Supplemental Material: Exclusion and genomic relatedness methods for assignment of parentage using genotyping-by-sequencing data

K. G. Dodds, J. C. McEwan, R. Brauning, T. C. van Stijn, S. J. Rowe, K. M. McEwan and S. M. Clarke

AgResearch, Invermay Agricultural Centre, Private Bag 50034, Mosgiel 9053, New Zealand

Contents

Expected mismatch rate for parent-offspring trios

Probabilities for each combination of observed and true parent-offspring trio genotypes are shown in Table S1. The probability of an apparent mismatch is the sum of the probabilities for the cases where there is a Y in the mismatch (MM) column. This simplifies to

$$p(1-p)\ (1-2p(1-p))\ [\ 2K_m + 2K_o + 2K_f - K_m K_o - K_f K_o\ ] + \ p^2(1-p)^2[\ 4K_o + 2K_m + 2K_f + 4K_f K_o + 4K_m K_o + 2K_f K_m - 12K_f K_m K_o]$$

For any particular offspring, we have the information on their apparent genotype, and therefore we need to condition on this information:
Probability(apparent mismatch | observed genotype) = P(apparent mismatch, observed genotype) / P (observed genotype)

The probabilities of each of the possible observed (progeny) genotypes are

$$P(\mathrm{AA}^*) = \ p^2 + 2p(1-p)K_o$$
$$P(\mathrm{AB}^*) = \ 2p(1-p)(1-2K_o)$$
$$P(\mathrm{BB}^*) = \ (1-p)^2 + 2p(1-p)K_o$$

$$P(\text{apparent mismatch}, \mathrm{AA}^*) = p^3(1-p)\ (\ K_m + K_f)\ (1+K_o) + \ p^2(1-p)^2[2K_o + K_m + K_f - K_f K_m + 2K_m K_o + 2K_f K_o - 2K_f K_m K_o]\ + 2p(1-p)^3 K_o$$

$$P(\text{apparent mismatch, AB}^{*}) = p(1-p)(1-2K_o)\,[(1-2p(1-p))\,(K_m+K_f)+4\,p(1-p)\,K_fK_m]$$

$$P(\text{apparent mismatch, BB}^{*}) = 2p^3(1-p)K_o + p^2(1-p)^2[2K_o+K_m+K_f-K_fK_m+2K_mK_o+2K_fK_o - 2K_fK_mK_o] + p(1-p)^3(K_m+K_f)(1+K_o)$$

Table S1. Probabilities for combinations of observed and true parent-offspring trio genotypes. Probabilities are calculated assuming parents are in Hardy-Weinberg equilibrium.

True Genotype			Observed Genotype			Probability[1]	MM[2]
Father	Mother	Offspring	Father	Mother	Offspring		
AA	AA	AA	AA	AA	AA	p^4	
AA	AB	AA	AA	AA	AA	$p^3(1\text{-}p)K_m$	
AA	AB	AA	AA	AB	AA	$p^3(1\text{-}p)(1\text{-}2K_m)$	
AA	AB	AA	AA	BB	AA	$p^3(1\text{-}p)K_m$	Y
AA	AB	AB	AA	AA	AA	$p^3(1\text{-}p)\,K_m\,K_o$	
AA	AB	AB	AA	AA	AB	$p^3(1\text{-}p)\,K_m(1\text{-}2K_o)$	Y
AA	AB	AB	AA	AA	BB	$p^3(1\text{-}p)\,K_m\,K_o$	Y
AA	AB	AB	AA	AB	AA	$p^3(1\text{-}p)\,(1\text{-}2K_m)\,K_o$	
AA	AB	AB	AA	AB	AB	$p^3(1\text{-}p)\,(1\text{-}2K_m)\,(1\text{-}2K_o)$	
AA	AB	AB	AA	AB	BB	$p^3(1\text{-}p)\,(1\text{-}2K_m)\,K_o$	Y
AA	AB	AB	AA	BB	AA	$p^3(1\text{-}p)\,K_m\,K_o$	Y
AA	AB	AB	AA	BB	AB	$p^3(1\text{-}p)\,K_m(1\text{-}2K_o)$	
AA	AB	AB	AA	BB	BB	$p^3(1\text{-}p)\,K_m\,K_o$	Y
AA	BB	AB	AA	BB	AA	$p^2(1\text{-}p)^2\,K_o$	Y
AA	BB	AB	AA	BB	AB	$p^2(1\text{-}p)^2(1\text{-}2K_o)$	
AA	BB	AB	AA	BB	BB	$p^2(1\text{-}p)^2\,K_o$	Y
AB	AA	AA	AA	AA	AA	$p^3(1\text{-}p)\,K_f$	
AB	AA	AA	AB	AA	AA	$p^3(1\text{-}p)\,(1\text{-}2K_f)$	
AB	AA	AA	BB	AA	AA	$p^3(1\text{-}p)\,K_f$	Y
AB	AA	AB	AA	AA	AA	$p^3(1\text{-}p)\,K_f\,K_o$	
AB	AA	AB	AA	AA	AB	$p^3(1\text{-}p)\,K_f(1\text{-}2K_o)$	Y
AB	AA	AB	AA	AA	BB	$p^3(1\text{-}p)\,K_f\,K_o$	Y
AB	AA	AB	AB	AA	AA	$p^3(1\text{-}p)\,(1\text{-}2K_f)\,K_o$	
AB	AA	AB	AB	AA	AB	$p^3(1\text{-}p)\,(1\text{-}2K_f)\,(1\text{-}2K_o)$	
AB	AA	AB	AB	AA	BB	$p^3(1\text{-}p)\,(1\text{-}2K_f)\,K_o$	Y
AB	AA	AB	BB	AA	AA	$p^3(1\text{-}p)\,K_f\,K_o$	Y
AB	AA	AB	BB	AA	AB	$p^3(1\text{-}p)\,K_f(1\text{-}2K_o)$	
AB	AA	AB	BB	AA	BB	$p^3(1\text{-}p)\,K_f\,K_o$	Y
AB	AB	AA	AA	AA	AA	$p^2(1\text{-}p)^2\,K_f\,K_m$	
AB	AB	AA	AA	AB	AA	$p^2(1\text{-}p)^2\,K_f\,(1\text{-}2K_m)$	
AB	AB	AA	AA	BB	AA	$p^2(1\text{-}p)^2\,K_f\,K_m$	Y
AB	AB	AA	AB	AA	AA	$p^2(1\text{-}p)^2\,(1\text{-}2K_f)K_m$	
AB	AB	AA	AB	AB	AA	$p^2(1\text{-}p)^2(1\text{-}2K_f)\,(1\text{-}2K_m)$	
AB	AB	AA	AB	BB	AA	$p^2(1\text{-}p)^2\,(1\text{-}2K_f)K_m$	Y
AB	AB	AA	BB	AA	AA	$p^2(1\text{-}p)^2\,K_f\,K_m$	Y
AB	AB	AA	BB	AB	AA	$p^2(1\text{-}p)^2\,K_f\,(1\text{-}2K_m)$	Y

