

Non-equilibrium theory of the allele frequency spectrum

Steven N. Evans

Departments of Statistics and Mathematics
and Graduate Group in Computational and Genomic Biology
University of California at Berkeley

`evans@stat.berkeley.edu`

`http://www.stat.berkeley.edu/users/evans`

(with Yelena Shvets and Montgomery Slatkin)

Research supported in part by NSF and NIH grants

Contemporary traces of past human population growth?

It is suggested in Reich and Lander (2001) and Williamson et al. (2005) that the effective size of the modern human population was stable at around 10^4 until 1.5×10^5 years ago.

The current effective size is 1.6×10^9 .

If the population grew according to a given schedule over this period (e.g. exponential growth at a constant rate), how would this manifest itself in the genetic composition of contemporary humans?

Equilibria under stable conditions

If mutations are not reversible at a polymorphic locus, then one allele will eventually fix there (i.e. no other alleles will be present in the population).

However, the distribution of allele frequencies across polymorphic loci (the *frequency spectrum*) reaches a non-trivial equilibrium if both population size and selection intensities are constant.

The theory for the equilibrium frequency spectrum under irreversible mutation was developed in Fisher (1930), Wright (1938), Kimura (1964). Kimura (1969), Sawyer and Hartl (1992), Bustamante et al. (2001), Williamson et al. (2004), ...

What if population size or selection intensities change?

Nei et al. (1975) showed that rapid growth resulted in more low frequency alleles than expected under neutrality.

Tajima (1989) confirmed that conclusion and examined the effect of past population growth on other aspects of the frequency spectrum.

Griffiths and Tavaré (1998) developed the coalescent theory for the frequency spectrum of neutral alleles in a population that has experienced arbitrary changes in population size (related work by Nielsen (2000), Wooding and Rogers (2002) and Polanski and Kimmel (2003)).

What if population size or selection intensities change? – continued

Griffiths (2003): the frequency spectrum in a **neutral** population of **variable size** could be derived from the spectrum for a population of constant size when a transformation of the time scale reduces a suitable backwards equation to one for a population of constant size.

We present a forward equation approach that computes the frequency spectrum for **arbitrary population growth and changes in selection intensity**.

Discrete time finite population models

Monoecious, randomly-mating, diploid population containing $N(t)$ individuals in generation $t \in \{0, 1, 2, \dots\}$.

Independent loci.

At each locus there are only two alleles **A**, the **derived allele**, and **a** the **ancestral allele**. **Mutation** only occurs from **ancestral to derived**, at loci that haven't seen mutations before (**infinite sites assumption**).

Large number of loci – effectively infinite pool.

Discrete time finite population models – continued

Put $f_j(t) :=$ **expected number of loci** at which **A** is found on j **chromosomes**, $1 \leq j \leq 2N(t)$.

Put $p_{ij}(t) :=$ **conditional probability** that a locus with i copies of **A** in generation t will have j copies in generation $t + 1$.

The change in $f_j(t)$ because of genetic drift and **mutation at rate** μ is described by the set of “forward” difference equations

$$f_j(t + 1) = \sum_{i=1}^{2N(t)} f_i(t)p_{ij}(t) + 2N(t)\mu\delta_{1,j}, \quad 1 \leq j \leq 2N(t + 1).$$

An analogous forward equation

Recall

$$f_j(t+1) = \sum_{i=1}^{2N(t)} f_i(t)p_{ij}(t) + 2N(t)\mu\delta_{1,j}, \quad 1 \leq j \leq 2N(t+1).$$

If $\pi_j(t) :=$ probability of a given locus having j copies of **A** in generation t , then

$$\pi_j(t+1) = \sum_{i=1}^{2N(t)} \pi_i(t)p_{ij}(t).$$

The diffusion limit for a single locus

Assume for now that $p_{ij}(t) = p_{ij}$ – time-homogeneous case.

Assume for now \neq chromosomes constant at $2N\rho$.

Suppose that if we shrink space by a factor of $2N\rho$ and speed time up by a factor of $2N$, then the chain converges to a diffusion process on $[0, 1]$ (call it the 0-diffusion) with generator $\mathcal{G} = a(x)\frac{d}{dx} + \frac{1}{2}b(x)\frac{d^2}{dx^2}$.

This is the scaling regime that is appropriate for models such as Wright-Fisher with or without selection.

The single locus diffusion forward equation

Write $\pi(y, t)$ for the **probability density** of the 0-diffusion at frequency $y \in (0, 1)$ and time $t > 0$.

The Kolomogorov forward equation

$$\frac{\partial}{\partial t}\pi(y, t) = -\frac{\partial}{\partial y}[a(y)\pi(y, t)] + \frac{1}{2}\frac{\partial^2}{\partial y^2}[b(y)\pi(y, t)].$$

holds with suitable boundary conditions.

The diffusion frequency spectrum without mutation

Suppose at time 0 that there are countably many loci at which derived alleles are present, with respective frequencies x_1, x_2, \dots

Assume **for now** there is no further mutation from the ancestral state.

After passage to the diffusion limit, the frequency spectrum at time t is just the **intensity measure** of the **point process** that comes from starting independent copies of the 0-diffusion process at each of the x_i and letting them run to time t .

The forward equation without mutation

The frequency spectrum is obtained by taking a sum of point masses at the x_i and moving it forwards with the 0-diffusion semigroup.

