1 **A method for allocating low-coverage sequencing**

2 **resources by targeting haplotypes rather than**

3 **individuals**

4

5 Roger Ros-Freixedes, Serap Gonen, Gregor Gorjanc, John M Hickey[§]

6

7 The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of

8 Edinburgh, Easter Bush, Midlothian, Scotland, UK

9

10 [§]Corresponding author

11

12 Email addresses:

13 RRF: roger.ros@roslin.ed.ac.uk

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

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

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

17

Abstract

Background

18   This paper describes a heuristic method for allocating low-coverage

19   sequencing resources by targeting haplotypes rather than individuals. Low-coverage

20   sequencing assembles high-coverage sequence information for every individual by

21   accumulating data from the genome segments that they share with many other

22   individuals into consensus haplotypes. Deriving the consensus haplotypes accurately

23   is critical for achieving a high phasing and imputation accuracy. In order to enable

24   accurate phasing and imputation of sequence information for the whole population we

25   allocate the available sequencing resources among individuals with existing phased

26   genomic data by targeting the sequencing coverage of their haplotypes.

Results

27   Our method, called AlphaSeqOpt, prioritizes haplotypes using a score function

28   that is based on the frequency of the haplotypes in the sequencing set relative to the

29   target coverage. AlphaSeqOpt has two steps: (1) selection of an initial set of

30   individuals by iteratively choosing the individuals that have the maximum score

31   conditional to the current set, and (2) refinement of the set through several rounds of

32   exchanges of individuals. AlphaSeqOpt is very effective for distributing a fixed

33   amount of sequencing resources evenly across haplotypes, which results in a

34   reduction of the proportion of haplotypes that are sequenced below the target

35   coverage. AlphaSeqOpt can provide a greater proportion of haplotypes sequenced at

36   the target coverage by sequencing less individuals, as compared with other methods

37   that use a score function based on the haplotypes population frequency. A refinement

38   of the initially selected set can provide a larger more diverse set with more unique

39   individuals, which is beneficial in the context of low-coverage sequencing. We extend

40 the method with an approach to filter rare haplotypes based on their flanking

41 haplotypes, so that only those that are likely to derive from a recombination event are

42 targeted.

Conclusions

43 We present a method for allocating sequencing resources so that a greater

44 proportion of haplotypes are sequenced at a coverage that is sufficiently high for

45 population-based imputation with low-coverage sequencing. The haplotype score

46 function, the refinement step, and the new approach of filtering rare haplotypes make

47 AlphaSeqOpt more effective for that purpose than methods reported previously for

48 reducing sequencing redundancy.

Introduction

49   This paper describes a heuristic method for allocating low-coverage

50   sequencing resources by targeting haplotypes rather than individuals so that

51   haplotypes have a coverage that is sufficiently high for population-based imputation.

52   The use of whole-genome sequencing data has great potential in livestock

53   breeding programs. It may increase the power of discovery of causative variants [1–3]

54   and may enable more accurate and persistent predictions of breeding values than

55   marker array genotypes [4,5]. To capture the full potential of sequence data in

56   livestock, sequence and phenotype data on a large number, perhaps millions, of

57   individuals may be required to accurately estimate the effects of the large number of

58   causative variants that underlie quantitative traits [6].

59   Low-cost sequencing strategies combined with imputation can be utilised to

60   generate the required amount of sequence information for a large number of

61   individuals at an affordable cost [7–11]. The strategies for low-cost sequencing can be

62   classified into three groups: (1) to sequence a certain number of key individuals at

63   high coverage, as in the 1,000 Bull Genomes project (**KeySires**) [2,5]; (2) to sequence

64   a larger number of individuals at low coverage (**LCSeq**) [6,12,13]; and (3) to

65   sequence a set of chosen individuals at a wide range of coverages (**VarCoverage**)

66   [14].

67   The LCSeq approach exploits the fact that the population structures that are

68   typical in livestock breeding result in individuals being sufficiently related to share

69   large genome segments. LCSeq focuses sequencing on the haplotypes in the

70   population rather than on any individual. LCSeq sequences individuals at low

71   coverage and assembles high-coverage sequence information for every haplotype by

72   accumulating the low-coverage sequence data from the genome segments that are

73   shared between many individuals to derive the 'consensus haplotypes'. The consensus

74   haplotypes are then used to impute the sequence data of the individuals. Deriving the

75   consensus haplotypes accurately is critical for achieving a high phasing and

76   imputation accuracy under the LCSeq strategy.

77   With the LCSeq approach potentially many more individuals can be sequenced

78   than with the KeySires or VarCoverage approaches. This provides three advantages to

79   the LCSeq approach: (1) higher variant discovery rates, particularly for low-frequency

80   variants [15]; (2) inclusion of rare haplotypes; and (3) a more precise capture of the

81   recombination events that have occurred in the population, which would enable better

82   definition of the haplotypes that are present in the population and thus better

83   imputation of these haplotypes into the individuals that carry them.

84   There are methods to optimise the selection of individuals for sequencing for

85   the three alternate sequencing approaches. Most of these methods focus only on the

86   choice of which individuals to sequence with the aim to impute their sequence

87   information into their relatives [5,16–18]. Recently, Gonen et al. [14] proposed a

88   method that identifies the individuals with the largest genetic footprint on the

89   population and optimises the allocation of sequence resources across these focal

90   individuals and their ancestors with the aim to maximise phasing accuracy of their

91   sequenced haplotypes when using family-based phasing methods.

92   Although LCSeq could be used alone, we envisage a sequencing strategy in

93   two stages for facilitating the imputation of sequence data. The first stage uses the

94   method developed by Gonen et al. [14] with the aim of producing a set of accurately

95   phased haplotypes that are shared by a lot of individuals in the population. The second

96   stage seeks to complement the first stage by applying the LCSeq approach as

97   described above to spread low-coverage sequence data across the population so that

98   whole-genome sequence data can be imputed to the whole population, which in turn

99   will be enhanced by the phasing of the most common haplotypes achieved in the first

100   stage. To do this effectively a method for optimising the allocation of sequencing

101   resources under the LCSeq approach should be developed.

102   We hypothesise that such a method should maximise the sequencing coverage

103   of the maximum possible number of haplotypes because this would enable

104   population-based phasing and imputation methods rather than family-based

105   imputation methods to accurately phase and impute the data to all individuals. For

106   such population-based phasing and imputation methods, a certain level of sequence

107   coverage must be accumulated for accurate inference of a consensus haplotype. With

108   a prototype of such a population-based phasing and imputation method we observed

109   that there is a positive relationship between the coverage that a particular haplotype

110   accumulates across individuals and the imputation accuracy of a consensus haplotype

111   (for a description, see Additional file 1: Figure S1). A random allocation of

112   sequencing resources under the LCSeq approach results in some haplotypes being

113   sequenced many times, some rarely, and some not at all. To optimise the allocation of

114   sequencing resources under LCSeq we need to maximise the proportion of haplotypes

115   that are sequenced at the target coverage and minimise the proportion of haplotypes

116   that are under- or over-sequenced. Similarly, we need to minimise the sequencing

117   resources allocated to haplotypes that are too rare to have consensus haplotypes

118   inferred or their effects estimated accurately.

119   The objective of this work was to develop a method that uses haplotypes

120   derived from existing phased marker array genotypes to identify which individuals

121   should be sequenced, and at what coverage, to maximize the proportion of consensus

122   haplotypes sequenced at a minimum target coverage. Our method uses a score

123   function to identify a set of individuals based on the coverage at which their

124   haplotypes are sequenced and then it refines the initial set of individuals through

125   rounds of exchanges. We extend the method with an approach to filter out rare

126   haplotypes so that we only target those that are likely to derive from the

127   recombination of common haplotypes. We tested the performance of the algorithm

128   using simulated data and the results showed that our method is efficient in distributing

129   the sequencing resources evenly across a large proportion of the haplotypes observed

130   in the population.

131

Materials and Methods

132

Description of the AlphaSeqOpt method

133   Our method utilises existing phased marker array genotypes to identify which

134   individuals should be sequenced, and at what coverage, so that the maximum

135   proportion of haplotypes are sequenced at any minimum target coverage with a fixed

136   sequencing budget. The method has two main steps. In the first step, referred as

137   'initial set selection', an initial set of individuals is selected by iteratively choosing the

138   individuals that are the most complementary to the ones already in the set according

139   to a score function. In the second step, referred as 'set refinement', the initial set of

140   individuals is refined through several rounds of exchanges. The method was

141   implemented in a software package called AlphaSeqOpt, which also implements the

142   method of Gonen et al [14]. Throughout the rest of the paper, AlphaSeqOpt is used

143   when referring to our method.

144

145   **<u>Initialisation step:</u>**

146   0a: Construct a haplotype library for the population using existing phased

147   marker array genotypes. Split each chromosome into c cores of length m markers. A

148   'core' is each of the strings of m consecutive marker positions used to determine the

149   haplotypes. Within a core, strings of alleles (previously phased) are compared to

150   define which haplotype each individual carries in each parental chromosome. Strings

151   of alleles that are identical between two individuals are defined as a unique haplotype

152   and strings with multiple mismatches are defined as different haplotypes. A

153   predefined number of mismatches can be allowed before two strings of alleles are

154   defined as different haplotypes to account for sequencing errors.

155   0b: Calculate the maximum size of the sequencing set. Assuming linear

156   sequencing costs, the sequencing budget divided by the cost of 1x sequencing

157   determines the total amount of sequencing coverage that could be produced,

158   represented by the number of slots of the sequencing set. A 'slot' is each of the

159   positions in the set, which can be assigned to any given individual following the steps

160   below. Each slot corresponds to 1x sequencing.

161

162   **Initial set selection (step 1):**

163   1a: Calculate a score for each haplotype in each core. We derived a score

164   function that prioritizes the haplotypes that are closer to reaching the target coverage.