AB	AB	AA	BB	BB	AA	$p^2(1-p)^2 K_f K_m$	Y
AB	AB	AB	AA	AA	AA	$2p^2(1-p)^2 K_f K_m K_o$	
AB	AB	AB	AA	AA	AB	$2p^2(1-p)^2 K_f K_m(1-2K_o)$	Y
AB	AB	AB	AA	AA	BB	$2p^2(1-p)^2 K_f K_m K_o$	Y
AB	AB	AB	AA	AB	AA	$2p^2(1-p)^2 K_f (1-2K_m) K_o$	
AB	AB	AB	AA	AB	AB	$2p^2(1-p)^2 K_f (1-2K_m) (1-2K_o)$	
AB	AB	AB	AA	AB	BB	$2p^2(1-p)^2 K_f (1-2K_m) K_o$	Y
AB	AB	AB	AA	BB	AA	$2p^2(1-p)^2 K_f K_m K_o$	Y
AB	AB	AB	AA	BB	AB	$2p^2(1-p)^2 K_f K_m(1-2K_o)$	
AB	AB	AB	AA	BB	BB	$2p^2(1-p)^2 K_f K_m K_o$	Y
AB	AB	AB	AB	AA	AA	$2p^2(1-p)^2 (1-2K_f)K_m K_o$	
AB	AB	AB	AB	AA	AB	$2p^2(1-p)^2 (1-2K_f)K_m(1-2K_o)$	
AB	AB	AB	AB	AA	BB	$2p^2(1-p)^2 (1-2K_f)K_m K_o$	Y
AB	AB	AB	AB	AB	AA	$2p^2(1-p)^2(1-2K_f) (1-2K_m) K_o$	
AB	AB	AB	AB	AB	AB	$2p^2(1-p)^2(1-2K_f) (1-2K_m) (1-2K_o)$	
AB	AB	AB	AB	AB	BB	$2p^2(1-p)^2(1-2K_f) (1-2K_m) K_o$	
AB	AB	AB	AB	BB	AA	$2p^2(1-p)^2 (1-2K_f)K_m K_o$	Y
AB	AB	AB	AB	BB	AB	$2p^2(1-p)^2 (1-2K_f)K_m(1-2K_o)$	
AB	AB	AB	AB	BB	BB	$2p^2(1-p)^2(1-2K_f) K_m K_o$	
AB	AB	AB	BB	AA	AA	$2p^2(1-p)^2 K_f K_m K_o$	Y
AB	AB	AB	BB	AA	AB	$2p^2(1-p)^2 K_f K_m(1-2K_o)$	
AB	AB	AB	BB	AA	BB	$2p^2(1-p)^2 K_f K_m K_o$	Y
AB	AB	AB	BB	AB	AA	$2p^2(1-p)^2 K_f (1-2K_m) K_o$	Y
AB	AB	AB	BB	AB	AB	$2p^2(1-p)^2 K_f (1-2K_m) (1-2K_o)$	
AB	AB	AB	BB	AB	BB	$2p^2(1-p)^2 K_f (1-2K_m) K_o$	
AB	AB	AB	BB	BB	AA	$2p^2(1-p)^2 K_f K_m K_o$	Y
AB	AB	AB	BB	BB	AB	$2p^2(1-p)^2 K_f K_m(1-2K_o)$	Y
AB	AB	AB	BB	BB	BB	$2p^2(1-p)^2 K_f K_m K_o$	
AB	AB	BB	AA	AA	BB	$p^2(1-p)^2 K_f K_m$	Y
AB	AB	BB	AA	AB	BB	$p^2(1-p)^2 K_f (1-2K_m)$	Y
AB	AB	BB	AA	BB	BB	$p^2(1-p)^2 K_f K_m$	Y
AB	AB	BB	AB	AA	BB	$p^2(1-p)^2 (1-2K_f)K_m$	Y
AB	AB	BB	AB	AB	BB	$p^2(1-p)^2(1-2K_f) (1-2K_m)$	
AB	AB	BB	AB	BB	BB	$p^2(1-p)^2(1-2K_f) K_m$	
AB	AB	BB	BB	AA	BB	$p^2(1-p)^2 K_f K_m$	Y
AB	AB	BB	BB	AB	BB	$p^2(1-p)^2 K_f (1-2K_m)$	
AB	AB	BB	BB	BB	BB	$p^2(1-p)^2 K_f K_m$	
AB	BB	AB	AA	BB	AA	$p(1-p)^3 K_f K_o$	Y
AB	BB	AB	AA	BB	AB	$p(1-p)^3 K_f(1-2K_o)$	
AB	BB	AB	AA	BB	BB	$p(1-p)^3 K_f K_o$	Y
AB	BB	AB	AB	BB	AA	$p(1-p)^3(1-2K_f) K_o$	Y
AB	BB	AB	AB	BB	AB	$p(1-p)^3(1-2K_f) (1-2K_o)$	
AB	BB	AB	AB	BB	BB	$p(1-p)^3(1-2K_f) K_o$	
AB	BB	AB	BB	BB	AA	$p(1-p)^3 K_f K_o$	Y
AB	BB	AB	BB	BB	AB	$p(1-p)^3 K_f(1-2K_o)$	Y

