

1 **Extending long-range phasing and haplotype library** 2 **imputation algorithms to very large and** 3 **heterogeneous datasets**

4 Daniel Money¹, David Wilson¹, Janez Jenko¹, Gregor Gorjanc¹, & John M. Hickey^{1§}

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

7 [§]Corresponding author

8 Email addresses:

9 DM: daniel.money@roslin.ed.ac.uk

10 DW: david.wilson@roslin.ed.ac.uk

11 JJ: janez.jenko@roslin.ed.ac.uk

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

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

14 **Abstract**

15 **Background**

16 This paper describes the latest improvements to the long-range phasing and
17 haplotype library imputation algorithms that enable them to successfully phase both
18 datasets with one million individuals and datasets genotyped using different sets of
19 single nucleotide polymorphisms (SNPs). Previous publicly available
20 implementations of long-range phasing could not phase large datasets due to the
21 computational cost of defining surrogate parents by exhaustive all-against-all
22 searches. Further, both long-range phasing and haplotype library imputation were not
23 designed to deal with large amounts of missing data, which is inherent when using
24 multiple SNP arrays.

25 **Methods**

26 Here, we developed methods which avoid the need for all-against-all searches
27 by performing long-range phasing on subsets of individuals and then combining results.
28 We also extended long-range phasing and haplotype library imputation algorithms to
29 enable them to use different sets of markers, including missing values, when
30 determining surrogate parents and identifying haplotypes. We implemented and tested
31 these extensions in an updated version of our phasing software AlphaPhase.

32 **Results**

33 A simulated dataset with one million individuals genotyped with the same set
34 of 6,711 SNP for a single chromosome took two days to phase. A larger dataset with
35 one million individuals genotyped with 49,579 SNP for a single chromosome took 14
36 days to phase. The percentage of correctly phased alleles at heterozygous loci was

37 respectively 90.5% and 90.0% for the two datasets, which is comparable to the
38 accuracy achieved with previous versions of AlphaPhase on smaller datasets.

39 The phasing accuracy for datasets with different sets of markers was generally
40 lower than that for datasets with one set of markers. For a simulated dataset with three
41 sets of markers 2.8% of alleles at heterozygous positions were phased incorrectly
42 whereas the equivalent figure with one set of markers was 0.6%.

43 **Conclusions**

44 The improved long-range phasing and haplotype library imputation algorithms
45 enable AlphaPhase to quickly and accurately phase very large and heterogeneous
46 datasets. This will enable more powerful breeding and genetics research and
47 application.

48

49 **Keywords:** Phasing, Large datasets, Heterogeneous datasets

Background

Here we describe the latest improvements to the Long-Range Phasing (LRP) and Haplotype Library Imputation (HLI) algorithms to phase genotypes for hundreds of thousands of individuals that have been genotyped on different platforms. Phasing genotypes is the process of inferring the parental origin of individual's alleles. This process resolves the inheritance of chromosome segments in a population and is as such a cornerstone technique in genetics. For example, it is useful for making genotype calls, imputing genotypes, detecting phenotype-genotype associations in the presence of effects such as allele-specific expression, and inferring recombination points and demographic history [1].

The size of genomic datasets has grown rapidly in recent years as genotype data is collected on an increasing number of individuals. In agriculture this growth has been driven by the increased value of genomic selection [2–4], whereas in human genetics it has been driven by the increased power of genome-wide association studies [5–7] and genomic prediction in human medicine [8]. Examples of such large datasets include the UK Biobank [9], which has recently released genotype data from approximately half a million people [10], and the US Dairy Cattle and Irish Cattle Breeding Federation Databases that each host well over a million of genotyped animals [4,11,12].

In many cases these datasets have been collected over several years and have been genotyped using different single nucleotide polymorphism (SNP) arrays [4,12]. Methods such as SNPchiMp [13] have been developed to allow the manipulation of different sets of markers from multiple SNP arrays, but their main aim is to ensure

74 that the different sets are combined correctly rather than to perform analyses of the
75 combined dataset.

76 Several methods for phasing genotype data have been developed based on
77 probabilistic methods, such as those implemented in fastPHASE [14] and Beagle [15].
78 Others, such as AlphaPhase [16] and findHap [17], are based on heuristic methods.
79 Recent developments in probabilistic methods e.g., SHAPEIT3 [18] and Beagle [15],
80 have enabled phasing of very large datasets, potentially containing over one million
81 individuals [19]. Heuristic methods are fast when compared to statistical methods and,
82 in many cases, more accurate. Thus, enabling their use on large datasets will be
83 beneficial for large scale genomic studies.

84 AlphaPhase [16] is a heuristic method that combines LRP [20] and HLI. LRP infers
85 parental origin of alleles by finding surrogate parents of an individual, that is
86 individuals who likely have the same haplotype as the individual. If a surrogate parent
87 is homozygous then it can be used to phase the individual's genotype. When a
88 homozygous surrogate parent cannot be found, surrogate parents of the heterozygous
89 surrogate parent can be used. This process is repeated, with increasingly remote
90 surrogate parents, until the individual's genotype can be phased.

91 HLI infers the phase of a genotype by creating a library of haplotypes that are fully
92 phased. Partially phased haplotypes can be fully phased by matching with library
93 haplotypes.

94 Existing publicly available LRP algorithms cannot efficiently phase large datasets as
95 finding surrogate parents amongst all individuals in a population involves comparing
96 every individual with every other individual. Both runtime and memory usage quickly

97 become impractical with large datasets as they scale with the square of the number of
98 individuals.

99 Additionally, existing publicly available algorithms for LRP and HLI could not phase
100 heterogeneous datasets with different sets of markers as they were not designed to
101 cope with large amounts of missing data. Combining data from multiple SNP arrays
102 can lead to large amounts of missing data.

103 In this paper we introduce improvements that allow a) LRP of large datasets and b)
104 LRP and HLI to work with missing data. These improvements enabled us to quickly
105 and accurately phase large heterogeneous simulated datasets. We phased one million
106 individuals, genotyped for 49,579 SNPs, in 14 days using modest computing
107 resources and correctly phased over 90% of alleles at heterozygous loci. We were also
108 able to phase a dataset consisting of individuals genotyped with three different arrays
109 and correctly phased 95% of alleles at heterozygous loci. The percentage of
110 incorrectly phased alleles at heterozygous loci was respectively only 1.0% and 2.8%
111 for the two examples. Our results show that it is possible to quickly and accurately
112 phase large heterogeneous datasets and our algorithm improvements will be of benefit
113 to those conducting large scale genomic studies.

114

115 **Methods**

116 **Previous LRP and HLI Algorithms**

117 Both LRP and HLI operate on genome regions called cores. A core is a set of
118 consecutive SNPs for which phasing is being attempted. For further details see Hickey
119 *et al.* [16].

120 LRP infers the phase of an individual by using other individuals known to share a
121 haplotype with the individual. Individuals sharing a haplotype are called “surrogate
122 parents” (shortened here to “surrogates”) and are identified by finding no opposing
123 homozygote markers at any position within a core. These surrogates are then
124 partitioned into either paternal or maternal surrogates of the individual using pedigree
125 information, if it is available. If pedigree information is not available, this assignment
126 is arbitrary.

127 If a surrogate is homozygous at a position then it enables phasing of the individual at
128 that position. If no homozygous surrogate is found, then it may be possible to phase
129 the individual by using surrogates of surrogates. This process can be continued to an
130 arbitrary depth. In practice, the consensus of several homozygous surrogates is taken
131 to allow for error in determining surrogates or genotype data.

132 HLI infers phase by matching partially phased haplotypes to a library of known
133 haplotypes. In the existing algorithm the initial library is constructed from the fully
134 phased haplotypes found during LRP and by adding new haplotypes as they are
135 discovered. New haplotypes are discovered when one haplotype of an individual is
136 inferred, because this haplotype together with genotype information determines the
137 other haplotype of the individual. This process is iterated until no new haplotypes are
138 found.

139 **Extending long-range phasing to large datasets**

140 To address the problem of scaling LRP to large datasets we modified the algorithm so
 141 that it is performed on subsets of individuals and the results from each subset are
 142 combined. By performing LRP on subsets the runtime can be vastly reduced, because
 143 search for surrogates has quadratic runtime scaling, and in the worst case involves all-
 144 against-all search for surrogates. When datasets are very large, all-against-all search
 145 for surrogates on the full dataset is too time consuming, while splitting the data into
 146 subsets limits the search time. Subsets of individuals, without replacement, are chosen
 147 randomly so that every individual is in a subset. These subsets are then merged and
 148 HLI is run on the complete dataset. We refer to this as the sub-setting method.

