Testing similarity metrics

We developed a number of metrics to evaluate the accuracy of fitness landscapes generated by ALFA-K. Metrics were developed using ALFA-K fits to data from simulations, so that ground truth fitnesses would be available from the GRF artificial landscapes. Across all fits we saw a Pearson correlation median 0.83 between predicted and true fitness. Accuracy was positively correlated with the number of sample timepoints (Fig. S1A) and with λ (Fig. S1B), was negatively correlated with distance from frequent karyotypes (Fig. S1C), but had little correlating with the frequency threshold - a hyperparameter used in fitting (Fig. S1D). ALFA-K did fail to produce accurate fits in some cases (Fig. S1E). Our main concern was to design metrics to identify such instances, for applications where no ground truth fitness estimates will be available.

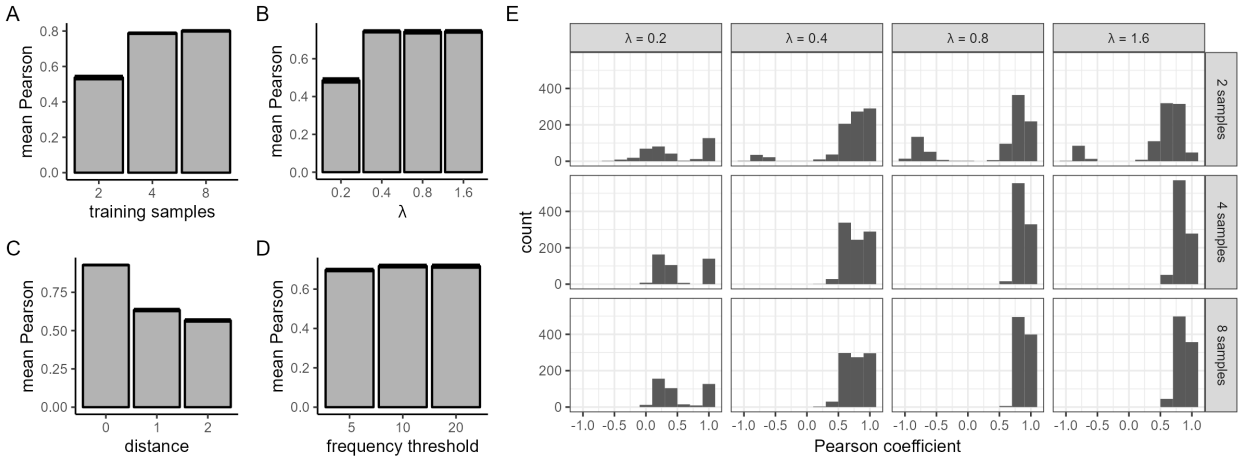

Figure S1. A-D) Pearson correlation coefficient of ALFA-K estimated fitness with ground truth fitness, grouped by A) Distance of karyotype from the nearest 'frequent karyotype' estimated in step 1 of the fit; B) Number of longitudinal samples used to train ALFA-K; C) frequency threshold for frequent karyotypes; D) GRF wavelength parameter λ. E) Pearson coefficients between actual fitness and ALFA-K estimated fitness of karyotypes.

We employed a leave-one-out cross validation procedure to test the ability of ALFA-K to estimate the fitness of karyotypes not present in the input data (Fig. S2). The fitness estimates from the first step of the ALFA-K pipeline (based on frequency changes of common karyotypes) are treated as "ground truth" and used as inputs for the cross validation test.

The cross validation metric performed extremely well at filtering out poor fits, with almost no fits having a positive R^2 but negatively correlating with the ground truth. However, the metric was perhaps overzealous in the sense that many simulations scored poorly despite having predicted fitnesses that correlated extremely strongly with the ground truth (it would be nice to explain why this happens, but I can't figure out why).

We also sought to identify metrics based on the combined ability of the ABM and ALFA-K to predict future karyotype population evolution. We developed and tested two metrics: the Wasserstein metric, and the angle metric. Suppose two initially indentical karyotype populations evolve independently across two fitness landscapes, which could be either identical, or different (as shown, Fig. S3A-B). Two vectors are computed connecting the centroid of the initial population to the centroids of each of the evolved populations (Fig. S3C). The angle metric is the angle between these two vectors. Alternatively the Wasserstein distance

