README

Richard J Beck

2024-04-10

Validating ALFA-K against output of ABM simulations

Results

Clonal evolution on random fitness landscapes

The ALFA-K methodology predicts single-cell karyotype evolution using longitudinal data as input. Sequencing data from individual cells is used to identify temporal changes in karyotype frequency. The method estimates the fitness of common karyotypes from changes in their observed frequencies and extends these estimates to rarer ones, constructing a local fitness landscape via Gaussian process regression (Fig. 1A). To test the ability of ALFA-K to infer fitness landscapes, we developed ABM simulations of karyotype diversification and selection on randomly generated fitness landscapes. We used Gaussian random fields (GRF) as fitness landscapes, which were generated via summation of multiple spherical waves (Fig. 1B). Varying the wavelength parameter (λ) allows control of the complexity of the resulting landscape (Fig. 1C). ABM simulations of cell populations evolving on the GRF landscapes are characterized by expansion and contraction of karyotype-defined subclones (Fig. 1D,E) as fitter clones are generated and the fitness of the population increases (Fig. 2A). In these test simulations, populations on landscapes with low λ (Fig.1D) tended to experience punctuated evolution, whereas populations evolving on landscapes with high λ exhibited gradual increases in fitness over time (Fig.1E).

A) Schematic flowchart of steps in ALFA-K pipeline. B) GRF are generated by summing multiple spherical waves. The interference patterns generated by the waves result in complex unpredictable landscapes. C) Increasing the wavelength (λ) results in GRF with decreasing complexity. D-E) Example simulation output for ABM cell populations evolving on GRF fitness landscape with D) $\lambda = 0.1$ or E) $\lambda = 1.6$. Each coloured line represents the longitudinal frequency of a different karyotype.

Data from the first 200 days of each simulated population was used to train ALFA-K. We trained ALFA-K varying both the number of longitudinal samples and the hyperparameter N, then evaluated the results using a cross validation procedure (Fig.2B) which tests the ability of ALFA-K to infer the fitness of karyotypes withheld from the input data (see Methods). ALFA-K performance was not sensitive to the value of the hyperparameter N within the tested range. It was however sensitive to landscape complexity and the number of longitudinal samples in the input, with at least 4 samples needed to obtain satisfactory results. We next tested the ability of ALFA-K to predict population evolution for the time period from 200-300 days that was withheld from the training data. We used the angle metric (SX Fig.) to evaluate predictive performance, in which values below 90 degrees are taken as good predictions. Landscapes with good cross-validation scores ($R^2 > 0.3$) predicted future population evolution well (Fig. 2C). The results from the forward prediction tests agreed with the cross validation test in terms of sensitivity to landscape complexity, number of sampled timepoints, and the value of the hyperparameter N (Fig. 2D). Finally, we evaluated the robustness of ALFA-K to different values of missegregation rate (Fig. 2E). The procedure was robust to a wide range of missegregation rates up to a threshold value, which occurred when karyotype became too unstable to

