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. As the roles of DNA and genotype data become more pervasive in animal agriculture, a systematic method for estimating the breed composition (the proportions of an animal's genome originating from ancestral pure breeds) has utility for a variety of downstream analyses including the estimation of genomic breeding values for crossbred animals, the estimation of quantitative trait locus effects, and heterosis and heterosis retention in advanced generation composite animals. Currently, there is no automated or semi-automated ancestry estimation platform for cattle and the objective of this study was to evaluate the utility of extant public software for ancestry estimation and determine the effects of reference population size and composition and number of utilized single nucleotide polymorphism loci on ancestry estimation. We also sought to develop an analysis pipeline that would simplify this process for members of the livestock genomics research community.

Results

We developed and tested a tool, "CRUMBLER", to estimate the global ancestry of cattle using ADMIXTURE and SNPweights based on a defined reference panel. CRUMBLER, was developed and evaluated in cattle, but is a species agnostic pipeline

that facilitates the streamlined estimation of breed composition for individuals with potentially complex ancestries using publicly available global ancestry software and a specified reference population SNP dataset. We developed the reference panel from a large cattle genotype data set and breed association pedigree information using iterative analyses to identify purebred individuals that were representative of each breed. We also evaluated the numbers of markers necessary for breed composition estimation and simulated genotypes for advanced generation composite animals to evaluate the precision of the developed tool.

Conclusion

The developed CRUMBLER pipeline extracts a specified subset of genotypes that is common to all current commercially available genotyping platforms, processes these into the file formats required for the analysis software, and predicts admixture proportions using the specified reference population allele frequencies.

Background

Estimation of the breed composition of individuals with complex ancestries has utility for estimating breed direct and heterosis effects as well as for the estimation of the additive genetic merit of these individuals. It also has value for identifying the breed composition of training populations used for genomic selection and hence the identification of target breeds in which the developed prediction equations may have some relevance. Visual classification of cattle based on breed characteristics suffers from similar problems as the self-identification of ethnicity in humans [1], as most breed characteristics are

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 colours 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 [1,2] and the visual classification of black-hided animals for branded beef programs can result in the marketing of animals with vastly different Angus genome content.

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 the previous 50 years, parentage has been validated by each breed association using blood or, more recently, 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. Pedigree errors that occurred prior to, or that were not identified following 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, random assortment of chromosomes into gametes and selection can also lead to considerable variation in the extent of identity by descent between relatives separated by more than a single meiosis and can

also lead to admixture proportions that differ substantially from expectation based on pedigree.

Crossbreeding is extensively used in commercial beef production and in other livestock species production systems to capitalize on the effects of breed complementarity and heterosis resulting in herds of females that may have 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 [3,4] or bovine respiratory disease [5], the presence of extensive admixture in the phenotyped and genotyped animals may impact the GWAA [3,4] 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.

As the number of genotyped beef animals has increased, the need to classify the breed composition of these animals has necessitated the development of a precise and accurate method for estimating breed composition in cattle based on single nucleotide polymorphism (SNP) data. Iterative ancestry estimation analyses performed using

115 different software input parameters may identify those that cause output sensitivity and
116 can lead to an interpretation of population structure that is close to the truth [6]. We
117 developed the CRUMBLER analysis pipeline to streamline the genomic estimation of
118 breed composition of crossbred cattle using high-density SNP genotype data, publicly
119 available software, and a reference panel containing genotypes for members of cattle
120 breeds that are numerically important in North America. The CRUMBLER pipeline is
121 species agnostic and could be adapted for breed composition estimation in other
122 species. CRUMBLER and the reference panel data are available on GitHub
123 (https://github.com/tamarcrum/CRUMBLER). This pipeline tool is released under the
124 GNU General Public License.
125

126 Materials and Methods

127

128 Genotype data

129 From among the numerically most important cattle breeds in North America, in terms of
130 their annual numbers of animal registrations, a list was compiled to define the target
131 breeds for reference panel development. Composite breeds, such as Brangus and
132 Braford, were not included in this list due to lack of available genotype data, but the
133 progenitor Angus, Hereford and Brahman breeds were included. Breeds such as
134 N'Dama, representing African taurine, and Nelore and Brahman, representing *Bos*
135 *taurus indicus* cattle, were included. We also initially included breeds that were likely to
136 be involved in early crossbreeding of cattle in the U.S. (Texas Longhorn).
137

138 From the 170,544 cattle with high-density SNP genotypes stored within the University of
139 Missouri Animal Genomics genotype database, we extracted genotypes for 48,776
140 animals identified as being registered with one of the numerically important U.S. Breed
141 Associations or belonging to other world breeds. Pedigree data were also obtained for
142 these animals from each of the Breed Associations, where available (Table 1). These
143 individuals had been genotyped using at least one of 9 different genotyping platforms
144 currently used internationally to genotype cattle including the GeneSeek (Lincoln, NE)
145 GGP-90KT, GGP-F250, GGP-HDV3, GGP-LDV1, GGP-LDV3, and GGP-LDV4 assays,
146 the Illumina (San Diego, CA) BovineHD and BovineSNP50 assays, and the Zoetis
147 (Kalamazoo, MI) i50K assay. The numbers of variants queried by each assay and the
148 number of individuals genotyped using each platform are shown in Table 2.
149
150 **Marker set determination**
151 To maximize the utility of the developed breed assignment tool, we identified the
152 intersection set of SNP markers located on the bovine assays for which we had
153 available genotype data (Table 2). However, during the process of identifying the
154 animals that would define the breed reference panel, only 16 individuals had been
155 genotyped using the GGP-LDV4 (n=2) and GGP-LDV3 (n=14) assays and no animals
156 had been genotyped using the GGP-LDV1 assay. To retain as many SNP markers as
157 possible for subsequent analysis, we identified the intersection of markers present on
158 the GGP-90KT, GGP-F250, GGP-HDV3, GGP-LDV3, GGP-LDV4, BovineHD,
159 BovineSNP50 and i50K assays. This intersection set included 6,799 SNP markers
160 (BC7K). The intersection of the markers representing 5 assays (GGP-90KT, GGP-
161 F250, GGP-HDV3, BovineHD, and BovineSNP50) was 13,291 markers (BC13K). By

162 removing only the 16 individuals from the breed reference panel that had been
163 genotyped on the GGP-LDV3 and GGP-LDV4 assays, we were able to compare
164 ancestry predictions using two marker set densities (BC13K and BC7K).
165
166 **Pipeline**
167 The developed CRUMBLER pipeline integrates the tools and the computational
168 efficiency of publicly available software, PLINK [7,8], EIGENSOFT [9,10] and
169 SNPweights [11] to generate ancestry estimates (Fig. 1). The pipeline integrates the
170 often cumbersome processes of data reformatting and sequentially processing the data
171 using analytical tools to generate ancestry proportions for targeted individuals based on
172 a curated breed reference panel.
173
174 **PLINK**
175 PLINK PED formatted genotypes are required as input to the pipeline. PLINK
176 (v1.90b3.31) was used for data filtering and formatting. Genotypes can arise from any
177 of the common bovine genotyping platforms (Table 2), provided that a PLINK
178 compatible MAP file is provided for each assay and data produced using only a single
179 genotyping assay is included in each PED file. The pipeline utilizes the PLINK marker
180 filtering tool (--extract) to extract the user-specified marker subset for ancestry analysis.
181 For analyses of animals genotyped on different genotyping platforms, the marker list
182 representing the intersection of the platforms can be provided to extract the markers
183 that are 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 based on genome-wide SNP weights computed using genotype data for an external panel of reference individuals. To obtain SNP weights, the matrix ($\mathbf{g_{ij}}$) of reference panel genotypes for SNP $\mathbf{i=1, \ldots, M}$ and individual $\mathbf{j=1, \ldots, N}$ is normalized by subtracting the mean $\boldsymbol{\mu_i} = N^{-1}\sum_j \mathbf{g_{ij}}$ and dividing by the standard deviation $\mathbf{[p_i(1-p_i)]^{0.5}}$ for each SNP, where $\mathbf{p_i = \mu_i/2}$, 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

207 and corresponding eigenvectors of the kinship matrix [11]. The SNP weights file only
208 needs to be recalculated if the reference panel is changed. EIGENSTRAT formatted
209 target animal genotypes are input into SNPweights, along with the precomputed
210 reference panel SNP weights. The SNP weights are then applied to the target
211 individuals to estimate their ancestry proportions [11].
212
213 **Reference panel development**
214 The definition of a set of reference individuals that define the genotype frequencies at
215 each SNP variant for each reference breed is technically demanding, but vitally
216 important to the process of defining ancestry. This process assumes that selection has
217 not operated to change gene frequencies between target and reference population
218 animals, and that each population is sufficiently large that drift has not impacted allele
219 frequencies. It also assumes that migration between different countries does not
220 influence population allele frequencies when registered animals are imported or
221 exported. FastSTRUCTURE [12] analysis and iterations of animal filtering using
222 SNPweights was performed using the genotypes of candidate reference panel
223 individuals to remove individuals with significant evidence of admixture from the
224 reference breed panel. An overview of the processes and iterations of filtering
225 conducted in the development of this reference panel set is shown in Fig. S1 and Table
226 1.
227
228 **FastSTRUCTURE analysis to identify candidate reference panel individuals**

229 Genotype data for 48,776 individuals produced by one of 8 different genotyping assays
230 were available for fastSTRUCTURE analysis (Table 1) [12]. We initially performed
231 focused fastSTRUCTURE analyses using small numbers of reference breeds including
232 Angus and Simmental, Angus and Gelbvieh, Angus and Limousin, Angus and Red
233 Angus, Red Angus, Hereford, Shorthorn and Salers, Red Angus, Hereford and
234 Shorthorn, and N'Dama, Nelore and Brahman (Figs. S2-S8). Individuals possessing an
235 ancestry assignment of at least 97% to their designated breed were retained for
236 subsequent analysis (see Supplementary Methods and Table 1). Following filtering
237 based on fastSTRUCTURE breed assignment, 17,852 individuals representing 19 of the
238 original breeds remained for further analysis (Supplementary Methods and Figs. S2-
239 S8). All of the Salers animals were removed in this filtering analysis which is consistent
240 with previous work that found that Salers and Limousin were very similar [4]. Variation in
241 reference population sample sizes has been shown to substantially influence the
242 estimation of the number of ancestral populations (K) in ancestry analyses [6,13,14]. To
243 minimize this effect and produce similar sample sizes for each of the reference breeds,
244 we randomly sampled 200 individuals from each reference breed for which at least 200
245 individuals remained after filtering on an ancestry assignment of at least 97%, otherwise
246 all remaining individuals were included for the breed (Table 1). Following
247 fastSTRUCTURE analysis using K=19 after removal of Salers and using the BC7K
248 marker set, Texas Longhorn was also removed from the reference panel breed list due
249 to the inability to distinguish Texas Longhorn as a distinct population (Figure 2).
250 Further, due to the known common ancestry [15] and similarity between Nelore and
251 Brahman (Figure 2), the breeds were combined to represent *Bos taurus indicus*.

252

253 **SNPweights analyses to refine and validate reference panel members**

254 Random sampling of reference breed individuals was performed to create sample sets
255 containing ≤n individuals per breed, for n = 50, 100, 150 and 200 individuals (Figs. 3 a-b
256 and Figs. S9-S10). Sampling was performed such that if a reference breed had ≥n
257 candidates then n individuals were randomly sampled, otherwise, all available
258 individuals were sampled. An analysis was performed using the BC7K marker set,
259 SNPweights was used to assign reference breed ancestries to the same sample of
260 individuals that was used to produce the SNP weights for each of the four samples of
261 individuals (Figs. 3 a-b and Figs. S9-S10). In the self-assignment analyses conducted
262 using the reference breed sample sets of ≤100 individuals per breed and ≤50 individuals
263 per breed, 7 individuals were removed due to their estimated breed ancestry being
264 ≤60% to their registry breed (Holstein n=3, Jersey n=1, Japanese Black n=3) (Figs. 3 a-
265 b).