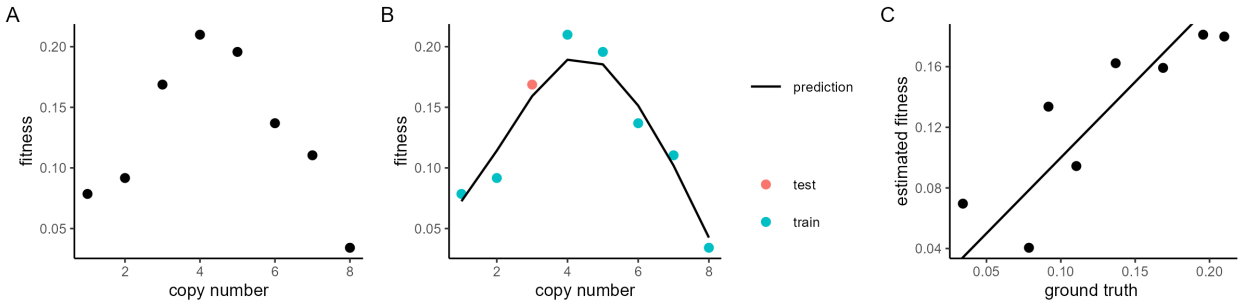

Figure S2 A) A simplified hypothetical example in which the fitness of a karyotype depends on the copy number of a single chromosome. B) In the cross validation procedure, each data-point is omitted from the training set in turn. The model is trained on the reduced dataset and then used to predict the fitness of the omitted clone. C) Cross validation predicted fitnesses are compared to the ground truth.

between the two evolved populations can be computed (Fig. S3D, d_3). If the populations are evolving slowly relative to the timescale of measurement, the distance d_3 will automatically be low. Therefore we normalise by the mean distance each population has travelled from the founder, giving the Wasserstein metric $M_w = 2d_3(d1 + d2)^{-1}$.

To aid intepretation of the Wasserstein and angle metrics, we generated pairs of vectors uniformly distributed on the surface of the 22-dimensional hypersphere and computed the values of the metrics for each vector pair (Fig. S4). The expected value of the angle metric was 90° while the Wasserstein metric was $\sqrt{2}$. Therefore one can say that recovering values less than these thresholds indicates better agreement than random chance.

We tested the ability of each metric to characterise whether an ALFA-K estimated landscape was accurate or not. We applied the R^2 metric to ALFA-K landscapes fit to data from our ABM test simulations. These fitted landscapes were also used as input to a second round of ABM simulations, to which we applied the angle and Wasserstein metrics. All three metrics were able to discriminate between landscapes that correlated strongly with the ground truth and those that did not.

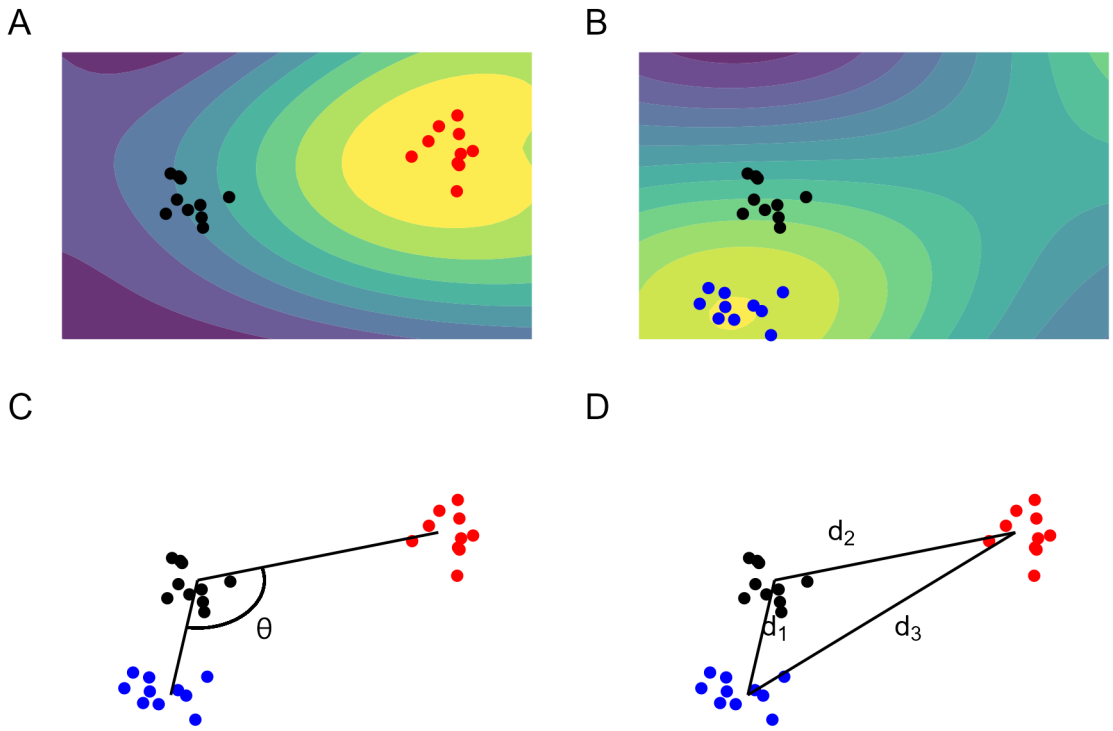

Figure S3. Metrics comparing similarity of population evolution. A-B) Examples of an initially identical founder population (black) evolving on two different landscapes. C) The vector angle between the centroids of the two evolved populations gives the angle metric. D) The Wasserstein metric is computed as the Wasserstein distance between the two evolved populations (d_3) normalised by the average distance the populations have travelled (d_1,d_2).

