Estimation of genome-wide and locus-specific breed composition in pigs[1]

Scott A. Funkhouser,* Ronald O. Bates,† Catherine W. Ernst,† Doug Newcom,‡ and Juan Pedro Steibel†[2]

*Genetics Graduate Program, Michigan State University, East Lansing 48824; †Department of Animal Science, Michigan State University, East Lansing 48824; and ‡National Swine Registry, West Lafayette, IN 47906

ABSTRACT: Advances in pig genomic technologies enable implementation of new methods to estimate breed composition, allowing innovative and efficient ways to evaluate and ensure breed and line background. Existing methods to test for homozygosity at key loci involve test mating the animal in question and observing phenotypic patterns among offspring, requiring extensive resources. In this study, whole-genome pig DNA microarray data from over 8,000 SNP was used to profile the composition of U.S. registered purebred pigs using a refined linear regression method that enhances the interpretation of coefficients. In a simulation analysis, a strong correlation between true and estimated breed composition was observed (R^2 = 0.94). Applying these methods to 930 Yorkshire animals registered with the National Swine Registry, 95% were estimated to have a "genome-wide" Yorkshire breed composition of at least 0.825 or 82.5%, with similar performance for evaluating datasets of registered Duroc (n = 88) Landrace (n = 129), and Hampshire (n = 17) breeds. We also developed new methods to evaluate locus-based breed probabilities. Such methods have been applied to multi-locus SNP genotypes flanking the *KIT* gene known to predominantly control coat color, thereby inferring the probability that an animal has haplotypes in the *KIT* region that are predominant in white breeds. These methods have been adopted by the National Swine Registry as a means to identify purebred Yorkshire animals.

Key words: breed composition, *KIT*, pigs

Transl. Anim. Sci. 2017.1:36–44
doi:10.2527/tas2016.0003

INTRODUCTION

Breed registries are used to develop and maintain certain conformational, performance and coat color characteristics for a specified population, along with cataloging the pedigree of every animal that is approved for registry within that breed. Purebred Yorkshire pigs, distinguished by their white coat color and erect ears, along with their reputation for superior maternal performance (Buchanan and Stalder, 2011), are registered in the National Swine Registry (NSR). Test matings are traditionally carried out to determine if a Yorkshire boar will only sire white progeny and thus can be safely assumed homozygous for the dominant white allele of the *KIT* gene (Marklund et al., 1998, Giuffra et al., 1999). This procedure requires substantial time and resources, so the swine industry has considered the potential use of genomic data to efficiently estimate the breed composition of selection candidates. SNP genotypes may be used to estimate the presence of the dominant white allele, and genome-wide genotypes may be used to indicate an animal's overall breed composition (Huang et al., 2014).

In this study, we further develop methodology that quantifies overall breed composition of an animal using genome-wide data, as well as the means to estimate the probability that the animal has haplotypes in the KIT region that are unique or predominate in white breeds using a limited number of SNP genotypes around *KIT*. Both methods, used in combination, have been adopted

[1]This project was supported by the National Swine Registry; Part of the genotyping for this study was funded by Agriculture and Food Research Initiative Competitive Grant No. 2010-65205-20342 from the USDA National Institute of Food and Agriculture.

[2]Corresponding author: steibelj@msu.edu

Received September 27, 2016.
Accepted October 18, 2016.

by the NSR to screen purebred Yorkshire pigs by requiring they possess a minimum Yorkshire genome-wide breed composition (GWBC) and minimum "*KIT*-based breed probability" (KBP) of being a homozygous white animal. Use of genotype information to estimate breed composition requires a fraction of the expense of traditional test matings, may remove errors due to subjectivity or human error, and integrates well with other common uses of genomic technologies such as genomic prediction and parentage testing.

MATERIALS AND METHODS

Animal Genotyping Datasets

Genotyping data was provided by the National Swine Registry or the USDA Meat Animal Research Center (MARC), or was obtained from previous studies (Institutional Animal Care and Use Committee approval AUF# 03/09-046-00).

All analyses used genotypes obtained from a panel of purebred reference animals to interpret the genome-wide or locus-based genotypes of a test animal. Purebred animals were chosen to widely represent breeds registered with the NSR (Yorkshire, Landrace, Hampshire, and Duroc), and with no common ancestors in the last 2 generations. This resource provided 74 Duroc, 68 Hampshire, 64 Landrace, and 75 Yorkshire pigs used as reference animals (Badke et al., 2012). Genome-wide breed composition and KBP methods were tested on datasets of purebred animals provided by USDA-MARC (*MARCduroc, n* = 88*; MARChampshire, n* = 17*; MARClandrace*, *n* = 65*; and MARCyorkshire, n* = 113), purebred Yorkshire provided by NSR (*YorkshirePure, n* = 930), Yorkshire barrows and showpigs provided by NSR (*ShowPigs, n* = 26) and known Yorkshire crossbred animals provided by NSR (*YorkhsireCross*, *n* = 12). All animals were genotyped using the Illumina PorcineSNP60 BeadChip (SNP60) version 2 (Illumina, San Diego, CA). Physical positions of SNPs in the Sus scrofa reference genome version 10.2 were obtained from the Pig Genome Data Repository (http://www.animalgenome.org/repository/pig/).