Set $f^o(y, t) :=$ frequency spectrum, then

$$\frac{\partial}{\partial t} f^o(y, t) = -\frac{\partial}{\partial y} [a(y) f^o(y, t)] + \frac{1}{2} \frac{\partial^2}{\partial y^2} [b(y) f^o(y, t)].$$

with suitable boundary conditions.

Introducing mutation from the ancestral type

In the Markov chain model, new mutants arise at rate $2N\rho\mu = \frac{\theta}{2}\rho$ per generation.

The initial number of mutants at a locus is 1.

Hence mutants appear at rate $2N\frac{\theta}{2}\rho$ per unit of rescaled time, with the initial proportion of mutants at a locus being $\frac{1}{2N\rho}$.

Introducing mutation from the ancestral type – continued

Pass to the diffusion limit for the allele frequencies, but **for now** still work with a finite N for the description of the appearance of new mutants.

The evolving point process has new points added at location $\frac{1}{2N\rho}$ at rate $\frac{\theta}{2}2N\rho$.

The points then evolve as independent copies of the diffusion.

Introducing mutation from the ancestral type – continued

Set $P_t(x, dy) := 0$ -diffusion **semigroup**.

Contribution to the frequency spectrum at $y \in (0, 1)$ from mutations that appear after time 0 =

$$2N \frac{\theta}{2\rho} \int_0^t P_{t-s} \left(\frac{1}{2N\rho}, dy \right) ds.$$

Entrance boundary theory

Choose a **scale function** s for the 0-diffusion such that $s(0) = 0$ and $s'(0) = 1$.

The **Doob h -transform**

$$P_u^\uparrow(x, dy) := \frac{1}{s(x)} P_u(x, dy) s(y), \quad 0 < x, y \leq 1,$$

is the semigroup of a diffusion that never hits 0 (the **\uparrow -diffusion**)

The \uparrow -semigroup can be extended to allow starting at 0 by setting

$$P_u^\uparrow(0, dy) = \lim_{x \downarrow 0} P_u^\uparrow(x, dy).$$

The extended process can start at 0 but it never returns to 0.

Entrance boundary theory – continued

Put

$$\lim_{N \rightarrow \infty} 2N\rho P_u \left(\frac{1}{2N\rho}, dy \right) = \frac{P_u^\uparrow(0, dy)}{s(y)} =: \lambda_u(dy),$$

then $\int \lambda_s(dx) P_t(x, dy) = \lambda_{s+t}(dy)$ and $(\lambda_u)_{u>0}$ has densities that satisfy the forward equation.

Hence

$$\lim_{N \rightarrow \infty} 2N \frac{\theta}{2\rho} \int_0^t P_{t-s} \left(\frac{1}{2N\rho}, dy \right) ds = \frac{\theta}{2} \int_0^t \lambda_{t-s}(dy) ds$$

also has densities $\phi_t(x)$ that satisfy the forward equation.

What are the boundary conditions?

Entrance boundary theory – continued

The \downarrow -diffusion is the 0-diffusion conditioned to hit 0 before 1. It has the Doob h -transform semigroup

$$P_t^\downarrow(x, dy) = \left(1 - \frac{s(x)}{s(1)}\right)^{-1} P_t(x, dy) \left(1 - \frac{s(y)}{s(1)}\right).$$

From Williams (1974), the \uparrow -diffusion started at 0 and killed at the last time it visits $y > 0$ is the time-reversal of the \downarrow -diffusion started at y and killed when it first hits 0.

Write $(Q_t^\downarrow)_{t \geq 0} :=$ semigroup of killed \downarrow -diffusion.

Finding the boundary condition

Since $s(y) \approx y$ for y close to 0,

$$\begin{aligned}\lim_{y \downarrow 0} y \phi_t(y) &= \lim_{y \downarrow 0} s(y) \phi_t(y) \\ &= \lim_{y \downarrow 0} s(y) \int_0^t \frac{\theta}{2} \frac{\lambda_{t-s}(dy)}{dy} ds \\ &= \frac{\theta}{2} \lim_{y \downarrow 0} \int_0^t \frac{P_{t-s}^\uparrow(0, dy)}{dy} ds \\ &= \frac{\theta}{2} \lim_{y \downarrow 0} \int_0^t \frac{P_s^\uparrow(0, dy)}{dy} ds \\ &= \frac{\theta}{2} \lim_{y \downarrow 0} \int_0^\infty \frac{P_s^\uparrow(0, dy)}{dy} ds \\ &= \frac{\theta}{2} \lim_{y \downarrow 0} \int_0^\infty \frac{Q_s^\downarrow(y, dy)}{dy} ds.\end{aligned}$$

Finding the boundary condition – continued

If $(B_t)_{t \geq 0}$ is a standard Brownian motion and $T := \inf\{t \geq 0 : B_t = 0\}$, then

$$\begin{aligned} \lim_{y \downarrow 0} \int_0^\infty \frac{\mathbb{P}^y\{B_s \in dy, T > s\}}{dy} ds &= \lim_{y \downarrow 0} \int_0^\infty \frac{1}{\sqrt{2\pi s}} - \frac{1}{\sqrt{2\pi s}} e^{-(2y)^2/2s} ds \\ &= 2y. \end{aligned}$$

By Itô–McKean theory,

$$\frac{\theta}{2} \lim_{y \downarrow 0} \int_0^\infty \frac{Q_s^\downarrow(y, dy)}{dy} ds = \theta \lim_{y \downarrow 0} \frac{y}{b(y)}.$$

Conclusion for the time-homogeneous case

Put $f(x, t) :=$ for the frequency spectrum of the model with mutation from ancestral type, with everything time-homogeneous.

