



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Continuous approximations for optimizing allele trajectories

Citation for published version:

Liu, AYH & Woolliams, JA 2010, 'Continuous approximations for optimizing allele trajectories', *Genetics Research*, vol. 92, no. 2, pp. 157-166.

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Early version, also known as pre-print

Published In:

Genetics Research

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Continuous Approximations for Optimising Allele Trajectories

by

A. Y. H. Liu and J. A. Woolliams

Roslin Institute, Midlothian EH25 9PS, United Kingdom

24 **Short Title:** Optimum Allele Trajectories

25 **Key Phrases:**

26 Continuous approximation, genotype assisted selection, GAS, QTL fixation,
27 frequency trajectory

28 **Proofs to be send to:**

29 Ariel Yi-Hsuan Liu

30 Roslin Institute & Royal (Dick) School of Veterinary Science

31 Roslin BioCentre, Roslin, Midlothian EH25 9PS, United Kingdom

32 E-mail: ariel.liu@roslin.ed.ac.uk

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47 **Summary**

48 The incorporation of genetic information such as quantitative trait loci (QTL) data into
49 breeding schemes has become feasible as DNA technologies have advanced. Such
50 strategies allow the frequency of desirable QTL to be controlled over a predefined time
51 frame, allowing the allele trajectory for QTL to be manipulated. A continuous
52 approximation to changes in allele frequency was developed to approximate the selection
53 procedure as a continuous rather than a discrete process, and analytical solutions were
54 obtained which shed light on how allele trajectories behave under different objective
55 functions. Three different objectives were considered: (1) minimising the total selection
56 intensity; (2) minimising the sum of squared selection intensities; and (3) equalising the
57 selection intensity applied over time. Simulations and genetic algorithms were performed
58 to test the accuracy and robustness of the continuous approximation. Theory shows that
59 firstly the total selection intensity required for moving an allele from a starting frequency
60 to another frequency point can be predicted independent of its trajectory, and secondly
61 that objective (2) and (3) are equivalent as the number of selection opportunities (T)
62 becomes large. The prediction of total selection intensity provides good fit for these two
63 objectives, with the accuracy of prediction improving as T increases. However, for (1) the
64 continuous approximation does not fit due to the existence of a discontinuous solution in
65 which the continuous approximation is applied before the frequency of selected allele
66 reaches 0.5 followed by rapid fixation.

67

68 1. INTRODUCTION

69 As identification of quantitative trait loci (QTL) becomes routine, genotype assisted
70 selection (GAS) has become possible and even desirable for populations with managed
71 breeding. GAS is where the frequency of a known allele, which affects the trait of
72 selection, is managed generation by generation within the population, often to fixation.
73 One of the known hurdles of the application of GAS is the Gibson effect, a phenomenon
74 whereby GAS results in higher short-term genetic gain but lower long-term genetic gain
75 than conventional selection methods which ignore the information on QTL (Gibson 1994).
76 The explanation is, although GAS can fix the QTL in shorter time, the loss of variation
77 on polygenes associated with the strong positive selection of QTL will lead to reduced
78 selection response on polygenes, which can not be fully recovered. Various authors have
79 shown that this effect can be ameliorated by optimising the selection procedures for the
80 QTL over multiple generations, i.e. the optimisation of allele trajectory (Dekkers & van
81 Arendonk 1998; Dekkers & Chakraborty 2001; Villanueva *et al.* 2002; Meuwissen &
82 Sonesson 2004; Villanueva *et al.* 2004; Sanchez *et al.* 2006).

83

84 This optimisation process has led to a variety of approaches to manage the trajectory:
85 maximising progress over the long term (Pong-Wong & Woolliams 1998; Villanueva *et*
86 *al.* 2004), with pre-defined time horizons (Dekkers & van Arendonk 1998); or
87 constrained to a constant rate of inbreeding (Villanueva *et al.* 2002). These studies make
88 selection decisions based on estimated breeding values in one form or another, so
89 optimum allele trajectory is therefore defined only implicitly. However the method of
90 Dekkers & van Arendonk (1998) allows the optimised trajectory to be defined explicitly

91 as a set of time points for the allele frequency, so defining the selection pressure that is
92 directly applied to the allele to be analysed. Based on an observation of Dekkers & van
93 Arendonk (1998), Meuwiseen & Sonesson (2004) directly defined the allele trajectory by
94 making selection intensity on the allele constant over the period of selection. More
95 recently, Sanchez et al (2006) pointed out that the effective population size was inversely
96 proportional to the square of selection intensity, so that the optimum trajectory to
97 minimise the accumulated inbreeding due to fixation should minimise the average
98 squared selection intensity on the major gene over generations up to the given fixation
99 time, in another words it should minimise the sum of squared selection intensities
100 (simplified as sum of squared intensities thereafter) applied to the allele over generations.

101

102 A common theme to these studies of allele trajectories is that they use discrete generation
103 models. This discrete time model imposes limitations on obtaining analytical solutions
104 for the problem of approximating the optimal pathway, and the lack of analytical solution
105 leaves unresolved the degree to which these approaches are distinct. Furthermore, the
106 iterative solutions from the equalising selection intensities method leave open questions
107 such as what is the total selection intensity (simplified as total intensity thereafter)
108 required to fix the QTL given the circumstances. Therefore, this study establishes a
109 continuous time model of the process of fixing an allele, and explicitly optimises the
110 trajectory with respect to various objective functions of the selection intensity applied to
111 the gene using the calculus of variations. This continuous model serves as a common
112 platform allowing further investigation and comparison between different optimising
113 objectives. The predictions from the continuous time model are compared to

114 optimisations using discrete generations to quantify the precision of the continuous time
115 model.

116

117 **2. METHOD**

118 (i) Theory

119 *Continuous Approximations*

120 Consider the process of moving a desired allele Q from frequency p_0 at time 0 to p_T at
121 time T in discrete generations and assume for simplicity of notation that there is only one
122 other allele, q, at that locus. The trajectory consists of the set of frequency points
123 $\{p_t, t = 0, \dots, T\}$ and optimisation of the trajectory is the set of p_t that maximise a certain
124 objective function. Commonly, when considering fixation of alleles in GAS $p_0 = (2N)^{-1}$
125 and $p_T = 1$, as it models the fixing of a new mutation occurring in a diploid population of
126 size N . This scenario is equivalent to the situation of eliminating a known allele from a
127 population ($0 < p_0 < 1$ and $p_T = 0$), as removing one allele forces the frequency of all
128 alternative alleles to 1. However the theory developed here will not be specific to these
129 starting and finishing frequencies.

