*P. monodon* Parentage Analysis – G1T1701

22nd October 2019

David B. Jones

Contents

[Preface and Objective 1](#_Toc22213563)

[Summary 2](#_Toc22213564)

[Data provided 3](#_Toc22213565)

[Revision of the G1T1701 Progeny and Broodstock lists for analysis 3](#_Toc22213566)

[Data manipulation 3](#_Toc22213567)

[DARTQC Filtering 4](#_Toc22213568)

[Exploration of SNP metrics between broodstock and progeny samples 4](#_Toc22213569)

[Broodstock Sex 4](#_Toc22213570)

[High genetic similarity between sample pairs 5](#_Toc22213571)

[Parentage analysis 9](#_Toc22213572)

[Cervus 9](#_Toc22213573)

[Colony 9](#_Toc22213574)

[APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY (MISSING ~2K PROGENY) 11](#_Toc22213575)

[APPENDIX 2 – ‘Haplotyping errors’ observed when merging DARTseq & DARTcap data 22](#_Toc22213576)

## Preface and Objective

This report outlines the revision of the input files, respective QC and parentage analysis for G1T1701.

The primary Attempt of the G1T1701 parentage was complete on the 18th Sep 2019, however, this analysis was missing 1,884 samples and needed to be redone. Details of initial parentage are still included in Appendix 1 for reference. A secondary attempt was also completed on the 14th October, but there were issues with the merging of DARTseq and DARTcap data as illustrated in Appendix 2. Corrections were made and incorporated within the current report.

## Summary

* Dataset
  + Initial merged genotypic data (DARTcap progeny and DARTseq broodstock) consisted of 1,867 common SNPs, 5,832 progeny and 411 potential broodstock
  + A total of 93 SNPs were removed as they produced ‘haplotyping’ errors during merger of the DARTcap and DARTseq datasets.
  + Moderate genotype filtering in DARTQC resulted in a genotypic dataset of 987 SNPs (read count silencing > 5 reads, count comparisons 0.05 – 0.1, SNP call rate 0.88, MAF > 0.02, repeatability > 0.9 and sequence clustering at 95%).
  + A total of 59 broodstock were removed as they had less than 8% (78) genotype calls across the 987 SNPs.
  + Pairwise genetic identity was conducted (PLINK IBS genetic distance) resulting in the removal of an additional 32 broodstock who were > 0.95 identical to each other.
  + Final SNP data for parentage analysis consisted of 987 SNPs, 321 broodstock and 5,768 progeny.
* Parentage analysis
  + Analysis was conducted initially in Cervus, and confirmed with Colony.
  + Cervus
    - A total of 4,440 parent-parent-child trios and 1,154 parent-child duo relationships were identified in Cervus and utilized as input into Colony (987 SNPs, 317 broodstock and 5,768 progeny utilized).
  + Colony
    - A total of 4,464 progeny had both parents assigned with an additional 507 only with a father assigned and 702 with only a mother assigned (987 SNPs, 147 potential fathers, 170 potential mothers and 5,768 progeny utilized).
    - A moderate skew in parental contribution was observed.
    - Parentage assignment was compared between Cervus and Colony and the Colony assignments were retained. They can be found at: “G1T1701\_ParentalAssignments\_Oct2019.csv”.
* Final notes:
  + Some broodstock had ‘near-identical’ samples based on highly congruent genotypes. Therefore assignments to one of the ‘near-identical’ samples could be interchangeable to their duplicates listed in Table 2.

## Data provided

The following raw data files were provided by Agnes, Kyall, Mehar and Jarrod on the 3rd September:

* G1T1701\_BR\_DArTSeq\_SNPs\_CloneIDAdded.csv
* G1T1701\_BR\_DArTSeq\_ReadCount.csv
* G1T1701BR\_SampleIDs.csv
* G1T1701\_ProgenyIDs.csv
* pool\_samples\_with\_ped\_v31.xlsx
* G1T1701\_Sex.csv

In addition, respective DARTcap data was downloaded from the ITRH database under ‘/genotypes/63/ORIGINAL/RC&GenotypeFiles\_11K\_CoAnalysedAllinclDBtp19-4081’:

* Report\_DBtp19-4081\_12\_moreOrders\_SNP\_2\_11K\_AllelleIDCorrected.csv
* SEQ\_SNPs\_counts\_0\_Target.csv

### Revision of the G1T1701 Progeny and Broodstock lists for analysis

Progeny list – Since the initial progeny list was missing some samples, the new progeny list consisted of all 3,948 initial progeny in “G1T1701\_ProgenyIDs.csv” plus the additional 1,884 progeny from Mehars “pool\_samples\_with\_ped\_v31.xlsx” file were merged together to make “G1T1701\_ProgenyIDs\_V2.xslx”.

Broodstock list – No additional novel broodstock were detected in addition to the 420 already in the initial broodstock list “G1T1701BR\_SampleIDs.csv”

These files were circulated to WP5 to ensure the inclusion of all broodstock and progeny was correct before proceeding.

### Data manipulation

An R script was utilized to:

* Find common SNPs between the four DARTcap and DARTseq genotypes and read count files.
* Merge the DARTcap and DARTseq datasets. A total of 1,867 SNPs were in common between the DARTcap and DARTseq dataset (by AlleleID).
* Ensure that the REF and SNP allele sequences of respective AlleleIDs matched between the DARTseq and DARTcap datasets. A total of 93 SNPs were identified as having a ‘haplotyping error’ as described in Appendix 2 and were subsequently removed from the data.
* Compare replicate samples within the merged data (by genotype ID), generate statistics and retain only the replicates with the highest call rates.

A total of 151 duplicate broodstock samples (based on IDs) were identified as being present in both DARTcap and DARTseq datasets. Concordance between these duplicate broodstock samples was 98.04% (± SD 2.44%) across the remaining common 1,774 SNPs.

## DARTQC Filtering

The effect of various DARTQC filtering thresholds were explored previously in the initial analysis (Appendix 1). Therefore, the same thresholds were retained for this revised analysis, however, SNP call rate was increased to 0.88 to prioritize the best SNPs for parentage analysis (max SNPs 1,000). These thresholds were:

* Genotypes < 5 reads were silenced,
* Count comparison thresholds between homozygous and heterozygous calls of 0.05 – 0.1,
* SNP call rate > 0.88,
* Repeatability of 0.9,
* MAF > 0.02,
* Clustering of allele sequences at 95% identity

A summary of the number of genotypes and SNPs silenced is presented in Table 1. PLINK .ped and .map files were created for ongoing analysis.



*Table 1*: Final DARTQC filtering parameters and thresholds for the common 1,774 SNPs between the DArTseq and DArTcap datasets for cohort G1T1701.

## Broodstock Sex

In order to split the broodstock into potential mothers and potential fathers for parentage assignment, broodstock sex needs to be assigned. This was done suing the “G1T1701\_Sex.csv” file and the *--update-sex* flag in PLINK (“plink --file G1T1701\_932 –recode 12 --update-sex G1T1701\_Sex.txt --out G1T1701\_932\_recode”; the recode function was necessary for Cervus input). This successfully assigned 197 males and 209 females.

## High genetic similarity between sample pairs

The initial parentage analysis (Appendix 1) returned some ambiguous results due to high genetic similarity between potential broodstock. This was influenced by two factors, 1) the low number of genotype calls for some broodstock samples (as a result of low read depth in DARTseq genotyping), and 2) genuine genetic similarity between samples (particularly broodstock).

To investigate and reduce these ambiguous assignments:

1. Sample call rates and read depths over the common filtered 987 SNPs were investigated for progeny and broodstock samples.
2. A total of 59 broodstock had less than 8% (78) genotype calls across the 987 SNPs and were therefore removed from subsequent parentage analysis to reduce ambiguity in parental assignments.
3. Pairwise identity by state (IBS) genetic distances were calculated in PLINK (plink --file G1T1701\_1556\_recode --mind 0.92 --genome --min 0.5) to identify sample pairs with high genetic similarity. As a result, 32 pairs of broodstock were identified as being near-identical (similarity of > 0.95). For broodstock pairs that returned a genetic distance higher than 0.95 (1 minus IBS\_Dist), the broodstock with the highest number of genotype calls was retained and the other was removed from subsequent parentage analysis (Table 2).