Note that $f(x, t) = f^o(x, t) + \phi_t(x)$.

Hence

$$\frac{\partial}{\partial t} f(x, t) = -\frac{\partial}{\partial x} [a(x) f(x, t)] + \frac{1}{2} \frac{\partial^2}{\partial x^2} [b(x) f(x, t)].$$

with $\lim_{x \downarrow 0} x f(x, t) = \theta \lim_{x \downarrow 0} \frac{x}{b(x)}$ and $\lim_{x \uparrow 1} f(x, t)$ finite.

Conclusion for the time-inhomogeneous case

Suppose that everything is allowed to depend on time.

Write $f(x, t)$ for the frequency spectrum.

Allowing a , b , θ and ρ to be **piecewise constant**, using the above analysis, and then **taking limits** gives

$$\frac{\partial}{\partial t} f(y, t) = -\frac{\partial}{\partial y} [a(y, t) f(y, t)] + \frac{1}{2} \frac{\partial^2}{\partial y^2} [b(y, t) f(y, t)]$$

with $\lim_{y \downarrow 0} y f(y, t) = \theta(t) \lim_{x \downarrow 0} \frac{x}{b(x, t)}$ and $\lim_{y \uparrow 1} f(y, t)$ finite.

Structured populations

The same approach works for **systems** of populations interacting via **migration**.

A single PDE with boundary conditions is replaced by a family of coupled PDE with boundary conditions.

Example

Take $\theta(t) = \theta$, $a(x, t) = Sx(1 - x)$, and $b(x, t) = x(1 - x)/\rho(t) \equiv$ Wright-Fisher diffusion with constant mutation, additive selection, and varying population size ($2N\rho(t)$ in generation $2Nt$).

Set $g(x, t) := x(1 - x)f(x, t)$.

The forward equation is

$$\frac{\partial}{\partial t} g(x, t) = -Sx(1 - x) \frac{\partial}{\partial x} [g(x, t)] + \frac{x(1 - x)}{2\rho(t)} \frac{\partial^2}{\partial x^2} [g(x, t)]$$

with $\lim_{x \downarrow 0} g(x, t) = \theta\rho(t)$.

Example – continued

Put $\mu_n(t) := \int_0^1 x^n g(x, t) dx = \int_0^1 x^n x(1-x)f(x, t) dx$.

Integrating by parts gives the coupled system of ODEs

$$\mu_0'(t) = \frac{\theta}{2} - \frac{1}{\rho(t)}\mu_0(t) + S(\mu_0(t) - 2\mu_1(t))$$

and

$$\begin{aligned} \mu_n'(t) = & \frac{1}{2\rho(t)} [(n+1)n\mu_{n-1}(t) - (n+2)(n+1)\mu_n(t)] \\ & + S((n+1)\mu_n(t) - (n+2)\mu_{n+1}(t)), \quad n \geq 1. \end{aligned}$$

Frequency spectrum in a finite sample

In a sample of n chromosomes the expected number of loci at which the derived allele is found on i chromosomes is

$$\begin{aligned} f_i(t) &= \binom{n}{i} \int_0^1 x^i (1-x)^{n-i} f(x, t) dx \\ &= \binom{n}{i} \sum_{j=0}^{n-i-1} (-1)^j \binom{n-i-1}{j} \mu_{j+i-1}(t). \end{aligned}$$

Recent human population growth

Assume an effective size $N_0 = 10^4$ until 1.5×10^5 ya ($t = 0$).

Assume a generation time of 25 years.

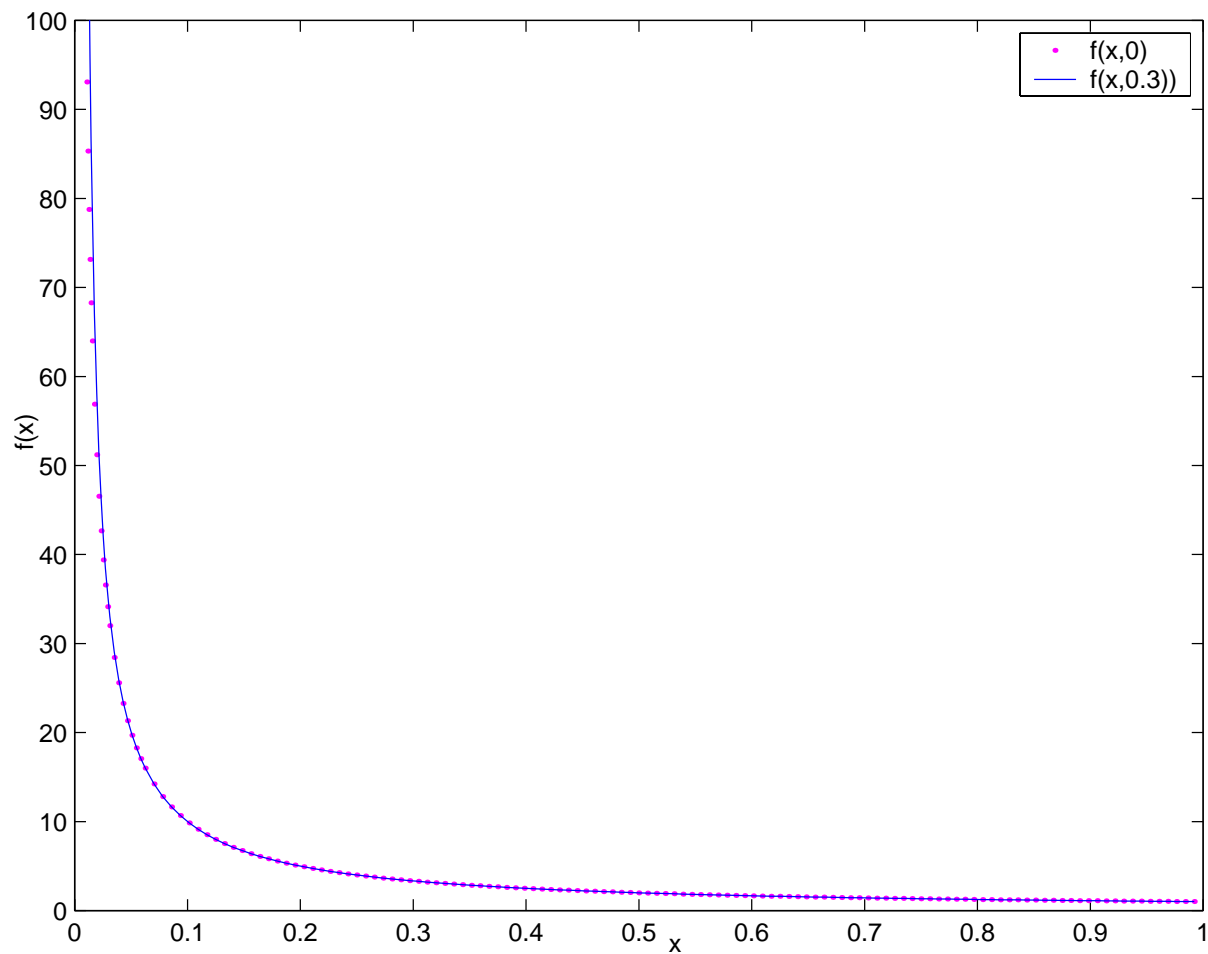
Measuring time in units of $2N_0$, the present is at $t = 0.3$.