AB	BB	AB	BB	BB	BB	$p(1\text{-}p)^3 K_f K_o$	
AB	BB	BB	AA	BB	BB	$p(1\text{-}p)^3 K_f$	Y
AB	BB	BB	AB	BB	BB	$p(1\text{-}p)^3(1\text{-}2K_f)$	
AB	BB	BB	BB	BB	BB	$p(1\text{-}p)^3 K_f$	
BB	AA	AB	BB	AA	AA	$p^2(1\text{-}p)^2 K_o$	Y
BB	AA	AB	BB	AA	AB	$p^2(1\text{-}p)^2(1\text{-}2K_o)$	
BB	AA	AB	BB	AA	BB	$p^2(1\text{-}p)^2 K_o$	Y
BB	AB	AB	BB	AA	AA	$p(1\text{-}p)^3 K_m K_o$	Y
BB	AB	AB	BB	AA	AB	$p(1\text{-}p)^3 K_m(1\text{-}2K_o)$	
BB	AB	AB	BB	AA	BB	$p(1\text{-}p)^3 K_m K_o$	Y
BB	AB	AB	BB	AB	AA	$p(1\text{-}p)^3(1\text{-}2K_m) K_o$	Y
BB	AB	AB	BB	AB	AB	$p(1\text{-}p)^3(1\text{-}2K_m) (1\text{-}2K_o)$	
BB	AB	AB	BB	AB	BB	$p(1\text{-}p)^3(1\text{-}2K_m) K_o$	
BB	AB	AB	BB	BB	AA	$p(1\text{-}p)^3 K_m K_o$	Y
BB	AB	AB	BB	BB	AB	$p(1\text{-}p)^3 K_m(1\text{-}2K_o)$	Y
BB	AB	AB	BB	BB	BB	$p(1\text{-}p)^3 K_m K_o$	
BB	AB	BB	BB	AA	BB	$p(1\text{-}p)^3 K_m$	Y
BB	AB	BB	BB	AB	BB	$p(1\text{-}p)^3(1\text{-}2K_m)$	
BB	AB	BB	BB	BB	BB	$p(1\text{-}p)^3 K_m$	
BB	BB	BB	BB	BB	BB	$(1\text{-}p)^4$	

[1] p is the frequency of allele A; K_x with $x = o,m,f$ is the value of K for the offspring, putative mother and putative father, respectively.
[2] MM: mismatch - values with a Y are combinations where there is an apparent mismatch.

