Supporting Information

Muir *et al.* 10.1073/pnas.0806569105

SI Methods

AS_BIAS Fortran Program to Correct for Ascertainment Bias. Input to the AS_BIAS Fortran program to correct for ascertainment bias using the formula of Nielson and Clark is an ASCI file called "input.txt," each line of which contains NA, the allele frequency, and NRT, which is the number of loci with this allele frequency. The output file contains the corrected allele frequency distribution.

```
program Main
INTEGER N,D
DOUBLE PRECISION, ALLOCATABLE :: PASK(:)
DOUBLE PRECISION, ALLOCATABLE :: P(:)
DOUBLE PRECISION, ALLOCATABLE :: NR(:)
DOUBLE PRECISION C1,C2,C3,SUMPR
OPEN(UNIT = 10,FILE = 'INPUT.txt') !CATEGORIES AND FREQUENCES
OPEN(UNIT = 11,FILE = 'OUTPUT.txt') !CORRECTED DISTRIBUTION
WRITE(*,*) 'TOTAL NUMBER OF INDIVIDUALS N'
READ (*,*) NS !NUMBER OF INDIVIDUALS GENOTYPED
n = 2*NS !N number of total alleles = 2N
ALLOCATE(PASK(N))
ALLOCATE(P(N))
ALLOCATE(NR(N))
pask = 0.0
P = 0.0
!D INDIVIDUALS SEQUENCED, the depth.
NF = 0
NR = 0.0
1 READ(10,*,END = 22) NA,NRT
NF = NF + 1
NR(NA) = NRT
GO TO 1
22 CONTINUE
!NA IS THE ALLELE FREQUENCY
!NRT IS THE NUMBER OF ALLELES WITH THIS FREQUENCY
!NR(NA) = NUMBER OF ALLELES WITH FREQ NA
DO D = 2,2
SUMPR = 0.D0
CALL RLCOMB(N,d,C3)
DO K = 1,(N-1)/2
IF(K .GT. D) THEN
CALL RLCOMB(K,d,C1)
CALL RLCOMB(N-K,d,C2)
PASK(K) = 1.0-DEXP(C1-C3)-DEXP(C2-C3)
ELSE
CALL RLCOMB(N-K,d,C2)
PASK(K) = 1.0-DEXP(C2-C3)
END IF
if(pask(k) .gt. 0) then
SUMPR = SUMPR+NR(K)/PASK(K)
end if
END DO
DO K = 1,(N-1)/2
if(pask(k) .gt. 0) then
P(K) = (NR(K)/PASK(K))/SUMPR
ELSE
P(K) = 0.0
END IF
if(p(k) .gt. 0) then
WRITE(*,*) N,D,K,P(K)
WRITE(11,*) N,D,K,P(K)
end if
END DO
END DO
STOP
END
Following computes LOG(n!/(d!)(n-d)!)
SUBROUTINE RLCOMB(n,d,X)
INTEGER N,D
DOUBLE PRECISION X,Y,Z,W
IF(N .GE. D) THEN
CALL Rlfact(N,Y)
CALL Rlfact(D,Z)
CALL Rlfact(N-D,W)
X = Y-Z-W
! X = DEXP(X)
ELSE
X = 0.D0
END IF
RETURN
END
Following COMPUTES THE LOG OF A FACTORIAL
SUBROUTINE Rlfact(n,Y)
DOUBLE PRECISION Y
IF(N .EQ. 0) THEN
Y = DLOG(1.D0)
ELSE IF (N .GT. 0) THEN
Y = 0.D0
DO I = 1,N
Y = Y+DLOG(DFLOAT(I))
END DO
ELSE
Y = -1.D0
STOP
END IF
RETURN
END
```

