**Genotyping Analysis Notes**

**Study design**: Genotyping happened in two phases:

a) Pilot Study (192 samples)

b) Full cohort study (3,168 samples divided as 2968 GRLS & 200 Oldies).

**Analysis protocol**: All the samples were analyzed as one batch. Analysis was done using Applied Biosystems Array Power Tools (APT) version 1.18 on the High-Performance Computer at UC Davis. The analysis was done using a new version of the array annotation files (r2) that were obtained from Thermofisher to replace the publicly available version on the company website.

The analysis pipeline (summarized below) followed the best practice workflow recommended by Thermofisher with some modification to minimize data loss and to handle some unexpected problems of the array design and annotation.

1. **Exclusion of samples with Dish QC (DQC) < 0.82**: DQC is a QC metric based on intensities of probe sequences at non-polymorphic genome locations.
2. **Exclusion of samples with low call rate values < 95%**: Call rate was calculated by **preliminary genotyping** using a recommended subset of probe sets with known high performance.
3. **No plates were excluded for low average calling rates**: The recommended Average call rate threshold is 98.5%. Applying this threshold would exclude one plate from Array A and another 3 plates from array B. The plate on array A has a borderline call rate value (98.3873 %). The 3 plates from array B have very few numbers of samples that does not allow accurate estimation of reliable average and would not affect the quality of the caller. Therefore, none of the 3 plates were excluded.
4. **Genotyping** was done using the recommended options of the best practice workflow for large batches.
5. **Probe set classification and prioritization to identify the final recommended list of SNPs**: The probe sets are typically classified based on a group quality metrics to identify those with high quality. When more than one probe set has been designed to interrogate a SNP, the “best” probe set is identified. This was important for Array A which has 643641 probe sets for 444805 SNPs. Array B has 625277 probe sets with each probe set representing a unique SNP. The probe sets were prioritized based on this order of classes: PolyHighResolution, NoMinorHom, MonoHighResolution, OTV, UnexpectedGenotypeFreq, CallRateBelowThreshold, and Other. The X, Y and MT chromosome markers were properly annotated using the provided annotation from the company to be classified into these groups instead of being isolated in a hemizygous group. The best probe set per SNP in the top three groups were extracted to make a recommended list of SNPs for downstream analysis. The analysis resulted in **401249 and 547232** recommended markers from Array A and B respectively.
6. **Export to PLINK files**: using the SNP ids (not the probe set ids)
7. **QC check:** 
   1. A homemade script was developed to identify the actual reference allele based on genome position (CanFam3 assembly and an additional Y chromosome assembly from the Genbank with accession: KP081776.1). A SNP was kept if this allele matches one of two design alleles. **This pipeline discarded 519 markers with unknown chromosomal location and another 138 markers with bad annotation on array A**.
   2. Another sanity check showed that there are **2907 markers** on Array A with duplicate entries. Each duplicate record has 2 different SNP IDs. The Marker with higher genotyping rate was kept while its duplicate was deleted.
   3. Computed genders on both arrays were compared to the input metadata.
      1. Comparing the computed genders on the 2 arrays, the computed genders were always concordant except for two samples.
      2. Comparing the computed genders with the input metadata showed another 9 discordant samples. In addition to the identifying the gender of 6 dogs with unknown gender in the input metadata

The gender in the output PLINK files were updated to match the computed gender. In the case of the two discordant samples, the gender matching metadata was chosen.

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual ID** | **Metadata Sex** | **Computed Sex** | |
| **Array A** | **Array B** |
| **Samples with discordant computed gender on the two arrays** | | | |
| S019740 | female | female | Male |
| S005865 | Male | unknown | Male |
| **Samples concordant on both arrays but different from metadata** | | | |
| S000054 | male | female | female |
| S000055 | female | male | male |
| S016242 | female | male | male |
| S020301 | male | female | female |
| S020302 | female | male | male |
| S020391 | male | female | female |
| S020913 | female | male | male |
| S028037 | female | male | male |
| S028073 | male | female | female |
| **Samples concordant on both arrays with unknow gender in the metadata** | | | |
| S040085 | unknown | female | female |
| S040135 | unknown | male | male |
| S040195 | unknown | female | female |
| S040228 | unknown | male | male |
| S040234 | unknown | female | female |
| S040259 | unknown | male | male |