CRUMBLER: A tool for the Prediction of Ancestry in Cattle

Tamar E. Crum[1], Robert D. Schnabel[1,2], Jared E. Decker[1,2], Luciana CA Regitano[3], and Jeremy F. Taylor[1]

[1]Division of Animal Sciences, University of Missouri, Columbia, MO, USA 65211

[2]Informatics Institute, University of Missouri, Columbia, MO, USA 65211

[3]Embrapa Pecuária Sudeste, São Carlos, SP 13560-970, Brazil

Corresponding author: Jeremy F. Taylor. E-mail: taylorjerr@missouri.edu.

Short Title: Predicting the Ancestry of Cattle

E-mail addresses:

TEC: tamar.crum@mail.missouri.edu

RDS: schnabelr@missouri.edu

JED: deckerje@missouri.edu

LCAR: luciana.regitano@embrapa.br

JFT: taylorjerr@missouri.edu

Abstract

Background

In many beef and some dairy production systems, crossbreeding is used to take advantage of breed complementarity and heterosis. Admixed animals are frequently identified by their coat color and body conformation phenotypes, however, without pedigree information it is not possible to identify the expected breed composition of an admixed animal and in the presence of selection, the actual composition may differ from expectation.

Results

We tested an approach to estimate the global ancestry of individuals using ADMIXTURE and SNPweights. ADMIXTURE estimates ancestry using a model-based approach applied to large single nucleotide polymorphism (SNP) genotype datasets. Individuals are assumed to be unrelated and a supervised analysis can be performed using reference animals sampled to represent ancestral populations. SNPweights infers ancestry using weights estimated by principal component analysis for genome-wide SNP panels that have been genotyped in the reference panel animals. We constructed analysis pipelines to determine the ancestry of individuals with potentially complex ancestries using both methods and a specified reference population SNP dataset. The reference population was constructed using breed association pedigree information and an iterative analysis to identify sets of purebred individuals representative of each breed.

Conclusion

The finally adopted CRUMBLER pipeline extracts a subset of genotypes that are common to all current commercially available genotyping platforms and processes these into the file formats required for the analysis software and predicts admixture proportions using the reference population dataset.

Background

Ethnic differences in human disease susceptibility and health phenotypes complicate prognoses and risk predictions in admixed individuals [1]. The estimation of population ancestry is important for genome-wide association analysis (GWAA) [1] and for the development of genomic prediction models [2]. Population subdivision due to geographic or other forms of barriers followed by drift over substantial numbers of generations leads to the formation of genetically distinct populations, which may then admix if barriers to isolation are removed [3]. Cattle were domesticated about 10,000 years ago in the Fertile Crescent and Indus Valley giving rise to two currently recognized subspecies *Bos taurus taurus* and *Bos taurus indicus*. As humans migrated throughout Europe and Asia, small groups of herded cattle became geographically isolated and ultimately led to the formation of geographically isolated populations and the subsequent formation of distinct breeds about 200 years ago [4, 5]. These breeds were selected to have distinct coat colors, presence or absence of horns and specific horn shapes and also to become specialized as meat or milk producers. Crossbreeding was also extensively used during this early period of breed development to combine desirable characteristics of each of the ancestral breeds in the formation of new breeds.

For example, the British Shorthorn has extensively influenced the development of many modern breeds of cattle [5]. More recently, crossbreeding has been used, particularly, in beef production to capitalize on the benefits of heterosis for fitness, adaptive and production traits.

In humans, self-identified ethnicity has not been found to be reliable and the use of these data in association studies can fail to remove population stratification effects which inflate rates of false positive discovery of associations [6]. Visual classification of cattle based on breed characteristics suffers from the same problems as self-identification of ethnicity in humans, as these breed characteristics are frequently determined by alleles at relatively few loci. For example, recent extensive crossing with Angus cattle in the U.S. produces a black hided animal which masks all other solid coat colors found in other breeds and requires only a single dominant allele at the *MC1R* locus. As a consequence, black hided cattle have a “cryptic” population structure [6, 7] and Angus branded beef products sold in the U.S. may originate from animals that actually have relatively little Angus ancestry.

In the U.S. and many other countries, the breed of an animal is associated with its being registered with a breed association which requires that both parents of the animal be identified and also registered with the association. For 50 years, parentage has been validated by each breed association using blood or DNA typing. Many breed associations have closed herdbooks which means, in theory, that the pedigrees of all animals can be traced back to the animals that founded the breed’s herdbook. Other

breed associations have open herdbooks, which means that crossbred animals can be registered with the breed if they have been graded up by crossbreeding to have the expectation of a certain percentage of their genome (e.g. 15/16ths) originating from the respective breed based upon pedigree records and parentage validation. The term "fullblood" is used to identify cattle for which every ancestor is registered in the herdbook and can be traced back to the breed founders. The term "purebred" refers to animals that have been graded up to purebred status. Pedigree errors that occurred prior to, or that were not identified with the implementation of blood typing and DNA testing, lead to admixed animals being incorrectly classified as fullblood and incorrectly identified admixture proportions in purebred animals. The effects of recombination and random assortment of chromosomes into gametes leads to considerable variation in the extent of identity by descent between relatives separated by more than a single meiosis [8] and can also lead to admixture proportions that differ substantially from expectation based on pedigree in purebred animals.

In commercial production of beef in the U.S., crossbreeding is extensively used to capitalize on the effects of breed complementarity and heterosis resulting in herds of females with very complex ancestries that frequently use fullblood or purebred bulls sourced from registered breeders. Changes in the decision as to which breed of bull to use can result in large changes in admixture proportions of replacement cows and marketed steers between years and large differences can occur between herds for the same reason. When commercially sourced animals are used to generate resource populations to study the genomics of economically important traits such as feed

conversion efficiency [9, 10] or bovine respiratory disease [11], the presence of extensive admixture in the phenotyped and genotyped animals may impact the GWAA [9, 10] and leads to the training of genomic prediction models in populations for which the breed composition is not understood. As a consequence, the utility of these models in other industry populations, including the registered breeds in which the majority of genetic improvement is generated is also not understood.

The inability to accurately identify the breed composition of crossbred animals necessitates the development of a reproducible method to generate estimates of breed composition in cattle based on single nucleotide polymorphism (SNP) data. We developed the CRUMBLER analysis pipeline to estimate the breed ancestry of crossbred cattle using high-density SNP genotype data, publicly available software, and a reference panel containing genotypes for members of cattle breeds that are numerically important in North America. CRUMBLER scripts and the reference panel data are available on GitHub (https://github.com/tamarcrum/CRUMBLER). This software is released under the GNU General Public License.

Materials and Methods

Genotype data

From among the numerically most important cattle breeds in North America, in terms of their annual numbers of animal registrations, a list was compiled to define the target breeds for reference panel development. Composite breeds, such as Brangus and

Braford, were not included in this list but the progenitor Angus, Hereford and Brahman breeds were included. We also included breeds in this list that would broaden the representation of world cattle breeds (N'Dama representing African taurine and Nelore representing indicine in addition to Brahman which was formed in the U.S. as a composite [18]) and those breeds that were likely to be involved in early crossbreeding in the U.S. (Texas Longhorn).

From the 170,544 cattle with high-density SNP genotypes stored within the University of Missouri Animal Genomics genotype database, we extracted genotypes for 48,776 animals identified as being registered with one of the numerically important U.S. Breed Associations or belonging to other world breeds. Pedigree data were also obtained for these animals from each of the Breed Associations, where available (Table 1).

These individuals had been genotyped using at least one of 9 different genotyping platforms currently used internationally to genotype cattle including the GeneSeek (Lincoln, NE) GGP-90KT, GGP-F250, GGP-HDV3, GGP-LDV1, GGP-LDV3, and GGP-LDV4 assays, the Illumina (San Diego, CA) BovineHD and BovineSNP50 assays, and the Zoetis (Kalamazoo, MI) i50K assay. The numbers of variants queried by each assay and the number of individuals genotyped using each platform are shown in Table 2.

Marker set determination

To maximize the utility of the developed breed assignment tool, we identified the intersection set of markers located on the bovine assays for which we had available genotype data (Table 2). However, during the process of identifying the animals that would define the breed reference panel, only 16 individuals had been genotyped using the GGP-LDV4 (n=2) and GGP-LDV3 (n=14) assays and no animals had been genotyped using the GGP-LDV1 assay. To retain as many SNP markers as possible for subsequent analysis, we identified the intersection of markers present on the GGP-90KT, GGP-F250, GGP-HDV3, GGP-LDV3, GGP-LDV4, BovineHD, BovineSNP50 and i50K assays. This intersection set included 6,799 SNP markers (BC7K). The intersection of the markers representing 5 assays (GGP-90KT, GGP-F250, GGP-HDV3, BovineHD, and BovineSNP50) was 13,291 markers (BC13K). By removing the 16 individuals from the breed reference panel that had been genotyped on the GGP-LDV3 and GGP-LDV4 assays we were able to compare ancestry predictions using two marker set densities.

Pipeline

The developed CRUMBLER pipeline integrates the tools and the computational efficiency of publicly available software, PLINK [12, 13], EIGENSOFT [14, 15] and SNPweights [16] to generate ancestry estimates (Fig. 1). The pipeline integrates the processes of data reformatting and sequentially processing the data using these analytical tools to generate ancestry proportions for targeted individuals using a curated breed reference panel.

PLINK

PLINK PED formatted genotypes are required as input to the pipeline. PLINK (v1.90b3.31) was used for data filtering and reformatting for all project needs. Genotypes can arise from any of the common bovine genotyping platforms (Table 2), provided that a PLINK compatible MAP file is provided for each assay and data produced using only a single genotyping assay is included in each PED file. The pipeline utilizes the PLINK marker filtering tool (--extract) to extract the BC7K marker subset used for ancestry analysis. For analyses of animals genotyped on different genotyping platforms, the list of BC7K markers can be provided to extract markers common to all assays. The pipeline allows multiple input genotype files and uses the PLINK merge genotype files tool (--merge) to combine genotypes into a single file for downstream analysis.

EIGENSOFT

The EIGENSOFT convertf package is used to convert all genotypes from PLINK PED format into EIGENSTRAT format which is required by the SNPweights software. To process the reference panel data, principal component analysis using EIGENSOFT smartpca is used to generate the eigenvalues and eigenvectors that are required to calculate SNP weights using SNPweights. However, the smartpca package included in EIGENSOFT versions beyond 5.0.2 are not compatible with SNPweights. SNPweights requires an input variable, "*trace*", to be located in the log file output from the smartpca analysis. For versions of EIGENSOFT beyond 5.0.2, the source code can be edited to

ensure that the log file output is compatible with the SNPweights software (See Supplementary Information).

SNPweights

SNPweights implements an ancestry inference model using genome-wide SNP weights computed using genotype data for external reference panel individuals. To obtain SNP weights, the reference panel genotypes are normalized by SNP to improve the results of the subsequent PCA analysis from which a kinship matrix is generated [15]. A principal component decomposition is then used to generate the eigenvalues and corresponding eigenvectors of the kinship matrix [16]. The SNP weights file only needs to be recalculated if the reference panel is changed. EIGENSTRAT formatted target animal genotypes are input into SNPweights, along with the precomputed reference panel SNP weights. The SNP weights are then applied to the target individuals to estimate their ancestry proportions [16].

Reference panel development

The definition of a set of reference individuals that define the genotype frequencies at each SNP variant for each reference breed is technically demanding, but vitally important to the process of defining ancestry. This process assumes that selection has not operated to change gene frequencies between target and reference population animals, and that each population is sufficiently large that drift has not impacted allele frequencies. It also assumes that migration between different countries does not influence population allele frequencies when registered animals are imported or

exported. FastSTRUCTURE [17] analysis and iterations of animal filtering using SNPweights was performed using the genotypes of candidate reference panel individuals to remove individuals with significant evidence of admixture from the reference breed panel. An overview of the processes and iterations of filtering conducted in the development of this reference panel set is shown in Fig. S1 and Table 1.

FastSTRUCTURE analysis to identify candidate reference panel individuals

Genotype data for 48,776 individuals produced by one of 8 different genotyping assays were available for fastSTRUCTURE analysis (Table 1). We initially performed focused fastSTRUCTURE analyses using small numbers of reference breeds including Angus and Simmental, Angus and Gelbvieh, Angus and Limousin, Angus and Red Angus, Red Angus, Hereford, Shorthorn and Salers, Red Angus, Hereford and Shorthorn, and N'Dama, Nelore and Brahman (Figs. S2-S8). Individuals possessing an ancestry assignment of at least 97% to their designated breed were retained for subsequent analysis (see Supplementary Methods and Table 1). Following filtering based on breed assignment, 17,852 individuals representing 19 of the original breeds remained for further analysis (Supplementary Methods and Figs. S2-S8). All of the Salers animals were removed in this filtering analysis which is consistent with previous work that found that Salers and Limousin were very similar [4]. To produce similar sample sizes for the comparison of each of the reference breeds, we randomly sampled 200 individuals from each reference breed for which at least 200 individuals remained after filtering on an ancestry assignment of at least 97%, otherwise all remaining individuals were included

for the breed (Table 1). Following the fastSTRUCTURE analysis using K=19 after removal of Salers and using the BC7K marker set, Texas Longhorn was also removed from the reference panel breed list due to the inability to distinguish Texas Longhorn as a distinct population (Figure 2). Further, due to the known common ancestry [18] and similarity between Nelore and Brahman (Figure 2), the breeds were combined to represent *Bos taurus indicus*.

