Predicting novel karyotypes

We asked whether ALFA-K can help predict the emergence of novel karyotypes. Denote Θ the set of karyotypes with fitness estimates from ALFA-K, ζ the subset of Θ observed in a given longitudinal sample. Then we would like to predict Ψ (the subset of Θ 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 ζ that were between 1-5 missegregations distant (Fig. 4C). These were used as variables which, together with the fitness estimate, were used to preduct 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 ζ that were distance-1 neighbours (d_1) was the most significant predictor of novel karyotype emergence. We found that in 39/45 tests d1 had a significant ($P<0.01$) positive predictor of novel karyotype emergence (Fig. 4D), and in 30/45 tests d1 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.

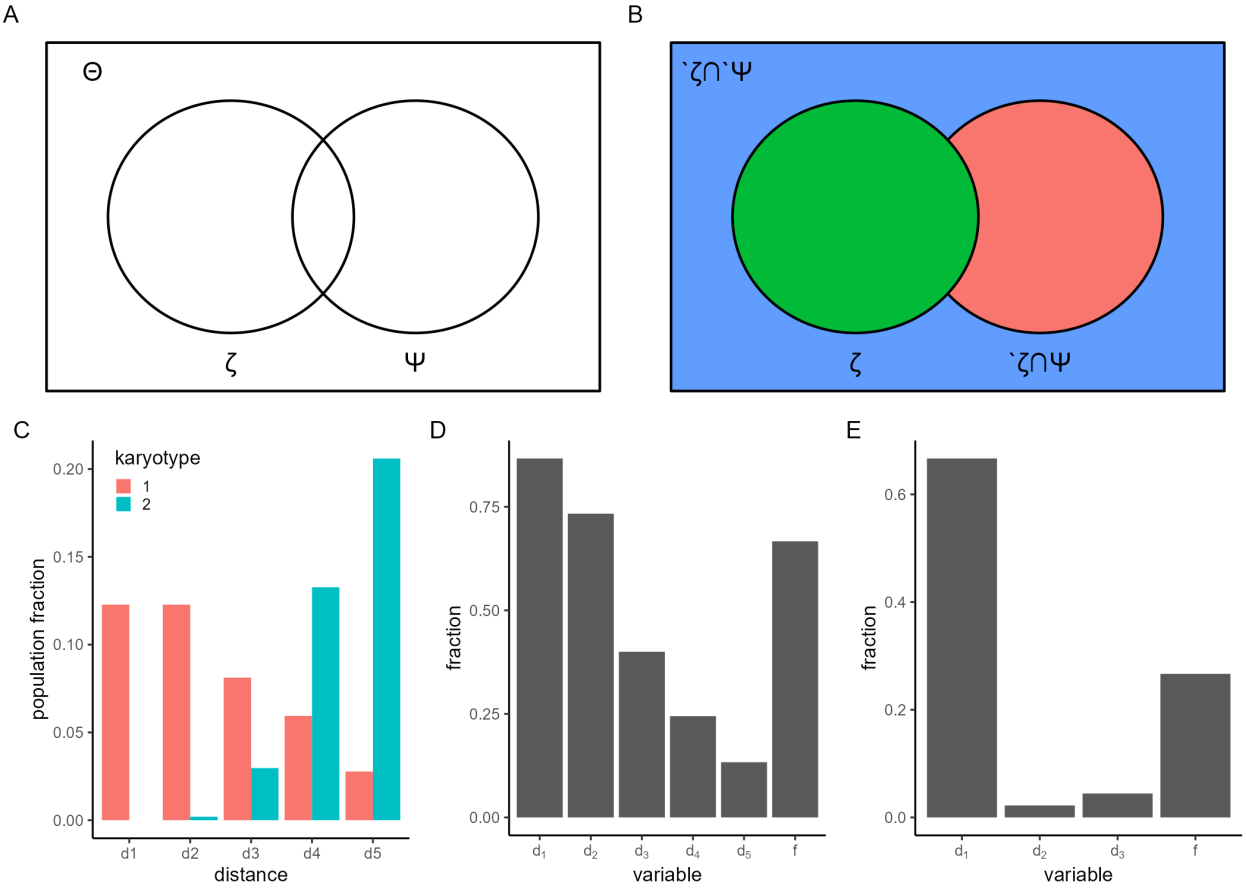

Figure 4 Predicting novel karyotypes. A) Venn diagram representing all karyotypes in the fitted landscape (Θ), the subset observed in the latest sample (ζ), and the subset that will be present in a future sample (Ψ). B) Θ is separated into 3 disjoint subsets. C) Each karyotype is assigned a feature vector (d_{1-5}). d_i is the fraction of karyotypes in ζ 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.