149 Preliminary analysis showed that including individuals in multiple subsets did not
 150 offer a significant improvement in accuracy, but increased runtime significantly (data
 151 not shown). Including related individuals in subsets also decreased accuracy (data not
 152 shown).

153 **Extending long-range phasing and haplotype library imputation to** 154 **heterogeneous datasets**

155 The LRP algorithm was modified to enable the identification of surrogates in the
 156 presence of missing data. Missing data hinders the identification of opposing
 157 homozygotes and so has the potential to wrongly identify surrogates. To alleviate this
 158 problem we introduced an additional parameter that defines the required number of
 159 shared markers by two individuals before surrogacy is tested.

160 The HLI algorithm required more complex modifications. In a multiple SNP array
 161 setting it is likely that most, or even all, individuals will have been genotyped with
 162 one array, so will not have a data for all array markers. Consequently, LRP cannot

infer parental origin of alleles at missing markers. We developed methods that allowed partially inferred haplotypes to be included in the haplotype library and to be used to infer other haplotypes.

Allowing for partially inferred haplotypes in the haplotype library makes matching a new partially inferred haplotype to a library haplotype much more difficult. It is necessary to ensure that the two haplotypes have enough markers with non-missing information to be confident they are the same haplotype. Thus, we added a parameter to the HLI algorithm that specifies the required number of shared alleles to match two haplotypes (Figure 1a).

In some cases it is possible that the new haplotype matches more than one library haplotype. In these cases it is possible that the new haplotype identifies duplicated library haplotypes (Figure 1b). In this situation we merge the new haplotype and library haplotypes, replace the incomplete library haplotypes with the merged haplotype, and update individuals known to carry the original incomplete library haplotypes.

If a new haplotype matches multiple library haplotypes and these matches cannot be the same haplotype, due to opposing homozygotes between the library haplotypes, then we add the new haplotype to the library.

Software engineering the new algorithms

Several changes were made to AlphaPhase to optimise it for speed and memory use. AlphaPhase was modified to store haplotypes and genotypes as bits and to use bit operations to operate on multiple SNPs at once wherever possible. It was also parallelised in several places to exploit high performance computing clusters.

186 **Test Datasets**

187 The performance of our new improvements to the LRP and HLI algorithms were
188 tested on large and heterogeneous simulated datasets.

189 **Simulated Data**

190 AlphaSim [21] was used to simulate test datasets. We followed the simulation scheme
191 from [22] which we describe briefly and show in Figure 2. AlphaSim first uses MaCS
192 [23] to simulate base population haplotypes. We simulated a single ‘breed’ that split
193 into three breeds 400 generations ago. 50 generations ago each of these breeds split
194 again into either three or four breeds to give ten breeds.

195 AlphaSim was then used to simulate two datasets consisting of ten equal-sized
196 ‘breeds’ and ten generations of selective breeding were simulated for each of these
197 ‘breeds’ (Figure 2). Selection was based on a single trait that had 10,000 quantitative
198 trait nucleotides with normally distributed effects. For the first dataset, for each breed
199 and for each generation, we selected 25 sires and 500 dams and generated 1,000
200 offspring. This resulted in a dataset of 100,000 animals (**100k dataset**). The second
201 dataset was created using 10,000 offspring for each breed and for each generation to
202 create a total dataset of one million animals (**one million dataset**). For both datasets
203 one chromosome worth of SNP data was generated and SNPs with a minor allele
204 frequency of at least 0.05 were chosen as possible candidates for inclusion on SNP
205 arrays.

206 SNPchiMp [13] was used to obtain information on the SNPs on different arrays and
207 the overlap between arrays. Across the bovine arrays there are 8,771 unique SNPs on
208 chromosome 1. We selected this number of SNPs from the candidate SNPs generated

209 by AlphaSim and then assigned SNPs to different arrays following the same pattern as
210 reported by SNPchiMp for bovine arrays.

211 We then used the assigned arrays to create scenarios (Table 1), where individuals
212 were genotyped with different arrays. There were two scenarios with **Homogeneous**
213 **Arrays**, where all individuals were genotyped with either the medium density (**MD**)
214 Bovine Illumina 50Kv2 or the high density (**HD**) Illumina HD SNP arrays. There
215 were five scenarios with **Heterogeneous Arrays**, where individuals were genotyped
216 with a set of partially overlapping combinations of SNP arrays. Three of these
217 scenarios were based on different MD chips. The **Two Illumina** scenario included
218 two different versions of the Illumina MD chip (Illumina 50K v1 and Illumina 50K
219 v2). The **Two Mixed** scenario combined one Illumina chip (Illumina 50K v2) and one
220 other chip (IDBv3). The **Three MD** scenario combined the Illumina 50Kv2 chip with
221 the IDBv3 chip and the GSeekHD chip. The **mixed MD / HD** scenario combined a
222 MD Illumina chip (Illumina 50K v2) with a HD Illumina chip (Illumina HD).

223 We created a further scenario that was not based on existing arrays, as in the future it
224 is likely that individuals will be genotyped on a wider range of SNP arrays. The **Ten**
225 **Array scenario** comprises five HD arrays and five MD arrays. We based the first HD
226 and first MD arrays on the Bovine Illumina HD and Bovine Illumina 50Kv2 arrays,
227 respectively.

228 We then created three further HD and three further MD arrays based on these two
229 arrays by splitting SNPs into four categories, those on both the original HD and MD
230 arrays (**HD/MD set**), those just on the HD array (**HD set**), those just on the MD array
231 (**MD set**) and those on neither (**unused set**). We removed ten percent of the SNPs
232 from each of the HD/MD, HD and MD sets and replaced them with randomly

233 sampled SNPs from the unused set. From these new sets we then created a second HD
234 and second MD array. We generated the third and fourth HD and MD arrays from the
235 previous arrays in the same way. Rather than removing the exact number of SNPs that
236 were to be added (or removed) we sampled based on a probability that resulted in the
237 expected number of SNPs being added (or removed). Because of this sampling the
238 resulting arrays were of slightly different in size.

239 We created the fifth HD and MD arrays to represent arrays from a different family of
240 arrays, perhaps from a different manufacturer. These arrays were simulated in a
241 similar way to the other arrays but by removing and replacing fifty percent of the
242 SNPs from the first HD and MD arrays.

243 For all scenarios we simulated individuals as being genotyped on different arrays by
244 assigning them to arrays in proportions we might expect to see in real datasets (Table
245 1).

246 **Phasing parameters used for AlphaPhase**

247 AlphaPhase has several parameters that control phasing of alleles. Two of these were
248 expected to have a significant effect on the performance of AlphaPhase – the existing
249 parameter controlling core length (defined as the number of SNP in each core) and a
250 new parameter that controlled the size of phasing subsets to speedup phasing of a
251 large dataset.

252 The length of the cores can have a significant effect on phasing accuracy [16]. To find
253 the best core length for both of the MD and HD scenarios we tested different core
254 lengths. For the Illumina 50Kv2 scenarios we tested core lengths in the same range as
255 those tested in Hickey *et al.* [16] for a similar size array: 50, 100, 200, 500, and 1,000

256 SNPs. For the Illumina HD scenario we tested core lengths of 500, 1,000, 2,000,
257 5,000, and 10,000 SNPs because the Illumina HD array contains approximately ten
258 times as many SNPs as the Illumina50Kv2 array.

259 We tested different sizes of the phasing subsets as this was expected to have an effect
260 on phasing accuracy. Tested values were 500, 1,000, 2,000, 5,000, and 10,000
261 individuals. For the Illumina 50Kv2 scenario we tested all combinations of core
262 length and subset size. We only report subset size results for a fixed core length of
263 500 SNPs since the interaction between core length and subset size was minimal (data
264 not shown). For the Illumina HD scenario we set the core length to 5,000 SNPs when
265 testing subset size.

266 For evaluating the Heterogeneous Array scenarios we set the core length to 500 SNPs
267 when the dataset consisted of only MD arrays since this value gave good performance
268 in Homogeneous MD Array scenarios. Similarly, for datasets containing HD arrays
269 we set core length to 5,000 SNPs. In both scenarios we set subset size to 5,000 SNPs.

270 AlphaPhase had several other parameters for which fixed default values were used.
271 Specifically, we fixed the maximum number of surrogates used to ten and allowed
272 10% of the marker genotypes to disagree between pairs of surrogates. We also set the
273 number of allele mismatches for clustering pairs of nearly identical library haplotypes
274 to be zero.