SNPweights analyses to refine and validate reference panel members

Random sampling of reference breed individuals was performed to create sample sets containing ≤n individuals per breed, for n = 50, 100, 150 and 200 individuals. Sampling was performed such that if a reference breed had ≥n candidates then n individuals were randomly sampled, otherwise, all available individuals were sampled. An analysis was performed using the BC7K marker set, SNPweights was used to assign reference breed ancestries to the same sample of individuals that was used to produce the SNP weights for each of the four samples of individuals (Figs. 3 a-b and Figs. S9-S10). In the self-assignment analyses conducted using the reference breed sample sets of ≤100 individuals per breed and ≤50 individuals per breed, 7 individuals were removed due to their estimated breed ancestry being ≤60% to their registry breed (Holstein n=3, Jersey n=1, Japanese Black n=3) (Figs. 3 a-b).

Breeds with open herdbooks

For the Gelbvieh, Limousin, Shorthorn, Simmental, and Braunvieh breeds that have open herdbook registries, fullblood or 100% ancestry individuals were identified based

on pedigree data obtained from the respective breed associations (Table 1). Charolais also has an open herdbook registry, however, access to Full French imported Charolais breed members was limited. As a result, all individuals identified as purebred in the association registry were retained for downstream analysis, however, these individuals could contain up to 1/32 introgression from another breed. From among these individuals, a random sample of 200 individuals was taken for each breed with more than 200 identified fullblood individuals, otherwise all animals were sampled. Individuals previously included in the candidate reference panel following preliminary fastSTRUCTURE filtering for the open herd book breeds were removed and replaced with the fullblood individuals.

Additional reference panel filtering using SNPweights

After filtering animals identified to not be fullblood based on their pedigree information, we randomly sampled ≤50 individuals per reference breed and utilized SNPweights to estimate weights for each sample and also to estimate breed ancestries for members of the same sample that was used to generate the SNP weights. Based on these analyses, we created 5 overlapping reference breed sets, each containing individuals with ≥90%, ≥85%, ≥80%, ≥75%, or ≥70% ancestry assignment to their registry breeds (Table 3).

Results and Discussion

296 The concept of breed and breed membership is man-made and does not inherently
297 exist in nature. Moreover, the formation of breeds of cattle is very recent. Cattle
298 domestication began about 10,000 years ago but the formation of herdbooks has
299 occurred only during the last 200-250 years [5]. Nevertheless, the effects of drift and
300 human selection over the last 200 years have caused sufficient divergence among
301 breeds that breed differences are identifiable at the molecular level. Such signals are
302 essential for breed ancestry analyses to be effective in modern admixed animals.
303
304 **Reference panel development**
305 FastSTRUCTURE analyses performed using the candidate individuals for each of the
306 19 reference breeds suggested population subdivision in both the Hereford and
307 Simmental (Figure 2). Pedigree analysis for the Herefords within each subpopulation
308 indicated that the subpopulations comprised animals from the highly inbred USDA Miles
309 City Line 1 Hereford population (L1) and other individuals representing broader U.S.
310 Hereford pedigrees. The Miles City L1 Hereford cattle were derived from two bulls, both
311 sired by Advance Domino 13 (AHA registration number 1668403) and 50 Hereford
312 foundation cows. Since the founding of the L1 Herefords, the migration of germplasm
313 has been unidirectional from L1 into the broader U.S. industry, as the L1 population has
314 been closed since its founding [19]. However, the L1 Herefords have profoundly
315 influenced the U.S. Hereford population. L1 Herefords do not segregate for recessive
316 dwarfism, which has been a threat to Hereford breeders since the 1950s, and this has
317 led to L1 cattle becoming popular in the process of purging herds of the defect [20]. In

1980, the average proportion of U.S. registered Herefords influenced by L1 genetics was 23%. By 2008, this proportion had increased to 81% [19].

The apparent subpopulation division within the Simmental breed (Figure 2) represents a large sire family within the Simmental breed. Members of this sire family are present in both subpopulations, however, in one subpopulation the family members are all fullblood and in the other they are all purebred or percentage Simmental animals. This indicates the need to remove clusters of highly related individuals in the formation of the reference panel.

By randomly sampling individuals from the candidate reference breed set and using SNPweights to assign these individuals to reference populations, we found that reference panel breed sample sizes of ≤50 or ≤100 individuals appeared to capture the diversity within each breed and appropriately determined the ancestry of the tested individuals (Fig. 3 a-b). For each breed, the percent ancestry predicted for the tested reference samples was, on average, 3.86% higher when the SNP weights were estimated using ≤50 individuals per breed than when ≤100 individuals per breed were used (Table 4). This clearly reflects the increased homogeneity of individuals within smaller samples of individuals from each breed.

After the replacement of reference breed individuals with those identified to be fullblood based on pedigree analysis for the open herdbook Gelbvieh, Simmental, Limousin, Braunvieh, Shorthorn, and Charolais breeds, additional self-assignment analyses were conducted to evaluate the effects of marker set size on ancestry

prediction. Breed reference panels were again constructed by randomly sampling ≤50 individuals per breed and SNP weights were calculated using both the BC13K markers and BC7K markers. The estimated SNP weights were then used to self-assign ancestry to members of the reference panel animals representing the reference breed set. The ancestry predictions for the reference breed individuals using either the BC7K (Fig. 4a; Fig. S11) or BC13K (Fig. 4b; Fig. S12) marker sets indicate that use of the BC13K marker set did not significantly impact the ancestry predictions. Consequently, the use of all markers common to the 8 commercially available genotyping platforms appears to be sufficient to assign breed ancestry for the majority of animals produced in the U.S.

We next examined the effects of reference breed homogeneity on ancestry assignment by identifying reference panel members that had been assigned to their breed of registry using SNPweights with probabilities of ancestry of ≥90%, ≥85%, ≥80%, ≥75%, and ≥70%, respectively (Table 3). From these individuals, reference breed panels were obtained by randomly sampling ≤50 individuals per breed, until each individual was represented in at least one sample set. SNP weights were then estimated using the BC7K marker set and ancestry was assigned for these individuals using SNPweights (Figs. 5-6 and Figs. S13-S15). Limiting the reference breed panel members to those individuals with ≥90% ancestry assigned to their breed of registry produced a reference panel that did not represent the extent of diversity within each of the breeds (Fig. 5). On the other hand, using an ancestry assignment of ≥85% clearly captured greater diversity within each breed (Fig. 6) and maximized the self-assignment of ancestry to the breed of registration (Table 5).

To examine whether the specific individuals represented in the reference panel sample influenced the self-assignment of ancestry to the sampled individuals, a second sample of ≤50 distinct individuals per breed was obtained from the individuals with ≥85% assignment to their breed of registration and analyzed with SNPweights (Fig. 7). This analysis indicates that the ability to predict ancestry was not influenced by the individuals sampled from the set of animals with ≥85% ancestry to their breed of registration. .

Figs. 6 and 7 suggest that the use of a reference breed panel constructed by the random sampling of ≤50 individuals per breed from individuals with ≥85% self-assigned ancestry to their breed of registration maintained sufficient within-breed diversity to accurately estimate the ancestry of target individuals. However, these figures also reveal small amounts of apparent introgression from other reference panel breeds within each of the breeds. This does not appear to be an issue of marker resolution since the analyses performed with the BC7K and BC13K marker sets generated similar results (Fig. 4). Consequently, we conclude that these apparent introgressions can be due to lack of power or may represent the presence of common ancestry among the breeds prior to the formation of breed herdbooks ~200 years ago. Molecular evidence for this shared ancestry exists, for example, Hereford and Angus cattle share the *Celtic* polled allele [21] and the segmental duplication responsible for the white anterior, ventral and dorsal coat color pattern occurs only in Hereford and Simmental cattle and

387 their crosses [22]. These data clearly indicate that crossbreeding was widespread prior
388 to the formal conceptualization of breeds.
389
390 **Reference panel validation**
391 To evaluate the ability of the selected reference breed panel (Table 1) to identify breed
392 composition, ancestry estimates were produced for several different sets of animals with
393 breed composition that was known *a priori* based on pedigree information. The
394 validation individuals included 27 animals from the Beefmaster breed, an admixed
395 breed of cattle predicted to contain ¼ Shorthorn, ¼ Hereford, and ½ Brahman descent
396 based upon pedigree. A second analysis was performed for 238 advanced generation
397 crossbred individuals with *a priori* pedigree estimates of 50% Angus and 50%
398 Simmental composition.
399 The Beefmaster breed utilizes mating strategies that produce individuals that are
400 expected to possess 50% Brahman, 25% Shorthorn, and 25% Hereford ancestry.
401 However, registerable animals are ultimately advanced generation composites and so
402 drift, meiotic sampling of parental chromosomes and selection are all expected to create
403 individual variation in these ancestry proportions. SNP weights were estimated for the
404 reference breed panel individuals (≤50 individuals per breed with ancestry self-
405 assignment percentages ≥85% to their breed of registry, Fig. 6). Using these SNP
406 weights, SNPweights software was used to estimate the ancestry proportions for the 27
407 registered Beefmaster individuals (Fig. 8). On average, SNPweights estimated the
408 Beefmaster individuals to contain 30.5% *Bos taurus indicus*, 16.5% Hereford, and
409 19.5% Shorthorn and the remaining 33.5% was distributed among the remaining breeds

represented in the reference panel (Table 6). SNPweights appeared to underestimate the pedigree expected breed compositions for Brahman, Shorthorn, and Hereford within the Beefmaster breed. However, if we normalize the proportions that were assigned to these breeds (30.5% *Bos taurus indicus*, 16.5% Hereford, and 19.5% Shorthorn) we obtain relative ancestries of 45.9% *Bos taurus indicus*, 24.9% Hereford and 29.2% Shorthorn which are close to expectation. Ancestry assignments of registered fullblood animals to their breed of registration were generally in the range 31-99%, with 93% of animals being assigned to their breed of registration with genome proportions of >50%. The remaining genome proportions again appear to identify common ancestry between the breeds that predates breed formation (Fig. S16).

To test the effects of using different reference panel individuals to assign ancestry for the Beefmaster individuals, we formed three independent reference panels with ≤50 individuals per breed from the set of individuals with ≥85% assignment to their breed of registration. SNP weights were estimated for each reference panel and ancestry was estimated for the 27 Beefmaster animals (Fig. 9). Variation in ancestry clearly exists between the Beefmaster individuals but little variation was detected within individual Beefmasters based upon the use of different reference panels. This indicates that the filtering and animal selection process for the formation of the reference panel led to robust estimates of breed composition.

Ancestry estimates for 238 crossbred individuals were obtained using SNPweights. From their available pedigree records, these crossbred individuals were

commercial, advanced generation animals with an expected 50% Angus and 50% Simmental ancestry. Results of the SNPweights analysis support the pedigree data (Fig. 10). The presence of Red Angus ancestry in these animals reveals the inability of the methodology to fully differentiate between Angus and Red Angus, which only diverged in the U.S. in 1954, and also the influence of Red Angus in the U.S. Simmental breed.

To assess the ability of CRUMBLER to distinguish between Red Angus and Angus influence, we conducted an analysis using SNPweights software either excluding Red Angus or Angus, respectively, from the reference panel for the estimation of SNP weights, but retained the omitted animals for the estimation of ancestry. When Red Angus was excluded from the reference panel, the registered Red Angus individuals cluster with Angus but show introgression from other breeds. When Angus individuals were excluded from the reference panel for SNP weight calculations, the registered Angus individuals clustered with Red Angus but with introgression from other breeds. While the two breeds share ancestors and foundation individuals, the breeds have still partially differentiated. (Fig. S17).

Admixture

We also tested the Admixture software for ancestry estimation using the same reference breed panel that was developed for use with SNPweights. Admixture uses maximum likelihood estimation to fit the same statistical model as STRUCTURE, however, STRUCTURE does not allow the specification of individuals of known descent

456 to be used as a reference panel [23]. Admixture allows a supervised analysis, in which
457 the user can specify a reference set of individuals, by specifying the "--supervised" flag
458 and requires an additional file with a ".pop" suffix to specify the genotypes of the
459 reference population individuals [23]. Unlike SNPweights, the reference population
460 individuals' genotypes must be provided in a genotype file for each analysis.
461
462 We first conducted an Admixture analysis in which we self-assigned ancestry for
463 the animals in the reference breed set formed with ≤50 individuals per breed from the
464 individuals that had ≥85% assignment to the breed of registration (Fig. 11). The results
465 shown in Fig. 11 are similar to those in Fig. 6 for the same reference panel, albeit with
466 perhaps less evidence of background introgression. We next conducted an analysis
467 using the reference panel used in Fig. 11 merged with 2,005 crossbred target
468 individuals that had been identified by pedigree as being high percentage ancestry
469 Hereford animals (FE Herefords). The results shown in Fig. 12, reveal a significant
470 change in the ancestry proportions estimated for the reference panel Guernsey,
471 Gelbvieh and Romagnola individuals between the two analyses which used exactly the
472 same reference panel, but differed only in the number of individuals for which ancestry
473 was to be estimated. This suggests that Admixture may use the target individuals to
474 update information provided by the reference panel individuals specified in the ".pop"
475 file. Consequently, the Admixture estimated ancestry proportions appear to be context
476 dependent and may vary based on the other individuals included in the analysis.
477

Moreover, the order in which the target individuals appear in the genotype input file also seems to affect the Admixture estimates of ancestry proportions for the target individuals. Fig. 13 shows the results of an Admixture analysis in which the target individuals were identical to those shown in Fig. 12, but the order of the reference individuals and the 2,005 Hereford crossbred individuals was reversed in the input files. In Fig. 12, the reference individuals appear before the 2,005 Hereford crossbred individuals in the input file, whereas in Fig. 13, the 2,005 Hereford crossbred individuals appeared before the reference individuals in the input file. The results reveal a significant change in ancestry proportions for Guernsey and Gelbvieh, but the Romagnola now appear to be non-admixed. Finally, we performed an Admixture analysis for these animals in which the order of animals in the input genotype file was completely randomized (Fig. 14). Following analysis, the individuals were sorted to generate Fig. 14. Again, the ancestry proportions for the Guernsey, Gelbvieh and Romagnola individuals suggest these breeds to be admixed. Because of these inconsistencies between results, we chose to not use Admixture for ancestry estimation.

