Supporting Information

Table S1. Subspecies, breeds, and animals used in this study**.**

Table S2. Summary of genetic diversity identified among 169 individuals derived from 5 breeds.

Table S3. Top 117 regions and overlapping genes in five cattle breeds.

Table S4. DAVID function analysis.

Figure S1. A. Multidimensional scaling plots of the identity by state matrices. Plots for the genetics structure of five breeds using the 1st and 2nd dimension. The label of each axis shows the proportion of the dimensions. B. Individual and breed relationships among 169 samples. Phylogenetic tree created from 72,945 SNPs and rooted by BRM. Genetic distance between pairwise combination of individuals was calculated using PLINK based on identical by state (IBS). C. Bar plots represent STRUCTURE inferences of individual assignments (K= 2–5). Each color represents the most likely ancestry of the cluster and each vertical bar represents one individual genotyped.

Figure S2. Decay of average r^2 over distance (bp). Average r^2 between markers in Holstein, Angus, Charolais, Brahman, N'Dama at various distances in base pairs ranging from 0 to 1Mb. LD (r^2) between SNPs was calculated in sliding 1-Mb windows. LD decay along genomic distance was plotted. The LD decay rate was estimated as the genetic distance at which the average pairwise correlation coefficient (r^2) dropped to half of its maximum value. While LD increased markedly in HOL and ANG, LD decay rates of BRM and NDA were estimated at ~40 kb and ~100 kb, where the r^2 drops to approximately 0.26 and 0.25, respectively. Compared to HOL and ANG which still had the highest r^2 values among the five breeds at 0.5 to 1Mb, the indicine BRM breed had low r^2 values at 50 to 100 kb and intermediate r^2 values at 0.5 to 1Mb. NDA had the highest r^2 values at 50 to 100 kb and the lowest r^2 at 1Mb, indicating that NDA was derived from a relatively small ancestral population and not subjected to very narrow recent bottlenecks.

Figure S3. Genome-wide distribution of positive selection regions. The outermost circle displays the cattle chromosomes. The inner circles represent the genome-wide distribution of selection regions. The height of the histogram bins indicates the density of selection regions as defined as $\log_2(di)$. The circles from inside going outside represent HOL, ANG, CHL, BRM and NDA (red).

Figure S4. Association test of *SAR1B* with HOL milk production traits and haplotype analysis in HOL, BRM and NDA.

Figure S5. Recombination rates of selected loci. We detected four patterns: (1) HOL, ANG and CHL shared similar peaks while BRM either has no peak or has its own unique peak near *GHR* (Figure 5) and *ABCA12* (Figure S5); (2) CHL had a unique internal peak near *KIT*; (3) BRM has significant peaks in both ends near *MC1R*; (4) all breeds were similar and had little or low level of recombination near *LAP3* and *LCORL*. CHL and NDA showed a low level of recombination near *WIF1*. As expected, there were strong correlations between high LD and low recombination or vice versa in these regions. We detected evidence for historical recombination within *GHR* and *ABCA12*, resulting in more than one LD block within each of them. It is noted that *ABCA12* shows most extreme divergence in human populations (Tennessen and Akey, 2011). *ABCA12*

encodes an ATP-binding cassette (ABC) transporter found in the lamellar granules of keratinocytes, which have an central role in the regulation of lipid trafficking in human studies (Lefevre et al., 2003; Akiyama, 2006). Cole et al. recently reported that *ABCA12* is associated with calf birth weight in Holsteins, and there may be divergence among the five breeds for this locus based on differences in calf sizes at birth (Cole et al., 2014). *ABCA12* also has been validated in previous studies as the causative mutation contributing to trait Ichthyosis fetalis in human and cattle (Thomas et al., 2006; Charlier et al., 2008; Goddard and Hayes, 2009).

Figure S6. Haplotype networks for selected loci.

Figure S7. The minor allele frequency distribution of nine SNPs in the RXFP2 gene region from SNP ARS-BFGL-NGS-76809 (chr12:29,238,096) to SNP BovineHD1200008655 (chr12:29,280,146) across five cattle breeds.

Figure S8. The minor allele frequency distribution of 88 SNPs in the Polled region on BOS1 ranging from SNP BovineHD0100000536 (chr1:1,670,513) to SNP ARS-BFGL-NGS-29653 (chr1:2,049,400) across five cattle breeds.

Table S1. Subspecies, breeds, and animals used in this study**.**

Breed	**Acronym**	**Animals**	**Subspecies**	**Usage**	**Breed region**	**Characteristics**
Taurine						
Holstein	HOL	44	*Bos taurus*	Dairy	North Holland and Friesland	Black and white coat, high milk yield
Angus	ANG	39	*Bos taurus*	Beef	Augus, Scotland	Black coat, high meat quality
Charolais	CHL	35	*Bos taurus*	Beef Multi-purpose	Charolles and Nievre	White to cream coat, large body size
Indicine						
Brahman	BRM	30	Mainly *Bos indicus*	Beef	Bred in America using Guzerat, Kankrej, Gir and etc. as founders.	Gray coat, humped. Heat and disease tolerance
African						
N'Dama	NDA	21	*Bos taurus*	Multi-purpose	Guinea	Fawn coat, small size. Trypanosome resistance
Total		169				

Table S2. Summary of genetic diversity identified among 169 individuals derived from 5 breeds.

	Acronym	Aninals	Het	F	Ar	pAr
Taurine						
	HOL	44	0.3460	0.0526	1.6347	0.0000
	ANG	39	0.3370	0.0857	1.6276	0.0001
	CHL	35	0.3559	0.0315	1.5275	0.0000
Indicine						
	BRM	30	0.2561	0.3098	1.5743	0.0000
African						
	NDA	21	0.2520	0.3121	1.6117	0.0000

Genetic diversity were measured in 5 breeds. **N** is the number of individuals used in this genetic diversity estimation; expected heterozygosity ($\boldsymbol{H}_{\mathbf{e}}$); the inbreeding coefficient (**F**); allelic richness ($\boldsymbol{A}_{\mathbf{r}}$) and private allele richness ($\mathbf{p}\boldsymbol{A}_{\mathbf{r}}$).

Table S3. Top 117 regions and overlapping genes in five cattle breeds. (As a separated excel file)

Table S4. DAVID function analysis. (As a separated excel file)

Figure S1. A. Multidimensional scaling plots of the identity by state matrices. Plots for the genetics structure of five breeds using the 1st and 2nd dimension. The label of each axis shows the proportion of the dimensions. B. Individual and breed relationships among 169 samples. Phylogenetic tree created from 72,945 SNPs and rooted by BRM. Genetic distance between pairwise combination of individuals was calculated using PLINK based on identical by state (IBS). C. Bar plots represent STRUCTURE inferences of individual assignments (K= 2–5). Each color represents the most likely ancestry of the cluster and each vertical bar represents one individual genotyped.

