Brownian motion models and phylogenies

Friday Harbor Laboratories, 8 June 2017

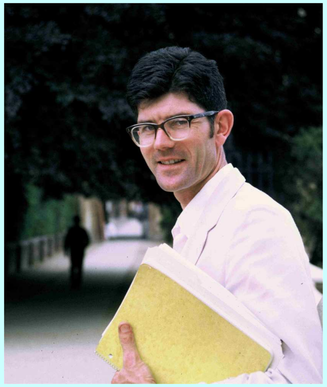
Joe Felsenstein

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?

Edwards and Cavalli-Sforza's approximation

Luca Cavalli-Sforza (and Edwards), 1963



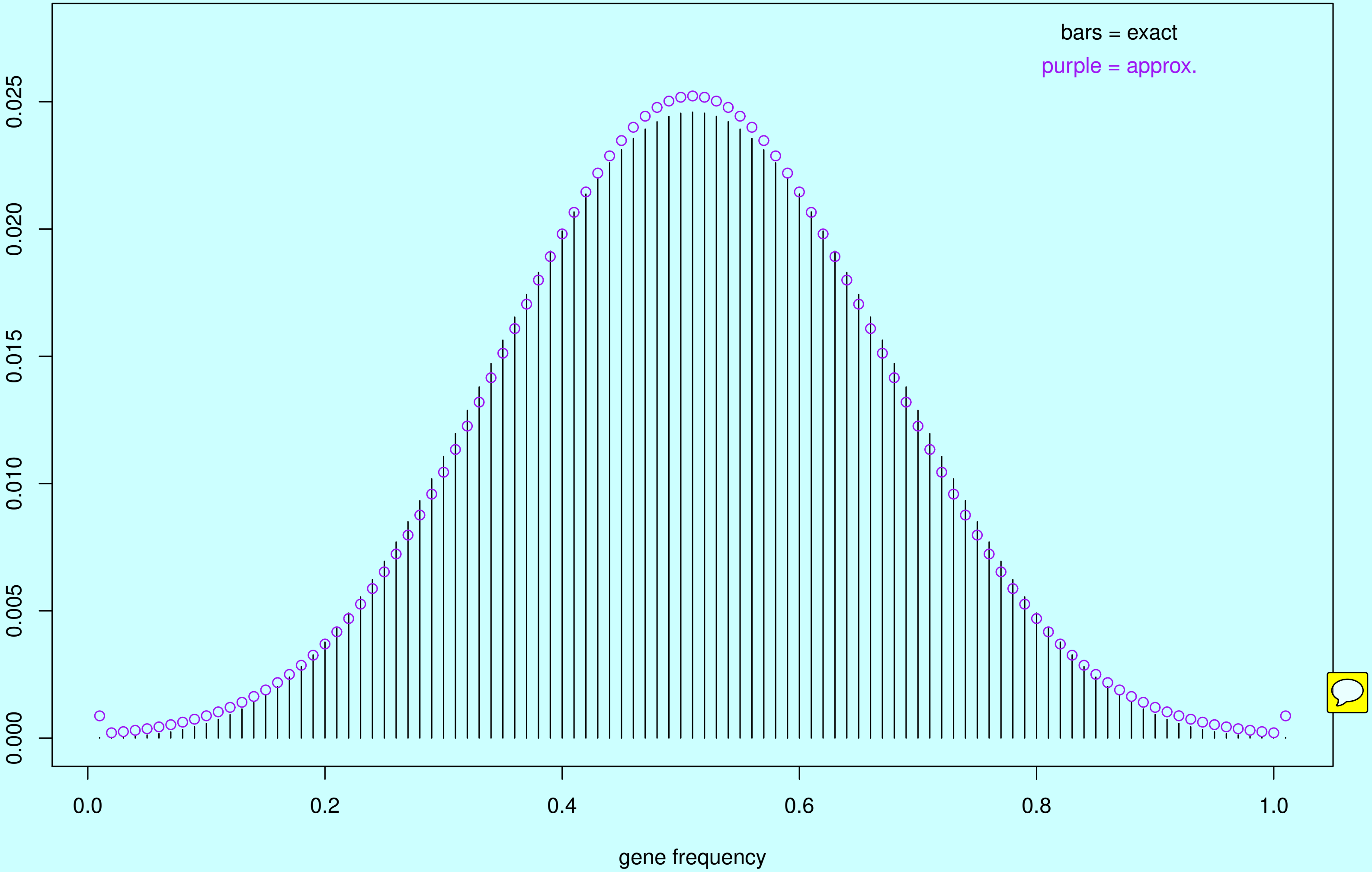
Anthony Edwards, 1970

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