Broader application using additional commercially available assays

To broaden the spectrum of data from different commercially available assays that can be evaluated, an additional intersection of markers was obtained using 11 commercially available bovine assays including the GGP-90KT, GGP-F250, GGP-HDV3, GGP-LDV3, GGP-LDV4, BovineHD, BovineSNP50, i50K, Irish Cattle Breeding Federation (Cork, Ireland) IDBv3, and GeneSeek (Lincoln, NE) BOVG50v1 assays. The intersection SNP set included 6,363 SNPs (BC6K). A SNPweights self-assignment analysis using the

reference set of individuals with ≥85% assignment to their breed of registration was conducted to assess the effects of the reduction in markers used for ancestry assignment. The ancestry proportions assigned based on the BC6K marker set (Fig. 15) do not differ significantly from those obtained using the BC7K marker set (Fig. 6). This result indicates the utility of CRUMBLER and the reference panel breed set across the spectrum of commercially available genotyping platforms.

Conclusions

The determination of a set of reference population breeds and individuals that define allele and genotype frequencies at each variant for each of the breeds is arguably the most important, yet technically difficult step in the process of ancestry estimation. We employed several iterations of filtering to remove recently admixed individuals and identify a relatively homogeneous set of individuals that nevertheless represented the variation that might be expected among individuals within a breed. Once determined, the reference panel genotype data need only be processed once to obtain SNP weights removing the need to share genotype data for reference individuals in subsequent studies [16]. The upfront development of an external reference breed panel capitalizes on the rich ancestry information available in large available datasets, and relatedness, variation in sample sizes and diversity among the target individuals does not affect the inference of ancestry [16].

In cattle, the visual evaluation of breed characteristics is a poor method for evaluating the ancestry of individuals. Breed association pedigrees can be used to

524 estimate expected breed compositions, however, the random assortment of
525 chromosomes into gametes and selection can lead to ancestry proportions that differ
526 from those expected based upon pedigree. Moreover, the vast majority of commercial
527 beef cattle in the U.S. have no or very limited pedigree information and since these
528 animals are frequently used for genomic research [9–11], there is a need for a robust
529 tool that can provide ancestry estimates for downstream use in GWAA or other genetic
530 studies.
531
532 We tested Admixture and SNPweights and found that results from Admixture
533 appear to depend on the ancestry and order of appearance of individuals within the
534 genotype input file. We therefore developed an analysis pipeline, CRUMBLER, based
535 upon PLINK, EIGENSOFT and SNPweights to automate the process of ancestry
536 estimation. The pipeline utilizes the 6,799 SNPs present on 8 commercially utilized
537 bovine SNP genotyping assays and results using these SNPs appear to be robust when
538 compared to results when 13,291 SNPs were used. From an available 48,776
539 genotyped individuals, we also developed a reference panel of 806 individuals sampled
540 from 17 breeds to have ≤50 individuals per breed that had ≥85% assignment to their
541 breed of registration. This panel appears to allow the robust estimation of the ancestry
542 of advanced generation admixed animals, however, all breeds share some common
543 ancestry which predates the recent development of breed association herdbooks and
544 also due to the complex history of development of cattle breeds [4, 5]. CRUMBLER
545 pipeline scripts and reference panel breed SNP weights are available on GitHub
546 (https://github.com/tamarcrum/CRUMBLER).

Additional files

Supplementary Information (PDF). This file contains the source code changes in SMARTPCA within versions of EIGENSOFT beyond 5.0.2 to enable compatibility with SNPweights.

Supplementary Methods (PDF). This file describes the preliminary fastSTRUCTURE analyses conducted on subsamples of breeds in the development of the reference breed panel.

Supplementary Figures (PDF). This file contains Supplementary Figures S1-S17.

Fig. S1 An overview of the processes and iterations of filtering conducted in the development of the reference panel.

Fig. S2 Preliminary FastSTRUCTURE analysis of candidate Angus and Simmental reference population animals.

Fig. S3 Preliminary fastSTRUCTURE analysis of candidate Angus and Gelbvieh reference population animals.

Fig. S4 Preliminary fastSTRUCTURE analysis of candidate Angus and Limousin reference population animals.

Fig. S5 Preliminary fastSTRUCTURE analysis of candidate Angus and Red Angus reference population animals.

Fig. S6 Preliminary fastSTRUCTURE analysis of candidate Red Angus, Hereford, Shorthorn and Salers reference population animals.

Fig. S7 Preliminary fastSTRUCTURE analysis of candidate Red Angus, Hereford and Shorthorn reference population animals.

Fig. S8 Preliminary fastSTRUCTURE analysis of candidate N'Dama, Nelore and Brahman reference population animals.

Fig. S9 SNPweights self-assignment analysis for the reference sample set containing ≤200 individuals per breed analyzed using the BC7K marker set.

Fig. S10 SNPweights self-assignment analysis for the reference sample set containing ≤150 individuals per breed analyzed using the BC7K marker set.

Fig. S11 SNPweights self-assignment analysis for the reference sample set containing ≤50 individuals per breed analyzed using the BC7K marker set.

Fig. S12 SNPweights self-assignment analysis for the reference sample sets containing ≤50 individuals per breed analyzed using the BC13K marker set.

Fig. S13 SNPweights self-assignment analysis for the reference sample set with ≥80% ancestry to breed of registry and ≤50 individuals per breed using the BC7K marker set.

Fig. S14 SNPweights self-assignment analysis for reference sample set with ≥75% ancestry to breed of registry and ≤50 individuals per breed using the BC7K marker set.

Fig. S15 SNPweights self-assignment analysis for the reference sample set with ≥70% ancestry to breed of registry and ≤50 individuals per breed using the BC7K marker set.

Fig. S16 Analysis of animals identified as fullblood based on pedigree for the open herdbook breeds. (a) SNPweights analysis of fullblood Shorthorn (n=178), Gelbvieh (n=48), Braunvieh (n=69), Simmental (n=337), Charolais (n=1482), and Limousin (n=294). (b) Distribution of ancestry assignment to the respective breed by the SNPweights analysis for the registered fullblood animals.

Fig. S17 SNPweights self-assignment analyses using a reference panel with ≤50 individuals per breed and sampling from the individuals with ≥85% assignment to their breed of registry but with (a) Red Angus or (b) Angus excluded from the reference panel.

Authors' contributions

TC conceived the study and managed the project. TC, RS, JD, and JT contributed to defining the research questions and analytical approaches and interpretation of the results. TC programmed the CRUMBLER pipeline and carried out the data analyses. TC and JT drafted the manuscript. LR provided the Nelore samples but did not have involvement in the scientific direction. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Project Name: CRUMBLER

Project Home Page: https://github.com/tamarcrum/CRUMBLER

Programming Language: Python

Other Requirements: PLINK, EIGENSOFT, and SNPweights

License: GNU GPL

636 **Ethics approval and consent to participate**

637 Not applicable.

638

639 **Funding**

640 JT and RS appreciate the support of NIH-USDA Dual Purpose with Dual Benefit
641 Program grant number NIH 1R01HD084353. JT and RS are supported by USDA-NIFA
642 grants 2013-68004-20364, 2015-67015-23183, 2016-67015-24923 and 2017-67015-
643 26760.

644

645 **Consent for Publication**

646 Not applicable.

647

648 **Acknowledgements**

649 Not applicable.

650

References

[1.] Liu Y, Nyunoya T, Leng S, Belinsky SA, Tesfaigzi Y, Bruse S. Softwares and methods for estimating genetic ancestry in human populations. Hum Genomics. 2013;7:1.

[2.] Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.

[3.] Tishkoff SA, Kidd KK. Implications of biogeography of human populations for'race'and medicine. Nat Genet. 2004. https://www.nature.com/ng/journal/v36/n11s/full/ng1438.html.

[4.] Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, Chen K, et al. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proc Natl Acad Sci U S A. 2009;106:18644–9.

[5.] Decker JE, McKay SD, Rolf MM, Kim J, Molina Alcalá A, Sonstegard TS, et al. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet. 2014;10:e1004254.

[6.] Burnett MS, Strain KJ, Lesnick TG, de Andrade M, Rocca WA, Maraganore DM. Reliability of self-reported ancestry among siblings: implications for genetic association studies. Am J Epidemiol. 2006;163:486–92.

[7.] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.

671 [8.] Rolf MM, Taylor JF, Schnabel RD, McKay SD, McClure MC, Northcutt SL, et al.
672 Impact of reduced marker set estimation of genomic relationship matrices on genomic
673 selection for feed efficiency in Angus cattle. BMC Genet. 2010;11:24.

674 [9.] Saatchi M, Beever JE, Decker JE, Faulkner DB, Freetly HC, Hansen SL, et al. QTLs
675 associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency
676 have little overlap across 4 beef cattle studies. BMC Genomics. 2014;15:1004.

677 [10.] Seabury CM, Oldeschulte DL, Saatchi M, Beever JE, Decker JE, Halley YA, et al.
678 Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle.
679 BMC Genomics. 2017;18:386.

680 [11.] Neibergs HL, Seabury CM, Wojtowicz AJ, Wang Z, Scraggs E, Kiser JN, et al.
681 Susceptibility loci revealed for bovine respiratory disease complex in pre-weaned
682 holstein calves. BMC Genomics. 2014;15:1164.

683 [12.] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al.
684 PLINK: a tool set for whole-genome association and population-based linkage analyses.
685 Am J Hum Genet. 2007;81:559–75.

686 [13.] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-
687 generation PLINK: rising to the challenge of larger and richer datasets. Gigascience.
688 2015;4:7.

689 [14.] Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS
690 Genet. 2006;2:e190.

691 [15.] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
692 components analysis corrects for stratification in genome-wide association studies. Nat
693 Genet. 2006;38:904–9.

694 [16.] Chen C-Y, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved
695 ancestry inference using weights from external reference panels. Bioinformatics.
696 2013;29:1399–406.

697 [17.] Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational Inference of
698 Population Structure in Large SNP Data Sets. Genetics. 2014;197:573–89.

699 [18.] Sanders JO. History and Development of Zebu Cattle in the United States. J Anim
700 Sci. 1980;50:1188–200.

701 [19.] Leesburg VLR, MacNeil MD, Neser FWC. Influence of Miles City Line 1 on the
702 United States Hereford population. J Anim Sci. 2014;92:2387–94.

703 [20.] McCann LP. battle of bull runts. 1974. http://agris.fao.org/agris-
704 search/search.do?recordID=US201300539215.

705 [21.] Wiedemar N, Tetens J, Jagannathan V, Menoud A, Neuenschwander S,
706 Bruggmann R, et al. Independent polled mutations leading to complex gene expression
707 differences in cattle. PLoS One. 2014;9:e93435.

708 [22.] Whitacre L. Structural variation at the KIT locus is responsible for the piebald
709 phenotype in Hereford and Simmental cattle. 2014.
710 http://search.proquest.com/openview/45eba5fa3c5757a2c4c2ab18af1a8a98/1?pq-

origsite=gscholar&cbl=18750&diss=y.

[23.] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.

715 **Table 1. Genotype data for 48,776 registered individuals from 20 breeds were used to**
716 **establish the reference population.**
717

Breed	No. Registered Individuals	No. FullBlood Individuals[a]	No. Individuals Assigned to Breed[b]	Sampled Individuals[c]	No. Individual After Pedigree and SNPweight
Angus	5552	5552	485	200	200
Hereford	969	969	348	200	200
Limousin	2734	321	367	200	200
Charolais	1542	1489	1542	200	200
Simmental	15858	337	1583	200	196
Japanese Black	97	97	97	97	94
Braunvieh	148	69	148	148	69
Gelbvieh	12835	51	6000	200	51
Romagnola	37	37	37	37	37
Salers	68	68	0	0	0
Texas Longhorn	45	45	45	0	0
Shorthorn	291	178	166	166	178
Red Angus	1377	1377	124	124	124
Holstein	5816	5816	5816	200	197
Jersey	119	119	119	119	118
Brown Swiss	92	92	92	92	90
Guernsey	30	30	30	30	30
N'Dama	98	98	59	59	59
Brahman	127	127	86	86	50
Nelore	941	941	708	200	50
Total	48776	17813	17852	2558	2143

718 [a]Number of registered animals determined by pedigree analysis to be fullblood for breed associations with
719 open herdbooks.