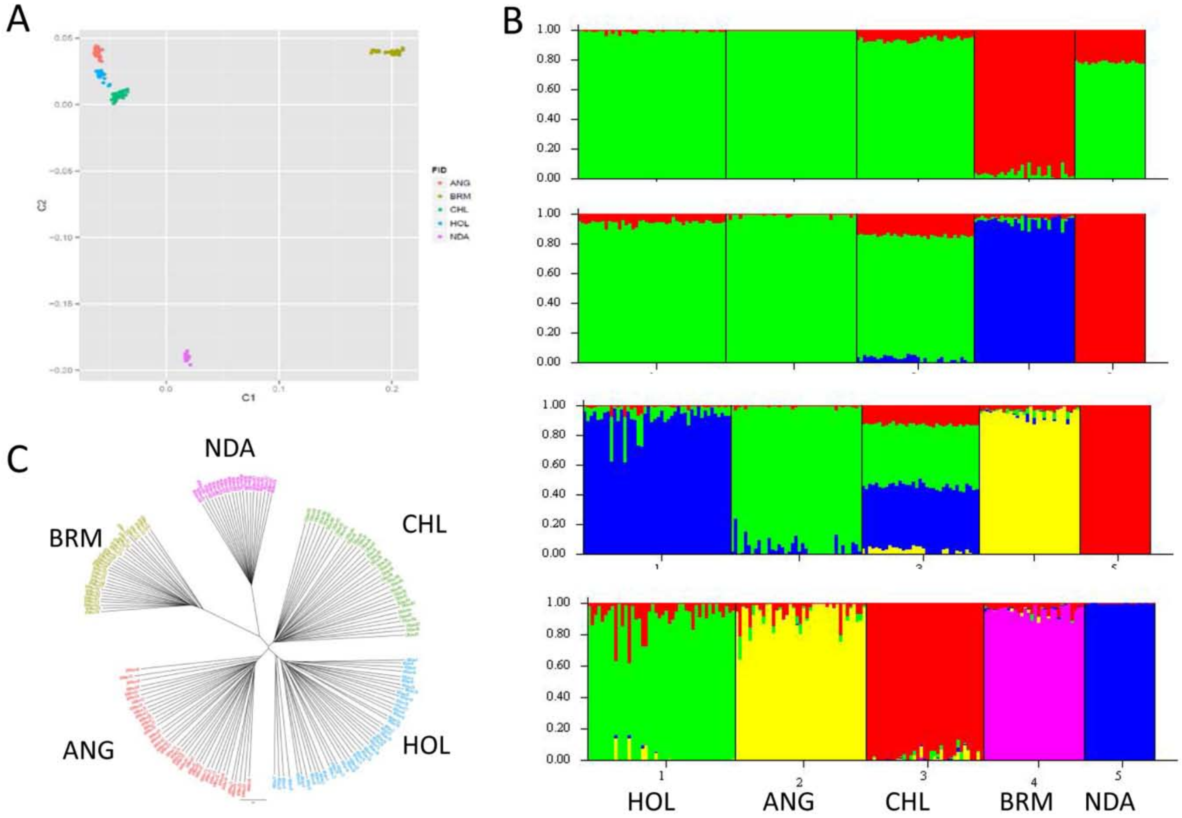


Figure S2. Decay of average r^2 over distance (bp). Average r^2 between markers in Holstein, Angus, Charolais, Brahman, N'Dama at various distances in base pairs ranging from 0 to 1Mb. LD (r^2) between SNPs was calculated in sliding 1-Mb windows. LD decay along genomic distance was plotted. The LD decay rate was estimated as the genetic distance at which the average pairwise correlation coefficient (r^2) dropped to half of its maximum value. While LD increased markedly in HOL and ANG, LD decay rates of BRM and NDA were estimated at ~40 kb and ~100 kb, where the r^2 drops to approximately 0.26 and 0.25, respectively. Compared to HOL and ANG which still had the highest r^2 values among the five breeds at 0.5 to 1Mb, the indicine BRM breed had low r^2 values at 50 to 100 kb and intermediate r^2 values at 0.5 to 1Mb. NDA had the highest r^2 values at 50 to 100 kb and the lowest r^2 at 1Mb, indicating that NDA was derived from a relatively small ancestral population and not subjected to very narrow recent bottlenecks.

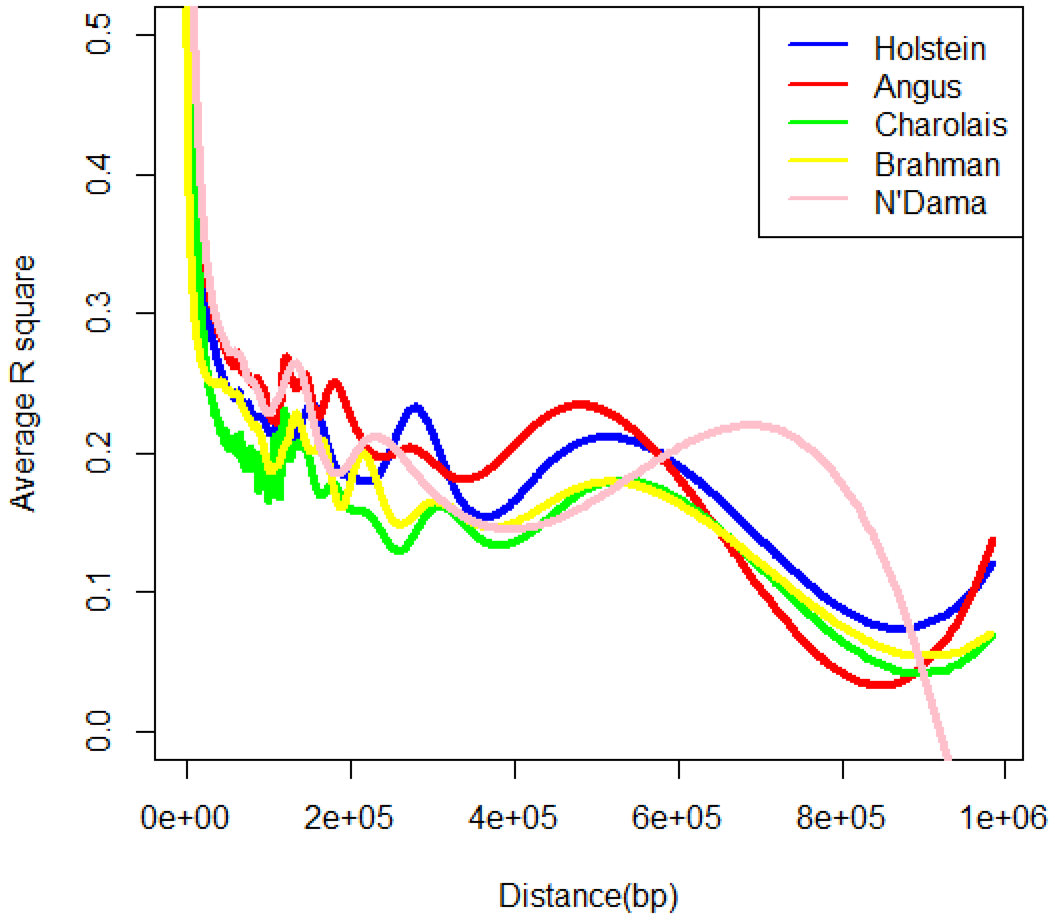


Figure S3. Genome-wide distribution of positive selection regions. The outermost circle displays the cattle chromosomes. The inner circles represent the genome-wide distribution of selection regions. The height of the histogram bins indicates the density of selection regions as defined as $\log_2(di)$. The circles from inside going outside represent HOL, ANG, CHL, BRM and NDA (red).

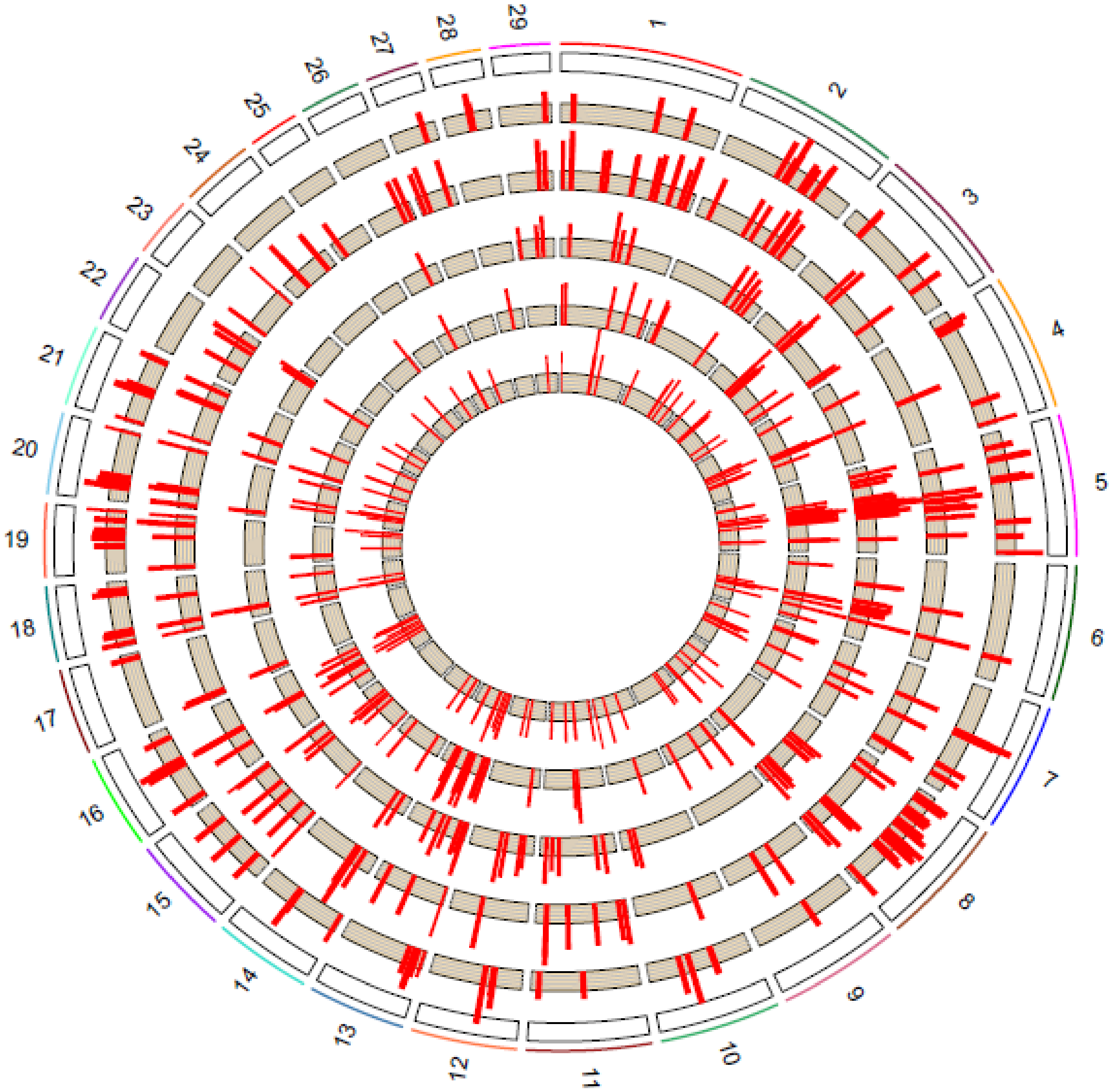


Figure S4. Association test of *SAR1B* with HOL milk production traits and haplotype analysis in HOL, BRM and NDA.

