PREDICTION OF OFFSPRING PROBABILITIES FOR OBSERVED HAPLOTYPE VARIANTS USING PARSIMONY- AND PROBABILITY-BASED METHODS

Genevieve Mae B. Aquino[1] and May Anne E. Mata[2]

ABSTRACT

Offspring probabilities for a cattle breeding population were calculated from historical genotypic data on closely-linked single nucleotide polymorphisms (SNPs) using a "haplotype-centric" approach. The number of haplotypes and their corresponding frequencies were estimated using manual parsimony methods and the probability-based methods of HAPLOVIEW (ver. 3.32) and PHASE (ver. 2.1). All methods identified the same set of haplotypes in the population. The Bayes theorem was applied on calculated haplotype frequencies to determine probable haplotypes and their corresponding frequencies for cases of incomplete genotype information (i.e. two out of six loci genotyped), with the assumption of Hardy-Weinberg equilibrium and the absence of recombination. The most probable haplotype frequencies for each incomplete genotype allowed the prediction of offspring probabilities for all possible crosses between individuals. Results show that the minimal set of haplotypes in a population can be determined by different methods. Moreover, the true haplotype of an individual can be predicted even when only a fraction of the SNPs was genotyped by applying Bayesian statistics on the known haplotype frequencies in the population.

Key words: Bayesian statistics, haplotype analysis, offspring probability, population genetics, SNP

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are used as molecular markers in high-density arrays because of their association with traits of economic interest in livestock (Schmid and Bennewitz 2017). Genetic marker panels in "SNP chips" are available for cattle (Dash *et al*., 2018), swine (Bertolini *et al*., 2018), chickens (Huang *et al*., 2018), water buffalo (Iamartino *et al*., 2017), and goats (Qiao *et al*., 2017). Furthermore, genotyping-by-sequencing allows for the identification of SNP genotypes in animal species with no commercially-available SNP chips (Zhu *et al*., 2016). Candidate gene alleles are characterized by multiple SNPs associated with varying biological effects. A set of SNP genotypes can be analyzed as an individual haplotype on a chromosome (Niu, 2004). The "haplotype-centric"

[1]Philippine Genome Center - Program for Agriculture, Livestock, Fisheries and Forestry, Office of the Vice Chancellor for Research and Extension, University of the Philippines – Los Baños, Laguna, Philippines, [2]Department of Mathematics, Physics, and Computer Science, University of the Philippines – Mindanao, Davao City, Philippines (email: gbaquino@up.edu.ph).

approach is limited by the non-independent inheritance of markers, the problematic phase determination for large loci numbers, and the minimum number of initially identified haplotypes.

Haplotypes from unphased genotype data can be inferred by various algorithms that are classified based on the underlying statistical method (Schmid and Bennewitz, 2017). Parsimony algorithms are deterministic rule-based methods that can quickly assign the least number of haplotypes from observed genotypes. Pairwise haplotypes have been determined in cattle using parsimony (Banos and Coffey, 2010). Expectation-maximum (EM) and Bayesian methods are stochastic statistical approaches, which are based on likelihood or conditional probability; these computationally exhaustive methods are suitable for complex pedigrees and have been applied in cattle populations by Zhang *et al*. (2016) and Krag *et al*. (2013), respectively.

This study aimed to infer the SNP haplotypes for a breeding population by applying parsimony, EM, and Bayesian algorithms. Conditional probabilities of haplotypes were determined from calculated haplotype frequencies for a set of observed SNP genotypes. The resulting probabilities were used to calculate offspring probabilities. The study supports the value of including SNP genotypes to predict linked offspring traits.

MATERIALS AND METHODS

The genotype data of six closely-linked SNPs in the leptin gene were obtained for 535 unrelated individuals in a cattle (*Bos taurus*) breeding population at the Roslin Institute, UK were used in this study (Table 1). The archival data was assumed to be correct and free of errors; the individuals were genotyped as described by Wooliams *et al*. (2006). Allele and genotype frequencies were calculated for the dataset, with each locus checked for Hardy-Weinberg equilibrium (HWE). The allele frequencies p and q are expected to remain constant over each generation for a biallelic locus in HWE, such that expected genotype frequencies (E) can be predicted using the equation $p^2 + 2pq + q^2 = 1$. Observed genotype frequencies for each SNP (O) were tested for HWE using the χ^2 goodness-of-fit test (Weir, 1996), with one degree of freedom at a 95% confidence interval, as:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

The *P*-values for χ^2 were subsequently computed in Microsoft Excel.