The current effective population size is $1.6 \times 10^9 = 10^5 \times e^{40 \times 0.3}$.

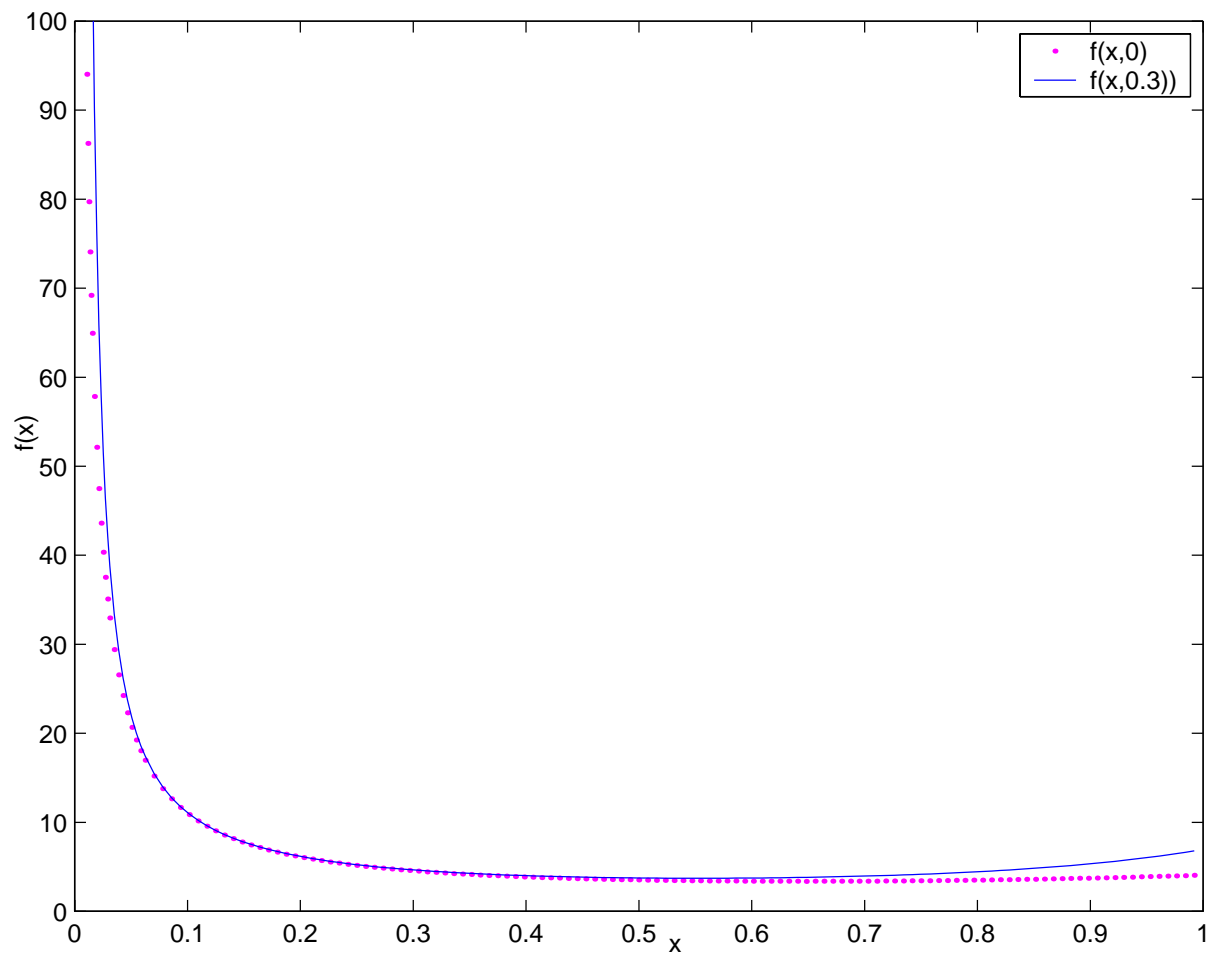
Assume exponential growth, so $\rho(t) = e^{40t}$, and $\theta(t) = 1$.

