Parentage assignment with low density array data and low coverage sequence data

Andrew Whalen, Gregor Gorjanc, and John M Hickey

The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, Scotland, UK

Abstract

In this paper we developed and analysed a parentage assignment method for low density array data or low coverage sequence data. We tested the algorithm on a series of simulated datasets varying the genotyping error rate, number of potential sires, and availability of maternal information. We found that for array data, to obtain more than 99% sire assignment accuracy 250 SNPs were sufficient when the dam was unknown and 100 SNPs were sufficient when the dam was known and genotyped. . These results replicated on a real pig dataset. We found that for sequence data, to obtain 99% sire assignment accuracy 500 reads on genotyped sites were sufficient when the dam was unknown, and 200 reads were sufficient when the dam was known. We also investigated the benefit of pre-correcting genotyping errors using an imputation algorithm, and utilizing linkage information to increase accuracy. We found that pre-correcting genotyping errors could drastically increase parentage assignment accuracy, but that there was little benefit to using linkage information when typed or sequence loci were distributed across multiple chromosomes. We provide this parentage assignment method as a standalone program called AlphaAssign.

Introduction

In this paper we develop and analyse a parentage assignment method that handles potentially sparse, noisy, and linked genome-wide data. For many years, pedigrees lay at the heart of the animal breeder's toolkit. Although the importance of pedigree information has lessened since the advent of affordable genome-wide marker genotyping, it is still essential for some routine breeding operations such as reducing the cost of genotyping through imputation (Huang et al., 2012), reducing the bias of genomic estimates of breeding values (Solberg et al., 2009), and combining genotyped and non-genotyped individuals into a joint analysis (Legarra et al., 2009). In many cases, particularly in intensely managed populations, or where artificial insemination is common, accurate parentage is readily available. However, in some breeding populations where natural service sires are used over artificial insemination, e.g., beef cattle, sheep, and goats, or where polyspermic mating is used, e.g., commercial pig production accurate paternal (and in some cases maternal) information may be useful. When parentage is unknown, but genotype data is available, parentage assignment methods can be used to reconstruct a pedigree (e.g., Jones et al., 2010; Huisman, 2017).

Low-density genotype information is often sufficient to accurately assign parents (Fisher et al., 2009), but for more distant relationships, high-density genotype information is required (Riester et al., 2009). Traditionally, microsatellite markers were used to perform parentage assignment, but have generally been replaced by SNP markers as a cost-effective alternative (Rohrer et al., 2007; Fisher et al., 2009; Riester et al., 2009; Tokarska et al., 2009). The number of SNPs required to determine parentage tends to be low, e.g., 50-700 SNPs (Rohrer et al., 2007; Strucken et al., 2016; Fisher et al., 2009; Tortereau et al., 2017). The exact number of SNPs required varies from case to case and in general depends on population diversity, degree of

crossbreeding, marker allele frequencies, and genotyping error rate. Accounting for genotyping error has received a large amount of attention with the low density microsatellite data (e.g., Kalinowski et al., 2007). Although less important when high-density SNP array data is used, accounting for errors is still important when low-density SNP arrays (e.g. hundreds of SNPs) are used (Strucken et al., 2016; Boerner, 2017).

Most modern parentage assignment methods assume that all of the observed markers segregate independently from one another (e.g., Hayes, 2011; Huisman, 2017). It may be possible to increase parentage assignment information by taking linkage information into account. This can be easily done if one or both of the parents of an individual are phased. The use of linkage information in parentage assignment and pedigree reconstruction is not a new idea. Several studies have shown that linkage information is useful for finding distant relatives when high-density SNP array data is available (Hill and White, 2013; Sun et al., 2016; Morimoto et al., 2018), but when high-density SNP array data is available, parentage assignment is trivial even without utilizing linkage information (Hayes, 2011; Huisman, 2017). However, use of a high-density SNP array on a large number of individuals may be prohibitively expensive in commercial livestock settings (Huang et al., 2012; Gorjanc et al., 2015). In such a setting a low-density SNP array or alternative low-coverage genotype by sequencing (Gorjanc et al., 2015; Dodds et al., 2015) might be used for indiviudals, whose parents might be (partially) unknown. The question in this case is: in the context of a low-density SNP array or low-coverage sequencing, can accounting for genotyping error and linkage increase the accuracy of parentage assignment?

In this paper we evaluate the effect of accounting for genotyping error and linkage on parentage assignment in a livestock population genotyped with a low-density SNP arrays or low-

74 coverage sequencing. We first present the method and then test it on a series of simulated data
75 sets and a real data set. We conclude by offering suggestions for implementing parentage
76 assignment in a breeding program.

77 **Materials and Methods**

78 Existing parentage assignment methods can be split into three loose categories: (1)
79 methods based on exclusion criteria, such as opposing homozygous markers (Hayes, 2011;
80 Ferdosi and Boerner, 2014; Huisman, 2017), (2) methods based on the assessed degree of
81 relationship (Hill and White, 2013; Morimoto et al., 2018), and (3) probabilistic methods which
82 evaluate the likelihood of an individual's genotype given their putative parent's genotype
83 (Kalinowski et al., 2007; Riester et al., 2009; Cussens et al., 2013). In this study, we deploy a
84 probabilistic method because it can jointly handle noisy genotype data and linkage between
85 markers.

