Brownian motion models, multiple characters, and phylogenies

13 August 2015.

Joe Felsenstein

NIMBioS Evol Quant Gen tutorial

What will approximate change of quantitative characters?

- ... when it occurs by genetic drift of pre-existing alleles?
- ... when it also occurs by mutation to new alleles?
- ... when variable selection affects the alleles at each locus?
- ... when selection is on the fitness based on the whole phenotype?

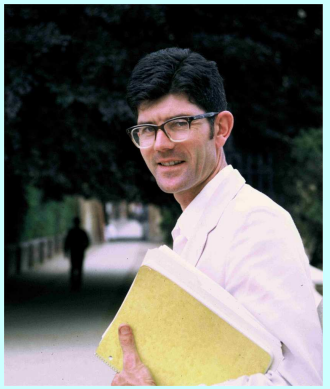
Approximating genetic drift of two alleles

Can we compute transition probabilities for genetic models such as the Wright-Fisher model?

- Can we do this analytically? **No.** (although the right eigenvectors and the eigenvalues are known) the full set of left eigenvalues has never been derived.
- We can take such a model for a given (not-too-big) population size N and compute the transition probability matrix, then either power it up numerically or get its eigenvalues and eigenvectors
- OK, what about the diffusion approximation. Aren't they very close approximations? Yes, they and Kimura (1955a, 1955b) derived transition probabilities for the diffusion process as sums of series in Gegenbaur polynomials. **But** they are difficult to work with.

Edwards and Cavalli-Sforza's approximation

Luca Cavalli-Sforza (and Edwards), 1963



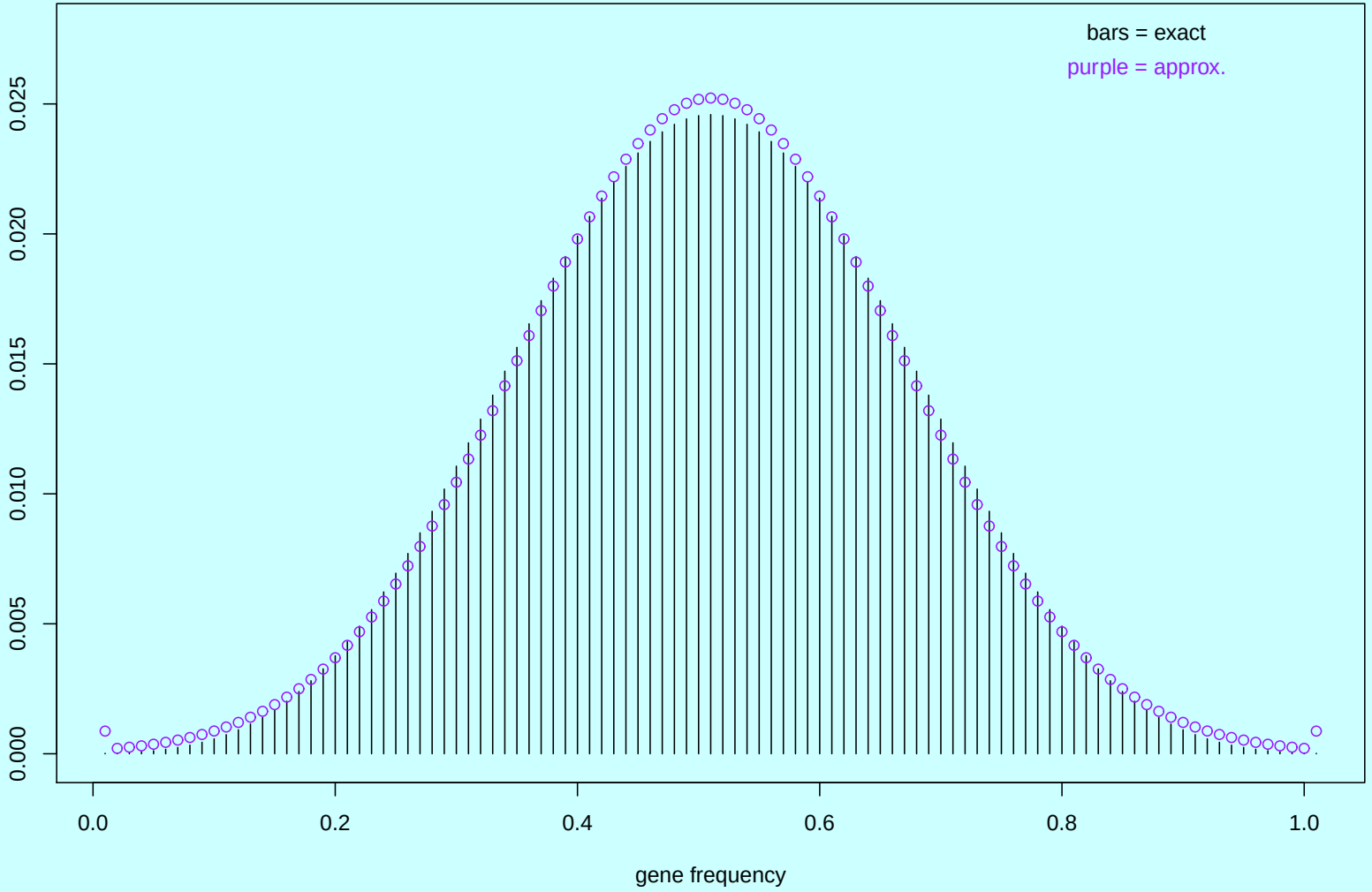
Anthony Edwards, 1970

The expectation of gene frequency change in one generation (under pure genetic drift without mutation) is zero. The variance is the binomial variance

$$E\left[(\Delta p)^2\right] = \frac{p(1-p)}{2N_e}$$

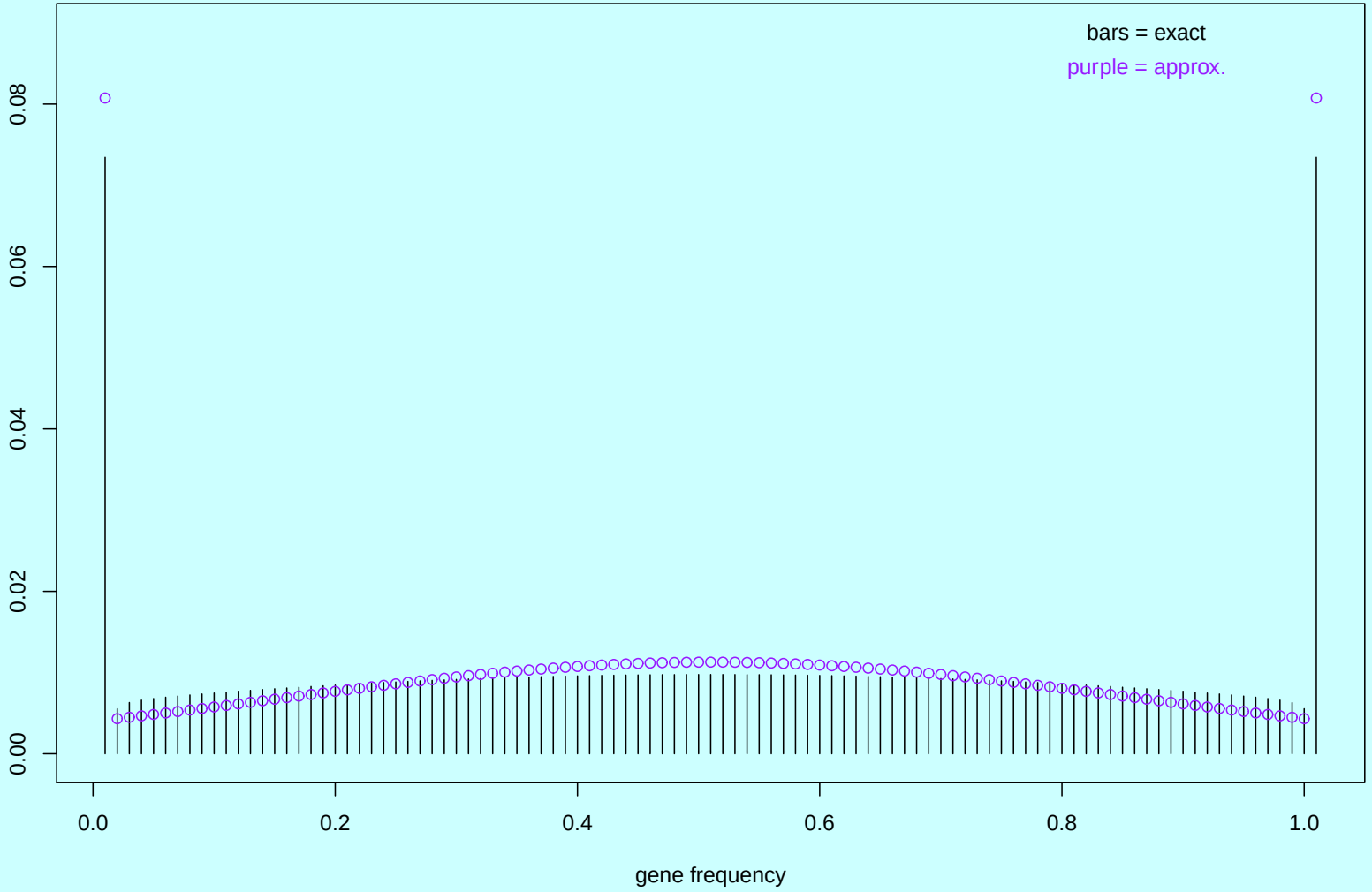
That variance is not constant: it varies with p (in a parabola), but maybe we can roughly approximate it by dealing with the case where all populations have roughly similar gene frequencies, so the variances are nearly the same. Maybe. Roughly.

