

# ALFA-K results

## Clonal evolution on random fitness landscapes

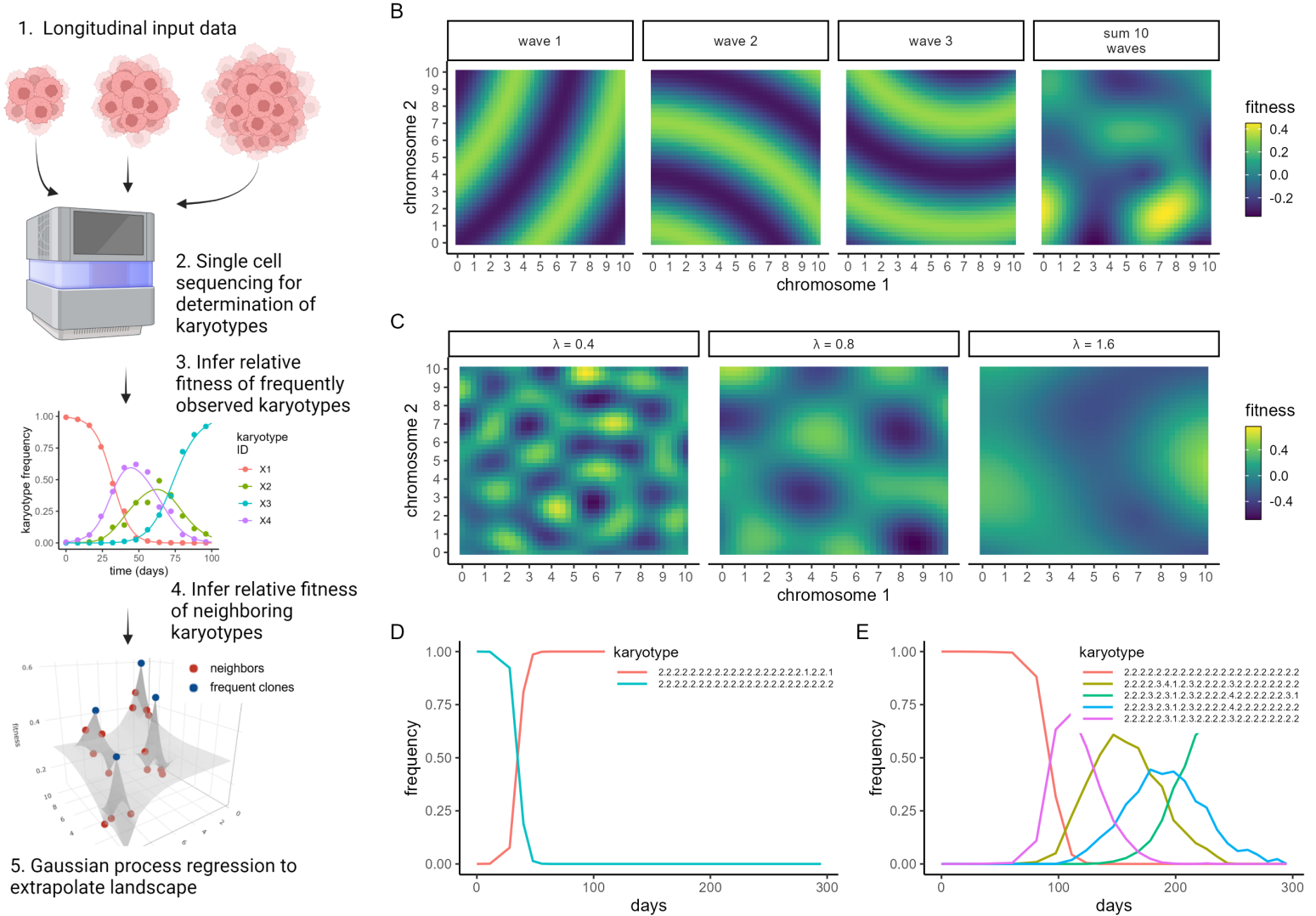


Figure 1: Clonal evolution on random fitness landscapes. A) Schematic flowchart of steps in ALFA-K pipeline. B) GRF are generated by summing multiple spherical waves. The interference patterns generated by the waves result in complex unpredictable landscapes. C) Increasing the wavelength ( $\lambda$ ) 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.

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 ( $\lambda$ ) 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  $\lambda$  (Fig. 1D) tended to experience punctuated evolution, whereas populations evolving on landscapes with high  $\lambda$  exhibited gradual increases in fitness over time (Fig. 1E).

**Validating ALFA-K against output of ABM simulations** 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 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.