Assume that the spectrum at $t = 0$ is the equilibrium

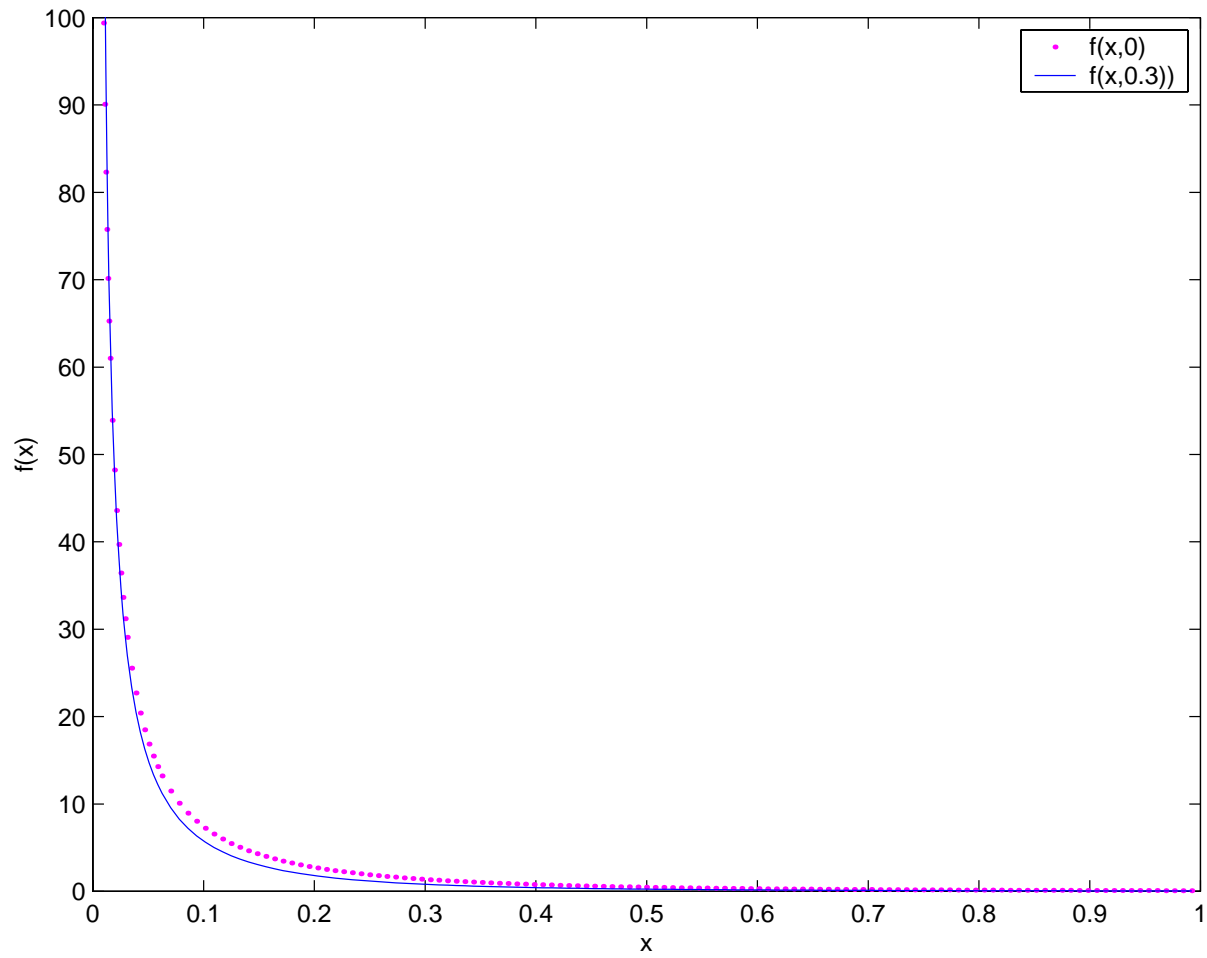
$$f(x, 0) = \frac{e^{2S} (1 - e^{-2S(1-x)})}{(e^{2S} - 1) x(1-x)}$$