## Missing 43 samples

In the processing of the data, 43 samples were identified as missing. These samples were traced back into their respective projects, (Genotypes 26, 30 and 31) and were found to be listed as ‘non-reported’ samples. These samples therefore failed genotyping and have not been included in parentage analysis (list in Appendix 2).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| IID1 | IID2 | DST |  | IID1 | IID2 | DST |
| Progeny pairs > 0.95 IBS | | |  | Broodstock pairs > 0.95 IBS | | |
| G1T1701\_149\_010\_H6\_G | G1T1701\_149\_010\_H7\_G | 1 |  | G1T1701\_BR\_013\_P | G1T1701\_BR\_329\_GMP | 1 |
| G1T1701\_156\_009\_F10\_G | G1T1701\_156\_009\_F11\_G | 1 |  | G1T1701\_BR\_347\_M | G1T1701\_BR\_353\_MP | 1 |
| G1T1701\_157\_002\_H10\_G | G1T1701\_157\_002\_H9\_G | 1 |  | G1T1701\_BR\_070\_P | G1T1701\_BR\_170\_MP | 0.998741 |
| G1T1701\_160\_002\_C2\_G | G1T1701\_160\_002\_D2\_G | 1 |  | G1T1701\_BR\_016\_P | G1T1701\_BR\_288\_MP | 0.998454 |
| G1T1701\_161\_001\_E5\_G | G1T1701\_161\_001\_E6\_G | 1 |  | G1T1701\_BR\_024\_P | G1T1701\_BR\_306\_GMP | 0.997951 |
| G1T1701\_149\_002\_C3\_G | G1T1701\_149\_002\_D3\_G | 0.99949 |  | G1T1701\_BR\_050\_GP | G1T1701\_BR\_112\_MP | 0.99793 |
| G1T1701\_152\_002\_D10\_G | G1T1701\_152\_002\_E10\_G | 0.999486 |  | G1T1701\_BR\_030\_P | G1T1701\_BR\_256\_GMP | 0.997908 |
| G1T1701\_150\_006\_H7\_G | G1T1701\_150\_006\_H8\_G | 0.999485 |  | G1T1701\_BR\_054\_P | G1T1701\_BR\_058\_P | 0.997441 |
| G1T1701\_155\_006\_E7\_G | G1T1701\_155\_006\_F7\_G | 0.999485 |  | G1T1701\_BR\_022\_P | G1T1701\_BR\_322\_MP | 0.997412 |
| G1T1701\_156\_001\_F10\_G | G1T1701\_156\_001\_G10\_G | 0.999485 |  | G1T1701\_BR\_058\_P | G1T1701\_BR\_068\_P | 0.996929 |
| G1T1701\_150\_006\_G7\_G | G1T1701\_150\_006\_G8\_G | 0.999484 |  | G1T1701\_BR\_042\_P | G1T1701\_BR\_270\_GMP | 0.996921 |
| G1T1701\_156\_009\_B5\_G | G1T1701\_156\_009\_B6\_G | 0.999483 |  | G1T1701\_BR\_025\_P | G1T1701\_BR\_293\_MP | 0.996917 |
| G1T1701\_150\_008\_G7\_G | G1T1701\_150\_008\_G8\_G | 0.999482 |  | G1T1701\_BR\_036\_P | G1T1701\_BR\_258\_MP | 0.996875 |
| G1T1701\_150\_006\_F7\_G | G1T1701\_150\_006\_F8\_G | 0.99948 |  | G1T1701\_BR\_061\_P | G1T1701\_BR\_082\_MP | 0.996868 |
| G1T1701\_155\_002\_C6\_G | G1T1701\_155\_002\_D6\_G | 0.999479 |  | G1T1701\_BR\_461\_P | G1T1701\_BR\_408\_GMP | 0.996583 |
| G1T1701\_156\_003\_C8\_G | G1T1701\_156\_003\_D8\_G | 0.999443 |  | G1T1701\_BR\_054\_P | G1T1701\_BR\_068\_P | 0.996421 |
| G1T1701\_152\_001\_D8\_G | G1T1701\_152\_001\_E8\_G | 0.998979 |  | G1T1701\_BR\_010\_P | G1T1701\_BR\_301\_MP | 0.996395 |
| G1T1701\_149\_007\_B5\_G | G1T1701\_149\_007\_B6\_G | 0.998968 |  | G1T1701\_BR\_003\_P | G1T1701\_BR\_438\_MP | 0.996331 |
| G1T1701\_150\_003\_G8\_G | G1T1701\_150\_003\_G9\_G | 0.998967 |  | G1T1701\_BR\_340\_MP | G1T1701\_BR\_457 | 0.996139 |
| G1T1701\_156\_009\_E5\_G | G1T1701\_156\_009\_E6\_G | 0.998963 |  | G1T1701\_BR\_062\_P | G1T1701\_BR\_182\_MP | 0.995833 |
| G1T1701\_156\_009\_H5\_G | G1T1701\_156\_009\_H6\_G | 0.998962 |  | G1T1701\_BR\_055\_P | G1T1701\_BR\_159\_MP | 0.995772 |
| G1T1701\_149\_006\_G10\_G | G1T1701\_149\_006\_G11\_G | 0.998957 |  | G1T1701\_BR\_007\_P | G1T1701\_BR\_436\_GMP | 0.995253 |
| G1T1701\_149\_010\_E6\_G | G1T1701\_149\_010\_E7\_G | 0.998954 |  | G1T1701\_BR\_026\_P | G1T1701\_BR\_299\_MP | 0.99484 |
| G1T1701\_156\_009\_G5\_G | G1T1701\_156\_009\_G6\_G | 0.998947 |  | G1T1701\_BR\_014\_P | G1T1701\_BR\_430\_MP | 0.994709 |
| G1T1701\_155\_010\_B7\_G | G1T1701\_155\_010\_B8\_G | 0.998942 |  | G1T1701\_BR\_056\_P | G1T1701\_BR\_088\_MP | 0.994301 |
| G1T1701\_161\_003\_E2\_G | G1T1701\_161\_003\_E3\_G | 0.99894 |  | G1T1701\_BR\_032\_P | G1T1701\_BR\_455\_GMP | 0.993681 |
| G1T1701\_156\_009\_D5\_G | G1T1701\_156\_009\_D6\_G | 0.998446 |  | G1T1701\_BR\_051\_P | G1T1701\_BR\_145\_M | 0.993519 |
| G1T1701\_156\_009\_E10\_G | G1T1701\_156\_009\_E11\_G | 0.998446 |  | G1T1701\_BR\_060\_P | G1T1701\_BR\_090\_M | 0.991667 |
| G1T1701\_155\_008\_A7\_G | G1T1701\_155\_008\_B7\_G | 0.998442 |  | G1T1701\_BR\_198\_M | G1T1701\_BR\_071\_P | 0.991311 |
| G1T1701\_156\_008\_E10\_G | G1T1701\_156\_008\_E8\_G | 0.998441 |  | G1T1701\_BR\_059\_P | G1T1701\_BR\_196\_M | 0.987952 |
| G1T1701\_149\_008\_D5\_G | G1T1701\_149\_008\_D6\_G | 0.998429 |  | G1T1701\_BR\_057\_GP | G1T1701\_BR\_103\_MP | 0.976247 |
| G1T1701\_149\_010\_F6\_G | G1T1701\_149\_010\_F7\_G | 0.998428 |  | G1T1701\_BR\_063\_P | G1T1701\_BR\_136\_M | 0.966216 |
| G1T1701\_149\_006\_H10\_G | G1T1701\_149\_006\_H11\_G | 0.998424 |  |  |  |  |
| G1T1701\_155\_006\_D9\_G | G1T1701\_155\_006\_E9\_G | 0.998421 |  |  |  |  |
| G1T1701\_149\_005\_B3\_G | G1T1701\_149\_005\_B4\_G | 0.998419 |  |  |  |  |
| G1T1701\_150\_004\_E2\_G | G1T1701\_150\_004\_E3\_G | 0.998416 |  |  |  |  |
| G1T1701\_155\_010\_A6\_G | G1T1701\_155\_010\_A8\_G | 0.998408 |  |  |  |  |
| G1T1701\_156\_002\_F4\_G | G1T1701\_156\_002\_F5\_G | 0.998391 |  |  |  |  |
| G1T1701\_156\_009\_F5\_G | G1T1701\_156\_009\_F6\_G | 0.997925 |  |  |  |  |
| G1T1701\_157\_002\_E7\_G | G1T1701\_157\_002\_F7\_G | 0.997925 |  |  |  |  |
| G1T1701\_149\_006\_E9\_G | G1T1701\_149\_006\_F9\_G | 0.997917 |  | KEY |  |  |
| G1T1701\_156\_009\_D10\_G | G1T1701\_156\_009\_D11\_G | 0.99791 |  | Excluded broodstock |  |  |
| G1T1701\_156\_010\_C2\_G | G1T1701\_156\_010\_C3\_G | 0.997888 |  |  |  |  |
| G1T1701\_156\_003\_E8\_G | G1T1701\_156\_003\_F8\_G | 0.997875 |  |  |  |  |
| G1T1701\_156\_004\_H10\_G | G1T1701\_156\_004\_H9\_G | 0.997875 |  |  |  |  |
| G1T1701\_149\_005\_G8\_G | G1T1701\_149\_005\_G9\_G | 0.997868 |  |  |  |  |
| G1T1701\_149\_005\_H8\_G | G1T1701\_149\_005\_H9\_G | 0.997849 |  |  |  |  |
| G1T1701\_160\_001\_F3\_G | G1T1701\_160\_001\_G3\_G | 0.997417 |  |  |  |  |
| G1T1701\_155\_008\_D4\_G | G1T1701\_155\_008\_E4\_G | 0.997379 |  |  |  |  |
| G1T1701\_149\_010\_G6\_G | G1T1701\_149\_010\_G7\_G | 0.997371 |  |  |  |  |
| G1T1701\_150\_008\_F7\_G | G1T1701\_150\_008\_F8\_G | 0.997323 |  |  |  |  |
| G1T1701\_155\_004\_B10\_G | G1T1701\_155\_004\_C10\_G | 0.997309 |  |  |  |  |
| G1T1701\_152\_001\_C6\_G | G1T1701\_152\_001\_C7\_G | 0.99697 |  |  |  |  |
| G1T1701\_156\_002\_D2\_G | G1T1701\_156\_002\_D3\_G | 0.996881 |  |  |  |  |
| G1T1701\_155\_008\_B2\_G | G1T1701\_155\_008\_B3\_G | 0.996859 |  |  |  |  |
| G1T1701\_150\_003\_F6\_G | G1T1701\_150\_003\_F7\_G | 0.996805 |  |  |  |  |
| G1T1701\_149\_003\_A4\_G | G1T1701\_149\_003\_B4\_G | 0.996767 |  |  |  |  |
| G1T1701\_149\_005\_H10\_G | G1T1701\_149\_005\_H11\_G | 0.996767 |  |  |  |  |
| G1T1701\_149\_005\_E7\_G | G1T1701\_149\_005\_F7\_G | 0.99675 |  |  |  |  |
| G1T1701\_152\_002\_F7\_G | G1T1701\_152\_002\_F8\_G | 0.996618 |  |  |  |  |
| G1T1701\_152\_002\_G5\_G | G1T1701\_152\_002\_G6\_G | 0.996369 |  |  |  |  |
| G1T1701\_155\_006\_C3\_G | G1T1701\_155\_006\_C4\_G | 0.996312 |  |  |  |  |
| G1T1701\_149\_005\_G10\_G | G1T1701\_149\_005\_G11\_G | 0.996292 |  |  |  |  |
| G1T1701\_156\_009\_A5\_G | G1T1701\_156\_009\_A6\_G | 0.995798 |  |  |  |  |
| G1T1701\_150\_010\_D11\_G | G1T1701\_150\_010\_E11\_G | 0.994781 |  |  |  |  |
| G1T1701\_156\_008\_E4\_G | G1T1701\_156\_008\_F4\_G | 0.994737 |  |  |  |  |
| G1T1701\_156\_009\_C5\_G | G1T1701\_156\_009\_C6\_G | 0.993011 |  |  |  |  |
| G1T1701\_156\_008\_H10\_G | G1T1701\_156\_008\_H8\_G | 0.990042 |  |  |  |  |
| G1T1701\_161\_001\_C8\_G | G1T1701\_161\_001\_C9\_G | 0.978434 |  |  |  |  |
| G1T1701\_155\_008\_H6\_G | G1T1701\_155\_008\_H7\_G | 0.973456 |  |  |  |  |
| G1T1701\_161\_004\_C9\_G | G1T1701\_161\_004\_D9\_G | 0.961493 |  |  |  |  |
| G1T1701\_150\_004\_A10\_G | G1T1701\_150\_004\_B10\_G | 0.959659 |  |  |  |  |
| G1T1701\_BR\_048\_P | G1T1701\_BR\_125\_M | 0.957944 |  |  |  |  |
| G1T1701\_156\_003\_G3\_G | G1T1701\_156\_003\_H5\_G | 0.95735 |  |  |  |  |
| G1T1701\_150\_004\_C10\_G | G1T1701\_150\_004\_D10\_G | 0.955347 |  |  |  |  |
| G1T1701\_150\_004\_F10\_G | G1T1701\_150\_004\_G10\_G | 0.954435 |  |  |  |  |
| G1T1701\_150\_004\_B10\_G | G1T1701\_150\_004\_C10\_G | 0.952103 |  |  |  |  |
| G1T1701\_156\_010\_B4\_G | G1T1701\_156\_010\_B5\_G | 0.95132 |  |  |  |  |
| G1T1701\_150\_004\_G10\_G | G1T1701\_150\_004\_H10\_G | 0.950932 |  |  |  |  |