How good is this?



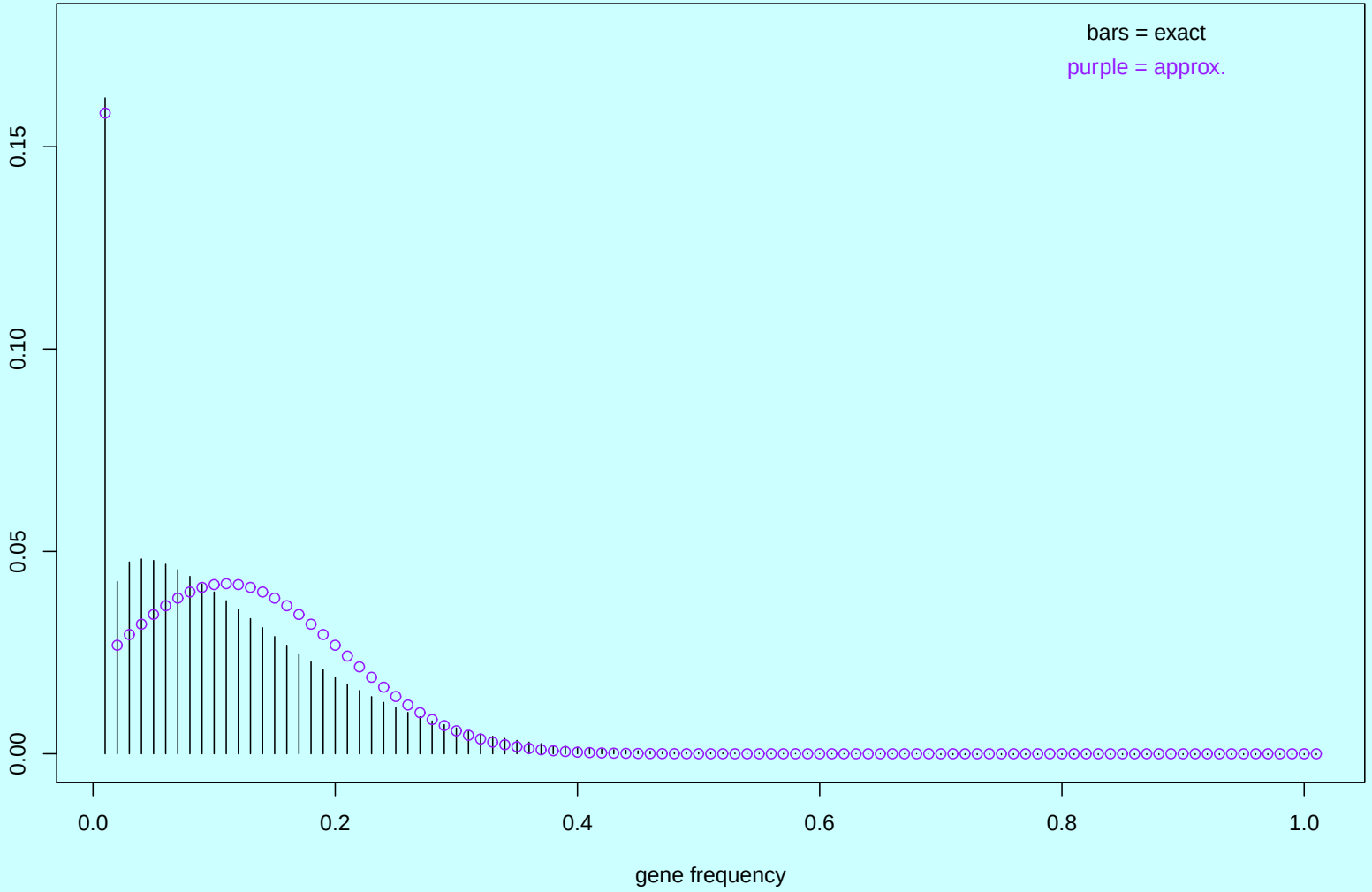
Starting with $p = 0.5$, after 10 generations in a population of size 50.

How good is this?



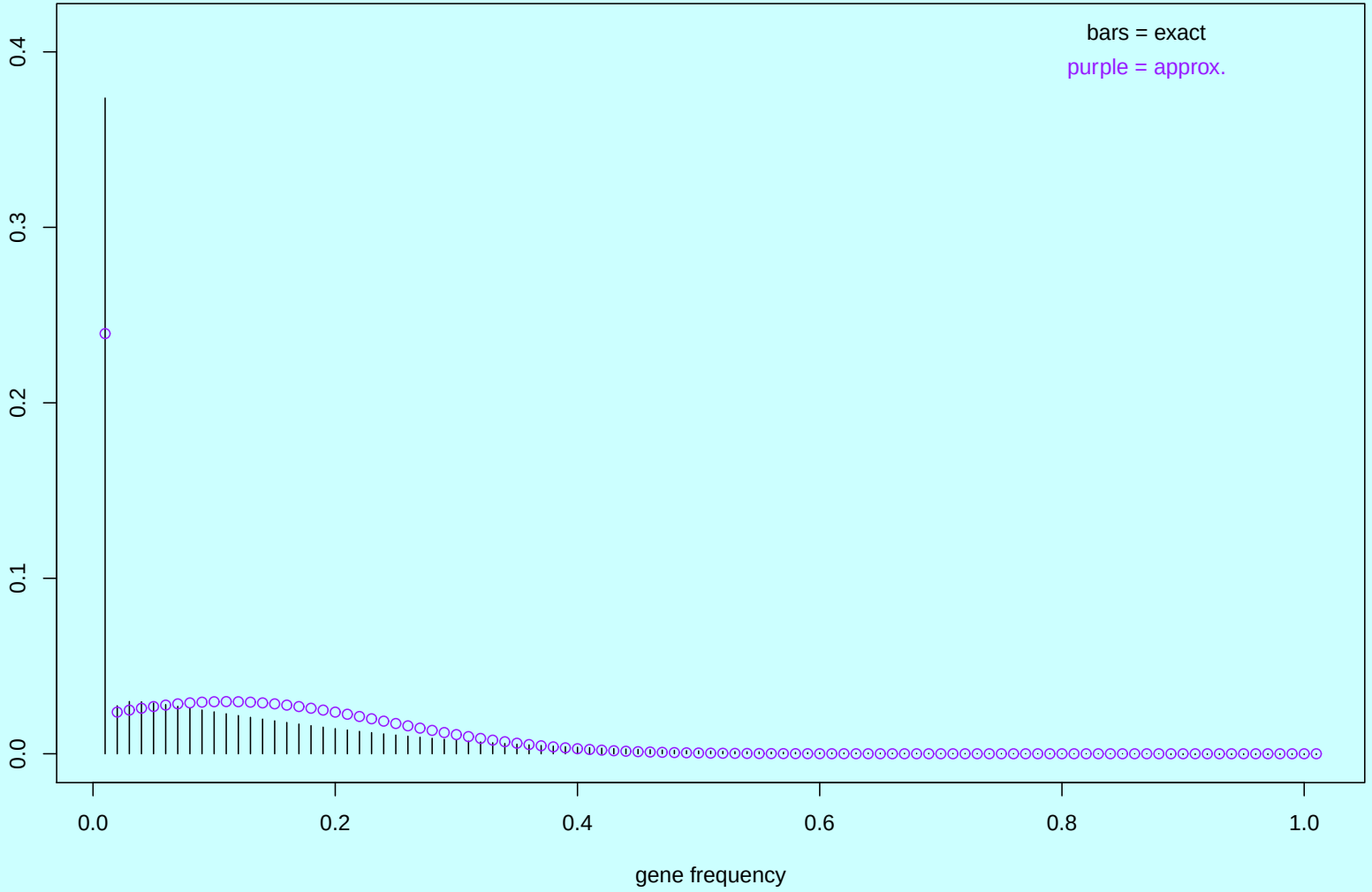
Starting with $p = 0.5$, after 50 generations in a population of size 50.