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

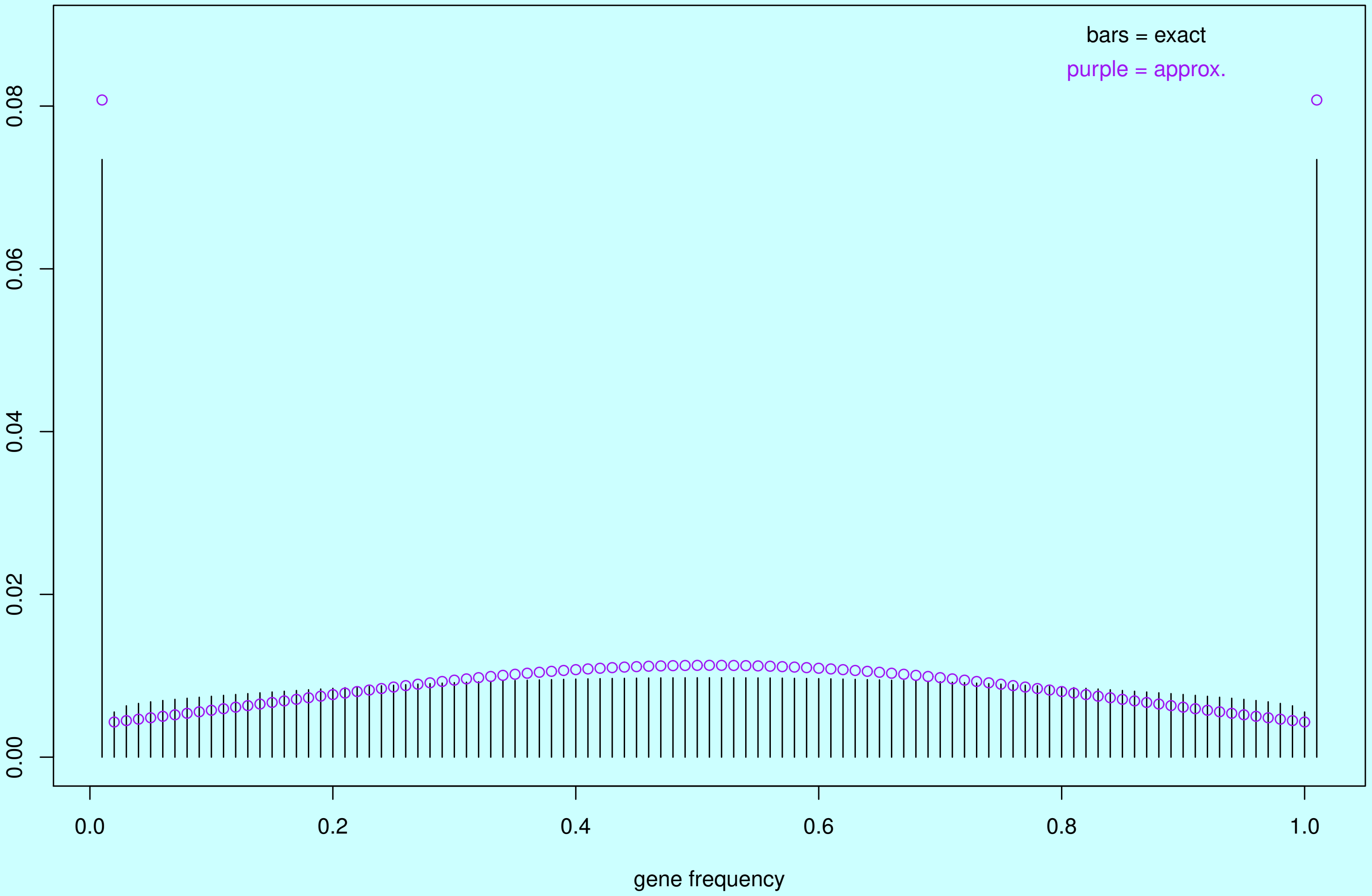
That variance is not constant: it varies with p (in a parabola), but maybe we can roughly approximate it by dealing with the case where all populations have roughly similar gene frequencies, so the variances are nearly the same. Maybe. Roughly.

How good is this?