Clustering Based on PCA. Let g_{ij} be the genotype for SNP i of individual j, where $i = 1$–M and $j = 1$–N. The g_{ij} were centered and normalized by subtracting the average allele frequency at that locus (p_i) and dividing by $\sqrt{(p_i)(1- p_i)}$. An NxN covariance matrix, ψ, was constructed among individuals based on the centered normalized genotypes, where $\psi_{jj'}$ is the covariance between individuals j and j'. Price *et al.* (13) defined the kth axis of variation as the kth largest eigenvalue of ψ. They also defined the ancestry, a_{jk}, of individual j along the kth axis of variation as the jth element of the kth eigenvector. They used the ancestry values as covariates to adjust phenotype and candidate gene data for admixture. In our application, we used the eigenvalues and ancestry coefficients to construct an index of shared ancestry among the strains to quantify strata among the samples. The ancestry coefficients were weighted by their associated eigenvalues λ_k, for all λ_k 1 and by 0 otherwise. Because there is always some shared ancestry between lineages, this index of ancestry provides a continuous scale for classification, which was divided into 10 bins or strata.

PCA analysis along the first two axis of variation are shown in supporting information (SI) Fig. S4. These results suggest that the RJF and Chinese Silkie are the most divergent of all samples

with all other breeds at the opposite extreme to these two. But seven axes of variation existed with eigenvalues 1. When all seven of these axes were combined into an index of weighted ancestry, the breeds within the center were clearly differentiable. These are shown in Fig. S5*A*, with bins constituting strata shown in Fig. S5*B*. It is interesting that all white egg layers, regardless of source, are considered to be from the same strata; similarly, all broiler populations constitute another strata. Thus, despite company differences in breeding goals, their populations are not really that different when considering the more global reference.

Results also are shown in Fig. S7. When inbreeding was calculated based on UPGMA clustering rather than on PCA, the UPGMA estimates were about 3% less than the PCA-derived values.

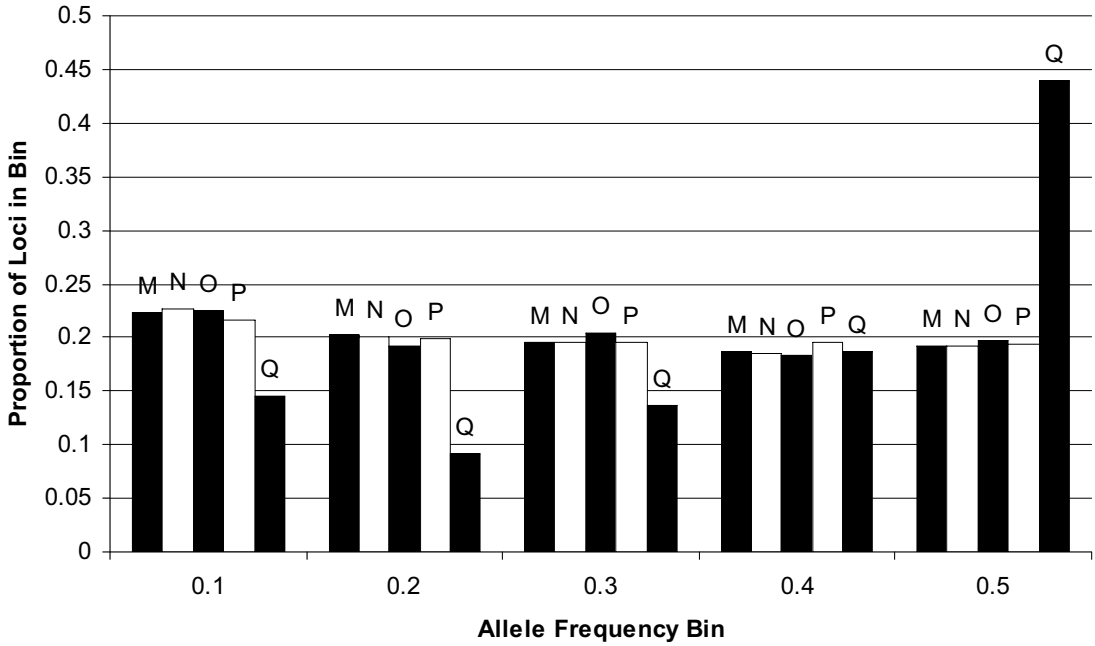

Fig. S1. Frequency distribution resulting from clustering at alternative distances.