[b]Number of registered animals assigned to their identified breed with P≥0.97 by fastSTRUCTURE in preliminary analyses and retained for subsequent analyses.

[c]A random sample of 200 individuals was obtained for breeds with >200 individuals after fastSTRUCTURE analysis and all individuals were sampled for breeds with ≤200 per breed and the data were again analyzed by fastSTRUCTURE with K=19 after removal of the Salers.

[d]Animals that were determined to not be fullblood by pedigree analysis and animals assigned with P≤0.60 by SNPweights to their breed of registry were removed.

Table 2. The number of variants queried by each assay and the number of individuals from the 20 reference breeds genotyped using each assay.

Assay	No. of Variants	No. of Registered Individuals
BovineSNP50	58336	20485
BovineHD	777962	2303
GGP-F250	227234	3068
GGP-90KT	76999	4407
GGP-LDV3	26504	6065
GGP-HDV3	139977	3630
GGP-LDV4	30105	8653
GGP-LDV1	8762	165
Zoetis i50K	59825	0
ICBF IDBv3	53450	0
BOVGv1	47843	0
Total		48776

778 **Table 3. Number of individuals for each reference breed assigned to their breed of**
779 **registration by minimum ancestry threshold.**
780

	Breed Assignment Probability				
Breed	**≥90%**	**≥85%**	**≥80%**	**≥75%**	**≥70%**
Angus	51	136	184	199	200
Hereford	58	136	184	200	200
Limousin	93	127	144	162	173
Charolais	52	92	119	132	147
Simmental	21	43	81	103	121
Japanese Black	52	73	78	83	86
Braunvieh	37	57	63	65	68
Gelbvieh	23	31	39	43	43
Romagnola	10	25	32	36	37
Shorthorn	34	98	159	170	177
Red Angus	48	88	110	120	123
Holstein	39	119	172	193	196
Jersey	52	77	91	108	116
Brown Swiss	38	64	73	82	86
Guernsey	12	22	29	30	30
N'Dama	27	45	59	59	59
Brahman	15	40	50	50	50
Nelore	32	50	50	50	50
Total	694	1323	1717	1885	1962

781
782

783 **Table 4. Ancestry proportion statistics for the self-assignment of reference panel members**
784 **from samples of ≤50 or ≤100 individuals from the candidate reference breed individuals.**
785

Breed	Min % (≤50)	Avg % (≤50)	Max % (≤50)	Min % (≤100)	Avg % (≤100)	Max % (≤100)
Angus	86.22	90.40	95.54	78.49	87.05	94.13
Hereford	79.75	90.08	95.05	73.41	87.39	96.81
Limousin	69.52	88.53	98.16	18.36	86.40	98.81
Charolais	78.14	90.19	99.82	48.93	77.46	93.96
Simmental	81.06	90.37	97.66	61.36	73.05	88.11
Japanese Black	81.44	90.00	97.07	24.51	86.50	98.95
Braunvieh	71.59	89.46	98.61	65.46	88.36	98.70
Gelbvieh	73.03	76.27	81.63	60.92	74.59	80.33
Romagnola	75.05	87.18	96.66	74.79	85.99	95.12
Shorthorn	84.42	88.69	94.54	70.71	85.27	96.35
Red Angus	79.00	89.60	96.33	68.07	86.83	97.38
Holstein	85.82	90.30	97.51	62.95	86.97	97.81
Jersey	78.55	89.28	95.93	61.23	86.54	97.18
Brown Swiss	80.10	89.22	96.40	61.68	86.02	98.42
Guernsey	79.53	89.19	95.85	77.40	88.31	94.36
N'Dama	80.67	89.25	96.90	78.91	87.78	95.67
Bos taurus indicus	87.83	91.91	97.75	81.43	89.79	97.60

786
787
788
789
790
791

Table 5. Average predicted ancestry and variance in predicted ancestry for candidate reference breed individuals when filtered on minimum predicted ancestry.

Breed	Avg % (70%)	Var (70%)	Avg % (75%)	Var (75%)	Avg % (80%)	Var (80%)	Avg % (85%)	Var (85%)	Avg % (90%)	Var (90%)
Angus	86.50	0.21	87.95	0.19	87.33	0.22	88.86	0.13	72.34	0.97
Hereford	86.99	0.22	87.09	0.23	87.48	0.19	88.25	0.13	84.62	0.43
Limousin	86.77	0.55	89.03	0.44	87.92	0.38	88.48	0.43	80.62	1.19
Charolais	80.18	2.16	85.03	1.77	86.28	0.99	88.56	0.52	81.54	0.76
Simmental	72.73	0.89	78.45	0.58	83.81	0.36	89.65	0.15	87.82	0.50
Japanese Black	87.85	0.52	88.04	0.39	88.46	0.27	88.74	0.21	80.06	0.61
Braunvieh	87.01	0.37	87.84	0.36	87.33	0.38	88.71	0.21	80.47	1.24
Gelbvieh	86.68	0.41	87.10	0.43	87.52	0.34	88.43	0.34	83.31	1.25
Romagnola	86.16	0.33	86.37	0.32	87.16	0.32	86.22	0.29	86.38	1.16
Shorthorn	85.97	0.26	87.03	0.22	86.80	0.14	87.38	0.07	83.00	0.70
Red Angus	86.41	0.53	87.08	0.48	87.40	0.35	87.46	0.23	23.37	0.66
Holstein	86.44	0.27	87.82	0.21	87.54	0.13	88.77	0.12	79.71	0.61
Jersey	87.01	0.46	86.93	0.44	87.86	0.24	87.98	0.27	80.52	0.71
Brown Swiss	86.22	0.47	86.73	0.51	88.24	0.26	88.11	0.20	82.23	0.70
Guernsey	86.46	0.23	87.64	0.19	87.50	0.25	88.02	0.51	80.43	2.36
N'Dama	87.76	0.19	87.91	0.21	87.89	0.15	89.25	0.17	86.40	0.52
Bos taurus indicus	87.68	0.07	88.24	0.09	87.55	0.11	88.53	0.09	84.89	0.38
Average	85.58	0.48	86.84	0.41	87.30	0.30	88.32	0.24	78.69	0.87

802 **Table 6. Average breed ancestry percentages assigned to 27 Beefmaster individuals.**
803

Breed	**Avg Ancestry %**
Angus	3.17
Hereford	16.54
Limousin	1.76
Charolais	7.32
Simmental	2.73
Japanese Black	1.06
Braunvieh	0.73
Gelbvieh	3.47
Romagnola	1.95
Shorthorn	19.45
Red Angus	2.60
Holstein	3.20
Jersey	1.03
Brown Swiss	1.43
Guernsey	2.67
N'Dama	0.34
Bos taurus indicus	30.54

804

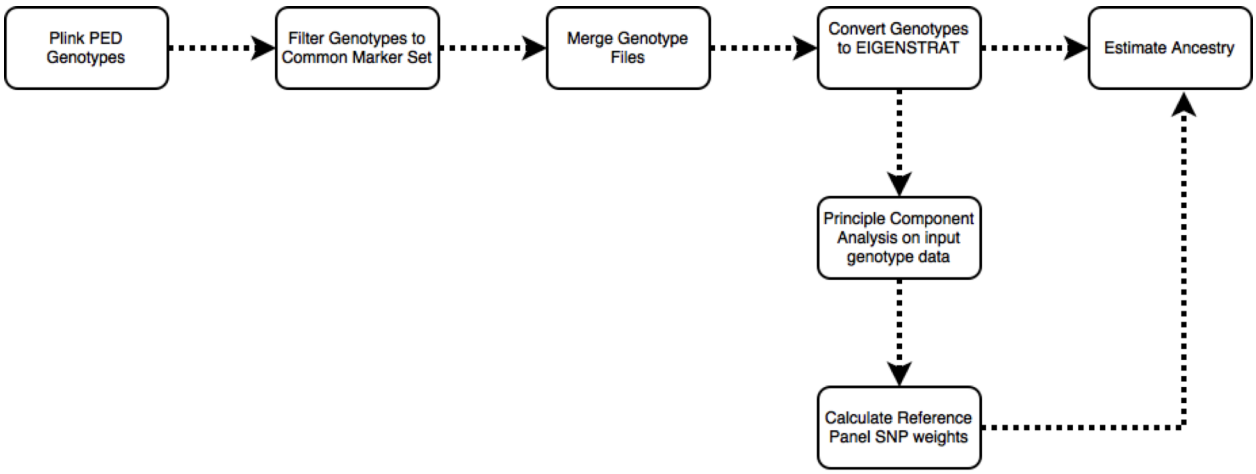


Fig. 1 Flow diagram of the breed composition pipeline.

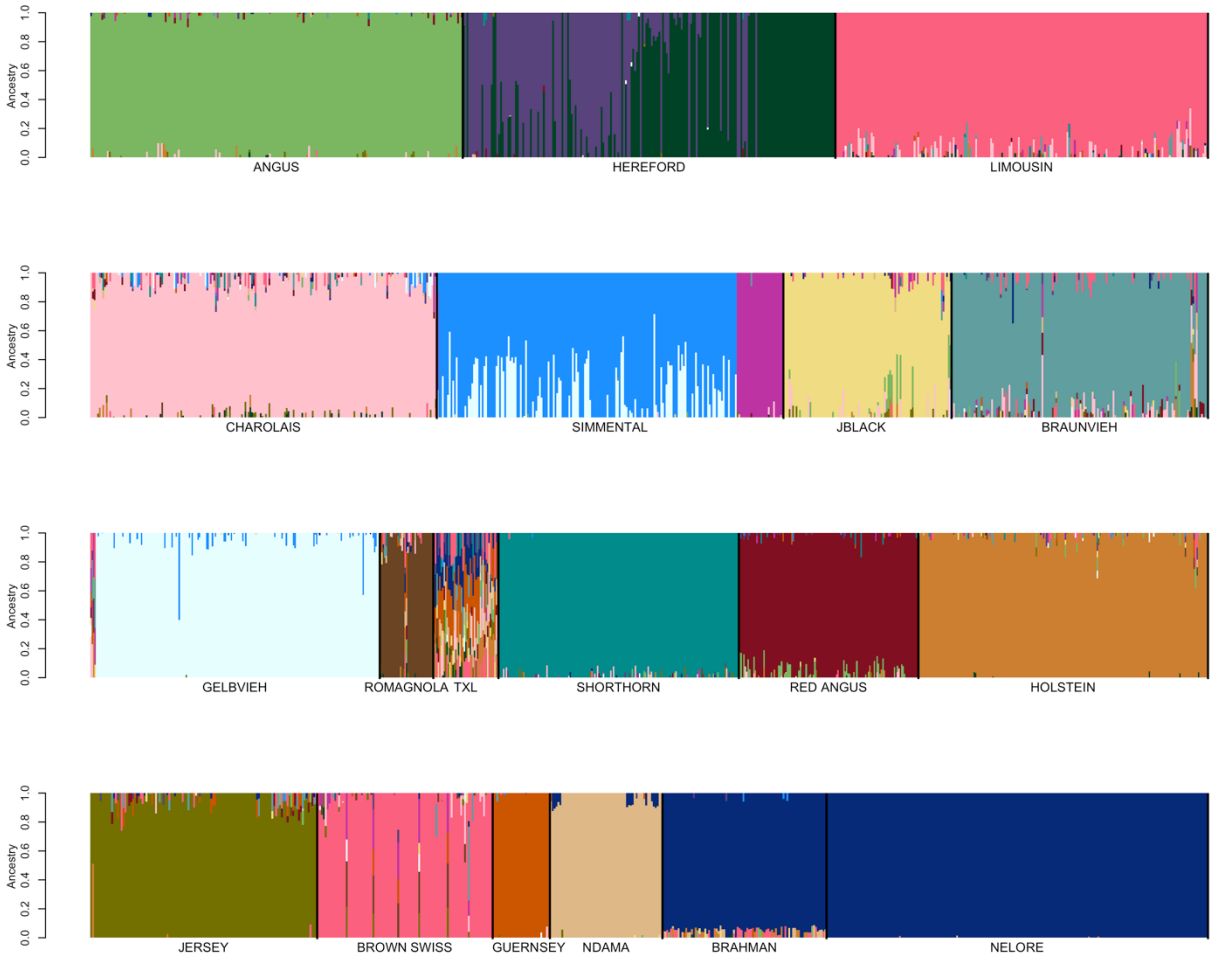
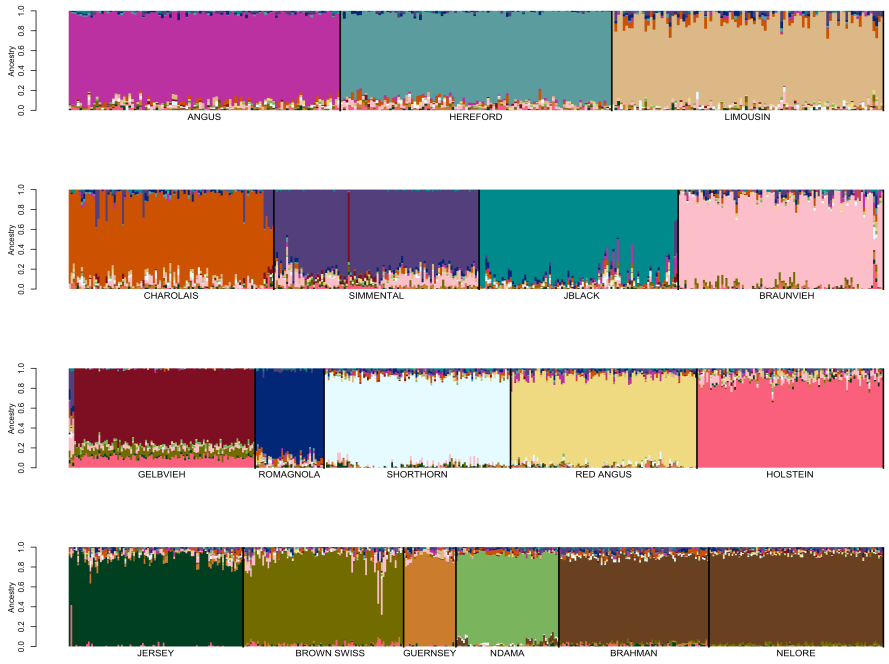
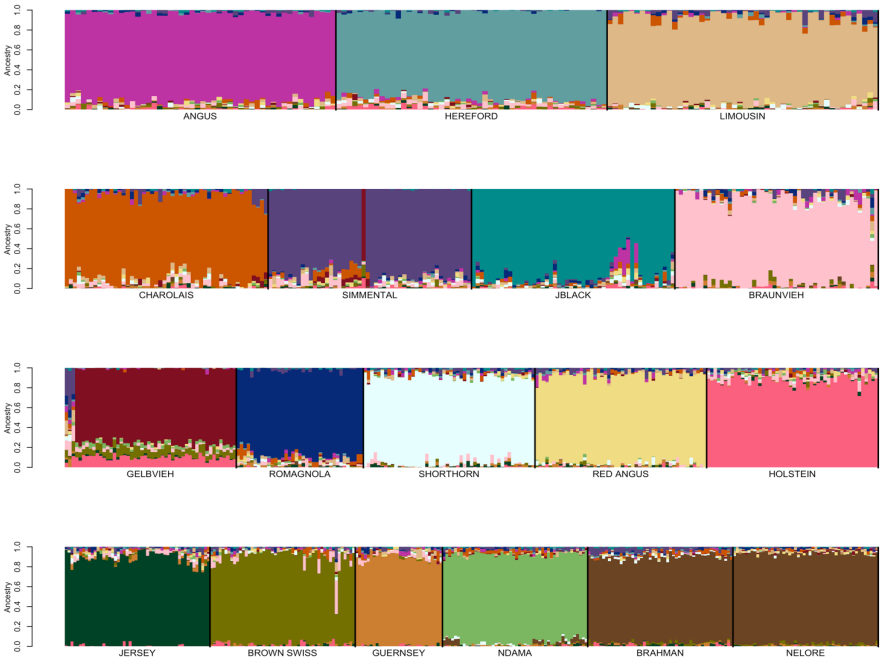