The minimum number of haplotypes and their corresponding frequencies were determined by applying the parsimony algorithm of Clark (1990). The initial set of "resolved" haplotypes was determined from homozygote individuals without missing data. The single heterozygotes were used determine unresolved genotypes that are composites of a known haplotype and a complementary haplotype. Complementary haplotypes that segregate in the population were added to the list of "resolved" haplotypes. This process was sequentially performed for all genotypes in the data set. Haplotypes were inferred from data with missing alleles to account for genotypes in the data set that could not be explained by the "resolved" haplotypes alone. The minimum number of haplotypes required to resolve the observed genotypes, along with their corresponding haplotype frequencies were computed.

Table 1. Genotype frequencies for 6 SNPs and χ2 test for HWE for 535 individuals.

Locus	Alleles (1)/(2)	Minor Allele Freq (MAF)	Minor Allele	Genotype Frequencies*			Test for HWE	
				(1) (1)	(1) (2)	(2) (2)	χ^2	*P*-value
SNP1	T/C	0.4701	C	0.2617	0.4991	0.2206	0.2786	0.8700
SNP2	A/G	0.3505	G	0.4187	0.4467	0.1271	0.2187	0.8964
SNP3	G/A	0.3430	A	0.4000	0.4393	0.1234	0.2428	0.8857
SNP4	A/G	0.0935	G	0.7832	0.1607	0.0131	0.0834	0.9591
SNP5	C/G	0.0252	G	0.9458	0.0505	0.0000	0.0282	0.9860
SNP6	A/G	0.3486	G	0.4262	0.4505	0.1234	0.2198	0.8959

*NOTE: (1)(1) for homozygous for major allele, (1)(2) for heterozygous, and (2)(2) for homozygous for minor allele.

Haplotyping was performed on the same dataset using HAPLOVIEW (ver. 3.32) (Barrett *et al.*, 2005). All the individuals in the analysis were included by setting the minimum number of allowed missing genotypes to >50%. Preliminary marker checks were performed, including an exact test for HWE (Wigginton *et al.*, 2005). An accelerated EM algorithm based on the partition/ligation method by Qin *et al.* (2002) estimated gamete frequencies of phased haplotypes based on the maximum likelihood of unphased genotype data. Haplotype frequencies greater than 0.01% were also computed. A linkage disequilibrium (LD) plot was constructed from all pairwise computations of the D' statistic (Weir, 1996).

A model-based Bayesian method was used in PHASE (ver. 2.1) to compute the distribution of unobserved haplotypes from the observed genotype data (Stephens and Scheet, 2005). Analysis was performed for 200,000 iterations with a burn-in period of 100,000 and a thinning interval of 100 between iterations.

The most likely haplotypes for genotypes with incomplete marker information was determined using Bayesian statistics, with the haplotype frequencies computed by HAPLOVIEW used as prior information. Given the current set of haplotypes (n = 9), the probability of observing haplotype *j* ($1 \leq j \leq$ n) in an individual *g* (g_i, …, g_n) with the genotype *z* was determined. The known frequency of haplotype *j* was taken as the prior probability Pr (g_j), and the expected frequency of the genotype *z* in an individual g_j with haplotype *j* was represented as Pr ($z \mid g_j$). During normalization, the resulting posterior probability was divided by the total expected frequency of observing a genotype *z* in the population Pr (*z*), computed as the sum of the joint probabilities of the observing genotype *z* in any individual g_i. Using Bayes' theorem (Weir, 1996), the posterior probability that an individual *g* with the genotype *z* possess haplotype *j* is:

$$\Pr(g_i|z) = \frac{\Pr(z|g_i)\,\Pr(g_i)}{\Pr(z)} = \frac{\Pr(z|g_i)}{\sum_{i=1}^{n}\Pr(z|g_i)\,\Pr(g_i)}$$

This equation was used to compute for the posterior probability that a haplotype could account for a particular observation when the first two markers, SNP1 and SNP2, have been genotyped. The observed genotypes are listed in Tables 2 to 4. For each observation, the total number of possible genotypes and the haplotype combination with the highest probability of being observed with a particular genotype were identified. The probability of randomly observing a haplotype in an individual with an incomplete genotype was computed as the joint probability of all the possible haplotype combinations which included the said haplotype. The most-probable haplotypes for each incomplete genotype was then used to compute the corresponding offspring probabilities for all possible crosses in the population.

