1 **A family-based phasing algorithm for sequence data**

2 Mara Battagin¹, Serap Gonen¹, Roger Ros-Freixedes¹, Andrew Whalen¹, Gregor

3 Gorjanc¹, John M Hickey^{1§}

4 ¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University
5 of Edinburgh, Easter Bush, Midlothian, Scotland, UK

6 [§]Corresponding author

7 Email addresses:

8 MB: mara.battagin@roslin.ed.ac.uk

9 SG: serap.gonen@roslin.ed.ac.uk

10 RR: roger.ros@roslin.ed.ac.uk

11 AW: awhalen@roslin.ed.ac.uk

12 GG: gregor.gorjanc@roslin.ed.ac.uk

13 JMH: john.hickey@roslin.ed.ac.uk

14

Abstract

This paper describes a family-based phasing algorithm, for variable-coverage sequence data, that first minimises phasing errors and then maximises the proportion of alleles phased. This algorithm is one of the essential tools that underpin an overall strategy for generating highly accurate sequence data on whole populations at low cost.

The algorithm is called AlphaFamSeq. It uses sequence data on the focal individual and at least two generations of ancestors to phase alleles. In the first step, AlphaFamSeq calculates allele probabilities using iterative peeling. In subsequent steps, the alleles are phased using heuristics deriving information from the sequence data of parents, grandparents and progenies and, if available, from other families in the pedigree. AlphaFamSeq was tested on a range of simulated data sets.

AlphaFamSeq gives low phasing error rates and, if there is sufficient sequence information and haplotype sharing amongst individuals, it can give a high yield of correctly phased alleles.

The allele threshold had a large effect and window size had a small effect on performance. When all individuals in a single family were sequenced at different coverages the highest correctly phased alleles reached 90% of the possible maximum (98.9%) at ~1/6 of the maximum aggregate coverage. Adding sequence information from other related individuals increased the percentage of correctly phased alleles. Imputation performance was high across all allele frequencies (average correlation by marker of 0.94), except for a slight decrease at very low frequencies (≤ 0.01 MAF).

Within an overall strategy for generating highly accurate sequence data on whole populations at low cost the role of AlphaFamSeq is to provide very accurately

39 phased haplotypes on focal individuals, who are individuals whose haplotypes are
40 very common in the population.

41

Background

This paper describes a family-based phasing algorithm for variable-coverage sequence data that first minimises phasing errors and then maximises the proportion of alleles phased. This design enables accurate imputation of sequence data in livestock populations. Superficially the work of Robin Thompson has tenuous links to phasing of sequence data. He was the first to recognise that the statistical modelling of breeding values could be partitioned into parent average and Mendelian sampling terms. Phasing, imputation and sequencing strategies exploit the same genetic principles. This paper describes one such method.

In a livestock population, sequence data has a number of potential advantages compared to classical marker genotype data including increased power to discover causative variants and more accurate genomic predictions. Sequence data has enabled the discovery of causative variants for qualitative traits (e.g. for embryonic lethality in the 1,000 Bulls Project [1]). However, sequence data has only shown modest benefits for quantitative traits, and a small increases in the accuracy of genomic prediction [2,3].

There are two likely reasons for the lack of large benefit for quantitative traits. First, data from millions of individuals may be required to capture the full potential of sequence data in livestock [4]. With fewer individuals, there will not be enough recombination events to estimate the effects of the large numbers of causative variants that underlie a quantitative trait. Second, the imputation methods used to generate large datasets may not be accurate enough. There may not be enough information in the generated sequence data to enable accurate phasing. Existing phasing algorithms are sub-optimal because the animals that provide the sequence data are insufficiently related to the animals that have sequence data imputed.

67 We believe that these issues can be addressed if effective sequencing strategies
68 and accurate phasing/imputation algorithms are developed jointly, which requires
69 strategies for the optimal distribution of sequence resources across a population and
70 accurate phasing/imputation algorithms. Separate strategies and algorithms are needed
71 for low coverage sequence data and for variable coverage sequence data. We have
72 recently developed AlphaSeqOpt, which implements the required strategies for the
73 optimal distribution of sequence resources [5]. Now we need accurate phasing
74 algorithms, specifically designed to meet the requirements of AlphaSeqOpt.

75 All imputation-based sequencing strategies involve: (i) collecting sequence
76 data on a relatively small number of individuals; (ii) accurately phasing these
77 individuals to resolve their haplotypes; and (iii) imputing the phased haplotypes into a
78 large number of unsequenced individuals by inferring the combination of haplotypes
79 that each unsequenced individual carries. In this context, the accurate phasing of the
80 haplotypes is essential: phasing errors lead to incorrect imputation. In populations
81 with hundreds of thousands of individuals this leads to large numbers of errors
82 because haplotypes are carried by thousands of individuals.

83 Many phasing algorithms or analogous imputation algorithms have been
84 developed (e.g., AlphaImpute [6], Findhap [7], Fimpute [8]) or used (e.g., MaCH [9],
85 Beagle [10,11], Impute2 [12]) in livestock populations. These algorithms were
86 originally developed for SNP genotype data, but some can also be used for sequence
87 data. However, we believe that these algorithms are suboptimal, because they do not
88 prioritise accurate phasing. Instead they seek to phase and impute alleles for all
89 individuals at all genome positions. This approach generates good phasing for most of
90 the markers, but fails in phasing low frequency alleles, which are common in

91 sequence data sets. The error rates are especially high in algorithms that preferentially
 92 determine phase and impute alleles based on haplotypes that have high frequency (i.e.,
 93 MaCH [9], Beagle [10,11], Impute2 [12], Findhap [7], Fimpute [8]).

94 The objective of this research was to develop a family-based phasing
 95 algorithm for variable-coverage sequence data that first minimises phasing errors and
 96 then maximises the proportion of alleles phased. This algorithm was designed to work
 97 with the sequencing strategy proposed by Gonen et al [5], which distributes a fixed
 98 amount of sequencing resources across a population by identifying focal individuals,
 99 determining the genomic footprint of these focal individuals and determining the
 100 optimal distribution of sequencing resources across these focal individuals and their
 101 ancestors (parents and grandparents) according to the genomic footprint of the focal
 102 individuals on the population.

103 The resulting algorithm performs well. It can work within a given family
 104 separately, can work with multiple families simultaneously by utilising pedigree
 105 information that connects them and can handle variable coverage sequencing data.
 106 The algorithm was implemented in a new software package called AlphaFamSeq. It
 107 gives low phasing error rates and, if there is sufficient sequence information and
 108 haplotype sharing amongst individuals, it can give a high yield of correctly phased
 109 alleles.

