# Establishment in a new habitat by polygenic adaptation

Nick Barton
Alison Etheridge


# A mathematical prelude: the infinitesimal model

*Luminy, June 2018*

# The basic model

$$\text{Trait value} = \underbrace{\text{genetic}}_{Z} + \underbrace{\text{non-genetic}}_{E}$$

For today's purposes we ignore environmental component $E$.

Genetic component normally distributed; mean average of values in parents;

$$Z \sim \mathcal{N}\left(\frac{z_1 + z_2}{2}, V_0\right)$$

In a large outcrossing population, $V_0 =$ constant, otherwise decreases in proportion to relatedness.

# The simplest case

Large outcrossing population. $Z \sim \mathcal{N}\left(\frac{z_1+z_2}{2}, V_0\right)$.

With purely random mating (neutral trait), the population as a whole rapidly converges to a Gaussian with variance $2V_0$ (Bulmer).

If variance in parental population is $V_1$, that of offspring is

$$\frac{V_1}{2} + V_0,$$

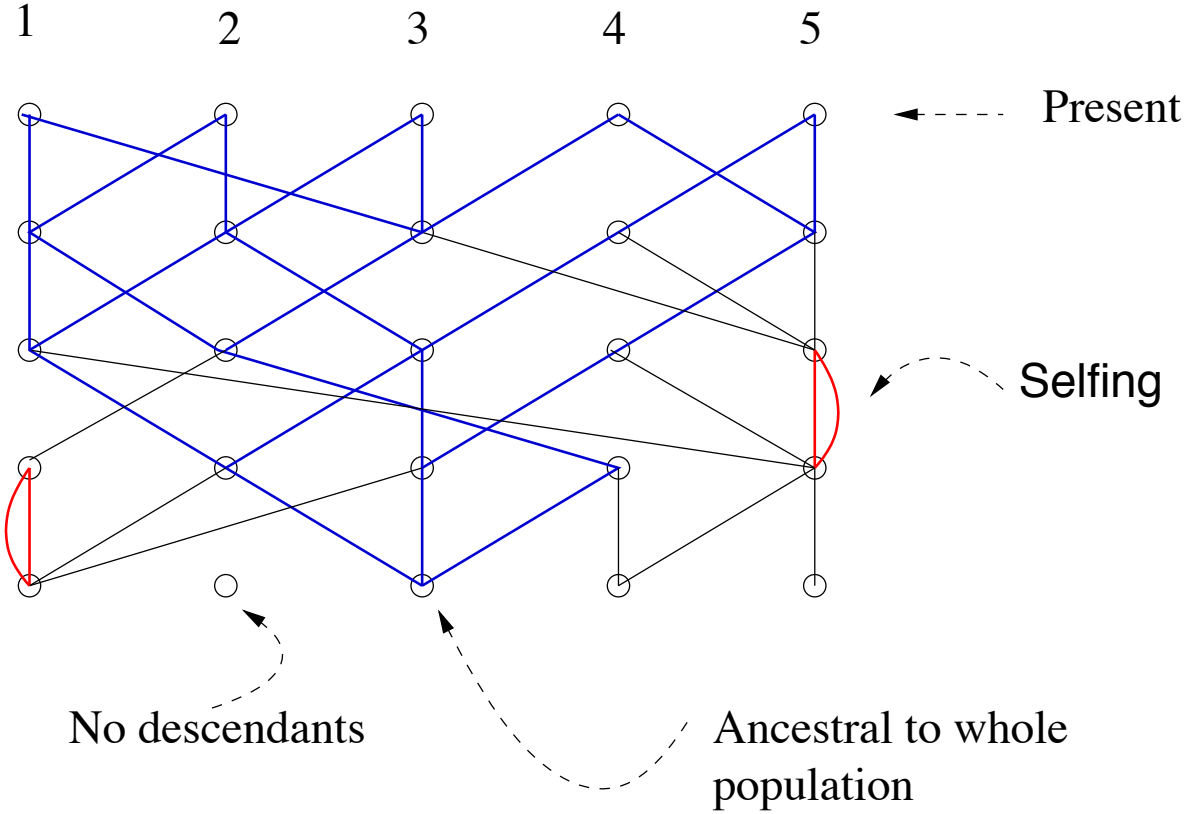
At equilibrium, this is $V_1$, so $V_1 = 2V_0$.

Half variance is within families, half between.

IN GENERAL THE INFINITESIMAL MODEL ONLY SAYS THAT THE GENETIC COMPONENTS WITHIN FAMILIES ARE NORMALLY DISTRIBUTED. THE DISTRIBUTION ACROSS THE WHOLE POPULATION MAY BE FAR FROM NORMAL.

# Pedigrees

Each individual has *two* parents in the previous generation.



# An aside on common ancestors

**Theorem (Chang 1999)**
Let $\tau_N$ be time to MRCA of population size $N$ evolving according to diploid Wright-Fisher model (fixed population size, parents picked uniformly at random with replacement).

$$\frac{\tau_N}{\log_2 N} \xrightarrow{\mathbb{P}} 1 \quad \text{as } N \to \infty.$$

**Theorem (Chang 1999)**
Let $\mathcal{U}_N$ be time until all ancestors are either common to whole population or have no surviving progeny.