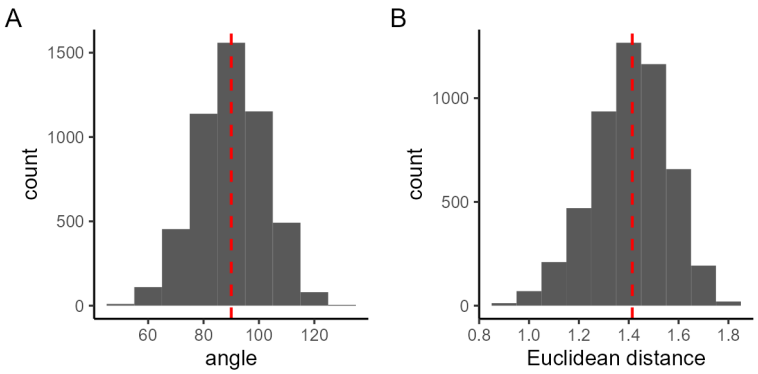

Figure S4. Metric expectation. Pairs of vectors were generated uniformly on the surface of a 22-dimensional hypersphere. The distribution of A) angles, and B) distances between each pair of vectors is shown.

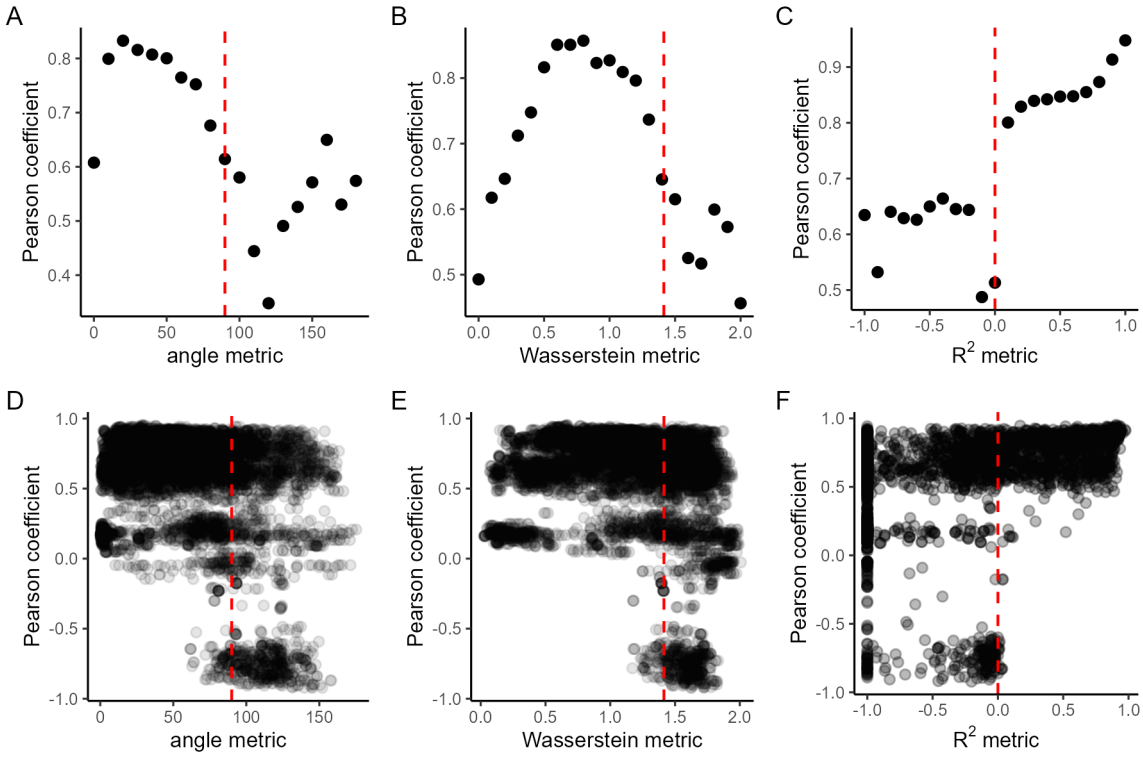

Figure S5. Evaluation of metrics. A-C) The mean value of Pearson correlation between ground truth and ALFA-K fitted landscape, for corresponding values of A) angle metric, B) Wasserstein metric, and C) R^2 metric. D-F) Comparison of Pearson correlation and D) angle metric, E) Wasserstein metric, and F) R^2 metric for all fitted landscapes.