SUPPLEMENTARY NOTES

Sequence encoding and model coefficients

Sequences are parsed as binary strings which are used as input to the model. Each locus in the sequence is parsed into a binary string with length equal to the number of possible alleles at that locus. At each position in this string a 1 indicates the presence of an allele and a 0 its absence. In theory it is possible to have different numbers of alleles for each locus (as was the case in the data analyzed in Hinkley et al. (2011)). An RNA sequence of length \( n \) has \( n \) loci, each with 4 possible alleles, resulting in a binary representation of length \( 4n \).

Sequence fitness is expressed using equation 3 in the main text,

\[
\log(y(s)) = I + \sum_{ij} m_{ij} s_{ij} + \sum_{ijkl} \epsilon_{ijkl} s_{ij} s_{kl},
\]

which defines for sequences of \( n \) loci, with 4 alleles per locus the following coefficients,

- 1 intercept, \( I \).
- \( 4n \) main effects, \( m_{ij} \).
- \( \binom{n}{2} \cdot 4^2 \) pairwise epistatic effects, \( \epsilon_{ijkl} \).

Note that the coefficients in equation 3 are not determined uniquely without the regularization parameter \( \lambda \).

The generalized kernel ridge-regression (GKRR) implementation we use (Hinkley et al., 2011), is applied directly to the binary encoding of each sequence, without any information about the number of loci per allele. Thus, it defines \( \binom{n}{2} \cdot 4^2 \) pairwise coefficients, resulting in \( 6n \) extra “nonsense” coefficients. These additional coefficients represent the pairwise effects of simultaneously observing 2 different alleles at the same locus, which is not allowed to happen in our situation. (However, in a hypothetical scenario this could be used when different alleles are present at known frequencies). These “nonsense” coefficients are not explicitly set to 0, but are in practice close to 0 (personal communication, Trevor Hinkley). Thus, for the sequences used here, where \( n \) is 217, the model has 377,147 coefficients, of which 1,302 are superfluous.

In the sequence encoding used here sequences are encoded independent of any reference sequence, as opposed to other representations where sequences are encoded as mutations made to a wild type sequence (see for instance Otwinowski and Plotkin (2014)). Our representation results in sequences that are longer and a model with more parameters, but it allows for a more general treatment of the sequence landscape in that the sequence representation is not tied to an arbitrary sequence.
Measuring predictive power

Deviance is a standard measure for generalized models and is analogous to the coefficient of determination, $R^2$, of linear models with normal error structures. The deviance of a model is defined as the difference between the log-likelihoods of the model and a complete model (a model with a parameter for every observation such that it fits the data perfectly), multiplied by $-2$ (Nelder and Wedderburn, 1972). GKRR assumes a Poisson error structure (Supplement, Section 1.2, Hinkley et al. (2011)), which results in the following formula for the deviance,

$$D = 2 \sum_{i=0}^{N} y_i \log\left( \frac{\hat{y}_i}{y_i} \right) - (y_i - \hat{y}_i),$$

where $N$ is the number of data points.

Predictive power is measured as the fraction of the deviance explained. This is measured as the improvement over a null model, which is equivalent to fitting only the mean of the data points (i.e. $\hat{y}_i = \bar{y}$ for all $i$) and thus maximizes the deviance. If the deviance of the null model is $D_N$ then the fraction of the deviance explained by a model is given by $\frac{D_N - D}{D_N}$.

Extrapolating from the independent effects of single and pairwise mutations

The quadratic model that we use assumes that only main effects and pairwise epistatic interactions contribute to the fitness of a sequence. In this section we investigate if it is possible to extrapolate from the independent fitness effects of single and pairwise mutations and thereby assess the contribution of higher-order interactions to sequence fitness. We train a quadratic model on the dataset containing all sequences reachable from the focal genotype within its 2-mutational neighbourhood (defined as $D_2$ in Methods). This dataset contains 211,576 sequences representing the independent fitness effects of all possible single and double mutants of the focal genotype. Although these are less sequences than the number of coefficients in the model (see above), assuming that the fitness landscape contains only main effects and pairwise epistatic interactions, we would expect that it is possible to extrapolate to sequences with higher numbers of mutations using a model trained on this dataset. On the other hand, if higher-order interactions play a significant role, we expect to observe a decrease in the quality of the prediction as more mutations are added to the test sequences.