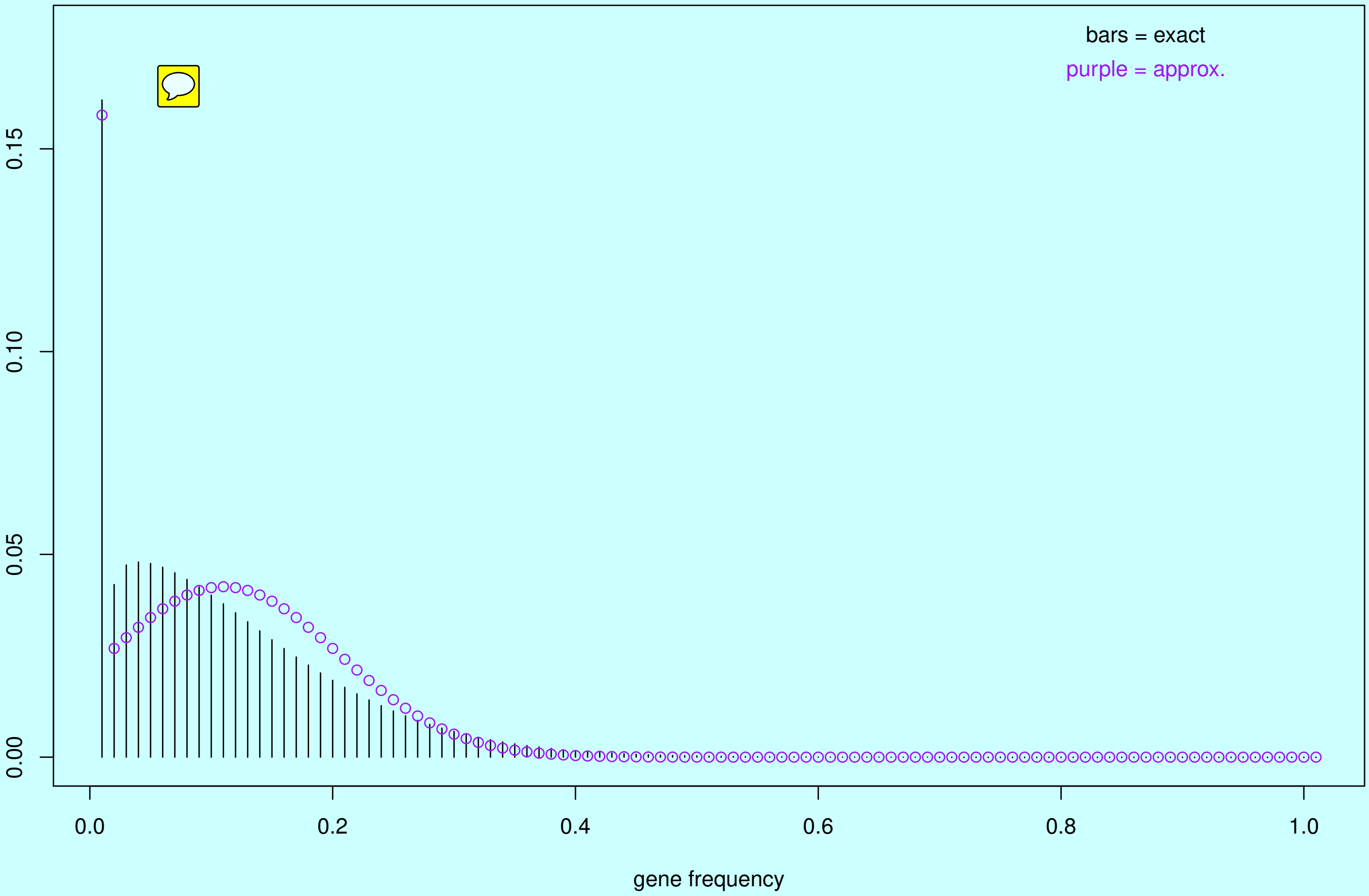
Starting with $p = 0.5$, after 10 generations in a population of size 50.

How good is this?