86 **Parentage assignment with noisy and linked markers**

87 In our method, we calculate a likelihood score for each set of putative parents which
88 gives the likelihood of an individual's genotype given the genotype of their putative parent(s).
89 We then compare this likelihood score across all potential parents and select as parent the one
90 with the highest likelihood score. This can be done when one or both of the parents are unknown.
91 We have packaged this method into an easy to use parentage assignment program called
92 AlphaAssign, which is available at www.alphagenes.roslin.ed.ac.uk/AlphaAssign. In the
93 following we describe this method.

94 For a given individual, i, and two putative parents, f and m, the likelihood score is the
95 probability that the individual obtained multi-locus genotype g_i given their parents genotypes g_f
96 and g_m. To evaluate this score, we assume that we observe potentially noisy observations of the

genotypes of the individual and putative parent's $\boldsymbol{d} = (d_i, d_f, d_m)$. In the case of SNP array data, the observed data is in the form of bi-allelic genotypes. In the case of sequence data, the observed data is in the form of read counts for reference and alternative alleles. Focusing on bi-allelic loci we consider four possible true genotype states at each locus (aa, aA, Aa, AA), where the first letter represents the allele inherited from the father, and the second letter indicates the allele inherited from the mother.

At a given locus *l* we want to evaluate the likelihood score $L_{i,mf,l} = p(g_{i,l}|g_{f,l}, g_{m,l})$, where $p(g_{i,l}|g_{f,l}, g_{m,l})$ is the Mendelian transmission function, i.e. the probability that the individual has genotype $g_{i,l}$ given that the parents have genotypes $g_{f,l}$ and $g_{m,l}$. For example, this probability is either 1, if the individual and both parents are homozygous for the same allele, .5 if one of the parents is heterozygous but the other parent and the individual are homozygous, and 0 if the parents are homozygous for an allele, but the individual is homozygous for another allele.

In the case of noisy data, the score is the expected value of the likelihood term $p(g_{i,l}|g_{f,l}, g_{m,l})$, found by marginalizing over the possible true genotypes given the observed data:

$$L_{i,mf,l} = \mathrm{E}\left[p(g_{i,l}|g_{f,l}, g_{m,l})\right]_{\mathrm{g_i,g_f,g_m}|\mathbf{d}} = \sum_{g_{i,l}} \sum_{g_{f,l}} \sum_{g_{m,l}} p(g_{i,l}|g_{f,l}, g_{m,l}) p(g_{i,l}, g_{f,l}, g_{m,l}|\boldsymbol{d})$$

The term $p(g_{i,l}, g_{f,l}, g_{m,l}|\boldsymbol{d}) = \prod_{x=i,f,m} p(g_{x,l}|d_{x,l})$ is the probability of individual's and parents' genotype states based on the observed data. In the case of SNP array data, we set $p(g_{x,l}|d_{x,l}) = 1 - e$ if the genotype state $g_{x,l}$ is consistent with the data $d_{x,l}$ and *e* otherwise, where *e* represents a small genotyping error. In this study we use *e=0.05,* although other values can be used. In the case of sequencing data, we set

$$p(g_{x,l}|d_{x,l}) = \begin{matrix} (1-e)^{n_{ref}} e^{n_{alt}} & if\ g_{x,l} = 0 \\ .5^{n_{ref}+n_{alt}} & if\ g_{x,l} = 1 \\ (1-e)^{n_{alt}} e^{n_{ref}} & if\ g_{x,l} = 2 \end{matrix}$$

where e is the sequencing error rate, n_{ref} is the number of reference alleles and n_{alt} is the number of alternative alleles. The summation is then performed over all possible combinations of individual's and parent's genotype states. This formulation is consistent with the model used previously (e.g., Kalinowski et al., 2007).

When the genotype states are unphased, we can assume that each locus segregates independently, and we compute the multi-locus likelihood as the product of the likelihood scores at each marker:

$$L_{i,mf} = \prod_l L_{i,mf,l}.$$

When genotype states are phased, we can exploit the fact that loci do not segregate independently. To account for linkage between neighbouring loci, we construct a vector *s* such that s_l describes a segregation state at a locus *l*. There are four possible segregation states based on whether an individual inherits the grand-maternal or grand-paternal chromosome from each parent. We then compute the multi-locus likelihood of individual's genotype, based on their potential parents' phased genotypes as:

$$L_{i,m,f} = \sum_l \sum_{s_l} \sum_{g_{i,l}} \sum_{g_{f,l}} \sum_{g_{m,l}} p(g_{i,l}|g_{f,l}, g_{m,l}, s_l)\, p(g_{i,l}, g_{f,l}, g_{m,l}|\boldsymbol{d}) p(s_l),$$

where the term $p(g_{i,l}|g_{f,l}, g_{m,l}, s_l)$ is either 1 if $g_{i,l}$ is consistent with $g_{f,l}$ and $g_{m,l}$ given the segregation pattern and 0 otherwise. The probability of a segregation vector is $p(s) = (1-\delta)^{n-k} \delta^k$, where δ is the recombination rate, *n* is the number of markers, and *k* is the number of recombination events in *s*. The summation over *s* can be calculated recursively in a manner similar to a traditional forward-backward algorithm for hidden Markov models.