110

111 **Methods**

112 We developed and tested a family-based phasing algorithm for variable-
 113 coverage sequence data that first minimises phasing errors and then maximises the

114 proportion of alleles phased. The algorithm is called AlphaFamSeq; it uses sequence
115 data on the focal individual and a pedigree that includes at least the parents and
116 grandparents of the focal individual and requires that some of the individuals in this
117 pedigree have sequence data. AlphaFamSeq processes individuals from the oldest to
118 the youngest and iterates until convergence. In the first step, AlphaFamSeq calculates
119 allele probabilities using iterative peeling [13]. In subsequent steps, the genotypes are
120 phased using heuristics that derive information from the genomic data of
121 grandparents, parents, and progeny and, if available, from other families in the
122 pedigree.

123 In the next sections we give a detailed description of the data used by
124 AlphaFamSeq, the algorithm and its parameters, the simulation of the data used to test
125 the algorithm, and the metrics used in the tests.

126 **Data used by AlphaFamSeq**

127 AlphaFamSeq uses pedigree information and genomic data, which can be
128 sequence data or genotype data. Sequencing data can cover the whole genome or parts
129 of it. The latter is common with the reduced representation sequencing approaches
130 such as genotyping-by-sequencing [14]. The pedigree information must include the
131 parents and grandparents of at least one focal individual. There is no upper limit on
132 pedigree complexity. Within the pedigree individuals can be sequenced to any
133 coverage or unsequenced. The sequence data are presented to the algorithm in the
134 form of the number of reads for the reference and the alternative alleles at each bi-
135 allelic variant.

136 **The AlphaFamSeq algorithm**

137 AlphaFamSeq iterates across five main steps until convergence. These steps
138 are summarised in Figure 1.

139 *1. Iterative peeling.*

140 The first step calculates allele probabilities for each variant of each individual
141 in the pedigree, using all the pedigree information and the genomic data available.
142 This step is based on a modified version of the iterative peeling method described by
143 Kerr and Kinghorn [13]. The penetrance function is modified so that it accepts both
144 observed genotypes or observed number of reads for the reference and the alternative
145 alleles. The penetrance function is based on the binomial distribution with a given
146 number of trials (number of sequence reads) and user-defined error rate [9]. This
147 method accumulates information from parents, the individual itself, and progeny
148 (accounting for mates' information), which enables propagation of genotype or
149 sequence information across large and complex pedigrees typical of livestock
150 populations [13]. The peeling step is a single locus phasing where the outputs are
151 allele probabilities. If an allele probability, is above a user-defined threshold the allele
152 is called. In AlphaFamSeq the peeling is performed once, in the first iteration, and its
153 output is used differently at each subsequent iteration by dynamically relaxing the
154 user-defined allele thresholds. This allows the algorithm to utilise only the very
155 informative variants in the early iterations to ensure that errors are not generated and
156 subsequently propagated but to relax this stringency in later iterations to enable the
157 yield to increase.

158 *2. Phasing based on parents informative markers.*

159 In the second step a Mendelian inheritance rule is used to phase alleles of an
160 individual if one or both of its parents are deemed homozygous either from
161 sequencing reads or from information derived from a previous iteration or a previous
162 step. The rule is: if a parent is homozygous at a variant, then the phase of the progeny
163 for the gamete inherited from that parent can be assigned.

164 3. *Phasing based on parent-progeny informative markers.*

165 In the third step, a Mendelian inheritance rule is used to phase alleles of an
166 individual on the basis of marker information gathered from parent to progeny
167 inheritance. The rule is: if an individual has one of its two gametes phased and the
168 evidence from the progeny indicates that it inherited the other allele, then on the basis
169 of progeny information the allele can be phased for the relevant gamete of the parent.

170 4. *Founder Assignment and haplotype definition.*

171 In the fourth step, rules are used to identify trios of individuals (a focal
172 individual, its relevant parent and grandparent) that share a haplotype. The
173 grandparents are considered the founders (within the context of this three generation
174 sub-pedigree) and the aim is to assign the correct founder to each allele in the focal
175 individual. In this step AlphaFamSeq allows the user to divide the chromosome into
176 different windows. Windows are used in AlphaFamSeq for two reasons:

- 177 (i) to speed up the steps 4 and 5 of AlphaFamSeq by working on multiple
178 windows in parallel; and
- 179 (ii) to limit the search space when individuals are sequenced at low-
180 coverage and there are not enough informative variants to detect recombinations.

181 At each heterozygous locus of the parent, if the allele that the focal individual
182 inherited from that parent is known, the grandparent from which the focal individual
183 inherited the allele is determined. Once this is complete for each locus, each gamete
184 of the focal individual is examined from the beginning to the end and from the end to
185 the beginning of a user defined window. If the same founder is identified across the
186 entire window by the examinations in both directions that founder is assigned for the
187 whole window. If, due to recombination, two founders give rise to different sections
188 of the window, the examinations in both directions will not agree in regions where a
189 recombination has occurred. Where there is agreement a founder is assigned and
190 where there is no agreement a founder is not assigned.

191 5. *Build the consensus haplotype and impute missing alleles.*

192 In the fifth step, the founder assignments defined in step 4 are used to build a
193 consensus haplotype for the focal individual, its relevant parent and grandparent and
194 to impute alleles for the parent and progeny. This imputation is done if an allele was
195 inferred on the shared haplotype for two of the three individuals in a previous step.

196 6. *Multiple family phasing.*

197 AlphaFamSeq works from the oldest to the youngest individual in the
198 pedigree. When multiple and related families are provided, the information of related
199 individuals are accumulated in step 1 (single locus peeler) and in steps 4 and 5. The
200 haplotypes generated for the focal individual, its parents and grandparents (steps 4 to
201 5) are dropped down into the pedigree if one of the focal pedigree members is related
202 with other individuals in the pedigree.

203

204 ***Examples of method implementation: Description of datasets***

205 The algorithm was tested on a range of simulated data sets that were designed
206 to reflect different use cases of AlphaFamSeq, and quantify the scenarios where
207 AlphaFamSeq performs well or performs poorly. Simulated data was generated using
208 AlphaSim [15], AlphaSeqOpt [5] and associated programs and tools. The simulation
209 and analysis of this data involved four steps: (i) Simulation of genomes for animals in
210 a pedigree; (ii) Simulation of whole genome sequence data; (iii) Simulation of a range
211 of sequencing scenarios; (iv) Phasing of this sequence data with a range of algorithm
212 parameters; and (v) Assessment of phasing accuracy. Ten replicates of each scenario
213 were simulated and the results are the average of these replicates.