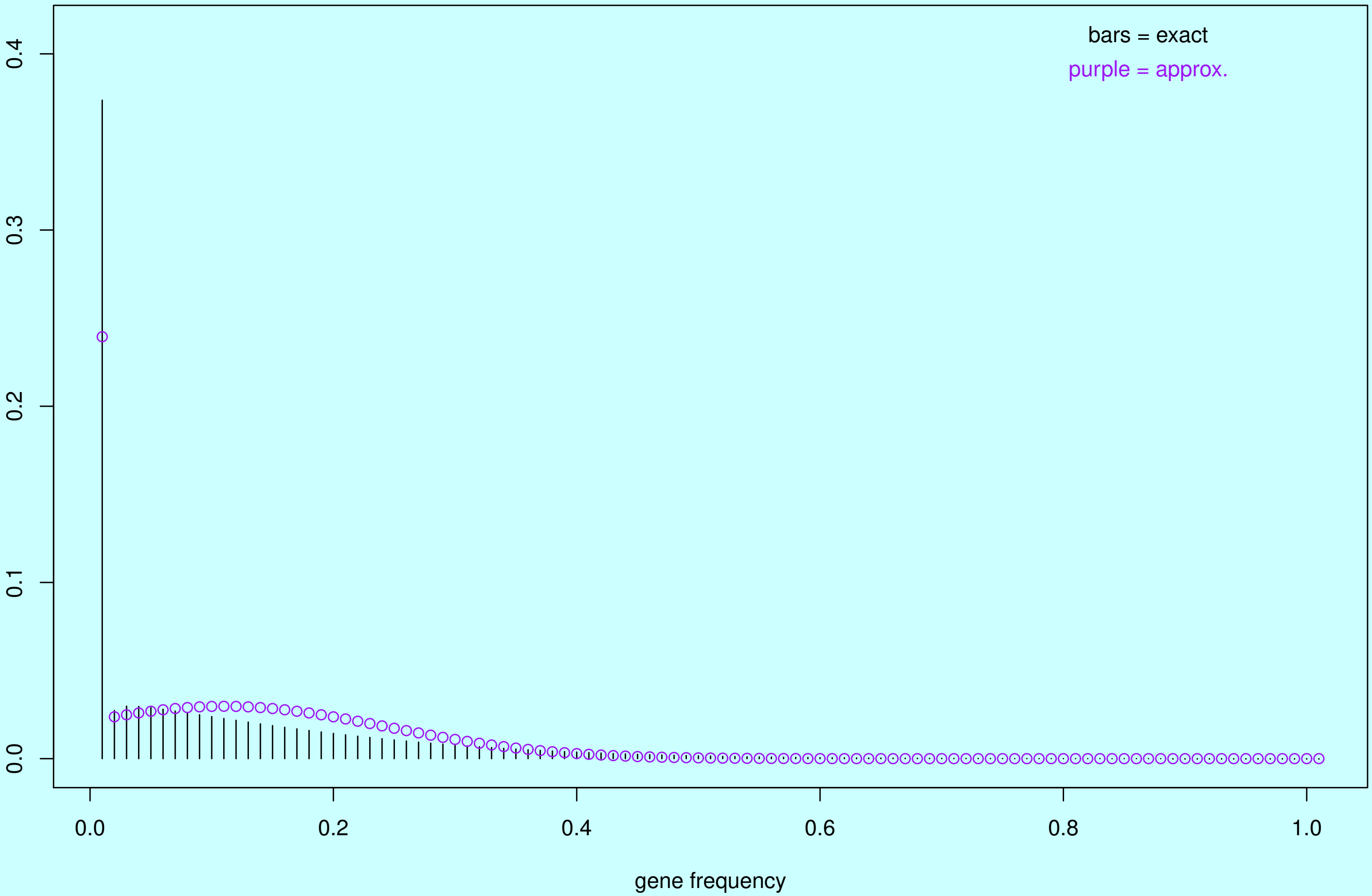
Starting with $p = 0.5$, after 50 generations in a population of size 50.

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.

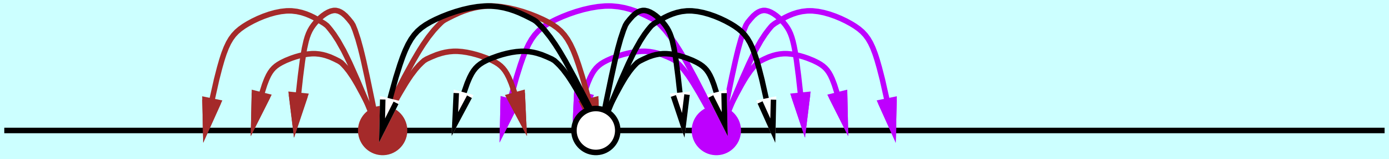
What about a quantitative character?

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.

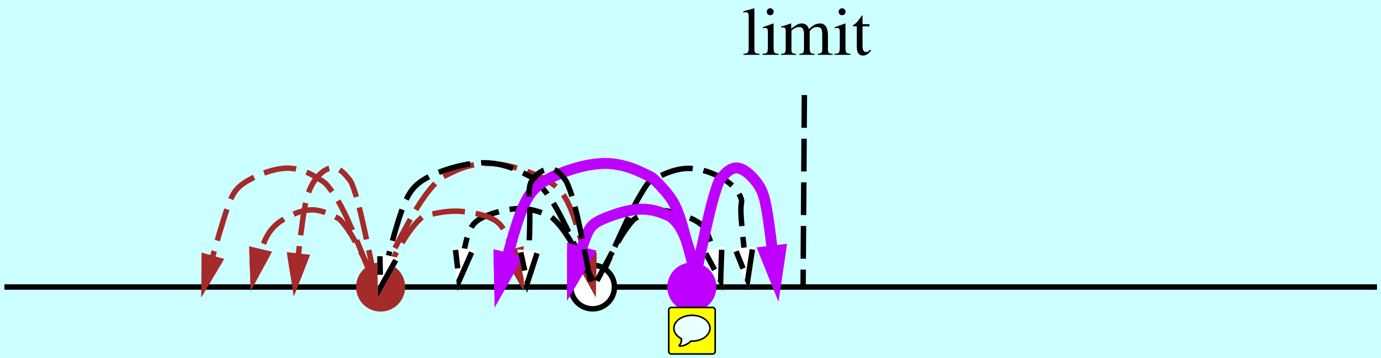
- if there is any dominance, there will be some nonlinearity and the approximation will be less good.
- Epistasis can cause even more trouble.