Fig. S1 shows the result of evaluating the model on 6 datasets of 5,000 sequences each, randomly sampled at increasing Hamming distances from the focal genotype. For comparison, the sixfold cross-validated result of the training set is also shown (produced by randomly sampling 6 sets of 5,000 sequences from the dataset for testing, and training on the remaining sequences). We observe a steep decline in the
predictive power for datasets consisting of sequences with more mutations than the training sequences, associated with a parallel rise in the correlation between the size of the residuals and the true fitness, indicating a tendency to underestimate high fitnesses and overestimate low fitnesses. For datasets sampled at more than 5 mutations from the focal genotype the model cannot explain any of the variance. Thus, it is apparent that higher-order interactions are present within the fitness landscape and that it is impossible to represent the fitness of sequences with higher numbers of mutations as a combination of independent fitness effects of single and pairwise mutations.
SUPPLEMENTARY FIGURES AND TABLES

**FIG. S1.** Ability of a quadratic model to extrapolate when trained on the independent fitness effects of all single and pairwise mutations of the focal genotype. The model was trained on all one and two step mutants of the focal genotype (211,576 sequences) and tested on 6 datasets of 5,000 sequences each, randomly sampled at Hamming distances of 3, 4 and 5 from the focal genotype. Data points are the mean predictive power (blue) and correlation between true fitnesses and residuals (orange) among replicates. Error bars indicate the standard error of the mean. For comparison the sixfold cross-validated results are shown for a Hamming distance of 2, where the training set was randomly split into training and test sets of 206,576 and 5,000 sequences respectively. Because of the cross validation the prediction is not perfect and there is a substantially higher correlation between the true fitness and the residuals.

**Table S1.** Statistics of the four datasets used to assess the predictive power under different sampling regimes. All datasets contain 65,000 (65,536 for Complete Subset) sequences of length 217 bp. Values for the Hamming distances and Fitness values are the median followed by the 95 percent confidence intervals.

|                  | Unique sequences | Mutable loci | Hamming distance from target | Pairwise distance | Fitness             |
|------------------|------------------|--------------|-------------------------------|-------------------|--------------------|
| Random           | 65000            | 217          | 163[150,175]                  | 164[151,176]      | 0.175[0.0586,0.425]|
| Random Neighbourhood | 65000          | 217          | 8[8,8]                        | 17[15,17]         | 0.754[0.209,3.8]   |
| Complete Subset  | 65536            | 8            | 6[3,8]                        | 7[4,9]            | 3.28[0.911,8.79]   |
| Evolved          | 65000            | 217          | 97[84,105]                    | 46[16,66]         | 57.3[17.7,83.7]    |
FIG. S2. Trajectories of three populations evolved under strong selection. Each population initially contains a monomorphic population composed of 1,000 copies of the fittest sequence among 100 randomly drawn sequences at a Hamming distance of 20 from the focal genotype of the fitness landscape. Population size was kept constant and populations were evolved for 20,000 generations with a genomic mutation rate of 0.001. The first 10,000 generations (left of the dotted red line) were discarded and subsequently sampled unique sequences from every 50 generations until 65,000 unique sequences were collected. The dataset produced from (A) in this manner was used for Evolved, shown in Fig. 1 and 2.
du Plessis et al. · doi:10.1093/molbev/mst012

FIG. S3. Effect of the training set size on the predictive power of a quadratic model trained on sequences sampled from Random Neighbourhood and Evolved. Each model was tested on 5,000 unseen sequences. Sixfold cross-validation was used to assess the predictive power and biases of the model. Error bars indicate the standard error of the mean.