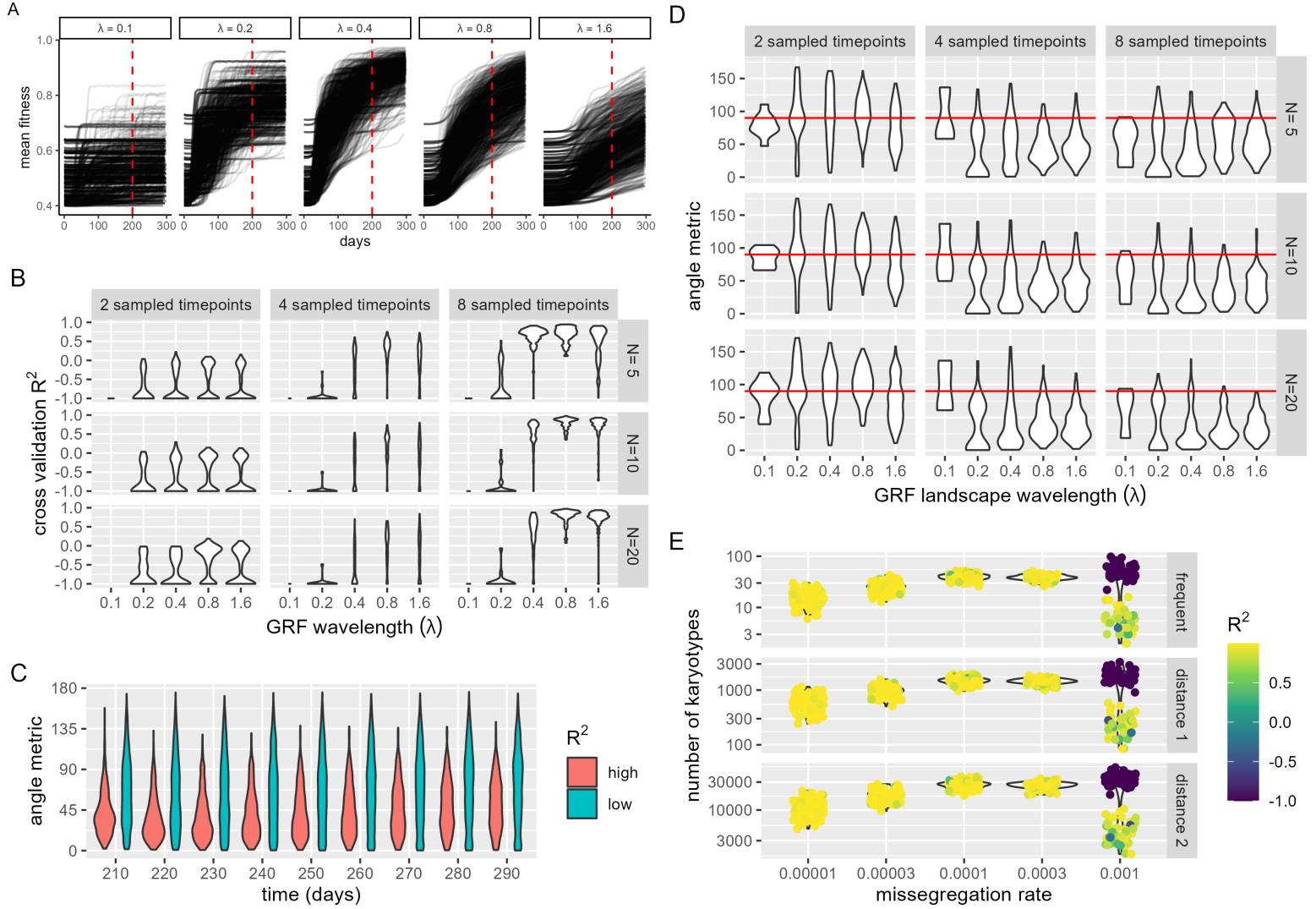


Figure 2: Validation of ALFA-K against synthetic data. A) Mean fitness of ABM cell populations evolving on artificial fitness landscapes of varying complexity (as determined by  $\lambda$ ). Data from the first 200 days of each simulation was demarcated by the red vertical lines. B) Cross validation results of ALFA-K for varying numbers of sampled timepoints and values of the hyperparameter  $N$ . C) Evolutionary prediction results of ALFA-K aggregated 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)

**Fitting ALFA-K model to passaging data** ALFA-K was fit to all lineages of data from passaging experiments by Salehi et al. A lineage was defined as two or more successive samples from the same cell line. Thus e.g. for SA532 there are 7 2-sample lineages, 6 3-sample lineages etc. Fig. A shows the relationship between samples for all 5 cell lines. Points are colored according to the  $R^2$  metric score obtained after fitting ALFA-K to the longest lineage terminating at that point. We were not able to obtain good fits to SA532 or SA906 lineage B. Examining  $R^2$  metric scores for all lineages obtained from the data, we found in agreement with our ABM results

that longer lineages result in better fits. We next applied ALFA-K to lineages with descendants, simulated population evolution, then compared the simulated populations to the withheld samples from subsequent passages using the angle metric. We were able to obtain good predictions for the next passage, with the quality of the predictions improved when there was a larger number of samples used to train ALFA-K or when ALFA-K cross validation had a positive  $R^2$  score.

