## SHORT TERM SCIENTIFIC MISSION (STSM) SCIENTIFIC REPORT

**This report is submitted for approval by the STSM applicant to the STSM coordinator**

**Action number:** FA1302 – Large-scale Methane Measurements On Individual Ruminants for Genetic Evaluations
**STSM title:** Imputing missing genotypes in heterogeneous ruminant populations and its impact on genomic predictions for methane emissions

**STSM start and end date:** 22/08/2017 to 12/09/2017
**Grantee name:** Jody Leigh Edmunds

### PURPOSE OF THE STSM:

(Max.200 words)

The purpose of this short-term scientific mission (STSM) was to estimate the effect of imputation on population/breed parameters in livestock, using genomic array data and corresponding phenotypic methane data.  The work was aimed at estimating the effects of statistical imputation on breed and population parameters, drawing a focus on imputing missing genotypes in heterogeneous populations and its impact on the accuracy of genomic predictions. The project addressed the idea that population structure would impact the accuracy of the imputation of genotypes thus, determining assumptions for the imputation accuracy of genomic data. An important aspect of this work was to assess the effect of imputation on the genomic predictions on methane production in cattle. Therefore, if we can impute genomic data for animals that have been genotyped with low-density SNP arrays or those with missing genotypes, we can then use this information to predict their methane production ability without sequencing that animal entirely. This work strengthens research in sustainable livestock breeding and facilitates the mitigation of anthropogenic green house gas emissions.

### DESCRIPTION OF WORK CARRIED OUT DURING THE STSM

(Max.500 words)

The aim of the STSM was to develop a workflow to assess the accuracy of imputation in livestock using a large bovine dataset (Holsteins and Nordic reds) of around 769 animals with approximately 37000 SNP loci. At first, training was an important aspect of the project familiarising the participant with script development and data preparation using specific subsets of the main data. The key part of the work began with the development of a computational workflow to perform the entirety of the work from start to finish. This work flow is highlighted in figure 1 and shows the four key stages of the project; Data Preparation, Injection of missing data, imputation of missing data, accuracy assessments and then the phenotype assessment or presentation of accuracies. Initial data preparation involved the use of PLINK (Purcell *et al.*,

Funded by the Horizon 2020 Framework Programme
of the European Union

2007), a genomics analysis program, to assess datasets for the number of missing genotypes present  (missing data threshold of zero) and the minor allele frequency (MAF).

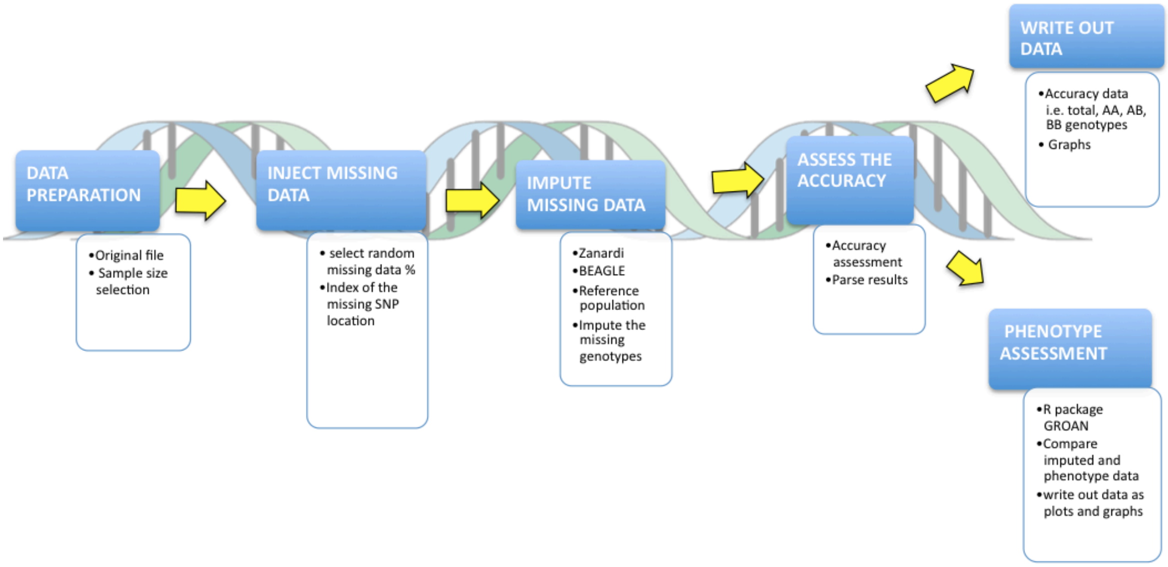
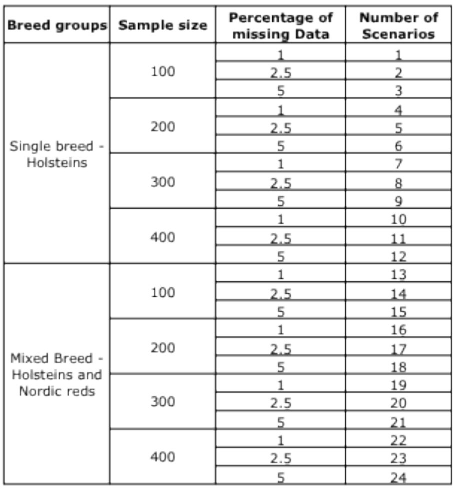