Fig. 2 FastSTRUCTURE results for a random sample of ≤200 individuals per breed from the pool of 17,852 potential reference individuals at K=19. Breed identification is shown below each colored block and each animal is represented as a vertical line within the block.

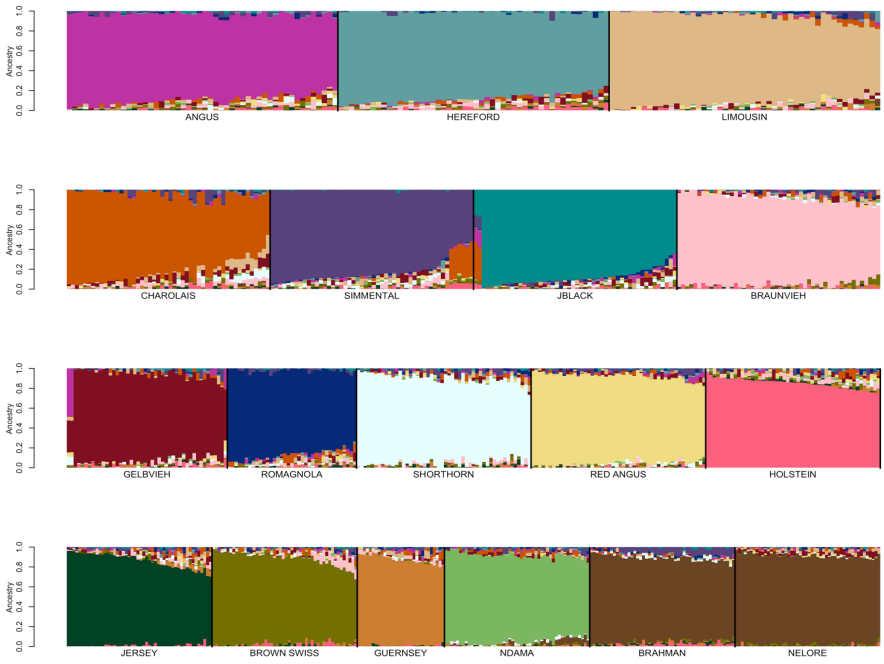


a.

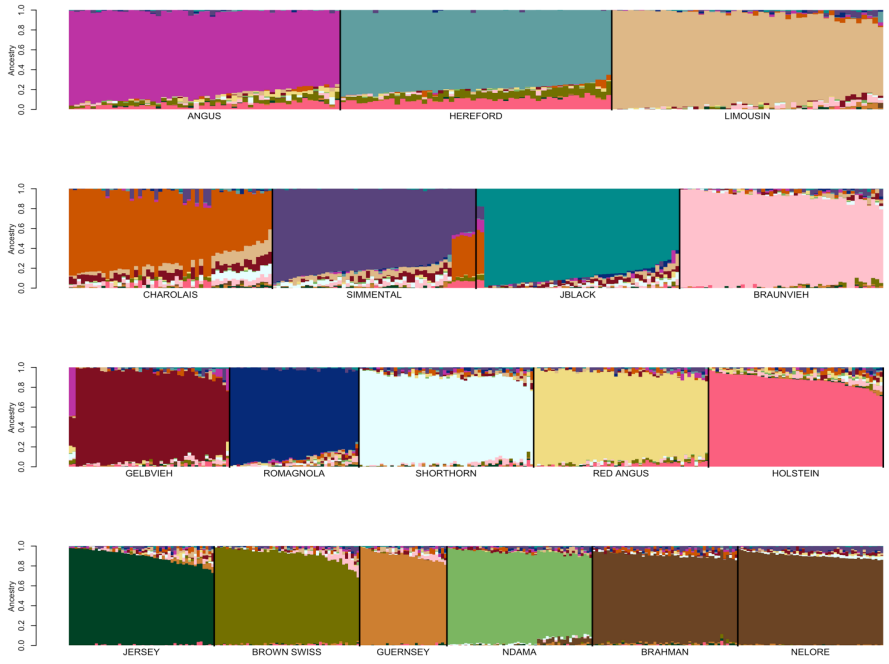


b.

Fig. 3 SNPweights self-assignment analysis results for reference panel sample sets consisting of: (a) ≤100 individuals per breed, or (b) ≤50 individuals per breed. Seven individuals were filtered for ≤60% ancestry to their breed of registry (Holstein n=3, Jersey n=1, Japanese Black n=3).



a.



b.

Fig. 4 SNPweights self-assignment of ancestry for candidate reference breed individuals following evaluation of open herdbook breeds using: (a) the BC7K, or (b) the BC13K marker panels. Reference breed panels were constructed by random sampling ≤50 individuals per breed and SNP weights were estimated using the BC7K and BC13K marker sets.

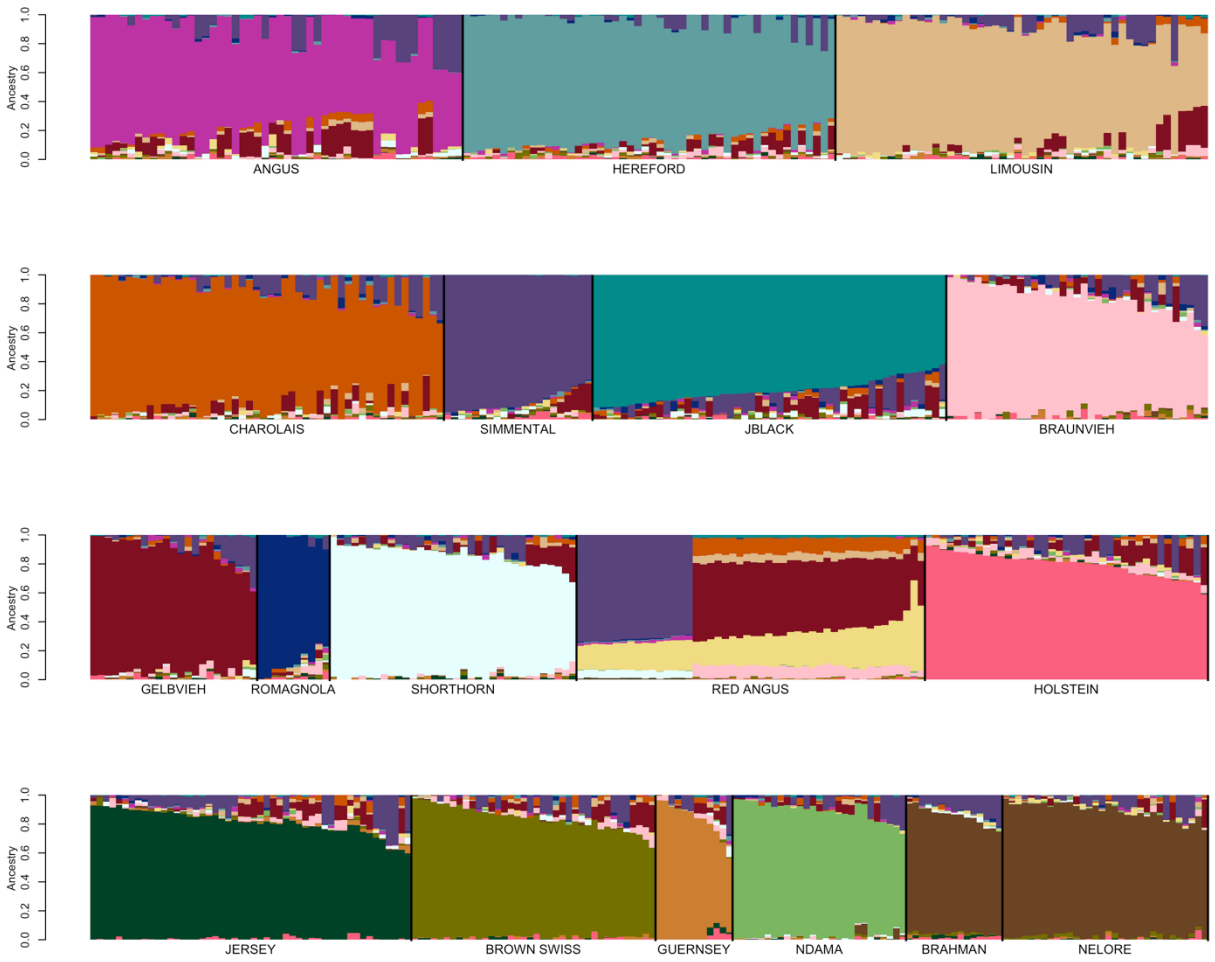


Fig. 5 Reference breed panel constructed by the random sampling of ≤50 individuals per breed from individuals with ≥90% ancestry was self-assigned to reference breed ancestry using the BC7K marker set.

