Data employed:

A random sample of size 8000 from Krish's overall file

A random sample of size 8000 from Krish's ancient repeat file

Neutral g scores are defined as:

sqrt(neut_n)*(neut_pid-ave(neut_pid)) /

sd(sqrt(neut_n)*(neut_pid-ave(neut_pid))

All g scores are defined as:

sqrt(all_n)*(all_pid-ave(neut_pid)) /

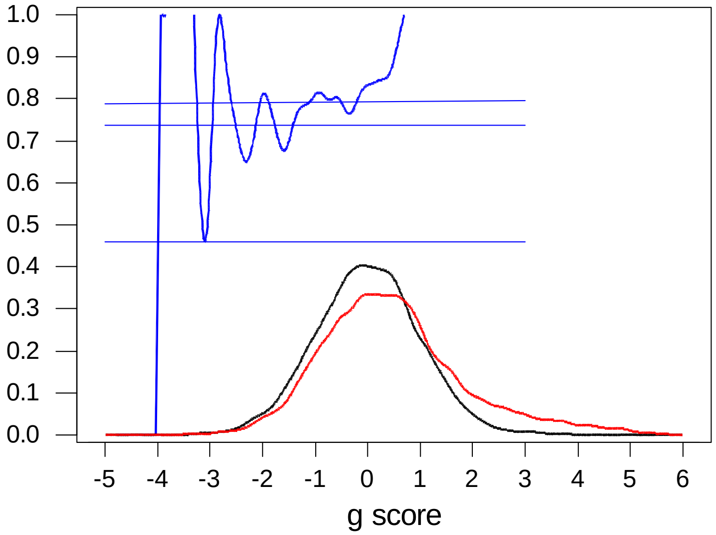
sd(sqrt(neut_n)*(neut_pid-ave(neut_pid))

(using neutral centering and rescaling coefficients)

The mixture model is:

f_all(g) = p_o f_neut(g) + (1-p_o) f_sel(g)

Using the observations available from f_all and f_neut, we create smoothed densities, and thus estimate p_o and the unobserved f_sel.

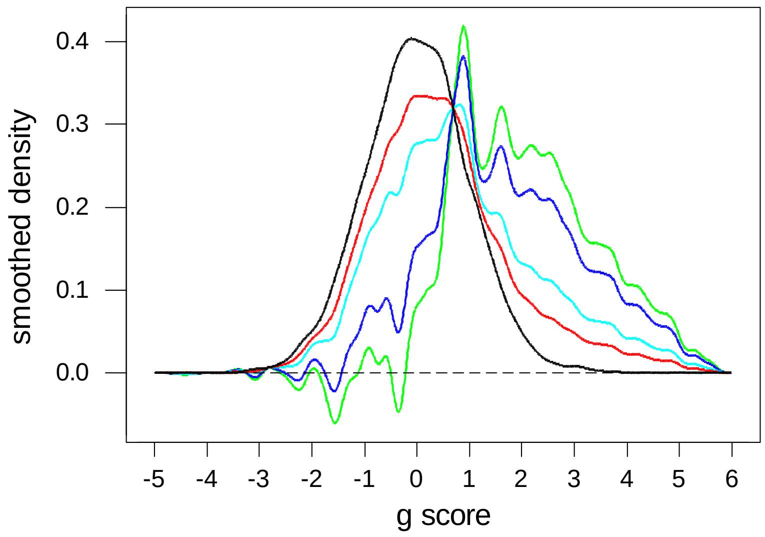


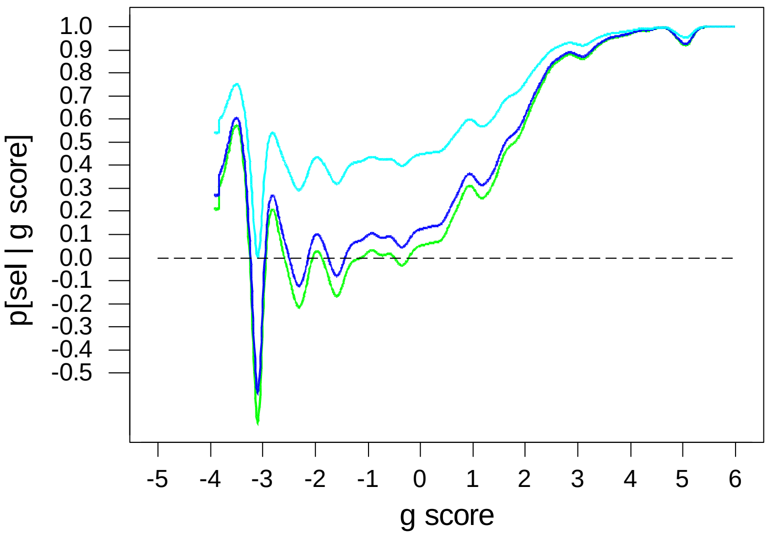
Gaussian Kernel smoothed densities for neutral g (black) and all g (red), implemented in S+ with width parameter=0.5 (associated st err ~ 0.5/4). The blue curve is the ratio of overall density to neutral density. The blue horizontal line represent minimum (0.46034), first quartile (0.73760) and median (0.79626) of ratio values below 1 that fall in the "relevant range" (where the neutral density is not ~0).

Theoretically, the share of all g compatible with neutrality (p_o) ought to be the minimum value of the ratio:

p_o = min f_all(g) / f_neut(g)

A pragmatic estimate of it should fall between 0.46 and 0.79, and reasonably be around 0.73. This would correspond to estimating the functional share, for the portion of the genome covered by the alignment data, to be between 0.54 and 0.21, reasonably around 0.27.





The black and red curve represent again neutral and all g smoothed densities. The other curves represent densities for functional g, when using p_o=median (green), first quartile (blue) and min (cyan). These are computed as

$$f\_sell(g) = (f\_all\ (g) – p\_o\ f\_neut(g)) / (1-p\_o)$$

Notice how the median estimate would put a non negligible portion of the curve below 0.

David's estimate rationale corresponds to "mirroring" the branch of the red curve (all g) that is on the left of 0 to the right, and measure what area remains unexplained on the right of 0. This is even more conservative than the median estimate here, producing a p_o above 0.80 (and thus a figure for the functional share below 0.20).

The curves represent the probability of being under selection (coming from f_sel) at various levels of g, computed as

$$p[sel\ |\ g\ ] = 1 – p\_o\ (f\_neut(g) / f\_all(g))$$

with the median (green), first quartile (blue) and min (cyan) estimate of p_o. I would ignore the oscillations on the left of –3 (very small relative oscillations in f_all and f_neut, see graph on left), but notice again how the median estimate would put a non negligible portion of the curve below 0 between –3 and 0.

NOTE: the f_sell (g) and p[sel | g] curves will become smoother as we increase the degree of smoothing producing the f_all and f_neut curves. The degree of smoothing in these calculations was very low. We will also repeat the calculations using a t-distribution instead of a non-parametric smooth for the neutral density.