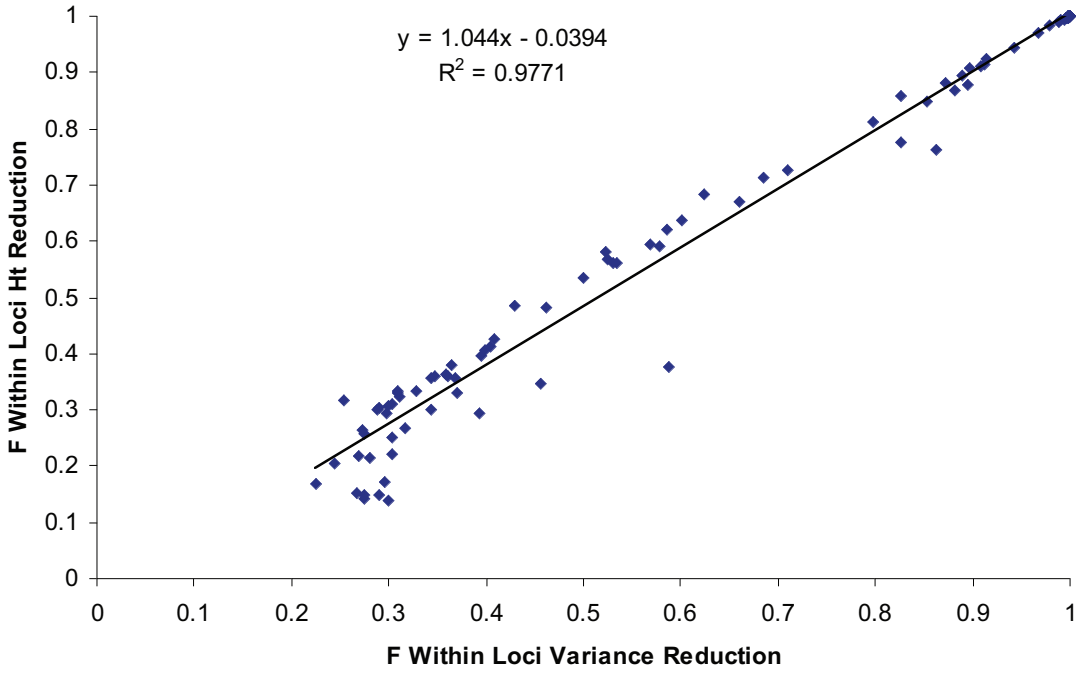

Fig. S2. Regression of F_{IT} estimated within loci as heterozygosity reduction on F_{IT} estimated within loci as variance reduction.