Given the condition that only the first two SNPs were genotyped, the array of offspring probabilities for all 45 possible crosses was computed by assuming HWE and the absence of recombination. From the inferred frequencies of the most likely haplotypes (H_1, …, H_9) for each observed genotype, the binomial expansion:

$$(H_1 + H_2 + H_3 + H_4 + H_5 + H_6 + H_7 + H_8 + H_9)^2 = 1$$

was used to calculate offspring probabilities, regardless of the sex of the parents. non-missing genotypes for all SNPs.

RESULTS AND DISCUSSION

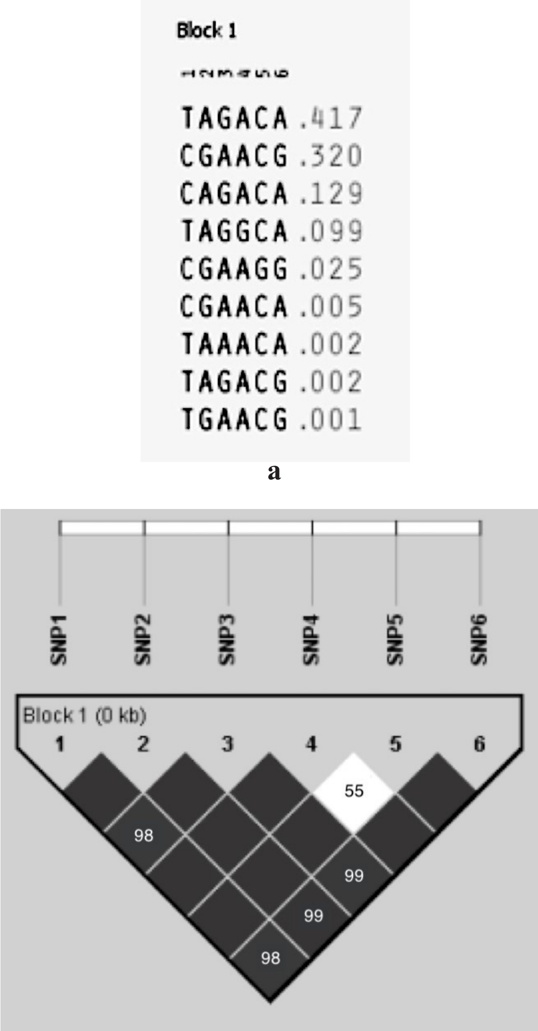
Genotype frequencies observed in the data set are described in Table 1. A total of 40 unique genotypes were observed in the data set, with more than 90% non-missing genotypes for all SNPs. All markers were in HWE according to the χ^2 goodness-of-fit test and the exact test for HWE in HAPLOVIEW.

Given the six loci, eight haplotypes were manually identified by parsimony. TAGACA, TAGGCA, CAGACA, and CGAACG were resolved directly from homozygotes whereas CGAACA and CGAAGG were resolved unambiguously from the single heterozygotes. TAAACA and TAGACG were inferred from individuals with missing genotypes. The ninth haplotype T?AACG could not be identified with certainty by parsimony because of the missing information on SNP2.

The same haplotype set was identified by HAPLOVIEW, with the ninth haplotype resolved as TGAACG (Figure 1a). Figure 1b shows the LD plot representing the degree of LD between any two markers. All pairwise comparisons had D’ >98%, except for SNP4 versus SNP5 (D’=55%). Pairwise comparisons with SNP5 had relatively lower LOD values than those with the other loci. PHASE identified the same set of haplotypes from the data but gave the standard error of the haplotype frequencies (Table 2).