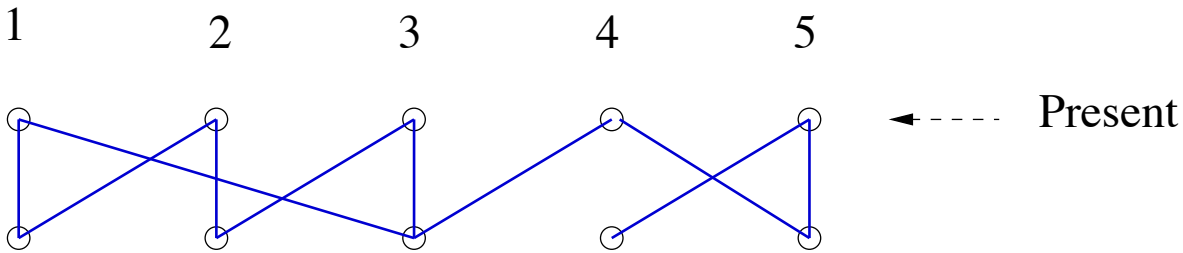
$$\frac{\mathcal{U}_N}{1.77 \log_2 N} \xrightarrow{\mathbb{P}} 1 \quad \text{as } N \to \infty.$$

There are many routes through the pedigree from ancestor to present.

# The pedigree as matrices

Pedigree spanning $t$ generations $\Leftrightarrow$ random matrices $M_0, \ldots, M_{t-1}$.

The $i$th row of $M_t$ specifies parents of individual labelled $i$ in generation $t$ before the present.

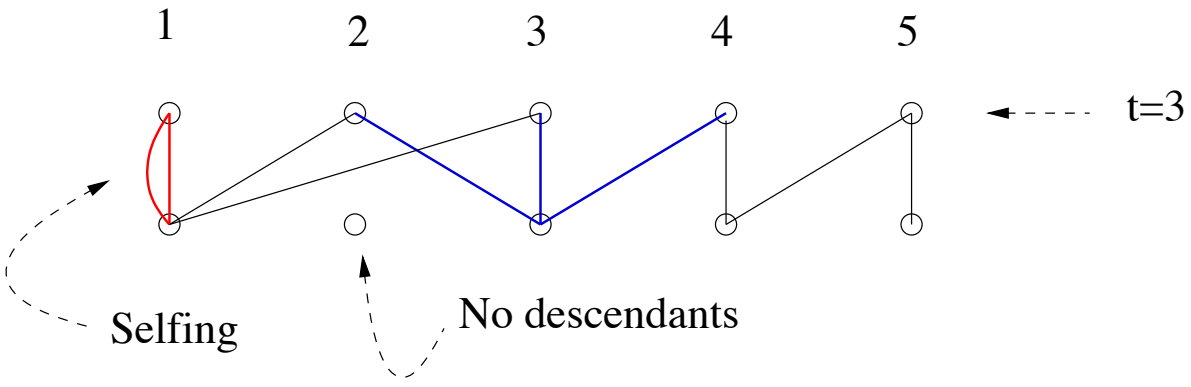



No need for constant population size

$$M_0 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

# Selfing

. . . or when there is selfing




$$M_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

# The probability of identity

Write $F_{ij}(t)$ for probability homologous genes in individuals labelled $i$, $j$ in generation $t$ descend from the same ancestral gene.

- Haploid case

$$F_{ij}(t) = \sum_{k,l} M_{ik}(t) M_{jl}(t) F_{kl}(t-1),$$

- Diploid case

$$F_{ij}(t) = \sum_{k,l} M_{ik}(t) M_{jl}(t) F^*_{kl}(t),$$

$$F^*_{kl} = F_{kl} \text{ if } k \neq l, \quad F^*_{kk} = \frac{1}{2}\left(1 + F_{kk}\right).$$

# The infinitesimal model

Let

1. $\mathcal{P}^{(t)}$ denote the *pedigree* relationships between all individuals up to and including generation $t$;
2. $Z^{(t)}$ denote the *traits* of all individuals in the pedigree up to and including the $t$th generation.

The distribution of trait values in generation $t$, conditional on knowing $\mathcal{P}^{(t)}$ and $Z^{(t-1)}$, is multivariate normal.

THIS IS A STATEMENT ABOUT DISTRIBUTION WITHIN FAMILIES, NOT ACROSS THE WHOLE POPULATION

The population need not be neutral. The pedigree captures selection, population structure etc.

# Why might it be a reasonable model?

Additive traits in haploids (no mutation)
$M$ = number of (unlinked) loci affecting trait.

- Trait value in individual $j$:

$$Z_j = \bar{z}_0 + \sum_{l=1}^{M} \frac{1}{\sqrt{M}} \eta_{jl},$$

where $\bar{z}_0$ = average value in ancestral population.

# Why might it be a reasonable model?

Additive traits in haploids (no mutation)
$M$ = number of (unlinked) loci affecting trait.

- Trait value in individual $j$:

$$Z_j = \bar{z}_0 + \sum_{l=1}^{M} \frac{1}{\sqrt{M}} \eta_{jl},$$

where $\bar{z}_0$ = average value in ancestral population.

- Ancestral population. $\hat{\eta}_{jl}$ i.i.d (for different $j$), say.