266

267 **Breeds with open herdbooks**

268 For the Gelbvieh, Limousin, Shorthorn, Simmental, and Braunvieh breeds that have
269 open U.S. herdbook registries, fullblood or 100% ancestry individuals were identified
270 based on pedigree data obtained from the respective breed associations (Table 1). The
271 term “fullblood” is used to identify cattle for which every ancestor is registered in the
272 herdbook and can be traced back to the breed founders. The term “purebred” refers to
273 animals that have been graded up via crossbreeding to purebred status. Charolais also
274 has an open herdbook registry in the U.S., however, access to Full French imported

275 Charolais breed members was limited. As a result, all Charolais individuals identified as
276 purebred in the association registry were retained for downstream analysis, however,
277 these individuals could contain up to 1/32 introgression from another breed. A random
278 sample of 200 individuals was taken for each breed with more than 200 identified
279 fullblood individuals, otherwise all animals were sampled. Individuals previously
280 included in the candidate reference panel following preliminary fastSTRUCTURE
281 filtering for the open herd book breeds were removed and replaced with the fullblood
282 individuals.
283
284 **Additional reference panel filtering using SNPweights**
285 After filtering animals identified to not be fullblood based on their pedigree information,
286 we randomly sampled ≤50 individuals per reference breed and utilized SNPweights to
287 estimate weights for each sample and also to estimate breed ancestries for members of
288 the same sample that was used to generate the SNP weights. Based on these
289 analyses, we created 5 overlapping reference breed sets, each containing individuals
290 with ≥90%, ≥85%, ≥80%, ≥75%, or ≥70% ancestry assignment to their registry breeds
291 (Table 3).
292
293 **Simulated Genotypes**
294 Using the phased BC7K genotypes for the final reference population of 803 individuals
295 (3 Nelore genotyped with the BovineHD assay were removed because they were
296 determined to cause the phasing software problems), we simulated genotypes for 803
297 individuals each generation (N = 1, 3, 5 and 10) by randomly sampling two individuals

298 as parents from generation N-1 and using a Poisson distribution to sample at random a
299 single recombinant chromosome from each parent. The number of recombination
300 events for each sampled chromosome was sampled from a Poisson distribution with
301 mean equal to chromosome length in Mb/100 (i.e. 1.58 Morgans for chromosome 1).
302 Simulated genotypes were produced for individuals 1 generation removed from the
303 fullblood/purebred reference population animals (i.e., 50% breed A and 50% breed B),
304 3, 5, and 10, generations, respectively, to evaluate the ability of CRUMBLER to detect
305 large through to small admixture proportions in animals with increasing numbers of
306 breeds represented in their ancestry. Breed composition estimates for these animals
307 were obtained by tracing the breed of origin of every allele present in each generation N
308 animal. For each marker, we attributed the genomic fragment from the centre points of
309 the intervals on each side of each marker to the breed of origin of the two alleles at
310 each marker and summed these across all loci. Finally, we normalized these sums by
311 dividing by the autosomal genome size using UMD3.1 coordinates.
312

313 ## Results and Discussion

314

315 The concept of breed and breed membership is man-made and does not inherently
316 exist in nature. Moreover, the formation of breeds of cattle is very recent, as cattle
317 domestication began about 10,000 years ago but the formation of herdbooks has
318 occurred only during the last 200-250 years [16]. Nevertheless, the effects of drift and
319 human selection over the last 200 years have caused sufficient divergence among
320 breeds that breed differences are identifiable at the molecular level. Such signals are

321 essential for breed ancestry analyses to be effective in modern admixed animals.
322 Previous work on assigning breed composition in admixed cattle utilized 50K genotype
323 data and a reference panel of 16 breeds, with the basis for reference panel inclusion
324 being breed association registration [17]. However, the continual evolution of
325 genotyping assays has led to content changes resulting in only a relatively small
326 proportion of markers in common among assays. Consequently, there is a need to
327 evaluate whether these markers are sufficient for breed content estimation, leading to
328 their conservation in the design of future assays. Furthermore the development of an
329 analytical pipeline based on these markers would simplify analysis for end-users and
330 the use of a single reference panel would allow the direct comparison of results
331 between applications.
332
333 **Reference panel development**
334 Previously developed cattle reference panels have relied on pedigree accuracy and
335 breed association registration for their definition [17]. Conversely, we used an iterative
336 approach for reference population curation that was able to validate the accuracy of the
337 pedigree information used to identify candidates. FastSTRUCTURE analyses
338 performed using the candidate individuals for each of the initial 19 reference breeds
339 suggested population subdivision in both the Hereford and Simmental (Figure 2).
340 Pedigree analysis for the Herefords within each subpopulation indicated that the
341 subpopulations comprised animals from the highly inbred USDA Miles City Line 1
342 Hereford population (L1) and other individuals representing broader U.S. Hereford
343 pedigrees. The Miles City L1 Hereford cattle were derived from two bulls, both sired by

344 Advance Domino 13 (AHA registration number 1668403) and 50 Hereford foundation
345 cows. Since the founding of the L1 Herefords, the migration of germplasm has been
346 unidirectional from L1 into the broader U.S. industry, as the L1 population has been
347 closed since its founding [18]. However, the L1 Herefords have profoundly influenced
348 the U.S. Hereford population. L1 Herefords do not segregate for recessive dwarfism,
349 which has been a threat to Hereford breeders since the 1950s, and this has led to L1
350 cattle becoming popular in the process of purging herds of the defect [19]. In 1980, the
351 average proportion of U.S. registered Herefords influenced by L1 genetics was 23%.
352 By 2008, this proportion had increased to 81% [18].
353
354 The detected subpopulation division within the Simmental breed (Figure 2) represents
355 the differentiation between purebred and fullblood animals. For example, progeny of a
356 popular Simmental sire are present in both subpopulations, however, in one
357 subpopulation the family members are all fullblood and in the other they are all purebred
358 or percentage Simmental animals. This result supports the need to identify fullblood
359 animals as reference panel breed representatives for breeds with open herdbooks.
360
361 Reference population sample size
362 By randomly sampling individuals from the candidate reference breed set and using
363 SNPweights to assign these individuals to reference populations, we found that
364 reference panel breed sample sizes of ≤50 or ≤100 individuals appeared to capture the
365 diversity within each breed and appropriately determined the ancestry of the tested
366 individuals (Fig. 3 a-b). For each breed, the percent ancestry predicted for the tested

367 reference samples was, on average, 3.86% higher when the SNP weights were
368 estimated using ≤50 individuals per breed than when ≤100 individuals per breed were
369 used (Table 4). This reflects the increased homogeneity of individuals within each
370 breed and a greater genetic distance between individuals from different breeds as
371 smaller samples of individuals from each breed are used to define the reference panel.
372 Further, due to limitations in the number of genotyped individuals for some breeds
373 (Table 1), as the sample size was increased globally, imbalances were created between
374 the reference panel breed sample sizes which impacted breed composition estimation
375 (Fig. S9-S10). It has previously been shown that the power to detect population
376 structure improves as the reference population sample sizes become more similar
377 [6,14].
378
379 Marker density
380 After the replacement of reference breed individuals with those identified to be fullblood
381 based on pedigree analysis for the open herdbook Gelbvieh, Simmental, Limousin,
382 Braunvieh, Shorthorn, and Charolais breeds, additional self-assignment analyses were
383 conducted to evaluate the effects of marker set size on ancestry prediction. Breed
384 reference panels were again constructed by randomly sampling ≤50 individuals per
385 breed and SNP weights were calculated using both the BC13K markers and BC7K
386 markers. The estimated SNP weights were then used to self-assign ancestry to
387 members of the reference panel animals representing the reference breed set. The
388 ancestry predictions for the reference breed individuals using either the BC7K (Fig. 4a;
389 Fig. S11) or BC13K (Fig. 4b; Fig. S12) marker sets indicate that use of the BC13K

390 marker set did not significantly impact the ancestry predictions. Consequently, the use
391 of the 6,799 markers common to the 8 commercially available genotyping platforms
392 appears to be sufficient to assign breed ancestry for the majority of animals produced in
393 the U.S. The CRUMBLER pipeline can accommodate samples genotyped using
394 alternative assays ,however, the produced breed composition estimates will be based
395 on the intersection of markers on the assay and the BC7K marker set.
396
397 Assignment thresholds
398 We next examined the effects of reference breed homogeneity on ancestry assignment
399 by identifying reference panel members that had been assigned to their breed of
400 registry using SNPweights with probabilities of ancestry of ≥90%, ≥85%, ≥80%, ≥75%,
401 and ≥70%, respectively (Table 3). From these individuals, reference breed panels were
402 obtained by randomly sampling ≤50 individuals per breed, until each individual was
403 represented in at least one sample set. SNP weights were then estimated using the
404 BC7K marker set and ancestry was assigned for these individuals using SNPweights
405 (Figs. 5-6 and Figs. S13-S15). Limiting the reference breed panel members to those
406 individuals with ≥90% ancestry assigned to their breed of registry produced a reference
407 panel that did not represent the extent of diversity within each of the breeds (Fig. 5). On
408 the other hand, using an ancestry assignment of ≥85% clearly captured greater diversity
409 within each breed (Fig. 6) and maximized the self-assignment of ancestry to the breed
410 of registration (Table 5).
411
412 Reference panel definition

413 To examine whether the specific individuals represented in the reference panel sample
414 influenced the self-assignment of ancestry to the sampled individuals, a second sample
415 of ≤50 distinct individuals per breed was obtained from the individuals with ≥85%
416 assignment to their breed of registration and analyzed with SNPweights (Fig. 7). Fig. 7
417 indicates that the ability to predict ancestry was not influenced by the specific individuals
418 sampled from the set of animals with ≥85% ancestry to their breed of registration.
419
420 Additionally, Figs. 6 and 7 suggest that the use of a reference breed panel constructed
421 by the random sampling of ≤50 individuals per breed from individuals with ≥85% self-
422 assigned ancestry to their breed of registration maintained sufficient within-breed
423 diversity to accurately estimate the ancestry of target individuals. However, these
424 figures also reveal small amounts of apparent introgression from other reference panel
425 breeds within each of the breeds. This does not appear to be an issue of marker
426 resolution since the analyses performed with the BC7K and BC13K marker sets
427 generated similar results (Fig. 4). We conclude that these apparent introgressions are
428 either due to a lack of power to discriminate among breeds using the common markers
429 designed onto commercial genotyping platforms, or represent the presence of common
430 ancestry among the breeds prior to the formation of breed herdbooks ~200 years ago.
431 Molecular evidence for this shared ancestry exists, for example, Hereford and Angus
432 cattle share the *Celtic* polled allele [20] and the segmental duplication responsible for
433 the white anterior, ventral and dorsal coat color pattern occurs only in Hereford and
434 Simmental cattle and their crosses [21]. These data clearly indicate that crossbreeding
435 was widespread prior to the formal conceptualization of breeds.

436

437 **Reference panel validation**

438 To evaluate the ability of the selected reference breed panel to identify breed
439 composition, an analysis was conducted for all 170,544 samples in the database which
440 required 60 processor minutes (Fig 6-7). We extracted animals with pedigree
441 information including fullblood and purebred animals registered with open herdbook
442 breed associations and 2,243 crossbred animals with varying degrees of admixture.
443 Considering the amount of available data, the number of pedigreed admixed animals
444 was very limited and the purebred animals all had similar expected admixture
445 proportions. Consequently, we next simulated genotypes for animals by assuming the
446 random mating of members of the reference breed panel for 1, 3, 5 and 10 generations
447 assuming non-overlapping generations to generate generations of animals with different
448 numbers of breeds and breed proportions represented in their genomes.

449

450 <u>Registered fullblood animals</u>

451 For the Gelbvieh, Limousin, Shorthorn, Simmental, and Braunvieh breeds that have
452 open herdbook registries, fullblood or 100% ancestry individuals were identified based
453 on pedigree data obtained from the respective breed associations (Table 1).
454 CRUMBLER estimates were obtained for these fullblood individuals and the distribution
455 of estimates by breed are in Fig. 8. For all breeds except Charolais, >50% of the
456 individuals had CRUMBLER estimated percentages of ≥80% to their respective breeds.
457 Average percentage estimates for fullblood Gelbvieh, Limousin, Shorthorn, Simmental,
458 and Braunvieh individuals were 76%, 78%, 83%, 79%, and 85%, respectively (Fig. 8b).

459 However, the number of genotyped imported Full French Charolais animals was limited
460 and so we also analyzed all purebred Charolais individuals which could contain up to
461 1/32nd of their genome introgressed from another breed. The average Charolais breed
462 assignment was 72% and the distribution of estimates was more variable than for the
463 fullblood animals from the other breeds (Fig. 8b).
464
465 Pedigreed crossbred animals
466 Based on pedigree, 2,005 individuals were identified as being primarily Hereford but
467 with varying degrees of Red Angus, Salers, Angus or unknown other breed influence.
468 The analysis results agreed with the pedigree data (Fig. 9a) To investigate the
469 correlations between pedigree and CRUMBLER estimated breed proportions, we
470 removed proportions for breeds that were less than 3% and normalized the remaining
471 values. CRUMBLER estimates were then correlated with the pedigree predicted
472 estimates of the proportion of Hereford in these individuals (Fig. 9c). CRUMBLER
473 tended to underestimate the Hereford proportion as the pedigree estimated Hereford
474 proportion tended to 100%.
475
476 The remaining 238 crossbred individuals were commercial, advanced generation
477 animals with an expected 50% Angus and 50% Simmental ancestry based on pedigree
478 data. Results of the CRUMBLER analysis support the pedigree data (Fig. 10). The
479 presence of Red Angus ancestry in these animals reveals the inability of the analysis to
480 fully differentiate between Angus and Red Angus, which only diverged in the U.S. in
481 1954, and also the influence of Red Angus in the U.S. Simmental breed (Fig. S16).

482

483 Simulated genotypes

484 Genomes were simulated using the phased genotypes for 803 individuals from the
485 reference breed panel to contain varying breed numbers and admixture proportions
486 after 1, 3, 5, and 10 generations of random mating with nonoverlapping generations. In
487 generation 1, the admixed individuals were F_1 individuals with a 50:50 autosomal
488 genome composition unless both parents were randomly sampled from the same breed.
489 CRUMBLER estimates of breed composition using the simulated genotypes were
490 strongly correlated with the simulated compositions, especially for generations 1 and 3
491 (Fig 11). As the number of generations increases, the number of breeds represented in
492 the simulated genomes tends to increase and the proportion of the genome originating
493 from any one breed tends to decrease and the correlation between the simulated
494 proportions and CRUMBLER estimates decreased. Nevertheless, by generation 10
495 44% of animals had their genome proportions estimated with a correlation of at least
496 70%. In the U.S. commercial crossbreeding does not usually involve the use of more
497 than 3-4 breeds of cattle and while the number of generations of crossbreeding may
498 very well be 10 or perhaps more, many generations will involve the mating of animals
499 with similar genome ancestries and the proportions for each breed will be much greater
500 than present in the generation 10 animals in Fig 11. Consequently, the achieved
501 accuracies are likely to be closer to the generation 3 or 5 results where 99% and 68% of
502 animals, respectively, had their genome proportions estimated with a correlation of
503 greater than 80%.