**Predicting novel karyotypes** We asked whether ALFA-K can help predict the emergence of novel karyotypes. Denote  $\Theta$  the set of karyotypes with fitness estimates from ALFA-K,  $\zeta$  the subset of  $\Theta$  observed in a given longitudinal sample. Then we would like to predict  $\Psi$  (the subset of  $\Theta$  that will be present in a future sample) (Fig. 4A). In particular we wish to predict  $\zeta \cap \Psi$ , which are the *novel* karyotypes (Fig. 4B). The probability of any novel karyotype actually emerging presumably depends both on its fitness and its number of neighbours in the preceding generation. Therefore for each member of  $\zeta$  we computed the fraction of karyotypes in  $\zeta$  that were between 1-5 missegregations distant (Fig. 4C). These were used as variables which, together with the fitness estimate, were used to predict whether the karyotype would emerge. For prediction we used binomial logistic regression, then assessed whether each predictor variable was significantly correlated with the response variable. Across 45 tests, the fraction of  $\zeta$  that were distance-1 neighbours ( $d_1$ ) was the most significant predictor of novel karyotype emergence. We found that in 39/45 tests  $d_1$  had a significant ( $P < 0.01$ ) positive predictor of novel karyotype emergence (Fig. 4D), and in 30/45 tests  $d_1$  was the most significant predictor (Fig. 4E). The fitness estimate was also a strong predictor, significant in 30/45 tests (Fig. 4D) and most significant in 12/45 tests (Fig. 4E). These results indicate that ALFA-K fitness estimates can help predict emergence of new karyotypes.

**Karyotypic background determines fitness effects of CNAs** We sought to compare fitness landscapes derived from the various in vitro and PDX datasets. For individual karyotypes within these landscapes, we computed copy number alteration profiles (CNAPs) i.e., the fitness deltas ( $\Delta f$ ) resulting from each possible CNA (Fig.5A). To assess whether there were any broad changes in CNAPs across experimental conditions, we aggregated the  $\Delta f$  for all karyotypes and CNAs per condition. To assess whether there were any broad changes in CNAPs across experimental conditions, we aggregated the  $\Delta f$  for karyotypes per condition. We observed larger  $\Delta f$  in PDX vs. in vitro landscapes (Fig.5B). In PDX landscapes, cisplatin-treated lineages showed greater fitness deltas ( $\Delta f$ ) than those untreated (Fig.5C). This indicates the strongest selection pressure is in cisplatin-treated PDX landscapes, followed by untreated PDX, and is least within the in vitro landscapes. Overall, the results indicate that ALFA-K fitted landscapes could be used to compare the stringency of selection various contexts.

We then explored how a CNA's impact on fitness varies with the karyotypic context using our model landscapes. We compared pairs of karyotypes within the same landscape and analyzed the correlation between their CNAPs. The correlation strength was inversely proportional to the manhattan distance between the copy number profiles of the karyotype pair (Fig.5D). These results indicate that within a fixed biological context, similar karyotypes will experience similar fitness effects from a given CNA. However, if the karyotypes are not similar the fitness effects of a particular CNA are unlikely to correlate. We asked whether CNAPs from similar karyotypes derived from different fitness landscapes would correlate. We first used landscapes fitted to data from different PDX cell lines (Fig.5E). We saw a small but significant positive correlation, of comparable magnitude to karyotype pairs taken from identical landscapes. Unfortunately however, there were no karyotypes separated by a manhattan distance less than 5, so we could not assess whether the higher correlations observed for very similar karyotypes were also present across cell lines. We also checked whether CNAPs from different fitness landscapes fitted within the same cell line would correlate. For the SA609 cell line we had three available non-overlapping lineages, two untreated and one treated. CNAPs across the two untreated lineages exhibited much more evidence of correlation than CNAPS compared between either untreated lineage and the cisplatin treated lineage (Fig.5F). For SA535 we had available two lineages, one untreated and the other cisplatin treated. CNAPs exhibited significant correlation across these two lineages (Fig.5G). Finally we used the angle metric to compare the similarity of the evolutionary paths taken by karyotype populations from the previously analysed lineages (i.e. Figs 5F-G). For SA535, the low angle metric values indicate similar evolutionary paths between untreated and cisplatin treated lineages (Fig. 5H). For SA609, low angle metrics between the untreated lineages indicate similar evolutionary paths whilst high angle metrics around 90 degrees indicated a different evolutionary path for the untreated lineage (Fig.5I). Overall, our findings strongly suggest that a CNA's fitness effects rely on the parent cell's karyotype. Our results are less clear regarding how CNA fitness effects depend on biological context. More data from populations with similar karyotypes in varying biological contexts would be useful to address this question.