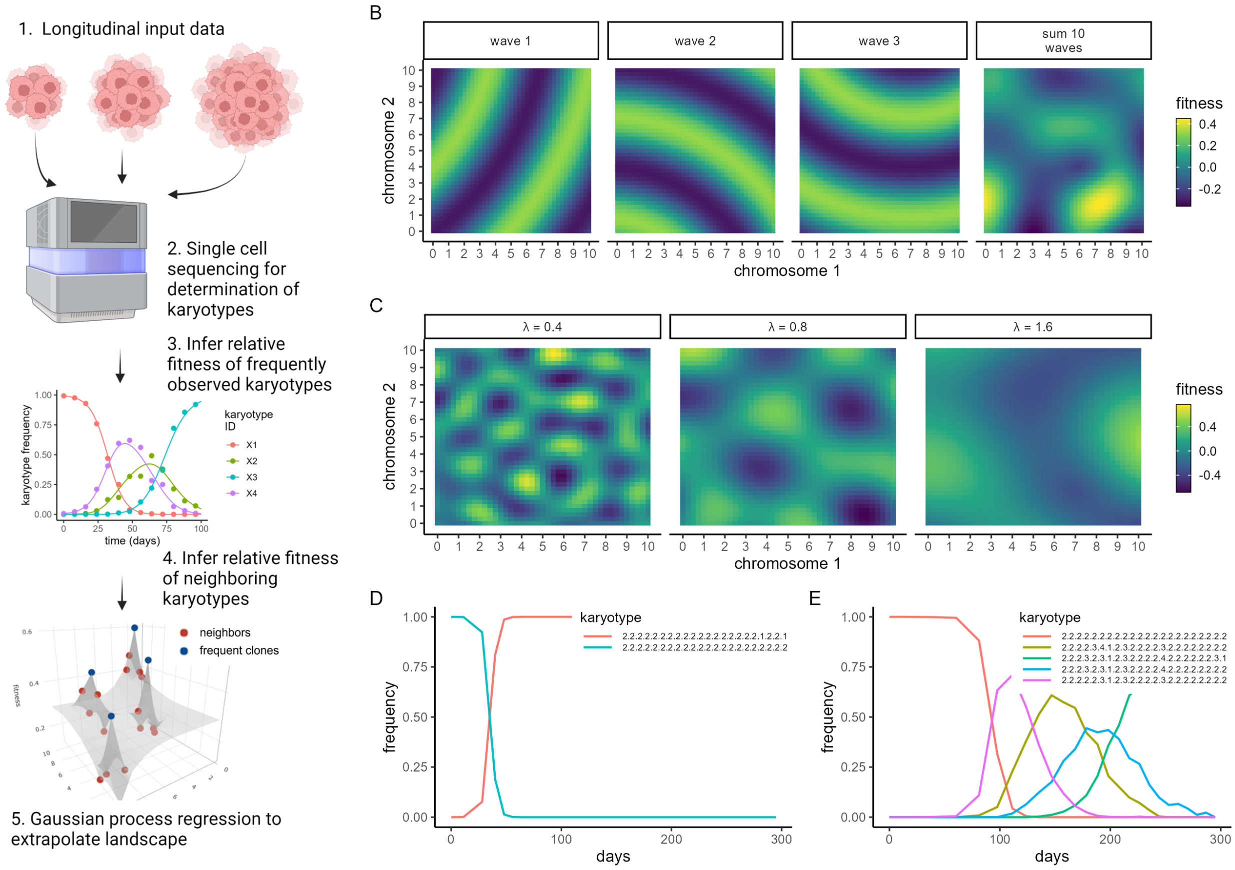

Figure 1: Figure 1: Clonal evolution on random fitness landscapes

estimate the fitness of subpopulations across multiple longitudinal timepoints. Up to this threshold however, increasing missegregation rate benefits ALFA-K by allowing a larger region of the fitness landscape to be charted.

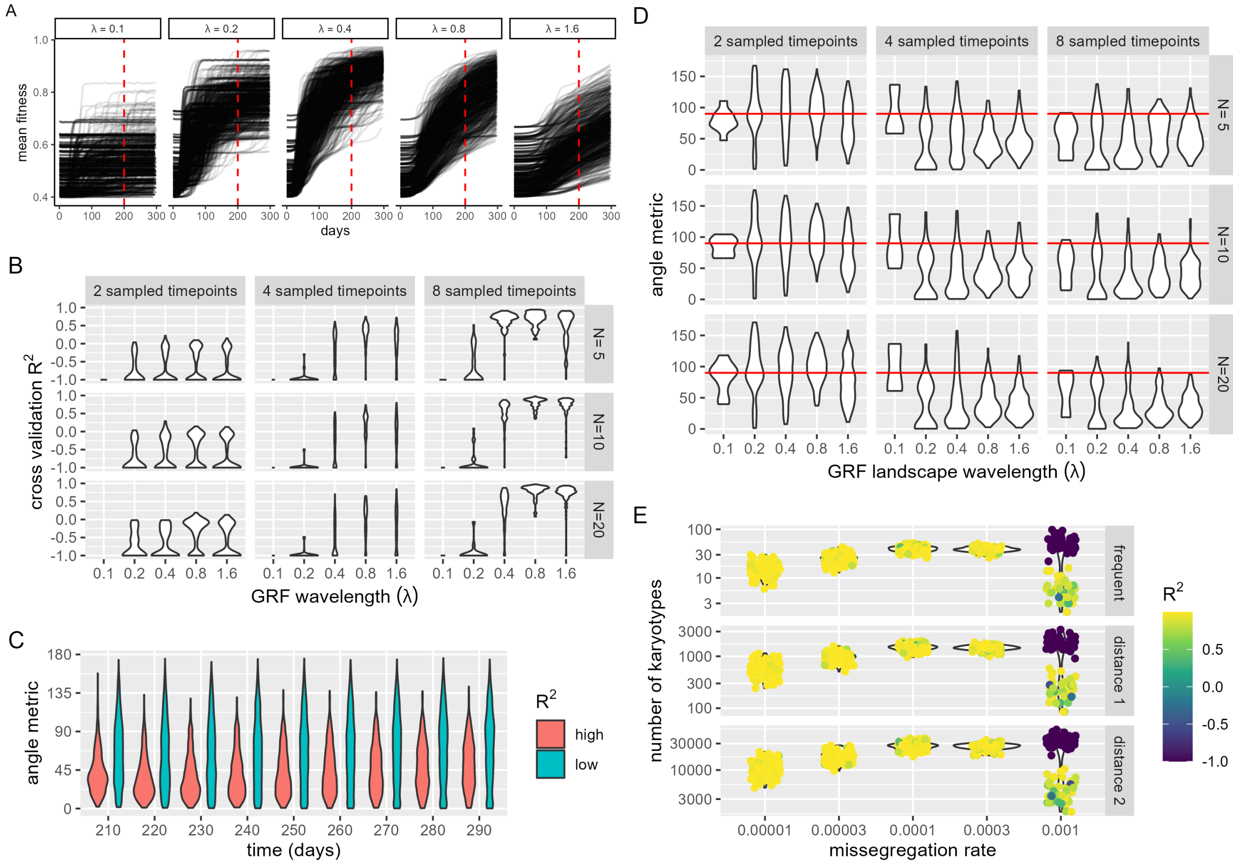

Figure 2: Figure 2: Validation of ALFA-K against synthetic data.

A) Mean fitness of ABM cell populations evolving on artificial fitness landscapes of varying complexity (as determined by λ). Data from the first 200 days of each simulation was demarcated by the red vertical lines was used to train ALFA-K. B) Cross validation results of ALFA-K for varying numbers of sampled timepoints and values of the hyperparameter N. c) Evolutionary prediction results of ALFA-K agggregated across all landscapes at various times in the validation period. Results are grouped by performance on the cross validation test and summarized by the angle metric. D) Evolutionary prediction results of ALFA-K for varying numbers of sampled timepoints and values of the hyperparameter N. Prediction results are summarized by the angle metric. E)

Notes:

TO DO Add details on how to reproduce ABM sweep outputs. Might be wise to have most of the code below tucked in scripts which are called from this RMD document? Some info about hwo to run sweeps is in the (silenced) comments of this markdown file (see esp. the misseg sweep for instructions on how to run a sweep). Tidy away pieces of code & data in the subdirectory that houses this RMD document