504

505 **Violation of model assumptions**

506 The ancestry model assumes that neither drift or selection has acted to alter the allele
507 frequencies from those created by the initial admixture proportions. We examined
508 CRUMBLER estimates of breed composition for advanced generation members of the
509 Brangus (n=13,842), Beefmaster (n=4,010) and Santa Gertrudis (n=2,056) composite
510 breeds where selection has had the opportunity to change breed composition from
511 expectations at breed formation and Brangus individuals are expected to be ⅝ Angus
512 and ⅜ Brahman, Beefmaster individuals ¼ Hereford, ¼ Shorthorn and ½ Brahman, and
513 Santa Gertrudis ⅝ Shorthorn and ⅜ Brahman, respectively. However, CRUMBLER
514 assigned relatively large proportions of Holstein ancestry to these advanced generation
515 composite animals (Fig. 12). These American breeds do not contain any Holstein
516 introgression and they do not contain ancestry from a “Ghost Population”, a population
517 not present in the reference set which would lead to a breed assignment to a reference
518 breed that it most closely resembles [6]. We speculate that this effect is caused by
519 selection creating a deviation from the founder breed proportions which is best
520 explained by an introgression from a distantly related breed, in this case, Holstein.
521 Stratifying these genotyped animals according to the number of generations from
522 foundation fullblood animals and examining the extent of estimated Holstein
523 introgression, which would be expected to increase with generation number, would
524 enable this to be tested, however, we did not have access to the necessary data.
525 However, this hypothesis is supported by the fact that the Santa Gertrudis had the least
526 Holstein introgression and the breed has published estimates of additive genetic merit
527 for many fewer years than the Beefmaster or Brangus.

528

529 **Admixture**

530 We also tested the ADMIXTURE software [22] for ancestry estimation and integration
531 into the CRUMBLER pipeline using the same reference breed panel that was developed
532 for use with SNPweights. ADMIXTURE uses maximum likelihood estimation to fit the
533 same statistical model as STRUCTURE, however, STRUCTURE does not allow the
534 specification of individuals of known descent to be used as a reference panel [22].
535 ADMIXTURE allows a supervised analysis, in which the user can specify a reference
536 set of individuals, by specifying the "--supervised" flag and requires an additional file
537 with a ".pop" suffix to specify the genotypes of the reference population individuals [22].
538 Unlike SNPweights, the reference population individuals' genotypes must be provided in
539 a genotype file for each analysis.

540

541 We first conducted an ADMIXTURE analysis in which we self-assigned ancestry for the
542 animals in the reference breed set formed with ≤50 individuals per breed from the
543 individuals that had ≥85% assignment to the breed of registration (Fig. 13). The results
544 shown in Fig. 13 are similar to those in Fig. 6 for the same reference panel, albeit with
545 perhaps less evidence of background introgression. We next conducted an analysis
546 using the reference panel used in Fig. 13 merged with data for the 2,005 high
547 percentage crossbred Herefords animals. The results shown in Fig. 14, reveal a
548 significant change in the ancestry proportions estimated for the reference panel
549 Guernsey, Gelbvieh and Romagnola individuals between the two analyses which used
550 exactly the same reference panel, but differed only in the number of individuals for

551 which ancestry was to be estimated. This suggests that ADMIXTURE may use the
552 target individuals to update information provided by the reference panel individuals
553 specified in the ".pop" file. Consequently, the ADMIXTURE estimated ancestry
554 proportions appear to be context dependent and may vary based on the other
555 individuals included in the analysis.
556
557 Moreover, the order in which the target individuals appear in the genotype input file also
558 appears to affect ADMIXTURE estimates of ancestry proportions for the target
559 individuals. Fig. 15 shows the results of an ADMIXTURE analysis in which the target
560 individuals were identical to those shown in Fig. 14, but for which the order of the
561 reference individuals and the 2,005 Hereford crossbred individuals was reversed in the
562 input files. In Fig. 14, the reference individuals appear before the 2,005 Hereford
563 crossbred individuals in the input file, whereas in Fig. 15, the 2,005 Hereford crossbred
564 individuals appeared before the reference individuals in the input file. The results reveal
565 a significant change in ancestry proportions for Guernsey and Gelbvieh, but the
566 Romagnola now appear to be non-admixed. Finally, we performed an ADMIXTURE
567 analysis for these animals in which the order of animals in the input genotype file was
568 completely randomized (Fig. 16). Following analysis, the individuals were sorted to
569 generate Fig. 16. Again, the ancestry proportions for the Guernsey, Gelbvieh and
570 Romagnola individuals suggest these breeds to be admixed.
571
572 STRUCTURE and ADMIXTURE are widely used for characterising admixed populations
573 [6], however, we have not found any reports in the literature that indicate that the

software is sensitive to the input order of individuals. However, we suspect that the majority of users would have no need or motivation to run the software with permuted data input files. Nevertheless, because of these inconsistencies between results, we chose to not use ADMIXTURE for ancestry estimation within the CRUMBLER pipeline.

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 number of markers used for ancestry assignment. The ancestry proportions assigned based on the BC6K marker set (Fig. 17) did not differ appreciably 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

597 employed several iterations of filtering to remove recently admixed individuals and
598 identify a relatively homogeneous set of individuals that nevertheless represented the
599 variation that might be expected among individuals within a breed. Once determined,
600 the reference panel genotype data need only be processed once to obtain SNP weights
601 removing the need to share genotype data for reference individuals in subsequent
602 studies [11]. The upfront development of an external reference breed panel capitalizes
603 on the rich ancestry information available in large available datasets, and relatedness,
604 variation in sample sizes and diversity among the target individuals does not affect the
605 inference of ancestry [11].
606
607 In cattle, the visual evaluation of breed characteristics is a poor method for evaluating
608 the ancestry of individuals. Breed association pedigrees can be used to estimate
609 expected breed compositions, however, the random assortment of chromosomes into
610 gametes and selection can lead to ancestry proportions that differ from those expected
611 based upon pedigree. Moreover, the vast majority of commercial beef cattle in the U.S.
612 have no or very limited pedigree information and since these animals are frequently
613 used for genomic research [3–5], there is a need for a tool that can routinely provide
614 ancestry estimates for downstream use in GWAA or other genetic studies.
615
616 We tested ADMIXTURE and SNPweights and found that results from ADMIXTURE
617 appear to depend on the ancestry and order of appearance of individuals within the
618 genotype input file. We therefore developed an analysis pipeline, CRUMBLER, based
619 upon PLINK, EIGENSOFT and SNPweights to automate the process of ancestry

620 estimation. The developed bovine pipeline utilizes the 6,799 SNPs present on all 8
621 commercially utilized bovine SNP genotyping assays and results using these SNPs are
622 consistent with results obtained when 13,291 SNPs were used. From an available
623 48,776 genotyped individuals, we also developed a reference panel of 806 individuals
624 sampled from 17 breeds to have ≤50 individuals per breed that had ≥85% assignment
625 to their breed of registration. This panel appears to allow the robust estimation of the
626 ancestry of advanced generation admixed animals, however, all breeds share some
627 common ancestry which predates the recent development of breed association
628 herdbooks [16,23].
629
630 CRUMBLER is not limited to application in cattle and with the provision of suitable
631 reference breed allele frequencies can be applied to other species for ancestry
632 estimation. CRUMBLER pipeline scripts and reference panel breed SNP weights are
633 available on GitHub (https://github.com/tamarcrum/CRUMBLER).

634 **Additional files**

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

638

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

642

643 **Supplementary Figures (PDF).** This file contains Supplementary Figures S1-S16.

644

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

647

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

650

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

653

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

656

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 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

Ethics approval and consent to participate

Not applicable.

Funding

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

Consent for Publication

726 Not applicable.

727

728 **Acknowledgements**

729 Not applicable.

730

731 ## References

732 1. Burnett MS, Strain KJ, Lesnick TG, de Andrade M, Rocca WA, Maraganore DM.
733 Reliability of self-reported ancestry among siblings: implications for genetic
734 association studies. Am J Epidemiol. 2006 Mar 1;163(5):486–92.

735 2. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using
736 multilocus genotype data. Genetics. 2000 Jun;155(2):945–59.

737 3. Saatchi M, Beever JE, Decker JE, Faulkner DB, Freetly HC, Hansen SL, et al.
738 QTLs associated with dry matter intake, metabolic mid-test weight, growth and feed
739 efficiency have little overlap across 4 beef cattle studies. BMC Genomics. 2014
740 Nov 20;15:1004.

741 4. Seabury CM, Oldeschulte DL, Saatchi M, Beever JE, Decker JE, Halley YA, et al.
742 Genome-wide association study for feed efficiency and growth traits in U.S. beef
743 cattle. BMC Genomics. 2017 May 18;18(1):386.

744 5. Neibergs HL, Seabury CM, Wojtowicz AJ, Wang Z, Scraggs E, Kiser JN, et al.
745 Susceptibility loci revealed for bovine respiratory disease complex in pre-weaned
746 holstein calves. BMC Genomics. 2014 Dec 22;15:1164.

747 6. Lawson DJ, van Dorp L, Falush D. A tutorial on how not to over-interpret
748 STRUCTURE and ADMIXTURE bar plots. Nat Commun. 2018 Aug 14;9(1):3258.

749 7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al.
750 PLINK: a tool set for whole-genome association and population-based linkage

751 analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

752 8. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-
753 generation PLINK: rising to the challenge of larger and richer datasets.
754 Gigascience. 2015 Feb 25;4:7.

755 9. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS
756 Genet. 2006 Dec;2(12):e190.

757 10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
758 components analysis corrects for stratification in genome-wide association studies.
759 Nat Genet. 2006 Aug;38(8):904–9.

760 11. Chen C-Y, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved
761 ancestry inference using weights from external reference panels. Bioinformatics.
762 2013 Jun 1;29(11):1399–406.

763 12. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational Inference of
764 Population Structure in Large SNP Data Sets. Genetics. 2014 Jun 1;197(2):573–
765 89.

766 13. Wang J. The computer program structure for assigning individuals to populations:
767 easy to use but easier to misuse. Mol Ecol Resour. 2017 Sep;17(5):981–90.

768 14. Puechmaille SJ. The program structure does not reliably recover the correct
769 population structure when sampling is uneven: subsampling and new estimators
770 alleviate the problem. Mol Ecol Resour. 2016 May;16(3):608–27.

771 15. Sanders JO. History and Development of Zebu Cattle in the United States. J Anim
772 Sci. 1980 Jun 1;50(6):1188–200.

773 16. Decker JE, McKay SD, Rolf MM, Kim J, Molina Alcalá A, Sonstegard TS, et al.
774 Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle.
775 PLoS Genet. 2014 Mar;10(3):e1004254.

776 17. Kuehn LA, Keele JW, Bennett GL, McDaneld TG, Smith TPL, Snelling WM, et al.
777 Predicting breed composition using breed frequencies of 50,000 markers from the
778 US Meat Animal Research Center 2,000 Bull Project. J Anim Sci. 2011
779 Jun;89(6):1742–50.

780 18. Leesburg VLR, MacNeil MD, Neser FWC. Influence of Miles City Line 1 on the
781 United States Hereford population. J Anim Sci. 2014 Jun;92(6):2387–94.

782 19. McCann LP. battle of bull runts. 1974; Available from: http://agris.fao.org/agris-
783 search/search.do?recordID=US201300539215

784 20. Wiedemar N, Tetens J, Jagannathan V, Menoud A, Neuenschwander S,
785 Bruggmann R, et al. Independent polled mutations leading to complex gene
786 expression differences in cattle. PLoS One. 2014 Mar 26;9(3):e93435.

787 21. Whitacre L. Structural variation at the KIT locus is responsible for the piebald
788 phenotype in Hereford and Simmental cattle. 2014; Available from:
789 http://search.proquest.com/openview/45eba5fa3c5757a2c4c2ab18af1a8a98/1?pq-
790 origsite=gscholar&cbl=18750&diss=y

791 22. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
792 unrelated individuals. Genome Res. 2009 Sep;19(9):1655–64.

793 23. Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, Chen K, et al. Resolving
794 the evolution of extant and extinct ruminants with high-throughput phylogenomics.
795 Proc Natl Acad Sci U S A. 2009 Nov 3;106(44):18644–9.

