

Dividing the sequence data into five mutation classes allows the inference of robust fitness values as we show below. Reducing the number of mutation classes can lead to significant problems. For example, if we were to assume that there are only three different mutation classes then about 50% of the sequences would end up in the last mutation class for the example of SBW25, the fitness of which we set to 0 (they do not duplicate). But since this mutation class does contain sequences that duplicate, the duplication rate for mutation class 0 and 1 would be inferred to be lower than they actually are as their duplicative capacity is relatively lower to that of mutation class 2. The inferred lower fitness values for lower number of mutation classes supports this argument (Figure 1). The inferred fitness values do further increase for six mutation classes (Figure 1). However, if we increase the number of mutation classes to six the estimates become less robust because the number of sequences in the last mutation class is very low. The large standard deviations and the overestimation of the fitness for a simulation of six mutation classes demonstrate this point (Figure 2). To further test this hypothesis we varied the number of the last mutation class by adding and subtracting a single sequence and then showed the resulting inferred fitness. For six mutation classes the fitness estimates are much more variable than for five mutation classes (Figure 3), showing that using five mutation classes provides us with the best and most robust fitness estimate given our data.

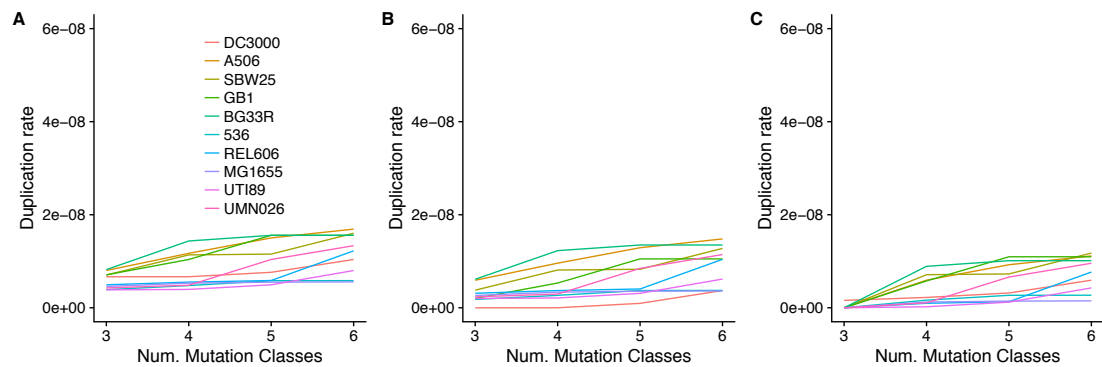
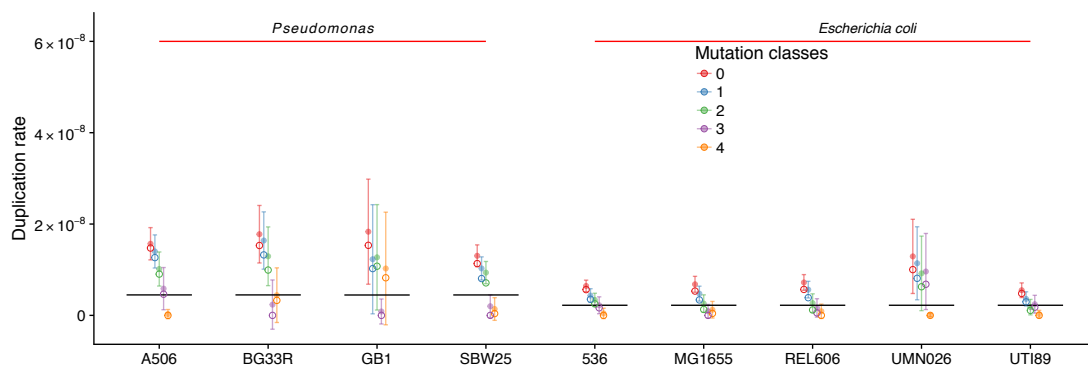


Figure 1. Inferred duplication rates increase with increasing numbers of mutation classes. Panels A-C show duplication rates for mutation classes 0 to 2, respectively. Different colored lines show mutation rates from different REPIN populations from different strains. The x-axis shows the number of mutation classes we used for our inference.