214 **1. Simulation of genomes for animals in a pedigree**

215 Sequence data were generated using the Markovian Coalescent Simulator
216 (MaCS) [16] and AlphaSim [15] for 1,000 base haplotypes for 1 chromosome of 1M
217 in length. The simulator used a per site mutation rate of 2.5×10^{-8} , a per site
218 recombination rate of 1×10^8 , and an effective population size (N_e) that varied over
219 time in accordance with estimates for a livestock population [17]. The resulting
220 sequences had 750,300 segregating sites in total. From the whole sequence, 700,000
221 ($MAF \geq 0.001$) segregating sites were selected to represent whole genome sequence
222 data for the subsequent analysis. After the simulation of base haplotypes, AlphaSim
223 [15] was used to drop them through a pedigree with 24 generations containing 59,194
224 individuals of which were 2,210 sires and 13,421 dams.

225 **2. Simulation of whole genome sequence data**

226 Sequence data were simulated for each individual by sampling sequencing
227 reads for the 700,000 segregating sites per chromosome. This was done using a

Poisson-Gamma process which allowed the number of sequence reads per locus to vary along the genome and to vary between individuals [9]. First sequenceability (γ_j) of each of the 700,000 loci along the genome was sampled according to a gamma distribution, with shape and scale parameters equal to α and $1/\alpha$, respectively. The number of reads (r) per individual i at locus j was then sampled from a Poisson distribution with mean equal to $\mu_{i,j} = x\gamma_j$, where x was the targeted (average) coverage and γ_j was the sequenceability for the locus j . Each read was generated by randomly sampling one of the alleles from the two gametes of individual i at locus j , accounting for a sequencing error. All individuals in the analyses (within and across replicates) had the same shape of coverage distribution per locus (γ_j). The α parameter in the gamma distribution was set to 4. The assumed error rate was 1‰ ($\epsilon = 0.001$).

3. Simulation of a range of sequencing scenarios

Two sets of sequencing scenarios were simulated.

The first set of scenarios was designed to assess the performance of AlphaFamSeq for a focal individual when sequence data was available only on the focal individual, its parents and its grandparents and there was no other pedigree information. We refer to this set of scenarios as **Single Family Phasing**.

The second set of scenarios was designed to assess the performance of AlphaFamSeq when sequence data was available for several focal individuals that were interrelated in ways that are typically found in livestock populations. Interconnected pedigree information and sequence data were available for many individuals within the pedigree. We refer to this set of scenarios as **Multiple Family Phasing**.

Single Family Phasing

252 The Single Family Phasing scenarios had modest computational requirements
253 (on average 175.95Mb of RAM and 51.1 seconds run time), which enabled the
254 performance of the algorithm to be assessed with different values for the user-defined
255 parameters and when the sequencing coverage of the individuals varied. To set up the
256 Single Family Phasing scenarios, ten families, including a focal individual and their
257 parents and grandparents, were randomly extracted from a pedigree of 59,194
258 individuals. The focal individuals were then phased using AlphaFamSeq using only
259 the sequence data on individuals in the family.

260 We tested the impact of the allele threshold and window size on the
261 performance of AlphaFamSeq. In these scenarios, all the 7 members were sequenced
262 at the same coverage and we only tested 5 sequencing coverages (1x, 2x, 5x, 15x, and
263 30x). We tested six values for the allele threshold (≥ 0.60 , ≥ 0.80 , ≥ 0.90 , ≥ 0.95 , 0.99
264 and ≥ 0.999) and ten values for window size (1, 100, 500, 1000, 5000, 10000, 50000,
265 100000, 350000 and 700000). The window size represents the number of variants per
266 each window, where “1” means that single locus phasing was performed (i.e., steps 4
267 and 5 of AlphaFamSeq were overpassed) and “700000” means that a single window,
268 with all the variants in the chromosome, was used.

269 We also tested the performance of AlphaFamSeq when individuals were
270 sequenced with variable coverage. In these simulations we fixed the allele probability
271 threshold and window size to the best performing values in the previously described
272 simulations. All combinations of six sequencing coverages were tested (0, 1, 2, 5, 15
273 and 30x). Because there were 7 members of a family (focal individual, parents and
274 grandparents) this resulted in 6^7 (i.e., 279,936) sequencing scenarios. We used the
275 same 10 families as replicates and averaged their results.

276 *Multiple Family Phasing*

277 The Multiple Family Phasing scenarios were designed to test the performance
278 of AlphaFamSeq when applied to what might be a typical livestock scenario in which
279 multiple families are connected via a large pedigree and some members of many of
280 these families are sequenced at different coverages. A second motivation for the
281 Multiple Family Phasing scenarios was to test its complementarity to AlphaSeqOpt.
282 AlphaSeqOpt is an algorithm for distributing sequencing resources across such a
283 pedigree [5]. AlphaSeqOpt distributes sequencing resources in proportion to the
284 genomic footprint of a focal individual and assigns some resources to the parents and
285 grandparents of focal individuals in order to enable the sequence data of the focal
286 individual to be phased.

287 To test the Multiple Family Phasing we used AlphaSeqOpt [5] to identify 500
288 focal individuals from the pedigree of 59,194 individuals and to distribute £317,800
289 worth of sequencing resources (on average 2x/individual) across the families of these
290 focal individuals with each individual being sequenced at either 0x, 1x, 2x, 5x, 15x or
291 30x. The available budget, number of focal individuals, and number of possible
292 sequencing coverages resulted in a total of $279,936^{500}$ possible ways of distributing
293 the sequencing resources across the focal individuals and their parents and
294 grandparents. The focal individuals were chosen by AlphaSeqOpt based on the
295 genotype data for 2,100 markers. The assumed sequencing costs were £40 for library
296 preparation and £80 for each 1x whole genome sequence. The full set of parameters
297 for AlphaSeqOpt [5] are reported in the additional file 1. The top 500 focal
298 individuals, their parents and grandparents encompassed a pedigree of 1,589
299 individuals that were sequenced at a range of coverages. This data was phased using
300 AlphaFamSeq with the allele threshold set to equal or greater than 0.90 and the

301 window size to 100000 variants. These parameters were chosen based on the results
302 of the earlier scenarios.