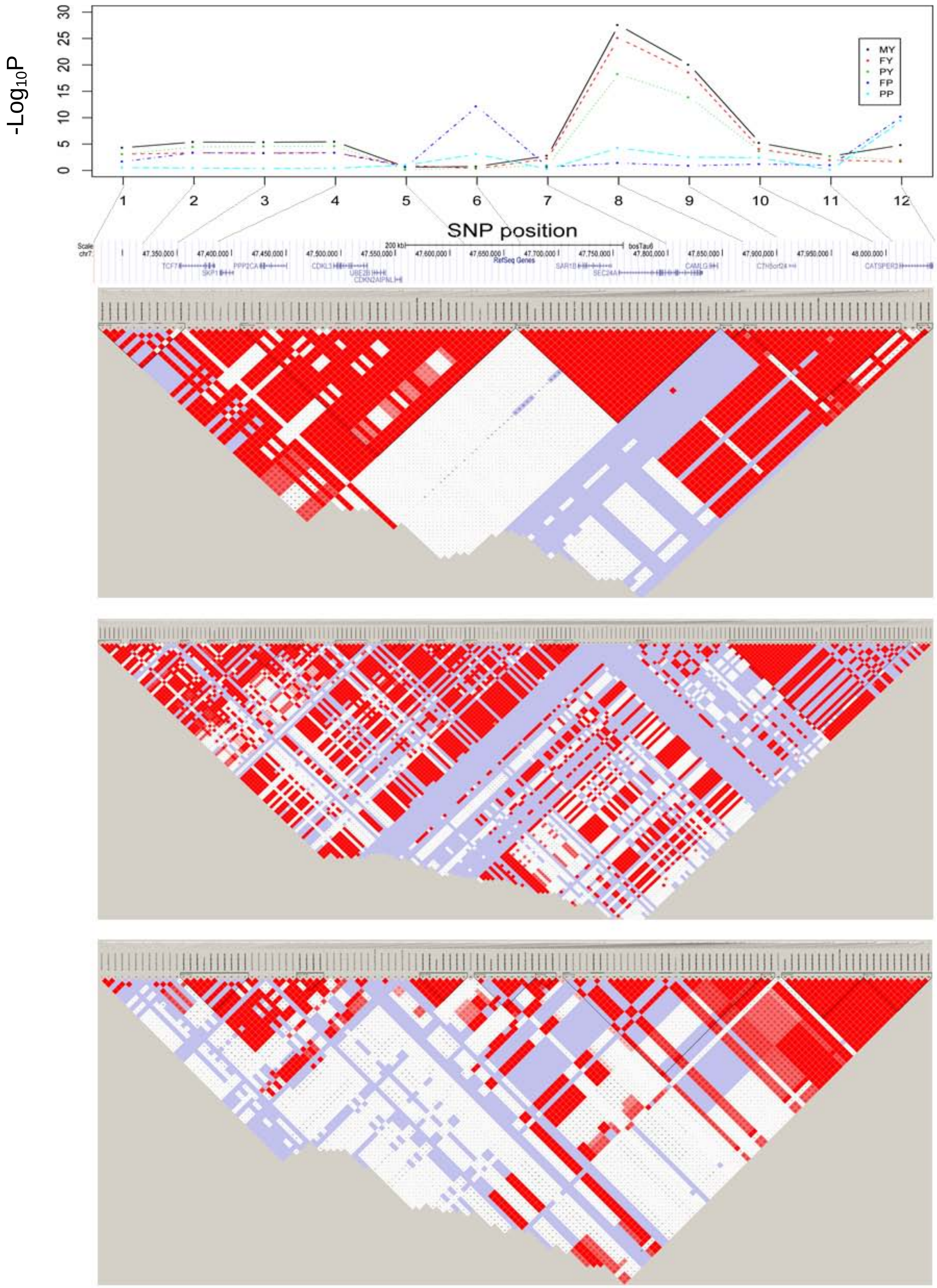
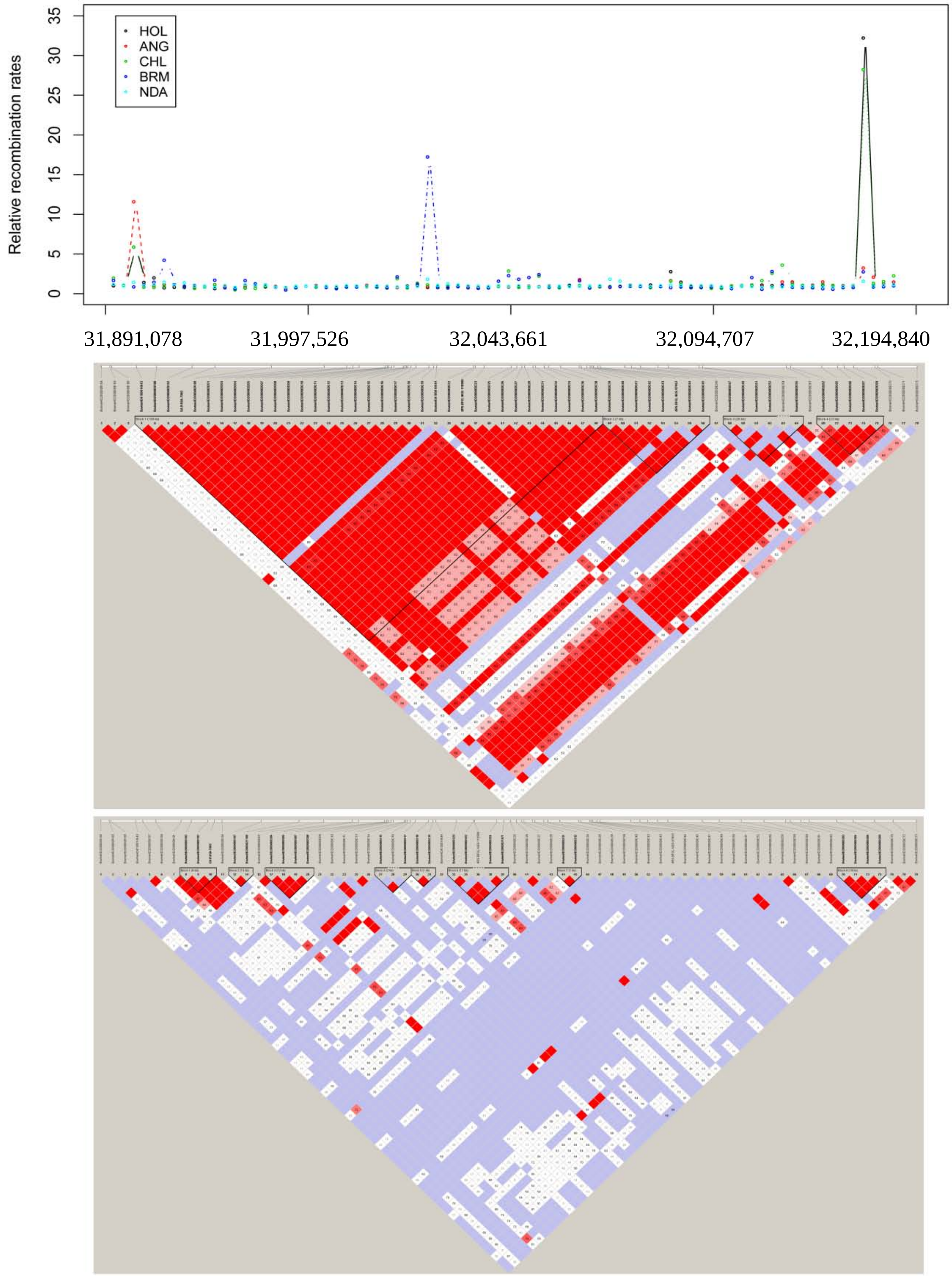
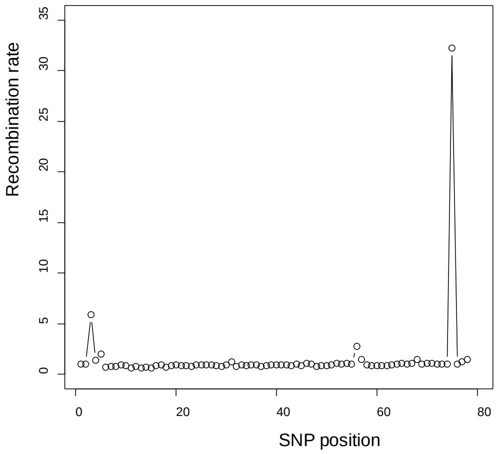