140 **Simulated data**

141 The simulated data modelled a livestock population. We initially sampled a set of
142 genomes with 20 chromosomes using the Markovian coalescent simulator MaCS (Chen et al.,
143 2009). For this we assumed that each chromosome is 10^8 bp long, a per site mutation rate is
144 2.5×10^{-8}, a per site recombination rate is 1.0×10^{-8}, and that effective population size changed
145 over time. Based on estimates for the Holstein cattle population (Villa-Angulo et al., 2009), we
146 set the effective population size to 100 in the final generation of the coalescent simulation and to
147 1,256, 4,350, and 43,500 at respectively 1000, 10,000, and 100,000 generations ago, with linear
148 changes in between. We then used the sampled chromosomes to initiate a population of 1,000
149 animals with equal sex proportions. We bred this population for 5 generations. In each
150 generation, we selected 100 males and mated them at random to all 500 females. All animals
151 were genotyped with 500 SNPs per chromosome (10,000 SNPs per genome). In addition to SNP
152 array data, we have generated low-coverage sequence data, but only for the last generation of
153 individuals. We assumed that a genotype by sequencing method targets the same loci as SNP
154 array and generated between 50 to 10,000 sequence reads at these loci per individual. This gives
155 a range of sequencing coverages between 0.005x to 1x. We generated sequence read data by first
156 randomly sampling the 10,000 SNPs with replacement and then randomly sampling SNP alleles.
157 The latter included the probability of sequencing error, which was to .001 (0.1%). The
158 sequencing data was passed into AlphaAssign by first converting sequencing reads into genotype
159 probabilities, using the underlying sequencing error rate. We performed the whole simulation
160 using the simulation R package AlphaSimR (Gaynor et al.), which is available at
161 www.alphagenes.roslin.ed.ac.uk/AlphaSimR.

162 **Imputation and phasing**

163 We imputed and phased the genotypes of the first 4,000 animals using the pedigree based
164 peeling program AlphaPeel (Whalen et al., 2017), which is available at
165 www.alphagenes.roslin.ed.ac.uk/AlphaPeel. All 500 SNPs per chromosome were used for this
166 task. The goal of imputation was not to infer missing genotypes – there were no missing
167 genotypes – but instead to use pedigree and linkage information to correct for genotyping errors
168 and to phase the parental genotypes. SNP array genotypes or sequence reads of the last 1000
169 individuals were neither imputed nor phased.

170 **Scenarios**

171 We evaluated the accuracy of sire assignment for the last generation of 1000 individuals
172 in a variety of situations by varying: (a) the type of input data for parents (raw genotypes,
173 imputed genotypes, or phased genotypes), (b) the number of chromosomes genotyped, either 1 or
174 20, (c) dam status, (d) the number of SNPs for parentage assignment from between 50 to 500, (e)
175 the number of potential sires from between 5 and 100, (f) the genotyping error rate from between
176 0 to 10% (errors were introduced by randomly changing true genotypes to any possible
177 genotype), and (g) using sequencing data for individuals with unknown parents with sequencing
178 coverage ranging from 50 to 10,000 reads per individual. Each condition was replicated 10 times
179 and results were averaged across replicates.

180 Our goal with these simulations was fourfold. First, we evaluated how accuracy could be
181 gained by using pre-imputed genotypes (correcting for genotyping errors) and exploiting linkage
182 information. Second, we evaluated the required number of SNPs to obtain high accuracy (>99%)
183 of parentage assignment at different genotyping error rates, numbers of potential sires, and the
184 availability of genotypes for a known parent. Third, for a fixed number of SNPs, we evaluated
185 the accuracy of parentage assignment at different genotyping error rates, the numbers of potential

186 sires, and the availability of genotypes for a known parent. Forth, we evaluated accuracy of
187 parentage assignment with sequence data with different sequencing coverage per individual.

188 **Real data**

189 To evaluate the method on a real dataset, we assigned parentage for 1,000 individuals
190 from a commercial pig breeding program. These animals had true sire known and verified (based
191 on 15,000 markers). We evaluated the accuracy of sire assignment under different scenarios. The
192 animals and their parents were genotyped on a mixture of SNP arrays and we only used the
193 overlapping 13,000 SNPs across 18 chromosomes (the sex chromosome was excluded). To phase
194 parental genotypes, we extracted data on the individual's parents, grandparents, great-
195 grandparents, and great-great grandparents. We imputed and phased these ancestral genotypes
196 using AlphaPeel. We then evaluated the accuracy of assigning sires with between 50 and 500
197 loci and between 5 to 45 potential sires. No artificial genotyping error was introduced into the
198 real data.