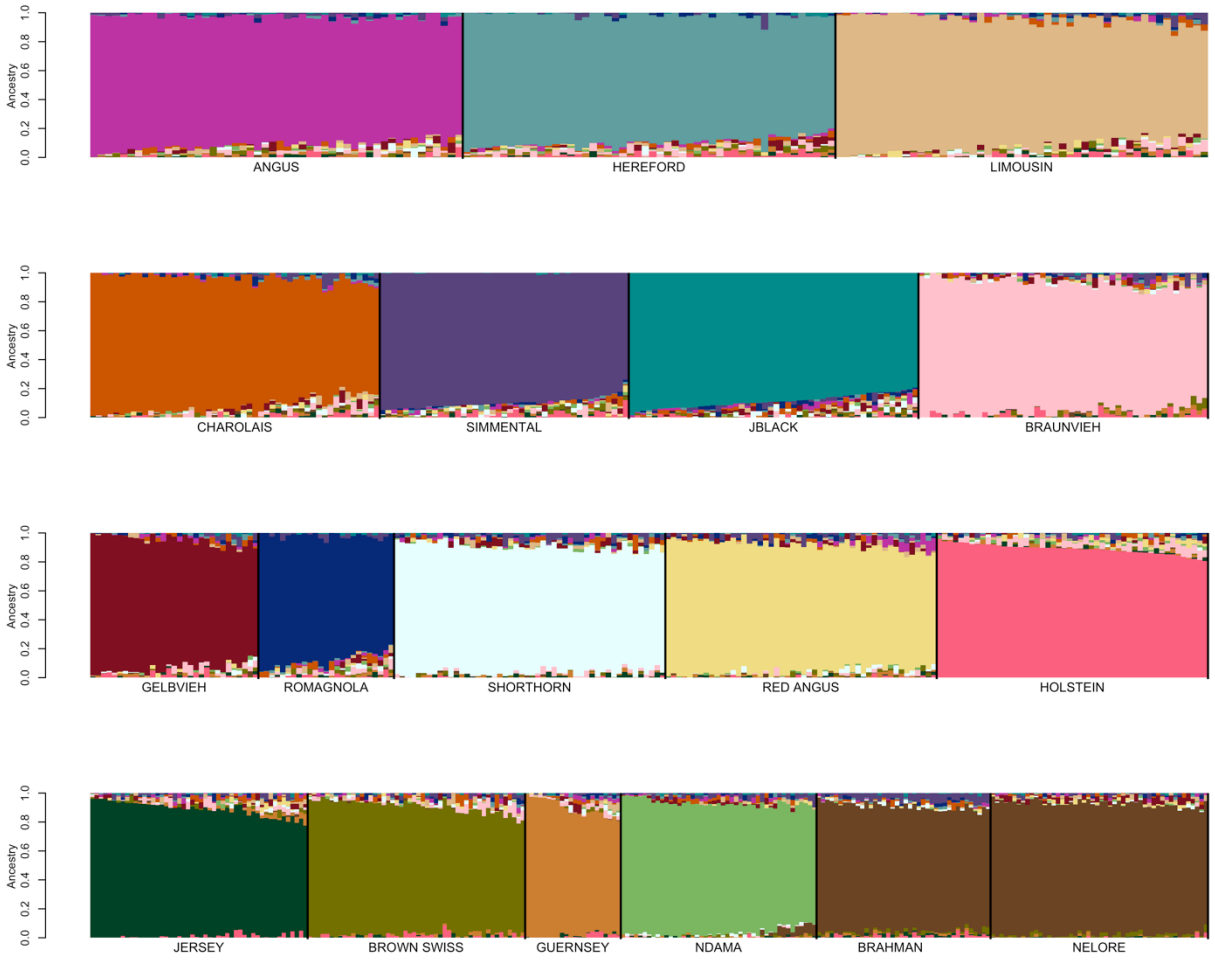


Fig. 6 Reference breed panel constructed by the random sampling of ≤50 individuals per breed from individuals with ≥85% ancestry was self-assigned to reference breed ancestry using the BC7K marker set.

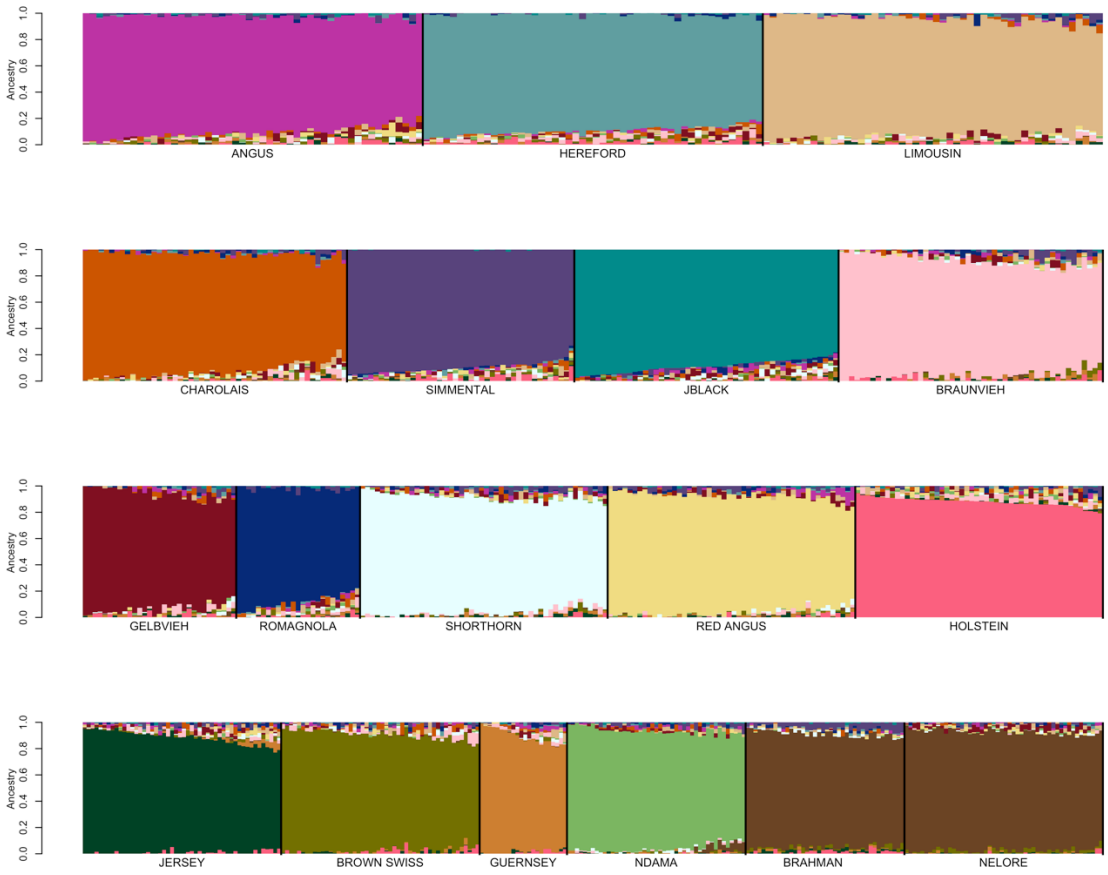


Fig. 7 Reference breed panel constructed by the independent random sampling of a second sample of ≤50 individuals per breed from individuals with ≥85% ancestry after eliminating individuals represented in the first sample was self-assigned to reference breed ancestry using the BC7K marker set.

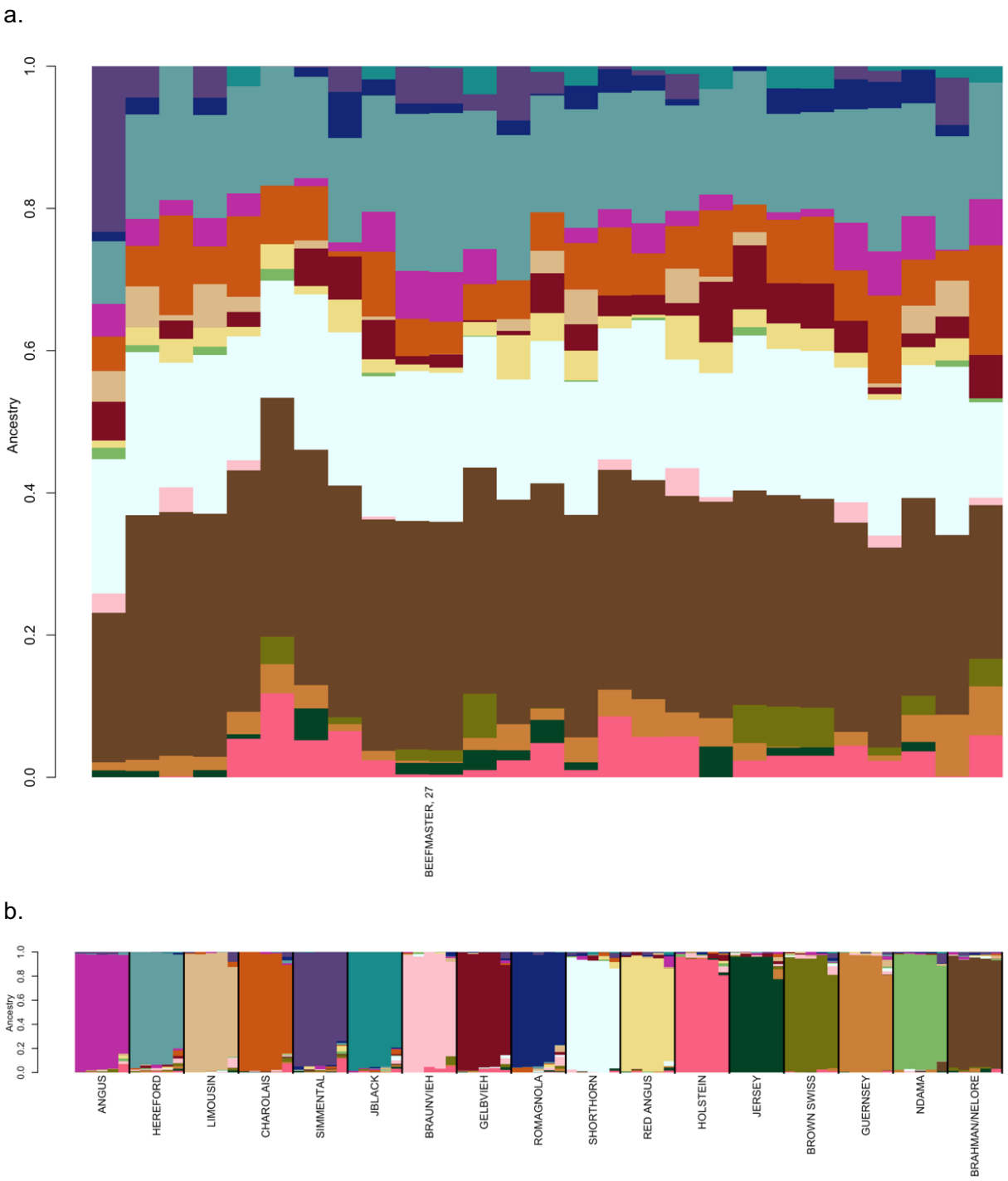
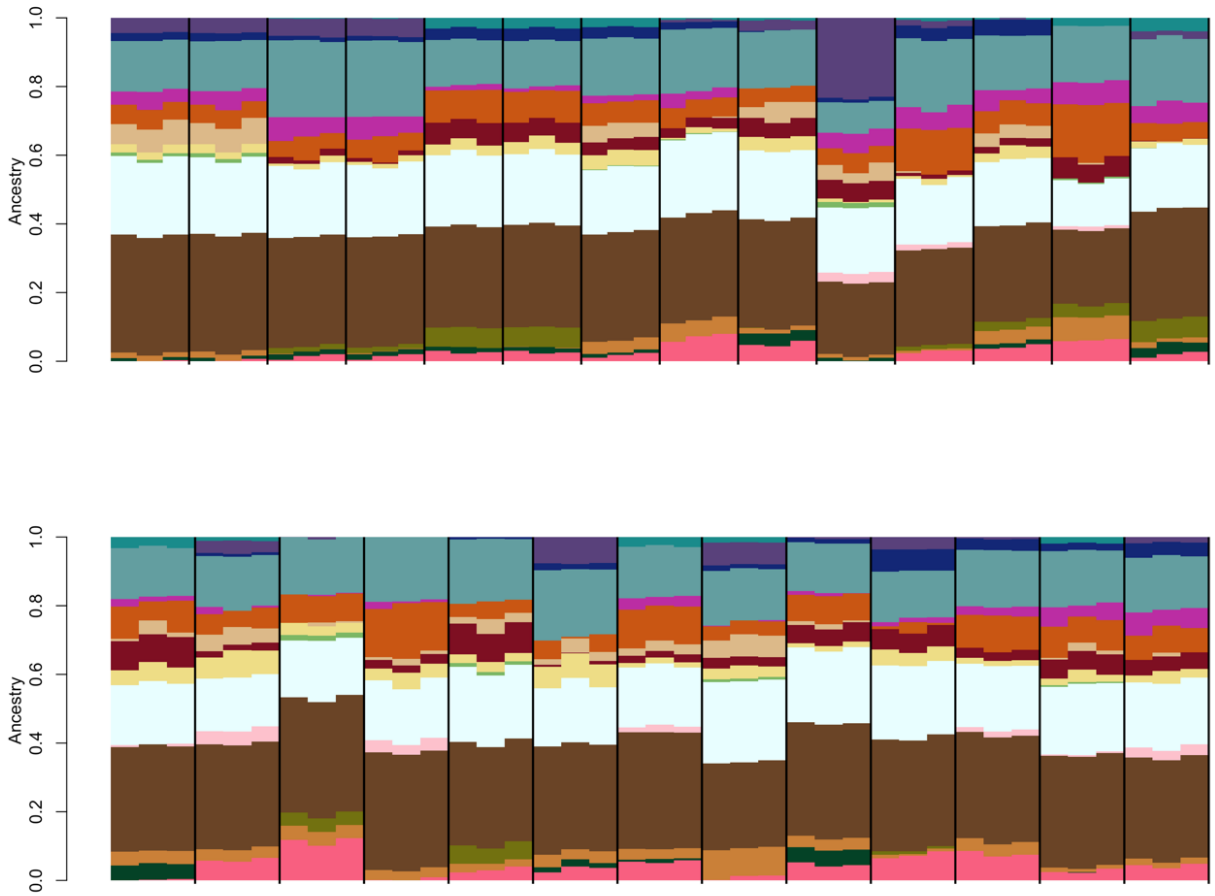


Fig. 8 (a) SNPweights ancestry assignment results for 27 Beefmaster individuals based on a reference panel sampled to contain ≤50 individuals per breed with ancestry self-assignment percentages of ≥85% to their breed of registry. (b) Breed assignment for the Beefmaster individuals can be determined using this reference breed key.

a.



b.