FIG. S4. Effect of the training set size on the predictive power of a linear model trained on sequences sampled from Random Neighbourhood and Evolved. Each model was tested on 5,000 unseen sequences. Sixfold cross-validation was used to assess the predictive power and biases of the model. Error bars indicate the standard error of the mean.

FIG. S5. Predictive capacities of quadratic and linear models trained on 10 replicate Complete Subset datasets where the 8 mutable loci were chosen at random. Each dataset contains 65,536 sequences and was randomly split into training and test sets of 60,000 and 5,536 sequences respectively. Sixfold cross-validation was used to assess the predictive power and biases of the linear and quadratic models on the simulated datasets. Error bars indicate the standard error of the mean.
FIG. S6. The distributions of the 95% intervals of the relative residuals for models fit on different datasets with the linear and quadratic models. The figure shows only the distributions of one of the cross-validated replicates. The red lines indicate the median and the interquartile range.

FIG. S7. The distributions of the 95% intervals of the mean scaled residuals for models fit on different datasets with the linear and quadratic models. The figure shows only the distributions of one of the cross-validated replicates. The red lines indicate the median and the interquartile range.

FIG. S8. Correlation between the true fitness value and the residuals under different sampling regimes. Four datasets were sampled from the same quasi-empirical RNA fitness landscape using different sampling regimes. Each dataset contains 65,000 sequences (65,536 for Complete subset) and was randomly split into training and test sets of 60,000 and 5,000 sequences (5,536 for Complete Subset) respectively. Sixfold cross-validation was used to assess the predictive power and biases of the linear and quadratic models on the simulated datasets. Error bars indicate the standard error of the mean.
**FIG. S9.** Scatterplots of predicted against true fitness for the different datasets using the linear model. The distributions at the top and right of plots indicate respectively the distributions of true and predicted fitnesses. The solid green line indicates a perfect prediction. The shaded red area contains 95% of the points. The solid red line indicates the median bias in the residuals.
FIG. S10. The effect of neutral drift on the ability of models trained on an evolved ensemble to correctly predict the fitness of unseen sequences. (A) Statistics of the evolved population as it is evolved past the last sequence in the training set (Evolved). Distributions of the pairwise Hamming distance between every sequence in the population and 5,000 randomly sampled sequences in the training set as well as the fitness distribution for the population are shown. Red lines indicate medians and interquartile ranges. (B) Predictive power of a quadratic model trained on 60,000 sequences (blue) as well as the correlation between true fitness and residuals (orange) are shown when applying the model on subsequent generations of the evolved population. (C) Predictive power of a linear model trained on 60,000 sequences (blue) as well as the correlation between true fitness and residuals (orange) are shown when applying the model on subsequent generations of the evolved population.
FIG. S11. The distribution of fitness effects for all single mutants of 100 sequences sampled at random from Random (A), Random Neighbourhood (B), Complete Subset (C) and Evolved (D) respectively. The figures show the true distributions as well as the approximations by a quadratic model. Each bar represents the mean value for the respective bin in the histograms of real and predicted fitness effects across the 100 sequences. Error bars represent the 95% confidence intervals of the means estimated from 1000 bootstrap replicates.
FIG. S12. The distribution of fitness effects for all single mutants of 100 sequences sampled at random from Random (A), Random Neighbourhood (B), Complete Subset (C) and Evolved (D) respectively. The figures show the true distributions as well as the approximations by a linear model. Each bar represents the mean value for the respective bin in the histograms of real and predicted fitness effects across the 100 sequences. Error bars represent the 95% confidence intervals of the means estimated from 1000 bootstrap replicates.
FIG. S13. The distribution of fitness effects for all double mutants of 100 sequences sampled at random from Random (A), Random Neighbourhood (B), Complete Subset (C) and Evolved (D) respectively. The figures show the true distributions as well as the approximations by a quadratic model. Each bar represents the mean value for the respective bin in the histograms of real and predicted fitness effects across the 100 sequences. Error bars represent the 95% confidence intervals of the means estimated from 1000 bootstrap replicates.
FIG. S14. The distribution of fitness effects for all double mutants of 100 sequences sampled at random from Random (A), Random Neighbourhood (B), Complete Subset (C) and Evolved (D) respectively. The figures show the true distributions as well as the approximations by a linear model. Each bar represents the mean value for the respective bin in the histograms of real and predicted fitness effects across the 100 sequences. Error bars represent the 95% confidence intervals of the means estimated from 1000 bootstrap replicates.
FIG. S15. The prevalence of different types of epistasis among all beneficial double mutants of 100 sequences sampled at random from Random (A), Random Neighbourhood (B), Complete Subset (C) and Evolved (D) respectively. The figures show the true distributions as well as the approximations by a quadratic model. Each bar represents the mean value for the respective bin in the histograms of real and predicted pairwise epistatic effects across the 100 sequences. Error bars represent the 95% confidence intervals of the means estimated from 1000 bootstrap replicates. Note that a similar amount of a given type of epistasis does not necessarily mean that it occurs on the same pairs of loci in the truth and prediction or that the effects are of a similar magnitude.
FIG. S16. The prevalence of different types of epistasis among all deleterious double mutants of 100 sequences sampled at random from Random (A), Random Neighbourhood (B), Complete Subset (C) and Evolved (D) respectively. The figures show the true distributions as well as the approximations by a quadratic model. Each bar represents the mean value for the respective bin in the histograms of real and predicted pairwise epistatic effects across the 100 sequences. Error bars represent the 95% confidence intervals of the means estimated from 1000 bootstrap replicates. Note that a similar amount of a given type of epistasis does not necessarily mean that it occurs on the same pairs of loci in the truth and prediction or that the effects are of a similar magnitude.
FIG. S17. Effect of the sampling density on the ability of a linear model to approximate a fitness landscape from randomly sampled sequences. Datasets are composed of 65,000 sequences randomly sampled within successively higher Hamming distances from the focal genotype. Datasets were randomly split into training and test sets of 60,000 and 5,000 sequences respectively and sixfold cross-validation was used to assess the predictive power and biases of the linear model on the simulated datasets. Data points are the mean predictive power (blue) and correlation between true fitnesses and residuals (orange) among replicates. Error bars indicate the standard error of the mean. For comparison Complete Subset and Evolved (shown in Fig. 3 and S9) are also shown.

