**Evolutionary model**

We developed a Wright-Fisher model that describes the evolutionary dynamics, similar to our previous study (Avecilla et al., 2022). This is a discrete-time, non-overlapping generations model with a constant population size. Our model follows the change in frequency of four genotypes (Fig 2A): , the wildtype genotype; , a cell with a non-CNV beneficial mutation; , a genotype with two copies of GAP1 and two copies of the fluorescent reporter; and , a genotype with two copies of GAP1 but only a single copy of the fluorescent reporter. CNV and non-CNV mutations are formed at a rate of and and have a fitness effect of and , respectively.

Unlike and , which may increase in frequency due to both mutation and selection, we assume that the genotype is not generated after generation 0 (as experimental results suggest that the reporter is working properly). Hence, the genotype only increases in frequency due to selection, with as its fitness effect. We assume has an initial frequency .  
The model is fully described by the following derivations.

Initial genotype frequencies are:

The change in genotype frequencies due to mutation is expressed by

The change in genotype frequencies due to selection is expressed by

where for is the relative fitness of genotype and is the population mean fitness.

The change in genotype frequencies due to random genetic drift is given by

where is the number of cells with genotype *i*, and therefore the genotype frequencies in the next generation are

where is the population size, set to be constant at , the effective population size in the chemostat (Avecilla et al., 2022). Overall, we have five model parameters: .

The empirical observations show the frequency of cells with multiple GAP1 reporters over time. This is in our model. Therefore, we can use the evolutionary model to simulate dynamics in the shape of the empirical observations, which we will then use for inference with NPE.

**Competition model.** Similarly, we can simulate a competition assay between two genotypes. The model steps are as follows:

Initial genotype frequencies are:

The change in genotype frequencies due to mutation is expressed by

The change in genotype frequencies due to selection is expressed by

where for and is the relative fitness of genotype and is the population mean fitness.

Note that here we have different formation rates and selection coefficients for the two competing genotypes.

Genetic drift will be applied in the same manner as in the original model.

We’re interested in the competition’s outcome, i.e., .

**Collective posterior distribution**

The above stochastic evolutionary model can be represented by a likelihood function with *X* as a vector of genotype frequencies and as a vector of model parameters. In our case, *X* is actually a summary of genotype frequencies, as X(t) gives the frequency of cells with two or more GAP1 copies with reporters: .

Simulation-based Bayesian inference methods, such as ABC, produce a posterior distribution over model parameters by comparing model simulations to the empirical data when the likelihood is intractable. However, when a collection of observations is assumed to arrive from the same distribution, such as in repeated biological experiments, we are provided with more information that could be used for collective conclusions.  
Dietrich (Dietrich, 2010) showed an axiomatic approach to finding a group belief in a Bayesian setting, that is, a collective belief based on all observations, relying on probability rules to form the probability of a hypothesis H given a collection of beliefs - .

Given an evolutionary model, we could use a similar axiomatic approach to form – a collective posterior distribution for model parameters θ, conditioned on all observations .

Let be observations from the model such that . Assuming we have already estimated using NPE, we wish to estimate .

Using Bayes’ theorem,

Assuming are independent after conditioning on , i.e., , we have

Using Bayes’ theorem again:

Next, recalling the law of total probability,

Using Bayes’ theorem once more,

As is positive and independent of , we can take it out of the integral and then cancel it from both sides giving

Using Bayes’ theorem in eq. 4,

Using commutativity of multiplication,

where , the prior distribution, is known to us (we chose it), and , the indivdual posteriors, are estimated by NPE. Therefore, only the single and collective marginal likelihoods (i.e., ‘evidence’) and are unknown, yet both are positive and independent of the parameters vector . We can therefore define and write

Thus, plugging C in eq. 3 (and changing the symbol for the integration variable) will result in the collective posterior,

However, as could be infinitesimal, a single idiosyncratic observation could potentially reject a parameter value with a high probability for the other observations. We want the collective posterior to be robust to such unrepresentative or idiosyncratic observations. Therefore, we define so that individual posteriors give at least a value of to all parameter values. Thus, the collective posterior is

For a good choice of , the mode of the collective posterior distribution should reflect a value with high posterior density for multiple observations, rather than a value that no observation completely rejects (which can occur when the correction is not applied).

**Genetic diversity**

We want to estimate the CNV diversity of the population, and specifically, the probability to draw two consecutive cells with identical CNVs from a chemostat, given the collective MAP estimations for the four different genetic backgrounds ().

We assume that each newly formed CNV is different in its specific genotype from all CNV formed before it. Thus, *Q(t)* is the number of new CNV cells at generation *t*, that is, the number of CNV cells in generation *t* that have a non-CNV parent cell in generation *t-1*. *P(t)* is the number of cells in the last generation *T* that are descendants of a new CNV cell from generation *t* (i.e., that had a non-CNV parent in generation *t-1*). To compute Q(t) and P(t) given the results of an evolutionary simulation, , , , and :

At each generation :

1. Formation of new CNVs:
2. Computation of mean fitness at generation *t*:
3. For :
   * Application of natural selection:

Next, we define p(t) as the probability that two randomly chosen cells in the last generation are both descendants of CNV cells that were formed in generation *t* (i.e., have a non-CNV parent in generation *t-1*). We also define *q(t)* as the probability that two randomly chosen cells in the last generation that are both descendants of CNV cells formed in generation *t* are actually descended from the same cell. To compute p(t) and q(t):

For each generation :

Then, we compute S, the probability to draw two cells from the last generation that have the same CNV:

Note that the denominator has the square of the total number of CNVs generated over the entire simulation or experiment.

We compute the CNV diversity as follows:

In generation , lineages of size have formed, resulting in a CNV cells in generation *T*. The Shannon diversity index, given by where *fi* is the frequency of genotype *I*, (Jost 2006), is therefore