# Activity 3: Filtering data using the TASSEL GUI

In this exercise, you will explore, summarize, and filter GBS data sets from black raspberry and meadowfoam. You will need to have installed TASSEL 5 and Java 1.8.

**PART I. DATA EXPLORATION**

**Step 1: Using a file transfer client (e.g. WinSCP or CyberDuck), download the two HapMap files named “BlackraspHapMap1.txt” and “MeadowfoamHapMap.txt” from here:**/nfs1/Teaching/data/viningk/GBS/HapMaps

**Step 2: Load the HapMap files into TASSEL by selecting Data>Load.   
The data sets will appear in the data tree in the Sequence subfolder.**

**Step 3: Select each of the HapMap files and look at the overall summary of the data in the left column, middle panel.  
How many individuals are in each population?   
How many SNP calls comprise each data set?**

**Step 4: Produce summary data for each of the HapMaps. With a HapMap highlighted, select Analysis> Geno Summary. Leave all three boxes checked in the pop-up window that appears.**

Four files will appear in the data tree in the Result>Genotype Summary subfolder. Each is a data table. \*\*HINT: When looking at these data tables, you can click and drag on column headings to make columns wider. You can also click on the column headings to sort, and then reverse sort.   
  
Look at the OverallSummary tables for the black raspberry and meadowfoam SNP data.   
What is the overall percentage of missing data for each?   
What is the overall proportion of heterozygous SNPs?   
  
Look at the AlleleSummary table for each data set.   
Are the proportions of A:T:C:G similar overall?

Look at the TaxaSummary table for the two data sets. Click the “Proportion Heterozygous” column heading to sort from high to low.   
What is the highest proportion of heterozygous SNPs in an individual from the black raspberry population?  
What is the highest proportion of heterozygous SNPs in an individual from the meadowfoam population?

**Step 5: For each data set, select the SiteSummary table. Scroll right until you find the Proportion Missing column, and then click the column heading to sort this column from high to low.   
How many SNPs have no missing data in each data set?**

**What is the highest proportion of missing data in each data set?**

**Step 6: With the SiteSummary table selected, go to Results>Chart. In the resulting pop-up window, the default graph type you should see is “histogram.” For Series 1, scroll to select Minor Allele Frequency. Change the number of bins to 20. Notice that you can also click on the Properties button to modify details of the chart display. Repeat the graphing steps for the other data set, so that you have both graphs displayed side by side. Save both graphs so that you can bring them up again if needed.**

**For the black raspberry data, roughly how many SNPs are not segregating in this population (minor allele frequency near 0)? Why do you think these were called SNPs by TASSEL?**

**Can you tell anything about allele segregation in the black raspberry population from looking at the minor allele frequency distribution? If so, what?**

**How does the meadowfoam minor allele frequency distribution differ from that of black raspberry?**

**Step 7: For each graph from step 5, change Series 1 to Proportion Heterozygous. Now, you are looking at the proportion of SNPs that exhibit various levels of heterozygosity. The x-axis shows the proportion of individuals in the population that are heterozygous for a particular SNP. For example, a heterozygosity proportion of 0.25 means that 25% of the individuals in the population are heterozygous for that SNP, and 75% of the individuals in population are homozygous. The y-axis shows the total number of SNPs at that heterozygosity level.**

**Approximately how many black raspberry SNPs are in a homozygous state in all individuals in the population (proportion heterozygous=0)?**

**How is the meadowfoam SNP distribution different from that of black raspberry?**

**Step 8. Now, change Series 1 on each graph to Proportion Missing. Change the number of bins to 5.**

**How are the two data sets different?**

**PART II. DATA FILTERING.  
Perform the following steps for the black raspberry and meadowfoam data sets.**

**Step 1. Filter by taxa. This filter is intended to remove individuals with a high proportion of missing data. Select Filter>Taxa. Set “Min Proportion of Sites Present” to 0.2. This setting is very conservative – only individuals with >=80% missing data will be filtered out. Leave the other settings at their defaults. New, filtered hapMaps will appear at the “Sequence” node on the data tree. Their names will include the number of remaining individuals.**

**How many individuals remain in each data set after filtering?**

**Step 2. Filter by site. This filter is intended to remove low quality SNPs, which may be the result of sequencing errors. IMPORTANT: Perform this operation on the taxa-filtered hapMaps produced in the previous step. Set the minimum count to 20% of the number of individuals in the data set. This means that SNPs must be present in at least 20% of the individuals in the population. Leave the “Minimum Frequency” at 0.05. This will filter out SNPs with minor allele frequencies < 5%. Leave all other settings blank or at their defaults. New hapMaps will be created under the Sequence node of the data tree.**

**How many SNPs remain in each data set?**

**Step 3. Produce new summary data for the two taxa+site-filtered HapMaps. With each HapMap highlighted, select Analysis> Geno Summary.**

**Compare your answers below to those in Part I, Step 4.**

Look at the OverallSummary tables for the black raspberry and meadowfoam SNP data.   
What is the overall percentage of missing data for each?

What is the overall proportion of heterozygous SNPs?

Look at the TaxaSummary table for the two data sets. Click the Proportion Heterozygous column heading to sort from high to low.   
What is the highest proportion of heterozygous SNPs in an individual from the black raspberry population?

What is the highest proportion of heterozygous SNPs in an individual from the meadowfoam population?

**Step 4: Repeat what you did in Part I, Step 6, but this time do it with the filtered HapMaps: with the SiteSummary table selected, go to Results>Chart. In the resulting pop-up window, select graph type “histogram.” For Series 1, scroll to select Minor Allele Frequency. Change the number of bins to 20. Repeat the graphing steps for both data sets, so that you have both graphs displayed side by side.**

**How do the meadowfoam and black raspberry graphs compare to the graphs of unfiltered data?**