Table 2: Sample pairs with extremely high pairwise genetic identity for broodstock pairs and progeny pairs (> 0.95).

## Parentage analysis

### Cervus

Final SNP data for parentage analysis consisted of 987 SNPs.

1. Allele frequency analysis was run with all 6,089 samples (321 broodstock & 5,768 progeny).
2. Simulation of parentage analysis was run as parent pair with known sexes across 6,000 offspring, 200 candidate mothers (0.9 proportion sampled), 200 candidate father (0.9 proportion sampled), proportion loci typed 0.9, proportion of loci mistyped 0.05, minimum loci typed 200 (run time was 13.5 hours).
3. Parentage analysis for parent pairs (known sex) was run by specifying all potential mothers and fathers (8.5 hrs run time). All parent trios (PPCs) with positive LOD scores were reported. The ‘primary’ trio assignment was one made for each progeny with the highest trio LOD score. Other positive trio LOD score assignments for offspring were classed as ‘secondary’.
4. Approximately 200 parental trio-relationship LOD scores were cross checked to ensure no spurious trio assignments were made. The majority of ‘secondary’ assignments were to broodstock with a low number of genotyped SNPs (from 200-250) compared to the ~980 in ‘primary’ assignments. In all cases, the ‘primary’ assignment was justified over the ‘secondary’ parental assignments, therefore all assignments with the highest Trio LOD score were trusted and retained for the next step.
5. Based in the initial data exploration of Cervus results (Appendix 1), individual parental assignments < LOD 10 were removed before proceeding (progeny vs single parent).
6. Also based on the mating design of this cohort, one broodstock could only have contributed to 2-3 families. Assignments were crosschecked and no broodstock were assigned to more than three family groupings.

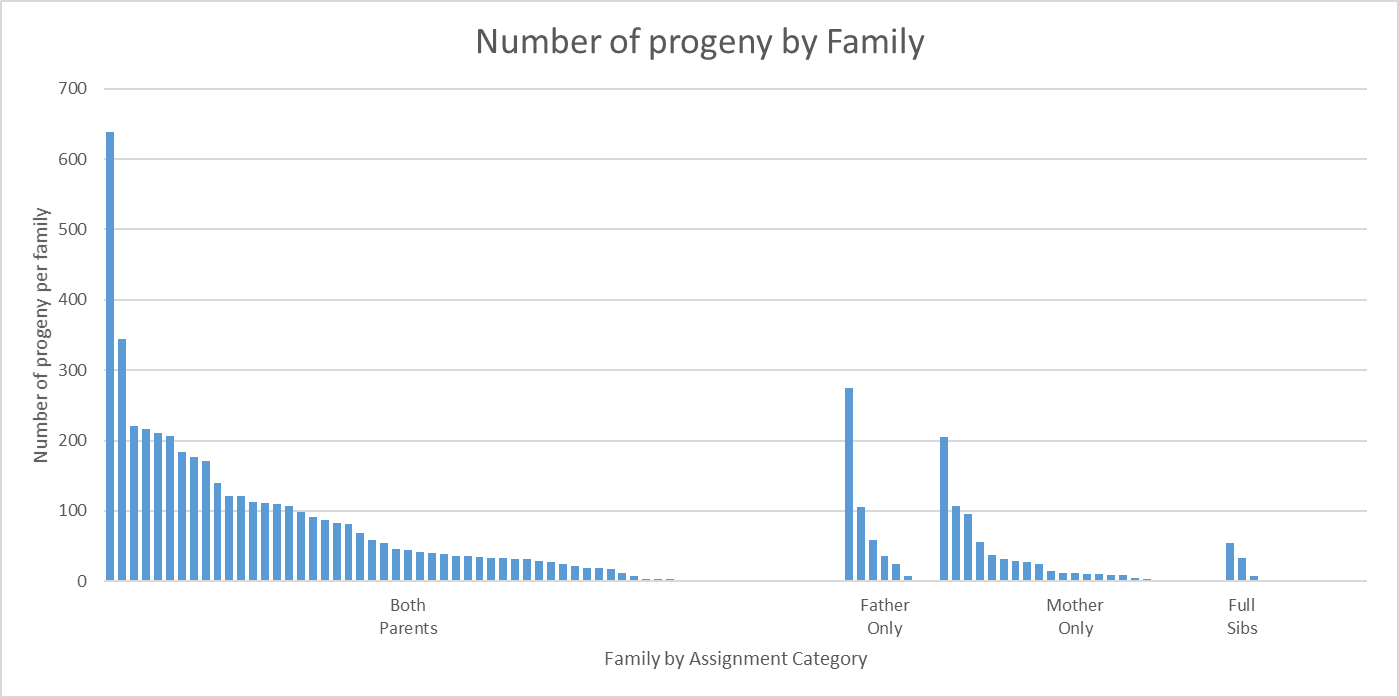
### Colony

1. All Cervus assigned parental relationships (4,440 PPCs and 1,154 additional PCs after individual parents with assignments < LOD 10 were removed) were imported into Colony alongside a dataset containing 987 SNPs, 147 potential fathers, 170 potential mothers and 5,768 progeny.
2. Settings and parameters consisted of: updating allele frequencies, dioecious species, no inbreeding, diploid, polygamy for both sexes, no clones, scaling of full sib relationships, weak sibship priors, average sibship size of 60, unknown population allele frequency, 1 run, short run length, run monitored by every 10,000th iterate, run in dos mode, full-likelihood, medium precision and maternal and paternal known sibships were included as inputs with a mismatch of 80 SNPs allowed (up to a 8% mismatch accepted in known parental relationships).
3. *Note*: the input Colony.dat file was created in the Colony GUI, but edited and run in dos mode (mpiexec -localonly -n 40 colony2p.exe) to allow multithreading over 40 cores. Total runtime was 2 days and 2 hours.
4. Individual paternity and maternity assignments as well as BestCluster assignments with probabilities were compared to the initial Colony results to determine congruency.
5. Colony parentage assignment, fullsib and halfsib cluster determination was much more robust than Cervus and therefore all Colony assignments were retained.
6. A total of 4,464 progeny had both parents assigned with an additional 507 only with a father assigned and 702 with only a mother assigned (Table 5). In addition, Colony has assigned fullsib clusters among progeny with or without parental assignments.
7. A moderate skew in was observed with parental contributions across the parental assignment categories (Figure 4).
8. All parentage assignments are listed in “G1T1701\_ParentalAssignments\_Oct2019.csv”.

|  |  |  |
| --- | --- | --- |
| Colony Category | #Progeny | #Family Groups |
| Trio Confirmed | 4464 | 62 |
| Father Confirmed | 507 | 8 |
| Mother Confirmed | 702 | 24 |
| FullSib Cluster Confirmed | 95 | 3 |
| TOTAL | 5786 | 89 |

Table 5: Summary of parental assignments across progeny and parent pair family groups.

Figure 4: Proportions of progeny assigned to parents as categorized in parental pairs, father only assignments, mother only assignments and fullsib clusters.