275 When phasing multiple arrays we added an additional parameter to AlphaPhase. This
276 parameter governs the minimum required number of matching alleles before two
277 haplotypes can be identified as the same haplotype. If all SNPs were independent of
278 each other, i.e., there was no linkage between them, we would expect the optimal

279 value of this parameter to remain unchanged irrespective of SNP density. Our results
280 (data not shown) suggest that the presence of linkage does not have a significant
281 effect for the SNP densities considered here and that requiring a match of 200 alleles
282 between two haplotypes is an appropriate value for this parameter. If SNP arrays with
283 greater density are considered then the value of this parameter may need to be revised.

284 **Performance Testing**

285 To test the performance of the new improvements to the LRP and HLI algorithms on
286 large datasets we used the data from the Homogeneous Array scenarios for both the
287 100k and one million datasets. To test the scenario where parents are known and
288 genotype information is available for them we evaluated phasing accuracy within each
289 of the ten breeds individually using data from all generations. Similarly, to test the
290 scenario where no parentage information is available we evaluated phasing accuracy
291 for each of the ten generations individually (Figure 2). We report average results
292 across either all ten families or all ten generations.

293 To test the speed and memory usage of AlphaPhase on large datasets we tested
294 multiple combinations of number of generation and families from both the 1000k and
295 the one million datasets using Homogeneous Array scenarios. To test the performance
296 of the new improvements to the LRP and HLI algorithms on heterogeneous datasets
297 we used the data from the Heterogeneous Array scenarios on the 100k dataset.

298 We report three phasing statistics – percentage of correctly phased alleles, percentage
299 of unphased alleles, and percentage of incorrectly phased alleles. Unless explicitly
300 stated otherwise, we report these statistics for heterozygous loci only. We also report
301 on memory usage and runtimes. Runs were performed on computers with an Intel
302 Xeon Processor E5-2630 v3 (2.4 GHz) and between 64 and 256GB of RAM.

Results

Long Range Phasing and Haplotype Library Imputation of Large Datasets

Core Length

To determine the accuracy of our new sub-setting method we first determined the optimal core length for each of the Illumina 50Kv2 and Illumina HD scenarios. Figure 3 and Tables S1-S2 show the accuracy on the Illumina 50Kv2 scenario for a variety of core lengths. Figure 3a shows the percentage of correctly phased heterozygous loci for the Illumina 50Kv2 array per family scenario. The percentage of correctly phased alleles increased as the core length increased, although the difference in accuracy between a core length of 500 SNPs (92.7%) and 1,000 SNPs (93.3%) was small. For the per generation scenario the percentage of correctly phased alleles peaked at a core length of 500 SNPs (92.3%) before dropping significantly for a core length of 1,000 SNPs (89.8%). The pattern for the number of incorrectly phased alleles for the Illumina 50Kv2 (Figure 3b) was less clear although there was a significant increase in the number of incorrectly phased alleles for a core length of 1,000 SNPs. Using a core length of 500 SNP the percentage of alleles incorrectly phased was 0.6% (per family scenario) and 0.9% (per generation scenario).

Figure 3 and Tables S3-S4 show that for the Illumina HD scenario the percentage of correctly phased alleles at heterozygous loci peaked at either a core length of 2,000 SNPs (per generation) or 5,000 SNPs (per family). For this scenario the number of incorrectly phased alleles was minimised at a core length of 1,000 SNPs (0.3% per family; 0.5% per generation), a shorter core length than that which maximised the number of correctly phased markers. Using a core length of 5,000 SNPs 93.5% (per

family) or 92.1% (per generation) of alleles were phased correctly, while 0.8% (per family) or 1.3% (per generation) were phased incorrectly.

In all scenarios runtime was inversely proportional to core length (Tables S1-S4). We chose to study core lengths of 500 SNPs (for MD scenarios) and 5,000 SNPs (for HD scenarios) as a reasonable trade-off between accuracy and runtime. For these core lengths runtime was around three minutes for both arrays and for both the per generation and per family scenarios. Memory usage was 2.5GB for the Illumina 50Kv2 array and 5.3GB for the Illumina HD array (Tables S1-S4).

Subset Size

Subset size can be expected to have a significant effect on the accuracy of phasing as it will directly influence the number of surrogates that are found. To test this we evaluated subset sizes of between 500 and 10,000 individuals (Figure 4, Tables S5-S8). For both the Illumina 50Kv2 and Illumina HD arrays accuracy increased as the subset size increased. For the Illumina 50Kv2 per family scenario the percentage of correctly phased alleles at heterozygous loci increased from 86.9% to 97.9% as subset size increased from 500 to 10,000 individuals. For the Illumina HD per family scenario it increased from 87.0% (500 individuals) to 97.8% (10,000 individuals). The results for phasing in the per generation scenarios were similar. The percentage of correctly phased alleles increased from 84.8% to 95.8% for the Illumina 50 Kv2 array and from 84.0% to 94.9% for the Illumina HD array as the subset size increased from 500 to 10,000 individuals.

In nearly all scenarios runtime was proportional to subset size (Tables S5-S8). The exception was for the Illumina 50Kv2 per family scenario in which runtime for a subset size of 10,000 individuals was noticeably shorter than runtime for a subset size

350 of 5,000 individuals. This was likely due to nearly all animals having had both parents
351 included in the subset when the subset size was 10,000, which can significantly
352 decrease the time required to partition surrogates into maternal and paternal
353 surrogates. Memory usage also increases as subset size grows. We chose to use a
354 subset size of 5,000 for the remainder of this study as a reasonable trade-off between
355 accuracy and runtime. With subsets of this size the percentage of correctly phased
356 alleles at heterozygous loci was 92.7% (per family) or 92.3% (per generation) for the
357 Illumina 50Kv2 scenario and 93.5% (per family) or 92.1% (per generation) for the
358 Illumina HD scenarios. The percentage phased incorrectly was 0.1% (per family) or
359 0.9% (per generation) for the Illumina 50Kv2 scenarios and 0.8% (per family) or
360 1.3% (per generation) for the Illumina HD scenarios.

361 **Accuracy, Runtime, and Memory Usage on Different Dataset Sizes**

362 To test the performance of our new improvements to the LRP and HLI algorithms on
363 datasets of different sizes we created multiple different sized scenarios from the 100k
364 and the one million datasets. Phasing accuracy was broadly comparable to the phasing
365 accuracy observed when investigating optimal core length and subset size (Tables S9
366 and S10). Figure 5 shows that runtimes scaled approximately linearly with the number
367 of individuals in a dataset. For the Illumina 50Kv2 dataset memory usage varied
368 between 0.6GB for the smallest dataset of 1,000 individuals to 76GB for a dataset of
369 one million individuals (Figure 6 and Table S19). Comparable figures for the Illumina
370 HD dataset were 0.9GB and 325GB (Figure 6 and Table S10).

371 **Heterogeneous Datasets**

372 Table 2 shows phasing accuracy, runtime and memory requirements for each of the
373 five Heterogeneous Arrays per family scenarios. For the scenarios involving only MD

arrays the percentage of alleles at heterozygous loci phased correctly was between 93.8% and 95.2% with between 1.9% and 2.8% phased incorrectly. For the two array scenarios (Two Illumina and Two Mixed) runtime was approximately two minutes. For the Three MD scenario runtime was approximately five minutes. For the two array scenarios memory usage was around 2.6GB whereas for the Three MD scenario memory usage was approximately 3GB.

We also tested a mixture of one MD array and one HD array with nine individuals genotyped on the MD array for every individual genotyped on the HD array (Mixed MD/HD). As expected the percentage of correctly phased alleles at heterozygous loci was lower than in other scenarios, but was still 93.4%. Runtime was around six hours and memory usage was 5.3GB. For the Ten Array scenario the percentage of correctly phased alleles was slightly higher at 95.4%, although memory usage was also higher at 7.6GB.

Table 3 shows results for the per generation scenarios. Results were broadly comparable to the per family results. Across the scenarios containing only MD arrays the percentage of correctly phased alleles at heterozygous loci was between 93.1% and 95.6% with between 3.1% and 3.6% incorrectly phased. Runtime was similar to the per family scenarios taking around three minutes for the two array scenarios and two minutes for the Three MD scenarios. Memory usage was the same as that of the per family scenarios.

The Mixed HD/MD per generation scenario showed a lower percentage of correctly phased alleles at heterozygous loci (88.4%) compared to the per family scenario (93.4%). Runtime and memory requirements were similar. The Ten Array per generation scenario also had lower accuracy than the per family scenario with 93.8%

398 of alleles correctly phased compared with 95.4% in the per family scenario. Runtime
399 and memory usage were similar for the per family and the per generation results.

400

Discussion

In this paper we introduced improvements to the LRP and HLI algorithms of AlphaPhase [16] to enable phasing of very large and heterogeneous datasets in which individuals have been genotyped on differing sets of markers. We tested the revised algorithms' performance on a range of simulated datasets and show that AlphaPhase can be used to accurately phase datasets that contain up to one million individuals and that have been genotyped with multiple different SNP arrays. In what follows we discuss the effect of: (i) core length and (ii) subset size on phasing accuracy and computational runtime and memory use; and (iii) the impact of these improvements on the phasing of large and heterogeneous datasets.