**Figure 1 The work flow developed during the STSM**

The workflow was designed to read and convert data sets to the required format, (i.e. .ped .raw) and samples a specific sample size (100, 200, 300 or 400) from the original data set. Random missing data (i.e. missing SNP values) was then injected into the sampled dataset at certain percentages (i.e. 1%, 2.5% or 5%) using a Boolean matrix. The script allowed for the generation of two data files; one a .ped file contains the genetic information and missing data and the other detailing the exact loci of this missing data. The file containing the missing data was then subject to imputation using the Zanardi (Marras *et al.,* 2016) and Beagle  (Browning and Browning, 2007) programs to obtain a complete dataset. The accuracy of this imputation was then assessed in R for the total accuracy and the accuracy of AA, AB and BB genotypes. Additionally computation time and MAF values were also calculated.

Twenty-four scenarios were selected for assessment of imputation accuracy (Figure 2). These include variation in sample size and proportion of missing data and two distinct populations; a single breed (Holsteins) and a multi breed (Holsteins and Nordic reds).



| Breed groups | Sample size | Percentage of missing Data | Number of Scenarios |
|---|---|---|---|
| Single breed - Holsteins | 100 | 1 | 1 |
| | | 2.5 | 2 |
| | | 5 | 3 |
| | 200 | 1 | 4 |
| | | 2.5 | 5 |
| | | 5 | 6 |
| | 300 | 1 | 7 |
| | | 2.5 | 8 |
| | | 5 | 9 |
| | 400 | 1 | 10 |
| | | 2.5 | 11 |
| | | 5 | 12 |
| Mixed Breed - Holsteins and Nordic reds | 100 | 1 | 13 |
| | | 2.5 | 14 |
| | | 5 | 15 |
| | 200 | 1 | 16 |
| | | 2.5 | 17 |
| | | 5 | 18 |
| | 300 | 1 | 19 |
| | | 2.5 | 20 |
| | | 5 | 21 |
| | 400 | 1 | 22 |
| | | 2.5 | 23 |
| | | 5 | 24 |

**Figure 2 The 24 scenarios implemented in the project.**

The impact of imputation accuracy on the prediction of methane emission was an essential aspect of this work. The R package GROAN (https://cran.r-project.org/web/packages/GROAN/) was used to assess the prediction performance requiring both the imputed genomic data and the phenotypic data in terms of ruminal methane production (g/d).

The work was presented as a seminar for the host institution (National research council, Italy) on Tuesday 12th September and was happily received and discussed by colleagues at the Italian National Research Council.  Additionally, as part of the project a github repository (https://github.com/filippob/heterogeneousImputation/) was created to hold the workflow and all scripts that were developed for the imputation accuracy analysis and the methane prediction analysis.

**DESCRIPTION OF THE MAIN RESULTS OBTAINED**
(Max.500 words)

The accuracy of imputation was first described using total accuracy as detailed below in figure 3. It is clear that a single population of Holstein animals had a higher mean accuracy and less variation around the mean than that of a mixed breed population (P<0.05).

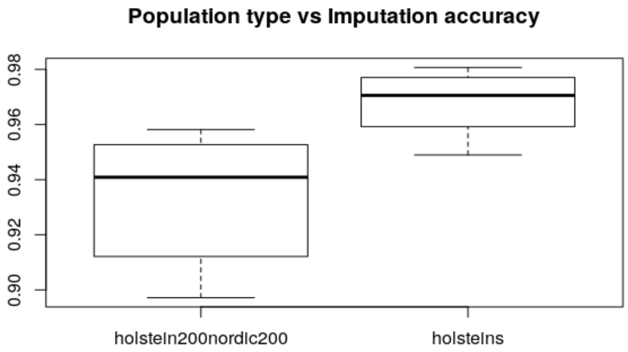


**Figure 3 A box plot detailing the Total accuracy for the two populations**

There was a distinct trend between sample size and total accuracy described as; as the sample size increased, the accuracy also increased and the variation around the mean values decreased (figure 4). The largest sample size of 400 had the highest accuracy for each population of 0.98 and 0.96 for the single and multi breed populations respectively. Figure 4 shows that the two populations (single and multi-breed) remained separated by accuracy with the single breed (Holsteins) exhibiting the higher distribution for all sample sizes.