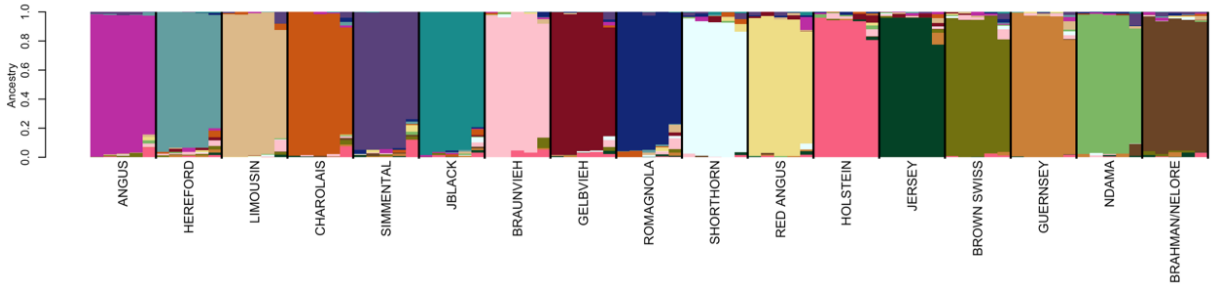
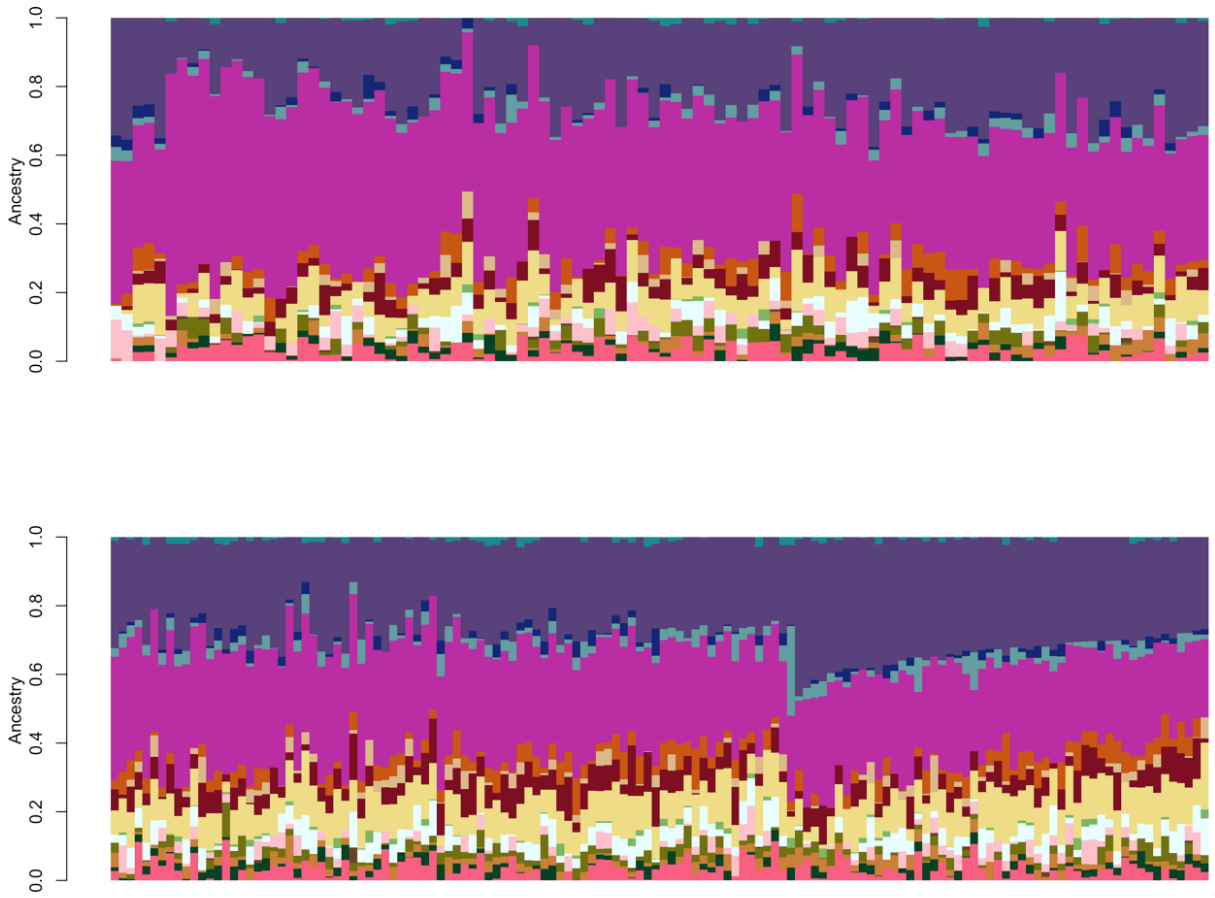


Fig. 9 (a) SNP weights were calculated using three independent reference panels with ≤50 individuals per breed sampled from the individuals with ≥85% assignment to their breed of registry. Ancestry results for the 27 Beefmaster animals are shown as 27 sets of three columns demarcated by solid lines. Each column within a demarcated set, represents results for each of the three reference breed panels. (b) Breed assignment for the Beefmaster individuals can be determined using this reference breed key.

a.



b.

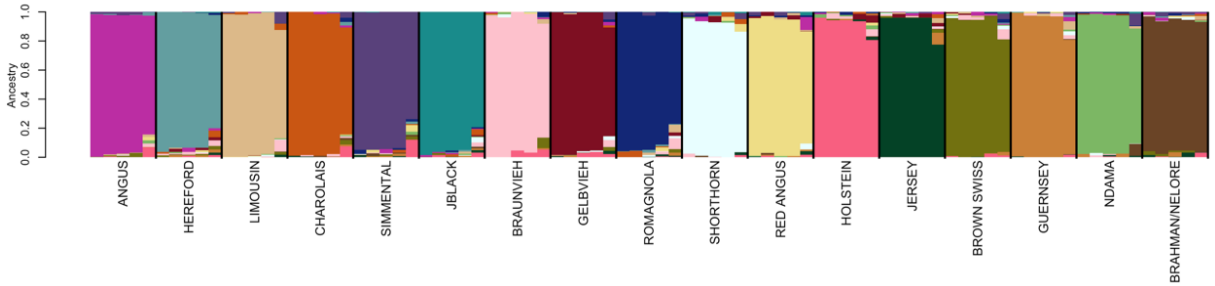


Fig. 10 (a) SNPweights ancestry results for 238 crossbred individuals with *a-priori* breed composition estimates of 50% Angus and 50% Simmental based on a reference panel with ≤50 individuals per breed sampled from individuals with ≥85% assignment to their breed of registry. (b) Breed assignment for the crossbred individuals can be determined using this reference breed key.

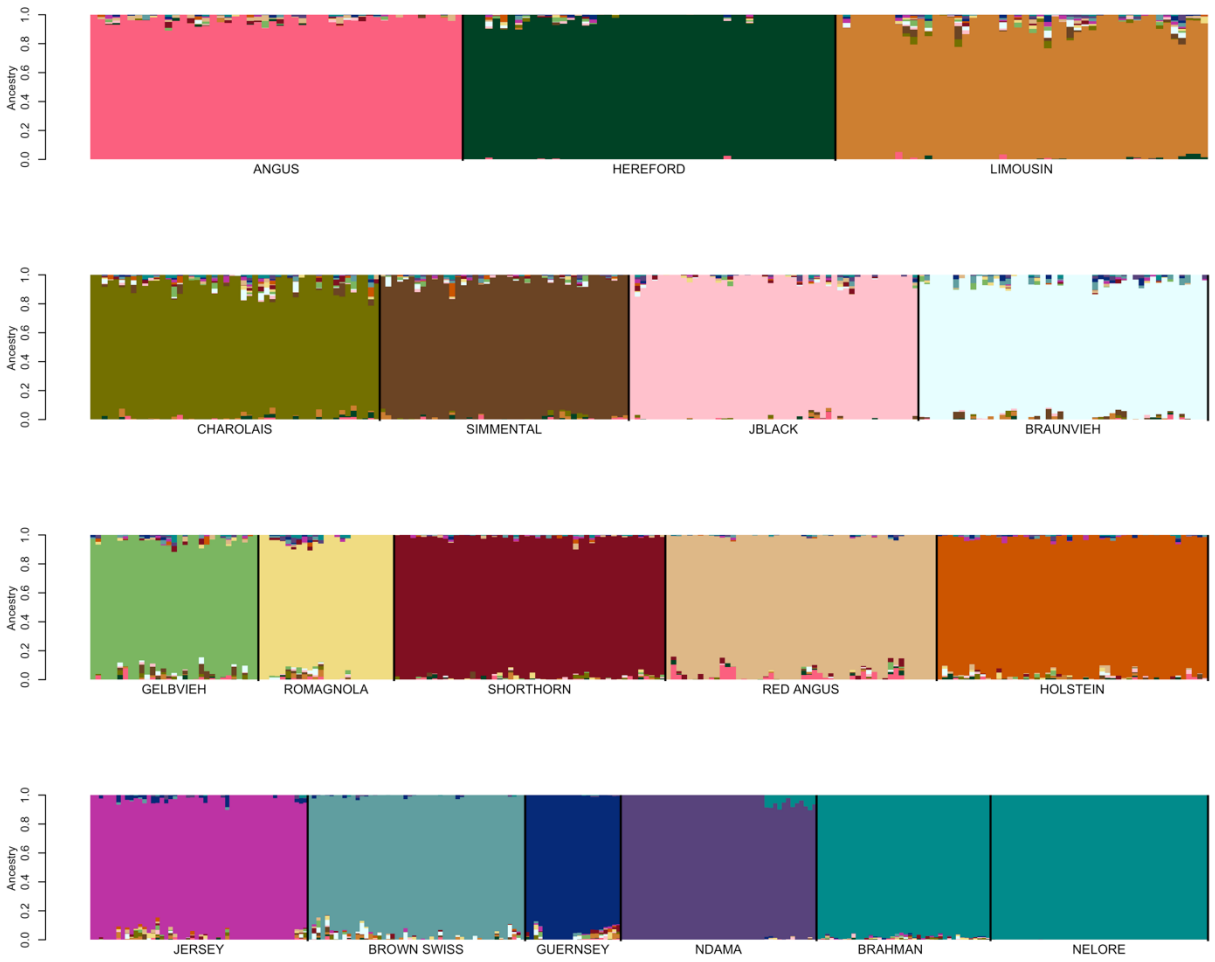


Fig. 11 Self-assignment of ancestry for the animals in the reference breed set formed with ≤50 individuals per breed from the individuals that had ≥85% assignment to their breed of registration using Admixture.

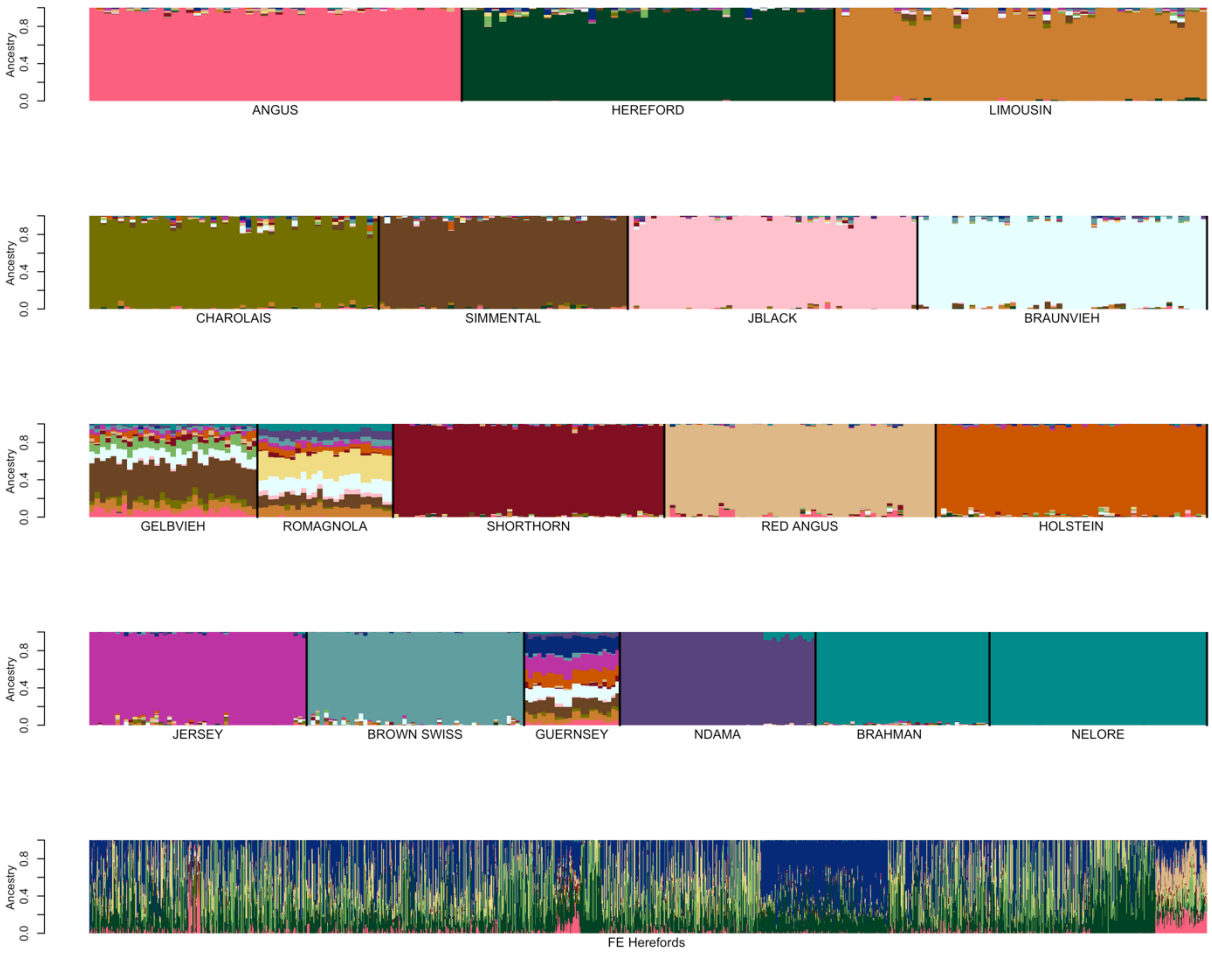


Fig. 12 Admixture analysis conducted using the same data as shown in Figure 11 (first four rows), merged with an additional 2,005 high percentage crossbred Hereford target individuals (last row).