130

131 In this paper, following Meuwissen & Sonesson (2004) and Sanchez et al (2006), the
132 objective functions considered are functions of the selection intensity applied directly to
133 the allele. Let $p_{t,k}$ be the frequency of Q allele of individual k born at time t , so $p_{t,k}$ will
134 take values 0, $\frac{1}{2}$ or 1 depending on whether k has genotype qq, qQ or QQ. Using $p_{t,k}$ as
135 the definition of an additive trait of selection, the population mean is p_t , the variance is
136 $\frac{1}{2} p_t(1 - p_t)$, the selection intensity i_t can be defined as:

137
$$i_t = (p_{t+1} - p_t) / \sqrt{\frac{1}{2} p_t (1 - p_t)} \quad (1)$$

138 for $t = 0, \dots, T - 1$, and the trajectory is a sequence of points $\{p_t, t = 0, \dots, T\}$.

139

140 The trajectory can be considered in continuous time rather than as a set of discrete
141 generations. It is an assumption, to be tested later, that the use of continuous time will
142 approximate the original problem better as the selection opportunities for changing allele
143 frequency become greater, i.e. when T is large. Let the trajectory over time be given
144 by $p(t)$, which is assumed to be a differentiable function of time t , then

145 $\delta p = p_{t+\delta t} - p_t \approx p'(t)\delta t$ when $p'(t) = \frac{dp}{dt}$, and $i_t \approx p'(t)\delta t / \sqrt{\frac{1}{2} p(t)(1 - p(t))}$. Therefore

146 provided the trajectory $p(t)$ is differentiable so that its derivative, $p'(t)$, exists, the sums
147 over of the trajectory may be approximated by integrals.

148

149 Different objective functions that optimise the trajectory are considered and analysed
150 using the continuous approximation, including: (1) the trajectory that minimises the total
151 intensity; (2) the trajectory that minimises the sum of squared intensities; and (3) the
152 trajectory that equalises selection intensity. Due to the amount of mathematical details
153 involved, only the essential information and core equations are shown in this section,
154 however, more details can be found in Appendix 1.

155

156 *Minimising the Total Intensity*

157 The total intensity for fixing an allele with a trajectory $p(t)$ as T becomes large can be
158 given by:

159
$$\sum_{t=0}^T i_t \approx \int_0^T \frac{p'(t)dt}{\sqrt{\frac{1}{2}p(t)(1-p(t))}} \quad (2)$$

160 Transformation and integration of the above equation gives the following:

161
$$\sqrt{2}(\sin^{-1}(1-2p_0) - \sin^{-1}(1-2p_T)) \quad (3)$$

162 Note that this solution only depends on the starting point p_0 and the ending point p_T ,
 163 suggesting that there is no such thing as minimising the total intensity if the
 164 approximation is valid – the total intensity is fixed between a pair of frequency points
 165 regardless of its trajectory or the value of T . For a new mutation moving to fixation,
 166 $p_0 = (2N)^{-1}$ and $p_T = 1$, the total intensity applied to the allele during fixation is
 167 $\sqrt{2} [\frac{1}{2}\pi + \sin^{-1}(1 - N^{-1})]$, which tends to $\sqrt{2}\pi$ as N becomes large, i.e. when starting
 168 frequency approaches zero.

169

170 *Minimising Sum of Squared Intensities*

171 The specific optimisation considered by Sanchez et al (2006) was the trajectory of the
 172 allele frequencies required to minimise the impact of the process on accumulated
 173 inbreeding during the fixation. It is assumed here that the accumulated inbreeding can be
 174 well-approximated from summed rates of inbreeding (ΔF) achieved in each generation,
 175 and that the allele can be fully identified throughout the process so its frequency can be
 176 explicitly managed over time. The value of ΔF will vary according to the impacts of all
 177 the different selection advantages inherent in a selection scheme, not only the carrier
 178 status of individuals for the allele of interest (Woolliams & Bijma 2000), and will depend
 179 upon the square of the selection intensities applied (Woolliams *et al.* 1993). Therefore the

180 objective for minimising the impact of the fixation is to minimise $\sum_{t=0}^T i_t^2$, which can be
 181 shown as follows:

$$182 \quad \sum_{t=0}^T i_t^2 \approx \int_0^T \frac{p'(t)^2}{\frac{1}{2}p(t)(1-p(t))} dt \quad (4)$$

183 Solving the above equation gives $\sin^{-1}(1-2p) = At + B$ or, equivalently as:

$$184 \quad p(t) = \frac{1}{2}[1 - \sin(At + B)] \quad (5)$$

185 , where A and B are constants of integration and vary depending on p_0 , p_T , and T . Values
 186 of A and B can be obtained by substituting these parameters into Equation (5). For
 187 example, assume that fixation is desired from a new mutation, i.e., $p_0 = (2N)^{-1} \approx 0$ for
 188 large N , and $p_T = 1$. With these conditions $B = \pi/2$ and $A = -\pi/T$ to give
 189 $p(t) = \frac{1}{2}(1 - \sin[\frac{1}{2}\pi(1 - 2tT^{-1})])$. The optimal trajectory for minimising the sum of
 190 squared intensities applied to the allele is therefore a segment of a sine wave.

191

192 *Equalising Selection Intensities*

193 Based on the observation from Dekkers & van Arendonk (1998) that the selection
 194 intensities achieved in each generation are roughly constant in their simulated result with
 195 best long term gain, Meuwissen & Sonesson (2004) suggest optimising the trajectory to
 196 maximise the cumulative selection response is by making the selection intensities
 197 constant over time. Applying this objective in the continuous approximation gives a
 198 differential equation that is identical to that obtained above for the objective of
 199 minimising the sum of squared intensities. This indicates that the objective from
 200 Meuwissen & Sonesson gives an optimum trajectory identical to that from minimising
 201 the sum of squared intensities. This conclusion is analogous to the minimisation of sum

202 of squares for n numbers whose sum is fixed to some value c – the solution has all
203 numbers equal to c/n . Therefore the theory suggests that as the continuous approximation
204 provides becomes more apt, so the distinction between the objectives of Sanchez et al
205 (2006) and Meuwissen & Sonesson (2004) disappears. The question remains over how
206 close an approximation.

