Inference from allele frequency time series

Steven N. Evans

Department of Mathematics & Department of Statistics
University of California at Berkeley

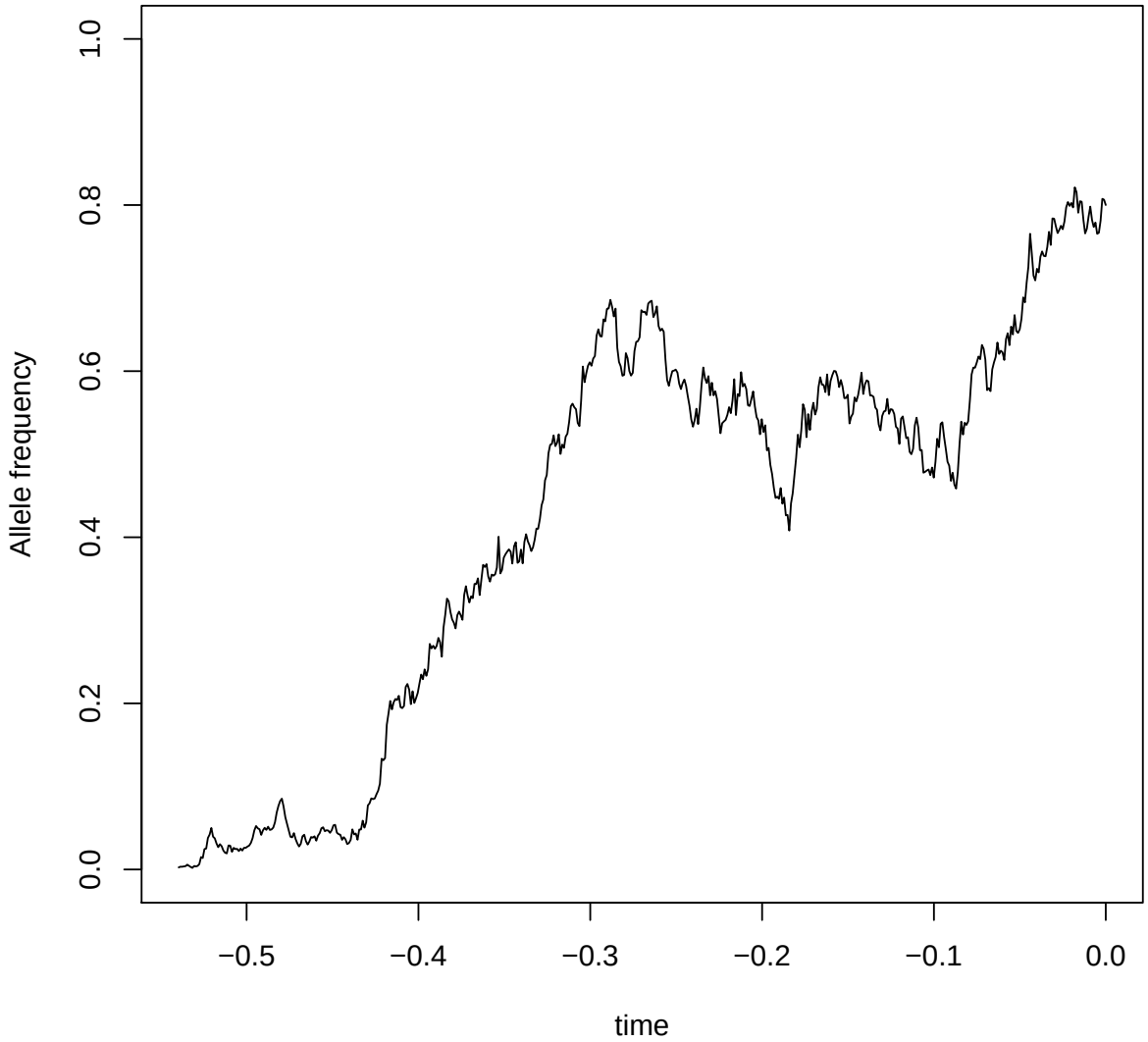
June, 2018

Collaborators

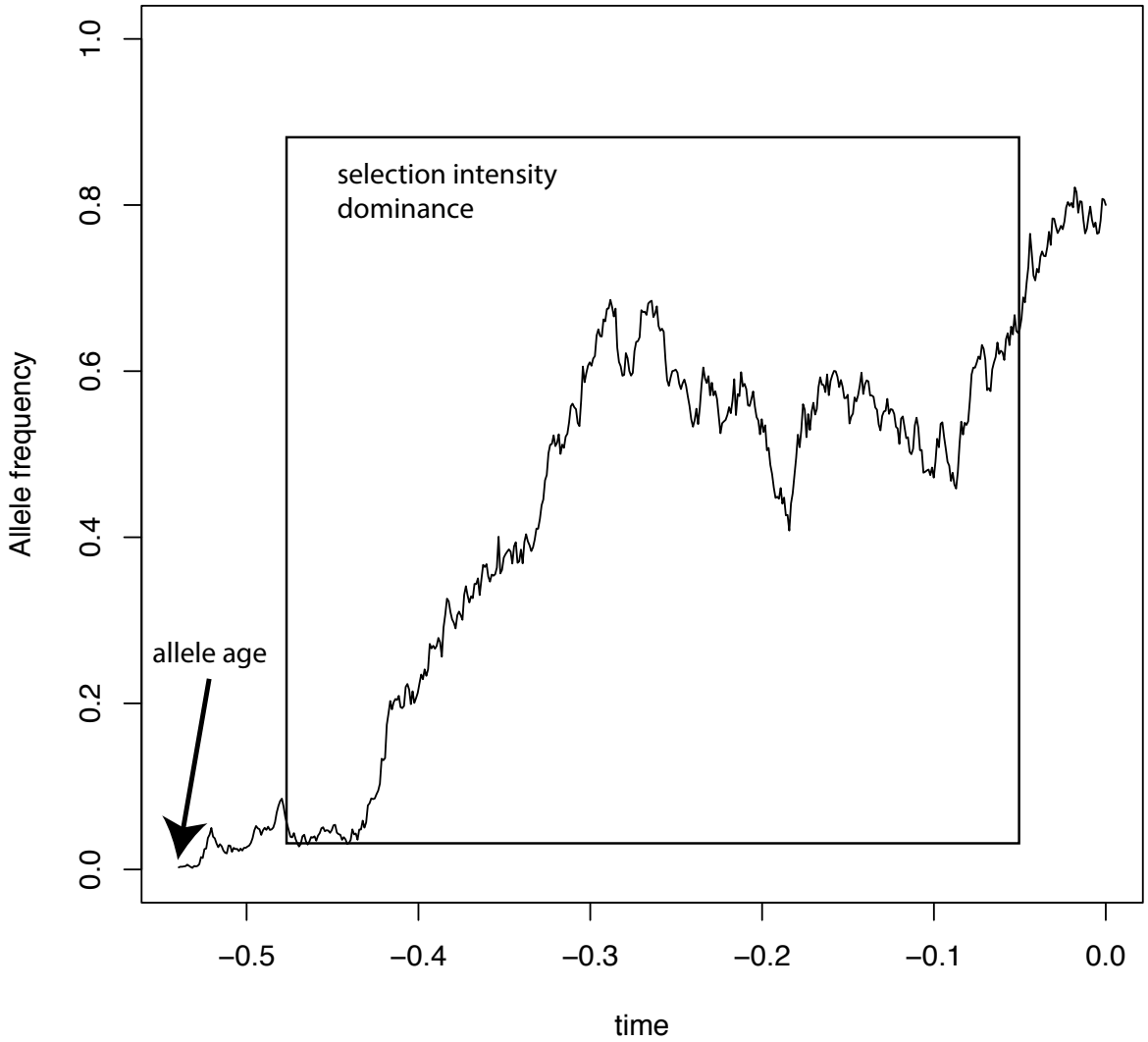
Joshua G. Schraiber, Department of Biology, Temple University