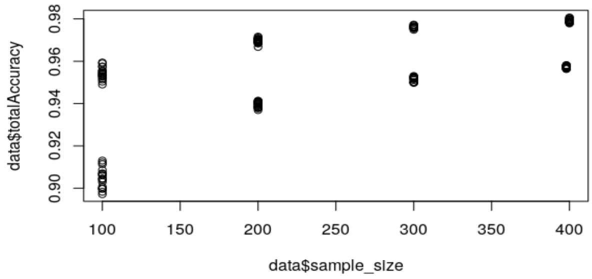


**Figure 4 A scatter plot displaying accuracy versus sample size for each population.**

In addition to the total accuracy, accuracy was assessed based on the genotype frequency; the major homozygous (AA), the heterozygous (AB) and the minor homozygous (BB) (figure 5). Figure 5 shows the effect of sample size against the accuracy for both the multi and single breed populations. It is apparent that accuracies were higher for the single, Holstein breed when compared the multi-breed population. For both populations the major homozygous allele (AA) had the best accuracies exceeding 95% for both, and the minor allele (BB) had the lowest accuracies. A prominent trend is repeated in figure 5 describing; the greater the sample size the better the accuracy (and lower the variation) with the single breed population displaying the better accuracy values.

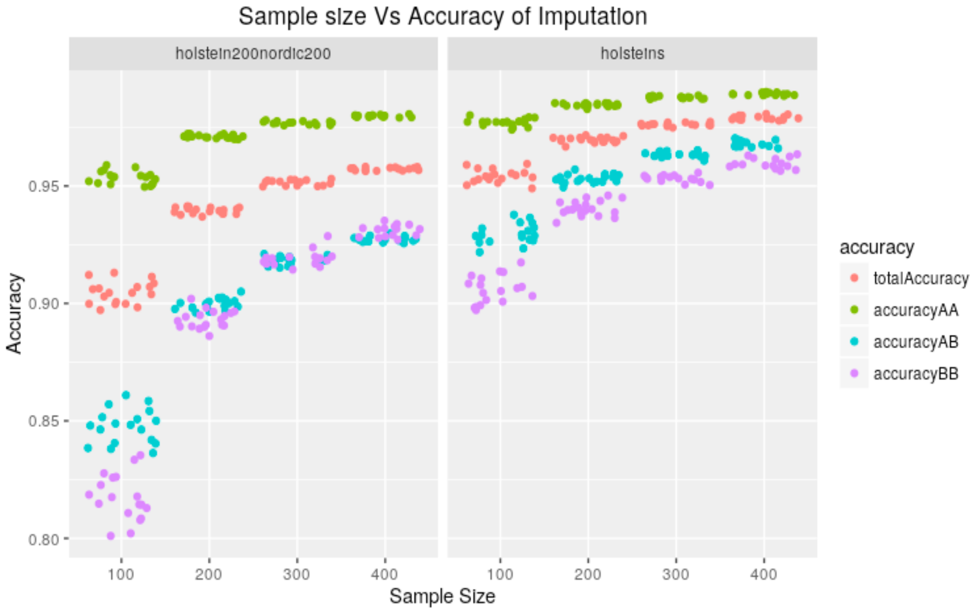


**Figure 5 A ggplot2 graph showing the accuracy against sample size for each population displaying varying types of accuracy in different colours.**

In terms of proportion of missing data, figure 6 shows the varying levels of missing genotypes used versus the accuracy of imputation per population. No significant effects (>0.05%) were found between the proportions of missing genotypes used (1%, 2.5% and 5%) and each accuracy category, including total accuracy.

