

Convergence notes

¹ Center for Population Biology, University of California, Davis.

² Department of Evolution and Ecology, University of California, Davis

³ New York Genome Center

To whom correspondence should be addressed: jjberg@ucdavis.edu, gmcoop@ucdavis.edu

September 16, 2016

Evolution (journal) has a commentary section, seems like something like that could work.

Why are we impressed by convergent evolution? Well it's because we get to see selection act repeatedly to shape new mutations and standing variation into adaptations in similar ways. This is particularly impressive when we see convergence to a particular environment. Convergence helps us build evidence that the phenotype is an adaptation, and form evidence that adaptation is 'for' increasing survival/fitness in particular environment.

How much do we say about genetic drift as a source of apparent convergence (and/or release from constraint).

Two distinct, but related questions, can we say that a phenotype/variant has been under selection. How many distinct instances of selection have there been. Related Q: as the more independent instances, the more evidence we have that it's selected being selected.

Formalizing the overlap in selection, in terms of selective births/deaths When can we say that selection has been convergent in among populations? Can we formalize this Q by thinking about the overlap in selective deaths/births underlying the adaptative evolution of a particular trait, allele, or set of alleles within a population?

One locus:

The selective load (L) for directional selection (where fittest homozygote has selection coeff s) move an allele from frequency x_1 to x_2 in t_1 to t_2 generations is

$$L = \sum_{t=t_1}^{t_2} s(1 - x_t) \quad (1)$$

$$\approx \int_{t_1}^{t_2} s(1 - x_t) dt \quad (2)$$

The number of selective deaths, in a population size of N , is NL . Assuming an additive model ($dx/dt = (s/2)x(1 - x)$)

$$L = \int_{t_1}^{t_2} s(1 - x_t) dt \quad (3)$$

$$= 2 \int_{x_1}^{x_2} \frac{1}{x} dx \quad (4)$$

$$= 2 \log(x_2/x_1) \quad (5)$$

So if the ancestral population goes from frequency x_1 to x_2 , and the descendant populations (1 and 2) move this to $x_3^{(1)}$ and $x_3^{(2)}$, then their shared deaths are $2N \log(x_2/x_1)$ and separate deaths are $2N \log(x_3^{(1)}/x_2)$ and $2N \log(x_3^{(2)}/x_2)$. The fraction of shared deaths would be ratio

$$\frac{\log(x_2/x_1)}{\log(x_3^{(1)}/x_1)} \quad \text{and} \quad \frac{\log(x_2/x_1)}{\log(x_3^{(2)}/x_1)} \quad (6)$$

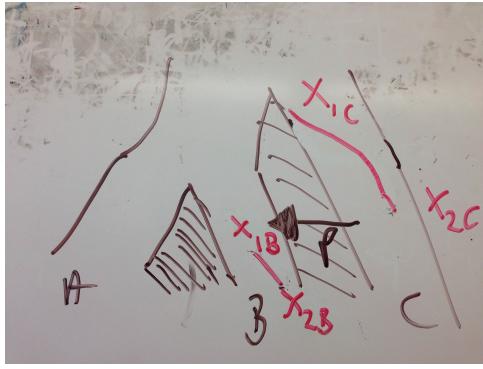


Figure 1: BLAH

Interestingly this is mostly determined by deaths when the allele when it is rare (wonder if this means HH will be informative).

Phenotype:

The response to selection, for a phenotype with heritability h^2 and phenotypic variance V_p , is

$$R = h^2 i(p) \sqrt{V_p} \quad (7)$$

where $i(p) = S/\sqrt{V_p}$, $i(p)$ is the selection intensity. So if the population achieves a phenotype shift R_T in T generations, then the implied selection intensity per generation is

$$i(p) = \frac{R_T}{h^2 T \sqrt{V_p}} \quad (8)$$

Under truncation selection (Lande says it the form of selection that implies weakest selection) we can obtain the fraction (p) of individuals who were parents from $i(p)$, but this isn't a nice simple expression. Perhaps just talking about fraction of phenotypic change would be enough, and point out that this is linked to the intensity of selection.

Why is this easier when we have changes polarized on a phylogeny? Why is this question easier when we have a resolved phylogeny with no-incomplete lineage sorting. Well in that case the selective births and deaths are clearly independent. E.g. the individuals who lived/died, due to differential predation pressure, to drive the adaptation of light coloured fur in arctic foxes and hares were clearly different sets of individuals (being foxes and hares respectively).

If we have a tree of drift (w. no gene flow), seeing non-sisters sharing a selective event can be "enough". E.g. if both populations show a sweep (A & C), not shared with sisters (B & D), then we have evidence that selection has occurred in both pops independently.