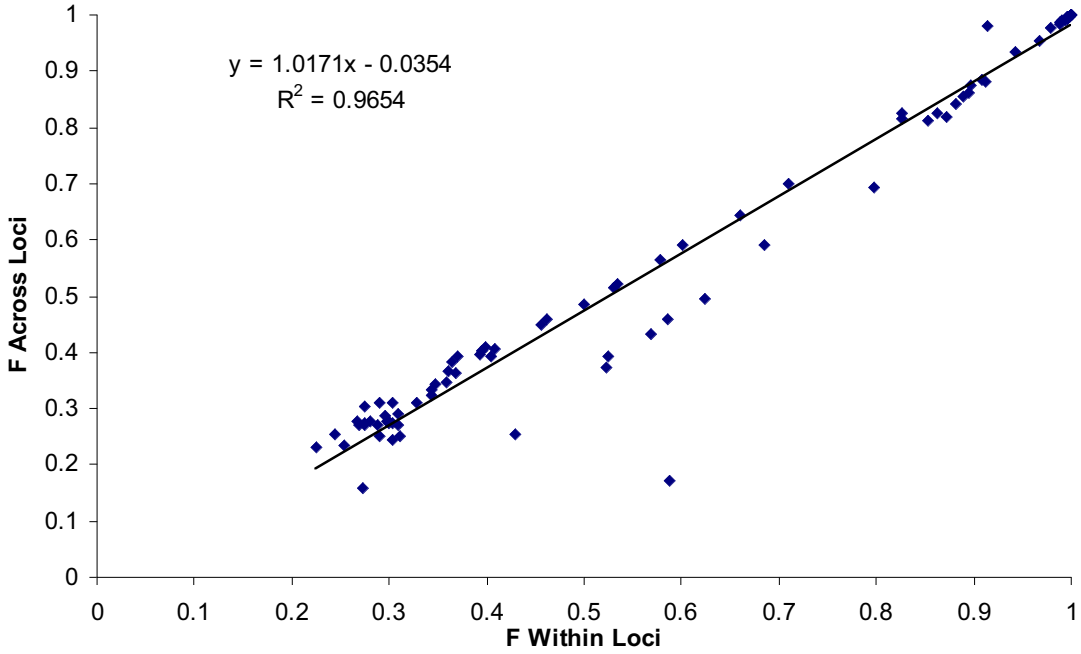

Fig. S3. Regression of F_{IT} estimated *across* loci as heterozygosity reduction on F_{IT} estimated within loci as variance reduction.

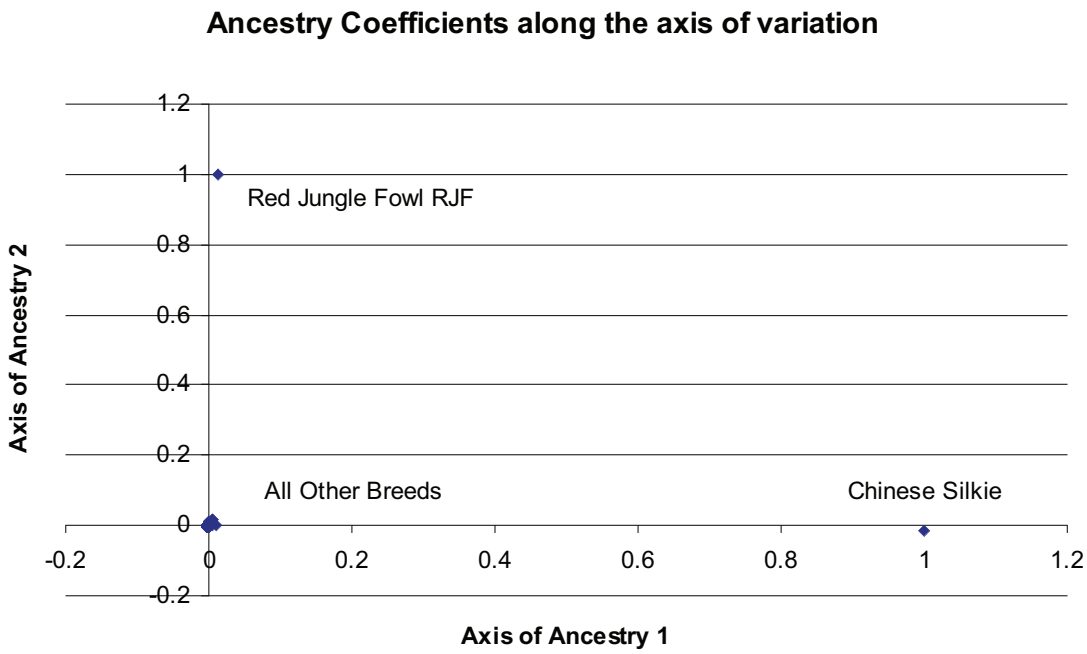

Fig. S4. Ancestry coefficients along the axis of variation.