Effect of core length on phasing performance

Both LRP and HLI break the genome into smaller sections of consecutive SNPs called cores. Each of these cores is then phased independently of each other. Core length, defined as the number of SNPs in each core, has previously been shown to have a significant effect on accuracy [16]. In light of our new improvements to the LRP and HLI algorithms and the availability of much denser SNP arrays we further investigated the impact of this parameter.

As expected we found that core length has a significant effect on phasing accuracy. Short cores showed similar levels of phasing accuracy. Accuracy started to deteriorate notably as cores get longer. We also found that phasing accuracy is a function of the length of the core as a proportion of the length of the chromosome, rather than the number of SNPs it contains, and that the Illumina 50Kv2 and Illumina HD arrays show a very similar pattern of results when core length is expressed as a proportion of chromosome length. This is to be expected, as phasing accuracy is likely to be highly

425 affected by the presence of recombinations within a core. The latter can be reliably
 426 measured by the relative size of a core versus the whole chromosome, and less so by
 427 the number of SNP array markers in a core. The reduction in accuracy observed as the
 428 core length increased is likely due to the increased chance of a core containing a
 429 recent recombination. This would reduce the number of surrogates and thus, reduce
 430 the information available for phasing.

431 Both runtime and memory usage were significantly affected by core length with
 432 runtime being approximately proportional to the number of cores and therefore,
 433 inversely proportional to core length. Memory usage also increased as the number of
 434 cores increased although the effect was less pronounced than for runtime. For these
 435 reasons we recommend the use of the longest possible cores that do not result in an
 436 unacceptable drop in accuracy. For the bovine arrays considered in this paper, our
 437 results suggest a core length of 500 when only MD arrays are used and a core length
 438 of 5,000 when HD arrays have also been used.

439 **Effect of subset size on phasing performance**

440 Our new improvements of the LRP algorithm partition a large dataset into subsets and
 441 then performing LRP on each of the subsets. We introduced a new parameter to the
 442 LRP algorithm that controls the size of these subsets and our results show that this
 443 parameter can have a significant effect on phasing accuracy. A subset size of 10,000
 444 gave both the highest percentage of correctly phased alleles and the lowest percentage
 445 of incorrectly phased ones. As the subset size decreased the proportion of correctly
 446 phased alleles decreased and the proportion of incorrectly phased alleles loci
 447 increased. This decrease in phasing performance was very likely due to the reduction

448 in the number of surrogates that would be expected in a smaller subset which, in turn,
449 led to less information with which to accurately phase.

450 The increase in phasing accuracy that resulted from increasing subset size had a
451 significant cost in terms of runtime. We found an approximately linear relationship
452 between the size of the subset and runtime with the used runtime parameters.
453 Consequently, there was a trade-off between runtime and accuracy. As the size of
454 subsets got larger the number of incorrectly phased alleles appeared to begin to
455 plateau for a subset size of 5,000 or greater and runtime started to increase
456 significantly. For the datasets we tested the subset size for which this value occurred
457 seemed to be largely invariant to total dataset size and so we suggest that a subset size
458 of 5,000 is an appropriate compromise. The optimal size for datasets with a different
459 structure, such as a human populations in which individuals are likely to be less
460 related than the ones considered here, warrants further investigation.

461 Phasing of large datasets is likely to be computationally expensive for any phasing
462 techniques due to the large number of individuals involved. Our results suggest that
463 there is sufficient information in small subsets from a larger dataset to allow a
464 significant number of alleles to be phased accurately. This suggests that for other
465 phasing methods, such as those based on probabilistic models, it could also be
466 beneficial to break the phasing of large datasets into subsets before merging the
467 results.

468 **Ability of AlphaPhase to phase large heterogeneous datasets**

469 Our results show that it is viable to run heuristic phasing on very large datasets, such
470 as those now available for humans [10] or cattle [4,11,12]. AlphaPhase took two days
471 and 76GB of memory to phase one million animals genotyped on a simulated Illumina

472 50Kv2 array. To phase one million animals genotyped on a simulated Illumina HD
473 array took 14 days and 325GB of memory.

474 The ability to phase datasets genotyped using multiple different arrays is important as
475 datasets are increasingly likely to consist of individuals genotyped using different
476 arrays due to the increase in the number of available arrays for commonly genotyped
477 species. Results from the analysis of the Heterogenous Arrays scenarios show that
478 similar phasing accuracy can be achieved for heterogenous datasets, consisting of
479 individuals genotyped on multiple MD arrays, as can be achieved for homogeneous
480 datasets. In general, accuracy was slightly worse than for the single array scenarios
481 that were tested, although in many scenarios the Heterogeneous Arrays phased
482 slightly more alleles correctly. However, this increase in percentage of correctly
483 phased alleles came at the cost of phasing more alleles incorrectly as well.

484 AlphaPhase can now also accurately phase individuals genotyped on a mixture of MD
485 and HD SNP arrays. The phasing of such datasets is likely to become increasingly
486 common as it is desirable to continue to use the data already collected using MD
487 arrays even as the use of HD arrays grows. Although the phasing accuracy for
488 heterogeneous datasets was often lower than when individuals were genotyped on a
489 single SNP array, the percentage of correctly phased alleles was still over 93% in all
490 scenarios tested other than the Mixed MD/HD per generation scenario. In this
491 scenario many individuals had high amounts of missing data due to them being
492 genotyped on the MD rather than HD array and so we would expect phasing to be
493 more difficult.

494 **Conclusions**

495 We have modified the LRP and HLI algorithms to allow phasing of large
496 heterogeneous datasets. These modifications are implemented in AlphaPhase version
497 1.3.7 (available from <http://alphagenes.roslin.ed.ac.uk/>) and allow the accurate
498 phasing of millions of individuals genotyped on multiple SNP arrays.

499 **Declarations**

500 **Ethics approval and consent to participate**

501 Not applicable

502 **Consent for publication**

503 Not applicable

504 **Availability of data and materials**

505 An implementation of our algorithm, in the software package AlphaPhase is available
506 from the authors' website, <https://alphagenes.roslin.ed.ac.uk/wp/software/alphaphase/>,
507 and is free for academic use.

508 **Competing interests**

509 The authors declare that they have no competing interests.

510 **Funding**

511 The authors acknowledge the financial support from the BBSRC ISPG to The Roslin
512 Institute BB/J004235/1, from Genus PLC, and from Grant Nos. BB/M009254/1,
513 BB/L020726/1, BB/N004736/1, BB/N004728/1, BB/L020467/1, BB/N006178/1 and
514 Medical Research Council (MRC) Grant No. MR/M000370/1.

515 **Authors' contributions**

516 JMH and DM designed the updates to the algorithm and this study. DM and DW
517 implemented the updates. JJ and GG provided the simulation strategy. DM conducted
518 the study. GG provided critical comments throughout the study. All authors read and
519 approved the final manuscript.

520 **Acknowledgements**

521 This work has made use of the resources provided by the Edinburgh Compute and
522 Data Facility (ECDF) (<http://www.ecdf.ed.ac.uk>).

523 **References**

- 524 1. Browning SR, Browning BL. Haplotype phasing: existing methods and new
525 developments. *Nat Rev Genet.* 2011;12:703–14.
- 526 2. Gorjanc G, Cleveland MA, Houston RD, Hickey JM. Potential of genotyping-by-
527 sequencing for genomic selection in livestock populations. *Genet Sel Evol.*
528 2015;47:12.
- 529 3. Meuwissen T, Hayes B, Goddard M. Genomic selection: A paradigm shift in
530 animal breeding. *Anim Front.* 2016;6:6–14.
- 531 4. Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic Selection in Dairy
532 Cattle: The USDA Experience. *Annu Rev Anim Biosci.* 2017;5:309–27.
- 533 5. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery.
534 *Am J Hum Genet.* 2012;90:7–24.
- 535 6. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10
536 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet.*
537 2017;101:5–22.
- 538 7. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase
539 information for human genomics. *Nat Rev Genet.* 2011;12:215.
- 540 8. Visscher PM. Human Complex Trait Genetics in the 21st Century. *Genetics.*
541 2016;202:377–9.
- 542 9. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank:
543 An Open Access Resource for Identifying the Causes of a Wide Range of Complex
544 Diseases of Middle and Old Age. *PLOS Med.* 2015;12:e1001779.

10. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*. 2017;166298.
11. Two Million Genotypes in U.S. Dairy Database [Internet]. Dairyherd. [cited 2018 Feb 5]. Available from: <https://www.dairyherd.com/article/two-million-genotypes-us-dairy-database>
12. McClure MC, McCarthy J, Flynn P, McClure JC, Dair E, O'Connell DK, et al. SNP Data Quality Control in a National Beef and Dairy Cattle System and Highly Accurate SNP Based Parentage Verification and Identification. *Front Genet* [Internet]. 2018 [cited 2018 Aug 9];9. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00084/full#SM1>
13. Nicolazzi EL, Caprera A, Nazzicari N, Cozzi P, Strozzi F, Lawley C, et al. SNPchiMp v.3: integrating and standardizing single nucleotide polymorphism data for livestock species. *BMC Genomics*. 2015;16:283.
14. Scheet P, Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *Am J Hum Genet*. 2006;78:629–44.
15. 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.
16. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JH. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet Sel Evol*. 2011;43:12.
17. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. *J Dairy Sci*. 2013;96:668–78.
18. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. *Nat Genet*. 2016;48:817.
19. Whalen A, Gorjanc G, Ros-Freixedes R, Hickey JM. Assessment of the performance of different hidden Markov models for imputation in animal breeding. *bioRxiv*. 2017;227157.
20. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet*. 2008;40:1068–75.
21. 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.
22. Jenko J, Whalen A, Gaynor R, Dadousis C, Gorjanc G, Hickey J. Identification of causal variants using one million individuals with whole-genome sequence information. In: *Proceedings of the World Congress on Genetics Applied to Livestock Production: Auckland*; 591, February 2018; 2018.

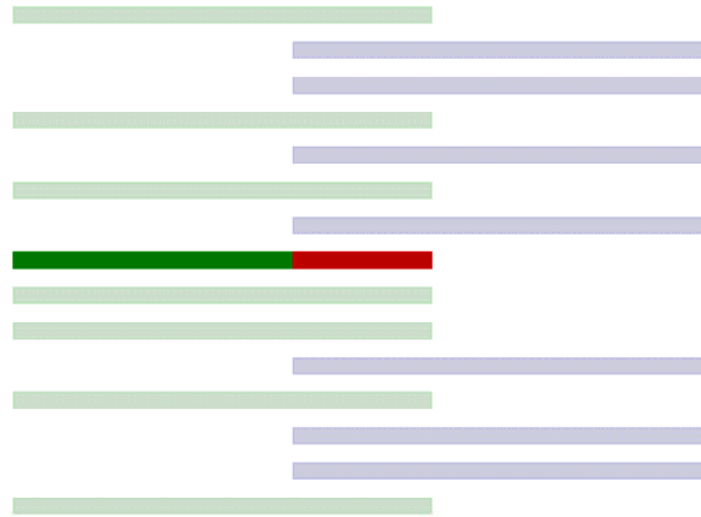
- 584 23. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence
585 data. *Genome Res.* 2009;19:136–42.

Figures

Figure 1 - New improvements to the LRP and HLI algorithms for dealing with library haplotypes with missing data. a) In this example we have generated haplotypes using two different SNP arrays indicated by green and blue haplotypes. If the shared markers between two haplotypes are identical (shown in red) then the two haplotypes can be merged into one haplotype. To ensure the two haplotypes are the same haplotype we set a minimum number of alleles that must be shared. Note that in reality blue and green markers will both be distributed along the length of the haplotype. b) In this example we have generated haplotypes using three different SNP arrays. Finding the new purple haplotype allows us to recognise that the green, purple, and blue haplotype are actually the same haplotype.

598 a)

Library Haplotypes



New Haplotype



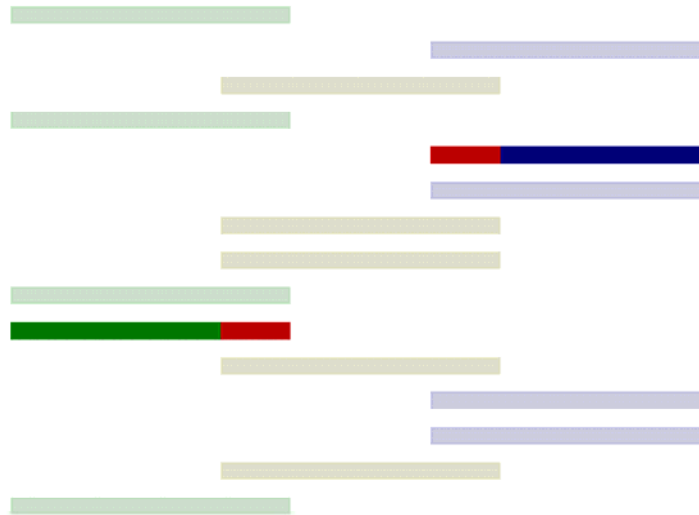
Merged Haplotype



599
600

601 b)

Library Haplotypes



New Haplotype

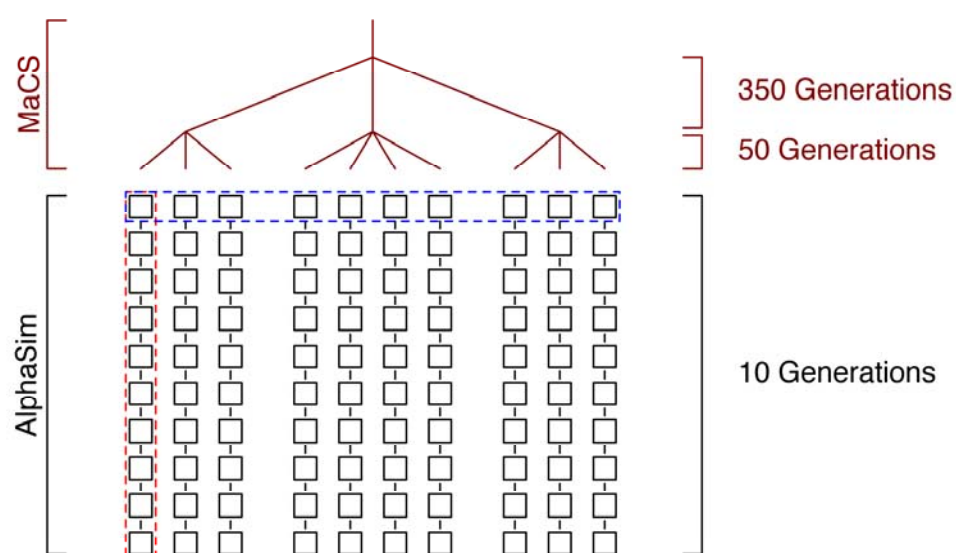


Merged Haplotype



602

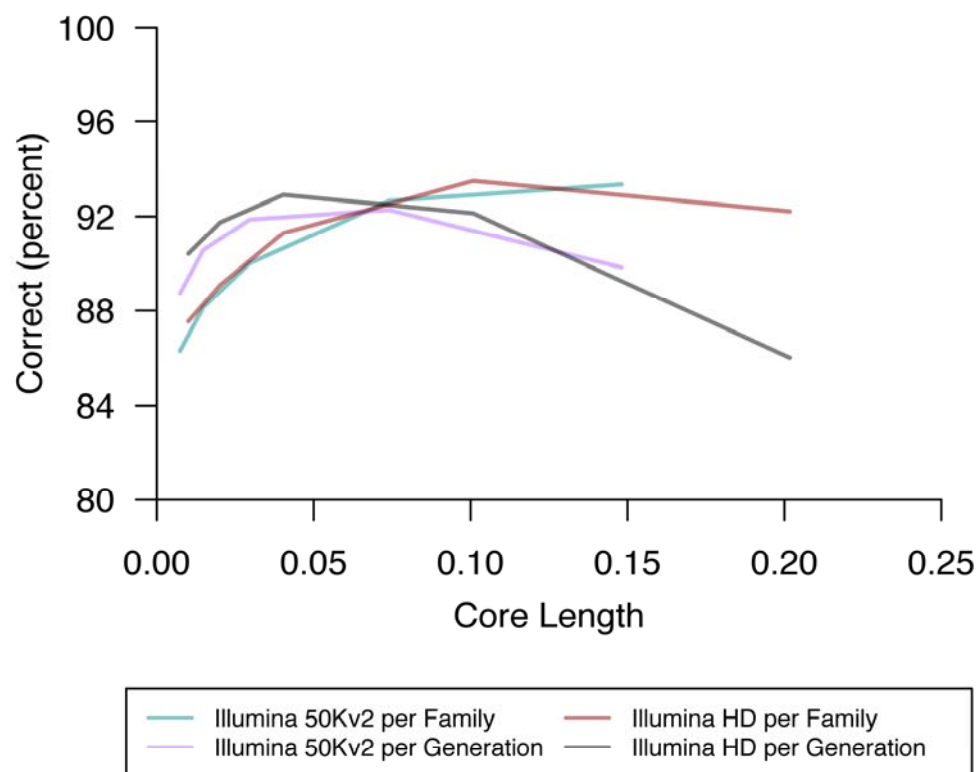
Figure 2 - Simulation structure. MaCS is used to simulate a base population. This base population is generated from a single ‘breed’ that split into three breeds 400 generations ago. 50 generations ago each of these breeds split again into either three or four breeds to give ten breeds. Each of these ten breeds then undergoes ten generations of selection using AlphaSim. The dotted blue line shows an example for the “per generation” scenario, while the dotted red line shows an example for the “per family” scenario.