207

208 (ii) Simulation Methods

209 Two types of simulation methods are included in this section, the first a genetic algorithm
210 with small population size ($N=10$), and the other a simulation of breeding populations
211 with large population size ($N=500$). Together they test the validity and robustness of the
212 continuous approximation under various scenarios.

213

214 *Genetic Algorithm*

215 The genetic algorithm used differential evolution (Shepherd & Kinghorn 1992) to
216 optimise the allele frequency in order to find the optimal trajectories with $N=10$, for the
217 three objectives considered above: (i) equalising selection intensities; (ii) minimising sum
218 of squared intensities; and (iii) minimising total intensity. Equalising the selection
219 intensities was achieved by minimising the sum of all squared differences among the
220 selection intensities.

221

222 *Simulations of Breeding Schemes*

223 Computer simulations of the breeding schemes start with a base population ($t=0$) of 500
224 diploid individuals and this population size was maintained through out the simulation.

225 One individual from the base population was randomly chosen to carry a single copy of
226 positive allele (initial frequency $p_0 = (2N)^{-1}$) with allelic effect a , which equals 0.5 as
227 the addition or removal of one positive allele result a change of 0.5 in terms of frequency.
228 Random mating with possible selfing was assumed for simplicity, i.e. the genetic make-
229 up of the offspring was randomly assigned from selected parents with replacement. As
230 the theory shows that the objective of minimising sum of squared intensities resembles
231 the objective of equalising selection intensities when T is large, only the objective of
232 equalising selection intensities is used for its ease to execute. In addition, other selection
233 strategies with oscillating intensities in a saw tooth pattern, i.e. intensity profiles of the
234 form $\{0.3, 0.1, 0.3, 0.1\dots\}$, were also employed to test whether the continuous
235 approximation still holds under more extreme conditions.

236

237 One should note that the time unit applied in this study was the opportunities for selection
238 and mating. Hence the word cohort will be used hereafter to represent a group of animals
239 which are the direct result of last selection and mating. The frequency of the positive
240 allele was then calculated and recorded for each cohort, and simulation ended when the
241 positive allele was either fixed ($p_t \geq (2N-1)/2N$) or lost ($p_t \leq (2N)^{-1}$). In the case of the
242 allele being lost, the data was excluded from the final data set as we considered the
243 pathway of allele fixation only. One thousand simulations were run for each set of
244 parameters and the average number of cohorts required to fix selected allele was obtained
245 to be compared to the expected number of cohorts required from the approximation.

246

247 *Discrete generation:* A pre-defined constant selection intensity (i) was applied over every
248 cohort by restricting the average frequency of the selected individuals. Calculation was
249 then performed for each cohort to obtain the target p_{t+1} from the p_t :

$$250 \quad p_{t+1} = p_t + i\sqrt{\frac{1}{2}p_t(1-p_t)} \quad (6)$$

251 Selection candidates were composed of all individuals from the current cohort and were
252 ranked according to their allelic value. Selection candidates were then removed
253 sequentially from lower rank until the target p_{t+1} was achieved. However, as mating
254 between selected parent are random, the average allele frequency in the resultant
255 population could not be guaranteed and may deviate from the target p_{t+1} . For oscillating
256 intensities, a similar procedure as described above was used, except that the intensity is
257 not constant over every cohort.

258

259 *Overlapping generation:* The overlapping generation model was largely identical to the
260 discrete generation model except that the candidates available for selection were not only
261 restricted to the current cohort, but also extended to include 2 previous cohorts. For
262 selection candidates with same allelic value, a randomisation process was used to
263 determine which candidate would become a parent. Generation interval (L) was
264 calculated as the age of parents (in units of cohorts) when the offspring born.

265

266 When several cohorts contribute to the selection, the genetic variance is higher than
267 shown in Equation (6), i.e. $\frac{1}{2}p_t(1-p_t)$. Apart from the variance within all selected
268 cohorts, the true genetic variance also contains an additional term for the variance
269 between different cohorts:

270
$$V_{total} = \frac{1}{2}E[p](1 - E[p]) + \frac{1}{2}(E[p^2] - E[p]^2) \quad (7)$$

271 where $E[p]$ denotes expectations over the selected cohorts. Simulations were carried out
272 using Equation (6) with V_{total} replacing $\frac{1}{2}p_t(1 - p_t)$. This was compared to using Equation
273 (6) without modification.

274

275 **3. RESULTS**

276 As shown in the theory, the continuous approximation provides a prediction for the total
277 intensity required to move a target allele from a specific frequency to another. The
278 prediction is only affected by the starting and ending frequencies alone, and is
279 independent of T or N , although in the case of new mutation, the starting frequency is
280 inversely related to the population size. Assuming fixation is the goal (i.e. ending
281 frequency is 1), the predicted total intensity is $\sqrt{2\pi}$ (≈ 4.44) for fixing a mutation in a
282 large population, and 3.80 for a starting frequencies of 0.05.

283

284 (i) Goodness of fit for small T , using genetic algorithm

285 Table 1 summarises and compares the results obtained from different GA evolutions and
286 the continuous approximation for $N = 10$, i.e. $p_0 = 0.05$, with small T values up to 11. For
287 these parameters the predicted total intensity from continuous approximation is 3.80
288 regardless of trajectory, in the other words, regardless of the objective functions of the
289 GA evolution. When equalising intensities across generations, the precision of predicting
290 total intensity was very good initially with an error of 1.7% at $T = 2$, deteriorating as T
291 increases, and then improving again, with the greatest error of predicting the total
292 intensity being 9.2% at $T = 5$. The continuous approximation introduces marginally

293 greater errors to the predicted total intensity when minimising sum of squared intensities,
294 with errors peaking at 12.3% for $T = 5$ and reducing to 10.4% for $T = 11$. Note the similar
295 trend on the goodness of fit of the continuous approximation varies with T for both
296 objectives. A very different trend was observed for the objective of minimising total
297 intensity, with total intensity continuing to reduce with T to 3.06 at $T = 11$ which is very
298 different from the prediction of 3.80. Reasons leading to this observation will be
299 explained in the discussion section.

