**QC/QA Documentation for microsatellite and SNP genotype data**

(Version 0.1, 4/22/13)

[Last updated: 7/22/2013 by KKM; Project status: Complete]

**Project name: Global analysis of melon-headed whale population structure**

**Report prepared by: Karen Martien, Brittany Hancock-Hanser**

**Laboratory Processing Project design:**

**PI Responsibilities**

**Project overview:** Provide a brief description of the project’s purpose, species, number of markers etc. In **Table 1** list the samples included in the project (from the beginning), including which ones were used as marker optimization samples, controls, random replicates and targeted replicates. If this report is being prepared for external use, consider including more information about each sample (i.e., where and when collected, tissue type).

**Targeted replication** of some samples in a data set may also be appropriate. For instance, targeted replication may be pursued if there are samples of particular interest or potential sensitivity based on projected outcome of results, or if samples are known or expected to be of low quality or DNA quantity. For historical samples, has DNA concentration been evaluated by qPCR? If so, are samples replicated from the multiplex step sufficient number of times to ensure accurate genotype? Typically samples with >20copies/µl only require 3 replicates, and those with <20copies/µl require at least 5 replicates/sample. (attach qPCR results and replicate numbers). Specify targeted replicates in **Table 1**.

**Random Replication**: In addition to the targeted replication, approximately 10% of the samples in each data set should be chosen for random replication. An easy way to choose samples for random replication is to assign random numbers to the sample list in Excel, sort, and select the first 10%. Replication should start with the extracted DNA. Genescan results for the 10% replication should be analyzed blind (i.e., called first, then checked against expected genotype). This is from SNP QA/QC – do we need it?: “Controls should represent 5-10% of total number of samples (including controls if they are all genotyped from gDNA or separate multiplex PCRs rather than from a single multiplex PCR)” This step may be completed either by the PI or Laboratory Scientist. Specify random replicates in **Table 1**.

**Laboratory Scientist Responsibilities**

**Replication**: Confirm that targeted and random replication has been specified in table 1.

**Controls**: Is this a new project? If so, evaluate samples run during initial runs to select 1-2 samples as **controls** for future runs, or if not, select 1-2 samples from the previous project’s sample set to use as controls and for normalization of allele calls. Indicate the controls in **Table 1**. If this project is adding new samples to an existing data set generated using different technology (i.e. different genetic analyzer or sequencing chemistry), additional controls will be needed for use in calibration.

**Plate design**: All plates should include a set of known genotype-controls (usually 7 for SNPs, 1-2 for microsatellites) and at least three negative controls (PCR blanks and multiplex PCR blanks combined). Attach plate maps at the end of this document.

**Laboratory Marker Evaluation:**

**Optimization data set**:Identify 10-20 samples that represent the geographic range of the study to use for optimization (if the markers have not previously (recently) been used by this lab to genotype this species). Samples used for marker evaluation should be chosen in consultation with the PI to ensure that the geographic range of the study is appropriately represented. Record optimization samples in **Table 1**.

**List of markers**: Create list of potential markers **Table 2**. Record details about the markers as they are evaluated during optimization and other processing steps. As a general rule, markers selected for the project should be perfect repeats and have 5 or more alleles. For markers identified as not appropriate to use in analyses, describe why and what software were used to review the marker data in the “Why not used” column of **Table 2**.

**Laboratory Data Review**

Are there any assays or plates that had mis-called controls? List control samples/assays that were miscalled. **NO**

Are there any assays or plates that had contaminated negative controls (i.e., the negatives amplified to the point where they could not be easily separated from the samples, and confidently called as non-amplifying relative to the samples)? Was it consistent or sporadic contamination?

List NTC’s and assays that were contaminated or “positive”. **NO PLATES had contaminated negatives**

Have genotypes been called by a single individual, or double-blind called? If a second genotyper was used, what percentage was double-called? **Yes a second genotyper was used. A random twenty percent of each plate was blind called.**

**Problematic samples**: In **Table 3** list the samples for which genotyping was inaccurate or troublesome. Specify how many extractions and runs were attempted, what extraction methods were used and a general description of the problems encountered (e.g signal was low in runs, tissue was of poor quality, etc).

**Upload all the data**, including replicated and duplicated sample runs, to the Genotypes database. Update the “Use” column in the database to true or false as appropriate for each record.

**Import second caller genotypes into Genotypes QC database**, then run query to check these genotypes against the primary genotypes. Export these results and check them with the second caller.

**Import changes to be made based on second caller results into the Genotypes QC database**, then run query to implement the changes to the primary database.

**Review of Replicates/Duplicates and Estimation of Error Rates:**

Use the Genotypes database switchboard query to export all duplicates for the species or project of interest (**Table 4**). Use the attached protocol to identify mismatched alleles among the duplicates.