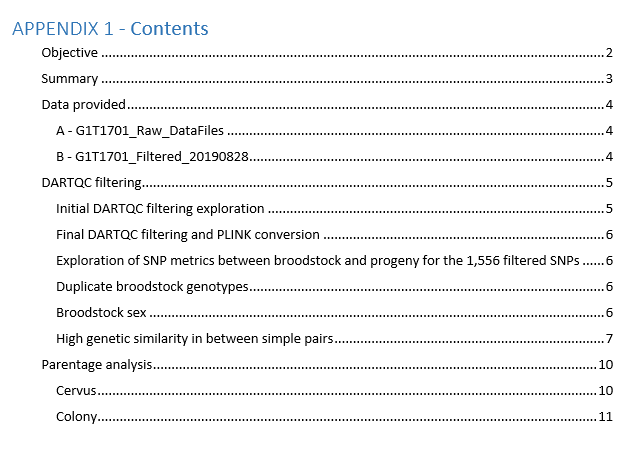
# APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY (MISSING ~2K PROGENY)

APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

*P. monodon* Parentage Analysis – G1T1701

17th September 2019

David B. Jones



**Objective**

Assign parentage to the G1T1701 cohort based on the data and lists of broodstock and progeny provided.

Initial complications with the DARTseq and DARTcap genotype data emerged and were addressed as outlined throughout this report.

**Summary**

APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

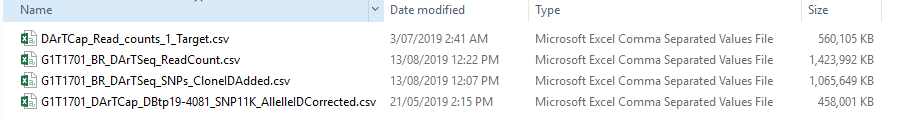
* *Dataset*
  + Initial merged genotypic data (DARTcap progeny and DARTseq broodstock) consisted of 1,944 common SNPs, 3,671 progeny and 411 potential broodstock.
  + Moderate genotype filtering in DARTQC resulted in a genotypic dataset of 1,556 SNPs (read count silencing > 5 reads, count comparisons 0.05 – 0.1, SNP call rate 0.8, MAF > 0.02, repeatability > 0.9 and sequence clustering at 95%).
  + A total of 64 broodstock were removed as they had less than 100 genotype calls across the 1,556 SNPs.
  + Pairwise genetic identity was conducted (PLINK IBS genetic distance) resulting in the removal of an additional 30 broodstock who were > 0.95 identical to each other.
  + Final SNP data for parentage analysis consisted of 1,556 SNPs, 317 broodstock and 3,671 progeny.
* *Parentage analysis*
  + Analysis was conducted initially in Cervus, and confirmed with Colony.
  + *Cervus*
    - A total of 2,624 parent-parent-child trios and 493 parent-child duo relationships were identified in Cervus and utilized as input into Colony (1,556 SNPs, 317 broodstock and 3,671 progeny utilized).
  + *Colony*
    - A total of 2,729 progeny had both parents assigned with an additional 417 only with a father assigned and 384 with only a mother assigned (1,000 SNPs, 148 potential fathers, 168 potential mothers and 3,671 progeny utilized).
    - The dataset was reduced from 1,556 SNPs to 1,000 as a constraint of Colony’s computational capacity.
    - A moderate skew in parental contribution was observed.
  + All results including working comparisons sheets can be found in the files: “ColonyCervusComparison.xlsx” and “G1T1701\_ParentalAssignments.csv”.
* *Final notes:* 
  + If you want to be conservative, remove all parental assignments to BR\_148\_M, BR\_067\_GP, BR\_183\_MP and BR\_272\_MP.
  + Some broodstock had ‘near-identical’ samples based on highly congruent genotypes. Therefore assignments to one of the ‘near-identical’ samples could be interchangeable to their duplicates listed in Table 4.

**Data provided**

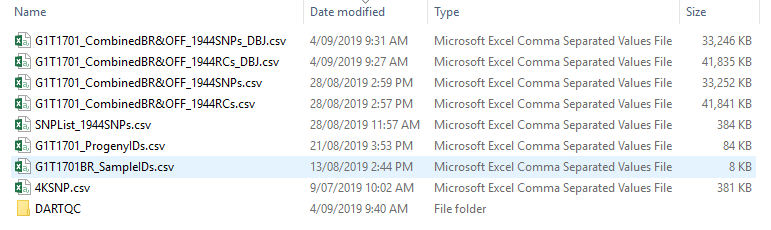
APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

Data was provided by Agnes, Kyall and Jarrod on the 3rd September (Table 1).

*A - G1T1701\_Raw\_DataFiles*



*B - G1T1701\_Filtered\_20190828*



*Table 1*: list of files provided for parentage analysis of G1T1701 in Sep 2019. A) The raw data files for DARTseq and DARTcap data. B) Files related to the merged DARTseq and DARTcap datasets containing the 1,944 common SNPs.

All analysis was conducted using the “G1T1701\_Filtered\_20190828” files. These files contain the SNPs that are common between the DArTseq and DArTcap files provided in “G1T1701\_Raw\_DataFiles”. The “G1T1701\_Filtered\_20190828” files with the extra extension of “\_DBJ” were modified slightly, allowing for import into DARTQC to investigate filtering parameters and export PLINK files. In addition to these files above, “G1T1701\_Sex.csv.csv” was provided by Jarrod as it is required to determine potential mothers from fathers during parentage analysis.

**DARTQC filtering**

APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

*Initial DARTQC filtering exploration*

The files “G1T1701\_CombinedBR&OFF\_1944SNPs\_DBJ.csv and “G1T1701\_CombinedBR&OFF\_1944RCs\_DBJ.csv” were imported into DARTQC and run through a series of filtering thresholds (Table 2). The aim of this was to generate SNP metrics and optimize the dataset for parentage.

A ‘baseline’ filter set was run with relaxed thresholds (read count silencing > 5 reads, count comparisons 0.05 – 0.1, SNP call rate 0.5, MAF > 0.01, Repeatability > 0.9 and sequence clustering at 95%). Then filter parameters were changed independently to see their effects. Optimal filtering thresholds were then locked in for respective parameters. Since there was no sequence clustering of the SNP probe sequences at the first baseline filtering thresholds, clustering was not run over subsequent filtering sets. Also no SNPs were removed due to having an average reproducibility < 0.9 in subsequent filter sets so this parameter was not reported.

Since parentage analysis is sensitive to genotyping error rates and missing data, genotype silencing was moderate (5 read counts), SNPs with higher call rate (> 0.8) were prioritized, as well as SNPs with MAF > 0.02. This resulted in 1,556 SNPs being retained for parentage analysis (Table 3). Increasing these thresholds lead to a much higher proportion of genotypes being removed and a subsequent drop in the number of broodstock available for parentage analysis.



*Table 2*: The number of SNPs and genotype calls silenced during initial DARTQC filtering exploration of the common 1,944 SNPs between the DArTseq and DArTcap datasets for cohort G1T1701. (A) Baseline filtering parameters and thresholds, (B) effect of increase call rate thresholds, (C) effect of increasing MAF, (D) effect of increasing read count silencing.

**Final DARTQC filtering and PLINK conversion**

After initial DARTQC data exploration, the following thresholds were applied to the data (Table 3) and filtered data was output in PLINK format (.ped and .map file). As always, filtering was a compromise between silencing too many genotypes and reducing genotype errors. In this case, since all broodstock had low call rates, we couldn’t be too strict with genotype silencing.

APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY



*Table 3*: Final DARTQC filtering parameters and thresholds for the common 1,944 SNPs between the DArTseq and DArTcap datasets for cohort G1T1701.

**Exploration of SNP metrics between broodstock and progeny 1,556 filtered SNPs**

Considering the broodstock were all run on DArTseq technology and the progeny were all run on the DArTcap platform, some sample specific metrics (call rate and average read depth) were calculated over the filtered 1,556 SNPs to quantify differences in the genotyping platforms. The average read depth across broodstock samples was 18.8 (± 16.8 SD) compared to 28.7 (± 8.5 SD) for the progeny samples. In addition, the average sample call rate across broodstock samples was 75.8% (± 33.3 SD) compared to 96.1% (± 3.1 SD) for the progeny samples. This is expected considering the enrichment conducted in the DArTcap platform leading to higher read depth and call rates in progeny samples.

**Duplicate broodstock genotypes**

As a result of being genotyped on both the DARTseq and DARTcap platforms, there are 151 duplicate broodstock samples in the merged data (G1T1701\_CombinedBR&OFF\_1944SNPs\_DBJ.csv). Even though the call rate of duplicate samples were higher in DARTcap data vs DARTseq data, the average concordance of genotype calls between duplicate samples run on DARTseq vs DARTcap was 98.2% (± 2.4 SD) after DARTQC filtering. As a result, all broodstock genotypes produced on the DARTcap platform were removed (*n* = 151), leaving just the DARTseq genotypes for all broodstock.

**Broodstock sex**

In order to split the broodstock into potential mothers and potential fathers for parentage assignment, broodstock sex needs to be assigned. This was done using the “G1T1701\_Sex.csv” file and the *--update-sex* flag in PLINK (“plink --file G1T1701\_1556 --recode12 --update-sex G1T1701\_Sex.txt --out G1T1701\_1556\_recode”; the recode function was necessary for Cervus input). This successfully assigned 197 males and 209 females.

**High genetic similarity in between sample pairs**

Initial parentage with the filtered dataset above returned some ambiguous results in broodstock assignment due to high genetic similarity between potential broodstock (multiple parental assignments with ± 1 mismatch). This was influenced by two factors, 1) the low number of genotype calls for some broodstock samples (as a result of low read depth in DARTseq genotyping), and 2) genuine genetic similarity between samples (particularly broodstock).

