SNP Panels/Imputation

Convener: Dr. George R. Wiggans
Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA
george.wiggans@ars.usda.gov

Secretary: Dr. Stephen P. Miller
Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON N1G 2W1, Canada
miller@uoguelph.ca

Thirteen participants from Australia, Belgium, Canada, France, Ireland, Italy, New Zealand, Norway, Poland, Spain, Sweden, and the United States participated in discussing the services that Interbull can perform or recommendations that Interbull can make to promote harmonization and assist member countries in improving their genomic evaluations in regard to SNP panels and imputation. Prior to convening the group, the participants were asked to consider the following questions:

- Should Interbull provide a test data set for imputation from 3K (low density) to 50K to assist countries in determining the accuracy of their programs, or should Interbull test available programs and provide statistics on their success?

- Should Interbull serve as a clearinghouse on SNP quality or store names of SNP that are names from a panel and the reason for rejection?

- Should Interbull collect improvement in the SNP map between major releases (e.g., UMD3) to assist in improving imputation accuracy?

- What role should Interbull play to facilitate sharing of genotypes? Is further work needed in defining formats? Should lists of genotyped bulls be maintained to facilitate trading genomic information?

- Can Interbull assist with any issues related to using Illumina and Affymetrics HD (high density) SNP chips and full-sequence data?

When the group convened, the following topics were discussed, and those discussions summarized for the plenary session.

Full Sequence Data

Mike Goddard (Australia) introduced this topic. He explained that accuracy of imputation will rely on the number of animals with more complete genotyping. For example, if imputing from 50K to HD, the number of animals with HD will be the limiting factor.

Ultimately, imputing to full sequence would be the most powerful. However, because sequencing is still expensive, relatively little is done. Therefore, combining sequence data across research projects is desirable to provide a larger data set with sequence data that can be used for imputation purposes. Although sequencing is dropping in price, individual data sets of perhaps 100 sequenced individuals will still be too small for effective imputation.

A repository of sequence data across breeds and countries in a single database would provide a useful resource for validation purposes. When comparing sequence data, small chromosomal segments can be compared. Those small segments are conserved across breeds; for example, the same segment would exist in both Holsteins and Jerseys without a recombination. For that reason, sequencing of all cattle is useful. Bulls and cows from all breeds are useful contributions to the database to enable imputation.

If one of the limitations to contributing sequence data to a common database is animal identity, a workable solution could be to submit the sequence without the animal identity. Those anonymous samples would still have value.

Different animals likely will be sequenced to different depths of coverage, depending on the individual projects from which they originate. The human 1000 Genomes Project is running with 4X coverage. If 2 different countries were to sequence the same animal to 4X coverage and both submitted the sequence to the database, the effective coverage for that animal in the database would then be 8X.

Goddard shared the method used in Melbourne to determine candidates for sequencing based on their independent genetic contributions to the current population with phenotypes. The pedigree relationship matrix determines the additive relationship between all animals. A multiple regression analysis then determines the independent additive relationship of founders to the current population.

General experiences among the group with regards to imputation were shared. Imputing HD to sequence is what is desired. Imputation from 50K to 800K must be done within breed, which has been done and works well (Goddard). Imputing 50K to sequence has not been done. Imputing from 3K to 50K has worked for Holsteins, but Goddard did not have any experience with other breeds. The experience in France has been that Montbeliarde and Normande genotypes could not be effectively imputed from the current 3K to 50K. Those breeds along with Nordic Red, Fleckvieh, and others as well as Holstein are of interest to the French.

The requirements to process sequence data effectively cannot be underestimated. As the cost of sequencing declines, the processing of the data is becoming a limiting factor. Sequence data contain errors and even with a low error frequency, the total number of errors per animal can be quiet large because of the massive volume of data in a given sequence. The errors need to be corrected utilizing sequence information from other animals.

A database is required to correct and phase all sequence data. As a service, a research group could submit a full sequence for phasing and correction, and the database would retain submitted sequence data to increase the power of the database to correct and phase submissions. With HD genotypes in addition to sequence data, you could phase the genotypes, which would provide an incentive for submission of HD genotypes. This service likely would be too difficult for individual groups to re-create; therefore, participation in the database would be an obvious choice. Ideally the database would be open and public, and contributing groups could have unrestricted access to the sequence data. A more conservative approach would be for clients to receive only corrected and phased sequence data for their animals. However, sequence information from all submitted animals would remain in the database. Because such a service would require considerable resources, a fee would be required to cover computing and software costs. Possible venues for delivery include:

- Victoria Department of Primary Industries – a pipeline was developed out of necessity and could be transformed into a service.
- Interbull.
- Companies such as Pfizer Animal Genetics.
- Companies such as Illumina and Affymetrix could offer this service to complement their HD chip sales and sequence offerings. Customers could submit HD genotypes to receive back-imputed full sequence. A delivery model would need to be established that did not preclude genotyping with any given company.