303 **4. Assessment of phasing accuracy**

304 Phasing was performed for each scenario using the parameters previously
305 described. The performance of AlphaFamSeq was measured as: (i) the percentage of
306 alleles across all variants that were correctly phased (**%Correct**); and (ii) the
307 percentage of alleles across all variants that were incorrectly phased (**%Error**). Most
308 of the results were presented as %Correct and %Error for the focal individual. Where
309 specified in the results section, these statistics were measured also for the parents and
310 grandparents. Moreover the %Correct and %Error were calculated for the imputed
311 genotypes to test the effect of the user-parameters in imputation of the homozygotes
312 and heterozygote genotypes. In the multiple family imputation, correlations between
313 true and imputed genotypes were calculated and binned by minor allele frequency.

314

Results

AlphaFamSeq correctly phased high percentages of alleles (%Correct) and incorrectly phased very low percentages (%Error). AlphaFamSeq always performed well when sequencing coverage was high and our aim was to identify the conditions under which it performs well at low or intermediate coverage.

The results show four things: (i) in Single Family Phasing, allele threshold had a big effect and window size had a very small and non-monotonic effect; (ii) when all individuals in a single family were sequenced at different coverages the best possible %Correct reached 90% of the possible maximum (98.9%) at $\sim 1/6$ of the maximum aggregate coverage; (iii) adding sequence information from other related individuals increased the %Correct; and (iv) imputation performance was good across all allele frequencies (average correlation by marker of 0.94), except for a slight decrease at for very low frequencies (≤ 0.01 MAF).

Impact of AlphaFamSeq parameters on performance of Single Family Phasing when all individuals were sequenced at the same coverage

In Single Family Phasing, allele threshold had a large effect and window size had a small effect. The %Correct ranged from 0.47% to 99.06% and the %Error ranged from 0.0001% to 8.68%.

User-defined allele probability threshold. The allele threshold affected both the %Correct (Figure 2) and %Error (Figure 3) more when individuals were sequenced at low coverage. Relaxing the allele thresholds (i.e., moving leftwards on the x-axis of Figure 2 and Figure 3) increased both the %Correct and the %Error for the focal individual.

339 Unsurprisingly, the impact of relaxing the allele threshold depended on the
340 sequencing coverage. Relaxing the allele threshold had higher impact when the
341 individuals in the family were sequenced at low-coverage (1x) than when they were
342 sequenced at high-coverage (30x). Relaxing the highest (0.999) to the lowest allele
343 threshold (0.6) increased the %Correct in the focal individual by 72.58% for a family
344 sequenced at 1x and by 1.81% for a family sequenced at 30x (Figure 2), and increased
345 the %Error by 8.63% for a family sequenced at 1x and by 0.08% for a family
346 sequenced at 30x (Figure 3).

347 Figure 2 and Figure 3 show that an allele threshold of 0.90 gave high
348 %Correct and sufficiently low %Error (<1.00%). For this reason, in the next section
349 of results we fix the allele threshold at 0.90 and explore the impact of window size for
350 the 50 sub-scenarios of the Single Family Phasing.

351 *Window size.* Within a given sequencing coverage, changing window sizes
352 produced small differences in %Correct and tiny and non-monotonic differences in
353 %Error (Figure 4). The differences in %Correct were only visible at 5x (4.01% of
354 differences) or higher sequencing coverage. Increasing window size increased the
355 %Correct. This is because bigger window sizes allowed more haplotypes to be
356 identified as being shared by the focal individual, its relevant parent and grandparent.
357 At these coverages the differences in %Error were still tiny ($\leq 0.18\%$).

358 When a family was sequenced at 1x, differences were too small to see: less
359 than 0.47% for %Correct and less than 0.01% for %Error. At 2x the differences were
360 also very small: less than 1.72% for %Correct and less than 0.05 for %Error. One of
361 the reasons for these trends is that coverage less than or equal to 2x on the 7 family
362 members produced a very low number of phased alleles at heterozygote variants in a

363 given window and thus there were not enough phased alleles to build the consensus
364 haplotypes even if the search space (i.e., window size) was large.

365 Figure 4 shows that a window size of 100,000 variants gave similar %Correct
366 and %Error to bigger window sizes but made it possible to parallelise the computation
367 of step 4 and 5. For this reason, in the next section of results we set the window size to
368 100,000 variants.

369 *Phasing behaviour for focal individual, parents and grandparents.*
370 AlphaFamSeq is designed to phase the sequence data of the focal individual. Some
371 phasing is achieved for the parents and grandparents but this is a by-product of the
372 steps taken to phase the focal individual and we expected to have less yield. Figure 5
373 shows that for all the sequencing coverages the focal individuals had the highest
374 %Correct, with the parents next, and the grandparents last. The differences in
375 %Correct between focal individual, parents and grandparents tended to decrease as
376 the sequencing coverage increased. At 1x the percentage of correctly phased alleles
377 was 30.62% (focal individual), 23.51% (parents) and 2.04% (grandparents). Whereas
378 at 30x the percentage of correctly phased alleles is 98.9% (focal individual), 93.83%
379 (parents) and 68.96% (grandparents).

380 The %Error behaved differently. The grandparents always had the lowest
381 %Error ($\leq 0.22\%$), followed by the focal individuals ($\leq 0.85\%$) and the parents
382 ($\leq 0.95\%$).

383 *Genotype imputation.* Figure 6 shows the results of imputation for the
384 homozygote genotypes “0” and “2”, and the heterozygote genotype “1”. At all the
385 sequencing coverages tested, the heterozygote genotypes always had the highest
386 %Error, reaching 1.63% at 5x and 1.55 at 2x. The two homozygote genotypes had
387 similar %Error ($\leq 0.25\%$).

Impact of AlphaFamSeq parameters on Single Family Phasing when all individuals in the family were sequenced at different coverages

When all individuals in a single family were sequenced at different coverages, the best possible %Correct reached 90% of the possible maximum at $\sim 1/6$ of the maximum aggregate coverage. Figure 7 summarizes the %Correct for each of the 279,936 scenarios of Single Family Phasing when all individuals in the family were sequenced at different coverages. The total aggregate coverage for one family of 7 members is shown on the x-axis and the %Correct is shown on the y-axis. The %Correct for the focal individual ranged from 0% to 98.9%. The %Error ranged from 0% to 3.63%. The horizontal dotted line represents the scenarios with the biggest difference in term of aggregate coverage (from 14x to 180x) that produced the same %Correct (69.6%), and highlights that the same phasing accuracy can be achieved at very different aggregate coverages. By way of example, in the scenario with 14x of aggregate sequencing coverage the focal individual was sequenced at 5x, the sire at 1x, the dam at 5x, the paternal grandsire at 0x, the paternal granddam at 1x, the maternal grandsire at 1x, the maternal granddam at 1x, while in the in the scenario with 180x of aggregate sequencing coverage the focal individual was not sequenced, and the parents and grandparents were sequenced at 30x.