610

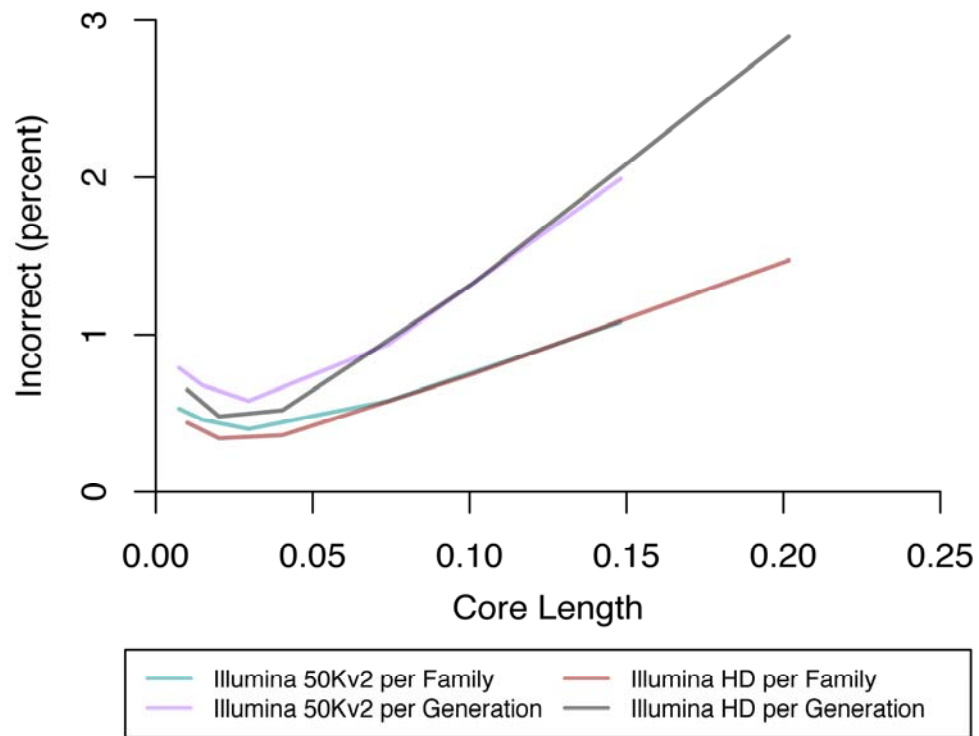
611 **Figure 3 – a)** Percentage of correctly phased alleles at heterozygous loci for a range
612 of core lengths. b) Percentage of incorrectly phased alleles at heterozygous loci for a
613 range of core lengths. Core lengths are given as a proportion of the total
614 chromosome length.

615 a)



616
617

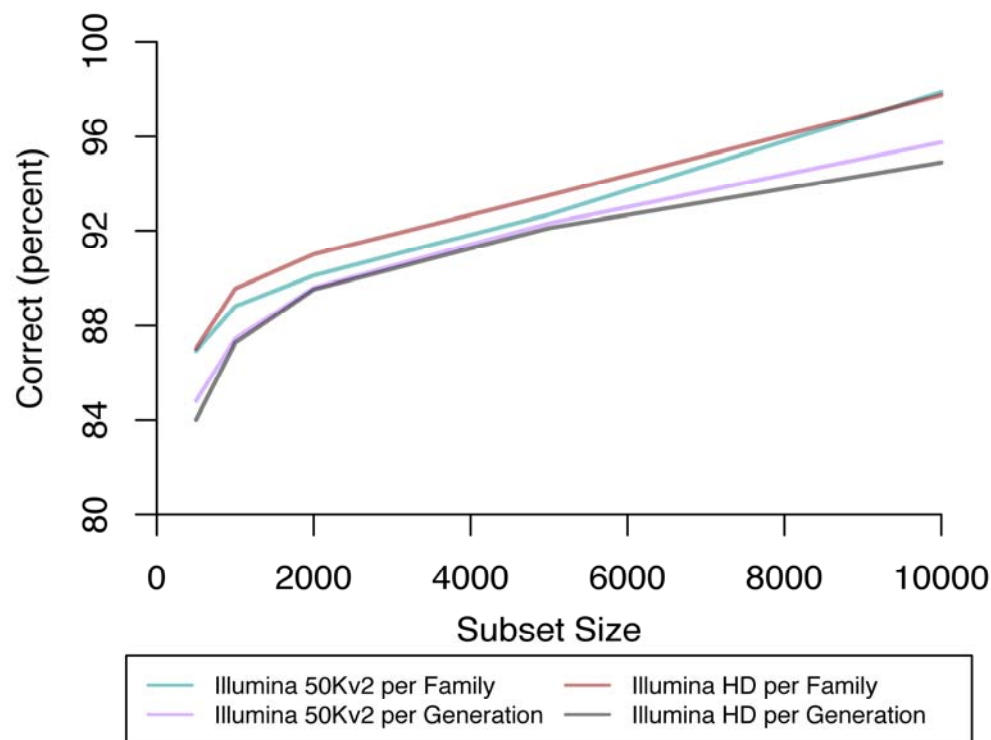
618 b)



619

620 **Figure 4 - a)** Percentage of correctly phased alleles at heterozygous loci for a range
 621 of subset sizes. b) Percentage of incorrectly phased alleles at heterozygous loci for a
 622 range of subset sizes.

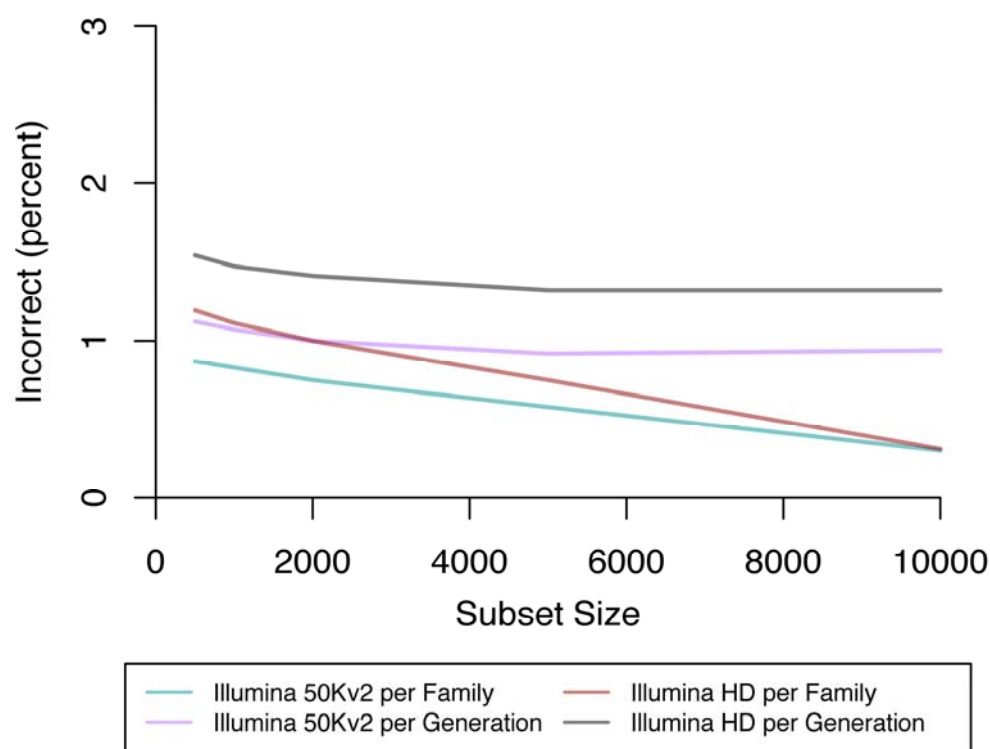
623 a)