300

301 Looking at the profile of these different GA evolutions reveals more detail about them.
302 The intensity profile of equalising intensities objective is quite similar to minimising sum
303 of squared intensities objective with their intensity achieved each generation became
304 more and more uniform over time (Figure 1a and 1b). This illustrates the derivation
305 showing that the solutions for the two objectives converge given the validity of the
306 continuous approximation.

307

308 Assuming the convergence of objectives of equalising intensities and minimising sum of
309 squared intensities, the minimum sum of squared intensities predicted from the
310 continuous approximation is equal to $3.80^2/T$ since $(\frac{i}{T})^2 T = \frac{i^2}{T}$. Figure 1a shows that as
311 T increases up to 11, the selection intensities become much more uniform, although Table
312 1 shows the prediction of minimum sum of squared intensities still has significant error at
313 $T = 11$ despite that the magnitude of the error is reducing. It might be expected that
314 minimising the sum of squared intensities will have approximately twice the error of
315 minimising total intensity (see Appendix 2).

316

317 (ii) Goodness of fit for large T , using simulations

318 The simulations allowed the goodness of fit to be tested for large T by varying the
319 selection intensity applied. For the results presented in this section, p_0 is 0.001 ($N=500$),
320 with the predicted total intensity being 4.35 from the continuous approximation.

321

322 *Discrete Generation with Constant Selection Intensity*

323 The comparisons between simulation of breeding with discrete generation and the
324 continuous approximation for a range of different but constant selection intensities
325 applied are shown in Figure 2. The results are presented as the mean number of cohorts
326 required to fixation with the expected number of cohort being calculated by dividing the
327 expected total intensity with the constant intensity applied during the simulation. Figure 2
328 shows that for constant intensities > 0.5 , where $T < 10$, the scale of errors agrees with the
329 result shown in Table 1. However, the simulations show that the approximation fits the
330 results progressively more closely for all intensities < 0.5 . For all intensities < 0.75 , the
331 differences between prediction and actual results are less than one cohort.

332

333 *Discrete Generation with Oscillating Selection Intensity*

334 The independence of the total intensity applied to trajectory was further tested by
335 oscillating selection intensities across cohorts as in a saw tooth pattern. Table 2 shows
336 comparison of total intensity applied for oscillating selection intensities patterns
337 compared to constant selection intensity with same pair-wise average. Results show that
338 the prediction errors are only slightly larger for oscillating selection intensities compared

339 to constant selection intensities with comparable average selection intensity. The
340 approximation still provides good prediction under such conditions, with errors around
341 2.3% for oscillating selection intensities $\{0.3, 0.1\}$ and increased to 11.7% for selection
342 intensities $\{0.6, 0.4\}$. The increase in error with higher selection intensities and lower
343 fixation times would be expected from the result of constant selection intensities. There
344 were only small differences between complementary patterns, i.e. $\{0.3, 0.1\}$ compared to
345 $\{0.1, 0.3\}$ (results not shown).

346

347 In all breeding simulations, the prediction often appears as an under-estimation of the
348 simulated result, which is unsurprising because that the selection intensity applied in the
349 simulation could not always be achieved, i.e. in the last few cohorts the target p could
350 exceed 1.0 in order to achieve the selection intensity applied – which is not possible. This
351 is particularly important for large i selection, when only small selection intensity might
352 have been required to move the frequency to 1.

353

354 *Overlapping Generation*

355 Table 3 summaries the results for simulations with overlapping generations. It shows total
356 intensity required for fixation is predictable from the continuous approximation for low
357 selection intensities but the % errors increase as selection intensity applied per cohort
358 increases. When the unmodified Equation (6) was used, the result is almost identical to
359 those shown in Figure 2. However the use of V_{total} , which represents the full genetic
360 variance, introduces an additional error. The 8.4% error for intensity of 0.5 represents

361 approximately 1 cohort difference between predicted and observed time to fixation. In
362 this case the mean actual number of cohorts was 9.4.

363

364 **4. DISCUSSION**

365 The theory developed in this paper shows that providing the continuous approximation is
366 valid, then the total intensity applied to move between two frequencies is directly
367 proportional to the difference between the arcsines of $(1-2p)$ for the end points p_0 and p_T
368 irrespective of trajectory – including standard logistic trajectories, $dp/dt = sp(1-p)$.
369 For fixation of a rare mutant, a frequent subject of interest, as p_0 tends to 0 and p_T tends to
370 1, the total intensity tends to $\sqrt{2}\pi$. Further the strategies of (i) equalising selection
371 intensities throughout the trajectory (Sonesson & Meuwissen, 2004), and (ii) minimising
372 the sum of squared intensities (Sanchez et al, 2006) converge to the same optimal
373 trajectory which is a function of time described by a segment of a sine wave. The results
374 showed that the goodness of fit of the continuous approximation became progressively
375 better as T increased, with prediction errors for total intensity reducing becoming
376 reasonable as $T \sim 10$, or average $i \sim 0.4$ during the period. Further this result remained
377 true for trajectories in which i was varied over time rather than constant, or where
378 generations were overlapping rather than discrete.

379

380 The continuous approximation will have a lack of fit for two reasons. First, a smooth
381 curve is used to approximate a step function; second, the dominator for $p_{t+1} - p_t$ in \dot{i}_t is
382 related to $p_t(1-p_t)$, not $p_{t+\frac{1}{2}}(1-p_{t+\frac{1}{2}})$ which would be more natural for the use of the
383 continuous approximation. This affects the goodness of fit under positive selection since

384 i_t is greater than expected from approximation by $p'(t + \frac{1}{2})$ when $p(t) < p(t + \frac{1}{2}) < 0.5$,
385 but less than the approximation when $p(t + \frac{1}{2}) > p(t) > 0.5$. The sizes of error are
386 comparable for the pair of $p(t)$ that are in equal deviation from 0.5. These trends are most
387 extreme for p close to 0 or 1, or for small T when $p(t)$ changes rapidly, and there are
388 greater opportunities for cancelling when trajectories move from $p < 0.5$ to $p > 0.5$.