Montgomery Slatkin, Department of Integrative Biology, University of California at Berkeley

Allele frequency trajectory



Quantities of interest



Sampling alleles

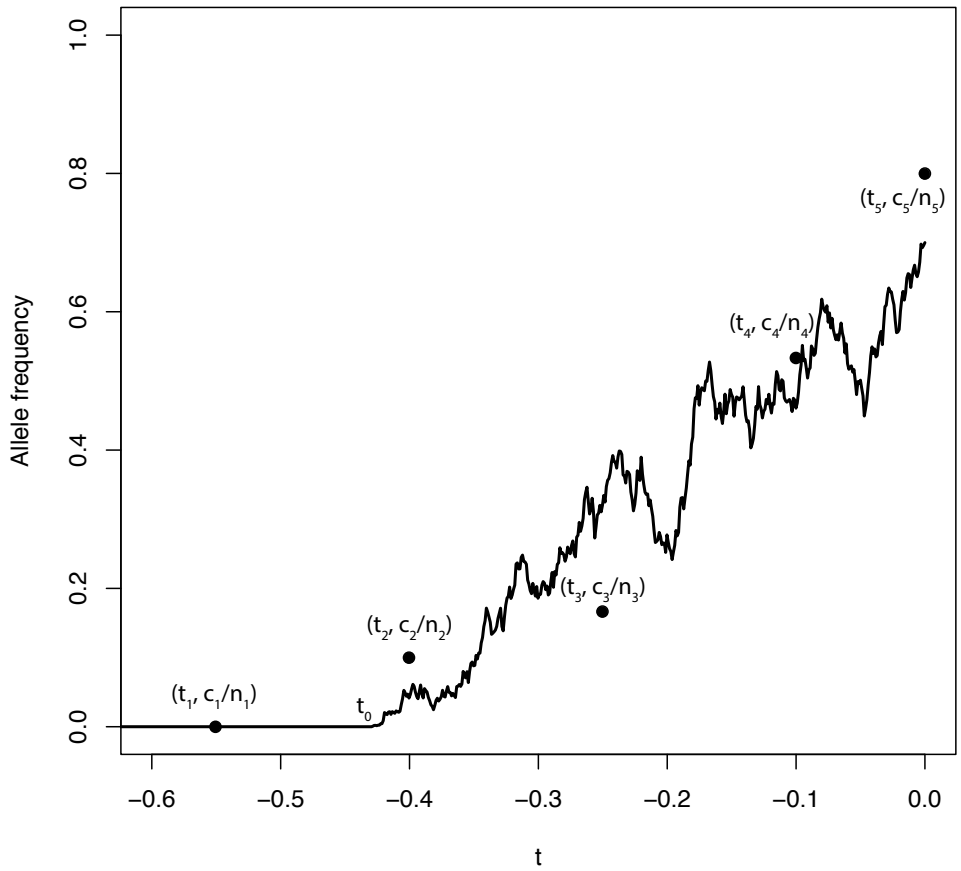
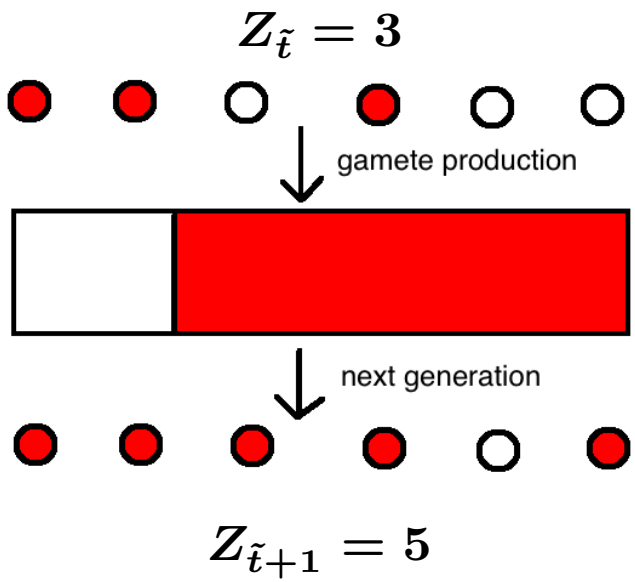


Figure: At each time, t_i, a sample of size n_i chromosomes is taken and c_i copies of the derived allele are observed. Note that t_1 is more ancient than the allele age, t_0.