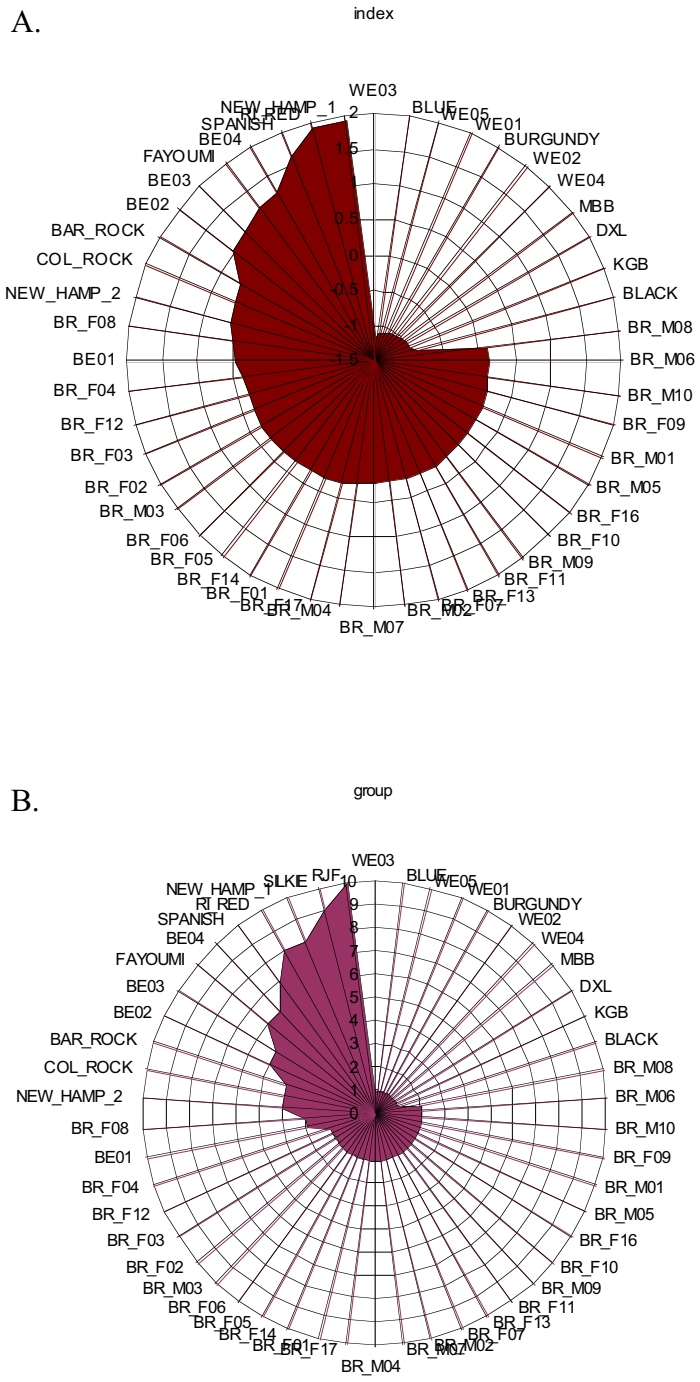

Fig. S5. Similarity of breeds based on index of ancestry from PCA analysis but excluding RJF and Chinese Silkie from the graph for better scale resolution of remaining breeds (*A*) and lines assigned to strata including all breeds (*B*). Each concentric circle represents a different strata.