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.

Brownian motion from moving adaptive peaks

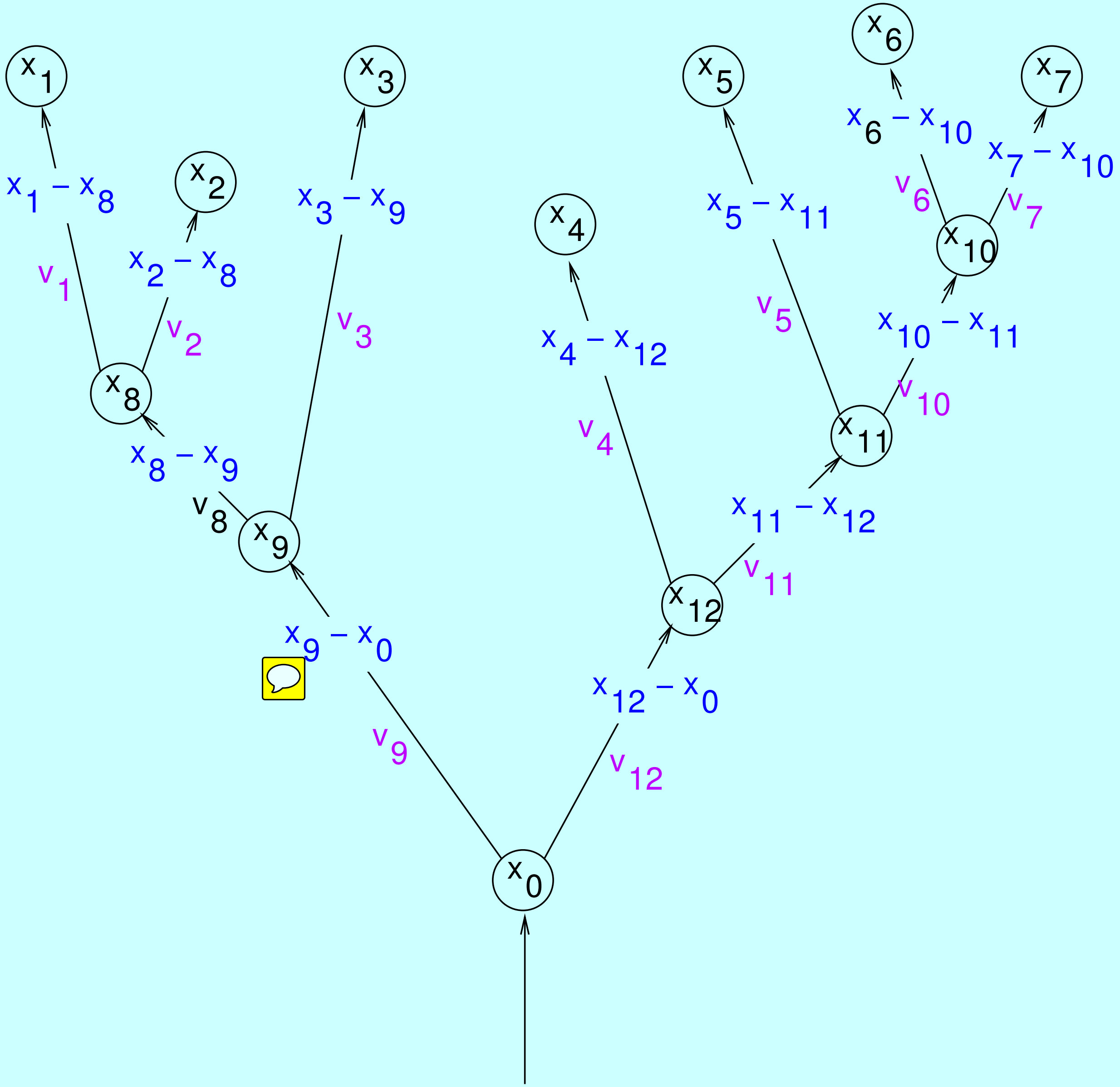
A reminder: we have seen earlier, and will hear about it later: a process close to Brownian motion can also arise from natural selection toward an optimum which is itself moving.

Brownian motion is mathematically tractable

You can easily compute transition probabilities from one value to another, since the net change after “time” $\mathtt{t}$ is normal. with mean zero and variance $\sigma^2\mathtt{t}$, and changes in successive time intervals are independent.