Figure S5. Recombination rates of selected loci. We detected four patterns: (1) HOL, ANG and CHL shared similar peaks while BRM either has no peak or has its own unique peak near *GHR* and *ABCA12* (Figure S5); (2) CHL had a unique internal peak near *KIT*; (3) BRM has significant peaks in both ends near *MC1R*; (4) all breeds were similar and had little or low level of recombination near *LAP3* and *LCORL*. CHL and NDA showed a low level of recombination near *WIF1*. As expected, there were strong correlations between high LD and low recombination or vice versa in these regions. We detected evidence for historical recombination within *GHR* and *ABCA12,* resulting in more than one LD block within each of them. It is noted that *ABCA12* shows most extreme divergence in human populations (Tennessen and Akey, 2011). *ABCA12* encodes an ATP-binding cassette (ABC) transporter found in the lamellar granules of keratinocytes, which have an central role in the regulation of lipid trafficking in human studies (Lefevre et al., 2003; Akiyama, 2006). Cole et al. recently reported that *ABCA12* is associated with calf birth weight in Holsteins, and there may be divergence among the five breeds for this locus based on differences in calf sizes at birth (Cole et al., 2014). *ABCA12* also has been validated in previous studies as the causative mutation contributing to trait Ichthyosis fetalis in human and cattle (Thomas et al., 2006; Charlier et al., 2008; Goddard and Hayes, 2009).

GHR: Relative recombination rates (Estimated recombination rate/background recombination rate) near *GHR* and haplotype analyses in HOL and BRM.

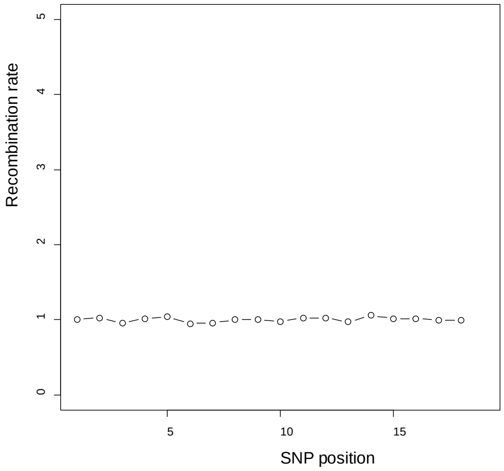


GHR


Recombination rate
0
5
10
15
20
25
30
35
0
20
40
60
80
SNP position

ABCA12

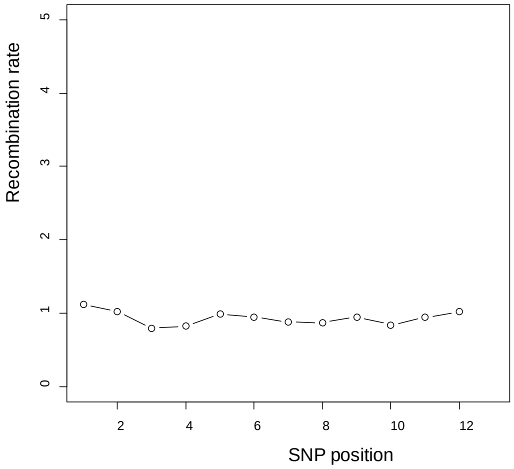
MC1R (block)



KIT

WIF1

LAP3



LCORL

SAR1B

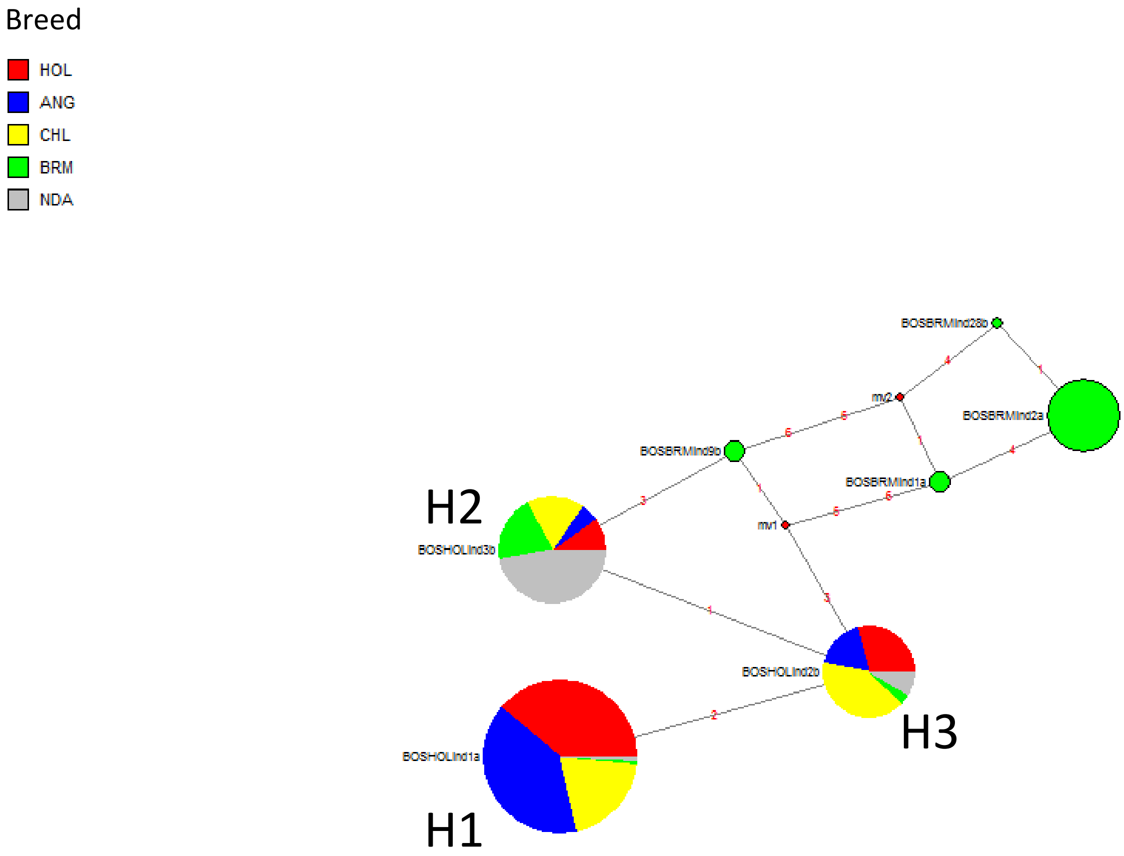
Figure S6. Haplotype networks for selected loci.

Growth and body size: We studied two regions near *LCORL* and *LRIG3* that were involved in growth rates and body size. For *LCORL*, we identified three top haplotypes (H1 at 18.34%, H2 at 15.34%, H3 at 12.46% and H4 at 11.24%) (Figure S6). H4 was a NDA-specific haplotype with a population-specific frequency of 90.48%. We also identified two BRM-specific haplotypes with frequency of 39.99% and 35.00%, respectively. Most of BRM haplotypes were clustered together, although H1 in BRM showed some degree of mixture with HOL, ANG, and CHL. For *LRIG3*, the top haplotype with a total frequency 36.50% included HOL (12.91%), ANG (33.82%), CHL (77.09%), BRM (6.67%), and NDA (65.90%) (Figure S6). The second haplotype (with frequency of 11.62%) contained HOL (1.85%), ANG (43.98%), and NDA (7.89%). Other genes related to body size were *FGF*5 (Cadieu et al., 2009) and *NCAPG* (Petersen et al., 2013). Their haplotype networks were shown in Figure S6.