389

390 The difference in the sign of errors when p is greater than or less than 0.5 helps to explain
391 the results found for minimising the total intensity, since for all T a trajectory with total
392 intensity less than predicted by the continuous approximation can be found (Table 1).
393 Figure 3 shows the trajectories that minimise total intensity shown in Table 1, and it is
394 seen the trajectory resembles a continuous curve for $p < 0.5$ with a jump in the final
395 generation from close to 0.5 directly to 1. As T increases this represents a discontinuity in
396 the trajectory, which can be seen as a combination of the continuous approximation from
397 p_0 (assumed < 0.5) to 0.5 and a direct jump from 0.5 to 1, For $T = 11$ used in Table 1, the
398 expected value from the discontinuous solution is 3.02 (c.f. 3.06) affirming the
399 continuous approximation can fit well to intervals that do not span both sides of 0.5.

400 .

401 The existence of the discontinuous solution for minimising the total intensity creates a
402 distinction between minimising sum of squared intensities and equalising selection
403 intensities. The sum of squared intensities can be broken down into two components: the
404 sum of selection intensity and the variance of selection intensity:

405
$$\sum_{t=1}^T i_t^2 = T E[i_t]^2 + T Var(i_t) = T^{-1}(\sum_{t=1}^T i_t)^2 + T Var(i_t)$$

406 The strategy of equalising selection intensities promotes reduction in the sum of squared
407 intensities by having no variance term, while the discontinuous solution is effective
408 through reducing total intensity. For small T , the trajectory minimising the sum of
409 squared intensities is temporarily effective in reducing the sum of squared intensities by
410 reducing the total intensity acquired and therefore allowing some variance. However, as T
411 increases the benefits from reducing total intensity become less than the penalty from the
412 variance among the selection intensities, and the optimum trajectory moves towards the
413 trajectory of equalising intensities (see Appendix 4).

414

415 Genomics is at the start of giving values to many small segments of chromosomes,
416 sometimes with QTL identified, and sometimes simply marked. Simultaneously we are
417 also at the threshold of being able to manage inbreeding at the level of the segment, i.e.
418 requiring slow change in diversity, or wishing to reduce the impact of negative LD on
419 what segments can be fixed in the population. Therefore we envisage the field of
420 “designer genomes” where the target trajectories of multiple loci are mapped out on a
421 genome-wide scale. This is not a problem with only one locus. However, to achieve
422 targets on frequency and inbreeding at multiple loci we need to understand in the long
423 term what is required to fix/eradicate an allele or to move from a frequency point to
424 another, and hence consider how closely the designed genome can be achieved. It is
425 precisely this approximation that allows such predictions over time to be made in a
426 simple fashion albeit that it is but one step towards achieving the wider goal.

427

428 One of the possible uses of this approximation is on the removal of the recessive mutant
429 allele that causes foal immunodeficiency syndrome (FIS), more commonly known as the
430 Fell pony syndrome. This fatal condition affects not only Fell ponies but also Dales
431 ponies, and the causal mutant has recently been identified (personal communication: June
432 Swinburne). Although the eradication of this mutant allele is highly desirable, two
433 reasons makes the execution difficult: first, the frequency of carrier is high within the
434 population (~ 0.4 in the Fell breed, personal communication: June Swinburne), and
435 second, the Fell breed is a small breed. In the other word, this allele is wide-spread in a
436 small gene pool; hence options such as culling of all carriers are not sensible as they
437 might lead to the loss of genetic diversity and the emergence of new recessives. Therefore
438 it is necessary to plan the removal of this mutant allele over a prior time scale to minimise
439 the impact on diversity. Theoretically the process of eradication should be carried out
440 slowly and carefully in order to minimise the reduction on genetic diversity within the
441 breeds. The approximation in this study can provide a simple means of getting a series of
442 stage goals for moving the frequency to zero, i.e. target frequency points, to be achieved
443 over the pre-determined horizon whilst minimising the diversity loss. With the mutant
444 allele frequency ~ 0.25 , the total intensity required to remove the mutant allele is ~ 1.48 ,
445 and the intensity in each generation is $1.48/T$.

446

447 Aspects of the results may be generalised to more than one QTL, and there is a synergy
448 with the results of Goddard (2009), where trajectories for two QTL are optimised with
449 respect to a profit function. The study of Goddard recognises allele frequencies do not
450 change linearly with the selection intensity applied, and uses a transformation to a scale

451 (denoted z in the paper) upon which linearity holds – this requires the continuous
452 approximation to hold since derivatives are required. Appendix 3 shows that the scale,
453 $z(p)$, can be interpreted as being directly proportional to accumulated selection intensity
454 applied to the locus for moving from an infinitesimally small frequency to p . This study
455 shows that to move m loci from \mathbf{p}_0 to \mathbf{p}_T whilst minimising inbreeding at a neutral locus
456 (and one that is affected by selection through the development of the pedigree only)
457 constant selection intensity is required to be simultaneously applied at each locus – albeit
458 with intensity differing among loci. This trajectory is represented by straight lines in an
459 m -dimensional z -space with the relative strength of selection on each locus determining
460 direction, and the line is traversed in T segments of equal length. However the actual
461 inbreeding accumulated will depend on T , the size of the population, and also upon the
462 linkage disequilibrium among the loci being selected.

463

464 In conclusion, the continuous approximation shows that: (i) the optimising approaches of
465 equalising intensities (Meuwissen & Sonesson, 2004) and minimising sum of squared
466 intensities (Sanchez et al., 2006) have the same limiting form and converge over time to a
467 sine wave; and (ii) the total intensity required to move an allele from a given frequency
468 point to another can be very closely approximated and only depends on the starting and
469 end frequency.

