A method to infer positive selection from marker dynamics in an asexual population
Christopher J. R. Illingworth and Ville Mustonen

SUPPLEMENTARY INFORMATION
Qualitative guide to the method

System We consider a population of cells which is divided by neutral markers into two subpopulations, which we term the ‘blue’ and ‘red’ marker populations. Over time, we make repeated observations of the system, counting the number of cells in the blue and red marker populations, and use these to generate time-resolved marker frequencies $\tilde{x}_{\text{red}}(t_k)$ and $\tilde{x}_{\text{blue}}(t_k)$, where the $t_k$ represent observation times. We construct a model to generate predicted marker frequencies, $x_{\text{red}}(t_k)$ and $x_{\text{blue}}(t_k)$, to fit the observations.

We model the initial system using two haplotypes, the haplotype $a = 0$ denoting the blue marker population, and the haplotype $a = 1$ denoting the red marker population. We consider a system in which there arises one beneficial mutation, with mutant haplotype $a = 2$.

Initial haplotypes Before the arrival of the beneficial mutation, the only haplotypes in the system are $a = 0$ and $a = 1$. In the language of the main text, these haplotypes comprise the set $H_0$. As described in Eq. 2 of the main text, these haplotypes grow proportional to $q_0(0)e^{f_0 t}$ and $q_1(0)e^{f_1 t}$. As the difference between these haplotypes is only a neutral marker, however, these haplotypes have equal fitness, so that

$$f_1 = f_0$$

As all that matters in the relative sizes of the marker populations is the difference in fitness, we can set

$$f_1 = f_0 = 0$$

Further, we know that the frequencies sum to one

$$q_0(0) + q_1(0) = 1$$

Together these equations give

$$q_0(t) = q_0(0)$$
$$q_1(t) = 1 - q_0(t)$$

These values are constant; until the emergence of the first mutation, the haplotype frequencies do not change. As the model marker frequencies are calculated by adding up the frequencies of the haplotypes with each marker, the model marker frequencies, too, do not change.

Mutant haplotype We now suppose that the mutant, represented by the haplotype $a = 2$, reaches a frequency of 0.001 at time $t_2$. At the time $t_2$, then, we set the haplotype frequency $q_2(t_2) = 0.001$. Other haplotype frequencies are rescaled at this time in a linear fashion. If $q_0(t_2')$ and $q_1(t_2')$ represent the frequencies of the haplotypes $a = 0$ and $a = 1$ at a moment instantaneously before this time, the rescaled haplotype frequencies can be expressed as

$$q_0(t_2) = (1 - 0.001)q_0(t_2')$$
$$q_1(t_2) = (1 - 0.001)q_1(t_2')$$

Supposing the mutant haplotype has fitness $f_2 > 0$, assuming its evolution to be deterministic, then from the time $t_2$, it will grow in size proportional to $e^{f_2 t}$. The frequency of the mutant haplotype, then, will grow proportional to $e^{f_2(t-t_2)}$, for $t > t_2$ (See Eq. 2 of the main text for comparison). As a result, if the mutant haplotype occurs in the red marker population, the red marker frequency will then increase in size, while if the mutant haplotype occurs in the blue marker population, the blue marker frequency will increase in size. The fitness $f_2$ determines the rate of increase, larger values of $f_2$ leading to faster increases in marker population frequency.

Concluding remarks In the above, we have described a deterministic model for the evolution of the marker frequencies, parameterised by $q_0$, the mutant haplotype properties $t_2, f_2$, and the marker population in which the mutant appears, $\chi_2$. Optimising these four parameters gives the model which generates marker frequencies $x_{\text{red}}(t_k)$ and $x_{\text{blue}}(t_k)$, which best fit the observations $\tilde{x}_{\text{red}}(t_k)$ and $\tilde{x}_{\text{blue}}(t_k)$. Measuring the fit between the model and the observations gives a likelihood score, indicating the goodness of fit of the model. This score can be compared to models generated with different numbers of mutants. Were more mutants to be added to the model, this would generate more $t_k, f_k$ and $\chi_k$ terms to be fitted. Considering more general classes of experiment, more markers could also be added to the model, requiring more $q_i(0)$ to be fitted. In either case, the principle of optimising a model to fit the observations is the same.