Expected mismatch rate for parent-offspring pair

Probabilities for each combination of observed and true parent-offspring genotypes are shown in Table S2. The probability of an apparent mismatch is the sum of the probabilities for the cases where there is a Y in the mismatch (MM) column. This simplifies to $p(1-p)\big(K_o + K_p + 2K_pK_o\big)$.

The probabilities conditional on observed genotype of the offspring are calculated similarly to the parent-pair case, but with the joint probabilities calculated from the single parent table:

$$P(\text{apparent mismatch, AA}^*) = p(1-p)\big[pK_p + K_pK_o + (1-p)K_o\big]$$

$$P(\text{apparent mismatch, AB}^*) = 0$$

$$P(\text{apparent mismatch, BB}^*) = p(1-p)\big[pK_o + K_pK_o + (1-p)K_p\big]$$

Table S2. Probabilities for combinations of observed and true parent-offspring pair genotypes.

True Genotype		Observed Genotype			
Parent	Offspring	Parent	Offspring	Probability[1]	MM[2]
AA	AA	AA	AA	p^3	
AA	AB	AA	AA	$p^2(1\text{-}p) K_o$	

AA	AB	AA	AB	$p^2(1\text{-}p)\ (1\text{-}2K_o)$	
AA	AB	AA	BB	$p^2(1\text{-}p)\ K_o$	Y
AB	AA	AA	AA	$p^2(1\text{-}p)\ K_p$	
AB	AA	AB	AA	$p^2(1\text{-}p)\ (1\text{-}2K_p)$	
AB	AA	BB	AA	$p^2(1\text{-}p)\ K_p$	Y
AB	AB	AA	AA	$p(1\text{-}p)\ K_p\ K_o$	
AB	AB	AA	AB	$p(1\text{-}p)\ K_p(1\text{-}2K_o)$	
AB	AB	AA	BB	$p(1\text{-}p)\ K_p\ K_o$	Y
AB	AB	AB	AA	$p(1\text{-}p)\ (1\text{-}2K_p)\ K_o$	
AB	AB	AB	AB	$p(1\text{-}p)\ (1\text{-}2K_p)\ (1\text{-}2K_o)$	
AB	AB	AB	BB	$p(1\text{-}p)\ (1\text{-}2K_p)\ K_o$	
AB	AB	BB	AA	$p(1\text{-}p)\ K_p\ K_o$	Y
AB	AB	BB	AB	$p(1\text{-}p)\ K_p(1\text{-}2K_o)$	
AB	AB	BB	BB	$p(1\text{-}p)\ K_p\ K_o$	
AB	BB	AA	BB	$p(1\text{-}p)^2\ K_p$	Y
AB	BB	AB	BB	$p(1\text{-}p)^2\ (1\text{-}2K_p)$	
AB	BB	BB	BB	$p(1\text{-}p)^2\ K_p$	
BB	AB	BB	AA	$p(1\text{-}p)^2\ K_o$	Y
BB	AB	BB	AB	$p(1\text{-}p)^2(1\text{-}2K_o)$	
BB	AB	BB	BB	$p(1\text{-}p)^2\ K_o$	
BB	BB	BB	BB	$(1\text{-}p)^3$	

[1] *p* is the frequency of allele A; K_x with $x = o,p$ is the value of *K* for the offspring, and putative parent, respectively.

[2] MM: mismatch - values with a Y are combinations where there is an apparent mismatch.

Sheep parentage example

The methods in the paper are applied here to a sheep dataset. The dataset consists of 198 offspring born in 2016, their 125 dams and 10 sires from the "methane yield selection flock" (Jonker *et al*. 2018). These offspring were recorded for parentage in the field within a day of their birth. All these animals have also been genotyped with SNP arrays (either 50K or 600K) and an overlapping set of 41020 autosomal SNPs were used here for pedigree validation.

GBS Genotyping was undertaken using similar methods to that used for the deer example in the accompanying paper. A GBS library for these 333 animals plus four positive controls was prepared using a double digest with the *Pst*I and *Msp*I restriction enzymes and then sequenced on a single lane of an Illumina Hiseq2500 and the SNPs and genotypes were called using the UNEAK pipeline.

This yielded 76,285 SNPs. One dam had insufficient results (mean depth of 0.18 at these SNPs). SNPs with a Hardy-Weinberg disequilibrium below -0.05 were removed with 75,825 remaining. The remaining individuals had a mean depth of 1.19 with a call rate (non-missing rate) of 56% for these SNPs.

The methods described here were applied to these data with the default settings in the KGD software and ignoring the recorded parentage (i.e., an attempt was made to assign the parents from within the set of sires and dams). This gave both parents assigned (and consistent with the recorded parents) for 195 offspring, a father-only assignment (consistent with the recorded father) for two offspring and one assignment failing the inbreeding threshold. One of the father-only assignments was due to the recorded mother failing the genotyping, the other appears to be due to a mis-recorded parentage (confirmed by the SNP array results) with the true mother not being included in these data. The offspring failing the inbreeding threshold had $r_{FM}-2F_O$ = 0.2004, very close to the threshold used (0.2). The initially assigned parents were the recorded parents. This threshold may be too strict at the depth of sequencing in this example. In summary, the all parents assigned were the recorded (likely true) parents, while 99.5% of true parents in the data were assigned.