# Reproduction

$[1]$ and $[2]$ refer to the first and second parents of an individual.

- $\eta_{jl}[1]$ is the scaled allelic effect at locus $l$ in the first parent of the $j$th individual. Similarly, $Z_j[1]$ will denote the trait value of the first parent of individual $j$.

- Write $X_{jl} = 1$ if the allelic type at locus $l$ in the $j$th individual is inherited from the 'first parent' of that individual; otherwise it is zero. $\mathbb{P}[X_{jl} = 1] = 1/2 = \mathbb{P}[X_{jl} = 0]$.

$$Z_j = \bar{z}_0 + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \{X_{jl}\eta_{jl}[1] + (1 - X_{jl})\eta_{jl}[2]\}\,.$$

# Conditioning

We would like to derive the distribution of trait values in generation $t$ conditional on knowing $\mathcal{P}^{(t)}$ and $Z^{(t-1)}$.

$$Z_j = \bar{z}_0 + \frac{1}{\sqrt{M}} \sum_{l=1}^{M} \{X_{jl}\eta_{jl}[1] + (1 - X_{jl})\eta_{jl}[2]\}.$$

Key: Need to be able to calculate the distribution of $\eta_{jl}[1]$ *conditional on* $Z^{(t-1)}$ and show that it is almost unaffected by the conditioning.

Then $\mathbb{E}[(\eta_{jl}^{[1]} - \eta_{jl}^{[2]})^2] \approx 2(1 - F_{[1][2]})\mathrm{var}(\hat{\eta}_l) \rightsquigarrow$ variance among offspring reduced proportional to probability of identity.

# A toy example

Suppose $\eta_l$ are i.i.d. with $\eta_l = \pm 1$ with equal probability, $\bar{z}_0 = 0$.

$$
\begin{aligned}
\mathbb{P}[\eta_1 = 1 | Z = k/\sqrt{M}] &= \frac{\mathbb{P}\left[\sum_{l=1}^{M} \eta_l = k \middle| \eta_1 = 1\right]}{\mathbb{P}\left[\sum_{l=1}^{M} \eta_l = k\right]} \mathbb{P}\left[\eta_1 = 1\right] \\
&= \frac{\mathbb{P}\left[\sum_{l=2}^{M} \eta_l = (k-1)\right]}{\mathbb{P}\left[\sum_{l=1}^{M} \eta_l = k\right]} \mathbb{P}\left[\eta_1 = 1\right] \\
&= \frac{\frac{1}{2^{M-1}}}{\frac{1}{2^{M}}} \frac{\binom{M-1}{(M+k-2)/2}}{\binom{M}{(M+k)/2}} \mathbb{P}\left[\eta_1 = 1\right] \\
&= \left(1 + \frac{k}{M}\right) \mathbb{P}\left[\eta_1 = 1\right].
\end{aligned}
$$

# Toy example continued

If scaled allelic effects are i.i.d. Bernoulli,

$$\mathbb{P}\left[\eta_1 = 1 \middle| Z = \frac{k}{\sqrt{M}}\right] = \left(1 + \frac{k}{M}\right)\mathbb{P}\left[\eta_1 = 1\right].$$

For a 'typical' trait value, $k/M = \mathcal{O}(1/\sqrt{M})$.

For extreme values ($k = \pm M$), the trait gives complete information about the allelic effect at each locus.

For 'typical' $k$, the distribution of $\eta_1$ is almost unchanged because there are so many different configurations of allelic effects that correspond to the same trait value.

# The infinitesimal model

Conditional on $\mathcal{P}^{(t)}$ and $Z^{(t-1)}$,

$$\left( Z_j - \frac{Z_j[1] + Z_j[2]}{2} \right)_{j=1,\ldots,N_t}$$

converges (in distribution) to mean zero multivariate normal with diagonal covariance matrix $\Sigma_t$.

$(\Sigma_t)_{jj}$ = *segregation variance* among offspring of the parents of individual $j$.

OVER TO NICK

Can a population establish in a new habitat ?

- migration from a source population
  - evolutionary rescue/sympatric speciation
- growth requires adaptation
- chance that a single migrant establishes
- time to establishment with steady migration
- stationary distribution of trait & N

Use the *infinitesimal model*

Barton, Etheridge & Véber, Theor. Pop. Biol. 2018
Barton & Etheridge, Theor. Pop. Biol. 2018

Can a population establish in a new habitat ?

- growth rate depends on a trait, z
- Poisson # of offspring, mean $e^{\beta z}$
- under the *infinitesimal model,*

  offspring have mean ~ midparent

  variance ~ V(1 - F)
- large source population has variance 2V, F=0

Haploid parents i,j: $V(1 - F_{i,j})$

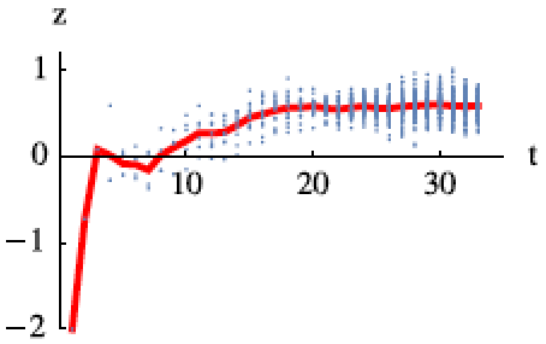
Diploid parents i,j: $V(1 - (F_{i,i}+F_{j,j})/2)$

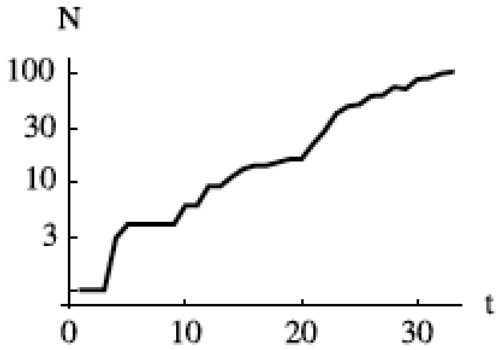
# What is the chance that one individual establishes?