How good is this?



Starting with $p = 0.1$, after 10 generations in a population of size 50.

How good is this?



Starting with $p = 0.1$, after 20 generations in a population of size 50.

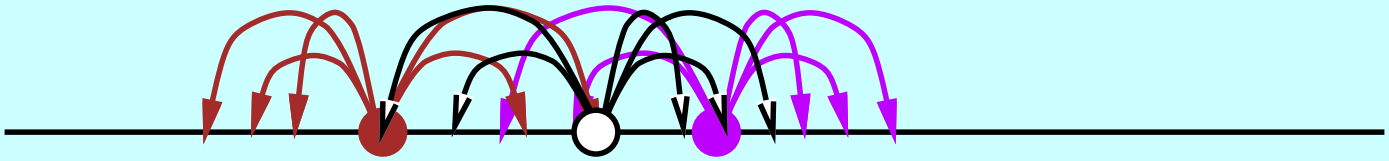
What about a quantitative character?

If a quantitative character is a sum of contributions from a number of loci, then if the individual locus gene frequencies have their change approximated by Brownian Motion, the linear combination will also change by Brownian motion. This works for multiple alleles.

- if there is any dominance, there will be some nonlinearity and the approximation will be less good.
- Epistasis can cause even more trouble.

First discussed by me (Felsenstein, 1973).

But, if there are mutations making incremental changes ..



... as we saw with the discussion of quantitative characters, if a relatively constant genetic variance is maintained, and mutations have additive effects, then genetic drift will cause the mean to change in a random walk close to Brownian Motion.

However, if one approaches some limit where most mutations oppose movement to it, and there are no mutations allowing you to go past that limit, this approximation will be poor.

Is the Brownian Motion approximation tractable?

Yes.

You can easily compute transition probabilities from one value to another.

This makes it possible to compute likelihoods on phylogenies, which allows both likelihood inference and Bayesian inference.

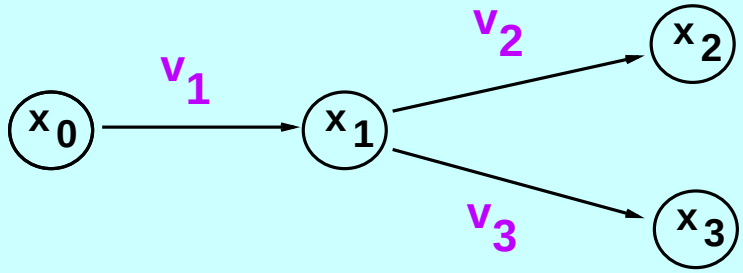
Brownian motion along a tree

Brownian motion for time t has expectation zero and variance (say) σ^2 per unit time.

So the displacement after time t is normal with expectation 0 and variance $\sigma^2 t$.

Displacements in successive periods are independent, and when they are added up, the overall expectation is still 0 and the variance is $\sigma^2 (t_1 + t_2 + t_3)$.

Brownian motion along a tree



Where $\Delta x_1 = x_1 - x_0$, $\Delta x_2 = x_2 - x_1$, etc.

Note that $x_2 = x_0 + \Delta x_1 + \Delta x_2$ and $x_3 = x_0 + \Delta x_1 + \Delta x_3$

where each of the displacements Δx_i is normally distributed, independently, with mean 0 and variance v_i.

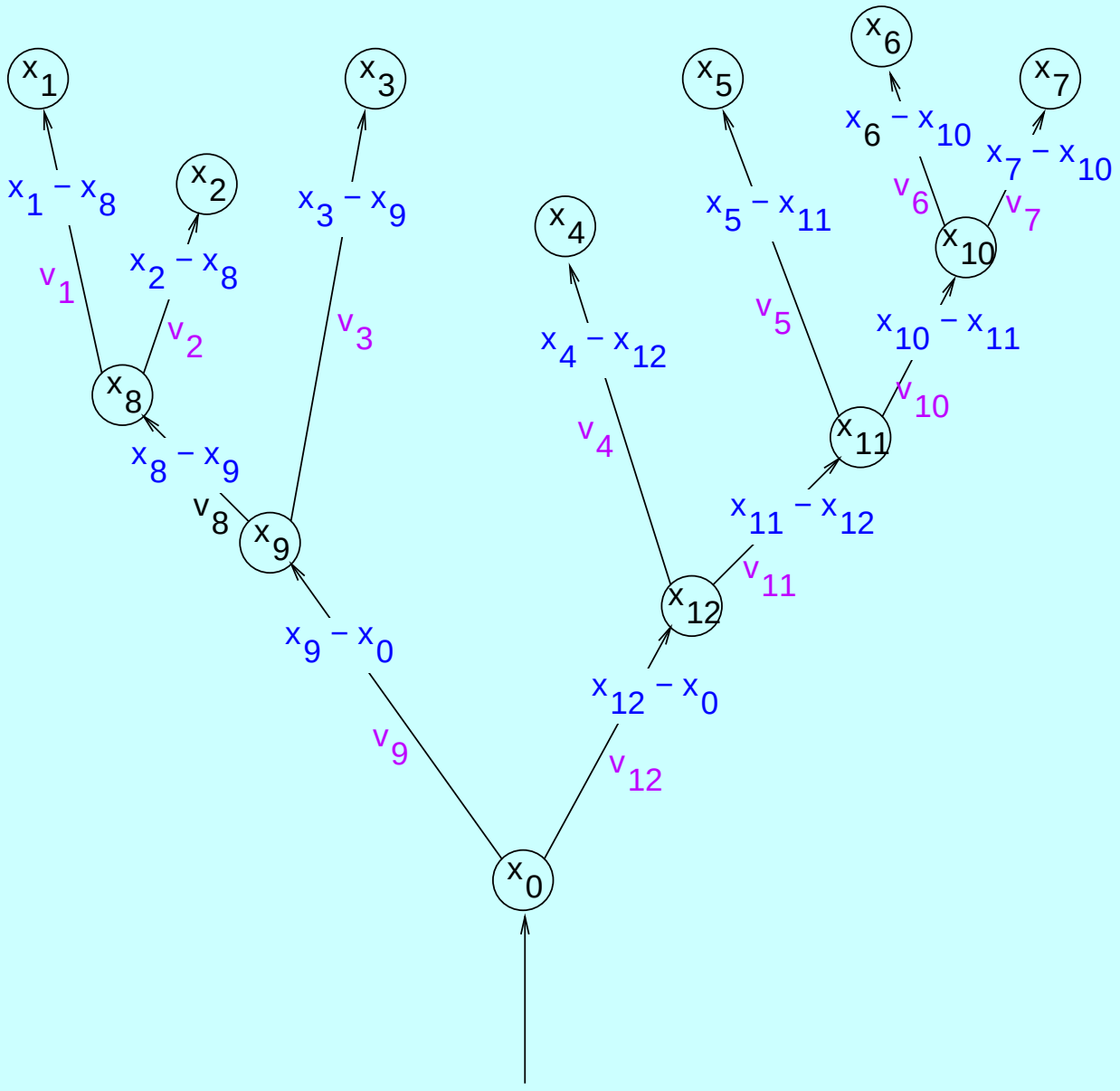
The covariance (noting that x_0 is constant and drops out) is

$$\begin{aligned} \mathrm{Cov}(x_2, x_3) &= \mathrm{Cov}(\Delta x_1 + \Delta x_2,\ \Delta x_1 + \Delta x_3) \\ &= \mathrm{Cov}(\Delta x_1,\ \Delta x_1) + \mathrm{Cov}(\Delta x_1, \Delta x_2) \\ &\quad + \mathrm{Cov}(\Delta x_1, \Delta x_3) + \mathrm{Cov}(\Delta x_2, \Delta x_3) \end{aligned}$$

and, since changes in different branches are independent, this is

$$\mathrm{Cov}(x_2, x_3) = \mathrm{Cov}(\Delta x_1, \Delta x_1) + 0 + 0 + 0 = \mathrm{Var}(\Delta x_1) = v_1$$