The relatedness threshold used was 0.4. The highest estimated relatedness between an offspring and one of the sires other than its recorded sire was 0.29. The highest EMM to a 2nd most related sire was 0.024 (threshold of 0.01). There were three cases where an offspring had relatedness to a non-dam greater than 0.4, but in all cases the relatedness with the recorded dam was higher. Two of these 2nd best dams also passed the EMM threshold, but no other 2nd best dams did. All three cases of a 2nd best dam passing the relatedness threshold also passed the trio EMM threshold (0.02) with that dam. Relatedness estimates from the SNP chip results gave similar results (for the best two dams for these three offspring), but parentage with the 2nd best dams could be ruled out due to a sufficiently high mismatch rate (at least 1.4%, compared to a maximum of 0.015% for a true parent). This shows that the assignment procedure (with default settings) would have assigned an incorrect dam for two offspring if their true dam had not been genotyped.

If the SNPs were filtered as in Thrasher *et al*. (2018) but without the MAF or Hardy-Weinberg filters (i.e., read depth of at least 10, SNP call rate on remaining results of at least 95%,) this would leave 32 SNPs. Only six of these 32 had Hardy Weinberg disequilibrium > -0.05. This is clearly insufficient for a parentage analysis.

Supplementary References

Jonker, A., S.M. Hickey, S.J. Rowe, P.H. Janssen, G.H. Shackell *et al.*, 2018 Genetic parameters of methane emissions determined using portable accumulation chambers in lambs and ewes grazing pasture and genetic correlations with emissions determined in respiration chambers. *Journal of Animal Science* 96 (8):3031-3042.

Thrasher, D.J., B.G. Butcher, L. Campagna, M.S. Webster, and I.J. Lovette, 2018 Double-digest RAD sequencing outperforms microsatellite loci at assigning paternity and estimating relatedness: A proof of concept in a highly promiscuous bird. *Molecular Ecology Resources* 18 (5):953-965.