199 **Results**

200 We found that for array data, to obtain more than 99% sire assignment accuracy 250
201 SNPs were sufficient when the dam was unknown and 100 SNPs were sufficient when the dam
202 was known and genotyped. We found that for sequence data, to obtain 99% sire assignment
203 accuracy 500 reads on genotyped sites were sufficient when the dam was unknown, and 200
204 reads were sufficient when the dam was known. The low density SNP results were replicated on
205 a real pig dataset. We explore these results in more detail below.

206

207 **Exploiting imputation based error correction and linkage information**

208 The accuracy of parentage assignment with 50 SNPs distributed across either 1 or 20
209 chromosomes is given in Figure 1. The accuracy of parentage assignment increased when
210 markers were distributed across 20 chromosomes compared to being distributed across a single
211 chromosome, and increased when the dam was known compared to when the dam was unknown.
212 In addition, the usage of AlphaPeel to correct genotyping errors increased pedigree assignment
213 accuracy in all cases. The usage of phase information increased accuracy when markers were
214 distributed across a single chromosome, but provided limited benefits when markers were
215 distributed across 20 chromosome.
216 Because of these results, the remaining simulations focus on the case where markers are
217 distributed across 20 chromosomes, and where AlphaPeel is used to correct spontaneous
218 genotyping errors.
219
220 **Number of SNPs for accurate parentage assignment**
221 The minimum number of SNPs required to obtain accuracy of sire assignment of above
222 99% is given in Figure 2. In all cases, 250 SNPs were sufficient to obtain 99% accuracy when
223 the dam was unknown, and 100 SNPs were sufficient to obtain 99% accuracy when the dam was
224 known and genotyped. In the best case scenario when there were no genotyping errors and only 5
225 potential sires, 100 SNPs were required when the dam was unknown, and 50 SNPs were required
226 when the dam was unknown. When there were no genotyping errors and 100 potential sires, 150
227 SNPs were required when the dam was unknown and 100 SNPs were required when the dam was
228 known. These values remained constant as long as the genotyping error rate was relatively low
229 (<3%).
230

231 **Accuracy of sire assignment for fixed numbers of SNPs**

232 We plot the accuracy of sire assignment when either 50, 100, or 150 SNPs were used in
233 Figure 3. We obtained the lowest accuracies when 50 SNPs were used and the dam was
234 unknown. In this case the accuracy ranged from 95% with no genotyping error and 5 potential
235 sires to 45% with 10% genotyping error and 100 potential sires. The accuracy in the latter
236 scenario increased to 80% when 100 SNPs were available, and to 95% when 150 SNPs were
237 available. When there was 1% genotyping error and the dam was unkown the parentage
238 assignment accuracy was 62% with 100 potential sires and 50 SNPs, 95% with 100 SNPs, and
239 100% with 150 SNPs. When the dam was known, the accuracy increased to 95% with 100
240 potential sires and 50 SNPs and 100% with either 100 or 150 SNPs.

241

242 **Accuracy of sire assignment with low-coverage sequence data**

243 The accuracy of parentage assignment when low-coverage sequence data was used is
244 given in Figure 4. The overall assignment accuracy with sequence data closely mimicked that of
245 a low-density SNP array where every two reads on a typed marker lead to an equivalent accuracy
246 as having a single SNP. We found that 500 overlapping reads were required to obtain perfect
247 parentage assignment when the dam was unknown and there were 100 potential sires. This value
248 lowered to 200 reads when the dam was known.

249 **Accuracy of sire assignment in real data**

250 The accuracy of sire assignment in the real dataset closely followed that of the simulated
251 data. We found that 100 SNPs were required to obtain high accuracy (>99%) parentage
252 assignment when there were either 5, 10, 25, or 45 potential sires and the dam was unknown.
253 Only 50 SNPs were required when there were 5, 10, or 25 potential sires and dam was known

254 and genotyped. As with simulated data, we found that using the error-corrected genotypes gave
255 consistently higher accuracy compared to using the raw genotypes, and that exploiting linkage
256 information gave limited benefit.

257 **Discussion**

258 In this paper we outline a versatile method for performing parentage assignment with
259 potentially noisy and linked genome-wide markers. We show that the method can accommodate
260 genotyping errors, use uncertain (noisy) sequence data, and exploit linkage between markers to
261 increase accuracy of parentage assignment under different scenarios. We also investigated the
262 number of SNPs required for accurate parentage assignment. We found that between 50 to 250
263 SNPs or between 100 to 500 targeted sequence reads per individual were generally required to
264 achieve an accuracy above 99% depending on the genotyping error rate, number of potential
265 sires, and dam status.