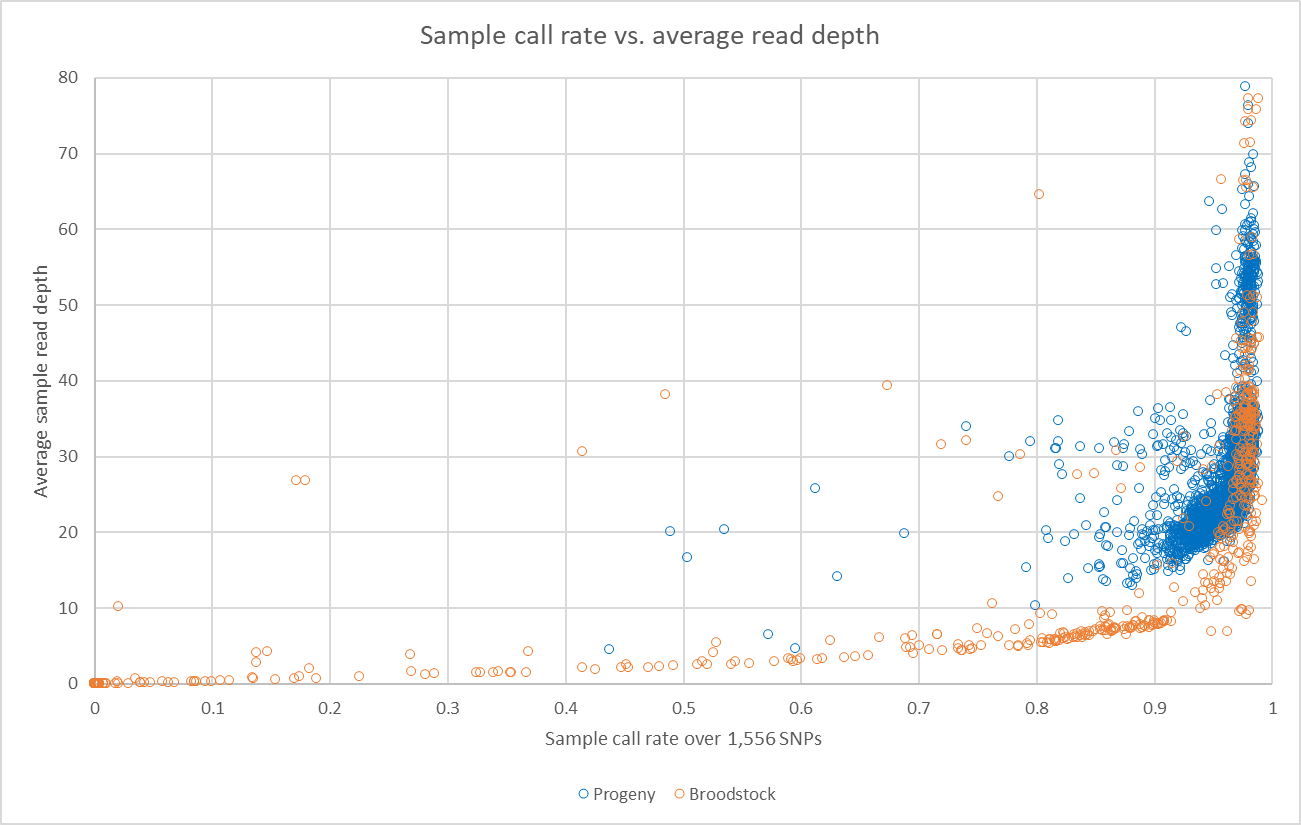
To investigate and reduce these ambiguous assignments:

1. Sample call rates and read depths over the common filtered 1,556 SNPs were plotted by progeny and broodstock (Figure 1). A total of 64 broodstock has less than 100 genotype calls across the 1,556 SNPs and were therefore removed from subsequent parentage analysis to reduce ambiguity in parental assignments.

APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

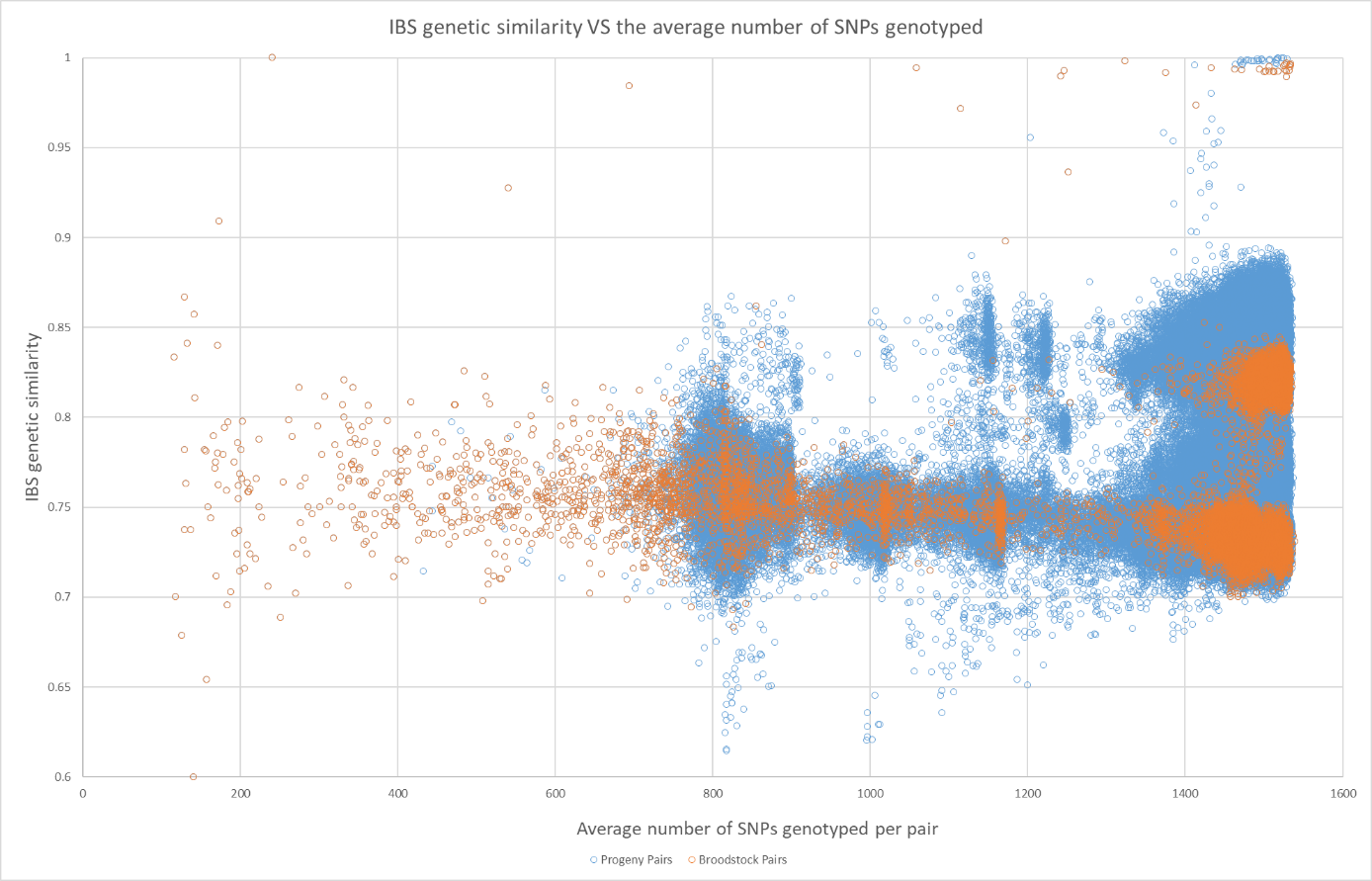
1. Pairwise identity by state (IBS) genetic distances were calculated in PLINK (plink --file G1T1701\_1556\_recode --genome --min 0.05) and plotted against the average number of SNPs genotyped to identify sample pairs with high genetic similarity (Figure 2). As a result, 30 pairs of broodstock were identified as being near-identical (similarity of > 0.95). For broodstock pairs that returned a genetic distance higher than 0.95 (1 minus IBS\_Dist), the broodstock with the highest number of genotype calls was retained and the other was removed from subsequent parentage analysis (Figure 2, Table 4).

Figure 1: Sample call rate and average read depth for the 1,556 filtered SNPs across broodstock and progeny (broodstock orange and progeny blue).



APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

Figure 2: Pairwise IBS genetic similarities and the average number of observed genotype calls between each pairwise comparison (broodstock pairs orange and progeny pairs blue). The cluster of high pairwise genetic similarity and high call rate indicates near-identical samples.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| IID1 | FID1#Genos | IID2 | FID2#Genos | av#genos | PI\_HAT | IBA\_DST |
| Broodstock Pairs > 0.95 Genetic Distance | | | | | | |
| G1T1701\_BR\_353\_MP | 271 | G1T1701\_BR\_347\_M | 210 | 240.5 | 1.000 | 1.000 |
| G1T1701\_BR\_170\_MP | 1412 | G1T1701\_BR\_070\_P | 1236 | 1324 | 0.995 | 0.998 |
| G1T1701\_BR\_024\_P | 1531 | G1T1701\_BR\_306\_GMP | 1526 | 1528.5 | 0.991 | 0.997 |
| G1T1701\_BR\_058\_P | 1535 | G1T1701\_BR\_054\_P | 1533 | 1534 | 0.988 | 0.997 |
| G1T1701\_BR\_061\_P | 1528 | G1T1701\_BR\_082\_MP | 1526 | 1527 | 0.988 | 0.996 |
| G1T1701\_BR\_058\_P | 1535 | G1T1701\_BR\_068\_P | 1533 | 1534 | 0.987 | 0.996 |
| G1T1701\_BR\_054\_P | 1533 | G1T1701\_BR\_068\_P | 1533 | 1533 | 0.987 | 0.996 |
| G1T1701\_BR\_293\_MP | 1526 | G1T1701\_BR\_025\_P | 1524 | 1525 | 0.986 | 0.995 |
| G1T1701\_BR\_003\_P | 1530 | G1T1701\_BR\_438\_MP | 1484 | 1507 | 0.985 | 0.995 |
| G1T1701\_BR\_301\_MP | 1534 | G1T1701\_BR\_010\_P | 1533 | 1533.5 | 0.984 | 0.995 |
| G1T1701\_BR\_145\_M | 1113 | G1T1701\_BR\_051\_P | 1005 | 1059 | 0.984 | 0.994 |
| G1T1701\_BR\_461\_P | 1523 | G1T1701\_BR\_408\_GMP | 1344 | 1433.5 | 0.983 | 0.994 |
| G1T1701\_BR\_050\_GP | 1526 | G1T1701\_BR\_112\_MP | 1400 | 1463 | 0.981 | 0.994 |
| G1T1701\_BR\_060\_P | 1528 | G1T1701\_BR\_090\_M | 1527 | 1527.5 | 0.980 | 0.993 |
| G1T1701\_BR\_022\_P | 1525 | G1T1701\_BR\_322\_MP | 1464 | 1494.5 | 0.979 | 0.994 |
| G1T1701\_BR\_032\_P | 1533 | G1T1701\_BR\_455\_GMP | 961 | 1247 | 0.978 | 0.993 |
| G1T1701\_BR\_299\_MP | 1526 | G1T1701\_BR\_026\_P | 1417 | 1471.5 | 0.978 | 0.993 |
| G1T1701\_BR\_014\_P | 1526 | G1T1701\_BR\_430\_MP | 1479 | 1502.5 | 0.978 | 0.993 |
| G1T1701\_BR\_329\_GMP | 1531 | G1T1701\_BR\_013\_P | 1526 | 1528.5 | 0.978 | 0.992 |
| G1T1701\_BR\_288\_MP | 1531 | G1T1701\_BR\_016\_P | 1505 | 1518 | 0.977 | 0.992 |
| G1T1701\_BR\_030\_P | 1527 | G1T1701\_BR\_256\_GMP | 1498 | 1512.5 | 0.977 | 0.992 |
| G1T1701\_BR\_055\_P | 1526 | G1T1701\_BR\_159\_MP | 1488 | 1507 | 0.977 | 0.992 |
| G1T1701\_BR\_007\_P | 1527 | G1T1701\_BR\_436\_GMP | 1475 | 1501 | 0.977 | 0.992 |
| G1T1701\_BR\_036\_P | 1526 | G1T1701\_BR\_258\_MP | 1500 | 1513 | 0.976 | 0.992 |
| G1T1701\_BR\_062\_P | 1534 | G1T1701\_BR\_182\_MP | 1530 | 1532 | 0.975 | 0.993 |
| G1T1701\_BR\_042\_P | 1519 | G1T1701\_BR\_270\_GMP | 1232 | 1375.5 | 0.975 | 0.991 |
| G1T1701\_BR\_198\_M | 1530 | G1T1701\_BR\_071\_P | 955 | 1242.5 | 0.969 | 0.990 |
| G1T1701\_BR\_088\_MP | 1529 | G1T1701\_BR\_056\_P | 1528 | 1528.5 | 0.967 | 0.990 |
| G1T1701\_BR\_059\_P | 1261 | G1T1701\_BR\_196\_M | 128 | 694.5 | 0.953 | 0.984 |
| G1T1701\_BR\_057\_GP | 1530 | G1T1701\_BR\_103\_MP | 1299 | 1414.5 | 0.921 | 0.973 |
| G1T1701\_BR\_063\_P | 1526 | G1T1701\_BR\_136\_M | 705 | 1115.5 | 0.915 | 0.972 |
| Progeny Pairs > 0.95 Genetic Distance | | | | | | |
| G1T1701\_161\_001\_E5\_G | 1525 | G1T1701\_161\_001\_E6\_G | 1525 | 1525 | 0.999 | 1.000 |
| G1T1701\_155\_002\_D6\_G | 1524 | G1T1701\_155\_002\_C6\_G | 1521 | 1522.5 | 0.999 | 1.000 |
| G1T1701\_152\_002\_E10\_G | 1523 | G1T1701\_152\_002\_D10\_G | 1514 | 1518.5 | 0.999 | 1.000 |
| G1T1701\_156\_001\_F10\_G | 1520 | G1T1701\_156\_001\_G10\_G | 1515 | 1517.5 | 0.999 | 1.000 |
| G1T1701\_152\_001\_E8\_G | 1532 | G1T1701\_152\_001\_D8\_G | 1527 | 1529.5 | 0.998 | 0.999 |
| G1T1701\_149\_005\_G11\_G | 1502 | G1T1701\_149\_005\_G10\_G | 1482 | 1492 | 0.998 | 0.999 |
| G1T1701\_150\_004\_E3\_G | 1502 | G1T1701\_150\_004\_E2\_G | 1493 | 1497.5 | 0.998 | 0.999 |
| G1T1701\_161\_003\_E3\_G | 1529 | G1T1701\_161\_003\_E2\_G | 1486 | 1507.5 | 0.997 | 0.999 |
| G1T1701\_157\_002\_H9\_G | 1519 | G1T1701\_157\_002\_H10\_G | 1514 | 1516.5 | 0.997 | 0.999 |
| G1T1701\_149\_005\_B3\_G | 1518 | G1T1701\_149\_005\_B4\_G | 1477 | 1497.5 | 0.997 | 0.999 |
| G1T1701\_160\_001\_F3\_G | 1517 | G1T1701\_160\_001\_G3\_G | 1515 | 1516 | 0.996 | 0.999 |
| G1T1701\_152\_002\_G6\_G | 1514 | G1T1701\_152\_002\_G5\_G | 1500 | 1507 | 0.996 | 0.999 |
| G1T1701\_150\_003\_F6\_G | 1481 | G1T1701\_150\_003\_F7\_G | 1480 | 1480.5 | 0.996 | 0.999 |
| G1T1701\_155\_004\_B10\_G | 1481 | G1T1701\_155\_004\_C10\_G | 1475 | 1478 | 0.996 | 0.999 |
| G1T1701\_156\_003\_C8\_G | 1484 | G1T1701\_156\_003\_D8\_G | 1458 | 1471 | 0.996 | 0.999 |
| G1T1701\_150\_003\_G8\_G | 1499 | G1T1701\_150\_003\_G9\_G | 1499 | 1499 | 0.995 | 0.998 |
| G1T1701\_156\_003\_E8\_G | 1498 | G1T1701\_156\_003\_F8\_G | 1490 | 1494 | 0.995 | 0.998 |
| G1T1701\_149\_005\_G8\_G | 1491 | G1T1701\_149\_005\_G9\_G | 1484 | 1487.5 | 0.995 | 0.998 |
| G1T1701\_156\_004\_H10\_G | 1484 | G1T1701\_156\_004\_H9\_G | 1484 | 1484 | 0.995 | 0.998 |
| G1T1701\_156\_002\_F4\_G | 1495 | G1T1701\_156\_002\_F5\_G | 1489 | 1492 | 0.994 | 0.998 |
| G1T1701\_149\_005\_E7\_G | 1475 | G1T1701\_149\_005\_F7\_G | 1465 | 1470 | 0.993 | 0.998 |
| G1T1701\_149\_005\_H8\_G | 1480 | G1T1701\_149\_005\_H9\_G | 1470 | 1475 | 0.992 | 0.997 |
| G1T1701\_157\_002\_F7\_G | 1522 | G1T1701\_157\_002\_E7\_G | 1507 | 1514.5 | 0.991 | 0.997 |
| G1T1701\_156\_002\_D2\_G | 1519 | G1T1701\_156\_002\_D3\_G | 1515 | 1517 | 0.990 | 0.997 |
| G1T1701\_149\_005\_H11\_G | 1486 | G1T1701\_149\_005\_H10\_G | 1460 | 1473 | 0.989 | 0.996 |
| G1T1701\_152\_002\_F7\_G | 1529 | G1T1701\_152\_002\_F8\_G | 1399 | 1464 | 0.988 | 0.996 |
| G1T1701\_149\_003\_B4\_G | 1475 | G1T1701\_149\_003\_A4\_G | 1471 | 1473 | 0.988 | 0.996 |
| G1T1701\_152\_001\_C6\_G | 1523 | G1T1701\_152\_001\_C7\_G | 1302 | 1412.5 | 0.987 | 0.996 |
| G1T1701\_161\_001\_C8\_G | 1462 | G1T1701\_161\_001\_C9\_G | 1404 | 1433 | 0.941 | 0.980 |
| G1T1701\_161\_004\_C9\_G | 1448 | G1T1701\_161\_004\_D9\_G | 1421 | 1434.5 | 0.898 | 0.966 |
| G1T1701\_150\_004\_B10\_G | 1452 | G1T1701\_150\_004\_A10\_G | 1440 | 1446 | 0.878 | 0.959 |
| G1T1701\_150\_004\_C10\_G | 1433 | G1T1701\_150\_004\_D10\_G | 1421 | 1427 | 0.878 | 0.959 |
| G1T1701\_150\_004\_H10\_G | 1395 | G1T1701\_150\_004\_G10\_G | 1351 | 1373 | 0.875 | 0.958 |
| G1T1701\_156\_003\_G3\_G | 1481 | G1T1701\_156\_003\_H5\_G | 926 | 1203.5 | 0.867 | 0.956 |
| G1T1701\_150\_004\_F10\_G | 1419 | G1T1701\_150\_004\_G10\_G | 1351 | 1385 | 0.861 | 0.953 |
| G1T1701\_150\_004\_B10\_G | 1452 | G1T1701\_150\_004\_C10\_G | 1433 | 1442.5 | 0.859 | 0.953 |
| G1T1701\_150\_004\_B10\_G | 1452 | G1T1701\_150\_004\_D10\_G | 1421 | 1436.5 | 0.857 | 0.952 |

Table 4: Sample pairs with extremely high pairwise genetic identity for broodstock pairs and progeny pairs (> 0.95).

APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

**Parentage analysis**

*Cervus*

APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

Final SNP data for parentage analysis consisted of 1,556 SNPs, 317 broodstock and 3,671 progeny.

1. Allele frequency analysis was run with all 3,988 samples.
2. Simulation of parentage analysis was run as parent pair with known sexes across 5,000 offspring, 200 candidate mothers (0.9 proportion sampled), 200 candidate father (0.9 proportion sampled), proportion loci typed 0.9, proportion of loci mistyped 0.01, minimum loci typed 200 (run time was 17 hours).
3. Parentage analysis for parent pairs (known sex) was run by specifying all potential mothers and fathers (8.5 hrs run time). All parent trios (PPCs) with positive LOD scores were reported. The ‘primary’ trio assignment was one made for each progeny with the highest trio LOD score. Other positive trio LOD score assignments for offspring were classed as ‘secondary’.
4. Approximately 200 parental trio-relationship LOD scores were cross checked to ensure no spurious trio assignments were made. The majority of ‘secondary’ assignments were to broodstock with a low number of genotyped SNPs (from 200-500) compared to the ~1,500 in ‘primary’ assignments. In all cases, the ‘primary’ assignment was justified over the ‘secondary’ parental assignments, therefore all assignments with the highest Trio LOD score were trusted and retained for the next step.
5. Even though all trio assignments (PPCs) reported were positive, there were numerous single parent assignments (PCs) that had negative LODs (Figure 3). A scatterplot of the LOD scores of each parent assigned to each progeny indicates some incorrect single parent assignments have been made (Figure 3). As a result, each individual parent with < LOD 10 was removed.

*Figure 3*: A scatterplot of assignment LODs of individual fathers and mothers returned from the initial Cervus parentage assignment run (Parent Pair Known Sexes). Grey – Parents assigned as the Primary Trio for a progeny (reliable assignments), Yellow – Parents assigned as the secondary Trio for a progeny (one parent known to be wrong), Orange – Parents with only one Trio assignment. This plot highlights that each individual parent with < LOD 10 is questionable and should be removed.



**Colony**

1. All Cervus assigned parental relationships (2,624 PPCs and 493 additional PCs after individual parents with assignments < LOD 10 were removed) were imported into Colony alongside a dataset containing a random 1,000 SNPs, 148 potential fathers, 168 potential mothers and 3,671 progeny. However, there was a problem with the ‘maternal sibships’ of mother ‘BR\_067\_GP’ (N=139), these were therefore removed for the Colony run.
2. Settings and parameters consisted of; updating allele frequencies, dioecious species, no inbreeding, diploid, polygamy for both sexes, no clones, scaling of full sib relationships, weak sibship priors, unknown population allele frequency, 1 run, medium run length, run monitored by every 10,000th iterate, run in dos mode, full-likelihood, medium precision and maternal and paternal known sibships were included as inputs with a mismatch of 80 SNPs allowed (up to a 8% mismatch accepted in known parental relationships).

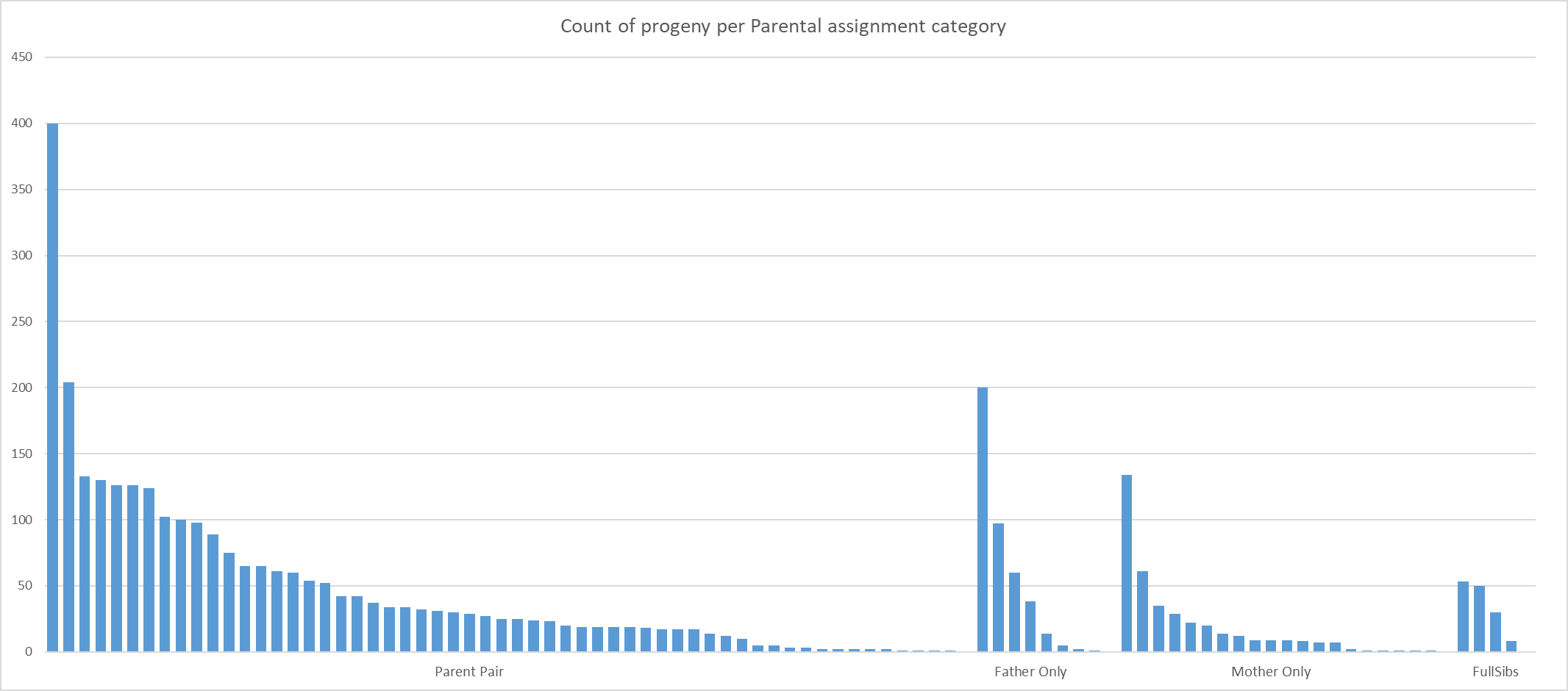
APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

1. *Note*: the input .dat file was created in the Colony GUI, but edited and run in dos mode (mpiexec -localonly -n 38 colony2p.exe) to allow multithreading over 38 cores. Total runtime was four days and 18 hours.
2. Individual paternity and maternity assignments as well as BestCluster assignments with probabilities were compared to the initial Colony results to determine congruency.
3. Colony parentage assignment, fullsib and halfsib cluster determination was much more robust than Cervus and therefore all Colony assignments were retained. However, below there is a list of parental assignments that did not agree between Cervus and Colony that could be removed if you want more *conservative* assignments:
   1. Remove all maternal assignments to BR\_148\_M, BR\_067\_GP and BR\_183\_MP
   2. Remove all paternal assignments to BR\_272\_MP
4. A total of 2,729 progeny had both parents assigned with an additional 417 only with a father assigned and 384 with only a mother assigned (Table 5). In addition, Colony has assigned fullsib clusters among progeny with or without parental assignments.
5. A moderate skew in was observed with parental contributions across the parental assignment categories (Figure 4).
6. All results including working comparisons sheets can be found in the file: “ColonyCervusComparison.xlsx” and “G1T1701\_ParentalAssignments.csv”.

|  |  |  |
| --- | --- | --- |
| Colony Category | #Progeny | #Family Groups |
| Trio Confirmed | 2729 | 57 |
| Father Confirmed | 417 | 8 |
| Mother Confirmed | 384 | 20 |
| FullSib Cluster Confirmed | 141 | 4 |
| TOTAL | 3671 | 89 |

Table 5: Summary of parental assignments across progeny and parent pair family groups.

Figure 4: Proportions of progeny assigned to parents as categorized in parental pairs, father only assignments, mother only assignments and fullsib clusters.



APPENDIX 1 – INITIAL PARENTAGE ANALYSIS FOR REFERENCE ONLY

# APPENDIX 2 – ‘Haplotyping errors’ observed when merging DARTseq & DARTcap data

**From:** Jones, Dave <david.jones3@jcu.edu.au>  
**Sent:** 16 October 2019 13:41  
**To:** Mehar Khatkar <mehar.khatkar@sydney.edu.au>; Herman Raadsma <herman.raadsma@sydney.edu.au>; Nick.Wade@csiro.au <Nick.Wade@csiro.au>; Jerry, Dean <dean.jerry@jcu.edu.au>  
**Cc:** Zenger, Kyall <kyall.zenger@jcu.edu.au>; Le Port, Agnes <agnes.leport@jcu.edu.au>; Jarrod Guppy <jarrod.guppy@my.jcu.edu.au>; Bajema, Casey <casey.bajema1@jcu.edu.au>  
**Subject:** Re: P. monodon - G1T1701 - Parentage analysis

Hi Mehar,

I have added some notes below in blue to address your comments.

Apologies for the lengthy email, but I wanted to address an important pitfall that may be of general interest when handling DARTseq vs DARtcap data.

It boils down to making sure we not only match on AlleleID, but also the REF and SNP sequences (in two row format) when comparing DARTseq and DARTcap data.

I will start a re-run on the parentage ASAP without some troublesome SNPs as described below. Lower iterations should be finished by Monday. Third time is a charm right?