796

797 **Table 1. Genotype data for 48,776 registered individuals from 20 breeds were**
798 **used to establish the reference population.**
799

Breed	No. Registered Individuals	No. FullBlood Individuals[a]	No. Individuals Assigned to Breed[b]	Sampled Individuals[c]	No. Individuals After Pedigree and SNPweights[d]
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

800 [a]Number of registered animals determined by pedigree analysis to be fullblood for breed
801 associations with open herdbooks.
802 [b]Number of registered animals assigned to their identified breed with P≥0.97 by
803 fastSTRUCTURE in preliminary analyses and retained for subsequent analyses.
804 [c]A random sample of 200 individuals was obtained for breeds with >200 individuals after
805 fastSTRUCTURE analysis and all individuals were sampled for breeds with ≤200 per breed and
806 the data were again analyzed by fastSTRUCTURE with K=19 after removal of the Salers.
807 [d]Animals that were determined to not be fullblood by pedigree analysis and animals
808 assigned with P≤0.60 by SNPweights to their breed of registry were removed.
809

810 **Table 2. The number of variants queried by each assay and the number of**
811 **individuals from the 20 reference breeds genotyped using each assay.**
812

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

813

814 **Table 3. Number of individuals for each reference breed assigned to their breed of**
815 **registration by minimum ancestry threshold.**
816

	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

817

818 **Table 4. Ancestry proportion statistics for the self-assignment of reference panel**
819 **members from samples of ≤50 or ≤100 individuals from the candidate reference**
820 **breed individuals.**
821

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

822

823 **Table 5. Average predicted ancestry and variance in predicted ancestry for**
824 **candidate reference breed individuals when filtered on minimum predicted**
825 **ancestry.**
826

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

827

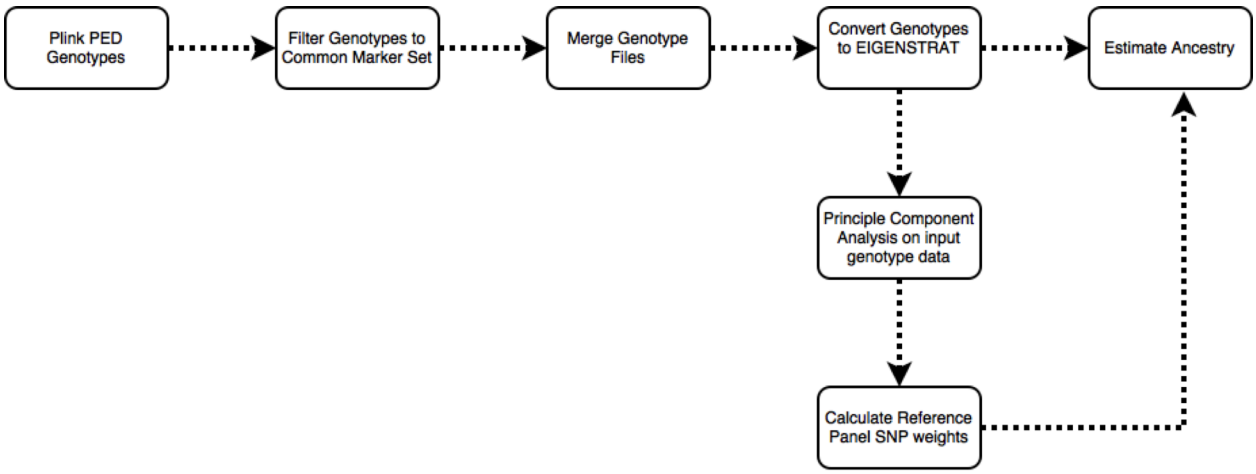


Fig. 1 Flow diagram of the breed composition pipeline.

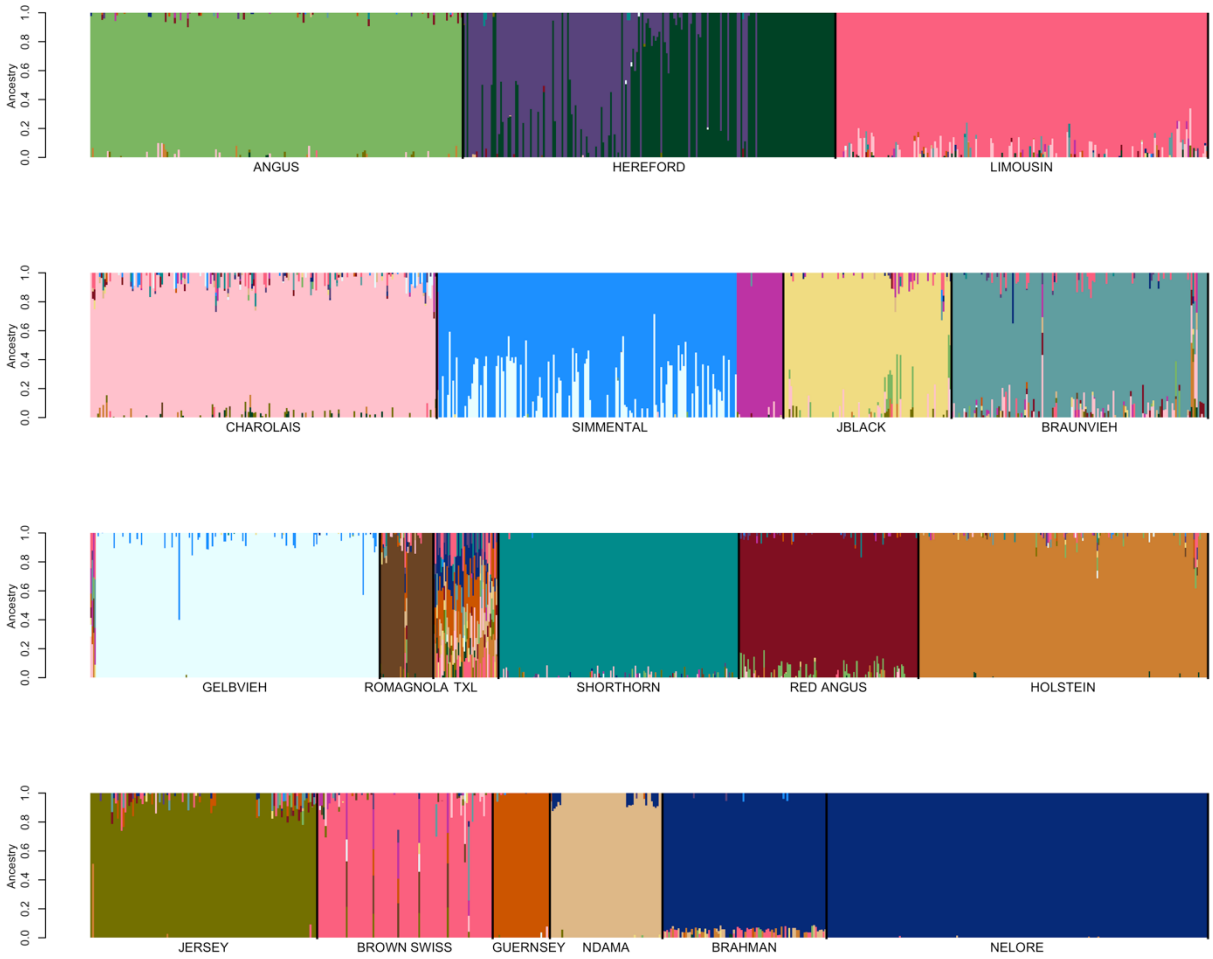
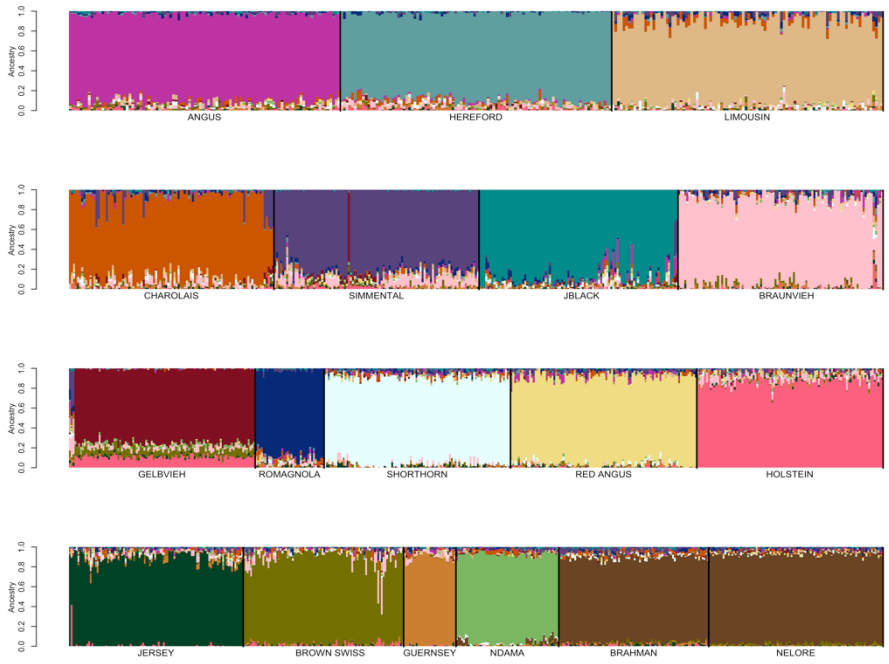
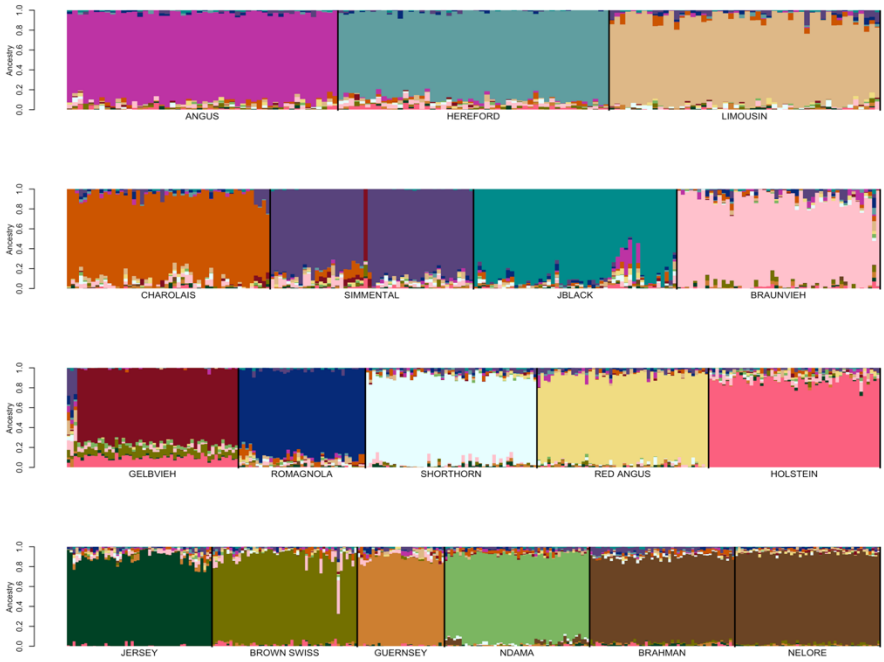


Fig. 2 FastSTRUCTURE results for a random sample of ≤200 individuals per breed from the pool of 17,852 potential reference individuals at K=19. Breed identification is shown below each colored block and each animal is represented as a vertical line within the block.

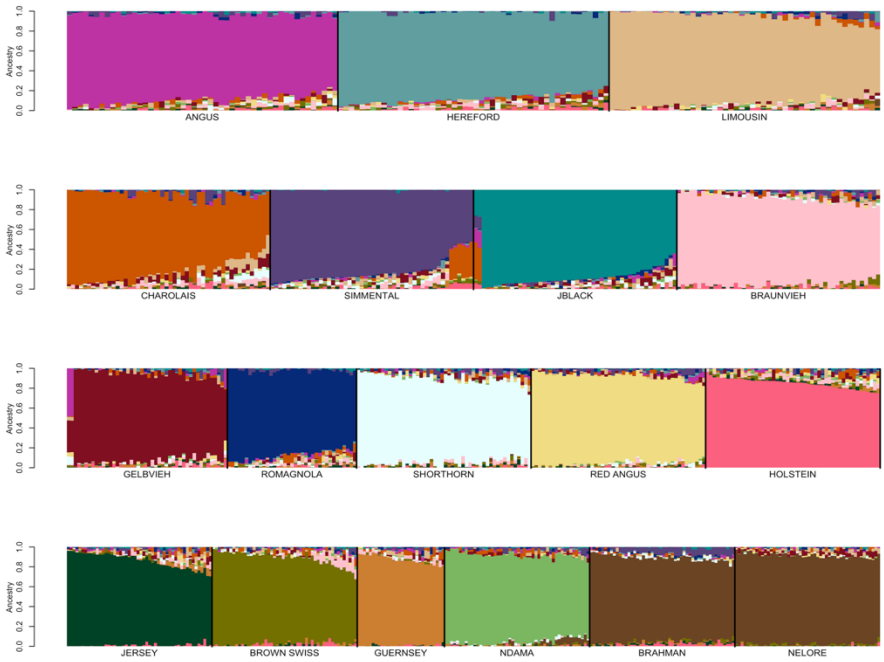


a.

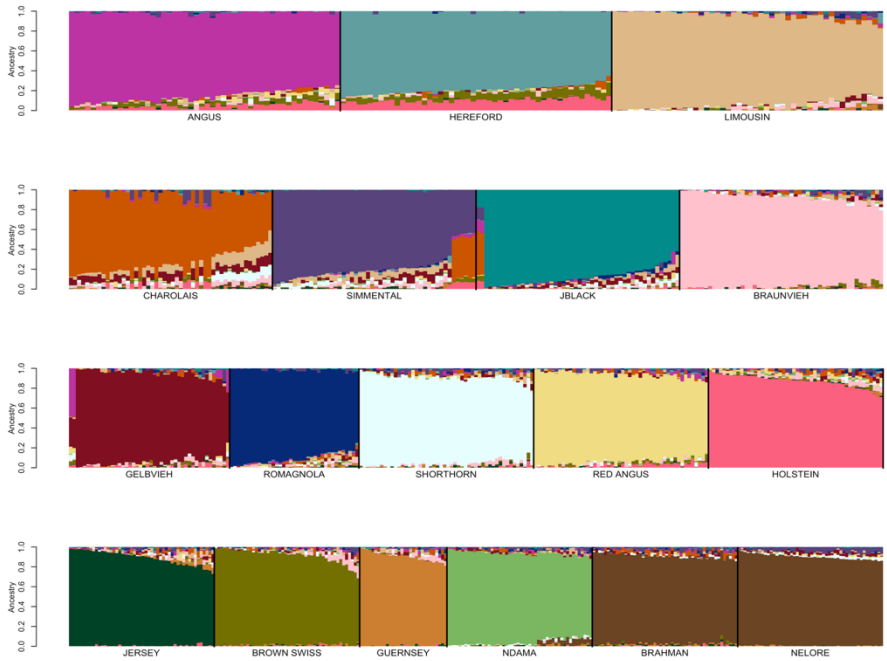


b.

