Emhimad Abdalla
Postdoctoral Fellow
University of Guelph
Centre for Genetic Improvement of Livestock
50 Stone Road East, Room 21
Guelph, ON Canada N1G 2W1
Email: alamroony@gmail.com

Accuracy of imputation:

It is important to assess quality of imputation as it would indicate how accurate the imputed data might be. It is also useful to compare different imputation scenarios and find out the best scenario to be applied. Imputation accuracy could be assessed using a cross-validation technique. Here I am providing some examples with codes for this procedure. In addition, a fully automated evaluation process that could be run for x times and outputs the average of the x cross-validation runs. This including picking up a random set of markers and a different size of reference populations, imputing the data and then calculating the accuracy parameters.

These are the main steps to perform imputation accuracy evaluation:

1. Drop some (%) of your genotype calls.
2. Impute
3. Calculate one or both of these parameters:
    A. The average proportion of genotypes correctly imputed.
    B. The average correlation between true and imputed genotypes.

Example:

Note: In this example I used BEAGLE to impute the data, but of course any other imputation software can be used.

Let see that we need to assess the accuracy of imputing a low (say 20K) density chip to a higher one (say 60K). To do the cross-validation we need to use the 60K chip. Here are the steps:

1. Decide which subset of animals would be a reference and which subset would be the target/validation population. The size of these groups depends on how big the data is. It could be 1 or up to 10% of the total data size. It makes more sense to choose the youngest animals to be the target subset.
2. Using PLINK (or any other software) create a vcf file for the reference (we may also use vcf-tools if we already have vcf files).
3. Using PLINK (or vcf tools) create a vcf file for a target subset (let's name it tarFull). From this file create another file while you mask all SNPs, but those in the low-density chip (20K in this case). Let's name it tarReduced.
4. Impute the reference file (since the ref should not have any missing SNPs and in order to be able to use BEAGLE, the separator between markers must be | not /.
5. Next, impute the tarReduced to the higher density (60K in this case).
6. Compare the imputed tarReduced with the tarFull file in terms of one or both of these parameters:

Emhimad Abdalla
Postdoctoral Fellow
University of Guelph
Centre for Genetic Improvement of Livestock
50 Stone Road East, Room 21
Guelph, ON Canada N1G 2W1
Email: alamroony@gmail.com

- The average proportion of genotypes correctly imputed.
- The average correlation between true and imputed genotypes.

These two parameters can be obtained using any software. What I do is transforming the vcf files into genotypes in 0125 (5= missing , if any) format. For example, 5 SNPs and 3 animals:

	Original		
	Animals		
SNPs	1	0	2
	2	0	1
	1	2	0
	0	5	0
	2	2	0

	Imputed		
	Animals		
SNPs	2	0	1
	2	2	5
	0	2	1
	1	2	0
	2	1	1

Then use the FORTRAN codes icorrect and icorrelation. In addition to the average percentage of SNP matched, icorrect creates the following file, which is similar to the one created by the vcf tools:

```
SNP_ID
Number_SNP_compared
Number_SNP_matched
Number_SNP_did_NOT_match
Percentage_SNP_matched
```

The icorrelation, however, re-order the SNPs from the two files such that they are listed on two columns (so it is easier to calculate the correlation between them). The icorrelation calculates the correlation and gives these outputs:

```
mean of original genotypes =   0.923076928     mean of imputed genotypes =    1.15384614
SD for original genotypes =   0.954073548     SD for imputed genotypes =   0.800640762
Covariance of original genotypes with imputed genotypes =   0.346153855

*********************************************************************************
Correlation =   0.453157961
```

And a file with four columns:

```
SNP ID (i)
```

Emhimad Abdalla
Postdoctoral Fellow
University of Guelph
Centre for Genetic Improvement of Livestock
50 Stone Road East, Room 21
Guelph, ON Canada N1G 2W1
Email: alamroony@gmail.com

Animal ID (j)
The SNP (i) in the original file for that animal (j)
The SNP (i) in the imputed file for that animal (j)

The two files above will be as follows:

1	1	1	2
1	2	0	0
1	3	2	1
2	1	2	2
2	2	0	2
3	1	1	0
3	2	2	2
3	3	0	1
4	1	0	1
4	3	0	0
5	1	2	2
5	2	2	1
5	3	0	1

Missing markers (5s) are ignored in all of the calculations above.