**Missegregation rate influences karyotype dominance** In most circumstances the fittest karyotype is expected to eventually become dominant within a population. We asked whether this status quo could be altered simply by changing the missegregation rate. We envisage two ways this could occur (Fig.6A): 1) karyotypes with more high-fitness neighbors should have an advantage when missegregation rate is high, since their numbers are bolstered by a continuous (albeit small) influx from their neighbors. 2) Low ploidy cells should have an advantage when missegregation rate is high, since their reduced total number of chromosomes mean a reduced per-cell risk of missegregation, so their populations should be more stable. Overall if the fittest karyotype has few neighbors or has high ploidy, it may not be dominant when the missegregation rate is high.

To identify possible missegregation dependent switches in karyotype distribution, we generated approximate transition matrices connecting all karyotypes in the charted region of each ALFA-K fitted landscape. We then determined the dominant eigenvectors of these transition matrices at varying missegregation rates to find the steady state karyotype distributions. For a subset of fitted landscapes we indeed found a missegregation rate dependent switch in karyotype distribution (Fig.6 B,C), where karyotypes that were fitter (Fig. 6 D,E) were dominant at lower missegregation rates but were displaced by less fit karyotypes at increasing missegregation rates. For SA906\_x57\_a the fitter group had significantly higher ploidy (Fig. 6F) but similar sized fitness peak to the less fit group (Fig.