624

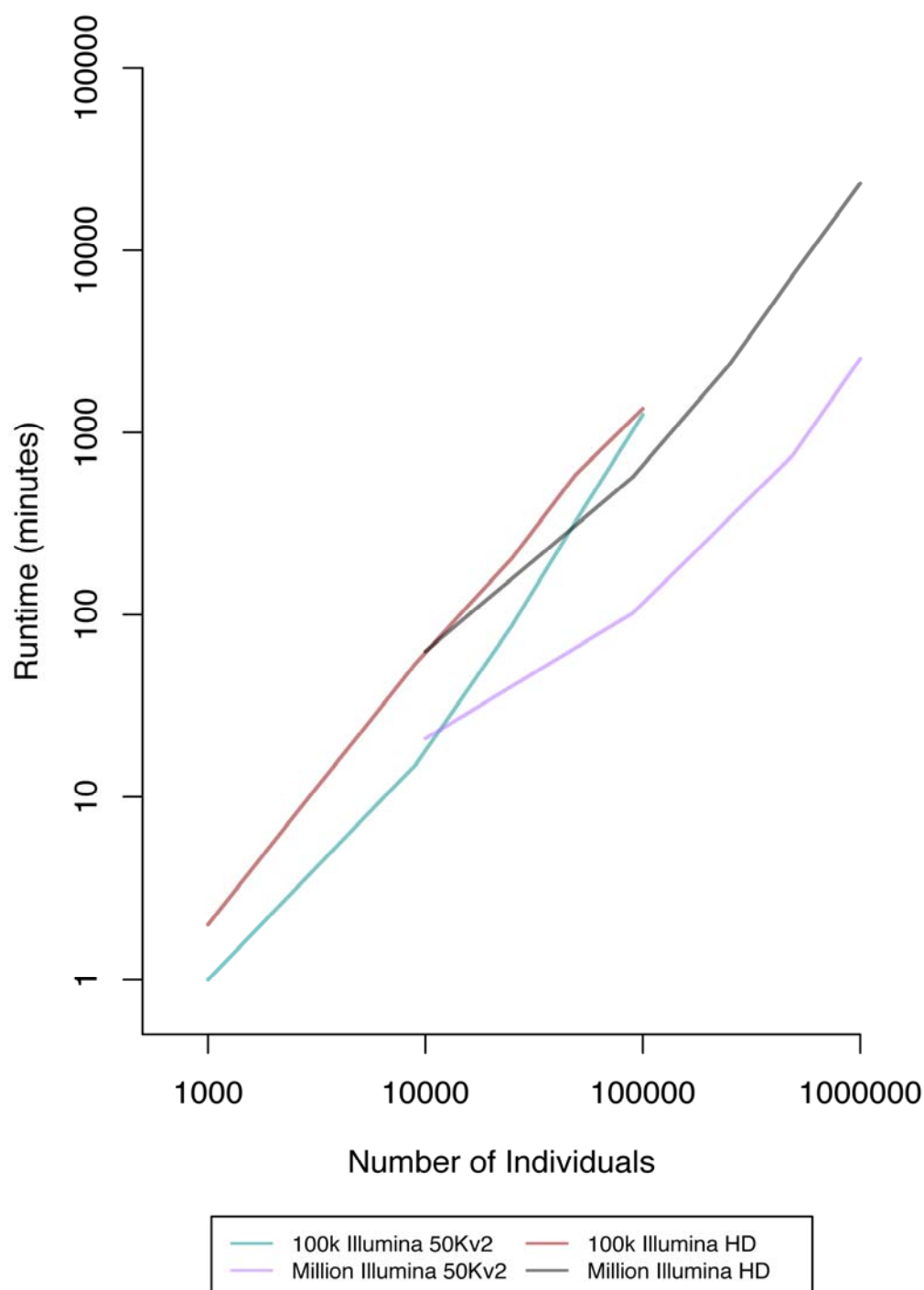
625

626 b)



627

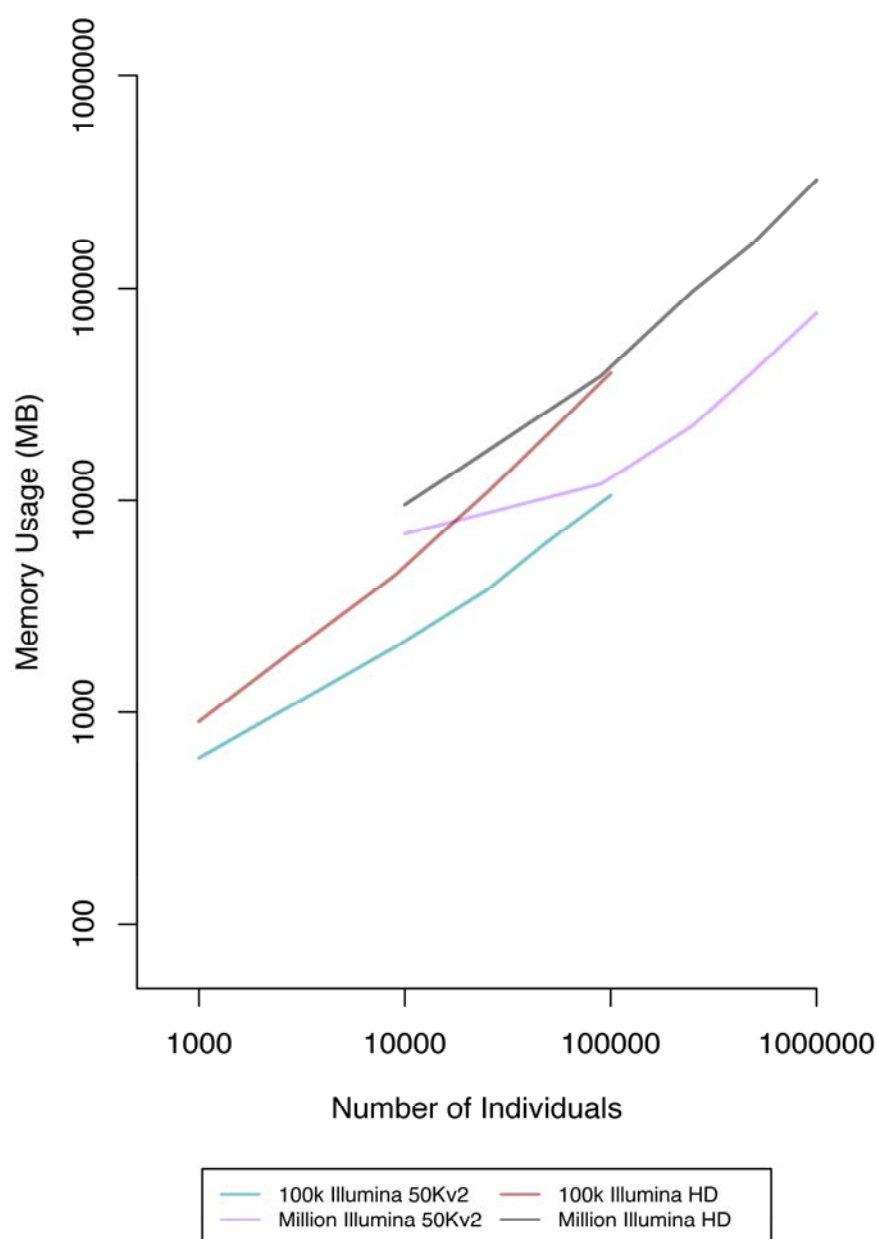
628 **Figure 5** – Runtime of AlphaPhase for a range of dataset sizes genotyped on two
629 different SNP arrays.



630

631

632 **Figure 6** – Memory usage of AlphaPhase for a range of dataset sizes genotyped on
633 two different SNP arrays.