Brownian motion along a tree



Covariances of species on the tree

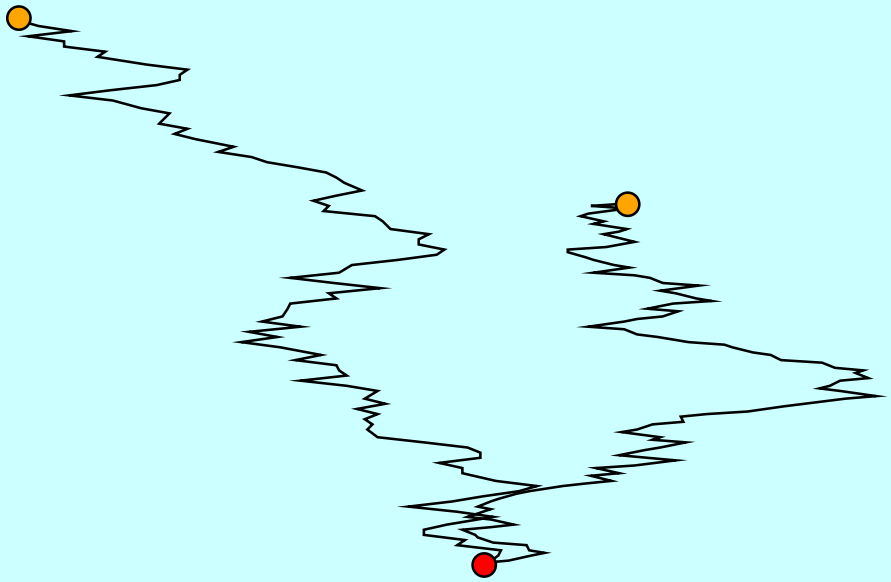
$$\begin{bmatrix}
v_1+v_8+v_9 & v_8+v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_8+v_9 & v_2+v_8+v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_9 & v_9 & v_3+v_9 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & v_4+v_{12} & v_{12} & v_{12} & v_{12} \\
0 & 0 & 0 & v_{12} & v_5+v_{11}+v_{12} & v_{11}+v_{12} & v_{11}+v_{12} \\
0 & 0 & 0 & v_{12} & v_{11}+v_{12} & v_6+v_{10}+v_{11}+v_{12} & v_{10}+v_{11}+v_{12} \\
0 & 0 & 0 & v_{12} & v_{11}+v_{12} & v_{10}+v_{11}+v_{12} & v_7+v_{10}+v_{11}+v_{12}
\end{bmatrix}$$

Covariances are of form

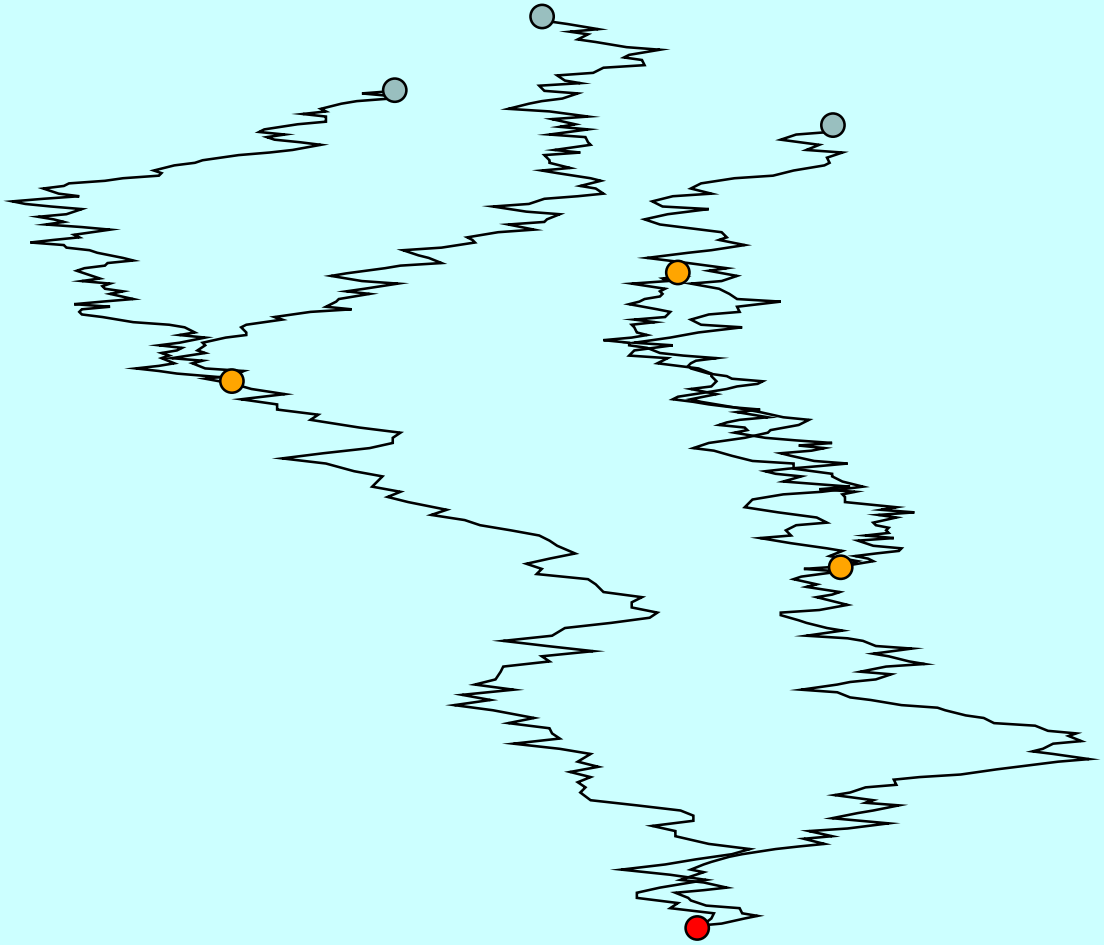
$$\left[\begin{array}{ccc|cccc} a & b & c & 0 & 0 & 0 & 0 \\ b & d & c & 0 & 0 & 0 & 0 \\ c & c & e & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & f & g & g & g \\ 0 & 0 & 0 & g & h & i & i \\ 0 & 0 & 0 & g & i & j & k \\ 0 & 0 & 0 & g & i & k & l \end{array}\right]$$