Genome-wide breed composition (GWBC)

In order to estimate the genome-wide breed composition of a test animal, a linear regression is performed similarly to previous studies in bulls (Kuehn et al., 2011) and in pigs (Huang et al., 2014). In this approach a test animal's genotypes are regressed onto allele frequencies derived from the reference animals. For this analysis, only those single nucleotide polymorphisms (SNP) present on the low-density GeneSeek Genomic Profiler for Porcine LD (GGP-LD; GeneSeek a Neogen Corp., Lincoln, NE) platform consisting of 8,826 SNP are considered for downstream use. Single nucleotide polymorphisms are further removed using the following procedure: genotypes from all reference animals are assembled, and SNP are removed if they have genotype call rate of 0.9 or less or are fixed across all reference animals, resulting in 8,341 SNP. In matrix form the regression can be represented as:

$$y = X\beta + \varepsilon \qquad [1]$$

Where **y** is a vector of length 8,341 containing a single test animal's genotypes for each filtered SNP, expressed as the dosage of the "B" allele divided by 2 (0, 0.5, 1), $\boldsymbol{X}$ is a 8,341 × 4 matrix with allele frequencies (frequency of the "B" allele) of each filtered SNP (in rows) derived from each reference breed (in columns), and $\boldsymbol{\epsilon}$ is a vector of residuals for each SNP assumed to be individually and identically distributed. A vector of estimated genome-wide breed compositions $\hat{\beta}$ is obtained using ordinary least squares (OLS):

$$\hat{\boldsymbol{\beta}} = \min_{\beta}\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \qquad [2]$$

Unlike previous studies that have used this regression method for estimating breed composition (Huang et al., 2014; Kuehn et al., 2011), we put a set of linear constraints on the solutions for $\hat{\beta}$ so that $\sum_{i=1}^{4}\hat{\beta}_i = 1$ and each $\hat{\beta}_i$ is between 0 and 1. This restriction was enforced by minimizing (2) using quadratic programming as implemented in the quadprog R package (Weingessel, 2013). The consequence of using this type of constrained OLS solution for Eq. 2 is that the estimated coefficients can be interpreted as "% of composition" for each breed. For demonstration, when a Yorkshire animal is tested using this method, its GWBC results ($\hat{\beta}'$) may read 0.01, 0.02, 0.02, 0.95 (or any non-negative set of values that sum to 1). Such results would indicate that the genome of the tested animal is estimated to be composed of 1% Duroc, 2% Hampshire, 2% Landrace, and 95% Yorkshire.

Simulation to assess GWBC accuracy

We assessed the performance of GWBC using a simulation study. In short, genotype blocks were randomly sampled from purebred animals in the MARC datasets (*MARCduroc, MARChampshire, MARClandrace* or *MARCyorkshire*) and stitched together to form a new "synthetic mosaic" genome