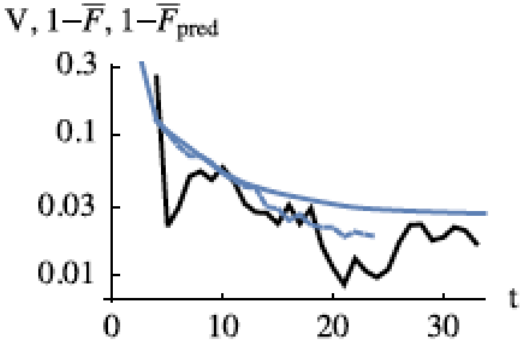
- Random mating, including selfing
  - ignore inbreeding depression

Variance in source is 2V = 1

Source mean is $z_0 = -2$, $\beta = 0.25$; $e^{-\beta z} = 0.61$

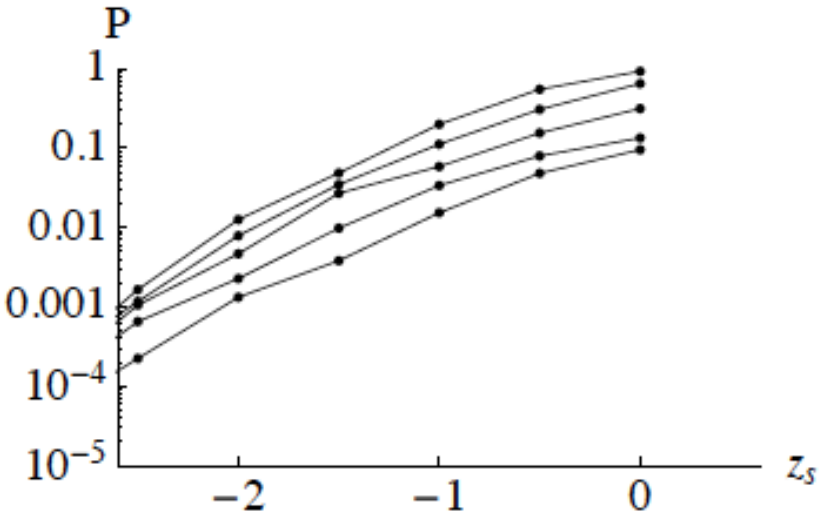


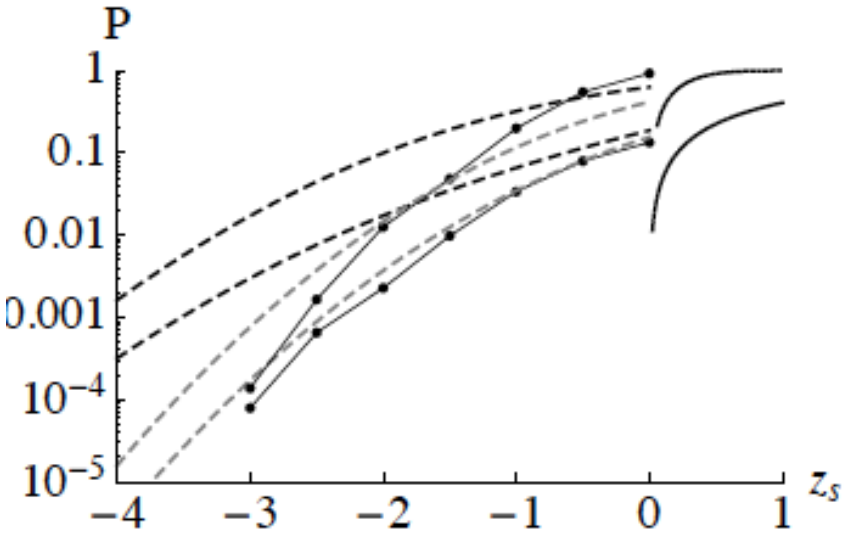



# What is the chance that one individual establishes?

- Random mating, including selfing
  - ignore inbreeding depression



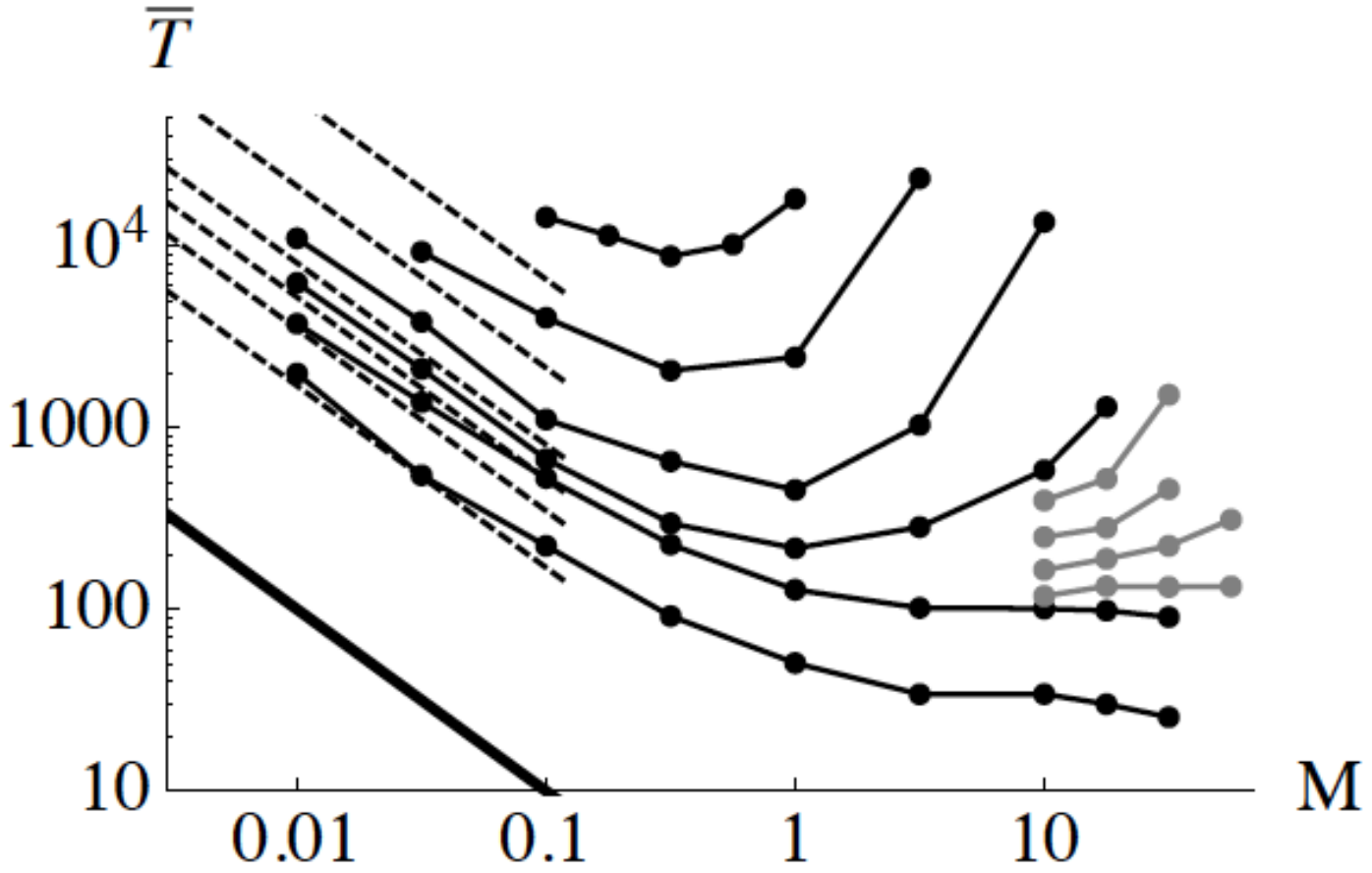

β = 0.125, 0.25,…, 1, 2 (bottom to top)




β = 0.25, 2 (bottom to top)
Solid curves at right: homozygous
Solid curve (dotted): individual, z
Dashed curve: random draw, mean z
Grey dashed: random draw, mean z, homozygous

Time to establishment with migration rate M

$\beta = 0.25$, $z_s = -1, -1.5, \ldots, -3.5$ (bottom to top)




For $z_s < -1.57$, migration pulls the sink popl'n back

# Recap of Nick's model

Large source population, trait values $\sim \mathcal{N}(\bar{z}_s, 2V)$.

$M$ (unrelated) migrants enter population in each generation.

$N(t)$ population size in generation $t$, $\overline{z}(t)$ mean trait value.

Before migrants arrive, number in next generation Poisson with expectation $N(t)\overline{W}$, where $\overline{W}$ mean fitness across offspring of random matings.

# Nick's model continued

Offspring of individuals $i$, $j$, have mean trait value given by the midparent value, variance:

- haploid parents $V_{ij} = V(1 - F_{ij})$,
- diploid parents $V_{ij} = V(1 - (F_{ii} + F_{jj})/2)$.

Fitness if trait value $z$ is $e^{\beta z}$, so

$$N(t)\overline{W} = \frac{1}{N(t)} \sum_{i,j} \exp\left(\beta \frac{(z_i + z_j)}{2} + \frac{\beta^2}{2} V_{ij}\right).$$

# A 'deterministic' model

Assume that the trait distribution across the whole population is Gaussian. NOT a consequence of using the infinitesimal model.

First approximation: suppose population size and trait mean/variance evolve deterministically.

Each diploid migrant carrries half of the genetic variance in the source population, so modest rates of migration into a small 'sink' population can maintain high genetic variance.

Denote within family variance by $V^*$, assumed constant irrespective of origin of parents. (i.e. Assume $F = 0$, but can be a bit more sophisticated. Recall variance across population will then be $2V^*$.)