The model

- Population of 2N chromosomes
- Two alleles: A/a
  - Keep track of $Z_{\tilde{t}}$, number of A chromosomes in generation $\tilde{t}$
- Infinite pool of gametes
  - Under neutrality, each chromosome contributes equally
  - With natural selection, chromosomes contribute unequally depending on allelic state
- 2N Bernoulli draws from the gamete pool produce the next generation



Allele frequency change

- On average, general diploid selection

Genotype	AA	Aa	aa
Fitness	$1+s$	$1+hs$	1

Expected allele frequency change

$$\mathbb{E}(Z_{\tilde{t}+1}|Z_{\tilde{t}}) = \frac{(1+s)Z_{\tilde{t}}^2 + (1+sh)Z_{\tilde{t}}(2N - Z_{\tilde{t}})}{(1+s)Z_{\tilde{t}}^2 + 2(1+sh)Z_{\tilde{t}}(2N - Z_{\tilde{t}}) + (2N - Z_{\tilde{t}})^2}$$

- Variance due to binomial

Variance in allele frequency change

$$\mathrm{Var}(Z_{\tilde{t}+1}|Z_{\tilde{t}}) \approx Z_{\tilde{t}}(2N - Z_{\tilde{t}})$$

Diffusion limit

- Define $X_t = \frac{Z_{\tilde{t}/2N}}{2N}$.
  - The allele frequency with time measured in units of $2N$ generations.
- Take a limit as $N \uparrow \infty$ and $|s| \downarrow 0$ such that $2Ns \to \alpha$.
- Get a diffusion limit.

Wright-Fisher SDE

$$dX_t = \alpha X_t(1 - X_t)(X_t + h(1 - 2X_t))\, dt + \sqrt{X_t(1 - X_t)}\, dB_t,$$

where B is a standard Brownian motion.

Diffusion limit with varying population size

- Suppose in rescaled time that the population size after $2Nt$ generations is $2N\rho(t)$.
- Still get a diffusion limit.

Wright-Fisher SDE with varying population size

$$dX_t = \alpha X_t(1-X_t)(X_t + h(1-2X_t))\,dt + \sqrt{\frac{X_t(1-X_t)}{\rho(t)}}\,dB_t,$$

where B is a standard Brownian motion.

Inference from a time series

- The likelihood of the data given the allele frequency path, $\boldsymbol{\alpha}$, $\boldsymbol{h}$ and $\boldsymbol{t_0}$ only depends on the allele frequency path and is easy to compute (just binomials).
- Computing the likelihood of the data given $\boldsymbol{\alpha}$, $\boldsymbol{h}$ and $\boldsymbol{t_0}$ is hard – it involves integrating out the allele frequency path, and this is equivalent to solving a PDE with no explicit solution.
- Various approaches:
  - Bollback *et al.* (2008): solve the PDE numerically.
  - Malaspinas *et al.* (2012): approximate the diffusion using a birth-and-death type Markov chain (essentially the same).
  - Steinrücken *et al.* (2013): use orthogonal polynomials

Another solution: infer the whole path!

- Treat the unknown allele frequency path as another parameter that has to be estimated.
- If we use Bayesian inference, then to be consistent we should ideally use a prior such that the conditional distribution of the allele frequency path given α, h and t_0 and segregation at the present is what it should be under the W-F model.
- An imputation of the allele frequency path is interesting in its own right.

Diffusion path likelihoods?

- The obvious way to obtain the posterior is to use a Markov chain Monte Carlo method such as Metropolis-Hastings.
- This requires a prior with a density against some fixed reference measure.
- What measure should we use on path space?

Girsanov's theorem

Write $\mathbb{W}_{t_*,x}$ for the distribution of a Brownian motion that starts at time t_* at position x.

Diffusion path likelihoods

Suppose that a diffusion satisfies the SDE

$$dX_t = a(X_t, t)dt + dB_t, \quad X_{t_*} = x,$$

and let $\mathbb{P}_{t_*,x}$ be the corresponding distribution on paths indexed by $[t_*, \infty)$. The likelihood under $\mathbb{P}_{t_*,x}$ relative to $\mathbb{W}_{t_*,x}$ for paths indexed by $[t_*, t]$ is

$$\frac{d\mathbb{P}_{t_*,x}}{d\mathbb{W}_{t_*,x}}(X) = \exp\left\{\int_{t_*}^{t} a(X_s, s)\, dX_s - \frac{1}{2}\int_{t_*}^{t} a^2(X_s, s)\, ds\right\}.$$

Can we use Girsanov?

- The Wright-Fisher SDE

$$dX_t = \alpha X_t(1 - X_t)(X_t + h(1 - 2X_t))\, dt + \sqrt{\frac{X_t(1 - X_t)}{\rho(t)}}\, dB_t$$

  is NOT of the form

$$dX_t = a(X_t, t)dt + dB_t,$$

  so how can we use Girsanov?

- Answer: We transform the time and space scales.

Changing the time scale

Apply the time transformation $\tau = f(t)$ with

$$f(t) = \int_0^t \frac{1}{\rho(s)}\, ds$$

to the W-F diffusion to obtain a new SDE

$$dX_\tau = \alpha\rho(f^{-1}(\tau))X_\tau(1-X_\tau)(X_\tau + h(1-2X_\tau))\, d\tau + \sqrt{X_\tau(1-X_\tau)}\, dB_\tau.$$

Changing the space scale: the Fisher transformation

Next, apply the space transformation

$$Y_\tau = \arccos(1 - 2X_\tau)$$

and note that the result is an SDE

$$dY_\tau = \frac{1}{4}\left(\alpha\rho(f^{-1}(\tau))\sin(Y_\tau)(1 + (2h-1)\cos(Y_\tau)) - 2\cot(Y_\tau)\right)\, d\tau + dB_\tau$$

to which Girsanov applies.

Bad things happen at the boundaries

- The process Y_τ lives on $(0, \pi)$ and is absorbed at the boundaries (corresponding to loss or fixation).
- However, Brownian motion lives on all of $\mathbb{R}$.
- This causes the drift of Y to blow up at the boundaries.
- This will create problems for inferring allele age.
- We need a better reference measure.