Estimate error rate for processing sample set. Follow the duplicates protocol to determine how many mismatched alleles there are among the random replicates and save this as a separate worksheet (“uncorrected random replicates”) in your Duplicates Excel workbook before proceeding to resolve the differences. **The replicates that had mismatched alleles were all “other duplicates”, they were not part of the random set I chose, the targeted set that KKM chose, or control sample.**

Per-allele error rate

= total number of allele differences among random replicates

=Allele 1 mismatch + Allele 2 mismatch/ (2 \* total # of random replicate genotypes)

% error rate = (per-allele error rate) \* 100

**These stats are at the bottom of sheet 4 in the excel spreadsheet.**

In a copy of your duplicates worksheet called “corrected duplicates” check all mismatched genotypes and either change the mis-called alleles or change the “use” field to “FALSE” for those genotypes deemed incorrect or unreliable. Sort these by the “update DB” field (see protocol). Ensure that all duplicate records are identical if the “Use” field is set to “true.” Record the number of genotypes that were changed or excluded among all duplicate genotypes (controls, random replicates, targeted replicates, and other duplicates).

Review available sequence and sex data for sample set. How many prompted review of all available data? (create table of notes if needed to document this here).

**Final Data Set Review** – to be completed by PI

Are there **markers** with >10% missing data? Document whether and why markers are kept or removed from a data set. **Once samples with incomplete genotypes were removed, all loci had <5% missing data**

Check whether there are duplicate individuals in the sample set. Use a program (e.g. DROPOUT (protocol attached), MStools or GenAlEx6) to look for samples that are exact or close matches for all genotypes. Exact matches are likely to be duplicates (depending on the number of loci used), and near matches may represent replicates with genotyping errors. Check all samples that are more similar than the vast majority of samples. For example, for 10-20 msat loci, different animals would not be expected to share more than 5-6 loci. For >20 SNPs, most samples should not share more than ~10-12, but some will share all but 1-3 genotypes. Add the mis-called or unresolved differences to the calculation of error rates above. Remove duplicated samples from the list of samples to use (after verifying the genotypes to know which is correct). Choice of which sample(s) to retain may be influenced by number of completed genotypes or other data collected for the sample. Summarize findings (list of perfect matches and samples that match at all but 1 or 2 loci and the resolution of differences) in Tab 5. **74882 and 104018 match at 12 out of 13 loci, but have different haplotypes. Need to follow up.**

**Incomplete genotypes:** Check for **samples** that have <80% completed genotypes (see protocols). This will reveal poor quality samples. Check each one’s history to ensure they were tried multiple times. Poor samples frequently fail when others on the same plate work fine and when they are tried multiple times. Samples with <80% completed genotypes should be flagged in the appropriate column in **Table 1** and “use (Y/N)” should be set to “N”. Insert graph (created in protocol) showing frequency of samples with missing loci under **Figure 1**. Record under ‘Processing comments’ (**Table 1**) why samples are considered to be of poor quality (e.g., failed to amplify in multiple assays, replicated inconsistently). **27 samples were excluded due to incomplete genotypes. Distribution of number of complete genotypes is shown in Excel workbook sheet Figure 1.**

Calculate % homozygosity for each sample (see protocols). This is currently a manual process of adding up homozygote genotypes for all loci, then dividing by the number of completed genotypes. Sort so that the samples with the highest % homozygosity are at the top of the list (i.e., descending order). Are there samples that are clear outliers? If not, then choose the top ~5% to check the homozygous genotypes to see if there is evidence of poor quality (e.g., they were low-signal amplifications). Samples that appear to be of poor quality and should be removed from the analysis should be flagged in the appropriate column in **Table 1** and “use (Y/N)” should be set to “N”. Insert graph (created under protocol) showing the frequency of samples with different % homozygosities under **Figure 2**. **Four samples had homozygosity >0.9. They were already excluded due to missing genos. Figure 2 appears in Excel spreadsheet.**

Perform jackknife analysis (see protocols) on the remaining samples. Samples that have log likelihood values of >2 indicate that they have a genotype at the indicated marker that is unlikely in the population and causes the population to move out of HWE for that marker. Check the indicated genotypes for accuracy and confidence, and re-call, re-genotype, or set the “Use” field in the Genotypes database to “false” for that genotype. Attach a list of samples that have an odds ratio of >2, and note whether any changes were made to the genotypes (**Table 6**). Samples that have loci that are outliers for the jackknife analysis should also be flagged in the appropriate column in **Table 1.**

Create a list of all remaining samples with their putative populations of origin (using the most inclusive likely populations). Import that list into the Analysis Set table of the Genotypes database (using the canned function in the “analysis functions” section of the switchboard). If there are multiple strata combinations, import them as separate analysis sets. List analysis sets and descriptions in **Table 7.**

For the final data set QA step before analyses, complete tests for Hardy Weinberg equilibrium and linkage disequilibrium. Download the Analysis Set from the Genotypes database using the “export analysis set” option and re-format it for GenePop using the program CREATE.

Using GenePop, perform the Hardy Weinberg Equilibrium exact test. List results in **Table 8**.

* For loci that are significantly out of HWE, recheck genotypes to see if there are genotyping errors or problems with the loci**.**

Using GenePop, analyze linkage disequilibrium for each pair of loci.List results in **Table 9.** Identify loci that are in significant LD across populations.

* For microsatellites, determine whether loci need to be analyzed as linked or if one locus should be excluded from the analysis.
* For SNPs, those loci that are linked in all populations or those loci known to be linked should be phased into haplotypes for population analysis.

MSAT PLATE DESIGN