Table S2. Statistics of the datasets used to assess the predictive power under different landscape sizes. Because the number of viable sequences increases exponentially with the Hamming distance most sequences in the datasets have the maximum allowed Hamming distance from the focal genotype. Hence, the median pairwise Hamming distance in the datasets is roughly twice the maximum allowed Hamming distance (but this effect decreases as the Hamming distance is increased) All datasets contain 65,000 sequences of length 217 bp. Values for the Hamming distances and Fitness values are the median followed by the 95 percent confidence intervals.

| Neighbourhood | Unique sequences | Unique loci | Hamming distance from target | Fitness value |
|---------------|-----------------|-------------|------------------------------|---------------|
| 2 bp          | 54,229          | 217         | 2 [2,2]                      | 4.25 [0.814,15.7] |
| 5 bp          | 65,000          | 217         | 5 [5,5]                      | 1.29 [0.343,8.09] |
| 10 bp         | 65,000          | 217         | 10 [10,10]                   | 0.604 [0.169,2.66] |
| 20 bp         | 65,000          | 217         | 20 [19,20]                   | 0.311 [0.0926,1.02] |
| 30 bp         | 65,000          | 217         | 30 [29,30]                   | 0.229 [0.0714,0.663] |
| 40 bp         | 65,000          | 217         | 40 [39,40]                   | 0.195 [0.0631,0.525] |
| 50 bp         | 65,000          | 217         | 50 [49,50]                   | 0.179 [0.0574,0.462] |
| 100 bp        | 65,000          | 217         | 100 [98,100]                 | 0.155 [0.0496,0.392] |