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.

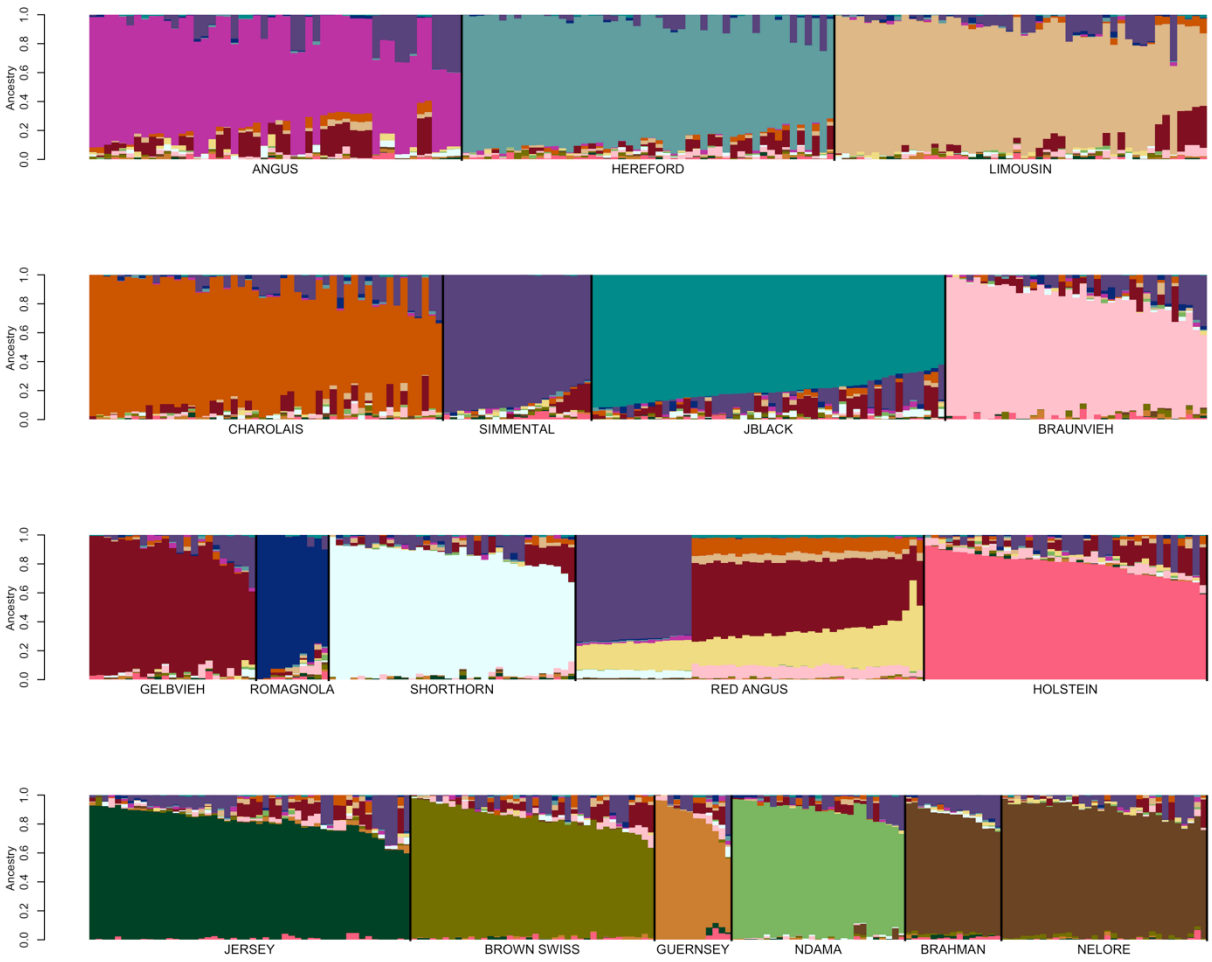


Fig. 5 Reference breed panel constructed by the random sampling of ≤50 individuals per breed from individuals with ≥90% ancestry was self-assigned to reference breed ancestry using the BC7K marker set.

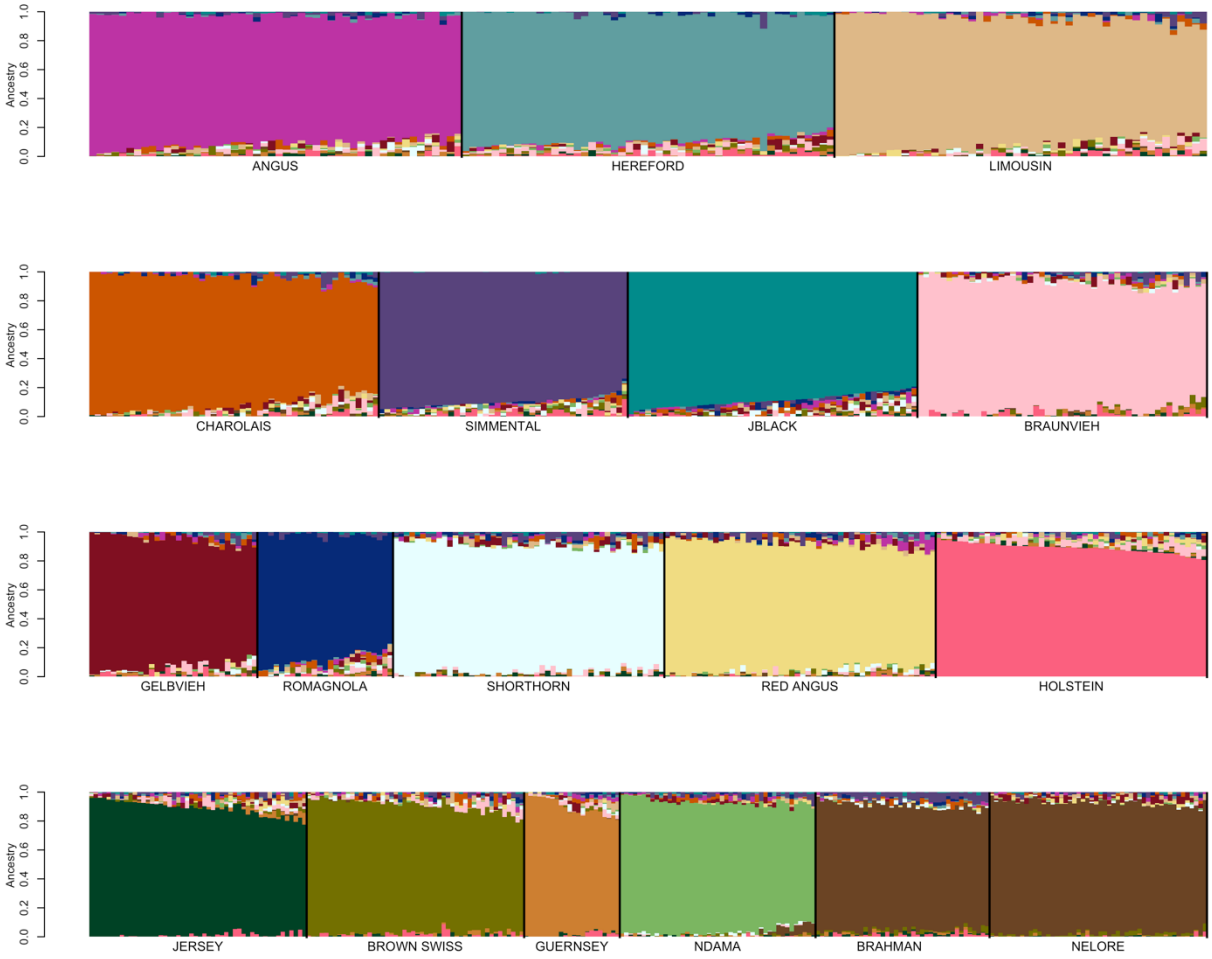


Fig. 6 Reference breed panel constructed by the random sampling of ≤50 individuals per breed from individuals with ≥85% ancestry was self-assigned to reference breed ancestry using the BC7K marker set.

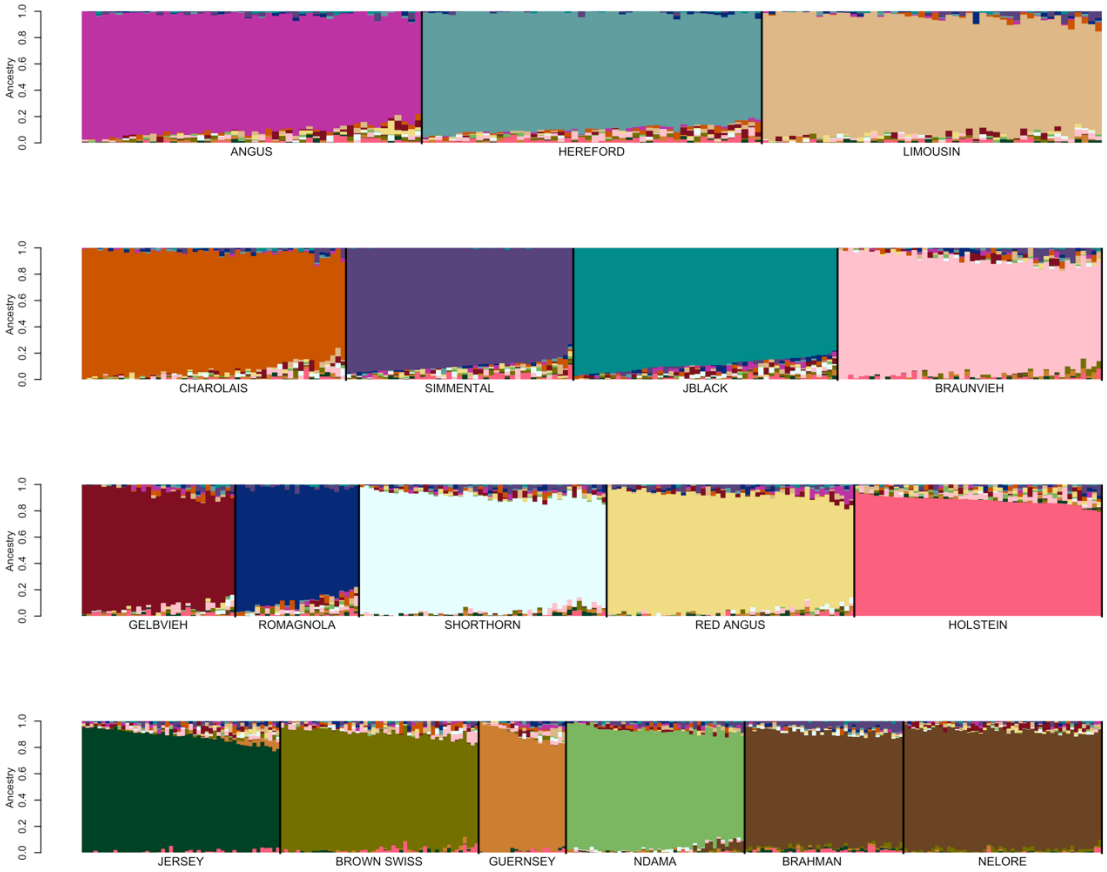
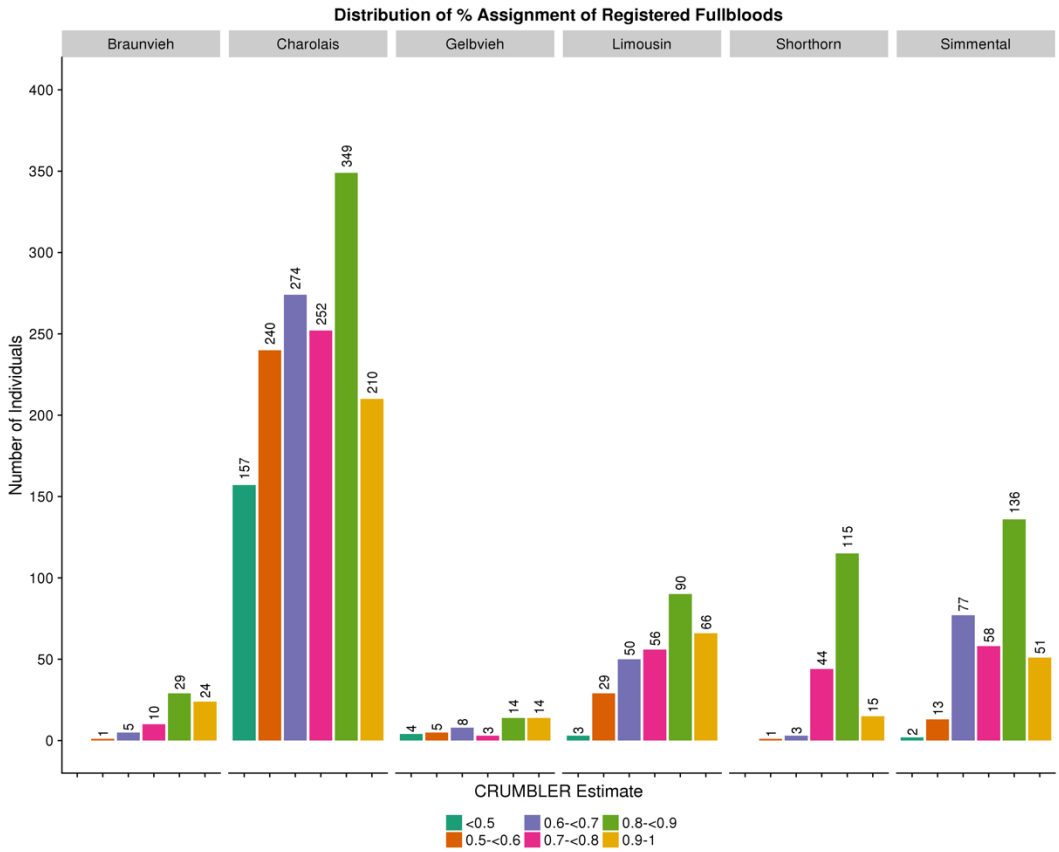


Fig. 7 Reference breed panel constructed by the independent random sampling of a second sample of ≤50 individuals per breed from individuals with ≥85% ancestry after eliminating individuals represented in the first sample was self-assigned to reference breed ancestry using the BC7K marker set.

a.



b.