# A recursion

The distribution of traits across the population $\sim \mathcal{N}(\bar{z}, 2V^*)$, so

$$\overline{W} = \exp\left(\beta\bar{z} + \beta^2 V^*\right),$$

After reproduction and the subsequent migration,

$$N(t+1) = M + N(t)\exp\left(\beta\bar{z}(t) + \beta^2 V^*\right);$$

$$\bar{z}(t+1) = \frac{1}{N(t+1)}\left(M\bar{z}_s + N(t)\mathbb{E}[ze^{\beta z}]\right),$$

(expectation is w.r.t. distribution of trait among offspring before selection, calculated by differentiating $\overline{W}$ w.r.t. $\beta$).

# New coordinates

$$N(t+1) = M + N(t)\exp\left(\beta\bar{z}(t) + \beta^2 V^*\right);$$

$$\bar{z}(t+1) = \bar{z}(t) + 2\beta V^*\left(1 - \frac{M}{N(t+1)}\right) - \frac{M}{N(t+1)}\left(\bar{z}(t) - \bar{z}_s\right).$$

Set $n = N/M$, $\alpha = \beta\sqrt{2V^*}$ and $y = (\bar{z} - \bar{z}_s)/\sqrt{2V^*}$.

$$n(t+1) = 1 + n(t)W_s e^{\alpha y(t)}, \quad y(t+1) = (y(t) + \alpha)\left(1 - \frac{1}{n(t+1)}\right),$$

$$W_s = \exp\left(\beta\bar{z}_s + \beta^2 V^*\right)$$

(mean growth rate of the source population in the new conditions)

# Critical behaviour

$$n(t+1) = 1 + n(t)W_s e^{\alpha y(t)}, \quad y(t+1) = (y(t) + \alpha)\left(1 - \frac{1}{n(t+1)}\right),$$

- If $W_s > W_{s,\mathrm{crit}}$, population size and trait increase together, regardless of $M$.
- If $W_s < W_{s,\mathrm{crit}}$, two equilibria, one stable and one unstable. Population may be unable to grow, regardless of how large is $M$; instead, it is maintained by migration as a poorly adapted 'sink'.

# The critical value

$$n(t+1) = 1 + n(t)W_s e^{\alpha y(t)}, \quad y(t+1) = (y(t)+\alpha)\left(1 - \frac{1}{n(t+1)}\right),$$

At equilibrium $y(t) = y(t+1) = \alpha(n-1)$, i.e., $y_{\text{crit}} = \alpha(n_{\text{crit}} - 1)$.

Writing $f(n) = 1 + nW_s e^{\alpha^2(n-1)}$, must solve

$$n = f(n), \qquad 1 = f'(n).$$

Yields quadratic in $n$, whose positive solution is

$$n_{\text{crit}} = \frac{\alpha^2 + \sqrt{\alpha^4 + 4\alpha^2}}{2\alpha^2} = \frac{1}{2}\left(1 + \sqrt{1 + 4/\alpha^2}\right).$$

# Back to original variables

$$N_{\rm crit} = \frac{M}{2}\left(1+\sqrt{1+2/(\beta^2 V^*)}\right),$$

$$W_{s,\rm crit} = \frac{n_{\rm crit}-1}{n_{\rm crit}} e^{-\alpha^2(n_{\rm crit}-1)} = \left(1-\frac{M}{N_{\rm crit}}\right) e^{-\alpha^2(N_{\rm crit}-M)/M},$$

$$\beta\bar{z}_{s,\rm crit} = -\frac{1}{2}\alpha\left(\sqrt{4+\alpha^2}\right) - \log\left(\frac{\alpha+\sqrt{4+\alpha^2}}{-\alpha+\sqrt{4+\alpha^2}}\right).$$

For $\alpha = \beta\sqrt{2V^*} \ll 1$, $\beta\bar{z}_{s,\rm crit} \sim -2\alpha$.

For $\alpha \gg 1$, $\beta\bar{z}_{s,\rm crit} \approx -\alpha^2/2 - 2\log\alpha$.

# A continuous time approximation

$$N(t+1) = M + N(t)\exp\left(\beta\bar{z}(t) + \beta^2 V^*\right);$$

$$\bar{z}(t+1) = \bar{z}(t) + 2\beta V^*\left(1 - \frac{M}{N(t+1)}\right) - \frac{M}{N(t+1)}\left(\bar{z}(t) - \bar{z}_s\right).$$

Assume $\beta\bar{z} + \beta^2 V^*$ is small, and ignore second order term $\beta^2 V^*$:

$$\begin{aligned}\frac{dN(t)}{dt} &= M + \beta\bar{z}(t)N(t);\\ \frac{d\bar{z}(t)}{dt} &= 2\beta V^*\left(1 - \frac{M}{2N(t)}\right) - \frac{M}{N(t)}(\bar{z}(t) - \bar{z}_s).\end{aligned}$$

In fact, to accumulate an error of order at most $\beta^2(M+N)$ per generation should also replace immigration rate $M$ in the first equation by $M(1-\beta\bar{z}(t)/2)$. With this equation the error is order $\beta\bar{z}M + \beta^2(M+N)$.