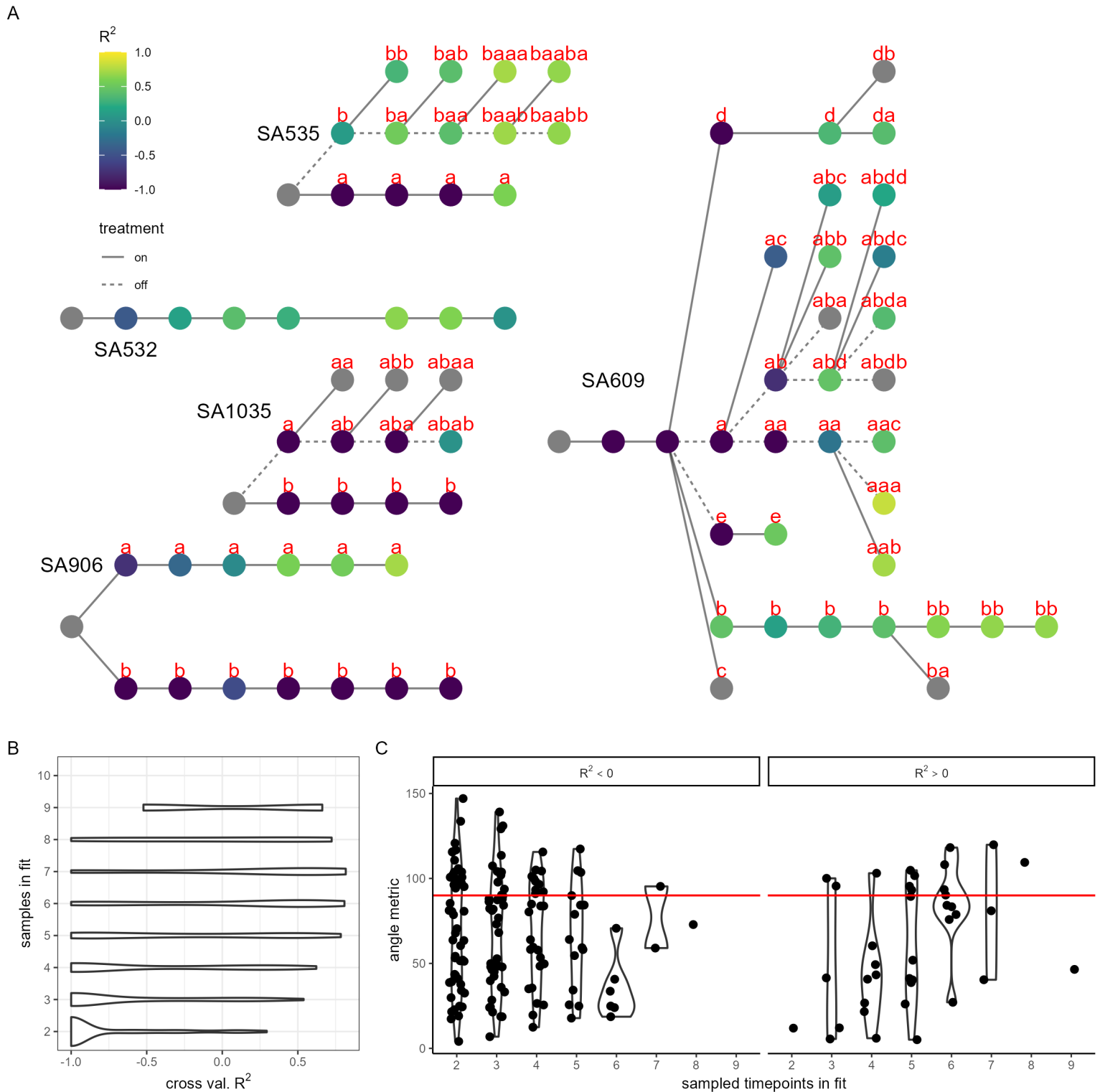


Figure 3. Fits to cell passing data A) Maps showing the relationship cell passages that were sequenced and used as inputs to ALFA-K. Each filled circle represents a sequenced sample. Lines connect parental passages to children, with line type indicating the presence or absence of cisplatin treatment in the intervening period. Circles are colored according to the  $R^2$  metric value for fitted landscapes including that sample and all its ancestors. B)  $R^2$  metric scores for all