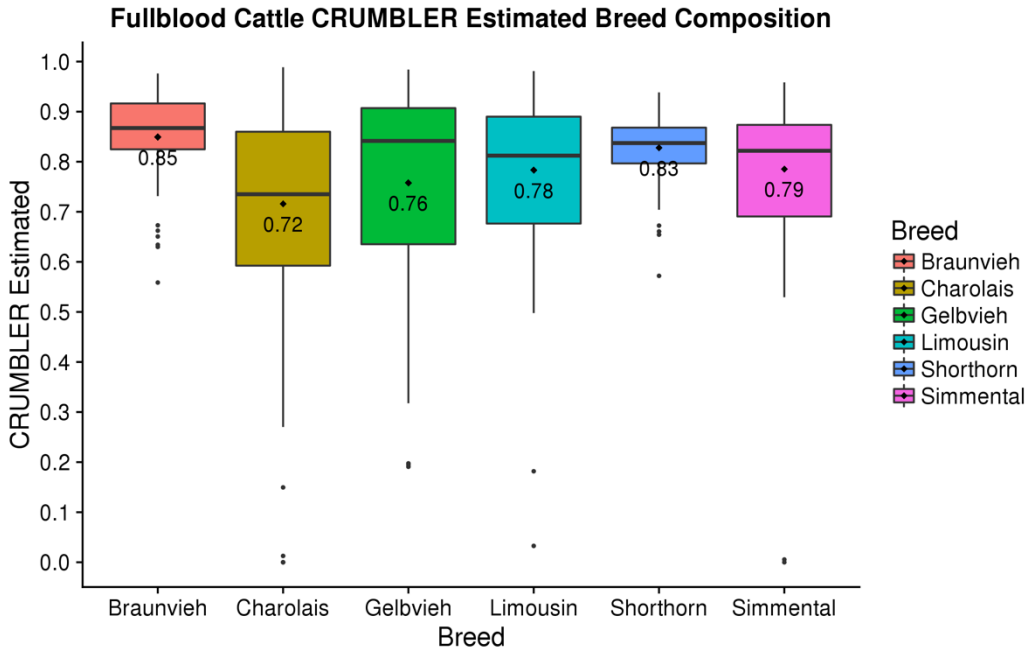
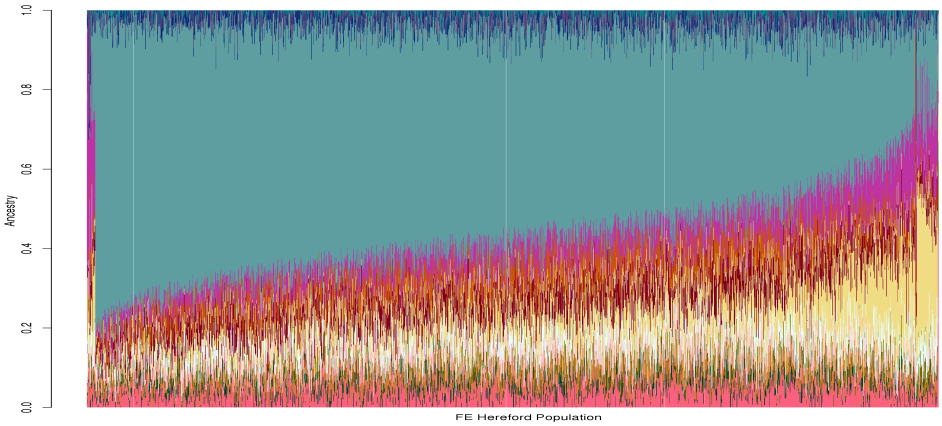
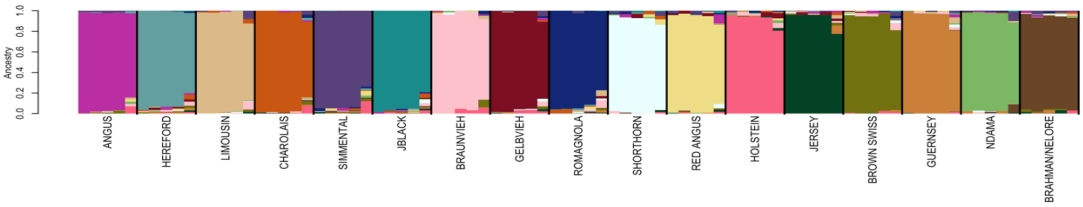


Fig. 8 (a) Distribution by breed of SNPweights ancestry assignment results for 2,408 registered fullblood animals from open herd book breeds. (b) Pictorial representation of CRUMBLER estimates for 2,408 registered fullblood animals from open herd book breeds.

a.



b.



c.

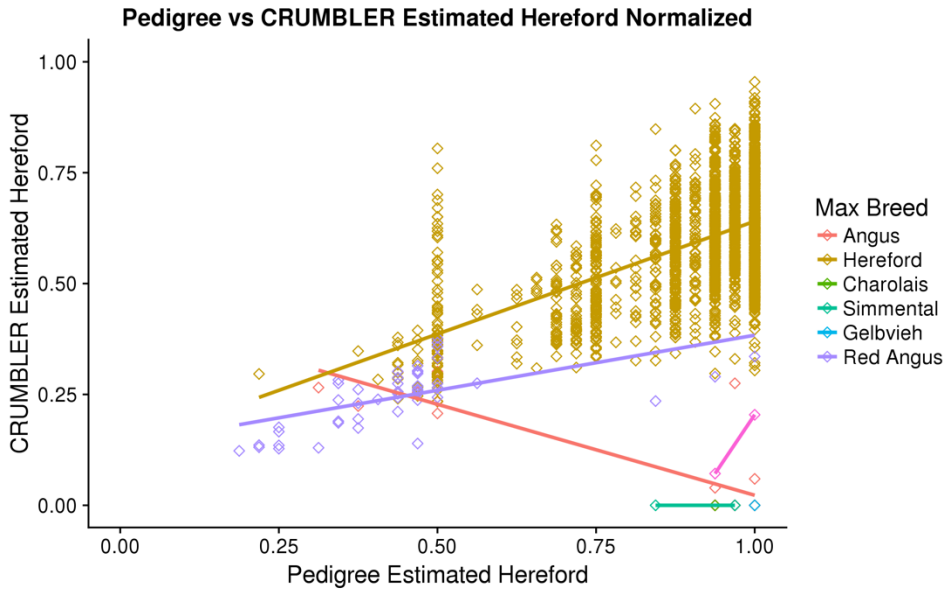
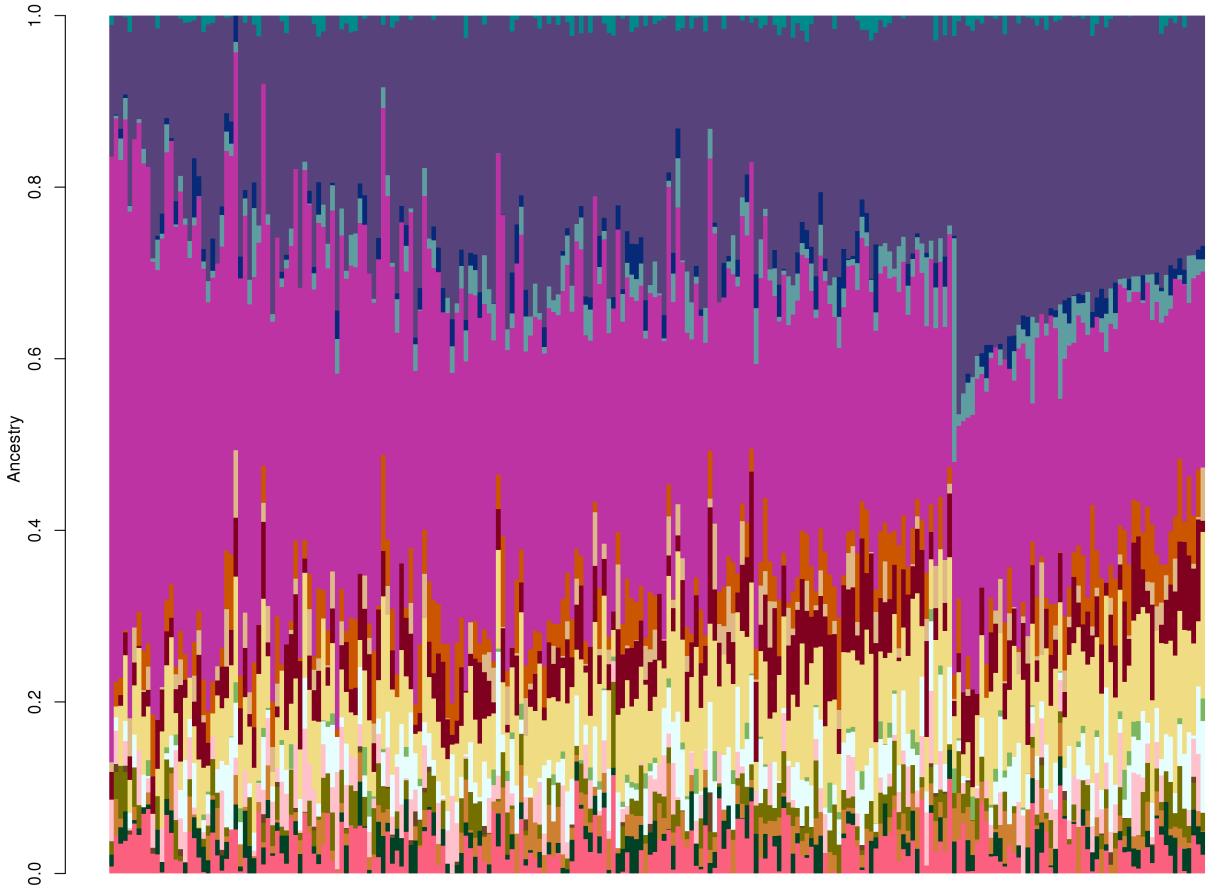


Fig. 9 (a) SNPweights ancestry results for 2,005 crossbred Hereford individuals with *a-priori* breed composition estimates determined by pedigree. (b) Breed assignment reference breed key. (c) Hereford SNPweights estimated proportions using CRUMBLER are plotted against the pedigree estimates. Data point color indicates the breed for which SNPweights assigned the highest proportion for each individual.

a.



b.

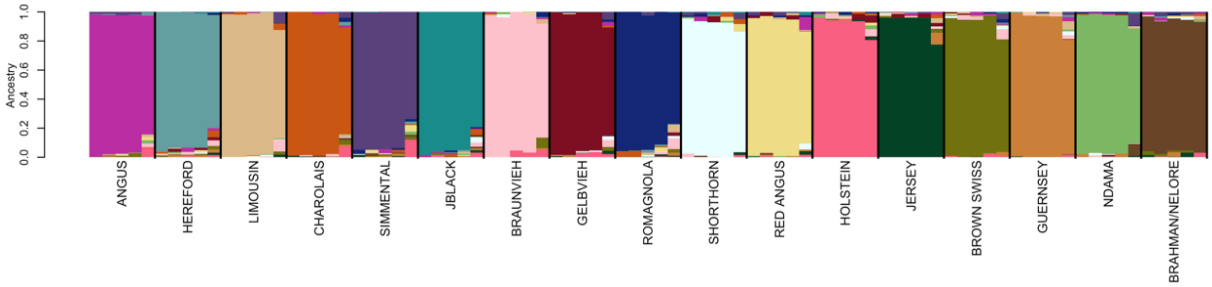


Fig. 10 (a) SNPweights ancestry results for 238 crossbred individuals with *a-priori* breed composition estimates of 50% Angus and 50% Simmental based on a reference panel with ≤50 individuals per breed sampled from individuals with ≥85% assignment to their breed of registry. (b) Breed assignment for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