What do we mean, or can we say about, convergence when we only see the selected allele
 Consider the case, shown in Figure 1. The selected allele has changed frequency in population C, and there is gene flow from C into B (w. admixture prop. p), but not into A. The question is, is how much convergence has occurred in population B? Assume first that allele is absent in A and was absent in B prior to admixture. The change in C is from x_{1C} to x_{2C} . Assume that x_{2C} was the frequency when admixture occurred, an assumption conservative against finding convergence. Then x_{1B} , the frequency immediately following admixture is pX_{2C} . Then the selective deaths in population C and B are

$$2N \log(x_{2C}/x_{1C}) \quad \text{and} \quad 2N \log(x_{2B}/pX_{2C}) \quad (9)$$

So most of the selective death will have been in population C when $pX_{2C} \gg x_{1C}$. I.e. most of the action is getting the allele to appreciable frequency, the selection after admixture is minor if the admixture prop. is not low.

Seems like there's two questions: how much of the selective death is shared? Is there evidence of selection at all in population B or is drift and admixture sufficient to explain the frequency of the allele in population B , with selection only occurs in population C .

What do we mean, or can we say about, convergence when we see the linked variation/sweep? Knowledge of variation linked to a selected site may enable us to address questions of convergence that would not be possible otherwise with certain phylogenies or only observing the selected allele.

If we observe sister populations sharing a selective event, we can ask if this is convergent (i.e. shared deaths not overlapping) or if this selective event occurred in their ancestor (i.e. fraction of shared deaths greater than zero [maybe come up with variable for this term?](#)). If the selected event is shared and the sweep is recent, we expect the regions around the selected site to be more similar than if the sweeps happened independently and are truly convergent. We can formalize this in terms of coancestry between populations.

Thinking about sister populations in Figure 1, A and B , we can define the coancestry due to both neutral processes and selection as $(f + \omega)_{ab}$ where f_{ab} alone specifies the coancestry between the populations due to drift and admixture. If the selective events occurred independently from new mutation of the beneficial allele in both populations A and B , the coancestry between them at loci near the selected allele is simply what we expect under neutrality, f_{ab} .

The coancestry will increase if selection is on the same ancestral standing variant present in the ancestor of the populations, and is a function of the frequency of the standing variant, g and the amount of time, t , once populations A and B have split from their ancestor before selection started. Assuming the standing variant was previously neutral or was maintained in the population at some low frequency by balancing selection,

$$(f + \omega)_{ab} = y^2 \left(e^{-2rt} \left(\frac{1}{1 + 4Nrg} + \frac{4Nrg}{1 + 4Nrg} f_{ab} \right) + (1 - e^{-2rt}) f_{ab} \right) + (1 - y^2) f_{ab}. \quad (10)$$

where y^2 represents the probability of both linked lineages failing to recombine off the beneficial allele. This can be approximated as e^{-rt_s} where t_s is the duration of the sweep phase. This increase in coancestry, that decays with distance from the selected site, is due to the fact the region around the beneficial allele looks more similar if the variant is shared. However, as the amount of time the beneficial allele is standing independently in the sister populations before selection occurs increases, we expect this similarity to decrease as recombination is occurring independently in the populations ([something tying this to idea of shared deterministic vs. stochastic events Jeremy was talking about?](#)).

If the selective event is shared completely (i.e. the sweep occurs in the ancestor of populations A and B before they split), there is a further increase in coancestry. In this case the fraction of shared deaths is 1.

$$(f + \omega)_{ab} = y^2 + (1 - y^2) f_{ab} \quad (11)$$

Even if the selective event is partially shared such that selection occurs in the ancestral population (for time t_1) and continues independently in the daughter populations (for time t_2), this takes the same form as (13) if $t_1 + t_2 = t_s$.

$$\begin{aligned} (f + \omega)_{ab} &= e^{-rt_2} (e^{-rt_1} + (1 - e^{-rt_1}) f_{ab}) + (1 - e^{-rt_2}) f_{ab} \\ &= y^2 + (1 - y^2) f_{ab} \end{aligned} \quad (12)$$

Therefore is may not be possible to distinguish cases where fraction of shared death is greater than zero such that selective event is shared completely or partially. However, it is possible to detect if truly convergent i.e. there is no overlap in selected deaths.

Additionally, knowledge of linked variation can enable us to parse whether the increase of selected variant in population B is due to drift and admixture with population C alone (Figure 1) or is convergence. We can estimate coancestry between B and C , f_{bc} , from unlinked neutral sites that incorporate effects of drift and admixture. There will be increased coancestry around the selected site if selection is occurring in both populations. Now,

$$(f + \omega)_{ac} = y^2 e^{-r\delta} + (1 - y) f_{bc} + y(1 - y e^{-r\delta}) f_{cc} \quad (13)$$

where δ is the delay in the sweep time between populations B and C , a function of the strength of selection and migration rate.

What do we mean by convergence when we are thinking of quantitative traits? Should we also cover when phenotypes alone are seen? in that case using covar matrix can help in QST style analysis. Idea of double sign test?

Issues about meaning of convergence when there is incomplete lineage sorting

Long term signal? Can long term selection/drift on trait or variant, erode signal of shared history. E.g. if stabilizing selection acts separately on a shared trait in two (now independent populations) can we fix alternate solutions, even though we original shared pool. Similar Q about shared haplotypes whittled away by recom/migration. –One Q is do we care?