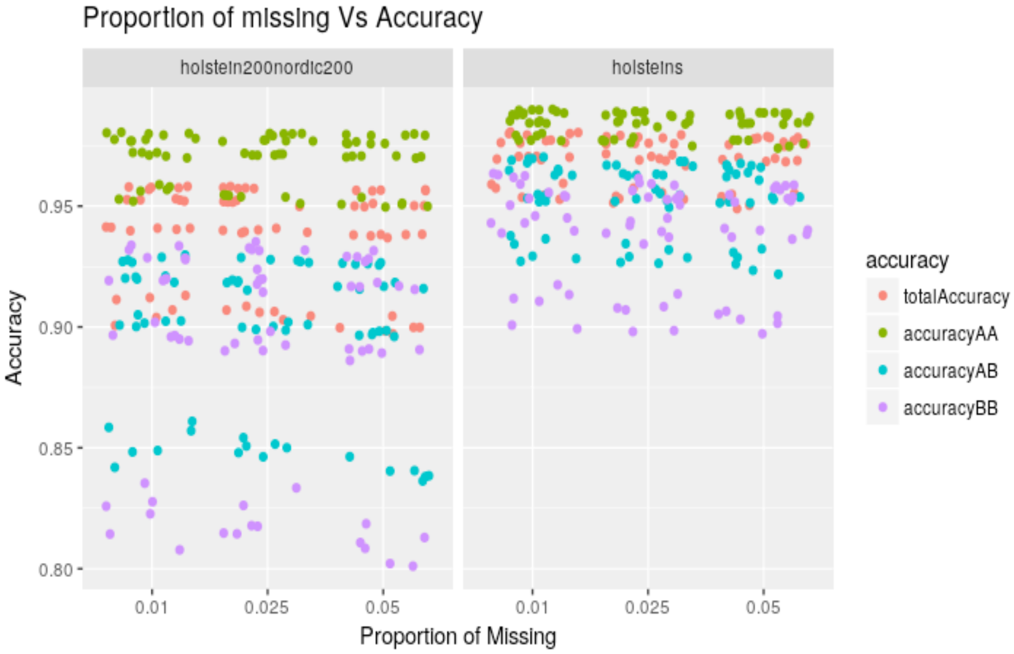


**Figure 6 A ggplot2 graph plotting the accuracy against the proportion of missing data for each population.**

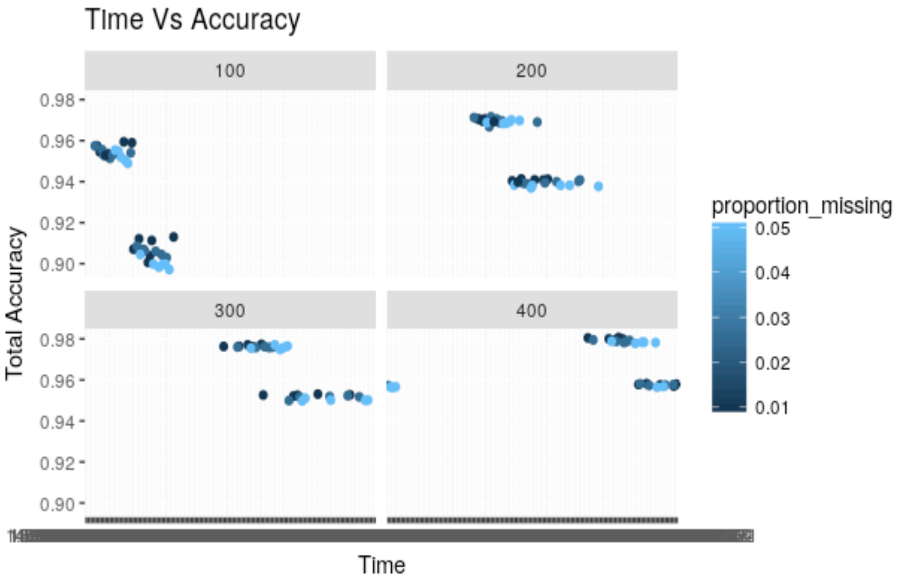


**Figure 7 A ggplot2 graph showing the total accuracy against computation time for each sample size used.**

Figure 7 explains how computation time increased with sample size and in terms of population; the single breed had a quicker computation time than that of the multiple breeds. However, the proportion of missing data (genotypes) did not have a significant effect on time and time was fairly varied for each proportion of 'missingness' analysis.

Phenotype assessment revealed that prediction accuracy methane output was generally low and revealed an ambiguous trend in that accuracy was lower for the single breed population than that of the mixed breed population at both 1% missing and 10% missing genotypes. The variation of results was greater for the higher proportion of missing genotypes, 10%, than the lower percentage (1%).

**FUTURE COLLABORATIONS (if applicable)**

The participant and Cardiff university will continue to work with The Institute of Agricultural Biology and Biotechnology at the National Research Council, Italy and Dr Biscarini to investigate the accuracy of imputation in livestock populations' i.e. divergent breeds and closely related breeds in greater detail. Future work will focus on various different scenarios, areas of varying genetic variation and the use of different SNP chip arrays. In addition to livestock, the collaboration will also look at imputation in other animal species including wild populations.