An outcome of Brownian motion on a 5-species tree

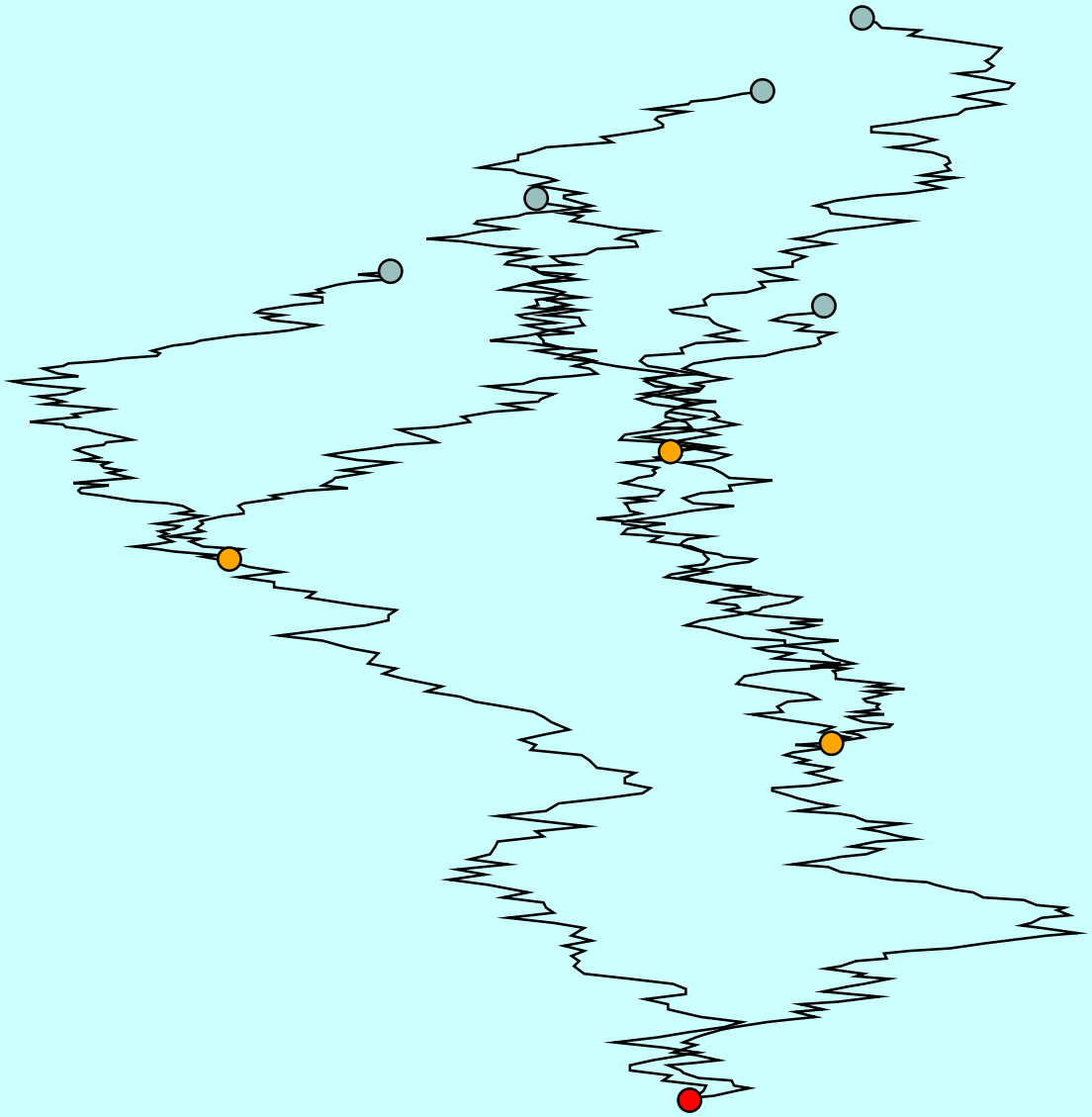
An outcome of Brownian motion on a 5-species tree



An outcome of Brownian motion on a 5-species tree



An outcome of Brownian motion on a 5-species tree



“Pruning” a tree in the Brownian motion case

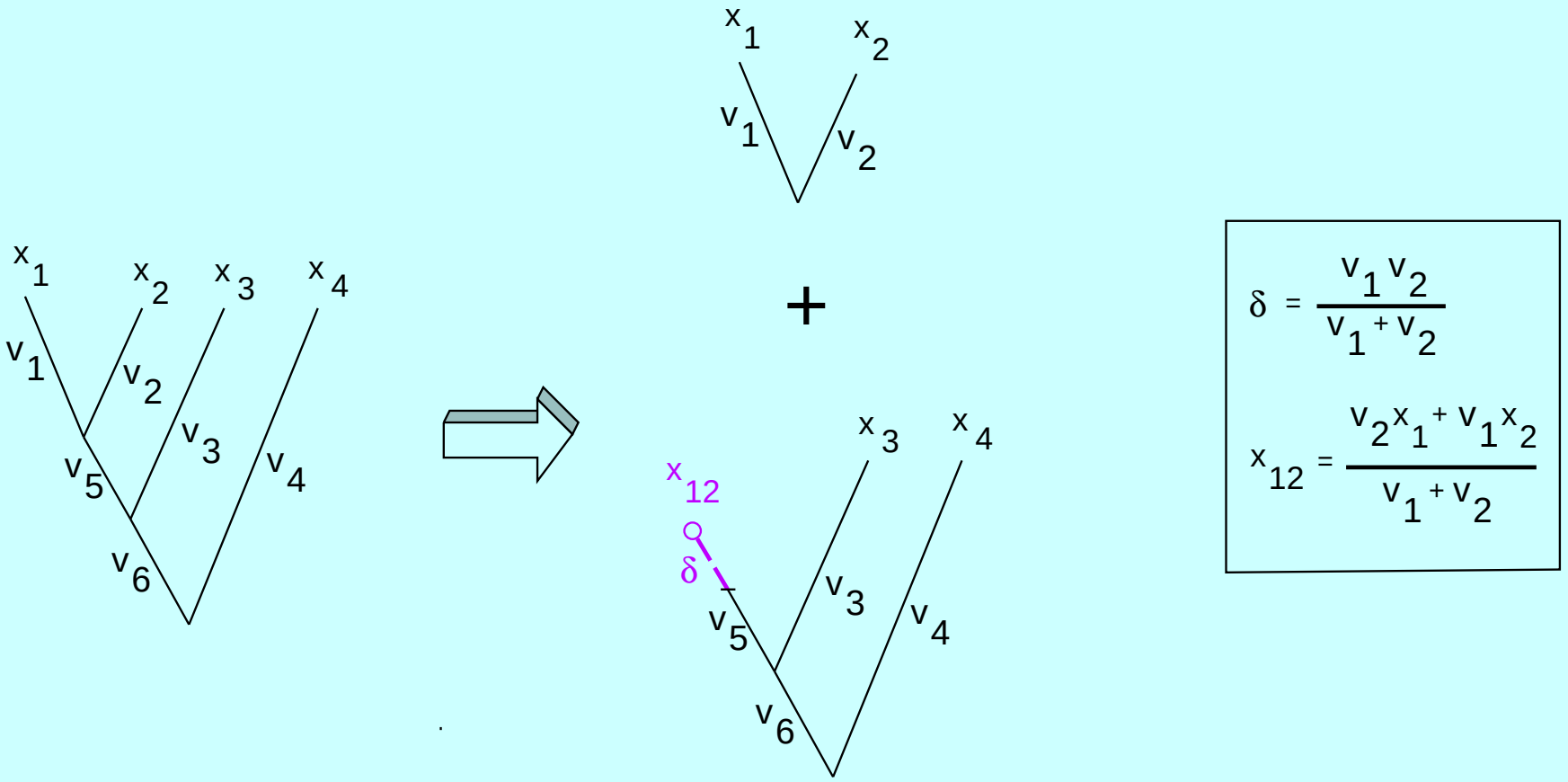
One can take two neighboring tips, and consider their difference $x_1 - x_2$ as well as a weighted average $ax_1 + (1-a)x_2$. Using weights $a : 1-a = 1/v_1 : 1/v_2$, the weighted average is independent of the difference, and the difference is also independent of the rest of the tree.

In fact, this weighted average behaves like a tip: Its covariances with the other species are the same as those of x_1 and x_2. It acts just as if the tree were pruned, cutting off species 1 and 2, leaving a single species whose variance is a bit bigger.

$$\mathrm{Var}[ax_1 + (1-a)x_2] = v_8 + v_9 + \frac{v_1 v_2}{v_1 + v_2}$$

so in effect, a small extra amount of branch length is added.

“Pruning” a tree in the Brownian motion case



(True in the sense that the REML likelihoods – which are a bit different than the usual likelihoods – add up).

Can decompose the tree into $n-1$ two-species trees

- As we decompose the tree of n species into $n-1$ two-species trees, we do so simultaneously at all characters.