266 Overall, the results of this paper are consistent with past analyses on the number of SNPs
267 required for accurate parentage assignment. For example, Fisher et al. (2009) found that in cattle
268 populations, a carefully selected set of 40 SNPs would accurately assign parentage. Strucken et
269 al. (2016) placed the number of SNPs at 200 for pure breed cattle, 500 for composite breeds, and
270 700 to exclude false positives. We found that between 250-500 SNPs would be sufficient to have
271 high accuracy parentage assignments across a wide range of scenarios in simulated data and in
272 real data. Also, in line with Fisher et al. (2009) we found that including dam genotypes
273 substantially increased parentage assignment accuracy.

274 The novelty of this study is in the parentage assignment method that can flexibly
275 accommodate genotyping errors, use uncertain (noisy) sequence data, and exploit linkage
276 between the markers. We achieve this in two ways. First, we pre-process raw parental genotypes

277 with an imputation and phasing program that can exploit pedigree and linkage information to
278 infer the true parental genotypes and potentially also parental haplotypes. Second, we use a
279 versatile likelihood method that can either use raw genotypes or sequence reads or pre-processed
280 genotypes. The pre-processed genotypes can be either supplied just as imputed genotypes or as
281 imputed and phased genotypes (haplotypes). We stress that the purpose of imputation in this
282 context is not necessarily in inferring missing genotypes but instead to infer the true genotype,
283 i.e., to remove errors in the raw genotypes. However, as demonstrated in our analysis of the real
284 data, imputation can be used to infer the missing genotypes in this context too.

285 We found that exploiting linkage information using phased parental genotypes did not
286 increase the accuracy compared to using imputed genotypes. This result is somewhat surprising
287 given that previous work has found that linkage information can be used to resolve distant
288 relationships (Hill and White, 2013; Sun et al., 2016; Morimoto et al., 2018). We did find that if
289 markers from a single chromosome were assayed, use of linkage information improved
290 parentage assignment. However, when markers from multiple chromosomes were assayed, the
291 use of linkage information did not yield substantial benefits. In retrospect, these findings make
292 sense. When a large number of markers are typed on a single chromosome, there is a large
293 amount of linkage information available to use. However, when the same number of markers are
294 spread across multiple chromosomes, the amount of available linkage information is
295 substantially decreased. We would likely see a benefit of exploiting linkage information when
296 typing multiple chromosomes if the overall density of markers on each chromosome was higher.
297 However, at the point where there are a sufficient number of markers to exploit linkage
298 information, the overall number of markers is high enough that a phasing-free parentage
299 assignment method has close to no errors.

300 One of the largest factors we observed, in line with past studies (Fisher et al. 2009), is the
301 benefit that having genotype information on an individual's dam. We found that including
302 genotype information on the dam of an individual effectively halved the number of SNPs or the
303 number of marker reads required for determining parentage. The use of dam information may not
304 be restricted to cases where the dam is genotyped. If the dam is known, but ungenotyped, the
305 genotypes of their parents or siblings (if known) could be used to impute some of their loci (via,
306 e.g. a pedigree based peeling method). Using imputed data for a dam would likely increase
307 parentage assignment accuracy, albeit to a lesser extent than using the dam's own genotype.
308 Results from Figure 1 suggest that imputed dam genotypes would give us intermediate accuracy
309 between accuracies with the dam unknown and the dam known and genotyped.
310 Overall, the results of this study offer some recommendations for routine parentage
311 assignment in a livestock breeding program. We demonstrate that between 250-500 SNPs should
312 be sufficient to determine parentage in most livestock breeding programs, even if genotyping
313 error rate is as high as 10%. Alternatively we showed that between 100 to 500 targeted sequence
314 reads per individual could be used in lieu of genotype information. To handle such high
315 genotyping error rates or low coverage sequence data, the raw parental genotypes must be first
316 pre-processed by a phasing and imputation program to infer the true parental genotypes. A
317 likelihood based method can use pre-processed parental data to assign parentage. We
318 investigated using linkage information to improve parentage assignment. We demonstrated that
319 in theory it could increase accuracy, particularly if animals were assayed on a single, or a small
320 number of chromosomes. However, in realistic cases where the markers are spread out across all
321 chromosomes, exploiting linkage information in the parentage assignment step did not show
322 meaningful improvement and can be even detrimental in case of genome assembly errors. We

provide an easy to use parentage assignment program AlphaAssign, which is available at www.alphagenes.roslin.ed.ac.uk/AlphaAssign.

Acknowledgements

The authors acknowledge the financial support from the BBSRC ISPG to The Roslin Institute BB/J004235/1, from Genus PLC and from Grant Nos. BB/M009254/1, BB/L020726/1, BB/N004736/1, BB/N004728/1, BB/L020467/1, BB/N006178/1 and Medical Research Council (MRC) Grant No. MR/M000370/1. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk).

References

Boerner, V. (2017). On marker-based parentage verification via non-linear optimization. Genet. Sel. Evol. *49*, 50.

Chen, G.K., Marjoram, P., and Wall, J.D. (2009). Fast and flexible simulation of DNA sequence data. Genome Res. *19*, 136–142.