When two lineages share a period of common ancestry, the resulting tip species have phenotypes that covary, the covariance being the variance expected during their shared ancestry.

Brownian motion along a tree

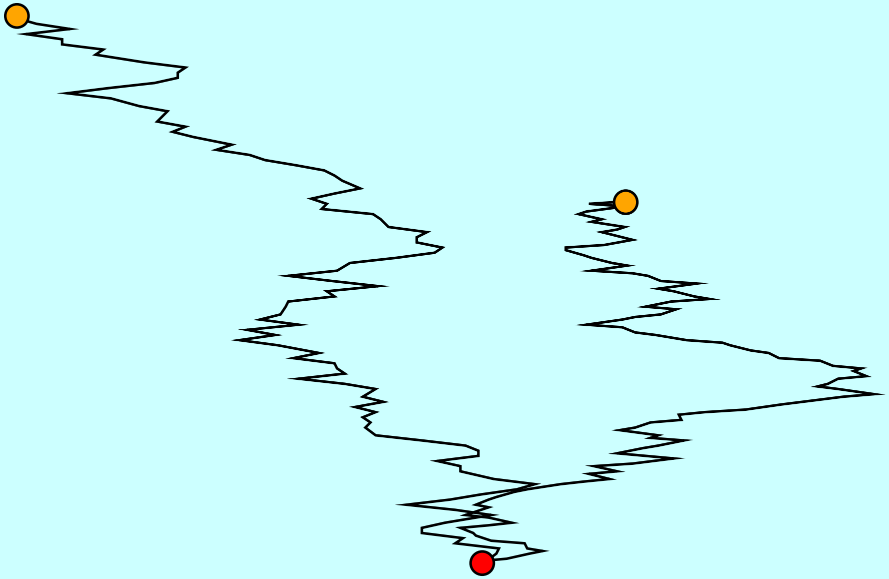


Covariances of species on the tree

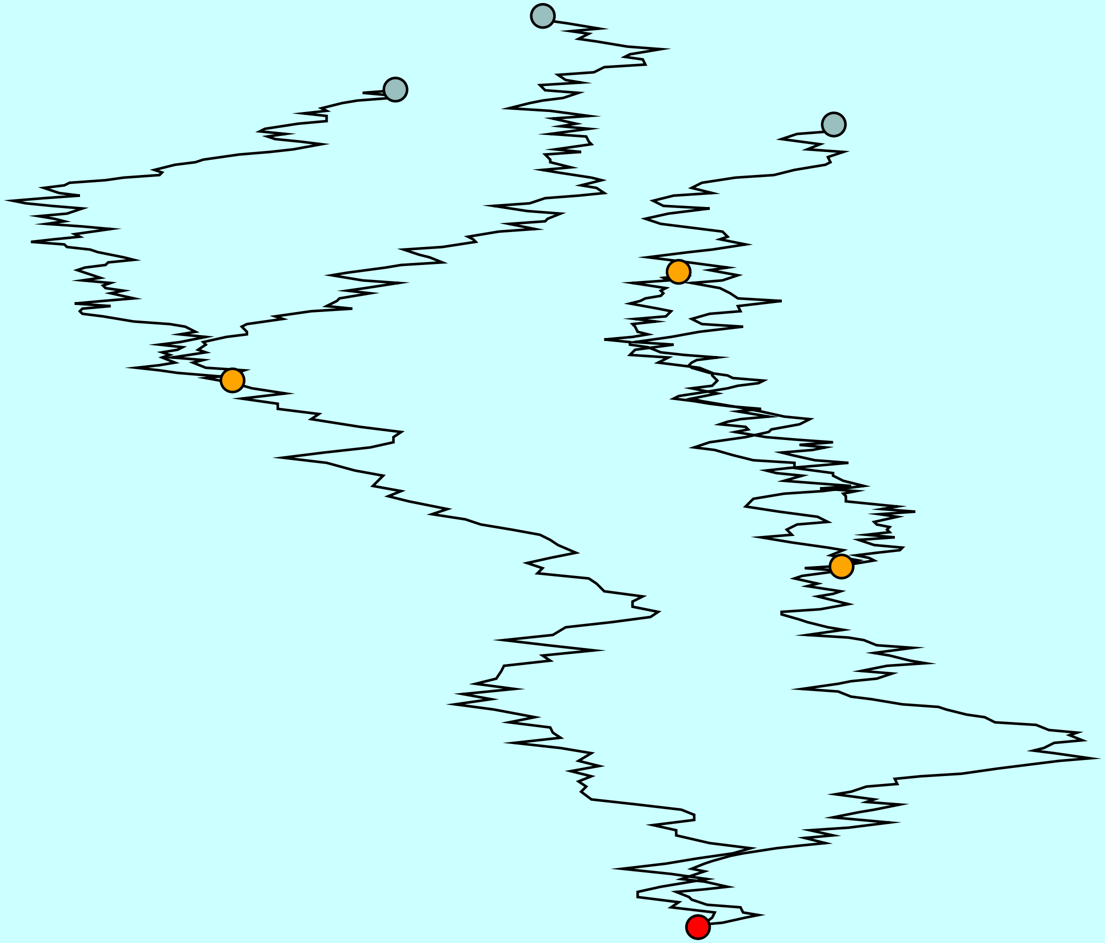
$$
\left.\begin{matrix}
v_1 + v_8 + v_9 & v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_8 + v_9 & v_2 + v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_9 & v_9 & v_3 + v_9 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & v_4 + v_{12} & v_{12} & v_{12} & v_{12} \\
0 & 0 & 0 & v_{12} & v_5 + v_{11} + v_{12} & v_{11} + v_{12} & v_{11} + v_{12} \\
0 & 0 & 0 & v_{12} & v_{11} + v_{12} & v_6 + v_{10} + v_{11} + v_{12} & v_{10} + v_{11} + v_{12} \\
0 & 0 & 0 & v_{12} & v_{11} + v_{12} & v_{10} + v_{11} + v_{12} & v_7 + v_{10} + v_{11} + v_{12}
\end{matrix}\right]
$$