The vertical dotted line represents the scenarios with the biggest difference in terms of %Correct (from 0% to 95.2%) at the same aggregate coverage of 60x, and highlights the fact that the same aggregate sequencing coverage can deliver very different phasing accuracies. Good phasing accuracy for the focal individual could be achieved with many different investments and distributions of investments. For example, by investing 100x aggregate coverage (the focal individual at 30x, the parents at 15x, 2 of the grandparents at 15x, and the other 2 grandparents at 5x), a

413 %Correct of 97.75% could be achieved. In contrast sequencing all seven family
414 members at 30x, more than doubled the aggregate coverage (210x), only produced a
415 1.15% increase in %Correct (98.9%).

416 Overall, the cloud of scenarios in Figure 7 shows that there are different
417 combinations of sequencing coverage in a 7 member family that have different
418 aggregate coverage, and thus costs, and give different %Correct. This cloud is taken
419 as input by AlphaSeqOpt when it is optimising the distribution of sequencing
420 resources across a population.

421

422 **Performance of Multiple Family Phasing**

423 Phasing multiple families simultaneously, increased the %Correct and slightly
424 increased the average %Error.

425 AlphaSeqOpt was instructed to find 500 focal individuals from the pedigree of
426 59,194 individuals and to assign £317,800 worth of sequencing resources to these
427 individual and their ancestors in an optimal way. It assigned sequence resources to
428 1,137 individuals on 1,589 in total. Of these 176 individuals sequenced at 1x, 626
429 individuals sequenced at 2x, 305 individuals sequenced at 5x and 30 individuals
430 sequenced at 15x (Figure 8A), and the total budget used was £317,720. To quantify
431 the increase in phasing performance produced by phasing multiple connected families
432 simultaneously, the data was phased by AlphaFamSeq using three generation pedigree
433 that connected the multiple families and by treating each of the 500 families of the
434 focal individuals as independent unconnected families. Based on the single family
435 phasing results we chose an allele probability threshold of 0.90 and a window size of
436 100,000 variants for the analysis of the Multiple Family Phasing scenario.

437 The phasing results for the focal individuals are reported in Figure 8B and
 438 Figure 8C. Figure 8B shows %Correct for the focal individuals when treating the
 439 families as independent and unconnected is shown on the x-axis. The %Correct for
 440 the focal individuals when treating the families as connected is shown on the y-axis.
 441 The average %Correct for the focal individuals increased from 47.08% to 65.25%
 442 when the families were treated as connected. 486 of the 500 focal individuals had a
 443 higher %Correct and the greatest increase in %Correct was from 10.03% to 66.34%.
 444 The focal individuals who gained most from connecting families were the 133 focal
 445 individuals not sequenced (from 22.70% to 48.15%), followed by the 53 focal
 446 individuals sequenced at 1x (from 40.31% to 63.67%), the 206 focal individuals
 447 sequenced at 2x (from 51.8% to 68.33%), the 55 focal individuals sequenced at 5x
 448 (from 69.67% to 79.93%) and finally the 9 focal individuals sequenced at 15x (from
 449 83.3% to 90.08%). The individual with the greatest reduction in %Correct by the
 450 connecting the families only had a small reduction (-1.58 %).

451 Figure 8C shows that the average %Error increased by connecting the families
 452 in the analysis from 0.84% on average to 1.51%. The greatest increase in %Error was
 453 for the 133 focal individuals not sequenced (from 0.32% to 2.39%), followed by those
 454 sequenced at 1x (from 0.75% to 1.14%), at 2x (from 1.02% to 1.21%) and at 5x (from
 455 1.19% to 1.24%) whereas for the 9 focal individuals sequenced at 15x the average
 456 %Error decreased from 0.86% to 0.81%.

457

458 **Impact of allele frequencies on imputation results.**

459 Imputation performance was good across all allele frequencies, except for a
 460 slight decrease for very low frequencies ≤ 0.01 - Figure 9). On average the correlation

461 by marker between true and imputed genotype was 0.94. As we would expect, the
462 accuracy of imputation is low at low allele frequencies: 0.63 at MAF bin ≤ 0.01 .

463 Discussion

464 In this paper we developed a family-based phasing algorithm for variable-
465 coverage sequence data that first minimises phasing errors and then maximises the
466 proportion of alleles phased. This algorithm performed well when variable coverage
467 sequence data was present for an individual and their parents and grandparents.
468 Accuracy was increased when additional information was available on more distant
469 relatives. Performance was stable across a range of user-defined parameters. We next
470 discuss the algorithm's performance, possible improvements to it, and its place in a
471 toolkit for imputing and phasing livestock sequence data.

472 The performance of the algorithm

473 In the single family phasing scenarios where focal individuals were phased
474 based on only their sequence data and that of their parents and grandparents, we found
475 that increasing the coverage increased our accuracy, measured by the percentage of
476 correctly phased alleles (%Correct) regardless of the parameters. At 30x the average
477 %Correct was 97.6% and the average %Error was 0.03%.

478 Scenarios with extremely high or extremely low allele thresholds performed
479 poorly in families sequenced at low coverage. For example, in single families
480 sequenced at 1x, a threshold that was too strict (i.e. 0.999) resulted in almost no
481 phased alleles (0.47 %Correct), whereas a threshold that was too relaxed (i.e. 0.6)
482 showed a high %Error (8.68%). For this reason we choose to use an allele threshold of
483 0.9, because it was the one that give a sufficiently low %Error (<1%) and a

484 sufficiently high %Correct (>30%). The allele threshold had a lower impact on
485 scenarios with family sequenced at high coverage. The reason is that individuals
486 sequenced at high coverage have enough observed reads for the variants and so it is
487 much easier to accurately phase their alleles and to impute any missing alleles based
488 on information from family members. On the other hand, individuals sequenced at
489 low coverage do not have enough observed reads to accurately phase the alleles.
490 Decreasing the allele threshold to low values increases the chance of phasing the
491 alleles, but increases the chance of phasing them incorrectly.