165   The score function is based on the frequency of a haplotype in the sequencing set

166   relative to the target coverage. The score function is:

167
$$\text{Score} = \begin{cases} \exp\left(\left(\frac{\text{HapCount}}{2\cdot\text{TargetCov}-1}\right)^2\right), & \text{if HapCount} < 2\cdot\text{TargetCov} \\ 0, & \text{if HapCount} \geq 2\cdot\text{TargetCov} \end{cases},$$

168   where HapCount is the number of times that a haplotype appears in the current

169   sequencing set and TargetCov is the target haplotype coverage (Figure 1). The score

170   increases every time that an individual that carries a given haplotype is added to the

171   sequencing set. When the haplotype count in the set reaches twice the target coverage,

172   which is the haplotype count required to produce the target coverage assuming that

173   for each x of coverage of an individual there is a probability of 0.5 of reading either

174   the paternal or maternal haplotype, the score is set to 0 to prevent over-sequencing of

175   well-covered haplotypes in favour of allocating sequencing resources to other

176   haplotypes.

177   1b: Calculate the total score for every individual as the sum of the scores of

178   the haplotypes that each individual carries at each core.

179   1c: Add the individual with the maximum score to the first available slot of the

180   initial set. If there is more than one individual satisfying this condition, one individual

181   is selected at random amongst those individuals with the maximum score. Repetition

182   of individuals in several slots of the set is allowed. The number of slots occupied by

183   the same individual indicates at what coverage it should be sequenced (i.e., an

184   individual that appears n times in the set should be sequenced at nx).

185   1d: Calculate the total cost of sequencing the current set as the cost of library

186   preparation times the number of individuals in the set plus the cost of 1x sequencing

187   times the total sequencing coverage produced.

188   1e: Repeat steps 1a to 1d until the initial set is complete (i.e., we have a set of

189   individuals at variable coverage that exhausts all the sequencing resources). Because

190   some resources are used for library preparation some slots will be left empty.

191

192   **Set refinement (step 2):**

193   2a: Choose randomly a predetermined number of slots of the set. Remove the

194   individuals assigned to these slots from the set.

195   2b: Repeat steps 1a to 1d to fill the emptied slots. Individuals removed in step

196   2a can go back into the set if they have the maximum individual score.

197   2c: If the exchanges result in the same or a greater percentage of unique

198   haplotypes sequenced at (or above) the target coverage, keep the new set. Otherwise,

199   discard the new set in favour of the previous set.

200   2d: Repeat steps 2a to 2c for a predefined number of exchange rounds.

201

202   If there are individuals that have been sequenced previously, AlphaSeqOpt can

203   account for the available sequence data easily by adding the pre-existing coverage of

204   their haplotypes to HapCount during the calculation of the haplotype scores in step 1a.

205   If there are not any individuals that have been sequenced previously, all haplotypes

206   will have the same starting score and the first individual will be selected at random

207   amongst those that have more non-missing haplotypes.

208   For any given target haplotype coverage, AlphaSeqOpt will produce a set of

209   individuals to be sequenced from 1x to a maximum coverage equal to twice the target

210   coverage. To ensure that all individuals are sequenced at a low coverage and that a

211   larger number of individuals is sequenced it is also possible to restrict the coverage

212   for the individuals in the set to a desired maximum (e.g., to 1x or 2x).

213   In the implementation of AlphaSeqOpt we are making two assumptions

214   regarding the yield of data from the sequencer: (1) that sequencing coverage is

215   uniform across the genome; and (2) that for each x of coverage of an individual there

216   is a probability of 0.5 of reading either the paternal or maternal haplotype and

217   therefore each haplotype receives half the coverage. Even though these assumptions

218   contradict empirical observations [19], there is no straightforward way of accounting

219   for variation of coverage across genome or between alleles prior to performing

220   sequencing. Regarding the sequencing costs, we are assuming that when we increase

221   the sequencing coverage we incur a linear increase of the sequencing costs.

222   AlphaSeqOpt can also account for non-linear cost structures by modifying the cost

223   equation used in step 1d.

224

Algorithm testing

225   The proposed method was tested against our implementation of the Inverse

226   Weight Selection (IWS) method as described by Bickhart et al. [17], our adaptation of

227   the IWS method to obtain more comparable results, and a method that selects the

228   individuals randomly (referred to as Random).

229   The IWS method as described by Bickhart et al. [17] follows the step 1 as

230   described above but in step 1a it uses an inverted parabolic score function $f_i^2 - 2f_i + 1$,

231   where f_i is the population frequency of the haplotype. Note that this function uses the

232   population frequency, while the score function that we propose uses the frequency of

233   the haplotype in the sequencing set relative to the target coverage. The two score

234   functions are compared in Figure 1. Another major difference with AlphaSeqOpt is

235   that Bickhart et al. [17] proposed targeting only homozygous haplotype cores based

236   on the marker array genotypes. Thus, the IWS method only scores such haplotypes

237   and it stops after the initial set is constructed, without a step of refinement.

238   Our adaptation of IWS mirrored the method that we propose more closely,

239   including a step of refinement of the initially selected set, with the only difference

240   being the score function used. This method follows both steps 1 and 2 as described

241   above but in step 1a it uses the inverted parabolic function $f_i^2 - 2f_i + 1$. We did not

242   follow the suggestion of targeting only the haplotypes at cores that are predicted to be

243   homozygous based on the marker array genotypes, because this would disadvantage

244   the adapted IWS method.

245   The Random method also used the algorithm described but individuals were

246   selected randomly instead of according to a score function. In the refinement step,

247   random exchanges of individuals were performed.

248   All methods were tested in a range of scenarios. The scenarios varied in the

249   target haplotype coverage (5x, 10x, or 15x) and in the total available sequencing

250   resources (£400,000, £800,000, or £1,600,000 GBP). We calculated the cost of each

251   scenario assuming a cost in library preparation of £40 and a cost in 1x sequencing of

252   £80. The tested sequencing resources would produce a total of 5,000x, 10,000x, or

253   20,000x whole-genome reads, respectively, if cost of library preparation was ignored.

254   Haplotypes observed only once or twice in the population were excluded from the

255   analyses unless stated otherwise. Additional tests were performed with a restriction of

256   maximum individual coverage of 1x, for different numbers of exchanges per round,

257   ranging from 1 slot to the total size of the set, and for different costs of library

258   preparation, ranging from no cost to £40. We performed 10 repetitions for all

259   analyses. The percentage of unique haplotypes sequenced at (or above) the target

260   coverage was used as the main criterion, together with the number of individuals

261   sequenced.

262   For simplicity, in some instances we will focus on the scenarios with a target

263   haplotype coverage of 10x but the algorithm can be used with any desired target

264   coverage.

265

Filtering of rare haplotypes based on flanking context

266   A new approach for filtering the rare haplotypes included in the analyses was

267   also developed. In this approach we filtered the rare haplotypes so that only those rare

268   haplotypes that are likely to derive from a recombination event between two common

269   haplotypes were targeted.

270   The filtering was based on two assumptions: (1) rare haplotypes that were

271   derived from a recombination event between common haplotypes will be flanked by

272   common haplotypes; and (2) there will be no other individuals that carry the same

273   combination of haplotypes at the cores that flank the rare recombined haplotype. The

274   second assumption could be false if, for example, there had been multiple

275   recombination events at different positions of the same core that produced multiple

276   rare recombinant haplotypes from the same two common haplotypes, but note that

277   this is a method for directing the sequencing resources among rare haplotypes, not an

278   exact method for capturing all recombination events. Note also that combinations of

279   consecutive cores with rare haplotypes could indicate either genomes that are

280   unrelated to the population or phasing errors.

281   We implemented the above filtering approach according to the population

282   count of the haplotypes at each core. In any given core, haplotypes with population

283   count below a predefined threshold are included in the analysis only if all of the

284   following conditions are met: (1) the rare haplotype is not at the first or last core of a

285   chromosome; (2) the counts of the flanking haplotypes are greater than a predefined

286   threshold (FlankCount); and (3) there are less than a predefined number (nComb) of

287   individuals carrying the same combination of haplotypes flanking the rare haplotype.

288   In our implementation of AlphaSeqOpt we used this filtering approach on

289   those rare haplotypes with population count ≤ 2 (observed only once in the population,

290   referred to as 'singletons', or twice, referred to as 'doubletons') using FlankCount=2

291   and nComb=3. The same method could be applied for any population count. This

292   approach for filtering the rare haplotypes was tested against the reference case with no

293   filtering and against the approach in which all singletons and doubletons were filtered

294   out.

295

Simulated dataset

296   To demonstrate the implementation of the algorithm, a testing dataset was

297   simulated to mimic a typical livestock population with known structured pedigree.

298   Sequence data was generated for 1,000 base haplotypes for each of ten

299   chromosomes using the Markovian Coalescent Simulator [20] and AlphaSim [21,22].

300   Chromosomes were simulated to be 100 cM and 10^8 base pairs in length, with a per

301   site mutation rate of 2.5×10^{-8} and a per site recombination rate of 1.0×10^{-8}. The

302   effective population size (N_e) was set to specific values during the simulation based

303   on previously estimated N_e values within the Holstein cattle population [23]. These

304   set values were: 100 in the base generation, 1,256 at 1,000 years ago, 4,350 at 10,000

305   years ago, and 43,500 at 100,000 years ago, with linear changes in between. The

306   resulting sequence had approximately 650,000 segregating SNP loci across the ten

307   chromosomes.

308   To enable the selection of sires for the generation of a pedigree, a quantitative

309   trait influenced by 10,000 QTN distributed equally across the ten chromosomes was

310   simulated. QTN positions were randomly chosen from the 650,000 segregating

311   sequence loci and their effect sizes sampled from a normal distribution with a mean of

312   zero and standard deviation of 0.01 (1.0 divided by the square root of the number of

313 QTN). The QTN effects were used to compute the true breeding value (TBV) for each

314 individual.