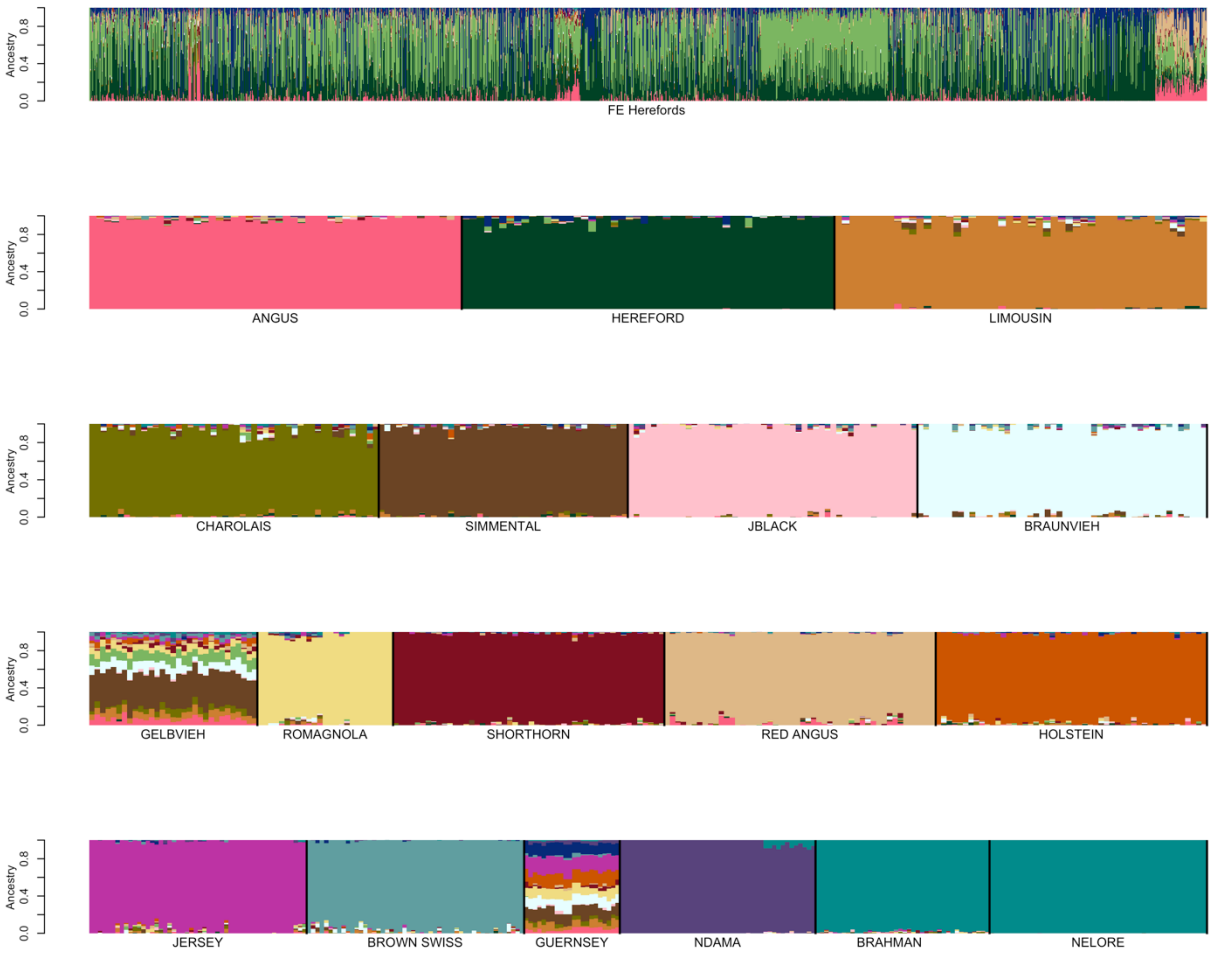


Fig. 13 Admixture analysis conducted using the same data as shown in Fig. 12 where the reference individuals appear before the 2,005 Hereford crossbred individuals in the input file. Here, the 2,005 Hereford crossbred individuals appear before the reference individuals in the input file. The first row represents the 2005 Hereford crossbred samples. Rows 2 to 5 show the reference panel individuals.

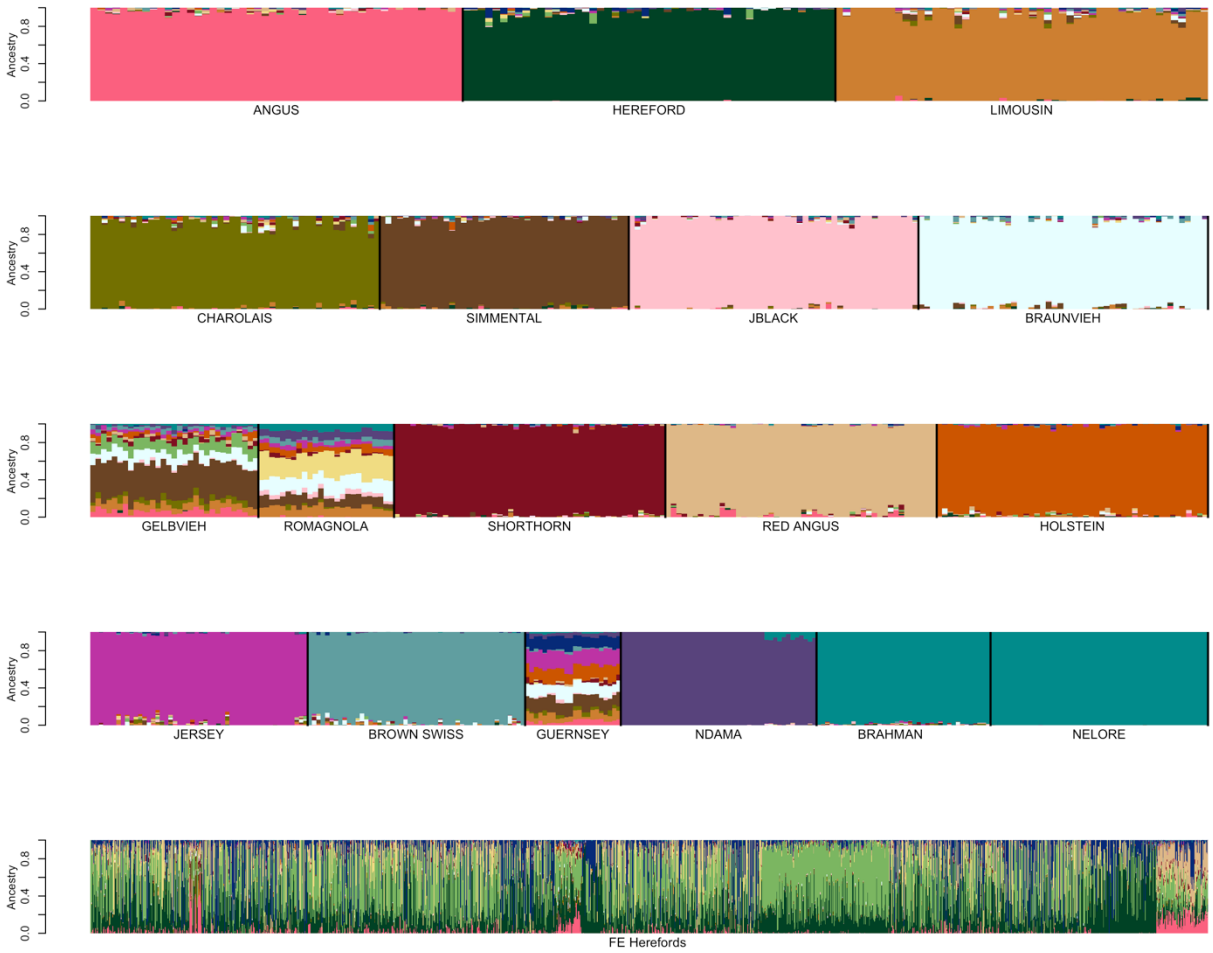


Fig. 14 Admixture analysis conducted using the same data as shown in Figs. 12 and 13, but with the order of the individuals in the input genotype file randomized. The first four rows represent the reference panel individuals, the fifth row shows the 2,005 Hereford crossbred animals.

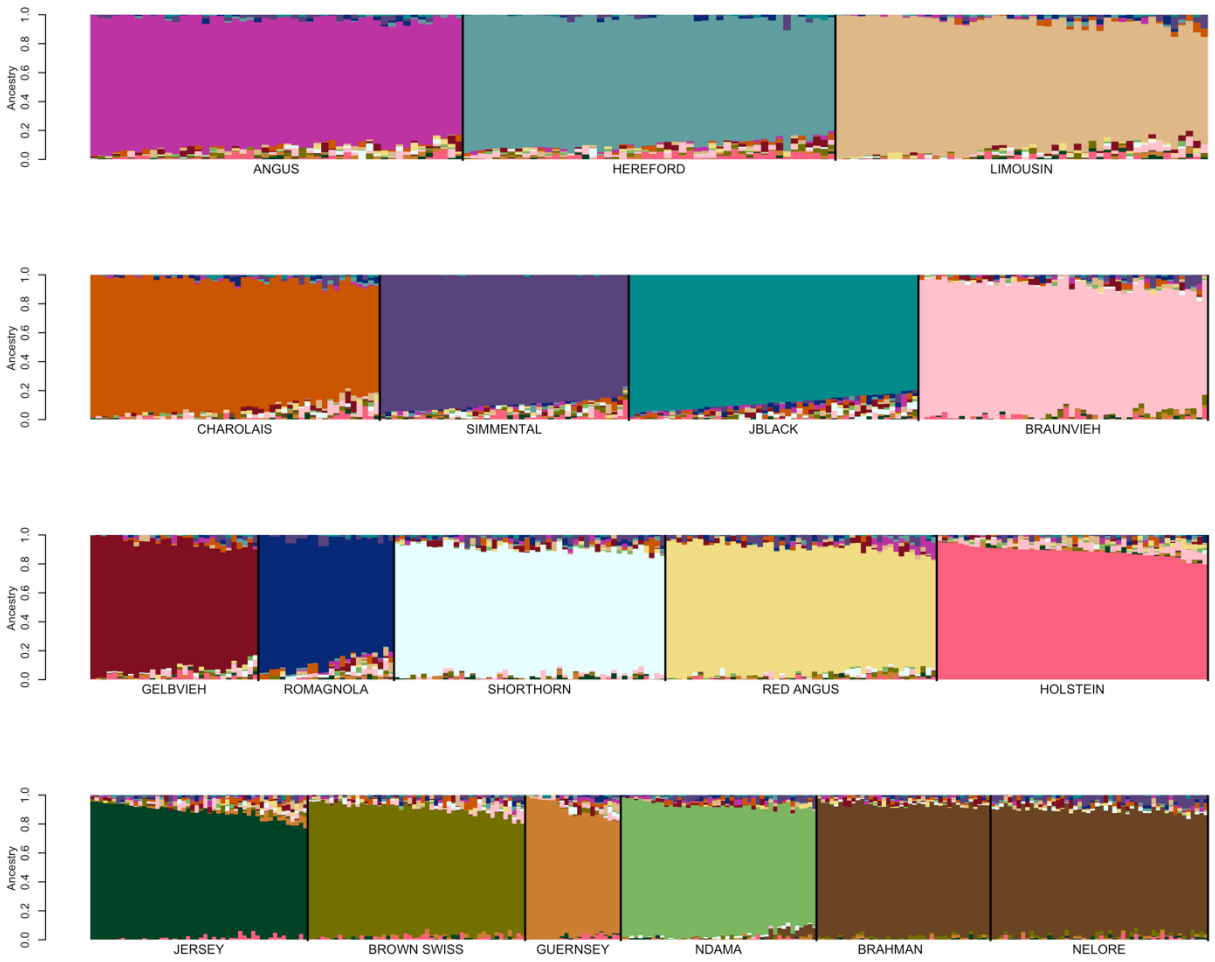


Fig. 15 Reference breed panel constructed by the random sampling of ≤50 individuals per breed from individuals with ≥85% ancestry was self-assigned to reference breed ancestry using the BC6K marker set.