A consequence of Girsanov

- Suppose that $\mathbb{P}_{t_*,x}$ is the distribution of a diffusion satisfying the SDE

$$dX_t = a(X_t, t)\, dt + dB_t, \quad X_{t_*} = x.$$

- Suppose that $\mathbb{Q}_{t_*,x}$ is the distribution of a diffusion satisfying the SDE

$$dX_t = b(X_t, t)\, dt + dB_t, \quad X_{t_*} = x.$$

- The likelihood under $\mathbb{P}_{t_*,x}$ relative to $\mathbb{Q}_{t_*,x}$ for paths indexed by $[t_*, t]$ is

$$\begin{aligned}\frac{d\mathbb{P}_{t_*,x}}{d\mathbb{Q}_{t_*,x}}(X) &= \frac{d\mathbb{P}_{t_*,x}}{d\mathbb{W}_{t_*,x}}(X) \Big/ \frac{d\mathbb{Q}_{t_*,x}}{d\mathbb{W}_{t_*,x}}(X) \\ &= \exp\left\{ \int_{t_*}^{t} (a(X_s, s) - b(X_s, s))\, dX_s \right. \\ &\quad \left. - \frac{1}{2} \int_{t_*}^{t} (a^2(X_s, s) - b^2(X_s, s))\, dt \right\}.\end{aligned}$$

Matching the singularities at 0

- Suppose that $\mathbb{P}_{t_*,x} = \mathbb{P}^{\alpha,h}_{t_*,x}$ is the distribution of the time and space transformed process (Y_τ) started from x at time t_*.
- Because

$$\frac{1}{4}\left(\alpha\rho(f^{-1}(\tau))\sin(Y_\tau)(1+(2h-1)\cos(Y_\tau)) - 2\cot(Y_\tau)\right)$$
$$= -\frac{1}{2Y_\tau} + O(Y_\tau)$$

  when Y_τ is small, a good choice for $\mathbb{Q}_{t_*,x}$ would be one where $b(x,t) \approx -\frac{1}{2x}$ as $x \downarrow 0$.
- Practically, such a choice is only helpful in a Metropolis-Hastings algorithm if we can easily sample from distributions related to $\mathbb{Q}_{t_*,x}$.

Bessel process of dimension d

Suppose that $W_t = (W_t^1, \ldots, W_t^d)$ is a d-dimensional standard Brownian motion. The radial part process

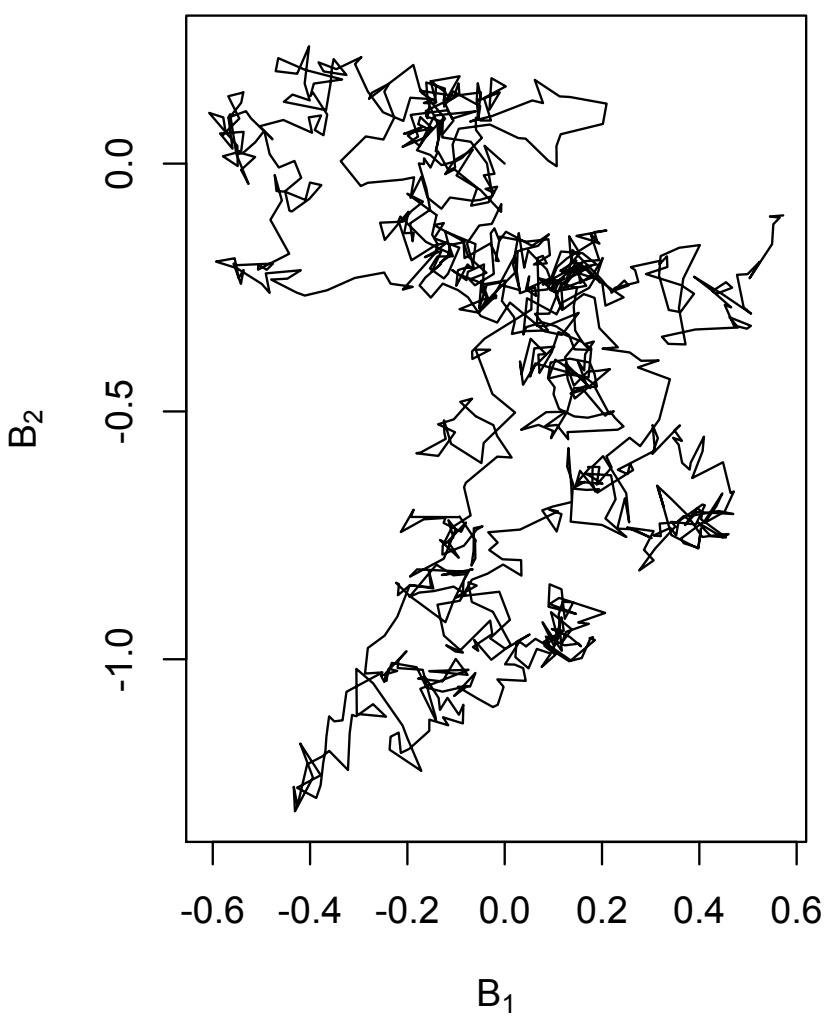
$$R_t = \sqrt{\sum_{i=1}^{d} (W_t^i)^2}$$