The missing samples were not present in Genotype 63. Are you able to send them through, or direct me to the complete DARTcap dataset for G1T1701? Unless of course they failed genotyping as Agnes suggests (through metadata and QC gel checks).

More details on all points below. Please let me know your thoughts.

Best regards,  
Dave

**From:** Mehar Khatkar <mehar.khatkar@sydney.edu.au>  
**Sent:** 16 October 2019 07:31  
**To:** Jones, Dave <david.jones3@jcu.edu.au>; Herman Raadsma <herman.raadsma@sydney.edu.au>; Nick.Wade@csiro.au <Nick.Wade@csiro.au>; Jerry, Dean <dean.jerry@jcu.edu.au>  
**Cc:** Zenger, Kyall <kyall.zenger@jcu.edu.au>; Le Port, Agnes <agnes.leport@jcu.edu.au>; Jarrod Guppy <jarrod.guppy@my.jcu.edu.au>; Bajema, Casey <casey.bajema1@jcu.edu.au>  
**Subject:** Re: P. monodon - G1T1701 - Parentage analysis

Hi Dave,

Thanks for doing this.

Wonder if you could also upload the combined genotypic datasets. May be versions with 1867 SNPs and also with 932 SNPs.

I will work with Casey to upload these datasets early this coming week (once all other points are addressed).

I guess duplicate BRs are between dartseq and dartcap, the concordance between duplicates seem low with wide spread SD.

Yes, the duplicate broodstock were a result of combining the DARTcap and DARTseq datasets. The reason why the concordance is low is to do with two factors:

1/ general poor sample performance / low call rate (over the common 1867 SNPs). The reason I produced concordance on all samples before filtering for call rate was to allow for the selection of the best performing sample from the duplicate pairs, then exclude samples later if it did not meet the call rate criteria for parentage (when the 932 SNP dataset was created). Some of these duplicate pairs had call rates of ~10-50% which is atrocious. These bring the concordance down.

2/ I found some ‘haploytping’ effects as described below when checking the allele orientations in common SNPs between DARTseq and DARTcap datasets. These change concordance dramatically.

Was it possible to check allele orientation while combining the two datasets i.e. if the alleles for the common SNPs were in the same orientation between dartseq and dartcap?

I didn't specifically check for the orientation of REF vs SNP allele between DARTcap and DARTseq datasets. All datasets were merged on Allele ID which was unique for the REF and SNP rows across the data. Going back, I crosschecked the AlleleID paired with the Allele sequence and respective Broodstock x SNP genotypes in regards to the merger between DARTseq and DARTcap datasets.

I am confident that the same SNPs have been merged between the DARTseq and DARTcap datasets, and that their orientation also matches, however I found something else… bear with me:

If I extract the genotype calls of the common 151 broodstock samples from the DARTcap and DARTseq datasets across the common 1867 SNPs, match by allele ID, then compare Allele sequences, this is what I get:

* 1774 SNPs matched by AlleleIDs have identical Allele sequences (REF and SNP alleles)
* 93 SNPs matched by AlleleIDs have matching REF seqs, however they do not have matching SNP seqs. For these SNPs, there were additional SNP/s identified within the allele sequence originating from the DARTseq dataset compared to the DARTcap dataset.

Interestingly, the average concordance between DARTseq and DARTcap genotypes of the 1774 SNPs is 98.04% (SD 2.44), but it drops to 87.44% (SD 19.06) in the 93 troublesome SNPs (based on the 151 duplicate broodstock).

I am confident that I have merged the same SNP and that the allele orientation is the same, however, I think there is a slight ‘haplotyping’ effect leading to the errors.

Even though the REF sequence is identical, the SNP calling algorithm is searching for a different SNP/alternative allele sequence between the DARTseq and DARTcap datasets due to the presence of the additional flanking SNP variation in DARTseq data.

EXAMPLE:

For SNP 24053774|F|0-14:C>A-14:C>A (SNP in bold, additional DARTseq variation highlighted)

REF allele (DARTseq and DARTcap) – ‘C’ Allele

TGCAGTTCAGGGTG**C**CTTCTTT**AG**TGCCTACCCTATTGGTACTATGCCTAGAACATGTAACTTCCTCAG

DARTseq SNP allele – ‘A’ Allele

TGCAGTTCAGGGTG**A**CTTCTTT**TA**TGCCTACCCTATTGGTACTACGCCTAGAACATGTAACTTCCTCAG

DARTcap SNP allele – ‘A’ Allele

TGCAGTTCAGGGTG**A**CTTCTTT**AG**TGCCTACCCTATTGGTACTACGCCTAGAACATGTAACTTCCTCAG

Often the SNP allele sequence from the DARTseq dataset is not identified (read count = 0) leading to a homozygous call when in fact it is heterozygous in the DARTcap dataset.

Unfortunately, all 93 of these troublesome SNPs were retained in the 932 SNP dataset, so I will start a re-run on parentage ASAP without them.

Cheers

Mehar

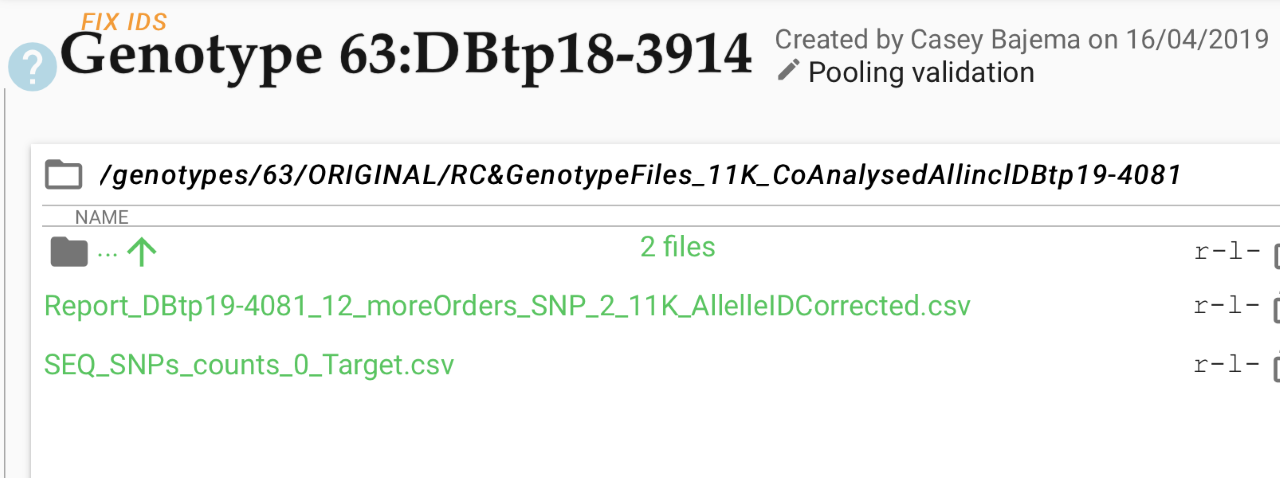
There are following 43 samples which are not in the pedigree file (I can send these in a file if you need):

I traced these samples back and they were present in my revised SampleID list, however, they were not present in Genotype 63 I downloaded from the database and therefore were not included in the analysis.

Is Genotype 63 the right DARTcap genotype for G1T1701 Parentage (below)?

I also checked all previous genotypes provided and these IDs do not appear in them.

If you have the missing genotype data, or the correct complete DARTcap dataset please send it through.



[1] "G1T1701\_149\_001\_D2\_G"    "G1T1701\_149\_002\_B3\_G"

[3] "G1T1701\_149\_002\_E2\_G"    "G1T1701\_149\_002\_F2\_G"

[5] "G1T1701\_149\_002\_G2\_G"    "G1T1701\_149\_002\_G9\_G"

[7] "G1T1701\_149\_002\_H2\_G"    "G1T1701\_149\_002\_H8\_G"

[9] "G1T1701\_149\_002\_H9\_G"    "G1T1701\_149\_004\_C1\_G"

[11] "G1T1701\_149\_005\_E8\_G"    "G1T1701\_149\_006\_G4\_G"

[13] "G1T1701\_149\_009\_D1\_G"    "G1T1701\_149\_009\_E2\_G"

[15] "G1T1701\_149\_009\_E8\_G"    "G1T1701\_149\_009\_F5\_G"

[17] "G1T1701\_149\_009\_G8\_G"    "G1T1701\_149\_009\_H8\_G"

[19] "G1T1701\_150\_004\_F9\_G"    "G1T1701\_150\_008\_G1\_G"

[21] "G1T1701\_150\_009\_G3\_G"    "G1T1701\_150\_010\_G5\_G"

[23] "G1T1701\_155\_001\_H7\_G"    "G1T1701\_155\_002\_B12\_G"

[25] "G1T1701\_155\_002\_C12\_G"   "G1T1701\_155\_002\_D12\_G"

[27] "G1T1701\_155\_002\_H3\_G"    "G1T1701\_155\_007\_D9\_G"

[29] "G1T1701\_155\_008\_C12\_G"   "G1T1701\_155\_008\_D9\_G"

[31] "G1T1701\_155\_009\_G10\_MP"  "G1T1701\_156\_001\_H2\_G"

[33] "G1T1701\_156\_003\_C7\_G"    "G1T1701\_156\_005\_D5\_G"

[35] "G1T1701\_156\_007\_A7\_G"    "G1T1701\_156\_007\_F4\_G"

[37] "G1T1701\_156\_008\_G8\_G"    "G1T1701\_156\_009\_A9\_G"

[39] "G1T1701\_RD2\_094\_PL"      "HUB17\_03\_TNA12\_G01"

[41] "G1T1701\_155\_009\_E10\_GMP" "G1T1701\_155\_009\_F10\_GMP"

[43] "G1T1701\_155\_009\_F12\_GMP"