Supplemental Figures

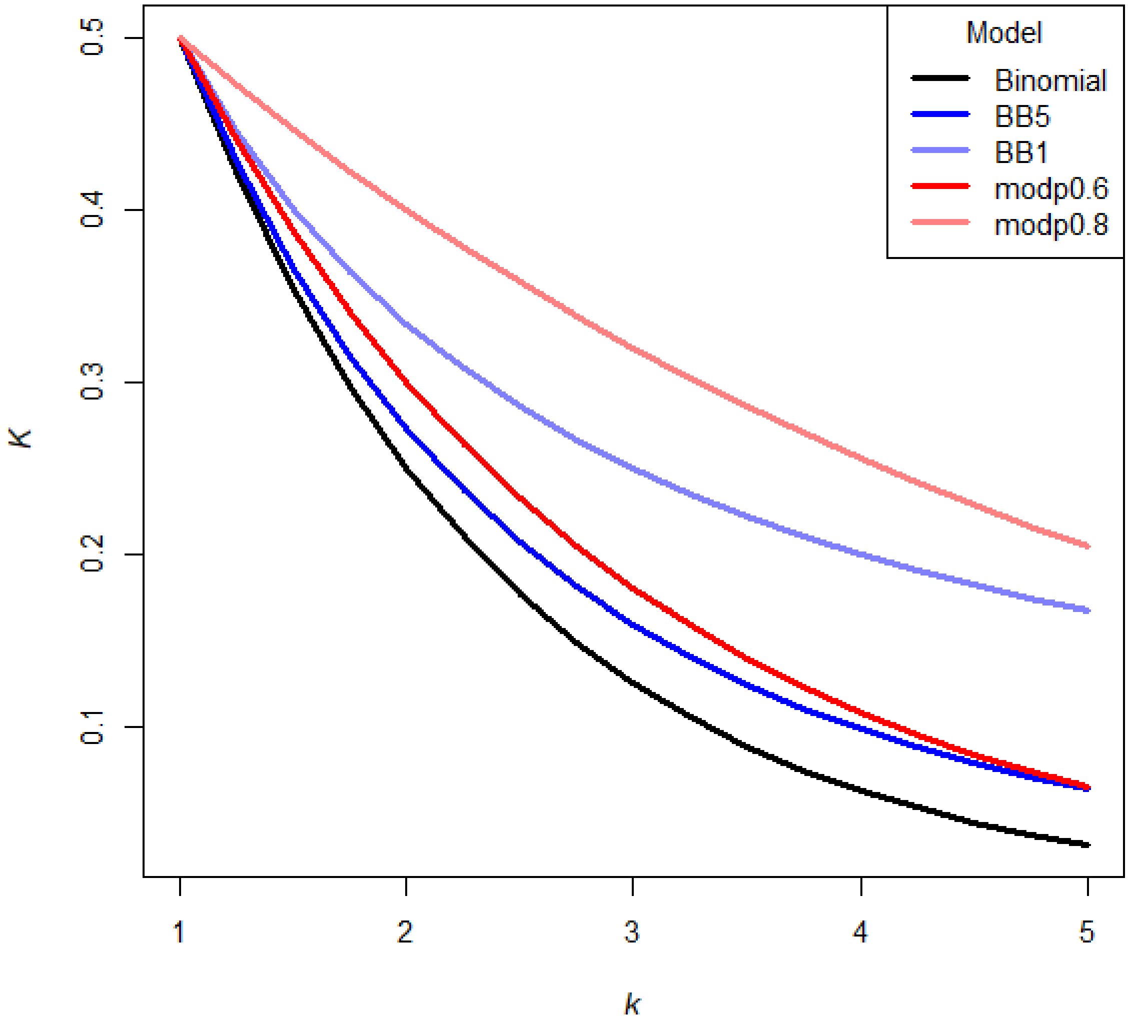

Figure S1 Relationship between *K* and *k* for different models. BB*t* refers to the beta-binomial sampling model with α = *t*; modp*t* refers to the modified p sampling model with *p*′ = *t*.

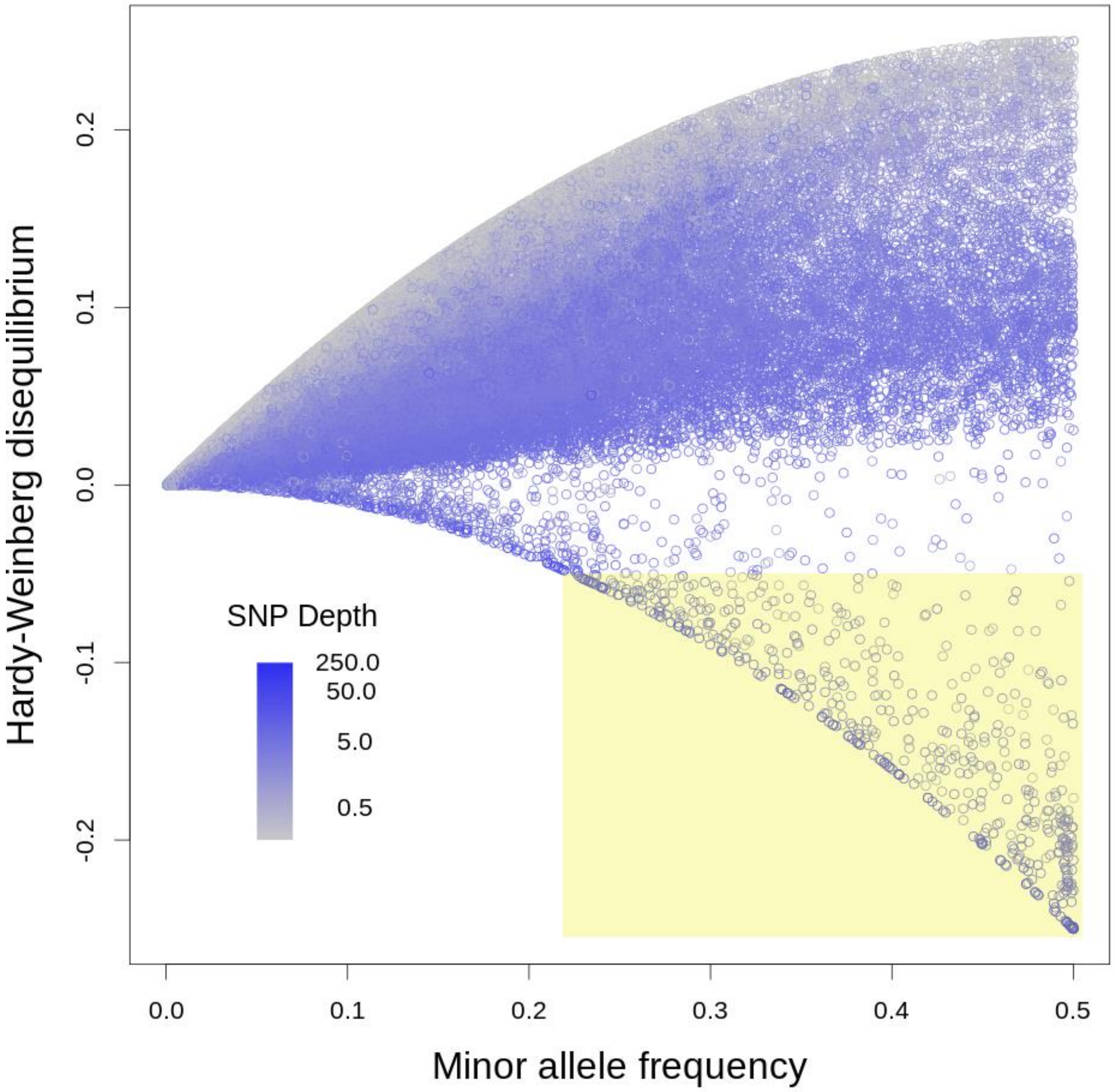

Figure S2 Fin plot of the combined breed data. SNPs within the yellow rectangle were removed from the analysis.