470

471 **Acknowledgements**

472 The BBSRC are gratefully acknowledged for funding.

473

474
475

References

- 476 Dekkers, J. C. M., and R. Chakraborty, 2001 Potential gain from optimizing
477 multigeneration selection on an identified quantitative trait locus. *Journal of*
478 *Animal Science* **79**: 2975-2990.
- 479 Dekkers, J. C. M., and J. A. M. van Arendonk, 1998 Optimizing selection for quantitative
480 traits with information on an identified locus in outbred populations. *Genetical*
481 *Research* **71**: 257-275.
- 482 Gibson J.P. Short term gain at the expense of long term response with selection on
483 identified loci. Proceedings of the Fifth World Congress on Genetics Applied to
484 Livestock Production , 201-204. 1994. 7-8-1994.
485 Ref Type: Conference Proceeding
- 486 Meuwissen, T. H. E., and A. K. Sonesson, 2004 Genotype-assisted optimum contribution
487 selection to maximize selection response over a specified time period. *Genetical*
488 *Research* **84**: 109-116.
- 489 Pong-Wong, R., and J. A. Woolliams, 1998 Response to mass selection when an
490 identified major gene is segregating. *Genetics Selection Evolution* **30**: 313-337.

- 491 Sanchez, L., A. Caballero, and E. Santiago, 2006 Palliating the impact of fixation of a
492 major gene on the genetic variation of artificially selected polygenes. *Genetical*
493 *Research* **88**: 105-118.
- 494 Shepherd, R. K., and B. P. Kinghorn, 1992 Optimizing Multitier Open Nucleus Breeding
495 Schemes. *Theoretical and Applied Genetics* **85**: 372-378.
- 496 Villanueva B., Dekkers J.C.M., Woolliams J.A. & Settar P. Maximising genetic gain with
497 QTL information and control of inbreeding. Proceedings of the Seventh World
498 Congress on Genetics Applied to Livestock Production . 2002. 19-8-2002.
499 Ref Type: Conference Proceeding
- 500 Villanueva, B., J. C. M. Dekkers, J. A. Woolliams, and P. Settar, 2004 Maximizing
501 genetic gain over multiple generations with quantitative trait locus selection and
502 control of inbreeding. *Journal of Animal Science* **82**: 1305-1314.
- 503 Weisstein E.W. Euler-Lagrange differential equation.
504 <http://mathworld.wolfram.com/euler-lagrangedifferentialequation.html> . 2005. 4-
505 10-2006.
506 Ref Type: Electronic Citation
- 507 Woolliams, J. A., and P. Bijma, 2000 Predicting rates of inbreeding: in populations
508 undergoing selection. *Genetics* **154**: 1851-1864.

509 Woolliams, J. A., N. R. Wray, and R. Thompson, 1993 Prediction of Long-Term
510 Contributions and Inbreeding in Populations Undergoing Mass Selection.
511 *Genetical Research* **62**: 231-242.
512
513
514

515 Table 1. The total intensity ($Sum i_t$) and sum of squared intensities ($Sum i_t^2$) required for
 516 $N=10$ and a range of T values, using three optimisation strategies: 1) equalising selection
 517 intensities across generations, 2) minimising $Sum i_t^2$, and 3) minimising $Sum i_t$, and 4)
 518 calculated from the continuous approximation.

Strategy	Criterion	$T=2$	% error	$T=5$	% error	$T=8$	% error	$T=11$	% error
Equalise i_t	$Sum i_t$	3.869	-1.7	3.456	9.2	3.474	8.7	3.509	7.8
	$Sum i_t^2$	7.486	-3.4	2.390	17.5	1.509	16.6	1.119	15.0
Minimise $Sum i_t^2$	$Sum i_t$	3.790	0.4	3.338	12.3	3.364	11.6	3.411	10.4
	$Sum i_t^2$	7.296	-0.8	2.300	20.6	1.460	19.3	1.088	17.3
Minimise $Sum i_t$	$Sum i_t$	3.748	1.5	3.166	16.8	3.088	18.8	3.059	19.6
	$Sum i_t^2$	7.670	-6.0	3.154	-8.9	2.604	-43.9	2.404	-82.7
Prediction	$Sum i_t$	3.805		3.805		3.805		3.805	
	$Sum i_t^2$	7.239		2.896		1.810		1.316	

519

1 Table 2. Comparison between the total intensity ($Sum i_t$) required to fix an allele under
 2 simulations with discrete generations and predicted from continuous approximation for a
 3 range of different selection intensities. The selection intensity can be either constant all
 4 through the simulation or oscillating between a pair of different values (shown as {a, b}).
 5 Population size (N) equals 500 in all cases.

6

Selection Intensity	0.2	{0.3,0.1}	0.3	{0.4,0.2}	0.5	{0.6,0.4}
Predicted $Sum i_t$	4.35	4.35	4.35	4.35	4.35	4.35
Simulated $Sum i_t$	4.40	4.45	4.50	4.55	4.68	4.86
% error	1.1	2.3	3.4	4.6	7.6	11.7

7

8

9

10

11

12

13

14

15

16

17

18

19

1 Table 3. A comparison between the total intensity ($Sum i_t$) required to fix an allele in
 2 simulations with overlapping generations for different constant selection intensities
 3 applied and for genetic variance calculated by different methods. In “Unmodified”
 4 Equation (6) was used directly, but in “Modified” the true genetic variance V_{total} replaced
 5 $\frac{1}{2}p_t(1 - p_t)$ in Equation (6). In all cases population size (N) equals 500 and predicted Sum
 6 $i_t = 4.35$. Standard errors, % error in prediction, and generation interval (L) are also
 7 shown.

