Supplementary Information - Additional Material and Methods

empiricIST: An Integrated Software and analysis Tool for analyzing deep-mutational scanning data (such as empiric)

Outlier detection and data imputation

First, amplification error during the sequencing procedure – amongst other factors -- can result in sequencing errors at single time points. To identify and correct for these individual false data points, Bank *et al.* (2014) showed that estimation accuracy can be optimized using the DFBETA statistic with a (conservative) cutoff of 2, based on the log ratio of the mutant's read number to the total number of reads at each individual time point (i.e., $log\,(\frac{n_{i,t}}{n_t})$). In particular, this regression-based statistic measures the difference in each parameter estimate when a specific data point is excluded. If this difference surpasses a chosen threshold, the corresponding data point has a large effect on the parameter estimates and is called influential, and indicative of an outlier. In our analyses, these data points (for the respective mutant and time point) were removed (i.e., set to zero). Excluded data points were summed up and added in an additional row for each time point at the end of the data set to preserve the overall read number for the multinomial sampling (and hence the multinomial probabilities). There is reason to believe that not all of these removed data points are true outliers, but biologically interesting non-log-linear effects. However, their majority should be sequencing errors that affect the accuracy of all growth rate estimates, rendering their exclusion necessary. Note that this approach uses the so-called total normalization (i.e., $log\,(\frac{n_{i,t}}{n_t})$), which has been shown to bias growth rate estimates (Matuszweski *et al.* 2016). This uniquely allows to detect outliers in the reference genotype, which, if undetected, would introduce a systematic bias in the growth rate estimates of all other mutants. If the reference genotype is sampled with great accuracy (i.e., does not contain any outliers), detecting outliers based on log ratio using the so-called wild-type normalization (i.e., $log\,(\frac{n_{i,t}}{n_{1,t}})$) should yield more accurate results. However, since data from previous DMS studies suggest (Bank *et al.* 2014) that wild-type sequences are often consistently mismeasured and since these mismeasurements can only be detected by using the total normalization, only this approach has been implemented into empiricIST at this point.

If there is only a single outlier in the data set, this data point is imputed (i.e., corrected) by updating the corresponding log-ratio. Particularly, the imputed log-ratio for mutant i at time point t, $\tilde{x}_{i,t}$, is given by

$$\tilde{x}_{i,t} \tilde{} N(s_{x_{i,t}}\, t,\, \epsilon_{x_{i,t}}),$$

where $s_{x_{i,t}}$ and $\epsilon_{x_{i,t}}$ denote the slope respectively residual variance of the linear regression excluding the outlier. The updated log-ratio is then translated into an updated read number. Note that due to the total normalization, updating a log-ratio changes all other log-ratios as well. This becomes problematic when there are multiple identified outliers, as it raises the questions which outlier should be imputed first (as this will affect all others too), and how to deal with all other data points. Here, we

took again a conservative approach and always imputed the outlier with the largest absolute studentized residual. All log-ratios were then updated and potential outliers re-identified. Note that due to swamping (i.e., the misclassification of a regular data point as outlier), imputing a single outlier might reduce the number of total outliers by more than one. In contrast, imputing a single outlier might also make previously masked outliers visible. We chose to impute only those data points that have initially been classified as outliers to minimize changes to the original experimental data. Thus, once a data point is classified as a non-outlier it will no longer be a candidate for being imputed.

DFE tail-shape estimation

All three domains are contained in the so-called generalized Pareto distribution (GDP; Pickands 1975), whose probability density function is given by:

$$f(\kappa, \xi) = \{ \frac{1}{\xi}(1 + \frac{\kappa x}{\xi})^{-(1+\frac{1}{\kappa})},\ x \geq 0, \textit{if } \kappa > 0 \ \frac{1}{\xi}(1 + \frac{\kappa x}{\xi})^{-(1+\frac{1}{\kappa})},\ 0 \leq x < -\frac{\xi}{\kappa}, \textit{if } \kappa < 0 \ \frac{1}{\xi} exp\,(-\frac{x}{\xi}),\ x \geq 0,\ \textit{if } \kappa = 0$$

where κ and ξ denote the shape and scale parameter, respectively. The domain of attraction of the GDPis solely determined by the shape parameter κ. In particular, if $\kappa > 0$, the beneficial tail of the DFE falls into the (heavy-tailed) Fréchet domain, if $\kappa < 0$ it falls into the (truncated) Weibull domain, and if $\kappa = 0$ it falls into the Gumbel domain containing the exponential distribution.

Using the re-parameterization $\sigma = \frac{\xi}{\kappa}$, a direct maximum likelihood estimate of the shape parameter κ of the underlying GPD can be obtained by maximizing the the profile-log-likelihood function of the GPD

(S1) $$log\ L(\sigma) = n(-log\ (\kappa(\sigma)\ \sigma) + \kappa(\sigma) - 1)$$

with

(S2) $$\kappa(\sigma) = \frac{1}{n} \sum log\ (1 - \frac{x_i}{\sigma}),$$

for $x = (x_1, \ldots, x_n)$ data points with respect to σ (Castillo & Serra 2015).
To account for unobserved small-effect mutations ("censored data"), we furthermore implemented a shifting data approach (see Beisel *et al.* 2007), where all (beneficial) selection coefficients are shifted by the smallest observed beneficial selection coefficient.

Finally, we implemented a likelihood-ratio test, to assess whether the beneficial tail of the DFE follows an exponential distribution (i.e., whether $H_o : \kappa = 0$). For that we first calculated the likelihood-ratio test statistic

(S3) $$-2log\ (\Lambda) = 2(L(X|\sigma) - (L(X|\sigma))$$

and then generated data sets under the null model using the estimated parameter $\kappa(\hat{\sigma})$. Finally, the likelihood-ratio test statistic $-2log(\Lambda)$ is calculated over all simulated data sets and an approximate p-value is obtained by comparing the empirical distribution of the test statistic to that obtained from the data (Beisel *et al.* 2007).

Bibliography

Bank C, Hietpas RT, Wong A, Bolon DN, Jensen JD (2014). A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: Uncovering the potential for adaptive walks in challenging environments. Genetics 196: 841–852.

Beisel CJ, Rokyta DR, Wichman HA, Joyce P (2007). Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics 176: 2441–2449.

Castillo Jd, Serra, I, Likelihood inference for generalized Pareto distribution, Computational Statistics & Data Analysis, Volume 83, 2015, Pages 116-128.

Matuszewski S, Hildebrandt ME, Ghenu AH, Jensen JD, Bank C (2016). A statistical guide to the design of deep mutational scanning experiments. Genetics 204: 77–87.

Pickands, J. (1975) Statistical Inference Using Extreme Order Statistics. The Annals of Statistics 3 (1): 119-131.