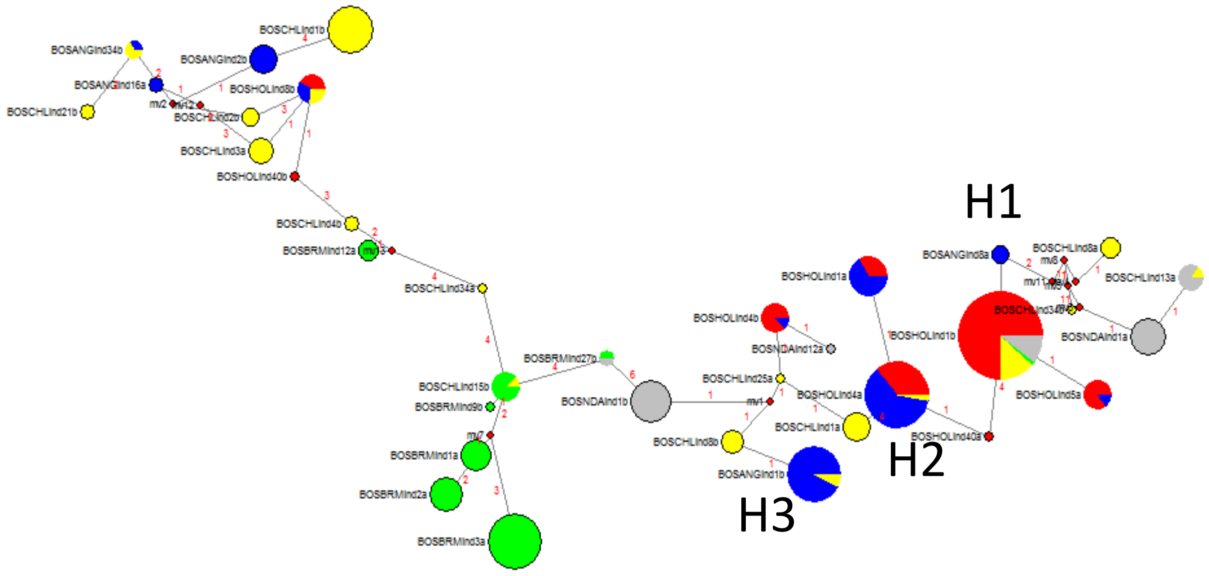
Coat color: we constructed haplotype networks in the regions containing *KIT* (55 haplotypes) and *MC1R* (33 haplotypes) (Figure S6). For *MC1R*, the prominent haplotype H1 consists of 19 SNPs with a total frequency of 72.90%, which contains most of the taurine samples including HOL (82.40%), ANG (96.15%), CHL (88.57%), and NDA (78.57%) and a few of the indicine BRM (6.49%). The second haplotype, H2, contained samples from four taurine breeds with various frequencies, but none from BRM. Based on the MAF information, we found little SNP diversity for most of SNPs in this region among HOL, ANG, CHL, and NDA. However, more SNP diversity was observed in BRM, as indicated by multiple distinct haplotypes. Additionally in BRM, we detected one recombination hotspot in the middle, and two hotspots located at both up- and down-stream of the *MC1R* region (Figure S5).

The *KIT* haplotype network was roughly divided into three main branches: the left branch mainly for the CHL haplotypes, the middle one for the BRM haplotypes, and the right branch mainly for the HOL and ANG haplotypes (Figure S6). It is interesting to note that most BRM haplotypes cluster exclusively together in the middle branch. The *KIT* gene has one haplotype, H1 (18.85%), overrepresented in HOL (56.14%), CHL (12.42%), and NDA (10.52%). H2 was overrepresented in both HOL (17.72%) and ANG (30.44%), but significantly less by either NDA (6.13%) or CHL (1.82%). Among the network, five haplotypes were specific to CHL, six to BRM, and four to NDA. One significant recombination hotspot was found in CHL, while another nearby hotspot was identified in ANG, CHL, and BRM (Figure S5). The reticulated pattern in the *KIT* haplotpye network suggest that recurrent events (selection, recombination and mutation) occurred across breeds.