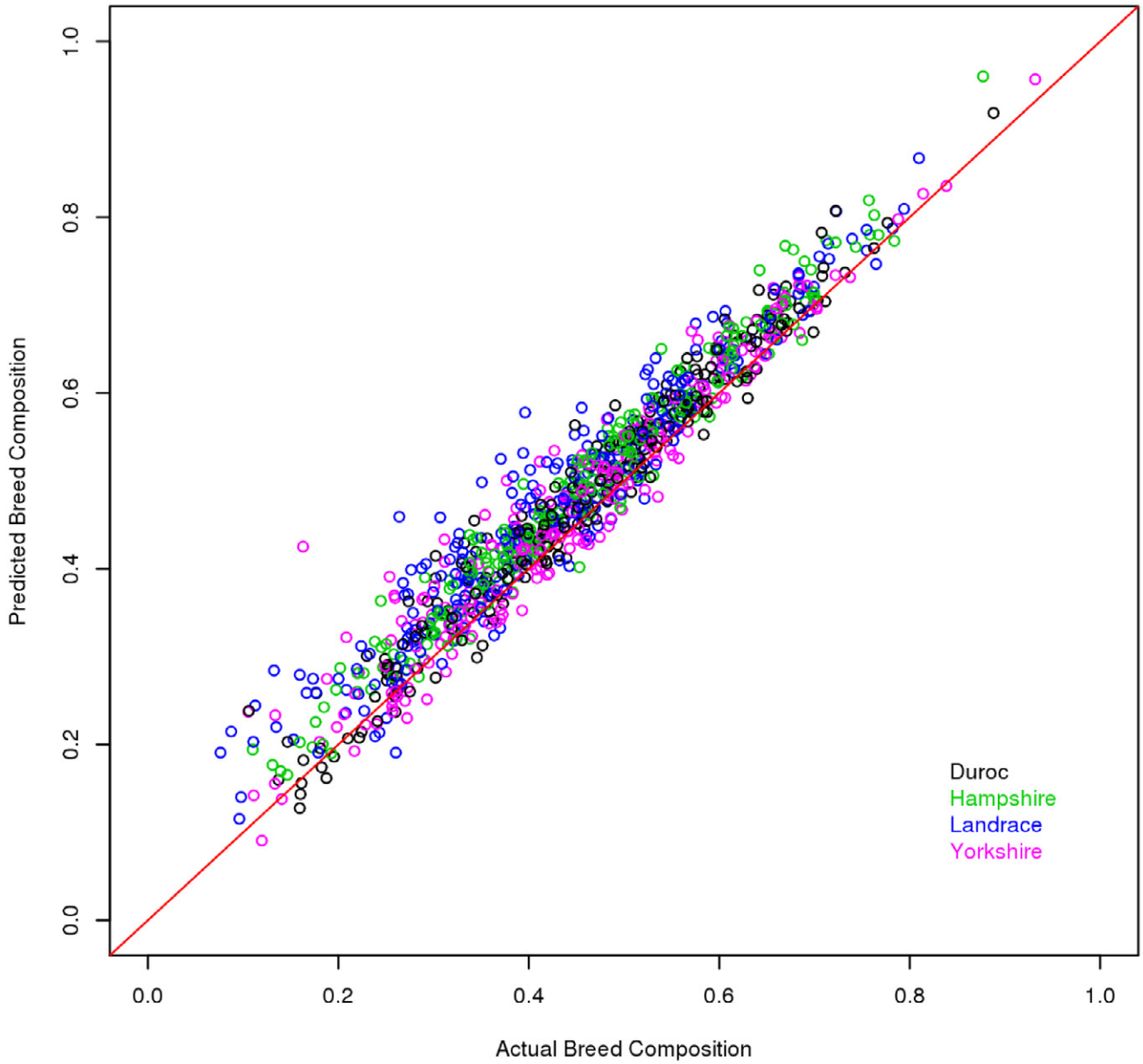

Figure 1. Estimated genome-wide breed composition compared to actual genome-wide breed composition of 1,000 simulated mosaic genomes. Color of each point indicates the breed of 1 of the parent genotypes used to construct the hybrid, and the composition that is estimated for that breed (the rest could be any other breed). Red line has a y-intercept of 0 and slope of 1.

(Supplemental Figure 1) of known breed composition. We replicated the simulation 1,000 times to create as many "synthetic mosaics" and we compared their estimated composition obtained using GWBC to their true breed composition (Fig. 1). Our proposed algorithm to estimate GWBC appears to accurately reflect true composition, as the observed correlation between actual and estimated breed composition was 0.97.

To test how GWBC performs when used to estimate the composition of an animal with ancestry not represented in the reference panel, we performed a similar simulation involving Pietrain animals from the Michigan State University resource population dataset (Edwards et al., 2008). A random Pietrain animal was chosen and "hybridized" in the same way as before with a random animal from one of the MARC datasets. Again this process was repeated 1,000 times and actual non-Pietrain breed composition was compared to non-Pietrain composition estimated with GWBC (Supplemental Figure 2). Again the correlation between actual and estimated breed composition remained high (r = 0.86), but considerable bias was observed in that Yorkshire and Landrace compositions were overestimated. The same is true for the Duroc and Hampshire breeds but to a lesser degree. Using these 1000 simulated Pietrain hybrids, we observed that the

regression model fit (R^2) increases as the proportion of Pietrain in the hybrids decreases, indicating that the presence of Pietrain in the test animal reduces model fit (Supplemental Figure 3). Together, these simulations suggest that GWBC performs well as long as the animal being tested is composed of the breeds included in the reference panel. If this is not true, GWBC estimates may be upwardly biased, likely due to the fact that any non-reference breed within the test animal is forced to be a reference breed, per the GWBC algorithm. Moreover, in some cases, the R^2 statistic of a regression could signal at this possibility, as the average R^2 tends to be lower than when all ancestral breeds are part of the reference panel.

Locus-specific breed probability applied to KIT

To estimate breed probabilities for multi-SNP genotypes surrounding the *KIT* locus, we first phased SNP around *KIT* to construct a reference haplotype database. This was accomplished by phasing all SNPs in the Illumina PorcineSNP60 that mapped to Porcine Chromosome 8 using FImpute software (Sargolzaei et al., 2014). These haplotypes were expected to be highly accurate because the reference animals and their progeny possess genotypic data for use in pedigree-based phasing of genotypes (Badke et al., 2013). After chromosome-wide phasing, SNP-haplotypes of seven SNP surrounding *KIT* (SSC 8:43-44Mb; ALGA0047798, ALGA0047807, ALGA0047809, ALGA0102731, ALGA0115258, ALGA0123881, MARC0034580) were extracted, with physical positions SSC 8:43730377, SSC 8:43878303, SSC 8:43916646, SSC 8:43462399, SSC 8:43651639, SSC 8:43068687, and SSC 8:43425758, respectively. Moreover, these seven SNP can be genotyped using either the Illumina PorcineSNP60 or the GeneSeek Genomic Profiler for Porcine LD arrays. Haplotype frequencies were then estimated for each haplotype in each breed (Table 1). Next, the probability of observing a particular test animal genotype is estimated for each breed [$P(genotype \mid breed)$] by assuming independent association of haplotypes and summing over all products of relative frequencies of haplotype combinations that lead to the genotype in question. Finally, for each possible genotype we compute breed probabilities [$P(genotype \mid breed)$] using Bayes theorem: $P(breed \mid genotype) \propto P(breed)\ P(genotype \mid breed)$. We assumed a "flat prior":