Cussens, J., Bartlett, M., Jones, E.M., and Sheehan, N.A. (2013). Maximum likelihood pedigree reconstruction using integer linear programming. Genet. Epidemiol. *37*, 69–83.

Dodds, K.G., McEwan, J.C., Brauning, R., Anderson, R.M., van Stijn, T.C., Kristjánsson, T., and Clarke, S.M. (2015). Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics *16*, 1047.

Ferdosi, M.H., and Boerner, V. (2014). A fast method for evaluating opposing homozygosity in large SNP data sets. Livest. Sci. *166*, 35–37.

Fisher, P.J., Malthus, B., Walker, M.C., Corbett, G., and Spelman, R.J. (2009). The number of single nucleotide polymorphisms and on-farm data required for whole-herd parentage testing in dairy cattle herds. J. Dairy Sci. *92*, 369–374.

Gaynor, R.C., Gorjanc, G., Wilson, D.L., Money, D., and Hickey, J.M. AlphaSimR: An R Package for Breeding Program Simulations. Manuscr. Prep.

Gorjanc, G., Cleveland, M.A., Houston, R.D., and Hickey, J.M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. *47*, 12.

Hayes, B.J. (2011). Technical note: Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J. Dairy Sci. *94*, 2114–2117.

Hill, W.G., and White, I.M.S. (2013). Identification of Pedigree Relationship from Genome Sharing. G3 GenesGenomesGenetics *3*, 1553.

Huang, Y., Hickey, J.M., Cleveland, M.A., and Maltecca, C. (2012). Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. *44*, 25.

Huisman, J. (2017). Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. Mol. Ecol. Resour. *17*, 1009–1024.

Jones, A.G., Small, C.M., Paczolt, K.A., and Ratterman, N.L. (2010). A practical guide to methods of parentage analysis. Mol. Ecol. Resour. *10*, 6–30.

Kalinowski, S.T., Taper, M.L., and Marshall, T.C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Mol. Ecol. *16*, 1099–1106.

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. J. Dairy Sci. *92*, 4656–4663.

Morimoto, C., Manabe, S., Fujimoto, S., Hamano, Y., and Tamaki, K. (2018). Discrimination of relationships with the same degree of kinship using chromosomal sharing patterns estimated from high-density SNPs. Forensic Sci. Int. Genet. *33*, 10–16.

371 Riester, M., Stadler, P.F., and Klemm, K. (2009). FRANz: reconstruction of wild multi-
372 generation pedigrees. Bioinformatics *25*, 2134–2139.

373 Rohrer, G.A., Freking, B.A., and Nonneman, D. (2007). Single nucleotide
374 polymorphisms for pig identification and parentage exclusion. Anim. Genet. *38*, 253–258.

375 Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Ødegard, J., and Meuwissen, T.H.
376 (2009). Persistence of accuracy of genome-wide breeding values over generations when
377 including a polygenic effect. Genet. Sel. Evol. *41*, 53.

378 Strucken, E.M., Lee, S.H., Lee, H.K., Song, K.D., Gibson, J.P., and Gondro, C. (2016).
379 How many markers are enough? Factors influencing parentage testing in different livestock
380 populations. J. Anim. Breed. Genet. *133*, 13–23.

381 Sun, M., Jobling, M.A., Taliun, D., Pramstaller, P.P., Egeland, T., and Sheehan, N.A.
382 (2016). On the use of dense SNP marker data for the identification of distant relative pairs. New
383 Dev. Relatedness Relatsh. Estim. *107*, 14–25.

384 Tokarska, M., Marshall, T., Kowalczyk, R., Wójcik, J.M., Pertoldi, C., Kristensen, T.N.,
385 Loeschcke, V., Gregersen, V.R., and Bendixen, C. (2009). Effectiveness of microsatellite and
386 SNP markers for parentage and identity analysis in species with low genetic diversity: the case of
387 European bison. Heredity *103*, 326.

388 Tortereau, F., Moreno, C.R., Tosser-Klopp, G., Servin, B., and Raoul, J. (2017).
389 Development of a SNP panel dedicated to parentage assignment in French sheep populations.
390 BMC Genet. *18*, 50.

Villa-Angulo, R., Matukumalli, L.K., Gill, C.A., Choi, J., Tassell, C.P.V., and Grefenstette, J.J. (2009). High-resolution haplotype block structure in the cattle genome. BMC Genet. *10*, 19.

Whalen, A., Ros-Freixedes, R., Wilson, D.L., Gorjanc, G., and Hickey, J.M. (2017). Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. BioRxiv.

