# **STAT 510: Consulting Seminar Handout**

November 15, 2016

**Group Members:**

* Andrea Mack
* Paul Harmon
* Nnamdi Ezike

**Client:**

Jamin Smitchger, PhD Candidate, Department of Plant Sciences and Plant Pathology

# Introduction:

Jamin Smitchger, our client, researches pea-plants at two sites in Bozeman and Moccasin, Montana. Data were collected over a 4 year period, leading to six site-year combinations of data:

* 4 Field Seasons in Bozeman from 2013-2016
* 2 Field Seasons in Moccasin from 2015-2016

Of interest to our client is to determine which genetic markers are associated with the most phenotypic variation. The phenotype of interest for Jamin in pea plants called **lodging**. Lodging refers to pea plants falling over, either due to root or stem failure. The more plants that exhibit lodging, the lower the yield. Although environmental factors can greatly influence whether a pea plant exhibits lodging or not, genetic factors are able to be controlled. Our client intends to find genetic markers associated with reduced lodging so that he can then breed generations of plants that are less susceptible to this trait.

# Background Information and Research Questions:

Some understanding of genetics is necessary to understand what is going on in this process. Basically, a **genotype** refers to the expression of a gene – because thousands of genes are located on any individual chromosome, it can be difficult to measure genes themselves. We instead track **genetic markers**. These markers are then correlated with observable **phenotypes** such as the lodging trait, plant height, etc. Gene location is important because genes that are located physically close together tend to be inherited together; this phenomenon is referred to as **linkage**. By analyzing the linkage in progeny of a hybridization – essentially mating two plants and examining the expressed phenotypes of their offspring, we can determine which genes are most associated with variation in phenotypes. The goal of our analysis is to assist our client in some aspects of a Quantitative Trait Locus (QTL) mapping of markers relative to phenotypic variation.

Our client has three research questions we need to answer prior to him conducting the QTL analysis.

1. Are there any phenotypic traits that are highly correlated with each other, and which type of correlation is most useful for these data?
2. Are the datasets from each year/site combination similar enough to be combined into a single dataset? Do year and site interact with each other or do we need to do separate QTL analyses for each dataset?
3. Because QTL analysis involves simultaneous hypothesis testing, what is an appropriate significance threshold (in LOD distance) for tests to be considered “statistically significant?”

# The Data:

QTL analyses require two datasets: A **Genotypic Dataset** which contains information about distances between genetic markers on a chromosome, and a **Phenotypic Dataset** which contains information about specific plants' phenotypic expression, measured at the two sites.

### The Genotype Data:

The Genotype dataset includes 1.2 million observations, most of which is not useful for the analysis due to sequencing errors and other measurement issues. Our client has been sorting through much of the “junk” in the genotypic dataset. "Junk" may include genes that are not expressed. Genotypic data indicates which **allele** of a gene has been inherited from each parent. This dataset looks different than the usual for several reasons: 1) there are more covariates than observations and 2)the data refer to the allele, so every observation is either “A” or “B”.

**Example of the Genotypic Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | white flow 2015 +2016 CHECK | white flower LG 2 A=white dwf only-updated 10-25-2016 | Afila lg1 Dwf only-2015 | Afila 1=YES=DELTA=A 2015 |
| Delta | A | A | A | A |
| RER | B | B | B | B |
| 1 | A | - | - | B |
| 2 | B | B | A | A |

### The Phenotype Data:

The phenotypic dataset includes 50,000 observations; in general, it is quantitative in nature. This looks more like the type of data that we are used to seeing; these data track expressed phenotypes observed in the progeny such as percent lodging, plant height, tendril width/length, etc. This dataset is used for the analysis of interactions and correlations between phenotypic traits. The percent lodging trait is highlighted in the data example below.

**Example of the Phenotypic Data:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **line #** | **average lodging Moccasin 2015** | **AVERAGE HEIGHT Moccasin 2015** | **Yield** | **% GERMINATION AVG with bad germinating lines left in and adjusted for seeds planted** | **% germ avg with bad germ lines removed and adjusted for seeds planted** | **# of branch** | **plants harvested** | **% Lodging** | **% Lodging with bad germinating lines included** |
| **DELTA** | 19.71705 | 45.42857 | 28.88814 | 98.57143 | 98.57143 | 1.8 | 3.9 | 30.901011 | 30.90101 |
| **RER** | 36.58911 | 61.78571 | 33.49136 | 87.14286 | 87.14286 | 3.16 | 3.43 | 57.178192 | 57.17819 |
| **1** | 31.99218 | 69.61111 | 33.95567 | 83.33333 | 83.33333 | 2.33 | 3.67 | 57.433633 | 57.43363 |
| **2** | 27.80952 | 39 | 22.95265 | 86.66667 | 86.66667 | 3.13 | 3 | 36.841987 | 36.84199 |
| **3** | 35.07338 | 37.61111 | 21.15533 | 100 | 100 | 1.58 | 4 | 36.995751 | 36.99575 |
| **4** | 29.93146 | 32.83333 | 24.6165 | 100 | 100 | 2.17 | 4 | 23.857089 | 23.85709 |

# Statistical Analysis: What we plan to do

We are involved in much of the prior work leading up to the QTL analysis; however, we are not actually performing QTL tests during this semester. Instead, we are helping our client to answer the aforementioned questions in the previous section. Our proposed strategy to answer them is:

* **Interaction between year/site combinations (phenotypic data)**. Two issues here need to be considered. Our client wants to combine all of his data from each site/year combination into a single dataset. We need to control for the variation in year/site combinations, possibly by fitting random effects into the models we fit. If year and site interact with each other, we may need to consider fitting these models with some interaction term. Note that we do not necessarily intend to fit these models, rather, we simply need to assess whether we can additively include year and environment or not. For each of the i genetic markers, the model we believe that is fit is:
* **Correlations between Phenotypic Data:** Our client has already created a scatterplot matrix (using pairs.panels in the ‘psych’ package) of correlations between covariates in the phenotypic data, in order to determine which traits are highly correlated with each other. We will remake the scatterplot matrix our client brought to us, but look at different ways to display the correlations so that the plot is useful. Also, since our client used Pearson correlations, we may want to re-examine the phenotype data to determine if there are other types of correlation (Spearman, Kendall’s Rank, etc.) that may be more useful in this context. Doing one QTL on several traits is more powerful than doing a separate QTL for each phenotypic trait. Since our client is only concerned with the pea lodging, it is unclear to us why he is looking at all of the phenotypic data.
* **Significance Threshold**: Because QTL relies on simultaneous testing, we need to develop a significance threshold that our client can use to interpret the statistical significance of his tests. However, rather than doing p-value based tests, we instead calculate a genome-wide adjusted LOD score at each loci (marker) and determine significance based on this distance threshold. This is feasible because distance between loci can be interpreted probabilistically; loci that are closer together are likely to be inherited together. LOD refers to ratio comparing the null that there is not a QTL to the alternative, that there is.
* **Quantitative Trait Locus (QTL) Mapping:** Although we are not performing the QTL analysis, all of this culminates in a series of hypothesis tests known as QTL analysis/mapping that will be performed by our client.