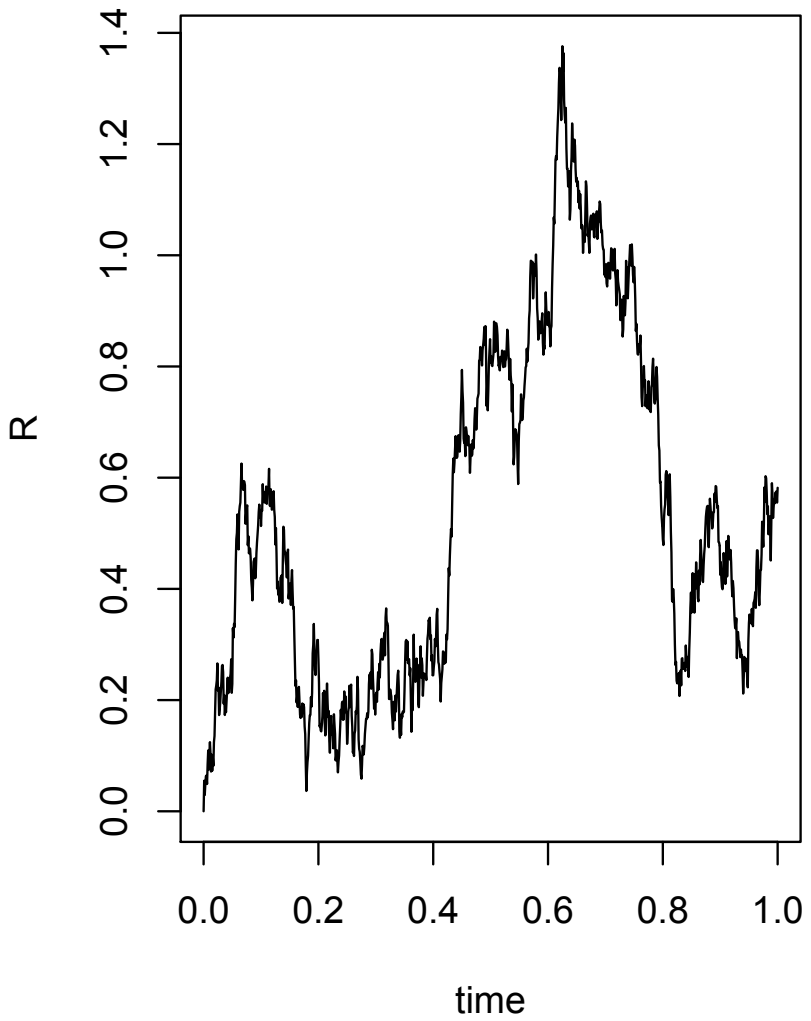
is the Bessel process of dimension d, written Bes(d). The diffusion R satisfies the SDE

$$dR_t = \frac{d-1}{2R_t}\, dt + dB_t.$$

Note that this SDE makes sense for all $d \geq 0$, even if $d \notin \mathbb{N}$, and the resulting process is still called Bes(d).

Bes(2)


B2
0.0
-0.5
-1.0
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
B1


R
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.2
0.4
0.6
0.8
1.0
time

Bes(0)

- Near $\mathbf{0}$, the drift of the time and space transformed Wright-Fisher diffusion looks like -$\frac{\mathbf{1}}{\mathbf{2}y}$
- The Bes(0) diffusion satisfies

$$\boldsymbol{dR_t = -\frac{1}{2R_t}\,dt + dB_t}$$

- This is exactly what we want.

Avoiding Itô integrals

- Recall that the likelihood under $\mathbb{P}^{\alpha,h}_{t_*,x}$ relative to $\mathbb{Q}_{t_*,x}$ for paths indexed by $[t_*, t]$ is

$$\frac{d\mathbb{P}^{\alpha,h}_{t_*,x}}{d\mathbb{Q}_{t_*,x}}(X) = \exp\left\{ \int_{t_*}^{t} (a(X_s, s) - b(X_s, s))\, dX_s - \frac{1}{2} \int_{t_*}^{t} (a^2(X_s, s) - b^2(X_s, s))\, dt \right\}$$

  for suitable a and b.

- It appears that computing the likelihood requires numerical evaluation of an Itô integral.
- This is known to be hard.
- An integration-by-parts type trick developed by the group around Gareth Roberts circumvents this.

Escaping from $\mathbf{0}$

- The time and space transformed Wright-Fisher diffusion and the Bes(0) diffusion both stay at $\mathbf{0}$ if they are started there.
- We need a prior that puts mass on paths that start at $\mathbf{0}$ but escape.

Entrance laws

Suppose that $q(s, x; t, y)$ is the transition density function with respect to Lebesgue measure of a time-inhomogeneous Markov process with state-space I, where I is some interval. An entrance law for this process relative to some entrance time t_* is a function $n(t_*; t, x)$, $t > t_*$, $x \in I$, such that $\int_I n(t_*; s, x) q(s, x; t, y)\, dx = n(t_*; t, y)$ for $t_* < s < t$.

Entrance law for Bes(0)

- The transition density function for the Bessel(0) process is

$$q(s,x;t,y) = \frac{y}{t-s} \exp\left(-\frac{x^2+y^2}{2(t-s)}\right) I_1\left(\frac{xy}{2(t-s)}\right),$$

where I_1 is the modified Bessel function of the first kind with index 1.

- Because $I_1(z) \sim \frac{z}{2}$ as $z \to 0$, it follows that

$$\lim_{x \downarrow 0} \frac{q(s,x;t,y)}{q(s,x;s+1,1)} = \frac{y^2}{t-s} \exp\left(-\frac{y^2}{2(t-s)}\right) \exp\left(\frac{1}{2}\right),$$

and so

$$n(t_*;t,x) = \frac{x^2}{t-t_*} \exp\left(-\frac{x^2}{2(t-t_*)}\right)$$

is an entrance law for the entrance time t_*.

Starting Bes(0) from 0

- Let $q(s,x;t,y)$ be the transition density function for the Bessel(0) process and $n(t_*;t,x)$ the entrance law for the entrance time t_*.
- There is a σ-finite measure $\mathbb{Q}_{t_*,\uparrow}$ on the paths indexed by (t_*,∞) such that

$$\mathbb{Q}_{t_*,\uparrow}\{X_{s_1}\in dx_1,\dots,X_{s_k}\in dx_k\} = n(t_*;s_1,x_1)q(s_1,x_1;s_2,x_2)\cdots q(s_{k-1},x_{k-1};s_k,x_k)\,dx_1\cdots dx_k$$

for $t_*<s_1<\ldots<s_k$.