sublineages in the dataset, grouped according to the number of samples used in the fitting. C) Angle metric scores for all lineages with at least one descendent, grouped according to the number of samples used in the fitting and the  $R^2$  metric score. Red horizontal line indicates the expected value of the angle under the null hypothesis (i.e. that the fitted and true landscape are different).

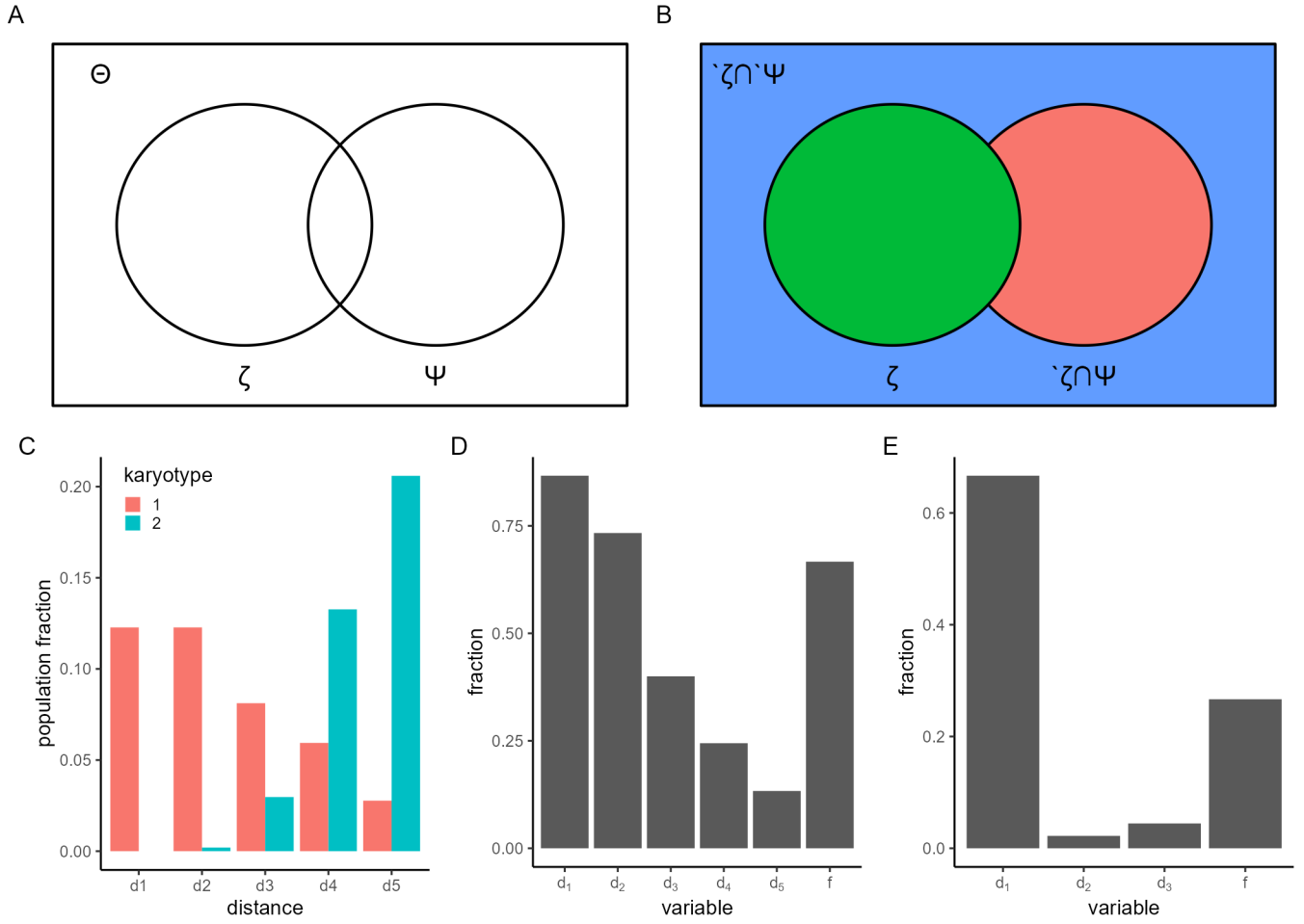
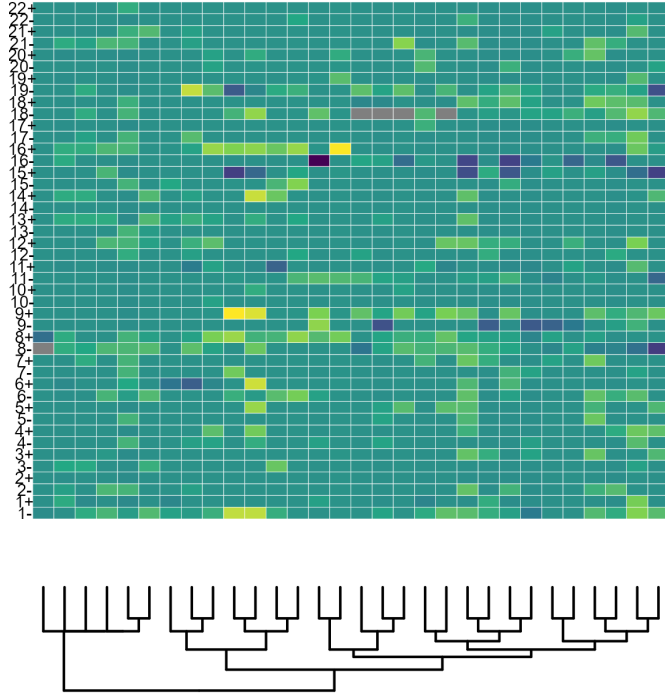
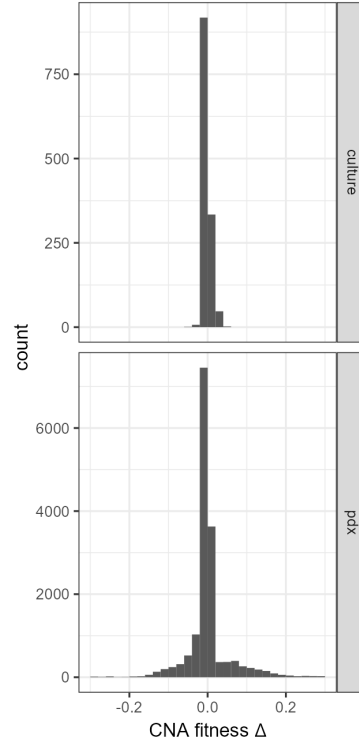