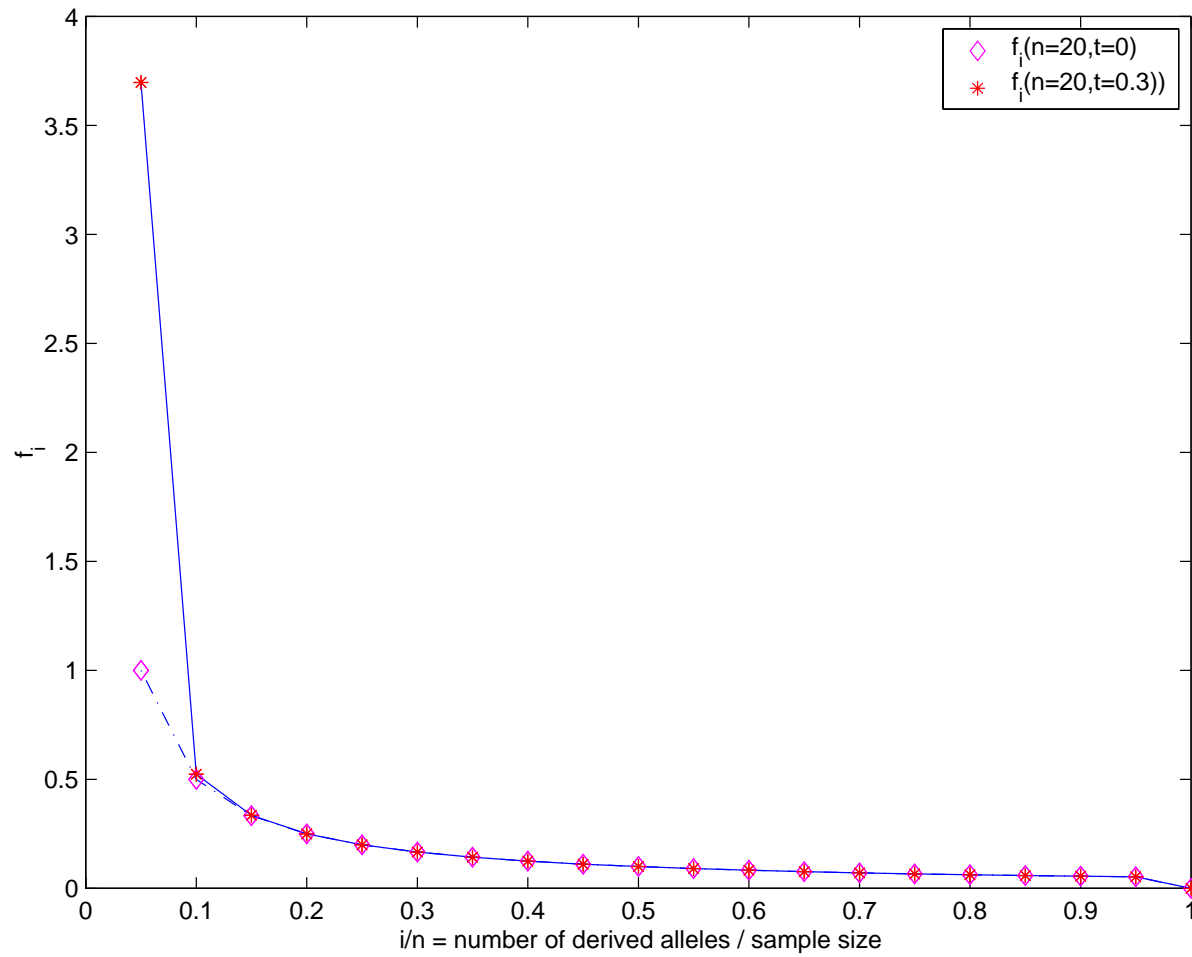
Plots of $f(x, t)$ for $S = 0$



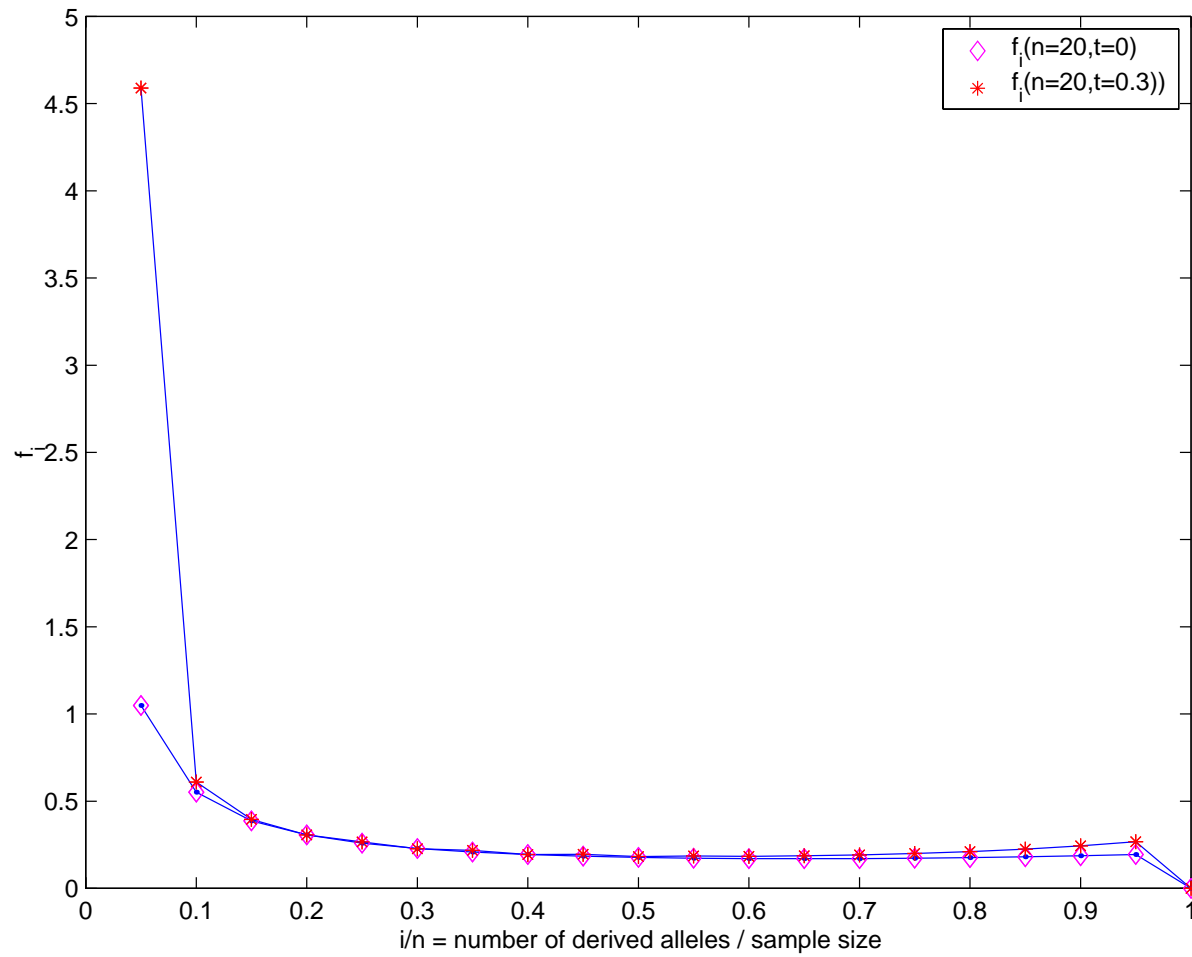
Plots of $f(x, t)$ for $S = +2$



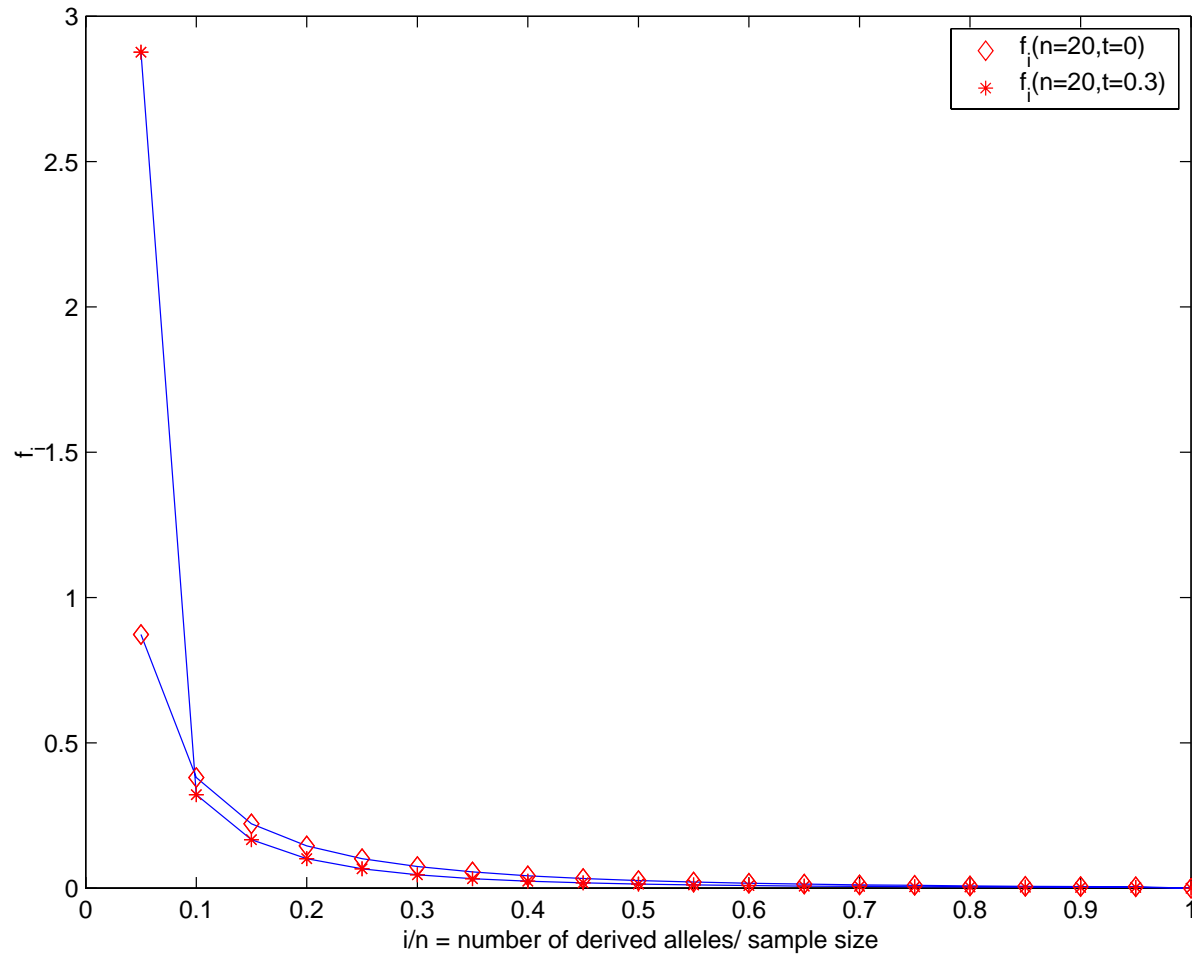
Plots of $f(x, t)$ for $S = -2$



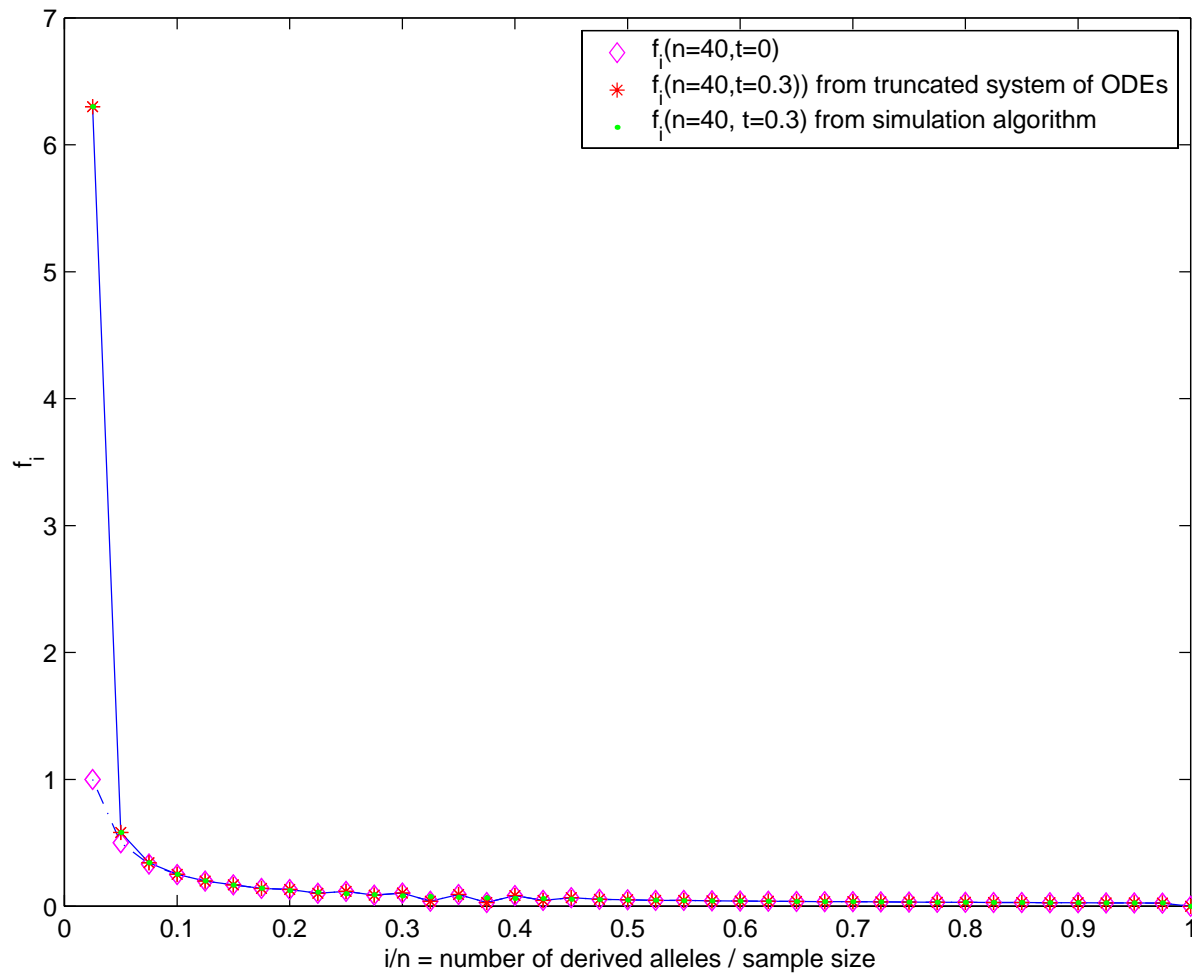
Plots of $f_i(t)$ for $S = 0$ and $n = 20$



Plots of $f_i(t)$ for $S = +2$ and $n = 20$



Plots of $f_i(t)$ for $S = -2$ and $n = 20$



Plots of $f_i(t)$ for $S = 0$ and $n = 40$

*References

Bustamante, C. D., Wakeley, J., Sawyer, S., Hartl, D. L., 2001. Directional selection and the site-frequency spectrum. *Genetics* 159 (4), 1779–1788.

Fisher, R. A., 1930. The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh* 50, 205–220.

Griffiths, R. C., 2003. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoretical Population Biology* 64 (2), 241–251.