Starting W-F from 0

- Let $\mathbb{P}^{\alpha,h}_{t_*,x}$ be the distribution of W-F process with parameters α, h started at time t_* at position x.
- Let $\mathbb{Q}_{t_*,x}$ be the distribution of Bes(0) started at time t_* at position x.
- Recall that the likelihood of a path distributed according to $\mathbb{P}^{\alpha,h}_{t_*,x}$ relative to $\mathbb{Q}_{t_*,x}$ for paths indexed by $[t_*, t]$ is

$$\begin{aligned}\frac{d\mathbb{P}^{\alpha,h}_{t_*,x}}{d\mathbb{Q}_{t_*,x}}(X) &= \exp\left\{\int_{t_*}^{t} (a(X_s,s) - b(X_s,s))\, dX_s \right.\\ &\qquad \left. - \frac{1}{2}\int_{t_*}^{t} (a^2(X_s,s) - b^2(X_s,s))\, ds\right\}\\ &=: \Phi^{\alpha,h}_{t_*,t}(X),\end{aligned}$$

  for suitable a and b.
- Set

$$\mathbb{P}^{\alpha,h}_{t_*,\uparrow}\ (dX) := \Phi^{\alpha,h}_{t_*,t}(X)\mathbb{Q}_{t_*,\uparrow}\ (dX)$$

  for a path X indexed by $(t_*, t]$.

Conditional prior for allele frequency path

- With a slight abuse of notation, think of $\mathbb{P}^{\alpha,h}_{t_*,\uparrow}$ as a measure that lives on paths indexed by $\mathbb{R}$ that are zero on $(-\infty, t_*]$.
- Take the (improper) conditional prior on the time and space transformed allele frequency path Y given (α, h, t_0) to be $\mathbb{P}^{\alpha,h}_{f(t_0),\uparrow}$.

Full prior

- Take any convenient distributions as the marginal priors on the selection coefficient α and the dominance coefficient h.
- Take the (improper) marginal prior on the allele age t_0 to be the measure with density ρ, where $\rho(t)$ is the relative population size at time t.
- Take the (improper) marginal prior on (α, h, t_0) to be the product of the respective marginal priors.
- This completely specifies the full prior on (t_0, Y, α, h).

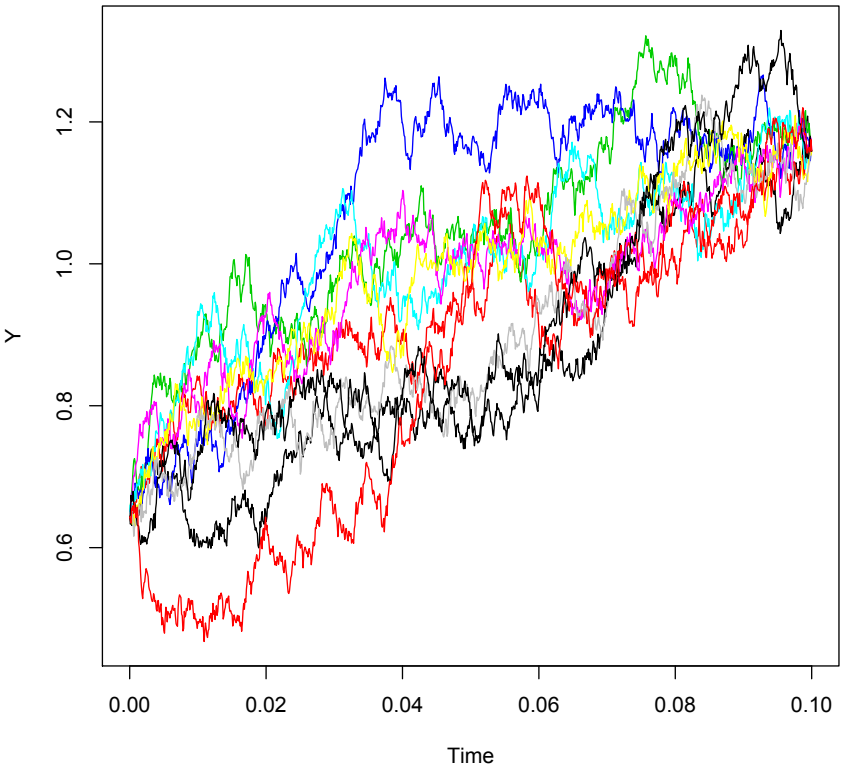
A Markov chain Monte Carlo algorithm

For each step of the algorithm:

1. Decide whether to update one of:
    a. a bit of the path,
    b. allele age,
    c. the current allele frequency,
    d. selection coefficient,
    e. dominance coefficient.
2. Update the chosen parameter.
3. Compute the proposal ratio.
4. Compute the prior ratio.
5. Compute the likelihood ratio.
6. Accept or reject the update according to the Metropolis-Hastings criterion.

Path updates

- Choose a piece of the path with random endpoints to update.
- Ideally, proposed path updates would come from the posterior, but that's hard.
- Simulate a new piece of path that follows a Bes(0) process conditioned to keep the same values at the endpoints as the original piece of path.
  - Such a Bes(0) bridge is not too different from a time and space transformed Wright-Fisher bridge.
  - That is, the proposal ratio $\approx \mathbf{1}$.
  - It's easy to simulate a Bes(0) bridge (WHY?).



Simulating Bes(0) bridges

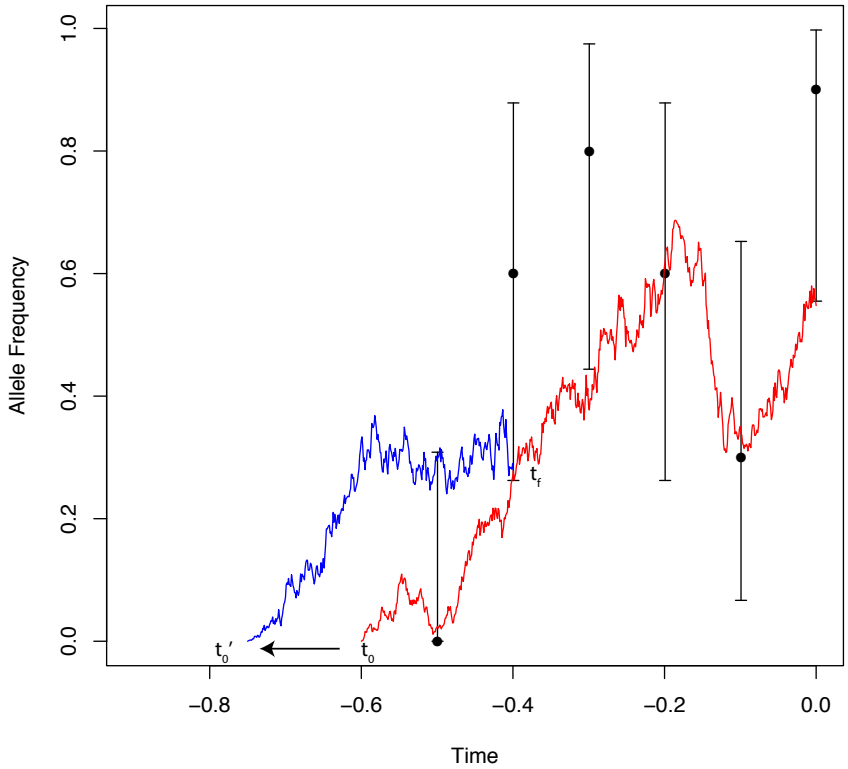
- Simulate a Bes(0) bridge from x at time 0 to y in time t as follows.
  1. Take the 4-dimensional vector $u = (0, 0, 0, x)^T$.
  2. Sample a vector $V \sim$ von Mises-Fisher $\left(\frac{u}{x}, \frac{xy}{t}\right)$
  3. Sample a 4-dimensional standard Brownian motion $\{B_s, 0 \leq s \leq t\}$.
  4. Construct the 4-dimensional Brownian bridge from u to V,