# Demographic stochasticity/sampling drift?

Add random perturbations $< \zeta_N^2 >= N$; $< \zeta_{\bar{z}}^2 >= \frac{2V^*}{N}$.

Introduce the potential, $U$:

$$U = M \log N + \beta(N - \frac{M}{2})\bar{z} - \frac{M}{4V^*}\left(\bar{z} - \bar{z}_s\right)^2.$$

$$\begin{aligned}
\frac{dN}{dt} &= N\frac{\partial U}{\partial N} + \zeta_N = M + \beta\bar{z}N + \zeta_N, \\
\frac{d\bar{z}}{dt} &= \frac{2V^*}{N}\frac{\partial U}{\partial \bar{z}} + \zeta_{\bar{z}} = 2\beta V^*\left(1 - \frac{M}{2N}\right) - \frac{M}{N}\left(\bar{z} - \bar{z}_s\right) + \zeta_{\bar{z}}.
\end{aligned}$$

# The 'stationary distribution'

If there *were* a stationary distribution, it would satisfy

$$\psi \propto \frac{e^{2U}}{N} = N^{2M-1} \exp\left(\beta(2N-M)\bar{z} - \frac{M}{2V^*}\left(\bar{z}-\bar{z}_s\right)^2\right).$$

Diverges for large $N$, $\bar{z}$; should approximate the density near to a stable 'sink' equilibrium, when that exists.

- $N^{2M-1}$, migration that increases population size;
- $e^{\beta(2N-M)\bar{z}}$, directional selection on the trait;
- $e^{-M(\bar{z}-z_s)^2/2V^*}$, gene flow that pulls the trait mean towards the source.

# More on the stationary distribution

For given $N$, the trait mean is normally distributed, with variance $V^*/M$, and mean

$$\mathbb{E}\left[\bar{z}\right] = \bar{z}_s + \beta V^*(2N - M)/M;$$

Deterministic equilibrium in which selection $2\beta V^*(1 - M/2N)$ increases the trait mean, but is opposed by gene flow at rate $M/N$.

Integrating over $\bar{z}$, distribution of $N$ proportional to

$$N^{2M-1} \exp\left(\beta^2(2N - M)^2 \frac{V^*}{2M} + \beta(2N - M)\bar{z}_s\right).$$

If $M > 1/2$ and $\bar{z}_s < -2\sqrt{V^*(1 - 1/(2M))} + \beta V^*/2 \sim -2\sqrt{V^*}$, distribution has a peak at low density, and with $\bar{z} < 0$.
Metastable 'sink' population maintained by gene flow despite maladaptation.

BACK TO NICK

# Diffusion approximation for N, z

– assuming Gaussian with constant variance $V^*$

$$dN(t) = (M + \beta\bar{z}(t)N(t))\,dt + \sqrt{N(t)}dB_t^1,$$

$$d\bar{z}(t) = \left(2\beta V^*\left(1 - \frac{M}{2N(t)}\right) - \frac{M}{N(t)}(\bar{z}(t) - \bar{z}_s)\right)dt + \sqrt{\frac{2V^*}{N(t)}}dB_t^2,$$

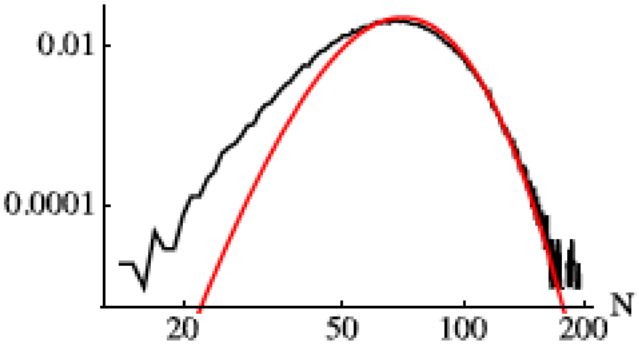
Stationary distribution for a sink population:

