Estimation of the probabilities of recurrent mutations in the same gene.

Let us consider a random 28 × 6075 matrix (the rows of the matrix represent independent sequenced genomes, the columns represent the number of functional yeast genes used in this study). If there is a mutation in the gene *j* from the genome *i*, a_{ij} = 1, otherwise a_{ij} = 0. Thus, the entries a_{ij} are independent variables taking the value 1 with the probability *p* and the value 0 with the probability 1-*p*. Note that the probability is the same for all *i* and *j*. Our goal is to estimate the probability $P_{\geq 5}$ that there is a gene where at least 5 independent mutations were found (the total number of mutations found in 28 independent genomes is 197, **Suppl. Table 6**). Let us first estimate *p*. Since the variables are independent, the expected value of the total number of mutations is

$$E\left(\sum_{i=1}^{28}\sum_{j=1}^{6075} a_{ij}\right) = \left(\sum_{i=1}^{28}\sum_{j=1}^{6075} E(a_{ij})\right) = 28 \times 6075 \times p = 197$$

whence

$$p = \frac{197}{28 \times 6075} = 1.16 \times 10^{-3}$$

Now let us calculate the probability of at least 5 mutations in a fixed gene *j*. The probability of exactly *k* mutations in this gene is

$$\binom{28}{k} p^{k} (1-p)^{28-k}.$$

Thus, the probability of at most 4 mutations is

$$p_{\leq 4} = \sum_{k=0}^{4} \binom{28}{k} p^{k} (1-p)^{28-k}$$

and the probability of at least 5 mutations is

$$p_{\geq 5} = 1 - p_{\leq 4} \approx 2 \times 10^{-10}.$$

Finally, since the events “at least five mutations in the *j*-th gene" are independent for different *j*, the probability $P_{\geq 5}$ is

$$P_{\geq 5} = 1 - (1 - p_{\geq 5})^{6075} = 1.2 \times 10^{-6}.$$

The same logic was applied to 3 and 4 multiple independent mutations.

3 mutations, P = 0.0302

4 mutations, P = 0.0002

5 mutations, $P = 1.2 \times 10^{-6}$

Thus, the probability of observing 3 independent mutations in the same gene is marginally significant, whereas 4 mutations are highly significant.

We also performed a simulation experiment as a control; we randomly populated the 28×6075 matrix with 197 mutations and estimated a weight

$$W_{random} = \underset{j=1}{\overset{6075}{MAX}}\left(\sum_{i=1}^{28} a_{ij}\right)$$

We repeated this procedure 1000 times and did not find any simulated matrix with the weight $W_{random} \geq 5$. The result suggested that the probability of observing 5 or more mutations in a gene is less than 0.001. This is consistent with our analytical estimates ($P_{\geq 5} \approx 1.2 \times 10^{-6}$).

Thus, the probability of finding 5 or more mutations in a gene is extremely small, suggesting that this event is likely to be biologically important.