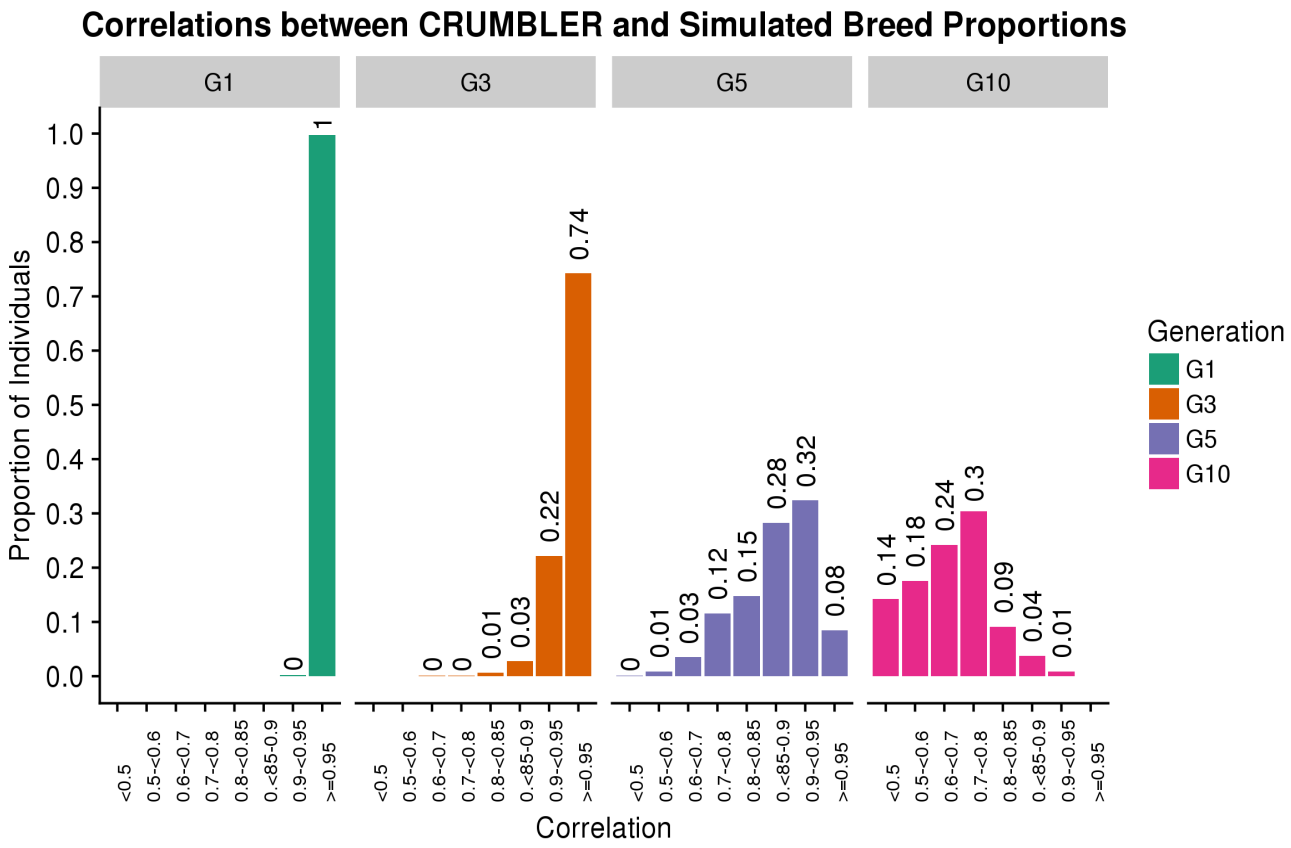
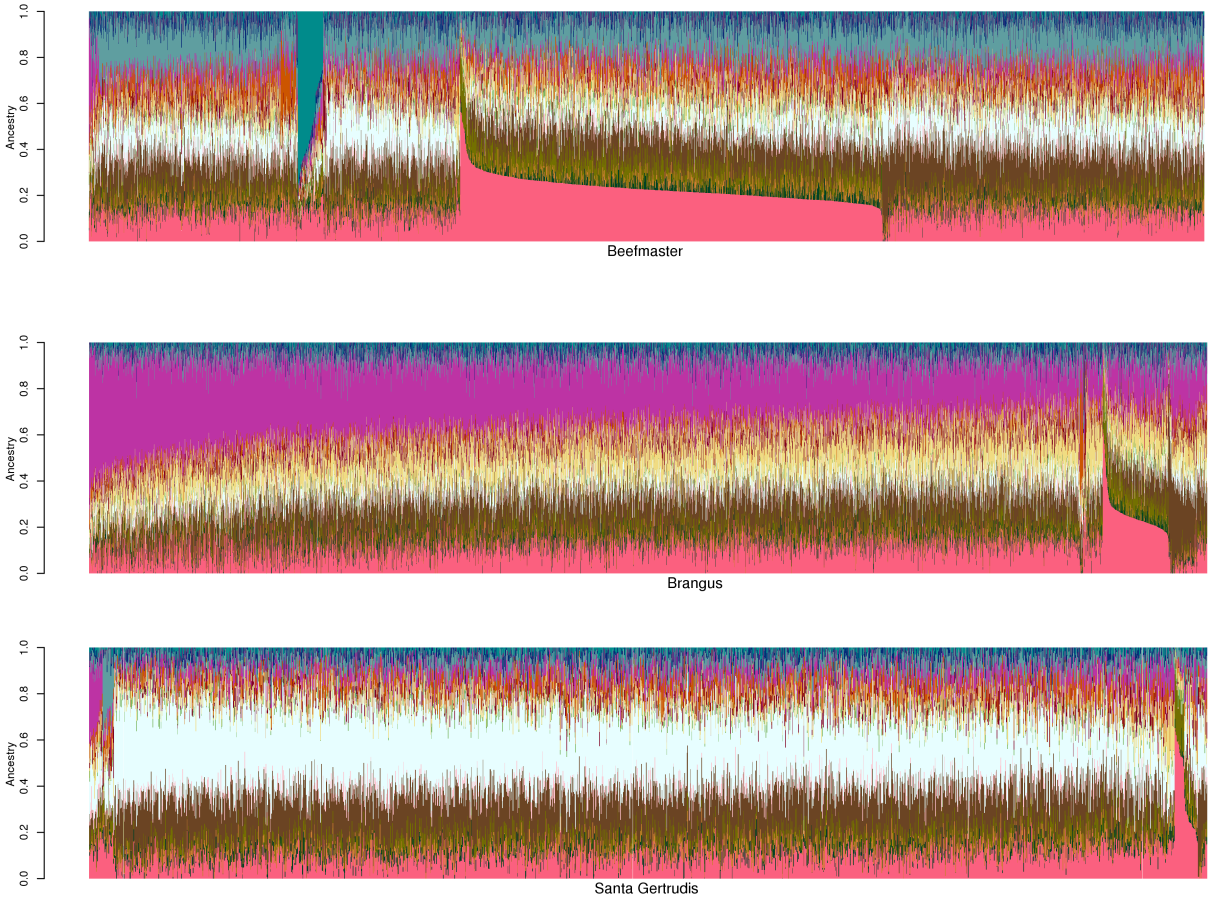


Fig. 11 Genotypes were simulated for the indicated number of generations of random mating, with generation 1 (G1) animals being 50:50 proportion except when two parents from the same breed were mated. SNPweights results were obtained using CRUMBLER pipeline parameters correlations between these estimates and the known simulated breed compositions were produced and the proportion of individuals within each correlation class is indicated.

a.



b.

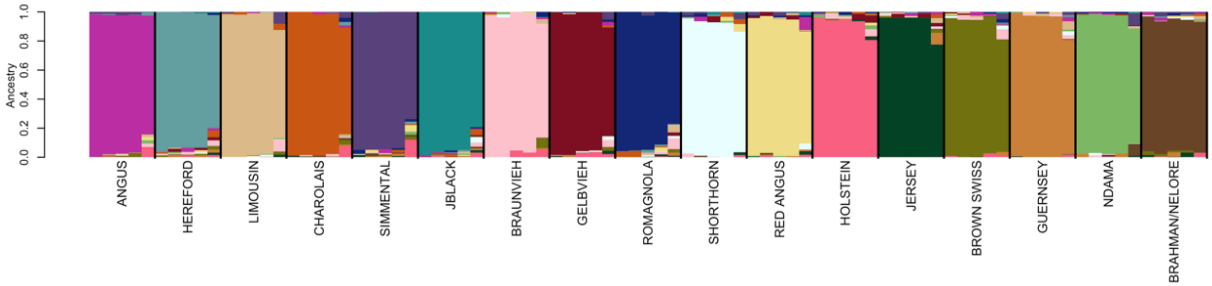


Fig. 12 (a) SNPweights ancestry results using CRUMBLER pipeline for 4,010 Beefmaster, 13,842 Brangus, and 2,056 Santa Gertrudis individuals. (b) Breed assignment for these advanced generation composite animals can be determined using this reference breed key.

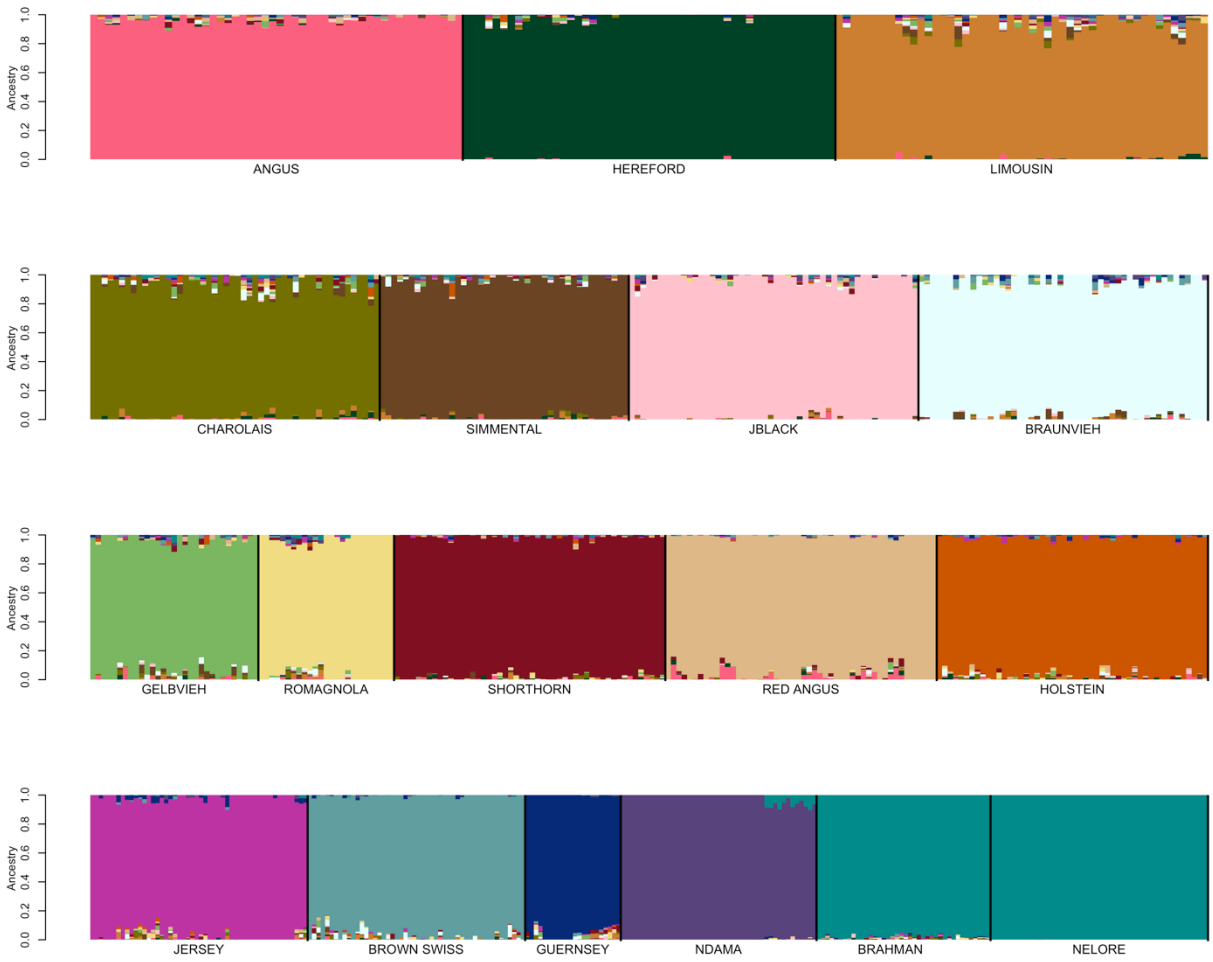


Fig. 13 Self-assignment of ancestry for the animals in the reference breed set formed with ≤50 individuals per breed from the individuals that had ≥85% assignment to their breed of registration using ADMIXTURE.

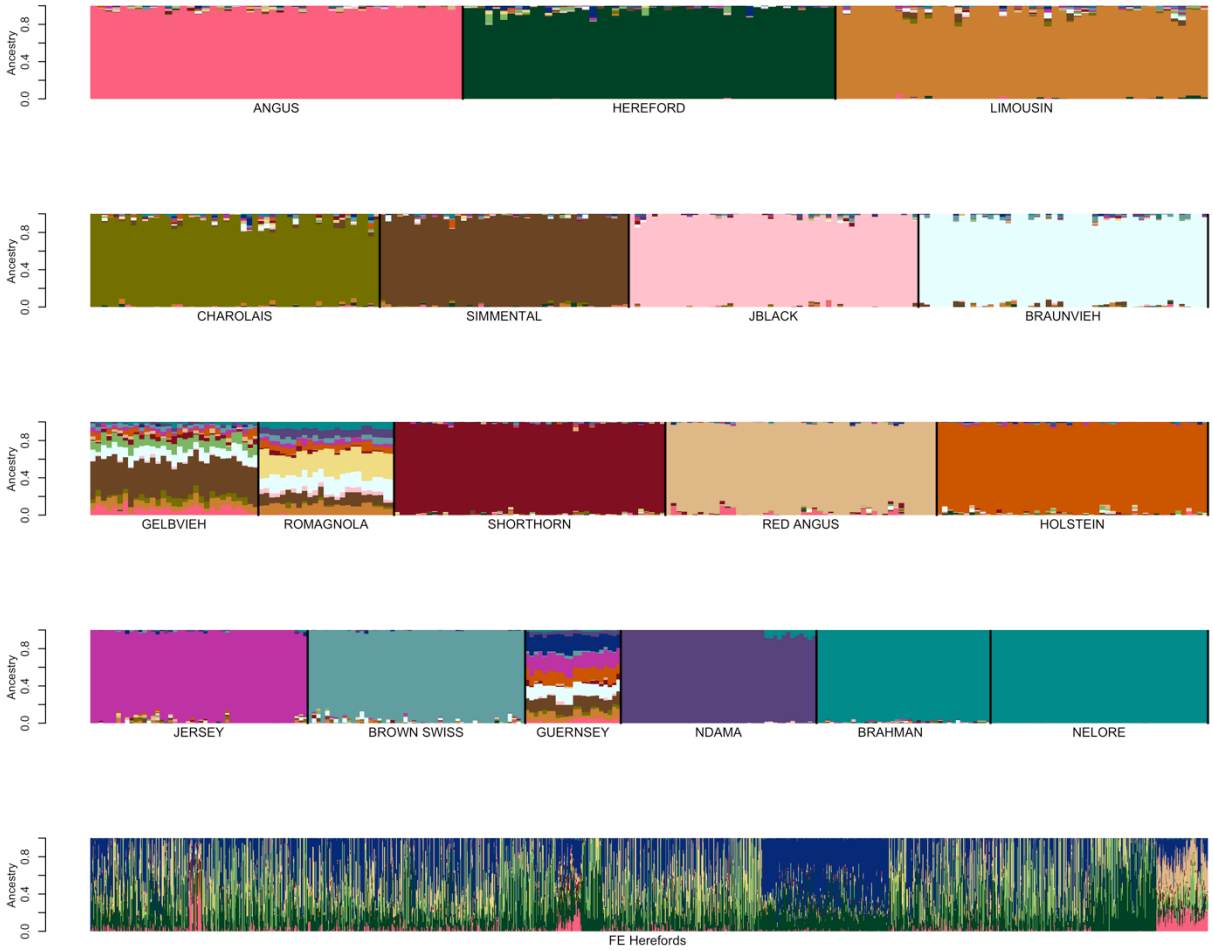


Fig. 14 ADMIXTURE analysis conducted using the same data as shown in Figure 11 (first four rows), merged with an additional 2,005 high percentage crossbred Hereford target individuals (last row).