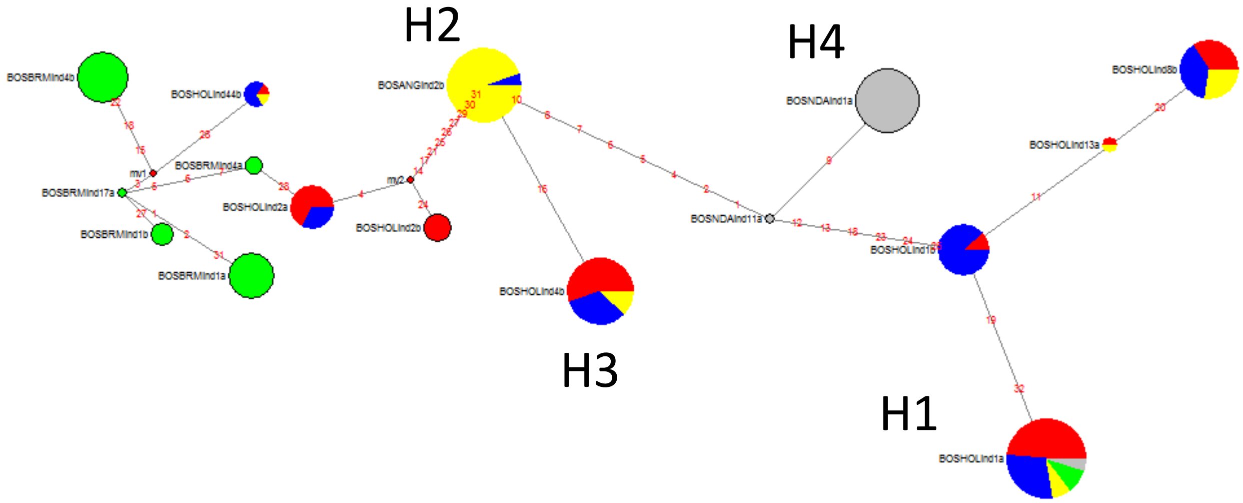
Growth, body size: *FGF5*



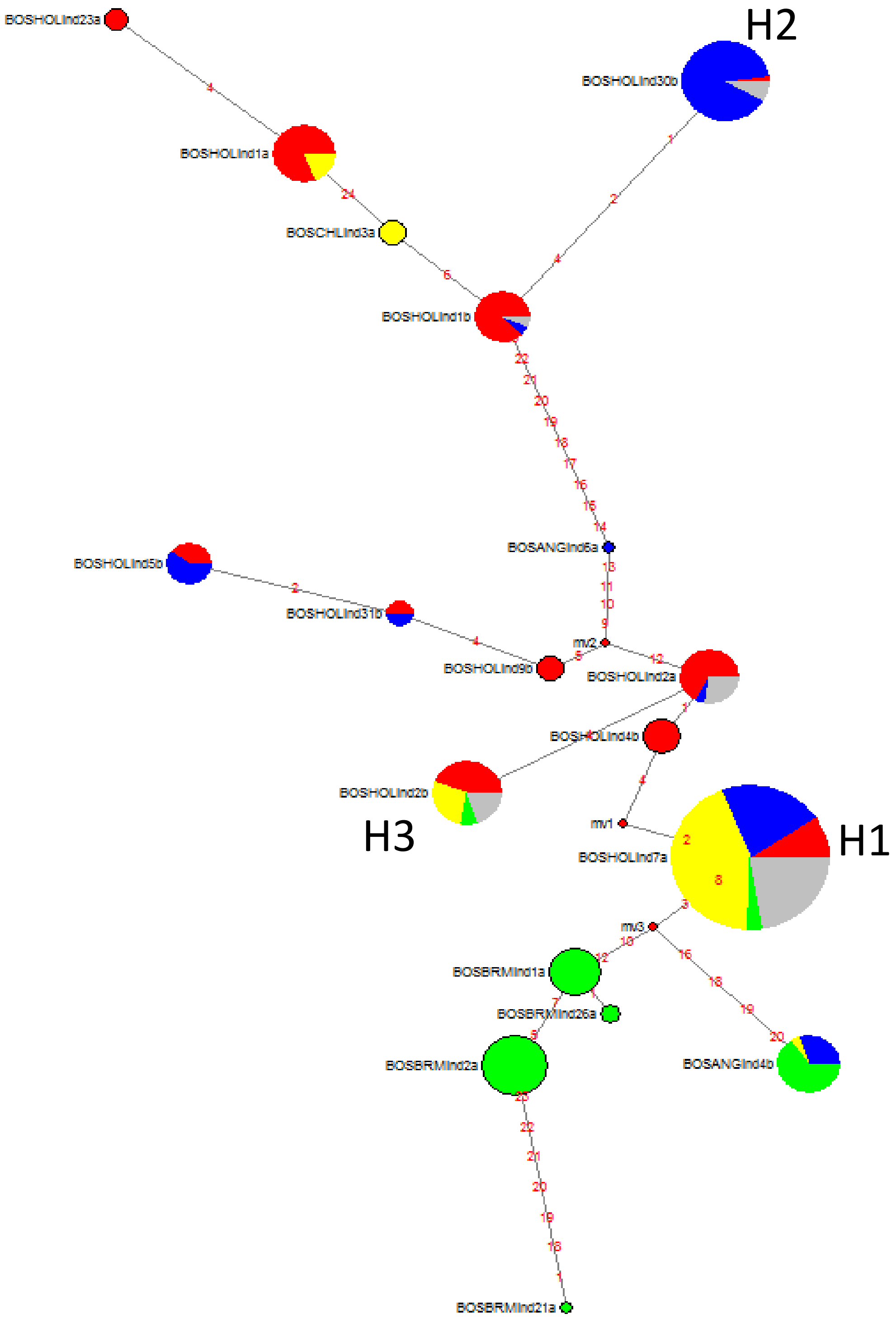
Coat color: *KIT*



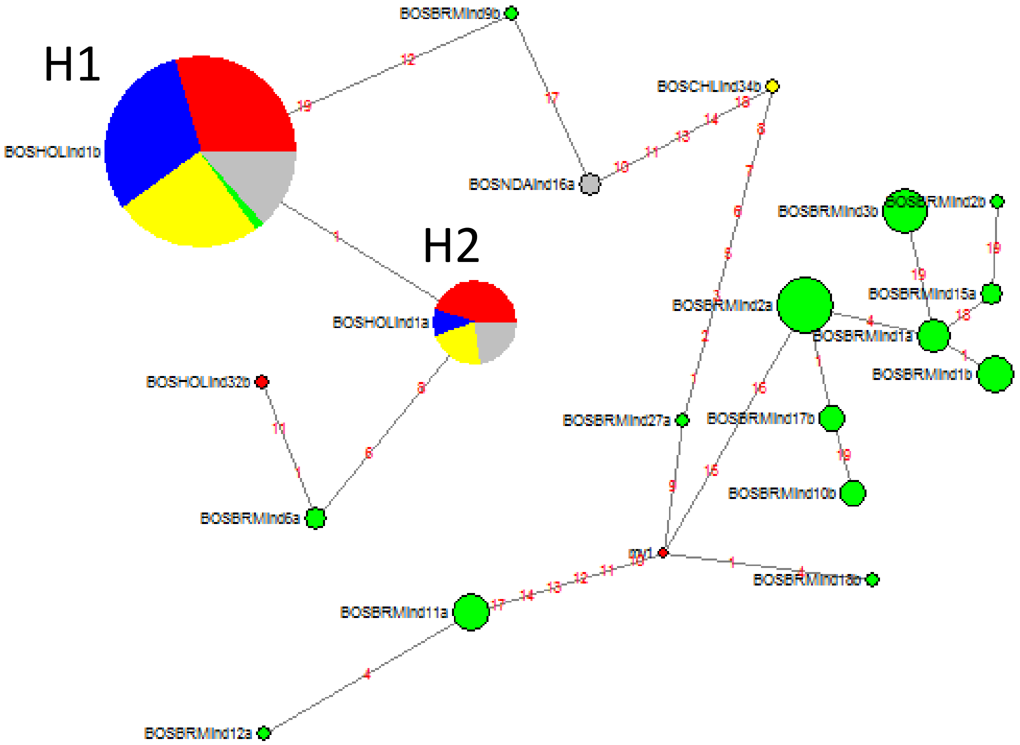
Body size: *LCORL*



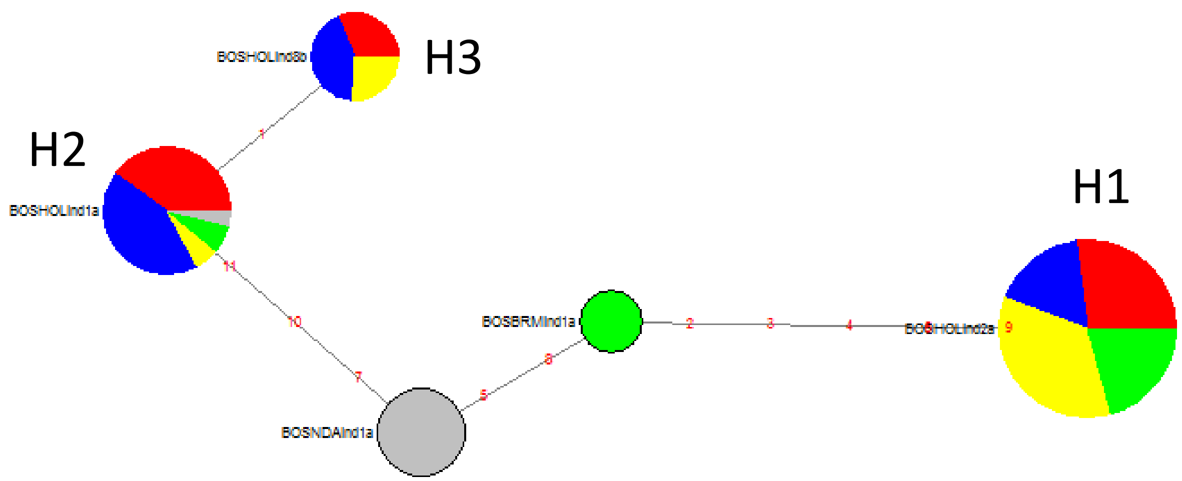
Growth, body size: *LRIG3*



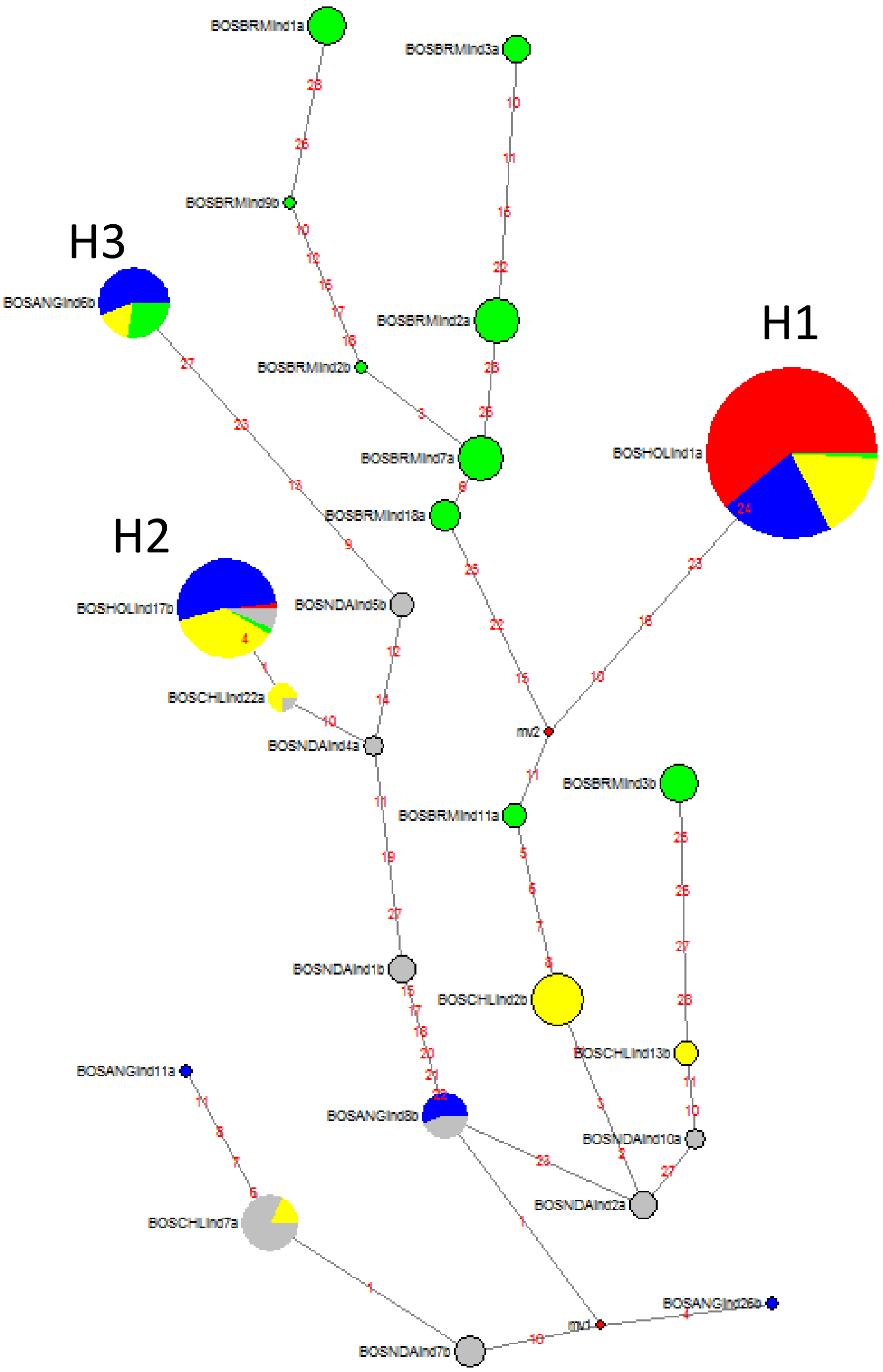
Coat color: *MC1R*



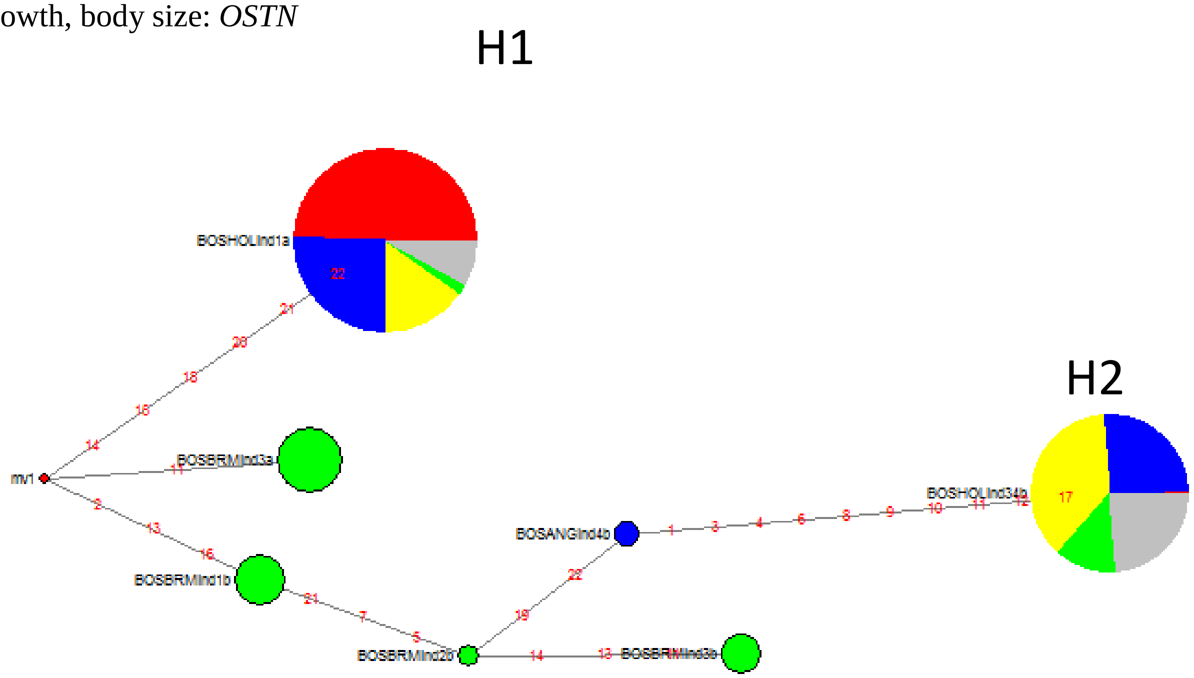
Growth, body size: *NCAPG*



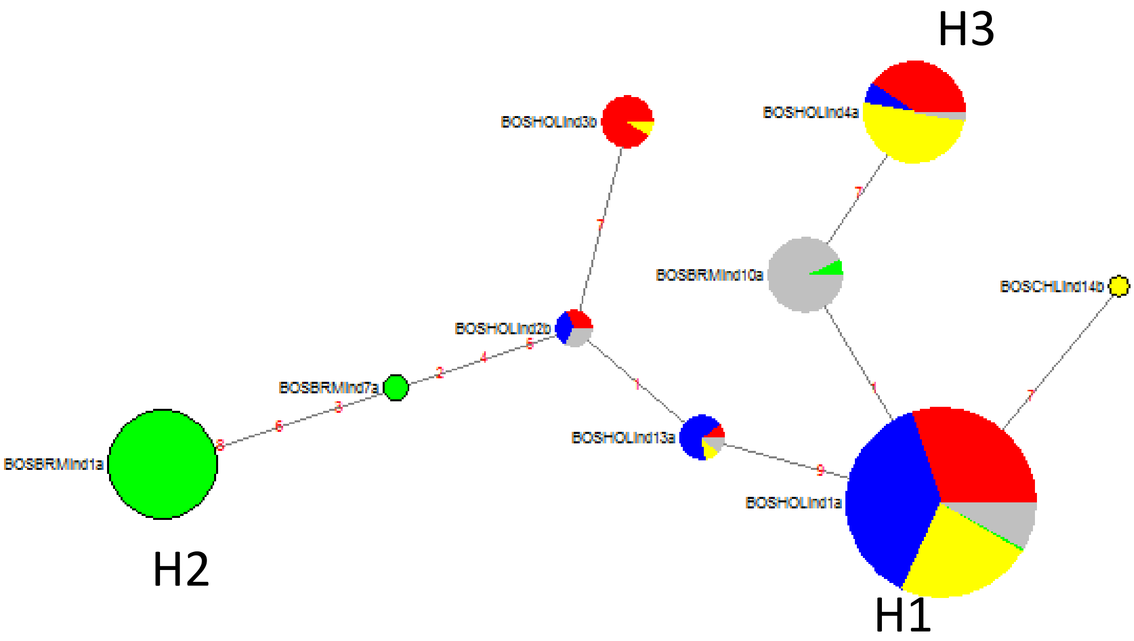