315   To emulate livestock breeding populations, a pedigree of 15 generations was

316 simulated. Each generation comprised 1,000 individuals in equal sex ratio (i.e., 500

317 males and 500 females). In the first generation, chromosomes for each individual

318 were sampled from the 1,000 sequence haplotypes in the base generation. In

319 subsequent generations, chromosomes of each individual were sampled from parental

320 chromosomes, assuming recombination with no interference. In each generation, the

321 25 males with the highest TBVs were selected as sires of the next generation. No

322 selection was performed on females, and all 500 females were used as parents.

323   All individuals were assumed to be genotyped with a panel of 10,000 SNP

324 markers distributed equally across the ten chromosomes. Marker genotypes of all

325 individuals were phased using AlphaPhase [24–26] as input for AlphaSeqOpt. The

326 parameters used for determining the population haplotype libraries were: (1)

327 population haplotype libraries were created using individuals and SNPs with at least

328 90% phased genotype data; (2) sharing of haplotypes was determined as 100%

329 identity matches; and (3) core lengths were set to 100 SNPs per chromosome.

330   In summary, the algorithm was tested using a dataset with 15,000 individuals.

331 Individuals had 10 chromosomes and 10 cores per chromosome. The total number of

332 haplotypes in the population was 8850 (on average, 88.5 haplotypes per core). Further

333 details on the simulated dataset can be found in Gonen et al. [14].

334

Software availability

335   The method has been implemented in the AlphaSeqOpt software package.

336 AlphaSeqOpt is available for download at

337   http://www.alphagenes.roslin.ed.ac.uk/alphaseqopt/, along with a detailed user

338   manual.

339

Results

340

Performance of algorithm

341   AlphaSeqOpt allocated sequencing resources to enable a greater percentage of

342   haplotypes in the population to be sequenced at the target coverage than other

343   methods previously reported.

344   Figure 2 shows the comparison of AlphaSeqOpt with IWS, the adapted IWS,

345   and Random when the target haplotype coverage was 10x. We tested different

346   scenarios in which the total available sequencing resources were £400,000, £800,000,

347   or £1,600,000. Figure 2a shows the percentage of haplotypes that would be sequenced

348   at (or above) the target coverage of 10x by sequencing the set of individuals selected

349   with AlphaSeqOpt. Figure 2b shows the number of individuals selected for

350   sequencing in each of the scenarios. AlphaSeqOpt delivered the highest percentage of

351   haplotypes sequenced at the target coverage, followed by the adapted IWS method,

352   which achieved a lower percentage even though it sequenced a number of individuals

353   similar to AlphaSeqOpt. The IWS method resulted in only a very small set of

354   individuals being sequenced and these individuals captured only a small percentage of

355   haplotypes sequenced at the target coverage. This result occurred because, as done by

356   Bickhart et al. [17], we only targeted haplotypes that appeared in a homozygous state

357   in at least one animal, which represent a small proportion of the haplotypes observed

358   in the population, and therefore the IWS method did not exhaust all the available

359   sequencing resources in any of the cases tested. The Random method sequenced a

360   very large set of individuals but it was inefficient for obtaining the haplotypes

361   sequenced at the target coverage.

362   The AlphaSeqOpt method was further tested to assess the effect of its main

363   features on the percentage of haplotypes sequenced at (or above) the target coverage,

364   the number of individuals sequenced, the performance under restriction of the

365   maximum coverage per individual, and the performance of the refinement step with

366   different number of exchanges per round.

367

368   **<u>Percentage of haplotypes sequenced at the target coverage:</u>**

369   The advantage provided by the AlphaSeqOpt score function and the step of

370   refinement over the adapted IWS method is shown in Figure 3. Figure 3a shows the

371   percentage of the haplotypes that would be sequenced at (or above) the target

372   coverage by sequencing the set of individuals selected with AlphaSeqOpt. We tested

373   nine scenarios in which the target coverage was 5x, 10x, or 15x and the total available

374   sequencing resources were £400,000, £800,000, or £1,600,000. Each scenario was

375   tested with either the AlphaSeqOpt score function or the IWS score function (adapted

376   IWS method), and both the initial and refined sets were examined.

377   The AlphaSeqOpt score function provided a greater percentage of haplotypes

378   sequenced at the target coverage than the IWS score function in all scenarios. The

379   AlphaSeqOpt score function gave 1.8 to 6.6% more haplotypes sequenced at the

380   target coverage than the IWS score function. The advantage of the AlphaSeqOpt score

381   function was observed both in the initial and refined sets. The refinement step

382   increased the percentage of haplotypes sequenced at the target coverage by 1.0 to

383   3.1% with the AlphaSeqOpt score function and 1.4% to 4.7% with the IWS score

384   function. In total, using the AlphaSeqOpt score function and a refinement step

385   delivered 6.6 to 9.3% more haplotypes sequenced at the target coverage than using the

386   IWS score function without a refinement step.

387   AlphaSeqOpt performed better because it was more efficient at allocating the

388   sequencing resources so that there were very few haplotypes that received some, but

389   insufficient, sequencing coverage.

390   Figure 4a shows the distribution of the population count of the haplotypes and

391   Figure 4b the distribution of the sequencing coverage that the haplotypes receive by

392   sequencing the set of individuals selected with each method. Note that the x-axis in

393   Figure 4b is half that of Figure 4a because for each x of coverage of an individual

394   there is a probability of 0.5 of reading either the paternal or maternal haplotype in the

395   diploid species that was simulated. Because the results for all scenarios were similar,

396   for illustration purposes from here onwards we only show results for the scenario in

397   which the target coverage was 10x and the sequencing resources were £800,000. Also

398   note that the haplotypes with population count ≤2 are shown in Figure 4a but were

399   excluded from the analyses shown in Figure 4b.

400   As a reference, choosing individuals randomly followed by random exchanges

401   of individuals followed the distribution of the population frequencies, with a large

402   percentage of haplotypes sequenced at coverages below the target 10x (54.0% of the

403   haplotypes had sequence coverage between 0.5x and 9.5x). The AlphaSeqOpt score

404   function reduced this percentage to only 6.3% in the initial set and 5.6% in the refined

405   set. This percentage was greater with the adapted IWS method than with

406   AlphaSeqOpt in both sets (17.3% in the initial set was reduced to 14.7% in the refined

407   set). The percentage of haplotypes that received no coverage at all in the refined set

408   were 19.2% for AlphaSeqOpt, 14.9% for the adapted IWS, and 6.3% for Random.

409

<u>Number of individuals sequenced:</u>

410

411   The initial sets that were selected by AlphaSeqOpt produced greater

412   percentages of haplotypes at the target coverage by sequencing less animals than the

413   sets selected by the adapter IWS method. The refinement step with the AlphaSeqOpt

414   score function produced sequencing sets that contained a larger number of unique

415   individuals than with the IWS score function. The extent to which the size of the

416   sequencing set was increased depended on the cost of library preparation and the

417   amount of sequencing resources available.

418   Figure 3b shows the number of individuals in the sets selected in each of the

419   scenarios explored in Figure 3a. The initial set was smaller with the AlphaSeqOpt

420   score function than with the IWS score function by between 122 and 340 individuals.

421   During the refinement step with the AlphaSeqOpt score function, the set maintained

422   approximately the same size when a small amount of sequencing resources was

423   available but increased by up to 457 individuals when more sequencing resources

424   were available. In contrast, during the refinement with the IWS score function, the

425   size of the sequencing set decreased when few sequencing resources were available

426   but remained more stable with a large amount of sequencing resources.

427   Figure 5 shows the effect of the cost of library preparation on the percentage

428   of haplotypes sequenced at (or above) the target coverage (Figure 5a) and the number

429   of unique individuals (Figure 5b) in the refined set produced with the AlphaSeqOpt

430   score function or the IWS score function. With both score functions, the percentage of

431   haplotypes sequenced at the target coverage increases linearly with decreasing library

432   costs. When library cost is low, the AlphaSeqOpt score function produces larger sets

433   with more unique individuals than the IWS score function, and these larger sets

434   produce greater percentages of haplotypes sequenced at the target coverage. When the

435   library costs are high, the difference between the sizes of the sets obtained with the

436   two score functions is reduced. Figure 5c shows the distribution of the sequencing

437   coverage across sequenced individuals in the refined set produced with the

438   AlphaSeqOpt score function considering two extreme library costs. Low library costs

439   allowed for the sequencing of more individuals at low coverage while high library

440   costs resulted in a greater number of individuals being sequenced at twice the target

441   coverage of the haplotypes. With a library cost of £5 the number of individuals

442   sequenced was 1307.6 (302.2 at 1x to 124.3 at 20x) and with a library cost of £40 it

443   decreased to 1036.4 (136.6 at 1x to 176.7 at 20x).

444

445   **Restriction of individual coverage:**

446   The size of the sequencing set can be maximised by restricting the maximum

447   coverage that each individual can get, so that the target coverage of the haplotypes is

448   achieved by accumulating individuals sequenced only at or below a certain coverage.

449   Figure 6 shows the comparison of AlphaSeqOpt with the adapted IWS when the

450   maximum individual coverage is restricted to 1x. Under this restriction, only

451   haplotypes with a population count ≥10, ≥20, and ≥30 can reach the target coverages

452   of 5x, 10x, and 15x, respectively, and therefore haplotypes with lower population

453   counts were excluded from the analyses. Figure 6a shows the percentage of targeted

454   haplotypes that would be sequenced at (or above) the three levels of target coverage

455   by sequencing the set of individuals selected with AlphaSeqOpt. Figure 6b shows the

456   number of individuals selected for sequencing in each of the scenarios.

457   With a budget of £400,000 a total of 3,333 individuals could be sequenced at

458   1x. Under this setting, AlphaSeqOpt delivered greater percentages of haplotypes

459   sequenced at the target coverage than the adapted IWS method. If the budget was

460   unrestricted, IWS selected a smaller set than AlphaSeqOpt to sequence all the targeted

461   haplotypes at the desired coverage.