Figures

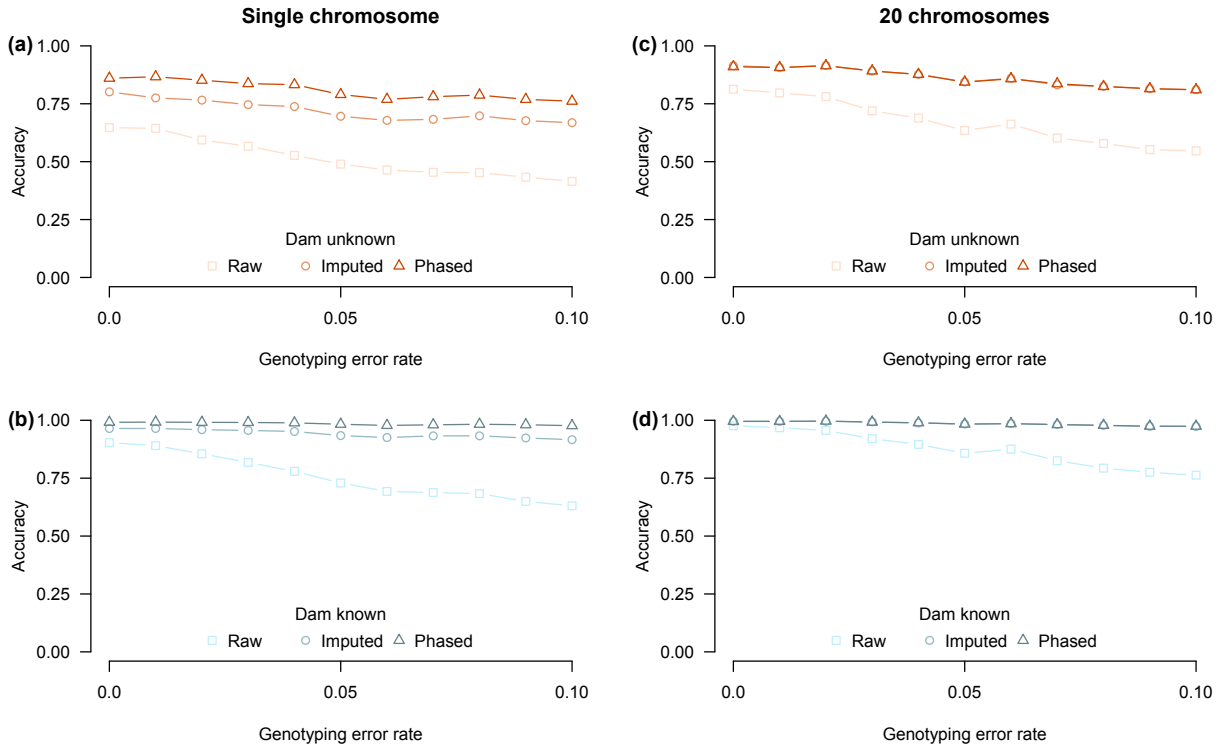

Figure 1. Sire assignment accuracy as a function of the genotyping error rate. Panels (a) and (c) present the accuracy when 50 SNPs are drawn from a single chromosome and there were 5 potential sires. Panels (b) and (d) present the accuracy when 50 SNPs are drawn across 20 chromosomes and there were 5 potential sires. Panels (a) and (b) assume dam was unknown. Panels (c) and (d) assume dam was known and genotyped.

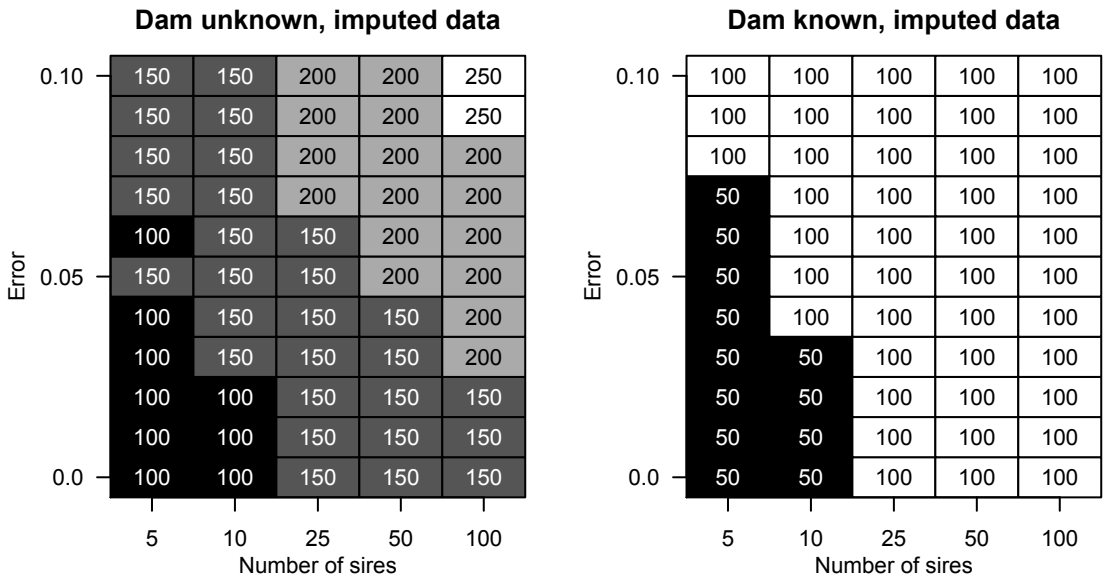

Figure 2. The number of SNPs required to obtain over 99% sire assignment accuracy when either (a) the dam was unknown, or (b) the dam was known and genotyped. Markers were drawn from across 20 chromosomes.