Mitosis: *NUDCD3*



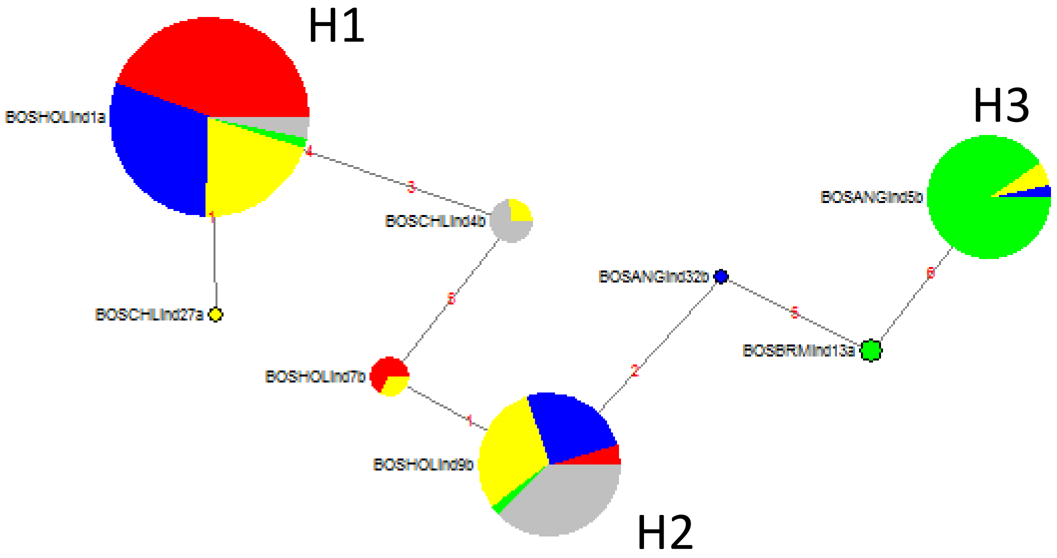
Growth, body size: *OSTN*



Polled: *RXFP2*



Milk production: *SAR1B*



Development: *WIF1*

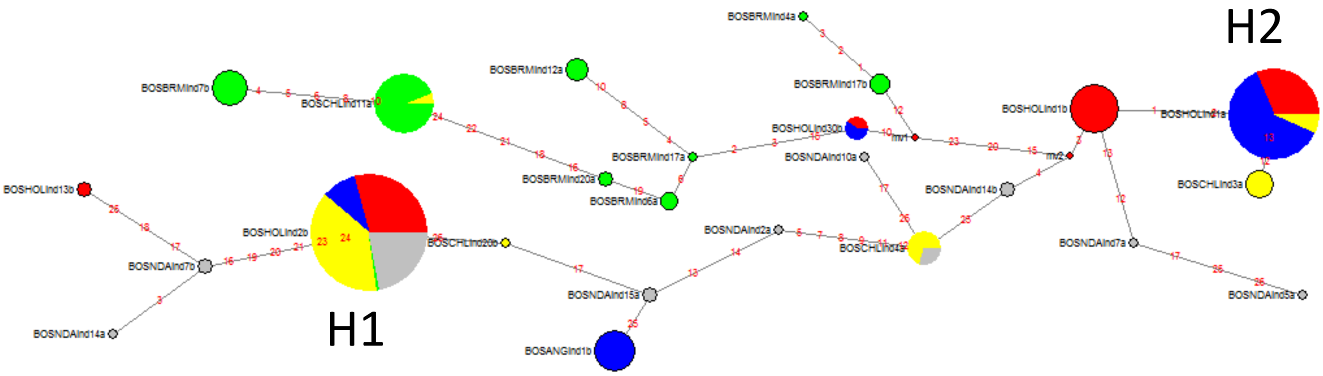


Figure S7. The minor allele frequency distribution of nine SNPs in the *RXFP2* gene region from ARS-BFGL-NGS-76809 (chr12:29,238,096) to BovineHD1200008655 (chr12:29,280,146) across five cattle breeds.