$$P(breed) = \frac{1}{Number\ of\ breeds}.$$

Table 1. Haplotype frequencies for each reference breed for the *KIT* region

Haplotype[1]	Landrace	Yorkshire	Hampshire	Duroc
0000011	0.016	0.007	-	-
0001001	0.016	-	-	-
0001011	-	0.007	0.007	-
0010001	0.063	0.073	-	-
0010011	0.164	0.333	0.007	-
0010101	-	-	0.007	-
0010111	-	-	0.007	-
0011001	0.219	0.133	-	-
0011011	0.156	0.320	0.888	-
0101011	-	0.007	-	-
1000001	0.031	-	-	-
1000011	-	-	0.007	-
1001001	0.023	-	-	-
1010001	0.008	-	-	-
1010010	-	-	-	0.007
1010011	0.055	-	0.067	0.171
1010111	-	-	-	0.007
1011001	0.039	0.013	-	-
1011011	0.023	-	-	0.021
1011111	-	-	0.007	0.623
1100001	0.102	0.013	-	-
1100011	-	0.007	-	0.014
1101001	0.086	0.087	-	-
1101111	-	-	-	0.158

[1]SNP-haplotype composed of 7 loci within SSC 8:43-44Mb.

Technically speaking, generating P(breed|genotype) from a h × b haplotype frequency matrix $\boldsymbol{H}$:

$$\boldsymbol{H} = \begin{pmatrix} hap_{ij} & \cdots & hap_{ib} \\ \vdots & \ddots & \vdots \\ hap_{hj} & \cdots & hap_{hb} \end{pmatrix} \quad [3]$$

where hap_{ij} is the haplotype frequency of the ith haplotype within the jth breed, was performed using the following operations:

1) For each pairwise breed combination j and j', including when $j = j'$:

a) Genotype probabilities between breed j and j' were calculated with $H_{*j}H'_{*j'}$ with H_{*j} denoting the jth column of $\boldsymbol{H}$.

b) Probabilities associated with the same genotype within the resulting matrix $H_{*j}H'_{*j'}$ are summed so that each genotype has only one probability.

2) Genotype probabilities for all breed combinations j and j' were assembled into a g × c matrix of genotype probabilities $\mathbf{G}$[P($genotype \mid breed$)]:

$$\boldsymbol{G} = \begin{pmatrix} gen_{ij} & \cdots & gen_{ic} \\ \vdots & \ddots & \vdots \\ gen_{gj} & \cdots & gen_{gc} \end{pmatrix} \quad [4]$$

where g is the number of unique genotypes that can be computed from the haplotypes in $\boldsymbol{H}$ and c is the number of within breed and between breed combinations. Using our reference panel of 74 Duroc, 68 Hampshire, 64 Landrace, and 75 Yorkshire animals, $g = 221$ and $c = 10$.

3) Each row of the $g \times c$ breed probability matrix $\boldsymbol{B}$ [P(*breed* | *genotype*)] is computed, where $\boldsymbol{B}_{i*} = \boldsymbol{G}_{i*} / \sum \boldsymbol{G}_{i*}$. At this stage, each row of the $\boldsymbol{B}$ matrix sums to 1 and contains breed probabilities associated with the ith genotype made possible from the reference haplotypes.

Obtaining KBP results for a test animal is then straightforward; assuming the test animal's multi-locus genotype (using the same SNPs within SSC 8: 43-44Mb) corresponds to the ith row of $\boldsymbol{B}$, then B_i* contains KBP results for the test animal:

$$\boldsymbol{B}_{i*} = [d_i \;\; h_i \;\; l_i \;\; y_i \;\; (d\_h)_i \;\; (d\_l)_i \;\; (d\_y)_i \;\; (h\_l)_i \;\; (h\_y)_I \;\; (l\_y)_i] \quad [5]$$

Where d_i, h_i, l_i, and y_i represent the probability that the test animal is homozygous for haplotypes of Duroc, Hampshire, Landrace, and Yorkshire origin (respectively) in the region of *KIT*. Likewise $(d\_h)_i$ and $(l\_y)_i$ represent the probabilities that the animal has a haplotype of Duroc origin and a haplotype of Hampshire origin, and a haplotype of Landrace origin and a haplotype of Yorkshire origin, respectively, in the region of *KIT*. Homozygous origin and heterozygous origin probabilities involving Yorkshire and Landrace are summed accordingly to provide "homozygous white" probabilities. In this way,

$$\boldsymbol{B}_{i*} = [d_i \;\; h_i \;\; w_i \;\; (d\_h)_i \;\; (d\_w)_i \;\; (h\_w)_i] \quad [6]$$

Where w_i is the probability that the 2 haplotypes in the *KIT* region are of Yorkshire or Landrace origin. If a test animal's multi-locus genotype cannot be found in $\boldsymbol{B}$, then KBP results for the animal cannot be computed with the given reference panel. This may indicate a genotyping error or a haplotype combination not well represented in the reference panel. If an animal has missing genotype calls in its multi-locus *KIT* genotype, then we provided KBP results that are the average across all possible genotypes that the test animal could have.