8

	Selection intensity/cohort = 0.2		Selection intensity/cohort = 0.5	
	<i>Unmodified</i>	<i>Modified</i>	<i>Unmodified</i>	<i>Modified</i>
$Sum i_t$	4.33 ± 0.012	4.30 ± 0.006	4.51 ± 0.016	4.72 ± 0.012
% error	-0.5	-1.1	3.7	8.4
L	2.29	2.32	2.06	2.21

9

10

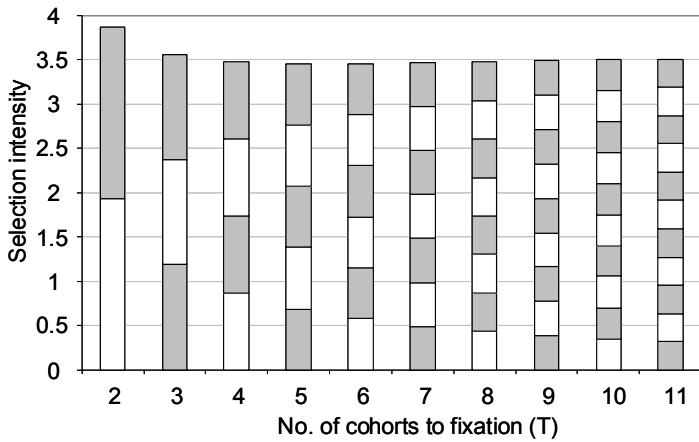
1 Figure 1. The composition of total intensity obtained from GA with the objective of (a)
 2 minimising sum of squared intensities and (b) equalising selection intensities. Each block
 3 represents the amount of selection intensity achieved in a single mating/frequency change.
 4 Shading is for the purpose of illustration only.

5 (a)



6

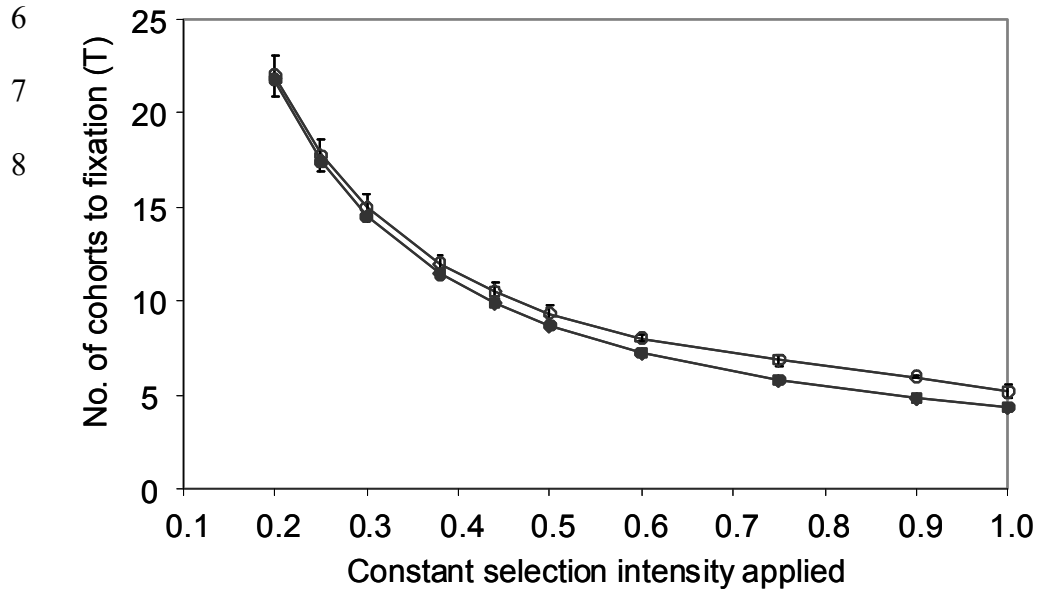
7 (b)



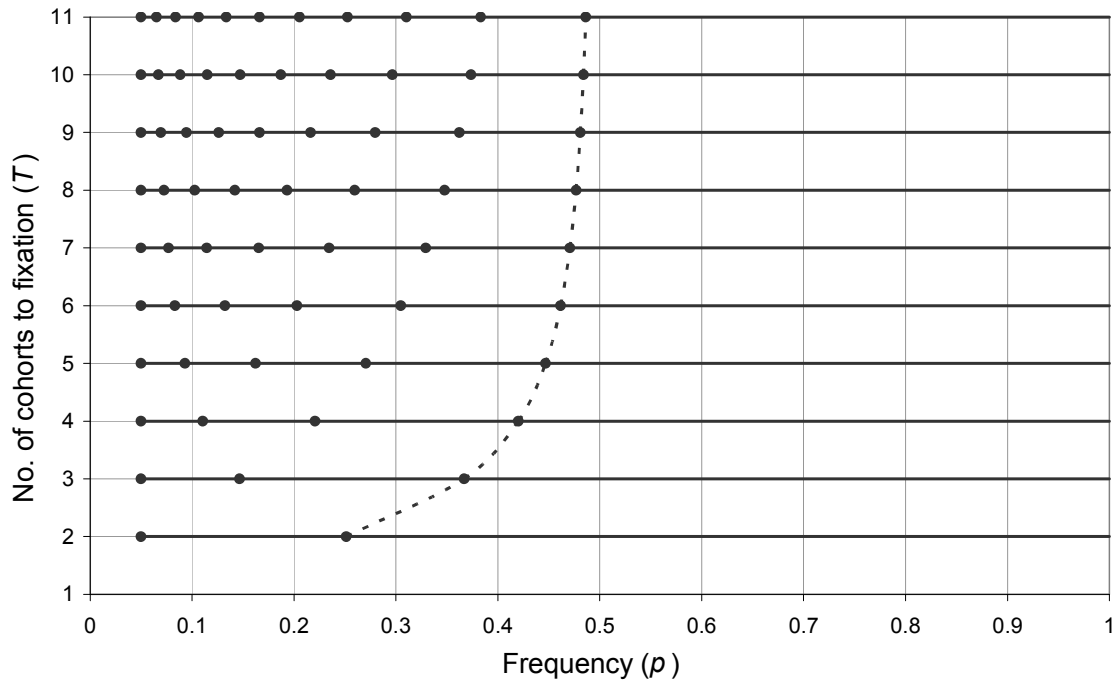
8

9

1 Figure 2. Comparison between the numbers of cohorts required to fix an allele for a range
2 of different selection intensities, for (a) simulations with discrete generation (open circle)
3 and (b) continuous approximation (filled circle). Population size (N) equals 500 in all
4 cases. The standard deviations are shown as error bars and the standard errors are
5 negligible.



1 Figure 3. The frequency path (trajectory) obtained by GA with the objective of
 2 minimising total intensity for different T values. For each profile with different T , the
 3 frequency points are shown as the solid circles along the horizontal line, with the first
 4 frequency point being at $p=0.05$. The last frequency points, p_{T-1} , from all profiles are
 5 joined by dashed line to illustrate how p_{T-1} approaches 0.5 as T increases.