492 The size of the imputation window had a very small effect in a single family.
493 When a family was sequenced at $\leq 2x$ differences were too small to see. When a
494 family was sequenced at $5x$ coverage or higher increasing the window size increased
495 the %Correct. One of the reasons for this trends is that coverage less than or equal to
496 $2x$ on the 7 family members produced a very low number of phased alleles at
497 heterozygous loci in a given window and thus there were not enough phased alleles to
498 build the consensus haplotypes even if the search space (i.e., window size) was large.
499 The highest %Error occurred at intermediate window sizes. One of the possible
500 explanations is that, at each iteration, AlphaFamSeq was set to relax the allele
501 threshold to a greater degree, from a value of 1.00 in the first iteration to a value of
502 0.90 in the final iterations. Thus, in the early iterations the algorithm used highly
503 informative variants and, with large window sizes, there were sufficient informative
504 variants to build and to impute highly accurate haplotypes for the focal individual and
505 its parents and grandparents. With small window sizes relaxing the allele threshold
506 increased the number of informative genotypes for each window and thus enabled
507 consensus haplotypes to be built and imputed, however the accuracy of these
508 haplotypes was lower, which increased the %Error.

Results on single families sequenced at fixed coverage lead us to believe that it is better to sequence individuals at high coverage to increase the phasing results. But results in Figure 7 show that our algorithm performs well with variable coverage. In fact, the clouds of results in Figure 7 shows that the maximum %Correct was 98.9% when all the 7 family members were sequenced at 30x (aggregate of 210x), but that a similar accuracy could be obtained with only an aggregate of 122x across a family, which is on average 17.4x per individual.

Our algorithm also performs well when multiple families are phased in the context of a larger, connected pedigree. Connecting families sequenced at variable coverage increased the %Correct for the focal individuals. In the scenario with an average sequencing coverage of 2x the %Error were still acceptable (on average 1.52%), although it increased substantially for some individuals without sequence data (up to 11.3 %Error). Moreover, imputation was good at all allele frequencies (average correlation by marker of 0.94), although there was a slight decrease for low allele frequency (≤ 0.01 MAF).

Possible improvements to the algorithm

Sequence data is noisy for a variety of reasons (i.e., quality of the reference genome [18,19], misalignment of the reads [20], index switching [21,22], reference allele bias [23]). Low coverage sequence data also comes with the risk of calling genotypes incorrectly due to a small number of reads (i.e., heterozygous genotypes that are called as homozygous because only one of the alleles has read information). Being able to call alleles at a high accuracy while minimizing the number of incorrectly called alleles is essential for using low coverage sequence data. In populations that share segments identical-by-descent, there is an advantage in

533 performing the allele- and genotype-calling accumulating the sequence information of
534 related individuals, as AlphaFamSeq does. This increases the accuracy of phasing for
535 individuals that are sequenced at low-coverage.

536 In the first step of AlphaFamSeq we utilize an iterative peeling algorithm to
537 initially pass and accumulate read information between relatives [13]. This peeling
538 algorithm works only on a single allele at a time, utilizing family relationships, and
539 does not use linkage information. We anticipate that it would be possible to increase
540 the yield of the initial iterative peeling steps by using a multi-locus version of the
541 iterative peeling algorithm, such as LDMIP [24]. Multi locus peeling has the
542 advantage of taking linkage information into account, allowing for a more accurate
543 sharing of read information. However, currently LDMIP does not scale well to whole
544 chromosome sequence data, and so an approach such as that was not used. Moreover,
545 the computational cost of peeling algorithms increase when the complexity and size of
546 the pedigree increases.

547 The sequence data use by AlphaFamSeq are the observed number of reads for
548 the reference and the alternative alleles. One of the possible improvements in the
549 algorithm is to use the information from the raw sequence reads (i.e., those stored in
550 bam or sam files [25]) as input data. These raw sequence reads have the advantage
551 that they may cover two or more heterozygote variants and so provide information on
552 the physical linkage of these variants. This linkage information can be used to
553 improve the phasing accuracy and reduce the switching errors [26–28]. An extension
554 of this would be to use some long read information from the single-molecule long-
555 read sequencing [28]. This technology can generate reads which are much longer
556 compare to the reads length of the NGS, but on the other hand, the sequencing errors

are high [29]. Modifying AlphaFamSeq to handle such errors and utilise long read information would increase its phasing accuracy.

In our simulations we tested only small pedigrees of related individuals and we only supplied simulated sequence data to AlphaFamSeq. The incorporation of more sequence data on more animals and incorporation of cheaper genotype data on many animals would improve the phasing results. Genotype data are available nowadays for large numbers of animals, and even if they represent a small portion of SNP of the whole genome sequence, they can help to reconstruct the haplotypes of not sequenced individuals or individuals sequenced at low coverage.

An overall strategy for imputing sequence data for whole populations and the role of AlphaFamSeq

We believe that the key to perform whole genome sequencing for large livestock populations can be broken up into four steps. First, animals that carry a large proportion of haplotypes in the population need to be identified and sequenced, potentially at variable coverage to optimise use of sequencing resources. Second, the sequence data needs to be phased to provide a reference panel for downstream imputation. Third, these two steps can be complemented with the judicious use of low-coverage sequence on some individuals. Fourthly, an imputation algorithm needs to be used to pass the phased haplotypes to the remaining individuals in the population. This sequencing strategy stems from the fact that cheap genotype arrays are a fraction of the cost of sequencing individuals, and livestock populations have a high degree of relatedness allowing for large shared haplotypes and enabling high accuracy imputation.

AlphaSeqOpt [5] solves the first problem by determining which individuals in a population should be sequenced and the coverages that they should be sequenced at. AlphaFamSeq solves the second problem by providing a set of reference haplotypes that can be used for downstream imputation. The last problem could be handled by either heuristic (e.g. AlphaImpute [6], Findhap [7], Fimpute [8]) or probabilistic (e.g. MaCH [9], Beagle [10,11], Impute2 [12]) imputation algorithms or a combination of both [30]. We discuss these steps in more detail below.

The first step of sequencing is to determine which animals (i.e., focal individuals) contain a large number of the haplotypes shared with many other animals in the population and determining the optimal distribution of sequencing resources across these focal individuals and their ancestors (parents and grandparents) according to the genomic footprint of the focal individuals on the population. Methods such as AlphaSeqOpt [5] use phased genotype data from genotype arrays to identify high-frequency haplotypes in the population and identify focal individuals who carry these haplotypes. Part of the sequencing resources is then spent on the focal individual and their parents and grandparents to genotype and phase the haplotypes of the focal individuals. To determine how to spend sequencing resources AlphaSeqOpt requires a function that estimates the phasing accuracy for a focal individual depending on the individual's coverage and that of their family. AlphaFamSeq fills this gap by providing a way to estimate the expected phasing accuracy for individuals based on a family's coverage.

The next step of our sequencing strategy is to take the variable coverage sequence information and phase the haplotypes of focal individuals. AlphaFamSeq, as described in this paper, performs this step and can phase alleles for focal animals. The

high accuracy haplotypes generated by AlphaFamSeq are key for downstream imputation.

