BY THE NUMBERS

by Duc Lu, Angus Genetics Inc.

From SNP to Haplotypes

Genomic selection refers to selecting animals with the assistance of a large number of SNPs. This has been practiced in the Angus breed since 2010.

Single nucleotide polymorphisms (SNPs) occur normally throughout an individual's DNA, and recent whole genome sequencing research has identified about 20 million SNPs in Angus cattle. These variations in an individual are inherited from their parents, with the exception of new mutations that might arise spontaneously in the progeny. Thus, SNPs can be used to track relationships between individuals, which is an improvement over using pedigree alone.

The roughly 50,000 to 75,000 SNPs found on the genotyping arrays used in the Angus genetic evaluation are subsets of the 20 million SNPs in Angus. Some of these are found in the DNA between genes, while others occur within a gene or in a regulatory region near a gene. Therefore, these are expected to play a role in the development or production performance of an individual by affecting the gene's function.

Many SNPs have no direct effect on the traits from which Angus Genetics Inc. (AGI®) generates expected progeny differences (EPDs) but are useful to track relationships between animals and may be linked to genetic variants that do affect traits directly. Some of the SNPs that are included on the genotyping arrays are the genetic variants that have been proven to be important for certain traits such as growth and meat quality. Ideally, one would want to identify these causal variants for all the traits of interest and focus on selecting individuals that carry favorable alleles.

Unfortunately, the number of causal variants found in cattle over the past two decades is very limited despite numerous association studies conducted by researchers around the world. One of the many reasons for that could be insufficient numbers of SNPs genotyped on the cattle in those studies. Increasing the number of SNPs that ultimately lead to identifying actual causal variants will come with a higher genotyping and sequencing cost. A viable alternative to genotyping large numbers of animals for the very high-density SNP (millions of SNPs) genotypes required is to use the moderate-density SNP information (50,000 SNPs) to track haplotypes in the population.

Haplotypes

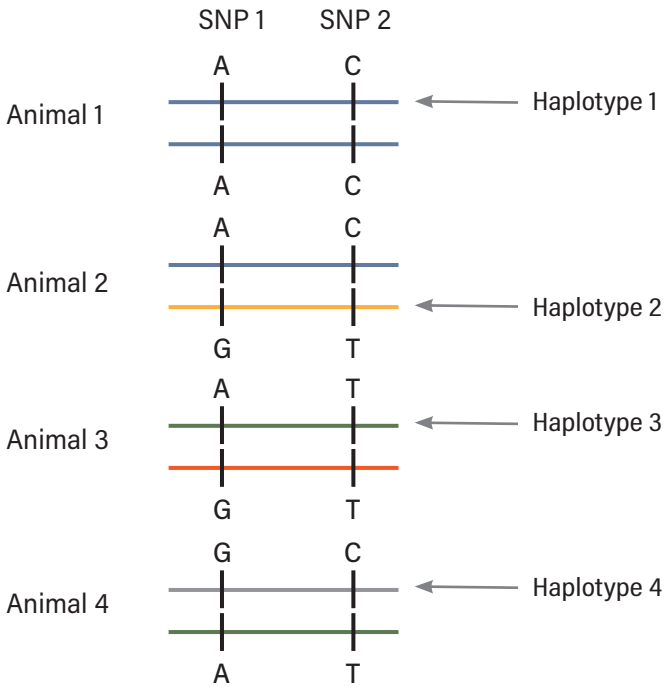
According to the National Human Genome Research Institute, a haplotype in its most general sense refers to a set of DNA variations (SNPs) along a chromosome that tend to be inherited together because they are close to each other. Considering the shortest chromosome segment consisting of two consecutive SNPs: SNP 1 with two alleles Adenine (A) and Guanine (G), and SNP 2 with alleles Cytosine (C) and Thymine (T), the four possible haplotypes are A – C (haplotype 1), G – T (haplotype 2), A – T (haplotype 3) and G – C (haplotype 4). Each animal carries two copies of a chromosome, and therefore two haplotypes are present for each individual when considering a chromosomal segment. If a bull carries two copies of haplotype 1, he is homozygous at that chromosome region. Another bull carrying one copy of haplotype 1 and one copy of haplotype 2 is heterozygous for that segment (see Figure 1 on page 46). Frequency of a haplotype can be used to analyze its association with production traits.

One advantage of using haplotypes over individual SNPs relies on the variation in haplotype frequency when a mutation occurs. According to Curtis et al. (2001), when a mutation occurs in the DNA between two SNPs, it tends to create linkage disequilibrium with nearby SNPs, causing major changes in the haplotype frequencies while

Continued on page 46

 continued from page 44

Figure 1: ***Examples of haplotypes.***



frequencies of alleles at surrounding SNPs are most likely unchanged. As a result, an association analysis using haplotypes might be able to pinpoint the location of the mutation better than a single SNP analysis.

Another benefit of haplotypes is their usefulness in detecting chromosome regions potentially carrying recessive lethal alleles. A lethal allele results in the death of the animal before it reaches reproduction age. Most lethal alleles cause death at an early development stage (i.e. embryonic stage) or right after birth. A single non-interacting lethal allele can be dominant or recessive. Dominant lethal alleles cannot be observed in the population because they are removed immediately as the dominant mutation occurs. It is expected that for recessive lethal alleles there should be no homozygous genotypes in the entire reproductive population, or even in the entire live population if the allele causes embryonic death, but heterozygous genotypes may exist. Therefore, one can search for homozygous deficiency at the SNP or haplotype level.

The best and most accurate strategy to identify recessive lethal alleles is to sequence a good number of dead embryos, which is not practical, or calves, especially those that are closely related, and then search for homozygous regions appearing in dead animals but never in parents and ancestors or even the population. This approach requires both phenotypic and genomic data, and is difficult to achieve, especially with embryos. Alternatively, where phenotypes are not available, a search for genomic regions where no cattle are found to carry two copies of a certain haplotype is feasible. Such haplotypes are then used in association analyses with fertility or health traits to verify their mode of action.

The use of haplotypes in animal breeding has potentially huge benefits. However it contains technical difficulties, including:

1) SNP array density is important in fine-tuning chromosome segments of focus. Lower density SNP chips might result in larger segments of chromosomes being tracked, which undermines the power of association analysis.

2) Haplotype analyses require that the individual SNP alleles that have been inherited from the dam vs. the sire to be known, a processing known as phasing, which is a time-consuming process.

3) Given a chromosome segment, the number of haplotypes available is normally larger than the number of SNPs in that region; and therefore association analyses will take longer to complete.

Regardless of these challenges the AGI research team, along with other partners, have been analyzing haplotypes within the three-quarters of a million animals with SNP genotypes. We have been working together to derive predictions of traits important to Angus breeders such as fertility and survival, which are anticipated to become part of genomic testing in the future. AJ

dlu@angus.org

Reference: Curtis, D., North, B.V., and Sham, P.C. 2001. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. Ann. Hum. Genet. 65: 95-107.