GWBC and KBP implementation

Our implementation of both GWBC and KBP are included in the package *breedTools.* Please see installation and usage instructions online at: https://github.com/funkhou9/breedTools. Reference panel data used to generate the results in this paper are included in the package. Functions include *solve_composition()* to calculate GWBCs from a GWBC reference panel and a matrix of test animal genotypes, *build_KBP()* to calculate a new KBP reference panel (breed probability matrix), and *screen_purity()*, a wrapper function used to provide GWBC and KBP results for any number of input genotypes.

Table 2. The 0.05 and 0.95 quantiles for each GWBC among each population tested

Dataset	n[1]	Duroc[2]	Hampshire[3]	Landrace[4]	Yorkshire[5]
YorkshirePure	930	(0.000, 0.055)	(0.000, 0.054)	(0.000, 0.113)	(0.825, 1.000)
Showpigs	26	(0.000, 0.033)	(0.002, 0.387)	(0.000, 0.037)	(0.572, 0.998)
MARCduroc	88	(0.907, 1.000)	(0.000, 0.035)	(0.000, 0.039)	(0.000, 0.041)
MARChampshire	17	(0.000, 0.013)	(0.946, 1.000)	(0.000, 0.054)	(0.000, 0.014)
MARClandrace	65	(0.000, 0.062)	(0.000, 0.066)	(0.811, 1.000)	(0.000, 0.161)
MARCyorkshire	113	(0.000, 0.076)	(0.000, 0.044)	(0.000, 0.111)	(0.816, 1.000)

[1]Sample size.

[2–5]Quantiles for β_1, β_2, β_3, and β_4 from GWBC regression, represented as (0.05 quantile, 0.95 quantile).

RESULTS AND DISCUSSION

GWBC testing

We applied our GWBC constrained regression method to several real datasets: *YorkhsirePure, Showpigs, MARCduroc, MARChampshire, MARClandrace, and MARCyorkshire* (Table 2). These datasets include purebred and crossbred animals of known breed composition as well as Showpigs that phenotypically look like Yorkshire pigs but may or may not be registered. For non-registered animals that phenotypically resemble Yorkshire, their actual breed composition is not certain but is expected to include significant Yorkshire ancestry. For each animal in each dataset we calculated breed compositions for 4 reference breeds (i.e., calculated all 4 GWBC regression coefficients). As expected, the breed identity within each dataset seems to be accurately estimated with GWBC. For instance, 95% of purebred commercial Yorkshire pigs (*YorkshirePure, MARCyorkshire*) have Yorkshire GWBC estimates of at least 0.825 and 0.816, respectively. Likewise, GWBC appears to perform similarly for Duroc, Landrace, and Hampshire datasets; for *MARCduroc, MARChampshire,* and *MARClandrace* datasets, 95% of animals have at least a 90.7% Duroc, 94.6% Hampshire, and 81.1% Landrace composition, respectively.

To further evaluate a distribution of Yorkshire GWBC estimates and R^2 values, we inspected both distributions