Covarying character change along a lineage

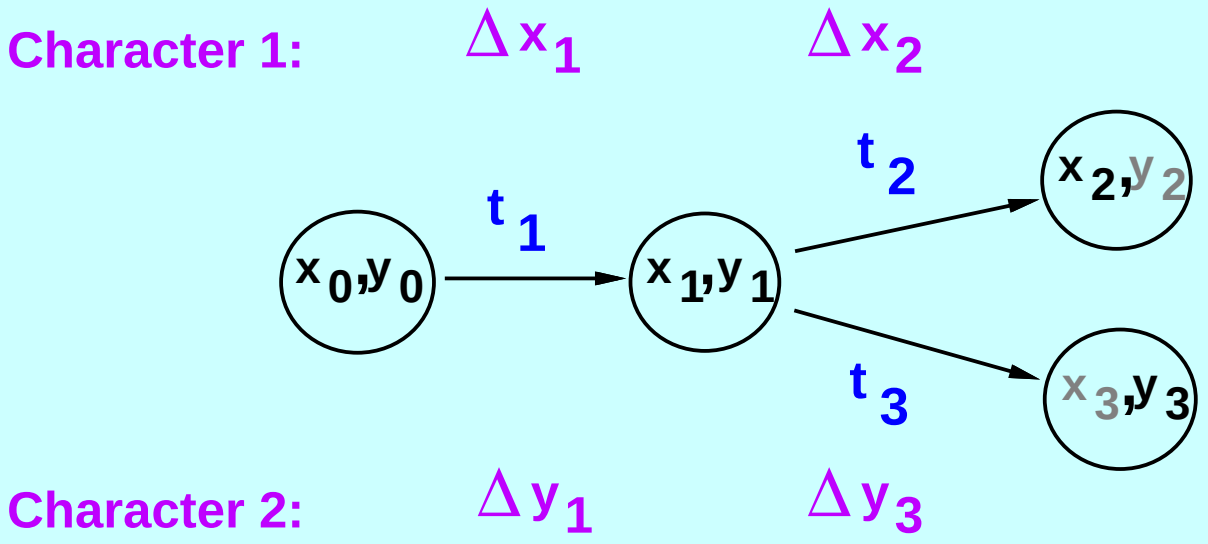
What is the distribution of changes in multiple characters (say p of them) along a lineage? Simply the appropriate multiple of the infinitesimal rate of change per unit branch length.

If a set of characters $\mathbf{x}$, changes under covarying Brownian motion, in time t (or a pseudo-time branch-length t) the change will be distributed as

$$\mathbf{\Delta x} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}t),$$

(where $\mathbf{V}$ is the covariance matrix of the infinitesimal change of the Brownian Motion).

Joint distribution for multiple species, characters



Consider change of two characters, each assessed in a different species. Say character x and character y, the first measured in species 2, the second in species 3. The result will give us the pattern for any two characters measured in any two species.

What causes change in quantitative characters?

For neutral mutation and genetic drift, can show that for a quantitative character with additive genetic variance V_A and population size N the genetic (additive) value of the population mean is:

$$\mathrm{Var}(\Delta\bar{g}) = V_A/N$$

If mutation and drift are at equilibrium:

$$E\left[V_A^{(t+1)}\right] = V_A^{(t)}\left(1 - \frac{1}{2N}\right) + V_M$$

In neutral traits additive genetic variance rules

so that

$$\mathsf{E}\left[\mathsf{V_A}\right] \;=\; 2\mathsf{N}\mathsf{V_M}$$

whereby

$$\mathrm{Var}[\Delta\bar{\mathsf{g}}] \;=\; \left(2\mathsf{N}\mathsf{V_M}\right)/\mathsf{N} \;=\; 2\mathsf{V_M}$$

an analog of Kimura's result for neutral mutation.

Thus to transform characters to independent Brownian motions of equal evolutionary variance, we could use the additive genetic variance $\mathsf{V_A}$.

With multiple characters ...

There is a precise analogue of this for multiple characters:

$$\mathsf{E}\left[\mathbf{A}^{(t+1)}\right] \;=\; \mathbf{A}^{(t)}\left(1-\frac{1}{2N}\right)+\mathbf{M}$$

where $\mathbf{A}$ is the additive genetic covariances, and $\mathbf{M}$ is the covariance matrix of pleiotropic effects of mutation.

$$\mathsf{E}\left[\mathbf{A}\right] \;=\; 2N\,\mathbf{M}$$

and

$$\mathrm{Var}[\boldsymbol{\Delta}\bar{\mathbf{g}}] \;=\; (2N\mathbf{M})\,/N \;=\; 2\mathbf{M}$$

so as long as mutations cause expected change zero (i.e. they are not near some biological limit), the effect of genetic drift is that the mean phenotype wanders according to the mutational covariances. The constant additive genetic variance assumption was used by Russ Lande.

Let’s see ...

(Pause to run simulation of mutation and genetic drift in two characters).

With selection ... life is harder

There is the "Breeder's Equation" of Wright and Fisher (1920's)

$$\Delta z \;=\; h^2 S$$

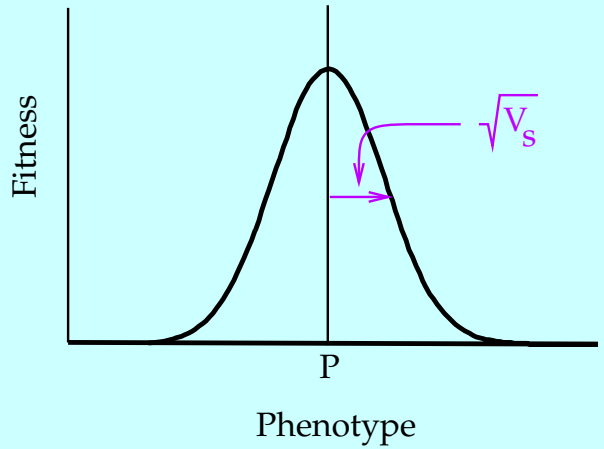
and Russ Lande's (1976) recasting of that in terms of slopes of mean fitness surfaces:

$$S \;=\; V_P \, \frac{d \log(\bar{w})}{d\bar{x}}$$

$$\Delta z \;=\; (V_A/V_P)\, V_P \, \frac{d \log(\bar{w})}{d\bar{x}} \;=\; V_A \, \frac{d \log(\bar{w})}{d\bar{x}}$$

Note – it's heritability times the slope of log of *mean* fitness with respect to *mean* phenotype. There is an exact multivariate analog of this equation.

Selection towards an optimum



If fitness as a function of phenotype is:

$$w(x) = \exp\left[-\frac{(x-p)^2}{2V_s}\right]$$

Then after some completing of squares and integrating, the change of mean phenotype "chases" the optimum:

$$m' - m = \frac{V_A}{V_s + V_P}(p - m)$$

(There is an exact matrix analog of this for multiple characters).

Sources of evolutionary correlation among characters

Variation (and covariation) in change of characters occurs for two reasons:

- **Genetic covariances**. (the same loci affect two or more traits)

Sources of evolutionary correlation among characters

Variation (and covariation) in change of characters occurs for two reasons:

- **Genetic covariances**. (the same loci affect two or more traits)
- **Selective covariances** (Tedin, 1926; Stebbins 1950). The same environmental conditions select changes in two or more traits – even though they may have no genetic covariance.

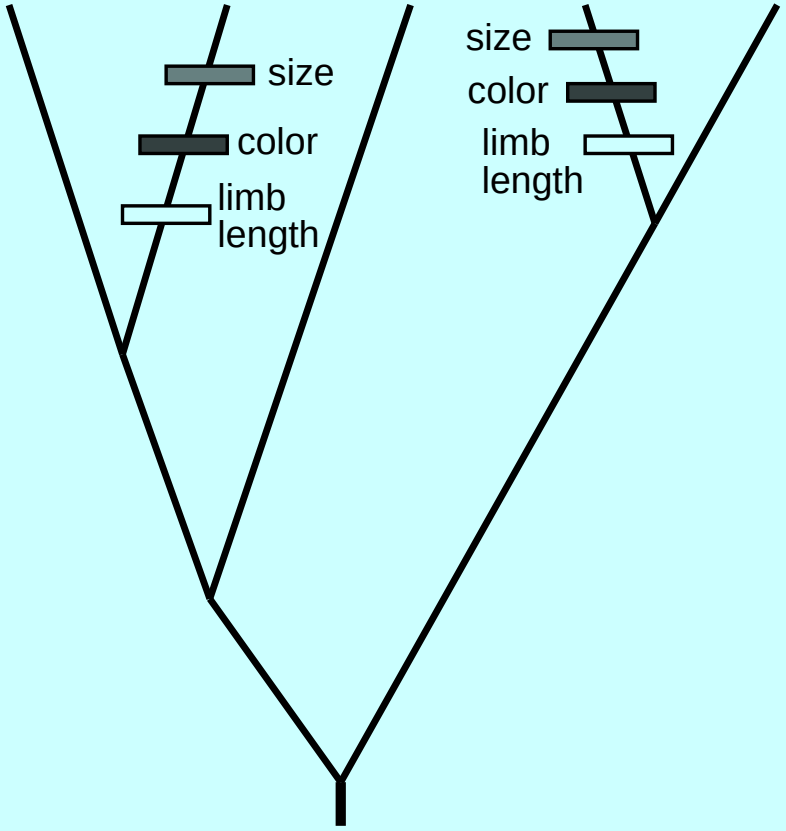
A simple example of selective covariance

covariation due not to genetic correlation
but to covariation of the selection pressure