6
 7
 8
 9
 10
 11
 12
 13
 14

1 **Appendix 1**

2 *Minimising the Total Intensity*

3 By noting that $p'(t)dt$ can be replaced by dp , and then substituting p for $p(t)$, Equation (2)
4 can be transformed to the following equation, where its direct integration leads to
5 Equation (3):

$$\sum_0^T i_t \approx \int_{p_0}^{p_T} \frac{dp}{\sqrt{\frac{1}{2}p(1-p)}} \tag{A1}$$

7
8 *Minimising Properties of Allele Trajectories*

9 An important methodology for optimising trajectories is the calculus of variations
10 (Weisstein 2005). When the function to be optimised is of the form:

$$\int_0^T f[p, p', t] dt \tag{A2}$$

11
12 then the solution can be obtained from the Euler-Lagrange equations providing
13 trajectories $p(t)$ are differentiable. This equation states that the optimum trajectory
14 satisfies:

$$\partial f / \partial p - d[\partial f / \partial p'] / dt = 0$$

15
16 This solution can be further simplified if $f[p, p', t]$ is independent of explicit dependence
17 on t , i.e. the partial derivative of f with respect to t is 0 (i.e. $\delta f / \delta t = 0$), then the
18 condition may be simplified to the Beltrami identity: $f[p, p', t] - p' \delta f / \delta p' = C$, where C
19 is a constant of integration.

20

1 *Minimising Sum of Squared Intensity:* The function $f[p]$ required to minimise the sum of
 2 squared intensity is as follows:

$$3 \int_0^T f[p, p', t] dt = \int_0^T \frac{p'^2}{0.5p(1-p)} dt \quad (\text{A3})$$

4 where p' is the derivative of p with respect to t , and $f[p, p', t] = p'^2[0.5p(1-p)]^{-1}$,
 5 representing the square of the selection intensity at time t . Applying the method of
 6 calculus of variation (Weisstein 2005) to the sum of squared intensities gives the
 7 following result: $f[p, p', t] - p' \delta f / \delta p' = -p'^2[0.5p(1-p)]^{-1} = C$ and p must satisfy:

$$8 p' = [0.5Cp(1-p)]^{\frac{1}{2}} \quad (\text{A4})$$

9 Solving this differential equation gives $\sin^{-1}(1-2p) = At + B$ or, equivalently,
 10 $p(t) = \frac{1}{2}[1 - \sin(At + B)]$, where A and B are constants of integration. This comes from
 11 noting that $[p(1-p)]^{\frac{1}{2}} = \frac{1}{2}[1 - u^2]^{\frac{1}{2}}$, where $u = (1-2p)$ to convert the function into a
 12 recognizable standard integral form. A and B are determined by the desired change from
 13 $t = 0, \dots, T$ and have units of radians (not degrees). The optimal trajectory for
 14 minimising the sum of squared intensity applied to the allele is therefore a segment of a
 15 sine wave.

16

17 *Equalising Selection Intensities:* The objective function of equalising the selection
 18 intensities is equivalent to making the selection intensity constant i.e.
 19 $p'[0.5p(1-p)]^{-\frac{1}{2}} = C$, which is the same differential equation as that obtained above for
 20 the criterion of minimising sum of squared intensities (Equation A4).

21

22 **Appendix 2**

1 Because the error associated with minimising sum of squared intensity (as a percentage to
 2 the total), can be simplified as: $\frac{(i_t + \delta(i_t))^2 - i_t^2}{i_t^2} \approx \frac{2i_t \delta(i_t)}{i_t^2} = \frac{2\delta(i_t)}{i_t}$ given $(\delta(i_t))^2$ can
 3 be neglected; which is twice the error of minimising total intensity ($\frac{\delta(i_t)}{i_t}$).

4

5 **Appendix 3**

6 This study considers the total intensity required to move from p_0 to p_T , as:

$$7 \int_{p_0}^{p_T} \frac{dp}{\sqrt{0.5p(1-p)}}$$

8 Note that in Goddard (2009) $z(p) = \sin^{-1}(\sqrt{p}) = \pi/4 - \frac{1}{2} \sin^{-1}(1-2p)$ and that

$$9 \int_{p_0}^{p_T} \frac{dp}{\sqrt{0.5p(1-p)}} = \int_{z_0}^{z_T} \frac{dp}{dz} \frac{dz}{\sqrt{0.5p(1-p)}} = \int_{z_0}^{z_T} dz = z_T - z_0$$

10 Therefore the increment in z is the accumulated selection intensity applied to the locus.

11

12 **Appendix 4**

13 For large N , the total intensity for the continuous solution $\approx \sqrt{2}\pi$, while the total intensity
 14 for the discontinuous solution approaches $\sqrt{2}(1 + \pi/2)$. This is obtained by calculating
 15 separately the intensity from p_0 (assumed < 0.5) to 0.5 and the intensity from 0.5 to 1.
 16 The first of these is $\sqrt{2} \sin^{-1}(1 - N^{-1})$, which tends to $\pi/\sqrt{2}$ as N becomes large,
 17 whilst the second is $\sqrt{2}$, giving a minimum of $\sqrt{2}(1 + \pi/2)$ for large N .

18

19 Hence, for continuous solution the sum of squared intensities $\sum i^2 = T(\frac{\sqrt{2}\pi}{T})^2 = \frac{2\pi^2}{T}$ and
 20 for discontinuous solution $\sum i^2 = (\sqrt{2})^2 + (T-1)(\frac{\pi}{\sqrt{2}(T-1)})^2 = 2 + \frac{\pi^2}{2(T-1)}$.

21

22 The sum of squared intensities from the two solutions for a range of T values are
 23 summarised below. When $T = 7$, the two solutions yield roughly equal results, and for $T >$
 24 7, the continuous solution performs better than the discontinuous solution.

25

26

27

<i>e</i>	Continuous	Discontinuous
2	9.87	6.93
3	6.58	4.47
4	4.93	3.64
5	3.95	3.23
6	3.29	2.99
7	2.82	2.82
8	2.47	2.70
9	2.19	2.62
10	1.97	2.55

1