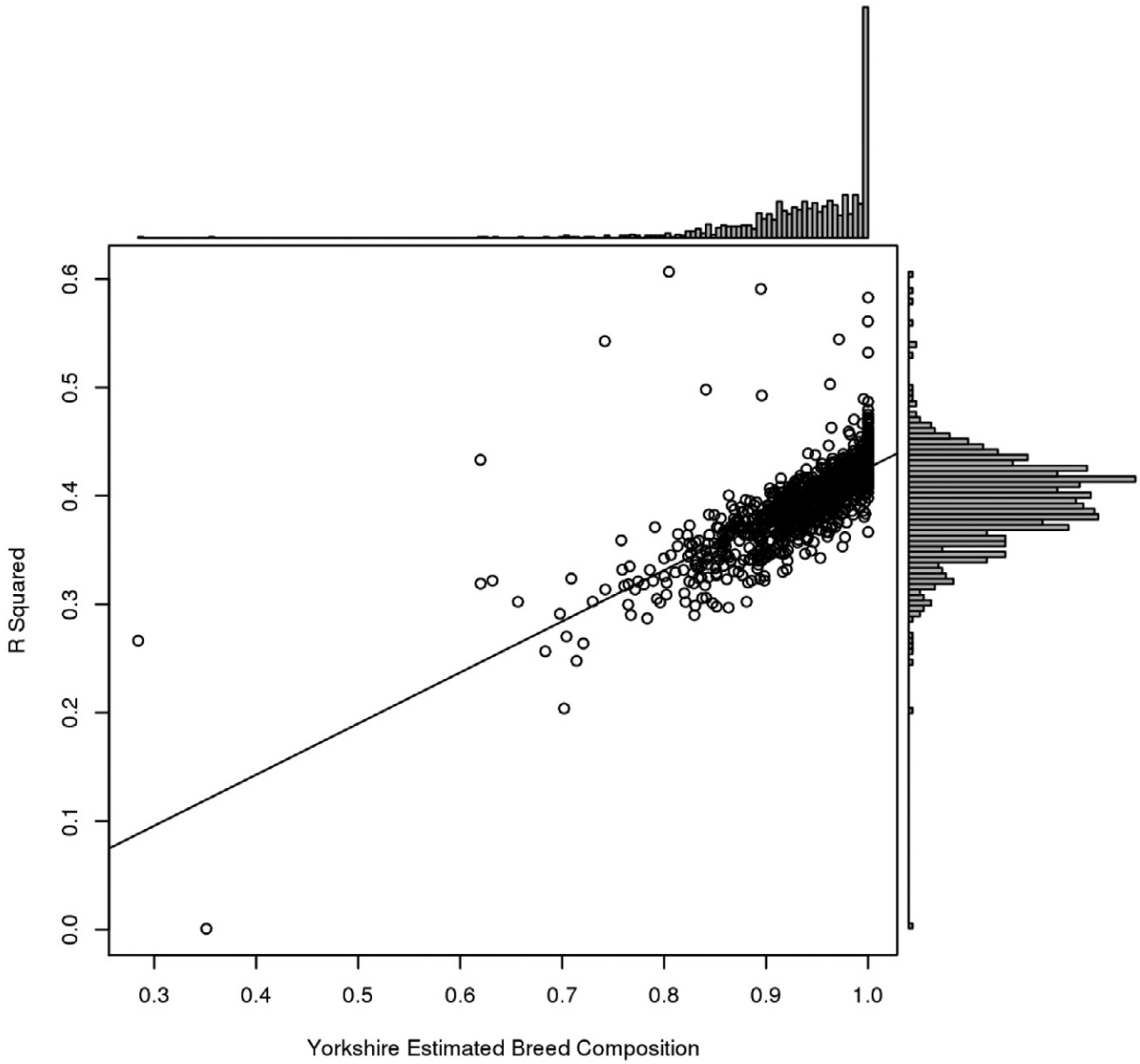

Figure 2. Joint distribution of R^2 values and estimated genome-wide breed compositions from the *YorkshirePure* dataset. Marginal distributions of each shown on axes.

using the *YorkshirePure* dataset consisting of 930 registered Yorkhsire animals (Fig. 2). The marginal distribution of Yorkshire GWBC values indicates that most animals have a Yorkshire GWBC of 1.0 (100% Yorkshire), and that the distribution is negatively skewed. Two animals have surprisingly low Yorkshire GWBCs of 0.351 and 0.284. While there is no obvious reason for this, these GWBC estimates suggest that the breed ancestry of these 2 boars may be in question. All other animals have Yorkshire GWBCs higher than 0.6. A noticeable feature of GWBC is that R2 values are roughly normally distributed around 0.4. These low R2 values overall may be a consequence of regressing discrete values (test animal genotypes) onto decimal values (reference animal allele frequencies). Nevertheless, the R2 values may still indicate relative differences in model predictive accuracy between test animals (in other words, relative differences in the appropriateness of the reference panel in evaluating the test animal).

The *YorkshireCross* dataset, containing 12 animals that are known Yorkshire crossbred animals, was evaluated separately (Table 3). These animals either possessed non-white (non-Yorkshire) phenotypes such as dark spots on top of white coat color or appeared as Yorkshire but were known to fail a test mating by producing litters with non-white piglets. Note that for each animal, Duroc GWBC, Hampshire GWBC, Landrace GWBC and Yorkshire GWBC sum to 1, as designed by our regression method. Although the average Yorkshire

Table 3. GWBC estimates for each animal in the *YorkshireCross* dataset

Test Animal	Duroc[1]	Hampshire[2]	Landrace[3]	Yorkshire[4]	R squared
A	0.075	0.145	0.114	0.665	0.328
B	0.008	0.199	0.020	0.773	0.359
C	0.000	0.248	0.000	0.752	0.376
D	0.000	0.396	0.017	0.587	0.380
E[5]	0.012	0.068	0.000	0.920	0.385
F	0.000	0.134	0.000	0.866	0.388
G	0.000	0.336	0.026	0.638	0.390
H	0.033	0.126	0.000	0.841	0.411
I	0.000	0.273	0.006	0.721	0.423
J[5]	0.000	0.000	0.000	1.000	0.445
K	0.112	0.235	0.108	0.545	0.463
L[5]	0.000	0.054	0.000	0.946	0.477

[1–4]Estimates for β_1, β_2, β_3, and β_4 from GWBC regression.

[5]Animals with high Yorkhsire GWBC despite being known crossbreds.

GWBC of these animals is 0.77, 3 of these animals have Yorkhsire GWBCs greater than 0.9. Should GWBC alone be used to screen purebred animals, it would likely incorrectly identify such animals as purebred. Theoretically, these animals may be predominantly Yorkshire, but may possess non-Yorkshire alleles at major-effect QTL that determine color patterns.

KBP testing

Recognizing that GWBC operates well genome-wide, but it has low resolution, we recommend that GWBC be supplemented with additional locus-based tests around QTL known to have strong effects on breed-distinguishing phenotypes. We developed an additional procedure called *KIT*-based breed probability (KBP) testing, which can estimate the probability that a test animal has haplotypes in the SNP surrounding the *KIT locus* that are unique or most frequent among white breeds.