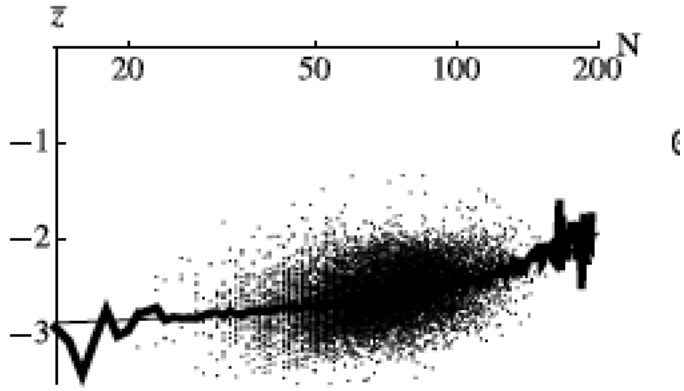
M=10

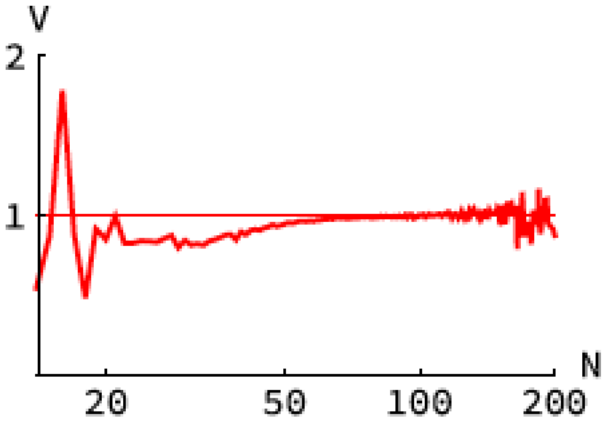
$z_s$ = -3

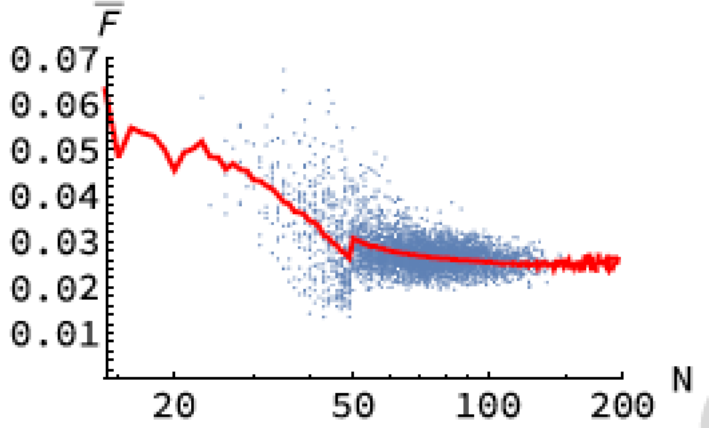
$\beta$ = 0.05

$2V^*$ = 1



# Diffusion approximation for N, z

– density regulation $-\gamma N^2$

- stabilising selection $-s(z-z_{opt})^2/2$

## Stationary distribution:

M=10

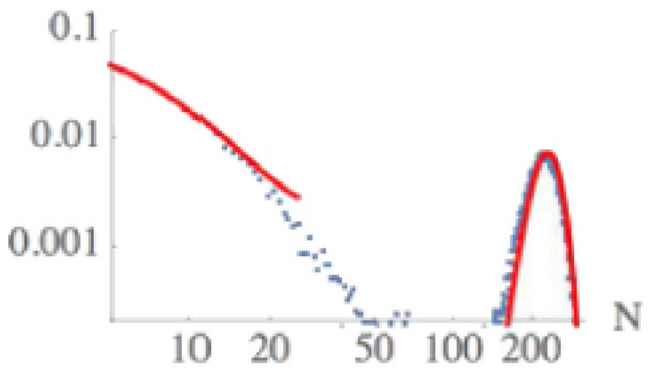
$z_s = -3$

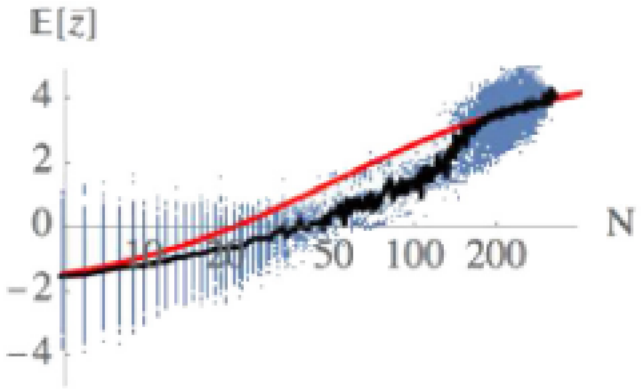
$\beta = 0.05$

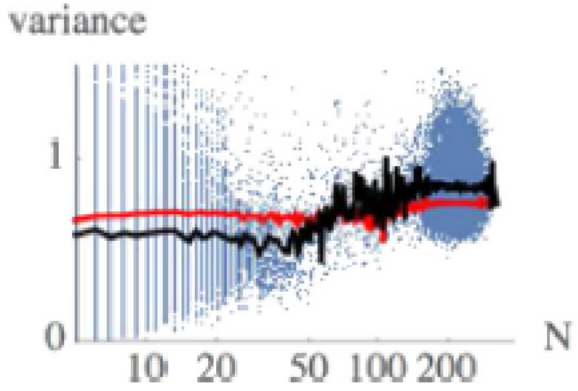
$s = 0.02$ $z_{opt} = 5$

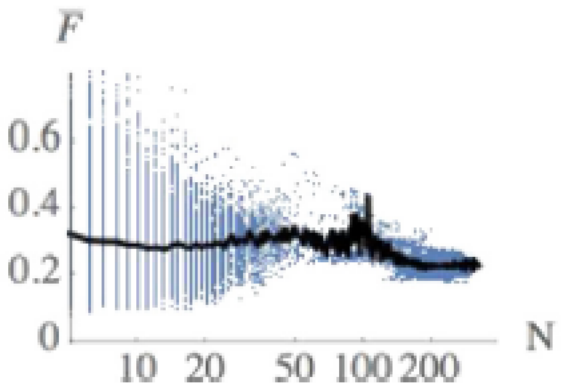
$\gamma = 0.001$

$2V^* = 1-F$



# Summary:

- one individual can establish if mean > 4 s.d. below threshold
- a ‘sink’ population may be trapped if source mean is too low
- a population can escape, and adapt to a new optimum
- it will then be (partly) reproductively isolated

- a model of speciation, due to adaptation despite gene flow