$$B_s^{(x,y,t)} = \left(1 - \frac{s}{t}\right) u + \frac{s}{t} y V + \left(B_s - \frac{s}{t} B_t\right), \quad 0 \leq s \leq t.$$

  5. Compute the Euclidean norm of $B^{(x,y,t)}$.
- The result is a Bes(4) bridge from x at time 0 to y in time t.
- However, the Bes(4) process is the same as the Bes(0) process conditioned to never hit 0.
- Moreover, the bridges of a conditioned diffusion are the same as the unconditioned diffusion.
- So the result is a Bes(0) bridge.

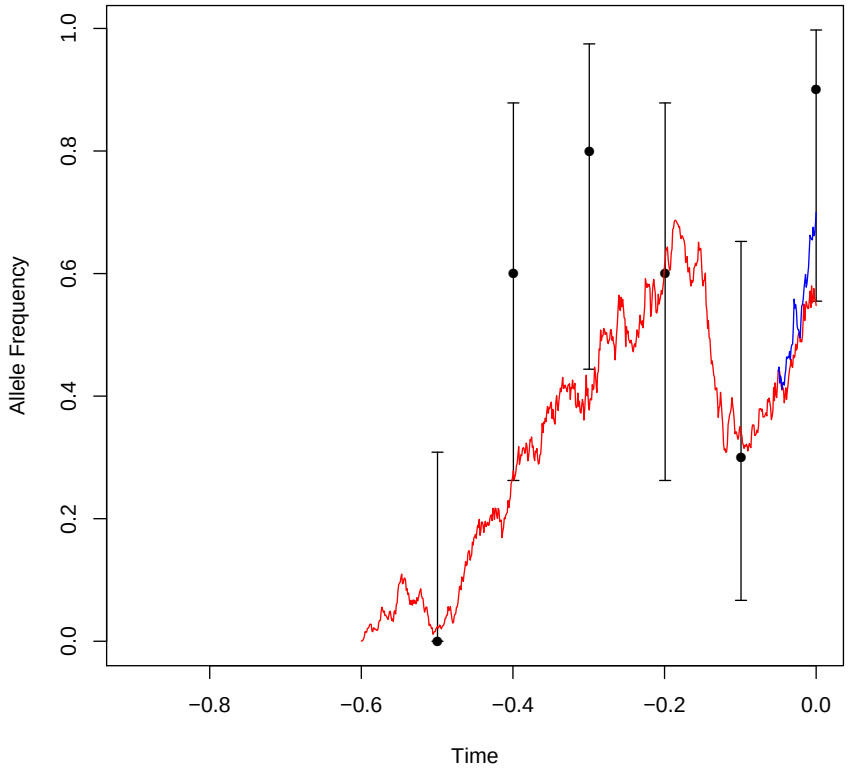
Allele age updates

- Choose a new age from some proposal distribution $q(t_0'|t_0)$.
- Generate a bridge from $(t_0', 0)$ to (t_f, Y_{t_f}) where t_f is the time of the first non-zero observation.
- The proposal ratio needs to account for the densities of paths that go from $(t_0, 0)$ to (t_f, Y_{t_f}) and from $(t_0', 0)$ to (t_f, Y_{t_f})



Current frequency updates

- Choose a new current frequency from some proposal distribution $q(Y'_{t_k}|Y_{t_k})$
- Generate a bridge from (t_s, Y_{t_s}) to (t_k, Y_{t_k}) where t_s is a (fixed) time in the final interval.
- The proposal ratio needs to account for the densities of Bes(0) paths that go from (t_s, Y_{t_s}) to (t_k, Y_{t_k}) and from (t_s, Y_{t_s}) to (t_k, Y'_{t_k}).