Griffiths, R. C., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. *Stochastic Models* 14, 273–295.

Kimura, M., 1964. Diffusion models in population genetics. *Journal of Applied Probability* 1, 177–232.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61, 893–903.

Nei, M., Maruyama, T., Chakraborty, R., 1975. The bottleneck effect and genetic variability in populations. *Evolution* 29, 1–10.

Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*. 154 (2), 931–942.

Polanski, A., Kimmel, M., Sep 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 165 (1), 427–436.

Reich, D. E., Lander, E. S., 2001. On the allelic spectrum of human disease. *Trends in Genetics* 17 (9), 502–510.

Sawyer, S. A., Hartl, D. L., 1992. Population genetics of polymorphism and divergence. *Genetics* 132 (4), 1161–1176.

Tajima, F., 1989. The effect of change in population size on DNA polymorphism. *Genetics* 123 (3), 597–602.

Williams, D., 1974. Path decomposition and continuity of local time for one-dimensional diffusions. I. *Proc. London Math. Soc.* (3) 28, 738–768.

Williamson, S., Fledel-Alon, A., Bustamante, C. D., September 1, 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* 168 (1), 463–475.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., Bustamante, C. D., May 19, 2005. Simultaneous

inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences USA*. 102, 7882–7887.

Wooding, S., Rogers, A., 2002. The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics* 161 (4), 1641–50.

Wright, S., 1938. The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the United States of America* 24, 253–259.