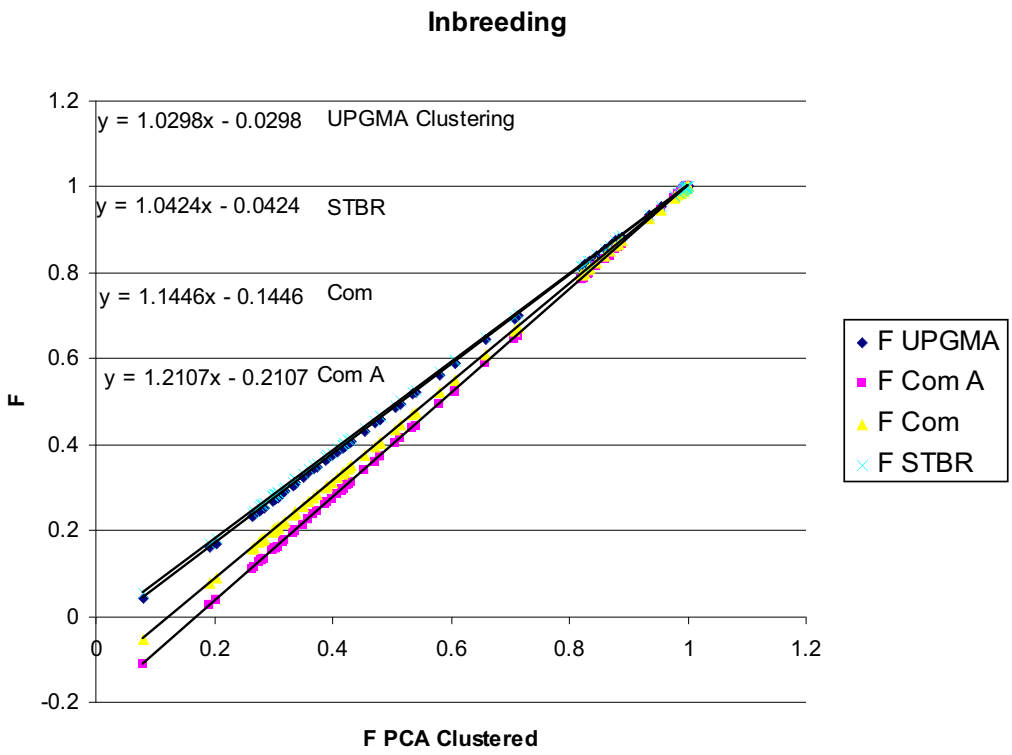

Fig. S6. Estimated inbreeding based on different reference populations (UPGMA, groups based on the UPGMA clustering method; Com A, group based only on birds from commercial company A; Com, groups based on all commercial birds across four companies; STBR, grouping based on standard breed identifications).