Recommendation: A mechanism to share sequence data should be developed. Interbull could define the objective and monitor the service, which would be provided by a group experienced with full sequence data.

Validation of Imputation

Participant experiences with imputation were discussed. In general, the success with imputation has been quite high.

- Semex Alliance/Bovitech – Mehdi Sargolzaei (Centre for Genetic Improvement of Livestock) has developed an imputation program (Fimpute) that uses both population and family information and is working well for Holsteins. Incorporating family-based information resulted in accuracies of 98.5% for imputation from 3K to 50K. That same level of accuracy was not possible with population-based imputation only.

- USDA – FindHap has recently been used for nearly a year. It does a better job than Fimpute when pedigree is missing. It has been recently revised to reprocess the genotypes with decreasing segment lengths to improve both accuracy and call rate. A combination of FImpute followed by Findhap was found to give the most accurate results.

- Australia – Different approaches have been attempted (FastPhase and Beagle)., and its own custom software Beagle has done a very good job. Beagle does not use pedigree information and requires considerably more processing time.

- New Zealand – Beagle is used. Most evaluations only process new genotypes because of the processing time required.

- Ireland – Donagh Berry has attained 98% accuracy when imputing from 3K to 50K with Beagle. He has a method to get Beagle to use pedigree information to improve accuracy by 2 percentage points.

Genotype probabilities should be provided for imputed genotypes.

Recommendation: A benchmarking trial for the usefulness of the different software methods should be considered. Given the experiences shared in the workshop, the best approach may depend on the specific scenario in which imputation is employed. The experiment should consider different data sets to see how the programs perform in different scenarios (parents known, unknown, etc.; different breeds). Interbull could make a number of test data sets available for testing imputation software. Such data sets could be based on actual data with an error rate on genotypes introduced to simulate the error rate that will be experienced in practice.

SNP Quality

Every analysis of SNP data requires an editing step to remove SNPs with Hardy-Weinberg (HW) equilibrium problems, low minor allele frequencies, excessive parent-progeny conflicts, etc. No standards exist for those edits.

Discussion regarding HW edits indicated that only those far from equilibrium should be removed. The HW deviation could be a calling problem or the result of families or selection. Sargolzaei indicated that SNPs around the SNP that is out of HW equilibrium can be used to determine if a calling error has occurred. If the original SNP problem is the result of HW disequilibrium, the surrounding SNPs also will be out of HW equilibrium.

Recommendation: Interbull should develop guidelines for editing SNP data. Interbull could host a database to compile problem SNPs from multiple studies. Knowledge of SNP performance across studies would enable more effective editing of SNPs prior to analysis.

SNP Maps

A better SNP map is required. Imputation requires an accurate SNP location. A mechanism is required to share map improvements and reduce redundancy in multiple groups repairing the map. The UMD3 and BT4 are 2 competing maps. The conclusion was that UMD3 is likely the best.

Recommendation: The bovine community could take some responsibility for the map. They could put all the developing sequence data together, and develop a resource for the community as a whole. The SNP map could become integrated into the SNP quality project (see above).

Genotype Sharing

Sharing genotypes between countries (or country groups) is increasing. One problem is having many small genotype files with multiple formats.

Another application for genotype sharing is in the implementation of the Affymetrix HD BOS array. The availability of this product provides competition for Illumina, and the BOS array could be more powerful. Imputation problems will ensue with 2 chips. To develop imputation ability, HD genotypes from both chips could be shared. The structure of the dairy population results in widespread sharing of haplotypes, so genotyping the same animals with both chips probably is not necessary. Sharing genotypes could assist in making the use of both Affymetrix and Illumina HD chips possible for evaluation centers.

Recommendation: Interbull could play a role in making genotypes available to incorporate into national evaluations. If a bull is to be marketed internationally, Interbull could make that genotype available for to the evaluation centers in the inporting countries. Interbull could encourage downloading data for all animals from the genotype database.

Conclusions

Interbull can assist national evaluation centers by hosting or sponsoring the sharing of genotypes and full-sequence data, establishing editing standards for SNP quality along with a database of problem SNPs, and providing data sets to test validation methods.