Performance of the method under high beneficial mutation rates

While in the main text, we evaluate the performance of the method under mutation rates up to $5 \times 10^{-7}$, some populations may have beneficial mutation rates substantially higher than this. We therefore carried out further simulations of a population with exponentially distributed selection coefficients with parameter $s_0 = 0.025$ and $U = 10^{-5}$. At this level of mutation, our method effectively reproduced the marker trajectories, but assessment of the selection coefficients and establishment times of the inferred haplotypes presents a challenge. At lower mutation rates, a mapping between inferred events and individual real events can easily be made by matching events based on their respective establishment times. However, at high levels of $U$, multiple establishment events an be seen within a short space of time, so that such a matching is prone to
error. Below, we discuss an example of a simulation at high mutation rate. While a far greater number of beneficial haplotypes arise within each marker population than can be inferred by the method, where multiple haplotypes arise, it is likely that the most beneficial of these will have the largest effect on the marker. As such, a reasonably good fit can be made between the inferred haplotypes, and the primary haplotypes within each marker population.

An example of a simulated population derived at $U = 10^{-5}$ is shown in Supplementary Figure S3, which is presented in a format similar to that of Figure 2a in the main text. Multiple beneficial haplotypes reach establishment in each population, leading to the relative growth of first the blue, then the red marker populations, before the eventual fixation of the blue population after around 500 generations. Use of our method produced a good fit between the marker frequencies and an inferred model with four beneficial mutations (Supplementary Figure S4).

Comparison of the details of the haplotypes inferred by our model showed a reasonably good correspondence between real and inferred haplotype fitnesses. Four beneficial haplotypes were inferred, one in the red marker population with establishment time and fitness $(t_a, \sigma_a)$ equal to (130, 0.098), and three in the blue marker population, with establishment times and fitnesses (78, 0.067), (166, 0.111), and (325, 0.171). The fitnesses compared favourably with those of the largest haplotypes in the population.

Examining the detail of the simulation, after 230 generations, the blue population is largely comprised of two haplotypes, having fitnesses 0.065 and 0.076 respectively, roughly matching the first beneficial haplotype inferred to reach establishment. At this time, the red population is primarily comprised of a single haplotype with fitness 0.098, equivalent to the second inferred haplotype. Moving forward in the simulation, after 380 generations, the largest haplotype in the blue population has selection coefficient 0.128, close to that of the third inferred haplotype, while the red population is finally overtaken by two haplotypes which make up the blue population at time 480 generations, having selection coefficients 0.185 and 0.225 respectively.

While the second beneficial haplotype to emerge in the red population is missed by our method, the haplotypes we do infer are generally representative of the most beneficial observed haplotypes in the simulation. As mentioned in the main text, sequencing of the population would be required to recapture the full haplotype details.

**Fig. S1.** Establishment times and fitnesses of real and inferred mutant haplotypes (growing population). Establishment times of inferred events paired with establishment times of significant real haplotypes. Paired event times are shown for the first (red), second (green) and third (blue) mutants inferred by the method. Each pair is shown by a rectangle, the height/width ratio of which is equal to the ratio between the real and inferred mutant selection coefficients. Data are shown for simulations in which mutations have exponentially distributed beneficial selection, and in which the population undergoes successive periods of growth and bottlenecks. Data is shown for $\sigma_0 = 0.05$ and $U = 1 \times 10^{-7}$.

**Fig. S2.** Estimates of the beneficial mutation rate $U$ for simulations with constant selection coefficients. Inferred beneficial mutation rates (blue dots) were calculated across sets of simulations with different selection coefficients using the formula $U = n_m/2\sigma T$ given in the main text. The red line indicates the line of perfect agreement between the real and inferred beneficial mutation rates.
Inference of positive selection from marker dynamics

**Fig. S3.** Haplotype frequencies for an example simulated population with high mutation rate. Data from a simulation in which selection coefficients were exponentially distributed with $\sigma_0 = 0.025$, and $U = 10^{-5}$. Times within marker populations at which a new mutation reaches $q_0$ are shown by red or blue dots with black outline according to marker. Frequencies are shown in a cumulative way, with black lines separating haplotypes within each markere population.

**Fig. S4.** Fitting of marker frequencies for an example simulated population with high mutation rate. The inferred marker frequency (blue line) is fitted to sample marker frequencies (red dots) from the population detailed in Supplementary Figure S3. The fitted model infers the establishment of four beneficial haplotypes.