634
635

636 Tables

637 **Table 1 – The different genotyping scenarios tested**

Scenario	Description
Illumina 50Kv2	Illumina 50Kv2 (all)
Illumina HD	Illumina HD (all)
Two Illumina	Illumina 50Kv1 and Illumina 50Kv2 in a 1:1 ratio
Two Mixed	Illumina 50Kv2 and IDBv3 in a 1:1 ratio
Three MD	Illumina 50Kv2, GSeekHD and IDBv3 in a 1:1:1 ratio
Mixed MD/HD	Illumina 50Kv2 and Illumina HD in a 9:1 ratio
Ten Array	Five MD and Five HD with equal numbers of individuals genotyped on each array

638 **Table 2 – Different genotype scenarios per family results**

Scenario	All			Heterozygous			Time (minutes)	Memory (MB)
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect		
Illumina 50Kv2	98.30	1.59	0.11	92.65	6.77	0.58	143	2,585
Illumina HD	98.40	1.46	0.14	93.48	5.77	0.75	154	5,279
Two Illumina	94.12	3.18	2.70	93.83	4.25	1.92	126	2,570
Two Mixed	96.67	2.77	0.56	94.41	3.26	2.33	153	2,609
Three MD	96.66	2.41	0.93	95.15	2.05	2.79	322	3,034
Mixed MD/HD	93.64	5.12	1.23	93.41	4.38	2.20	342	5,311
Ten Array	95.12	4.05	0.83	95.37	3.66	0.97	562	7,671

639 **Table 3 – Different genotype scenarios per generation results**

640

Scenario	All			Heterozygous			Time (minutes)	Memory (MB)
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect		
Illumina 50Kv2	98.29	1.55	0.16	92.27	6.79	0.94	183	2,534
Illumina HD	98.40	1.46	0.14	93.50	5.75	0.75	165	5,262
Two Illumina	95.04	2.08	2.88	93.06	3.82	3.13	147	2,544
Two Mixed	97.51	1.70	0.80	94.02	2.38	3.60	196	2,591
Three MD	97.13	1.66	1.22	95.59	1.26	3.15	107	3,040
Mixed MD/HD	91.15	6.86	1.99	88.36	6.72	4.92	340	5,257
Ten Array	94.36	4.35	1.29	93.80	4.56	1.64	568	7,732

Supplementary tables

Table S1: Illumina 50Kv2 per family results for a range of core lengths

Core Length	All Loci			Heterozygous Loci			Time	Memory
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect	(minutes)	(MB)
50	97.37	2.53	0.10	86.30	13.16	0.53	3,247	5,930
100	97.66	2.25	0.09	88.14	11.40	0.46	1,392	4,286
200	97.93	2.00	0.07	90.02	9.58	0.40	561	3,271
500	98.30	1.59	0.11	92.65	6.77	0.58	143	2,585
1,000	98.39	1.40	0.21	93.33	5.60	1.08	39	2,060

Table S2: Illumina 50Kv2 per generation results for a range of core lengths

Core Length	All Loci			Heterozygous Loci			Time	Memory
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect	(minutes)	(MB)
50	97.92	1.93	0.15	88.75	10.46	0.79	4,017	5,946
100	98.18	1.69	0.13	90.57	8.76	0.68	1,729	4,263
200	98.34	1.56	0.10	91.87	7.55	0.58	706	3,233
500	98.29	1.55	0.16	92.27	6.79	0.94	183	2,534

1,000	97.82	1.84	0.34	89.84	8.17	1.99	46	2,047
-------	-------	------	------	-------	------	------	----	-------

644 **Table S3: Illumina HD per family results for a range of core lengths**

Core Length	All Loci			Heterozygous Loci			Time	Memory
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect	(minutes)	(MB)
500	97.56	2.36	0.08	87.55	12.02	0.44	2,633	7,707
1,000	97.77	2.16	0.06	89.09	10.58	0.34	1,151	6,705
2,000	98.10	1.83	0.07	91.30	8.34	0.36	494	5,974
5,000	98.40	1.46	0.14	93.48	5.77	0.75	154	5,279
10,000	98.23	1.49	0.28	92.21	6.32	1.47	78	4,904

645 **Table S4: Illumina HD per generation results for a range of core lengths**

Core Length	All Loci			Heterozygous Loci			Time	Memory
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect	(minutes)	(MB)
500	98.16	1.71	0.12	90.42	8.93	0.65	3,138	7,783
1,000	98.32	1.59	0.09	91.75	7.77	0.48	1,351	6,683
2,000	98.43	1.48	0.09	92.90	6.58	0.52	583	5,990

5,000	98.19	1.59	0.23	92.13	6.55	1.32	183	5,255
10,000	97.19	2.31	0.50	86.01	11.10	2.89	79	4,877

646 **Table S5: Illumina 50Kv2 per family results for a range of subset sizes**

Subset Size	All Loci			Heterozygous Loci			Time	Memory
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect	(minutes)	(MB)
500	96.98	2.86	0.16	86.93	12.20	0.16	4	1,246
1,000	97.39	2.45	0.16	88.82	10.34	0.16	8	1,360
2,000	97.69	2.17	0.14	90.11	9.13	0.14	23	2,153
5,000	98.31	1.58	0.11	92.68	6.75	0.11	138	2,579
10,000	99.60	0.34	0.06	97.89	1.81	0.06	56	3,277

647 **Table S6: Illumina 50Kv2 per generation results for a range of subset sizes**

Subset Size	All Loci			Heterozygous Loci			Time	Memory
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect	(minutes)	(MB)
500	96.40	3.40	0.20	84.80	14.07	1.12	4	1,248
1,000	97.00	2.81	0.19	87.45	11.48	1.07	8	1,367

648

2,000	97.55	2.27	0.17	89.60	9.40	1.00	24	1,955
5,000	98.30	1.54	0.16	92.31	6.77	0.92	168	2,530
10,000	99.28	0.56	0.16	95.76	3.30	0.94	703	3,788

649 **Table S7: Illumina HD per family results for a range of subset sizes**

Subset Size	All Loci			Heterozygous Loci			Time (minutes)	Memory (MB)
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect		
500	96.94	2.83	0.22	87.03	11.78	1.19	20	4,146
1,000	97.45	2.34	0.21	89.56	9.33	1.11	30	4,243
2,000	97.76	2.05	0.19	91.00	8.00	1.00	49	4,854
5,000	98.40	1.46	0.14	93.50	5.75	0.75	165	5,262
10,000	99.57	0.37	0.06	97.76	1.94	0.31	247	6,109

650 **Table S8: Illumina HD per generation results for a range of subset sizes**

Subset Size	All Loci			Heterozygous Loci			Time (minutes)	Memory (MB)
	Correct	Unphased	Incorrect	Correct	Unphased	Incorrect		
500	96.20	3.54	0.27	84.01	14.45	1.54	19	4,172
1,000	96.86	2.89	0.25	87.27	11.26	1.47	30	4,280
2,000	97.43	2.33	0.24	89.52	9.07	1.41	48	4,759
5,000	98.18	1.59	0.23	92.11	6.57	1.32	163	5,261
10,000	99.12	0.65	0.22	94.90	3.78	1.32	533	6,586

651 **Table S9: Illumina 50Kv2 results for scenarios of different sizes**

Dataset	Families × Generations	Number of Individuals	All Loci			Heterozygous Loci			Time (minutes)	Memory (MB)
			Correct	Unphased	Incorrect	Correct	Unphased	Incorrect		
100k	1 × 1	1,000	99.62	0.30	0.07	98.39	1.29	0.31	1	613
100k	3 × 3	9,000	98.08	1.86	0.06	93.19	6.55	0.26	15	2,022
100k	5 × 5	25,000	98.01	1.89	0.10	92.59	6.96	0.45	87	3,770
100k	7 × 7	49,000	97.98	1.88	0.14	91.82	7.50	0.67	326	6,326
100k	10 × 10	100,000	97.81	2.02	0.17	90.25	8.83	0.92	1,253	10,542
One million	1 × 1	10,000	98.59	1.35	0.06	95.48	4.27	0.25	21	6,927
One million	3 × 3	90,000	96.98	2.94	0.07	89.73	9.95	0.32	102	11,972
One million	5 × 5	250,000	97.62	2.27	0.11	91.40	8.13	0.47	341	22,455
One million	7 × 7	490,000	97.70	2.17	0.13	91.30	8.10	0.60	751	40,112
One million	10 × 10	1,000,000	97.60	2.22	0.18	90.35	8.84	0.81	2,534	76,305

652 **Table S10: Illumina HD results for scenarios of different sizes**

Dataset	Families × Generations	Number of Individuals	All Loci			Heterozygous Loci			Time (minutes)	Memory (MB)
			Correct	Unphased	Incorrect	Correct	Unphased	Incorrect		

100k	1 × 1	1,000	99.57	0.36	0.08	98.16	1.52	0.33	2	906
100k	3 × 3	9,000	98.10	1.83	0.07	93.89	5.82	0.29	54	4,449
100k	5 × 5	25,000	98.04	1.85	0.12	93.27	6.20	0.53	205	10,864
100k	7 × 7	49,000	98.01	1.82	0.17	92.41	6.76	0.83	587	20,319
100k	10 × 10	100,000	97.75	2.01	0.24	90.47	8.27	1.26	1,350	39,682
One million	1 × 1	10,000	98.61	1.32	0.07	95.84	3.85	0.31	63	9,545
One million	3 × 3	90,000	97.18	2.74	0.09	91.35	8.28	0.37	572	38,500
One million	5 × 5	250,000	98.07	1.81	0.12	93.23	6.22	0.55	2,346	96,974
One million	7 × 7	490,000	97.96	1.88	0.16	92.30	6.98	0.72	7,313	162,024
One million	10 × 10	1,000,000	97.53	2.26	0.21	90.00	9.03	0.97	23,223	325,217

653