Supplemental Information for:

‘Quantifying the risk of hemiplasy in phylogenetic inference

Rafael F. Guerrero & Matthew W. Hahn

This file contains an Appendix and Supplemental Figures 1-3

www.pnas.org/cgi/doi/10.1073/pnas.1811268115

Appendix: Probabilities of mutation along genealogies

To clarify the calculations of P_e and P_o in equations 1 and 2, we must specify how the probability of mutation is calculated for each branch. On random branch lengths (as in the case of coalescent times), the probability of mutation is $\int 1 - \mathrm{e}^{-\mu x} f(x) \mathrm{d}x$, where $f(x)$ is the probability density function of x (the random variable for branch length). All mutation probabilities $\nu(\lambda_i,\tau)$ have some version of that general form, varying the value of x and $f(x)$. Genealogies β and γ can only happen in the absence of coalescence between lineages B and C in the BC ancestor. In contrast, genealogy α can happen with or without coalescence in BC, and it is helpful to consider the two alternatives separately: α_o denotes the subset of α trees that coalesce in the ABC ancestor (i.e., without coalescence in BC), whereas α_+ are the trees that have the α topology and coalesced in BC.

The genealogies α_o, β, and γ are identical in length (although they have different tip identities), and their mutation probabilities can be described by the equations:

$$\nu_1 = \frac{1}{\Lambda}\int_0^{t_3} (1 - e^{-\mu\,(t_1+t_2+x)})\frac{3}{2}(e^{-x} - e^{-3x})\,\mathrm{d}x$$

$$\nu_2 = \frac{1}{\Lambda}\int_0^{t_3} (1 - e^{-\mu\,(t_1+t_2+x)})3e^{-3x}(1 - e^{-(t_3-x)})\,\mathrm{d}x$$

$$\nu_4 = \frac{1}{\Lambda}\int_0^{t_3} 3e^{-3y}\left(\int_0^{t_3-y} e^{-x}(1 - e^{-\mu x})\,\mathrm{d}x\right)\mathrm{d}y$$

$$\nu_5 = \frac{1}{\Lambda}\int_0^{t_3} (1 - e^{-\mu\,(t_3-x)})3e^{-3x}\,\mathrm{d}x$$

In the above, $\Lambda = 1 + \frac{1}{2}e^{-3t_3} - \frac{3}{2}e^{-t_3}$ and represents the probability of coalescence of A, B, and C in the ABC ancestor (i.e., the cumulative distribution function of the coalescent for a sample of size 3). Each of these four probabilities correspond to multiple branches in Fig. 1B. Specifically, $\nu_1 = \nu(\lambda_1, \alpha_o) = \nu(\lambda_2, \beta) = \nu(\lambda_3, \gamma)$, $\nu_2 = \nu(\lambda_2, \alpha_o) = \nu(\lambda_3, \alpha_o) = \nu(\lambda_1, \beta) = \nu(\lambda_3, \beta) = \nu(\lambda_1, \gamma) = \nu(\lambda_2, \gamma)$, $\nu_4 = \nu(\lambda_4, \alpha_o) = \nu(\lambda_4, \beta) = \nu(\lambda_4, \gamma)$, and $\nu_5 = \nu\,(\lambda_5, \alpha_o) = \nu\,(\lambda_5, \beta) = \nu\,(\lambda_5, \gamma)$.

The mutation probabilities for genealogy α_+ are:

$$\nu\,(\lambda_1, \alpha_+) = \frac{1}{1 - e^{-t_3}} \int_0^{t_3} (1 - e^{-\mu\,(t_1 + t_2 + x)}) e^{-x}\,\mathrm{d}x$$

$$\nu\,(\lambda_2, \alpha_+) = \nu\,(\lambda_3, \alpha_+) = \frac{1}{1 - e^{-t_2}} \int_0^{t_2} (1 - e^{-\mu\,(t_1 + x)}) e^{-x}\,\mathrm{d}x$$

$$\nu\,(\lambda_4, \alpha_+) = \int_0^{t_2} \frac{e^{-y}}{1 - e^{-t_2}} \left(\int_0^{t_3} (1 - e^{-\mu\,(t_2 - y + x)}) \frac{e^{-x}}{1 - e^{-t_3}}\,\mathrm{d}x \right) \mathrm{d}y$$

$$\nu\,(\lambda_5, \alpha_+) = \frac{1}{1 - e^{-t_3}} \int_0^{t_3} (1 - e^{-\mu\,(t_3 - x)}) e^{-x}\,\mathrm{d}x$$

We use these functions, together with the probabilities of their corresponding genealogies ($p(\alpha_+) = 1 - e^{-t_2}$, and $p(\alpha_o) = p(\beta) = p(\gamma) = \frac{1}{3}e^{-t_2}$), to obtain the values of P_e and P_o in equations 1 and 2.