Table S1. Correlation between real and inferred mutation establishment times, and haplotype fitness ratios. No result for correlation could be calculated in the case of exponentially distributed selection coefficients with $\sigma_0 = 0.01$ and $1 \times 10^{-8}$, only one mutation being identified across all 50 simulations, this one mutation having no real counterpart. We have applied a maximum cut off of 0.5 to inferred fitnesses, i.e. $f_i > 0.5 \rightarrow 0.5$, to avoid bias coming from cases where only a lower bound can be fixed for fitness as was explained in the text.

| Selection $\sigma_0$ | Mutation rate $U$ | Correlations ($t_i, t_r$) |
|----------------------|-------------------|-------------------------|
| Constant selection coefficients |  |  |
| $10^{-8}$ | $5 \times 10^{-8}$ | $10^{-7}$ | $2 \times 10^{-7}$ | $5 \times 10^{-7}$ |
| 0.025 | 0.959 | 0.792 | 0.781 | 0.592 | 0.679 |
| 0.050 | 0.995 | 0.977 | 0.960 | 0.942 | 0.970 |
| 0.100 | 0.999 | 0.953 | 0.983 | 0.926 | 0.963 |
| 0.200 | 0.999 | 0.994 | 0.975 | 0.974 | 0.984 |
| Mean of selection coefficient ratios $f_i/f_r$ |  |  |
| 0.025 | 1.027 | 0.980 | 0.940 | 1.114 | 1.108 |
| 0.050 | 0.973 | 1.139 | 1.016 | 1.254 | 1.096 |
| 0.100 | 1.002 | 0.987 | 0.965 | 0.925 | 1.042 |
| 0.200 | 1.029 | 1.040 | 0.940 | 0.939 | 0.966 |
| Exponentially distributed selection coefficients. |  |  |
| Correlations ($t_i, t_r$) |  |  |
| 0.010 | n/a | 0.931 | 0.600 | 0.609 | 0.852 |
| 0.025 | 0.995 | 0.986 | 0.995 | 0.954 | 0.908 |
| 0.050 | 0.999 | 0.992 | 0.989 | 0.989 | 0.938 |
| 0.100 | 0.999 | 0.994 | 0.985 | 0.983 | 0.912 |
| Mean of selection coefficient ratios $f_i/f_r$ |  |  |
| 0.010 | n/a | 0.960 | 1.231 | 1.030 | 1.483 |
| 0.025 | 1.015 | 1.021 | 1.002 | 1.094 | 1.134 |
| 0.050 | 0.984 | 1.053 | 1.058 | 1.121 | 1.160 |
| 0.100 | 1.022 | 1.085 | 1.060 | 1.024 | 1.069 |
Table S2. Fraction of new haplotypes called by the method Mean numbers of mutant haplotypes called by the method as a fraction of the mean number of haplotypes in simulated populations that reached a frequency of $10q_0 = 0.01$. Results are given for constant and exponentially distributed selection coefficients, for each set of mutation and selection parameters.

| Selection $\sigma$ | Mutation rate $U$ | $10^{-8}$ | $5 \times 10^{-8}$ | $10^{-7}$ | $2 \times 10^{-7}$ | $5 \times 10^{-7}$ |
|------------------|------------------|----------|----------------|----------|----------------|----------|
| Constant selection coefficients | | | | | | |
| 0.025 | 0.667 | 0.758 | 0.565 | 0.442 | 0.300 |
| 0.050 | 0.952 | 0.825 | 0.683 | 0.560 | 0.284 |
| 0.100 | 0.909 | 0.781 | 0.770 | 0.561 | 0.318 |
| 0.200 | 0.898 | 0.855 | 0.645 | 0.505 | 0.320 |
| Exponentially distributed selection coefficients | | | | | | |
| 0.010 | 0.500 | 0.875 | 0.667 | 0.725 | 0.443 |
| 0.025 | 0.889 | 0.824 | 0.797 | 0.760 | 0.650 |
| 0.050 | 0.812 | 0.839 | 0.747 | 0.737 | 0.602 |
| 0.100 | 0.923 | 0.800 | 0.870 | 0.714 | 0.682 |