Application of this method to real datasets provides similar performance as GWBC (Table 4), in that 688 of 836 *YorkshirePure* animals with a complete set of genotyped SNPs in the *KIT* region (SSC 8:43-44Mb) possess a "white" KBP value of 0.6 or greater, indicating that such animals have at least a 0.6 probability of having both haplotypes of Yorkshire or Landrace origin (either white breed) in the region of *KIT*. Since these are probabilities, the sum of all test animal breed probabilities for each animal will be 1. Inspecting KBP results from each animal in the *YorkshireCross* dataset (Table 5) reveals potential uses of KBP as a purity screening procedure. Although 3 of these animals had Yorkshire GWBC greater than 0.9, 2 of these animals possess "white" KBP of 0.446 while the third has a "white"

Table 4. The 0.05 and 0.95 quantiles for each KBP among each population tested

Dataset	n[1]	Duroc[2]	Hampshire[3]	White[4]	Duroc/Hampshire[5]	Duroc/White[6]	Hampshire/White[7]
YorkshirePure	836	(0.00, 0.00)	(0.00, 0.02)	(0.44, 1.00)	(0.00, 0.00)	(0.00, 0.02)	(0.00, 0.54)
MARClandrace	48	(0.00, 0.00)	(0.00, 0.02)	(0.27, 1.00)	(0.00, 0.20)	(0.00, 0.37)	(0.00, 0.54)
MARChampshire	15	(0.00, 0.00)	(0.24, 0.60)	(0.07, 0.13)	(0.00, 0.31)	(0.00, 0.19)	(0.16, 0.31)
MARCyorkshire	106	(0.00, 0.00)	(0.00, 0.02)	(0.44, 1.00)	(0.00, 0.00)	(0.00, 0.34)	(0.00, 0.55)
MARCduroc	80	(0.73, 0.99)	(0.00, 0.00)	(0.00, 0.00)	(0.01, 0.15)	(0.00, 0.12)	(0.00, 0.00)

[1]Sample size. In each dataset, only animals with a complete set of genotypes near *KIT* (SSC 8:43-44Mb) are represented.

[2–7]Quantiles for *KIT*-based breed probabilities, represented as (0.05 quantile, 0.95 quantile). "White" indicates the probability of being Yorkshire or Landrace in the region of *KIT*. "Duroc/Hampshire" indicates the probability of being a Duroc-Hampshire hybrid in the region of *KIT*.

Table 5. KBP estimates for each animal in the *YorkshireCross* dataset

Test Animal	Duroc[1]	Hampshire[2]	White[3]	Duroc/Hampshire[4]	Duroc/White[5]	Hampshire/White[6]
K[7]	0.000	0.000	0.568	0.000	0.150	0.282
C	0.000	0.000	0.178	0.382	0.241	0.200
E[8]	0.000	0.016	0.446	0.000	0.000	0.538
D	0.000	0.016	0.446	0.000	0.000	0.538
B	0.000	0.682	0.000	0.000	0.000	0.318
F	0.000	0.682	0.000	0.000	0.000	0.318
H	0.000	0.682	0.000	0.000	0.000	0.318
I	0.000	0.016	0.446	0.000	0.000	0.538
L[8]	0.000	0.002	0.768	0.000	0.000	0.229
J[8]	0.000	0.016	0.446	0.000	0.000	0.538
G	0.000	0.016	0.446	0.000	0.000	0.538
A[7]	0.000	0.000	1.000	0.000	0.000	0.000

[1–6]*KIT*-based breed probabilities. "White" indicates the probability of being Yorkshire or Landrace in the region of *KIT*. "Duroc/Hampshire" indicates the probability of being a Duroc-Hampshire hybrid in the region of *KIT*.

[7]Animals were missing at least one genotype within SSC 8:43-44Mb, so KBP results are the average across all possible values.

[8]Animals that had high (> 0.9) estimated Yorkshire GWBC.

KBP of 0.768. Together, both GWBC and KBP analyses provide evidence that animals in the *YorkshireCross* dataset are not confidently Yorkshire genome-wide and "white" in the region of *KIT*.

Implementation for screening purebred Yorkshire

Once we developed the methodology (GWBC and KBP) to estimate breed compositions and breed probabilities around specific loci, the NSR has been instrumental in deciding how such tools should be used to screen purebred animals. With NSR, we have pursued these methodologies as screening procedures for the Yorkshire breed, whereby certain Yorkshire GWBC and white KBP thresholds would need to be met for an animal to pass as purebred Yorkshire. Since initial testing of GWBC and KBP methods using the reference panel presented here, we have expanded the number of reference animals of each breed. The expanded reference panel currently consists of 806 Yorkshire, 129 Landrace, 159 Duroc, and 85 Hampshire, made possible by including animals from the *MARC* datasets and a subset of *YorkshirePure* animals (those whose sire is not genotyped or those that have at least 1 genotyped progeny). This expanded reference panel has been used to estimate GWBC of newly genotyped Yorkshire animals, while the original reference panel has been used to estimate KBP values.

