**StrataGUI User Guide**

**General Info/Background:**

StrataGUI is a graphic interface for working with the StrataG R package. This means that it is running code the same as when you use a script, but the code is behind the scenes. Additionally, this interface offers some user friendly visuals and features to interact with the data that can be helpful, especially during data exploration and QA/QC. It also produces a nice summary report at the end.

However, a con of this interface is that it can be very slow with larger datasets on some steps. If you have a large dataset and are comfortable in R, it may be better to just work directly in R with StrataG functions. Additionally, the shiny window sometimes crashes and then you must start from the beginning (you can load a gtypes object, but everything else you have to do all over). This prevents you from skipping QA/QC steps (which is good) but can be annoying when you run into bugs. It is also sometimes hard to scroll left and right in tables within the shiny window.

**General Steps for Use:**

1. Click select working directory and browse through your folders to choose the appropriate one. You also can also create a new working directory (This is nice because you don’t have to type complicated paths).
2. Choose whether you want to upload your data as a csv file or as a gtypes object. Then click on the Load data tab.
3. You should be able to see your csv file (or gtypes object). If you have uploaded a csv, select the appropriate columns from the drop down menus for Sample ID, stratification, first locus, etc.
4. Click load gytpe object (can be slow) and you can also select stratification scheme
5. Then click on the QA/QC tab
6. The first step is to remove samples that are missing data for too many loci. We want to remove samples that are missing data for too many loci. We traditionally generally err on the side of being conservative and only include samples with 80% of called genotypes in which we have confidence, but this may depend on the project and dataset. A sample that did not amplify for multiple markers may yield false results for markers for which it did amplify. Make this number of columns in your genotype data that are not alleles (aka sample IDs or strata)
7. You can use the slider to choose minimum percent of missing loci that a sample can have to be included in further analysis. The graph at the bottom allows you to visualize how many of your samples match the criteria you select. Once you have picked an appropriate threshold, click remove samples then click next step.
8. The second step is to remove samples with too many homozygous loci. This is important because if a sample has higher homozygosity than the population average, the sample itself may be degraded, and may yield false results. You can use the slider to choose minimum percent of homozygous genotypes a sample can have to be included in analysis. This may be useful if you suspect allelic dropout. The percent homozygous graph is also very useful, helpful to know how many of your samples you can keep with each threshold you select. Click remove samples and/or click next step.
9. The third step is to remove loci that are missing a lot of data. Generally it’s not advisable to include markers that did not amplify well in your sample set. Using markers with high proportions of missing data may make populations look more structured than they really are or cause other spurious results. Use slider to choose minimum percent of missing individuals a locus can have to be included in analysis. The graph helps you visualize how many of your samples fit the criteria you select. Then click remove samples then click next step.
10. The fourth step is to remove loci with too many homozygous genotypes. A marker with very little variation may or may not be due to allelic drop out and in any case, is not very useful. However, the specific appropriate threshold depends on your species, dataset and study goal. Use slider to choose minimum percent of homozygous individuals a locus can have to be included in analysis. The graph helps you visualize how many of your loci fit the threshold you select. Then click remove samples then click next step.
11. The fifth step is to check for duplicates in your sample set. This may be a factor of intentional duplication to check your genotyping error rate or it could be unintentional resampling of the same individual. Removing duplicates is critical because counting single individuals as multiple members of a subpopulation/strata greatly reduces calculated genetic variation within groups which will lead to erroneously large genetic distances between strata. This step compares each sample against each other, so it is one of the more computationally intensive steps, especially in larger datasets. This step can take a long time but will eventually show you pair of sample that share different percent of genotypes. You can then select samples to remove and click remove samples. Then click next step.
12. The Next step is to check for Hardy Weinberg Equilibrium within your strata. The Hardy-Weinberg theorem states that allele frequencies should be constant over time in the absence of selection, drift, migration or other processes that shift allele frequencies. This principle allows us to calculate expected levels of heterozygosity from measured allele frequencies. Populations that are not experiencing selection, drift, migration, nonrandom mating, or mutation should have observed heterozygosities similar to expected values and are said to be in Hardy-Weinberg equilibrium (HWE). Loci may be out of HWE for multiple reasons, but generally should not be used to measure gene flow/genetic structure and as they do not meet the assumptions required for informative neutral markers (N.B., especially for SNPs, this does not apply to study questions may involving adaptive variation vs. genetic draft/gene flow). If your populations have significant genetic structure an HWE test for all populations combined may reveal some deviation from HWE (indicating selection, drift, nonrandom mating, etc.) Therefore, it is good to check for deviation from HWE in each of the individual populations. Markers that are out of HWE in many populations should be excluded, depending on the questions and data set. You can choose your HWE method and alpha level. This can be very slow but when it is done it will show you HWE p-values for each strata so you can quickly see which ones are out of HWE. It will also show you p-values for each locus. It also shows you jackknife plots (generated by calculating HWE many times by excluding each sample), which shows you how the odds ratio increases as more samples are added.
13. The next step is to check for linkage disequilibrium between your markers. Linkage disequilibrium (LD) occurs when loci are not randomly assorted. This can be due to physical proximity on the chromosome or an artifact of selection. Markers that are linked should generally not be considered as independent when examining genetic structure because they may artificially inflate structure signals.
14. Importantly it is best to examine linkage disequilibrium within each strata, as genetic structure can inflate LD when considered across populations. If markers are linked in many of your strata you should pick only one of each pair of linked markers to use for population structure analyses or combine markers (e.g., PHASE for SNPs), so as not to overestimate genetic structure. If your populations have significant genetic structure, testing for LD in all populations combined may reveal artificially indicate LD as a product of selection, drift, nonrandom mating, etc. It’s good to check for LD in each of the individual populations. Markers that are linked in many populations should be excluded. You can use the slider to set acceptable level of LD. The screen clearly shows pairs of linked markers. You can then choose to remove markers.
15. The final step is to generate pairwise genetic distance values between each of your strata. You can select which method you want to calculate genetic distance. These genetic distances will indicate levels of differentiation between your strata and reveal genetic structure (or lack of genetic structure) in your study population. Each of the methods has strengths and weaknesses and you should think carefully about which method or combination of methods you want to use to answer your questions.
16. Report the results of your analysis in a way that makes sense according to your research question and data!