The two steps described above could be complemented by a sequencing strategy that we call LCSeq. LCSeq aims to generate accurate sequence for the haplotypes in the population. LCSeq sequences individuals at low-coverage and assembles high-coverage sequence information for every haplotype by accumulating the low-coverage sequence data from the genome segments that are shared between many individuals to derive the ‘consensus haplotypes’. LCSeq uses the consensus haplotypes to impute the sequence data of the individuals. Accurate derivation of the consensus haplotypes is critical for phasing and imputation accuracy. Efficient derivation of the consensus haplotypes requires distribution of sequence data across the population such that a maximal number of haplotypes are sequenced to a coverage (e.g., 20x) that is sufficient to enable them to be phased accurately, which requires an algorithm such as that developed by Ros et al. [31]. Within the context of LCSeq, AlphaFamSeq can provide set of accurate starting haplotypes from which many of the alternate haplotypes carried by individuals sequenced at low coverage can be derived. AlphaFamSeq will increase the numbers of haplotypes accurately phased by consensus because any individual that is sequenced at low coverage and who shares haplotypes with a focal individual will have those haplotypes phased and many of the other haplotypes that it carries will be phased either by being the complement of a shared haplotype or by low-coverage sequence reads.

The fourth step of our sequencing strategy involves taking the phased reference haplotypes generated by AlphaFamSeq, or LCSeq, and imputing them to the remaining individuals in the population. Several methods or combinations of methods

could be used for this (e.g. AlphaImpute [6], Findhap [7], Fimpute [8], MaCH [9], Beagle [10,11], Impute2 [12]). The availability of a pre-phased set of individuals or haplotypes greatly decreases the computational time of these algorithms and increases their accuracy [11,12,30]. AlphaFamSeq, with its low percentage of alleles that incorrectly phased and high percentage of alleles that are correctly phased can provide such pre-phased data.

Conclusions

This paper describes a family-based phasing algorithm, for variable-coverage sequence data, that first minimises phasing errors and then maximises the proportion of alleles correctly phased. We tested the algorithm in several scenarios that are typical of those found in livestock breeding and genetics. It can work within a given family separately, can work with multiple families simultaneously by utilising pedigree information that connects them and can handle variable coverage sequencing data. It gives low phasing error rates and, if there is sufficient sequence information and haplotype sharing amongst individuals, it can give a high yield of phased alleles. We envisage that AlphaFamSeq will be one of a number of essential tools that would underpin an overall strategy for generating highly accurate sequence data on whole populations at low cost. The role of AlphaFamSeq in this overall strategy is to provide very accurately phased haplotypes on focal individuals, who are individuals whose haplotypes are very common in the population.

Supporting data

Programs to simulate the true genotypes (AlphaSim version 1.08), the sequence reads (SimulateSequenceReads), to optimally distribute the sequencing

resources (AlphaSeqOpt version 1.00) and perform phasing (AlphaFamSeq version 1.00) used in this paper are available from the AlphaGenes website [32].

654

655 **Competing interests**

656 The authors declare that they have no competing interests.

657 **Authors' contributions**

658 JMH conceived the algorithm and supervised the study. MB further developed the
659 algorithm, implemented it in software, performed all of the analysis and wrote the first
660 draft of the paper. RRF, SG, GG and AW contributed to the development of the
661 method, the interpretation of the results and to the writing of the paper.

662 **Acknowledgements**

663 The authors acknowledge the financial support from the BBSRC ISPG to The Roslin
664 Institute "BB/J004235/1", from Genus PLC and from grant numbers
665 "BB/M009254/1", "BB/L020726/1", "BB/N004736/1", "BB/N004728/1",
666 "BB/L020467/1", "BB/N006178/1" and Medical Research Council (MRC) grant
667 number "MR/M000370/1". This work has made use of the resources provided by the
668 Edinburgh Compute and Data Facility (ECDF) (<http://www.ecdf.ed.ac.uk>). The
669 authors thank Dr Andrew Derrington (Scotland, UK) for assistance in refining the
670 manuscript.

671 John Hickey would like to acknowledge the outstanding contribution by Robin
672 Thompson to the fields of statistics, animal breeding and plant breeding. Robin has
673 been a great teacher, mentor, friend and inspiration to him over many years.

References

1. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 2014;46:858–65.
2. Veerkamp RF, Bouwman AC, Schrooten C, Calus MPL. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. *Genet. Sel. Evol.* 2016;48:95.
3. VanRaden PM, Tooker ME, O’Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol.* 2017;49:32.
4. Hickey JM. Sequencing millions of animals for genomic selection 2.0. *J. Anim. Breed. Genet.* 2013;130:331–2.
5. Gonen S, Ros-Freixedes R, Battagin M, Gorjanc G, Hickey JM. A method for the allocation of sequencing resources in genotyped livestock populations. *Genet. Sel. Evol.* 2017;49.
6. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genet. Sel. Evol.* 2012;44:11.
7. VanRaden PM, Sun C, O’Connell JR. Fast imputation using medium or low-coverage sequence data. *BMC Genet.* 2015;16:82.
8. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics.* 2014;15:478.
9. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 2010;34:816–34.
10. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 2007;81:1084–97.
11. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* 2016;98:116–26.
12. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009;5:e1000529.
13. Kerr RJ, Kinghorn BP. An efficient algorithm for segregation analysis in large populations. *J. Anim. Breed. Genet.* 1996;113:457–69.

14. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 2011;6:e19379.
15. Faux A-M, Gorjanc G, Gaynor RC, Battagin M, Edwards SM, Wilson DL, et al. AlphaSim: Software for Breeding Program Simulation. *Plant Genome*. 2016;9.
16. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. *Genome Res*. 2009;19:136–42.
17. MacLeod IM, Larkin DM, Lewin HA, Hayes BJ, Goddard ME. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. *Mol. Biol. Evol*. 2013;30:2209–23.
18. Servin B, Faraut T, Iannuccelli N, Zelenika D, Milan D. High-resolution autosomal radiation hybrid maps of the pig genome and their contribution to the genome sequence assembly. *BMC Genomics*. 2012;13:585.
19. Warr A, Robert C, Hume D, Archibald AL, Deeb N, Watson M. Identification of Low-Confidence Regions in the Pig Reference Genome (Sscrofa10.2). *Front. Genet*. 2015;6.
20. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*. 2011;43:491–8.
21. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*. 2017;125724.
22. Owens GL, Todesco M, Drummond EBM, Yeaman S, Rieseberg LH. A Novel Post Hoc Method For Detecting Index Switching Finds No Evidence For Increased Switching On The Illumina HiSeq X. *bioRxiv*. 2017;142356.
23. Chen X, Listman JB, Slack FJ, Gelernter J, Zhao H. Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next Generation Sequencing of Pooled Samples. *Genet. Epidemiol*. 2012;36:549–60.
24. Meuwissen T, Goddard M. The Use of Family Relationships and Linkage Disequilibrium to Impute Phase and Missing Genotypes in Up to Whole-Genome Sequence Density Genotypic Data. *Genetics*. 2010;185:1441–9.
25. Samtools. GitHub. <https://github.com/samtools>. Accessed 23 August 2017.
26. Delaneau O, Marchini J, McVean GA, Donnelly P, Lunter G, Marchini JL, et al. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun*. 2014;5:3934.
27. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet*. 2016.