The “observed” nine possible incomplete genotypes in the population represented 45 possible complete genotypes. The non-zero frequencies of these complete genotypes were predicted for the given set of observed genotypes (Table 3). Incomplete genotypes have haplotype combinations that can be unambiguously predicted by Bayesian methods. The most likely haplotype for the remaining incomplete genotypes had relatively high probabilities (0.6431–0.9143) of being the “true” haplotype. The probability that a particular haplotype is sampled from individuals with a known genotype was 0.4010–1.000. Therefore, a minimum of two correctly-typed loci can be used to infer haplotypes from incomplete genotypes in a population under HWE.

Non-zero frequencies of the complete genotypes were predicted for the given set of observed genotypes (Table 4). The most likely offspring were notably combinations of the most common haplotypes in the original data set.



Block 1

TAGACA	.417
CGAACG	.320
CAGACA	.129
TAGGCA	.099
CGAAGG	.025
CGAACA	.005
TAAACA	.002
TAGACG	.002
TGAACG	.001

a

b

Figure 1. Haplotype block (a) with the corresponding frequencies and LD plot (b) for the six loci as computed by HAPLOVIEW (Barrett *et al.*, 2005). Numbers in boxes in (b) indicate D' values less than 100%.

A single set of haplotypes was predicted for the dataset by rule-based and likelihood-based algorithms. Parsimony provided a rapid method to identify haplotypes but did not consistently give the minimum number of possible haplotypes in the population. The efficiency of the parsimony approach is limited by the presence of incomplete genotypes and the available homozygotes in the population used to create the "resolved" haplotype set. Parsimony also does not consider the existing genotype frequencies when identifying haplotypes (Niu, 2004). Algorithms based on likelihood probabilities are recommended despite the requirement of more computing power. The performance of EM methods in simulation studies is not strongly affected by the departures from HWE. However, EM algorithms determine locally optimal maximum likelihood estimates and may not identify unique haplotypes with very low population frequencies. Bayesian methods share the strengths of EM methods but their robustness can be determined by computing for standard errors (Stephens and Scheet, 2005).

The haplotype probabilities inferred from data with missing genotypes by Bayes theorem depend on prior information, such as haplotype frequencies in the population (Niu, 2004). The relatively high probability of inferred haplotypes is due to the assumption of

Table 2. Haplotype counts and their corresponding frequencies as inferred by PHASE (Stephens and Scheet, 2005).

Haplotype	Observed Frequency in Population	Predicted Frequency in Offspring, E(freq)	Standard Error
TAAACA	0.001869	0.001853	0.000420
TAGACA	0.425234	0.416989	0.002556
TAGACG	0.001869	0.001812	0.000257
TAGGCA	0.093458	0.098920	0.002123
TGAACG	0.000935	0.001008	0.000521
CAGACA	0.126168	0.128735	0.001542
CGAACA	0.004672	0.004686	0.000375
CGAACG	0.320561	0.320409	0.000534
CGAAGG	0.025234	0.025186	0.000256

Table 3. Most-probable haplotypes and genotypes of individuals with missing genotype information (only 2 of 6 SNPs genotyped).

Observed Genotype	SNP1					
	TT		TC		CC	
SNP 2	Genotype	Haplotype	Genotype	Haplotype	Genotype	Haplotype
AA	(10)	(4)	(4)	(5)	(1)	(1)
	TAGACA / TAGACA	TAGACA	CAGACA / TAGACA	CAGACA	CAGACA / CAGACA	CAGACA
	0.6431	0.8019	0.8019	0.5	1	1
AG	(4)	(5)	(13)	(9)	(3)	(4)
	TAGACA / TGAACG	TAGACA	CGAACG / TAGACA	CGAACG	CAGACA / CGAACG	CAGACA
	0.8019	0.4010	0.7327	0.4568	0.9143	0.5
GG	(1)	(1)	(3)	(4)	(6)	(3)
	TGAACG / TGAACG	TGAACG	CGAACG / TGAACG	TGAACG	CGAACG / CGAACG	CGAACG
	1	1	0.9143	0.5	0.8359	0.9143

NOTE: Numbers in parenthesis show total number of haplotypes/genotypes with non-zero probability of occurring the population.