An outcome of Brownian motion on a 5-species tree

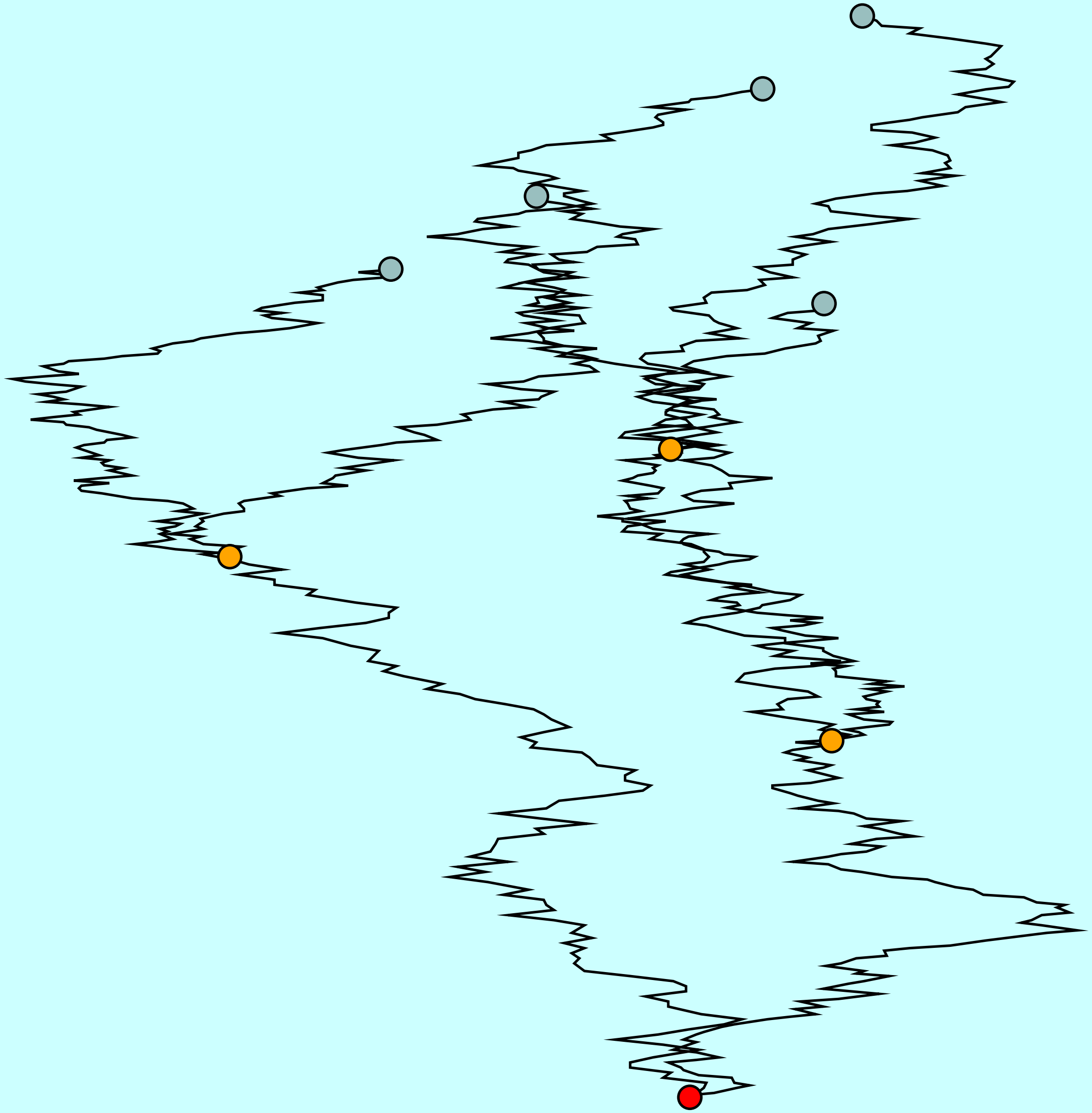
An outcome of Brownian motion on a 5-species tree