462

463   **Effect of the number of exchanges per round during refinement:**

464   For the refinement of the set, there was an optimum number of exchanges per

465   round that maximized the percentage of haplotypes sequenced at the target coverage

466   given a fixed total number of exchanges. Figure 7a shows the percentage of

467   haplotypes sequenced at (or above) the target coverage with a fixed number of total

468   exchanges but with different numbers of rounds and exchanges per round, considering

469   two extreme costs of library preparation. Figure 7b shows the size of the resultant set.

470   Doing 1 to 100 exchanges per round improved the percentage of haplotypes

471   sequenced at the target coverage of the refined set to similar values. In this case, the

472   set that produced the maximum percentage of haplotypes sequenced at the target

473   coverage was obtained by doing 10 exchanges per round. Even though this greater

474   percentage was generally achieved by increasing the number of unique sequenced

475   individuals, the size of the refined set slightly decreased when the library cost was

476   high and few exchanges per round were made. Doing more than 500 exchanges per

477   round did not improve the results of the initial set when library cost was £40 and

478   made the algorithm less robust when library cost was £5. However, the most extreme

479   scenario of exchanging the whole set, which is equivalent to selecting a new initial set

480   without any refinement in each round, provided the best improvement of the initial

481   percentage of haplotypes sequenced at the target coverage and the greatest reduction

482   of the sequencing set.

483

Filtering of rare haplotypes based on flanking context

484   As the target haplotype coverage increases, the least frequent haplotypes can

485   be sequenced at the target coverage only if a large amount of resources is available.

486   More sequencing resources can be focused on sequencing common haplotypes if the

487   number of rare haplotypes included in the analyses is reduced either by excluding

488   them all or by filtering them based on their flanking context.

489   Figure 8 shows the distribution of the sequencing resources depending on the

490   population count of the haplotypes with the three different approaches to deal with the

491   rare haplotypes: to include all singletons and doubletons in the analysis, to exclude

492   them, or to filter them based on their flanking context. Almost half of the haplotypes

493   in the test population were observed only once (singletons; 31.7%) or twice

494   (doubletons; 13.2%), making a total of 3,971 singletons and doubletons. Of these, 953

495   (19%) remained after filtering based on their flanking context and these were

496   considered as likely to have derived from a recombination event of two common

497   haplotypes. We only show results for the scenarios in which the target haplotype

498   coverage was 10x, with the total available sequencing resources being £400,000,

499   £800,000, or £1,600,000.

500   With £800,000, when all singletons and doubletons were included in the

501   analyses 72.4% of the haplotypes with population count ≥ 3 were sequenced at (or

502   above) 10x. This percentage increased to 75.3% when all singletons and doubletons

503   were excluded. This percentage also increased, but a little bit less, when they were

504   filtered based on their flanking context (74.8%). A similar trend was observed with

505   £400,000 and £1,600,000.

506   When we have a large amount of sequencing resources we may be interested

507   in targeting rare haplotypes as well as common haplotypes. By filtering based on their

508   flanking context we can target the rare haplotypes that are likely to derive from a

509   recombination of common haplotypes. With £1,600,000, a total of 38.6% of the 953

510   target singletons and doubletons were sequenced at 10x. Only 33.0% of these 953 was

511   sequenced at 10x when all singletons and doubletons were included in the analyses

512   without any restriction. This benefit of filtering by flanking context was not observed

513   when less sequencing resources were available, probably because in such scenarios

514   sequencing resources were implicitly focused on the common haplotypes.

515

Discussion

516   We have presented a method that identifies which individuals need to be

517   sequenced and at what coverage they should be sequenced when a given amount of

518   sequencing resources are available so that the maximum percentage of the haplotypes

519   present in the population are sequenced at (or above) a coverage that is sufficiently

520   high to ensure that the consensus haplotypes can be accurately derived. Deriving the

521   consensus haplotypes accurately is a critical requirement for achieving high

522   population-based imputation accuracy under the LCSeq strategy and we have

523   observed with a prototype of a novel population-based phasing and imputation

524   method that there is a relationship between the coverage that a particular haplotype

525   accumulates across individuals and the imputation accuracy of the consensus

526   haplotype (Additional file 1: Figure S1). We also developed and tested a new

527   approach to deal with rare haplotypes by filtering them based on their flanking

528   context rather than excluding them from the analysis. We compared AlphaSeqOpt

529   with previously published methods and hereafter discuss the advantages and

530   limitations of AlphaSeqOpt.

531

Advantages of AlphaSeqOpt over other methods

532   AlphaSeqOpt has two features that make it effective for its purpose: (1) a

533   score function based on the frequency of the haplotypes in the sequencing set relative

534   to the target coverage instead of on the population frequency of the haplotypes; and

535   (2) a step of refinement of the initial set.

536

537   **<u>Score function:</u>**

538   The score function that we propose allocates sequencing resources such that

539   the percentage of haplotypes sequenced at any target coverage is greater than with

540   other score functions based on the population frequency of the haplotype. The score

541   function based on the population frequency of the haplotype used in the IWS method

542   [17] was designed for producing the least redundant set that should be sequenced to

543   have all the targeted haplotypes sequenced. The reduction of redundancy with the

544   IWS method is achieved by giving a greater score to the least frequent haplotypes

545   and, therefore, selecting the individuals that carry less frequent haplotypes first.

546   Therefore, if the sequencing resources are sufficient for sequencing all the targeted

547   haplotypes, the IWS method does so by sequencing a smaller set than AlphaSeqOpt.

548   However, the IWS method is not ideal for identifying the set of individuals that would

549   provide a more even sequencing coverage of the largest percentage of population

550   haplotypes when the sequencing resources are limited and insufficient for sequencing

551   all the targeted haplotypes at the desired coverage. With a score function that uses the

552   population frequency the haplotype scores are constant until these haplotypes reach

553   the target coverage, at which point they are set to zero. A score function based on the

554   frequency of the haplotypes in the sequencing set relative to the target coverage like

555   the one used in AlphaSeqOpt performs better for this purpose because, in contrast, the

556   haplotype scores change as the sequencing resources are allocated. With the

557   AlphaSeqOpt score function all haplotypes start with an equal score of 1 and their

558   score increases exponentially as they approach the target coverage.

559   By doing this, the AlphaSeqOpt score function prioritizes the haplotypes that

560   are already closer to the target coverage and, implicitly, the individuals that carry a

561   larger number of these haplotypes. This reduces the percentage of haplotypes that are

562   sequenced at a suboptimal coverage, but it increases the percentage of haplotypes that

563   receive no coverage at all. With limited sequencing resources, AlphaSeqOpt selects a

564   set for sequencing with a larger percentage of population haplotypes at the target

565   coverage than the IWS method. These sequencing sets can be even smaller than the

566   ones produced with IWS score function if the initial set is not refined.

567

568   **<u>Refinement of the initial set:</u>**

569   The other main feature of AlphaSeqOpt is the step of refinement of the initial

570   set. The step of refinement adjusts the allocation of resources by replacing individuals

571   that have become redundant after the last additions to the set or by reducing the

572   sequencing coverage of these individuals. A refinement step as described here further

573   increases the percentage of haplotypes sequenced at the target coverage obtained with

574   the AlphaSeqOpt score function. A side benefit in the context of LCSeq is that the

575   refinement step achieves this increase by diversifying the set of individuals that are

576   sequenced. While the IWS score function restrains the number of sequenced

577   individuals, the AlphaSeqOpt score function benefits from low library costs relative to

578   the total amount of sequencing resources available to produce larger sets with more

579   unique individuals that are sequenced at lower coverage. This benefit is greater when

580   the cost of library preparation represents a small fraction of the total amount of

581   sequencing resources for LCSeq. Methods for reducing significantly the costs of

582   library preparation for high-throughput LCSeq have already been described [27].

583   Increasing the number of individuals sequenced would empower subsequent

584   imputation for more individuals (i.e., these individuals and their relatives) as well as

585   any downstream analyses [12].

586   The refinement step can be fine-tuned by adjusting parameters such as the

587   number of exchange rounds and the number of exchanges per round. The optimal

588   parameters may depend largely on the size and structure of each dataset, but the

589   following general observations were made:

590   - AlphaSeqOpt was very robust across repetitions. A stable solution was

591   produced after a relatively low number of exchange rounds (unpublished results).

592   Small further increases of the percentage of haplotypes sequenced at the target

593   coverage could be obtained by using a longer chain of exchange rounds, but the

594   benefit of this was little.

595   - To some extent, increasing the number of exchanges per round enables

596   greater mobility across possible sets. Consequently, the algorithm can retrieve a better

597   solution more easily. However, when too many exchanges are made per round, the

598   benefit of this refinement of the existing set is diluted due to the drift towards

599   solutions that are too divergent from each other and thus the final solution becomes

600   less reliable. Exchanging all the individuals in the set is an extreme case of this that is

601   equivalent to choosing the best of multiple initial sets without refinement. It can

602   produce good results in terms of percentage of haplotypes for small sequencing sets.

603

604   **Practical implications for real populations:**

605   Provided that the cost of library preparation is low enough or by restricting the

606   maximum coverage of the individuals, AlphaSeqOpt will produce large sets of

607   individuals with many unique individuals that are sequenced at low coverage.

608   The performance of AlphaSeqOpt will likely be influenced by structure of the

609   data, either intrinsic, like the number and size of the chromosomes in a species or the

610   degree of relatedness between individuals, or extrinsic, like the core length used to

611   define the haplotypes. AlphaSeqOpt assumes that coverage is uniform along the

612   genome but variation in coverage at the level of nucleobase should be expected, as

613   well as variation of coverage between samples.

