thus, there is a "tight rope" that breeders must walk to balance the demands for improved varieties while maintaining genetic diversity. When genotyping information derived from HTS data is available, scientists and breeders can make more informed decisions to mitigate these challenges and help to achieve food security.

Guiding Questions

1. Belamkar et al. (2016) estimated the genome size of *A. americana* to be approximately 1.65 Gbp. Your team needed to assay SNPs across the genome to examine genetic diversity in your new collection of *A. americana*. Which HTS platform would you have used for this project and why?

2. Summarize the findings from your team's project (Table 2 and Figure 1) on HTS-based characterization of genetic diversity in *A. americana*.

3. Given the context of the situation and summary of your findings, what would be the top three recommendations your team would make to the breeding community?

4. How can HTS analysis of genomic diversity help in addressing challenges associated with food security?

References:

Belamkar, V., A. Wenger, S. R. Kalberer, V. G. Bhattacharya, W. J. Blackmon, and S. B. Cannon (2015) Evaluation of phenotypic variation in a collection of *Apios americana*: An edible tuberous legume. *Crop Science* 55:712-726.

Belamkar, V., A. D. Farmer, N. T. Weeks, S. R. Kalberer, W. J. Blackmon, and S. B. Cannon (2016) Genomics-assisted characterization of a breeding collection of *Apios americana*, an edible tuberous legume. *Scientific Reports* 6:34908.

Blackmon, W. J. (1986) The crop potential of *Apios americana*—preliminary evaluations. *HortScience* 21:1334-1336.

Goodwin, S., J. D. Mcpherson, and W. R. McCombie (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17(6):333-351.

Rhoads, A., and K. F. Au (2015) PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics* 13(5):278-289.