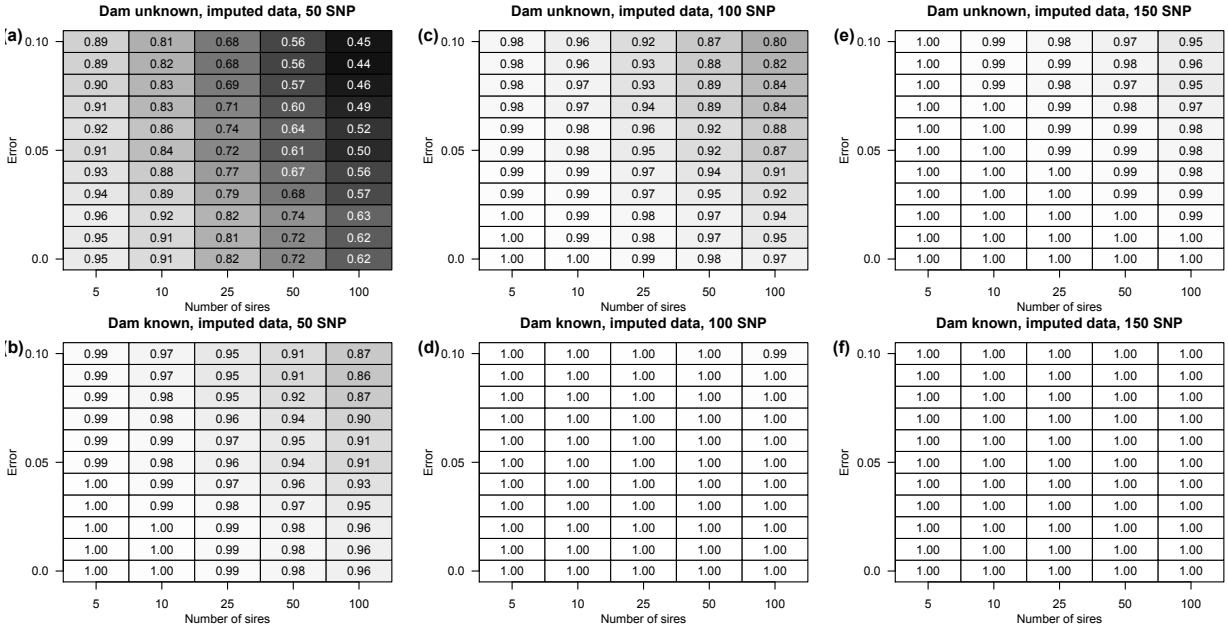

Figure 3. Sire assignment accuracy based on the number of potential sires, genotyping error rates, dam status, and the number of SNPs. Markers were drawn from across 20 chromosomes.

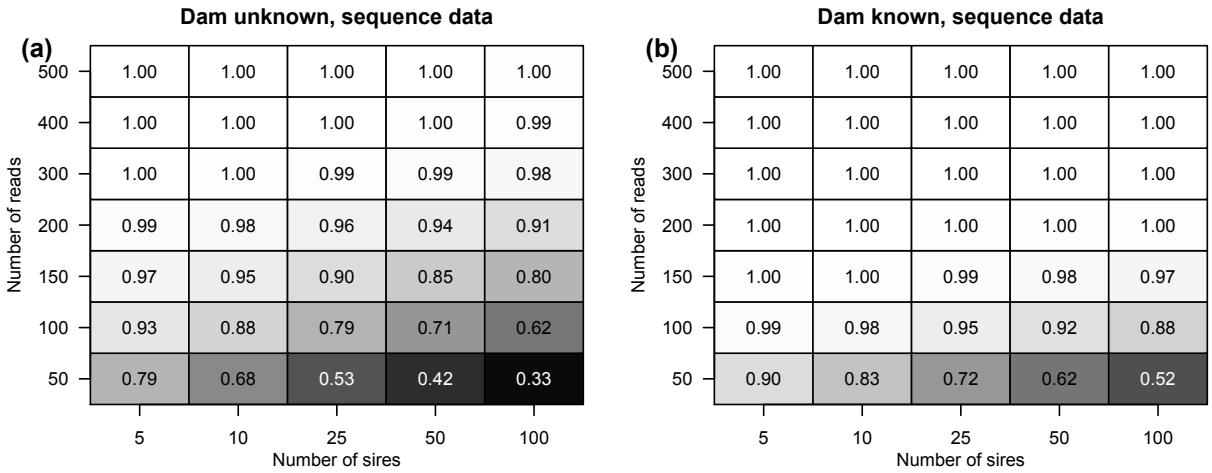

Figure 4. Sire assignment accuracy using low-coverage sequence data based on the number of potential sires, and number of reads at 10,000 loci. Sequence reads were drawn from across 20 chromosomes.