The joint use of GWBC and KBP could be illustrated using results from 78 animals (Supplementary Table 1) composed of 69 registered Yorkshire and 9 impure animals (either Landrace or animals known to have colored progeny). For instance, if thresholds of 0.9 Yorkshire GWBC and 0.6 white KBP, whereby an animal would need to pass both thresholds to be considered purebred Yorkshire. Among the 52 animals that passed both thresholds, only 1 was known to be impure, having previously farrowed colored pigs. Among the remaining 26 animals that failed to pass both thresholds were 18 registered Yorkshire animals, 15 of which were either Yorkshires imported into the US or sons of such imported Yorkshires. It is possible that imported animals may tend to fail thresholds that U.S. animals pass if they represent a subpopulation of Yorkshire that is underrepresented among reference animals. Subpopulations may differ in allele and haplotype frequencies, highlighting the importance of having a representative reference panel that captures the genetic variation existing among animals to be tested. Sometimes the results of the GWBC or KBP suggest possible reference panel deficiencies. For instance, the R^2 associated with GWBC may be abnormally low, indicating that the breed-specific allelic frequencies in the reference panel do not represent well the composition of the tested animal. This may be due to population substructure or to ancestry of a completely different breed, not represented in the panel. Similarly, the multi-SNP genotype around KIT may not be obtained by any known combination of haplotypes used in KBP calculation, which also may happen if there is substantial population substructure. In all cases, checking the tested animal pedigree may offer valuable information because it may reveal that the animal's registered ancestry is not well represented in the panel. We continuously check for the presence of subpopulations among registered animals and we assess the necessity of adding more animals to growing reference panels.

Ongoing work with NSR is being done to evaluate newly genotyped Yorkshire animals in order to improve the reference panels to better represent Yorkshires to be tested, and to develop the means to screen purebred animals of other breeds than Yorkshire. An important practical consideration for implementation of the proposed methods is the presence of missing genotypes. A small proportion of missing genotypes will likely not affect the computation of GWBC because the rows corresponding to SNP with missing genotypes can be dropped from Equation 1. However, even a single missing genotype in the SNP surrounding KIT may greatly affect KBP results. If there are missing genotypes in the SNP used in KBP estimation, then there are several possible multi-locus genotypes (for instance 3 possible genotypes if one SNP has a missing genotype). These multiple possible genotypes in the *KIT* region lead to multiple KBP estimates. Our recommendation is that if all KBP estimates (associated with different genotypes) pass (or do not pass) the established thresholds, the animal will not require re-genotyping because all genotypes lead to the same conclusion. However, if some outcomes lead to KBP values that fail to pass thresholds while others lead to passing thresholds, re-genotyping the animal is recommended.

By demonstrating new techniques to evaluate breed composition, we believe that such techniques may be used to save resources and provide new metrics with which to evaluate purity of animals. Moreover, all of this work is being done using the same genomic resources that could be used for genomic evaluations and parentage testing, thus adding value to the genotypes and minimizing further testing that would come at an extra cost to the breeders. Both GWBC and KBP methods are simple to implement and are efficient and suitable for routine use. The *KIT*-based breed probability approach may be adapted for other locus-specific tests at alternative loci to profile other QTL, and may be used in conjunction with GWBC for a variety of breeding or genomic evaluation objectives.

LITERATURE CITED

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics 13:24. doi:10.1186/1471-2164-13-24

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, J. Fix, C. P. Van Tassell, and J. P. Steibel. 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14:8. doi:10.1186/1471-2156-14-8

Buchanan, D. S., and K. Stalder. 2011. Breeds of pigs. In: M. F. Rothschild and A. Ruvinsky, editors, The genetics of the pig. 2nd ed. CPI Antony Rowe, Chippendam, UK. p. 445–472. doi:10.1079/9781845937560.0445

Edwards, D. B., C. W. Ernst, R. J. Tempelman, G. J. M. Rosa, N. E. Raney, M. D. Hoge, and R. O. Bates. 2008. Quantitative trait loci mapping in an F2 Duroc × Pietrain resource population: I. Growth traits. J. Anim. Sci. 86:241–253. doi:10.2527/jas.2006-625

Giuffra, E., G. Evans, A. To, R. Wales, A. Day, H. Looft, G. Plastow, and L. Andersson. 1999. The Belt mutation in pigs is an allele at the Dominant white (I/KIT) locus. Mamm. Genome 10:1132–1136. doi:10.1007/s003359901178

Huang, Y., R. O. Bates, C. W. Ernst, J. S. Fix, and J. P. Steibel. 2014. Estimation of U.S. Yorkshire breed composition using genomic data. J. Anim. Sci. 92:1395–1404. doi:10.2527/jas.2013-6907

Kuehn, L. A., J. W. Keele, G. L. Bennett, T. G. McDaneld, T. P. L. Smith, W. M. Snelling, T. S. Sonstegard, and R. M. Thallman. 2011. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. J. Anim. Sci. 89:1742–1750. doi:10.2527/jas.2013-6907

Marklund, S., J. Kijas, H. Rodriguez-Martinez, K. Funa, M. Moller, D. Lange, I. Edfors-Lilja, and L. Andersson. 1998. Pig molecular basis for the dominant white phenotype in the domestic pig. Genome Res. 8:826–833. doi: 10.1101/gr.8.8.826

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. doi:10.1186/1471-2164-15-478

Weingessel, A. 2013. Quadprog: Functions to solve Quadratic Programming Problems. R package version 1.5-5. http://CRAN.R-project.org/package=quadprog