Table 4. Most-likely offspring from all possible crosses of individuals with missing genotype information (only 4 of 6 loci genotyped).

Observed Genotype	**SNP 1**								
SNP 2	**TTAA**	**TTAG**	**TTGG**	**TCAA**	**TCAG**	**TCGG**	**CCAA**	**CCAG**	**CCGG**
TTAA	(10)	(14)	(4)	(14)	(30)	(16)	(4)	(16)	(12)
	TAGACA / TAGACA	TAGACA / TGAACG	TAGACA / TGAACG	CAGACA / TAGACA	CGAACG / TAGACA	TAGACA / TGAACG	CAGACA / TAGACA	CAGACA / TAGACA	CGAACG / TAGACA
	0.6431	0.4010	0.8019	0.4010	0.3663	0.4010	0.8019	0.4010	0.7332
TTAG		(14)	(5)	(19)	(35)	(20)	(5)	(20)	(15)
		TAGACA / TGAACG	TGAACG / TGAACG	CAGACA / TGAACG	CGAACG / TGAACG	TGAACG / TGAACG	CAGACA / TGAACG	CAGACA / TGAACG	CGAACG / TGAACG
		0.4010	0.5000	0.2500	0.2284	0.2500	0.5000	0.2500	0.4571
TTGG			(1)	(5)	(9)	(4)	(1)	(4)	(3)
			TGAACG / TGAACG	CAGACA / TGAACG	CGAACG / TGAACG	TGAACG / TGAACG	CAGACA / TGAACG	CAGACA / TGAACG	CGAACG / TGAACG
			1.000	0.5000	0.4568	0.5000	1.000	0.5000	0.9143
TCAA				(15)	(35)	(20)	(5)	(20)	(15)
				CAGACA / TAGACA	CAGACA / CGAACG	CAGACA / TGAACG	CAGACA / CAGACA	CAGACA / CAGACA	CAGACA / CGAACG
				0.4010	0.2284	0.2500	0.5000	0.2500	0.4571
TCAG					(45)	(30)	(9)	(30)	(24)
					CGAACG / TAGACA	CGAACG / TGAACG	CAGACA / CGAACG	CAGACA / CGAACG	CGAACG / CGAACG
					0.3661	0.2286	0.4568	0.2286	0.4177

Table 4. Continuation...

Observed Genotype	SNP 1								
SNP 2	**TTAA**	**TTAG**	**TTGG**	**TCAA**	**TCAG**	**TCGG**	**CCAA**	**CCAG**	**CCGG**
TCGG						(10)	(4)	(13)	(9)
						CGAACG / TGAACG	CAGACA / TGAACG	CAGACA / TGAACG	CGAACG / TGAACG
						0.4571	0.5000	0.2500	0.4571
CCAA							(1)	(4)	(3)
							CAGACA / CAGACA	CAGACA / CAGACA	CAGACA / CGAACG
							1.000	0.5000	0.9143
CCAG								(10)	(9)
								CAGACA / CGAACG	CAGACA / CGAACG
								0.4571	0.4571
CCGG									(6)
									CGAACG / CGAACG
									0.8359

NOTE: Numbers in parenthesis show total number of possible F1 offspring genotypes with non-zero probability of being observed the population. SNP1 and SNP2 haplotypes (with X representing an unknown genotype) are abbreviated as: TTAA for TTAAXXXXXXXXX, TCAA for TCAAXXXXXXXXX, CCAA for CCAAXXXXXXXXX, TTAG for TTAGXXXXXXXXX, TCAG for TCAGXXXXXXXXX, CCAG for CCAGXXXXXXXXX, TTGG for TTGGXXXXXXXXX, TCGG for TCGGXXXXXXXXX, and CCGG for CCGGXXXXXXXXX.

HWE and the absence of recombination. Population substructure and mutations in SNP loci would drastically change the set of predicted haplotypes.

In conclusion, the analysis reveals that parsimony is the most rapid approach for predicting offspring haplotypes. However, stochastic approaches are recommended because they are less sensitive to departures from HWE in real populations. Bayesian methods are recommended over EM algorithms for the detection of unique haplotypes with very low population frequencies. Other statistical tools for the prediction of offspring haplotypes in higher-level analyses, such as genome-wide association studies, are beyond the scope of this study.