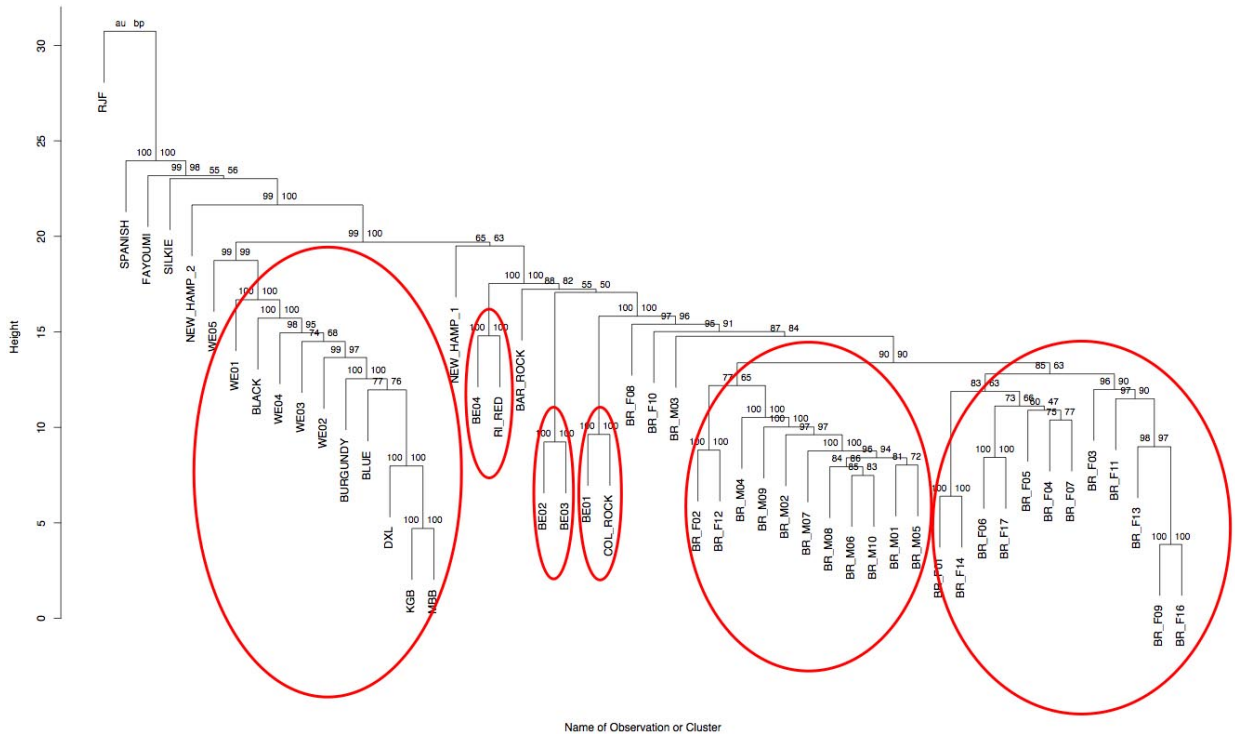

Fig. S7. Cluster analysis of commercial and STBR chicken lines.

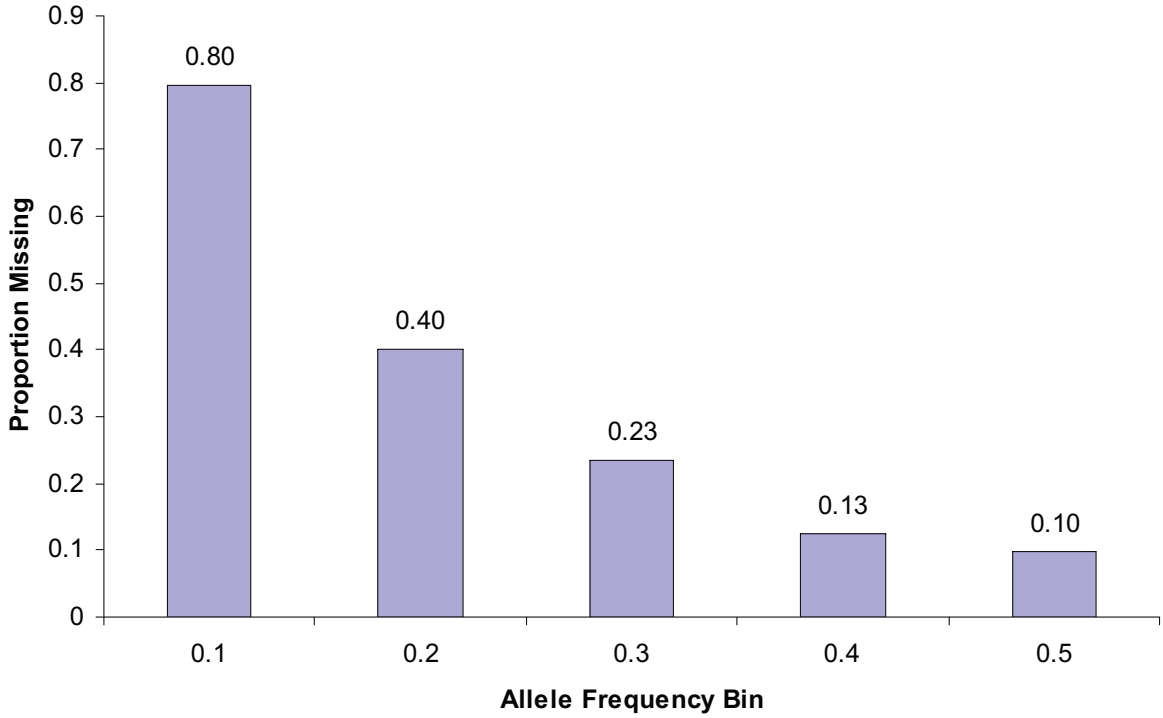

Fig. S8. Proportion of alleles missing in broiler line BR_F02 by allele frequency bin in the HAP.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Table S5
Table S6
Table S7