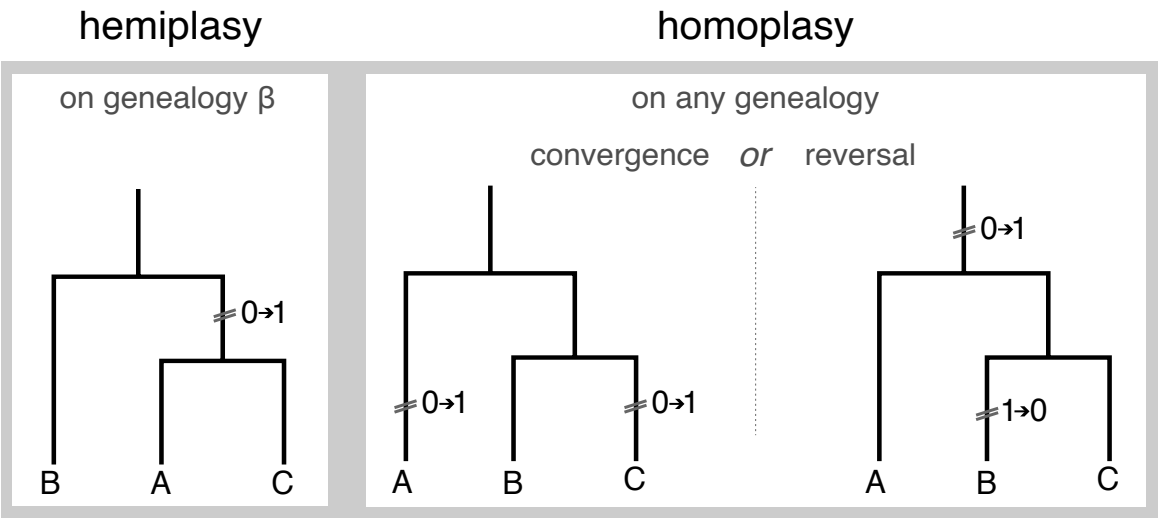

Figure S1. Diagram of possible scenarios of trait discordance.

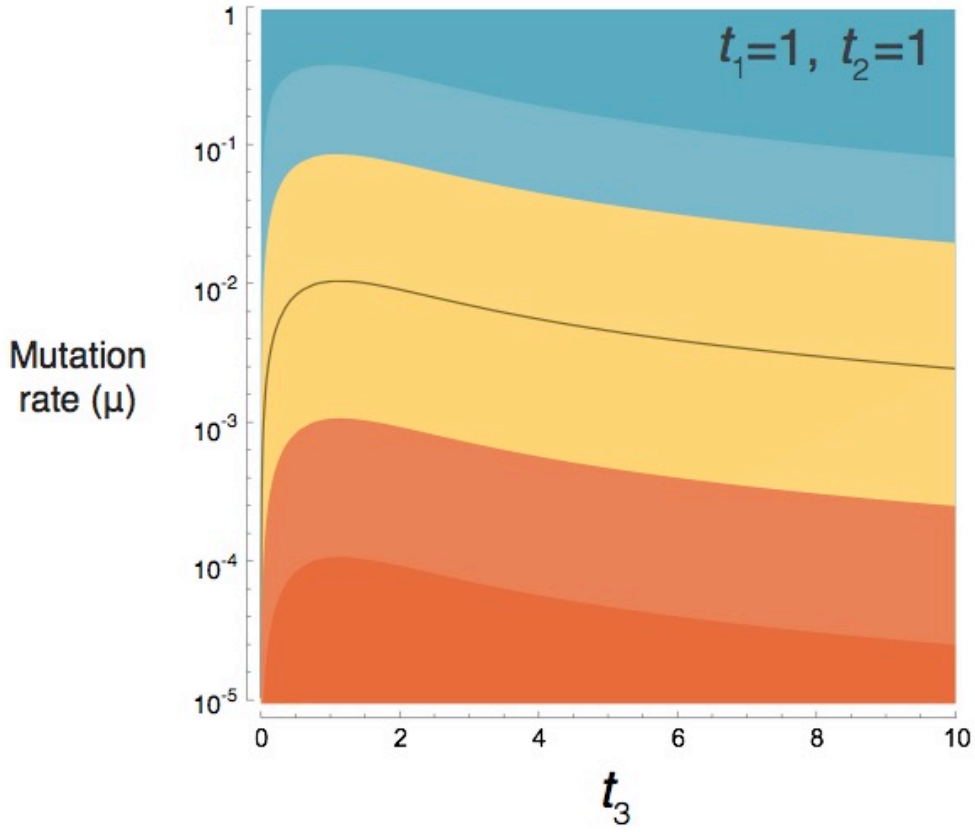

Figure S2. The relative probabilities of hemiplasy and homoplasy (P_e/P_o) are not dramatically affected by the length of the ancestral branch (t_3; in units of $2N$ generations). Contours are as for Figure 2 (dark orange = hemiplasy more than 100x more likely than homoplasy). Rate of mutation shown along the vertical axis (μ, per $2N$ generations). The black solid line represents parameter values for which hemiplasy and homoplasy are exactly equal ($P_e/P_o = 1$).