614   Although the criterion that is maximised in AlphaSeqOpt is the percentage of

615   unique haplotypes sequenced at (or above) the target coverage, the method also

616   provides good coverage in terms of total population haplotypes, i.e., haplotypes

617   weighted by their population frequencies. Implicitly, the scores of more frequent

618   haplotypes will increase faster than the scores of less frequent haplotypes because

619   they are more likely to be carried by the individuals that are added to the sequencing

620   set. In all scenarios tested, both AlphaSeqOpt and the adapted IWS method provided

621   total percentages of haplotypes >99%, but the percentage was consistently greater for

622   AlphaSeqOpt. Although both methods were similarly successful in covering the

623   haplotypes of most of the population, AlphaSeqOpt captured a greater diversity of

624   haplotypes at the desired coverage.

625   The resolution of the haplotype library will depend on the density of the

626   marker array used to construct it. However after sequencing the individuals it is

627   possible that haplotypes that were considered to form a single consensus haplotype

628   when defined with marker data actually correspond to a number of true haplotypes. In

629   such cases the sequence data can be clustered into the multiple consensus haplotypes

630   and the pedigree information could enhance their imputation.

631

Utility of filtering rare haplotypes based on flanking context

632   We proposed an approach that uses the haplotype population frequencies at the

633   cores flanking a particular core to identify those rare haplotypes that could have

634   derived from a recombination event. Although rare haplotypes may contain relevant

635   biological information, we may not be able to impute and estimate accurately the

636   effect of most rare haplotypes. The rationale behind the filtering approach that we

637   propose is that sequence data of those rare haplotypes that are potentially mosaic of

638   common haplotypes could enable a more precise capture of the recombination events

639   that have occurred in the population and that this sequence data would also contribute

640   to the consensus haplotypes of the haplotypes that gave rise to the mosaic. The new

641   approach that we propose, although not ideal, may be of a particular interest in cases

642   in which large amounts of sequencing resources are available.

643   In real populations we expect to identify large numbers of rare haplotypes.

644   Preliminary tests indicated that in real populations our filtering approach based on

645   flanking context can filter out around 92% of the singletons and doubletons observed,

646   with the other 8% retained as potentially mosaic (unpublished results).

647   The challenge of targeting rare mosaic haplotypes is that the individuals that

648   carry them must be sequenced at a greater coverage so that the rare haplotypes reach

649   the target coverage. Another approach, for which we do not show results here,

650   involves setting a lower secondary target coverage for less frequent haplotypes. This

651   is a compromise solution where reducing the sequencing coverage of the rare

652   haplotypes will reduce their imputation accuracy but will allow more rare haplotypes

653   to be sequenced. In the particular case of potentially mosaic rare haplotypes having

654   less coverage would be less critical because the information of the common

655   haplotypes from which they derive will be also available. Any of the approaches

656   discussed to filter low-frequency haplotypes can be combined using multiple

657   frequency thresholds.

658

Suitability of AlphaSeqOpt for low-coverage sequencing designs

659   A number of optimisation methods that use haplotypes derived from existing

660   phased marker array genotypes have already been proposed to identify which

661   individuals should be sequenced under the KeySires approach. Druet et al. [5]

662   proposed a method that maximizes the proportion of haplotypes observed in the

663   population that are sequenced. This method was more effective in detecting rare

664   variants (minor allele frequency <5%) than methods based solely on pedigree

665   information and it provided good imputation accuracies for both common and rare

666   variants. Bickhart et al. [17] proposed the IWS method, which reduces the number of

667   individuals that need to be sequenced in order to have all haplotypes above a certain

668   population frequency sequenced. The method by Gusev et al. [16] selects the

669   individuals that share the largest proportion of the population haplotypes with other

670   individuals identical-by-descent (IBD). More recently, Neuditschko et al. [18]

671   proposed a method based on the eigenvalue decomposition of a genomic relationship

672   matrix that identifies individuals that maximise genetic diversity within complex

673   population structures. These four methods identify which individuals should be

674   sequenced but do not make any decision on the coverage at which they should be

675   sequenced. Gonen et al. [14] proposed an approach that distributes sequence at

676   variable coverage across individuals in a population. This method accounts for

677   haplotype frequency and the ability to phase the resulting sequence data as criteria

678   when choosing individuals to sequence and assigning sequencing coverage to those

679   individuals and their recent ancestors. We have presented a method for optimizing the

680   allocation of sequencing resources for the LCSeq approach so that the imputation

681   accuracy of consensus haplotypes is high enough for enabling novel population-based

682   imputation methods.

683   In practice, it is likely that a combination of the three sequencing approaches

684   discussed in this paper (KeySires, LCSeq, and VarCoverage) would yield similar or

685   even better imputation accuracies than LCSeq alone [28]. AlphaSeqOpt is flexible in

686   that it can take into account the already available sequence information. Therefore,

687   AlphaSeqOpt for optimizing LCSeq can be used either alone or complementarily to

688   other existing methods to top-up the coverage of those haplotypes that are under-

689   sequenced after using any other method. In this later case, however, existing methods

690   or newly developed ones should be integrated to find the right allocation of resources

691   into each of the three sequencing approaches.

692

Conclusion

693   We have presented a method for optimizing the allocation of sequencing

694   resources so that the maximum proportion of population haplotypes are sequenced at

695   a coverage that is sufficiently high for population-based imputation with low-

696   coverage sequencing. The haplotype score function and the refinement step make

697   AlphaSeqOpt more effective for this purpose than methods reported previously for

698   reducing sequencing redundancy. AlphaSeqOpt can account for sequence information

699   already available for the population, which makes it a good complementary method to

700   increase coverage of haplotypes that are not sufficiently covered when other

701   optimisation methods are used. We also explored a new approach to deal with rare

702   haplotypes by targeting only those that are likely derived by recombination of

703   common haplotypes. This approach frees resources to sequence greater proportion of

704   distinct haplotypes, which can be useful particularly when large amounts of

705   sequencing resources are available.

706

707

Competing interests

708   The authors declare that they have no competing interests.

709

Authors' contributions

710   JMH and RRF designed the algorithm and the study; RRF performed the analyses;

711   RRF wrote the first draft; GG and SG assisted in the interpretation of the results and

712   provided comments on the manuscript. All authors read and approved the final

713   manuscript.

714

Acknowledgements

722

References

723 1. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al.
724 Extremely low-coverage sequencing and imputation increases power for genome-
725 wide association studies. Nat. Genet. 2012;44:631–5.

726 2. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF,
727 et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and
728 complex traits in cattle. Nat Genet. 2014;46:858–65.

729 3. Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, et al. Genome-
730 wide association of multiple complex traits in outbred mice by ultra-low-coverage
731 sequencing. Nat. Genet. 2016;48:912–8.

732 4. Meuwissen T, Goddard M. Accurate Prediction of Genetic Values for Complex
733 Traits by Whole-Genome Resequencing. Genetics. 2010;185:623–31.

734 5. Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome
735 sequence data: impact of sequencing design on genotype imputation and accuracy of
736 predictions. Heredity. 2014;112:39–47.

737 6. Hickey JM. Sequencing millions of animals for genomic selection 2.0. J. Anim.
738 Breed. Genet. 2013;130:331–2.

739 7. Hickey J, Kinghorn BP, Cleveland MA, Tier B, van der Werf JHJ. Recursive long
740 range phasing and long haplotype library imputation: Building a global haplotype
741 library for Holstein cattle. Proc. 9th World Congr. Genet. Appl. Livest. Prod.
742 WCGALP. Leipzig, Germany; 2010. p. 0934.

743 8. Brøndum R, Guldbrandtsen B, Sahana G, Lund M, Su G. Strategies for imputation
744 to whole genome sequence using a single or multi-breed reference population in
745 cattle. BMC Genomics. 2014;15:728.

746 9. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et
747 al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle.
748 Genet. Sel. Evol. 2014;46:41.

749 10. VanRaden PM, Sun C, O'Connell JR. Fast imputation using medium or low-
750 coverage sequence data. BMC Genet. 2015;16:82.

751 11. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence
752 without reference panels. Nat. Genet. 2016;48:965–9.

753 12. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing:
754 Implications for design of complex trait association studies. Genome Res.
755 2011;21:940–51.

756 13. Hickey JM, Gorjanc G, Cleveland MA, Kranis A, Jenko J, Mésázros G, et al.
757 Sequencing Millions of Animals for Genomic Selection 2.0. Proc. 10th World Congr.
758 Genet. Appl. Livest. Prod. WCGALP. Vancouver, BC, Canada; 2014. p. 377.

759   14. Gonen S, Ros-Freixedes R, Battagin M, Gorjanc G, Hickey JM. A method for
760   allocation of sequencing resources in genotyped livestock populations. Genet. Sel.
761   Evol. Submitted;

762   15. Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing
763   data on multiple diploid samples. Genome Res. 2011;21:952–60.

764   16. Gusev A, Shah MJ, Kenny EE, Ramachandran A, Lowe JK, Salit J, et al. Low-
765   Pass Genome-Wide Sequencing and Variant Inference Using Identity-by-Descent in
766   an Isolated Human Population. Genetics. 2012;190:679–89.

767   17. Bickhart DM, Hutchison JL, Null DJ, VanRaden PM, Cole JB. Reducing animal
768   sequencing redundancy by preferentially selecting animals with low-frequency
769   haplotypes. J. Dairy Sci. 2016;99:5526–34.

770   18. Neuditschko M, Raadsma HW, Khatkar MS, Jonas E, Steinig EJ, Flury C, et al.
771   Identification of key contributors in complex population structures. Chaubey G,
772   editor. PLOS ONE. 2017;12:e0177638.

773   19. Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N. Efficient study design
774   for next generation sequencing. Genet. Epidemiol. 2011;35:269–77.

775   20. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence
776   data. Genome Res. 2009;19:136–42.

777   21. Hickey JM, Gorjanc G. Simulated data for genomic selection and genome-wide
778   association studies using a combination of coalescent and gene drop methods. G3.
779   2012;2:425–7.

780   22. Faux A-M, Gorjanc G, Gaynor RC, Battagin M, Edwards SM, Wilson DL, et al.
781   AlphaSim: Software for Breeding Program Simulation. Plant Genome. 2016;9.