Figure 4 Predicting novel karyotypes. A) Venn diagram representing all karyotypes in the fitted landscape ( $\Theta$ ), the subset observed in the latest sample ( $\zeta$ ), and the subset that will be present in a future sample ( $\Psi$ ). B)  $\Theta$  is separated into 3 disjoint subsets. C) Each karyotype is assigned a feature vector ( $d_1$ – $d_5$ ).  $d_i$  is the fraction of karyotypes in  $\zeta$  that are  $i$  missegregations away from the current karyotype. D) Fraction of fitted lineages in which each variable was significant ( $P > 0.01$ ) E) Fraction of fitted lineages in which each variable was most significant.

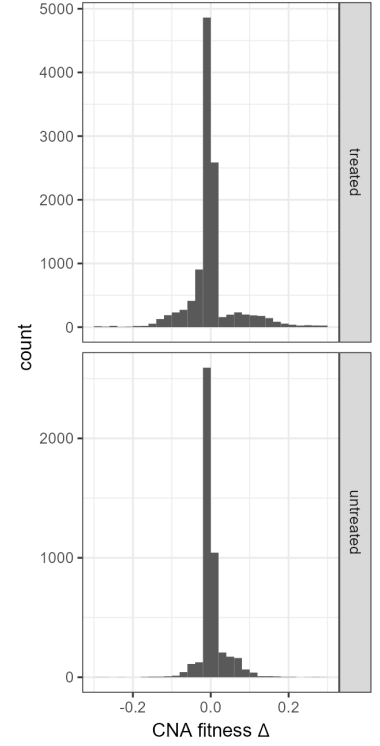
A



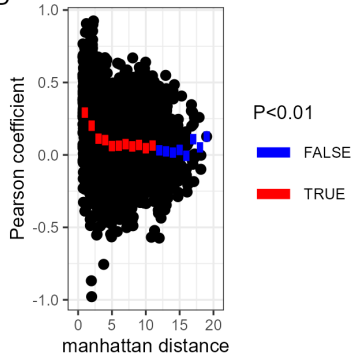
B



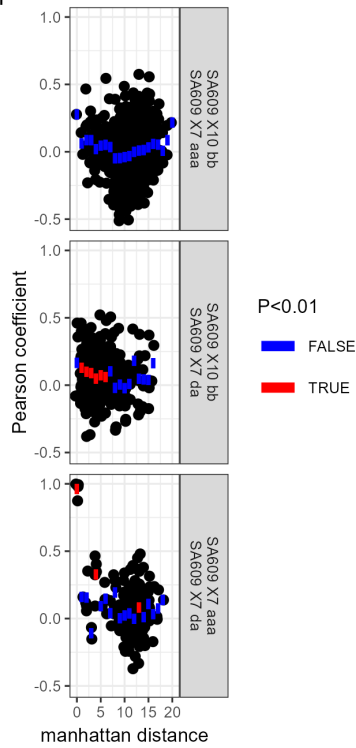
C



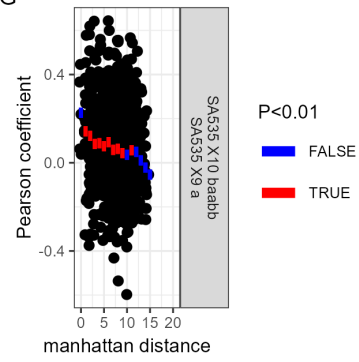
D



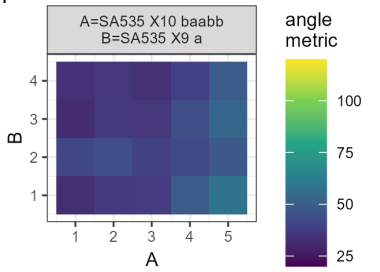
F



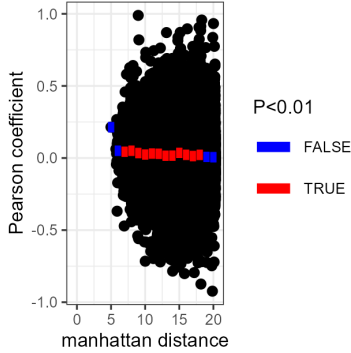
G



I



E



H

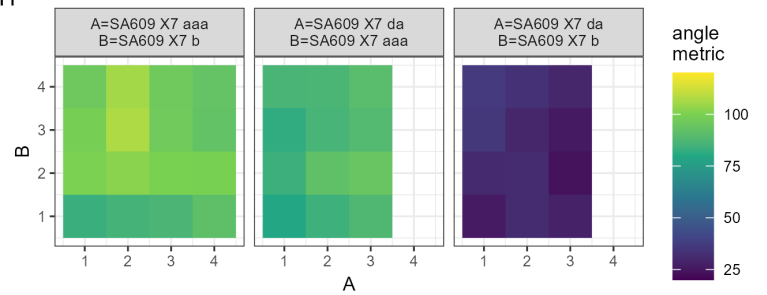


Figure 5. Comparison of fitted landscapes. A) CNAPs for karyotypes taken from a single fitness landscape. Each column in the image represents a CNAP. Karyotypes are clustered based on the Manhattan distance between their copy number profiles. B) ( $\Delta f$ ) comparison between in vitro (cultured) and PDX karyotypes. C) ( $\Delta f$ ) comparison between treated and untreated PDX karyotypes. D) Correlation between CNAPs within the same landscape. Each point represents the Pearson correlation coefficient between two CNAPs, presented according to the Manhattan distance between the two karyotypes in karyotype space. E) Correlation between CNAPs across cell lines. F) correlation of CNAPS across independent lineages of SA609. Untreated lineages are da and bb, cisplatin treated lineage is aaa. G) correlation of CNAPS across independent lineages of SA609. H) Angle metrics indicate similarity of evolutionary paths taken by SA609 lineages. I) Angle metrics indicate similarity of evolutionary paths taken by SA535 lineages.

6G), indicating that this was an example of a ploidy dependent frequency shift. For SA535\_X7\_bb both groups had similar ploidy (Fig. 6H) but the fitter group had less members (Fig. 6I), indicating that this was an example of peak size dependent frequency shift. Finally, we simulated population evolution in our ABM as a validation of our transition matrix based approximation (Fig. 6J). Overall, these results identify two ways which missegregation rate can interact with fitness to determine clonal dominance and found evidence of their occurrence in cultured cell lines.