- 746 28. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly
747 for diverse sequencing technologies. *Genome Res.* 2017;27:801–12.
- 748 29. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-
749 generation sequencing technologies. *Nat. Rev. Genet.* 2016;17:333–51.
- 750 30. Antolín R, Nettelblad C, Gorjanc G, Money D, Hickey JM. A hybrid method for
751 the imputation of genomic data in livestock populations. *Genet. Sel. Evol.*
752 2017;49:30.
- 753 31. Ros-Freixedes R, Gonen S, Gorjanc G, Hickey JM. A method for allocating low-
754 coverage sequencing resources by targeting haplotypes rather than individuals. *GSE*
755 Final Rev.
- 756 32. AlphaGenes. www.alphagenes.roslin.ed.ac.uk. Accessed 28 August 2017.

757 **Figures**

758

763

764 **Figure 2.** Average effect of different allele thresholds and sequencing coverage on the
765 percentage of correctly phased alleles (%Correct) for the focal individuals when all
766 the seven family members are sequenced at the same coverage (1x, 2x, 5x, 10x, 15x,
767 or 30x) within each analysis.

768

769 **Figure 3.** Average effect of different allele thresholds and sequencing coverage on the
770 percentage of incorrectly phased alleles (%Error) for the focal individuals when all
771 the seven family members are sequenced at the same coverage (1x, 2x, 5x, 10x, 15x,
772 or 30x) within each analysis.

Figure 4. Effect of different window sizes and sequencing coverage on the percentage of correctly phased alleles (%Correct) and on the percentage of incorrectly phased alleles (%Error) for the focal individuals when all the seven family members are sequenced at the same coverage (1x, 2x, 5x, 10x, 15x, or 30x) within each analysis and the allele threshold is set to 0.90.

780

781 **Figure 5.** Effect of generation and sequencing coverage on the percentage of correctly
782 phased alleles (%Correct) and on the percentage of incorrectly phased alleles
783 (%Error) for the focal individuals when all the seven family members are sequenced
784 at the same coverage (1x, 2x, 5x, 10x, 15x, or 30x) within each analysis, allele
785 threshold is set to 0.90 and window size is set to 100,000 variants.

786

787 **Figure 6.** Effect of sequencing coverage on the percentage of incorrectly imputed
788 genotypes (%Error) for the focal individuals when all the seven family members are
789 sequenced at the same coverage (1x, 2x, 5x, 10x, 15x, or 30x) within each analysis,
790 allele threshold is set to 0.90 and window size is set to 100,000 variants.

Figure 7. Average effect of the family aggregate coverage on the percentage of correctly phased alleles (%Correct) for the focal individuals when AlphaFamSeq is tested within independent families of 7 members. Allele threshold equal to 0.90 and window size is 100,000 variants. All the possible 279,936 combinations of sequencing depths (0, 1x, 2x, 5x, 15x and 30x) are tested.

805

806 **Figure 8. A)** Distribution of sequencing coverage for 1,589 individuals in the
807 pedigree when a budget of £317,720 was invested to sequence the individuals. **B)**
808 Expected correctly phased alleles (%Correct – independent families) against the
809 realised correctly phased alleles (%Correct – connected families) for the 500 focal
810 individuals. **C)** Expected incorrectly phased alleles (%Error – independent families)
811 against the realised incorrectly phased alleles (%Error – connected families) for the
812 500 focal individuals.

Figure 9. Performance of AlphaFamSeq on correlation by markers binned by minor allele frequency (MAF).

817 Additional file

818

819 File name: Additional file 1

820 Format: .txt

821 Title: Parameters used to run AlphaSeqOpt.

822 Description: parameters used to run AlphaSeqOpt (version 1.00). The parameters are
823 fully described in www.alphagenes.roslin.ed.ac.uk/alphasuite-software/alphaseqopt/

824 #####

825 # ALPHASEQOPTSPEC #

826 #####

827 OptimisationMethod ,Sequenceandcoverage

828 NumberOfChromosomes ,1

829 NumberOfSnps ,Constant,2100

830 IndividualPhasedThreshold ,0.9

831 SnpPhasedThreshold ,0.9

832 HaplotypeThresholdPhasing ,0.97

833 HaplotypeThresholdComparing ,0.97

834 HaplotypeMismatchAllowed ,3

835 HaplotypePrinting ,Summary

836 NumberOfIndividualsToSequence ,500

837 CoreLength ,100

838 ByPassSharedHaploDeinition ,No

839 PriorNumberOfCores ,0

840 SequencedIndividualFile ,None

841 IndividualsNotToSequenceFile ,None

842 MatrixOfScenariosFile ,UpdatedScenario.txt

843 MatrixOfMetricsFile ,UpdatedMetrics.txt

844 TopIndividualsFile ,TopIndividualsToSequence.txt

845 NumberOfFamiliesToSequence ,500

846	PriorSeqCostsFile	,None
847	OverallTotalCost	,317800
848	NumberOfCoverage	,6
849	AllowedCoverage	,0,1,2,5,15,30
850	EvolAlgProbOrMaxCoding	,1
851	EvolAlgPopSize	,100
852	EvolAlgOptimisationMethod	,EvolutionaryAlgorithm
853	EvolAlgOptimisationRounds	,10000
854	EvolAlgOptimisationRoundsConv	,10000
855	EvolAlgOptimisationRoundsPrint	,10
856	EvolAlgRecombinationType	,Regions
857	EvolAlgRecombination	,0.05,2.00
858	EvolAlgWeightFactor	,0.05,0.50
859	EvolAlgRoundsSwapGAorDE	,1
860	CostOfLibraryPerIndividual	,40
861	CostOfLibraryOneXSequencing	,80
862	Metric	,Accuracy