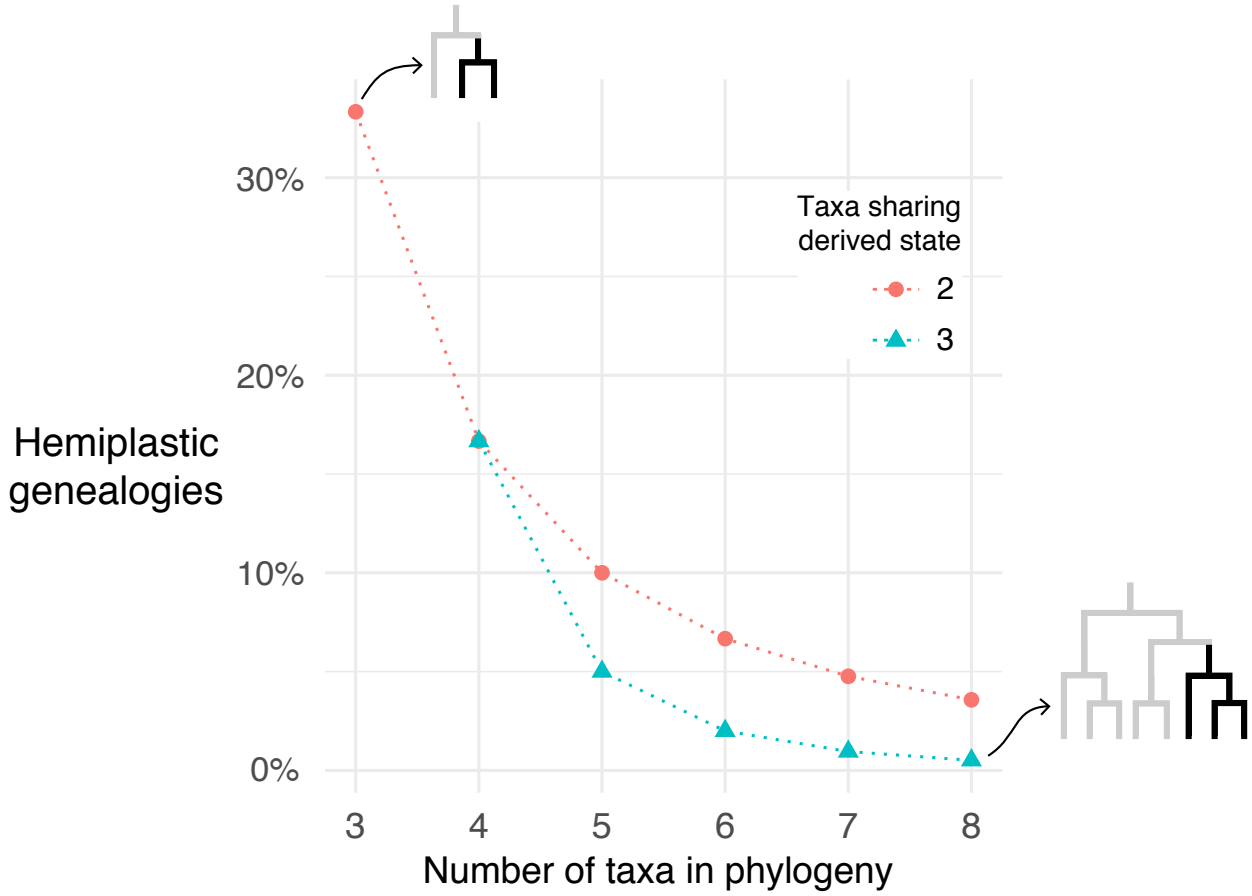

Figure S3. The fraction of genealogies that allow for a single hemiplastic substitution declines with the number of taxa in the phylogeny and the number of taxa that have the derived state (two in red, three in blue). The y-axis is the fraction of topologies that place the taxa sharing the derived state in a single clade ("hemiplastic genealogies"). This is calculated as $h(k)h(n-k+1)/h(n)$, where $h(x) = x!\,(x-1)!/2^{x-1}$, k is the number of taxa with the derived state, and n is the total number of taxa in the phylogeny. The numerator is the number of topologies that contain the subclade grouping the k taxa of interest, and the denominator is the total number of topologies for n taxa. For instance, in the top left corner one out of three possible topologies can lead to hemiplasy (as in the main text, genealogy β). As the tree gets larger or more taxa share the derived state, the fraction of trees that have a single branch upon which a shared hemiplastic substitution can occur gets smaller. These calculations assume a hard polytomy of n taxa, in which all samples coalesce in the ancestor (so all topologies are equally likely). It therefore represents a best-case scenario for hemiplasy.