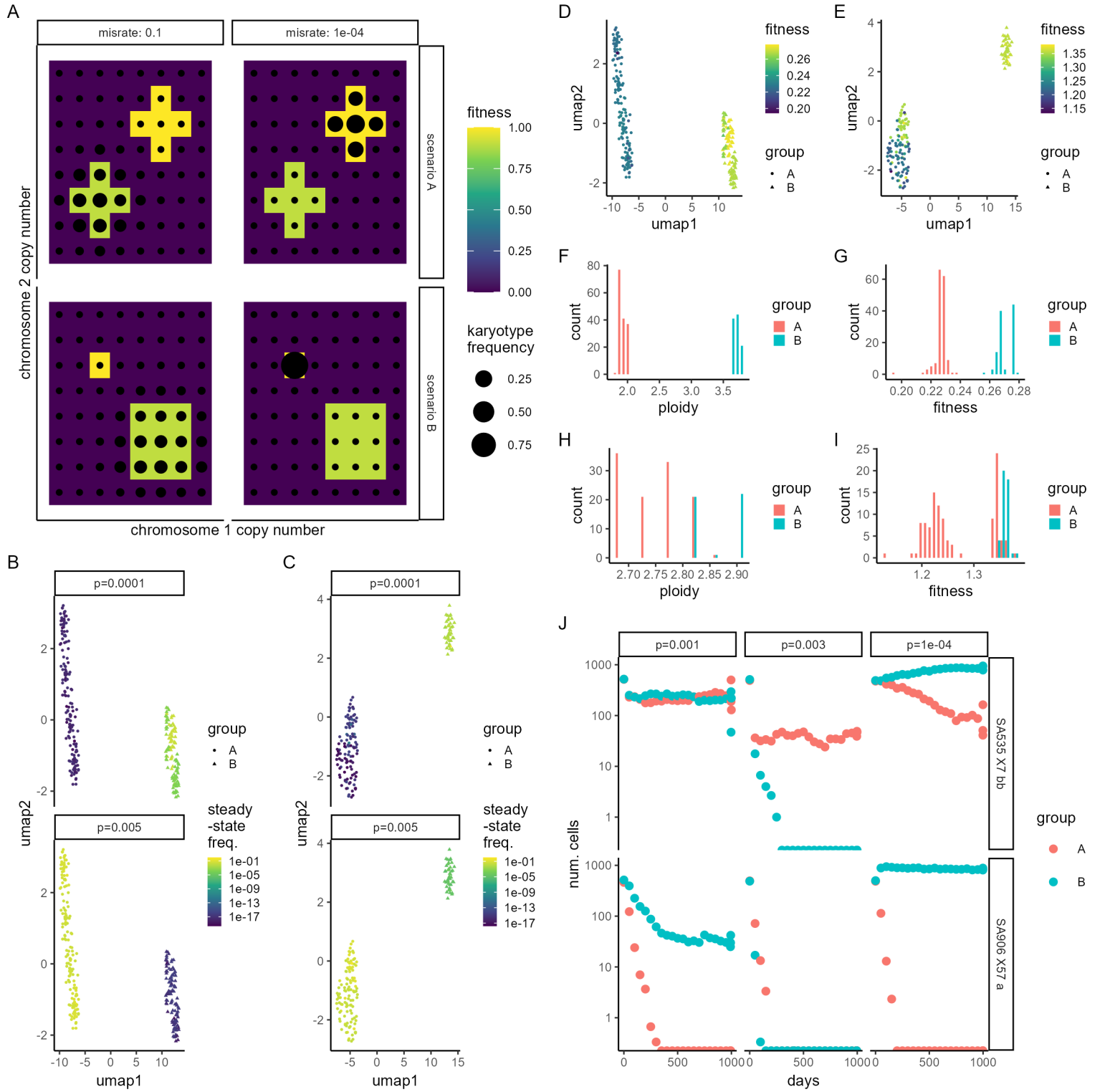


Figure 6 Influence of missegregation rate on karyotype selection. A) Simple hypothetical example fitness landscapes where karyotype frequency depends on missegregation rate. Karyotype frequency (point size) shifts from high to low ploidy (top row) or broader to narrower peaks (bottom row) as missegregation rate changes. (B,C) UMAPs showing the relative frequency of karyotypes representing at least 2% of the predicted steady-state population for SA906\_x57\_a (B) and SA535\_X7\_bb (C) at different missegregation rates. (D,E) ALFA-K estimated fitness of karyotypes shown in B-C for SA906\_x57\_a (D) and SA535\_X7\_bb (E). F) ploidy and G) fitness distributions for the top 2% most frequent clones in SA906\_x57\_a. H) ploidy and I) fitness distributions for the top 2% most frequent clones in SA535\_X7\_bb. J) ABM simulation of karyotype evolution on fitted landscapes for SA906\_x57\_a and SA535\_X7\_bb.