These are Bergmann's, Allen's and Glogler's Rules
They are presumably not the result of genetic correlations
but result from patterns of selection

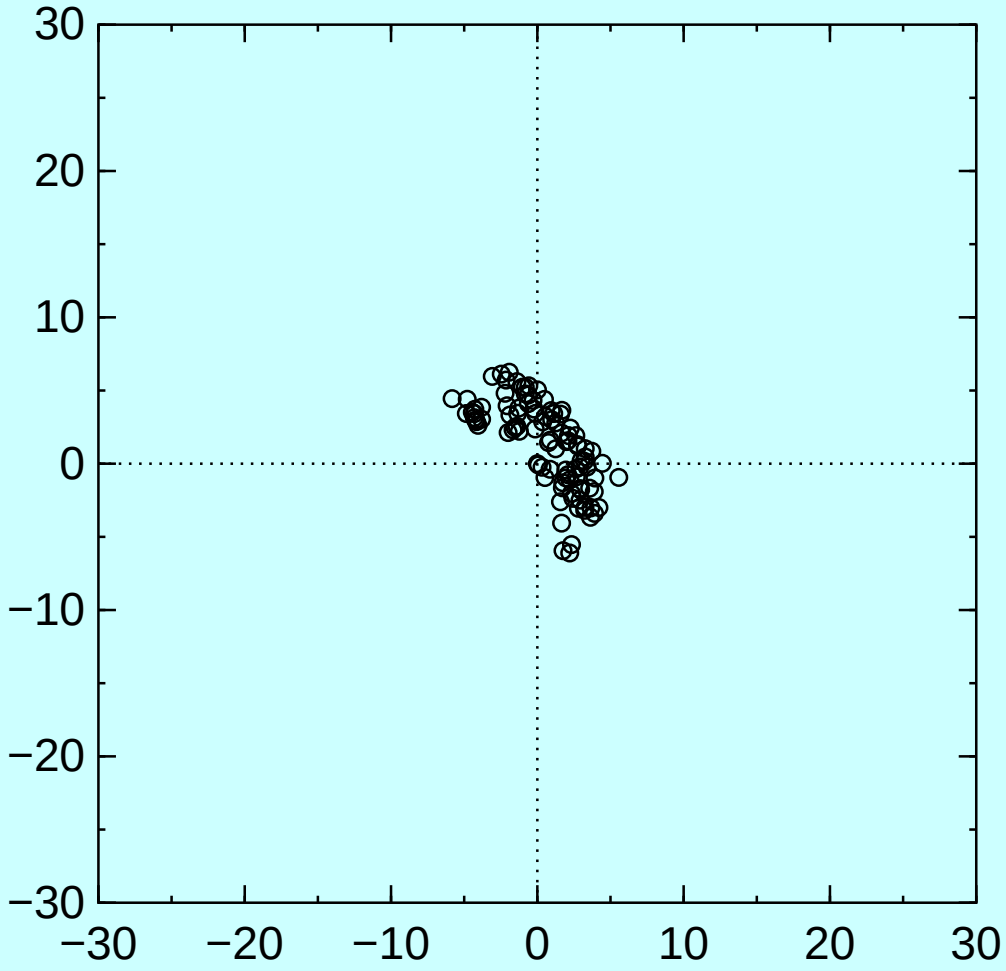
a simple example:

(temperate) (arctic) (temperate) (arctic) (temperate)



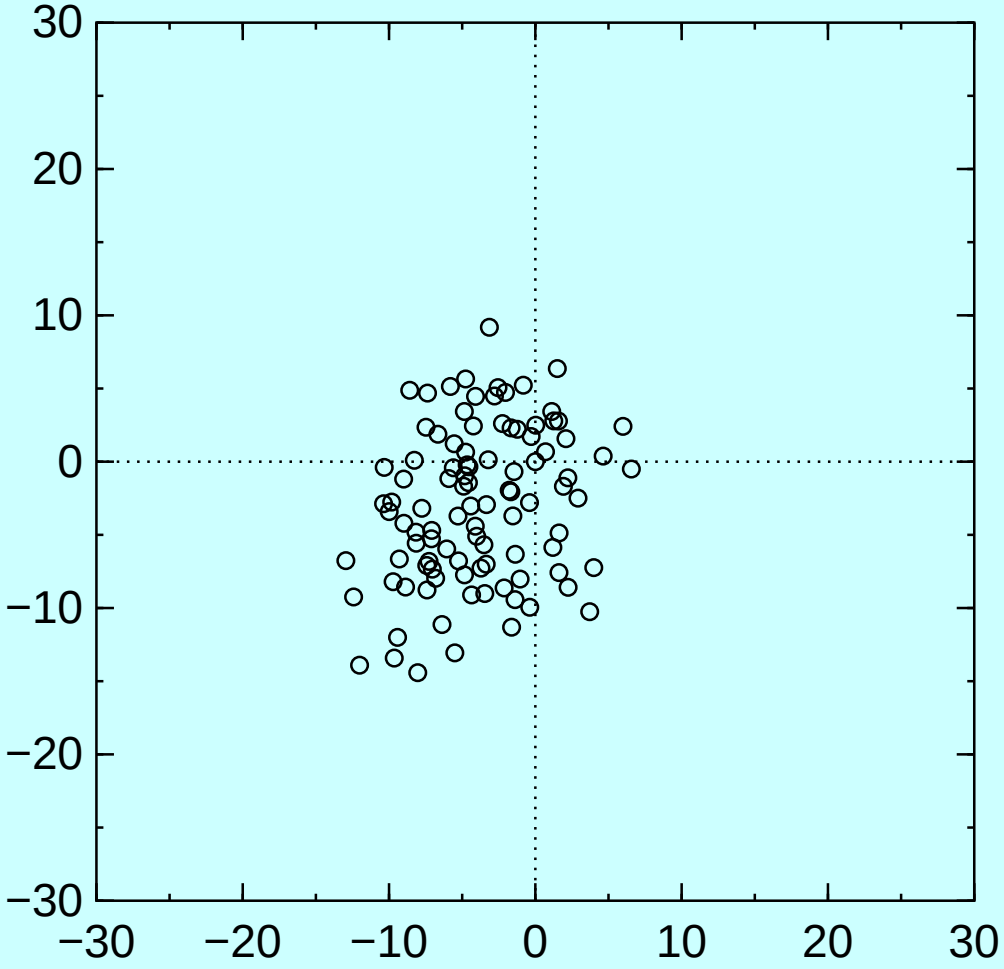
G. L. Stebbins. 1950. *Variation and evolution in plants.* Columbia Univ. Press, New York. page 121