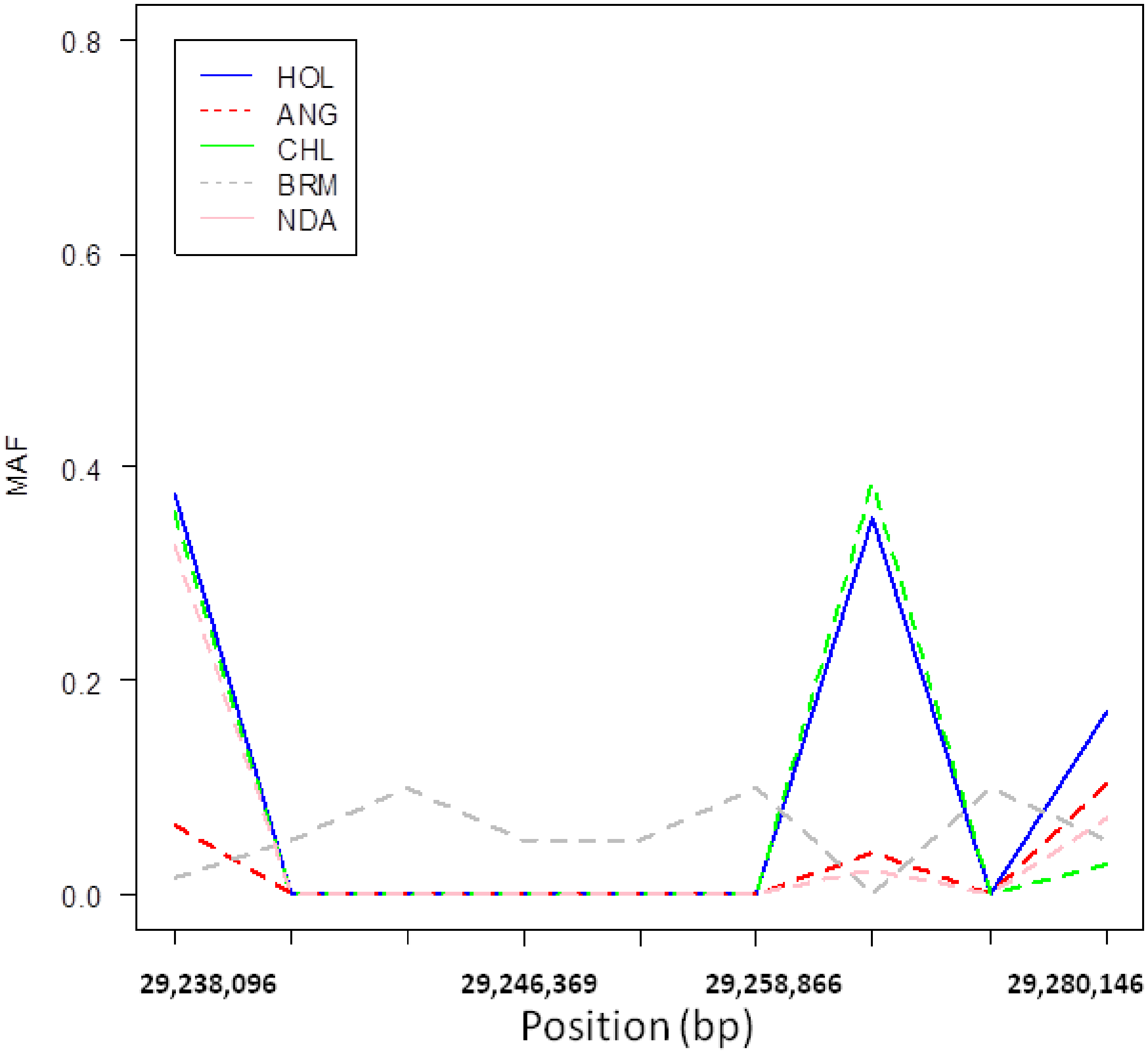
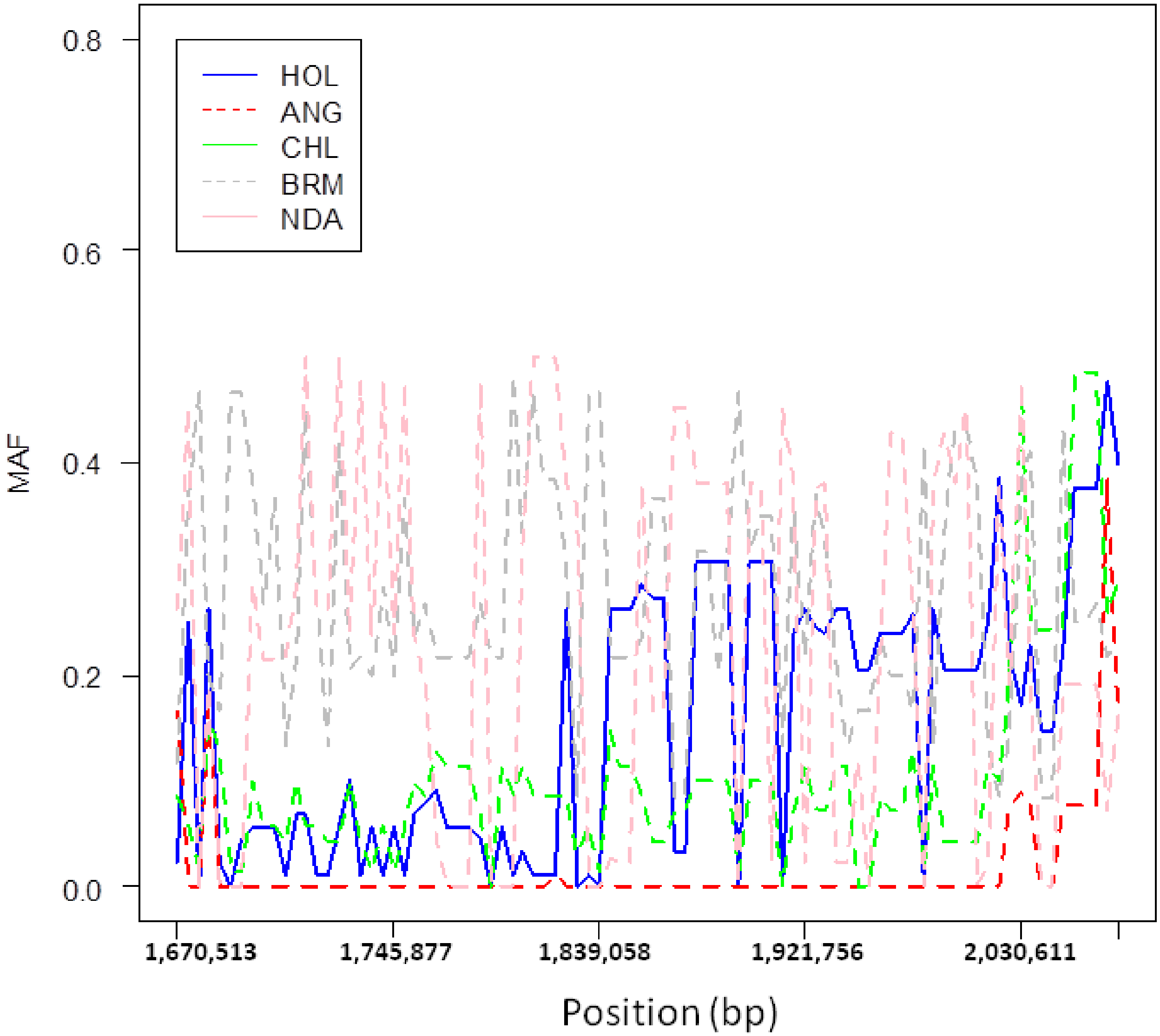


Figure S8. The minor allele frequency distribution of 88 SNPs in the polled region on chr1 ranging from BovineHD0100000536 (chr1:1,670,513) to ARS-BFGL-NGS-29653 (chr1:2,049,400) across five cattle breeds.


HOL
ANG
CHL
BRM
NDA
MAF
Position (bp)