782   23. Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Tassell CPV, Grefenstette JJ.
783   High-resolution haplotype block structure in the cattle genome. BMC Genet.
784   2009;10:19.

785   24. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, Werf JH van der. A
786   combined long-range phasing and long haplotype imputation method to impute phase
787   for SNP genotypes. Genet. Sel. Evol. 2011;43:12.

788   25. Hickey JM, Kranis A. Extending long-range phasing and haplotype library
789   imputation methods to impute genotypes on sex chromosomes. Genet. Sel. Evol.
790   2013;45:10.

791   26. Hickey JM, Gorjanc G, Varshney RK, Nettelblad C. Imputation of Single
792   Nucleotide Polymorphism Genotypes in Biparental, Backcross, and Topcross
793   Populations with a Hidden Markov Model. Crop Sci. 2015;55:1934–46.

794   27. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing libraries
795   for multiplexed target capture. Genome Res. 2012;22:939–46.

796   28. Xu C, Wu K, Zhang J-G, Shen H, Deng H-W. Low-, high-coverage, and two-
797   stage DNA sequencing in the design of the genetic association study. Genet.
798   Epidemiol. 2017;41:187–97.
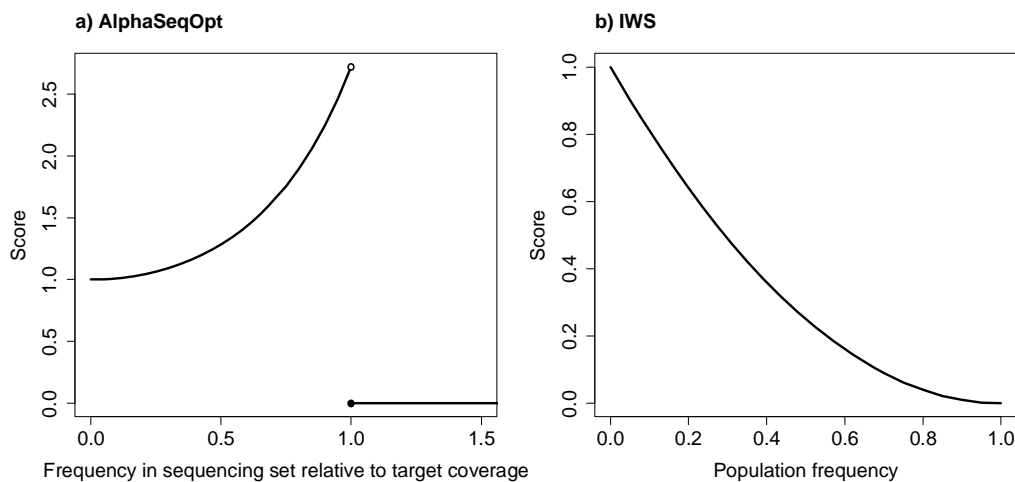
799

Figures

800

801

Figure 1. Score functions: (a) in AlphaSeqOpt; and (b) in the IWS method proposed by Bickhart et al. [17]. Note the different axes: in (a) scores range from 1 to e^k based on the frequency of the haplotype in the sequencing set relative to the target coverage, which is variable across rounds; in (b), scores range from 0 to 1 based on the population frequency, which is fixed.

807
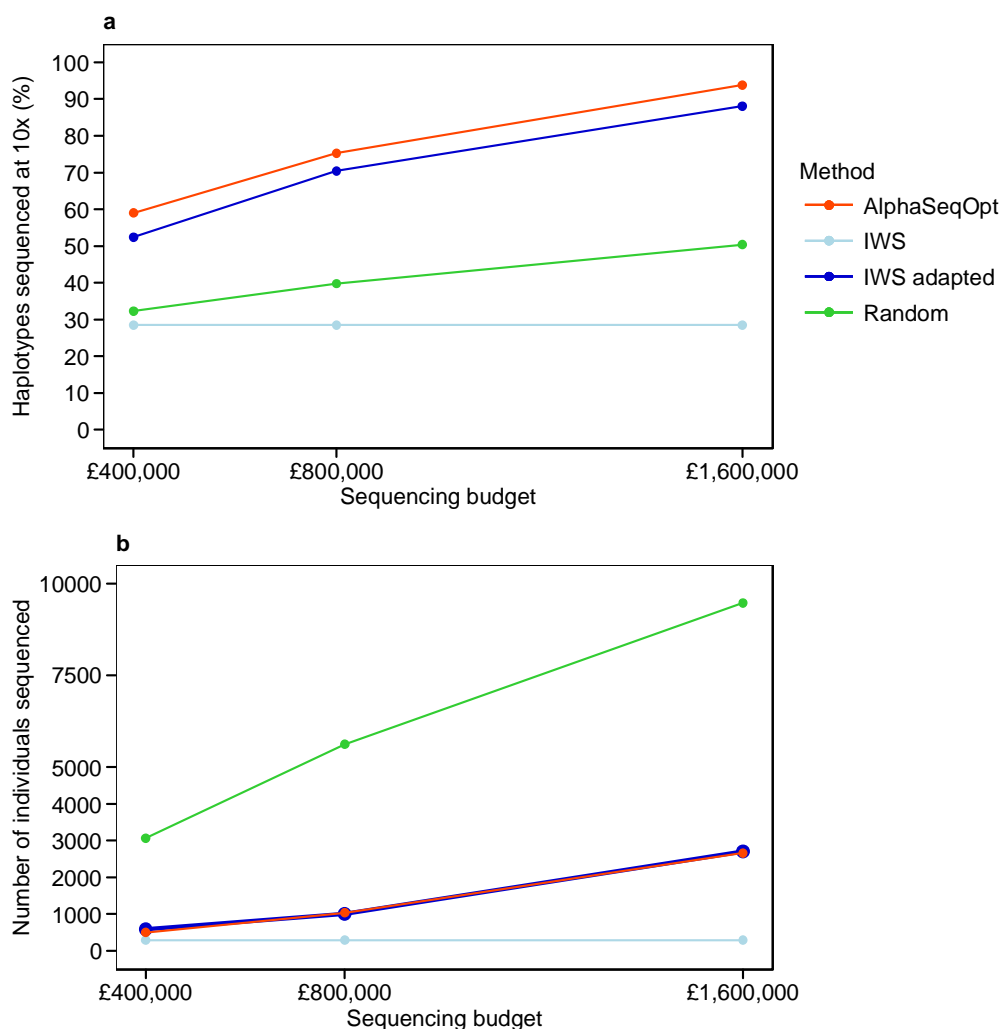
808
809 **Figure 2.** Performance of the four methods tested with different amounts of
810 sequencing resources, in terms of: (a) the percentage of population haplotypes
811 sequenced at (or above) the target coverage of 10x; and (b) the number of individuals
812 sequenced. Standard errors were less than 0.2% (a) and 25 (b).
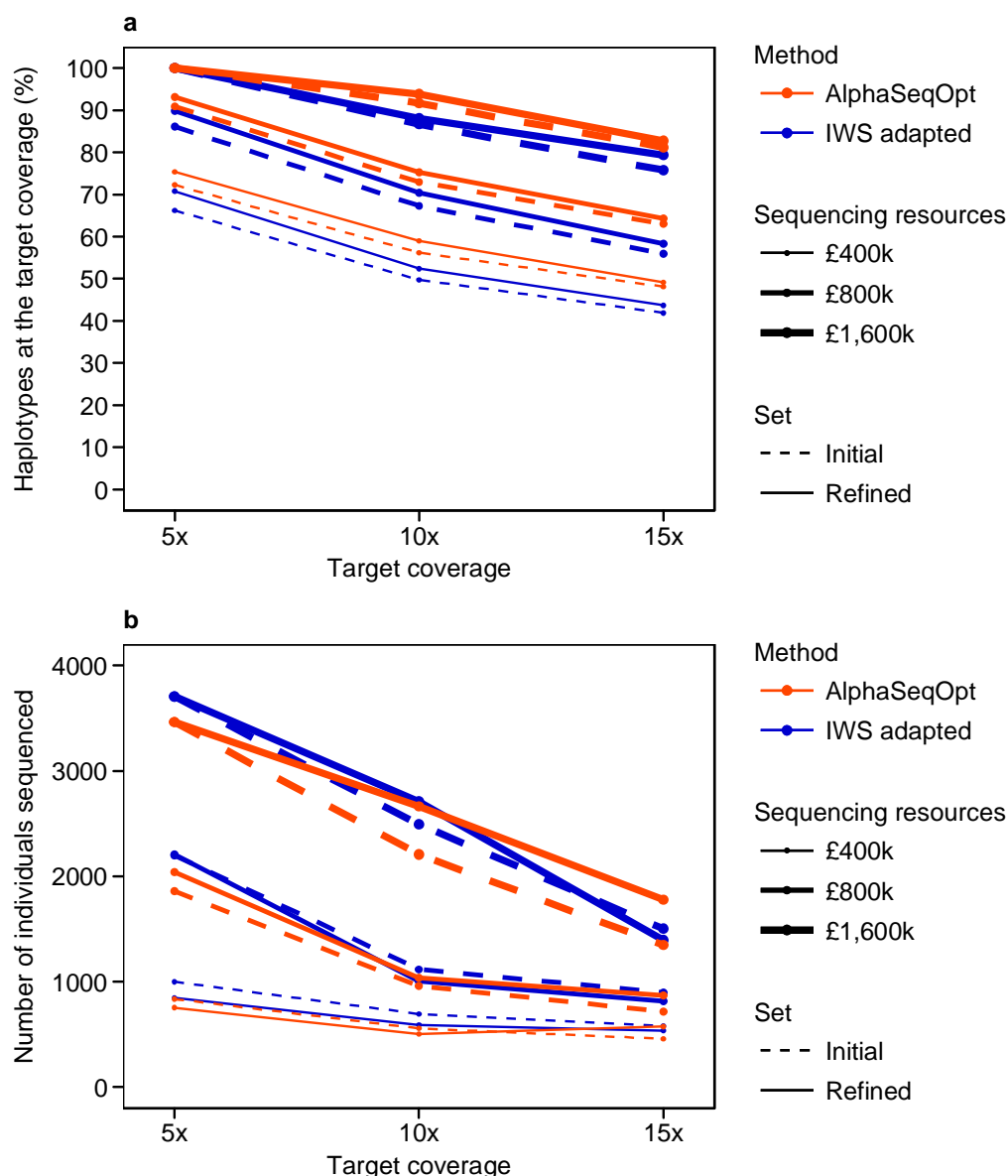
Figure 3. Performance of AlphaSeqOpt and the adapted IWS method with three levels of target haplotype coverage and different amounts of sequencing resources, in terms of: (a) the percentage of population haplotypes sequenced at (or above) the target coverage of 10x; and (b) the number of individuals sequenced. Both the initial and refined sets were examined. Standard errors were less than 0.4% (a) and 40 (b).