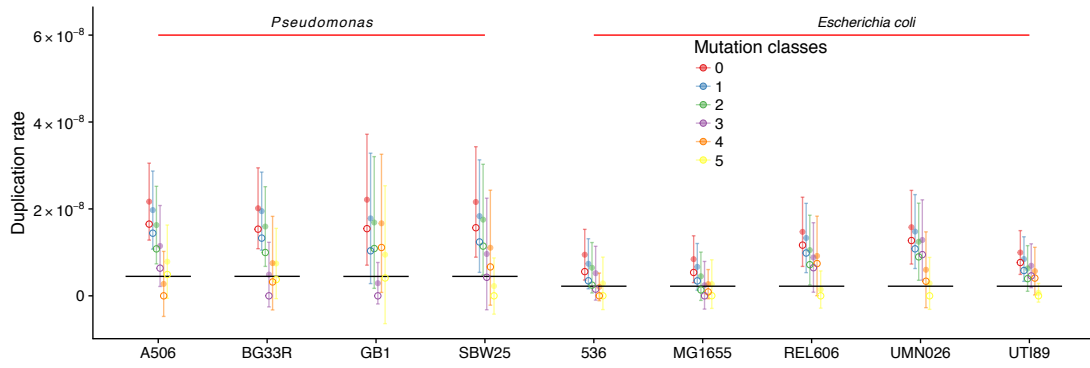


Figure 2. Fitness inferences from five (top) mutation classes are more robust than for six (bottom) mutation classes. The two plots show the inferred fitness values from the real data (empty circles) and for 20 time points of a simulation. The large standard deviations observed for six mutation classes as well as the overestimation of the fitness values supports our notion that fewer mutation classes enhance the robustness of our inferred fitness values.

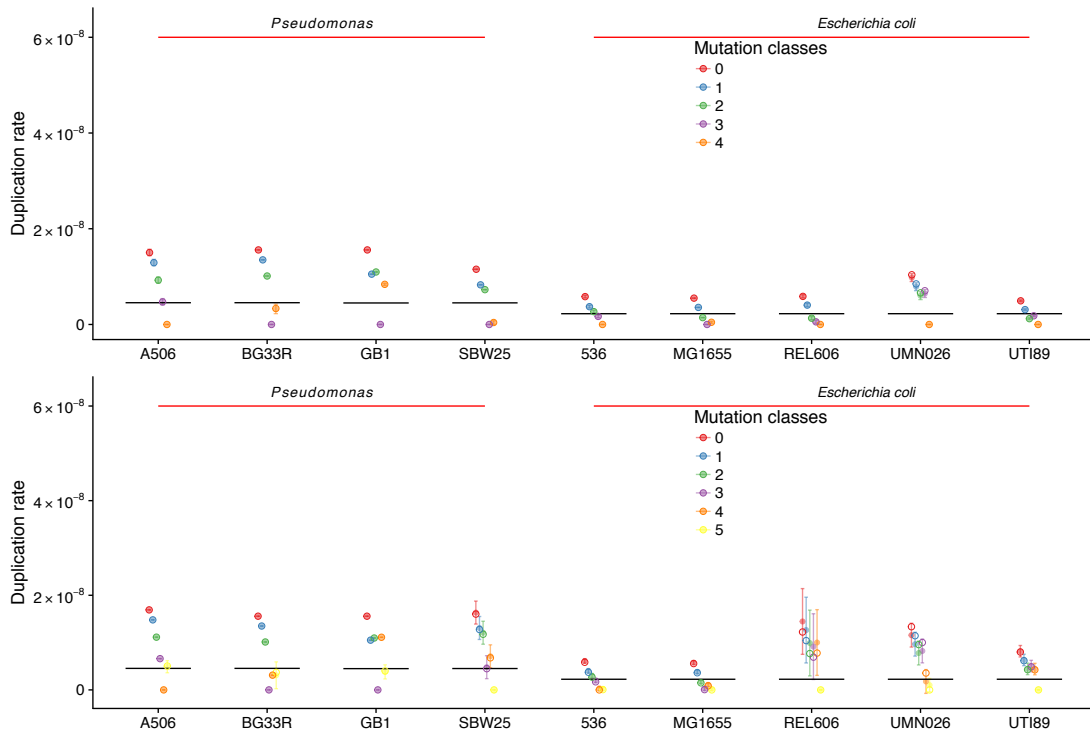


Figure 3. Five mutation classes (top) are more robust to changes than six mutation classes (bottom). In both figures we varied mutation class by one. The full circles show the mean inferred duplication rate. Empty circles show the inferred duplication rates from the data alone. If the mutation class with the largest hamming distance to the master sequence (four or five) is also the one with the lowest fitness the variation in this mutation class will affect only the inferred fitness from the other mutation classes because by definition the fitness of the lowest mutation class is set to one (e.g. REL606 bottom). If it is not the one with the lowest fitness then the variation will only affect the fitness inferred for mutation class five/six (e.g. BG33R bottom). Large standard deviations and fitness overestimation occur when the number of sequences in a mutation class is small. For example, if there are only two sequences in a mutation class, then adding a sequence will increase the abundance of this mutation class by fifty percent but will half the abundance if a sequence is subtracted.