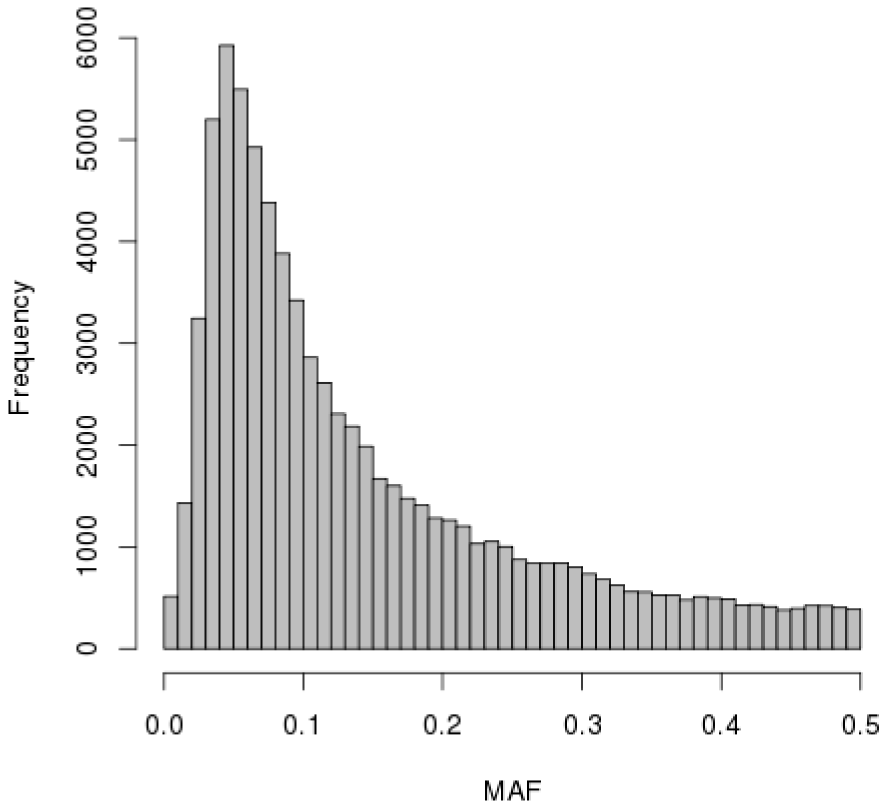

Figure S3 Distribution of minor allele frequencies (MAF) for 77,473 filtered SNPs in the combined breed data.

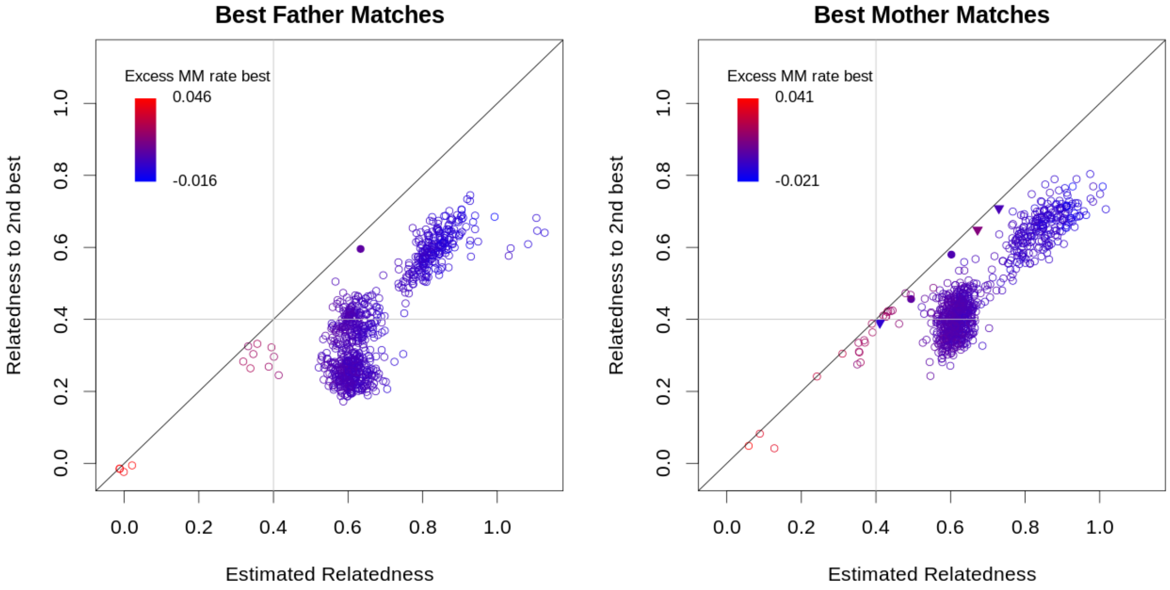

Figure S4 Comparison of relatedness for best and second best matching parents for the combined breed data. Points where these are within 0.05 are shown with filled symbols; those not reaching the 0.99 bootstrap support threshold are shown as triangles.