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

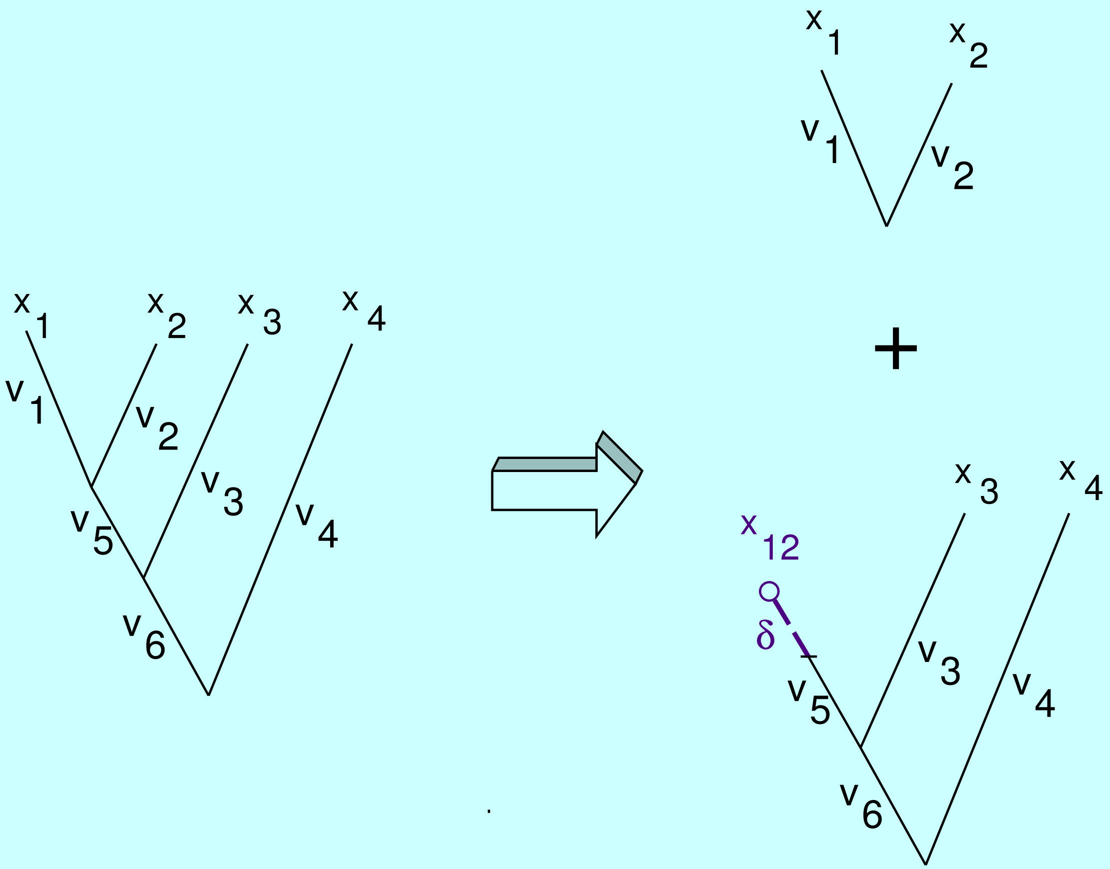
One can take two neighboring tips, and consider their difference $x_1 - x_2$ as well as a weighted average $ax_1 + (1 - a)x_2$. Using weights $a \; : \; 1 - a \;\; = \;\; 1/v_1 \; : \; 1/v_2$, the weighted average is independent of the difference, and the difference is also independent of the rest of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



$$\delta = \frac{v_1 v_2}{v_1 + v_2}$$

$$x_{12} = \frac{v_2 x_1 + v_1 x_2}{v_1 + v_2}$$

(True in the sense that the log-likelihoods – which are a bit different than the usual likelihoods – add up, since the likelihoods multiply).

References for genetic drift

ller, 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)]**

mura, 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 Gegenbauer polynomials for two-allele pure genetic drift in a diffusion process approximation]**

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

References for the Brownian Motion approximation

lwards, 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]**

lwards, 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]**

lsenstein, 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]**

elsen, 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

ompson, 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]**

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

lsenstein, 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]**