a) Population

b) Refined set
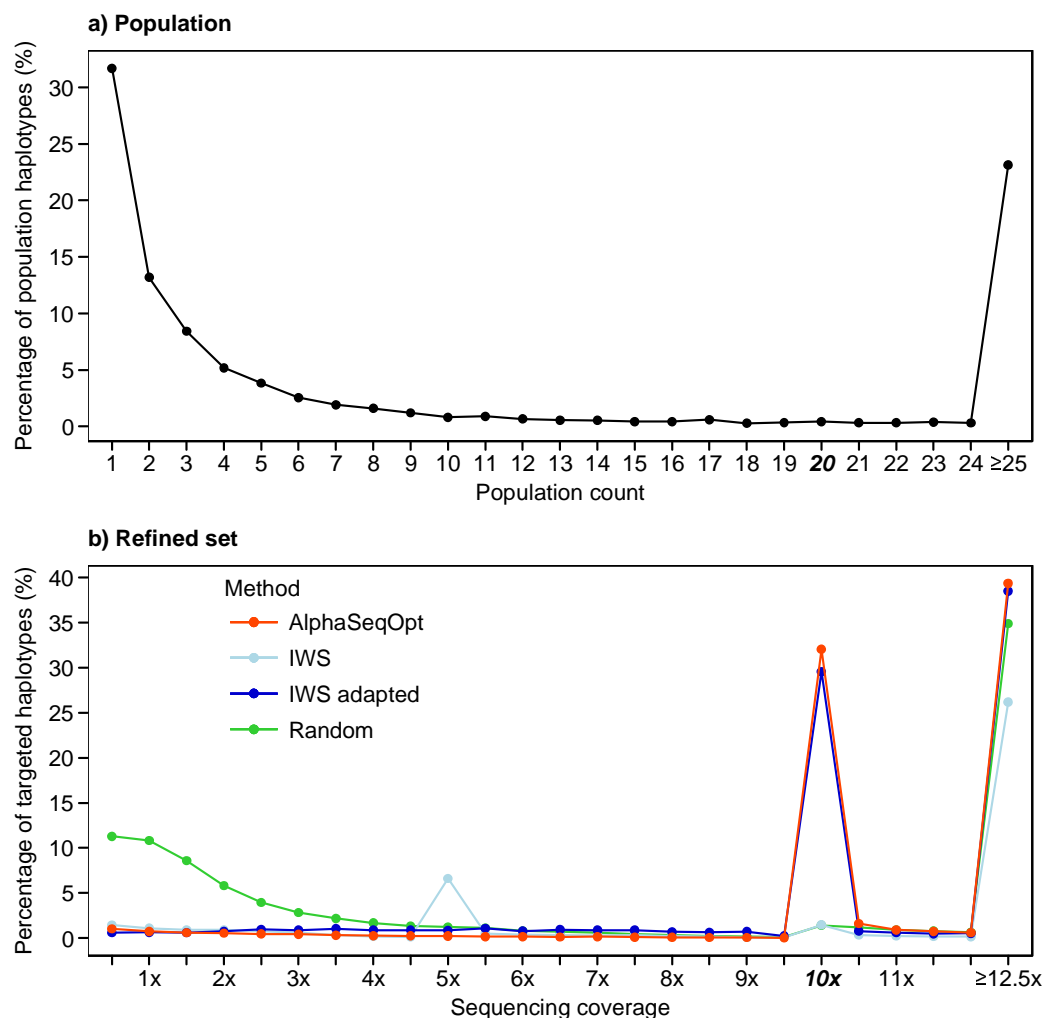
820
Figure 4. Distribution of: (a) the population count of the haplotypes; and (b) the sequencing coverage of the haplotypes using the four methods tested. The target haplotype coverage was 10x, the sequencing resources were set to £800,000 and haplotypes with population count ≤2 were excluded from the analyses. Standard errors were less than 14 (b).
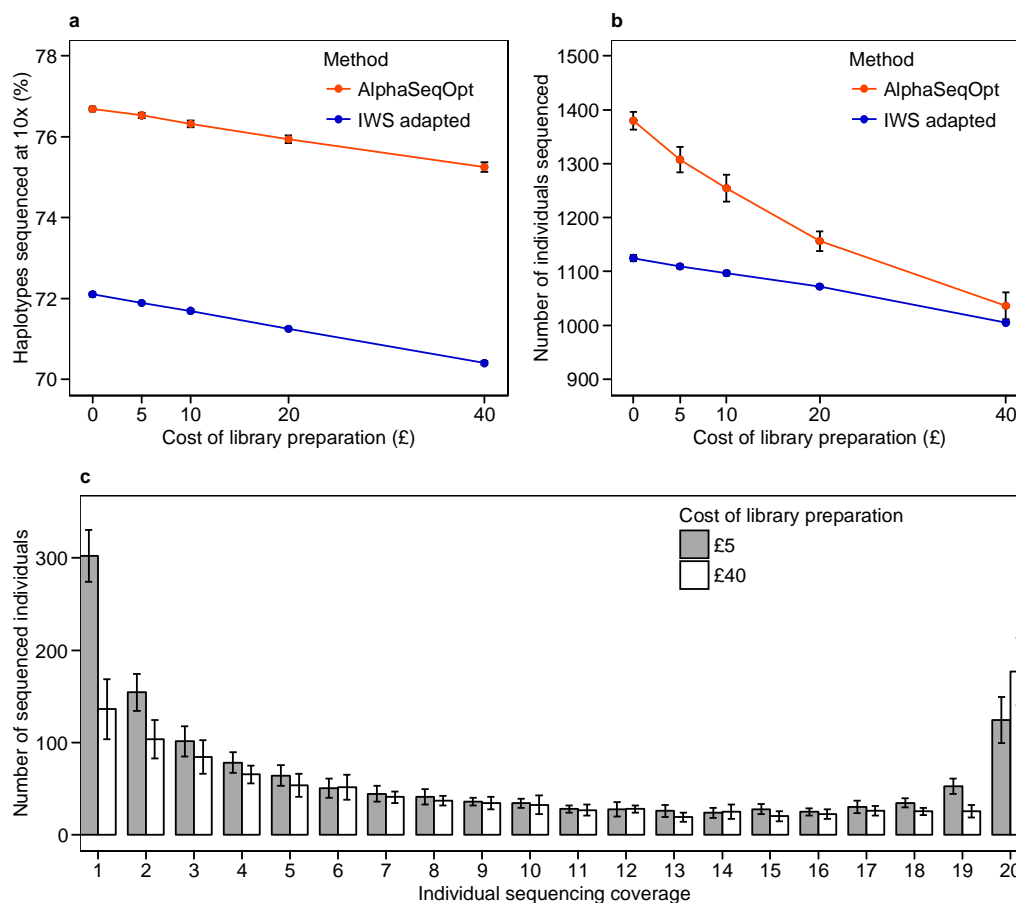
Figure 5. Effect of the cost of library preparation: (a) on the percentage of haplotypes sequenced at (or above) the target coverage of 10x; (b) on the number of individuals sequenced, using AlphaSeqOpt or the adapted IWS method; and (c) on the distribution of sequencing coverage of the individuals selected with AlphaSeqOpt. The sequencing resources were set to £800,000 and haplotypes with population count ≤2 were excluded from the analyses.
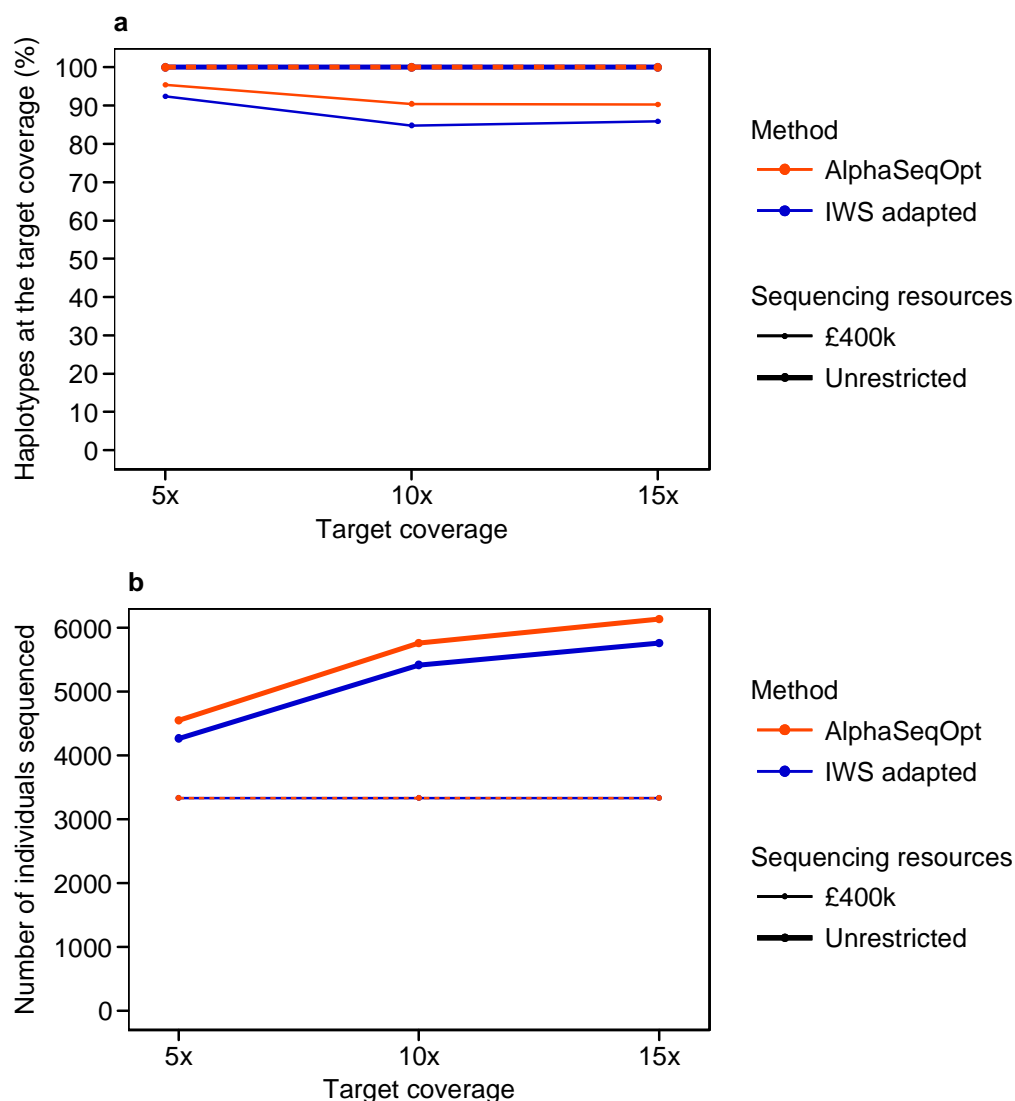
Figure 6. Performance of AlphaSeqOpt and the adapted IWS method when individual coverage is restricted to 1x, in terms of: (a) the percentage of population haplotypes sequenced at (or above) the target coverage; and (b) the number of individuals sequenced. When the target coverage was 5x, 10x, or 15x only haplotypes with population count ≥ 10, ≥ 20, or ≥ 30 were targeted, respectively.
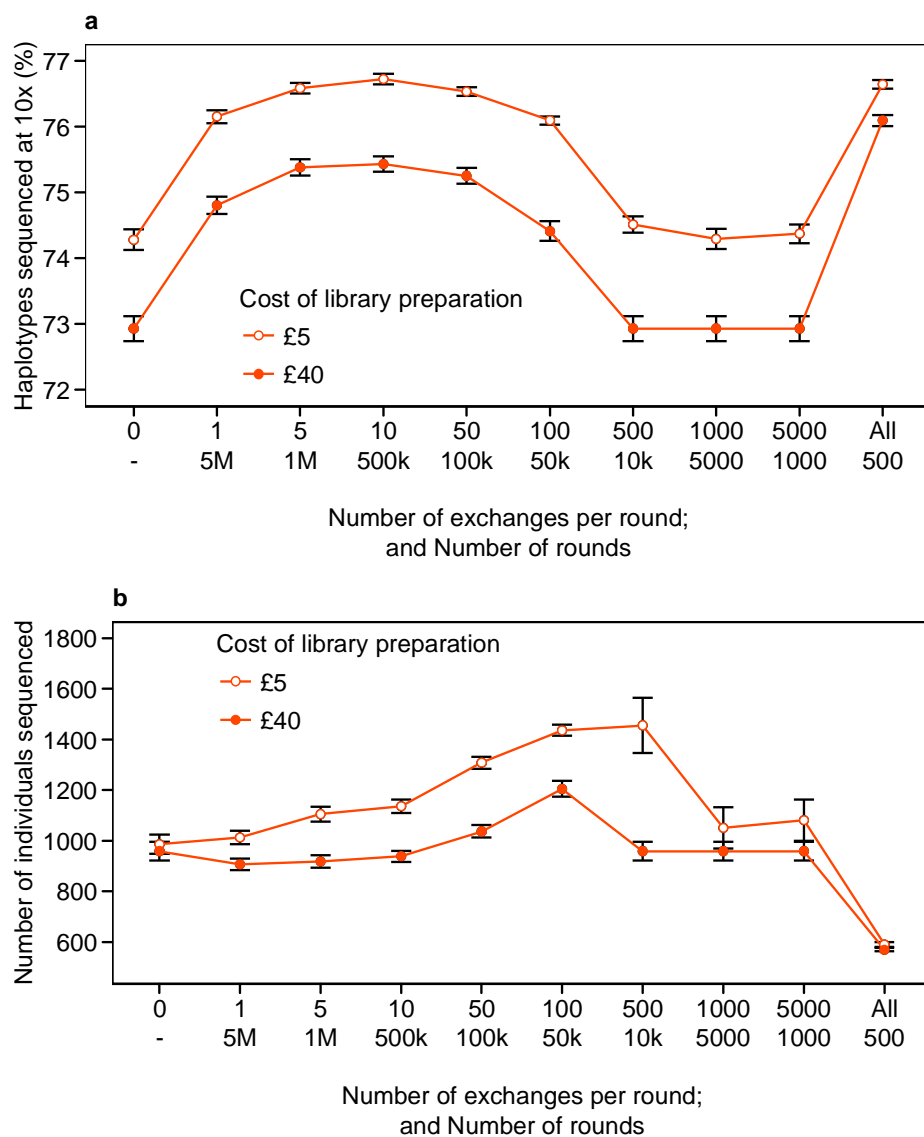
840
841 **Figure 7.** Effect of the number of exchanges per round: (a) on the percentage of
842 haplotypes sequenced at (or above) the target coverage of 10x; and (b) on the number
843 of individuals sequenced, with two costs of library preparation. The total number of
844 exchanges is 5 millions in all cases. The sequencing resources were set to £800,000
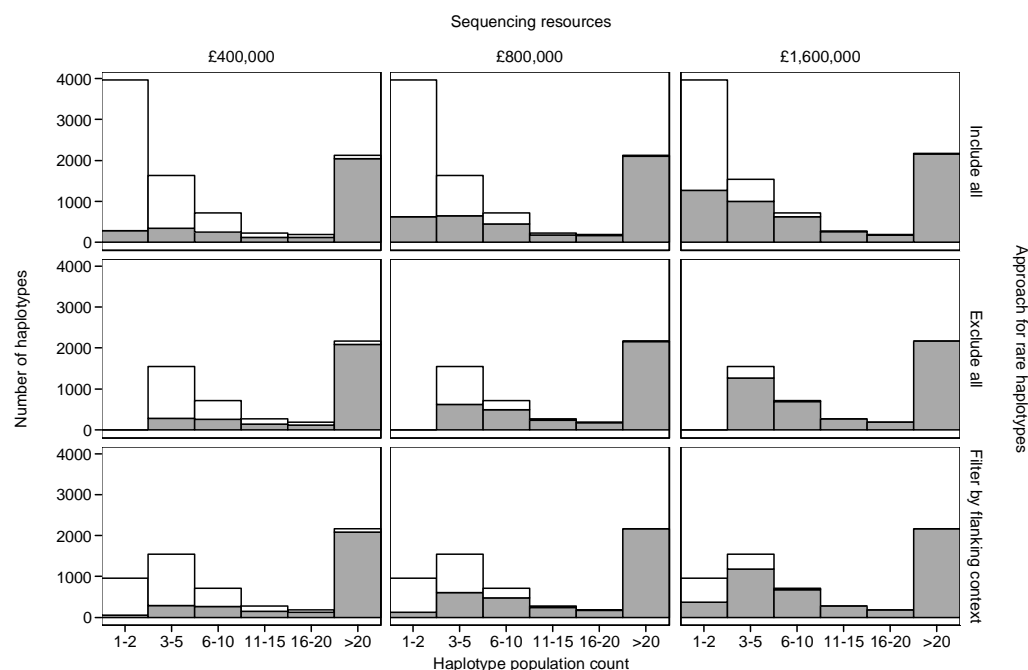845 and haplotypes with population count ≤2 were excluded from the analyses.
846