Summary of results for the ASIP locus in horses

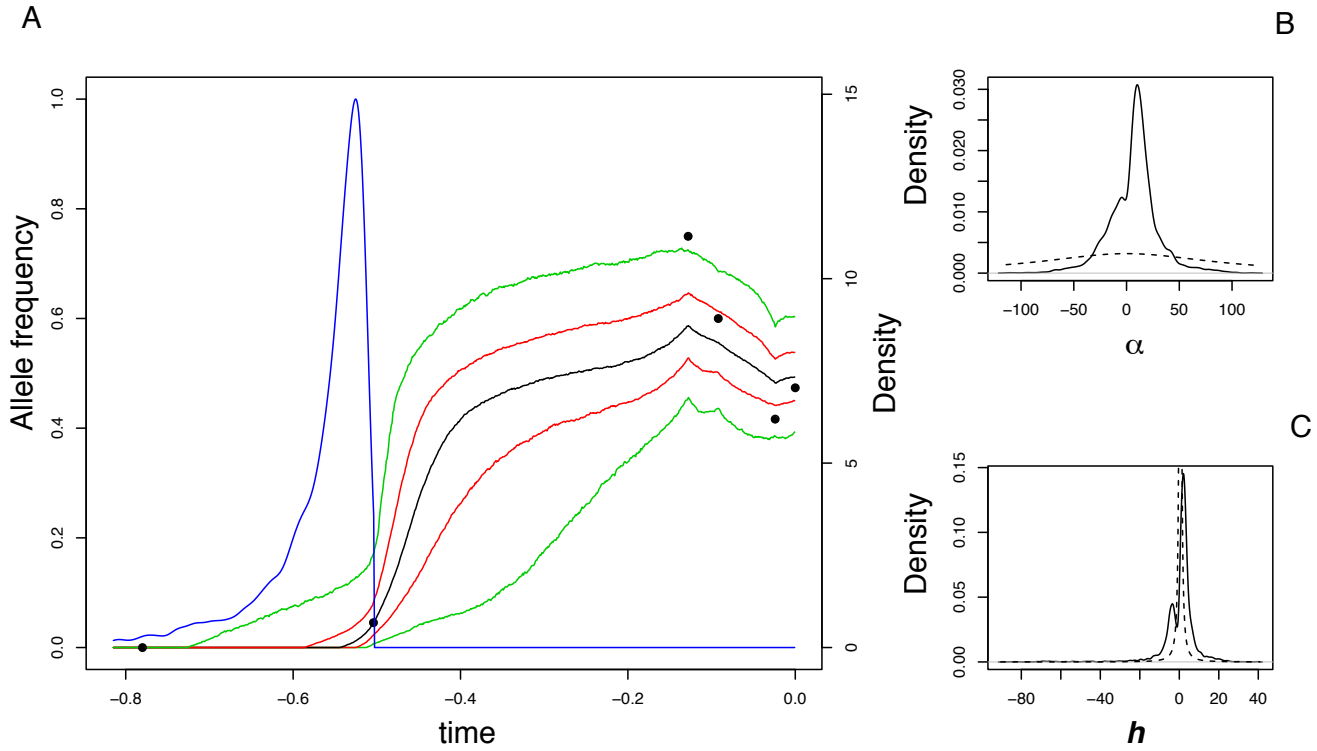


Figure: Panel A shows the posterior distribution of paths as well as the posterior distribution of allele age. Filled circles are the sample allele frequencies, while the solid black, red and green lines show the median, interquartile, and 95% credible intervals of the path, respectively. The blue curve shows the posterior distribution of the allele age. Time is measured in diffusion units relative to the most recent sample (so that 0.0 corresponds to 500 years BCE). Panel B and C show the posterior distribution of α and h, respectively. In both, solid lines are the posterior while dashed lines show the prior.

ASIP locus in horses continued

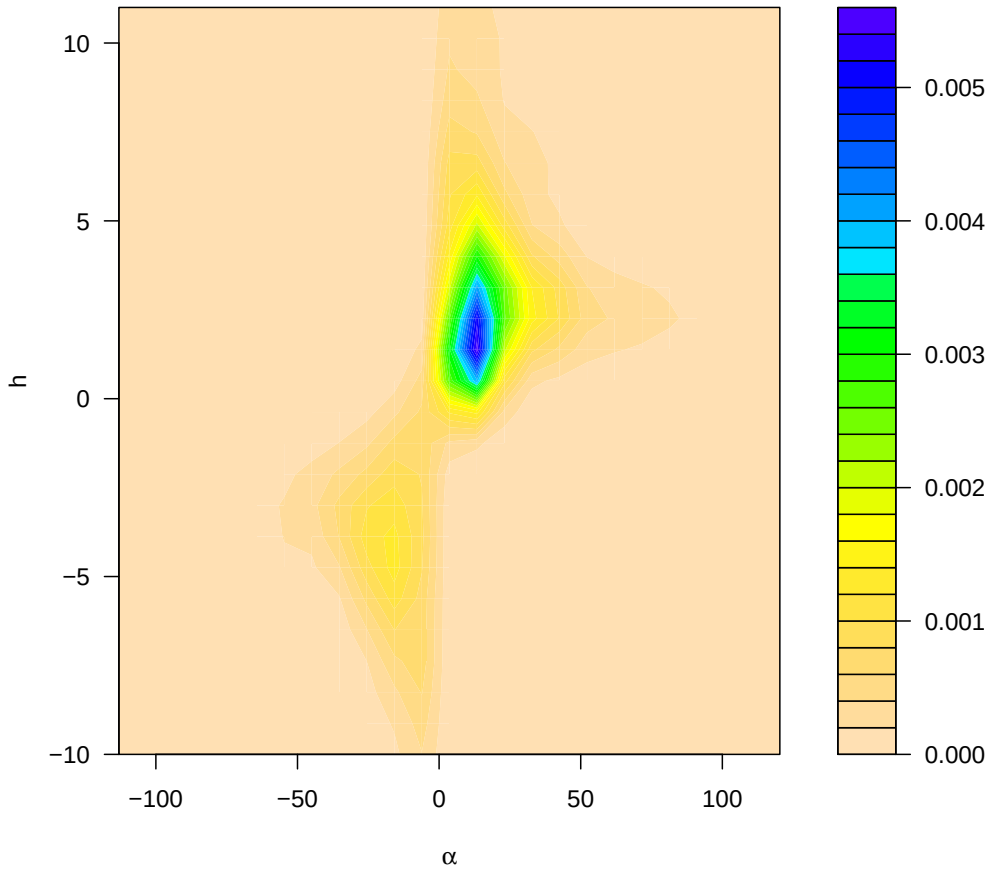


Figure: Joint posterior density of α and h for the ASIP locus in horses. Regions of highest posterior density are shown in blue.

ASIP locus in horses continued

- The most likely selective mechanism is overdominance ($\hat{h} = \mathbf{2.02}$, $\hat{\alpha} = \mathbf{10.23}$), in agreement with the conclusion reached by Steinrücken *et al.* (2013).
- The inferred allele frequency quickly rises to intermediate value and then stays approximately constant, a hallmark of overdominance.
- The allele almost certainly arose more recently than the most ancient time point, at which time zero copies of the derived allele were found ($\hat{t}_0 = \mathbf{-0.53}$, approximately $\mathbf{13,700}$ years BCE).

Acknowledgments

- Mason Liang
- Anand Bhaskar
- Matthias Steinrücken
- Yun Song.