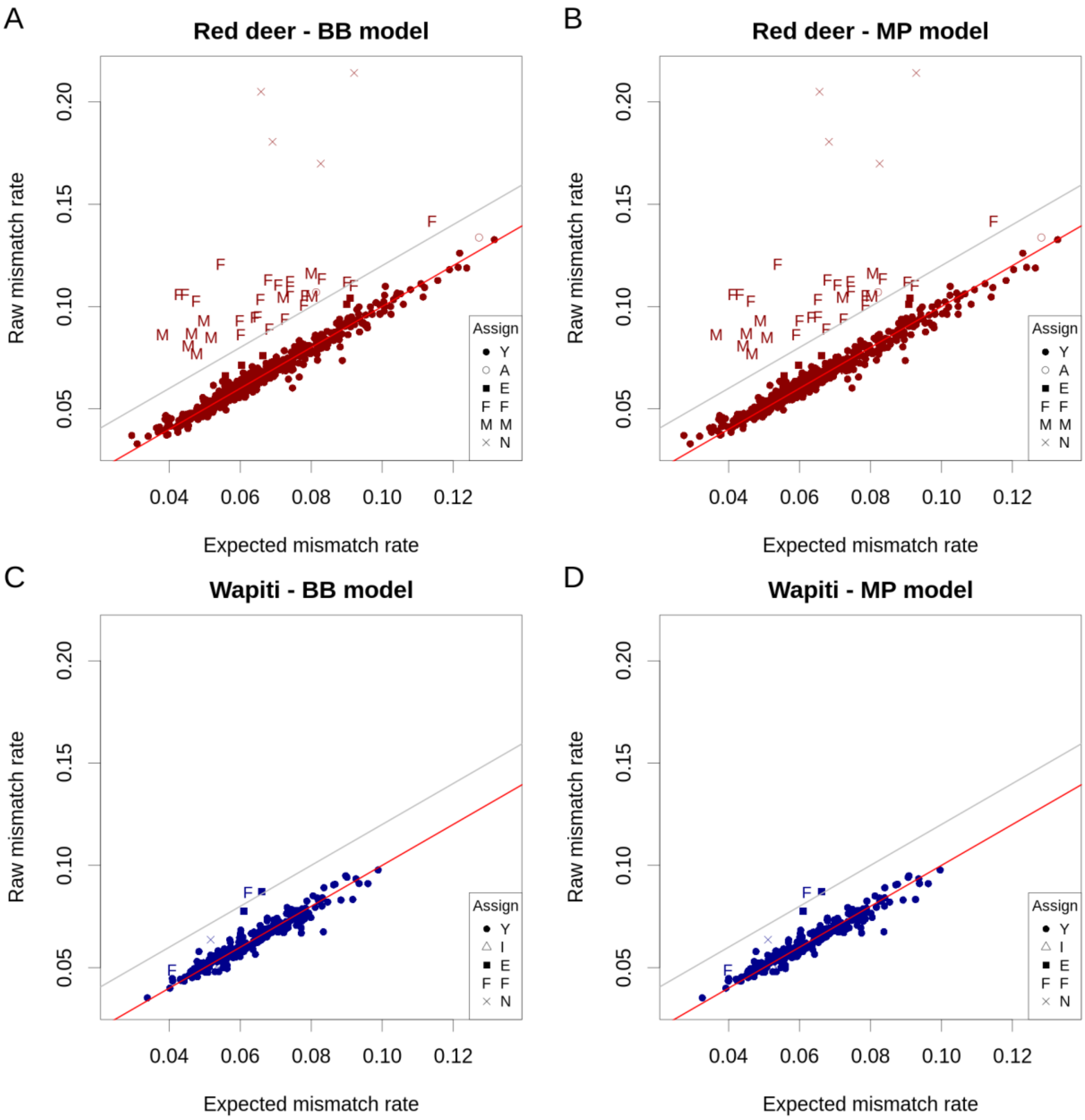

Figure S5 Comparison of raw and expected parent-offspring trio mismatch rates for the Red deer (A and B) and Wapiti (C and D) analyses using the BB (A and C) and MP (B and D) models. The red lines show where raw and expected rates are equal. The grey lines show the threshold used for excluding a trio from parentage. Assign codes shown are from the analysis using the binomial model. Assign codes are Y: assign parentage, A: an alternate parentage has lower EMM, I: fails the inbreeding criterion, E: exclude based on trio EMM, F: assign father only, M: assign mother only, N: do not assign either parent.