847
848 **Figure 8.** Number of haplotypes sequenced at (or above) the target coverage of 10x
849 (filled section) using three different approaches to handle the rare haplotypes: to
850 include all singletons and doubletons, to exclude all singletons and doubletons, or to
851 filter them based on flanking context. Numbers are shown by haplotype population
852 count and for different amounts of sequencing resources. The number of singletons
853 and doubletons for each approach were 3,971, 0, and 953, respectively.

Additional files

Additional file 1: Figure S1

854   Format: pdf

855   Title: Expected haplotype imputation accuracy against the accumulated haplotype
856   sequencing coverage, as estimated using a novel population-based imputation method
857   (Battagin and Hickey, unpublished).

858   Description: A description of the prototype algorithm developed for the imputation of
859   consensus haplotypes under the LCSeq approach and the simulated results on which
860   the AlphaSeqOpt method is based. We generated 1x sequence data for the sires from a
861   simulated population. The x-axis represents the expected accumulated coverage that
862   each haplotype would receive. The y-axis represents the percentage of alleles phased
863   and imputed for each haplotype. The imputation accuracy increased with the
864   accumulated haplotype coverage until it platooned. Haplotypes with a sequencing
865   coverage of 10x accumulated from 20 individuals sequenced at 1x were imputed to
866   the whole population with an accuracy of 0.88. Haplotypes with a sequencing
867   coverage of 15x or 20x accumulated from 30 or 40 individuals sequenced at 1x were
868   imputed to the whole population with an accuracy of 0.93 or 0.97, respectively. For
869   accurate inference of a consensus haplotype a certain amount of sequencing coverage
870   must be accumulated. According to the results above, 10x or 15x could be good target
871   coverages for the haplotypes.