Major advancements have been made in marker technology for progeny testing and offspring prediction. More genetic markers are associated with various traits, and individuals in populations can be efficiently genotyped for SNPs with minimum output and effort (Zhu *et al*., 2016; Schmid and Bennewitz, 2017). As more candidate genes are identified and included in commercially produced genetic marker panels (Bertolini *et al*., 2018), the power of these test kits must be evaluated whether the same amount of information on the haplotype as well as the associated phenotype of an individual can be obtained with less genotype information. The analytical methods in this study could also be extended for models where recombination, population substructure, and other deviations from HWE are present in the population.

ACKNOWLEDGMENTS

GMBA wishes to thank Prof. John Wooliams of the Roslin Institute in Edinburgh, UK for providing the test data set and software tools used in the analysis.

REFERENCES

Banos G and Coffey MP. 2010. Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. *J Dairy Sci* 93(6):2775-8.

Barrett JC, Fry B, Maller J and Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.

Bertolini F, Schiavo G, Galimberti G, Bovo S, D'Andrea M, Gallo M, Buttazzoni L, Rothschild MF and Fontanesi L. 2018. Genome-wide association studies for seven production traits highlight genomic regions useful to dissect dry-cured ham quality and production traits in Duroc heavy pigs. *Animal* 12(9):1777–1784.

Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7:111–122.

Dash S, Singh A, Bhatia AK, Jayakumar S, Sharma A, Singh S, Ganguly I and Dixit SP. 2018. Evaluation of bovine high-density SNP genotyping array in indigenous dairy cattle breeds. *Anim Biotech* 29(2):129–135.

Huang S, He Y, Ye S, Wang J, Yuan X, Zhang H, Li J, Zhang X and Zhang Z. 2018. Genome-wide association study on chicken carcass traits using sequence data imputed from SNP array. *J Appl Genet* 59(3):1-0.

Iamartino D, Nicolazzi EL, Van Tassell CP, Reecy JM, Fritz-Waters ER, Koltes JE, Biffani S, Sonstegard TS, Schroeder SG, Ajmone-Marsan P and Negrini R. 2017. Design and validation of a 90K SNP genotyping assay for the water buffalo (*Bubalus bubalis*). *PLoS One* 12(10):e0185220.

Krag K, Poulsen NA, Larsen MK, Larsen LB, Janss LL and Buitenhuis B. 2013. Genetic parameters for milk fatty acids in Danish Holstein cattle based on SNP markers using a Bayesian approach. *BMC Genetics* 14(1):79.

Niu T. 2004. Algorithms for inferring haplotypes. *Genet Epidemiol* 27:334–347.

Qiao X, Su R, Wang Y, Wang R, Yang T, Li X, Chen W, He S, Jiang Y, Xu Q and Wan W. 2017. Genome-wide target enrichment-aided chip design: a 66 K SNP chip for cashmere goat. *Sci Rep* 7(1):8621.

Qin ZS, Niu T and Liu JS. 2002. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* 71:1242–1247.

Schmid M and Bennewitz J. 2017. Genome-wide association analysis for quantitative traits in livestock – a selective review of statistical models and experimental designs. *Arch Anim Breed* 60:335–346.

Stephens M and Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* 76:449–462.

Weir B. 1996. *Genetic Data Analysis II*. USA: Sinaner Associates, Inc. Publishers

Wigginton JE, Cutler DJ and Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 76:887–893.

Wooliams JA, Banos G and Coffey MP. 2006. Impact of leptin, leptin receptor and growth hormone receptor genotypes on milk production, feed intake and body energy traits in dairy cattle. *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, Belo Horizonte, Minas Gerais, Brazil, pp. 1–13.

Zhang W, Li J, Guo Y, Zhang L, Xu L, Gao X, Zhu B, Gao H, Ni H and Chen Y. 2016. Multi-strategy genome-wide association studies identify the DCAF16-NCAPG region as a susceptibility locus for average daily gain in cattle. *Sci Rep* 6:38073.

Zhu F, Cui QQ, and Hou ZC. 2016. SNP discovery and genotyping using genotyping-by-sequencing in Pekin ducks. *Sci Rep* 6:36223.