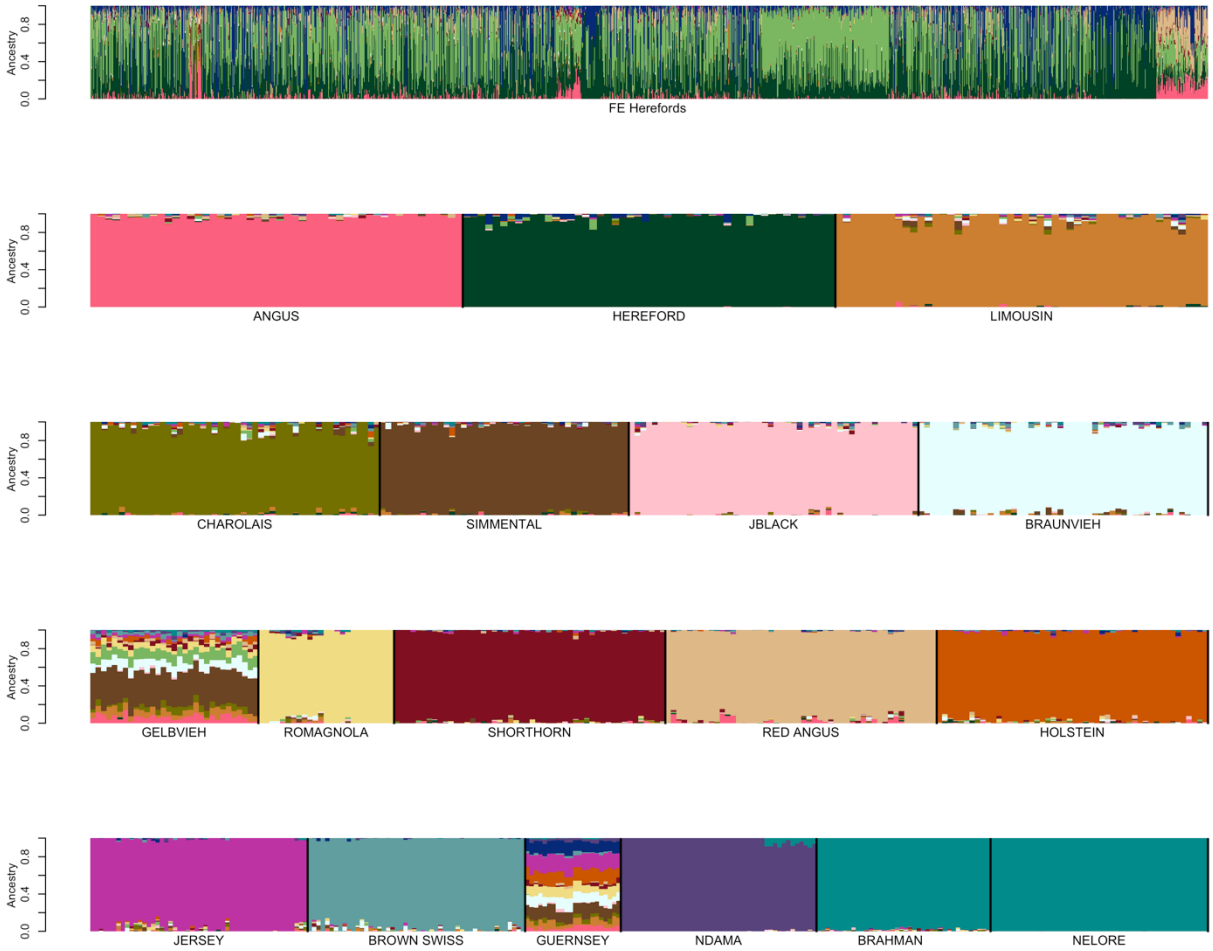


Fig. 15 ADMIXTURE analysis conducted using the same data as shown in Fig. 12 where the reference individuals appear before the 2,005 Hereford crossbred individuals in the input file. Here, the 2,005 Hereford crossbred individuals appear before the reference individuals in the input file. The first row represents the 2005 Hereford crossbred samples. Rows 2 to 5 show the reference panel individuals.

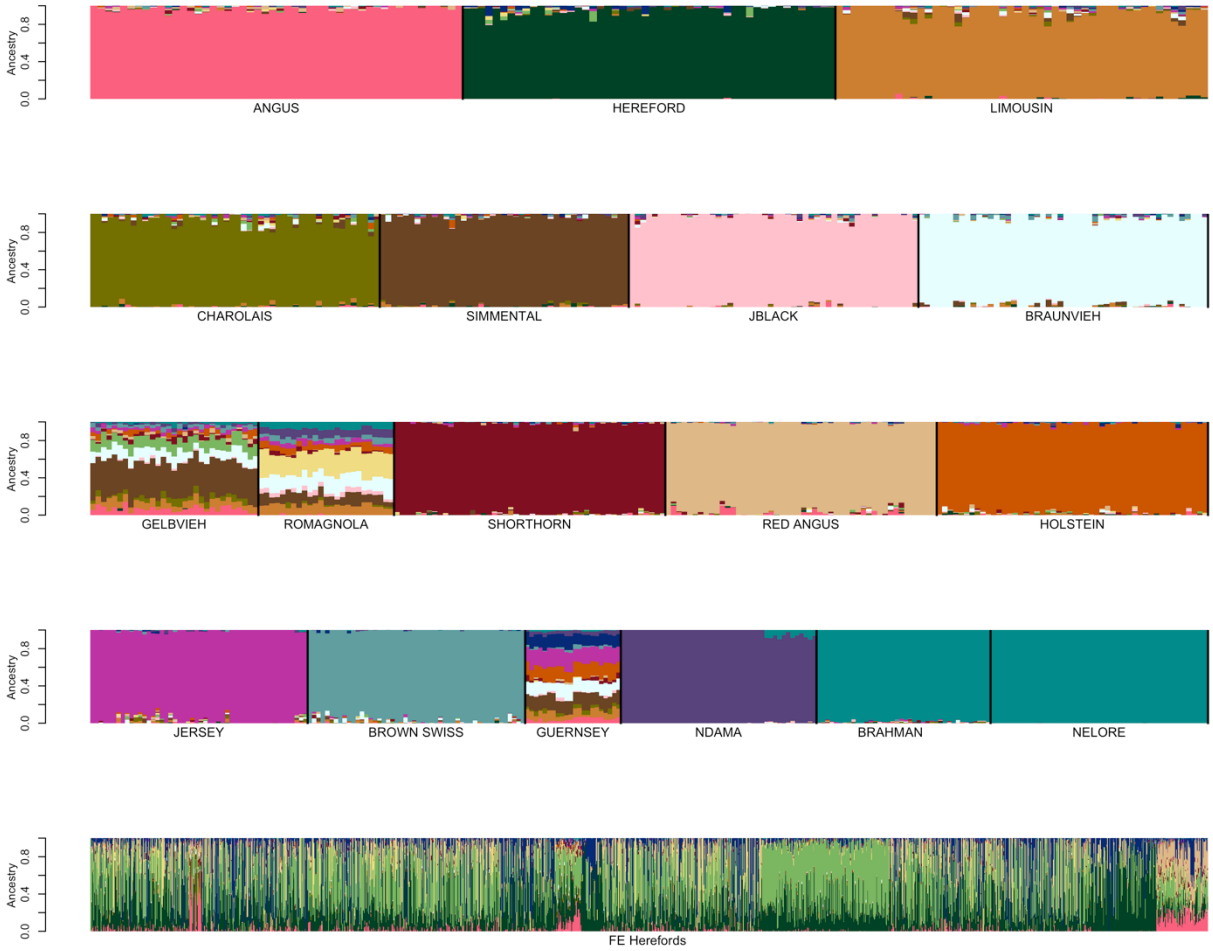


Fig. 16 ADMIXTURE analysis conducted using the same data as shown in Figs. 12 and 13, but with the order of the individuals in the input genotype file randomized. The first four rows represent the reference panel individuals, the fifth row shows the 2,005 Hereford crossbred animals.

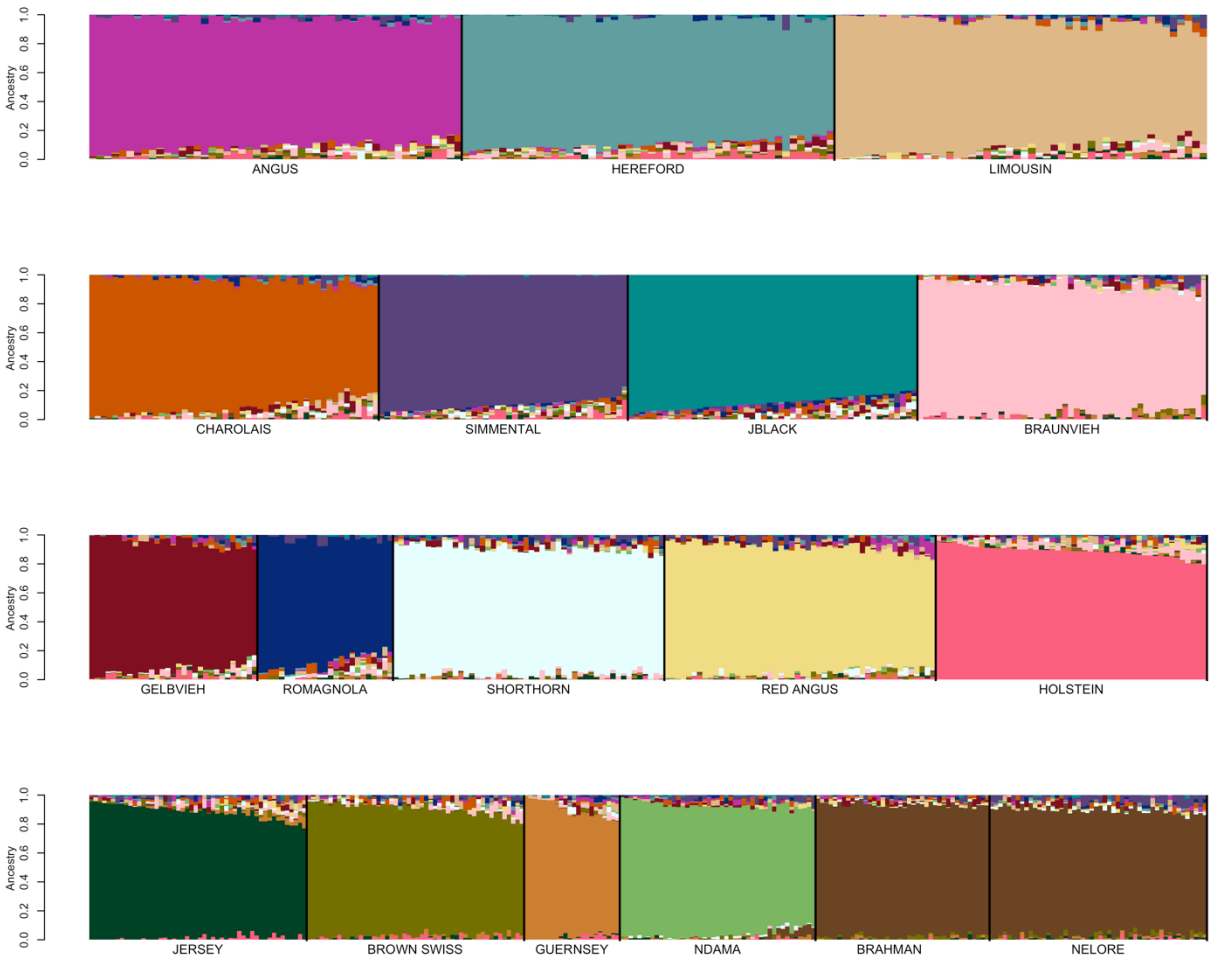


Fig. 17 Reference breed panel constructed by the random sampling of ≤50 individuals per breed from individuals with ≥85% ancestry was self-assigned to reference breed ancestry using the BC6K marker set.