Chasing a peak, simulated with two characters



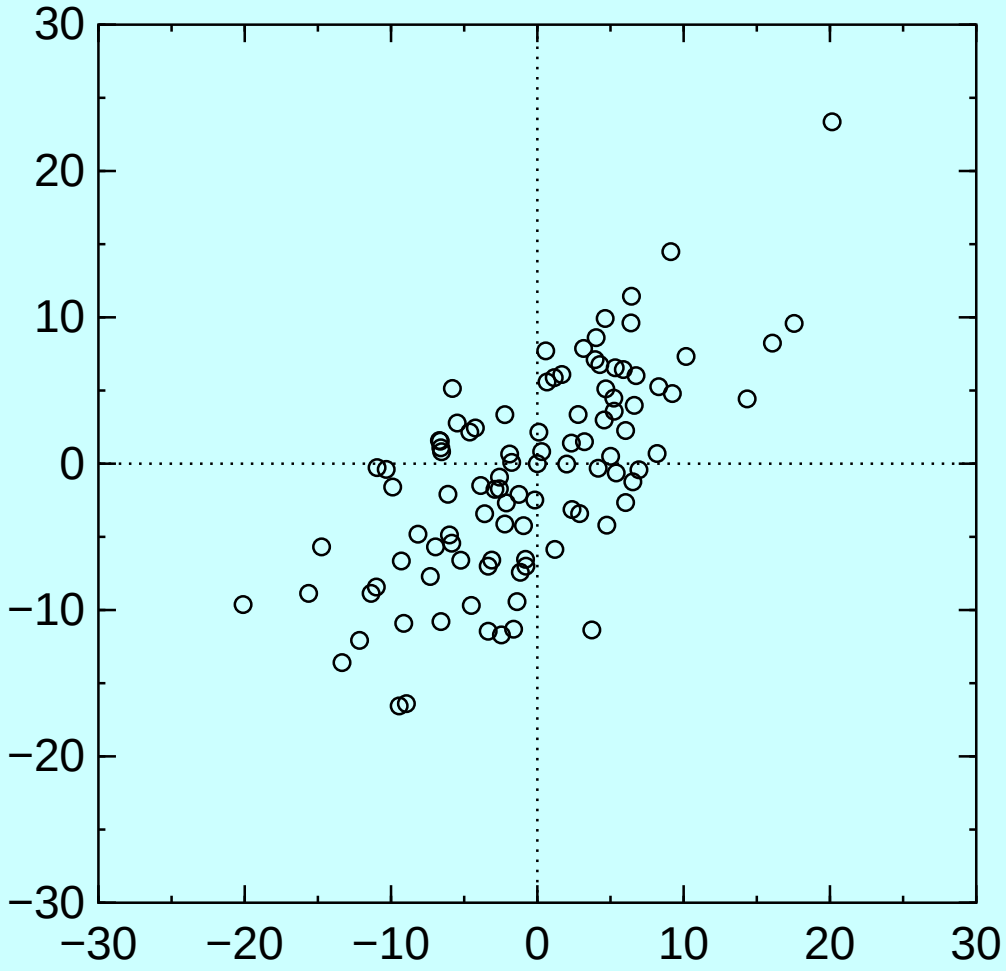
Genetic covariances assumed negative, but the wanderings of the adaptive peaks assumed positively correlated. In the first 100 generations the genetic covariances are most influential.

Chasing a peak, simulated with two characters



Genetic covariances assumed negative, but the wanderings of the adaptive peaks assumed positively correlated. After a while (every 10th generation up to generation 1000), the wanderings of the peaks start to be influential.

Chasing a peak, simulated with two characters



Genetic covariances assumed negative, but the wanderings of the adaptive peaks assumed positively correlated. In the long run (every 100th generation up to generation 10,000) the means go mostly where the peaks go.

A case that has received too little attention

- Suppose characters x and y are genetically correlated.
- and y is under optimum selection, but x is the one we observe.
- What will we see? In effect, the sum (actually, a weighted average) of an Ornstein-Uhlenbeck process and Brownian Motion.
- So Brownian motion restricted in the short run but not in the long run.

Most models so far do not allow for characters that are observed to covary with those that aren't observed.

A little algebra showing the effect of selective covariance

If we start from the familiar "Breeder's Equation" of quantitative genetics:

$$\Delta z = h^2 S$$

it has long been known to have a multivariate version:

$$\boldsymbol{\Delta z} = \mathbf{G P^{-1} S}$$

Multiplying $\boldsymbol{\Delta z}$ by its transpose:

$$\boldsymbol{\Delta z \Delta z}^{\mathsf{T}} = \mathbf{G P^{-1} S S^{\mathsf{T}} P^{-1} G}$$

and taking expectations (treating $\mathbf{G}$ and $\mathbf{P}$ as constants) we get for the mean squares:

$$\mathrm{E}[\boldsymbol{\Delta z \Delta z}^{\mathsf{T}}] = \mathbf{G P^{-1}} \mathrm{E}[\mathbf{S S^{\mathsf{T}}}] \mathbf{P^{-1} G}$$

(Felsenstein, 1988)

A research program?

What we could imagine doing is:

- We might hope to infer additive genetic covariances by doing quantitive genetics breeding experiments to infer them from covariances among relatives, perhaps even in multiple species.
- Infer the covariances of the changes along the phylogeny.
- From them, back-calculate the selective covariances.
- The genetic covariances may also be inferrable from differences between nearby tips on the tree if we do not have breeding experiments.
- There is little or no hope of inferring “selective correlations” directly without a complete understanding of the functional ecology.

References for genetic drift

Feller, W. 1951. Diffusion processes in genetics. pp. 227-246 in *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. University of California Press, Berkeley and Los Angeles. **[Feller's partial solution of the pure drift process for the Wright-Fisher model (and his famous proof that the process converges to the diffusion process)]**

Kimura, M. 1955a. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences* **41:** 144-150. **[Exact solution in Gegenbaur polynomials for two-allele pure genetic drift in a diffusion process approximation]**

Kimura, M. 1955b. Random drift in a multi-allelic locus. *Evolution* **9:** 419-435. **[The same, for three alleles]**

References for the Brownian Motion approximation

Edwards, A.W. F. and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. pp. 67–76 in *Phenetic and Phylogenetic Classifcation*, ed. V. H. Heywood and J. McNeill. Systematics Association Publ. No. 6, London.' **[The first paper on numerical approaches to phylogeny reconstruction; uses parsimony and proposes likelihood for gene frequency trees]**

Edwards, A.W. F. 1970. Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B* **32:** 155–174. **[More detailed consideration of the statistical properties of a maximum likelihood approach to gene frequency phylogenies]**

Felsenstein, J. 1973. Maximum likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* **25:** 471–492. **[REML approach to gene frequency phylogenies, including the contrasts algorithm for rapid computation of likelihood]**

Nielsen, R., J. L. Mountain, J. P. Huelsenbeck, and M. Slatkin. 1998. Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52:** 669-677. **[Little-noticed but much more exact method that would require MCMC machinery]**

References on likelihood of Brownian Motion trees

Thompson, E. A. 1975. *Human Evolutionary Trees*. Cambridge University Press, Cambridge **[Thesis monograph on how to infer ML phylogenies from gene frequencies, published because it won a Smith's Prize at Cambridge University]**

Felsenstein, J. 1981. Maximum likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* **25:** 471–492. **[Reworks the 1973 paper with more care and some additional algorithmics, including discussion of effect of character covariation]**

Felsenstein, J. 1985. Phylogenies from gene frequencies: A statistical problem. *Systematic Zoology* **34:** 300–311. **[Shows how gene frequency changes depart from being approximated by Brownian Motion]**

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. **[See particularly chapter 23]**

References for multivariate Brownian motion

Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **19:** 445-471. **[Review with mention of usefulness of threshold model]**

Felsenstein, J. 2002. Quantitative characters, phylogenies, and morphometrics.pp. 27-44 in *Morphology, Shape, and Phylogenetics*, ed. N. MacLeod. Systematics Association Special Volume Series 64. Taylor and Francis, London. **[Review repeating 1988 material and going into some more detail on the question of threshold models.]**

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. **[Mentions this model and also sample size issues in contrasts method]**.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30:** 314-334. **[Lande's classic paper on drift versus optimum selection]**

Lande, R. 1979. The quantitative genetic analysis of multivariate evolution, applied to brain-body size allometry. *Evolution* **33:** 402-416.

References

Lande, R. 1980. The genetic covariance between characters maintained by pleiotropic mutations. *Genetics* **94:** 203-215.

Lynch, M. and W. G. Hill. 1986. Phenotypic evolution by neutral mutation. *Evolution* **40:** 915-935.

Stebbins, G. L. 1950. *Variation and Evolution in Plants*. Columbia University Press, New York. **[Describes selective covariance and cites Tedin (1925) for it]**

Tedin, 0. 1925. Vererbung, Variation, und Systematik der Gattung *Camelina*. *Hereditas* **6:** 275-386.

Armbruster, W. S. 1996. Causes of covariation of phenotypic traits among populations. *Journal of Evolutionary Biology* **9:** 261-276. **[Good exposition of selective covariance]**